# Scaling Laws in Linear Regression: Compute, Parameters, and Data

**Licong Lin**
UC Berkeley
liconglin@berkeley.edu

**Jingfeng Wu**
UC Berkeley
uuujf@berkeley.edu

**Sham M. Kakade**
Harvard University
sham@seas.harvard.edu

**Peter L. Bartlett**
UC Berkeley and Google DeepMind
peter@berkeley.edu

**Jason D. Lee**
Princeton University
jasonlee@princeton.edu

## Abstract

Empirically, large-scale deep learning models often satisfy a neural scaling law: the test error of the trained model improves polynomially as the model size and data size grow. However, conventional wisdom suggests the test error consists of approximation, bias, and variance errors, where the variance error increases with model size. This disagrees with the general form of neural scaling laws, which predict that increasing model size monotonically improves performance.

We study the theory of scaling laws in an infinite dimensional linear regression setup. Specifically, we consider a model with $M$ parameters as a linear function of sketched covariates. The model is trained by one-pass stochastic gradient descent (SGD) using $N$ data. Assuming the optimal parameter satisfies a Gaussian prior and the data covariance matrix has a power-law spectrum of degree $a > 1$, we show that the reducible part of the test error is $\Theta(M^{-(a-1)} + N^{-(a-1)/a})$. The variance error, which increases with $M$, is dominated by the other errors due to the implicit regularization of SGD, thus disappearing from the bound. Our theory is consistent with the empirical neural scaling laws and verified by numerical simulation.

## 1 Introduction

Deep learning models, particularly those on a large scale, are pivotal in advancing the state-of-the-art across various fields. Recent empirical studies have shed light on the so-called *neural scaling laws* [see 26, 21, for example], which suggest that the generalization performance of these models improves polynomially as both model size, denoted by $M$, and data size, denoted by $N$, increase. The neural scaling law quantitatively describes the population risk as:

$$\mathcal{R}(M, N) \approx \mathcal{R}^* + \frac{c_1}{M^{a_1}} + \frac{c_2}{N^{a_2}}, \tag{1}$$

where $\mathcal{R}^*$ is a positive irreducible risk and $c_1$, $c_2$, $a_1$, $a_2$ are positive constants independent of $M$ and $N$. For instance, by fitting the above formula with empirical measurements in standard large-scale language benchmarks, Hoffmann et al. [21] estimated $a_1 \approx 0.34$ and $a_2 \approx 0.28$, while Besiroglu et al. [7] estimated that $a_1 \approx 0.35$ and $a_2 \approx 0.37$. Though the exact exponents depend on the tasks, neural scaling laws in (1) are observed consistently in practice and are used as principled guidance to build state-of-the-art models, especially under a compute budget [21].

From the perspective of statistical learning theory, (1) is rather intriguing. Standard statistical learning bounds [see 30, 41, for example] often decompose the population risk into the sum of irreducible

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

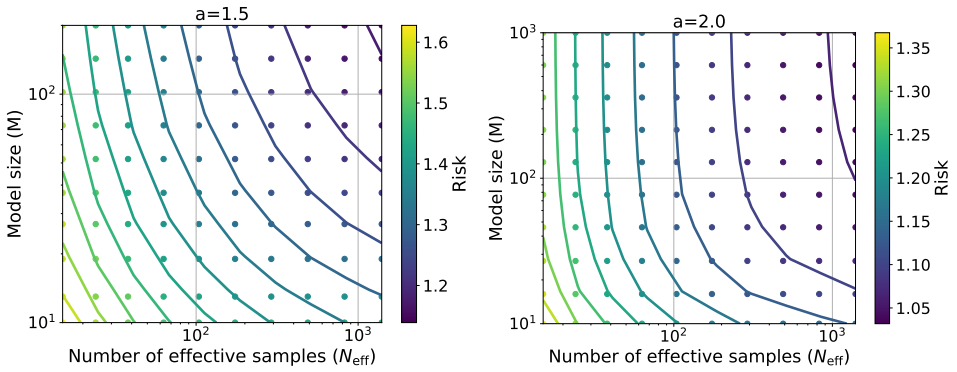



Figure 1: The expected risk (Risk) of the last iterate of (SGD) versus the effective sample size $N_{\texttt{eff}}$ and the model size $M$ for different power-law degrees $a$. The expected risk is computed by averaging over 1000 independent samples of $(\mathbf{w}^*, \mathbf{S})$. We fit the expected risk using the formula Risk $\sim \sigma^2 + c_1/M^{a_1} + c_2/N^{a_2}$ via minimizing the Huber loss as in [21]. Parameters: $\sigma = 1, \gamma = 0.1$. Left: For $a = 1.5$, $d = 20000$, the fitted exponents are $(a_1, a_2) = (0.54, 0.34) \approx (0.5, 0.33)$. Right: For $a = 2$, $d = 2000$, the fitted exponents are $(a_1, a_2) = (1.07, 0.49) \approx (1.0, 0.5)$. Note that the values of $(a_1, a_2)$ are close to our theoretical predictions $(a-1, 1-1/a)$ in both cases, verifying the sharpness of our risk bounds. More details can be found in Sections 4 and 5.

error, approximation error, bias error, and variance error (some theory replaces bias and variance errors by optimization and generalization errors, respectively) as in the form of

$$\mathcal{R}(M, N) = \mathcal{R}^* + \underbrace{\mathcal{O}\left(\frac{1}{M^{a_1}}\right)}_{\text{approximation}} + \underbrace{\mathcal{O}\left(\frac{1}{N^{a_2}}\right)}_{\text{bias}} + \underbrace{\mathcal{O}\left(\frac{c(M)}{N^{a_3}}\right)}_{\text{variance}}, \tag{2}$$

where $a_1, a_2, a_3$ are positive constants and $c(M)$ is a measure of *model complexity* that typically increases with the model size $M$. In (2), the approximation error is induced by the mismatch of the best-in-class predictor and the best possible predictor, hence decreasing with the model size $M$. The bias error is induced by the mismatch of the expected algorithm output and the best-in-class predictor, hence decreasing with the data size $N$. The variance error measures the uncertainty of the algorithm output, which decreases with the data size $N$ but increases with the model size $M$ (since the model complexity $c(M)$ increases).

**A mystery.** The empirical neural scaling law (1) is incompatible with the typical statistical learning theory bound (2). While the two error terms in the neural scaling law (1) can be explained by the approximation and bias errors in the theoretical bound (2) respectively, it is not clear why *the variance error is unobservable when fitting the neural scaling law empirically*. This difference must be reconciled, otherwise, the statistical learning theory and the empirical scaling law make conflict predictions: as the model size $M$ increases, the theoretical bound (2) predicts an increase of variance error that eventually causes an increase of the population risk, but the neural scaling law (1) predicts a decrease of the population risk. In other words, it remains unclear when to follow the prediction of the empirical scaling law (1) and when to follow that of the statistical learning bound (2).

Certain prior works provided risk upper bounds that do not grow with model size [see for example 36, 12]. Still, their results are insufficient for studying scaling law as those bounds require a large model size such that the approximation error is ignorable. Moreover, they do not provide instance-wise matching lower bounds to verify the tightness of the upper bounds. See a detailed discussion in Section 2.

**Our explanation.** We investigate this issue in an infinite dimensional linear regression setup. We only assume access to $M$-dimensional sketched covariates given by a fixed Gaussian sketch and their responses. We consider a linear predictor with $M$ trainable parameters, which is trained by one-pass *stochastic gradient descent* (SGD) with geometrically decaying stepsizes using $N$ sketched data. Assuming that the spectrum of the data covariance matrix satisfies a power-law of degree $a > 1$ and that the optimal model parameters satisfy a Gaussian prior, we derive matching upper and lower bounds on the population risk achieved by the SGD output (see Theorem 4.1). Specifically, we show

that

$$\mathcal{R}(M,N) = \mathcal{R}^* + \underbrace{\Theta\Big(\frac{1}{M^{a-1}}\Big) + \tilde{\Theta}\Big(\frac{1}{(N\gamma)^{(a-1)/a}}\Big)}_{\text{leading order given by the sum of Approx and Bias}}, \quad \mathsf{Var} = \underbrace{\tilde{\Theta}\Big(\frac{\min\{M,(N\gamma)^{1/a}\}}{N}\Big)}_{\text{higher order, thus unobservable}},$$

where $\gamma = \mathcal{O}(1)$ is the initial stepsize used in SGD and $\tilde{\Theta}(\cdot)$ hides $\log(N)$ factors. In our bound, the sum of the approximation and bias errors determines the order of the excess risk, while the variance error is of a strictly higher order and is therefore nearly unobservable when fitting $\mathcal{R}(M,N)$ as a function of $M$ and $N$ empirically. In addition, our analysis reveals that the small variance error is due to the implicit regularization effect of one-pass SGD [47]. Our theory suggests that the empirical neural scaling law (1) is a simplification of the statistical learning bound (2) in a special regime when strong regularization (either implicit or explicit) is employed.

Moreover, we generalize the above scaling law to (1) constant stepsize SGD with iterate average (see Theorem F.6), (2) cases where the optimal model parameter satisfies an anisotropic prior (see Theorem 4.2), and (3) where the spectrum of the data covariance matrix satisfies a logarithmic power law (see Theorem 4.3).

**Emprical evidence.** Based on our theoretical results, we conjecture that the clean neural scaling law (1) observed in practice is due to the disappearance of variance error caused by strong regularization. Two pieces of empirical evidence to support our understanding. First, large language models that follow the scaling law (1) are often *underfitted*, as the models are trained over a single pass or a few passes over the data [27, 31, 9, 39]. When models are underfitted, the variance error tends to be smaller. Second, when language models are trained with multiple passes (up to 7 passes), Muennighoff et al. [31] found that the clean scaling law in (1) no longer holds and they proposed a more sophisticated scaling law to explain their data. This can be explained by a relatively large variance error caused by multiple passes.

**Notation.** For two positive-valued functions $f(x)$ and $g(x)$, we write $f(x) \lesssim g(x)$ (and $f(x) = \mathcal{O}(g(x))$) or $f(x) \gtrsim g(x)$ (and $f(x) = \Omega(g(x))$) if $f(x) \le cg(x)$ or $f(x) \ge cg(x)$ holds for some absolute (if not otherwise specified) constant $c > 0$ respectively. We write $f(x) \eqsim g(x)$ (and $f(x) = \Theta(g(x))$) if $f(x) \lesssim g(x) \lesssim f(x)$. For two vectors $\mathbf{u}$ and $\mathbf{v}$ in a Hilbert space, we denote their inner product by $\langle \mathbf{u}, \mathbf{v} \rangle$ or $\mathbf{u}^\top \mathbf{v}$. For two matrices $\mathbf{A}$ and $\mathbf{B}$ of appropriate dimensions, we define their inner product by$\langle \mathbf{A}, \mathbf{B} \rangle := \mathrm{tr}(\mathbf{A}^\top \mathbf{B})$. We use $\|\cdot\|$ to denote the operator norm for matrices and $\ell_2$-norm for vectors. For a positive semi-definite (PSD) matrix $\mathbf{A}$ and a vector $\mathbf{v}$ of appropriate dimension, we write $\|\mathbf{v}\|_{\mathbf{A}}^2 := \mathbf{v}^\top \mathbf{A} \mathbf{v}$. For a symmetric matrix $\mathbf{A}$, we use $\mu_j(\mathbf{A})$ to refer to the $j$-th eigenvalue of $\mathbf{A}$ and $r(\mathbf{A})$ to refer to its rank. Finally, $\log(\cdot)$ refers to logarithm base 2.

## 2 Related work

**Empirical scaling laws.** In recent years, the scaling laws of deep neural networks in compute, sample size, and model size have been widely studied across different models and domains [20, 35, 26, 19, 21, 46, 31]. The early work by Kaplan et al. [26] first proposed the neural scaling laws of transformer-based models. They observed that the test loss exhibits a power-law decay in quantities including the amount of compute, sample size, and model size, and provided joint formulas in these quantities to predict the test loss. The proposed formulas were later generalized and refined in subsequent works [19, 21, 1, 10, 31]. Notably, Hoffmann et al. [21] proposed the Chinchilla law, that is, (1) with $a_1 \approx 0.34$ and $a_2 \approx 0.28$. The empirical observation guided them to allocate data and model size under a given compute budget. The Chinchilla law is further revised by Besiroglu et al. [7]. Motivated by the Chinchilla law, Muennighoff et al. [31] considered the effect of multiple passes over training data and empirically fitted a more sophisticated scaling law that takes account of the effect of data reusing.

**Theory of scaling laws.** Although neural scaling laws have been empirically observed over a broad spectrum of problems, there is a relatively limited literature on understanding these scaling laws from a theoretical perspective [37, 4, 28, 22, 42, 29, 23, 8, 2, 32, 17]. Among these works, [37] showed that the test loss scales as $N^{4/d}$ for regression on data with intrinsic dimension $d$. Hutter [22] studied a toy problem under which a non-trivial power of $N$ arises in the test loss. Jain et al. [23]

considered scaling laws in data selection. Bahri et al. [4] considered a linear teacher-student model under a power-law spectrum assumption on the covariates, and they showed that the test loss of the ordinary least square estimator decreases following a power law in sample size $N$ (resp. model size $M$) when the model size $M$ (resp. sample size $N$) is infinite. Bordelon et al. [8] considered a linear random feature model and analyzed the test loss of the solution found by (batch) gradient flow. They focused on the bottleneck regimes where two of the quantities $N, M, T$ (training steps) are infinite and showed that the risk has a power-law decay in the remaining quantity. The problem in Bahri et al. [4], Bordelon et al. [8] can be viewed as a sketched linear regression model similar to ours. It should be noted that both Bahri et al. [4] and Bordelon et al. [8] only derived the dependence of population risk on one of the data size, model size, or training steps in the asymptotic regime where the remaining quantities go to infinity, and their derivations are based on statistical physics heuristics. In comparison, we prove matching (ignoring constant factors) upper and lower risk bounds jointly depending on the finite model size $M$ and data size $N$.

**Implicit regularization of SGD.** One-pass SGD in linear regression has been extensively studied in both the classical finite-dimensional setting [34, 3, 14, 16, 25, 24, 18] and the modern high-dimensional setting [15, 6, 48, 47, 44, 45, 40]. In particular, Zou et al. [47] showed that SGD induces an implicit regularization effect that is comparable to, and in certain cases even more preferable than, the explicit regularization effect induced by ridge regression. This is one of the key motivations of our scaling law interpretation. From a technical perspective, we utilize the sharp finite-sample and dimension-free analysis of SGD developed by Zou et al. [48], Wu et al. [44, 45]. Different from them, we consider a sequence of linear regression models with an increasing number of trainable parameters given by data sketch. Our main technical innovation is to sharply control the effect of data sketch. Some of our intermediate results, for example, tight bounds on the spectrum of the sketched data covariance under the power law (see Lemma 6.2), might be of independent interest.

Prior works investigated linear regression with random features [36, 12], which can be viewed as a kind of sketched features via random coordinate selection. They mainly focused on the small approximation error regime, where the model size (or the number of features) is much larger than the data size. In comparison, we treat both model size and data size as free variables. Moreover, we provide matching upper and lower bounds while prior works mainly focused on upper bounds. These two innovations are crucial for studying scaling laws that predict test error as a function of both model size and data size. Finally, in the comparable regimes with small or zero approximation error, our excess risk bounds recover the bounds in prior works [36, 12, 33, 15, 13].

## 3 Setup

We use $\mathbf{x} \in \mathbb{H}$ to denote a feature vector, where $\mathbb{H}$ is a finite $d$-dimensional or countably infinite dimensional Hilbert space, and $y \in \mathbb{R}$ to denote its label. In linear regression, we measure the population risk of a parameter $\mathbf{w} \in \mathbb{H}$ by the mean squared error,

$$\mathcal{R}(\mathbf{w}) := \mathbb{E}\big(\langle \mathbf{x}, \mathbf{w}\rangle - y\big)^2, \quad \mathbf{w} \in \mathbb{H},$$

where the expectation is over $(\mathbf{x}, y) \sim P$ for some distribution $P$ on $\mathbb{H} \times \mathbb{R}$.

**Definition 1** (Data covariance and optimal parameter)**.** Let $\mathbf{H} := \mathbb{E}[\mathbf{x}\mathbf{x}^\top]$ be the data covariance. Assume that $\mathrm{tr}(\mathbf{H})$ and all entries of $\mathbf{H}$ are finite. Let $(\lambda_i)_{i\geq 0}$ be the eigenvalues of $\mathbf{H}$ sorted in non-increasing order. Let $\mathbf{w}^* \in \arg\min_{\mathbf{w}} \mathcal{R}(\mathbf{w})$ be the optimal model parameter[1]. Assume that $\|\mathbf{w}^*\|_{\mathbf{H}}^2 := (\mathbf{w}^*)^\top \mathbf{H} \mathbf{w}^*$ is finite.

We only assume access to $M$-dimensional sketched covariates and their responses, that is, $(\mathbf{S}\mathbf{x}, y)$, where $\mathbf{S} \in \mathbb{R}^M \times \mathbb{H}$ is a fixed *sketch* matrix. We focus on the Gaussian sketch matrix[2], that is, entries of $\mathbf{S}$ are independently sampled from $\mathcal{N}\big(0, 1/M\big)$. We then consider linear predictors with $M$ trainable parameters given by

$$f_{\mathbf{v}} : \mathbb{H} \to \mathbb{R}, \quad \mathbf{x} \mapsto \langle \mathbf{v}, \mathbf{S}\mathbf{x}\rangle,$$

where $\mathbf{v} \in \mathbb{R}^M$ are the trainable parameters. Varying $M$ should be viewed as a linear analog of varying the neural network model size. Our sketched linear regression setting is comparable to the teacher-student setting considered by Bahri et al. [4], Bordelon et al. [8].

[1] If $\arg\min \mathcal{R}(\cdot)$ is not unique, we choose $\mathbf{w}^*$ to be the minimizer with minimal $\mathbf{H}$-norm.

[2] Our results can be extended to other sketching methods [see 43, for example].

We consider the training of $f_{\mathbf{v}}$ via one-pass *stochastic gradient descent* (SGD), that is,

$$\begin{aligned}\mathbf{v}_t &:= \mathbf{v}_{t-1} - \gamma_t\big(f_{\mathbf{v}_{t-1}}(\mathbf{x}_t) - y_t\big)\nabla_{\mathbf{v}} f_{\mathbf{v}_{t-1}}(\mathbf{x}_t) \\ &:= \mathbf{v}_{t-1} - \gamma_t\big(\mathbf{x}_t^\top \mathbf{S}^\top \mathbf{v}_{t-1} - y_t\big)\mathbf{S}\mathbf{x}_t, \qquad t = 1, \dots, N,\end{aligned} \tag{SGD}$$

where $(\mathbf{x}_t, y_t)_{t=1}^N$ are independent samples from $P$ and $(\gamma_t)_{t=1}^N$ are the stepsizes. We consider a popular geometric decaying stepsize scheduler [18, 44],

$$\text{for } t = 1, \dots, N,\ \gamma_t := \gamma/2^\ell,\ \text{where } \ell = \lfloor t/(N/\log(N)) \rfloor. \tag{3}$$

Here, the initial stepsize $\gamma$ is a hyperparameter for the SGD algorithm. Without loss of generality, we assume the initial parameter is $\mathbf{v}_0 = 0$. The output of the SGD algorithm is the last iterate $\mathbf{v}_N$. Our proof techniques apply to other stepsize schedulers (e.g., polynomial decay) as well, but we focus on geometric decay as it is known to achieve near minimax-optimal excess risk for the last iterate of SGD [18].

Conditioning on a sketch matrix $\mathbf{S} \in \mathbb{R}^M \times \mathbb{H}$, each parameter $\mathbf{v} \in \mathbb{R}^M$ induces a sketched predictor through $\mathbf{x} \mapsto \langle \mathbf{S}^\top \mathbf{v}, \mathbf{x} \rangle$, and we denote its risk by

$$\mathcal{R}_M(\mathbf{v}) := \mathcal{R}(\mathbf{S}^\top \mathbf{v}) = \mathbb{E}\big(\langle \mathbf{S}\mathbf{x}, \mathbf{v} \rangle - y\big)^2, \quad \mathbf{v} \in \mathbb{R}^M.$$

By increasing $M$ and $N$, we have a sequence of datasets and trainable parameters of increasing sizes, respectively. This prepares us to study the scaling law (1) in the sketched linear regression problem, that is, to understand $\mathcal{R}_M(\mathbf{v}_N)$ as a function of both $M$ and $N$.

**Risk decomposition.** In a standard way, we decompose the risk achieved by $\mathbf{v}_N$, the last iterate of (SGD), to the sum of *irreducible risk*, *approximization error*, and *excess risk* as follows,

$$\mathcal{R}_M(\mathbf{v}_N) = \underbrace{\min \mathcal{R}(\cdot)}_{\text{Irreducible}} + \underbrace{\min \mathcal{R}_M(\cdot) - \min \mathcal{R}(\cdot)}_{\text{Approx}} + \underbrace{\mathcal{R}_M(\mathbf{v}_N) - \min \mathcal{R}_M(\cdot)}_{\text{Excess}}. \tag{4}$$

We emphasize that the irreducible risk is independent of $M$ and $N$ and thus can be viewed as a constant; the approximation error is determined by the sketch matrix $\mathbf{S}$, thus depends on $M$ but is independent of $N$; the excess risk depends on both $M$ and $N$ as it is determined by the algorithm.

## 4 Scaling laws

We first demonstrate a scaling-law behavior when the data spectrum satisfies a power law.

**Assumption 1** (Distributional conditions)**.** *Assume the following about the data distribution.*

A. ***Gaussian design.*** *Assume that* $\mathbf{x} \sim \mathcal{N}(0, \mathbf{H})$.

B. ***Well-specified model.*** *Assume that* $\mathbb{E}[y|\mathbf{x}] = \mathbf{x}^\top \mathbf{w}^*$. *Define* $\sigma^2 := \mathbb{E}(y - \mathbf{x}^\top \mathbf{w}^*)^2$.

C. ***Parameter prior.*** *Assume that* $\mathbf{w}^*$ *satisfies a prior such that* $\mathbb{E}(\mathbf{w}^*)^{\otimes 2} = \mathbf{I}$.

**Assumption 2** (Power-law spectrum)**.** *There exists* $a > 1$ *such that the eigenvalues of* $\mathbf{H}$ *satisfy* $\lambda_i \eqsim i^{-a}$, $i > 0$.

**Theorem 4.1** (Scaling law)**.** *Suppose that Assumptions 1 and 2 hold. Consider an* $M$*-dimensional sketched predictor trained by* (SGD) *with* $N$ *samples. Let* $N_{\texttt{eff}} := N/\log(N)$ *and recall the risk decomposition in* (4)*. Then there exists some* $a$*-dependent constant* $c > 0$ *such that when the initial stepsize* $\gamma \le c$*, with probability at least* $1 - e^{-\Omega(M)}$ *over the randomness of the sketch matrix* $\mathbf{S}$*, we have*

1. $\mathsf{Irreducible} := \mathcal{R}(\mathbf{w}^*) = \sigma^2$.
2. $\mathbb{E}_{\mathbf{w}^*}\mathsf{Approx} \eqsim M^{1-a}$.
3. *Suppose in addition* $\sigma^2 \gtrsim 1$. *The expected excess risk (*$\mathsf{Excess}$*) can be decomposed into a bias error (*$\mathsf{Bias}$*) and a variance error (*$\mathsf{Var}$*), namely,*

$$\mathbb{E}\mathsf{Excess} \eqsim \mathsf{Bias} + \sigma^2 \mathsf{Var},$$

*where the expectation is over the randomness of* $\mathbf{w}^*$ *and* $(\mathbf{x}_i, y_i)_{i=1}^N$. *Moreover,* Bias *and* Var *satisfy*

$$\begin{aligned}
&\mathsf{Bias} \lesssim \max\left\{M^{1-a},\ (N_{\mathtt{eff}}\gamma)^{1/a-1}\right\},\\
&\mathsf{Bias} \gtrsim (N_{\mathtt{eff}}\gamma)^{1/a-1} \text{ when } (N_{\mathtt{eff}}\gamma)^{1/a} \le M/c \quad \text{for some constant } c>0,\\
&\mathsf{Var} \eqsim \min\left\{M,\ (N_{\mathtt{eff}}\gamma)^{1/a}\right\}/N_{\mathtt{eff}}.
\end{aligned}$$

*In all results, the hidden constants only depend on the power-law degree* $a$. *As a direct consequence, when* $\sigma^2 \eqsim 1$, *it holds with probability at least* $1-e^{-\Omega(M)}$ *over the randomness of the sketch matrix* $\mathbf{S}$ *that*

$$\mathbb{E}\mathcal{R}_M(\mathbf{v}_N) = \sigma^2 + \Theta\bigg(\frac{1}{M^{a-1}}\bigg) + \Theta\bigg(\frac{1}{(N_{\mathtt{eff}}\gamma)^{(a-1)/a}}\bigg),$$

*where the expectation is over the randomness of* $\mathbf{w}^*$ *and* $(\mathbf{x}_i, y_i)_{i=1}^N$.

Theorem 4.1 shows a sharp (up to constant factors) scaling law risk bound under an isotroptic prior assumption and the power-law spectrum assumption. We emphasize that the scaling law bound in Theorem 4.1 holds for every $M, N \ge 1$. We also remark that the sum of approximization and bias errors dominates $\mathbb{E}\mathcal{R}_M(\mathbf{v}_N) - \sigma^2$, whereas the variance error is of strict higher order in terms of both $M$ and $N$, and is thus disappeared in the population risk bound.

**Optimal stepsize.** Based on the tight scaling law in Theorem 4.1, we can calculate the optimal stepsize that minimizes the risk. Specifically, the optimal stepsize is $\gamma \eqsim 1$ when $N_{\mathtt{eff}} \lesssim M^a$ and can be anything such that $M^a/N_{\mathtt{eff}} \lesssim \gamma \lesssim 1$ when $N_{\mathtt{eff}} \gtrsim M^a$. In both cases, choosing $\gamma \eqsim 1$ is optimal. When the sample size is large such that $N_{\mathtt{eff}} \gtrsim M^a$, the optimal stepsize is relatively robust and can be chosen from a range.

**Allocation of data and model sizes.** Following Hoffmann et al. [21], we measure the compute complexity by $MN$ as (SGD) queries $M$-dimensional gradients for $N$ times. Given a total compute budget of $MN = C$, from Theorem 6.1 and $N_{\mathtt{eff}} := N/\log(N)$, we see that the best population risk is achieved by setting $\gamma = \Theta(1)$, $M = \tilde{\Theta}(C^{1/(a+1)})$, and $N = \tilde{\Theta}(C^{a/(a+1)})$. Our theory suggests setting a data size slightly larger than the model size when the compute budget is the bottleneck.

**Comparison with [8].** The work by Bordelon et al. [8] considered the scaling law of batch gradient descent (or gradient flow) on a teacher-student model (see their equation (14)). Their teacher-student model can be viewed as our sketched linear regression model. However, we consider one-pass SGD, therefore in our setting the number of gradient steps is equivalent to the data size. When we equalize the number of gradient steps and the data size in their equation (14) and set the parameter prior as Assumption 1C, their prediction is consistent with ours. However, our analysis shows the computational advantage of SGD over batch GD since each iteration requires only $1/N$ the compute. Bordelon et al. [8] obtained the limit of the population risk as two out of the data size, model size, and the number of gradient steps go to infinity based on statistical physics heuristics. In comparison, we obtain upper and lower risk bounds that hold for any finite $M$ and $N$ and match ignoring a constant factor depending only on the spectrum power-law degree $a$.

**Average of the SGD iterates** Results similar to Theorem 4.1 can also be established for the average of the iterates of online SGD with constant stepsize [34, 16, 25, 24, 48]. All results will be the same once replacing the effective sample size $N_{\mathtt{eff}}$ in Theorem 4.1 to the sample size $N$. For more details see Theorem F.6 in Appendix F.

### 4.1 Scaling law under source condition

The isotropic parameter prior condition (Assumption 1C) in Theorem 4.1 can be generalized to the following anisotropic version [11].

**Assumption 3** (Source condition)**.** *Let* $(\lambda_i, \mathbf{v}_i)_{i>0}$ *be the eigenvalues and eigenvectors of* $\mathbf{H}$ *with* $(\lambda_i)_{i>0}$ *in non-increasing order. Assume* $\mathbf{w}^*$ *satisfies a prior such that*

$$\text{for } i \ne j,\ \ \mathbb{E}\langle \mathbf{v}_i, \mathbf{w}^*\rangle\langle \mathbf{v}_j, \mathbf{w}^*\rangle = 0;\ \ \text{and for } i>0,\ \ \mathbb{E}\lambda_i\langle \mathbf{v}_i, \mathbf{w}^*\rangle^2 \eqsim i^{-b},\ \ \text{for some } b>1.$$

A larger exponent $b$ implies a faster decay of signal $\mathbf{w}^*$ and thus corresponds to a simpler task [11]. Note that Assumption 1C satisfies Assumption 3 with $b = a$.

**Theorem 4.2** (Scaling law under source condition)**.** *In Theorem 4.1, suppose Assumption 1C is replaced by Assumption 3 with $1 < b < a + 1$. Then there exists some $a$-dependent constant $c > 0$ such that when $\gamma \le c$, with probability at least $1 - e^{-\Omega(M)}$ over the randomness of the sketch matrix $\mathbf{S}$, we have*

$$\mathbb{E}\mathcal{R}_M(\mathbf{v}_N) = \sigma^2 + \underbrace{\Theta\left(\frac{1}{M^{b-1}}\right) + \Theta\left(\frac{1}{(N_{\mathtt{eff}}\gamma)^{(b-1)/a}}\right)}_{\mathsf{Approx+Bias}} + \underbrace{\Theta\left(\frac{\min\left\{M,\ (N_{\mathtt{eff}}\gamma)^{1/a}\right\}}{N_{\mathtt{eff}}}\right)}_{\mathsf{Var}}.$$

*where the expectation is over the randomness of $\mathbf{w}^*$ and $(\mathbf{x}_i, y_i)_{i=1}^N$, and $\Theta(\cdot)$ hides constants that may depend on $(a, b)$.*

When $1 < b \le a$, the tasks are relatively hard (compared to when $b = a$), and the variance error is dominated by the sum of approximation and bias errors for all choices of $M$, $N$, and $\gamma \lesssim 1$. In this case, Theorem 4.2 gives the same prediction about optimal stepsize and optimal allocation of data and model sizes under compute budget as Theorem 4.1.

When $a < b < a + 1$, the tasks are relatively easy (compared to when $b = a$), and variance remains dominated by the sum of approximation and bias error if the stepsize is optimally tuned. Recall that $\gamma \lesssim 1$, thus we can rewrite the risk bound in Theorem 4.2 as

$$\begin{aligned}\mathbb{E}\mathcal{R}_M(\mathbf{v}_N) - \sigma^2 &\eqsim \frac{1}{\min\left\{M,\ (N_{\mathtt{eff}}\gamma)^{1/a}\right\}^{b-1}} + \frac{\min\left\{M,\ (N_{\mathtt{eff}}\gamma)^{1/a}\right\}}{N_{\mathtt{eff}}} \\ &\eqsim \begin{cases} \min\left\{M,\ (N_{\mathtt{eff}}\gamma)^{1/a}\right\}/N_{\mathtt{eff}} & M \gtrsim N_{\mathtt{eff}}^{1/b} \text{ and } N_{\mathtt{eff}}^{a/b-1} \lesssim \gamma \lesssim 1, \\ \min\left\{M,\ (N_{\mathtt{eff}}\gamma)^{1/a}\right\}^{1-b} & M \lesssim N_{\mathtt{eff}}^{1/b} \text{ or } \gamma \lesssim N_{\mathtt{eff}}^{a/b-1}. \end{cases}\end{aligned}$$

Therefore the optimal stepsize and the risk under the optimal stepsize is

$$\gamma \eqsim N_{\mathtt{eff}}^{a/b-1} \text{ if } M \gtrsim N_{\mathtt{eff}}^{1/b}, \quad \text{and} \quad M^a/N_{\mathtt{eff}} \lesssim \gamma \lesssim 1 \text{ if } M \lesssim N_{\mathtt{eff}}^{1/b}.$$

Under the *optimally tuned* stepsize, the population risk is in the form of

$$\min_\gamma \mathbb{E}\mathcal{R}_M(\mathbf{v}_N) = \sigma^2 + \Theta(N_{\mathtt{eff}}^{(1-b)/b}) + \Theta(M^{1-b}),$$

which is again in the scaling law form (1). This is expected since an optimally tuned stepsize controls the variance error by adjusting the strength of the implicit bias of SGD. Under a fixed compute budget $C = MN$, our theory suggests to assign $M = \tilde{\Theta}(C^{1/(b+1)})$ and $N = \tilde{\Theta}(C^{b/(b+1)})$, and set the stepsize to $\gamma \eqsim \tilde{\Theta}(C^{(a-b)/(b+1)})$.

When $b \ge a + 1$, the tasks are even simpler. We provide upper and lower bounds in Appendix D.3. However, there exists a gap between the bounds, fixing which is left for future work.

Moreover, we note that in the comparable regimes where $M$ is large, the results in Theorem 4.2 match existing bounds on the risk of SGD iterates and ridge estimators [33, 36].

### 4.2 Scaling law under logarithmic power law

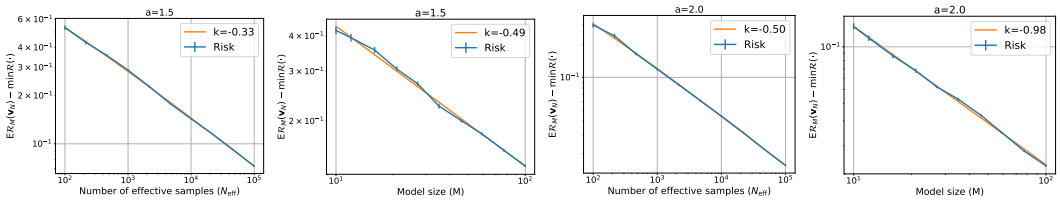
We also derive the risk formula when the data covariance has a logarithmic power-law spectrum [5].

**Assumption 4** (Logarithmic power-law spectrum)**.** *There exists $a > 1$ such that the eigenvalues of $\mathbf{H}$ satisfy $\lambda_i \eqsim i^{-1}\log^{-a}(i+1)$, $i > 0$.*

**Theorem 4.3** (Scaling law under logarithmic power spectrum)**.** *In Theorem 4.1, suppose Assumption 2 is replaced by Assumption 4. Then with probability at least $1 - e^{-\Omega(M)}$ over the randomness of the sketch matrix $\mathbf{S}$, we have*

$$\mathbb{E}\mathcal{R}_M(\mathbf{v}_N) = \sigma^2 + \Theta\left(\frac{1}{\log^{a-1}(M)}\right) + \Theta\left(\frac{1}{\log^{a-1}(N_{\mathtt{eff}}\gamma)}\right), \quad \mathsf{Var} \eqsim \frac{\min\left\{M,\ \frac{N_{\mathtt{eff}}\gamma}{\log^a(N_{\mathtt{eff}}\gamma)}\right\}}{N_{\mathtt{eff}}},$$

*where the expectation is over the randomness of $\mathbf{w}^*$ and $(\mathbf{x}_i, y_i)_{i=1}^N$.*




(a) $a = 1.5$ (b) $a = 1.5$ (c) $a = 2$ (d) $a = 2$

Figure 2: The expected risk of the last iterate of (SGD) minus the irreducible risk versus the effective sample size and model size. Parameters $\sigma = 1, \gamma = 0.1$. (a), (b): $a = 1.5, d = 10000$; (c), (d): $a = 2, d = 1000$. The error bars denote the $\pm 1$ standard deviation of estimating the expected risk using 100 independent samples of $(\mathbf{w}^*, \mathbf{S})$. We use linear functions to fit the expected risk under the log-log scale and report the slope of the fitted lines (denoted by $k$).

Theorem 4.3 provides a scaling law under the logarithmic power-law spectrum. Similar to Theorem 4.1, the variance error is dominated by the approximation and bias errors for all choices of $M$, $N$, and $\gamma$, and thus disappeared from the risk bound. Different from Theorem 4.1, here the population risk is a polynomial of $\log(M)$ and $\log(N_{\mathtt{eff}}\gamma)$.

## 5 Experiments

In this section, we examine the relation between the expected risk of the (SGD) output, the data size $N$, and the model size $M$ when the covariates satisfy a power-law covariance spectrum. Although our results in Section 4 hold with high probability over $\mathbf{S}$, for simplicity, we assume the expectation of the risk is taken over both $\mathbf{w}^*$ and $\mathbf{S}$ in our simulations. We adopt the model in Section 3 and train it using one-pass (SGD) with geometric decaying stepsize (3). We choose the dimension $d$ sufficiently large to approximate the infinite-dimensional case, and the data are generated so that Assumption 1 is satisfied. Moreover, we choose the covariance $\mathbf{H} \in \mathbb{R}^{d\times d}$ to be diagonal with $\mathbf{H}_{ii} \propto i^a$ and $\operatorname{tr}(\mathbf{H}) = 1$ for some $a > 1$. From Figure 1, we observe that the risk indeed follows a power-law formula jointly in the number of samples and the number of parameters. In addition, the fitted exponents are aligned with our theoretical predictions $(a-1, 1-1/a)$ in Theorem 4.1. Figure 2 shows the scaling of the expected risk in data size (or model size) when the model size (or data size) is relatively large. We see that the expected risk also satisfies a power-law decay with exponents matching our predictions. It is noteworthy that our simulations demonstrate stronger observations than the theoretical results in Theorem 4.1, which only establishes matching upper and lower bounds up to a constant factor. Additional simulation results on the risk of the average of (SGD) iterates can be found in Appendix F.

## 6 Risk bounds under a general spectrum

In this section, we present some general results on the upper and lower bounds of the risk of the output of (SGD). Due to the rotational invariance of the sketched matrix $\mathbf{S}$, without loss of generality, we assume the covariance $\mathbf{H}$ is diagonal with non-increasing diagonal entries. Our main results in Section 4 are directly built on the general bounds introduced here.

**Assumption 5** (General distributional conditions)**.** *Assume the following about the data distribution.*

A. ***Hypercontractivity.*** *There exists $\alpha \geq 1$ such that for every PSD matrix $\mathbf{A}$ it holds that*

$$\mathbb{E}\mathbf{x}\mathbf{x}^\top\mathbf{A}\mathbf{x}\mathbf{x}^\top \preceq \alpha \operatorname{tr}(\mathbf{H}\mathbf{A})\mathbf{H}.$$

B. ***Misspecified model.*** *There exists $\sigma^2 > 0$ such that $\mathbb{E}(y - \mathbf{x}^\top\mathbf{w}^*)^2\mathbf{x}\mathbf{x}^\top \preceq \sigma^2\mathbf{H}$.*

It is clear that Assumption 1 implies Assumption 5 with $\alpha = 3$.

**Excess risk decomposition.** Conditioning on the sketch matrix $\mathbf{S}$, the training of the sketched linear predictor can be viewed as an $M$-dimensional linear regression problem. We can therefore

invoke existing SGD analysis [44, 45] to sharply control the excess risk by controlling the bias and variance errors. Specifically, let us define the ($\mathbf{w}^*$-dependent) bias error as

$$\mathsf{Bias}(\mathbf{w}^*) := \left\| \prod_{t=1}^{N} \left(\mathbf{I} - \gamma_t \mathbf{S}\mathbf{H}\mathbf{S}^\top\right)\mathbf{v}^* \right\|_{\mathbf{S}\mathbf{H}\mathbf{S}^\top}^2, \quad \text{where } \mathbf{v}^* := (\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1}\mathbf{S}\mathbf{H}\mathbf{w}^*, \tag{5}$$

and the variance error as

$$\mathsf{Var} := \frac{\#\{\tilde{\lambda}_j \geq 1/(N_{\mathtt{eff}}\gamma)\} + (N_{\mathtt{eff}}\gamma)^2 \sum_{\tilde{\lambda}_j < 1/(N_{\mathtt{eff}}\gamma)} \tilde{\lambda}_j^2}{N_{\mathtt{eff}}}, \quad N_{\mathtt{eff}} := N/\log(N), \tag{6}$$

where $\left(\tilde{\lambda}_j\right)_{j=1}^M$ are eigenvalues of $\mathbf{S}\mathbf{H}\mathbf{S}^\top$. We also let $\mathsf{Bias} := \mathbb{E}\mathsf{Bias}(\mathbf{w}^*)$, where the expectation is over the prior of $\mathbf{w}^*$. Using the existing results on the output of (SGD) in Wu et al. [44, 45], we show that the excess risk in (4) can be exactly decomposed as the sum of bias and variance errors under weak conditions.

**Theorem 6.1** (Excess risk decomposition)**.** *Conditioning on the sketch matrix* $\mathbf{S}$*, consider the excess risk in* (4) *induced by the output of* (SGD)*. Assume* $\mathbf{v}_0 = 0$*. Then for any* $\mathbf{w}^* \in \mathbb{H}$*,*

1. *Under Assumptions 5A and 5B and suppose* $\gamma \leq 1/\left(c\alpha\,\mathrm{tr}(\mathbf{S}\mathbf{H}\mathbf{S}^\top)\right)$ *for some constant* $c > 0$*, we have*
$$\mathbb{E}\mathsf{Excess} \lesssim \mathsf{Bias}(\mathbf{w}^*) + \left(\alpha\|\mathbf{w}^*\|_{\mathbf{H}}^2 + \sigma^2\right)\mathsf{Var}.$$
2. *Under the stronger Assumptions 1A and 1B and suppose* $\gamma \leq 1/\left(c\alpha\,\mathrm{tr}(\mathbf{S}\mathbf{H}\mathbf{S}^\top)\right)$ *for some constant* $c > 0$*, we have*
$$\mathbb{E}\mathsf{Excess} \gtrsim \mathsf{Bias}(\mathbf{w}^*) + \sigma^2\mathsf{Var}.$$

*In both results, the expectations of* Excess *are taken over* $(\mathbf{x}_t, y_t)_{t=1}^N$.

Assuming that the signal-to-noise ratio is upper bounded, that is, $\|\mathbf{w}^*\|_{\mathbf{H}}^2/\sigma^2 \lesssim 1$, then the bias-variance decomposition of the excess risk is sharp up to constant factors.

The variance error is in a nice form and can be computed using the following important lemma on the spectrum of $\mathbf{S}\mathbf{H}\mathbf{S}^\top$. Similar results for logarithmic power-law are also established in Lemma G.6 in Appendix G.

**Lemma 6.2** (Power law)**.** *Under Assumption 2, it holds with probability at least* $1 - e^{-\Omega(M)}$ *that*

$$\mu_j(\mathbf{S}\mathbf{H}\mathbf{S}^\top) \eqsim \mu_j(\mathbf{H}) \eqsim j^{-a}, \quad j = 1, \ldots, M.$$

For any $0 \leq k^* \leq k^\dagger \leq \infty$, let $\mathbf{S}_{k^*:k^\dagger} \in \mathbb{R}^{M \times (k^\dagger - k^*)}$ denote the matrix formed by the $k^* + 1 - k^\dagger$-th columns of $\mathbf{S}$. We also abuse the notation $k^\dagger : \infty$ for $k^\dagger : d$ when $d$ is finite. We let $\mathbf{H}_{k^*:k^\dagger} \in \mathbb{R}^{(k^\dagger - k^*)\times(k^\dagger - k^*)}$ be the submatrix of $\mathbf{H}$ formed by the $k^* + 1 - k^\dagger$-th eigenvalues. For the approximation and bias error, we use the following upper and lower bounds to compute their values.

**Theorem 6.3** (A general upper bound)**.** *Suppose Assumption 5 holds. Assume* $\mathbf{v}_0 = 0$*,* $r(\mathbf{H}) \geq 2M$ *and the initial stepsize satisfies* $\gamma < 1/(c\alpha\,\mathrm{tr}(\mathbf{S}\mathbf{H}\mathbf{S}^\top))$ *for some constant* $c > 0$*. Then for any* $k_1, k_2 \leq M/3$*, with probability at least* $1 - e^{-\Omega(M)}$

$$\mathsf{Approx} \lesssim \|\mathbf{w}^*_{k_1:\infty}\|^2_{\mathbf{H}_{k_1:\infty}} + \left(\frac{\sum_{i>k_1}\lambda_i}{M} + \lambda_{k_1+1} + \sqrt{\frac{\sum_{i>k_1}\lambda_i^2}{M}}\right)\|\mathbf{w}^*_{0:k_1}\|^2,$$

$$\mathsf{Bias}(\mathbf{w}^*) \lesssim \frac{\|\mathbf{w}^*_{0:k_2}\|_2^2}{N_{\mathtt{eff}}\gamma} \cdot \left[\frac{\mu_{M/2}(\mathbf{S}_{k_2:\infty}\mathbf{H}_{k_2:\infty}\mathbf{S}_{k_2:\infty}^\top)}{\mu_M(\mathbf{S}_{k_2:\infty}\mathbf{H}_{k_2:\infty}\mathbf{S}_{k_2:\infty}^\top)}\right]^2 + \|\mathbf{w}^*_{k_2:\infty}\|^2_{\mathbf{H}_{k_2:\infty}}.$$

**Theorem 6.4** (A general lower bound)**.** *Suppose Assumption 1 holds. Assume* $\mathbf{v}_0 = 0$*,* $r(\mathbf{H}) \geq M$ *and the initial stepsize* $\gamma < 1/\left(c\,\mathrm{tr}(\mathbf{S}\mathbf{H}\mathbf{S}^\top)\right)$ *for some constant* $c > 0$*. Then*

$$\mathbb{E}_{\mathbf{w}^*}\mathsf{Approx} \gtrsim \sum_{i=M}^{d}\lambda_i, \quad \mathbb{E}_{\mathbf{w}^*}\mathsf{Bias}(\mathbf{w}^*) \gtrsim \sum_{i:\tilde{\lambda}_i < 1/(N_{\mathtt{eff}}\gamma)} \frac{\mu_i(\mathbf{S}\mathbf{H}^2\mathbf{S}^\top)}{\mu_i(\mathbf{S}\mathbf{H}\mathbf{S}^\top)}$$

*almost surely, where* $(\lambda_i)_{i=1}^d$ *are eigenvalues of* $\mathbf{H}$ *in non-increasing order,* $(\tilde{\lambda}_i)_{i=1}^d$ *are eigenvalues of* $\mathbf{S}\mathbf{H}\mathbf{S}^\top$ *in non-increasing order.*

## 7 Conclusion

We analyze neural scaling laws in infinite-dimensional linear regression. We consider a linear predictor with $M$ trainable parameters on the sketched covariates, which is trained by one-pass stochastic gradient descent with $N$ data. Under a Gaussian prior assumption on the optimal model parameter and a power law (of degree $a > 1$) assumption on the spectrum of the data covariance, we derive matching upper and lower bounds on the population risk minus the irreducible error, that is, $\Theta(M^{-(a-1)} + N^{-(a-1)/a})$. In particular, we show that the variance error, which increases with $M$, is of strictly higher order compared to the other errors, thus disappearing from the risk bound. We attribute the nice empirical formula of the neural scaling law to the non-domination of the variance error, which ultimately is an effect of the implicit regularization of SGD.

Many directions remain open for future study. First, our work is limited to the linear model; it would be interesting to see whether similar scaling laws can be derived in more complex models, such as random feature models or two-layer networks. Second, we focus on one-pass SGD training, and it is unclear if similar results hold for other optimization methods like accelerated SGD or Adam. Additionally, from a technical perspective, many results in our work depend on the Gaussian assumption and the source condition of the data. Investigating how these assumptions can be relaxed would also be valuable.

## Acknowledgements

We gratefully acknowledge the support of the NSF for FODSI through grant DMS-2023505, of the NSF and the Simons Foundation for the Collaboration on the Theoretical Foundations of Deep Learning through awards DMS-2031883 and #814639, and of the ONR through MURI award N000142112431. JDL acknowledges support of the NSF CCF 2002272, NSF IIS 2107304, and NSF CAREER Award 2144994. SMK acknowledges a gift from the Chan Zuckerberg Initiative Foundation to establish the Kempner Institute for the Study of Natural and Artificial Intelligence; support from ONR under award N000142212377, and NSF under award IIS 2229881.

## Bibliography

[1] Ibrahim M Alabdulmohsin, Behnam Neyshabur, and Xiaohua Zhai. Revisiting neural scaling laws in language and vision. *Advances in Neural Information Processing Systems*, 35:22300–22312, 2022.

[2] Alexander B Atanasov, Jacob A Zavatone-Veth, and Cengiz Pehlevan. Scaling and renormalization in high-dimensional regression. *arXiv preprint arXiv:2405.00592*, 2024.

[3] Francis Bach and Eric Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate $o(1/n)$. *Advances in neural information processing systems*, 26:773–781, 2013.

[4] Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. Explaining neural scaling laws. *arXiv preprint arXiv:2102.06701*, 2021.

[5] Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 2020.

[6] Raphaël Berthier, Francis Bach, and Pierre Gaillard. Tight nonparametric convergence rates for stochastic gradient descent under the noiseless linear model. *Advances in Neural Information Processing Systems*, 33:2576–2586, 2020.

[7] Tamay Besiroglu, Ege Erdil, Matthew Barnett, and Josh You. Chinchilla scaling: A replication attempt. *arXiv preprint arXiv:2404.10102*, 2024.

[8] Blake Bordelon, Alexander Atanasov, and Cengiz Pehlevan. A dynamical model of neural scaling laws. *arXiv preprint arXiv:2402.01092*, 2024.

[9] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[10] Ethan Caballero, Kshitij Gupta, Irina Rish, and David Krueger. Broken neural scaling laws. *arXiv preprint arXiv:2210.14891*, 2022.

[11] Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7:331–368, 2007.

[12] Luigi Carratino, Alessandro Rudi, and Lorenzo Rosasco. Learning with sgd and random features. *Advances in neural information processing systems*, 31, 2018.

[13] Leonardo Defilippis, Bruno Loureiro, and Theodor Misiakiewicz. Dimension-free deterministic equivalents for random feature regression. *arXiv preprint arXiv:2405.15699*, 2024.

[14] Alexandre Défossez and Francis Bach. Averaged least-mean-squares: Bias-variance trade-offs and optimal sampling distributions. In *Artificial Intelligence and Statistics*, pages 205–213, 2015.

[15] Aymeric Dieuleveut and Francis R. Bach. Non-parametric stochastic approximation with large step sizes. *The Annals of Statistics*, 2015.

[16] Aymeric Dieuleveut, Nicolas Flammarion, and Francis Bach. Harder, better, faster, stronger convergence rates for least-squares regression. *The Journal of Machine Learning Research*, 18 (1):3520–3570, 2017.

[17] Elvis Dohmatob, Yunzhen Feng, Pu Yang, Francois Charton, and Julia Kempe. A tale of tails: Model collapse as a change of scaling laws. *arXiv preprint arXiv:2402.07043*, 2024.

[18] Rong Ge, Sham M Kakade, Rahul Kidambi, and Praneeth Netrapalli. The step decay schedule: A near optimal, geometrically decaying learning rate procedure for least squares. *Advances in neural information processing systems*, 32, 2019.

[19] Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B Brown, Prafulla Dhariwal, Scott Gray, et al. Scaling laws for autoregressive generative modeling. *arXiv preprint arXiv:2010.14701*, 2020.

[20] Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017.

[21] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

[22] Marcus Hutter. Learning curve theory. *arXiv preprint arXiv:2102.04074*, 2021.

[23] Ayush Jain, Andrea Montanari, and Eren Sasoglu. Scaling laws for learning with real and surrogate data. *arXiv preprint arXiv:2402.04376*, 2024.

[24] Prateek Jain, Praneeth Netrapalli, Sham M Kakade, Rahul Kidambi, and Aaron Sidford. Parallelizing stochastic gradient descent for least squares regression: mini-batching, averaging, and model misspecification. *The Journal of Machine Learning Research*, 18(1):8258–8299, 2017.

[25] Prateek Jain, Sham M. Kakade, Rahul Kidambi, Praneeth Netrapalli, Venkata Krishna Pillutla, and Aaron Sidford. A Markov Chain Theory Approach to Characterizing the Minimax Optimality of Stochastic Gradient Descent (for Least Squares). In *37th IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science (FSTTCS 2017)*, 2018.

[26] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[27] Aran Komatsuzaki. One epoch is all you need. *arXiv preprint arXiv:1906.06669*, 2019.

[28] Alexander Maloney, Daniel A Roberts, and James Sully. A solvable model of neural scaling laws. *arXiv preprint arXiv:2210.16859*, 2022.

[29] Eric Michaud, Ziming Liu, Uzay Girit, and Max Tegmark. The quantization model of neural scaling. *Advances in Neural Information Processing Systems*, 36, 2024.

[30] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

[31] Niklas Muennighoff, Alexander M Rush, Boaz Barak, Teven Le Scao, Aleksandra Piktus, Nouamane Tazi, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. Scaling data-constrained language models. *arXiv preprint arXiv:2305.16264*, 2023.

[32] Yoonsoo Nam, Nayara Fonseca, Seok Hyeong Lee, and Ard Louis. An exactly solvable model for emergence and scaling laws. *arXiv preprint arXiv:2404.17563*, 2024.

[33] Loucas Pillaud-Vivien, Alessandro Rudi, and Francis Bach. Statistical optimality of stochastic gradient descent on hard learning problems through multiple passes. *Advances in Neural Information Processing Systems*, 31, 2018.

[34] Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855, 1992.

[35] Jonathan S Rosenfeld, Amir Rosenfeld, Yonatan Belinkov, and Nir Shavit. A constructive prediction of the generalization error across scales. *arXiv preprint arXiv:1909.12673*, 2019.

[36] Alessandro Rudi and Lorenzo Rosasco. Generalization properties of learning with random features. *Advances in neural information processing systems*, 30, 2017.

[37] Utkarsh Sharma and Jared Kaplan. A neural scaling law from the dimension of the data manifold. *arXiv preprint arXiv:2004.10802*, 2020.

[38] William Swartworth and David P Woodruff. Optimal eigenvalue approximation via sketching. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pages 145–155, 2023.

[39] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[40] Aditya Vardhan Varre, Loucas Pillaud-Vivien, and Nicolas Flammarion. Last iterate convergence of SGD for least-squares in the interpolation regime. In *Advances in Neural Information Processing Systems*, 2021.

[41] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

[42] Alexander Wei, Wei Hu, and Jacob Steinhardt. More than a toy: Random matrix models predict how real-world neural representations generalize. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, 2022.

[43] David P Woodruff et al. Sketching as a tool for numerical linear algebra. *Foundations and Trends® in Theoretical Computer Science*, 10(1–2):1–157, 2014.

[44] Jingfeng Wu, Difan Zou, Vladimir Braverman, Quanquan Gu, and Sham M. Kakade. Last iterate risk bounds of sgd with decaying stepsize for overparameterized linear regression. *The 39th International Conference on Machine Learning*, 2022.

[45] Jingfeng Wu, Difan Zou, Vladimir Braverman, Quanquan Gu, and Sham M. Kakade. The power and limitation of pretraining-finetuning for linear regression under covariate shift. *The 36th Conference on Neural Information Processing Systems*, 2022.

[46] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12104–12113, 2022.

[47] Difan Zou, Jingfeng Wu, Vladimir Braverman, Quanquan Gu, Dean P Foster, and Sham Kakade. The benefits of implicit regularization from sgd in least squares problems. *Advances in Neural Information Processing Systems*, 34:5456–5468, 2021.

[48] Difan Zou, Jingfeng Wu, Vladimir Braverman, Quanquan Gu, and Sham M Kakade. Benign overfitting of constant-stepsize sgd for linear regression. *Journal of Machine Learning Research*, 24(326):1–58, 2023.

# Appendix

## Table of Contents



## A Preliminary

In this section, we provide some preliminary discussions and a proof of Theorem 6.1. Concretely, in Section A.1 we discuss our data assumptions and introduce additional notations. In Section A.2, A.3 we derive intermediate results that contribute to the proof of Theorem 6.1. Finally, a complete proof of Theorem 6.1 is contained in Section A.4.

### A.1 Additional notations and comments on data assumptions

**Tensors.** For matrices $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, $\mathbf{D}$, and $\mathbf{X}$ of appropriate shape, it holds that

$$(\mathbf{B}^\top \otimes \mathbf{A}) \circ \mathbf{X} = \mathbf{A}\mathbf{X}\mathbf{B},$$

and that

$$(\mathbf{D}^\top \otimes \mathbf{C}) \circ (\mathbf{B}^\top \otimes \mathbf{A}) \circ \mathbf{X} = \big((\mathbf{D}^\top \mathbf{B}^\top) \otimes (\mathbf{C}\mathbf{A})\big) \circ \mathbf{X}$$

$$= \mathbf{CAXBD}.$$

For simplicity, we denote

$$\mathbf{A}^{\otimes 2} := \mathbf{A} \otimes \mathbf{A}.$$

**Comments on Assumption 2, 3 and 4** Due to the rotational invariance of the Gaussian sketched matrix $\mathbf{S}$, throughout the appendix, we assume w.l.o.g. that the covariance of the input covariates $\mathbf{H}$ is diagonal with the $(i,i)$-th entry being the $i$-th eigenvalue. Specifically, Assumption 3 can be rewritten as

**Assumption 6** (Source condition)**.** *Assume $\mathbf{H} = (h_{ij})_{i,j\geq 1}$ is a diagonal matrix with diagonal entries in non-increasing order, and $\mathbf{w}^*$ satisfies a prior such that*

$$\text{for } i \neq j,\ \ \mathbb{E}\mathbf{w}_i^*\mathbf{w}_j^* = 0;\ \ \text{and for } i > 0,\ \ \mathbb{E}\lambda_i \mathbf{w}_i^{*2} \eqsim i^{-b},\ \ \text{for some } b > 1.$$

Now that we assume $\mathbf{H}$ is diagonal. We make the following notations. Define

$$\mathbf{H}_{k^*:k^\dagger} := \operatorname{diag}(\lambda_{k^*+1}, \dots, \lambda_{k^\dagger}) \in \mathbb{R}^{(k^\dagger - k^*)^2},$$

where $0 \leq k^* \leq k^\dagger$ are two integers, and we allow $k^\dagger = \infty$. For example,

$$\mathbf{H}_{0:k} = \operatorname{diag}(\lambda_1, \dots, \lambda_i), \quad \mathbf{H}_{k:\infty} = \operatorname{diag}(\lambda_{k+1}, \dots).$$

Similarly, for a vector $\mathbf{w} \in \mathbb{H}$, we have

$$\mathbf{w}_{k^*:k^\dagger} := (\mathbf{w}_{k^*+1}, \dots, \mathbf{w}_{k^\dagger}^*)^\top \in \mathbb{R}^{k^\dagger - k^*}.$$

### A.2 Approximation error

Recall the risk decomposition in (4),

$$\mathcal{R}_M(\mathbf{v}_N) = \underbrace{\min \mathcal{R}(\cdot)}_{\text{Irreducible}} + \underbrace{\min \mathcal{R}_M(\cdot) - \min \mathcal{R}(\cdot)}_{\text{Approx}} + \underbrace{\mathcal{R}_M(\mathbf{v}_N) - \min \mathcal{R}_M(\cdot)}_{\text{Excess}}.$$

**Lemma A.1** (Approximization error)**.** *Conditional on the sketch matrix $\mathbf{S}$, the minimizer of $\mathcal{R}_M(\mathbf{v})$ is given by*

$$\mathbf{v}^* := (\mathbf{SHS}^\top)^{-1}\mathbf{SHw}^*,$$

*and the approximation error in (4) is*

$$\begin{aligned} \text{Approx} &:= \min \mathcal{R}_M(\cdot) - \min \mathcal{R}(\cdot) \\ &= \left\| \left( \mathbf{I} - \mathbf{H}^{\frac{1}{2}}\mathbf{S}^\top \left(\mathbf{SHS}^\top\right)^{-1} \mathbf{SH}^{\frac{1}{2}} \right) \mathbf{H}^{\frac{1}{2}}\mathbf{w}^* \right\|^2. \end{aligned} \tag{7}$$

*Moreover,* $\text{Approx} \leq \|\mathbf{w}^*\|_{\mathbf{H}}^2$ *almost surely over the randomness of* $\mathbf{S}$.

*Proof of Lemma A.1.* Recall that the risk

$$\mathcal{R}(\mathbf{w}) := \mathbb{E}\big(\langle \mathbf{x}, \mathbf{w}\rangle - y\big)^2$$

is a quadratic function and that $\mathbf{w}^*$ is the minimizer of $\mathcal{R}(\cdot)$, so we have

$$\left(\mathbb{E}\mathbf{x}^{\otimes 2}\right)\mathbf{w}^* = \mathbb{E}\mathbf{x}y \Leftrightarrow \mathbf{Hw}^* = \mathbb{E}\mathbf{x}y,$$

and

$$\begin{aligned} \mathcal{R}(\mathbf{w}) &= \mathbb{E}\big(\langle \mathbf{x}, \mathbf{w}\rangle - \langle \mathbf{x}, \mathbf{w}^*\rangle\big)^2 + \mathcal{R}(\mathbf{w}^*) \\ &= \|\mathbf{H}^{\frac{1}{2}}(\mathbf{w} - \mathbf{w}^*)\|^2 + \mathcal{R}(\mathbf{w}^*). \end{aligned}$$

Recall that the risk in a restricted subspace

$$\mathcal{R}_M(\mathbf{v}) := \mathcal{R}(\mathbf{S}^\top\mathbf{v}) = \mathbb{E}\big(\langle \mathbf{Sx}, \mathbf{v}\rangle - y\big)^2$$

is also a quadratic function, so its minimizer is given by

$$\mathbf{v}^* = \left(\mathbb{E}(\mathbf{Sx})^{\otimes 2}\right)^{-1}\mathbb{E}\mathbf{Sx}y$$

$$= \left(\mathbf{S}\mathbf{H}\mathbf{S}^\top\right)^{-1}\mathbf{S}\mathbf{H}\mathbf{w}^*.$$

Therefore, the approximation error is

$$\begin{aligned}
\mathsf{Approx} &:= \mathcal{R}_M(\mathbf{v}^*) - \mathcal{R}(\mathbf{w}^*) \\
&= \mathcal{R}(\mathbf{S}^\top\mathbf{v}^*) - \mathcal{R}(\mathbf{w}^*) \\
&= \|\mathbf{H}^{\frac{1}{2}}(\mathbf{S}^\top\mathbf{v}^* - \mathbf{w}^*)\|^2 \\
&= \|\mathbf{H}^{\frac{1}{2}}(\mathbf{S}^\top\left(\mathbf{S}\mathbf{H}\mathbf{S}^\top\right)^{-1}\mathbf{S}\mathbf{H}\mathbf{w}^* - \mathbf{w}^*)\|^2 \\
&= \left\|\left(\mathbf{I} - \mathbf{H}^{\frac{1}{2}}\mathbf{S}^\top\left(\mathbf{S}\mathbf{H}\mathbf{S}^\top\right)^{-1}\mathbf{S}\mathbf{H}^{\frac{1}{2}}\right)\mathbf{H}^{\frac{1}{2}}\mathbf{w}^*\right\|^2.
\end{aligned}$$

Finally, since

$$\left(\mathbf{I} - \mathbf{H}^{\frac{1}{2}}\mathbf{S}^\top\left(\mathbf{S}\mathbf{H}\mathbf{S}^\top\right)^{-1}\mathbf{S}\mathbf{H}^{\frac{1}{2}}\right)^2 = \mathbf{I} - \mathbf{H}^{\frac{1}{2}}\mathbf{S}^\top\left(\mathbf{S}\mathbf{H}\mathbf{S}^\top\right)^{-1}\mathbf{S}\mathbf{H}^{\frac{1}{2}} \preceq \mathbf{I},$$

it follows that $\mathsf{Approx} \le \|\mathbf{w}^*\|_{\mathbf{H}}^2$. □

### A.3 Bias-variance decomposition

The excess risk in (4) can be viewed as the SGD excess risk in an $M$-dimensional (misspecified) linear regression problem. We will utilize Corollary 3.4 in [45] to get a bias-variance decomposition of the excess risk. The following two lemmas check the related assumptions for Corollary 3.4 in [45] in our setup.

**Lemma A.2** (Hypercontractivity and the misspecified noise under sketched feature)**.** *Suppose that Assumptions 5A and 5B hold. Conditioning on the sketch matrix* $\mathbf{S}$*, for every PSD matrix* $\mathbf{A} \in \mathbb{R}^{M\times M}$*, we have*

$$\mathbb{E}(\mathbf{S}\mathbf{x})^{\otimes 2}\mathbf{A}(\mathbf{S}\mathbf{x})^{\otimes 2} \preceq \alpha\,\mathrm{tr}\left(\mathbf{S}\mathbf{H}\mathbf{S}^\top\mathbf{A}\right)\mathbf{S}\mathbf{H}\mathbf{S}^\top.$$

*Moreover, for the minimizer of* $\mathcal{R}_M(\mathbf{v})$*, that is,* $\mathbf{v}^*$ *defined in Lemma A.1, we have*

$$\mathbb{E}\left(y - \langle\mathbf{v}^*, \mathbf{S}\mathbf{x}\rangle\right)^2(\mathbf{S}\mathbf{x})^{\otimes 2} \preceq 2(\sigma^2 + \alpha\|\mathbf{w}^*\|_{\mathbf{H}}^2)\mathbf{S}\mathbf{H}\mathbf{S}^\top.$$

*The expectation in the above is over* $(\mathbf{x}, y)$*.*

*Proof of Lemma A.2.* The first part is a direct application of Assumption 5A:

$$\begin{aligned}
\mathbb{E}(\mathbf{S}\mathbf{x})^{\otimes 2}\mathbf{A}(\mathbf{S}\mathbf{x})^{\otimes 2} &= \mathbf{S}\left(\mathbb{E}\mathbf{x}\mathbf{x}^\top(\mathbf{S}^\top\mathbf{A}\mathbf{S})\mathbf{x}\mathbf{x}^\top\right)\mathbf{S}^\top \\
&\preceq \mathbf{S}\left(\alpha\,\mathrm{tr}\left(\mathbf{H}\mathbf{S}^\top\mathbf{A}\mathbf{S}\right)\mathbf{H}\right)\mathbf{S}^\top \\
&= \alpha\,\mathrm{tr}\left(\mathbf{S}\mathbf{H}\mathbf{S}^\top\mathbf{A}\right)\mathbf{S}\mathbf{H}\mathbf{S}^\top.
\end{aligned}$$

For the second part, we first show that

$$\begin{aligned}
\mathbb{E}\left(y - \langle\mathbf{v}^*, \mathbf{S}\mathbf{x}\rangle\right)^2\mathbf{x}^{\otimes 2} &\preceq 2\mathbb{E}\left(y - \langle\mathbf{w}^*, \mathbf{x}\rangle\right)^2\mathbf{x}^{\otimes 2} + 2\mathbb{E}\langle\mathbf{w}^* - \mathbf{S}^\top\mathbf{v}^*, \mathbf{x}\rangle^2\mathbf{x}^{\otimes 2} \\
&\preceq 2\sigma^2\mathbf{H} + 2\alpha\langle\mathbf{H}, (\mathbf{w}^* - \mathbf{S}^\top\mathbf{v}^*)^{\otimes 2}\rangle\mathbf{H},
\end{aligned}$$

where the last inequality is by Assumptions 5A and 5B. From the proof of Lemma A.1, we know that

$$\langle\mathbf{H}, (\mathbf{w}^* - \mathbf{S}^\top\mathbf{v}^*)^{\otimes 2}\rangle = \mathsf{Approx} \le \|\mathbf{w}^*\|_{\mathbf{H}}^2, \text{ almost surely.}$$

So we have

$$\mathbb{E}\left(y - \langle\mathbf{v}^*, \mathbf{S}\mathbf{x}\rangle\right)^2\mathbf{x}^{\otimes 2} \preceq 2(\sigma^2 + \alpha\|\mathbf{w}^*\|_{\mathbf{H}}^2)\mathbf{H}.$$

Left and right multiplying both sides with $\mathbf{S}$ and $\mathbf{S}^\top$, we obtain the second claim. □

**Lemma A.3** (Gaussianity and well-specified noise under sketched features)**.** *Suppose that Assumptions 1A and 1B hold. Conditional on the sketch matrix* $\mathbf{S}$*, we have*

$$\mathbf{S}\mathbf{x} \sim \mathcal{N}(0, \mathbf{S}\mathbf{H}\mathbf{S}^\top).$$

*Moreover, for the minimizer of* $\mathcal{R}_M(\mathbf{v})$*, that is,* $\mathbf{v}^*$ *defined in Lemma A.1, we have*

$$\mathbb{E}[y|\mathbf{S}\mathbf{x}] = \langle\mathbf{S}\mathbf{x}, \mathbf{v}^*\rangle, \quad \mathbb{E}(y - \langle\mathbf{S}\mathbf{x}, \mathbf{v}^*\rangle)^2 = \sigma^2 + \mathsf{Approx} \ge \sigma^2.$$

*Proof of Lemma A.3.* The first claim is a direct consequence of Assumption 1A.

For the second claim, by Assumption 1A and Lemma A.1, we have

$$\begin{aligned}\mathbb{E}[y|\mathbf{x}] &= \langle \mathbf{x}, \mathbf{w}^* \rangle \\ &= \langle \mathbf{x}, \mathbf{S}^\top \mathbf{v}^* \rangle + \langle \mathbf{x}, \mathbf{w}^* - \mathbf{S}^\top \mathbf{v}^* \rangle \\ &= \langle \mathbf{x}, \mathbf{S}^\top \mathbf{v}^* \rangle + \langle \mathbf{x}, \big[\mathbf{I} - (\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1}\mathbf{S}\mathbf{H}\big]\mathbf{w}^* \rangle \\ &= \langle \mathbf{H}^{-\frac{1}{2}}\mathbf{x}, \mathbf{H}^{\frac{1}{2}}\mathbf{S}^\top \mathbf{v}^* \rangle + \langle \mathbf{H}^{-\frac{1}{2}}\mathbf{x}, \big[\mathbf{I} - \mathbf{H}^{\frac{1}{2}}\mathbf{S}^\top(\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1}\mathbf{S}\mathbf{H}^{\frac{1}{2}}\big]\mathbf{H}^{\frac{1}{2}}\mathbf{w}^* \rangle \\ &= \langle \mathbf{S}\mathbf{H}^{\frac{1}{2}}\mathbf{H}^{-\frac{1}{2}}\mathbf{x}, \mathbf{v}^* \rangle + \langle \big[\mathbf{I} - \mathbf{H}^{\frac{1}{2}}\mathbf{S}^\top(\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1}\mathbf{S}\mathbf{H}^{\frac{1}{2}}\big]\mathbf{H}^{-\frac{1}{2}}\mathbf{x}, \mathbf{H}^{\frac{1}{2}}\mathbf{w}^* \rangle. \end{aligned} \tag{8}$$

Notice that

$$\mathbf{H}^{-\frac{1}{2}}\mathbf{x} \sim \mathcal{N}(0, \mathbf{I}),$$

by Assumption 1A and that

$$\mathbf{S}\mathbf{H}^{\frac{1}{2}}\big[\mathbf{I} - \mathbf{H}^{\frac{1}{2}}\mathbf{S}^\top(\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1}\mathbf{S}\mathbf{H}^{\frac{1}{2}}\big] = 0,$$

therefore

$$\mathbf{S}\mathbf{x} = \mathbf{S}\mathbf{H}^{\frac{1}{2}}\mathbf{H}^{-\frac{1}{2}}\mathbf{x} \text{ is independent of } \big[\mathbf{I} - \mathbf{H}^{\frac{1}{2}}\mathbf{S}^\top(\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1}\mathbf{S}\mathbf{H}^{\frac{1}{2}}\big]\mathbf{H}^{-\frac{1}{2}}\mathbf{x}.$$

Taking expectation over the second random vector in (8), we find

$$\mathbb{E}[y|\mathbf{S}\mathbf{x}] = \mathbb{E}\mathbb{E}[y|\mathbf{x}] = \langle \mathbf{S}\mathbf{H}^{\frac{1}{2}}\mathbf{H}^{-\frac{1}{2}}\mathbf{x}, \mathbf{v}^* \rangle = \langle \mathbf{S}\mathbf{x}, \mathbf{v}^* \rangle.$$

It remains to show

$$\mathbb{E}(y - \langle \mathbf{S}\mathbf{x}, \mathbf{v}^* \rangle)^2 = \sigma^2 + \mathsf{Approx}.$$

This follows from the proof of Lemma A.1. Specifically,

$$\begin{aligned}\mathbb{E}(y - \langle \mathbf{S}\mathbf{x}, \mathbf{v}^* \rangle)^2 &= \mathcal{R}(\mathbf{S}^\top \mathbf{v}^*) \\ &= \mathsf{Approx} + \mathcal{R}(\mathbf{w}^*) \\ &= \mathsf{Approx} + \sigma^2 \\ &\geq \sigma^2, \end{aligned}$$

where the second equality is by the definition of Approx and the third equality is by Assumption 1B. We have completed the proof. □

### A.4 Proof of Theorem 6.1

We now use the results in [44, 45] for SGD to obtain the following bias-variance decomposition on the excess risk.

**Theorem A.4** (Excess risk bounds)**.** *Consider the excess risk in (4) induced by the output of (SGD). Let*

$$N_{\mathtt{eff}} := N/\log(N), \quad \mathtt{SNR} := (\|\mathbf{w}^*\|_{\mathbf{H}}^2 + \|\mathbf{v}_0\|_{\mathbf{S}\mathbf{H}\mathbf{S}^\top}^2)/\sigma^2.$$

*Then conditioning on the sketch matrix* $\mathbf{S}$*, for any* $\mathbf{w}^* \in \mathbb{H}$

1. *Under Assumptions 5A and 5B, we have*

$$\mathbb{E}\mathsf{Excess} \lesssim \Big\| \prod_{t=1}^{N} \big(\mathbf{I} - \gamma_t \mathbf{S}\mathbf{H}\mathbf{S}^\top\big)(\mathbf{v}_0 - \mathbf{v}^*) \Big\|_{\mathbf{S}\mathbf{H}\mathbf{S}^\top}^2 + (1 + \alpha\mathtt{SNR})\sigma^2 \cdot \frac{D_{\mathtt{eff}}}{N_{\mathtt{eff}}}$$

   *when* $\gamma \lesssim \frac{1}{c\alpha\,\mathrm{tr}(\mathbf{S}\mathbf{H}\mathbf{S}^\top)}$ *for some constant* $c > 0$.

2. *Under Assumptions 1A and 1B, we have*

$$\mathbb{E}\mathsf{Excess} \gtrsim \Big\| \prod_{t=1}^{N} \big(\mathbf{I} - \gamma_t \mathbf{S}\mathbf{H}\mathbf{S}^\top\big)(\mathbf{v}_0 - \mathbf{v}^*) \Big\|_{\mathbf{S}\mathbf{H}\mathbf{S}^\top}^2 + \sigma^2 \cdot \frac{D_{\mathtt{eff}}}{N_{\mathtt{eff}}}$$

   *when* $\gamma \lesssim \frac{1}{c\,\mathrm{tr}(\mathbf{S}\mathbf{H}\mathbf{S}^\top)}$ *for some constant* $c > 0$.

*In both results, the expectation is over* $(\mathbf{x}_t, y_t)_{t=1}^N$*, and*

$$D_{\mathtt{eff}} := \#\{\tilde{\lambda}_j \geq 1/(N_{\mathtt{eff}}\gamma)\} + (N_{\mathtt{eff}}\gamma)^2 \sum_{\tilde{\lambda}_j < 1/(N_{\mathtt{eff}}\gamma)} \tilde{\lambda}_j^2,$$

*where* $(\tilde{\lambda}_j)_{j=1}^M$ *are eigenvalue of* $\mathbf{SHS}^\top$.

Theorem 6.1 follows immediately by Lemma A.1 and by setting $\mathbf{v}_0 = 0$ and plugging the definition of $\mathsf{Bias}(\mathbf{w}^*)$ and $\mathsf{Var}$ into Theorem A.4.

*Proof of Theorem A.4.* This follows from Corollary 3.4 in [45] for a linear regression problem with population data given by $(\mathbf{Sx}, y)$. Note that the data covariance becomes $\mathbf{SHS}^\top$ and the optimal model parameter becomes $\mathbf{v}^*$.

For the upper bound, Lemma A.2 verifies Assumptions 1A and 2 in [45], with the noise level being

$$\tilde{\sigma}^2 = 2(\sigma^2 + \alpha\|\mathbf{w}^*\|_{\mathbf{H}}^2).$$

Then we can apply the upper bound in Corollary 3.4 in [45] (setting their index set $\mathbb{K} = \emptyset$) to get

$$\mathbb{E}\mathsf{Excess} \lesssim \left\| \prod_{t=1}^N (\mathbf{I} - \gamma_t \mathbf{SHS}^\top)(\mathbf{v}_0 - \mathbf{v}^*) \right\|_{\mathbf{SHS}^\top}^2 + (\|\mathbf{v}^* - \mathbf{v}_0\|_{\mathbf{SHS}^\top}^2 + \tilde{\sigma}^2)\frac{D_{\mathtt{eff}}}{N_{\mathtt{eff}}}.$$

We verify that

$$\begin{aligned}
\|\mathbf{v}^* - \mathbf{v}_0\|_{\mathbf{SHS}^\top}^2 &\leq 2\|\mathbf{H}^{\frac{1}{2}}\mathbf{S}^\top\mathbf{v}^*\|^2 + 2\|\mathbf{v}_0\|_{\mathbf{SHS}^\top}^2 \\
&= 2\|\mathbf{H}^{\frac{1}{2}}\mathbf{S}^\top(\mathbf{SHS}^\top)^{-1}\mathbf{SHw}^*\|^2 + 2\|\mathbf{v}_0\|_{\mathbf{SHS}^\top}^2 \\
&\leq 2\|\mathbf{H}^{\frac{1}{2}}\mathbf{w}^*\|^2 + 2\|\mathbf{v}_0\|_{\mathbf{SHS}^\top}^2 \\
&= 2\|\mathbf{w}^*\|_{\mathbf{H}}^2 + 2\|\mathbf{v}_0\|_{\mathbf{SHS}^\top}^2,
\end{aligned}$$

which implies that

$$\begin{aligned}
(\|\mathbf{v}^* - \mathbf{v}_0\|_{\mathbf{SHS}^\top}^2 + \tilde{\sigma}^2) &\leq 2\|\mathbf{w}^*\|_{\mathbf{H}}^2 + 2\|\mathbf{v}_0\|_{\mathbf{SHS}^\top}^2 + 2(\sigma^2 + \alpha\|\mathbf{w}^*\|_{\mathbf{H}}^2) \\
&\lesssim (1 + \alpha\mathtt{SNR})\sigma^2.
\end{aligned}$$

Substituting, we get the upper bound.

For the lower bound, Lemma A.3 shows $\mathbf{Sx}$ is Gaussian, therefore it satisfies Assumption 1B in Wu et al. [45] with $\beta = 1$. Besides, Lemma A.3 shows that the linear regression problem is well-specified, with the noise level being

$$\tilde{\sigma}^2 = \sigma^2 + \mathsf{Approx} \geq \sigma^2.$$

Although the lower bound in Corollary 3.4 in Wu et al. [45] is stated for Gaussian additive noise (see their Assumption 2'), it is easy to check that the lower bound holds for any well-specified noise as described by Lemma A.3. Using the lower bound in Corollary 3.4 in Wu et al. [45], we obtain

$$\mathbb{E}\mathsf{Excess} \gtrsim \left\| \prod_{t=1}^N (\mathbf{I} - \gamma_t \mathbf{SHS}^\top)(\mathbf{v}_0 - \mathbf{v}^*) \right\|_{\mathbf{SHS}^\top}^2 + \tilde{\sigma}^2\frac{D_{\mathtt{eff}}}{N_{\mathtt{eff}}}.$$

Plugging in $\tilde{\sigma}^2 \geq \sigma^2$ gives the desired lower bound. □

### A.5 Proofs of Lemma 6.2, Theorem 6.3 and 6.4

Lemma 6.2 is proved in Lemma G.4. Theorem 6.3 follows from Lemma C.1 and D.1. Theorem 6.4 follows from Lemma C.2 and D.2.

## B Proofs in Section 4

### B.1 Proof of Theorem 4.1

**Proof of part 1.** By Assumption 1B and the definition of $\mathcal{R}(\cdot)$, we have

$$\begin{aligned}
\mathcal{R}(\mathbf{w}) &= \mathbb{E}(\langle \mathbf{x}, \mathbf{w}\rangle - y)^2 = \mathbb{E}(\langle \mathbf{x}, \mathbf{w}\rangle - \mathbb{E}[y \mid \mathbf{x}])^2 + \mathbb{E}(y - \mathbb{E}[y \mid \mathbf{x}])^2 \\
&= \mathbb{E}(\langle \mathbf{x}, \mathbf{w}\rangle - \langle \mathbf{x}, \mathbf{w}^*\rangle)^2 + \sigma^2 \geq \sigma^2.
\end{aligned}$$

Note that the equality holds if and only if $\mathbf{w} = \mathbf{w}^*$. Therefore we have $\min \mathcal{R}(\cdot) = \mathcal{R}(\mathbf{w}^*) = \sigma^2$.

**Proof of part 2.** Part 2 of Theorem 4.1 follows immediately from Lemma C.2.

**Proof of part 3.** We choose $\mathsf{Bias}(\mathbf{w}^*), \mathsf{Var}$ as defined in Eq. (5) and (6) and let $\mathsf{Bias} := \mathbb{E}_{\mathbf{w}^*}\mathsf{Bias}(\mathbf{w}^*)$. Part 3 of Theorem 4.1 follows directly from the decomposition of the excess risk in Theorem 6.1 (note that $\mathbb{E}\|\mathbf{w}^*\|_{\mathbf{H}}^2/\sigma^2 \lesssim 1$), and the matching bounds in Lemma D.3 and E.1.

It remains to verify the stepsize assumption required in Lemma D.3. Since we have from Lemma G.4 that

$$\frac{1}{\mathrm{tr}(\mathbf{SHS}^\top)} = \frac{1}{\sum_{i=1}^M \tilde{\lambda}_i} \geq \frac{c_1}{\sum_{i=1}^M \lambda_i} \geq \frac{c_2}{\sum_{i=1}^M i^{-a}} \geq c_3$$

for some $a$-dependent constants $c_1, c_2, c_3 > 0$ with probability at least $1 - e^{-\Omega(M)}$, it follows that for any constant $c > 1$, we can choose $\gamma \leq c_0$ for some $a$-dependent $c_0$ such that $\gamma \leq \frac{1}{c\,\mathrm{tr}(\mathbf{SHS}^\top)}$. Therefore, we have verified the stepsize assumption.

Finally, the last claim in Theorem 4.1 follows directly from combining the previous three parts and Theorem 6.1, noting $\sigma^2 \lesssim 1$. and

$$\mathsf{Var} \eqsim \frac{\min\{M,\ (N_{\mathtt{eff}}\gamma)^{1/a}\}}{N_{\mathtt{eff}}} \lesssim \frac{(N_{\mathtt{eff}}\gamma)^{1/a}}{N_{\mathtt{eff}}} \lesssim (N_{\mathtt{eff}}\gamma)^{1/a-1} \lesssim \mathsf{Bias} + \mathsf{Approx}$$

under the stepsize assumption $\gamma \lesssim 1$. Here the hidden constants may depend on $a$.

### B.2 Proof of Theorem 4.2

Similar to the proof of Theorem 4.1, we have $\min \mathcal{R}(\cdot) = \sigma^2$ under Assumption 1B. Moreover, by Lemma C.5, D.4 and E.1, we have with probability at least $1 - e^{-\Omega(M)}$ that

$$\begin{aligned}
\mathbb{E}_{\mathbf{w}^*}\mathsf{Approx} &\eqsim M^{1-b},\\
\mathsf{Bias} &\lesssim \max\left\{M^{1-b},\ (N_{\mathtt{eff}}\gamma)^{(1-b)/a}\right\},\\
\mathsf{Bias} &\gtrsim (N_{\mathtt{eff}}\gamma)^{(1-b)/a} \text{ when } (N_{\mathtt{eff}}\gamma)^{1/a} \leq M/3,\\
\mathsf{Var} &\eqsim \min\left\{M,\ (N_{\mathtt{eff}}\gamma)^{1/a}\right\}/N_{\mathtt{eff}},
\end{aligned}$$

when the stepsize $\gamma \leq c$ for some $a$-dependent constant $c > 0$. Here the hidden constants in the bounds may depend only on $(a, b)$. Combining the bounds on $\mathsf{Approx}, \mathsf{Bias}, \mathsf{Var}$ and noting

$$\mathsf{Var} \eqsim \frac{\min\{M,\ (N_{\mathtt{eff}}\gamma)^{1/a}\}}{N_{\mathtt{eff}}} \lesssim \frac{(N_{\mathtt{eff}}\gamma)^{1/a}}{N_{\mathtt{eff}}} \lesssim (N_{\mathtt{eff}}\gamma)^{(1-b)/a} \lesssim \mathsf{Bias} + \mathsf{Approx}$$

yields Theorem 4.2. Here in the second inequality, we use the assumption $b \leq a$.

### B.3 Proof of Theorem 4.3

Similar to the proof of Theorem 4.1, we have $\min \mathcal{R}(\cdot) = \sigma^2$ under Assumption 1B. Notice that we have $\gamma \lesssim 1$ implies $\gamma \lesssim 1/(\mathrm{tr}(\mathbf{SHS}^\top))$ with probability at least $1 - e^{-\Omega(M)}$ by Lemma G.6. It follows from Lemma C.6, D.5 and E.2 that

$$\begin{aligned}
\mathbb{E}_{\mathbf{w}^*}\mathsf{Approx} &\eqsim \log^{1-a} M,\\
\mathsf{Bias} &\lesssim \max\left\{\log^{1-a} M,\ \log^{1-a}(N_{\mathtt{eff}}\gamma)\right\},\\
\mathsf{Bias} &\gtrsim \log^{1-a}(N_{\mathtt{eff}}\gamma) \text{ when } (N_{\mathtt{eff}}\gamma)^{1/a} \leq M^c \text{ for some small constant } c > 0,\\
\mathsf{Var} &\eqsim \frac{\min\{M,\ (N_{\mathtt{eff}}\gamma)/\log^a(N_{\mathtt{eff}}\gamma)\}}{N_{\mathtt{eff}}} \lesssim \frac{(N_{\mathtt{eff}}\gamma)/\log^a(N_{\mathtt{eff}}\gamma)}{N_{\mathtt{eff}}\gamma} = \log^{-a}(N_{\mathtt{eff}}\gamma)
\end{aligned}$$

with probability at least $1 - e^{-\Omega(M)}$ when the stepsize $\gamma \leq c$ for some $a$-dependent constant $c > 0$. Since $\mathsf{Var} \lesssim \mathsf{Bias}$ and $\log^{1-a}(N_{\mathtt{eff}}\gamma) \lesssim \log^{1-a} M$ when $(N_{\mathtt{eff}}\gamma)^{1/a} \gtrsim M^c$, putting the bounds together gives Theorem 4.3.

## C Approximation error

In this section, we derive upper and lower bounds for the approximation error in (4) (and 7). We will also show that the upper and lower bounds match up to constant factors in several examples.

### C.1 An upper bound

**Lemma C.1** (An upper bound on the approximation error)**.** *Given any $k \le d$ such that $r(\mathbf{H}) \ge k+M$, the approximation error in (4) (and 7) satisfies*

$$\mathsf{Approx} \lesssim \|\mathbf{w}^*_{k:\infty}\|^2_{\mathbf{H}_{k:\infty}} + \langle [\mathbf{H}^{-1}_{0:k} + \mathbf{S}^\top_{0:k}\mathbf{A}_k^{-1}\mathbf{S}_{0:k}]^{-1}, \mathbf{w}^*_{0:k}\mathbf{w}^{*\top}_{0:k}\rangle$$

*almost surely, where $\mathbf{A}_k := \mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{S}^\top_{k:\infty}$. If in addition $k \le M/2$, then with probability $1 - e^{-\Omega(M)}$*

$$\mathsf{Approx} \lesssim \|\mathbf{w}^*_{k:\infty}\|^2_{\mathbf{H}_{k:\infty}} + \Big(\frac{\sum_{i>k}\lambda_i}{M} + \lambda_{k+1} + \sqrt{\frac{\sum_{i>k}\lambda_i^2}{M}}\Big)\|\mathbf{w}^*_{0:k}\|^2,$$

*where $(\lambda_i)_{i=1}^p$ are eigenvalues of $\mathbf{H}$ in non-increasing order.*

*Proof of Lemma C.1.* Write the singular value decomposition $\mathbf{H} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$, where $\mathbf{\Lambda} := \mathrm{diag}\{\lambda_1, \lambda_2, \ldots\}$ with $\lambda_1 \ge \lambda_2 \ge \ldots \ge 0$ and $\mathbf{U}\mathbf{U}^\top = \mathbf{I}$. Define $\tilde{\mathbf{S}} := \mathbf{S}\mathbf{U}, \tilde{\mathbf{w}}^* := \mathbf{U}^\top\mathbf{w}^*$. Then by Lemma A.1 the approximation error $\mathsf{Approx} = \mathsf{Approx}(\mathbf{S}, \mathbf{H}, \mathbf{w}^*)$ satisfies

$$\begin{aligned}
\mathsf{Approx}(\mathbf{S}, \mathbf{H}, \mathbf{w}^*) &= \Big\|\Big(\mathbf{I} - \mathbf{H}^{\frac12}\mathbf{S}^\top\big(\mathbf{S}\mathbf{H}\mathbf{S}^\top\big)^{-1}\mathbf{S}\mathbf{H}^{\frac12}\Big)\mathbf{H}^{\frac12}\mathbf{w}^*\Big\|^2 \\
&= \Big\|\Big(\mathbf{I} - \mathbf{U}\mathbf{\Lambda}^{\frac12}\tilde{\mathbf{S}}^\top\big(\tilde{\mathbf{S}}\mathbf{\Lambda}\tilde{\mathbf{S}}^\top\big)^{-1}\mathbf{S}\mathbf{\Lambda}^{\frac12}\mathbf{U}^\top\Big)\mathbf{U}\mathbf{\Lambda}^{\frac12}\mathbf{U}^\top\mathbf{w}^*\Big\|^2 \\
&= \Big\|\mathbf{U}\Big(\mathbf{I} - \mathbf{\Lambda}^{\frac12}\tilde{\mathbf{S}}^\top\big(\tilde{\mathbf{S}}\mathbf{\Lambda}\tilde{\mathbf{S}}^\top\big)^{-1}\mathbf{S}\mathbf{\Lambda}^{\frac12}\Big)\mathbf{\Lambda}^{\frac12}\mathbf{U}^\top\mathbf{w}^*\Big\|^2 \\
&= \Big\|\Big(\mathbf{I} - \mathbf{\Lambda}^{\frac12}\tilde{\mathbf{S}}^\top\big(\tilde{\mathbf{S}}\mathbf{\Lambda}\tilde{\mathbf{S}}^\top\big)^{-1}\mathbf{S}\mathbf{\Lambda}^{\frac12}\Big)\mathbf{\Lambda}^{\frac12}\tilde{\mathbf{w}}^*\Big\|^2 = \mathsf{Approx}(\tilde{\mathbf{S}}, \mathbf{\Lambda}, \tilde{\mathbf{w}}^*).
\end{aligned}$$

Since $\tilde{\mathbf{S}} \overset{d}{=} \mathbf{S}$ by rotational invariance of standard gaussian variables, it suffices to analyze the case where $\mathbf{H} = \mathbf{\Lambda}$ is a diagonal matrix, as the results may transfer to general $\mathbf{H}$ by replacing $\tilde{\mathbf{w}}^*$ with $\mathbf{w}^*$.

Therefore, from now on we assume w.l.o.g. that $\mathbf{H}$ is a diagonal matrix with non-increasing diagonal entries. Define $\mathbf{A} := \mathbf{S}\mathbf{H}\mathbf{S}^\top$.

By definition of Approx, we have

$$\begin{aligned}
\mathsf{Approx} &= \Big\|\Big(\mathbf{I} - \mathbf{H}^{\frac12}\mathbf{S}^\top\big(\mathbf{S}\mathbf{H}\mathbf{S}^\top\big)^{-1}\mathbf{S}\mathbf{H}^{\frac12}\Big)\mathbf{H}^{\frac12}\mathbf{w}^*\Big\|^2 \\
&= \big\langle [\mathbf{H}^{1/2}\mathbf{S}^\top\mathbf{A}^{-1}\mathbf{S}\mathbf{H}^{1/2} - \mathbf{I}_p]^{\otimes 2}, \mathbf{H}^{1/2}\mathbf{w}^*\mathbf{w}^{*\top}\mathbf{H}^{1/2}\big\rangle.
\end{aligned}$$

Moreover, for any $k \in [p]$

$$\begin{aligned}
\mathbf{H}^{1/2}\mathbf{S}^\top\mathbf{A}^{-1}\mathbf{S}\mathbf{H}^{1/2} - \mathbf{I}_p &= \begin{pmatrix}\mathbf{H}^{1/2}_{0:k}\mathbf{S}^\top_{0:k} \\ \mathbf{H}^{1/2}_{k:\infty}\mathbf{S}^\top_{k:\infty}\end{pmatrix}\mathbf{A}^{-1}\begin{pmatrix}\mathbf{S}_{k:\infty}\mathbf{H}^{1/2}_{k:\infty} & \mathbf{S}_{k:\infty}\mathbf{H}^{1/2}_{k:\infty}\end{pmatrix} - \mathbf{I}_p \\
&= \begin{pmatrix}\mathbf{H}^{1/2}_{0:k}\mathbf{S}^\top_{0:k}\mathbf{A}^{-1}\mathbf{S}_{0:k}\mathbf{H}^{1/2}_{0:k} - \mathbf{I}_k & \mathbf{H}^{1/2}_{0:k}\mathbf{S}^\top_{0:k}\mathbf{A}^{-1}\mathbf{S}_{k:\infty}\mathbf{H}^{1/2}_{k:\infty} \\ \mathbf{H}^{1/2}_{k:\infty}\mathbf{S}^\top_{k:\infty}\mathbf{A}^{-1}\mathbf{S}_{0:k}\mathbf{H}^{1/2}_{0:k} & \mathbf{H}^{1/2}_{k:\infty}\mathbf{S}^\top_{k:\infty}\mathbf{A}^{-1}\mathbf{S}_{k:\infty}\mathbf{H}^{1/2}_{k:\infty} - \mathbf{I}_{d-k}\end{pmatrix} \\
&=: \begin{pmatrix}\mathsf{U} & \mathsf{V} \\ \mathsf{V}^\top & \mathsf{W}\end{pmatrix} \qquad (9)
\end{aligned}$$

Therefore

$$[\mathbf{H}^{1/2}\mathbf{S}^\top\mathbf{A}^{-1}\mathbf{S}\mathbf{H}^{1/2} - \mathbf{I}_p]^{\otimes 2} = \begin{pmatrix}\mathsf{U}^2 + \mathsf{V}\mathsf{V}^\top & \mathsf{U}\mathsf{V} + \mathsf{V}\mathsf{W} \\ \mathsf{V}^\top\mathsf{U} + \mathsf{W}\mathsf{V}^\top & \mathsf{W}^2 + \mathsf{V}^\top\mathsf{V}\end{pmatrix} \preceq 2\begin{pmatrix}\mathsf{U}^2 + \mathsf{V}\mathsf{V}^\top & \mathbf{0} \\ \mathbf{0} & \mathsf{W}^2 + \mathsf{V}^\top\mathsf{V}\end{pmatrix},$$

and hence

$$\mathsf{Approx} \le 2\left\langle \begin{pmatrix}\mathsf{U}^2 + \mathsf{V}\mathsf{V}^\top & \mathbf{0} \\ \mathbf{0} & \mathsf{W}^2 + \mathsf{V}^\top\mathsf{V}\end{pmatrix}, \mathbf{H}^{1/2}\mathbf{w}^*\mathbf{w}^{*\top}\mathbf{H}^{1/2}\right\rangle$$

$$= 2\langle \mathsf{U}^2 + \mathsf{V}\mathsf{V}^\top, \mathbf{H}_{0:k}^{1/2}\mathbf{w}_{*,0:k}\mathbf{w}_{*,0:k}^\top\mathbf{H}_{0:k}^{1/2}\rangle + 2\langle \mathsf{W}^2 + \mathsf{V}^\top\mathsf{V}, \mathbf{H}_{k:\infty}^{1/2}\mathbf{w}_{*,k:\infty}\mathbf{w}_{*,k:\infty}^\top\mathbf{H}_{k:\infty}^{1/2}\rangle.$$

We claim the following results which we will prove at the end of the proof.

$$\langle \mathsf{W}^2 + \mathsf{V}^\top\mathsf{V}, \mathbf{H}_{k:\infty}^{1/2}\mathbf{w}_{*,k:\infty}\mathbf{w}_{*,k:\infty}^\top\mathbf{H}_{k:\infty}^{1/2}\rangle \le \|\mathbf{w}_{k:\infty}^*\|_{\mathbf{H}_{k:\infty}}^2, \tag{10}$$

$$\langle \mathsf{U}^2 + \mathsf{V}\mathsf{V}^\top, \mathbf{H}_{0:k}^{1/2}\mathbf{w}_{*,0:k}\mathbf{w}_{*,0:k}^\top\mathbf{H}_{0:k}^{1/2}\rangle = \langle [\mathbf{H}_{0:k}^{-1} + \mathbf{S}_{0:k}^\top\mathbf{A}_k^{-1}\mathbf{S}_{0:k}]^{-1}, \mathbf{w}_{0:k}^*\mathbf{w}_{0:k}^{*\top}\rangle. \tag{11}$$

Note that in claim (11) the inverse $\mathbf{A}_k^{-1}$ exists almost surely since $r(\mathbf{H}_{k:\infty}) \ge r(\mathbf{H}) - k \ge M$ by our assumption and $\mathbf{S}_{k:\infty} \in \mathbb{R}^{M\times(d-k)}$ is a random gaussian projection onto $\mathbb{R}^M$. First part of Lemma C.1 follows immediately from combining claim (10) and (11).

To prove the second part of Lemma C.1, first note that with probability $1 - e^{-\Omega(M)}$ we have

$$\mu_{\min}(\mathbf{A}_k^{-1}) = \|\mathbf{A}_k\|^{-1} \ge c/\Big(\frac{\sum_{i>k}\lambda_i}{M} + \lambda_{k+1} + \sqrt{\frac{\sum_{i>k}\lambda_i^2}{M}}\Big)$$

forc some constant $c > 0$ by Lemma G.2. Moreover, by the concentration of the Gaussian variance matrix (see e.g., Theorem 6.1 in [41]), we have $\mathbf{S}_{0:k}^\top\mathbf{S}_{0:k} \succeq \mathbf{I}_k/5$ with probability $1 - e^{-\Omega(M)}$ when $M/k \ge 2$. Combining the last two arguments, we obtain

$$\begin{aligned}\mathbf{S}_{0:k}^\top\mathbf{A}_k^{-1}\mathbf{S}_{0:k} &\succeq c\mathbf{S}_{0:k}^\top\mathbf{S}_{0:k}/\Big(\frac{\sum_{i>k}\lambda_i}{M} + \lambda_{k+1} + \sqrt{\frac{\sum_{i>k}\lambda_i^2}{M}}\Big)\\ &\gtrsim \mathbf{I}_k/\Big(\frac{\sum_{i>k}\lambda_i}{M} + \lambda_{k+1} + \sqrt{\frac{\sum_{i>k}\lambda_i^2}{M}}\Big),\end{aligned}$$

and therefore

$$\begin{aligned}\langle [\mathbf{H}_{0:k}^{-1} + \mathbf{S}_{0:k}^\top\mathbf{A}_k^{-1}\mathbf{S}_{0:k}]^{-1}, \mathbf{w}_{0:k}^*\mathbf{w}_{0:k}^{*\top}\rangle &\le \langle [\mathbf{S}_{0:k}^\top\mathbf{A}_k^{-1}\mathbf{S}_{0:k}]^{-1}, \mathbf{w}_{0:k}^*\mathbf{w}_{0:k}^{*\top}\rangle\\ &\le \|[\mathbf{H}_{0:k}^{-1} + \mathbf{S}_{0:k}^\top\mathbf{A}_k^{-1}\mathbf{S}_{0:k}]^{-1}\|\|\mathbf{w}_{0:k}^*\|^2\\ &\lesssim \Big(\frac{\sum_{i>k}\lambda_i}{M} + \lambda_{k+1} + \sqrt{\frac{\sum_{i>k}\lambda_i^2}{M}}\Big)\|\mathbf{w}_{0:k}^*\|^2 \end{aligned}\tag{12}$$

with probability $1 - e^{-\Omega(M)}$. Combining Eq. (12) with the first part of Lemma C.1 completes the proof.

**Proof of claim** (10) Note that

$$\begin{aligned}-\mathbf{I}_{d-k} \preceq \mathsf{W} &= \mathbf{H}_{k:\infty}^{1/2}\mathbf{S}_{k:\infty}^\top\mathbf{A}^{-1}\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}^{1/2} - \mathbf{I}_{d-k}\\ &= \mathbf{H}_{k:\infty}^{1/2}\mathbf{S}_{k:\infty}^\top(\mathbf{S}_{0:k}\mathbf{H}_{0:k}\mathbf{S}_{0:k}^\top + \mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{S}_{k:\infty}^\top)^{-1}\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}^{1/2} - \mathbf{I}_{d-k}\\ &\preceq \mathbf{H}_{k:\infty}^{1/2}\mathbf{S}_{k:\infty}^\top(\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{S}_{k:\infty}^\top)^{-1}\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}^{1/2} - \mathbf{I}_{d-k} \preceq \mathbf{0}_{d-k},\end{aligned}$$

where the last inequality uses the fact that the norm of projection matrices is no greater than one. Therefore, we have $\|\mathsf{W}\|_2 \le 1$. Now, it remains to show

$$\mathsf{W}^2 + \mathsf{V}^\top\mathsf{V} = -\mathsf{W}, \tag{13}$$

as claim (10) is a direct consequence of Eq. (13) and the fact that $\|\mathsf{W}\| \le 1$.

By definition of $\mathsf{W}$ in Eq. (9), we have

$$\begin{aligned}\mathsf{W}^2 &= (\mathbf{H}_{k:\infty}^{1/2}\mathbf{S}_{k:\infty}^\top\mathbf{A}^{-1}\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}^{1/2} - \mathbf{I}_{d-k})^2\\ &= \mathbf{I}_{d-k} - 2\mathbf{H}_{k:\infty}^{1/2}\mathbf{S}_{k:\infty}^\top\mathbf{A}^{-1}\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}^{1/2} + \mathbf{H}_{k:\infty}^{1/2}\mathbf{S}_{k:\infty}^\top\mathbf{A}^{-1}\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{S}_{k:\infty}^\top\mathbf{A}^{-1}\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}^{1/2}\\ &= \mathbf{I}_{d-k} - 2\mathbf{H}_{k:\infty}^{1/2}\mathbf{S}_{k:\infty}^\top\mathbf{A}^{-1}\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}^{1/2} + \mathbf{H}_{k:\infty}^{1/2}\mathbf{S}_{k:\infty}^\top\mathbf{A}^{-1}\mathbf{A}_k\mathbf{A}^{-1}\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}^{1/2}.\end{aligned}$$

By definition of $\mathsf{V}$ in Eq. (9), we have

$$\mathsf{V}^\top\mathsf{V} = \mathbf{H}_{k:\infty}^{1/2}\mathbf{S}_{k:\infty}^\top\mathbf{A}^{-1}(\mathbf{S}_{0:k}\mathbf{H}_{0:k}\mathbf{S}_{0:k}^\top)\mathbf{A}^{-1}\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}^{1/2}.$$

Since $\mathbf{S}_{0:k}\mathbf{H}_{0:k}\mathbf{S}_{0:k}^\top + \mathbf{A}_k = \mathbf{A}$, it follows that

$$\begin{aligned}\mathsf{W}^2 + \mathsf{V}^\top\mathsf{V} &= \mathbf{I}_{d-k} - 2\mathbf{H}_{k:\infty}^{1/2}\mathbf{S}_{k:\infty}^\top\mathbf{A}^{-1}\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}^{1/2} + \mathbf{H}_{k:\infty}^{1/2}\mathbf{S}_{k:\infty}^\top\mathbf{A}^{-1}\mathbf{A}\mathbf{A}^{-1}\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}^{1/2}\\ &= \mathbf{I}_{d-k} - \mathbf{H}_{k:\infty}^{1/2}\mathbf{S}_{k:\infty}^\top\mathbf{A}^{-1}\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}^{1/2} = -\mathsf{W}.\end{aligned}$$

**Proof of claim** (11) It suffices to show $\mathsf{U}^2+\mathsf{V}^\top\mathsf{V} = [\mathbf{H}_{0:k}^{-1}+\mathbf{S}_{0:k}^\top\mathbf{A}_k^{-1}\mathbf{S}_{0:k}]^{-1}$. Using the definition of $\mathsf{U}$ in Eq. (9), we obtain

$$\begin{aligned}\mathsf{U} &= \mathbf{H}_{0:k}^{1/2}\mathbf{S}_{0:k}^\top\mathbf{A}^{-1}\mathbf{S}_{0:k}\mathbf{H}_{0:k}^{1/2} - \mathbf{I}_k \\ &= \mathbf{H}_{0:k}^{1/2}\mathbf{S}_{0:k}^\top\mathbf{A}_k^{-1}\mathbf{S}_{0:k}\mathbf{H}_{0:k}^{1/2} - \mathbf{H}_{0:k}^{1/2}\mathbf{S}_{0:k}^\top\mathbf{A}_k^{-1}\mathbf{S}_{0:k}[\mathbf{H}_{0:k}^{-1}+\mathbf{S}_{0:k}^\top\mathbf{A}_k^{-1}\mathbf{S}_{0:k}]^{-1}\mathbf{S}_{0:k}^\top\mathbf{A}_k^{-1}\mathbf{S}_{0:k}\mathbf{H}_{0:k}^{1/2} - \mathbf{I}_k \\ &= \mathbf{H}_{0:k}^{1/2}\mathbf{S}_{0:k}^\top\mathbf{A}_k^{-1}\mathbf{S}_{0:k}[\mathbf{H}_{0:k}^{-1}+\mathbf{S}_{0:k}^\top\mathbf{A}_k^{-1}\mathbf{S}_{0:k}]^{-1}\mathbf{H}_{0:k}^{-1}\mathbf{H}_{0:k}^{1/2} - \mathbf{I}_k,\end{aligned}$$

where the second line uses Woodbury's matrix identity, namely

$$\mathbf{A}^{-1} = [\mathbf{S}_{0:k}\mathbf{H}_{0:k}\mathbf{S}_{0:k}^\top + \mathbf{A}_k]^{-1} = \mathbf{A}_k^{-1} - \mathbf{A}_k^{-1}\mathbf{S}_{0:k}[\mathbf{H}_{0:k}^{-1}+\mathbf{S}_{0:k}^\top\mathbf{A}_k^{-1}\mathbf{S}_{0:k}]^{-1}\mathbf{S}_{0:k}^\top\mathbf{A}_k^{-1}.$$

Continuing the calculation of $\mathsf{U}$, we have

$$\begin{aligned}\mathsf{U} &= \mathbf{H}_{0:k}^{1/2}\mathbf{S}_{0:k}^\top\mathbf{A}_k^{-1}\mathbf{S}_{0:k}[\mathbf{H}_{0:k}^{-1}+\mathbf{S}_{0:k}^\top\mathbf{A}_k^{-1}\mathbf{S}_{0:k}]^{-1}\mathbf{H}_{0:k}^{-1/2} - \mathbf{I}_k \\ &= \mathbf{H}_{0:k}^{1/2}(\mathbf{S}_{0:k}^\top\mathbf{A}_k^{-1}\mathbf{S}_{0:k}[\mathbf{H}_{0:k}^{-1}+\mathbf{S}_{0:k}^\top\mathbf{A}_k^{-1}\mathbf{S}_{0:k}]^{-1} - \mathbf{I}_k)\mathbf{H}_{0:k}^{-1/2} \\ &= -\mathbf{H}_{0:k}^{-1/2}[\mathbf{H}_{0:k}^{-1}+\mathbf{S}_{0:k}^\top\mathbf{A}_k^{-1}\mathbf{S}_{0:k}]^{-1}\mathbf{H}_{0:k}^{-1/2}.\end{aligned}$$

Therefore,

$$\mathsf{U}^2 = \mathbf{H}_{0:k}^{-1/2}[\mathbf{H}_{0:k}^{-1}+\mathbf{S}_{0:k}^\top\mathbf{A}_k^{-1}\mathbf{S}_{0:k}]^{-1}\mathbf{H}_{0:k}^{-1}[\mathbf{H}_{0:k}^{-1}+\mathbf{S}_{0:k}^\top\mathbf{A}_k^{-1}\mathbf{S}_{0:k}]^{-1}\mathbf{H}_{0:k}^{-1/2}. \tag{14}$$

Since

$$\begin{aligned}\mathbf{H}_{0:k}^{1/2}\mathbf{S}_{0:k}^\top\mathbf{A}^{-1} &= \mathbf{H}_{0:k}^{1/2}\mathbf{S}_{0:k}^\top\mathbf{A}_k^{-1} - \mathbf{H}_{0:k}^{1/2}\mathbf{S}_{0:k}^\top\mathbf{A}_k^{-1}\mathbf{S}_{0:k}[\mathbf{H}_{0:k}^{-1}+\mathbf{S}_{0:k}^\top\mathbf{A}_k^{-1}\mathbf{S}_{0:k}]^{-1}\mathbf{S}_{0:k}^\top\mathbf{A}_k^{-1} \\ &= \mathbf{H}_{0:k}^{-1/2}[\mathbf{H}_{0:k}^{-1}+\mathbf{S}_{0:k}^\top\mathbf{A}_k^{-1}\mathbf{S}_{0:k}]^{-1}\mathbf{S}_{0:k}^\top\mathbf{A}_k^{-1}\end{aligned}$$

by Woodbury's matrix indentity, it follows from the definition of $\mathsf{V}$ in Eq. (9) that

$$\begin{aligned}\mathsf{V}\mathsf{V}^\top &= \mathbf{H}_{0:k}^{1/2}\mathbf{S}_{0:k}^\top\mathbf{A}^{-1}\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{S}_{k:\infty}^\top\mathbf{A}^{-1}\mathbf{S}_{0:k}\mathbf{H}_{0:k}^{1/2} \\ &= \mathbf{H}_{0:k}^{-1/2}[\mathbf{H}_{0:k}^{-1}+\mathbf{S}_{0:k}^\top\mathbf{A}_k^{-1}\mathbf{S}_{0:k}]^{-1}\mathbf{S}_{0:k}^\top\mathbf{A}_k^{-1}(\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{S}_{k:\infty}^\top)\mathbf{A}_k^{-1}\mathbf{S}_{0:k}[\mathbf{H}_{0:k}^{-1}+\mathbf{S}_{0:k}^\top\mathbf{A}_k^{-1}\mathbf{S}_{0:k}]^{-1}\mathbf{H}_{0:k}^{-1/2} \\ &= \mathbf{H}_{0:k}^{-1/2}[\mathbf{H}_{0:k}^{-1}+\mathbf{S}_{0:k}^\top\mathbf{A}_k^{-1}\mathbf{S}_{0:k}]^{-1}\mathbf{S}_{0:k}^\top\mathbf{A}_k^{-1}\mathbf{S}_{0:k}[\mathbf{H}_{0:k}^{-1}+\mathbf{S}_{0:k}^\top\mathbf{A}_k^{-1}\mathbf{S}_{0:k}]^{-1}\mathbf{H}_{0:k}^{-1/2}. \qquad (15)\end{aligned}$$

Combining Eq. (14) and (15) yields

$$\mathsf{U}^2 + \mathsf{V}\mathsf{V}^\top = \mathbf{H}_{0:k}^{-1/2}[\mathbf{H}_{0:k}^{-1}+\mathbf{S}_{0:k}^\top\mathbf{A}_k^{-1}\mathbf{S}_{0:k}]^{-1}\mathbf{H}_{0:k}^{-1/2}, \tag{16}$$

and therefore

$$\langle \mathsf{U}^2 + \mathsf{V}\mathsf{V}^\top, \mathbf{H}_{0:k}^{1/2}\mathbf{w}_{*,0:k}\mathbf{w}_{*,0:k}^\top\mathbf{H}_{0:k}^{1/2}\rangle = \langle [\mathbf{H}_{0:k}^{-1}+\mathbf{S}_{0:k}^\top\mathbf{A}_k^{-1}\mathbf{S}_{0:k}]^{-1}, \mathbf{w}_{0:k}^*{\mathbf{w}_{0:k}^*}^\top\rangle.$$

□

### C.2 A lower bound

For the approximation error Approx, we have the following result.

**Lemma C.2** (Lower bound on the approximation error)**.** *When $r(\mathbf{H}) \geq M$, under Assumption 1C, the approximation error in* (4) *(and 7) satisfies*

$$\mathbb{E}_{\mathbf{w}^*}\mathsf{Approx} \gtrsim \sum_{i=M}^{d}\lambda_i,$$

*where $(\lambda_i)_{i=1}^d$ are eigenvalues of $\mathbf{H}$ in non-increasing order.*

*Proof of Lemma C.2.* For any $k \leq d$, following the proof of Lemma C.1, we have

$$\mathsf{Approx} = \langle [\mathbf{H}^{1/2}\mathbf{S}^\top\mathbf{A}^{-1}\mathbf{S}\mathbf{H}^{1/2} - \mathbf{I}_d]^{\otimes 2}, \mathbf{H}^{1/2}\mathbf{w}^*(\mathbf{w}^*)^\top\mathbf{H}^{1/2}\rangle$$

and

$$\mathbf{H}^{1/2}\mathbf{S}^\top\mathbf{A}^{-1}\mathbf{S}\mathbf{H}^{1/2} - \mathbf{I}_d = \begin{pmatrix}\mathbf{H}_{0:k}^{1/2}\mathbf{S}_{0:k}^\top\mathbf{A}^{-1}\mathbf{S}_{0:k}\mathbf{H}_{0:k}^{1/2} - \mathbf{I}_k & \mathbf{H}_{0:k}^{1/2}\mathbf{S}_{0:k}^\top\mathbf{A}^{-1}\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}^{1/2} \\ \mathbf{H}_{k:\infty}^{1/2}\mathbf{S}_{k:\infty}^\top\mathbf{A}^{-1}\mathbf{S}_{0:k}\mathbf{H}_{0:k}^{1/2} & \mathbf{H}_{k:\infty}^{1/2}\mathbf{S}_{k:\infty}^\top\mathbf{A}^{-1}\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}^{1/2} - \mathbf{I}_{d-k}\end{pmatrix}$$

$$=: \begin{pmatrix} \mathsf{U} & \mathsf{V} \\ \mathsf{V}^\top & \mathsf{W} \end{pmatrix}. \tag{17}$$

Therefore

$$[\mathbf{H}^{1/2}\mathbf{S}^\top\mathbf{A}^{-1}\mathbf{S}\mathbf{H}^{1/2} - \mathbf{I}_d]^{\otimes 2} = \begin{pmatrix} \mathsf{U}^2 + \mathsf{V}\mathsf{V}^\top & \mathsf{U}\mathsf{V} + \mathsf{V}\mathsf{W} \\ \mathsf{V}^\top\mathsf{U} + \mathsf{W}\mathsf{V}^\top & \mathsf{W}^2 + \mathsf{V}^\top\mathsf{V} \end{pmatrix}$$

and

$$\begin{aligned}
\mathbb{E}_{\mathbf{w}^*}\mathsf{Approx} &= \mathbb{E}_{\mathbf{w}^*}\langle \mathsf{U}^2 + \mathsf{V}\mathsf{V}^\top, \mathbf{H}_{0:k}^{1/2}\mathbf{w}_{0:k}^*\mathbf{w}_{0:k}^*\mathbf{H}_{0:k}^{1/2}\rangle + \mathbb{E}_{\mathbf{w}^*}\langle \mathsf{W}^2 + \mathsf{V}^\top\mathsf{V}, \mathbf{H}_{0:k}^{1/2}\mathbf{w}_{k:\infty}^*\mathbf{w}_{k:\infty}^*\mathbf{H}_{k:\infty}^{1/2}\rangle \\
&\quad + 2\mathbb{E}_{\mathbf{w}^*}\langle \mathsf{U}\mathsf{V} + \mathsf{V}\mathsf{W}, \mathbf{H}_{0:k}^{1/2}\mathbf{w}_{0:k}^*\mathbf{w}_{k:\infty}^*\mathbf{H}_{k:\infty}^{1/2}\rangle \\
&= \mathrm{tr}((\mathsf{U}^2 + \mathsf{V}\mathsf{V}^\top)\mathbf{H}_{0:k}) + \mathrm{tr}((\mathsf{W}^2 + \mathsf{V}^\top\mathsf{V})\mathbf{H}_{k:\infty}),
\end{aligned}$$

where the last line uses the fact that $\mathbb{E}_{\mathbf{w}^*}(\mathbf{w}^*)^{\otimes 2} = \mathbf{I}_d$. Using Eq. (13) and (16) in the proof of Lemma C.1, we further obtain

$$\begin{aligned}
\mathbb{E}_{\mathbf{w}^*}\mathsf{Approx} &= \mathrm{tr}(\mathbf{H}_{0:k}^{-1/2}[\mathbf{H}_{0:k}^{-1} + \mathbf{S}_{0:k}^\top\mathbf{A}_k^{-1}\mathbf{S}_{0:k}]^{-1}\mathbf{H}_{0:k}^{-1/2}\mathbf{H}_{0:k}) - \mathrm{tr}(\mathsf{W}\mathbf{H}_{k:\infty}) \\
&= \mathrm{tr}([\mathbf{H}_{0:k}^{-1} + \mathbf{S}_{0:k}^\top\mathbf{A}_k^{-1}\mathbf{S}_{0:k}]^{-1}) - \mathrm{tr}(\mathsf{W}\mathbf{H}_{k:\infty}) \\
&\geq -\mathrm{tr}(\mathsf{W}\mathbf{H}_{k:\infty}) =: T_3.
\end{aligned}$$

where $\mathbf{A}_k := \mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{S}_{k:\infty}^\top$. For $T_3$, we further have

$$\begin{aligned}
T_3 &= \mathrm{tr}(\mathbf{H}_{k:\infty}^{1/2}[\mathbf{I}_{d-k} - \mathbf{H}_{k:\infty}^{1/2}\mathbf{S}_{k:\infty}^\top\mathbf{A}^{-1}\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}^{1/2}]\mathbf{H}_{k:\infty}^{1/2}) \\
&\geq \mathrm{tr}(\mathbf{H}_{k:\infty}^{1/2}[\mathbf{I}_{d-k} - \mathbf{H}_{k:\infty}^{1/2}\mathbf{S}_{k:\infty}^\top\mathbf{A}_k^{-1}\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}^{1/2}]\mathbf{H}_{k:\infty}^{1/2}) \\
&\geq \sum_{i=1}^{d-k} \mu_i(\mathbf{I}_{d-k} - \mathbf{H}_{k:\infty}^{1/2}\mathbf{S}_{k:\infty}^\top\mathbf{A}_k^{-1}\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}^{1/2}) \cdot \mu_{d+1-k-i}(\mathbf{H}_{k:\infty}),
\end{aligned}$$

where the second line is due to $\mathbf{A} \succeq \mathbf{A}_k$ (and hence $-\mathbf{A}^{-1} \succeq -\mathbf{A}_k^{-1}$ ), the third line follows from Von-Neuman's inequality. Since $\mathbf{M} := \mathbf{I}_{d-k} - \mathbf{H}_{k:\infty}^{1/2}\mathbf{S}_{k:\infty}^\top\mathbf{A}_k^{-1}\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}^{1/2}$ is a projection matrix such that $\mathbf{M}^2 = \mathbf{M}$ and $\mathrm{tr}(\mathbf{I}_{d-k} - \mathbf{M}) = M$, it follows that $\mathbf{M}$ has $M$ eigenvalues 0 and $d-k-M$ eigenvalues 1. Therefore, we further have

$$T_3 \geq \sum_{i=1}^{d-k} \mu_i(\mathbf{M}) \cdot \mu_{d+1-k-i}(\mathbf{H}_{k:\infty}) \geq \sum_{i=k+M}^{d} \lambda_i$$

for any $k \leq d$. Letting $k = 0$ maximizes the lower bound and concludes the proof. □

### C.3 A lower bound under Assumption 3

**Lemma C.3** (Lower bound on the approximation error under Assumption 3)**.** *Under Assumption 3, the approximation error in* (4) *(and 7) satisfies*

$$\mathbb{E}_{\mathbf{w}^*}\mathsf{Approx} \gtrsim \sum_{i=M}^{d} \lambda_i i^{a-b},$$

*where $(\lambda_i)_{i=1}^d$ are eigenvalues of $\mathbf{H}$ in non-increasing order and the inequality hides some $(a,b)$-dependent constant.*

*Proof of Lemma C.3.* The proof is essentially the same as the proof of Lemma C.2 but we include it here for completeness. Let $\mathbf{H}^{\mathbf{w}} := \mathbb{E}[\mathbf{w}^*\mathbf{w}^{*\top}]$ be the covariance of the prior on $\mathbf{w}^*$. Following the proof of Lemma C.2, we have

$$\begin{aligned}
\mathbb{E}_{\mathbf{w}^*}\mathsf{Approx} &= \mathbb{E}_{\mathbf{w}^*}\langle \mathsf{U}^2 + \mathsf{V}\mathsf{V}^\top, \mathbf{H}_{0:k}^{1/2}\mathbf{w}_{0:k}^*\mathbf{w}_{0:k}^*\mathbf{H}_{0:k}^{1/2}\rangle + \mathbb{E}_{\mathbf{w}^*}\langle \mathsf{W}^2 + \mathsf{V}^\top\mathsf{V}, \mathbf{H}_{0:k}^{1/2}\mathbf{w}_{k:\infty}^*\mathbf{w}_{k:\infty}^*\mathbf{H}_{k:\infty}^{1/2}\rangle \\
&\quad + 2\mathbb{E}_{\mathbf{w}^*}\langle \mathsf{U}\mathsf{V} + \mathsf{V}\mathsf{W}, \mathbf{H}_{0:k}^{1/2}\mathbf{w}_{0:k}^*\mathbf{w}_{k:\infty}^*\mathbf{H}_{k:\infty}^{1/2}\rangle \\
&= \mathrm{tr}((\mathsf{U}^2 + \mathsf{V}\mathsf{V}^\top)\mathbf{H}_{0:k}\mathbf{H}_{0:k}^{\mathbf{w}}) + \mathrm{tr}((\mathsf{W}^2 + \mathsf{V}^\top\mathsf{V})\mathbf{H}_{k:\infty}\mathbf{H}_{k:\infty}^{\mathbf{w}}),
\end{aligned}$$

where the last line uses Assumption 3 and notice that $\mathbf{H}, \mathbf{H}^{\mathbf{w}}$ are both diagonal. Next, similar to the proof of Lemma C.2, using Eq. (13) and (16), we derive

$$\begin{aligned}\mathbb{E}_{\mathbf{w}^*}\mathsf{Approx} &= \mathrm{tr}(\mathbf{H}_{0:k}^{-1/2}[\mathbf{H}_{0:k}^{-1} + \mathbf{S}_{0:k}^\top \mathbf{A}_k^{-1}\mathbf{S}_{0:k}]^{-1}\mathbf{H}_{0:k}^{-1/2}\mathbf{H}_{0:k}\mathbf{H}_{0:k}^{\mathbf{w}}) - \mathrm{tr}(\mathsf{W}\mathbf{H}_{k:\infty}\mathbf{H}_{k:\infty}^{\mathbf{w}}) \\ &= \mathrm{tr}([\mathbf{H}_{0:k}^{-1} + \mathbf{S}_{0:k}^\top \mathbf{A}_k^{-1}\mathbf{S}_{0:k}]^{-1}\mathbf{H}_{0:k}^{\mathbf{w}}) - \mathrm{tr}(\mathsf{W}\mathbf{H}_{k:\infty}\mathbf{H}_{k:\infty}^{\mathbf{w}}) \\ &\geq -\mathrm{tr}(\mathsf{W}\mathbf{H}_{k:\infty}\mathbf{H}_{k:\infty}^{\mathbf{w}}) =: \tilde{T}_3\end{aligned}$$

where $\mathbf{A}_k := \mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{S}_{k:\infty}^\top$. For $\tilde{T}_3$, following the same argument for $T_3$ in the proof of Lemma C.2, we have

$$\begin{aligned}\tilde{T}_3 &\geq \sum_{i=1}^{d-k} \mu_i(\mathbf{I}_{d-k} - \mathbf{H}_{k:\infty}^{1/2}\mathbf{S}_{k:\infty}^\top \mathbf{A}_k^{-1}\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}^{1/2}) \cdot \mu_{d+1-k-i}(\mathbf{H}_{k:\infty}\mathbf{H}_{k:\infty}^{\mathbf{w}}) \\ &\geq \sum_{i=k+M}^{d} \mu_i(\mathbf{H}\mathbf{H}^{\mathbf{w}}) \gtrsim \sum_{i=k+M}^{d} i^{a-b}\lambda_i,\end{aligned}$$

for any $k \leq d$ where the last inequality uses Assumption 3. Setting $k = 0$ maximizes the lower bound and concludes the proof. □

### C.4 Examples on matching bounds for Approx

In this section, we derive matching upper and lower bounds for Approx (defined in Eq. 4 and 7) in three concrete examples: power-law spectrum (Lemma C.4), power-law spectrum with source condition (Lemma C.5) and logarithmic power-law spectrum (Lemma D.5).

**Lemma C.4** (Bounds on Approx under the power-law spectrum)**.** *Suppose Assumption 1C and 2 hold. Then with probability at least* $1 - e^{-\Omega(M)}$ *over the randomness of* $\mathbf{S}$

$$M^{1-a} \lesssim \mathbb{E}_{\mathbf{w}^*}\mathsf{Approx} \lesssim M^{1-a}.$$

*Here, the hidden constants only depend on the power-law degree* $a$*.*

*Proof of Lemma C.4.* For the upper bound, by Lemma C.1 and noting $\mathbb{E}\mathbf{w}_i^{*2} = 1$ for all $i$, we have with probability at least $1 - e^{-\Omega(M)}$

$$\begin{aligned}\mathbb{E}_{\mathbf{w}^*}\mathsf{Approx} &\lesssim \sum_{k>k_1} \lambda_i + \left(\frac{\sum_{i>k_1}\lambda_i}{M} + \lambda_{k_1+1} + \sqrt{\frac{\sum_{i>k_1}\lambda_i^2}{M}}\right) \cdot k_1 \\ &\lesssim k_1^{1-a} + \left(\frac{k_1^{1-a}}{M} + k_1^{-a} + \sqrt{\frac{k_1^{1-2a}}{M}}\right) k_1 \\ &\lesssim \left(\frac{k_1}{M} + 1\right) k_1^{1-a}\end{aligned}$$

for any given $k_1 \leq M/2$. Here the hidden constants depend on $a$. Therefore, letting $k_1 = M/2$ in the upper bound yields

$$\mathbb{E}_{\mathbf{w}^*}\mathsf{Approx} \lesssim M^{1-a}$$

with probability at least $1 - e^{-\Omega(M)}$.

For the lower bound, we have from Lemma C.2 that

$$\mathbb{E}_{\mathbf{w}^*}\mathsf{Approx} \gtrsim \sum_{i=M}^{\infty} i^{-a} \gtrsim M^{1-a}.$$

This completes the proof. □

**Lemma C.5** (Bounds on Approx under the source condition)**.** *Suppose Assumption 3 hold. Then with probability at least* $1 - e^{-\Omega(M)}$ *over the randomness of* $\mathbf{S}$

$$M^{1-b} \lesssim \mathbb{E}_{\mathbf{w}^*}\mathsf{Approx} \lesssim M^{1-b}.$$

*Here, the hidden constants only depend on the power-law degrees* $a, b$*.*

*Proof of Lemma C.5.* For the upper bound, by Lemma C.1 and noting $\mathbb{E}\mathbf{w}_i^{*2} \eqsim i^{a-b}$ for all $i$, we have with probability at least $1 - e^{-\Omega(M)}$

$$\begin{aligned}
\mathsf{Approx} &\lesssim \sum_{k>k_1} \lambda_i i^{a-b} + \left( \frac{\sum_{i>k_1} \lambda_i}{M} + \lambda_{k_1+1} + \sqrt{\frac{\sum_{i>k_1} \lambda_i^2}{M}} \right) \cdot k_1^{1+a-b} \\
&\lesssim k_1^{1-b} + \left( \frac{k_1^{1-a}}{M} + k_1^{-a} + \sqrt{\frac{k_1^{1-2a}}{M}} \right) k_1^{1+a-b} \\
&\lesssim \left( \frac{k_1}{M} + 1 \right) k_1^{1-b}
\end{aligned}$$

for any given $k_1 \leq M/2$. Here the hidden constants depend on $a, b$. Moreover, choosing $k_1 = M/2$ in the upper bound gives

$$\mathbb{E}_{\mathbf{w}^*} \mathsf{Approx} \lesssim M^{1-b}$$

with probability at least $1 - e^{-\Omega(M)}$.

For the lower bound, we have from Lemma C.3 that

$$\mathbb{E}_{\mathbf{w}^*} \mathsf{Approx} \gtrsim \sum_{i=M}^{\infty} i^{-a} \cdot i^{a-b} \gtrsim M^{1-b}.$$

This completes the proof. □

**Lemma C.6** (Bounds on Approx under the logarithmic power-law spectrum)**.** *Suppose Assumption 4 hold. Then with probability at least* $1 - e^{-\Omega(M)}$ *over the randomness of* $\mathbf{S}$

$$\log^{1-a} M \lesssim \mathbb{E}_{\mathbf{w}^*} \mathsf{Approx} \lesssim \log^{1-a} M.$$

*Here, the hidden constants only depend on the power-law degree* $a$.

*Proof of Lemma C.6.* For the upper bound, by Lemma C.1 and noting $\mathbb{E}\mathbf{w}_i^{*2} = 1$ for all $i$, we have with probability at least $1 - e^{-\Omega(M)}$

$$\begin{aligned}
\mathsf{Approx} &\lesssim \sum_{k>k_1} \lambda_i + \left( \frac{\sum_{i>k_1} \lambda_i}{M} + \lambda_{k_1+1} + \sqrt{\frac{\sum_{i>k_1} \lambda_i^2}{M}} \right) k_1 \\
&\lesssim \log^{1-a} k_1 + \left( \frac{\log^{1-a} k_1}{M} + k_1^{-1} \log^{-a} k_1 + \sqrt{\frac{k_1^{1-2a}}{M}} \right) k_1 \\
&\lesssim \left( 1 + \frac{k_1}{M} + \frac{1}{\log k_1} + \frac{1}{\log k_1} \sqrt{\frac{k_1}{M}} \right) \log^{1-a} k_1 \\
&\lesssim \log^{1-a} k_1
\end{aligned}$$

for any given $k_1 \leq M/2$, where the third line uses $\sum_{i>k_1} \lambda_i^2 \lesssim 1/(k_1 \log^{2a} k_1)$. Choosing $k_1 = M/2$, we obtain

$$\mathbb{E}_{\mathbf{w}^*} \mathsf{Approx} \lesssim \log^{1-a} M$$

with probability at least $1 - e^{-\Omega(M)}$. Here the hidden constants depend on $a, b$.

For the lower bound, we have from Lemma C.2 that

$$\mathbb{E}_{\mathbf{w}^*} \mathsf{Approx} \gtrsim \sum_{i=M}^{\infty} \lambda_i \gtrsim \sum_{i=M}^{\infty} i^{-1} \log^{-a} i \gtrsim \log^{1-a} M.$$

Therefore, we have established matching upper and lower bounds for Approx. □

## D Bias error

In this section, we derive upper and lower bounds for $\mathsf{Bias}(\mathbf{w}^*)$ defined in Eq. (5). Moreover, we show that the upper and lower bounds match up to constant factors in concrete examples.

### D.1 An upper bound

**Lemma D.1** (Upper bound on the bias term)**.** *Suppose the initial stepsize $\gamma \le \frac{1}{c\,\mathrm{tr}(\mathbf{SHS}^\top)}$ for some constant $c > 1$. Then for any $\mathbf{w}^* \in \mathbb{H}$ and $k \in [d]$ such that $r(\mathbf{H}) \ge k + M$, the bias term in (5) satisfies*

$$\mathsf{Bias}(\mathbf{w}^*) \lesssim \frac{1}{N_{\mathtt{eff}}\gamma}\|\mathbf{v}^*\|_2^2.$$

*Moreover, for any $k \le M/3$ such that $r(\mathbf{H}) \ge k + M$, the bias term satisfies*

$$\mathsf{Bias}(\mathbf{w}^*) \lesssim \frac{\|\mathbf{w}^*_{0:k}\|_2^2}{N_{\mathtt{eff}}\gamma} \cdot \left[\frac{\mu_{M/2}(\mathbf{A}_k)}{\mu_M(\mathbf{A}_k)}\right]^2 + \|\mathbf{w}^*_{k:\infty}\|^2_{\mathbf{H}_{k:\infty}}$$

*with probability $1 - e^{-\Omega(M)}$, where $\mathbf{A}_k := \mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{S}_{k:\infty}^\top$, $\{\mu_i(\mathbf{A}_k)\}_{i=1}^M$ denote the eigenvalues of $\mathbf{A}_k$ in non-increasing order for some constant $c > 1$.*

*Proof of Lemma D.1.* Similar to the proof of Lemma C.1, we can without loss of generality assume the covariance matrix $\mathbf{H} = \mathrm{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_d\}$ where $\lambda_i \ge \lambda_j$ for any $i \ge j$. Let $\mathbf{SH}^{1/2} = \tilde{\mathbf{U}}\begin{pmatrix}\tilde{\mathbf{\Lambda}}^{1/2} & \mathbf{0}\end{pmatrix}\tilde{\mathbf{V}}^\top$ be the singular value decomposition of $\mathbf{SHS}^\top$, where $\tilde{\mathbf{\Lambda}} := \mathrm{diag}\{\tilde{\lambda}_1, \tilde{\lambda}_2, \ldots, \tilde{\lambda}_d\}$ is a diagonal matrix diagonal entries in non-increasing order. Define $\mathbf{A}_k := \mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{S}_{k:\infty}^\top$. Then it follows from similar arguments as in Lemma C.1 that $\mathbf{A}_k$ is invertible.

Since

$$\|\gamma_t\mathbf{SHS}^\top\|_2 = \gamma_t\tilde{\lambda}_1 \le \gamma\tilde{\lambda}_1 \le \frac{\tilde{\lambda}_1}{c\sum_{i=1}^M \tilde{\lambda}_i} \le 1$$

for some constant $c > 1$ by the stepsize assumption, it follows that $\mathbf{I}_M - \gamma_t\mathbf{SHS}^\top \succ \mathbf{0}_M$ for all $t \in [N]$. Therefore, it can be verified that

$$\prod_{t=1}^N(\mathbf{I}_M - \gamma_t\mathbf{SHS}^\top)\mathbf{SHS}^\top\prod_{t=1}^N(\mathbf{I}_M - \gamma_t\mathbf{SHS}^\top) \preceq (\mathbf{I}_M - \gamma\mathbf{SHS}^\top)^{N_{\mathtt{eff}}}\mathbf{SHS}^\top(\mathbf{I}_M - \gamma\mathbf{SHS}^\top)^{N_{\mathtt{eff}}} =: \mathsf{M},$$

and by definition of $\mathsf{Bias}(\mathbf{w}^*)$ in Eq. (5), we have

$$\begin{aligned}\mathsf{Bias}(\mathbf{w}^*) \eqsim \left\|\prod_{t=1}^N\left(\mathbf{I} - \gamma_t\mathbf{SHS}^\top\right)\mathbf{v}^*\right\|^2_{\mathbf{SHS}^\top} &\le \left\|\left(\mathbf{I} - \gamma\mathbf{SHS}^\top\right)^{N_{\mathtt{eff}}}\mathbf{v}^*\right\|^2_{\mathbf{SHS}^\top} \\ &= \langle\mathsf{M}, \mathbf{v}^{*\otimes 2}\rangle. \end{aligned} \tag{18}$$

Note that the eigenvalues of $\mathsf{M}$ are $\{\tilde{\lambda}_i(1-\gamma\tilde{\lambda}_i)^{2N_{\mathtt{eff}}}\}_{i=1}^M$. Since the function $f(x) = x(1-\gamma x)^{2N_{\mathtt{eff}}}$ is maximized at $x_0 = 1/[(2N_{\mathtt{eff}} + 1)\gamma]$ for $x \in [0, 1/\gamma]$ with $f(x_0) \lesssim 1/(N_{\mathtt{eff}}\gamma)$, it follows that

$$\|\mathsf{M}\|_2 \le c/(N_{\mathtt{eff}}\gamma) \tag{19}$$

for some constant $c > 0$. The first part of Lemma D.1 follows immediately.

Now we prove the second part of Lemma D.1. Recall that $\mathbf{v}^* = \left(\mathbf{SHS}^\top\right)^{-1}\mathbf{SHw}^*$. Substituting $\mathbf{SH} = (\mathbf{S}_{0:k}\mathbf{H}_{0:k} \quad \mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty})$ into $\mathbf{v}^*$, we obtain

$$\begin{aligned}\langle\mathsf{M}, \mathbf{v}^{*\otimes 2}\rangle &= \langle\mathsf{M}, ((\mathbf{SHS}^\top)^{-1}\mathbf{SHw}^*)^{\otimes 2}\rangle \\ &= \mathbf{w}^{*\top}\mathbf{HS}^\top(\mathbf{SHS}^\top)^{-1}\mathsf{M}(\mathbf{SHS}^\top)^{-1}\mathbf{SHw}^* \\ &\le 2T_1 + 2T_2,\end{aligned}$$

where

$$T_1 := (\mathbf{w}^*_{0:k})^\top \mathbf{H}_{0:k}\mathbf{S}_{0:k}^\top (\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1}\mathsf{M}(\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1}\mathbf{S}_{0:k}\mathbf{H}_{0:k}\mathbf{w}^*_{0:k}, \tag{20}$$

$$T_2 := (\mathbf{w}^*_{k:\infty})^\top \mathbf{H}_{k:\infty}\mathbf{S}_{k:\infty}^\top (\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1}\mathsf{M}(\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1}\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{w}^*_{k:\infty}. \tag{21}$$

We claim the following results which we prove later. With probability $1 - e^{-\Omega(M)}$

$$T_1 \le \frac{c\|\mathbf{w}^*_{0:k}\|_2^2}{N_{\mathtt{eff}}\gamma} \cdot \left[\frac{\mu_{M/2}(\mathbf{A}_k)}{\mu_M(\mathbf{A}_k)}\right]^2 \tag{22a}$$

for some constant $c > 0$.

$$T_2 \le \|\mathbf{w}^*_{k:\infty}\|^2_{\mathbf{H}_{k:\infty}}. \tag{22b}$$

Combining Eq. (22a), (22b) gives the second part of Lemma D.1.

**Proof of claim** (22a) By definition of $T_1$, we have

$$T_1 \le \|\mathbf{H}_{0:k}\mathbf{S}_{0:k}^\top (\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1}\mathsf{M}(\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1}\mathbf{S}_{0:k}\mathbf{H}_{0:k}\|_2 \cdot \|\mathbf{w}^*_{0:k}\|_2^2.$$

Moreover,

$$\begin{aligned}
&\|\mathbf{H}_{0:k}\mathbf{S}_{0:k}^\top (\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1}\mathsf{M}(\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1}\mathbf{S}_{0:k}\mathbf{H}_{0:k}\|_2 \\
&\le \|\mathsf{M}\|_2 \cdot \|(\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1}\mathbf{S}_{0:k}\mathbf{H}_{0:k}\|_2^2 \\
&\le \frac{c}{N_{\mathtt{eff}}\gamma}\|(\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1}\mathbf{S}_{0:k}\mathbf{H}_{0:k}\|_2^2
\end{aligned}$$

for some constant $c > 0$, where the last line uses Eq. (19).

It remains to show

$$\|(\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1}\mathbf{S}_{0:k}\mathbf{H}_{0:k}\|_2 \le c \cdot \frac{\mu_{M/2}(\mathbf{A}_k)}{\mu_M(\mathbf{A}_k)} \tag{23}$$

for some constant $c > 0$ with probability $1 - e^{-\Omega(M)}$. Since $\mathbf{S}\mathbf{H}\mathbf{S}^\top = \mathbf{S}_{0:k}\mathbf{H}_{0:k}\mathbf{S}_{0:k}^\top + \mathbf{A}_k$, we have

$$\begin{aligned}
(\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1}\mathbf{S}_{0:k}\mathbf{H}_{0:k} &= (\mathbf{A}_k^{-1} - \mathbf{A}_k^{-1}\mathbf{S}_{0:k}[\mathbf{H}_{0:k}^{-1} + \mathbf{S}_{0:k}^\top\mathbf{A}_k^{-1}\mathbf{S}_{0:k}]^{-1}\mathbf{S}_{0:k}^\top\mathbf{A}_k^{-1})\mathbf{S}_{0:k}\mathbf{H}_{0:k} \\
&= \mathbf{A}_k^{-1}\mathbf{S}_{0:k}\mathbf{H}_{0:k} - \mathbf{A}_k^{-1}\mathbf{S}_{0:k}[\mathbf{H}_{0:k}^{-1} + \mathbf{S}_{0:k}^\top\mathbf{A}_k^{-1}\mathbf{S}_{0:k}]^{-1}\mathbf{S}_{0:k}^\top\mathbf{A}_k^{-1}\mathbf{S}_{0:k}\mathbf{H}_{0:k} \\
&= \mathbf{A}_k^{-1}\mathbf{S}_{0:k}[\mathbf{H}_{0:k}^{-1} + \mathbf{S}_{0:k}^\top\mathbf{A}_k^{-1}\mathbf{S}_{0:k}]^{-1}\mathbf{H}_{0:k}^{-1}\mathbf{H}_{0:k} \\
&= \mathbf{A}_k^{-1}\mathbf{S}_{0:k}[\mathbf{H}_{0:k}^{-1} + \mathbf{S}_{0:k}^\top\mathbf{A}_k^{-1}\mathbf{S}_{0:k}]^{-1},
\end{aligned} \tag{24}$$

where the second line uses Woodbury's identity. Since

$$\mathbf{H}_{0:k}^{-1} + \mathbf{S}_{0:k}^\top\mathbf{A}_k^{-1}\mathbf{S}_{0:k} \succeq \mathbf{S}_{0:k}^\top\mathbf{A}_k^{-1}\mathbf{S}_{0:k},$$

it follows that

$$\|[\mathbf{H}_{0:k}^{-1} + \mathbf{S}_{0:k}^\top\mathbf{A}_k^{-1}\mathbf{S}_{0:k}]^{-1}\|_2 \le \|[\mathbf{S}_{0:k}^\top\mathbf{A}_k^{-1}\mathbf{S}_{0:k}]^{-1}\|_2.$$

Therefore, with probability at least $1 - e^{-\Omega(M)}$

$$\begin{aligned}
\|\mathbf{A}_k^{-1}\mathbf{S}_{0:k}[\mathbf{H}_{0:k}^{-1} + \mathbf{S}_{0:k}^\top\mathbf{A}_k^{-1}\mathbf{S}_{0:k}]^{-1}\|_2 &\le \|\mathbf{A}_k^{-1}\|_2 \cdot \|\mathbf{S}_{0:k}\|_2 \cdot \|[\mathbf{H}_{0:k}^{-1} + \mathbf{S}_{0:k}^\top\mathbf{A}_k^{-1}\mathbf{S}_{0:k}]^{-1}\|_2 \\
&\le \|\mathbf{A}_k^{-1}\|_2 \cdot \|\mathbf{S}_{0:k}\|_2 \cdot \|[\mathbf{S}_{0:k}^\top\mathbf{A}_k^{-1}\mathbf{S}_{0:k}]^{-1}\|_2 \\
&\le \frac{\|\mathbf{A}_k^{-1}\|_2 \cdot \|\mathbf{S}_{0:k}\|_2}{\mu_{\min}(\mathbf{S}_{0:k}^\top\mathbf{A}_k^{-1}\mathbf{S}_{0:k})} \lesssim \frac{\|\mathbf{A}_k^{-1}\|_2}{\mu_{\min}(\mathbf{S}_{0:k}^\top\mathbf{A}_k^{-1}\mathbf{S}_{0:k})}
\end{aligned}$$

where the last inequality follows from the fact that $\|\mathbf{S}_{0:k}\|_2 = \sqrt{\|\mathbf{S}_{0:k}^\top\mathbf{S}_{0:k}\|_2} \le c$ for some constant $c > 0$ when $k \le M/2$ with probability at least $1 - e^{-\Omega(M)}$. Since $\mathbf{S}_{0:k}$ is independent of $\mathbf{A}_k$ and the distribution of $\mathbf{S}_{0:k}$ is rotationally invariant, we may write $\mathbf{S}_{0:k}^\top\mathbf{A}_k^{-1}\mathbf{S}_{0:k} = \sum_{i=1}^M \frac{1}{\tilde{\lambda}_{M-i}}\tilde{\mathbf{s}}_i\tilde{\mathbf{s}}_i^\top$,

where $\tilde{\mathbf{s}}_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \mathbf{I}_k/M)$ and $(\hat{\lambda}_i)_{i=1}^M$ are eigenvalues of $\mathbf{A}_k$ in non-increasing order. Therefore, for $k \le M/3$

$$\mathbf{S}_{0:k}^\top \mathbf{A}_k^{-1} \mathbf{S}_{0:k} = \sum_{i=1}^{M} \frac{1}{\hat{\lambda}_{M-i}} \tilde{\mathbf{s}}_i \tilde{\mathbf{s}}_i^\top \succeq \sum_{i=1}^{M/2} \frac{1}{\hat{\lambda}_{M-i}} \tilde{\mathbf{s}}_i \tilde{\mathbf{s}}_i^\top \succeq \frac{1}{\hat{\lambda}_{M/2}} \sum_{i=1}^{M/2} \tilde{\mathbf{s}}_i \tilde{\mathbf{s}}_i^\top \succeq \frac{c\mathbf{I}_k}{\hat{\lambda}_{M/2}} \tag{25}$$

for some constant $c > 0$ with probability at least $1 - e^{-\Omega(M)}$, where in the last line we again use the concentration properties of gaussian covariance matrices (see e.g., Theorem 6.1 in [41]). As a direct consequence, we have

$$\|\mathbf{A}_k^{-1} \mathbf{S}_{0:k} [\mathbf{H}_{0:k}^{-1} + \mathbf{S}_{0:k}^\top \mathbf{A}_k^{-1} \mathbf{S}_{0:k}]^{-1}\|_2 \le c \cdot \frac{\mu_{M/2}(\mathbf{A}_k)}{\mu_M(\mathbf{A}_k)}$$

with probability at least $1 - e^{-\Omega(M)}$ for some constant $c > 0$. This concludes the proof.

**Proof of claim** (22b) By definition of $T_2$ in Eq. (21), we have

$$\begin{aligned}
T_2 &= \mathbf{w}_{k:\infty}^{*\top} \mathbf{H}_{k:\infty} \mathbf{S}_{k:\infty}^\top (\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1/2} (\mathbf{I}_M - \gamma \mathbf{S}\mathbf{H}\mathbf{S}^\top)^{2N_{\mathtt{eff}}} (\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1/2} \mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty} \mathbf{w}_{k:\infty}^* \\
&\le \mathbf{w}_{k:\infty}^{*\top} \mathbf{H}_{k:\infty} \mathbf{S}_{k:\infty}^\top (\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1} \mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty} \mathbf{w}_{k:\infty}^* \\
&\le \|\mathbf{H}_{k:\infty}^{1/2} \mathbf{S}_{k:\infty}^\top (\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1} \mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty}^{1/2}\| \cdot \|\mathbf{w}_{k:\infty}^*\|_{\mathbf{H}_{k:\infty}}^2 \\
&\le \|\mathbf{w}_{k:\infty}^*\|_{\mathbf{H}_{k:\infty}}^2,
\end{aligned}$$

where the last line follows from

$$\begin{aligned}
\|\mathbf{H}_{k:\infty}^{1/2} \mathbf{S}_{k:\infty}^\top (\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1} \mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty}^{1/2}\|_2 &= \|\mathbf{H}_{k:\infty}^{1/2} \mathbf{S}_{k:\infty}^\top (\mathbf{S}_{0:k} \mathbf{H}_{0:k} \mathbf{S}_{0:k}^\top + \mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty} \mathbf{S}_{k:\infty}^\top)^{-1} \mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty}^{1/2}\|_2 \\
&\le \|\mathbf{H}_{k:\infty}^{1/2} \mathbf{S}_{k:\infty}^\top \mathbf{A}_k^{-1} \mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty}^{1/2}\|_2 \le 1.
\end{aligned}$$

□

### D.2 A lower bound

**Lemma D.2** (Lower bound on the bias term)**.** *Suppose $\mathbf{w}^*$ follows some prior distribution and the initial stepsize $\gamma \le \frac{1}{c\,\mathrm{tr}(\mathbf{S}\mathbf{H}\mathbf{S}^\top)}$ for some constant $c > 2$. Let $\mathbf{H}^{\mathbf{w}} := \mathbb{E}\mathbf{w}^*\mathbf{w}^{*\top}$. Then the bias term in Eq.* (5) *satisfies*

$$\mathbb{E}_{\mathbf{w}^*} \mathsf{Bias}(\mathbf{w}^*) \gtrsim \sum_{i:\tilde{\lambda}_i < 1/(\gamma N_{\mathtt{eff}})} \frac{\mu_i(\mathbf{S}\mathbf{H}\mathbf{H}^{\mathbf{w}}\mathbf{H}\mathbf{S}^\top)}{\mu_i(\mathbf{S}\mathbf{H}\mathbf{S}^\top)}$$

*almost surely, where* $\mathsf{M}_{\mathsf{N}} := \mathbf{S}\mathbf{H}\mathbf{S}^\top \big(\mathbf{I} - 2\gamma \mathbf{S}\mathbf{H}\mathbf{S}^\top\big)^{2N_{\mathtt{eff}}}$.

*Proof of Lemma D.2.* Adopt the notations in the proof of Lemma D.1. By definition of the bias term, we have

$$\begin{aligned}
\mathsf{Bias}(\mathbf{w}^*) &\eqsim \left\| \prod_{t=1}^{N} \big(\mathbf{I} - \gamma_t \mathbf{S}\mathbf{H}\mathbf{S}^\top\big) \mathbf{v}^* \right\|_{\mathbf{S}\mathbf{H}\mathbf{S}^\top}^2 \\
&= \langle \mathbf{S}\mathbf{H}\mathbf{S}^\top \prod_{t=1}^{N} (\mathbf{I} - \gamma_t \mathbf{S}\mathbf{H}\mathbf{S}^\top\big)^{2N_{\mathtt{eff}}}, \mathbf{v}^{*\otimes 2} \rangle \\
&\ge \langle \mathbf{S}\mathbf{H}\mathbf{S}^\top (\mathbf{I} - \sum_{t=1}^{N} \gamma_t \mathbf{S}\mathbf{H}\mathbf{S}^\top\big)^{2N_{\mathtt{eff}}}, \mathbf{v}^{*\otimes 2} \rangle \\
&\ge \langle \mathbf{S}\mathbf{H}\mathbf{S}^\top (\mathbf{I} - 2\gamma \mathbf{S}\mathbf{H}\mathbf{S}^\top\big)^{2N_{\mathtt{eff}}}, \mathbf{v}^{*\otimes 2} \rangle =: \langle \mathsf{M}_{\mathsf{N}}, \mathbf{v}^{*\otimes 2} \rangle,
\end{aligned} \tag{26}$$

where the third line uses $\mathbf{I}_M - 2\gamma_t \mathbf{S}\mathbf{H}\mathbf{S}^\top \succ \mathbf{0}_M$ for all $t \in [N]$ established in the proof of Lemma D.1, $\sum_{i=1}^{N} \gamma_i \le 2\gamma N_{\mathtt{eff}}$, and the fact that $(1-w)(1-v) \ge 1 - w - v$ for $w, v > 0$. Substituting the definition of $\mathbf{v}^*$ in Eq. (5) into the expression, we obtain

$$\mathbb{E}_{\mathbf{w}^*} \mathsf{Bias}(\mathbf{w}^*) \gtrsim \mathbb{E}_{\mathbf{w}^*} \langle \mathsf{M}_{\mathsf{N}}, \mathbf{v}^{*\otimes 2} \rangle = \mathbb{E}_{\mathbf{w}^*} \langle \mathsf{M}_{\mathsf{N}}, ((\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1} \mathbf{S}\mathbf{H}\mathbf{w}^*)^{\otimes 2} \rangle$$

$$
\begin{aligned}
&= \mathrm{tr}(\mathbf{H}\mathbf{S}^\top(\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1}\mathsf{M}_{\mathsf{N}}(\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1}\mathbf{S}\mathbf{H}\mathbf{H}^{\mathbf{w}}) \\
&= \mathrm{tr}((\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1}\mathsf{M}_{\mathsf{N}}(\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1}\mathbf{S}\mathbf{H}\mathbf{H}^{\mathbf{w}}\mathbf{H}\mathbf{S}^\top) \\
&\geq \sum_{i=1}^{M} \mu_{M-i+1}((\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1}\mathsf{M}_{\mathsf{N}}(\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1}) \cdot \mu_i(\mathbf{S}\mathbf{H}\mathbf{H}^{\mathbf{w}}\mathbf{H}\mathbf{S}^\top),
\end{aligned}
$$

where the last line uses Von Neumann's trace inequality. Continuing the calculation, we have

$$
\begin{aligned}
\mathbb{E}_{\mathbf{w}^*}\mathsf{Bias}(\mathbf{w}^*) &\gtrsim \sum_{i=1}^{M} \frac{\mu_i(\mathbf{S}\mathbf{H}\mathbf{H}^{\mathbf{w}}\mathbf{H}\mathbf{S}^\top)}{\mu_i((\mathbf{S}\mathbf{H}\mathbf{S}^\top)^2\mathsf{M}_{\mathsf{N}}{}^{-1})} \\
&= \sum_{i=1}^{M} \frac{\mu_i(\mathbf{S}\mathbf{H}\mathbf{H}^{\mathbf{w}}\mathbf{H}\mathbf{S}^\top)}{\mu_i\Big((\mathbf{S}\mathbf{H}\mathbf{S}^\top)(\mathbf{I}-2\gamma\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-2N_{\mathtt{eff}}}\Big)} \\
&\gtrsim \sum_{i:\tilde{\lambda}_i<1/(\gamma N_{\mathtt{eff}})} \frac{\mu_i(\mathbf{S}\mathbf{H}\mathbf{H}^{\mathbf{w}}\mathbf{H}\mathbf{S}^\top)}{\mu_i(\mathbf{S}\mathbf{H}\mathbf{S}^\top)},
\end{aligned}
$$

where the first inequality uses $\mu_{M+i-1}(A) = \mu_i^{-1}(A^{-1})$ for any positive definite matrix $A \in \mathbb{R}^{M\times M}$, and the second line follows from the definition of $\mathsf{M}_{\mathsf{N}}$ and the fact that $(1-\lambda\gamma N_{\mathtt{eff}})^{-2N_{\mathtt{eff}}} \lesssim 1$ when $\lambda < 1/(\gamma N_{\mathtt{eff}})$. □

### D.3 Examples on matching bounds for $\mathsf{Bias}(\mathbf{w}^*)$

In this section, we derive matching upper and lower bounds for $\mathsf{Bias}(\mathbf{w}^*)$ in (5) in three scenarios: power-law spectrum (Lemma D.3), power-law spectrum with source condition (Lemma D.4) and logarithmic power-law spectrum (Lemma D.5). Recall that we define $\mathsf{Bias} := \mathbb{E}_{\mathbf{w}^*}\mathsf{Bias}(\mathbf{w}^*)$.

**Lemma D.3** (Bounds on Bias under the power-law spectrum)**.** *Suppose Assumption 1C and 2 hold and the initial stepsize $\gamma \leq \frac{1}{c\,\mathrm{tr}(\mathbf{S}\mathbf{H}\mathbf{S}^\top)}$ for some constant $c > 2$. Then with probability at least $1 - e^{-\Omega(M)}$ over the randomness of $\mathbf{S}$*

$$
\mathbb{E}_{\mathbf{w}^*}\mathsf{Bias}(\mathbf{w}^*) \lesssim \max\left\{(N_{\mathtt{eff}}\gamma)^{1/a-1},\ M^{1-a}\right\},
$$

*and*

$$
\mathbb{E}_{\mathbf{w}^*}\mathsf{Bias}(\mathbf{w}^*) \gtrsim (N_{\mathtt{eff}}\gamma)^{1/a-1}
$$

*when $(N_{\mathtt{eff}}\gamma)^{1/a} \leq M/c$ for some constant $c > 0$. Here, all the (hidden) constants depend only on the power-law degree $a$.*

*Proof of Lemma D.3.* For the upper bound, using Lemma G.5, D.1 and the assumption that $\mathbb{E}\mathbf{w}_i^{*2} = 1$ for all $i > 0$, with probability at least $1 - e^{-\Omega(M)}$, we have

$$
\begin{aligned}
\mathbb{E}_{\mathbf{w}^*}\mathsf{Bias}(\mathbf{w}^*) &\lesssim \mathbb{E}_{\mathbf{w}^*}\left[\frac{\|\mathbf{w}^*_{0:k_2}\|_2^2}{N_{\mathtt{eff}}\gamma} + \|\mathbf{w}^*_{k_2:\infty}\|^2_{\mathbf{H}_{k_2:\infty}}\right] \\
&\lesssim \frac{k_2}{N_{\mathtt{eff}}\gamma} + \sum_{k>k_2}\lambda_i \\
&\eqsim \frac{k_2}{N_{\mathtt{eff}}\gamma} + k_2^{1-a} \\
&\lesssim \max\left\{(N_{\mathtt{eff}}\gamma)^{1/a-1},\ M^{1-a}\right\},
\end{aligned}
$$

where in the last inequality, we choose $k_2 = [M/3] \wedge (N_{\mathtt{eff}}\gamma)^{1/a}$ to minimize the upper bound.

When $(N_{\mathtt{eff}}\gamma)^{1/a} \leq M/3$, combining Lemma D.2 and G.4 gives the lower bound

$$
\mathbb{E}_{\mathbf{w}^*}\mathsf{Bias}(\mathbf{w}^*) \gtrsim \sum_{i:\tilde{\lambda}_i<1/(N_{\mathtt{eff}}\gamma)} \frac{\mu_i(\mathbf{S}\mathbf{H}\mathbf{H}^{\mathbf{w}}\mathbf{H}\mathbf{S}^\top)}{\mu_i(\mathbf{S}\mathbf{H}\mathbf{S}^\top)} = \sum_{i:\tilde{\lambda}_i<1/(N_{\mathtt{eff}}\gamma)} \frac{\mu_i(\mathbf{S}\mathbf{H}^2\mathbf{S}^\top)}{\mu_i(\mathbf{S}\mathbf{H}\mathbf{S}^\top)},
$$

$$\gtrsim \sum_{\tilde{\lambda}_i < 1/(N_{\mathtt{eff}}\gamma), i \le M} \frac{i^{-2a}}{i^{-a}} = \sum_{\lambda_i < 1/(N_{\mathtt{eff}}\gamma), i \le M} i^{-a} \gtrsim (N_{\mathtt{eff}}\gamma)^{1/a-1}$$

with probability at least $1 - e^{-\Omega(M)}$. Here, the hidden constants depend only on $a$. □

**Lemma D.4** (Bounds on Bias under the source condition)**.** *Suppose Assumption 3 hold and the initial stepsize $\gamma \le \frac{1}{c\,\mathrm{tr}(\mathbf{SHS}^\top)}$ for some constant $c > 2$. Then with probability at least $1 - e^{-\Omega(M)}$ over the randomness of $\mathbf{S}$*

$$\mathbb{E}_{\mathbf{w}^*}\mathsf{Bias}(\mathbf{w}^*) \lesssim \max\left\{(N_{\mathtt{eff}}\gamma)^{(1-b)/a},\ M^{1-b}\right\},$$

*and*

$$\mathbb{E}_{\mathbf{w}^*}\mathsf{Bias}(\mathbf{w}^*) \gtrsim (N_{\mathtt{eff}}\gamma)^{(1-b)/a}$$

*when $(N_{\mathtt{eff}}\gamma)^{1/a} \le M/c$ for some constant $c > 0$. In both results, the hidden constants depend only on $a, b$.*

*Proof of Lemma D.4.* For the upper bound, using Lemma G.5, D.1 and the assumption that (w.l.o.g.) $\mathbb{E}\mathbf{w}_i^{*2} \eqsim i^{a-b}$ for all $i > 0$, with probability at least $1 - e^{-\Omega(M)}$, we have

$$\begin{aligned}
\mathbb{E}_{\mathbf{w}^*}\mathsf{Bias}(\mathbf{w}^*) &\lesssim \mathbb{E}_{\mathbf{w}^*}\left[\frac{\|\mathbf{w}^*_{0:k_2}\|_2^2}{N_{\mathtt{eff}}\gamma} + \|\mathbf{w}^*_{k_2:\infty}\|^2_{\mathbf{H}_{k_2:\infty}}\right] \\
&\lesssim \frac{k_2^{1+a-b}}{N_{\mathtt{eff}}\gamma} + \sum_{k > k_2} \lambda_i \cdot i^{a-b} \\
&\lesssim \frac{k_2^{1+a-b}}{N_{\mathtt{eff}}\gamma} + k_2^{1-b} \\
&\lesssim \max\left\{(N_{\mathtt{eff}}\gamma)^{(1-b)/a},\ M^{1-b}\right\}
\end{aligned}$$

when $b < a + 1$, where in the last inequality, we choose $k_2 = [M/3] \wedge (N_{\mathtt{eff}}\gamma)^{1/a}$ to minimize the upper bound.

When $(N_{\mathtt{eff}}\gamma)^{1/a} \le M/c$ for some large constant $c > 0$, combining Lemma D.2 and G.4 yields the lower bound

$$\begin{aligned}
\mathbb{E}_{\mathbf{w}^*}\mathsf{Bias}(\mathbf{w}^*) &\gtrsim \sum_{i:\tilde{\lambda}_i < 1/(N_{\mathtt{eff}}\gamma)} \frac{\mu_i(\mathbf{SHH^w HS}^\top)}{\mu_i(\mathbf{SHS}^\top)} \eqsim \sum_{i:\tilde{\lambda}_i < 1/(N_{\mathtt{eff}}\gamma)} \frac{\mu_i(\mathbf{SH}^{(a+b)/a}\mathbf{S}^\top)}{\mu_i(\mathbf{SHS}^\top)}, \\
&\gtrsim \sum_{\tilde{\lambda}_i < 1/(N_{\mathtt{eff}}\gamma), i \le M} \frac{i^{-(a+b)}}{i^{-a}} = \sum_{\lambda_i < 1/(N_{\mathtt{eff}}\gamma), i \le M} i^{-b} \gtrsim (N_{\mathtt{eff}}\gamma)^{(1-b)/a}
\end{aligned}$$

with probability at least $1 - e^{-\Omega(M)}$. Here, the hidden constants depend only on $a, b$.

**Upper bound when $b \ge a + 1$.** Following the previous derivations, when $b = a + 1$, we have with probability at least $1 - e^{-\Omega(M)}$

$$\begin{aligned}
\mathbb{E}_{\mathbf{w}^*}\mathsf{Bias}(\mathbf{w}^*) &\lesssim \mathbb{E}_{\mathbf{w}^*}\left[\frac{\|\mathbf{w}^*_{0:k_2}\|_2^2}{N_{\mathtt{eff}}\gamma} + \|\mathbf{w}^*_{k_2:\infty}\|^2_{\mathbf{H}_{k_2:\infty}}\right] \\
&\lesssim \frac{\log k_2}{N_{\mathtt{eff}}\gamma} + k_2^{1-b} \\
&\lesssim \frac{\log(N_{\mathtt{eff}}\gamma)}{N_{\mathtt{eff}}\gamma} + M^{1-b}
\end{aligned}$$

where the last line follows by setting $k_2 = [M/3] \wedge (N_{\mathtt{eff}}\gamma)^{1/a}$. When $b > a + 1$, we have with probability at least $1 - e^{-\Omega(M)}$

$$\mathbb{E}_{\mathbf{w}^*}\mathsf{Bias}(\mathbf{w}^*) \lesssim \mathbb{E}_{\mathbf{w}^*}\left[\frac{\|\mathbf{w}^*_{0:k_2}\|_2^2}{N_{\mathtt{eff}}\gamma} + \|\mathbf{w}^*_{k_2:\infty}\|^2_{\mathbf{H}_{k_2:\infty}}\right]$$

$$\lesssim \frac{1}{N_{\mathtt{eff}}\gamma} + k_2^{1-b}$$
$$\lesssim \frac{1}{N_{\mathtt{eff}}\gamma} + M^{1-b},$$

where the last follows by chooing $k_2 = M/3$ to minimize the upper bound.

Note that there exist non-constant gaps between the upper and lower bounds on the bias term in the simple regime where $b \geq a+1$. We leave a more precise analysis of the bias term for future work.

□

**Lemma D.5** (Bounds on Bias under the logarithmic power-law spectrum)**.** *Suppose Assumption 4 hold and the initial stepsize $\gamma \leq \frac{1}{c\,\mathrm{tr}(\mathbf{SHS}^\top)}$ for some constant $c > 2$. Let $\mathsf{k} := \inf\{k : k\log^a k \geq N_{\mathtt{eff}}\gamma\}$. Then with probability at least $1 - e^{-\Omega(M)}$ over the randomness of $\mathbf{S}$*

$$\mathbb{E}_{\mathbf{w}^*}\mathsf{Bias}(\mathbf{w}^*) \lesssim \max\left\{\log^{1-a}(N_{\mathtt{eff}}\gamma), \log^{1-a} M\right\},$$

*and*

$$\mathbb{E}_{\mathbf{w}^*}\mathsf{Bias}(\mathbf{w}^*) \gtrsim \log^{1-a}(N_{\mathtt{eff}}\gamma)$$

*when $(N_{\mathtt{eff}}\gamma) \leq M^c$ for some sufficiently small constant $c > 0$. Here, all constants depend only on the power-law degree $a$.*

*Proof of Lemma D.5.* For the upper bound, using Lemma G.7, D.1 and the assumption that $\mathbb{E}\mathbf{w}_i^{*2} = 1$ for all $i > 0$, with probability at least $1 - e^{-\Omega(M)}$, we have

$$\begin{aligned}
\mathbb{E}_{\mathbf{w}^*}\mathsf{Bias}(\mathbf{w}^*) &\lesssim \mathbb{E}_{\mathbf{w}^*}\left[\frac{\|\mathbf{w}^*_{0:k_2}\|_2^2}{N_{\mathtt{eff}}\gamma} + \|\mathbf{w}^*_{k_2:\infty}\|^2_{\mathbf{H}_{k_2:\infty}}\right] \\
&\lesssim \frac{k_2}{N_{\mathtt{eff}}\gamma} + \sum_{k>k_2}\lambda_i \\
&\eqsim \frac{k_2}{N_{\mathtt{eff}}\gamma} + \log^{1-a} k_2 \\
&\lesssim \max\left\{\log^{1-a}(N_{\mathtt{eff}}\gamma), \log^{1-a} M\right\},
\end{aligned}$$

where in the last inequality, we choose $k_2 = [M/3] \wedge \left[(N_{\mathtt{eff}}\gamma)/\log^a(N_{\mathtt{eff}}\gamma)\right]$ to minimize the upper bound.

Recall $\mathsf{k}^* \eqsim M/\log M$ (for example we may define $\mathsf{k}^* = \inf\{k : k\log k \geq M\}$) in Lemma G.6. Combining Lemma D.2 and G.6 gives the lower bound

$$\begin{aligned}
\mathbb{E}_{\mathbf{w}^*}\mathsf{Bias}(\mathbf{w}^*) &\gtrsim \sum_{i:\tilde{\lambda}_i < 1/(N_{\mathtt{eff}}\gamma)} \frac{\mu_i(\mathbf{SHH^w HS}^\top)}{\mu_i(\mathbf{SHS}^\top)} \eqsim \sum_{i:\tilde{\lambda}_i < 1/(N_{\mathtt{eff}}\gamma)} \frac{\mu_i(\mathbf{SH}^2\mathbf{S}^\top)}{\mu_i(\mathbf{SHS}^\top)}, \\
&\gtrsim \sum_{\tilde{\lambda}_i < 1/(N_{\mathtt{eff}}\gamma), i\leq \mathsf{k}^*} \frac{i^{-2}\log^{-2a} i}{i^{-1}\log^{-a} i} = \sum_{\lambda_i < 1/(N_{\mathtt{eff}}\gamma), i\leq \mathsf{k}^*} i^{-1}\log^{-a} i \\
&\gtrsim \sum_{i=N_{\mathtt{eff}}\gamma}^{\mathsf{k}^*} i^{-1}\log^{-a} i \\
&\gtrsim \log^{1-a}(N_{\mathtt{eff}}\gamma) - \log^{1-a}(\mathsf{k}^*) \\
&\gtrsim \log^{1-a}(N_{\mathtt{eff}}\gamma) - c_1\log^{1-a}(M)
\end{aligned}$$

with probability at least $1 - e^{-\Omega(M)}$ for some constant $c_1 > 0$. Here, the (hidden) constants depend only on $a$. Therefore, when $(N_{\mathtt{eff}}\gamma)^{1/a} \leq M^c$ for some sufficiently small constant $c > 0$, we have

$$\mathbb{E}_{\mathbf{w}^*}\mathsf{Bias}(\mathbf{w}^*) \gtrsim \log^{1-a}(N_{\mathtt{eff}}\gamma) - c_1\log^{1-a}(M) \gtrsim \log^{1-a}(N_{\mathtt{eff}}\gamma).$$

with probability at least $1 - e^{-\Omega(M)}$. □

## E Variance error

In this section, we present matching upper and lower bounds on the variance term Var defined in (6) under the power-law or logarithmic power-law spectrum.

**Lemma E.1** (Matching bounds on Var under power-law spectrum)**.** *Under Assumption 2,* Var *defined in Eq.* (6) *satisfies*

$$\mathsf{Var} \eqsim \frac{\min\{M,\ (N_{\mathtt{eff}}\gamma)^{1/a}\}}{N_{\mathtt{eff}}}$$

*with probability at least* $1-e^{-\Omega(M)}$ *over the randomness of* $\mathbf{S}$*. Here, the hidden constants only depend on* $a$*.*

*Proof of Lemma E.1.* By the definition of Var in Eq. (6) and Lemma G.4, we have

$$\begin{aligned}\mathsf{Var} &= \frac{\#\{\tilde{\lambda}_j \ge 1/(N_{\mathtt{eff}}\gamma)\} + (N_{\mathtt{eff}}\gamma)^2 \sum_{\tilde{\lambda}_j < 1/(N_{\mathtt{eff}}\gamma)} \tilde{\lambda}_j^2}{N_{\mathtt{eff}}} \\ &\eqsim \frac{\min\left\{M,\ (N_{\mathtt{eff}}\gamma)^{1/a} + (N_{\mathtt{eff}}\gamma)^2 \cdot (N_{\mathtt{eff}}\gamma)^{(1-2a)/a}\right\}}{N_{\mathtt{eff}}} \\ &\eqsim \frac{\min\{M,\ (N_{\mathtt{eff}}\gamma)^{1/a}\}}{N_{\mathtt{eff}}}\end{aligned}$$

with probability at least $1-e^{-\Omega(M)}$ over the randomness of $\mathbf{S}$. Here the hidden constants may depend on $a$. □

**Lemma E.2** (Matching bounds on Var under logarithmic power-law spectrum)**.** *Under Assumption 4,* Var *defined in Eq.* (6) *satisfies*

$$\mathsf{Var} \eqsim \frac{\min\{M,\ \bar{\mathsf{k}}\}}{N_{\mathtt{eff}}} \eqsim \frac{\min\{M,\ (N_{\mathtt{eff}}\gamma)/\log^a(N_{\mathtt{eff}}\gamma)\}}{N_{\mathtt{eff}}}$$

*with probability at least* $1-e^{-\Omega(M)}$ *over the randomness of* $\mathbf{S}$*, where* $\bar{\mathsf{k}} := \inf\{k : k\log^a k \ge (N_{\mathtt{eff}}\gamma)\}$ *and* $\eqsim$ *hides constants that only depend on* $a$*.*

*Proof of Lemma E.2.* Define $\mathsf{k}^* = \inf\{k : k\log k \ge M\}$ and let $\tilde{D} := \#\{\tilde{\lambda}_j \ge 1/(N_{\mathtt{eff}}\gamma)\} + (N_{\mathtt{eff}}\gamma)^2 \sum_{\tilde{\lambda}_j < 1/(N_{\mathtt{eff}}\gamma)} \tilde{\lambda}_j^2$. By the definition of Var in Eq. (6) and Lemma G.6, we have

$$\begin{aligned}\mathsf{Var} &= \frac{\#\{\tilde{\lambda}_j \ge 1/(N_{\mathtt{eff}}\gamma)\} + (N_{\mathtt{eff}}\gamma)^2 \sum_{\tilde{\lambda}_j < 1/(N_{\mathtt{eff}}\gamma)} \tilde{\lambda}_j^2}{N_{\mathtt{eff}}} \\ &= \frac{\tilde{D}}{N_{\mathtt{eff}}} \eqsim \frac{\min\{M, \bar{\mathsf{k}}\}}{N_{\mathtt{eff}}}\end{aligned}$$

with probability at least $1-e^{-\Omega(M)}$ over the randomness of $\mathbf{S}$, where the second line follows from

$$\begin{aligned}&\tilde{D} \gtrsim \#\{\tilde{\lambda}_j \ge 1/(N_{\mathtt{eff}}\gamma)\} \eqsim \frac{N_{\mathtt{eff}}\gamma}{\log^a(N_{\mathtt{eff}}\gamma)}, \qquad \text{and} \\ &\tilde{D} \lesssim \frac{N_{\mathtt{eff}}\gamma}{\log^a(N_{\mathtt{eff}}\gamma)} + \frac{(N_{\mathtt{eff}}\gamma)^2}{\log^{2a}(N_{\mathtt{eff}}\gamma)} \cdot \sum_{j:\tilde{\lambda}_j < 1/(N_{\mathtt{eff}}\gamma)} \frac{1}{j^2} \lesssim \frac{N_{\mathtt{eff}}\gamma}{\log^a(N_{\mathtt{eff}}\gamma)}\end{aligned}$$

when $\bar{\mathsf{k}} \lesssim M$. □

## F Expected risk of the average of (SGD) iterates

In this section, we study the expected risk of the average of (SGD) iterates. Namely, we consider a fixed stepsize (SGD) procedure where $\gamma_t = \gamma$ and define $\bar{\mathbf{v}}_N := \sum_{i=0}^{N-1} \mathbf{v}_i/N$. Our goal is to derive matching upper and lower bounds $\mathcal{R}(\bar{\mathbf{v}}_N)$ in terms of the sample size $N$ and model size $M$.

Compared with the last iterate of (SGD) with geometrically decaying stepsizes, we show that the average of (SGD) iterates with a fixed stepsize achieves a better risk, in the sense that the effective sample size $N_{\texttt{eff}}$ is replaced by $N$ in the bounds (c.f. Theorem 4.1). This may give improvement up to logarithmic factors.

We start with invoking the following result in [48].

**Theorem F.1** (A variant of Theorem 2.1 and 2.2 in [48]). *Suppose Assumption 1 hold. Consider an $M$-dimensional sketched predictor trained by fixed stepsize (SGD) with $N$ samples. Let $\bar{\mathbf{v}}_N := \sum_{i=0}^{N-1} \mathbf{v}_i/N$ be the average of the iterations of SGD. Assume $\mathbf{v}_0 = \mathbf{0}$ and $\sigma^2 \gtrsim 1$. Conditional on $\mathbf{S}$ and suppose the stepsize $\gamma < 1/(c\,\mathrm{tr}(\mathbf{S}\mathbf{H}\mathbf{S}^\top))$ for some constant $c > 0$, then there exist* Approx, Bias, Var *such that*

$$\mathbb{E}\mathcal{R}_M(\bar{\mathbf{v}}_N) - \sigma^2 \eqsim \mathbb{E}_{\mathbf{w}^*}\mathsf{Approx} + \mathsf{Bias} + \sigma^2\mathsf{Var},$$

*where the expectation of $\mathcal{R}_M$ is over $\mathbf{w}^*$ and $(\mathbf{x}_i, y_i)_{i=1}^N$ and*

$$\begin{aligned}\mathsf{Approx} &:= \mathbb{E}\xi^2 \\ &= \left\| \left(\mathbf{I} - \mathbf{H}^{\frac{1}{2}}\mathbf{S}^\top \left(\mathbf{S}\mathbf{H}\mathbf{S}^\top\right)^{-1}\mathbf{S}\mathbf{H}^{\frac{1}{2}}\right)\mathbf{H}^{\frac{1}{2}}\mathbf{w}^* \right\|^2,\end{aligned}$$

$$\mathbb{E}_{\mathbf{w}^*}(T_1 + T_3) \lesssim \mathsf{Bias} \lesssim \mathbb{E}_{\mathbf{w}^*}(T_2 + T_4),$$

$$\mathsf{Var} \eqsim \frac{D_{\mathrm{eff},N}}{N},$$

*and*

$$T_1 := \frac{1}{\gamma^2 N^2}\,\mathrm{tr}\left(\left(\mathbf{I} - (\mathbf{I} - \gamma\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{N/4}\right)^2 (\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1}\mathbf{B}_0\right), \tag{27a}$$

$$T_2 := \frac{1}{\gamma^2 N^2}\,\mathrm{tr}\left(\left(\mathbf{I} - (\mathbf{I} - \gamma\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{N}\right)^2 (\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1}\mathbf{B}_0\right), \tag{27b}$$

$$T_3 := \frac{1}{\gamma N^2}\,\mathrm{tr}\left(\left(\mathbf{I} - (\mathbf{I} - \gamma\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{N/4}\right)\mathbf{B}_0\right) \cdot \mathrm{tr}\left(\left(\mathbf{I} - (\mathbf{I} - \gamma\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{N/4}\right)^2\right), \tag{27c}$$

$$T_4 := \frac{1}{\gamma N}\,\mathrm{tr}\left(\mathbf{B}_0 - (\mathbf{I} - \gamma\mathbf{S}\mathbf{H}\mathbf{S}^\top)^N\mathbf{B}_0(\mathbf{I} - \gamma\mathbf{S}\mathbf{H}\mathbf{S}^\top)^N\right) \cdot \frac{D_{\mathrm{eff},N}}{N}, \tag{27d}$$

$$\mathbf{B}_0 := \mathbf{v}^*\mathbf{v}^{*\top}, \tag{27e}$$

$$D_{\mathrm{eff},N} := \#\{\tilde{\lambda}_j \geq 1/(N\gamma)\} + (N\gamma)^2 \sum_{\tilde{\lambda}_j < 1/(N\gamma)} \tilde{\lambda}_j^2, \tag{27f}$$

*where $(\tilde{\lambda}_j)_{j=1}^M$ are eigenvalue of $\mathbf{S}\mathbf{H}\mathbf{S}^\top$.*

See Section F.2.1 for the proof.

For $T_i (i = 1, 2, 3, 4)$, we also have the following upper (and lower) bounds.

**Lemma F.2** (Lower bound on $T_1$). *Under the assumptions and notations in Theorem F.1, we have*

$$\mathbb{E}_{\mathbf{w}^*} T_1 \gtrsim \sum_{i:\tilde{\lambda}_i < 1/(\gamma N)} \frac{\mu_i(\mathbf{S}\mathbf{H}^2\mathbf{S}^\top)}{\mu_i(\mathbf{S}\mathbf{H}\mathbf{S}^\top)}$$

*almost surely, where $(\tilde{\lambda}_i)_{i=1}^N$ are eigenvalues of $\mathbf{S}\mathbf{H}\mathbf{S}^\top$ in non-increasing order.*

See the proof in Section F.2.2.

**Lemma F.3** (Upper bound on $T_2$). *Under the assumptions and notations in Theorem F.1, for any $k \leq M/3$ such that $r(\mathbf{H}) \geq k + M$, we have with probability at least $1 - e^{-\Omega(M)}$ that*

$$T_2 \lesssim \frac{1}{N\gamma}\left[\frac{\mu_{M/2}(\mathbf{A}_k)}{\mu_M(\mathbf{A}_k)}\right]^2 \cdot \|\mathbf{w}^*_{0:k}\|^2 + \|\mathbf{w}^*_{k:\infty}\|^2_{\mathbf{H}_{k:\infty}},$$

*where $\mathbf{A}_k := \mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{S}_{k:\infty}^\top$.*

See the proof in Section F.2.3.

**Lemma F.4** (Lower bound on $T_3$)**.** *Under the assumptions and notations in Theorem F.1, we have*

$$\mathbb{E}_{\mathbf{w}^*} T_3 \gtrsim \frac{D_{\mathrm{eff},N}}{N} \cdot \sum_{i:\tilde{\lambda}_i < 1/(\gamma N)} \frac{\mu_i(\mathbf{S}\mathbf{H}^2\mathbf{S}^\top)}{\mu_i(\mathbf{S}\mathbf{H}\mathbf{S}^\top)}$$

*almost surely, where $(\tilde{\lambda}_i)_{i=1}^M$ are eigenvalues of $\mathbf{S}\mathbf{H}\mathbf{S}^\top$ in non-increasing order.*

See the proof in Section F.2.4.

**Lemma F.5** (Upper bound on $T_4$)**.** *Under the assumptions and notations in Theorem F.1 and assume $r(\mathbf{H}) \geq M$, we have*

$$T_4 \lesssim \|\mathbf{w}^*\|_{\mathbf{H}}^2 \cdot \frac{D_{\mathrm{eff},N}}{N}$$

*almost surely, where $\mathbf{A}_k := \mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{S}_{k:\infty}^\top$.*

See the proof in Section F.2.5.

With these results at hand, we are ready to derive upper and lower bounds for the risk of the average of (SGD) iterates.

### F.1 Matching bounds for the average of (SGD) iterates under power-law spectrum

In this section, we derive upper and lower bounds for the expected risk under the power-law spectrum. Our main result (Theorem F.6) follows directly from Theorem F.1 and the bounds on $T_i (i = 1, 2, 3, 4)$ in Lemmas F.2 to F.5.

**Theorem F.6** (Scaling law for average iterates of SGD)**.** *Suppose Assumption 1 and 2 hold and $\sigma^2 \lesssim 1$. Then there exists some $a$-dependent constant $c > 0$ such that when $\gamma \leq c$, with probability at least $1 - e^{-\Omega(M)}$ over the randomness of the sketch matrix $\mathbf{S}$, we have*

$$\mathbb{E}\mathcal{R}_M(\bar{\mathbf{v}}_N) = \sigma^2 + \Theta\big(M^{1-a}\big) + \Theta\big((N\gamma)^{1/a-1}\big),$$

*where the expectation is over the randomness of $\mathbf{w}^*$ and $(\mathbf{x}_i, y_i)_{i=1}^N$, and $\Theta(\cdot)$ hides constants that may depend on $a$.*

See the proof in Section F.2.6.

Compared with Theorem 4.1, Theorem F.6 suggests that the average of (SGD) achieves a smaller risk in the sketched linear model—the $(N_{\mathtt{eff}}\gamma)^{1/a}$ is replaced by $(N\gamma)^{1/a}$ in the bound for the bias term. This is intuitive since the sum of stepsizes $\sum_t \gamma_t \eqsim N_{\mathtt{eff}}\gamma$ for the geometrically decaying stepsize scheduler while $\sum_t \gamma_t \eqsim N\gamma$ for the fixed stepsize scheduler.

We also verify the observations in Theorem F.6 via simulations. We adopt the same model and setup as in Section 5 but use the average of iterates of fixed stepsize (SGD) (denoted by $\bar{\mathbf{v}}_N$) as the predictor. From Figure 3 and 4 we see that the expected risk $\mathbb{E}\mathcal{R}(\bar{\mathbf{v}}_N)$ also scales following a power-law relation in both sample size $N$ and model size $M$. Moreover, the fitted exponents match our theoretical predictions in Theorem F.6.

### F.2 Proofs

#### F.2.1 Proof of Theorem F.1

Similar to the proof of Theorem A.4, we have the decomposition

$$\mathcal{R}(\bar{\mathbf{v}}_N) = \sigma^2 + \mathsf{Approx} + \|\bar{\mathbf{v}}_N - \mathbf{v}^*\|_{\mathbf{S}\mathbf{H}\mathbf{S}^\top}^2.$$

Note that $(\mathbf{v}_t)_{t=1}^N$ can also be viewed as the SGD iterates on the model $y = \langle \mathbf{S}\mathbf{x}, \mathbf{v}^* \rangle + \xi + \epsilon$, where the noise satisfies

$$\mathbb{E}(\xi + \epsilon)^2 = \mathcal{R}(\mathbf{v}^*) = \mathbb{E}\xi^2 + \sigma^2.$$

Therefore, the upper and lower bounds on $\mathsf{Bias}$, $\mathsf{Var}$ follow directly from the proof of Theorem 2.1, 2.2 and related lemmas (Lemma B.6, B.11, C.3, C.5) in [48].

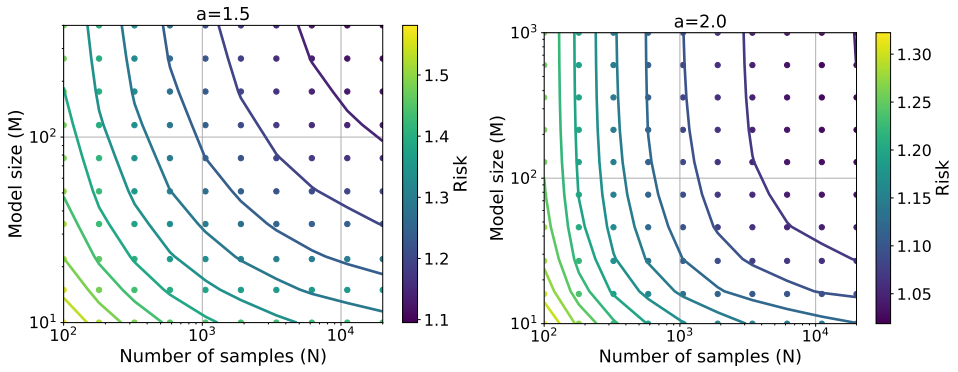
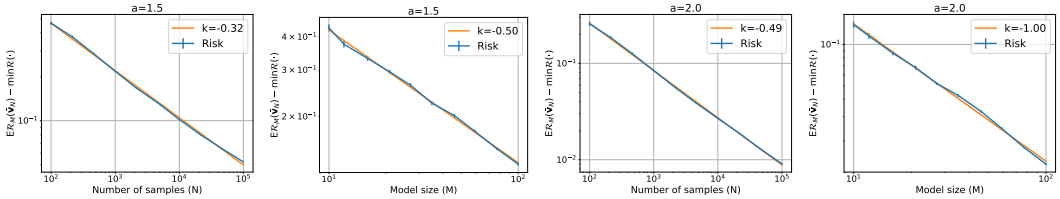



Figure 3: The expected risk (Risk) of the average of iterates of (SGD) versus the sample size $N$ and the model size $M$ for different power-law degrees $a$. The expected risk is computed by averaging over 1000 independent samples of $(\mathbf{w}^*, \mathbf{S})$. We fit the expected risk using the formula Risk $\sim \sigma^2 + c_1/M^{a_1} + c_2/N^{a_2}$ via minimizing the Huber loss as in [21]. Parameters: $\sigma = 1, \gamma = 0.1$. Left: For $a = 1.5$, $d = 20000$, the fitted exponents are $(a_1, a_2) = (0.59, 0.33) \approx (0.5, 0.33)$. Right: For $a = 2$, $d = 2000$, the fitted exponents are $(a_1, a_2) = (1.09, 0.49) \approx (1.0, 0.5)$. Note that the values of $(a_1, a_2)$ are close to our theoretical predictions $(a - 1, 1 - 1/a)$ in both cases.




(a) $a = 1.5$ (b) $a = 1.5$ (c) $a = 2$ (d) $a = 2$

Figure 4: The expected risk of the average of iterates of (SGD) minus the irreducible risk versus the effective sample size and model size. Parameters $\sigma = 1, \gamma = 0.1$. (a), (b): $a = 1.5, d = 10000$; (c), (d): $a = 2, d = 1000$. The errobars denotes the $\pm 1$ standard deviation of estimating the expected risk using 100 independent samples of $(\mathbf{w}^*, \mathbf{S})$. We use linear functions to fit the expected risk under log-log scale and report the slope of the fitted lines (denoted by $k$).

### F.2.2 Proof of Lemma F.2

Let $f_1(A) := (\mathbf{I} - (\mathbf{I} - \gamma A)^{N/4})^2 A^{-1}/\gamma^2/N^2$ for any positive definite matrix $A \in \mathbb{R}^{M \times M}$. Since $\gamma \le 1/(c\,\mathrm{tr}(\mathbf{S}\mathbf{H}\mathbf{S}^\top))$, we have $f_1(\mathbf{S}\mathbf{H}\mathbf{S}^\top) \succeq \mathbf{0}$. By definition of $T_1$ and recalling $\mathbf{v}^* = (\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1}\mathbf{S}\mathbf{H}\mathbf{w}^*$, we have with probability at least $1 - e^{-\Omega(M)}$ that

$$\begin{aligned}\mathbb{E}_{\mathbf{w}^*} T_1 &= \mathbb{E}_{\mathbf{w}^*}[{\mathbf{w}^*}^\top \mathbf{H}\mathbf{S}^\top (\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1} f_1(\mathbf{S}\mathbf{H}\mathbf{S}^\top)(\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1}\mathbf{S}\mathbf{H}\mathbf{w}^*] \\ &= \mathrm{tr}\left([(\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1} f_1(\mathbf{S}\mathbf{H}\mathbf{S}^\top)(\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1}](\mathbf{S}\mathbf{H}^2\mathbf{S}^\top)\right).\end{aligned}$$

Following the proof of Lemma D.2 (by Von Neumann's trace inequality), we have

$$\begin{aligned}\mathbb{E}_{\mathbf{w}^*} T_1 &\ge \sum_{i=1}^{M} \frac{\mu_i(\mathbf{S}\mathbf{H}^2\mathbf{S}^\top)}{\mu_i\left((\mathbf{S}\mathbf{H}\mathbf{S}^\top)^2 f_1(\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1}\right)} \\ &\ge \sum_{i:\tilde{\lambda}_i < 1/(\gamma N)} \frac{\mu_i(\mathbf{S}\mathbf{H}^2\mathbf{S}^\top)}{\mu_i\left((\mathbf{S}\mathbf{H}\mathbf{S}^\top)^2 f_1(\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1}\right)} \\ &\gtrsim \sum_{i:\tilde{\lambda}_i < 1/(\gamma N)} \frac{\mu_i(\mathbf{S}\mathbf{H}^2\mathbf{S}^\top)}{\mu_i(\mathbf{S}\mathbf{H}\mathbf{S}^\top)},\end{aligned}$$

where the third inequality is due to

$$\lambda/f_1(\lambda) \lesssim \frac{\lambda^2\gamma^2N^2}{(1-(1-\gamma\lambda)^{N/4})^2} \lesssim \frac{N^2}{(\sum_{i=0}^{N/4-1}(1-\gamma\lambda)^i)^2} \lesssim \frac{1}{(1-\gamma\lambda)^{2N}} \lesssim 1$$

when $\lambda < 1/(N\gamma)$.

### F.2.3 Proof of Lemma F.3

By definition of $T_2$, the fact that $1-x^N = (1-x)\sum_{i=0}^{N-1} x^i$, and recalling $\mathbf{v}^* = (\mathbf{SHS}^\top)^{-1}\mathbf{SHw}^*$, we have

$$\begin{aligned}
T_2 &= {\mathbf{w}^*}^\top \mathbf{H}\mathbf{S}^\top f_2(\mathbf{SHS}^\top)\mathbf{SHw}^*, \\
&\le 2[\underbrace{{\mathbf{w}^*_{0:k}}^\top \mathbf{H}_{0:k}\mathbf{S}_{0:k}^\top f_2(\mathbf{SHS}^\top)\mathbf{S}_{0:k}\mathbf{H}_{0:k}\mathbf{w}^*_{0:k}}_{T_{21}} + \underbrace{{\mathbf{w}^*_{k:\infty}}^\top \mathbf{H}_{k:\infty}\mathbf{S}_{k:\infty}^\top f_2(\mathbf{SHS}^\top)\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{w}^*_{k:\infty}}_{T_{22}}],
\end{aligned}$$

where $f_2(A) := [\sum_{i=0}^{N-1}(\mathbf{I}-\gamma A)^i]^2/A/N^2$ for any symmetric matrix $A \in \mathbb{R}^{M\times M}$. Moreover, we have

$$\begin{aligned}
T_{21} &= {\mathbf{w}^*_{0:k}}^\top \mathbf{H}_{0:k}\mathbf{S}_{0:k}^\top f_2(\mathbf{SHS}^\top)\mathbf{S}_{0:k}\mathbf{H}_{0:k}\mathbf{w}^*_{0:k} \\
&\le \|f_2(\mathbf{SHS}^\top)(\mathbf{SHS}^\top)^2\| \cdot \|(\mathbf{SHS}^\top)^{-1}\mathbf{S}_{0:k}\mathbf{H}_{0:k}\mathbf{w}^*_{0:k}\|^2.
\end{aligned}$$

Using the assumption on the stepsize that $\gamma \le 1/(c\,\mathrm{tr}(\mathbf{SHS}^\top))$, we have

$$\begin{aligned}
\|f_2(\mathbf{SHS}^\top)(\mathbf{SHS}^\top)^2\| &\le \max_{\lambda\in[0,1/\gamma]} \frac{1}{N^2}\Big[\sum_{i=0}^{N-1}(1-\gamma\lambda)^i\Big]^2\lambda \\
&= \max_{\lambda\in[0,1/\gamma]} \frac{1}{N^2\gamma}\Big[\sum_{i=0}^{N-1}(1-\gamma\lambda)^i\Big]\cdot(1-(1-\gamma\lambda)^N) \\
&\le \frac{1}{N^2\gamma}\cdot N\cdot 1 = \frac{1}{N\gamma}. \qquad (28)
\end{aligned}$$

Combining Eq. (28) with Eq. (23) in the proof of Lemma D.1 (note that we assume $k \le M/3$), we obtain

$$T_{21} \le c\frac{1}{N\gamma}\Big[\frac{\mu_{M/2}(\mathbf{A}_k)}{\mu_M(\mathbf{A}_k)}\Big]^2 \cdot \|\mathbf{w}^*_{0:k}\|^2$$

for some constant $c > 0$ with probability at least $1 - e^{-\Omega(M)}$. For $T_{22}$, we have

$$\begin{aligned}
T_{22} &= {\mathbf{w}^*_{k:\infty}}^\top \mathbf{H}_{k:\infty}\mathbf{S}_{k:\infty}^\top f_2(\mathbf{SHS}^\top)\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{w}^*_{k:\infty} \\
&\le \|f_2(\mathbf{SHS}^\top)\mathbf{SHS}^\top\| \cdot \|(\mathbf{SHS}^\top)^{-1/2}(\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}^{1/2})\mathbf{H}_{k:\infty}^{1/2}\mathbf{w}^*_{k:\infty}\|^2 \\
&\le \|f_2(\mathbf{SHS}^\top)\mathbf{SHS}^\top\| \cdot \|(\mathbf{SHS}^\top)^{-1/2}\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}^{1/2}\|^2 \cdot \|\mathbf{w}^*_{k:\infty}\|^2_{\mathbf{H}_{k:\infty}}.
\end{aligned}$$

Since $\|f_2(\mathbf{SHS}^\top)\mathbf{SHS}^\top\| = \|[\sum_{i=0}^{N-1}(\mathbf{I}-\gamma\mathbf{SHS}^\top)^i]^2/N^2\| \le 1$ by the assumption $\gamma \le 1/(c\,\mathrm{tr}(\mathbf{SHS}^\top))$, and

$$\begin{aligned}
\|(\mathbf{SHS}^\top)^{-1/2}\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}^{1/2}\|^2 &= \|\mathbf{H}_{k:\infty}^{1/2}\mathbf{S}_{k:\infty}^\top(\mathbf{SHS}^\top)^{-1}\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}^{1/2}\| \\
&= \|\mathbf{H}_{k:\infty}^{1/2}\mathbf{S}_{k:\infty}^\top(\mathbf{S}_{0:k}\mathbf{H}_{0:k}\mathbf{S}_{0:k}^\top + \mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{S}_{k:\infty}^\top)^{-1}\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}^{1/2}\| \\
&\le \|\mathbf{H}_{k:\infty}^{1/2}\mathbf{S}_{k:\infty}^\top(\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{S}_{k:\infty}^\top)^{-1}\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}^{1/2}\| = 1,
\end{aligned}$$

it follows that $T_{22} \le \|\mathbf{w}^*_{k:\infty}\|^2_{\mathbf{H}_{k:\infty}}$. Combining the bounds on $T_{21}, T_{22}$ completes the proof.

### F.2.4 Proof of Lemma F.4

Let $f_3(A) := (\mathbf{I}-(\mathbf{I}-\gamma\mathbf{A})^{N/4})/\gamma/N^2$ for any positive definite matrix $A \in \mathbb{R}^{M\times M}$. Following the same arguments as in the proof of Lemma F.2, we have $f_3(\mathbf{SHS}^\top) \succeq \mathbf{0}$ and

$$\mathbb{E}_{\mathbf{w}^*}\Big[\frac{1}{\gamma N^2}\,\mathrm{tr}\Big(\Big(\mathbf{I}-(\mathbf{I}-\gamma\mathbf{SHS}^\top)^{N/4}\Big)\mathbf{B}_0\Big)\Big]$$

$$= \mathbb{E}_{\mathbf{w}^*}[{\mathbf{w}^*}^\top \mathbf{H}\mathbf{S}^\top (\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1} f_3(\mathbf{S}\mathbf{H}\mathbf{S}^\top)(\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1}\mathbf{S}\mathbf{H}\mathbf{w}^*]$$
$$= \mathrm{tr}((\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1} f_3(\mathbf{S}\mathbf{H}\mathbf{S}^\top)(\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1}\mathbf{S}\mathbf{H}^2\mathbf{S}^\top).$$

Moreover,

$$\begin{aligned}
\mathbb{E}_{\mathbf{w}^*} T_1 &\geq \sum_{i=1}^{M} \frac{\mu_i(\mathbf{S}\mathbf{H}^2\mathbf{S}^\top)}{\mu_i\Big((\mathbf{S}\mathbf{H}\mathbf{S}^\top)^2 f_3(\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1}\Big)} \\
&\geq \sum_{i:\tilde{\lambda}_i < 1/(\gamma N)} \frac{\mu_i(\mathbf{S}\mathbf{H}^2\mathbf{S}^\top)}{\mu_i\Big((\mathbf{S}\mathbf{H}\mathbf{S}^\top)^2 f_3(\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1}\Big)} \\
&\gtrsim \frac{1}{N} \sum_{i:\tilde{\lambda}_i < 1/(\gamma N)} \frac{\mu_i(\mathbf{S}\mathbf{H}^2\mathbf{S}^\top)}{\mu_i(\mathbf{S}\mathbf{H}\mathbf{S}^\top)},
\end{aligned} \tag{29}$$

where the third inequality is due to

$$\lambda / f_3(\lambda) \lesssim \frac{\lambda\gamma N^2}{1-(1-\gamma\lambda)^{N/4}} \lesssim \frac{N^2}{\sum_{i=0}^{N/4-1}(1-\gamma\lambda)^i} \lesssim \frac{N}{(1-\gamma\lambda)^N} \lesssim N$$

when $\lambda < 1/(N\gamma)$. Note that

$$1-(1-\gamma\tilde{\lambda}_i)^{\frac{N}{4}} \geq \begin{cases} 1-(1-\frac{1}{N})^{\frac{N}{4}} \geq 1-e^{-\frac{1}{4}} \geq \frac{1}{5}, & \tilde{\lambda}_i \geq \frac{1}{\gamma N}, \\ \frac{N}{4}\cdot\gamma\tilde{\lambda}_i - \frac{N(N-4)}{32}\cdot\gamma^2\tilde{\lambda}_i^2 \geq \frac{N}{5}\cdot\gamma\tilde{\lambda}_i, & \tilde{\lambda}_i < \frac{1}{\gamma N}, \end{cases} \quad \geq \frac{1}{5}\min\{N\gamma\tilde{\lambda}_i, 1\}. \tag{30}$$

We thus have

$$\begin{aligned}
\mathrm{tr}\left(\left(\mathbf{I}-(\mathbf{I}-\gamma\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{N/4}\right)^2\right) &= \sum_{i=1}^{M}[1-(1-\gamma\tilde{\lambda}_i)^{\frac{N}{4}}]^2 \gtrsim \sum_{i=1}^{M}\min\{(N\gamma\tilde{\lambda}_i)^2, 1\} \\
&= \#\{\tilde{\lambda}_i \geq \frac{1}{N\gamma}\} + N^2\gamma^2 \sum_{\tilde{\lambda}_i < 1/(N\gamma)} \tilde{\lambda}_i^2 = D_{\mathrm{eff},N}.
\end{aligned} \tag{31}$$

Combining Eq. (31) and (29) completes the proof.

#### F.2.5 Proof of Lemma F.5

Substituting $\mathbf{v}^* = (\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1}\mathbf{S}\mathbf{H}\mathbf{w}^*$ in the expression of $T_4$ and noting $\mathbf{v}_0 = \mathbf{0}$, we have

$$\begin{aligned}
T_4 &= \frac{1}{\gamma N}\mathrm{tr}\left(\mathbf{B}_0 - (\mathbf{I}-\gamma\mathbf{S}\mathbf{H}\mathbf{S}^\top)^N\mathbf{B}_0(\mathbf{I}-\gamma\mathbf{S}\mathbf{H}\mathbf{S}^\top)^N\right)\cdot\frac{D_{\mathrm{eff},N}}{N} \\
&= \frac{1}{\gamma N}\mathrm{tr}\left({\mathbf{w}^*}^\top\mathbf{H}\mathbf{S}^\top(\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1}\left[\mathbf{I}-(\mathbf{I}-\gamma\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{2N}\right](\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1}\mathbf{S}\mathbf{H}\mathbf{w}^*\right)\cdot\frac{D_{\mathrm{eff},N}}{N} \\
&=: \mathrm{tr}\left({\mathbf{w}^*}^\top\mathbf{H}\mathbf{S}^\top f_4(\mathbf{S}\mathbf{H}\mathbf{S}^\top)\mathbf{S}\mathbf{H}\mathbf{w}^*\right)\cdot\frac{D_{\mathrm{eff},N}}{N},
\end{aligned} \tag{32}$$

where $f_4(A) := A^{-1}\left[\mathbf{I}-(\mathbf{I}-\gamma A)^{2N}\right]A^{-1}/(N\gamma)$ for any symmetric matrix $A \in \mathbb{R}^{M\times M}$. Moreover,

$$\begin{aligned}
&\mathrm{tr}\left({\mathbf{w}^*}^\top\mathbf{H}\mathbf{S}^\top f_4(\mathbf{S}\mathbf{H}\mathbf{S}^\top)\mathbf{S}\mathbf{H}\mathbf{w}^*\right) \\
&\leq \|f_4(\mathbf{S}\mathbf{H}\mathbf{S}^\top)\mathbf{S}\mathbf{H}\mathbf{S}^\top\|\cdot\|(\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1/2}\mathbf{S}\mathbf{H}^{1/2}\|^2\cdot\|\mathbf{w}_*\|_{\mathbf{H}}^2 \\
&\leq \|f_4(\mathbf{S}\mathbf{H}\mathbf{S}^\top)\mathbf{S}\mathbf{H}\mathbf{S}^\top\|\cdot\|\mathbf{w}_*\|_{\mathbf{H}}^2.
\end{aligned}$$

Since

$$\|f_4(\mathbf{S}\mathbf{H}\mathbf{S}^\top)\mathbf{S}\mathbf{H}\mathbf{S}^\top\| = \frac{1}{N}\Big\|\sum_{i=0}^{2N-1}(\mathbf{I}-\gamma\mathbf{S}\mathbf{H}\mathbf{S}^\top)^i\Big\| \leq 2$$

by our assumption on the stepsize, it follows that

$$\mathrm{tr}\left({\mathbf{w}^*}^\top\mathbf{H}\mathbf{S}^\top f_4(\mathbf{S}\mathbf{H}\mathbf{S}^\top)\mathbf{S}\mathbf{H}\mathbf{w}^*\right) \lesssim \|\mathbf{w}_*\|_{\mathbf{H}}^2. \tag{33}$$

Combining Eq. (32) and (33) we find

$$T_4 \lesssim \|\mathbf{w}_*\|_{\mathbf{H}}^2\cdot\frac{D_{\mathrm{eff},N}}{N}.$$

### F.2.6 Proof of Theorem F.6

First, by Lemma G.4 we have $1/\operatorname{tr}(\mathbf{SHS}^\top) \gtrsim c_1$ for some $a$-dependent $c_1 > 0$ with probability at least $1 - e^{-\Omega(M)}$. Therefore we may choose $c$ sufficiently small so that $\gamma \le c$ implies $\gamma \lesssim 1/\operatorname{tr}(\mathbf{SHS}^\top)$ with probability at least $1 - e^{-\Omega(M)}$. Now, suppose we have $\gamma \lesssim 1/\operatorname{tr}(\mathbf{SHS}^\top)$. Following the notations in Theorem F.1, we claim the following bounds on Approx, Bias, Var:

$$\mathbb{E}\mathsf{Approx} \eqsim M^{1-a} \tag{34a}$$

$$\mathsf{Var} \eqsim \min\left\{M,\ (N\gamma)^{1/a}\right\}/N. \tag{34b}$$

$$\mathsf{Bias} \lesssim \max\left\{M^{1-a},\ (N\gamma)^{1/a-1}\right\}, \tag{34c}$$

$$\mathsf{Bias} \gtrsim (N\gamma)^{1/a-1} \text{ when } (N\gamma)^{1/a} \le M/c \text{ for some constant } c > 0, \tag{34d}$$

with probability at least $1 - e^{-\Omega(M)}$. Putting the bounds together yields Theorem F.6.

**Proof of claim** (34a) Note that our definition of Approx in Thereom F.1 is the same as that in Eq. (4) (and 7). Therefore the claim follows immediately from Lemma C.4.

**Proof of claim** (34b) This follows from the proof of Lemma E.1 with $N_{\mathtt{eff}}$ replaced by $N$.

**Proof of claim** (34c) By Theorem F.1, Lemma F.3 and F.5, we have

$$\begin{aligned}\mathsf{Bias} &\lesssim \mathbb{E}_{\mathbf{w}^*}\frac{\|\mathbf{w}^*_{0:k_2}\|_2^2}{N\gamma}\cdot\left[\frac{\mu_{M/2}(\mathbf{S}_{k_2:\infty}\mathbf{H}_{k_2:\infty}\mathbf{S}_{k_2:\infty}^\top)}{\mu_M(\mathbf{S}_{k_2:\infty}\mathbf{H}_{k_2:\infty}\mathbf{S}_{k_2:\infty}^\top)}\right]^2 + \mathbb{E}_{\mathbf{w}^*}\|\mathbf{w}^*_{k_2:\infty}\|^2_{\mathbf{H}_{k_2:\infty}} + \sigma^2\frac{D_{\text{eff},N}}{N},\\ &\lesssim \frac{k_2}{N\gamma}\left[\frac{\mu_{M/2}(\mathbf{S}_{k_2:\infty}\mathbf{H}_{k_2:\infty}\mathbf{S}_{k_2:\infty}^\top)}{\mu_M(\mathbf{S}_{k_2:\infty}\mathbf{H}_{k_2:\infty}\mathbf{S}_{k_2:\infty}^\top)}\right]^2 + k_2^{1-a} + \frac{D_{\text{eff},N}}{N}\end{aligned}$$

with probability at least $1 - e^{-\Omega(M)}$ for any $k_2 \le M/3$. Choosing $k_2 = \min\{M/3, (N\gamma)^{1/a}\}$ and using Lemma G.4 and claim (34b), we obtain

$$\begin{aligned}\mathsf{Bias} &\lesssim \max\left\{M^{1-a},\ (N\gamma)^{1/a-1}\right\} + \frac{\min\left\{M,\ (N\gamma)^{1/a}\right\}}{N} \lesssim \max\left\{M^{1-a},\ (N\gamma)^{1/a-1}\right\} + (N\gamma)^{1/a-1}\\ &\lesssim \max\left\{M^{1-a},\ (N\gamma)^{1/a-1}\right\}\end{aligned}$$

with probability at least $1 - e^{-\Omega(M)}$.

**Proof of claim** (34d) By Theorem F.1 and Lemma F.2, we have

$$\mathbb{E}_{\mathbf{w}^*}\mathsf{Bias} \gtrsim \sum_{i:\tilde\lambda_i < 1/(\gamma N)} \frac{\mu_i(\mathbf{SH}^2\mathbf{S}^\top)}{\mu_i(\mathbf{SHS}^\top)}.$$

When $(N\gamma)^{1/a} \le M/c$ for some large constant $c > 0$, we have from Lemma G.4 that

$$\mathbb{E}_{\mathbf{w}^*}\mathsf{Bias} \gtrsim \sum_{i:\tilde\lambda_i < 1/(\gamma N)} \frac{i^{-2a}}{i^{-a}} = \sum_{i:\tilde\lambda_i < 1/(\gamma N)} i^{-a} \gtrsim [(N\gamma)^{1/a}]^{1-a} = (N\gamma)^{1/a-1}$$

with probability at least $1 - e^{-\Omega(M)}$.

# G Concentration lemmas

## G.1 General concentration results

**Lemma G.1.** *Suppose that* $\mathbf{S} \in \mathbb{R}^{M\times d}$ *is such that* [3]

$$\mathbf{S}_{ij} \sim \mathcal{N}(0, 1/M).$$

[3] We allow $d = \infty$.

*Let $(\lambda_i)_{i\ge 1}$ be the eigenvalues of $\mathbf{H}$ in non-increasing order. Let $(\tilde{\lambda}_i)_{i=1}^M$ be the eigenvalues of $\mathbf{SHS}^\top$ in non-increasing order. Then there exists a constant $c > 1$ such that for every $M \ge 0$ and every $0 \le k \le M$, with probability $\ge 1 - e^{-\Omega(M)}$, we have*

$$\text{for every } j \le M, \quad \left|\tilde{\lambda}_j - \left(\lambda_j + \frac{\sum_{i>k}\lambda_i}{M}\right)\right| \le c \cdot \left(\sqrt{\frac{k}{M}} \cdot \lambda_j + \lambda_{k+1} + \sqrt{\frac{\sum_{i>k}\lambda_i^2}{M}}\right).$$

*As a direct consequence, for $k \le M/c^2$, we have*

$$\text{for every } j \le M, \quad \left|\tilde{\lambda}_j - \left(\lambda_j + \frac{\sum_{i>k}\lambda_i}{M}\right)\right| \le \frac{1}{2} \cdot \left(\lambda_j + \frac{\sum_{i>k}\lambda_i}{M}\right) + c_1 \cdot \lambda_{k+1},$$

*where $c_1 = c + 2c^2$.*

*Proof of Lemma G.1.* We have the following decomposition motivated by Swartworth and Woodruff [38] (see their Section 3.4, Proof of Theorem 1).

$$\begin{aligned}\mathbf{SHS}^\top &= \mathbf{S}_{0:k}\mathbf{H}_{0:k}\mathbf{S}_{0:k}^\top + \mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{S}_{k:\infty}^\top \\ &= \mathbf{S}_{0:k}\mathbf{H}_{0:k}\mathbf{S}_{0:k}^\top + \frac{\sum_{i>k}\lambda_i}{M} \cdot \mathbf{I}_M + \mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{S}_{k:\infty}^\top - \frac{\sum_{i>k}\lambda_i}{M} \cdot \mathbf{I}_M.\end{aligned}$$

We remark that this decomposition idea has been implicitly used in Bartlett et al. [5] to control the eigenvalues of a Gram matrix. In fact, we will use techniques from Bartlett et al. [5] to obtain a sharper bound than that presented in Swartworth and Woodruff [38].

For the upper bound, we have

$$\begin{aligned}\mu_j(\mathbf{SHS}^\top) &\le \mu_j\left(\mathbf{S}_{0:k}\mathbf{H}_{0:k}\mathbf{S}_{0:k}^\top + \frac{\sum_{i>k}\lambda_i}{M} \cdot \mathbf{I}_M\right) + \left\|\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{S}_{k:\infty}^\top - \frac{\sum_{i>k}\lambda_i}{M} \cdot \mathbf{I}_M\right\|_2 \\ &= \mu_j(\mathbf{S}_{0:k}\mathbf{H}_{0:k}\mathbf{S}_{0:k}^\top) + \frac{\sum_{i>k}\lambda_i}{M} \cdot \mathbf{I}_M + \left\|\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{S}_{k:\infty}^\top - \frac{\sum_{i>k}\lambda_i}{M} \cdot \mathbf{I}_M\right\|_2 \\ &\le \mu_j(\mathbf{S}_{0:k}\mathbf{H}_{0:k}\mathbf{S}_{0:k}^\top) + \frac{\sum_{i>k}\lambda_i}{M} \cdot \mathbf{I}_M + c_1 \cdot \left(\lambda_{k+1} + \sqrt{\frac{\sum_{i>k}\lambda_i^2}{M}}\right),\end{aligned}$$

where the last inequality is by Lemma G.2. For $j \le k$, using Lemma G.3, we have

$$\mu_j(\mathbf{S}_{0:k}\mathbf{H}_{0:k}\mathbf{S}_{0:k}^\top) \le \lambda_j + c_2 \cdot \sqrt{\frac{k}{M}} \cdot \lambda_j.$$

For $k < j \le M$, we have

$$\mu_j(\mathbf{S}_{0:k}\mathbf{H}_{0:k}\mathbf{S}_{0:k}^\top) = 0 \le \lambda_j + c_2 \cdot \sqrt{\frac{k}{M}} \cdot \lambda_j.$$

Putting these together, we have the following for every $j = 1, \dots, M$:

$$\begin{aligned}\mu_j(\mathbf{SHS}^\top) &\le \mu_j(\mathbf{S}_{0:k}\mathbf{H}_{0:k}\mathbf{S}_{0:k}^\top) + \frac{\sum_{i>k}\lambda_i}{M} \cdot \mathbf{I}_M + c_1 \cdot \left(\lambda_{k+1} + \sqrt{\frac{\sum_{i>k}\lambda_i^2}{M}}\right) \\ &\le \lambda_j + \frac{\sum_{i>k}\lambda_i}{M} \cdot \mathbf{I}_M + c \cdot \left(\sqrt{\frac{k}{M}} \cdot \lambda_j + \lambda_{k+1} + \sqrt{\frac{\sum_{i>k}\lambda_i^2}{M}}\right).\end{aligned}$$

Similarly, we can show the lower bound. By the decomposition, we have

$$\begin{aligned}\mu_j(\mathbf{SHS}^\top) &\ge \mu_j\left(\mathbf{S}_{0:k}\mathbf{H}_{0:k}\mathbf{S}_{0:k}^\top + \frac{\sum_{i>k}\lambda_i}{M} \cdot \mathbf{I}_M\right) - \left\|\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{S}_{k:\infty}^\top - \frac{\sum_{i>k}\lambda_i}{M} \cdot \mathbf{I}_M\right\| \\ &= \mu_j(\mathbf{S}_{0:k}\mathbf{H}_{0:k}\mathbf{S}_{0:k}^\top) + \frac{\sum_{i>k}\lambda_i}{M} \cdot \mathbf{I}_M - \left\|\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{S}_{k:\infty}^\top - \frac{\sum_{i>k}\lambda_i}{M} \cdot \mathbf{I}_M\right\|\end{aligned}$$

$$\geq \mu_j\big(\mathbf{S}_{0:k}\mathbf{H}_{0:k}\mathbf{S}_{0:k}^\top\big) + \frac{\sum_{i>k}\lambda_i}{M}\cdot\mathbf{I}_M - c_1\cdot\left(\lambda_{k+1} + \sqrt{\frac{\sum_{i>k}\lambda_i^2}{M}}\right),$$

where the last inequality is by Lemma G.2. For $j \leq k$, using Lemma G.3, we have

$$\mu_j\big(\mathbf{S}_{0:k}\mathbf{H}_{0:k}\mathbf{S}_{0:k}^\top\big) \geq \lambda_j - c_2\cdot\sqrt{\frac{k}{M}}\cdot\lambda_j.$$

For $k < j \leq M$, we have

$$\mu_j\big(\mathbf{S}_{0:k}\mathbf{H}_{0:k}\mathbf{S}_{0:k}^\top\big) = 0 \geq \lambda_j - \lambda_{k+1} - c_2\cdot\sqrt{\frac{k}{M}}\cdot\lambda_j,$$

where the last inequality is due to $\lambda_j \leq \lambda_k$ for $j \geq k$. Putting these together, we have

$$\begin{aligned}\mu_j\big(\mathbf{S}\mathbf{H}\mathbf{S}^\top\big) &\geq \mu_j\big(\mathbf{S}_{0:k}\mathbf{H}_{0:k}\mathbf{S}_{0:k}^\top\big) + \frac{\sum_{i>k}\lambda_i}{M}\cdot\mathbf{I}_M - c_1\cdot\left(\lambda_{k+1} + \sqrt{\frac{\sum_{i>k}\lambda_i^2}{M}}\right)\\ &\geq \lambda_j + \frac{\sum_{i>k}\lambda_i}{M}\cdot\mathbf{I}_M - c\cdot\left(\sqrt{\frac{k}{M}}\cdot\lambda_j + \lambda_{k+1} + \sqrt{\frac{\sum_{i>k}\lambda_i^2}{M}}\right).\end{aligned}$$

So far, we have proved the first claim. To show the second claim, we simply apply

$$c\cdot\sqrt{\frac{k}{M}} \leq \frac{1}{2} \quad \text{for } k \leq M/c^2,$$

and

$$\begin{aligned}c\cdot\sqrt{\frac{\sum_{i>k}\lambda_i^2}{M}} &\leq c\cdot\sqrt{\frac{\sum_{i>k}\lambda_i}{M}\cdot\lambda_{k+1}}\\ &\leq \frac{1}{2}\cdot\frac{\sum_{i>k}\lambda_i}{M} + 2c^2\cdot\lambda_{k+1},\end{aligned}$$

in the first claim. □

**Lemma G.2** (Tail concentration, Lemma 26 in Bartlett et al. [5])**.** *For any $k \geq 1$, with probability at least $1 - e^{-\Omega(M)}$, we have*

$$\left\|\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{S}_{k:\infty}^\top - \frac{\sum_{i>k}\lambda_i}{M}\cdot\mathbf{I}_M\right\|_2 \lesssim \lambda_{k+1} + \sqrt{\frac{\sum_{i>k}\lambda_i^2}{M}}.$$

*Moreover, the minimum eigenvalue of $\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{S}_{k:\infty}^\top$ satisfies*

$$\mu_{\min}(\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{S}_{k:\infty}^\top) \gtrsim \lambda_{k+2M}$$

*with probability at least $1 - e^{-\Omega(M)}$.*

*Proof of Lemma G.2.* The first part of Lemma G.2 is a version of Lemma 26 in [5] (see their proof). We provide proof here for completeness.

We write $\mathbf{S} \in \mathbb{R}^{M\times p}$ as

$$\mathbf{S} = (\mathbf{s}_1 \quad \dots \quad \mathbf{s}_p), \quad \mathbf{s}_i \sim \mathcal{N}\left(0,\ \frac{1}{M}\cdot\mathbf{I}_M\right), \quad i \geq 1.$$

Since Gaussian distribution is rotational invariance, without loss of generality, we may assume

$$\mathbf{H} = \mathrm{diag}\{\lambda_1, \dots, \lambda_p\}.$$

Then we have

$$\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{S}_{k:\infty}^\top = \sum_{i>k}\lambda_i\mathbf{s}_i\mathbf{s}_i^\top.$$

Fixing a unit vector $\mathbf{v} \in \mathbb{R}^M$, then

$$\mathbf{v}^\top \mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty} \mathbf{S}_{k:\infty}^\top \mathbf{v} = \sum_{i>k} \lambda_i \left(\mathbf{s}_i^\top \mathbf{v}\right)^2,$$

where each $\mathbf{s}_i^\top \mathbf{v}$ is $(1/M)$-subGaussian. By Bernstein's inequality, we have, with probability $\geq 1-\delta$,

$$\left| \sum_{i>k} \lambda_i \left(\mathbf{s}_i^\top \mathbf{v}\right)^2 - \frac{\sum_{i>k} \lambda_i}{M} \right| \lesssim \frac{1}{M} \cdot \left( \lambda_{k+1} \cdot \log \frac{1}{\delta} + \sqrt{\sum_{i>k} \lambda_i^2 \cdot \log \frac{1}{\delta}} \right).$$

By a union bound and net argument on $\mathcal{S}^{M-1}$, we have, with probability $\geq 1-\delta$, for every unit vector $\mathbf{v} \in \mathbb{R}^M$,

$$\left| \sum_{i>k} \lambda_i \left(\mathbf{s}_i^\top \mathbf{v}\right)^2 - \frac{\sum_{i>k} \lambda_i}{M} \right| \lesssim \frac{1}{M} \cdot \left( \lambda_{k+1} \cdot \left( M + \log \frac{1}{\delta} \right) + \sqrt{\sum_{i>k} \lambda_i^2 \cdot \left( M + \log \frac{1}{\delta} \right)} \right).$$

So with probability at least $1 - e^{-\Omega(M)}$, we have

$$\begin{aligned} \left\| \mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty} \mathbf{S}_{k:\infty}^\top - \frac{\sum_{i>k} \lambda_i}{M} \cdot \mathbf{I}_M \right\|_2 &\lesssim \frac{1}{M} \cdot \left( \lambda_{k+1} \cdot M + \sqrt{\sum_{i>k} \lambda_i^2 \cdot M} \right) \\ &\eqsim \lambda_{k+1} + \sqrt{\frac{\sum_{i>k} \lambda_i^2}{M}}, \end{aligned}$$

which completes the proof of the first part of Lemma G.2.

To prove the second part of Lemma G.2, it suffices to note that

$$\mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty} \mathbf{S}_{k:\infty}^\top \succeq \sum_{i=k+1}^{2M+k} \lambda_i \mathbf{s}_i \mathbf{s}_i^\top \succeq \lambda_{2M+k} \cdot \sum_{i=k+1}^{2M+k} \mathbf{s}_i \mathbf{s}_i^\top \succeq c \lambda_{2M+k} \cdot \mathbf{I}_M$$

for some constant $c > 1$ with probability at least $1 - e^{-\Omega(M)}$, where the last line follows from concentration properties of Gaussian covariance matrices (see e.g., Thereom 6.1 [41]). □

**Lemma G.3** (Head concentration)**.** *With probability at least* $1 - e^{-\Omega(M)}$*, we have*

$$\textit{for every } j \leq k, \quad |\mu_j(\mathbf{S}_{0:k} \mathbf{H}_{0:k} \mathbf{S}_{0:k}^\top) - \lambda_j| \lesssim \sqrt{\frac{k}{M}} \cdot \lambda_j.$$

*Proof of Lemma G.3.* Note that the spectrum of $\mathbf{S}_{0:k} \mathbf{H}_{0:k} \mathbf{S}_{0:k}^\top$ is indentical to the spectrum of $\mathbf{H}_{0:k}^{1/2} \mathbf{S}_{0:k}^\top \mathbf{S}_{0:k} \mathbf{H}_{0:k}^{1/2}$. We will bound the latter. We start with bounding the spectrum of $\mathbf{S}_{0:k} \mathbf{S}_{0:k}^\top$. To this end, we write $\mathbf{S}_{0:k}^\top \in \mathbb{R}^{k \times M}$ as

$$\mathbf{S}_{0:k}^\top = \begin{pmatrix} \mathbf{s}_1 & \dots & \mathbf{s}_M \end{pmatrix}, \quad \mathbf{s}_i \sim \mathcal{N}\left(0, \frac{1}{M} \cdot \mathbf{I}_k\right), \quad i = 1, \dots, M.$$

Then repeating the argument in Lemma G.2, we have, with probability $\geq 1-\delta$, for every unit vector $\mathbf{v} \in \mathbb{R}^k$,

$$\begin{aligned} \left| \mathbf{v}^\top \mathbf{S}_{0:k}^\top \mathbf{S}_{0:k} \mathbf{v} - 1 \right| &= \left| \sum_{i=1}^M \left(\mathbf{s}_i^\top \mathbf{v}\right)^2 - 1 \right| \\ &\lesssim \frac{1}{M} \cdot \left( 1 \cdot \left( k + \log \frac{1}{\delta} \right) + \sqrt{M \cdot \left( k + \log \frac{1}{\delta} \right)} \right) \\ &\lesssim \sqrt{\frac{k + \log(1/\delta))}{M}}. \end{aligned}$$

So we have, with probability $\geq 1 - e^{-\Omega(M)}$,

$$\left\| \mathbf{S}_{0:k}^\top \mathbf{S}_{0:k} - \mathbf{I}_k \right\|_2 \lesssim \sqrt{\frac{k}{M}}.$$

This implies that

$$\mu_j\big(\mathbf{H}_{0:k}^{1/2}\mathbf{S}_{0:k}^\top\mathbf{S}_{0:k}\mathbf{H}_{0:k}^{1/2}\big) \le \mu_j\big(\mathbf{H}_{0:k}^{1/2}\mathbf{H}_{0:k}^{1/2}\big) + c_1\cdot\sqrt{\frac{k}{M}}\cdot\mu_j\big(\mathbf{H}_{0:k}^{1/2}\mathbf{H}_{0:k}^{1/2}\big)$$
$$= \lambda_j + c_1\cdot\sqrt{\frac{k}{M}}\cdot\lambda_j,$$

and that

$$\mu_j\big(\mathbf{H}_{0:k}^{1/2}\mathbf{S}_{0:k}^\top\mathbf{S}_{0:k}\mathbf{H}_{0:k}^{1/2}\big) \ge \mu_j\big(\mathbf{H}_{0:k}^{1/2}\mathbf{H}_{0:k}^{1/2}\big) - c_1\cdot\sqrt{\frac{k}{M}}\cdot\mu_j\big(\mathbf{H}_{0:k}^{1/2}\mathbf{H}_{0:k}^{1/2}\big)$$
$$= \lambda_j - c_1\cdot\sqrt{\frac{k}{M}}\cdot\lambda_j.$$

We have completed the proof. □

### G.2 Concentration results under power-law spectrum

**Lemma G.4** (Eigenvalues of $\mathbf{SHS}^\top$ under power-law spectrum)**.** *Suppose Assumption 2 hold. There exist $a$-dependent constants $c_2 > c_1 > 0$ such that*

$$c_1 j^{-a} \le \mu_j(\mathbf{SHS}^\top) \le c_2 j^{-a}$$

*with probability at least $1 - e^{-\Omega(M)}$.*

*Proof of Lemma G.4.* Let $(\tilde\lambda_i)_{i=1}^M$ denote the eigenvalues of $\mathbf{SHS}^\top$ in an non-increasing order. Using Lemma G.1 with $k = M/c$ for some sufficiently large constant $c$ and noting that $\sum_{i>k} i^{-a} \eqsim k^{1-a}$, we have

$$\frac{1}{2}\cdot(j^{-a} + \tilde c_1 M^{-a}) - \tilde c_2\cdot M^{-a} \le \tilde\lambda_j \le \frac{3}{2}\cdot(j^{-a} + \tilde c_1 M^{-a}) + \tilde c_2\cdot M^{-a}$$

for every $j \in [M]$ for some constants $\tilde c_i, i \in [2]$ with probability at least $1 - e^{-\Omega(M)}$. Therefore, for all $j \le M/\tilde c$ for some sufficiently large constant $\tilde c > 1$, we have

$$\tilde\lambda_j \in [\tilde c_3 j^{-a}, \tilde c_4 j^{-a}]$$

with probability at least $1 - e^{-\Omega(M)}$ for some constants $\tilde c_3, \tilde c_4 > 0$. For $j \in [M/\tilde c, M]$, by monotonicity of the eigenvalues, we have

$$\tilde\lambda_j \le \tilde\lambda_{\lfloor M/\tilde c\rfloor} \le \tilde c_4\Big(\Big\lfloor\frac{M}{\tilde c}\Big\rfloor\Big)^{-a} \le \tilde c_5 M^{-a} \le \tilde c_5 j^{-a}$$

for some sufficiently large constant $\tilde c_5 > \tilde c_4$ with probability at least $1 - e^{-\Omega(M)}$. Moreover, using Lemma G.2 with $k = 0$, we obtain

$$\tilde\lambda_j \ge \tilde\lambda_M \ge \mu_{\min}(\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{S}_{k:\infty}^\top) \ge \tilde c_6\tilde\lambda_{2M} \ge \tilde c_7(M/\tilde c)^{-a} \ge \tilde c_8 j^{-a}$$

with probability at least $1 - e^{-\Omega(M)}$ for some constants $\tilde c_6, \tilde c_7, \tilde c_8 > 0$. Combining the bounds for $j \le M/\tilde c$ and $j \in [M/\tilde c, M]$ completes the proof. □

**Lemma G.5** (Ratio of eigenvalues of $\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{S}_{k:\infty}^\top$ under power-law spectrum)**.** *Suppose Assumption 2 hold. There exists some $a$-dependent constant $c > 0$ such that for any $k \ge 1$, the ratio between the $M/2$-th and $M$-th eigenvalues*

$$\frac{\mu_{M/2}(\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{S}_{k:\infty}^\top)}{\mu_M(\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{S}_{k:\infty}^\top)} \le c$$

*with probability at least $1 - e^{-\Omega(M)}$.*

*Proof of Lemma G.5.* We prove the lemma under two scenarios where $k$ is relatively small (or large) compared with $M$.

Let $c > 0$ be some sufficiently large constant. Applying Lemma G.1 with $\mathbf{H}_{k:\infty}$ replacing $\mathbf{H}$, for $k_0 = M/c$, we have

$$\begin{aligned}\mu_{M/2}(\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{S}_{k:\infty}^\top) &\le \frac{3}{2}\cdot\left(\lambda_{M/2+k} + \frac{\sum_{i>k_0}\lambda_{i+k}}{M}\right) + c_1\cdot\lambda_{k_0+1+k},\\ &\lesssim \big(\frac{M}{2}+k\big)^{-a} + \frac{(k_0+k)^{1-a}}{M} + (k_0+1+k)^{-a}\\ &\lesssim (k\vee M)^{-a} + (k\vee M)^{-a}\big(1\vee\frac{k}{M}\big) + (k\vee M)^{-a}\\ &\lesssim (k\vee M)^{-a}\big(1\vee\frac{k}{M}\big) \end{aligned} \tag{35}$$

with probability at least $1-e^{-\Omega(M)}$ for some constant $c_1 > 0$.

**Case 1:** $k \lesssim M$ From Lemma G.2, we have

$$\mu_{\min}(\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{S}_{k:\infty}^\top) \gtrsim \lambda_{k+2M} \gtrsim (k\vee M)^{-a}.$$

with probability at least $1-e^{-\Omega(M)}$. Therefore

$$\frac{\mu_{M/2}(\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{S}_{k:\infty}^\top)}{\mu_M(\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{S}_{k:\infty}^\top)} \lesssim 1$$

with probability at least $1-e^{-\Omega(M)}$ when $k/M \lesssim 1$.

**Case 2:** $k \gtrsim M$ On the other hand, when $k$ is relatively large, using Lemma G.1 with $\mathbf{H}_{k:\infty}$ replacing $\mathbf{H}$ again, we obtain

$$\begin{aligned}\mu_M(\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{S}_{k:\infty}^\top) &\ge \frac{1}{2}\cdot\left(\lambda_{M+k} + \frac{\sum_{i>k_0}\lambda_{i+k}}{M}\right) - c_1\cdot\lambda_{k_0+1+k},\\ &\ge c_2\Big[\big(M+k\big)^{-a} + \frac{(k_0+k)^{1-a}}{M}\Big] - c_3\cdot(k_0+1+k)^{-a}\end{aligned}$$

with probability at least $1-e^{-\Omega(M)}$, where $c_1, c_2, c_3 > 0$ are some universal constants. Choosing $k_0 = M/c^2$ for some sufficiently large constant $c > 0$, we further obtain

$$\begin{aligned}\mu_M(\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{S}_{k:\infty}^\top) &\ge c_4\big(M+k\big)^{-a}\Big[1+\frac{k}{M}\Big] - c_5(M+k)^{-a}\\ &\ge c_6\big(M\vee k\big)^{-a}\Big[1\vee\frac{k}{M}\Big] - c_7(M\vee k)^{-a}\end{aligned} \tag{36}$$

with probability at least $1-e^{-\Omega(M)}$, where $(c_i)_{i=4}^7$ are $a$-dependent constants. Since

$$c_6\big(M\vee k\big)^{-a}\Big[1\vee\frac{k}{M}\Big] - c_7(M\vee k)^{-a} \ge \frac{c_6}{2}(k\vee M)^{-a}\big(1\vee\frac{k}{M}\big)$$

when $k$ is large, i.e., $k/M > \tilde{c}$ for some sufficiently large $a$-dependent constant $\tilde{c} > 0$ that may depend on $(c_i)_{i=1}^7$, we have from Eq. (35) and (36) that

$$\frac{\mu_{M/2}(\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{S}_{k:\infty}^\top)}{\mu_M(\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{S}_{k:\infty}^\top)} \lesssim \frac{(k\vee M)^{-a}\big(1\vee\frac{k}{M}\big)}{(k\vee M)^{-a}\big(1\vee\frac{k}{M}\big)} \lesssim 1$$

with probability at least $1-e^{-\Omega(M)}$. □

### G.3 Concentration results under logarithmic power-law spectrum

**Lemma G.6** (Proof of Theorem 6 in [5])**.** *Suppose Assumption 4 hold. Then there exist some $a$-dependent constants $c, \tilde{c} > 0$ such that, with probability at least $1-e^{-\Omega(M)}$*

$$\mu_j(\mathbf{SHS}^\top) \in \begin{cases}[c\cdot j^{-1}\log^{-a}(j+1), \tilde{c}\cdot j^{-1}\log^{-a}(j+1)] & j \le \mathsf{k}^*,\\ [c\cdot M^{-1}\log^{1-a}(M), \tilde{c}\cdot M^{-1}\log^{1-a}(M)] & \mathsf{k}^* < j \le M,\end{cases}$$

*where* $\mathsf{k}^* \eqsim M/\log(M)$. *Also, there exists some* $a$*-dependent constants* $c_1, c_2 > 0$ *such that*

$$\frac{c_1}{j\log^{2a}(j+1)} \le \mu_j(\mathbf{S}\mathbf{H}^2\mathbf{S}^\top) \le \frac{c_2}{j\log^{2a}(j+1)}$$

*with probability at least* $1 - e^{-\Omega(M)}$.

*Proof of Lemma G.6.* The proof is adapted from the proof of Theorem 6 in [5]. We include it here for completeness.

**First part of Lemma G.6.** In Lemma G.1, for some constant $c > 1$, choose

$$\mathsf{k}^* := \min\left\{k \ge 0 : \sum_{i>k}\lambda_i \ge c\cdot M\cdot\lambda_{k+1}\right\}.$$

Then with probability $\ge 1 - e^{-\Omega(M)}$, we have:

$$\text{for every } 1 \le j \le M, \quad \frac{1}{c_1}\cdot\left(\lambda_j + \frac{\sum_{i>\mathsf{k}^*}\lambda_i}{M}\right) \le \tilde{\lambda}_j \le c_1\cdot\left(\lambda_j + \frac{\sum_{i>\mathsf{k}^*}\lambda_i}{M}\right),$$

where $c_1 > 1$ is a constant.

When $\lambda_j \eqsim j^{-1}\log^{-a}(j+1)$, we have

$$\mathsf{k}^* \eqsim M/\log(M),$$

and

$$\sum_{i>\mathsf{k}^*}\lambda_i \eqsim \log^{1-a}(\mathsf{k}^*) \eqsim \log^{1-a}(M).$$

Therefore, we have

$$\begin{aligned}\tilde{\lambda}_j &\eqsim \lambda_j + \frac{\sum_{i>\mathsf{k}^*}\lambda_i}{M}\\ &\eqsim \begin{cases} j^{-1}\log^{-a}(j+1) & j \le \mathsf{k}^*,\\ M^{-1}\log^{1-a}(M) & \mathsf{k}^* < j \le M,\end{cases}\end{aligned}$$

where $\mathsf{k}^* \eqsim M/\log(M)$.

**Second part of Lemma G.6.** Let $\bar{\lambda}_i$ denote the $i$-th eigenvalue of $\mathbf{S}\mathbf{H}^2\mathbf{S}^\top$ for $i \in [M]$. Using Lemma G.1 with $k = M/c$ for some sufficiently large constant $c_0$ and noting that $\sum_{i>k}\lambda_i^2 \eqsim \sum_{i>k} i^{-2}\log^{-2a}(i+1) \lesssim k^{-1}\log^{-2a} k$, we have

$$\begin{aligned}&\frac{1}{2}\cdot j^{-2}\log^{-2a}(j+1) - \tilde{c}_2\cdot M^{-2}\log^{-2a} M\\ &\quad\le \bar{\lambda}_j \le \frac{3}{2}\cdot(j^{-2}\log^{-2a}(j+1) + \tilde{c}_1 M^{-2}\log^{-2a} M) + \tilde{c}_2\cdot M^{-2}\log^{-2a} M\end{aligned}$$

for every $j \in [M]$ for some constants $\tilde{c}_i, i \in [2]$ with probability at least $1 - e^{-\Omega(M)}$. Therefore, for all $j \le M/\tilde{c}$ for some sufficiently large constant $\tilde{c} > 1$, we have

$$\bar{\lambda}_j \in [\tilde{c}_3\cdot j^{-2}\log^{-2a}(j+1), \tilde{c}_4\cdot j^{-2}\log^{-2a}(j+1)]$$

with probability at least $1 - e^{-\Omega(M)}$ for some constants $\tilde{c}_3, \tilde{c}_4 > 0$. For $j \in [M/\tilde{c}, M]$, by monotonicity of the eigenvalues, we have

$$\bar{\lambda}_j \le \bar{\lambda}_{\lfloor M/\tilde{c}\rfloor} \le \tilde{c}_4\left(\left\lfloor\frac{M}{\tilde{c}}\right\rfloor\right)^{-2}\log^{-2a}\left(\left\lfloor\frac{M}{\tilde{c}}\right\rfloor\right) \le \tilde{c}_5 M^{-2}\log^{-2a} M \le \tilde{c}_6\cdot j^{-2}\log^{-2a}(j+1)$$

for some constants $c_5, c_6 > 0$ with probability at least $1 - e^{-\Omega(M)}$. Moreover, using Lemma G.2 with $k = 0$, we obtain

$$\bar{\lambda}_j \ge \bar{\lambda}_M \ge \mu_{\min}(\mathbf{S}_{k:\infty}\mathbf{H}^2_{k:\infty}\mathbf{S}^\top_{k:\infty}) \ge \tilde{c}_7\bar{\lambda}_{2M} \ge \tilde{c}_8\cdot j^{-2}\log^{-2a}(j+1)$$

with probability at least $1 - e^{-\Omega(M)}$ for some constants $\tilde{c}_7, \tilde{c}_8 > 0$ when $j \in [M/\tilde{c}, M]$. Combining the bounds for $j \le M/\tilde{c}$ and $j \in [M/\tilde{c}, M]$ completes the proof. □

**Lemma G.7** (Ratio of eigenvalues of $\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{S}_{k:\infty}^\top$ under logarithmic power-law spectrum)**.** *Suppose Assumption 4 hold. There exists some $a$-dependent constant $c > 0$ such that for any $k \geq 1$, the ratio between the $M/2$-th and $M$-th eigenvalues*

$$\frac{\mu_{M/2}(\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{S}_{k:\infty}^\top)}{\mu_M(\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{S}_{k:\infty}^\top)} \leq c$$

*with probability at least $1 - e^{-\Omega(M)}$.*

*Proof of Lemma G.7.* Similar to the proof of Lemma G.5, we prove the lemma under two scenarios where $k$ is relatively small (or large) compared with $M$.

Let $c > 0$ be some sufficiently large constant. Applying Lemma G.1 with $\mathbf{H}_{k:\infty}$ replacing $\mathbf{H}$, for $k_0 = M/c$, we have

$$\begin{aligned}
\mu_{M/2}(\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{S}_{k:\infty}^\top) &\leq \frac{3}{2}\cdot\left(\lambda_{M/2+k} + \frac{\sum_{i>k_0}\lambda_{i+k}}{M}\right) + c_1\cdot\lambda_{k_0+1+k}, \\
&\lesssim \left(\frac{M}{2}+k\right)^{-1}\log^{-a}\left(\frac{M}{2}+k\right) + \frac{\log^{1-a}(k_0+k)}{M} + \frac{\log^{-a}(k_0+1+k)}{k_0+1+k} \\
&\lesssim \frac{\log^{-a}(M+k)}{(M+k)} + \frac{\log^{1-a}(M+k)}{M} \lesssim \frac{\log^{1-a}(M+k)}{M} \qquad (37)
\end{aligned}$$

with probability at least $1 - e^{-\Omega(M)}$ for some constant $c_1 > 0$.

**Case 1: $k \lesssim M$.** Applying Lemma G.1 with $\mathbf{H}_{k:\infty}$ replacing $\mathbf{H}$, for $k_0 = M/c$, we have

$$\begin{aligned}
\mu_M(\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{S}_{k:\infty}^\top) &\gtrsim \frac{1}{2}\cdot\left(\lambda_{M+k} + \frac{\sum_{i>k_0}\lambda_{i+k}}{M}\right) - c_1\cdot\lambda_{k_0+1+k}, \\
&\gtrsim \left(M+k\right)^{-1}\log^{-a}\left(M+k\right) + \frac{\log^{1-a}(k_0+k)}{M} - c\frac{\log^{-a}(k_0+1+k)}{k_0+1+k} \\
&\gtrsim \frac{\log^{-a}(M+k)}{(M+k)} + \frac{\log^{1-a}(M+k)}{M} - \tilde{c}\frac{\log^{-a}(M)}{M} \\
&\gtrsim \frac{\log^{1-a}M}{M}
\end{aligned}$$

with probability at least $1 - e^{-\Omega(M)}$. Therefore,

$$\frac{\mu_{M/2}(\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{S}_{k:\infty}^\top)}{\mu_M(\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{S}_{k:\infty}^\top)} \lesssim \left[\frac{\log^{1-a}(M+k)}{M}\right]\Big/\left[\frac{\log^{1-a}M}{M}\right] \lesssim 1$$

with probability at least $1 - e^{-\Omega(M)}$ when $k/M \lesssim 1$.

**Case 2: $k \gtrsim M$.** On the other hand, when $k$ is relatively large, using Lemma G.1 with $\mathbf{H}_{k:\infty}$ replacing $\mathbf{H}$ and $k_0 = M/c$ again, we obtain

$$\begin{aligned}
\mu_M(\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{S}_{k:\infty}^\top) &\geq \frac{1}{2}\cdot\left(\lambda_{M+k} + \frac{\sum_{i>k_0}\lambda_{i+k}}{M}\right) - c_1\cdot\lambda_{k_0+1+k}, \\
&\gtrsim \left(M+k\right)^{-1}\log^{-a}\left(M+k\right) + \frac{\log^{1-a}(k_0+k)}{M} - c_2\frac{\log^{-a}(k_0+1+k)}{k_0+1+k} \\
&\gtrsim k^{-1}\log^{-a}\left(k\right) + \frac{\log^{1-a}(M+k)}{M} - c_3\frac{\log^{-a}(M+k)}{M+k} \\
&\gtrsim \frac{\log^{1-a}(M+k)}{M}
\end{aligned}$$

with probability at least $1 - e^{-\Omega(M)}$, where $c_1, c_2, c_3 > 0$ are some $a$-dependent constants. Therefore,

$$\frac{\mu_{M/2}(\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{S}_{k:\infty}^\top)}{\mu_M(\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{S}_{k:\infty}^\top)} \lesssim \left[\frac{\log^{1-a}(M+k)}{M}\right]\Big/\left[\frac{\log^{1-a}(M+k)}{M}\right] \lesssim 1$$

with probability at least $1 - e^{-\Omega(M)}$ when $k/M \gtrsim 1$. □

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims are accurate and reflect the contributions and scope of the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of our results.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We provide all the details for our theorems.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide the details for our experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: Our experiments are numerical simulations and easy to reproduce.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide the details for our experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We provide error bars.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Our experiments are numerical simulations and can run with light compute resources.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: We conform with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: Our paper is theoretical and we do not expect direct societal impacts.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our paper is theoretical and we do not expect risk of such.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: Our paper is theoretical and does not use existing assets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: Our paper is theoretical and does not create new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Our paper is theoretical and does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Our paper is theoretical and does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.