The Benefits of Implicit Regularization from SGD in Least Squares Problems

Difan Zou[*]
University of California, Los Angeles
knowzou@cs.ucla.edu

Jingfeng Wu[*]
Johns Hopkins University
uuujf@jhu.edu

Vladimir Braverman
Johns Hopkins University
vova@cs.jhu.edu

Quanquan Gu
University of California, Los Angeles
qgu@cs.ucla.edu

Dean P. Foster
Amazon
dean@foster.net

Sham M. Kakade
University of Washington & Microsoft Research
sham@cs.washington.edu

Abstract

Stochastic gradient descent (SGD) exhibits strong algorithmic regularization effects in practice, which has been hypothesized to play an important role in the generalization of modern machine learning approaches. In this work, we seek to understand these issues in the simpler setting of linear regression (including both underparameterized and overparameterized regimes), where our goal is to make sharp instance-based comparisons of the implicit regularization afforded by (unregularized) average SGD with the explicit regularization of ridge regression. For a broad class of least squares problem instances (that are natural in high-dimensional settings), we show: (1) for every problem instance and for every ridge parameter, (unregularized) SGD, when provided with *logarithmically* more samples than that provided to the ridge algorithm, generalizes no worse than the ridge solution (provided SGD uses a tuned constant stepsize); (2) conversely, there exist instances (in this wide problem class) where optimally-tuned ridge regression requires *quadratically* more samples than SGD in order to have the same generalization performance. Taken together, our results show that, up to the logarithmic factors, the generalization performance of SGD is always no worse than that of ridge regression in a wide range of overparameterized problems, and, in fact, could be much better for some problem instances. More generally, our results show how algorithmic regularization has important consequences even in simpler (overparameterized) convex settings.

1 Introduction

Deep neural networks often exhibit powerful generalization in numerous machine learning applications, despite being *overparameterized*. It has been conjectured that the optimization algorithm itself, e.g., *stochastic gradient descent* (SGD), implicitly regularizes such overparameterized models [29]; here, (unregularized) overparameterized models could admit numerous global and local minima (many of which generalize poorly [29, 21]), yet SGD tends to find solutions that generalize well, even in the absence of explicit regularizers [22, 29, 19]. This regularizing effect due to the choice of the optimization algorithm is often referred to as *implicit regularization* [22].

[*]Equal Contribution

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

Before moving to the non-convex regime, we may hope to start by understanding this effect in the (overparameterized) convex regime. At least for linear models, there is a growing body of evidence suggesting that the implicit regularization of SGD is closely related to an explicit, ℓ_2-type of (ridge) regularization [25]. For example, (multi-pass) SGD for linear regression converges to the *minimum-norm interpolator*, which corresponds to the limit of the ridge solution with a vanishing penalty [29, 14]. Tangential evidence for this also comes from examining gradient descent, where a continuous time (gradient flow) analysis shows how the optimization path of gradient descent is (pointwise) closely connected to an explicit, ℓ_2-regularization [24, 1]. Similar results [2] have been further extended to SGD, where a (early-stopped) continuous-time SGD is demonstrated to perform similarly to ridge regression with certain regularization parameters.

However, as of yet, a precise comparison between the implicit regularization afforded by SGD and the explicit regularization of ridge regression (in terms of the *generalization performance*) is still lacking, especially when the hyperparameters (e.g., stepsize for SGD and regularization parameter for ridge regression) are allowed to be tuned. This motivates the central question in this work:

How does the generalization performance of SGD compare with that of ridge regression in least square problems?

In particular, even in the arguably simplest setting of linear regression, we seek to understand if/how SGD behaves differently from using an explicit ℓ_2-regularizer, with a particular focus on the overparameterized regime.

Our Contributions. Due to recent advances on sharp, *instance-dependent* excess risks bounds of both (single-pass) SGD and ridge regression for overparameterized least square problems [26, 30], a nearly complete answer to the above question is now possible using these tools. In this work, we deliver an *instance-based* risk comparison between SGD and ridge regression in several interesting settings, including one-hot distributed data and Gaussian data. In particular, for a broad class of least squares problem instances that are natural in high-dimensional settings, we show that

- For every problem instance and for every ridge parameter, (unregularized) SGD, when provided with *logarithmically* more samples than that provided to ridge regularization, generalizes no worse than the ridge solution, provided SGD uses a tuned constant stepsize.
- Conversely, there exist instances in our problem class where optimally-tuned ridge regression requires *quadratically* more samples than SGD to achieve the same generalization performance.

Quite strikingly, the above results show that, up to some logarithmic factors, the generalization performance of SGD is always no worse than that of ridge regression in a wide range of overparameterized least square problems, and, in fact, could be much better for some problem instances. As a special case (for the above two claims), our problem class includes a setting in which: (i) the signal-to-noise is bounded and (ii) the eigenspectrum decays at a polynomial rate $1/i^{\alpha}$, for $0 \leq \alpha \leq 1$ (which permits a relatively fast decay). This one-sided near-domination phenomenon (in these natural overparameterized problem classes) could further support the preference for the implicit regularization brought by SGD over explicit ridge regularization.

Several novel technical contributions are made to make the above risk comparisons possible. For the one-hot data, we derive similar risk upper bound of SGD and risk lower bound of ridge regression. For the Gaussian data, while a sharp risk bound of SGD is borrowed from [30], we prove a sharp lower bound of ridge regression by adapting the proof techniques developed in [26, 7]. By carefully comparing these upper and lower bound results (and exhibiting particular instances to show that our sample size inflation bounds are sharp), we are able to provide nearly complete conditions that characterize when SGD generalizes better than ridge regression.

Notation. For two functions $f(x) \geq 0$ and $g(x) \geq 0$ defined on $x > 0$, we write $f(x) \lesssim g(x)$ if $f(x) \leq c \cdot g(x)$ for some absolute constant $c > 0$; we write $f(x) \gtrsim g(x)$ if $g(x) \lesssim f(x)$; we write $f(x) \eqsim g(x)$ if $f(x) \lesssim g(x) \lesssim f(x)$. For a vector $\mathbf{w} \in \mathbb{R}^d$ and a positive semidefinite matrix $\mathbf{H} \in \mathbb{R}^{d \times d}$, we denote $\|\mathbf{w}\|_{\mathbf{H}} := \sqrt{\mathbf{w}^\top \mathbf{H} \mathbf{w}}$.

2 Related Work

In terms of making sharp risk comparisons with ridge, the work of [10] shows that OLS (after a PCA projection is applied to the data) is instance-wise competitive with ridge on fixed design problems.

The insights in our analysis are draw from this work, though there are a number of technical challenges in dealing with the random design setting. We start with a brief discussion of the technical advances in the analysis of ridge regression and SGD, and then briefly overview more related work comparing SGD to explicit norm-based regularization.

Excess Risk Bounds for Ridge Regression. In the underparameterized regime, the excess risk bounds for ridge regression has been well-understood [16]. In the overparameterized regime, a large body of works [12, 15, 28, 27] focused on characterizing the excess risk of ridge regression in the asymptotic regime where both the sample size N and dimension d go to infinite and $d/N \to \gamma$ for some finite γ. More recently, Bartlett et al. [7] developed sharp non-asymptotic risk bounds for ordinary least square in the overparameterized setting, which are further extended to ridge regression by Tsigler and Bartlett [26]. These bounds have additional interest because they are instance-dependent, in particular, depending on the data covariance spectrum. The risk bounds of ridge regression derived in Tsigler and Bartlett [26] is highly nontrivial in the overparameterized setting as it holds when the ridge parameter equals to zero or even being negative. This line of results build one part of the theoretical tools for this paper.

Excess Risk Bounds for SGD. Risk bounds for one-pass, constant-stepsize (average) SGD have been derived in the finite dimensional case [4, 9, 17, 18, 11, 1]. Very recently, the work of [30] extends these analyses, providing sharp *instance-dependent* risk bound applicable to the overparameterized regime; here, Zou et al. [30] provides nearly matching upper and lower excess risk bounds for constant-stepsize SGD, which are sharply characterized in terms of the full eigenspectrum of the population covariance matrix. This result plays a pivotal role in our paper.

Implicit Regularization of SGD vs. Explicit Norm-based Regularization. For least square problems, multi-pass SGD converges to the minimum-norm solution [22, 29, 14], which is widely cited as (one of) the implicit bias of SGD. However, in more general settings, e.g., convex but non-linear models, a (distribution-independent) norm-based regularizer is no longer sufficient to characterize the optimization behavior of SGD [3, 8, 23]. Those discussions, however, exclude the possibility of *hyperparameter tuning*, e.g., stepsize for SGD and penalty strength for ridge regression, and are not instance-based, either. Our aim in this paper is to provide instance-based excess risk comparison between the optimally tuned (one-pass) SGD and the optimally tuned ridge regression.

3 Problem Setup and Preliminaries

We seek to compare the generalization ability of SGD and ridge algorithms for *least square problems*. We use $\mathbf{x} \in \mathcal{H}$ to denote a feature vector in a (separable) Hilbert space $\mathcal{H}$. We use d to refer to the dimensionality of $\mathcal{H}$, where $d = \infty$ if $\mathcal{H}$ is infinite-dimensional. We use $y \in \mathbb{R}$ to denote a response that is generated by

$$y = \langle \mathbf{x}, \mathbf{w}^* \rangle + \xi,$$

where $\mathbf{w}^* \in \mathcal{H}$ is an unknown true model parameter and $\xi \in \mathbb{R}$ is the model noise. The following regularity assumption is made throughout the paper.

Assumption 3.1 (Well-specified noise)**.** *The second moment of* $\mathbf{x}$*, denoted by* $\mathbf{H} := \mathbb{E}[\mathbf{x}\mathbf{x}^\top]$*, is strictly positive definite and has finite trace. The noise* ξ *is independent of* $\mathbf{x}$ *and satisfies*

$$\mathbb{E}[\xi] = 0, \quad \textit{and} \quad \mathbb{E}[\xi^2] = \sigma^2.$$

In order to characterize the interplay between $\mathbf{w}^*$ and $\mathbf{H}$ in the excess risk bound, we introduce:

$$\mathbf{H}_{0:k} := \textstyle\sum_{i=1}^{k} \lambda_i \mathbf{v}_i \mathbf{v}_i^\top, \quad \text{and} \quad \mathbf{H}_{k:\infty} := \textstyle\sum_{i>k} \lambda_i \mathbf{v}_i \mathbf{v}_i^\top,$$

where $\{\lambda_i\}_{i=1}^{\infty}$ are the eigenvalues of $\mathbf{H}$ sorted in non-increasing order and $\mathbf{v}_i$'s are the corresponding eigenvectors. Then we define

$$\|\mathbf{w}\|^2_{\mathbf{H}^{-1}_{0:k}} = \sum_{i \le k} \frac{(\mathbf{v}_i^\top \mathbf{w})^2}{\lambda_i}, \quad \|\mathbf{w}\|^2_{\mathbf{H}_{k:\infty}} = \sum_{i>k} \lambda_i (\mathbf{v}_i^\top \mathbf{w})^2.$$

The least squares problem is to estimate the true parameter $\mathbf{w}^*$. Assumption 3.1 implies that $\mathbf{w}^*$ is the unique solution that minimizes the *population risk*:

$$L(\mathbf{w}^*) = \min_{\mathbf{w} \in \mathcal{H}} L(\mathbf{w}), \quad \text{where } L(\mathbf{w}) := \frac{1}{2} \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}} \big[(y - \langle \mathbf{w}, \mathbf{x} \rangle)^2 \big]. \tag{3.1}$$

Moreover we have that $L(\mathbf{w}^*) = \sigma^2$. For an estimation $\mathbf{w}$ found by some algorithm, e.g., SGD or ridge regression, its performance is measured by the *excess risk*, $L(\mathbf{w}) - L(\mathbf{w}^*)$.

Constant-Stepsize SGD with Tail-Averaging. We consider the constant-stepsize SGD with tail-averaging [4, 17, 18, 30]: at the t-th iteration, a fresh example $(\mathbf{x}_t, y_t)$ is sampled independently from the data distribution, and SGD makes the following update on the current estimator $\mathbf{w}_{t-1} \in \mathcal{H}$,

$$\mathbf{w}_t = \mathbf{w}_{t-1} + \gamma \cdot \big(y_t - \langle \mathbf{w}_{t-1}, \mathbf{x}_t \rangle\big)\mathbf{x}_t, \ t = 1, 2, \ldots, \qquad \mathbf{w}_0 = 0,$$

where $\gamma > 0$ is a constant stepsize. After N iterations (which is also the number of samples observed), SGD outputs the tail-averaged iterates as the final estimator:

$$\mathbf{w}_{\mathrm{sgd}}(N; \gamma) := \frac{2}{N} \sum_{t=N/2}^{N-1} \mathbf{w}_t.$$

In the underparameterized setting ($d < N$), constant-stepsize SGD with tail-averaging is known for achieving minimax optimal rate for least squares [17, 18]. More recently, Zou et al. [30] investigate the performance of constant-stepsize SGD with tail-averaging in the overparameterized regime ($d > N$), and establish *instance-dependent*, nearly-optimal excess risk bounds under mild assumptions on the data distribution. Notably, results from [30] cover underparameterized cases ($d < N$) as well.

Ridge Regression. Given N i.i.d. samples $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$, let us denote $\mathbf{X} := [\mathbf{x}_1, \ldots, \mathbf{x}_N]^\top \in \mathbb{R}^{N \times d}$ and $\mathbf{y} := [y_1, \ldots, y_N]^\top \in \mathbb{R}^d$. Then ridge regression outputs the following estimator for the true parameter [25]:

$$\mathbf{w}_{\mathrm{ridge}}(N; \lambda) := \arg\min_{\mathbf{w} \in \mathcal{H}} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 + \lambda \|\mathbf{w}\|_2^2, \tag{3.2}$$

where λ (which could possibly be negative) is a regularization parameter. We remark that the ridge regression estimator takes the following two equivalent form:

$$\mathbf{w}_{\mathrm{ridge}}(N; \lambda) = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_d)^{-1} \mathbf{X}^\top \mathbf{y} = \mathbf{X}^\top (\mathbf{X}\mathbf{X}^\top + \lambda \mathbf{I}_N)^{-1} \mathbf{y}. \tag{3.3}$$

The first expression is useful in the classical, underparameterized setting ($d < N$) [16]; and the second expression is more useful in the overparameterized setting ($d > N$) where the empirical covariance $\mathbf{X}^\top \mathbf{X}$ is usually not invertible [20, 26]. As a final remark, when $\lambda = 0$, ridge estimator reduces to the *ordinary least square estimator* (OLS) [13].

Generalizable Regime. In the following sections we will make instance-based risk comparisons between SGD and ridge regression. To make the comparison meaningful, we focus on regime where SGD and ridge regression are "generalizable", i.e, the SGD and the ridge regression estimators, with the optimally-tuned hypeparameters, can achieve excess risk that is smaller than the optimal population risk, i.e., σ^2. The formal mathematical definition is as follows.

Definition 1 (Generalizability)**.** *Consider an algorithm* `Alg` *and a least squares problem instance* `P`*. Let* $\texttt{Alg}(n, \boldsymbol{\theta})$ *be the output of the algorithm when provided with* n *i.i.d. samples from the problem instance* `P`*, and a set of hyperparameters* $\boldsymbol{\theta}$ *(that could be a function on* n*). Then we say that the algorithm* `Alg` *with sample size* n *and hyperparameters configuration* $\boldsymbol{\theta}$ *is* generalizable *on problem instance* `P`*, if*

$$\mathbb{E}_{\texttt{Alg},\texttt{P}}[L\big(\texttt{Alg}(n, \boldsymbol{\theta})\big)] - L(\mathbf{w}^*) \le \sigma^2,$$

where the expectation is over the randomness of `Alg` *and data drawn from the problem instance* `P`*.*

Clearly, the generalizable regime is defined by conditions on both the sample size, hyperparameter configuration, the problem instance, and the algorithm. For example, in the d-dimensional setting with $\|\mathbf{w}^*\|_2 = O(1)$, the ordinary least squares (OLS) solution (ridge regression with $\lambda = 0$), i.e., $\mathbf{w}_{\mathrm{ridge}}(N; 0)$ has $\mathcal{O}(d\sigma^2/N)$ excess risk, then we can say that the ridge regression with regularization parameter $\lambda = 0$ and sample size $N = \omega(d)$ is in the generalizable regime on all problem instances in d-dimension with $\|\mathbf{w}^*\|_2 = O(1)$.

Sample Inflation vs. Risk Inflation Comparisons. This work characterizes the *sample inflation* of SGD, i.e., bounding the required sample size of SGD to achieve an instance-based comparable excess risk as ridge regression (which is essentially the notion of Bahadur statistical efficiency [5, 6]). Another natural comparison would be examining the *risk inflation* of SGD, examining the instance-based increase in risk for any fixed sample size. Our preference for the former is due to the relative instability of the risk with respect to the sample size (in some cases, given a slightly different sample size, the risk could rapidly change.).

4 Warm-Up: One-Hot Least Squares Problems

Let us begin with a simpler data distribution, the *one-hot* data distribution. (inspired by settings where the input distribution is sparse). In detail, assume each input vector $\mathbf{x}$ is sampled from the set of natural basis $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_d\}$ according to the data distribution given by $\mathbb{P}\{\mathbf{x} = \mathbf{e}_i\} = \lambda_i$, where $0 < \lambda_i \leq 1$ and $\sum_i \lambda_i = 1$. The class of one-hot least square instances is completely characterized by the following problem set:

$$\big\{(\mathbf{w}^*; \lambda_1, \cdots, \lambda_d) : \ \mathbf{w}^* \in \mathcal{H}, \ \textstyle\sum_i \lambda_i = 1, \ 1 \geq \lambda_1 \geq \lambda_2 \geq \cdots > 0\big\}.$$

Clearly the population data covariance matrix is $\mathbf{H} = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$. The next two theorems give an instance-based sample inflation comparisons for this problem class.

Theorem 4.1 (Instance-wise comparison, one-hot data)**.** *Let $\mathbf{w}_{\mathrm{sgd}}(N; \gamma)$ and $\mathbf{w}_{\mathrm{ridge}}(N; \lambda)$ be the solutions found by SGD and ridge regression when using N training examples. Then for any one-hot least square problem instance such that the ridge regression solution is generalizable and any λ, there exists a choice of stepsize γ^* for SGD such that*

$$L\big[\mathbf{w}_{\mathrm{sgd}}(N_{\mathrm{sgd}}; \gamma^*)\big] - L(\mathbf{w}^*) \lesssim L\big[\mathbf{w}_{\mathrm{ridge}}(N_{\mathrm{ridge}}; \lambda)\big] - L(\mathbf{w}^*) < \sigma^2,$$

provided the sample size of SGD satisfies

$$N_{\mathrm{sgd}} \geq N_{\mathrm{ridge}}.$$

Theorem 4.1 suggests that for *every* one-hot problem instance, when provided with the same or more number of samples, the SGD solution with a properly tuned stepsize generalizes at most constant times worse than the optimally tuned ridge regression solution. In other words, with the same number of samples, SGD is *always* competitive with ridge regression.

Theorem 4.2 (Best-case comparison, one-hot data)**.** *There exists an one-hot least square problem instance satisfying $\|\mathbf{w}^*\|_{\mathbf{H}}^2 = \sigma^2$, and a SGD solution with constant stepsize and sample size N_{sgd}, such that for any ridge regression solution with sample size*

$$N_{\mathrm{ridge}} \leq \frac{N_{\mathrm{sgd}}^2}{\log^2(N_{\mathrm{sgd}})},$$

it holds that,

$$L\big[\mathbf{w}_{\mathrm{ridge}}(N_{\mathrm{ridge}}; \lambda)\big] - L(\mathbf{w}^*) \gtrsim L\big[\mathbf{w}_{\mathrm{sgd}}(N_{\mathrm{sgd}}; \gamma^*)\big] - L(\mathbf{w}^*).$$

Theorem 4.2 shows that for some one-hot least square instance, ridge regression, even with the optimally-tuned regularization, needs at least (nearly) quadratically more samples than that provided to SGD, in order to compete with the optimally-tuned SGD. In other words, ridge regression could be much worse than SGD for one-hot least squares problems.

Remark 4.3. *The above two results together indicate a* superior *performance of the implicit regularization of SGD in comparison with the explicit regularization of ridge regression, for one-hot least squares problems. This is not the only case that SGD is always no worse than ridge estimator. In fact, we will next turn to compare SGD with ridge regression for the class of Gaussian least square instances, where both SGD and ridge regression exhibit richer behaviors but SGD still exhibits superiority over the ridge estimator.*

5 Gaussian Least Squares Problems

In this section, we consider least squares problems with a Gaussian data distribution. In particular, assume the population distribution of the input vector $\mathbf{x}$ is Gaussian[2], i.e., $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{H})$. We further make the following regularity assumption for simplicity:

Assumption 5.1. $\mathbf{H}$ *is strictly positive definite and has a finite trace.*

Gaussian least squares problems are completely characterized by the following problem set $\big\{(\mathbf{w}^*; \mathbf{H}) : \ \mathbf{w}^* \in \mathcal{H}\big\}$.

The next theorem give an instance-based sample inflation comparison between SGD and ridge regression for Gaussian least squares instances.

[2]We restrict ourselves to the Gaussian distribution for simplicity. Our results hold under more general assumptions, e.g., $\mathbf{H}^{-1/2}\mathbf{x}$ has sub-Gaussian tail and independent components [7] and is symmetrically distributed.

Theorem 5.1 (Instance-wise comparison, Gaussian data). *Let $\mathbf{w}_{\mathrm{sgd}}(N;\gamma)$ and $\mathbf{w}_{\mathrm{ridge}}(N;\lambda)$ be the solutions found by SGD and ridge regression respectively. Then under Assumption 5.1, for any Gaussian least square problem instance such that the ridge regression solution is generalizable and any λ, there exists a choice of stepsize γ^* for SGD such that*

$$L\big[\mathbf{w}_{\mathrm{sgd}}(N_{\mathrm{sgd}};\gamma^*)\big] - L(\mathbf{w}^*) \lesssim L\big[\mathbf{w}_{\mathrm{ridge}}(N_{\mathrm{ridge}};\lambda)\big] - L(\mathbf{w}^*),$$

provided the sample size of SGD satisfies

$$N_{\mathrm{sgd}} \geq (1+R^2)\cdot\kappa(N_{\mathrm{ridge}})\cdot\log(a)\cdot N_{\mathrm{ridge}},$$

where

$$\kappa(n) = \frac{\mathrm{tr}(\mathbf{H})}{n\lambda_{\min\{n,d\}}}, \quad R^2 = \frac{\|\mathbf{w}^*\|_{\mathbf{H}}^2}{\sigma^2}, \quad a = \kappa(N_{\mathrm{ridge}})R\sqrt{N}.$$

Note that the result in Theorem 5.1 holds for arbitrary λ. Then this theorem provides a sufficient condition for SGD such that it provably performs no worse than optimal ridge regression solution (i.e., ridge regression with optimal λ). Besides, we would also like to point out that the SGD stepsize γ^* in Theorem 5.1 is only a function of the regularization parameter λ and $\mathrm{tr}(\mathbf{H})$, which can be easily estimated from training dataset without knowing the exact formula of $\mathbf{H}$.

Different from the one-hot case, here the required sample size for SGD depends on two important quantities: R^2 and $\kappa(N_{\mathrm{ridge}})$. In particular, $R^2 = \|\mathbf{w}^*\|_{\mathbf{H}}^2/\sigma^2$ can be understood as the *signal-to-noise* ratio. The quantity $\kappa(N_{\mathrm{ridge}})$ characterizes the flatness of the eigenspectrum of $\mathbf{H}$ in the top N_{ridge}-dimensional subspace, which clearly satisfies $\kappa(N_{\mathrm{ridge}}) \geq 1$. Let us further explain why we have the dependencies on R^2 and $\kappa(N_{\mathrm{ridge}})$ in the condition of the sample inflation for SGD.

A large R^2 emphasizes the problem hardness is more from the numerical optimization instead of from the statistic learning. In particular, let us consider a special case where $\sigma = 0$ and $R^2 = \infty$, i.e., there is no noise in the least square problem, and thus solving it is purely a numerical optimization issue. In this case, ridge regression with $\lambda = 0$ achieves *zero* population risk so long as the observed data can span the whole parameter space, but constant stepsize SGD in general suffers a non-zero risk in finite steps, thus cannot be competitive with the risk of ridge regression, which is as predicted by Theorem 5.1. From a learning perspective, a constant or even small R^2 is more interesting.

To explain why the dependency on $\kappa(N_{\mathrm{ridge}})$ is unavoidable, we can consider a 2-d dimensional example where

$$\mathbf{H} = \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{N_{\mathrm{ridge}}\cdot\kappa(N_{\mathrm{ridge}})} \end{pmatrix}, \quad \mathbf{w}^* = \begin{pmatrix} 0 \\ N_{\mathrm{ridge}}\cdot\kappa(N_{\mathrm{ridge}}) \end{pmatrix}.$$

It is commonly known that for this problem, ridge regression with $\lambda = 0$ can achieve $\mathcal{O}(\sigma^2/N_{\mathrm{ridge}})$ excess risk bound [13]. However, this problem is rather difficult for SGD since it is hard to learn the second coordinate of $\mathbf{w}^*$ using gradient information (the gradient in the second coordinate is quite small). In fact, in order to accurately learn $\mathbf{w}^*[2]$, SGD requires at least $\Omega(1/\lambda_2) = \Omega\big(N_{\mathrm{ridge}}\kappa(N_{\mathrm{ridge}})\big)$ iterations/samples, which is consistent with our theory.

Then from Theorem 5.1 it can be observed that when the signal-to-noise ratio is nearly a constant, i.e., $R^2 = \Theta(1)$, and the eigenspectrum of $\mathbf{H}$ does not decay too fast so that $\kappa(N_{\mathrm{ridge}}) \leq \mathrm{polylog}(N_{\mathrm{ridge}})$, SGD provably generalizes no worse than ridge regression, provided with logarithmically more samples than that provided to ridge regression. More specifically, the following corollary gives a family of problem instances that are in this regime.

Corollary 5.1. *Under the same conditions as Theorem 5.1, let N_{ridge} be the sample size of ridge regression. Consider the problem instance that satisfies $R^2 = \Theta(1)$, $d = O(N_{\mathrm{ridge}})$, and $\lambda_i = 1/i^\alpha$ for some $\alpha \leq 1$, then SGD, with a tuned stepsize γ^*, provably generalizes no worse than any ridge regression solution in the generalizable regime if*

$$N_{\mathrm{sgd}} \geq \log^2(N_{\mathrm{ridge}})\cdot N_{\mathrm{ridge}}.$$

We would like to further point out that the comparison made in Corollary 5.1 concerns the worst-case result regarding $\mathbf{w}^*$ (from the perspective of SGD), while SGD could perform much better if $\mathbf{w}^*$ has a nice structure. For example, considering the same setting in Corollary 5.1 but assuming that the ground truth $\mathbf{w}^*$ is drawn from a prior distribution that is rotation invariant, SGD can be no worse than ridge regression provided the same or larger sample size. We formally state this result in the following corollary.

Corollary 5.2. *Under the same conditions as Corollary 5.1, let $N_{\rm ridge}$ be the sample size of ridge regression. Consider the problem instance with random and rotation invariant $\mathbf{w}^*$, then SGD with a tuned stepsize γ^* provably generalizes no worse than any ridge regression solution in the generalizable regime if*

$$N_{\rm sgd} \geq N_{\rm ridge}.$$

The next theorem shows that, in fact, for some instances, SGD could perform much better than ridge regression, as for the one-hot least square problems.

Theorem 5.2 (Best-case comparison, Gaussian data)**.** *There exists a Gaussian least square problem instance satisfying $R^2 = 1$ and $\kappa(N_{\rm sgd}) = \Theta(1)$, and an SGD solution with a constant stepsize and sample size $N_{\rm sgd}$, such that for any ridge regression solution (i.e., any λ) with sample size*

$$N_{\rm ridge} \leq \frac{N_{\rm sgd}^2}{\log^2(N_{\rm sgd})},$$

it holds that,

$$L\big[\mathbf{w}_{\rm ridge}(N_{\rm ridge}; \lambda)\big] - L(\mathbf{w}^*) \gtrsim L\big[\mathbf{w}_{\rm sgd}(N_{\rm sgd}; \gamma^*)\big] - L(\mathbf{w}^*).$$

Besides the instance-wise comparison, it is also interesting to see under what condition SGD can provably outperform ridge regression, i.e., achieving comparable or smaller excess risk using the *same* number of samples. The following theorem shows that this occurs when the signal-to-noise ratio R^2 is a constant and there is only a small fraction of $\mathbf{w}^*$ living in the tail eigenspace of $\mathbf{H}$.

Theorem 5.3 (SGD outperforms ridge regression, Gaussian data)**.** *Let $N_{\rm ridge}$ be sample size of ridge regression and $k^* = \min\big\{k : \lambda_k \leq \frac{\mathrm{tr}(\mathbf{H})}{N_{\rm ridge}\log(N_{\rm ridge})}\big\}$, then if $R^2 = \Theta(1)$, and*

$$\sum_{i=k^*+1}^{N_{\rm ridge}} \lambda_i (\mathbf{w}^*[i])^2 \lesssim \frac{k^* \|\mathbf{w}^*\|_{\mathbf{H}}^2}{N_{\rm ridge}},$$

for any ridge regression solution that is generalizable and any λ, there exists a choice of stepsize γ^* for SGD such that

$$L\big[\mathbf{w}_{\rm sgd}(N_{\rm sgd}; \gamma^*)\big] - L(\mathbf{w}^*) \lesssim L\big[\mathbf{w}_{\rm ridge}(N_{\rm ridge}; \lambda)\big] - L(\mathbf{w}^*)$$

provided the sample size of SGD satisfies

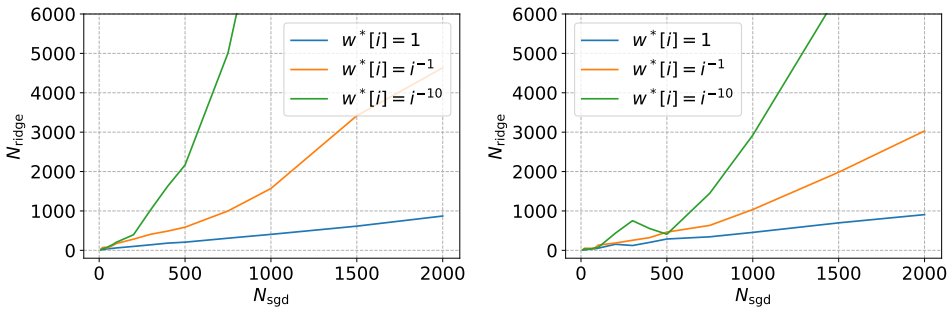
$$N_{\rm sgd} \geq N_{\rm ridge}.$$

Experiments. We perform experiments on Gaussian least square problem. We consider 6 problem instances, which are the combinations of 2 different covariance matrices $\mathbf{H}$: $\lambda_i = i^{-1}$ and $\lambda_i = i^{-2}$; and 3 different true model parameter vectors $\mathbf{w}^*$: $\mathbf{w}^*[i] = 1$, $\mathbf{w}^*[i] = i^{-1}$, and $\mathbf{w}^*[i] = i^{-10}$. Figure 1 compares the required sample sizes of ridge regression and SGD that lead to the same population risk on these 6 problem instances, where the hyperparameters (i.e., γ and λ) are fine-tuned to achieve the best performance. We have two key observations: (1) in terms of the worst problem instance for SGD (i.e., $\mathbf{w}^*[i] = 1$), its sample size is only worse than ridge regression up to nearly constant factors (the curve is nearly linear); and (2) SGD can significantly outperform ridge regression when the true model $\mathbf{w}^*$ mainly lives in the head eigenspace of $\mathbf{H}$ (i.e., $\mathbf{w}^*[i] = i^{-10}$). The empirical observations are pretty consistent with our theoretical findings and again demonstrate the benefit of the implicit regularization of SGD.

6 An Overview of the Proof

In this section, we will sketch the proof of main Theorems for Gaussian least squares problems. Recall that we aim to show that provided certain number of training samples, SGD is guaranteed to generalize better than ridge regression. Therefore, we will compare the risk *upper bound* of SGD [30] with the risk *lower bound* of ridge regression [26][3]. In particular, we first provide the following informal lemma summarizing the aforementioned risk bounds of SGD and ridge regression.

[3]The lower bound of ridge regression in our paper is a tighter variant of the lower bound in Tsigler and Bartlett [26] since we consider Gaussian case and focus on the expected excess risk. Tsigler and Bartlett [26] studied the sub-Gaussian case and established a high-probability risk bound.



(a) $\lambda_i = i^{-1}$ (b) $\lambda_i = i^{-2}$

Figure 1: Sample size comparison between SGD and ridge regression, where the stepsize γ and regularization parameter λ are fine-tuned to achieve the best performance. The problem dimension is $d = 200$ and the variance of model noise is $\sigma^2 = 1$. We consider 6 combinations of 2 different covariance matrices and 3 different ground truth model vectors. The plots are averaged over 20 independent runs.

Lemma 6.1 (Risk bounds of SGD and ridge regression, informal)**.** *Suppose Assumptions 3.1 and 5.1 hold and $\gamma \leq 1/\operatorname{tr}(\mathbf{H})$, then SGD has the following risk upper bound for arbitrary $k_1, k_2 \in [d]$,*

$$\text{SGDRisk} \lesssim \underbrace{\frac{1}{\gamma^2 N_{\text{sgd}}^2} \cdot \left\| \exp(-N_{\text{sgd}}\gamma\mathbf{H})\mathbf{w}^* \right\|_{\mathbf{H}_{0:k_1}^{-1}}^2 + \left\|\mathbf{w}^*\right\|_{\mathbf{H}_{k_1:\infty}}^2}_{\text{SGDBiasBound}}$$
$$\underbrace{+ (1+R^2)\sigma^2 \cdot \left(\frac{k_2}{N_{\text{sgd}}} + N_{\text{sgd}}\gamma^2 \sum_{i>k_2} \lambda_i^2 \right)}_{\text{SGDVarianceBound}}. \tag{6.1}$$

Additionally, ridge regression has the following risk lower bound for a constant $\widetilde{\lambda}$, depending on λ, N_{ridge}, and $\mathbf{H}$, and $k^* = \min\{k : N_{\text{ridge}}\lambda_k \lesssim \widetilde{\lambda}\}$

$$\text{RidgeRisk} \gtrsim \underbrace{\left(\frac{\widetilde{\lambda}}{N_{\text{ridge}}} \right)^2 \left\|\mathbf{w}^*\right\|_{\mathbf{H}_{0:k^*}^{-1}}^2 + \left\|\mathbf{w}^*\right\|_{\mathbf{H}_{k^*:\infty}}^2}_{\text{RidgeBiasBound}} + \underbrace{\sigma^2 \cdot \left(\frac{k^*}{N_{\text{ridge}}} + \frac{N_{\text{ridge}}}{\widetilde{\lambda}^2} \sum_{i>k^*} \lambda_i^2 \right)}_{\text{RidgeVarianceBound}}. \tag{6.2}$$

We first highlight some useful observations in Lemma 6.1.

1. SGD has a condition on the stepsize: $\gamma \leq 1/\operatorname{tr}(\mathbf{H})$, while ridge regression has no condition on the regularization parameter λ.
2. Both the upper bound of SGD and the lower bound of ridge regression can be decomposed into two parts corresponding to the head and tail eigenspaces of $\mathbf{H}$. Furthermore, for the upper bound of SGD, the decomposition is arbitrary (k_1 and k_2 are arbitrary), while for the lower bound of ridge estimator, the decomposition is fixed (i.e., k^* is fixed).
3. Regarding the SGDBiasBound and SGDVarianceBound, performing the transformation $N \to \alpha N$ and $\gamma \to \alpha^{-1}\gamma$ will decrease SGDVarianceBound by a factor of α while the SGDBiasBound remains unchanged.

Based on the above useful observations, we can now interpret the proof sketch for Theorems 5.1, 5.2, and 5.3. We will first give the sketch for Theorem 5.3 and then prove Theorem 5.2 for the ease of presentation. We would like to emphasize that the calculation in the proof sketch may not be the sharpest since they are presented for the ease of exposition. A preciser and sharper calculation can be found in Appendix.

Proof Sketch of Theorem 5.1. In order to perform instance-wise comparison, we need to take care of all possible $\mathbf{w}^* \in \mathcal{H}$. Therefore, by Observation 2, we can simply pick $k_1 = k_2 = k^*$ in the upper

bound (6.1). Then it is clear that if setting $\gamma = \widetilde{\lambda}^{-1}$ and $N_{\mathrm{sgd}} = N_{\mathrm{ridge}}$, we have

$$\mathrm{SGDBiasBound} \le \mathrm{RidgeBiasBound}$$
$$\mathrm{SGDVarianceBound} = (1+R^2)\cdot \mathrm{RidgeVarianceBound}.$$

Then by Observation 3, enlarging N_{sgd} by $(1+R^2)$ times suffices to guarantee

$$\mathrm{SGDBiasBound} + \mathrm{SGDVarianceBound} \le \mathrm{RidgeBiasBound} + \mathrm{RidgeVarianceBound}.$$

On the other hand, according to Observation 1, there is an upper bound on the feasible stepsize of SGD: $\gamma \le 1/\operatorname{tr}(\mathbf{H})$. Therefore, the above claim only holds when $\widetilde{\lambda} \ge \operatorname{tr}(\mathbf{H})$.

When $\widetilde{\lambda} \le \operatorname{tr}(\mathbf{H})$, the stepsize $\widetilde{\lambda}^{-1}$ is no longer feasible and instead, we will use the largest possible stepsize: $\gamma = 1/\operatorname{tr}(\mathbf{H})$. Besides, note that we assume ridge regression solution is in the generalizable regime, then it holds that $k^* \le N_{\mathrm{ridge}}$ since otherwise we have

$$\mathrm{RidgeRisk} \gtrsim \mathrm{RidgeVarianceBound} \ge \sigma^2.$$

Then again we set $k_1 = k_2 = k^*$ in SGDBiasBound and SGDVarianceBound. Applying the choice of stepsize $\gamma = 1/\operatorname{tr}(\mathbf{H})$ and sample size

$$N_{\mathrm{sgd}} = \frac{\log(R^2 N_{\mathrm{ridge}})}{\gamma \lambda_{k^*}} \le N_{\mathrm{ridge}} \cdot \kappa(N_{\mathrm{ridge}}) \cdot \log(R^2 N_{\mathrm{ridge}}),$$

we get

$$\begin{aligned}
\mathrm{SGDBiasBound} &\le \frac{(1-N_{\mathrm{sgd}}\gamma\lambda_{k^*})^{N_{\mathrm{sgd}}}}{\gamma^2 N_{\mathrm{sgd}}^2 \lambda_{k^*}^2} \cdot \|\mathbf{w}^*\|_{\mathbf{H}_{0:k^*}}^2 + \|\mathbf{w}^*\|_{\mathbf{H}_{k^*:\infty}}^2 \\
&\le \frac{\sigma^2}{N_{\mathrm{ridge}}} + \|\mathbf{w}^*\|_{\mathbf{H}_{k^*:\infty}}^2 \\
&\le \mathrm{RidgeBiasBound} + \mathrm{RidgeVarianceBound}. \qquad (6.3)
\end{aligned}$$

Moreover, we can also get the following bound on SGDVarianceBound,

$$\begin{aligned}
\mathrm{SGDVarianceBound} &\le (1+R^2)\sigma^2 \cdot \left(\frac{k^*}{N_{\mathrm{ridge}}} + \frac{\log(R^2 N_{\mathrm{ridge}})}{\lambda_{k^*} \operatorname{tr}(\mathbf{H})} \sum_{i>k^*} \lambda_i^2 \right) \\
&\le (1+R^2)\log(R^2 N_{\mathrm{ridge}}) \cdot \mathrm{RidgeVarianceBound},
\end{aligned}$$

where in the second inequality we use the fact that

$$\frac{N_{\mathrm{ridge}}}{\widetilde{\lambda}^2} \ge \frac{1}{\lambda_{k^*}\widetilde{\lambda}} \ge \frac{1}{\lambda_{k^*}\operatorname{tr}(\mathbf{H})}.$$

Therefore by Observation 3 again we can enlarge N_{sgd} properly to ensure that SGDVarianceBound remains unchanged and SGDVarianceBound $\le$ RidgeVarianceBound. Then combining this and (6.3) we can get

$$\mathrm{SGDBiasBound} + \mathrm{SGDVarianceBound} \le 2 \cdot \mathrm{RidgeBiasBound} + 2 \cdot \mathrm{RidgeVarianceBound},$$

which completes the proof.

Proof Sketch of Theorem 5.3. Now we will investigate in which regime SGD will generalizes no worse than ridge regression when provided with same training sample size. For simplicity in the proof we assume $R^2 = 1$. First note that we only need to deal with the case where $\widetilde{\lambda} \le \operatorname{tr}(\mathbf{H})$ by the proof sketch of Theorem 5.1.

Unlike the instance-wise comparison that consider all possible $\mathbf{w}^* \in \mathcal{H}$, in this lemma we only consider the set of $\mathbf{w}^*$ that SGD performs well. Specifically, as we have shown in the proof of Theorem 5.1, in the worst-case comparison (in terms of $\mathbf{w}^*$), we require SGD to be able to learn the first k^* (where $k^* \le N_{\mathrm{ridge}}$) coordinates of $\mathbf{w}^*$ in order to be competitive with ridge regression, while SGD with sample size N_{sgd} can only be guaranteed to learn the first k_{sgd}^* coordinates of $\mathbf{w}^*$, where $k_{\mathrm{sgd}}^* = \min\{k : N_{\mathrm{ridge}}\lambda_k \le \operatorname{tr}(\mathbf{H})\}$. Therefore, in the instance-wise comparison we need to enlarge N_{sgd} to $N_{\mathrm{ridge}} \cdot \kappa(N_{\mathrm{ridge}})$ to guarantee the learning of the top k^* coordinates of $\mathbf{w}^*$.

However, this is not required for some good $\mathbf{w}^*$'s that have small components in the $k^*_{\rm sgd}$-k^* coordinates. In particular, as assumed in the theorem, we have $\sum_{i=\widehat{k}+1}^{N_{\rm ridge}} \lambda_i(\mathbf{w}^*[i])^2 \le \widehat{k}\|\mathbf{w}^*\|^2_{\mathbf{H}}/N_{\rm ridge}$, where $\widehat{k} := \min\{k : \lambda_k N_{\rm sgd} \le \mathrm{tr}(\mathbf{H})\cdot\log(N_{\rm sgd})\}$ satisfies $\widehat{k} \le k^*_{\rm sgd} \le k^*$. Then let $k_1 = \widehat{k}$ in SGDBiasBound, we have

$$\begin{aligned}
\text{SGDBiasBound} &= \frac{1}{\gamma^2 N^2_{\rm ridge}} \cdot \big\| \exp(-N_{\rm ridge}\gamma\mathbf{H})\mathbf{w}^* \big\|^2_{\mathbf{H}^{-1}_{0:\widehat{k}}} + \|\mathbf{w}^*\|^2_{\mathbf{H}_{\widehat{k}:\infty}} \\
&\le (1 - N_{\rm ridge}\gamma\lambda_{\widehat{k}})^{N_{\rm ridge}} \cdot \|\mathbf{w}^*\|^2_{\mathbf{H}_{0:k^*}} + \|\mathbf{w}^*\|^2_{\mathbf{H}_{\widehat{k}:\infty}} \\
&\overset{(i)}{\le} \frac{R^2\sigma^2(\widehat{k}+1)}{N_{\rm ridge}} + \|\mathbf{w}^*\|^2_{\mathbf{H}_{k^*:\infty}} \\
&\le 2\cdot\text{RidgeVarBound} + \text{RidgeBiasBound}.
\end{aligned}$$

where (i) is due to the condition that $\sum_{i=\widehat{k}+1}^{N_{\rm ridge}} \lambda_i(\mathbf{w}^*[i])^2 \le \widehat{k}\|\mathbf{w}^*\|^2_{\mathbf{H}}/N_{\rm ridge}$. Moreover, it is easy to see that given $N_{\rm sgd} = N_{\rm ridge}$ and $\gamma = 1/\mathrm{tr}(\mathbf{H}) \le 1/\widetilde{\lambda}$, we have $\text{SGDVarianceBound} \le 2\cdot\text{RidgeVarianceBound}$. As a consequence we can get

$$\text{SGDBiasBound} + \text{SGDVarianceBound} \le 3\cdot\text{RidgeBiasBound} + 3\cdot\text{RidgeVarianceBound}.$$

Proof Sketch of Theorem 5.2. We will consider the best $\mathbf{w}^*$ for SGD, which only has nonzero entry in the first coordinate. For example, consider a true model parameter vector with $\mathbf{w}^*[1] = 1$ and $\mathbf{w}^*[i] = 0$ for $i \ge 2$ and a problem instance whose spectrum of $\mathbf{H}$ has a flat tail with $\sum_{i\ge N_{\rm ridge}} \lambda_i^2 = \Theta(1)$ and $\sum_{i\ge 2}\lambda_i^2 = \Theta(1)$. Then according to Lemma 6.1, we can set the stepsize as $\gamma = \Theta(\log(N_{\rm sgd})/N_{\rm sgd})$ and get

$$\begin{aligned}
\text{SGDRisk} &\lesssim \text{SGDBiasBound} + \text{SGDVarianceBound} \\
&= O\bigg(\frac{1}{N_{\rm sgd}} + \frac{\log^2(N_{\rm sgd})}{N_{\rm sgd}}\bigg) = O\bigg(\frac{\log^2(N_{\rm sgd})}{N_{\rm sgd}}\bigg).
\end{aligned}$$

For ridge regression, according to Lemma 6.1 we have

$$\begin{aligned}
\text{RidgeRisk} &\gtrsim \text{RidgeBiasBound} + \text{RidgeVarianceBound} \\
&= \Omega\bigg(\frac{\widetilde{\lambda}^2}{N^2_{\rm ridge}} + \frac{N_{\rm ridge}}{\widetilde{\lambda}^2}\bigg) \qquad \text{since } \sum_{i\ge k^*}\lambda_i^2 = \Theta(1) \\
&= \Omega\bigg(\frac{1}{N^{1/2}_{\rm ridge}}\bigg). \qquad \text{by the fact that } a+b \ge \sqrt{ab}
\end{aligned}$$

Therefore, it is evident that ridge regression is guaranteed to be worse than SGD if $N_{\rm ridge} \le N^2_{\rm sgd}/\log^2(N_{\rm sgd})$. This completes the proof.

7 Conclusions

We conduct an instance-based risk comparison between SGD and ridge regression for a broad class of least square problems. We show that SGD is always no worse than ridge regression provided logarithmically more samples. On the other hand, there exist some instances where even optimally-tuned ridge regression needs quadratically more samples to compete with SGD. This separation in terms of sample inflation between SGD and ridge regression suggests a provable benefit of implicit regularization over explicit regularization for least squares problems. In the future, we will explore the benefits of implicit regularization for learning other linear models and potentially nonlinear models.

Acknowledgments and Disclose of Funding

We would like to thank the anonymous reviewers and area chairs for their helpful comments. DZ is supported by the Bloomberg Data Science Ph.D. Fellowship. JW is supported in part by NSF CAREER grant 1652257. VB is supported in part by NSF CAREER grant 1652257, ONR Award N00014-18-1-2364 and the Lifelong Learning Machines program from DARPA/MTO. QG is supported in part by the National Science Foundation awards IIS-1855099 and IIS-2008981. SK acknowledges funding from the National Science Foundation under Award CCF-1703574. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agencies.

References

[1] Alnur Ali, J Zico Kolter, and Ryan J Tibshirani. A continuous-time view of early stopping for least squares regression. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1370–1378. PMLR, 2019.

[2] Alnur Ali, Edgar Dobriban, and Ryan Tibshirani. The implicit regularization of stochastic gradient flow for least squares. In *International Conference on Machine Learning*, pages 233–244. PMLR, 2020.

[3] Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. *arXiv preprint arXiv:1905.13655*, 2019.

[4] 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.

[5] R. R. Bahadur. Rates of convergence of estimates and test statistics. *Annals of Mathematical Statistics*, 38:303–324, 1967.

[6] R. R. Bahadur. *Some Limit Theorems in Statistics*. Society for Industrial and Applied Mathematics, 1971.

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

[8] Assaf Dauber, Meir Feder, Tomer Koren, and Roi Livni. Can implicit bias explain generalization? stochastic convex optimization as a case study. *arXiv preprint arXiv:2003.06152*, 2020.

[9] 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.

[10] Paramveer S Dhillon, Dean P Foster, Sham M Kakade, and Lyle H Ungar. A risk comparison of ordinary least squares vs ridge regression. *The Journal of Machine Learning Research*, 14(1): 1505–1511, 2013.

[11] 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.

[12] Edgar Dobriban, Stefan Wager, et al. High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics*, 46(1):247–279, 2018.

[13] Jerome Friedman, Trevor Hastie, Robert Tibshirani, et al. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

[14] Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning*, pages 1832–1841. PMLR, 2018.

[15] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019.

[16] Daniel Hsu, Sham M Kakade, and Tong Zhang. Random design analysis of ridge regression. In *Conference on learning theory*, pages 9–1. JMLR Workshop and Conference Proceedings, 2012.

[17] 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). *arXiv preprint arXiv:1710.09430*, 2017.

[18] 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.

[19] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

[20] Dmitry Kobak, Jonathan Lomond, and Benoit Sanchez. The optimal ridge penalty for real-world high-dimensional data can be zero or negative due to the implicit ridge regularization. *Journal of Machine Learning Research*, 21(169):1–16, 2020.

[21] Shengchao Liu, Dimitris Papailiopoulos, and Dimitris Achlioptas. Bad global minima exist and sgd can reach them. *arXiv preprint arXiv:1906.02613*, 2019.

[22] Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. *arXiv preprint arXiv:1412.6614*, 2014.

[23] Noam Razin and Nadav Cohen. Implicit regularization in deep learning may not be explainable by norms. *arXiv preprint arXiv:2005.06398*, 2020.

[24] Arun Suggala, Adarsh Prasad, and Pradeep K Ravikumar. Connecting optimization and regularization paths. *Advances in Neural Information Processing Systems*, 31:10608–10619, 2018.

[25] Andrei Nikolajevits Tihonov. Solution of incorrectly formulated problems and the regularization method. *Soviet Math.*, 4:1035–1038, 1963.

[26] Alexander Tsigler and Peter L Bartlett. Benign overfitting in ridge regression. *arXiv preprint arXiv:2009.14286*, 2020.

[27] Denny Wu and Ji Xu. On the optimal weighted \ell_2 regularization in overparameterized linear regression. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 10112–10123. Curran Associates, Inc., 2020.

[28] Ji Xu and Daniel Hsu. On the number of variables to use in principal component regression. *arXiv preprint arXiv:1906.01139*, 2019.

[29] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

[30] Difan Zou, Jingfeng Wu, Vladimir Braverman, Quanquan Gu, and Sham M Kakade. Benign overfitting of constant-stepsize sgd for linear regression. *arXiv preprint arXiv:2103.12692*, 2021.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] This paper focuses on theoretical explanations of the implicit regularization of SGD and its comparison to explicit regularization in ridge regression. It has no potential negative societal impact.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Proof of One-hot Least Squares

A.1 Excess risk bound of SGD

In this part we will mainly follow the proof technique in Zou et al. [30] that is developed to sharply characterize the excess risk bound for SGD (with tail-averaging) when the data distribution has a nice finite fourth-moment bound. However, such condition does not hold for the one-hot case so that their results cannot be directly applied here.

Before presenting the detailed proofs, we first introduce some notations and definitions that will be repeatedly used in the subsequent analysis. Let $\mathbf{H} = \mathbb{E}[\mathbf{x}\mathbf{x}^\top]$ be the covariance of data distribution. It is easy to verify that $\mathbf{H}$ is a diagonal matrix with eigenvalues $\lambda_1, \ldots, \lambda_d$. Let $\mathbf{w}_t$ be the t-th iterate of the SGD, we define $\boldsymbol{\eta}_t := \mathbf{w}_t - \mathbf{w}^*$ as the centered SGD iterate. Then we define $\boldsymbol{\eta}_t^{\text{bias}}$ and $\boldsymbol{\eta}_t^{\text{variance}}$ as the bias error and variance error respectively, which are described by the following update rule:

$$\begin{aligned} \boldsymbol{\eta}_t^{\text{bias}} &= \big(\mathbf{I} - \gamma \mathbf{x}_t \mathbf{x}_t^\top\big)\boldsymbol{\eta}_{t-1}^{\text{bias}}, \qquad \boldsymbol{\eta}_0^{\text{bias}} = \boldsymbol{\eta}_0, \\ \boldsymbol{\eta}_t^{\text{variance}} &= \big(\mathbf{I} - \gamma \mathbf{x}_t \mathbf{x}_t^\top\big)\boldsymbol{\eta}_{t-1}^{\text{bias}} + \gamma \xi_t \mathbf{x}_t, \qquad \boldsymbol{\eta}_0^{\text{variance}} = \mathbf{0}. \end{aligned} \tag{A.1}$$

Accordingly, we can further define the bias covariance $\mathbf{B}_t$ and variance covariance $\mathbf{C}_t$ as follows

$$\mathbf{B}_t = \mathbb{E}[\boldsymbol{\eta}_t^{\text{bias}} \otimes \boldsymbol{\eta}_t^{\text{bias}}], \qquad \mathbf{C}_t = \mathbb{E}[\boldsymbol{\eta}_t^{\text{variance}} \otimes \boldsymbol{\eta}_t^{\text{variance}}].$$

Regarding these two covariance matrices, the following lemma mathematically characterizes the upper bounds of the diagonal entries of $\mathbf{B}_t$ and $\mathbf{C}_t$.

Lemma A.1. *Under Assumptions 3.1, let $\bar{\mathbf{B}}_t = \operatorname{diag}(\mathbf{B}_t)$ and $\bar{\mathbf{C}}_t = \operatorname{diag}(\mathbf{C}_t)$, then if the stepsize satisfies $\gamma \leq 1$, we have*

$$\bar{\mathbf{B}}_t \preceq (\mathbf{I} - \gamma\mathbf{H})\bar{\mathbf{B}}_{t-1}, \qquad \bar{\mathbf{C}}_t \preceq (\mathbf{I} - \gamma\mathbf{H})\bar{\mathbf{C}}_{t-1} + \gamma^2\sigma^2\mathbf{H}.$$

Proof. According to (A.1), we have

$$\begin{aligned} \mathbf{B}_t = \mathbb{E}[\boldsymbol{\eta}_t^{\text{bias}} \otimes \boldsymbol{\eta}_t^{\text{bias}}] &= \mathbb{E}\big[(\mathbf{I} - \gamma \mathbf{x}_t \mathbf{x}_t^\top)\boldsymbol{\eta}_{t-1}^{\text{bias}} \otimes (\mathbf{I} - \gamma \mathbf{x}_t \mathbf{x}_t^\top)\boldsymbol{\eta}_{t-1}^{\text{bias}}\big] \\ &= \mathbf{B}_{t-1} - \gamma\mathbf{H}\mathbf{B}_{t-1} - \gamma\mathbf{B}_{t-1}\mathbf{H} + \gamma^2\mathbb{E}[\mathbf{x}_t\mathbf{x}_t^\top\mathbf{B}_{t-1}\mathbf{x}_t\mathbf{x}_t^\top]. \end{aligned} \tag{A.2}$$

Note that $\mathbf{x}_t = \mathbf{e}_i$ with probability λ_i, then we have

$$\begin{aligned} \mathbb{E}[\mathbf{x}_t\mathbf{x}_t^\top\mathbf{B}_{t-1}\mathbf{x}_t\mathbf{x}_t^\top] &= \sum_i \lambda_i \cdot \mathbf{e}_i\mathbf{e}_i^\top\mathbf{B}_{t-1}\mathbf{e}_i\mathbf{e}_i^\top \\ &= \sum_i \lambda_i \cdot \mathbf{e}_i^\top\mathbf{B}_{t-1}\mathbf{e}_i \cdot \mathbf{e}_i\mathbf{e}_i^\top \\ &= \bar{\mathbf{B}}_{t-1}\mathbf{H}. \end{aligned}$$

Plugging the above equation into (A.2) gives

$$\mathbf{B}_t = \mathbf{B}_{t-1} - \gamma\mathbf{H}\mathbf{B}_{t-1} - \gamma\mathbf{B}_{t-1}\mathbf{H} + \gamma^2\bar{\mathbf{B}}_{t-1}\mathbf{H}.$$

Then if only look at the diagonal entries of both sides, we have

$$\bar{\mathbf{B}}_t = \bar{\mathbf{B}}_{t-1} - 2\gamma\mathbf{H}\bar{\mathbf{B}}_{t-1} + \gamma^2\mathbf{H}\bar{\mathbf{B}}_{t-1} \preceq (\mathbf{I} - \gamma\mathbf{H})\bar{\mathbf{B}}_{t-1},$$

where in the first equation we use the fact that $\operatorname{diag}(\mathbf{H}\mathbf{B}) = \operatorname{diag}(\mathbf{B}\mathbf{H}) = \mathbf{H}\bar{\mathbf{B}}$ and the inequality follows from the fact that both $\bar{\mathbf{B}}_t$ and $\mathbf{H}$ are diagonal and $\gamma \leq 1$.

Similarly, regarding $\mathbf{C}_t$ the following holds according to (A.1),

$$\mathbf{C}_t = \mathbb{E}\big[(\mathbf{I} - \gamma \mathbf{x}_t \mathbf{x}_t^\top)\boldsymbol{\eta}_{t-1}^{\text{variance}} \otimes (\mathbf{I} - \gamma \mathbf{x}_t \mathbf{x}_t^\top)\boldsymbol{\eta}_{t-1}^{\text{variance}}\big] + \gamma^2\mathbb{E}[\xi_t^2\mathbf{x}_t\mathbf{x}_t^\top],$$

where we use the fact that $\mathbb{E}[\xi_t|\mathbf{x}_t] = 0$. Similar to deriving the bound for $\bar{\mathbf{B}}_t$, we have

$$\operatorname{diag}\big(\mathbb{E}\big[(\mathbf{I} - \gamma \mathbf{x}_t \mathbf{x}_t^\top)\boldsymbol{\eta}_t^{\text{variance}} \otimes (\mathbf{I} - \gamma \mathbf{x}_t \mathbf{x}_t^\top)\boldsymbol{\eta}_t^{\text{variance}}\big]\big) \preceq (\mathbf{I} - \gamma\mathbf{H})\bar{\mathbf{C}}_{t-1}.$$

Besides, under Assumption 3.1 we also have $\mathbb{E}[\xi_t^2\mathbf{x}_t\mathbf{x}_t^\top] = \sigma^2\mathbf{H}$, which is a diagonal matrix. Based on these two results, we can get the following upper bound for $\bar{\mathbf{C}}_t$,

$$\begin{aligned} \bar{\mathbf{C}}_t &= \operatorname{diag}\big(\mathbb{E}\big[(\mathbf{I} - \gamma \mathbf{x}_t \mathbf{x}_t^\top)\boldsymbol{\eta}_{t-1}^{\text{variance}} \otimes (\mathbf{I} - \gamma \mathbf{x}_t \mathbf{x}_t^\top)\boldsymbol{\eta}_{t-1}^{\text{variance}}\big] + \gamma^2\mathbb{E}[\xi_t^2\mathbf{x}_t\mathbf{x}_t^\top]\big) \\ &\preceq (\mathbf{I} - \gamma\mathbf{H})\bar{\mathbf{C}}_{t-1} + \gamma^2\sigma^2\mathbf{H}. \end{aligned}$$

This completes the proof. □

Lemma A.2 (Lemmas D.1 & D.2 in Zou et al. [30]). *Let $\bar{\mathbf{w}}_{N:2N}$ be the output of tail-averaged SGD, then if the stepsize satisfied $\gamma \le 1/\lambda_1$, it holds that*

$$\mathbb{E}[L(\bar{\mathbf{w}}_{N:2N})] - L(\mathbf{w}^*) \lesssim \mathrm{SGDBias} + \mathrm{SGDVariance},$$

where

$$\mathrm{SGDBias} \le \frac{1}{N^2}\sum_{t=0}^{N-1}\sum_{k=t}^{N-1}\left\langle (\mathbf{I}-\gamma\mathbf{H})^{k-t}\mathbf{H}, \mathbf{B}_{N+t}\right\rangle$$

$$\mathrm{SGDVariance} \le \frac{1}{N^2}\sum_{t=0}^{N-1}\sum_{k=t}^{N-1}\left\langle (\mathbf{I}-\gamma\mathbf{H})^{k-t}\mathbf{H}, \mathbf{C}_{N+t}\right\rangle$$

Lemma A.3. *Under Assumptions 3.1, if the stepsize satisfies $\gamma \le 1$ and set $\mathbf{w}_0 = \mathbf{0}$, then*

$$\mathbb{E}[L(\bar{\mathbf{w}}_{N:2N})] - L(\mathbf{w}^*) \le 2\cdot \mathrm{bias} + 2\cdot \mathrm{variance},$$

where

$$\mathrm{bias} \lesssim \frac{1}{N^2\gamma^2}\cdot \left\|(\mathbf{I}-\gamma\mathbf{H})^{N/2}\mathbf{w}^*\right\|_{\mathbf{H}^{-1}_{0:k_1}} + \left\|(\mathbf{I}-\gamma\mathbf{H})^{N/2}\mathbf{w}^*\right\|^2_{\mathbf{H}_{k_1:\infty}}$$

$$\mathrm{variance} \lesssim \sigma^2\cdot\left(\frac{k_2}{N} + N\gamma^2\sum_{i>k_2}\lambda_i^2\right)$$

for arbitrary $k_1, k_2 \in [d]$.

Proof. The first conclusion of this theorem can be directly proved via Young's inequality.

Note that $\mathbf{H}$ is a diagonal matrix, and thus $(\mathbf{I}-\gamma\mathbf{H})^{k-t}$ is also a diagonal matrix for all k and t. Therefore, by Lemma A.2, it is clear that in order to calculate the upper bound of the bias and variance error, it suffices to consider the diagonal entries of $\mathbf{B}_{N+t}$ and $\mathbf{C}_{N+t}$, denoted by $\bar{\mathbf{B}}_{N+t}$ and $\bar{\mathbf{C}}_{N+t}$ (which are obtained by setting all non-diagonal entries of $\mathbf{B}_{N+t}$ and $\mathbf{C}_{N+t}$ as zero). Then by Young's inequality, Lemma A.2 implies that

$$\mathrm{bias} \le \frac{1}{N^2}\sum_{t=0}^{N-1}\sum_{k=t}^{N-1}\left\langle (\mathbf{I}-\gamma\mathbf{H})^{k-t}\mathbf{H}, \bar{\mathbf{B}}_{N+t}\right\rangle$$

$$\mathrm{variance} \le \frac{1}{N^2}\sum_{t=0}^{N-1}\sum_{k=t}^{N-1}\left\langle (\mathbf{I}-\gamma\mathbf{H})^{k-t}\mathbf{H}, \bar{\mathbf{C}}_{N+t}\right\rangle. \tag{A.3}$$

Now we are ready to precisely calculate the above two bounds. In particular, by Lemma A.1 we have

$$\bar{\mathbf{B}}_t \preceq (\mathbf{I}-\gamma\mathbf{H})\bar{\mathbf{B}}_{t-1} \preceq (\mathbf{I}-\gamma\mathbf{H})^t\mathbf{B}_0, \tag{A.4}$$

$$\bar{\mathbf{C}}_t \preceq (\mathbf{I}-\gamma\mathbf{H})\bar{\mathbf{C}}_{t-1} \preceq \sum_{s=0}^{t-1}\sigma^2\gamma^2(\mathbf{I}-\gamma\mathbf{H})^s\mathbf{H} = \sigma^2\gamma\big(\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^t\big), \tag{A.5}$$

where in the second inequality we use the fact that $\mathbf{C}_0 = \boldsymbol{\eta}_t^{\mathrm{variance}}\otimes\boldsymbol{\eta}_t^{\mathrm{variance}} = \mathbf{0}$. Then plugging (A.4) into (A.3) gives

$$\begin{aligned}
\mathrm{bias} &\le \frac{1}{N^2}\sum_{t=0}^{N-1}\sum_{k=t}^{N-1}\langle(\mathbf{I}-\gamma\mathbf{H})^{k-t}\mathbf{H}, (\mathbf{I}-\gamma\mathbf{H})^{N+t}\mathbf{B}_0\rangle \\
&= \frac{1}{N^2}\left\langle\sum_{k=0}^{N-1-t}(\mathbf{I}-\gamma\mathbf{H})^k\mathbf{H}, \sum_{t=0}^{N-1}(\mathbf{I}-\gamma\mathbf{H})^{N+t}\mathbf{B}_0\right\rangle \\
&\le \frac{1}{N^2}\left\langle\sum_{k=0}^{N-1}(\mathbf{I}-\gamma\mathbf{H})^k\mathbf{H}, \sum_{t=0}^{N-1}(\mathbf{I}-\gamma\mathbf{H})^{N+t}\mathbf{B}_0\right\rangle \\
&= \frac{1}{N^2\gamma^2}\Big\langle\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^N, \mathbf{H}^{-1}(\mathbf{I}-\gamma\mathbf{H})^N\big(\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^N\big)\mathbf{B}_0\Big\rangle \\
&= \frac{1}{N^2\gamma^2}\Big\langle(\mathbf{I}-\gamma\mathbf{H})^N\big[\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^N\big]^2\mathbf{H}^{-1}, \mathbf{B}_0\Big\rangle
\end{aligned} \tag{A.6}$$

Note that $(1-x)^N \geq \min\{0, 1-Nx\}$ for all $x \in [0,1]$. Then for all i we have

$$\big[1-(1-\gamma\lambda_i)^N\big]^2\lambda^{-1} \leq \min\left\{\frac{1}{\lambda_i}, N^2\gamma^2\lambda_i\right\}$$

where we use the fact that $\gamma \leq 1 \leq 1/\lambda_i$ for all i. This further implies that

$$\big[\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^N\big]^2\mathbf{H}^{-1} \preceq \mathbf{H}_{0:k}^{-1} + N^2\gamma^2\mathbf{H}_{k:\infty}$$

for all $k \in [d]$. Plugging the above results into (A.6) leads to

$$\text{bias} \leq \frac{1}{N^2\gamma^2}\cdot\big\langle\mathbf{H}_{0:k}^{-1}, (\mathbf{I}-\gamma\mathbf{H})^N\mathbf{B}_0\big\rangle + \big\langle\mathbf{H}_{k:\infty}, (\mathbf{I}-\gamma\mathbf{H})^N\mathbf{B}_0\big\rangle \tag{A.7}$$

for all $k \in [d]$. Further note that $\mathbf{B}_0 = (\mathbf{w}_0-\mathbf{w}^*)\otimes(\mathbf{w}_0-\mathbf{w}^*) = \mathbf{w}^*\otimes\mathbf{w}^*$ as we pick $\mathbf{w}_0 = \mathbf{0}$. Thus (A.7) implies that

$$\text{bias} \leq \frac{1}{N^2\gamma^2}\cdot\big\|(\mathbf{I}-\gamma\mathbf{H})^{N/2}\mathbf{w}^*\big\|_{\mathbf{H}_{0:k}^{-1}}^2 + \big\|(\mathbf{I}-\gamma\mathbf{H})^{N/2}\mathbf{w}^*\big\|_{\mathbf{H}_{k:\infty}}^2.$$

Then we will deal with the variance error. Plugging (A.5) into (A.3) gives

$$\begin{aligned}
\text{variance} &\leq \frac{\sigma^2\gamma}{N^2}\sum_{t=0}^{N-1}\sum_{k=t}^{N-1}\big\langle(\mathbf{I}-\gamma\mathbf{H})^{k-t}\mathbf{H}, \mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{N+t}\big\rangle \\
&\leq \frac{\sigma^2\gamma}{N^2}\sum_{t=0}^{N-1}\left\langle\sum_{k=0}^{N-1}(\mathbf{I}-\gamma\mathbf{H})^k\mathbf{H}, \mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{N+t}\right\rangle \\
&= \frac{\sigma^2}{N^2}\sum_{t=0}^{N-1}\Big\langle\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^N, \mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{N+t}\Big\rangle \\
&\leq \frac{\sigma^2}{N}\Big\langle\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{2N}, \mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{2N}\Big\rangle.
\end{aligned}$$

We then use the inequality $(1-x)^N \geq \min\{0, 1-xN\}$ again and thus the above inequality further leads to

$$\begin{aligned}
\text{variance} &\leq \frac{\sigma^2}{N}\cdot\sum_i \min\{1, 4N^2\gamma^2\lambda_i^2\} \\
&\leq \frac{4\sigma^2}{N}\cdot\left(k + N^2\gamma^2\sum_{i>k}\lambda_i^2\right)
\end{aligned}$$

for any $k \in [d]$. □

A.2 Excess risk bound of ridge regression

Lemma A.4. *Let $\mathbf{X} \in \mathbb{R}^{N\times d}$ be the training data matrix and $\mathbf{w}_{\text{ridge}}(N;\lambda)$ be the solution of ridge regression with parameter λ and sample size N, then for any $\lambda > 0$*

$$\mathbb{E}[L(\mathbf{w}_{\text{ridge}}(N;\lambda))] - L(\mathbf{w}^*) = \text{bias} + \text{variance},$$

where

$$\begin{aligned}
\text{bias} &= \lambda^2\cdot\mathbb{E}\big[\mathbf{w}^{*\top}(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{H}(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{w}^*\big] \\
\text{variance} &= \sigma^2\cdot\mathbb{E}\big[\,\text{tr}\,\big((\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{H}\big)\big],
\end{aligned}$$

where the expectations are taken over the randomness of the training data matrix $\mathbf{X}$.

Proof. Recall that the solution of ridge regression takes form

$$\mathbf{w}_{\text{ridge}}(N;\lambda) = (\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y},$$

where $\mathbf{X}$ is the data matrix and $\mathbf{y}$ is the response vector. Then according to the definition of the loss function $L(\mathbf{w})$, we have

$$\begin{aligned}\mathbb{E}[L(\mathbf{w}_{\text{ridge}}(N;\lambda))] &= \mathbb{E}\Big[\big(y - \langle \mathbf{w}_{\text{ridge}}(N;\lambda), \mathbf{x}\rangle\big)^2\Big] \\ &= \mathbb{E}\Big[\big(\langle \mathbf{w}^*, \mathbf{x}\rangle - \langle \mathbf{w}_{\text{ridge}}(N;\lambda), \mathbf{x}\rangle\big)^2\Big] + \mathbb{E}\Big[\big(y - \langle \mathbf{w}^*, \mathbf{x}\rangle\big)^2\Big] \\ &\quad + 2\mathbb{E}\big[\big(\langle \mathbf{w}^*, \mathbf{x}\rangle - \langle \mathbf{w}_{\text{ridge}}(N;\lambda), \mathbf{x}\rangle\big) \cdot \big(y - \langle \mathbf{w}^*, \mathbf{x}\rangle\big)\big] \\ &= \mathbb{E}[\|\mathbf{w}_{\text{ridge}}(N;\lambda) - \mathbf{w}^*\|_{\mathbf{H}}^2] + L(\mathbf{w}^*),\end{aligned}$$

where the last equation is by Assumption 3.1. Then regarding $\mathbb{E}[\|\mathbf{w}_{\text{ridge}}(N;\lambda) - \mathbf{w}^*\|_{\mathbf{H}}^2]$, let $\boldsymbol{\xi} = \mathbf{y} - \mathbf{X}\mathbf{w}^*$ be the model noise vector, we have

$$\begin{aligned}\mathbb{E}[\|\mathbf{w}_{\text{ridge}}(N;\lambda) - \mathbf{w}^*\|_{\mathbf{H}}^2] &= \mathbb{E}\big[\big\|(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y} - \mathbf{w}^*\big\|_{\mathbf{H}}^2\big] \\ &= \mathbb{E}\big[\big\|(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top(\mathbf{X}\mathbf{w}^* + \boldsymbol{\xi}) - \mathbf{w}^*\big\|_{\mathbf{H}}^2\big] \\ &= \underbrace{\mathbb{E}\big[\big\|(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{X}\mathbf{w}^* - \mathbf{w}^*\big\|_{\mathbf{H}}^2\big]}_{\text{bias}} + \underbrace{\mathbb{E}\big[\big\|(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\boldsymbol{\xi}\big\|_{\mathbf{H}}^2\big]}_{\text{variance}}.\end{aligned}$$

where in the last inequality we again apply Assumption 3.1 that $\mathbb{E}[\boldsymbol{\xi}|\mathbf{X}] = \mathbf{0}$. More specifically, the bias error can be reformulated as

$$\begin{aligned}\text{bias} &= \mathbb{E}\big[\big\|\big((\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{X} - \mathbf{I}\big)\mathbf{w}^*\big\|_{\mathbf{H}}^2\big] \\ &= \lambda^2\mathbb{E}\big[\big\|(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{w}\big\|_{\mathbf{H}}^2\big] \\ &= \lambda^2\mathbb{E}\big[\mathbf{w}^{*\top}(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{H}(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{w}^*\big].\end{aligned}$$

In terms of the variance error, note that by Assumption 3.1 we have $\mathbb{E}[\boldsymbol{\xi}\boldsymbol{\xi}^\top|\mathbf{X}] = \sigma^2\mathbf{I}$, then

$$\begin{aligned}\text{variance} &= \mathbb{E}\big[\big\|(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\boldsymbol{\epsilon}\big\|_{\mathbf{H}}^2\big] \\ &= \mathbb{E}\big[\operatorname{tr}\big((\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\boldsymbol{\xi}\boldsymbol{\xi}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{H}\big)\big] \\ &= \sigma^2 \cdot \mathbb{E}\big[\operatorname{tr}\big((\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{H}\big)\big].\end{aligned}$$

□

Lemma A.5. *The solution of ridge regression with sample size N and regularization parameter λ satisfies*

$$\mathbb{E}[L(\mathbf{w}_{\text{ridge}}(N;\lambda))] - L(\mathbf{w}^*) = \text{RidgeBias} + \text{RidgeVariance},$$

where

$$\begin{aligned}\text{RidgeBias} &\gtrsim \max\bigg\{\sum_i (1-\lambda_i)^N \cdot \lambda_i \mathbf{w}^*[i]^2, \sum_{i=1}^{k^*} \frac{\lambda^2\lambda_i\mathbf{w}^*[i]^2}{(N\lambda_i + \lambda)^2} + \sum_{i>k^*}\lambda_i\mathbf{w}^*[i]^2\bigg\} \\ \text{RidgeVariance} &\gtrsim \sigma^2 \cdot \bigg(\sum_{i=1}^{k^*}\frac{N\lambda_i^2}{(N\lambda_i+\lambda)^2} + \sum_{i>k^*}\frac{N\lambda_i^2}{(1+\lambda)^2}\bigg),\end{aligned}$$

where $k^* = \min\{k : N\lambda_k \le 1\}$.

Proof. In the one-hot case, it is easy to verify that $\mathbf{X}^\top\mathbf{X} = \sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top$ is a diagonal matrix. Let $\mu_1, \mu_2, \ldots, \mu_d$ be the eigenvalues of $\mathbf{X}^\top\mathbf{X}$ corresponding to the eigenvectors $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_d$ respectively. Then by Lemma A.4, we have the following results for the bias and variance errors of ridge regression.

$$\begin{aligned}\text{RidgeBias} &= \lambda^2 \cdot \mathbb{E}\big[\mathbf{w}^{*\top}(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{H}(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{w}^*\big] \\ &= \lambda^2\sum_i \mathbb{E}_{\mu_i}\bigg[\frac{\lambda_i\mathbf{w}^*[i]^2}{(\mu_i+\lambda)^2}\bigg], \qquad \text{(A.8)}\end{aligned}$$

where the expectation in the first equation is taken over the training data $\mathbf{X}$ and in the second inequality the expectation is equivalently taken over the eigenvalues $\mu_1, \ldots, \mu_d$. Since $\mathbf{x}_i$ can only take on natural basis, the eigenvalue μ_i can be understood as the number of training data that equals $\mathbf{e}_i$. Note that the probability of sampling $\mathbf{e}_i$ is λ_i, then we can get that μ_i has a marginal distribution $\mathrm{Binom}(N, \lambda_i)$, where N is the sample size. Then in terms of each expectation in (A.8), we first have

$$\mathbb{E}_{\mu_i}\left[\frac{\lambda_i \mathbf{w}^*[i]^2}{(\mu_i+\lambda)^2}\right] \geq \frac{\lambda_i \mathbf{w}^*[i]^2}{(\mathbb{E}[\mu_i]+\lambda)^2} = \frac{\lambda_i \mathbf{w}^*[i]^2}{(N\lambda_i+\lambda)^2},$$

where the first inequality is by applying Jensen's inequality to the convex function $f(x) = 1/(x+\lambda)^2$. On the other hand, we also have

$$\mathbb{E}_{\mu_i}\left[\frac{\lambda_i \mathbf{w}^*[i]^2}{(\mu_i+\lambda)^2}\right] \geq \frac{\lambda_i \mathbf{w}^*[i]^2}{\lambda^2} \cdot \mathbb{P}(\mu_i = 0) = \frac{\lambda_i \mathbf{w}^*[i]^2}{\lambda^2} \cdot (1-\lambda_i)^N.$$

Therefore, combining the above two lower bounds, we can get the following lower bound on the bias error by (A.8)

$$\mathrm{RidgeBias} = \lambda^2 \sum_i \mathbb{E}_{\mu_i}\left[\frac{\lambda_i \mathbf{w}^*[i]^2}{(\mu_i+\lambda)^2}\right] \geq \sum_i \max\left\{\frac{\lambda^2 \lambda_i \mathbf{w}^*[i]^2}{(N\lambda_i+\lambda)^2}, \lambda_i \mathbf{w}^*[i]^2 \cdot (1-\lambda_i)^N\right\}. \quad \text{(A.9)}$$

Therefore, a trivial lower bound on the bias error of ridge regression is

$$\mathrm{RidgeBias} \geq \sum_i (1-\lambda_i)^N \cdot \lambda_i \mathbf{w}^*[i]^2.$$

Additionally, note that $(1-\lambda_i)^N \geq 0.25$ if $\lambda_i \leq 1/N$ and $N \geq 2$. Then let $k^* = \min\{k : N\lambda_k \leq 1\}$, (A.9) further leads to

$$\mathrm{RidgeBias} \geq \sum_{i=1}^{k^*} \frac{\lambda^2 \lambda_i \mathbf{w}^*[i]^2}{(N\lambda_i+\lambda)^2} + 0.25 \cdot \sum_{i>k^*} \lambda_i \mathbf{w}^*[i]^2.$$

This completes the proof of the lower bound of the bias error.

By Lemma A.4, we have

$$\begin{aligned}\mathrm{RidgeVariance} &= \sigma^2 \cdot \mathbb{E}\big[\,\mathrm{tr}\left((\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{H}\right)\big] \\ &= \sigma^2 \cdot \sum_i \mathbb{E}_{\mu_i}\left[\frac{\lambda_i \mu_i}{(\mu_i+\lambda)^2}\right], \quad \text{(A.10)}\end{aligned}$$

Regarding the variance error, we cannot use the similar approach since the function $g(x) = x/(x+\lambda)^2$ is no longer convex. Instead, we will directly make use of property of the binomial distribution of μ_i to prove the desired bound. In particular, note that $\mu_i \sim \mathrm{binom}(N, \lambda_i)$, by Bernstein inequality, we have

$$\mathbb{P}(|\mu_i - N\lambda_i| \leq t) \geq 1 - 2\exp\left(-\frac{t^2}{2(N\lambda_i + t/3)}\right).$$

If $N\lambda_i \geq 6$, by set $t = \sqrt{3N\lambda_i}$, we have

$$\mathbb{P}\big(\mu_i \in \big[N\lambda_i - \sqrt{3N\lambda_i}, N\lambda_i + \sqrt{3N\lambda_i}\big]\big) \geq 1 - 2e^{-1} \geq 0.2,$$

which further implies that

$$\mathbb{P}\big(\mu_i \in \big[0.25N\lambda_i, 2N\lambda_i\big]\big) \geq 0.2,$$

where we use the fact that $\sqrt{3N\lambda_i} \leq 0.75N\lambda_i$ if $N\lambda_i > 6$. Therefore, in this case, we can get

$$\mathbb{E}_{\mu_i}\left[\frac{\lambda_i \mu_i}{(\mu_i+\lambda)^2}\right] \geq 0.2 \min\left\{\frac{0.25N\lambda_i^2}{(0.25N\lambda_i+\lambda)^2}, \frac{2N\lambda_i^2}{(2N\lambda_i+\lambda)^2}\right\} \geq \frac{0.05N\lambda_i^2}{(N\lambda_i+\lambda)^2}. \quad \text{(A.11)}$$

Then we consider the case that $N\lambda_i < 6$. In particular, we have

$$\mathbb{E}_{\mu_i}\left[\frac{\lambda_i \mu_i}{(\mu_i+\lambda)^2}\right] \geq \frac{\lambda_i}{(1+\lambda)^2} \cdot \mathbb{P}(\mu_i = 1). \quad \text{(A.12)}$$

Note that μ_i follows $\text{Binom}(N, \lambda_i)$ distribution, which implies that

$$\mathbb{P}(\mu_i = 1) = N\lambda_i(1-\lambda_i)^{N-1} \geq N\lambda_i\big(1-\frac{6}{N}\big)^{N-1} \geq e^{-6}N\lambda_i.$$

Plugging this into (A.12) gives

$$\mathbb{E}_{\mu_i}\left[\frac{\lambda_i\mu_i}{(\mu_i+\lambda)^2}\right] \geq \frac{e^{-6}N\lambda_i^2}{(1+\lambda)^2}. \tag{A.13}$$

Therefore, let $k^* = \min\{k : N\lambda_k \leq 1\}$, then for all $i \leq k^*$, combining (A.11) and (A.13) gives

$$\mathbb{E}_{\mu_i}\left[\frac{\lambda_i\mu_i}{(\mu_i+\lambda)^2}\right] \geq \frac{e^{-6}N\lambda_i^2}{(N\lambda_i+\lambda)^2}.$$

For all $i > k^*$, we can directly apply (A.13) to get the lower bound. Therefore, according to (A.10), the variance error can be lower bounded as follows,

$$\begin{aligned}\text{RidgeVariance} &= \sigma^2 \cdot \sum_i \mathbb{E}_{\mu_i}\left[\frac{\lambda_i\mu_i}{(\mu_i+\lambda)^2}\right] \\ &\geq e^{-6}\sigma^2 \cdot \left(\sum_{i=1}^{k^*}\frac{N\lambda_i^2}{(N\lambda_i+\lambda)^2} + \sum_{i>k^*}\frac{N\lambda_i^2}{(1+\lambda)^2}\right).\end{aligned}$$

This completes the proof of the lower bound of the variance error. □

A.3 Proof of Theorem 4.1

Proof. In the beginning, we first recall the excess risk upper bound of SGD (see Lemma A.3) and excess risk lower bound of ridge (see Lemma A.3) as follows,

$$\mathbb{E}[L(\mathbf{w}_{\text{sgd}}(N_{\text{sgd}};\gamma))] - L(\mathbf{w}^*) \leq 2\cdot\text{SGDBias} + 2\cdot\text{SGDVariance},$$

where

$$\begin{aligned}\text{SGDBias} &\lesssim \frac{1}{N^2\gamma^2}\cdot\big\|(\mathbf{I}-\gamma\mathbf{H})^{N/2}\mathbf{w}^*\big\|_{\mathbf{H}_{0:k_1}^{-1}} + \big\|(\mathbf{I}-\gamma\mathbf{H})^{N/2}\mathbf{w}^*\big\|_{\mathbf{H}_{k_1:\infty}}^2 \\ \text{SGDVariance} &\lesssim \sigma^2\cdot\left(\frac{k_2}{N} + N\gamma^2\sum_{i>k_2}\lambda_i^2\right)\end{aligned} \tag{A.14}$$

for arbitrary $k_1, k_2 \in [d]$.

$$\mathbb{E}[L(\mathbf{w}_{\text{ridge}}(N;\lambda))] - L(\mathbf{w}^*) = \text{RidgeBias} + \text{RidgeVariance},$$

where

$$\begin{aligned}\text{RidgeBias} &\gtrsim \max\left\{\sum_i(1-\lambda_i)^N\cdot\lambda_i\mathbf{w}^*[i]^2, \sum_{i=1}^{k^*}\frac{\lambda^2\lambda_i\mathbf{w}^*[i]^2}{(N\lambda_i+\lambda)^2} + \sum_{i>k^*}\lambda_i\mathbf{w}^*[i]^2\right\} \\ \text{RidgeVariance} &\gtrsim \sigma^2\cdot\left(\sum_{i=1}^{k^*}\frac{N\lambda_i^2}{(N\lambda_i+\lambda)^2} + \sum_{i>k^*}\frac{N\lambda_i^2}{(1+\lambda)^2}\right),\end{aligned} \tag{A.15}$$

where $k^* = \min\{k : N\lambda_k \leq 1\}$.

Next, we will show that the excess risk of SGD can be provably upper bounded (up to constant factors) by the excess risk of ridge regression respectively, given the sample size of ridge regression N_{ridge} (which we will use N in the remaining proof for simplicity). In particular, we consider two cases regarding different λ: **Case I** $\lambda < 1$ and **Case II** $\lambda \geq 1$.

For **Case I**, (A.15) gives the following bias lower bound for ridge regression,

$$\begin{aligned}
\text{RidgeBias} &\gtrsim \sum_i (1-\lambda_i)^N \cdot \lambda_i \mathbf{w}^*[i]^2 \\
&\gtrsim \sum_{i>k^*} \lambda_i \mathbf{w}^*[i]^2 \\
\text{RidgeVariance} &\gtrsim \sigma^2 \cdot \left(\sum_{i=1}^{k^*} \frac{N\lambda_i^2}{(N\lambda_i+\lambda)^2} + \sum_{i>k^*} \frac{N\lambda_i^2}{(1+\lambda)^2} \right) \\
&\overset{(i)}{\eqsim} \sigma^2 \cdot \left(\frac{k^*}{N} + N \sum_{i>k^*} \lambda_i^2 \right),
\end{aligned}$$

where in (i) we use the fact that $N\lambda_i + \lambda \eqsim N\lambda_i$ for all $i \le k^*$.

Then let $R^2 = \|\mathbf{w}^*\|_2^2/\sigma^2$ denotes the signal-to-noise ratio, let's consider the following configuration for SGD:

$$N_{\text{sgd}} = N, \quad \gamma = 1.$$

Then by (A.14) and setting $k_1 = 0$ and $k_2 = k^*$, we get

$$\begin{aligned}
\text{SGDBias} &\lesssim \sum_i (1-\lambda_i)^N \cdot \lambda_i \mathbf{w}^*[i]^2 \\
\text{SGDVariance} &\lesssim \sigma^2 \cdot \left(\frac{k^*}{N_{\text{sgd}}} + N_{\text{sgd}}\gamma^2 \sum_{i>k^*} \lambda_i^2 \right) \\
&\overset{(i)}{\lesssim} \sigma^2 \cdot \left(\frac{k^*}{N} + N \sum_{i>k^*} \lambda_i^2 \right).
\end{aligned}$$

Therefore, given such choice of N_{sgd} and γ, we have

$$\begin{aligned}
\mathbb{E}[L(\mathbf{w}_{\text{sgd}}(N_{\text{sgd}};\gamma))] - L(\mathbf{w}^*) &\lesssim \text{SGDBias} + \text{SGDVariance} \\
&\lesssim \sum_i (1-\lambda_i)^N \cdot \lambda_i \mathbf{w}^*[i]^2 + \sigma^2 \cdot \left(\frac{k^*}{N} + N \sum_{i>k^*} \lambda_i^2 \right) \\
&\lesssim \text{RidgeBias} + \text{RidgeVariance} \\
&= \mathbb{E}[L(\mathbf{w}_{\text{ridge}(N;\lambda)})] - L(\mathbf{w}^*).
\end{aligned}$$

For **Case II**, we can define $\widetilde{k}^* = \min\{k : N\lambda_k \le \lambda\}$, then (A.15) implies

$$\begin{aligned}
\text{RidgeBias} &\gtrsim \sum_{i=1}^{k^*} \frac{\lambda^2 \lambda_i \mathbf{w}^*[i]^2}{(N\lambda_i+\lambda)^2} + \sum_{i>k^*} \lambda_i \mathbf{w}^*[i]^2 \\
&\overset{(i)}{\eqsim} \sum_{i=1}^{\widetilde{k}^*} \frac{\lambda^2 \mathbf{w}^*[i]^2}{N^2\lambda_i} + \sum_{i>\widetilde{k}^*} \lambda_i \mathbf{w}^*[i]^2 \\
\text{RidgeVariance} &\gtrsim \sigma^2 \cdot \left(\sum_{i=1}^{k^*} \frac{N\lambda_i^2}{(N\lambda_i+\lambda)^2} + \sum_{i>k^*} \frac{N\lambda_i^2}{(1+\lambda)^2} \right) \\
&\overset{(ii)}{\eqsim} \sigma^2 \cdot \left(\frac{\widetilde{k}^*}{N} + \frac{N}{\lambda^2} \sum_{i>\widetilde{k}^*} \lambda_i^2 \right),
\end{aligned}$$

where (i) and (ii) are due to the fact that for every $i \le k^*$, we have

$$\frac{1}{(N\lambda_i+\lambda)^2} \eqsim \begin{cases} \frac{1}{N^2\lambda_i} & i \le \widetilde{k}^* \\ \frac{1}{\lambda^2} & \widetilde{k}^* < i \le k^*. \end{cases}$$

Therefore, we can apply the following configuration for SGD:

$$N_{\mathrm{sgd}} = N, \quad \gamma = 1/\lambda.$$

Then by (A.14) and set $k_1 = k_2 = \widetilde{k}^*$, we have

$$\begin{aligned}
&\mathbb{E}[L(\mathbf{w}_{\mathrm{sgd}}(N_{\mathrm{sgd}};\gamma))] - L(\mathbf{w}^*) \\
&\lesssim \mathrm{SGDBias} + \mathrm{SGDVariance} \\
&\lesssim \sum_{i=1}^{\widetilde{k}^*} \frac{(1-\gamma\lambda_i)^{N_{\mathrm{sgd}}}\mathbf{w}^*[i]^2}{\lambda_i N_{\mathrm{sgd}}^2\gamma^2} + \sum_{i>\widetilde{k}^*}\lambda_i\mathbf{w}^*[i]^2 + \sigma^2\cdot\left(\frac{\widetilde{k}^*}{N_{\mathrm{sgd}}} + N_{\mathrm{sgd}}\gamma^2\sum_{i>\widetilde{k}^*}\lambda_i^2\right) \\
&\eqsim \sum_{i=1}^{\widetilde{k}^*} \frac{\lambda^2\mathbf{w}^*[i]^2}{\lambda_i N^2} + \sum_{i>\widetilde{k}^*}\lambda_i\mathbf{w}^*[i]^2 + \sigma^2\cdot\left(\frac{\widetilde{k}^*}{N} + \frac{N}{\lambda^2}\sum_{i>\widetilde{k}^*}\lambda_i^2\right) \\
&\lesssim \mathrm{RidgeBias} + \mathrm{RidgeVariance} \\
&= \mathbb{E}[L(\mathbf{w}_{\mathrm{ridge}}(N;\lambda))] - L(\mathbf{w}^*).
\end{aligned}$$

Combining the results for these two cases completes the proof. □

A.4 Proof of Theorem 4.2

Proof. For simplicity we define $N := N_{\mathrm{sgd}}$ in the proof.

- The data covariance matrix $\mathbf{H}$ has the following spectrum

$$\lambda_i = \begin{cases} \frac{\log(N)}{N^{1/2}} & i = 1, \\ \frac{1-\log(N)/N^{1/2}}{N} & 1 < i \le N, \\ 0 & N < i \le d \end{cases}$$

- The true parameter $\mathbf{w}^*$ is given by

$$\mathbf{w}^*[i] = \begin{cases} \sigma\cdot\sqrt{\frac{N^{1/2}}{\log(N)}} & i = 1, \\ 0 & 1 < i \le d. \end{cases}$$

Then it is easy to verify that $\mathrm{tr}(\mathbf{H}) = 1$. For SGD, we consider setting the stepsize as $\gamma^* = N^{-1/2}$. Then by Lemma A.3 and choosing $k_1 = 1$, we have the following on the bias error of SGD,

$$\mathrm{SGDBias} \lesssim \sum_{i=1}^{k^*} \frac{(1-\gamma\lambda_i)^{N_{\mathrm{sgd}}}\mathbf{w}^*[i]^2}{\lambda_i N_{\mathrm{sgd}}^2\gamma^2} + \sum_{i>k^*}\lambda_i\mathbf{w}^*[i]^2 \lesssim \frac{(1-\log(N)/N)^N\sigma^2}{\log^2(N)} \lesssim \frac{\sigma^2}{N}.$$

For variance error, we can pick $k_2 = 1$ and get

$$\mathrm{SGDVariance} \lesssim \sigma^2\cdot\left(\frac{1}{N} + N\gamma^2\sum_{i>1}\lambda_i^2\right) \lesssim \sigma^2\left(\frac{1}{N} + \sum_{i>1}\lambda_i^2\right) \eqsim \frac{\sigma^2}{N}.$$

Now let us characterize the excess risk of ridge regression. In terms of the bias error, by Lemma A.5 we have

$$\mathrm{RidgeBias} \gtrsim \sum_{i=1}^{k^*} \frac{\lambda^2\lambda_i\mathbf{w}^*[i]^2}{(N_{\mathrm{ridge}}\lambda_i + \lambda)^2} + \sum_{i>k^*}\lambda_i\mathbf{w}^*[i]^2 \eqsim \frac{\lambda^2\sigma^2}{(N_{\mathrm{ridge}}\log(N)/N^{1/2} + \lambda)^2}, \tag{A.16}$$

where $k^* = \min\{k : N_{\mathrm{ridge}}\lambda_k \le 1\}$. Then it is clear for ridge regression we must have $\lambda \lesssim N_{\mathrm{ridge}}\log(N)/N^{1/2}$ since otherwise $\mathrm{RidgeBias} \gtrsim \sigma^2 \gtrsim \mathrm{SGDRisk}$. Regarding the variance, we have

$$\mathrm{RidgeVariance} \gtrsim \sigma^2\cdot\left(\sum_{i=1}^{k^*}\frac{N_{\mathrm{ridge}}\lambda_i^2}{(N_{\mathrm{ridge}}\lambda_i + \lambda)^2} + \sum_{i>k^*}\frac{N_{\mathrm{ridge}}\lambda_i^2}{(1+\lambda)^2}\right).$$

Then we will consider two cases: (1) $N_{\text{ridge}} \lesssim N$ and (2) $N_{\text{ridge}} \gtrsim N$. In the first case we can get $k^* = 1$ and then

$$\text{RidgeVariance} \gtrsim \sigma^2 \cdot \left(\frac{N_{\text{ridge}} \log^2(N)/N^2}{(N_{\text{ridge}} \log(N)/N + \lambda)^2} + \frac{N_{\text{ridge}}}{N^2(1+\lambda)^2} \right) \geq \frac{N_{\text{ridge}} \sigma^2}{N^2(1+\lambda^2)}.$$

In this case, we can get $k^* = 1$ and thus

$$\text{RidgeVariance} \gtrsim \sigma^2 \cdot \frac{N_{\text{ridge}} \log^2(N)/N^2}{(N_{\text{ridge}} \log(N)/N + \lambda)^2} \overset{(i)}{\eqsim} \frac{\sigma^2}{N_{\text{ridge}}} \overset{(ii)}{\gtrsim} \frac{\sigma^2}{N},$$

where (i) is due to we require $\lambda \lesssim N_{\text{ridge}} \log(N)/N^{1/2}$ to guarantee vanishing bias error and (ii) is due to in this case we have $N_{\text{ridge}} \lesssim N$. As a result, ridge regression cannot achieve smaller excess risk than SGD in this case.

In the second case we can get $k^* = N$ and then

$$\begin{aligned} \text{RidgeVariance} &\gtrsim \sigma^2 \cdot \left(\frac{N_{\text{ridge}} \log^2(N)/N^2}{(N_{\text{ridge}} \log(N)/N + \lambda)^2} + \frac{(k^* - 1) \cdot N_{\text{ridge}}/N^2}{(N_{\text{ridge}}/N + \lambda^2)} \right) \\ &\gtrsim \sigma^2 \cdot \frac{N N_{\text{ridge}}}{N_{\text{ridge}}^2 + N^2 \lambda^2}, \end{aligned} \tag{A.17}$$

where the second inequality is due to $k^* = N$. We will again consider two cases: (a) $N_{\text{ridge}} \gtrsim N\lambda$ and (b) $N_{\text{ridge}} \lesssim N\lambda$. Regarding Case (a) we have

$$\text{RidgeVariance} \geq \frac{N\sigma^2}{N_{\text{ridge}}},$$

and it is clear that for all $N_{\text{ridge}} \lesssim N^2$ we have $\text{RidgeVariance} \gtrsim \sigma^2/N \gtrsim \text{SGDRisk}$. Regarding Case (b), combining the lower bounds of bias (A.16) and variance (A.17) of ridge regression, we get

$$\text{RidgeRisk} \gtrsim \sigma^2 \cdot \left(\frac{\lambda^2 N}{N_{\text{ridge}}^2 \log^2(N)} + \frac{N_{\text{ridge}}}{N\lambda^2} \right) \gtrsim \frac{\sigma^2}{N_{\text{ridge}}^{1/2} \log(N)},$$

where the first inequality follows from the fact that $\lambda \lesssim N_{\text{ridge}} \log(N)/N^{1/2}$ and $N_{\text{ridge}} \lesssim N\lambda$, and the second inequality is by Cauchy-Schwartz inequality. This further suggests that $\text{RidgeRisk} \lesssim \sigma^2/N \lesssim \text{SGDRisk}$ if $N_{\text{ridge}} \leq N^2/\log^2(N)$, which completes the proof.

□

B Proof of Gaussian Least Squares

B.1 Excess risk bounds of SGD and ridge regression

We first recall the excess risk bounds for SGD (with tail averaging) and ridge regression as follows.

SGD with tail averaging

Theorem B.1 (Extension of Theorem 5.1 in Zou et al. [30]). *Consider SGD with tail-averaging with initialization $\mathbf{w}_0 = \mathbf{0}$. Suppose Assumption 5.1 holds and the stepsize satisfies $\gamma \lesssim 1/\operatorname{tr}(\mathbf{H})$. Then the excess risk can be upper bounded as follows,*

$$\mathbb{E}[L(\mathbf{w}_{\text{sgd}}(N;\gamma))] - L(\mathbf{w}^*) \leq \text{SGDBias} + \text{SGDVariance},$$

where

$$\begin{aligned} \text{SGDBias} &\lesssim \frac{1}{\gamma^2 N^2} \cdot \left\| (\mathbf{I} - \gamma\mathbf{H})^N \mathbf{w}^* \right\|_{\mathbf{H}_{0:k_1}^{-1}}^2 + \left\| (\mathbf{I} - \gamma\mathbf{H})^N \mathbf{w}^* \right\|_{\mathbf{H}_{k_1:\infty}}^2 \\ \text{SGDVariance} &\lesssim \frac{\sigma^2 + \|\mathbf{w}^*\|_{\mathbf{H}}^2}{N} \cdot \left(k_2 + N^2 \gamma^2 \sum_{i > k_2} \lambda_i^2 \right). \end{aligned}$$

where $k_1, k_2 \in [d]$ are arbitrary.

This theorem is a simple extension of Theorem 5.1 in Zou et al. [30]. In particular, we observe that though the original theorem is stated for some particular k^* and $k^\dagger$, based on the proof, their results hold for arbitrary k_1 and k_2, as stated in Theorem B.1.

Ridge regression. See Appendix C for a proof of the following theorem.

Theorem B.2 (Extension of Lemmas 2 & 3 in Tsigler and Bartlett [26])**.** *Suppose Assumption 5.1 holds. Let $\lambda \geq 0$ be the regularization parameter, n be the training sample size and $\widehat{\mathbf{w}}_{\mathrm{ridge}}(N;\lambda)$ be the output of ridge regression. Then*

$$\mathbb{E}\big[L(\mathbf{w}_{\mathrm{ridge}}(N;\lambda))\big] - L(\mathbf{w}^*) = \mathrm{RidgeBias} + \mathrm{RidgeVariance},$$

and there is some absolute constant $b > 1$, such that for

$$k^*_{\mathrm{ridge}} := \min\left\{k : b\lambda_{k+1} \leq \frac{\lambda + \sum_{i>k}\lambda_i}{n}\right\},$$

the following holds:

$$\mathrm{RidgeBias} \gtrsim \left(\frac{\lambda + \sum_{i>k^*_{\mathrm{ridge}}}\lambda_i}{N}\right)^2 \cdot \|\mathbf{w}^*\|^2_{\mathbf{H}^{-1}_{0:k^*_{\mathrm{ridge}}}} + \|\mathbf{w}^*\|^2_{\mathbf{H}_{k^*_{\mathrm{ridge}}:\infty}},$$

$$\mathrm{RidgeVariance} \gtrsim \sigma^2 \cdot \left\{\frac{k^*_{\mathrm{ridge}}}{N} + \frac{N\sum_{i>k^*_{\mathrm{ridge}}}\lambda_i^2}{\left(\lambda + \sum_{i>k^*_{\mathrm{ridge}}}\lambda_i\right)^2}\right\}.$$

B.2 Proof of Theorem 5.1

Proof. For simplicity, let us fix $N := N_{\mathrm{ridge}}$ and $k := k_{\mathrm{ridge}}$, we will next locate γ such that the risk of SGD competes with that of Ridge. Denote $\widetilde{\lambda} := \lambda + \sum_{i>k}\lambda_i$. Then

$$\begin{aligned}\mathrm{RidgeRisk} &= \mathrm{RidgeBias} + \mathrm{RidgeVariance}\\ &\gtrsim \left(\frac{\widetilde{\lambda}}{N}\right)^2 \|\mathbf{w}^*\|^2_{\mathbf{H}^{-1}_{0:k}} + \|\mathbf{w}^*\|^2_{\mathbf{H}_{k:\infty}} + \frac{\sigma^2}{N}\left(k + \left(\frac{N}{\widetilde{\lambda}}\right)^2\sum_{i>k}\lambda_i^2\right).\end{aligned}$$

Then for SGD we can set

$$N_{\mathrm{sgd}} = (1+R^2)\cdot N \cdot (1 \vee \kappa\log a),$$

where

$$\kappa := \frac{\mathrm{tr}(\mathbf{H})}{N\lambda_N}, \qquad a = \frac{\mathrm{tr}(\mathbf{H})}{\lambda + \sum_{i>N}\lambda_i} \wedge (\kappa R\sqrt{N}) = \frac{\mathrm{tr}(\mathbf{H})}{\lambda + \sum_{i>N}\lambda_i} \wedge \frac{\mathrm{tr}(\mathbf{H})R}{\sqrt{N}\lambda_N}.$$

Next we discuss two cases:

Case I, $\widetilde{\lambda}\cdot(1\vee\kappa\log a) \geq \mathrm{tr}(\mathbf{H})$. For SGD, let us set $k_{\mathrm{sgd}} = k$ and that

$$\gamma = \frac{1}{(1+R^2)\cdot\widetilde{\lambda}\cdot(1\vee\kappa\log a)} \leq \frac{1}{\mathrm{tr}(\mathbf{H})},$$

then

$$N_{\mathrm{sgd}}\cdot\gamma = \frac{N}{\widetilde{\lambda}}.$$

Thus we obtain that

$$\begin{aligned}\mathrm{SGDRisk} &\lesssim \frac{(1-\gamma\lambda_k)^{2N_{\mathrm{sgd}}}}{(\gamma N_{\mathrm{sgd}})^2}\|\mathbf{w}^*\|^2_{\mathbf{H}^{-1}_{0:k}} + \|\mathbf{w}^*\|^2_{\mathbf{H}_{k:\infty}} + \frac{(1+R^2)\sigma^2}{N_{\mathrm{sgd}}}\left(k + (\gamma N_{\mathrm{sgd}})^2\sum_{i>k}\lambda_i^2\right)\\ &= \frac{(1-\gamma\lambda_k)^{2N_{\mathrm{sgd}}}}{\left(N/\widetilde{\lambda}\right)^2}\|\mathbf{w}^*\|^2_{\mathbf{H}^{-1}_{0:k}} + \|\mathbf{w}^*\|^2_{\mathbf{H}_{k:\infty}} + \frac{\sigma^2}{N(1\vee\kappa\log a)}\left(k + \left(N/\widetilde{\lambda}\right)^2\sum_{i>k}\lambda_i^2\right)\\ &\leq \left(\frac{\widetilde{\lambda}}{N}\right)^2\|\mathbf{w}^*\|^2_{\mathbf{H}^{-1}_{0:k}} + \|\mathbf{w}^*\|^2_{\mathbf{H}_{k:\infty}} + \frac{\sigma^2}{N}\left(k + \left(\frac{N}{\widetilde{\lambda}}\right)^2\sum_{i>k}\lambda_i^2\right)\\ &\lesssim \mathrm{RidgeRisk}.\end{aligned}$$

Case II, $\widetilde{\lambda} \cdot (1 \vee \kappa \log a) < \mathrm{tr}(\mathbf{H})$. For SGD, let us set $k_{\mathrm{sgd}} = k$ and that

$$\gamma = \frac{1}{(1+R^2) \cdot \mathrm{tr}(\mathbf{H})} \leq \frac{1}{\mathrm{tr}(\mathbf{H})},$$

then

$$N_{\mathrm{sgd}} \cdot \gamma = \frac{N \cdot (1 \vee \kappa \log a)}{\mathrm{tr}(\mathbf{H})} \leq \frac{N}{\widetilde{\lambda}}.$$

We obtain that

$$\begin{aligned}
\mathrm{SGDRisk} &\leq \mathrm{SGDBias} + \mathrm{SGDVariance} \\
&\lesssim \frac{(1-\gamma\lambda_k)^{2N_{\mathrm{sgd}}}}{(\gamma N_{\mathrm{sgd}})^2} \|\mathbf{w}^*\|^2_{\mathbf{H}^{-1}_{0:k}} + \|\mathbf{w}^*\|^2_{\mathbf{H}_{k:\infty}} + \frac{(1+R^2)\sigma^2}{N_{\mathrm{sgd}}} \left(k + (\gamma N_{\mathrm{sgd}})^2 \sum_{i>k} \lambda_i^2 \right) \\
&\leq \frac{(1-\gamma\lambda_k)^{2N_{\mathrm{sgd}}}}{(\gamma N_{\mathrm{sgd}})^2} \|\mathbf{w}^*\|^2_{\mathbf{H}^{-1}_{0:k}} + \|\mathbf{w}^*\|^2_{\mathbf{H}_{k:\infty}} + \frac{\sigma^2}{N(1 \vee \kappa \log a)} \left(k + \left(N/\widetilde{\lambda} \right)^2 \sum_{i>k} \lambda_i^2 \right) \\
&\leq \frac{(1-\gamma\lambda_k)^{2N_{\mathrm{sgd}}}}{(\gamma N_{\mathrm{sgd}})^2} \|\mathbf{w}^*\|^2_{\mathbf{H}^{-1}_{0:k}} + \|\mathbf{w}^*\|^2_{\mathbf{H}_{k:\infty}} + \frac{\sigma^2}{N} \left(k + \left(\frac{N}{\widetilde{\lambda}} \right)^2 \sum_{i>k} \lambda_i^2 \right).
\end{aligned}$$

The second and the third terms match those of ridge error. As for the first term, notice that by the choice of γ and that $\lambda_k \geq \lambda_N$, we have that

$$\begin{aligned}
\frac{(1-\gamma\lambda_k)^{N_{\mathrm{sgd}}}}{\gamma N_{\mathrm{sgd}}} &\leq \left(1 - \frac{\lambda_N}{(1+R^2) \cdot \mathrm{tr}(\mathbf{H})} \right)^{N_{\mathrm{sgd}}} \cdot \frac{1}{\gamma N_{\mathrm{sgd}}} \\
&= \left(1 - \frac{1}{(1+R^2) \cdot N \cdot \kappa} \right)^{(1+R^2) \cdot N \cdot (1 \vee \kappa \log a)} \cdot \frac{\mathrm{tr}(\mathbf{H})}{N \cdot (1 \vee \kappa \log a)} \\
&\leq \left(1 - \frac{1}{(1+R^2) \cdot N \cdot \kappa} \right)^{(1+R^2) \cdot N \cdot \kappa \log a} \cdot \frac{\mathrm{tr}(\mathbf{H})}{N} \\
&\leq \frac{1}{a} \cdot \frac{\mathrm{tr}(\mathbf{H})}{N} = \frac{(\lambda + \sum_{i>N} \lambda_i) \vee (\sqrt{N}\lambda_N / R)}{\mathrm{tr}(\mathbf{H})} \cdot \frac{\mathrm{tr}(\mathbf{H})}{N} \\
&\leq \frac{\lambda + \sum_{i>k} \lambda_i}{N} \vee \frac{\lambda_k}{R \cdot \sqrt{N}} = \frac{\widetilde{\lambda}}{N} \vee \frac{\lambda_k}{R \cdot \sqrt{N}}.
\end{aligned}$$

If $\frac{(1-\gamma\lambda_k)^{N_{\mathrm{sgd}}}}{\gamma N_{\mathrm{sgd}}} \leq \frac{\widetilde{\lambda}}{N}$, then

$$\begin{aligned}
\mathrm{SGDRisk} &\lesssim \frac{(1-\gamma\lambda_k)^{2N_{\mathrm{sgd}}}}{(\gamma N_{\mathrm{sgd}})^2} \|\mathbf{w}^*\|^2_{\mathbf{H}^{-1}_{0:k}} + \|\mathbf{w}^*\|^2_{\mathbf{H}_{k:\infty}} + \frac{\sigma^2}{N} \left(k + \left(\frac{N}{\widetilde{\lambda}} \right)^2 \sum_{i>k} \lambda_i^2 \right) \\
&\leq \left(\frac{\widetilde{\lambda}}{N} \right)^2 \|\mathbf{w}^*\|^2_{\mathbf{H}^{-1}_{0:k}} + \|\mathbf{w}^*\|^2_{\mathbf{H}_{k:\infty}} + \frac{\sigma^2}{N} \left(k + \left(\frac{N}{\widetilde{\lambda}} \right)^2 \sum_{i>k} \lambda_i^2 \right) \\
&\lesssim \mathrm{RidgeRisk}.
\end{aligned}$$

If $\frac{(1-\gamma\lambda_k)^{N_{\mathrm{sgd}}}}{\gamma N_{\mathrm{sgd}}} \leq \frac{\lambda_k}{R \cdot \sqrt{N}}$, then

$$\frac{(1-\gamma\lambda_k)^{2N_{\mathrm{sgd}}}}{(\gamma N_{\mathrm{sgd}})^2} \|\mathbf{w}^*\|^2_{\mathbf{H}^{-1}_{0:k}} \leq \frac{\lambda_k^2}{R^2 \cdot N} \|\mathbf{w}^*\|^2_{\mathbf{H}^{-1}_{0:k}} \leq \frac{\|\mathbf{w}^*\|^2_{\mathbf{H}}}{R^2 \cdot N} \leq \frac{\sigma^2}{N},$$

and

$$\begin{aligned}
\mathrm{SGDRisk} &\lesssim \frac{(1-\gamma\lambda_k)^{2N_{\mathrm{sgd}}}}{(\gamma N_{\mathrm{sgd}})^2} \|\mathbf{w}^*\|^2_{\mathbf{H}^{-1}_{0:k}} + \|\mathbf{w}^*\|^2_{\mathbf{H}_{k:\infty}} + \frac{\sigma^2}{N} \left(k + \left(\frac{N}{\widetilde{\lambda}} \right)^2 \sum_{i>k} \lambda_i^2 \right) \\
&\leq \frac{\sigma^2}{N} + \|\mathbf{w}^*\|^2_{\mathbf{H}_{k:\infty}} + \frac{\sigma^2}{N} \left(k + \left(\frac{N}{\widetilde{\lambda}} \right)^2 \sum_{i>k} \lambda_i^2 \right) \\
&\lesssim 2 \cdot \mathrm{RidgeRisk}.
\end{aligned}$$

These complete the proof. □

B.3 Proof of Corollary 5.1

Proof. By Theorem 5.1, we only need to verify that $\kappa(N_{\text{ridge}}) \lesssim \log(N_{\text{ridge}})$. Recall that $\lambda_i = 1/i^\alpha$ for $0 < \alpha \leq 1$, and $d \lesssim N_{\text{ridge}}$. For $\alpha = 1$, then

$$\operatorname{tr}(\mathbf{H}) = \sum_{i=1}^{d} i^{-\alpha} \lesssim \log d \lesssim \log(N_{\text{ridge}}),$$

thus

$$\kappa(N_{\text{ridge}}) = \frac{\operatorname{tr}(\mathbf{H})}{N_{\text{ridge}}\lambda_{\min\{d, N_{\text{ridge}}\}}} \lesssim \frac{\log(N_{\text{ridge}})}{N_{\text{ridge}} \cdot N_{\text{ridge}}^{-1}} = \log(N_{\text{ridge}}).$$

For $\alpha < 1$, then

$$\operatorname{tr}(\mathbf{H}) = \sum_{i=1}^{d} i^{-\alpha} \lesssim d^{1-\alpha} \lesssim N_{\text{ridge}}^{1-\alpha},$$

thus

$$\kappa(N_{\text{ridge}}) = \frac{\operatorname{tr}(\mathbf{H})}{N_{\text{ridge}}\lambda_{\{N_{\text{ridge}}, d\}}} \lesssim \frac{N_{\text{ridge}}^{1-\alpha}}{N_{\text{ridge}} \cdot N_{\text{ridge}}^{-\alpha}} = 1.$$

□

B.4 Proof of Corollary 5.2

Proof. Note that given random $\mathbf{w}^*$, the expected risk considered in our paper will be including the expectation over both random data $\mathbf{x}$ and random ground-truth $\mathbf{w}^*$. Since the distribution of $\mathbf{w}^*$ is rotation invariant, the expectation of $\mathbf{w}^*[i]$ will be the same for all $i \in [d]$. Therefore, let $B = \mathbb{E}[(\mathbf{w}^*[i])^2]$, the following holds according to (6.2)

$$\begin{aligned}
\text{RidgeRisk} &\gtrsim \left(\frac{\widetilde{\lambda}}{N_{\text{ridge}}}\right)^2 \cdot \mathbb{E}\big[\|\mathbf{w}^*\|^2_{\mathbf{H}^{-1}_{0:k^*}}\big] + \mathbb{E}\big[\|\mathbf{w}^*\|^2_{\mathbf{H}_{k^*:\infty}}\big] + \sigma^2 \cdot \left(\frac{k^*}{N_{\text{ridge}}} + \frac{N_{\text{ridge}}}{\widetilde{\lambda}^2}\sum_{i>k^*}\lambda_i^2\right) \\
&= B\left(\frac{\widetilde{\lambda}}{N_{\text{ridge}}}\right)^2 \cdot \sum_{i=1}^{k^*} i^\alpha + B \cdot \sum_{i=k^*+1} i^{-\alpha} + \sigma^2 \cdot \left(\frac{k^*}{N_{\text{ridge}}} + \frac{N_{\text{ridge}}}{\widetilde{\lambda}^2}\sum_{i>k^*}\lambda_i^2\right)
\end{aligned}$$

where $k^* = \min\{k : N_{\text{ridge}}\lambda_k \leq \widetilde{\lambda}\}$. Then note that $\lambda_i = i^{-\alpha}$, we have $k^* = (N_{\text{ridge}}/\widetilde{\lambda})^{1/\alpha}$, which implies that

$$\begin{aligned}
\text{RidgeRisk} &\gtrsim B\left(\frac{\widetilde{\lambda}}{N_{\text{ridge}}}\right)^2 \cdot (k^*)^{1+\alpha} + B \cdot \big[d^{1-\alpha} - (k^*)^{1-\alpha}\big] + \sigma^2 \cdot \left(\frac{k^*}{N_{\text{ridge}}} + \frac{N_{\text{ridge}}}{\widetilde{\lambda}^2}\sum_{i>k^*}\lambda_i^2\right) \\
&\gtrsim N_{\text{ridge}}^{1-\alpha} \cdot B
\end{aligned}$$

where we use the fact that $d = \Theta(N)$. Note that constant SNR $R = \Theta(1)$ implies that

$$\sigma^2 \eqsim B\sum_{i=1}^{d}\lambda_i \eqsim N_{\text{ridge}}^{1-\alpha}B.$$

Then by (6.1) and set $N_{\text{sgd}} = N_{\text{ridge}} = N$ and $k_1 = k_2 = N_{\text{ridge}}$, we have

$$\begin{aligned}
\text{SGDRisk} &\lesssim \frac{1}{\gamma^2 N_{\text{sgd}}^2} \cdot \mathbb{E}\big[\big\|\exp(-N_{\text{sgd}}\gamma\mathbf{H})\mathbf{w}^*\big\|^2_{\mathbf{H}^{-1}_{0:k_1}}\big] + \mathbb{E}\big[\|\mathbf{w}^*\|^2_{\mathbf{H}_{k_1:\infty}}\big] \\
&\quad + (1+R^2)\sigma^2 \cdot \left(\frac{k_2}{N_{\text{sgd}}} + N_{\text{sgd}}\gamma^2\sum_{i>k_2}\lambda_i^2\right) \\
&= \frac{1}{\gamma^2 N^2} \cdot \mathbb{E}\big[\big\|\mathbf{w}^*\big\|^2_{\mathbf{H}^{-1}_{0:N}}\big] + \mathbb{E}\big[\|\mathbf{w}^*\|^2_{\mathbf{H}_{N:d}}\big] + BN^{1-\alpha} \cdot \left(1 + N\gamma^2\sum_{i>N}\lambda_i^2\right).
\end{aligned}$$

Note that we have

$$\mathbb{E}\big[\|\mathbf{w}^*\|^2_{\mathbf{H}^{-1}_{0:N}}\big] = BN^{1+\alpha},\ \mathbb{E}\big[\|\mathbf{w}^*\|^2_{\mathbf{H}_{N:d}}\big] = BN^{1-\alpha}.$$

Then we can set $\gamma \eqsim 1/\operatorname{tr}(\mathbf{H}) \eqsim N^{\alpha-1}$ and get

$$\begin{aligned}\text{SGDRisk} &\lesssim \frac{B}{N^{2\alpha}} \cdot N^{1+\alpha} + BN^{1-\alpha} + BN^{1-\alpha} \cdot \left(1 + N\gamma^2 \sum_{i>N} \lambda_i^2\right) \\ &\lesssim BN^{1-\alpha} \\ &\lesssim \text{RidgeRisk}.\end{aligned}$$

This implies that SGD can be no worse than ridge regression as long as provided same or larger sample size, which completes the proof.

□

B.5 Proof of Theorem 5.2

Proof. For simplicity we fix $N := N_{\text{sgd}}$. Let us consider the following problem instance:

- The data covariance matrix $\mathbf{H}$ has the following spectrum

$$\lambda_i = \begin{cases} 1 & i = 1, \\ \frac{1}{N \log N} & 1 < i \le N^2, \\ 0 & N^2 < i \le d \end{cases}$$

  where we require the dimension $d \ge N^2$. We note that $\operatorname{tr}(\mathbf{H}) = 1 + N/\log N \eqsim N/\log N$.

- The true parameter $\mathbf{w}^*$ is given by

$$\mathbf{w}^*[i] = \begin{cases} \sigma & i = 1, \\ 0 & 1 < i \le d. \end{cases}$$

Then for SGD, we choose stepsize as $\gamma = \log(N)/(2N) \le 1/\operatorname{tr}(\mathbf{H})$. By Lemma B.1, we have the following excess risk bound for $\mathbf{w}_{\text{sgd}}(N;\gamma^*)$,

$$L\big[\mathbf{w}_{\text{sgd}}(N;\gamma)\big] - L(\mathbf{w}^*) \le \text{SGDBias} + \text{SGDVariance},$$

where

$$\begin{aligned}\text{SGDBias} &\lesssim \sigma^2 \cdot \frac{(1-\gamma)^N}{(\gamma N)^2} \lesssim \sigma^2 \cdot \log^2 N \cdot \left(1 - \frac{\log N}{2N}\right)^N \lesssim \frac{\sigma^2 \log^2 N}{N^2} \lesssim \frac{\sigma^2}{N}, \\ \text{SGDVariance} &\lesssim \frac{\sigma^2}{N} \cdot \left(1 + (N\gamma)^2 \sum_{i>1} \lambda_i^2\right) \eqsim \frac{\sigma^2}{N},\end{aligned}$$

where we use the fact that $\sum_{i>1} \lambda_i^2 = \frac{1}{\log^2 N}$. This implies that SGD with sample size N achieves at most $\mathcal{O}(\sigma^2/N)$ excess risk on this example.

Then we calculate the excess risk lower bound of ridge regression. By Lemma B.2 and let $\widetilde{\lambda} = \lambda + \sum_{i>k^*_{\text{ridge}}} \lambda_i$, we have

$$\begin{aligned}L\big[\mathbf{w}_{\text{ridge}}(N;\lambda)\big] - L(\mathbf{w}^*) &= \text{RidgeBias} + \text{RidgeVariance} \\ &\gtrsim \sigma^2 \cdot \left(\frac{\widetilde{\lambda}^2}{N^2_{\text{ridge}}} + \frac{k^*_{\text{ridge}}}{N_{\text{ridge}}} + \frac{N_{\text{ridge}} \sum_{i>k^*_{\text{ridge}}} \lambda_i^2}{\widetilde{\lambda}^2}\right).\end{aligned}$$

If $k^*_{\text{ridge}} > N$, then

$$L\big[\mathbf{w}_{\text{ridge}}(N;\lambda)\big] - L(\mathbf{w}^*) \gtrsim \frac{\sigma^2 k^*_{\text{ridge}}}{N_{\text{ridge}}} \ge \frac{\sigma^2 N}{N_{\text{ridge}}} \ge \frac{\sigma^2}{N}, \quad \text{for } N_{\text{ridge}} < \frac{N^2}{\log^2 N}.$$

If $k^*_{\text{ridge}} \le N$, then $\sum_{i>k^*_{\text{ridge}}} \lambda_i^2 \ge \sum_{N<i\le N^2} \frac{1}{N^2\log^2 N} \eqsim \frac{1}{\log^2 N}$, which implies that

$$\begin{aligned} L\big[\mathbf{w}_{\text{ridge}}(N;\lambda)\big] - L(\mathbf{w}^*) &\gtrsim \sigma^2 \cdot \left(\frac{\widetilde{\lambda}^2}{N_{\text{ridge}}^2} + \frac{N_{\text{ridge}}}{\widetilde{\lambda}^2}\cdot\frac{1}{\log^2 N}\right) \\ &\ge \frac{\sigma^2}{N_{\text{ridge}}^{1/2}\log N} \\ &\ge \frac{\sigma^2}{N}, \quad \text{for } N_{\text{ridge}} < \frac{N^2}{\log^2 N}. \end{aligned}$$

To sum up, we have show that

$$L\big[\mathbf{w}_{\text{ridge}}(N_{\text{ridge}};\lambda)\big] - L(\mathbf{w}^*) \gtrsim \frac{\sigma^2}{N} \gtrsim L\big[\mathbf{w}_{\text{sgd}}(N;\lambda)\big] - L(\mathbf{w}^*), \quad \text{for } N_{\text{ridge}} < \frac{N^2}{\log^2 N}.$$

This completes the proof. □

B.6 Proof of Theorem 5.3

Proof. The proof of Theorem 5.3 is similar to that of Theorem 5.1. In particular, we still consider two cases: (1) $\lambda \gtrsim \operatorname{tr}(\mathbf{H})$ and (2) $\lambda \lesssim \operatorname{tr}(\mathbf{H})$. For the first case, we can use the identical proof in Theorem 5.1 and get that SGD with sample size $N_{\text{sgd}} \eqsim (1+R^2)\cdot N_{\text{ridge}}$ to achieve better excess risk than ridge regression. Note that we have assumed $R^2 = \Theta(1)$, therefore, we can claim that SGD outperforms ridge regression, as long as the sample size is at least in the same order of N_{ridge}.

For the second case that $\lambda \lesssim \operatorname{tr}(\mathbf{H})$, for simplicity we denote $N := N_{\text{ridge}}$ and we can directly set $\gamma = 1/\operatorname{tr}(\mathbf{H})$ and $N_{\text{sgd}} = N$. Let $k^* = \min\big\{k : \lambda_k \le \frac{\operatorname{tr}(\mathbf{H})\log(N)}{N}\big\}$, then by the definition of k^*_{ridge} in Lemma B.2 and the assumption that ridge regression is in the generalizable regime, we have $k^* \le k^*_{\text{ridge}} \le N_{\text{ridge}}$. Therefore, applying Lemma B.1 with $k_1 = k^*$, we have the following bound on the effective bias of SGD,

$$\begin{aligned} \text{SGDBias} &\lesssim \sum_{i=1}^{k^*} \frac{(1-\gamma\lambda_i)^N (\mathbf{w}^*[i])^2}{\lambda_i \gamma^2 N^2} + \sum_{i>k^*} \lambda_i (\mathbf{w}[i])^2 \\ &\lesssim \sum_{i=1}^{k^*} \frac{\big(1-\frac{\log(N)}{N}\big)^N (\mathbf{w}^*[i])^2}{\lambda_i N^2} + \sum_{i>k^*} \lambda_i (\mathbf{w}[i])^2 \\ &\lesssim \frac{\|\mathbf{w}^*\|_{\mathbf{H}}^2}{N} + \sum_{i>k^*} \lambda_i (\mathbf{w}[i])^2. \end{aligned}$$

Then by our assumption that

$$\sum_{i=k^*}^{N_{\text{ridge}}} \lambda_i \big(\mathbf{w}[i]\big)^2 \lesssim \frac{k^* \|\mathbf{w}^*\|_{\mathbf{H}}^2}{N},$$

we further have

$$\begin{aligned} \text{SGDBias} &\lesssim \frac{\|\mathbf{w}^*\|_{\mathbf{H}}^2}{N} + \sum_{i>k^*} \lambda_i (\mathbf{w}[i])^2 \\ &\lesssim \sum_{i>k^*_{\text{ridge}}} \lambda_i (\mathbf{w}[i])^2 + \frac{(k^*_{\text{ridge}}+1)\|\mathbf{w}^*\|_{\mathbf{H}}^2}{N}, \end{aligned}$$

where in the second inequality we use the fact that $k^* \le k^*_{\text{ridge}} \le N_{\text{ridge}}$. Regarding the variance of SGD, applying Lemma B.1 with $k_2 = k^*_{\text{ridge}}$ gives

$$\begin{aligned} \text{SGDVariance} &\lesssim (\sigma^2 + \|\mathbf{w}^*\|_{\mathbf{H}}^2)\cdot\left(\frac{k^*_{\text{ridge}}}{N} + \frac{N}{(\operatorname{tr}(\mathbf{H})^2)}\cdot\sum_{i\ge k^*_{\text{ridge}}}\lambda_i^2\right) \\ &\lesssim (\sigma^2 + \|\mathbf{w}^*\|_{\mathbf{H}}^2)\cdot\left(\frac{k^*_{\text{ridge}}}{N} + \frac{N}{(\lambda + \sum_{i>k^*_{\text{ridge}}}\lambda_i)^2}\cdot\sum_{i\ge k^*_{\text{ridge}}}\lambda_i^2\right), \end{aligned}$$

where the last inequality is due to the fact that $\lambda \lesssim \operatorname{tr}(\mathbf{H})$. Combining the above upper bounds for the bias and variance of SGD, we have that the output of SGD, with sample size $N_{\rm sgd} = N$ and learning rate $\gamma = 1/\operatorname{tr}(\mathbf{H})$, satisfies

$$\begin{aligned}
\text{SGDRisk} &\lesssim \text{SGDBias} + \text{SGDVariance} \\
&\lesssim \sum_{i>k^*_{\rm ridge}} \lambda_i (\mathbf{w}[i])^2 + \frac{(k^*_{\rm ridge}+1)\|\mathbf{w}^*\|_{\mathbf{H}}^2}{N} \\
&\quad + (\sigma^2 + \|\mathbf{w}^*\|_{\mathbf{H}}^2) \cdot \left(\frac{k^*_{\rm ridge}}{N_{\rm ridge}} + \frac{N_{\rm ridge}\gamma^2}{(\lambda + \sum_{i>N_{\rm ridge}} \lambda_i)^2} \cdot \sum_{i \geq k^*_{\rm ridge}} \lambda_i^2 \right) \\
&\eqsim \sum_{i>k^*_{\rm ridge}} \lambda_i (\mathbf{w}[i])^2 + \frac{(k^*_{\rm ridge}+1)\|\mathbf{w}^*\|_{\mathbf{H}}^2}{N} \\
&\quad + \sigma^2 \cdot \left(\frac{k^*_{\rm ridge}}{N_{\rm ridge}} + \frac{N_{\rm ridge}\gamma^2}{(\lambda + \sum_{i>N_{\rm ridge}} \lambda_i)^2} \cdot \sum_{i \geq k^*_{\rm ridge}} \lambda_i^2 \right) \\
&\lesssim \text{RidgeBias} + \text{RidgeVariance}, \qquad \text{(B.1)}
\end{aligned}$$

where the last equality holds since we assume that $\|\mathbf{w}\|_{\mathbf{H}}^2/\sigma^2 = \Theta(1)$. Note that the R.H.S. of (B.1) is exactly the lower bound of the excess risk of ridge regression. Therefore, we can conclude that as long as $N_{\rm sgd} = N$, SGD with a tuned stepsize γ will be no worse than ridge regression for all λ (up to constant factors). This completes the proof.

□

C Proof of Theorem B.2

In this section we always make Assumption 5.1. The results and techniques are either explicitly or implicitly presented in [7, 26]. For self-completeness, we provide a formal proof here.

Notation. Following [26] and [7], we define the following notations:

- $\mathbf{v} := \mathbf{H}^{-\frac{1}{2}}\mathbf{x} \in \mathbb{R}^d$, then $\mathbf{v}$ is sub-Gaussian and has independent components.
- Let $\mathbf{X} := (\mathbf{x}_1, \dots, \mathbf{x}_n)^\top \in \mathbb{R}^{n\times d}$. Let $\mathbf{X} = (\mathbf{X}_{0:k}\ \mathbf{X}_{k:\infty})$
- Let $\mathbf{X} = \left(\sqrt{\lambda_1}\mathbf{z}_1, \dots, \sqrt{\lambda_d}\mathbf{z}_d\right) \in \mathbb{R}^{n\times d}$, then by Assumption 5.1, $\mathbf{z}_j$ is 1-sub-Gaussian and has independent components.
- Let $\widetilde{\mathbf{A}} := \mathbf{X}\mathbf{X}^\top = \sum_{i=1}^d \lambda_i \mathbf{z}_i \mathbf{z}_i^\top \in \mathbb{R}^{n\times n}$. Let $\mathbf{A} := \widetilde{\mathbf{A}} + \lambda_n \mathbf{I}_n = \mathbf{X}\mathbf{X}^\top + \lambda \mathbf{I}_n$.
- Let $\widetilde{\mathbf{A}}_k := \mathbf{X}_{k:\infty}\mathbf{X}_{k:\infty}^\top = \sum_{i\leq k} \lambda_i \mathbf{z}_i \mathbf{z}_i^\top \in \mathbb{R}^{n\times n}$. Let $\mathbf{A}_k := \widetilde{\mathbf{A}}_k + \lambda \mathbf{I}_n = \mathbf{X}_{k:\infty}\mathbf{X}_{k:\infty}^\top + \lambda \mathbf{I}_n$.
- Let $\widetilde{\mathbf{A}}_{-j} := \sum_{i\neq j} \lambda_i \mathbf{z}_i \mathbf{z}_i^\top \in \mathbb{R}^{n\times n}$. Let $\mathbf{A}_{-j} := \widetilde{\mathbf{A}}_{-j} + \lambda \mathbf{I}_n$.
- Let $\rho_k := \frac{\lambda + \sum_{i>k}\lambda_i}{\lambda_{k+1}}$.
- Let $\mathbf{C} := \mathbf{A}^{-1}\mathbf{X}\mathbf{H}\mathbf{X}^\top\mathbf{A}^{-1}$.
- Let $\mathbf{B} := \left(\mathbf{I}_d - \mathbf{X}^\top\mathbf{A}^{-1}\mathbf{X}\right)\mathbf{H}\left(\mathbf{I}_d - \mathbf{X}^\top\mathbf{A}^{-1}\mathbf{X}\right)$.
- We use $\mathbb{E}_{\mathbf{X}}[\cdot]$ and $\mathbb{E}_{\boldsymbol{\epsilon}}[\cdot]$ to denote the expectation with respect to the randomness of drawing $\mathbf{X}$ and the randomness of noise, respectively.

Under the above notations and from [7, 26], we have

$$\mathbb{E}_{\mathbf{X},\boldsymbol{\epsilon}}[\text{ridge error}] = \mathbb{E}_{\mathbf{X}}[\text{RidgeBias}] + \mathbb{E}_{\mathbf{X},\boldsymbol{\epsilon}}[\text{RidgeVariance}],$$

where

$$\text{RidgeBias} := (\mathbf{w}^*)^\top \mathbf{B}\mathbf{w}^*, \qquad \text{RidgeVariance} := \boldsymbol{\epsilon}^\top \mathbf{C}\boldsymbol{\epsilon}.$$

We next provide lower bounds for each terms respectively.

Lemma C.1 (Variant of Lemma 10 in [7])**.** *There are constants $b, c \geq 1$ such that for every $k \geq 0$, with probability at least 0.1,*

1. *for all* $i \geq 1$,

$$\mu_{k+1}(\mathbf{A}_{-i}) \leq \mu_{k+1}(\mathbf{A}) \leq \mu_1(\mathbf{A}_k) \leq c\left(\lambda + \sum_{j>k}\lambda_j + \lambda_{k+1}n\right),$$

2. *for all* $1 \leq i \leq k$,

$$\frac{1}{c}\left(\lambda + \sum_{j>k}\lambda_j\right) - c\lambda_{k+1}n \leq \mu_n(\mathbf{A}_k) \leq \mu_n(\mathbf{A}_{-i}) \leq \mu_n(\mathbf{A}),$$

3. *if* $\rho_k \geq bn$, *then*

$$\frac{1}{c}\lambda_{k+1}\rho_k \leq \mu_n(\mathbf{A}_k) \leq \mu_1(\mathbf{A}_k) \leq c\lambda_{k+1}\rho_k.$$

4. *if* $\rho_k \geq bn$, *then for all* $i > k$,

$$\mu_n(\mathbf{A}_{-i}) \geq \frac{1}{c}\lambda_{k+1}\rho_k$$

Proof. The first two claims are proved by noticing that $\mathbf{A} = \lambda\mathbf{I} + \widetilde{\mathbf{A}}$, $\mathbf{A}_k = \lambda\mathbf{I} + \widetilde{\mathbf{A}}_k$, $\mathbf{A}_{-i} = \lambda\mathbf{I} + \widetilde{\mathbf{A}}_{-i}$, and applying Lemma 10 in [7] to $\widetilde{\mathbf{A}}, \widetilde{\mathbf{A}}_k, \widetilde{\mathbf{A}}_{-j}$.

The third claim is proved by using the first two claims and that $\rho_k \geq bn$ to obtain that

$$\mu_1(\mathbf{A}_k) \leq c\left(\lambda + \sum_{i>k}\lambda_i + \lambda_{k+1}n\right) \leq \left(c + \frac{c}{b}\right)\cdot\left(\lambda + \sum_{i>k}\lambda_i\right),$$
$$\mu_n(\mathbf{A}_k) \geq \frac{1}{c}\left(\lambda + \sum_{i>k}\lambda_i\right) - c\lambda_{k+1}n \geq \left(\frac{1}{c} - \frac{c}{b}\right)\cdot\left(\lambda + \sum_{i>k}\lambda_i\right),$$

and by re-scaling the constants.

The fourth claim is used in Lemma 3 in [26], which can be proved under Assumption 5.1 as follows. Let $i > k$ and $\widetilde{\mathbf{A}}_{k,-i} = \sum_{j>k, j\neq i}\lambda_j \mathbf{z}_j\mathbf{z}_j^\top$. Then by Lemma 10 in [7] there is an absolute constant $c \geq 1$ such that

$$\mu_n(\widetilde{\mathbf{A}}_{-i}) \geq \mu_n(\widetilde{\mathbf{A}}_{k,-i}) \geq \frac{1}{c}\sum_{j>k, j\neq i}\lambda_j - c\lambda_{k+1}n$$

holds with probability at least $1 - 2e^{-n/c}$, which yields

$$\mu_n(\mathbf{A}_{-i}) \geq \lambda + \frac{1}{c}\sum_{j>k, j\neq i}\lambda_i - c\lambda_{k+1}n \geq \lambda + \frac{1}{2c}\sum_{j>k}\lambda_j - \left(c + \frac{1}{c}\right)\lambda_{k+1}n,$$

where the last inequality is because: (1) $\sum_{j>k, j\neq i}\lambda_j \geq \frac{1}{2}\sum_{j>k}\lambda_j$ if $i > k+1$, and (2) $\sum_{j>k, j\neq i}\lambda_j = \sum_{j>k}\lambda_j - \lambda_{k+1}$ if $i = k+1$. Finally, using the condition that $\rho_k \geq bn$ we obtain that for $i > k$,

$$\mu_n(\mathbf{A}_{-i}) \geq \lambda + \frac{1}{2c}\sum_{j>k}\lambda_j - (c + \frac{1}{c})\lambda_{k+1}n \geq \left(\frac{1}{2c} - \frac{c}{b} - \frac{1}{cb}\right)\cdot\left(\lambda + \sum_{j>k}\lambda_j\right),$$

which completes the proof by letting $b > 4c^2$ and $c \geq 1$

□

Variance Lower Bounds. According to Lemma 7 in [7], and note that $\boldsymbol{\epsilon}$ is independent of $\mathbf{X}$, has zero mean, and is σ-sub-Gaussian, we have that

$$\mathbb{E}_{\boldsymbol{\epsilon}}[\text{RidgeVariance}] = \mathbb{E}_{\boldsymbol{\epsilon}}[\boldsymbol{\epsilon}^\top \mathbf{C}\boldsymbol{\epsilon}] = \operatorname{tr}\left(\mathbf{C}\cdot\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top]\right) \geq \frac{1}{c}\sigma^2 \operatorname{tr}(\mathbf{C}) \tag{C.1}$$

for some constant $c > 1$. In the following we lower bound $\operatorname{tr}(\mathbf{C})$.

Lemma C.2 (Variant of Lemma 8 in [7])**.**

$$\operatorname{tr}(\mathbf{C}) = \sum_i \lambda_i^2 \mathbf{z}_i^\top \mathbf{A}^{-2}\mathbf{z}_i = \sum_i \frac{\lambda_i^2 \mathbf{z}_i^\top \mathbf{A}_{-i}^{-2}\mathbf{z}_i}{\left(1+\lambda_i \mathbf{z}_i^\top \mathbf{A}_{-i}^{-1}\mathbf{z}_i\right)^2}.$$

Proof. This is from the proof of Lemma 14 in [26], and can be proved in the same way as Lemma 8 in [7]. □

Lemma C.3 (Variant of Lemma 14 in [7])**.** *There is a constant c such that for any $i \geq 1$ with $\lambda_i > 0$, and any $0 \leq k \leq n/c$, with probability at least 0.1,*

$$\frac{\lambda_i^2 \mathbf{z}_i^\top \mathbf{A}_{-i}^{-2}\mathbf{z}_i}{\left(1+\lambda_i \mathbf{z}_i^\top \mathbf{A}_{-i}^{-1}\mathbf{z}_i\right)^2} \geq \frac{1}{cn}\cdot\left(1+\frac{\lambda_{k+1}}{\lambda_i}\cdot\left(1+\frac{\rho_k}{n}\right)\right)^{-2}$$

Proof. Let $\mathcal{L}_i$ be a random subspace if $\mathbb{R}^n$ of codimension k, then

$$\begin{aligned}
\mathbf{z}_i^\top \mathbf{A}_{-i}^{-1}\mathbf{z}_i &\geq \frac{1}{c_1}\cdot\frac{\|\Pi_{\mathcal{L}_i}\mathbf{z}_i\|_2^2}{\lambda+\sum_{j>k}\lambda_j+\lambda_{k+1}n} && \text{(by Lemma C.1)}\\
&\geq \frac{1}{c_2}\cdot\frac{n}{\lambda+\sum_{j>k}\lambda_j+\lambda_{k+1}n} && \text{(by Corollary 13 in [7])}\\
&= \frac{1}{c_2}\cdot\frac{n}{\lambda_{k+1}(\rho_k+n)},
\end{aligned}$$

where $c_1, c_2 > 1$ are constants. The above implies that

$$\begin{aligned}
\frac{\lambda_i^2 \mathbf{z}_i^\top \mathbf{A}_{-i}^{-2}\mathbf{z}_i}{\left(1+\lambda_i \mathbf{z}_i^\top \mathbf{A}_{-i}^{-1}\mathbf{z}_i\right)^2} &= \left(1+\left(\lambda_i \mathbf{z}_i^\top \mathbf{A}_{-i}^{-1}\mathbf{z}_i\right)^{-1}\right)^{-2}\cdot\frac{\left\|\mathbf{z}_i^\top \mathbf{A}_{-i}^{-1}\right\|_2^2}{\left(\mathbf{z}_i^\top \mathbf{A}_{-i}^{-1}\mathbf{z}_i\right)^2}\\
&\geq \left(1+\left(\lambda_i \mathbf{z}_i^\top \mathbf{A}_{-i}^{-1}\mathbf{z}_i\right)^{-1}\right)^{-2}\cdot\frac{1}{\|\mathbf{z}_i\|_2^2} \quad \text{(by Cauchy-Schwarz's inequality)}\\
&\geq \left(1+c_2\cdot\frac{\lambda_{k+1}(\rho_k+n)}{n\lambda_i}\right)^{-2}\cdot\frac{1}{\|\mathbf{z}_i\|_2^2}. \quad \text{(by the lower bound for } \mathbf{z}_i^\top \mathbf{A}_{-i}^{-1}\mathbf{z}_i)
\end{aligned}$$

According to Corollary 13 in [7], there is constant $c_3 > 1$ such that $\|\mathbf{z}_i\|_2^2 \leq \frac{1}{c_3}n$ holds with constant probability, inserting which into the above inequality and rescaling the constants complete the proof. □

Lemma C.4 (Variant of Lemma 16 in [7])**.** *There is constant c such that for any $0 \leq k \leq n/c$ and any $b > 1$ with probability at least 0.1,*

- *if $\rho_k < bn$, then $\operatorname{tr}(\mathbf{C}) \geq \frac{k+1}{cb^2 n}$;*
- *if $\rho_k \geq bn$, then $\operatorname{tr}(\mathbf{C}) \geq \frac{1}{cb^2}\min_{\ell\leq k}\left\{\frac{\ell}{n}+\frac{b^2 n\sum_{i>\ell}\lambda_i^2}{(\lambda_{k+1}\rho_k)^2}\right\}$.*

Proof. This is proved by repeating the proof of Lemma 16 in [7], where we replace Lemmas 8 and 14 in [7] with our Lemmas C.2 and C.3 respectively. □

Theorem C.5 (Restatement of Theorem B.2, variance part)**.** *There exist absolute constants $b, c, c_1 > 1$ for the following to hold: let*

$$k^* := \min\{k : \lambda + \sum_{i>k}\lambda_i \geq bn\lambda_{k+1}\},$$

then with probability at least 0.1:

- *if* $k^* \geq n/c_1$ *then*

$$\mathbb{E}_{\boldsymbol{\epsilon}}[\text{RidgeVariance}] \geq \frac{\sigma^2}{c};$$

- *if* $k^* < n/c_1$ *then*

$$\mathbb{E}_{\boldsymbol{\epsilon}}[\text{RidgeVariance}] \geq \frac{\sigma^2}{c}\left(\frac{k^*}{n} + \frac{n}{\lambda + \sum_{i>k^*}\lambda_i} \cdot \sum_{i>k^*}\lambda_i^2\right).$$

As a direct consequence, the expected ridge variance is lower bounded by

$$\mathbb{E}_{\mathbf{X},\boldsymbol{\epsilon}}[\text{RidgeVariance}] \geq \begin{cases} \frac{\sigma^2}{10c}, & k^* \geq n/c_1 \\ \frac{\sigma^2}{10c}\left(\frac{k^*}{n} + \frac{n}{\lambda+\sum_{i>k^*}\lambda_i} \cdot \sum_{i>k^*}\lambda_i^2\right), & k^* < n/c_1. \end{cases}$$

Proof. The high probability lower bound is proved by (C.1), our Lemma C.4, and Lemma 17 in [7]. The expectation lower bound follows immediately from the high probability lower bound by noticing the ridge variance error is non-negative. □

Bias Lower Bound. Recall the ridge bias error is [26]

$$\text{RidgeBias} = (\mathbf{w}^*)^\top \mathbf{B}\mathbf{w}^* = \sum_i (\mathbf{B})_{ii}(\mathbf{w}_i^*)^2 + 2\sum_{i>j}(\mathbf{B})_{ij}\mathbf{w}_i^*\mathbf{w}_j^*. \tag{C.2}$$

The following lemma shows the crossing terms are zero in expectation.

Lemma C.6. *For* $i \neq j$,

$$\mathbb{E}_{\mathbf{X}}[(\mathbf{B})_{ij}] = 0.$$

Proof. Recall that

$$\mathbf{B} := \left(\mathbf{I}_d - \mathbf{X}^\top\mathbf{A}^{-1}\mathbf{X}\right)\mathbf{H}\left(\mathbf{I}_d - \mathbf{X}^\top\mathbf{A}^{-1}\mathbf{X}\right).$$

Recall that $\mathbf{X} = \left(\sqrt{\lambda_1}\mathbf{z}_1, \ldots \sqrt{\lambda_d}\mathbf{z}_d\right)$, thus the i-th column of $\left(\mathbf{I}_d - \mathbf{X}^\top\mathbf{A}^{-1}\mathbf{X}\right)$ is

$$\left(\mathbf{I}_d - \mathbf{X}^\top\mathbf{A}^{-1}\mathbf{X}\right)_i = \mathbf{e}_i - \sqrt{\lambda_i}\mathbf{X}^\top\mathbf{A}^{-1}\mathbf{z}_i.$$

Moreover recall $\mathbf{H} = \text{diag}(\lambda_1, \ldots, \lambda_d)$, therefore

$$\begin{aligned}(\mathbf{B})_{ij} &= \mathbf{e}_i^\top\mathbf{B}\mathbf{e}_j = \left(\mathbf{e}_i - \sqrt{\lambda_i}\mathbf{X}^\top\mathbf{A}^{-1}\mathbf{z}_i\right)^\top\mathbf{H}\left(\mathbf{e}_j - \sqrt{\lambda_j}\mathbf{X}^\top\mathbf{A}^{-1}\mathbf{z}_j\right) \\ &= \mathbf{e}_i^\top\mathbf{H}\mathbf{e}_j - \sqrt{\lambda_i}\mathbf{e}_j^\top\mathbf{H}\mathbf{X}^\top\mathbf{A}^{-1}\mathbf{z}_i - \sqrt{\lambda_j}\mathbf{e}_i^\top\mathbf{H}\mathbf{X}^\top\mathbf{A}^{-1}\mathbf{z}_j + \sqrt{\lambda_i\lambda_j}\mathbf{z}_i^\top\mathbf{A}^{-1}\mathbf{X}\mathbf{H}\mathbf{X}^\top\mathbf{A}^{-1}\mathbf{z}_j \\ &= \mathbf{e}_i^\top\mathbf{H}\mathbf{e}_j - \left(\sqrt{\lambda_i\lambda_j}\lambda_j + \sqrt{\lambda_i\lambda_j}\lambda_i\right)\mathbf{z}_i^\top\mathbf{A}^{-1}\mathbf{z}_j + \sqrt{\lambda_i\lambda_j}\mathbf{z}_i^\top\mathbf{A}^{-1}\mathbf{X}\mathbf{H}\mathbf{X}^\top\mathbf{A}^{-1}\mathbf{z}_j.\end{aligned}$$

The first term is zero since $\mathbf{H}$ is diagonal and $i \neq j$. We next show the second term is zero in expectation. Indeed, let

$$F(\mathbf{z}_i) := \mathbf{z}_i^\top\mathbf{A}^{-1}\mathbf{z}_j = \mathbf{z}_i^\top\left(\mathbf{A}_{-i} + \lambda_i\mathbf{z}_i\mathbf{z}_i^\top\right)^{-1}\mathbf{z}_j,$$

where $\mathbf{A}_{-i}$ is independent of $\mathbf{z}_i$, then $F(\mathbf{z}_i) = -F(-\mathbf{z}_i)$. Also note that $\mathbf{z}_i$ follows a standard Gaussian which is symmetric, therefore $\mathbb{E}_{\mathbf{z}_i}F(\mathbf{z}_i) = 0$. In a similar manner, the third term is also zero in expectation. The proof is then completed.

□

Lemma C.7 (Part of the proof of Lemma 15 in [26])**.** *There exists absolute constant* $c > 1$*, such that with probability at least* 0.1*,*

$$(\mathbf{B})_{ii} \geq \frac{1}{c} \cdot \frac{\lambda_i}{\left(1 + \frac{\lambda_i}{\lambda_{k+1}} \cdot \frac{n}{\rho_k}\right)^2}.$$

As a direct consequence,

$$\mathbb{E}_{\mathbf{X}}[(\mathbf{B})_{ii}] \geq \frac{1}{10c} \cdot \frac{\lambda_i}{\left(1 + \frac{\lambda_i}{\lambda_{k+1}} \cdot \frac{n}{\rho_k}\right)^2}.$$

Proof. This lemma summarizes part of the proof of Lemma 15 in [26]. Recall that $\mathbf{H}$ is diagonal and $\mathbf{B} := \left(\mathbf{I}_d - \mathbf{X}^\top \mathbf{A}^{-1}\mathbf{X}\right)\mathbf{H}\left(\mathbf{I}_d - \mathbf{X}^\top \mathbf{A}^{-1}\mathbf{X}\right)$, thus

$$\begin{aligned}
(\mathbf{B})_{ii} &= \lambda_i \left\|\left(\mathbf{I}_d - \mathbf{X}^\top \mathbf{A}^{-1}\mathbf{X}\right)_i\right\|_2^2 \qquad \text{(since } \mathbf{H} \text{ is diagonal)} \\
&= \lambda_i \left\|e_i^\top - \sqrt{\lambda_i}\mathbf{z}_i^\top \mathbf{A}^{-1}\mathbf{X}\right\|_2^2 \qquad \left(\mathbf{X} = \left(\sqrt{\lambda_1}\mathbf{z}_1, \ldots, \sqrt{\lambda_j}\mathbf{z}_j, \ldots \sqrt{\lambda_d}\mathbf{z}_d\right)\right) \\
&= \lambda_i \left\|e_i^\top - \left(\sqrt{\lambda_i\lambda_1}\mathbf{z}_i^\top \mathbf{A}^{-1}\mathbf{z}_1, \ldots, \sqrt{\lambda_i\lambda_j}\mathbf{z}_i^\top \mathbf{A}^{-1}\mathbf{z}_j, \ldots \sqrt{\lambda_i\lambda_d}\mathbf{z}_i^\top \mathbf{A}^{-1}\mathbf{z}_d\right)\right\|_2^2 \\
&\geq \lambda_i \left(1 - \lambda_i \mathbf{z}_i^\top \mathbf{A}^{-1}\mathbf{z}_i\right)^2 \qquad \text{(use Pythagorean theorem)} \\
&= \frac{\lambda_i}{\left(1 + \lambda_i \mathbf{z}_i^\top \mathbf{A}_{-i}^{-1}\mathbf{z}_i\right)^2},
\end{aligned}$$

where in the last step we use $\mathbf{A} = \mathbf{A}_{-i} + \lambda_i \mathbf{z}_i \mathbf{z}_i^\top$ and that

$$\begin{aligned}
1 - \lambda_i \mathbf{z}_i^\top \mathbf{A}^{-1}\mathbf{z}_i &= 1 - \lambda_i \mathbf{z}_i^\top \left(\mathbf{A}_{-i} + \lambda_i \mathbf{z}_i \mathbf{z}_i^\top\right)^{-1}\mathbf{z}_i \\
&= 1 - \lambda_i \mathbf{z}_i^\top \left(\mathbf{A}_{-i}^{-1} - \lambda_i \mathbf{A}_{-i}^{-1}\mathbf{z}_i(1 + \mathbf{z}_i^\top \mathbf{A}_{-i}^{-1}\mathbf{z}_i)^{-1}\mathbf{z}_i^\top \mathbf{A}_{-i}^{-1}\right)\mathbf{z}_i \\
&= \frac{1}{1 + \lambda_i \mathbf{z}_i^\top \mathbf{A}_{-i}^{-1}\mathbf{z}_i}.
\end{aligned}$$

Now according to Corollary 13 in [7], there exists constant $c_1 > 1$ such that

$$\|\mathbf{z}_i\|_2^2 \leq c_1 n$$

holds with constant probability; and according to Lemma C.1, there exists constant $c_2 > 1$ such that for any $i \geq 1$,

$$\mu_n(\mathbf{A}_{-i}) \geq \frac{1}{c_2}\lambda_{k+1}\rho_k$$

holds with constant probability. These two facts imply that

$$\mathbf{z}_i^\top \mathbf{A}_{-i}^{-1}\mathbf{z}_i \leq \mu_n(\mathbf{A}_{-i})^{-1}\|\mathbf{z}_i\|_2^2 \leq c_1 c_2 \frac{n}{\lambda_{k+1}\rho_k},$$

inserting which into the bound of $(\mathbf{B})_{ii}$, we conclude that with constant probability,

$$(\mathbf{B})_{ii} \geq \frac{\lambda_i}{\left(1 + \lambda_i \mathbf{z}_i^\top \mathbf{A}_{-i}^{-1}\mathbf{z}_i\right)^2} \geq \frac{\lambda_i}{\left(1 + c_1 c_2 \cdot \frac{\lambda_i}{\lambda_{k+1}} \cdot \frac{n}{\rho_k}\right)^2}.$$

Finally a rescaling of the constants completes the proof. □

Theorem C.8 (Restatement of Theorem B.2, bias part)**.** *There exist absolute constants $b, c > 1$ for the following to hold: let*

$$k^* := \min\{k : \lambda + \sum_{i>k}\lambda_i \geq bn\lambda_{k+1}\},$$

then

$$\mathbb{E}_{\mathbf{X}}[\text{RidgeBias}] \geq \frac{1}{c}\left(\frac{\lambda + \sum_{i>k^*}\lambda_i}{n^2} \cdot \|\mathbf{w}^*\|_{\mathbf{H}_{0:k^*}^{-1}}^2 + \|\mathbf{w}^*\|_{\mathbf{H}_{k^*:\infty}}^2\right).$$

Proof. By (C.2), Lemmas C.6 and C.7, we have that,

$$\begin{aligned}
\mathbb{E}_{\mathbf{X}}[\text{RidgeBias}] &= \sum_i (\mathbf{B})_{ii}(\mathbf{w}_i^*)^2 \\
&\geq \frac{1}{c_1}\sum_i \frac{1}{\left(1 + \frac{\lambda_i}{\lambda_{k^*+1}} \cdot \frac{n}{\rho_{k^*}}\right)^2} \cdot \lambda_i (\mathbf{w}_i^*)^2 \qquad \text{(choose } k = k^*) \\
&\geq \frac{1}{c_1 b^2}\sum_i \frac{1}{\left(\frac{1}{b} + \frac{\lambda_i}{\lambda_{k^*+1}} \cdot \frac{n}{\rho_{k^*}}\right)^2} \cdot \lambda_i (\mathbf{w}_i^*)^2,
\end{aligned}$$

where $c_1, b > 1$ are all absolute constants. Note that for all $i \le k^*$, we must have $\lambda + \sum_{j>i-1} \lambda_j < bn\lambda_i$,

$$\frac{\lambda_i}{\lambda_{k^*+1}} \cdot \frac{n}{\rho_{k^*}} = \frac{\lambda_i n}{\lambda + \sum_{j>k^*} \lambda_j} \ge \frac{\lambda_i n}{\lambda + \sum_{j>i-1} \lambda_j} \ge \frac{1}{b},$$

and for all $i \ge k^* + 1$, we have

$$\frac{\lambda_i}{\lambda_{k^*+1}} \cdot \frac{n}{\rho_{k^*}} \le \frac{n}{\rho_{k^*}} \le \frac{1}{b},$$

then

$$\begin{aligned}
\mathbb{E}_{\mathbf{X}}[\mathrm{RidgeBias}] &\ge \frac{1}{c_1 b^2} \sum_i \frac{1}{\left(\frac{1}{b} + \frac{\lambda_i}{\lambda_{k^*+1}} \cdot \frac{n}{\rho_{k^*}}\right)^2} \cdot \lambda_i (\mathbf{w}_i^*)^2 \\
&\ge \frac{1}{2c_1 b^2} \cdot \left(\sum_{i \le k^*} \frac{1}{\left(\frac{\lambda_i}{\lambda_{k+1}} \cdot \frac{n}{\rho_k}\right)^2} \cdot \lambda_i (\mathbf{w}_i^*)^2 + \sum_{i>k^*} \frac{1}{(1/b)^2} \cdot \lambda_i (\mathbf{w}_i^*)^2 \right) \\
&\ge \frac{1}{c} \left(\sum_{i \le k^*} \frac{(\lambda_{k+1}\rho_k)^2}{n^2} \cdot \lambda_i^{-1} (\mathbf{w}_i^*)^2 + \sum_{i>k^*} \lambda_i (\mathbf{w}_i^*)^2 \right) \\
&= \frac{1}{c} \left(\frac{\lambda + \sum_{i>k^*} \lambda_i}{n^2} \cdot \|\mathbf{w}^*\|^2_{\mathbf{H}^{-1}_{0:k^*}} + \|\mathbf{w}^*\|^2_{\mathbf{H}_{k^*:\infty}} \right),
\end{aligned}$$

where $c > 1$ is an absolute constant. □