The Benefits of Implicit Regularization from SGD in Least Squares Problems

Difan Zou
University of California, Los Angeles
dknowzou@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

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, \( \ell_2 \)-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, \( \ell_2 \)-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:

\[ \text{How does the generalization performance of SGD compare with that of ridge regression in least }
\text{square problems?} \]

In particular, even in the arguably simplest setting of linear regression, we seek to understand
if/how SGD behaves differently from using an explicit \( \ell_2 \)-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 overpara-
meterized least squares 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 \( x \geq 0 \), 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 \( f(x) \asymp g(x) \) if \( f(x) \lesssim g(x) \leq f(x) \). For a vector \( w \in \mathbb{R}^d \) and a positive semidefinite matrix
\( H \in \mathbb{R}^{d \times d} \), we denote \( \|w\|_H := \sqrt{w^\top Hw} \).

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.
We use $x$ to denote a feature vector in a (separable) Hilbert space $H$. We use $d$ to refer to the dimensionality of $H$, where $d = \infty$ if $H$ is infinite-dimensional. We use $y \in \mathbb{R}$ to denote a response that is generated by
\[ y = \langle x, w^* \rangle + \xi, \]
where $w^* \in H$ 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 $x$, denoted by $H := \mathbb{E}[xx^T]$, is strictly positive definite and has finite trace. The noise $\xi$ is independent of $x$ and satisfies $\mathbb{E}[\xi] = 0$, and $\mathbb{E}[\xi^2] = \sigma^2$.

In order to characterize the interplay between $w^*$ and $H$ in the excess risk bound, we introduce:
\[ H_{0:k} := \sum_{i=1}^k \lambda_i v_i v_i^T, \quad \text{and} \quad H_{k:\infty} := \sum_{i>k} \lambda_i v_i v_i^T, \]
where $\{\lambda_i\}_{i=1}^\infty$ are the eigenvalues of $H$ sorted in non-increasing order and $v_i$'s are the corresponding eigenvectors. Then we define
\[ \|w\|^2_{H_{0:k}} = \sum_{i \leq k} \frac{(v_i^Tw)^2}{\lambda_i}, \quad \|w\|^2_{H_{k:\infty}} = \sum_{i> k} \lambda_i (v_i^Tw)^2. \]
The least squares problem is to estimate the true parameter $w^*$. Assumption 3.1 implies that $w^*$ is the unique solution that minimizes the population risk:
\[
L(w^*) = \min_{w \in H} L(w), \quad \text{where} \quad L(w) := \frac{1}{2} \mathbb{E}_{(x,y) \sim D} [(y - \langle w, x \rangle)^2]. \quad (3.1)
\]
Moreover we have that $L(w^*) = \sigma^2$. For an estimation $w$ found by some algorithm, e.g., SGD or ridge regression, its performance is measured by the excess risk, $L(w) - L(w^*)$.

**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 $(x_t, y_t)$ is sampled independently from the data distribution, and SGD makes the following update on the current estimator $w_{t-1} \in \mathcal{H}$,

$$w_t = w_{t-1} + \gamma \cdot (y_t - (w_{t-1}, x_t)) x_t, \quad t = 1, 2, \ldots, \quad w_0 = 0,$$

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:

$$w_{\text{sgd}}(N; \gamma) := \frac{2}{N} \sum_{t=N/2}^{N-1} w_t.$$  

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 $\{(x_i, y_i)\}_{i=1}^N$, let us denote $X := [x_1, \ldots, x_N]^T \in \mathbb{R}^{N \times d}$ and $y := [y_1, \ldots, y_N]^T \in \mathbb{R}^d$. Then ridge regression outputs the following estimator for the true parameter [25]:

$$w_{\text{ridge}}(N; \lambda) := \arg \min_{w \in \mathcal{H}} \|Xw - y\|^2 + \lambda\|w\|^2,$$  

(3.2)

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

$$w_{\text{ridge}}(N; \lambda) = (X^T X + \lambda I_d)^{-1} X^T y = X^T (XX^T + \lambda I_N)^{-1} y.$$  

(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 $X^T X$ 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 hyperparameters, can achieve excess risk that is smaller than the optimal population risk, i.e., $\sigma^2$. The formal mathematical definition is as follows.

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

$$\mathbb{E}_{\text{Alg,P}}[L(\text{Alg}(n, \theta))] - L(w^*) \leq \sigma^2,$$

where the expectation is over the randomness of $\text{Alg}$ 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 $\|w^*\|_2 = \mathcal{O}(1)$, the ordinary least squares (OLS) solution (ridge regression with $\lambda = 0$), i.e., $w_{\text{ridge}}(N; 0)$ 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 $\|w^*\|_2 = \mathcal{O}(1)$.

**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 $x$ is sampled from the set of natural basis $\{e_1, e_2, \ldots, e_d\}$ according to the data distribution given by $\mathbb{P}(x = e_i) = \lambda_i$, 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:

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

Clearly the population data covariance matrix is $\mathbf{H} = \text{diag}(\lambda_1, \ldots, \lambda_d)$. 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 $w_{\text{sgd}}(N; \gamma)$ and $w_{\text{ridge}}(N; \lambda)$ 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 $\lambda$, there exists a choice of stepsize $\gamma^*$ for SGD such that

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

provided the sample size of SGD satisfies

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

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 $\|w^*\|_H^2 = \sigma^2$, and a SGD solution with constant stepsize and sample size $N_{\text{sgd}}$, such that for any ridge regression solution with sample size

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

it holds that,

$$L[w_{\text{ridge}}(N_{\text{ridge}}; \lambda)] - L(w^*) \gtrsim L[w_{\text{sgd}}(N_{\text{sgd}}; \gamma^*)] - L(w^*).$$

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 $x$ is Gaussian\(^2\), i.e., $x \sim \mathcal{N}(0, \mathbf{H})$. 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

$$\{(w^*; \mathbf{H}) : w^* \in \mathcal{H}\}.$$

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., $\mathbf{H}^{-1/2}x$ has sub-Gaussian tail and independent components [7] and is symmetrically distributed.
Theorem 5.1 (Instance-wise comparison, Gaussian data). Let \( w_{\text{sgd}}(N; \gamma) \) and \( w_{\text{ridge}}(N; \lambda) \) 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 \( \lambda \), there exists a choice of stepsize \( \gamma^* \) for SGD such that

\[
L[w_{\text{sgd}}(N_{\text{sgd}}; \gamma^*)] - L(w^*) \leq L[w_{\text{ridge}}(N_{\text{ridge}}; \lambda)] - L(w^*),
\]

provided the sample size of SGD satisfies

\[
N_{\text{sgd}} \geq (1 + R^2) \cdot \kappa(N_{\text{ridge}}) \cdot \log(a) \cdot N_{\text{ridge}},
\]

where

\[
\kappa(n) = \frac{\text{tr}(H)}{n \lambda_{\min}\{n, d\}}, \quad R^2 = \frac{\|w^*\|_H^2}{\sigma^2}, \quad a = \kappa(N_{\text{ridge}}) R \sqrt{N}.
\]

Note that the result in Theorem 5.1 holds for arbitrary \( \lambda \). 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 \( \lambda \)). Besides, we would also like to point out that the SGD stepsize \( \gamma^* \) in Theorem 5.1 is only a function of the regularization parameter \( \lambda \) and \( \text{tr}(H) \), which can be easily estimated from training dataset without knowing the exact formula of \( H \).

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

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

It is commonly known that for this problem, ridge regression with \( \lambda = 0 \) can achieve \( O(\sigma^2/N_{\text{ridge}}) \) excess risk bound [13]. However, this problem is rather difficult for SGD since it is hard to learn the second coordinate of \( w^* \) using gradient information (the gradient in the second coordinate is quite small). In fact, in order to accurately learn \( w^*[2] \), SGD requires at least \( \Omega(N_{\text{ridge}} N_{\text{ridge}} \kappa(N_{\text{ridge}})) \) 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 \( H \) does not decay too fast so that \( \kappa(N_{\text{ridge}}) \leq \text{polylog}(N_{\text{ridge}}) \), 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 \( N_{\text{ridge}} \) be the sample size of ridge regression. Consider the problem instance that satisfies \( R^2 = \Theta(1) \), \( d = O(N_{\text{ridge}}) \), and \( \lambda = 1/i^\alpha \) for some \( \alpha \leq 1 \), then SGD, with a tuned stepsize \( \gamma^* \), provably generalizes no worse than any ridge regression solution in the generalizable regime if

\[
N_{\text{sgd}} \geq \log^2(N_{\text{ridge}}) \cdot N_{\text{ridge}}.
\]

We would like to further point out that the comparison made in Corollary 5.1 concerns the worst-case result regarding \( w^* \) (from the perspective of SGD), while SGD could perform much better if \( w^* \) has a nice structure. For example, considering the same setting in Corollary 5.1 but assuming that the ground truth \( w^* \) 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.
Theorem 5.2 (Best-case comparison, Gaussian data). There exists a Gaussian least square problem instance satisfying \( R^2 = 1 \) and \( \kappa(N_{\text{sgd}}) = \Theta(1) \), and an SGD solution with a constant stepsize and sample size \( N_{\text{sgd}} \), such that for any ridge regression solution (i.e., any \( \lambda \)) with sample size

\[
N_{\text{ridge}} \leq \frac{N_{\text{sgd}}^2 \log^2(N_{\text{sgd}})}{1},
\]

it holds that,

\[
L[w_{\text{ridge}}(N_{\text{ridge}}; \lambda)] - L(w^*) \gtrsim L[w_{\text{sgd}}(N_{\text{sgd}}; \gamma^*)] - L(w^*).
\]

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 \( w^* \) living in the tail eigenspace of \( H \).

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

\[
N_{\text{sgd}} \geq N_{\text{ridge}}.
\]

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 5.3 (SGD outperforms ridge regression, Gaussian data). Let \( N_{\text{ridge}} \) be sample size of ridge regression and \( k^* = \min \{ k : \lambda_k \leq \frac{\text{tr}(H)}{N_{\text{ridge}} \max(\lambda_{\text{ridge}})} \} \), then if \( R^2 = \Theta(1) \), and

\[
\sum_{i = k^* + 1}^{N_{\text{ridge}}} \lambda_i (w^*[i] - 1)^2 \lesssim \frac{k^* \|w^*\|_2^2}{N_{\text{ridge}}},
\]

for any ridge regression solution that is generalizable and any \( \lambda \), there exists a choice of stepsize \( \gamma^* \) for SGD such that

\[
L[w_{\text{sgd}}(N_{\text{sgd}}; \gamma^*)] - L(w^*) \lesssim L[w_{\text{ridge}}(N_{\text{ridge}}; \lambda)] - L(w^*)
\]

provided the sample size of SGD satisfies

\[
N_{\text{sgd}} \geq N_{\text{ridge}}.
\]

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 \( w^* \): \( w^*[i] = 1 \), \( w^*[i] = i^{-1} \), and \( w^*[i] = i^{-10} \). 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., \( \gamma \) and \( \lambda \)) 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., \( w^*[i] = 1 \)), 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 \( w^* \) mainly lives in the head eigenspace of \( H \) (i.e., \( w^*[i] = i^{-10} \)). 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]. In particular, we first provide the following informal lemma summarizing the aforementioned risk bounds of SGD and ridge regression.

---

3The 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.
Additionally, ridge regression has the following risk lower bound for a constant

\[ \text{RidgeRisk} \gtrsim \frac{1}{\lambda^2 N_{\text{ridge}}} \cdot \| \exp(-N_{\text{ridge}} \gamma H) w^* \|^2_{H^{-1}} + \| w^* \|^2_{H_{k^* \infty}} \]

\[ + (1 + R^2) \sigma^2 \cdot \left( \frac{k^*}{N_{\text{ridge}}} + N_{\text{ridge}} \frac{\gamma}{\lambda^2} \sum_{i > k^*} \lambda_i^2 \right). \]  

(6.1)

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

\[ \text{RidGeeRisk} \gtrsim \left( \frac{\tilde{\lambda}}{N_{\text{ridge}}} \right)^2 \left( \| w^* \|_{H^{-1}_k}^2 + \| w^* \|_{H_{k^* \infty}}^2 \right) + \sigma^2 \cdot \left( \frac{k^*}{N_{\text{ridge}}} + N_{\text{ridge}} \frac{\gamma}{\lambda^2} \sum_{i > k^*} \lambda_i^2 \right). \]  

(6.2)

We first highlight some useful observations in Lemma 6.1.

1. SGD has a condition on the stepsize: \( \gamma \leq 1 / \text{tr}(H) \), while ridge regression has no condition on the regularization parameter \( \lambda \).

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 \( H \). 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 SGD Bias Bound and SGD Variance Bound, performing the transformation \( N \to \alpha N \) and \( \gamma \to \alpha^{-1} \gamma \) will decrease SGD Variance Bound by a factor of \( \alpha \) while the SGD Bias Bound 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 \( w^* \in \mathcal{H} \). 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 $\gamma = \tilde{\lambda}^{-1}$ and $N_{\text{sgd}} = N_{\text{ridge}}$, we have

$$\text{SGDBiasBound} \leq \text{RidgeBiasBound}$$

$$\text{SGDVarianceBound} = (1 + R^2) \cdot \text{RidgeVarianceBound}.$$

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

$$\text{SGDBiasBound + SGDVarianceBound} \leq \text{RidgeBiasBound + RidgeVarianceBound}.$$ 

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

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

$$\text{RidgeRisk} \geq \text{RidgeVarianceBound} \geq \sigma^2.$$ 

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

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

we get

$$\text{SGDBiasBound} \leq \frac{(1 - N_{\text{sgd}} \gamma \lambda_{k^*}) N_{\text{sgd}}}{\gamma^2 N_{\text{sgd}}^{1/2} \lambda_{k^*}^2} \cdot \|w^*\|_{H_{0,k^*}}^2 + \|w^*\|_{H_{k^*,\infty}}^2$$

$$\leq \frac{\sigma^2}{N_{\text{ridge}}} + \|w^*\|_{H_{k^*,\infty}}^2$$

$$\leq \text{RidgeBiasBound + RidgeVarianceBound}. \quad (6.3)$$

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

$$\text{SGDVarianceBound} \leq (1 + R^2) \sigma^2 \cdot \left(\frac{k^*}{N_{\text{ridge}}} + \frac{\log(R^2 N_{\text{ridge}})}{\lambda_{k^*} \cdot \text{tr}(H)} \cdot \sum_{i > k^*} \lambda_i^2\right)$$

$$\leq (1 + R^2) \log(R^2 N_{\text{ridge}}) \cdot \text{RidgeVarianceBound},$$

where in the second inequality we use the fact that

$$\frac{N_{\text{ridge}}}{\lambda^2} \geq \frac{1}{\lambda_{k^*} \cdot \lambda} \geq \frac{1}{\lambda_{k^*} \cdot \text{tr}(H)}.$$ 

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

$$\text{SGDBiasBound + SGDVarianceBound} \leq 2 \cdot \text{RidgeBiasBound} + 2 \cdot \text{RidgeVarianceBound},$$

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 $\lambda \leq \text{tr}(H)$ by the proof sketch of Theorem 5.1.

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

9
However, this is not required for some good \( w^* \)'s that have small components in the \( k^*_{\text{sgd}} \cdot k^* \) coordinates. In particular, as assumed in the theorem, we have \( \sum_{i=k+1}^{N_{\text{ridge}}} \lambda_i (w^*[i])^2 \leq \hat{k} \|w^*\|_H^2 / N_{\text{ridge}}, \)
where \( \hat{k} := \min \{ k : \lambda_k N_{\text{sgd}} \leq \text{tr}(H) \cdot \log(N_{\text{sgd}}) \} \) satisfies \( \hat{k} \leq k^*_{\text{sgd}} \leq k^* \). Then let \( k_1 = \hat{k} \) in SGDBiasBound, we have
\[
\text{SGDBiasBound} = \frac{1}{\gamma^2 N_{\text{ridge}}^2} \cdot \| \exp(-N_{\text{ridge}} H) w^* \|_{H^{-1}_0}^2 + \| w^* \|_{H_{\infty}^0}^2 \\
\leq (1 - N_{\text{ridge}} \gamma \hat{k}) N_{\text{ridge}} \cdot \| w^* \|_{H^0_{\text{sgd}}}^2 + \| w^* \|_{H_{\infty}^0}^2 \\
\leq 2 \cdot \text{RidgeVarBound} + \text{RidgeBiasBound}.
\]

where \((i)\) is due to the condition that \( \sum_{i=k+1}^{N_{\text{ridge}}} \lambda_i (w^*[i])^2 \leq \hat{k} \|w^*\|_H^2 / N_{\text{ridge}} \). Moreover, it is easy to see that given \( N_{\text{sgd}} = N_{\text{ridge}} \) and \( \gamma = 1 / \text{tr}(H) \leq 1 / \hat{\lambda} \), we have SGDVarianceBound \( \leq 2 \cdot \text{RidgeVarBound} \). As a consequence we can get
\[
\text{SGDBiasBound} + \text{SGDVarianceBound} \leq 3 \cdot \text{RidgeBiasBound} + 3 \cdot \text{RidgeVarianceBound}.
\]

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

For ridge regression, according to Lemma 6.1 we have
\[
\text{RidgeRisk} \leq \text{RidgeBiasBound} + \text{RidgeVarianceBound} \\
= \Omega \left( \frac{\hat{\lambda}^2 N_{\text{ridge}}}{\lambda^2} + \frac{N_{\text{ridge}}}{\lambda^2} \right) \quad \text{since} \quad \sum_{i \geq \hat{k}} \lambda_i^2 = \Theta(1) \\
= \Omega \left( \frac{1}{N_{\text{ridge}}^{1/2}} \right) \quad \text{by the fact that} \quad a + b \geq \sqrt{ab}
\]
Therefore, it is evident that ridge regression is guaranteed to be worse than SGD if \( N_{\text{ridge}} \leq N_{\text{sgd}}^2 / \log^2(N_{\text{sgd}}) \). 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


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]