The Directional Bias Helps Stochastic Gradient Descent to Generalize in Kernel Regression Models

Yiling Luo, Xiaoming Huo, and Yajun Mei
School of Industrial and Systems Engineering
Georgia Institute of Technology
{yluo373, huo, ymei3}@gatech.edu

***Abstract*—We study the Stochastic Gradient Descent (SGD) algorithm in nonparametric statistics: kernel regression in particular. The directional bias property of SGD, which is known in the linear regression setting, is generalized to the kernel regression. More specifically, we prove that SGD with moderate and annealing step-size converges along the direction of the eigenvector that corresponds to the largest eigenvalue of the Gram matrix. In addition, the Gradient Descent (GD) with a moderate or small step-size converges along the direction that corresponds to the smallest eigenvalue. These facts are referred to as the directional bias properties; they may interpret how an SGD-computed estimator has a potentially smaller generalization error than a GD-computed estimator. The application of our theory is demonstrated by simulation studies and a case study that is based on the FashionMNIST dataset.**

***Index Terms*—directional bias, SGD, nonparametric regression**

I. Introduction

The Stochastic Gradient Descent (SGD) is a popular optimization algorithm that has a wide range of applications, including generalized linear model in statistics and deep Neural Network in machine learning. One main advantage of the SGD is the computational scalability due to low cost per iteration. Recent work also indicates that the SGD might also lead to outcomes that possess nice statistical properties under the linear regression framework, see [19].

In this paper, we study the statistical properties of the SGD under nonparametric regression models. We focus on the Reproducing Kernel Hilbert Space (RKHS) model, which is popular in both statistics and machine learning communities and is often simply referred to as the "kernel trick," see [2, 25]. The kernel method can be applied in various domains such as image processing [23] and text mining [12].

Our main approach is to analyze the directional bias of the SGD algorithm under the RKHS model. The directional bias might improve the efficiency of signal detection, and can explain why the outcome of SGD has good prediction performance. Directional bias, also referred to as implicit bias, means that an algorithm generates a solution path that is biased towards a certain direction, and it is also closely related to implicit regularization in deep learning [10]. Directional bias also means that algorithms prefer some directions over others even though they may have the same objective function value.

The state-of-the-art result on the directional bias of SGD can be divided into two categories, based on their underlying techniques, mostly under the linear regression model. The first category is the stochastic gradient flow method where one assumes an infinitesimal step-size in SGD and thus the parameter dynamic follows a stochastic differential equation, see [1, 4]. The second category is to analyze the discrete SGD sequence for a moderate step-size, which is also related with the convergence analysis of SGD, see [17, 11, 21, 27]. Our approach belongs to the second category. While our main technique is inspired by [27], there are a couple of significant difference in our analysis: (1) we extend the result from linear regression to non-parametric kernel model; (2) our SGD algorithm is different from that in [27].

We want to point out that there are more research to study the directional bias of the Gradient Descent (GD) than for the SGD. For instance, paper [28] analyzes the early stopped GD estimator in kernel regression; for Neural Networks in the 'lazy training' regime, paper [6] shows that GD converges in the direction of the smallest eigenvalue of the Neural Tangent Kernel.

Our contributions are two folded. First, we study the directional bias of (S)GD in a nonparametric regression model. Though the nonparametric regression is well studied in statistics, the directional bias is a relatively new concept [27]. Second, we unify the conditions to guarantee the directional bias of GD and SGD sequences. The main condition is the diagonally dominant Gram matrix, which covers a large class of kernel functions.

Our result can shed new light on deep learning [3]. By the state-of-the-art mathematical theory of Neural Networks (NN), kernel and/or nonparametric methods can approximate the functional space of neural networks, see for example the NTK theory [14], and the Radon bounded variance space description for ReLU NN [20]. These phenomena can lead to interesting future research.

Paper organization. The rest of the paper is organized as follows. In Section II, we formulate the problem, give the algorithms, and state our assumption. In Section III, we state our main theorems, including the directional bias result and

This project is partially supported by the Transdisciplinary Research Institute for Advancing Data Science (TRIAD), http://triad.gatech.edu, which is a part of the TRIPODS program at NSF and locates at Georgia Tech, enabled by the NSF grant CCF-1740776. Luo is supported in part by ARC fellowship. Huo is supported in part by NSF grant DMS-2015363. Mei is supported in part by NSF grant DMS-2015405.

its implication for generalization. In Section IV, we provide numerical experiments to support our theorems. In Section V, we discuss the finding in this paper, and propose some future research topics. Due to the page limit, we only include proof sketch and high-level description of the experiment in this paper, full details can be found in our arXiv paper [18].

II. Problem Formulation

In Section II-A we define the kernel regression; in Section II-B, we present the SGD and GD algorithms; in Section II-C, we state our assumption for later analysis. We also provide a simple example to justify our assumption.

A. Kernel Regression

Suppose that we have n data pairs $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$, where $\boldsymbol{x}_i \in \mathcal{X} \subset \mathcal{R}^p$, $y_i \in \mathcal{R}$ and y_i's are associated with $\boldsymbol{x}_i$'s through an unknown model $f(\boldsymbol{x}_i)$. The goal is to estimate the unknown model f from the data. One solution is to minimize the empirical risk function

$$\min_f \frac{1}{n}\sum_{i=1}^n \ell(y_i, f(\boldsymbol{x}_i)), \tag{1}$$

where ℓ is the loss function. A popular choice for the regression task is the squared loss $\ell(y, \boldsymbol{x}) = \frac{1}{2}(y - f(\boldsymbol{x}))^2$.

One can see that problem (1) is not well-defined, as there are infinitely many solutions to $\forall i : f(\boldsymbol{x}_i) = y_i$, and some of them do not generalize for a new test data. One way to fix it is to restrict $f \in \mathcal{H}$ and penalize $\|f\|_{\mathcal{H}}$ for smoothness, where $\mathcal{H}$ is a RKHS with reproducing kernel $K(\cdot,\cdot)$ and $\|\cdot\|_{\mathcal{H}}$ is the Hilbert norm. Adding these restrictions and applying Representer Theorem, problem (1) with the squared loss becomes

$$\min_{\boldsymbol{\alpha}\in\mathcal{R}^n} \frac{1}{2n}\sum_{i=1}^n (y_i - \boldsymbol{K}_i^T\boldsymbol{\alpha})^2 = \frac{1}{2n}\|\boldsymbol{y} - K\boldsymbol{\alpha}\|_2^2, \tag{2}$$

where $\boldsymbol{K}_i^T$ is the ith row of $K := K(X,X) = (K(\boldsymbol{x}_i, \boldsymbol{x}_j))_{i,j}$. For a parameter $\boldsymbol{\alpha}$, the corresponding estimator in $\mathcal{H}$ is $f(\cdot) = \sum_{i=1}^n \alpha_i K(\boldsymbol{x}_i, \cdot) := \boldsymbol{\alpha}^T K(\cdot, X)$.

Now when K is invertible, it is trivial that any algorithm on objective function (2) converges at the unique minimizer $\hat{\boldsymbol{\alpha}} = K(X,X)^{-1}\boldsymbol{y}$, so the RKHS functional estimator is

$$\hat{f}(\boldsymbol{x}) = K(\boldsymbol{x}, X)^T K(X,X)^{-1}\boldsymbol{y}, \tag{3}$$

where $K(\boldsymbol{x}, X)^T = (K(\boldsymbol{x}, \boldsymbol{x}_1), \ldots, K(\boldsymbol{x}, \boldsymbol{x}_n))$. Estimator (3) is the minimum norm interpolant, i.e.:

$$\arg\min_{f\in\mathcal{H}}\{\|f\|_{\mathcal{H}} : f(\boldsymbol{x}_i) = y_i, i = 1, \ldots, n\},$$

whose properties are studied in [16].

In this paper, we compare the convergence direction of SGD and GD to $\hat{\boldsymbol{\alpha}}$. Specifically, we consider a two-stage SGD with a phase transition from a larger step-size to a decreased step-size. Note that this matches the training scheme people always use in practice for SGD algorithms: decreasing the step-size after training for a few epochs. For that purpose, in the following sections, we define the one-step SGD/GD update and state our assumptions and notations for analysis.

B. One step SGD/GD update

For objective function (2), denote the parameter estimation at tth step as α_t, then SGD update $\boldsymbol{\alpha}_{t+1}$ as

$$\boldsymbol{\alpha}_{t+1} = \boldsymbol{\alpha}_t - \eta_t(\boldsymbol{K}_{i_t}^T\boldsymbol{\alpha}_t - y_{i_t})\cdot\boldsymbol{K}_{i_t}, \tag{4}$$

where i_t is uniformly random sampled from $\{1, \ldots, n\}$.

The GD update $\boldsymbol{\alpha}_{t+1}$ as

$$\boldsymbol{\alpha}_{t+1} = \boldsymbol{\alpha}_t - \frac{\eta_t}{n}K^T(K\boldsymbol{\alpha}_t - \boldsymbol{y}). \tag{5}$$

C. Assumptions and Notations

At high level, we assume the Gram matrix to be diagonal dominant. This happens when the high-dimension features are sparse, and is observed in a lot of practical problems [26], for example, linear or string kernels being applied to text data [12], domain-specific kernels being applied to image retrieval [24] and bioinformatics [22], and the Global Alignment kernel being applied to most datasets [9, 8].

We formally state our assumption as follows:

Assumption 1 (Diagonally dominant Gram matrix)**.** *Denote by $K = K(X,X)$ the Gram matrix, we assume that K is diagonally dominant. Specifically, suppose w.l.o.g. that $K_{1,1} \geq K_{2,2} \geq \ldots \geq K_{n,n} > 0$, then we have for a small value τ that*

$$|K_{i,j}| \leq \tau \ll K_{n,n}, \forall i \neq j.$$

One can justify that a Gram matrix is diagonally dominant by imposing proper assumptions on the kernel function $K(\cdot,\cdot)$ and the data distribution. In our arXiv paper [18], we show examples of diagonally dominant Gram matrix for some popular kernels. Due to page limit, we only include the bilinear kernel example in this paper as follows.

Proposition 1 (Lemma 1 in [27])**.** *Consider the bilinear kernel $K(\boldsymbol{x}, \boldsymbol{x}') := \langle \boldsymbol{x}, \boldsymbol{x}'\rangle$. Assume the data $\boldsymbol{x}_i, i = 1, \ldots, n$, are i.i.d. uniformly distributed on the unit sphere S^{d-1}, where $d \gg n$. When $d \geq 4\log(2n^2/\delta)$ for some $\delta \in (0,1)$. Then with probability at least $1-\delta$, we have*

$$|K_{i,j}| = |\langle \boldsymbol{x}_i, \boldsymbol{x}_j\rangle| < \tilde{\tau} := \tilde{\mathcal{O}}(1/\sqrt{d}) \ll K_{n,n} = 1, \forall i \neq j.$$

Though commonly exists, the diagonal dominance is undesired in classification and clustering tasks. It indicates that the data points are dissimilar to each other, which means not enough information for classification/clustering. There are some efforts for solving the issue of diagonal dominance in these cases, see for example [12, 15]. But for the regression task, the diagonal dominance, in other words, the dissimilarity of data points, may have benefits. One can find similar conditions such as Restricted Isometry Property and s-goodness that describes linearly dissimilar features in a regression literature [5, 7]. Such conditions are required for proving minimax optimality or exact recovery of a sparse signal in many settings. In our case, we adopt the dissimilarity concept and apply it to data points in high-dimensional nonlinear feature space. Later we will see that in our case, the directional bias drives SGD to select a good solution that generalizes well

among all solutions of the same level of empirical loss. In this way, our SGD estimator benefits from the diagonal dominance.

Notations. We use the following notations throughout the paper. For the Gram matrix K, let $K_{i,j}$ denote its (i,j)th element. Denote $\lambda_i = K_{i,i} = K(\boldsymbol{x}_i, \boldsymbol{x}_i)$, and assume w.l.o.g. that $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$. Denote the ith column of K as $\boldsymbol{K}_i$, let $K_{-1} = [K_2, \ldots, K_n]$. Denote P_{-1} the projection onto column space of K_{-1}, and $P_1 = I - P_{-1}$. And denote $\gamma_1 \geq \ldots \geq \gamma_n > 0$ the eigenvalues of K in non-increasing order.

III. Main result

The main results are presented in two subsections. Section III-A states the directional bias results of SGD and GD estimators, respectively. Section III-B shows that certain directional bias leads to good generalization performance, and applies this result to show that an outcome from SGD potentially generalizes better than an outcome from GD.

A. Directional bias

By our assumption, K will be full rank, (S)GD on (2) converges at $\hat{\boldsymbol{\alpha}} = K^{-1}\boldsymbol{y}$. We are interested in the direction at which $\boldsymbol{\alpha}_t$ converges to $\hat{\boldsymbol{\alpha}}$, i.e., the quantity

$$\boldsymbol{b}_t := \boldsymbol{\alpha}_t - \hat{\boldsymbol{\alpha}}.$$

With assumption 1 that the Gram matrix is diagonally dominant, we prove that a two-stage SGD has $\boldsymbol{b}_t$ converge in the direction that is aligned with the eigenvector associated with the largest eigenvalue of the Gram matrix K.

Theorem 1 (Directional bias of an SGD-based estimator)**.** *Assume Assumption 1 holds, run a two-stage SGD with a fixed step-size for each stage: stage 1 with step-size η_1 for steps $1, \ldots, k_1$, stage 2 with step-size η_2 for steps $k_1 + 1, \ldots, k_2$, such that*

$$\frac{2}{\lambda_1^2 - C_1\sqrt{n\tau}} < \eta_1 < \frac{2}{\lambda_2^2 + C_2\sqrt{n\tau}},$$
$$\eta_2 < \frac{1}{\lambda_1^2 + C_3\sqrt{n\tau}},$$

where C_1, C_2, C_3 are constants. For a small $\epsilon > 0$ such that $n\tau < poly(\epsilon)$ there exists $k_1 = \mathcal{O}(\log \frac{1}{\epsilon})$ and k_2 such that

$$(1 - 2\epsilon)\gamma_1 \leq E[\|K\boldsymbol{b}_{k_2}^{SGD}\|_2]/E[\|\boldsymbol{b}_{k_2}^{SGD}\|_2] \leq \gamma_1.$$

That is, $\boldsymbol{b}_{k_2}^{SGD}$ is close to the direction of eigenvector corresponding to the largest eigenvalue of K.

Remark 1. *One should assume τ in Assumption 1 to be small enough for ϵ to be very small if one would like the resulting estimator $\boldsymbol{b}_{k_2}^{SGD}$ to have the direction that corresponds to the largest eigenvalue of K. Later we will see that if one only wants different directional bias of SGD and GD estimators, a moderate ϵ is allowed, the assumption on τ is not that strong.*

Next, we see the different convergence direction of GD.

Theorem 2 (Directional bias of a GD-based estimator)**.** *Assume Assumption 1 holds, run GD with a fixed step-size η such that*

$$\eta < n/(\lambda_1 + n\tau)^2.$$

For a $\epsilon' > 0$, let $k = \mathcal{O}(\log \frac{1}{\epsilon'})$, we have the GD estimator after k steps satisfying:

$$\gamma_n \leq \|K\boldsymbol{b}_k^{GD}\|_2/\|\boldsymbol{b}_k^{GD}\|_2 \leq \sqrt{1+\epsilon'}\gamma_n.$$

That is, $\boldsymbol{b}_k^{GD}$ is close to the direction that corresponds to the smallest eigenvalue of K.

Remark 2. *The assumption (on τ) is mild for differentiating the directional bias of SGD and GD. Comparing Theorem 1 and 2,when $\gamma_n < (1-2\epsilon)\gamma_1$, taking k large enough we have*

$$\frac{\|K\boldsymbol{b}_k^{GD}\|_2}{\|\boldsymbol{b}_k^{GD}\|_2} < \frac{E\|K\boldsymbol{b}_{k_2}^{SGD}\|_2}{E\|\boldsymbol{b}_{k_2}^{SGD}\|_2}.$$

That is, one may expect $\boldsymbol{b}_{k_2}^{SGD}$ to be in the direction of larger eigenvalue compared with $\boldsymbol{b}_k^{GD}$. In the following subsection, we will see that the directional bias towards a larger eigenvalue of the kernel is good for generalization. That is, directional bias helps an SGD-based estimator to generalize.

Though Assumption 1 appears in Theorem 2, it is just used to bound the step-size so that GD converges; the diagonally dominant structure of K is not required. Moreover, the choice of ϵ' is independent of τ, then for an arbitrarily small $\epsilon' > 0$, run GD long enough then the theorem will apply. The estimator $\boldsymbol{b}_k^{GD}$ can be arbitrarily close to the eigenvector that correspond to the smallest eigenvalue.

B. Effect of directional bias

In this subsection, the estimator that is biased towards the largest eigenvalue of the Hessian is shown to be the best for parameter estimation, see Theorem 3. Later, we define a realizable problem setting of kernel regression where the generalization error depends on the parameter estimation error, and in this way, the directional bias helps an SGD-based estimator to generalize.

Theorem 3. *Consider minimizing the quadratic loss function*

$$L(\boldsymbol{w}) = \|A\boldsymbol{w} - \boldsymbol{y}\|_2^2.$$

Assume there is a ground truth $\boldsymbol{w}^*$ such that $\boldsymbol{y} = A\boldsymbol{w}^*$. For a fixed level of the quadratic loss, the parameter estimation error $\|\boldsymbol{w} - \boldsymbol{w}^*\|_2^2$ has a lower bound:

$$\forall \boldsymbol{w} \in \{\boldsymbol{w} : L(\boldsymbol{w}) = a\} : \|\boldsymbol{w} - \boldsymbol{w}^*\|_2^2 \geq a/\|A^TA\|_2.$$

Moreover, the equality is obtained when $\boldsymbol{w} - \boldsymbol{w}^*$ is in the direction of the eigenvector that corresponds to the largest eigenvalue of matrix A^TA.

Remark 3. *Theorem 3 implies that the directional bias towards the largest eigenvalue is good for parameter estimation. As discussed in Remark 2, the SGD-based estimator is biased towards a larger eigenvalue compared to the GD-based*

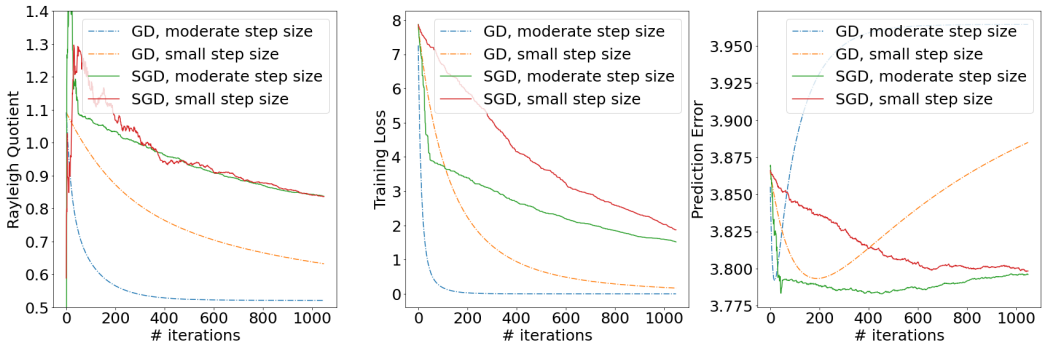


Fig. 1: Kernel regression on synthetic data. The first plot shows directional bias by Rayleigh Quotient(RQ):= $\|K\boldsymbol{b}\|_2^2/\|\boldsymbol{b}\|_2^2$. The SGD indeed converges in the direction of a larger RQ, which matches our Theorems 1 and 2. In the third plot we show the prediction error of the solution paths, and the SGD does have lower prediction error than GD, even GD has smaller training loss than SGD. This supports Theorem 4.

estimator, by Theorem 3 the SGD estimator potentialy better estimates the true parameter and thus generalizes better, which we will formalize later.

Suppose $\exists f^* \in \mathcal{H}$ such that $y = f^*(\boldsymbol{x})$. Consider the generalization error $L_D(f) := \|f - f^*\|_{\mathcal{H}}^2$. For an algorithm output f^{alg}, we decompose its generalization error as:

$$\begin{aligned} &L_D(f^{\text{alg}}) - \inf_{f\in\mathcal{H}} L_D(f) \\ &= \underbrace{L_D(f^{\text{alg}}) - \inf_{f\in\mathcal{H}_s} L_D(f)}_{:=\Delta(f^{\text{alg}}),\ \text{estimation error}} + \underbrace{\inf_{f\in\mathcal{H}_s} L_D(f) - \inf_{f\in\mathcal{H}} L_D(f)}_{\text{approximation error}}, \end{aligned}$$

where $\mathcal{H}_s$ is the hypothesis class that the output of the algorithm is restricted to. By formulation (2), we have $\mathcal{H}_s$:

$$\mathcal{H}_s = \{f \in \mathcal{H} : f = \boldsymbol{\alpha}^T K(\cdot, X), \boldsymbol{\alpha} \in \mathcal{R}^n\}.$$

We define the a-level set of training loss:

$$\nu_a = \{f \in \mathcal{H}_s : f = \boldsymbol{\alpha}^T K(\cdot, X), \frac{1}{2n}\|K\boldsymbol{\alpha} - \boldsymbol{y}\|_2^2 = a\},$$

and denote $\Delta_a^* := \inf_{f\in\nu_a} \Delta(f)$.

Note that the approximation error can not be improved unless we change the hypothesis class, which is, changing the problem formulation in our case. So we just minimize the estimation error for estimators that are in the a-level set. One can check the estimation error is given by

$$f \in \mathcal{H}_s : \Delta(f) = \boldsymbol{b}^T K \boldsymbol{b},$$

where $\boldsymbol{b} = \boldsymbol{\alpha} - \hat{\boldsymbol{\alpha}}$. Similar to Theorem 3, the estimation error is minimized when $\boldsymbol{b}$ is in the direction of the largest eigenvalue of K, so the directional bias towards a larger eigenvalue helps to generalize in kernel regression. We compare the estimation error of SGD and GD in following theorem.

Theorem 4 (Generalization performance)**.** *Follow Theorems 1 and 2, we have the following:*

- *The output of SGD has* $E[\Delta^{1/2}(f^{SGD})] \leq (1 + 4\epsilon)(\Delta_a^*)^{1/2}$*, where* a *is such that* $E[\|K\boldsymbol{\alpha}^{SGD} - y\|_2]^2 = 2na$ *and* ϵ *could be any positive small constant;*
- *The output of GD has* $\Delta(f^{GD}) \geq M\Delta_a^*$*, where* a *is the training loss of GD estimator, and* $M = \frac{\gamma_1}{\gamma_n}(1 - \epsilon') > 1$ *is a large constant.*

Remark 4. *This theorem indicates that* $E[\Delta^{1/2}(f^{SGD})] \leq \Delta^{1/2}(f^{GD})$ *when* $1+4\epsilon \leq M^{1/2}$*. Taking* $\epsilon < (\sqrt{\gamma_1/\gamma_n}-1)/4$ *in Theorem 1 and combining with Theorem 2 which states that* $\epsilon' \overset{k\to\infty}{\longrightarrow} 0$*, we will have* $1 + 4\epsilon \leq M^{1/2}$ *holds. In this way,* $\Delta(f^{SGD}) < \Delta(f^{GD})$ *with high probability. This finishes our claim that SGD generalizes better than GD.*

IV. Numeric Study

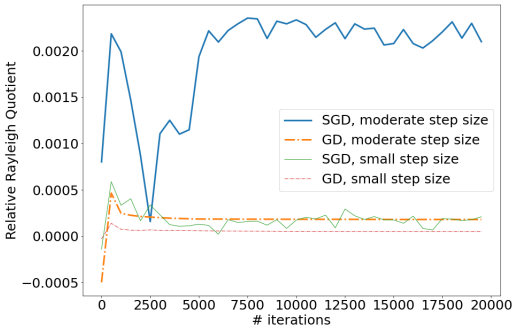
Simulation. We simulate data from a nonlinear regression model with Gaussian additive noise as $y_i = \sum_{j=1}^{100} \sin(x_{i,j}) + \epsilon_i$, where $x_{i,j} \sim N(0, 1)$ and $\epsilon_i \sim N(0, 0.01)$. We fit kernel regression using the polynomial kernel $K(\boldsymbol{x}_1, \boldsymbol{x}_2) = (\langle \boldsymbol{x}_1, \boldsymbol{x}_2\rangle + .01)^2$ on 10 training data and test the estimator on 5 testing data. We run both SGD and GD for two step-size schemes: small step-size, and moderate annealing step-size. The results are in Fig. 1.

Real data experiment. We run a 6-layer ResNet [13] on FashionMNIST. The network structure is

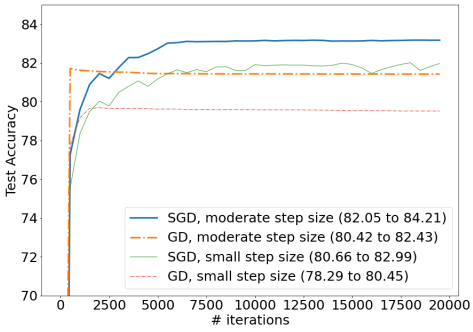
$$\begin{aligned} \text{Input} \ &\Rightarrow 7 \times 7 \text{ Conv} \ \Rightarrow \text{BatchNorm} \Rightarrow \text{ReLU} \\ &\Rightarrow 3 \times 3 \text{ MaxPool} \ \Rightarrow \text{ResBlock1} \Rightarrow \text{ResBlock2} \\ &\Rightarrow \text{Global AvePool} \Rightarrow \text{FC} \Rightarrow \text{output}. \end{aligned}$$

We run SGD and GD for two step-size schemes, similar to our simulation. There are $1,500$ training data and $10,000$ testing data in our experiment. The result is in Fig. 2.

Remark 5. *The purposes of experiment using a Neural Network (Fig. 2) are: first, the Neural Network results support our finding on kernel regression, since Neural Network is related to kernel regression through NTK theory [14]; second, our experiment indicates that our result may be empirically true for the more general deep learning framework [3].*



(a) Relative Rayleigh Quotient.



(b) Test accuracy

Fig. 2: The experiment of a small ResNet on FashionMNIST. In (a), we follow [27] to use the Relative Rayleigh Quotient(RRQ) as the measurement of the convergence direction. The SGD with moderate step-size has higher RRQ than the GD with either moderate step-size or small step-size, which supports the theory in Theorems 1 and 2. It is interesting to observe that SGD with a small step-size has a different directional bias compared with SGD with a moderate step-size, indicating that the directional bias studied in this paper does not hold for a general SGD. In (b), we plot the testing accuracy from 20 repetitions of experiments, the test accuracy (inside bracket) of SGD with moderate step-size is higher than the other cases, and we have Wilcoxon signed-rank test to confirm that the difference is significant at 0.01 level. The test accuracy validates Theorem 4.

V. Discussion and Further Work

We advance one more step towards understanding the directional bias of SGD in kernel learning. We discuss some implications of our results.

Implication to the SGD scheme: Our result shows the directional bias holds to SGD with annealing step-size. Specifically, the first stage of SGD with moderate step-size should run long enough, then in the second stage by decreasing step-size we have the directional bias towards the largest eigenvalue of the Hessian, which helps in obtaining a better generalization error bound. This explains a technique for tuning the learning rate that people adopt in practice: starting with a large step-size, running long enough until the error plateaus, then decreasing the step-size [13]. Although this technique is always used for speed convergence, we show that it also helps in predictive power, which becomes even better.

Implication to deep learning: Our assumption in the analysis implies certain structures for the deep learning models. As mentioned in section II-C, our assumption holds when the feature space is high dimensional and/or when features are possibly sparse. This matches the deep learning scenario where we have a highly overparameterized model and when the trained parameter estimator becomes sparse. In addition, considering that some deep learning tasks can be approximated by kernel learning [14], our results help to explain why the SGD-based estimator can perform better in an overparameterized deep learning setting.

Just as stated in [3], to understand deep learning one needs to understand kernel learning. This work improves our understanding in kernel learning. One may further generalize our result to neural networks through NTK theory, which can help to promote understanding for deep learning.

Appendix I: Proof Sketch for Theorem 1

We show the proof sketch for a special case, the proof for general case is similar subject to some modifications. Consider the case when $K = diag(\lambda_1, \ldots, \lambda_n)$ where $\lambda_1 > \lambda_2 \geq \ldots \geq \lambda_n$, the first stage of SGD with step-size $\eta_1 \in (\frac{2}{\lambda_1^2}, \frac{2}{\lambda_2^2})$ will have for the direction of the first data point:

$$(\boldsymbol{\alpha}_{t+1})_1 = (\boldsymbol{\alpha}_t)_1 - \eta_1\lambda_1(\lambda_1(\boldsymbol{\alpha}_t)_1 - \lambda_1(\hat{\boldsymbol{\alpha}})_1).$$

Thus

$$\begin{aligned}&(\boldsymbol{\alpha}_{t+1} - \hat{\boldsymbol{\alpha}})_1 = (1 - \eta_1\lambda_1^2)(\boldsymbol{\alpha}_t - \hat{\boldsymbol{\alpha}})_1\\ \Longrightarrow&|(\boldsymbol{\alpha}_{t+1} - \hat{\boldsymbol{\alpha}})_1| = |1 - \eta_1\lambda_1^2||(\boldsymbol{\alpha}_t - \hat{\boldsymbol{\alpha}})_1| > |(\boldsymbol{\alpha}_t - \hat{\boldsymbol{\alpha}})_1|,\end{aligned}$$

while for the other data points, we have:

$$|(\boldsymbol{\alpha}_{t+1} - \hat{\boldsymbol{\alpha}})_i| = |1 - \eta_1\lambda_i^2||(\boldsymbol{\alpha}_t - \hat{\boldsymbol{\alpha}})_i| < |(\boldsymbol{\alpha}_t - \hat{\boldsymbol{\alpha}})_i|.$$

That is, the first stage of SGD does not converge in the direction corresponding to λ_1 and converges for other directions. After run the first stage long enough, we will have all directions sufficiently fitted except the first eigenvector. In second stage, we decrease the step-size for convergence. Since the first eigenvector direction is the only direction that remains to be fitted, the estimator will converge in this direction.

Appendix II: Proof Sketch for Theorem 2

Denote the eigen decomposition of K:

$$K = G\Gamma G^T, \Gamma = diag(\gamma_1, \ldots, \gamma_n), G = [\boldsymbol{g}_1, \ldots, \boldsymbol{g}_n].$$

Denote $\boldsymbol{w}_t := G^T(\boldsymbol{\alpha}_t - \hat{\boldsymbol{\alpha}})$, we can rewrite GD update in $\boldsymbol{w}_t$:

$$\begin{aligned}&\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \frac{\eta}{n}\Gamma^2\boldsymbol{w}_t = (I - \frac{\eta}{n}\Gamma^2)\boldsymbol{w}_t\\ \Longrightarrow&(\boldsymbol{w}_t)_i = (1 - \eta\gamma_i^2/n)^t(\boldsymbol{w}_0)_i.\end{aligned}$$

So for $\eta < \frac{n}{\gamma_1^2}$, one has $|1-\eta\gamma_1^2/n| \leq \ldots \leq |1-\eta\gamma_n^2/n|$. The direction corresponding to larger eigenvalue is fitted faster, left the direction of smaller eigenvalue to be fitted later, which is the direction of convergence after several steps of GD.

References

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

[2] Peter L. Bartlett, Andrea Montanari, and Alexander Rakhlin. "Deep learning: a statistical viewpoint". In: *arXiv:2103.09177* (2021).

[3] Mikhail Belkin, Siyuan Ma, and Soumik Mandal. "To understand deep learning we need to understand kernel learning". In: *International Conference on Machine Learning*. PMLR. 2018, pp. 541–549.

[4] Guy Blanc, Neha Gupta, Gregory Valiant, and Paul Valiant. "Implicit regularization for deep neural networks driven by an Ornstein-Uhlenbeck like process". In: *Proceedings of Thirty Third Conference on Learning Theory*. 2020, pp. 483–513.

[5] Emmanuel Candes and Terence Tao. "The Dantzig selector: Statistical estimation when p is much larger than n". In: *The annals of Statistics* 35.6 (2007), pp. 2313–2351.

[6] Yuan Cao, Zhiying Fang, Yue Wu, Ding-Xuan Zhou, and Quanquan Gu. "Towards Understanding the Spectral Bias of Deep Learning". In: *arXiv:1912.01198* (2020).

[7] Shaobing Chen and David Donoho. "Basis pursuit". In: *Asilomar Conference on Signals, Systems and Computers*. Vol. 1. IEEE. 1994, pp. 41–44.

[8] Marco Cuturi. "Fast global alignment kernels". In: *International Conference on Machine Learning*. 2011, pp. 929–936.

[9] Marco Cuturi, Jean-Philippe Vert, Oystein Birkenes, and Tomoko Matsui. "A kernel for time series based on global alignments". In: *ICASSP'07*. IEEE. 2007.

[10] Michal Derezinski, Feynman T Liang, and Michael W Mahoney. "Exact expressions for double descent and implicit regularization via surrogate random design". In: *Advances in Neural Information Processing Systems*. 2020, pp. 5152–5164.

[11] Aymeric Dieuleveut and Francis Bach. "Nonparametric stochastic approximation with large step-sizes". In: *The Annals of Statistics* 44.4 (2016), pp. 1363–1399.

[12] Derek Greene and Pádraig Cunningham. "Practical solutions to the problem of diagonal dominance in kernel document clustering". In: *International Conference on Machine Learning*. 2006, pp. 377–384.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep Residual Learning for Image Recognition". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.

[14] Arthur Jacot, Franck Gabriel, and Clement Hongler. "Neural Tangent Kernel: Convergence and Generalization in Neural Networks". In: 2018.

[15] Jaz Kandola, Thore Graepel, and John Shawe-Taylor. "Reducing kernel matrix diagonal dominance using semi-definite programming". In: *Learning Theory and Kernel Machines*. Springer, 2003, pp. 288–302.

[16] Tengyuan Liang and Alexander Rakhlin. "Just Interpolate: Kernel 'Ridgeless' Regression Can Generalize". In: *The Annals of Statistics* 48.3 (2020), pp. 1329–1347.

[17] Junhong Lin and Lorenzo Rosasco. "Optimal rates for multi-pass stochastic gradient methods". In: *The Journal of Machine Learning Research* 18.1 (2017), pp. 3375–3421.

[18] Yiling Luo, Xiaoming Huo, and Yajun Mei. *The Directional Bias Helps Stochastic Gradient Descent to Generalize in Kernel Regression Models*. 2022. DOI: 10.48550/ARXIV.2205.00061. URL: https://arxiv.org/abs/2205.00061.

[19] Siyuan Ma, Raef Bassily, and Mikhail Belkin. "The Power of Interpolation: Understanding the Effectiveness of SGD in Modern Over-parametrized Learning". In: *arXiv:1712.06559* (2018).

[20] Rahul Parhi and Robert D. Nowak. "What Kinds of Functions do Deep Neural Networks Learn? Insights from Variational Spline Theory". In: *arXiv:2105.03361* (2021).

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

[22] Hiroto Saigo, Jean-Philippe Vert, Nobuhisa Ueda, and Tatsuya Akutsu. "Protein homology detection using string alignment kernels". In: *Bioinformatics* 20.11 (2004), pp. 1682–1689.

[23] Hiroyuki Takeda, Sina Farsiu, and Peyman Milanfar. "Kernel regression for image processing and reconstruction". In: *IEEE Transactions on image processing* 16.2 (2007), pp. 349–366.

[24] Qingping Tao, Stephen Scott, NV Vinodchandran, Thomas Takeo Osugi, and Brandon Mueller. "An extended kernel for generalized multiple-instance learning". In: *16th IEEE International Conference on Tools with Artificial Intelligence*. IEEE. 2004, pp. 272–277.

[25] Grace Wahba. *Spline models for observational data*. SIAM, 1990.

[26] Jason Weston, Bernhard Schölkopf, Eleazar Eskin, Christina Leslie, and William Stafford Noble. "Dealing with large diagonals in kernel matrices". In: *Annals of the institute of statistical mathematics* 55.2 (2003), pp. 391–408.

[27] Jingfeng Wu, Difan Zou, Vladimir Braverman, and Quanquan Gu. "Direction Matters: On the Implicit Bias of Stochastic Gradient Descent with Moderate Learning Rate". In: *International Conference on Learning Representations*. 2021.

[28] Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. "On early stopping in gradient descent learning". In: *Constructive Approximation* 26.2 (2007), pp. 289–315.