# DIRECTION MATTERS: ON THE IMPLICIT REGULARIZATION EFFECT OF STOCHASTIC GRADIENT DESCENT WITH MODERATE LEARNING RATE

**Anonymous authors**
Paper under double-blind review

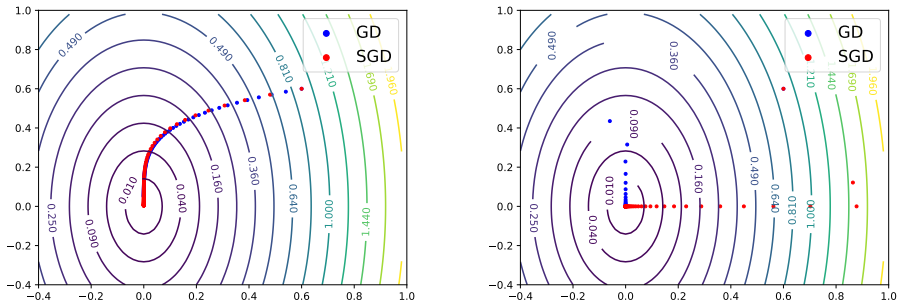
## ABSTRACT

Understanding the algorithmic regularization effect of *stochastic gradient descent* (SGD) is one of the key challenges in modern machine learning and deep learning theory. Most of the existing works, however, focus on *very small or even infinitesimal* learning rate regime, and fail to cover practical scenarios where the learning rate is *moderate and annealing*. In this paper, we make an initial attempt to characterize the particular regularization effect of SGD in the moderate learning rate regime by studying its behavior for optimizing an overparameterized linear regression problem. In this case, SGD and GD are known to converge to the unique minimum-norm solution; however, with the moderate and annealing learning rate, we show that they exhibit different *directional bias*: SGD converges along the large eigenvalue directions of the data matrix, while GD goes after the small eigenvalue directions. Furthermore, we show that such directional bias does matter when early stopping is adopted, where the SGD output is nearly optimal but the GD output is suboptimal. Finally, our theory explains several folk arts in practice used for SGD hyperparameter tuning, such as (1) linearly scaling the initial learning rate with batch size; and (2) overrunning SGD with high learning rate even when the loss stops decreasing.

## 1 INTRODUCTION

*Stochastic gradient descent* (SGD) and its variants play a key role in training deep learning models. From the optimization perspective, SGD is favorable in many aspects, e.g., scalability for large-scale models (He et al., 2016), parallelizability with big training data (Goyal et al., 2017), and rich theory for its convergence (Ghadimi & Lan, 2013; Gower et al., 2019). From the learning perspective, more surprisingly, overparameterized deep nets trained by SGD usually generalize well, even in the absence of explicit regularizers (Zhang et al., 2016; Keskar et al., 2016; Wu et al., 2020). This suggests that SGD favors certain "good" solutions among the numerous global optima of the overparameterized model. Such phenomenon is attributed to the *implicit regularization effect* of SGD. It remains one of the key theoretical challenges to characterize the algorithmic bias of SGD, especially with moderate and annealing learning rate as typically used in practice (He et al., 2016; Keskar et al., 2016).

In the *small learning rate regime*, the regularization effect of SGD is relatively well understood, thanks to the recent advances on the implicit bias of *gradient descent* (GD) (Gunasekar et al., 2017; 2018a;b; Soudry et al., 2018; Ma et al., 2018; Li et al., 2018; Ji & Telgarsky, 2019b;a; Ji et al., 2020; Nacson et al., 2019a; Ali et al., 2019; Arora et al., 2019; Moroshko et al., 2020; Chizat & Bach, 2020). According to classical stochastic approximation theory (Kushner & Yin, 2003), with a sufficiently small learning rate, the randomness in SGD is negligible (which scales with learning rate), and as a consequence SGD will behave highly similar to its deterministic counterpart, i.e., GD. Based on this fact, the regularization effect of SGD with small learning rate can be understood through that of GD. Take linear models for example, GD has been shown to be biased towards max-margin/minimum-norm solutions depending on the problem setups (Soudry et al., 2018; Gunasekar et al., 2018a; Ali et al., 2019); correspondingly, follow-ups show that SGD with small learning rate has the same bias (up to certain small uncertainty governed by the learning rate) (Nacson et al.,




(a) Small learning rate regime

(b) Moderate learning rate regime

Figure 1: Illustration for the 2-D example studied in Section 3. Here $\kappa = 4$ and $w_0 = (0.6, 0.6)$. **(a)**: Small learning rate regime. The small learning rate is $0.1/\kappa$. In this regime SGD and GD behave similarly and they both converge along $e_2$. **(b)**: Moderate learning rate regime. The initial moderate learning rate is $\eta = 1.1/\kappa$ and the decayed learning rate is $\eta' = 0.1/\kappa$. In this regime GD converges along $e_2$ but SGD converges along $e_1$, the larger eigenvalue direction of the data matrix. Please refer to Section 3 for further discussions.

2019b; Gunasekar et al., 2018a; Ali et al., 2020). The analogy between SGD and GD in the small learning rate regime is also demonstrated in Figure 1(a) and Figure 2.

However, the regularization theory for SGD with small learning rate cannot explain the benefits of SGD in the *moderate learning rate regime*, where the initial learning rate is moderate and followed by annealing (Li et al., 2019; Nakkiran, 2020; Leclerc & Madry, 2020; Jastrzebski et al., 2019). In particular, empirical studies show that, in the moderate learning rate regime, (small batch) SGD generalizes much better than GD/large batch SGD (Keskar et al., 2016; Jastrzębski et al., 2017; Zhu et al., 2019; Wu et al., 2020) (see Figure 2(b)). This observation implies that, instead of imitating the bias of GD as in the small learning rate regime, SGD in the moderate learning rate regime admits superior bias than GD — it requires a dedicated characterization for the implicit regularization effect of SGD with moderate learning rate.

In this paper, we reveal a particular regularization effect of SGD with moderate learning rate that involves *convergence direction*. In specific, we consider an overparameterized linear regression model learned by SGD/GD. In this setting, SGD and GD are known to converge to the unique minimum-norm solution (Gunasekar et al., 2018a) (see also Section 2.1). However, with a moderate and annealing learning rate, we show that SGD and GD favor different convergence directions: *SGD converges along the large eigenvalue directions of the data matrix*; in contrast, GD goes after the small eigenvalue directions. The phenomenon is illustrated in Figure 1(b). To sum up, we make the following contributions in this work:

1. For an overparameterized linear regression model, we show that SGD with moderate learning rate converges along the large eigenvalue directions of the data matrix, while GD goes after the small eigenvalue directions. To our knowledge, this result initiates the regularization theory for SGD in the moderate learning rate regime, and complements existing results for the small learning rate.
2. Furthermore, we show the particular directional bias of SGD with moderate learning rate benefits generalization when early stopping is used. This is because converging along the large eigenvalue directions (SGD) leads to nearly optimal solutions, while converging along the small eigenvalue directions (GD) can only give suboptimal solutions.
3. Finally, our results explain several folk arts for tuning SGD hyperparameters, such as (1) linearly scaling the initial learning rate with batch size (Goyal et al., 2017); and (2) overrunning SGD with high learning rate even when the loss stops decreasing (He et al., 2016).

## 2 Preliminary

Let $(x, y) \in \mathbb{R}^d \times \mathbb{R}$ be a pair of $d$-dimensional feature vector and 1-dimensional label. We consider a linear regression problem with square loss defined as $\ell(x, y; w) := (w^\top x - y)^2$, where $w \in \mathbb{R}^d$ is the model parameter. Let $\mathcal{D}$ be the population distribution over $(x, y)$, then the test loss is $L_{\mathcal{D}}(w) := \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\ell(x, y; w)\right]$. Let $\mathcal{S} := \{(x_i, y_i)\}_{i=1}^n$ be a training set of $n$ data points drawn i.i.d. from the

population distribution $\mathcal{D}$. Then the training/empirical loss is defined as the average of the individual loss over all training data points,

$$L_{\mathcal{S}}(w) := \frac{1}{n}\sum_{i=1}^{n} \ell_i(w), \quad \text{where } \ell_i(w) := \ell(x_i, y_i; w) = (w^\top x_i - y_i)^2. \tag{1}$$

We use $\{\eta_k\}$ to denote a *learning rate scheme* (LR). Then *gradient descent* (GD) iteratively performs the following update:

$$w_{k+1} = w_k - \eta_k \nabla L_{\mathcal{S}}(w_k) = w_k - \frac{2\eta_k}{n}\sum_{i=1}^{n} x_i(x_i^\top w_k - y_i). \tag{GD}$$

Next we introduce *mini-batch stochastic gradient descent* (SGD).[1] Let $b$ be the batch size. For simplicity suppose $n = mb$ for an integer $m$ (number of mini-batches). Then at each epoch, SGD first randomly partitions the training set into $m$ disjoint mini-batches with size $b$, and then sequentially performs $m$ updates using the stochastic gradients calculated over the $m$ mini-batches. Specifically, at the $k$-th epoch, let the mini-batch index sets be $\mathcal{B}_1^k, \mathcal{B}_2^k, \ldots, \mathcal{B}_m^k$, where $|\mathcal{B}_j^k| = b$ and $\bigcup_{j=1}^m \mathcal{B}_j^k = \{1, 2, \ldots, n\}$, then SGD takes $m$ updates as follows

$$w_{k,j+1} = w_{k,j} - \frac{\eta_k}{b}\sum_{i\in\mathcal{B}_j^k} \nabla \ell_i(w_{k,j}) = w_{k,j} - \frac{2\eta_k}{b}\sum_{i\in\mathcal{B}_j^k} x_i(x_i^\top w_{k,j} - y_i), \quad j = 1, \ldots, m. \tag{SGD}$$

We also write $w_{k+1} = w_{k,m+1}$ and $w_k = w_{k,1}$ to be consistent with notations in (GD).

### 2.1 The minimum-norm bias

Before presenting our results on the directional bias, let us first recap the well-known minimum-norm bias for SGD/GD optimizing linear regression problem (Gunasekar et al., 2018a; Belkin et al., 2019; Bartlett et al., 2020). We rewrite the training loss as $L_{\mathcal{S}}(w) = \frac{1}{n}\left\|X^\top w - Y\right\|_2^2$, where $X = (x_1, \ldots, x_n) \in \mathbb{R}^{d\times n}$ and $Y = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$. Then its global minima are given by $\mathcal{W}_* := \left\{w \in \mathbb{R}^d : Pw = PY\right\}$, where $P$ is the projection operator onto the data manifold, i.e., the column space of $X$. We focus on overparameterized cases where $\mathcal{W}_*$ contains multiple elements.

Notice that every gradient $\nabla \ell_i(w) = 2x_i(x_i^\top w - y_i)$ is spanned in the data manifold, thus (GD) and (SGD) can never move along the direction that is orthogonal to the data manifold. In other words, (GD) and (SGD) implicitly admit the following *hypothesis class*:

$$\mathcal{H}_{\mathcal{S}} = \left\{w \in \mathbb{R}^d : P_\perp w = P_\perp w_0\right\}, \tag{2}$$

where $w_0$ is the initialization and $P_\perp = I - P$ is the projection operator onto the orthogonal complement to the column space of $X$.

Putting things together, for any global optimum $w \in \mathcal{W}_*$ (hence $Pw = PY$), we have

$$\|w - w_0\|_2^2 = \|Pw - Pw_0\|_2^2 + \|P_\perp w - P_\perp w_0\|_2^2 = \|PY - Pw_0\|_2^2 + \|P_\perp w - P_\perp w_0\|_2^2,$$

where the right hand side is minimized when $P_\perp w = P_\perp w_0$, i.e., $w \in \mathcal{H}_{\mathcal{S}}$, thus $w$ is the solution found by SGD/GD in the non-degenerated cases (when the learning rate is set properly so that the algorithms can find a global optimum). In sum, SGD/GD is biased to find the global optimum that is *closest to the initialization*, which is referred as the "minimum-norm" bias in literature since the initialization is usually set to be zero.

## 3 Warming up: A 2-Dimensional Case Study

In this section we conduct a 2-dimensional case study to motivate our understanding on the directional bias of SGD in the moderate learning rate regime. Let us consider a training set consisting of two orthogonal points, $\mathcal{S} = \{(x_1,\ y_1 = 0),\ (x_2,\ y_2 = 0)\}$ where

$$x_1 = \sqrt{\kappa}\cdot e_1 = (\sqrt{\kappa},\ 0)^\top, \quad x_2 = e_2 = (0,\ 1)^\top, \quad \kappa > 2.$$

[1] In this work we focus on SGD without replacement. Nonetheless our results can easily carry over to SGD with replacement using a high probability argument.

Clearly $w_* = 0$ is the unique minimum of $L_{\mathcal{S}}(w)$. The Hessian of the empirical loss is $\nabla^2 L_{\mathcal{S}}(w) = x_1 x_1^\top + x_2 x_2^\top = \operatorname{diag}(\kappa, 1)$, which has two eigenvalues: the smaller one 1 is contributed by data $x_2$, and the larger one $\kappa$ contributed by data $x_1$. Hence $L_{\mathcal{S}}(w)$ is $\kappa$-smooth. Similarly the Hessian of the individual losses are $\nabla^2 \ell_1(w) = 2x_1 x_1^\top = \operatorname{diag}(2\kappa, 0)$ and $\nabla^2 \ell_2(w) = 2x_2 x_2^\top = \operatorname{diag}(0, 2)$. Thus $\ell_2(w)$ is 2-smooth, but $\ell_1(w)$, the individual loss for data $x_1$, is only $2\kappa$-smooth, which is more ill-conditioned compared to $L_{\mathcal{S}}(w)$ and $\ell_2(w)$.

Next we consider a moderate initial learning rate $\eta \in \left(\frac{1}{\kappa}, \frac{2}{1+\kappa}\right)$. According to convex optimization theory (Boyd et al., 2004), gradient step with such learning rate is convergent for $L_{\mathcal{S}}(w)$ and $\ell_2(w)$, but oscillating for $\ell_1(w)$. In other words, (GD) is convergent; and (SGD) is convergent along direction $x_2$ (or $e_2$), but oscillating along direction $x_1$ (or $e_1$). We also see this by analytically solving (GD) and (SGD) for this example:

$$w_k^{\mathrm{gd}} = \begin{pmatrix} (1-\eta\kappa)^k & \\ & (1-\eta)^k \end{pmatrix} w_0, \quad w_k^{\mathrm{sgd}} = \begin{pmatrix} (1-2\eta\kappa)^k & \\ & (1-2\eta)^k \end{pmatrix} w_0, \tag{3}$$

where $|1-\eta\kappa| < |1-\eta| < 1$ and $|1-2\eta| < 1 < |1-2\eta\kappa|$.

By Eq. (3), with moderate learning rate GD is convergent for both directions $e_1$ and $e_2$. Moreover, GD fits $e_1$ faster since the contraction parameter is smaller, i.e., $|1-\eta\kappa| < |1-\eta| < 1$. Thus observing the entire optimization path, GD approaches the minimum $w_* = 0$ along $e_2$, which corresponds to the smaller eigenvalue direction of $\nabla^2 L_{\mathcal{S}}(w)$. This is verified by the blue dots in Figure 1(b). We note this directional bias for GD also holds in the small learning rate regime, as shown in Figure 1(a).

As for SGD in the initial phase where the learning rate is moderate, Eq. (3) shows it converges along $e_2$ but oscillates along $e_1$ since $|1-2\eta| < 1 < |1-2\eta\kappa|$. In other words, SGD cannot fit $e_1$ before the learning rate decays; however when this happens, $e_2$ is already well fitted. Overall, SGD fits $e_2$ first then fits $e_1$, i.e., SGD converges to the minimum $w_* = 0$ along $e_1$, which corresponds to the larger eigenvalue direction of $\nabla^2 L_{\mathcal{S}}(w)$. This is verified by the red dots in Figure 1(b). We note this particular directional bias for SGD is dedicated to the moderate learning rate regime; in the small learning rate regime, as discussed before, SGD behaves similar to GD thus goes after the smaller eigenvalue direction, which is illustrated in Figure 1(a).

The above idea can be carried over to more general cases: the training loss usually has relatively smooth curvature because of the empirical averaging; yet some individual losses can possess bad smoothness condition, corresponding to the data points that contribute to the large eigenvalues of the Hessian/data matrix. Then with a moderate learning rate, while GD is convergent, SGD is convergent for the smooth individual losses but oscillating for the ill-conditioned individual losses. Thus SGD can only fit the latter losses after the learning rate anneals. Therefore, in the moderate learning rate regime, SGD tends to converge along the large eigenvalue directions while GD tends to go after the small eigenvalue directions. We will rigorously justify the above intuitions in the following section.

## 4 Main Theory

In this section we present our main theoretical results. The proofs are deferred to Appendix B.

We specify the population distribution of $(x, y) \in \mathbb{R}^d \times \mathbb{R}$ in the following manner: (1) we consider the feature vector given by $x = \zeta \cdot \xi$, where $\zeta \in \mathbb{R}^1$ is a magnitude random variable that takes value in $(0, 1]$, and $\xi \in \mathbb{R}^d$ is an angle random variable that follows a sphere uniform distribution, i.e., $\xi \sim \mathcal{U}(S^{d-1})$; (2) we consider a realizable setting where the label is given by $y = w_*^\top x$, i.e., there exists a true parameter $w_* \in \mathbb{R}^d$ that generates the label from the feature vector. Then the test loss is $L_{\mathcal{D}}(w) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[(w-w_*)^\top x x^\top (w-w_*)\right] = \mu \left\|w - w_*\right\|_2^2$, where $\mu = \mathbb{E}[\zeta]/d$. For an i.i.d. generated training set $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^n$, the training loss and the individual losses are

$$L_{\mathcal{S}}(w) = \frac{1}{n}(w-w_*)^\top X X^\top (w-w_*), \quad \ell_i(w) = (w-w_*)^\top x_i x_i^\top (w-w_*), \quad i = 1, \ldots, n,$$

where $X = (x_1, \ldots, x_n)$. We denote by $P$ the projection operator onto the column space of $X$ (the data manifold). For $i \in [n]$, we denote $\lambda_i := \|x_i\|_2^2$. By the previous data generalization process, we have $\lambda_i \in (0, 1]$. Without loss of generality, we assume $\{\lambda_i\}_{i\in[n]}$ are sorted in a descending order, i.e., $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. With the these preparations, we are ready to state our main theorems.

### 4.1 THE DIRECTIONAL BIAS OF SGD

We first present Theorem 1 and 2 that characterize the different directional biases of SGD and GD in the moderate learning rate regime.

**Theorem 1** (The directional bias of SGD with moderate LR, informal)**.** *Suppose $d \geq \operatorname{poly}(n)$, $\lambda_1 > \lambda_2 + \mathrm{o}(1)$ and $\lambda_n > \mathrm{o}(1)$*[2]*, where $\mathrm{o}(1)$ is a small positive constant depending on $n/d$, and the initialization is away from $w_*$. Consider (SGD) with the following moderate learning rate scheme*

$$\eta_k = \begin{cases} \eta \in \left(\frac{b}{\lambda_1 - \mathrm{o}(1)}, \frac{b}{\lambda_2 + \mathrm{o}(1)}\right), & k = 1, \ldots, k_1; \\ \eta' \in \left(0, \frac{b}{2\lambda_1}\right), & k = k_1 + 1, \ldots, k_2. \end{cases} \tag{4}$$

*Then for $\epsilon = \mathrm{o}(1)$, there exist $k_1 > \mathcal{O}\left(\log \frac{1}{\epsilon}\right)$ and $k_2$, such that with high probability the output of SGD $w^{\mathrm{sgd}}$ satisfies*

$$(1-\epsilon) \cdot \gamma_1 \leq \frac{\left(P(w^{\mathrm{sgd}} - w_*)\right)^\top \cdot XX^\top \cdot P\left(w^{\mathrm{sgd}} - w_*\right)}{\left\|P\left(w^{\mathrm{sgd}} - w_*\right)\right\|_2^2} \leq \gamma_1, \tag{5}$$

*where $\gamma_1$ is the largest eigenvalue of the data matrix $XX^\top$.*

**Theorem 2** (The directional bias of GD with moderate or small LR, informal)**.** *Suppose $d \geq \operatorname{poly}(n)$, $\lambda_{n-1} > \lambda_n + \mathrm{o}(1)$ and $\lambda_n > \mathrm{o}(1)$, where $\mathrm{o}(1)$ is a small positive constant depending on $n/d$, and the initialization is away from $w_*$. Consider (GD) with the following moderate or small learning rate scheme*

$$\eta_k \in \left(0, \frac{n}{2\lambda_1 + \mathrm{o}(1)}\right), \quad k = 1, \ldots, k_2. \tag{6}$$

*Then for any $\epsilon > 0$, if $k_2 > \mathcal{O}\left(\log \frac{1}{\epsilon}\right)$, then with high probability the output of GD $w^{\mathrm{gd}}$ satisfies*

$$\gamma_n \leq \frac{\left(P(w^{\mathrm{gd}} - w_*)\right)^\top \cdot XX^\top \cdot P\left(w^{\mathrm{gd}} - w_*\right)}{\left\|P\left(w^{\mathrm{gd}} - w_*\right)\right\|_2^2} \leq (1+\epsilon) \cdot \gamma_n, \tag{7}$$

*where $\gamma_n$ is the smallest eigenvalue of the data matrix $XX^\top$ restricted in the column space of $X$.*

A few remarks on Theorem 1 and 2 are in order.

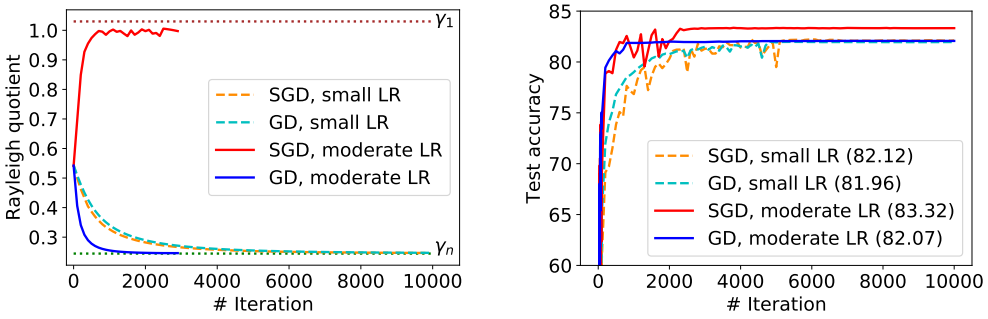
*Remark* 1. As the Rayleigh quotient (5) (resp. (7)) converges to its maximum (resp. minimum), the vector gets closer to the eigenvector of the largest (resp. smallest) eigenvalue (Trefethen & Bau III, 1997). Thus Theorem 1 and 2 suggest that, when projected onto the data manifold, SGD and GD converge to the optimum along the largest and smallest eigenvalue direction respectively. Here we are only interested in the projection onto the data manifold, since SGD/GD cannot move along the direction that is orthogonal to the data manifold as discussed in Section 2.1.

*Remark* 2. For the ease of presentation, we make strong gap assumptions ($\lambda_1 > \lambda_2 + \mathrm{o}(1)$ and $\lambda_{n-1} > \lambda_n + \mathrm{o}(1)$), and show SGD/GD converges along the largest/smallest eigenvalue direction. Nonetheless, one can weaken these requirements and extend our analysis to show a refined result where SGD/GD converges along the subspace spanned by the eigenvectors of the top/bottom eigenvalues.

*Remark* 3. We legitimately assume $b < \frac{n}{2} - \mathrm{o}(1)$ since it is not very meaningful to discuss SGD that uses more than (roughly) half of the training set as a mini-batch. Then the learning rate schedule in (4) intersects with that in (6), i.e., (6) *covers both moderate and small learning rate schemes*. And their intersection determines a moderate learning rate scheme, where SGD converges along the large eigenvalue directions while GD goes after the small eigenvalue directions. This justifies the regularization effect of SGD with moderate learning rate.

*Remark* 4. Technically, in Theorem 1 one can set $b = n$ to include GD as a special case, so that GD also follows the large eigenvalue directions. However, the initial learning rate in (4) needs to be at least $\frac{n}{\lambda_1} \geq n$ ignoring the small order term, which is way too large to be even numerically stable in practical big data circumstances. The typical learning rate for GD falls into learning rate schedule in (6). Thus it follows the small eigenvalue directions according to Theorem 2.

[2] Consider two sequences $\{x_n\}$ and $\{y_n\}$. We denote $x_n = \mathcal{O}(y_n)$ if there exist constants $C$ and $N$ such that $|x_n|/|y_n| \leq C$ for all $k \geq N$. We denote $x_n = \mathrm{o}(y_n)$ if for any $C > 0$ there exists a positive constant $N(C)$ such that $|x_n|/|y_n| \leq C$ for all $k \geq N(C)$. We denote $x_n \geq \operatorname{poly}(y_n)$ if there exist constants $C$, $D$, and $N$ such that $|x_n| \geq |y_n|^D + C$ for all $n \geq N$.




(a) Linear regression on synthetic data

(b) Neural network on small FashionMNIST

Figure 2: **(a)**: A linear regression example. We randomly draw 100 samples from a $10,000$-dimensional space as described in Section 4, where $\zeta \sim \mathcal{U}([0.5, 1])$. The small learning rate scheme is specified by $(\eta', k_2) = (0.2, 10^4)$, and the moderate learning rate scheme is specified by $(\eta, \eta', k_1, k_2) = (1.05, 0.1, 2\times 10^3, 3\times 10^3)$. Numerical results show the Rayleigh quotient converges to its maximum for SGD with moderate learning rate, and converges to its minimum for GD and SGD with small learning rate, which verifies Theorem 1, 2 and 3. **(b)**: A neural network example. We randomly draw $2,000$ samples from FashionMNIST as the training set. The model is a 5-layer convolutional neural network. The small learning rate scheme is specified by $(\eta', k_2) = (10^{-3}, 10^4)$, and the moderate learning rate scheme is specified by $(\eta, \eta', k_1, k_2) = (10^{-2}, 10^{-3}, 4 \times 10^3, 10^4)$. Experiment shows SGD with moderate learning rate achieves the highest test accuracy, and GD and SGD with small learning rate perform similarly, but are worse.

Note GD with small learning rate also converges along the small eigenvalue directions, since the learning rate schedule in (6) covers the small learning rate scheme. In complement, the following Theorem 3 shows that in the small learning rate regime, SGD is imitating GD and converges along the small eigenvalue directions as well. Theorem 1, 2 and 3 together show that, converging along the large eigenvalue directions is a distinct regularization effect that is unique to *SGD* with *moderate learning rate*.

**Theorem 3** (The directional bias of SGD with small LR, informal)**.** *Theorem 2 applies to* (SGD) *with the following small learning rate scheme*

$$\eta_k = \eta' \in \left(0,\ \frac{b}{2\lambda_1 + \mathrm{o}(1)}\right), \quad k = 1, \ldots, k_2. \tag{8}$$

**Experiment** We conduct a numerical experiment and plot the Rayleigh quotients (5) and (7) for the SGD and GD iterates in both moderate and small learning rate regimes in Figure 2(a). We can see that the Rayleigh quotient reaches its maximum for SGD with moderate learning rate, and reaches its minimum for both GD and SGD with small learning rate, which verifies Theorem 1, 2 and 3.

### 4.2 Effects of the directional bias

Next we justify the benefit of the particular directional bias of SGD with moderate learning rate.

Recall the hypothesis class $\mathcal{H}_{\mathcal{S}}$ (Eq. (2)) for SGD and GD. Then for an algorithm output $w^{\mathrm{alg}}$, we have the following generalization error decomposition (Shalev-Shwartz & Ben-David, 2014),

$$L_{\mathcal{D}}(w^{\mathrm{alg}}) - \inf_{w} L_{\mathcal{D}}(w) = \underbrace{L_{\mathcal{D}}(w^{\mathrm{alg}}) - \inf_{w' \in \mathcal{H}_{\mathcal{S}}} L_{\mathcal{D}}(w')}_{\Delta(w^{\mathrm{alg}}),\ \text{estimation error}} + \underbrace{\inf_{w' \in \mathcal{H}_{\mathcal{S}}} L_{\mathcal{D}}(w') - \inf_{w} L_{\mathcal{D}}(w)}_{\text{approximation error}}.$$

The *approximate error* is an intrinsic error determined by the hypothesis class, and is not improvable unless enlarging the hypothesis class. In contrast, the *estimation error* $\Delta(w^{\mathrm{alg}})$ is determined by the algorithm as well as its hyperparameters. Therefore we will use the estimation error to compare the generalization performance of the SGD and GD outputs in different learning rate regimes. In particular, we consider the *optimality within an $\alpha$-level set* (of the training loss) defined as follows.

**Definition** (Optimality within an $\alpha$-level set)**.** Let $\mathcal{W} := \left\{w \in \mathcal{H}_{\mathcal{S}} : L_{\mathcal{S}}(w) = \inf_{w \in \mathcal{H}_{\mathcal{S}}} L_{\mathcal{S}}(w) + \alpha\right\}$ be an $\alpha$-level set of the training loss. Let $\Delta_* := \inf_{w \in \mathcal{W}} \Delta(w)$ be the optimal estimation error within the $\alpha$-level set $\mathcal{W}$. Then for a parameter $w \in \mathcal{W}$, we say $w$ is *$\epsilon$-optimal*, if $\Delta(w) < (1+\epsilon)\cdot\Delta_*$ for small $\epsilon$; we say $w$ is *$M$-suboptimal*, if $\Delta(w) > M \cdot \Delta_*$ for constant $M > 1$.

*Remark* 5. In practice it is usually intractable and unnecessary to achieve the exact global minima of the training loss; instead we often *early stop* the algorithm once obtaining a small enough training loss, i.e., reaching an $\alpha$-level set. Thus in order to characterize the generalization ability of an algorithm, it is reasonable to compare its early stopped output with the best possible outcome within the $\alpha$-level set, as in our definition.

The following theorem justifies the benefit of convergence along the large eigenvalue directions, as adopted by SGD in the moderate learning rate regime.

**Theorem 4** (Effects of the directional bias, informal)**.** *Under the same conditions as Theorem 1 and 2, if $b < \frac{n}{2} - \mathrm{o}(1)$, then there exists $k_1$ and $k_2$ such that with high probability:*

- *SGD with moderate LR (4) outputs an $\epsilon$-optimal solution, for small $\epsilon = \mathrm{o}(1)$;*
- *GD with moderate or small LR (6) outputs an $M$-suboptimal solution, for constant $M > 1$;*
- *SGD with small LR (8) outputs an $M$-suboptimal solution, for constant $M > 1$.*

*Remark* 6. The second conclusion is also obtained by Nakkiran (2020) in special case. However they have not studied SGD, which is more challenging in theory and insightful for practice, nor have they characterized the directional bias as in Theorems 1 and 2.

From Theorem 4 we conclude: (1) in the moderate learning rate regime, early stopped SGD finds a nearly optimal solution thanks to its particular directional bias. In contrast, early stopped GD can only find a suboptimal one; (2) in the small learning rate regime, however, SGD no longer admits the dedicated directional bias for moderate learning rate. Instead it behaves similarly as GD, and hence outputs suboptimal solutions when early stopping is adopted.

**Experiment** Figure 2(b) shows the generalization performance of a neural network trained by SGD and GD in both moderate and small learning rate regimes. We can observe that (1) SGD with moderate learning rate generalizes the best, and (2) GD and SGD with small learning rate perform similarly, but both are worse than SGD with moderate learning rate. The empirical results suggest SGD with moderate learning has certain benign regularization effect. This is explained by the distinct directional bias for SGD with moderate learning rate shown in previous theorems.

## 5 Related Work

In this section, we review related works and discuss their similarities and differences to our work.

**The implicit bias of GD** The implicit bias of GD has been extensively studied in recent years. We summarize several representative results as follows. For homogeneous model with exponentially-tailed loss, GD converges along the max-margin direction (Gunasekar et al., 2017; 2018a;b; Soudry et al., 2018; Ma et al., 2018; Ji & Telgarsky, 2019a; Ji et al., 2020; Nacson et al., 2019a). For least square problem and its variants, GD is biased towards minimum-norm solution (Gunasekar et al., 2018a; Ali et al., 2019; Suggala et al., 2018). Note that this is also the foundation for the learning theory in the interpolation regime such as *double descent* (Belkin et al., 2019) and *benign overfitting* (Bartlett et al., 2020). For matrix factorization and linear network, Gunasekar et al. (2017); Li et al. (2018) show the GD solution minimizes nuclear norm in special cases; more generally, Arora et al. (2019); Ji & Telgarsky (2019b) show GD balances/aligns layers; Moroshko et al. (2020) suggest the GD bias relies on initialization scale and training accuracy. For infinite-width network, Chizat & Bach (2020) show GD finds a max-margin classifier in a functional space. The regularization theory for GD is fruitful; While some of them can be applied to SGD with small learning rate as have discussed (Nacson et al., 2019b; Gunasekar et al., 2018a; Ali et al., 2020), none of them can be carried over to SGD in the moderate learning rate regime. As far as we know, our paper is the first to study the regularization effect of SGD with moderate learning rate.

**The stability of SGD** *Stability* is another approach to justify the generalization ability of SGD, where a stable algorithm is guaranteed to have small generalization error (Bousquet & Elisseeff, 2002). Along this line, several works show that SGD is stable under certain assumptions and therefore generalizes well (Hardt et al., 2016; Kuzborskij & Lampert, 2018; Charles & Papailiopoulos, 2018; Bassily et al., 2020). More interestingly, Charles & Papailiopoulos (2018) show a simple example where SGD is stable but GD is not, which partly explains the empirically superior performance of SGD. We also aim to justify the benefits of SGD, but take a different approach from the algorithmic regularization.

**The escaping behavior of SGD** A popular theory (Jastrzębski et al., 2017; Zhu et al., 2019; Simsekli et al., 2019) attributes the regularization effect of SGD to its behavior of escaping from sharp minima. These works are built upon the continuous approximation of SGD via stochastic differential equations (Li et al., 2017; Hu et al., 2017; Xu et al., 2018; Simsekli et al., 2019). However it requires a small learning rate for the approximation to hold. In contrast, our main interest is in the moderate learning rate regime. Another related work in this line is (Wu et al., 2018), where they study the dynamical stability of minima and how SGD chooses them, but they do not show a directional bias introduced in our work.

**Non-small learning rate** The regularization effect of non-small learning rate has received increasing attentions recently. Theoretically, Li et al. (2019); HaoChen et al. (2020) study the generalization of certain stochastic dynamics equipped with annealing learning rate scheme. However, their results cannot cover vanilla SGD. Lewkowycz et al. (2020) study the role of initial large learning rate for infinity-width network. Our work is motivated by Nakkiran (2020), which shows that a large initial learning rate helps GD generalize. This result can be recovered from our theorems; more importantly, we characterize the directional bias of SGD. Empirically, Jastrzebski et al. (2019); Leclerc & Madry (2020) investigate the impact of two-phase learning rate for training with SGD, but they do not provide any theoretical analysis.

## 6 Discussions

**Linear scaling rule** Linearly enlarging the initial learning rate according to batch size, or the *linear scaling rule* (Goyal et al., 2017), is an important folk art for paralleling SGD with large batch size while maintaining a good generalization performance. Interestingly, the linear scaling rule arises naturally from Theorem 1, where LR (4) suggests the initial learning rate $\eta$ to scale linearly with batch size $b$ to guarantee the desired directional bias. Our theory thus partly explains the mechanism of the linear scaling rule.

**Overrunning SGD with high learning rate** Besides the moderate initial learning rate, another key ingredient in Theorem 1 is a sufficiently large $k_1$, i.e., SGD needs to run with moderate learning rate for sufficiently long time (to fit small eigenvalue directions). This requirement coincidentally agrees with the folk art to *overrun SGD with high learning rate*. For example, see Figure 4 in (He et al., 2016): from the $\sim 2 \times 10^5$-th to the $\sim 3 \times 10^5$-th iteration, even the training error seems to make no progress, practitioners let SGD run with a relatively high learning rate for nearly $10^5$ iterates. Our theory sheds light on understanding the hidden benefits in this "overrunning" phase: indeed the loss is not decreasing since SGD with high learning rate cannot fit the large eigenvalue directions, but on the other hand the overrunning lets SGD fit the small eigenvalue directions better, which in the end leads to the directional bias that SGD converges along the large eigenvalue directions according to Theorem 1. Thus overrunning SGD with high learning rate is helpful.

**Key technical challenges** With non-small learning rate, analyzing SGD is usually hard since measure concentration turns vacuous and as a result one cannot relate the SGD iterates to that of GD. Alternatively, we control the SGD iterates epoch by epoch, then bound their composition to characterize the long run behavior of SGD. However, controlling the epoch-wise update of SGD is highly non-trivial, since in different epochs the sequence of stochastic gradient steps varies, and they do not commute due to the issue of matrix multiplication. To overcome this difficulty, we adopt techniques from matrix perturbation theory (Horn & Johnson, 2012) and conduct an analysis in the overparameterized regime. We believe these techniques are of independent interest. As an initial attempt, our results are limited to linear model and high dimensional data, which is already insightful. It is left as future works to extend our results to nonlinear/nonconvex models and low dimensional data.

## 7 Conclusion

We characterize a distinct directional regularization effect of SGD with moderate learning rate, where SGD converges along the large eigenvalue directions of the data matrix. In contrast, neither GD nor SGD with small learning rate can achieve this effect. Moreover, we show this directional bias benefits generalization when early stopping is adopted. Finally, our theory explains several folk arts used in practice for SGD hyperparameter tuning.

## REFERENCES


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*, pp. 1370–1378, 2019.

Alnur Ali, Edgar Dobriban, and Ryan J Tibshirani. The implicit regularization of stochastic gradient flow for least squares. *arXiv preprint arXiv:2003.07802*, 2020.

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. In *Advances in Neural Information Processing Systems*, pp. 7413–7424, 2019.

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

Raef Bassily, Vitaly Feldman, Cristóbal Guzmán, and Kunal Talwar. Stability of stochastic gradient descent on nonsmooth convex losses. *arXiv preprint arXiv:2006.06914*, 2020.

Mikhail Belkin, Daniel Hsu, and Ji Xu. Two models of double descent for weak features. *arXiv preprint arXiv:1903.07571*, 2019.

Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.

Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Zachary Charles and Dimitris Papailiopoulos. Stability and generalization of learning algorithms that converge to global optima. In *International Conference on Machine Learning*, pp. 745–754. PMLR, 2018.

Lenaic Chizat and Francis Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. *arXiv preprint arXiv:2002.04486*, 2020.

Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

Robert Mansel Gower, Nicolas Loizou, Xun Qian, Alibek Sailanbayev, Egor Shulgin, and Peter Richtárik. Sgd: General analysis and improved rates. *arXiv preprint arXiv:1901.09401*, 2019.

Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

Suriya Gunasekar, Blake E Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems*, pp. 6151–6159, 2017.

Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In *ICML*, 2018a.

Suriya Gunasekar, Jason D Lee, Daniel Soudry, and Nati Srebro. Implicit bias of gradient descent on linear convolutional networks. In *Advances in Neural Information Processing Systems*, pp. 9461–9471, 2018b.

Jeff Z HaoChen, Colin Wei, Jason D Lee, and Tengyu Ma. Shape matters: Understanding the implicit bias of the noise covariance. *arXiv preprint arXiv:2006.08680*, 2020.

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning*, pp. 1225–1234, 2016.

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*, pp. 770–778, 2016.

Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

Wenqing Hu, Chris Junchi Li, Lei Li, and Jian-Guo Liu. On the diffusion approximation of nonconvex stochastic gradient descent. *arXiv preprint arXiv:1705.07562*, 2017.

Stanisław Jastrzębski, Zachary Kenton, Devansh Arpit, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amos Storkey. Three factors influencing minima in sgd. *arXiv preprint arXiv:1711.04623*, 2017.

Stanislaw Jastrzebski, Maciej Szymczak, Stanislav Fort, Devansh Arpit, Jacek Tabor, Kyunghyun Cho, and Krzysztof Geras. The break-even point on optimization trajectories of deep neural networks. In *International Conference on Learning Representations*, 2019.

Ziwei Ji and Matus Telgarsky. The implicit bias of gradient descent on nonseparable data. In *Conference on Learning Theory*, pp. 1772–1798, 2019a.

Ziwei Ji and Matus Jan Telgarsky. Gradient descent aligns the layers of deep linear networks. In *7th International Conference on Learning Representations, ICLR 2019*, 2019b.

Ziwei Ji, Miroslav Dudík, Robert E Schapire, and Matus Telgarsky. Gradient descent follows the regularization path for general losses. In *Conference on Learning Theory*, pp. 2109–2136, 2020.

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.

Harold Kushner and G George Yin. *Stochastic approximation and recursive algorithms and applications*, volume 35. Springer Science & Business Media, 2003.

Ilja Kuzborskij and Christoph Lampert. Data-dependent stability of stochastic gradient descent. In *International Conference on Machine Learning*, pp. 2815–2824. PMLR, 2018.

Guillaume Leclerc and Aleksander Madry. The two regimes of deep network training. *arXiv preprint arXiv:2002.10376*, 2020.

Aitor Lewkowycz, Yasaman Bahri, Ethan Dyer, Jascha Sohl-Dickstein, and Guy Gur-Ari. The large learning rate phase of deep learning: the catapult mechanism. *arXiv preprint arXiv:2003.02218*, 2020.

Qianxiao Li, Cheng Tai, et al. Stochastic modified equations and adaptive stochastic gradient algorithms. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 2101–2110. JMLR. org, 2017.

Yuanzhi Li, Tengyu Ma, and Hongyang Zhang. Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. In *Conference On Learning Theory*, pp. 2–47. PMLR, 2018.

Yuanzhi Li, Colin Wei, and Tengyu Ma. Towards explaining the regularization effect of initial large learning rate in training neural networks. In *Advances in Neural Information Processing Systems*, pp. 11674–11685, 2019.

Cong Ma, Kaizheng Wang, Yuejie Chi, and Yuxin Chen. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval and matrix completion. In *International Conference on Machine Learning*, pp. 3345–3354, 2018.

Edward Moroshko, Suriya Gunasekar, Blake Woodworth, Jason D Lee, Nathan Srebro, and Daniel Soudry. Implicit bias in deep linear classification: Initialization scale vs training accuracy. *arXiv preprint arXiv:2007.06738*, 2020.

Mor Shpigel Nacson, Jason Lee, Suriya Gunasekar, Pedro Henrique Pamplona Savarese, Nathan Srebro, and Daniel Soudry. Convergence of gradient descent on separable data. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 3420–3428. PMLR, 2019a.

Mor Shpigel Nacson, Nathan Srebro, and Daniel Soudry. Stochastic gradient descent on separable data: Exact convergence with a fixed learning rate. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 3051–3059. PMLR, 2019b.

Preetum Nakkiran. Learning rate annealing can provably help generalization, even for convex problems. *arXiv preprint arXiv:2005.07360*, 2020.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Umut Simsekli, Mert Gürbüzbalaban, Thanh Huy Nguyen, Gaël Richard, and Levent Sagun. On the heavy-tailed theory of stochastic gradient descent for deep neural networks. *arXiv preprint arXiv:1912.00018*, 2019.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19 (1):2822–2878, 2018.

Arun Suggala, Adarsh Prasad, and Pradeep K Ravikumar. Connecting optimization and regularization paths. In *Advances in Neural Information Processing Systems*, pp. 10608–10619, 2018.

Lloyd N Trefethen and David Bau III. *Numerical linear algebra*, volume 50. Siam, 1997.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Jingfeng Wu, Wenqing Hu, Haoyi Xiong, Jun Huan, Vladimir Braverman, and Zhanxing Zhu. On the noisy gradient descent that generalizes as sgd. *The 37th International Conference on Machine Learning*, 2020.

Lei Wu, Chao Ma, and E Weinan. How sgd selects the global minima in over-parameterized learning: A dynamical stability perspective. In *Advances in Neural Information Processing Systems*, pp. 8279–8288, 2018.

Pan Xu, Jinghui Chen, Difan Zou, and Quanquan Gu. Global convergence of langevin dynamics based algorithms for nonconvex optimization. In *Advances in Neural Information Processing Systems*, pp. 3122–3133, 2018.

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

Zhanxing Zhu, Jingfeng Wu, Bing Yu, Lei Wu, and Jinwen Ma. The anisotropic noise in stochastic gradient descent: Its behavior of escaping from sharp minima and regularization effects. *The 36th International Conference on Machine Learning*, 2019.