The Implicit Regularization of Momentum Gradient Descent in Overparametrized Models

Li Wang[1], Zhiguo Fu[1,*] Yingcong Zhou[1], Zili Yan[2]

[1]School of Computer Science and Information Technology & KLAS, Northeast Normal University, China
[2]School of Mathematics and Statistics, Beihua University, China
wangl024@nenu.edu.cn, zycong0821@163.com, mathyanzili@163.com, fuzg432@nenu.edu.cn

Abstract

The study of the implicit regularization induced by gradient-based optimization in deep learning is a long-standing pursuit. In the present paper, we characterize the implicit regularization of momentum gradient descent (MGD) in the continuous-time view, so-called momentum gradient flow (MGF). We show that the components of weight vector are learned for a deep linear neural networks at different evolution rates, and this evolution gap increases with the depth. Firstly, we show that if the depth equals one, the evolution gap between the weight vector components is linear, which is consistent with the performance of ridge. In particular, we establish a tight coupling between MGF and ridge for the least squares regression. In detail, we show that when the regularization parameter of ridge is inversely proportional to the square of the time parameter of MGF, the risk of MGF is no more than 1.54 times that of ridge, and their relative Bayesian risks are almost indistinguishable. Secondly, if the model becomes deeper, i.e. the depth is greater than or equal to 2, the evolution gap becomes more significant, which implies an implicit bias towards sparse solutions. The numerical experiments strongly support our theoretical results.

Introduction

The central question of deep learning, how neural networks generalize, still eludes full theoretical understanding. Recently, it has been shown that optimization may be a key to understanding the generalization mystery of deep learning (Zhang et al. 2016). This finding, along with a series of studies (Gunasekar et al. 2018; Soudry et al. 2018; Vardi and Shamir 2021; Zhang et al. 2021), suggest that while the hypothesis space itself is extremely complex, the search strategy of optimization favors certain structured solution as if some explicit regularization term appeared in its objective. This preference of optimization is called the implicit regularization of optimization and how to characterize it has developed as an open problem in deep learning theories.

Recent results state that increasing the depth of the neural networks will modify the learning rule of optimization and enhances the implicit regularization of optimization (Vaskevicius, Kanade, and Rebeschini 2019; Gidel, Bach, and Lacoste-Julien 2019; Li et al. 2021). For example, (Arora, Cohen, and Hazan 2018) claims increasing depth can speed up optimization. Based on this work,(Arora et al. 2019) find that adding depth to a matrix factorization enhances an implicit tendency towards low-rank solutions. Additionally, (Gissin, Shalev-Shwartz, and Daniely 2019) shows that a depth-2 model requires exponentially small initialization for incremental learning to occur, while deeper models only require the initialization to be polynomially small. Our work deals with a simple setting, allowing us to explore that how depth effect the implicit regularization of optimization.

In this paper, we aim to characterize the implicit regularization of MGD (Polyak 1964), which is one of the most popular optimization algorithms in practice because of its ability to accelerate learning, especially for the cases of high curvature, small but consistent gradients, or noisy gradients. Many other variants and improvements of MGD have been developed in (Cyrus et al. 2018; Lin, Li, and Fang 2020; Even et al. 2021), and their convergence behaviors have been analyzed. It has been empirically observed that MGD and its variants (e.g., Nesterov) perform well in deep learning. However, there is a lack of theoretical discussions to uncover how MGD affects generalization performance, which is our consideration in the present paper.

An important way to explore the implicit regularization of optimization is to compare the optimization paths with the explicit regularization paths. Recently, (Suggala, Prasad, and Ravikumar 2018; Ali, Kolter, and Tibshirani 2019) showed how the optimization path of GD is (point-wise) closely connected to an explicit ℓ_2 regularization. In a similar idea, (Ali, Dobriban, and Tibshirani 2020) studied the implicit regularization of stochastic gradient descent (SGD). Furthermore, (Zou et al. 2021) showed that the generalization performance of SGD is always no worse than that of ridge regression in a wide range of overparameterized problems. More results can be found in (Steinerberger 2021; Smith et al. 2021; Sheng and Ali 2022).

In the present paper, for the depth $N = 1$, we firstly study the implicit regularization of MGD by comparing its path to the path of ridge for the least squares problems. Note that MGD is a second-order iteration essentially, so the corresponding continuous-time form MGF is a second-order differential equation. We find the analytical solution of MGF by the singular value decomposition of the data matrix. But

[*]Corresponding author

the solution is involved and it is more challenging to show the coupling between MGF and ridge. For $N \geq 2$, it is difficult to give the analytical solution of MGF. Therefor, we will study the dynamics of MGF and the evolution gap of the components of the weight vector in the depth-N linear models. The main contributions of this paper are as follows:

- We give the analytical continuous-time form of MGD for the least squares problems, called MGF, and prove that MGD convergences to MGF as the step size $\epsilon \to 0$.
- We find that MGF can be expressed as the solutions to a sequence of ℓ_2 regularized least squares problems, and then set the calibration of early stopping $t = \sqrt{2/\lambda}$.
- We show that the risk of MGF is no more than 1.54 times that of ridge regression at tuning parameter $\lambda = 2/t^2$. And the ratio of the Bayes risk of MGF to that of ridge is between 1 and 1.035 under the optimal tuning.
- We show that the components of weight vector are learned for a depth-N linear networks at different evolution rates, and this evolution gap increases with depth. This tendency implies an implicit bias towards sparse solutions, and intensifies with depth.
- We carry out the numerical experiments to verify our theoretical results.

Dynamical Analysis of the Least Squares Problems

We begin by investigating the implicit regularization of MGF for a simple yet striking model — the least squares regression (i.e. depth-1 model). This will enable us to attain the closed-form expression for MGF and obtain fruitful analysis results. Specifically, we will see that the path of MGF is extremely close to ridge, and they have the similar generalization behavior.

Momentum Gradient Flow

Let $X \in \mathbb{R}^{n\times p}$, a column full-rank matrix, is the data matrix and $y \in \mathbb{R}^n$ is the response vector. We would like to analyse the learning by minimizing the loss function,

$$\min_{\beta\in\mathbb{R}^p} \ L(f(X;\beta), y), \tag{1}$$

where L is the loss function, $\beta \in \mathbb{R}^p$ is the weight vector and $f(X;\beta)$ is the predicted output when the input is X. We are particularly interested in the implicit regularization of MGD applied to (1). Consider the standard MGD iterations

$$\begin{aligned}\tilde{v}_{k+1} &= \tilde{\mu}\tilde{v}_k - \tilde{\epsilon} g(\beta_k),\\ \beta_{k+1} &= \beta_k + \tilde{v}_{k+1},\end{aligned}$$

where $g(\beta_k) = \nabla_\beta L(f(X;\beta), y)$; $\tilde{\epsilon} > 0$ is the step size; $\tilde{v}$ is momentum which is set to an exponentially decaying average of the negative gradient; and $\tilde{\mu} \in (0,1)$ is the momentum parameter that determines how quickly the contributions of previous gradients decay. To facilitate the following analysis, we consider a rescaled version of MGD. By redefining $\epsilon = \sqrt{\tilde{\epsilon}}, v_k = \frac{\tilde{v}_k}{\sqrt{\tilde{\epsilon}}}$ and $\mu = \frac{1-\tilde{\mu}}{\sqrt{\tilde{\epsilon}}}$, we have

$$\begin{aligned}v_{k+1} &= v_k - \mu\epsilon v_k - \epsilon g(\beta_k),\\ \beta_{k+1} &= \beta_k + \epsilon v_{k+1}.\end{aligned} \tag{2}$$

After rescaling, we have the momentum parameter $\mu \in (0, \epsilon^{-1/2})$ from $\tilde{\mu} \in (0,1)$. It follows that

$$\beta_{k+1} = \beta_k + \epsilon v_k - \mu\epsilon^2 v_k - \epsilon^2 g(\beta_k).$$

Moreover, let $v_k = \frac{\beta_k - \beta_{k-1}}{\epsilon}$, then we have

$$\beta_{k+1} = 2\beta_k - \beta_{k-1} - \mu\epsilon(\beta_k - \beta_{k-1}) - \epsilon^2 g(\beta_k). \tag{3}$$

(3) shows that MGD is a second-order iteration essentially. Rearranging (3) yields that

$$\frac{\beta_{k+1} + \beta_{k-1} - 2\beta_k}{\epsilon^2} + \mu\frac{(\beta_k - \beta_{k-1})}{\epsilon} = -g(\beta_k).$$

Letting $\epsilon \to 0$, we get the continuous-time form of MGD

$$\beta''(t) + \mu\beta'(t) = -g(\beta(t)), \tag{4}$$

over time $t \geq 0$. We call (4) MGF which are the second-order differential equations. For depth $N = 1$, we focus on the analysis of the implicit regularization of MGF through the least squares problem, which is

$$\min_{\beta\in R^p} \ L(f(X;\beta), y) = \frac{1}{2n}||y - X\beta||_2^2, \tag{5}$$

and $g(\beta_k) = \frac{1}{n}X^{\mathrm{T}}X\beta_k - \frac{1}{n}X^{\mathrm{T}}y$. Let $X = \sqrt{n}US^{\frac{1}{2}}V^{\mathrm{T}}$ be the singular value decomposition, thus $X^{\mathrm{T}}X/n = VSV^{\mathrm{T}}$ is the eigendecomposition, where $S = \mathrm{diag}(s_i)(i = 1, \cdots, p)$ and s_i are the eigenvalues of $X^{\mathrm{T}}X/n$ satisfying $s_1 \geq s_2 \geq \cdots s_p > 0$. We note that $X^{\mathrm{T}}X/n$ is a symmetric positive definite matrix, since X has the rank p. And then applying MGD to (5) initialized at $v_0 = -\frac{\epsilon X^{\mathrm{T}}y}{2n(1-\mu\epsilon)}$ and $\beta_0 = 0$ (which implies that $\beta_1 = \frac{1}{2n}\epsilon^2 X^{\mathrm{T}}y$ by (2)), we have

$$\begin{aligned}\beta_{k+1} =& 2\beta_k - \beta_{k-1} - \epsilon D(\mu)(\beta_k - \beta_{k-1})\\ &- \epsilon^2\left(VSV^{\mathrm{T}}\beta_k - \frac{1}{\sqrt{n}}VS^{\frac{1}{2}}U^{\mathrm{T}}y\right),\end{aligned} \tag{6}$$

where $D(\mu) = \mathrm{diag}(\mu)$ and the corresponding MGF is

$$\beta''(t) + D(\mu)\beta'(t) + VSV^{\mathrm{T}}\beta(t) = \frac{1}{\sqrt{n}}VS^{\frac{1}{2}}U^{\mathrm{T}}y \tag{7}$$

for $t \geq 0$, which subjects to the initial conditions $\beta(0) = 0$, $\beta'(0) = 0$. Now, we derive the exact solution of MGF.

Lemma 1. *Fix a response y and a data matrix X. The MGF (7), subject to $\beta(0) = 0$, $\beta'(0) = 0$ and $D(\mu) \succ 2S^{1/2}$ admits the exact solution*

$$\hat{\beta}^{\mathrm{mgf}}(t) = \frac{1}{\sqrt{n}}VS^{-1}\left(I - H(S,t)\right)S^{\frac{1}{2}}U^{\mathrm{T}}y \tag{8}$$

where

$$\begin{aligned}&H(S,t) =\\ &\left(2\sqrt{D(\mu)^2 - 4S}\right)^{-1}\Big[\left(D(\mu) + \sqrt{D(\mu)^2 - 4S}\right)\cdot\\ &\exp\left(\frac{1}{2}\left(-D(\mu) + \sqrt{D(\mu)^2 - 4S}\right)t\right)\\ &-\left(D(\mu) - \sqrt{D(\mu)^2 - 4S}\right)\cdot\\ &\exp\left(\frac{1}{2}\left(-D(\mu) - \sqrt{D(\mu)^2 - 4S}\right)t\right)\Big].\end{aligned}$$

Proof. The result follows from solving the second-order differential equations (7)-*see Supplement S.1.* □

Throughout this paper, $H(S,t)$ is defined as above, $\prec$ denotes the Loewner ordering on the matrices (i.e., $A \prec B$ means that $B - A$ is positive definite), $\|v\|$ denotes the Euclidean norm of a vector v and $\|A\|$ denotes the spectral norm of a matrix A. The following lemma shows that MGD (point-wise) convergences to MGF as the step size $\epsilon \to 0$.

Lemma 2. *For the least squares (5), consider MGD $\{\beta_k : k = 0, \cdots, n\}$ (6) initialized at $v_0 = -\frac{\epsilon X^{\mathrm{T}} y}{2n(1-\mu\epsilon)}$ and $\beta_0 = 0$, MGF $\{\beta(t) : t \in [0,T]\}$ (8) subjects to $\beta(0) = 0, \beta'(0) = 0$. Partition the interval of $[0,T]$ into a uniform mesh with the step size ϵ, i.e. $n+1 = \lfloor T/\epsilon \rfloor$. It holds that*

$$\|\hat{\beta}^{\mathrm{mgf}}(t_{k+1}) - \beta_{k+1}\| \leq \epsilon C, \tag{9}$$

where $\hat{\beta}^{\mathrm{mgf}}(t_{k+1})$ is the value of the exact solution of MGF at the $k+1$-th grid point and C is a positive constant.

Proof. The uniform bound is given by numerical analysis-*see Supplement S.2.* □

Lemma 2 ensures that we can use the exact solution of the continuous-time MGF to study the implicit regularization of MGD in the following analysis.

Basic Comparisons Between MGF and Ridge

Consider the ridge regression, the ℓ_2 regularized version of (5), that is

$$\min_{\beta \in R^p} \frac{1}{2n} ||y - X\beta||_2^2 + \lambda ||\beta||_2^2, \tag{10}$$

where $\lambda > 0$ is a tuning parameter. The explicit ridge solution is

$$\hat{\beta}^{\mathrm{ridge}}(\lambda) = (X^{\mathrm{T}} X + n\lambda I)^{-1} X^{\mathrm{T}} y. \tag{11}$$

To compare the paths of ridge (11) and MGF (8), it is helpful to rearrange them, on the scale of fitted values, to

$$\begin{aligned} X\hat{\beta}^{\mathrm{ridge}}(\lambda) &= US^{\frac{1}{2}} V^{\mathrm{T}} V (S + \lambda I)^{-1} S^{\frac{1}{2}} U^{\mathrm{T}} y \\ &= US(S + \lambda I)^{-1} U^{\mathrm{T}} y. \qquad (12) \\ X\hat{\beta}^{\mathrm{mgf}}(t) &= US^{\frac{1}{2}} V^{\mathrm{T}} V S^{-1} \left(I - H(S,t)\right) S^{\frac{1}{2}} U^{\mathrm{T}} y \\ &= U \left(I - H(S,t)\right) U^{\mathrm{T}} y. \qquad (13) \end{aligned}$$

Letting $u_i \in \mathbb{R}^n, i = 1, \cdots, p$ denote the columns of U, we note that (12) and (13) are both linear smoothers (linear functions of y) of the form $\sum_{i=1}^{p} \varphi(s_i, \kappa) \cdot u_i u_i^{\mathrm{T}} y$, for a spectral shrinkage map $\varphi(\cdot, \kappa) : [0, \infty) \to [0, \infty)$ and parameter κ. This map is $\varphi^{\mathrm{ridge}}(s, \lambda) = s/(s + \lambda)$ for ridge, and $\varphi^{\mathrm{mgf}}(s,t) = 1 - H(s,t)$ for MGF. We see that both apply more shrinkage for smaller values of s, i.e., lower-variance directions of $X^{\mathrm{T}} X/n$, but do so in apparently different ways. And the two shrinkage maps agree at the extreme ends (i.e., set $\lambda \to 0$ and $t \to \infty$, $\varphi(s, \cdot) \to 1$, or $\lambda \to \infty$ and $t \to 0$, $\varphi(s, \cdot) \to 0$). We note that the parametrization $\lambda = 2/t^2$ (the calibration setting is obtained by Taylor expansion and will be explained in the next subsection.) gives the two shrinkage maps similar behaviors: see Figure 1 for a visualization. Moreover, as we will show later, the two shrinkage maps (under the calibration $\lambda = 2/t^2$) lead to similar risk curves for MGF and ridge.

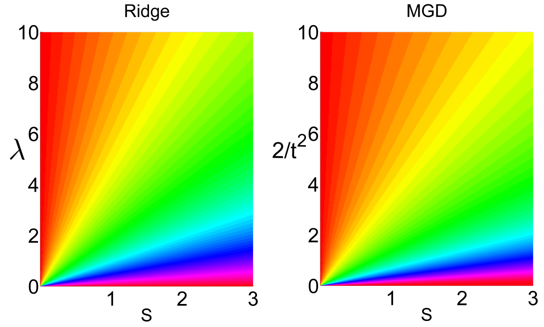

Figure 1: Comparison of MGF and ridge spectral shrinkage maps.

Underlying Regularization Problems

We are interested in the connection between MGF and ridge. It is natural to wonder whether MGF can be expressed as solutions to sequences of the regularized least squares. The following lemma confirms this conjecture.

Lemma 3. *Fix y and X. Under the initial conditions $\beta(0) = 0, \beta'(0) = 0$, for $t \geq 0$, MGF (8) uniquely solves the optimization problem*

$$\min_{\beta \in R^p} \frac{1}{2n} \|y - X\beta\|_2^2 + \beta^{\mathrm{T}} Q_t \beta, \tag{14}$$

where $Q_t = VS\left(H(S,t)^{-1} - I\right)^{-1} V^{\mathrm{T}}$.

Proof. The result readily follows from MGF (8) and the solution of (14)-*see Supplement S.3.* □

Remark 1. *Computing the first two orders of the Taylor's series of $H(S,t)^{-1}$ at the point $t = 0$, we have*

$$H(S,t)^{-1} \approx I + H'(S,t)^{-1} t + \frac{1}{2} H''(S,t)^{-1} t^2 \approx I + \frac{1}{2} t^2 S.$$

An application of the claim of Lemma 3 can give the expression of the regularization parameter

$$Q_t = VS(H(S,t)^{-1} - I)^{-1} V^{\mathrm{T}} \approx VS(\frac{1}{2} t^2 S)^{-1} V^{\mathrm{T}} = \frac{2}{t^2} I.$$

It shows that MGF is extremely close to ridge, under the calibration $\lambda = 2/t^2$.

Estimation Risk

For any feature matrix $X \in \mathbb{R}^{n \times p}$, we consider a generic response model,

$$y \mid \beta_0 \sim (X\beta_0, \sigma^2 I), \tag{15}$$

i.e., $E(y \mid \beta_0) = X\beta_0, \mathrm{Cov}(y \mid \beta_0) = \sigma^2 I$ for the underlying coefficient vector $\beta_0 \in \mathbb{R}^p$ and the error variance $\sigma^2 > 0$. For an estimator $\hat{\beta}$ (i.e., measurable function of X, y), we define its estimation risk (or simply, risk) as

$$\mathrm{Risk}(\hat{\beta}; \beta_0) = E[\|\hat{\beta} - \beta_0\|_2^2 \mid \beta_0]. \tag{16}$$

We consider a spherical prior,

$$\beta_0 \sim (0, \frac{r^2}{p} I), \tag{17}$$

for some signal strength $r^2 = E\|\beta_0\|_2^2 > 0$, and define the Bayes risk of an estimator $\hat{\beta}$ as

$$\mathrm{Risk}(\hat{\beta}) = E\|\hat{\beta} - \beta_0\|_2^2. \tag{18}$$

Now, we give the expressions for the risk and Bayes risk of MGF.

Lemma 4. *Under the data model (15), for $t \geq 0$, the risk of the MGF (8) is*

$$\mathrm{Risk}(\hat{\beta}^{\mathrm{mgf}}(t); \beta_0) = \sum_{i=1}^{p} \left[|\beta_0^{\mathrm{T}} v_i|^2 H^2(s_i, t) + \frac{\sigma^2}{n} \frac{(1 - H(s_i, t))^2}{s_i} \right], \tag{19}$$

and under the prior (17), the Bayes risk is

$$\mathrm{Risk}(\hat{\beta}^{\mathrm{mgf}}(t)) = \frac{\sigma^2}{n} \sum_{i=1}^{p} \left[\alpha H^2(s_i, t) + \frac{(1 - H(s_i, t))^2}{s_i} \right], \tag{20}$$

where $\alpha = r^2 n / (\sigma^2 p)$.

Proof. The results follow from the definitions of the risk, the Bayes risk and the bias-variance decomposition-*see Supplement S.4*. □

Remark 2. *Compare (19) to the risk of ridge,*

$$\mathrm{Risk}(\hat{\beta}^{\mathrm{ridge}}(\lambda); \beta_0) = \sum_{i=1}^{p} \left[|\beta_0^{\mathrm{T}} v_i|^2 \frac{\lambda^2}{(s_i + \lambda)^2} + \frac{\sigma^2}{n} \frac{s_i}{(s_i + \lambda)^2} \right], \tag{21}$$

and compare (20) to the Bayes risk of ridge,

$$\mathrm{Risk}(\hat{\beta}^{\mathrm{ridge}}(\lambda)) = \frac{\sigma^2}{n} \sum_{i=1}^{p} \left[\frac{\alpha \lambda^2 + s_i}{(s_i + \lambda)^2} \right], \tag{22}$$

where $\alpha = r^2 n / (\sigma^2 p)$. These ridge results follow from standard calculations, which can be found in many other papers; for completeness, we give the details in Supplement-see S.5.

Prediction Risk

In this section, we analyse the prediction risk. Let

$$x_0 \sim (0, \Sigma) \tag{23}$$

for a positive semidefinite matrix $\Sigma \in \mathbb{R}^{p \times p}$, and assume that x_0 is independent of $y \mid \beta_0$. We define the in-sample prediction risk and the out-of-sample prediction risk as

$$\mathrm{Risk}^{\mathrm{in}}(\hat{\beta}; \beta_0) = \frac{1}{n} E[\|X\hat{\beta} - X\beta_0\|_2^2 \mid \beta_0], \tag{24}$$

$$\mathrm{Risk}^{\mathrm{out}}(\hat{\beta}; \beta_0) = E[(x_0^{\mathrm{T}}\hat{\beta} - x_0^{\mathrm{T}}\beta_0)^2 \mid \beta_0], \tag{25}$$

respectively, and their Bayes versions as $\mathrm{Risk}^{\mathrm{in}}(\hat{\beta}) = \frac{1}{n}E[\|X\hat{\beta} - X\beta_0\|_2^2]$, $\mathrm{Risk}^{\mathrm{out}}(\hat{\beta}; \beta_0) = E[(x_0^{\mathrm{T}}\hat{\beta} - x_0^{\mathrm{T}}\beta_0)^2]$, respectively. Now, we give the expressions for the prediction risk and Bayes prediction risk of MGF.

Lemma 5. *Under (15) and (23), the out-of-sample prediction risk of MGF (8) is*

$$\mathrm{Risk}^{\mathrm{out}}(\hat{\beta}^{\mathrm{mgf}}(t); \beta_0) = \beta_0^{\mathrm{T}} V H(S,t) V^{\mathrm{T}} \Sigma V H(S,t) V^{\mathrm{T}} \beta_0 + \frac{\sigma^2}{n} \mathrm{tr}\left[S^{-1} (I - H(S,t))^2 \Sigma \right], \tag{26}$$

and under (17), the Bayes out-of-sample prediction risk is

$$\mathrm{Risk}^{\mathrm{out}}(\hat{\beta}^{\mathrm{mgf}}(t)) = \frac{\sigma^2}{n} \mathrm{tr}\left[\alpha H^2(S,t) \Sigma + S^{-1} (I - H(S,t))^2 \Sigma \right]. \tag{27}$$

Proof. The results follow from the definitions of the out-of-sample prediction risk, Bayes out-of-sample prediction risk and bias-variance decomposition- *see Supplement S.6*. □

Remark 3. *Similar to (26) and (27), we have the out-of-sample prediction risk and Bayes out-of-sample prediction risk of ridge*

$$\mathrm{Risk}^{\mathrm{out}}(\hat{\beta}^{\mathrm{ridge}}(\lambda); \beta_0) = \lambda^2 \beta_0^{\mathrm{T}} V (S + \lambda I)^{-1} V^{\mathrm{T}} \Sigma V (S + \lambda I)^{-1} V^{\mathrm{T}} \beta_0 + \frac{\sigma^2}{n} \mathrm{tr}[S (S + \lambda I)^{-2} \Sigma], \tag{28}$$

$$\mathrm{Risk}^{\mathrm{out}}(\hat{\beta}^{\mathrm{ridge}}(\lambda)) = \frac{\sigma^2}{n} \mathrm{tr}\left[\lambda^2 \alpha (S + \lambda I)^{-2} \Sigma + S (S + \lambda I)^{-2} \Sigma \right], \tag{29}$$

respectively. More details can be found in Supplement S.7.

Remark 4. *The results of the in-sample prediction risk of MGF can be expressed as*

$$\mathrm{Risk}^{\mathrm{in}}(\hat{\beta}^{\mathrm{mgf}}(t); \beta_0) = \sum_{i=1}^{p} \left[|\beta_0^{\mathrm{T}} v_i|^2 s_i H^2(s_i, t) + \frac{\sigma^2}{n} (1 - H(s_i, t))^2 \right], \tag{30}$$

and the Bayse prediction in-sample risk can be expressed as

$$\mathrm{Risk}^{\mathrm{in}}(\hat{\beta}^{\mathrm{mgf}}(t)) = \frac{\sigma^2}{n} \sum_{i=1}^{p} \left[\alpha s_i H^2(s_i, t) + (1 - H(s_i, t))^2 \right]. \tag{31}$$

Similarly, we can give the ridge results,

$$\mathrm{Risk}^{\mathrm{in}}(\hat{\beta}^{\mathrm{ridge}}(\lambda); \beta_0) = \sum_{i=1}^{p} \left[|\beta_0^{\mathrm{T}} v_i|^2 \frac{\lambda^2 s_i}{(s_i + \lambda)^2} + \frac{\sigma^2}{n} \frac{s_i^2}{(s_i + \lambda)^2} \right], \tag{32}$$

$$\mathrm{Risk}^{\mathrm{in}}(\hat{\beta}^{\mathrm{ridge}}(\lambda)) = \frac{\sigma^2}{n} \sum_{i=1}^{p} \left[\frac{\alpha \lambda^2 s_i + s_i^2}{(s_i + \lambda)^2} \right]. \tag{33}$$

The proof can be found in the Supplement-see S.8.

Relative Estimation Risk and Prediction Risk

In this section, we study the bound on the relative risk of MGF to ridge, under the calibration $\lambda = 2/t^2$. Firstly, we need to introduce two critical inequalities.

Lemma 6. *For $t \geq 0$ and $s_i > 0$, we have* (i) $H(s_i, t) < 1.24\frac{1}{1+s_it^2/2}$; (ii) $1 - H(s_i, t) < 1.04\frac{s_it^2/2}{1+s_it^2/2}$.

Proof. The results follow from the numerically computing-*see Supplement S.9.* □

The following theorem gives the bounds of the relative risk of MGF to ridge.

Theorem 1. *Consider the data model (15). For all $\beta_0 \in \mathbb{R}^p$ and $t \geq 0$, we have*

$$\mathrm{Risk}(\hat{\beta}^{\mathrm{mgf}}(t); \beta_0) < 1.5376 \cdot \mathrm{Risk}(\hat{\beta}^{\mathrm{ridge}}(2/t^2); \beta_0). \quad (34)$$

Moreover, (34) also holds if we replace the risk by the Bayes risk for any prior (17), the in-sample prediction risk, the Bayes in-sample prediction risk for any prior (17), or the Bayes out-of-sample prediction risk for any prior (17) and the feature distribution (23).

Proof. For the risk, set $\lambda = 2/t^2$ and denote the ith summand of (19) and (21) by a_i and b_i, respectively. Then we have

$$\begin{aligned}
a_i &= |v_i^{\mathrm{T}}\beta_0|^2 H^2(s_it) + \frac{\sigma^2}{n}\frac{[1 - H(s_i, t)]^2}{s_i} \\
&< |v_i^{\mathrm{T}}\beta_0|^2 1.24^2 \frac{1}{(1 + s_it^2/2)^2} + \frac{\sigma^2}{n} 1.04^2 \frac{s_i(t^2/2)^2}{(1 + s_it^2/2)^2} \\
&< 1.5376 \cdot \left[|v_i^{\mathrm{T}}\beta_0|^2 \frac{(2/t^2)^2}{(s_i + 2/t^2)^2} + \frac{\sigma^2}{n}\frac{s_i}{(s_i + 2/t^2)^2} \right] \\
&= 1.5376 \cdot b_i.
\end{aligned}$$

The first inequality follows from Lemma 6. Then the bound of the risk follows by summing over $i = 1, ..., p$.

We have the bound of the Bayes risk just by taking expectations on each side of (34).

For the in-sample prediction risk, we can get the bound by multiplying s_i to each summand in (34). By taking expectations for the in-sample prediction risk, we have the bound of the Bayes in-sample prediction risk.

Since $H(S, t)$ and S are diagonal matrices, the two inequalities in Lemma 6 can be extended to matrix operations, i.e. $H^2(S, t) < 1.5376(I + St^2/2)^{-2}$; $(I - H(S, t))^2 < 1.0816(S^2t^4/4)(I + St^2/2)^{-2}$. Note that $\Sigma \succeq 0$ in (27) and (29). Then for the Bayes out-of-sample prediction risk, we have

$$\begin{aligned}
&\alpha H^2(S, t)\Sigma + S^{-1}(I - H(S, t))^2\Sigma \\
&= \left(\alpha H^2(S, t) + S^{-1}(I - H(S, t))^2\right)\Sigma \\
&< 1.5376 \cdot \left[\alpha(2/t^2)^2(2/t^2 I + S)^{-2} + S(2/t^2 I + S)^{-2}\right]\Sigma.
\end{aligned}$$

□

Relative Risks at the Optima

Note that the Bayes risk (22), the Bayes prediction risk (27) and (33) of ridge are minimized at $\lambda^* = 1/\alpha$ (Dicker and Lee 2016). In the special case that the distributions of $y \mid \beta_0$ and the prior β_0 are normal, we know that $\hat{\beta}^{\mathrm{ridge}}(\lambda^*)$ is the Bayes estimator, which achieves the optimal Bayes risk (hence certainly the lowest Bayes risk over the whole ridge family). So the Bayes risk of $\hat{\beta}^{\mathrm{mgf}}(t)$, for $t \geq 0$, must be at least that of $\hat{\beta}^{\mathrm{ridge}}(\lambda^*)$. Applying the fact that $\lambda = 2/t^2$ and $\lambda^* = 1/\alpha$, we can set the optima time $t = \sqrt{2\alpha}$ for the MGF. The following inequality is a key step to obtain the relative Bayes risk and the Bayes prediction risk of MGF to ridge, when both are optimally tuned.

Lemma 7. *For all $s_i > 0$ and $\alpha > 0$, it holds that*

$$\alpha H^2\left(s_i, \sqrt{2\alpha}\right) + \frac{\left[1 - H(s_i, \sqrt{2\alpha})\right]^2}{s_i} < 1.035\frac{1}{\alpha(1 + s_i)}.$$

Proof. The result follows from the numerically computing-*see Supplement S.10.* □

Theorem 2. *Consider the data model (15), the prior (17) and the (out-of-sample) feature distribution (23). It holds that*

$$1 \leq \frac{\inf_{t>0} \mathrm{Risk}(\hat{\beta}^{\mathrm{mgf}}(\mathrm{t}))}{\inf_{\lambda>0} \mathrm{Risk}(\hat{\beta}^{\mathrm{ridge}}(\lambda))} < 1.035. \quad (35)$$

Moreover, (35) also holds if we replace the Bayes risk by the Bayes in-sample prediction risk, or the Bayes out-of-sample prediction risk.

Proof. Note that in the special case of a normal-normal likelihood-prior pair, the minimum of the Bayes risk of $\hat{\beta}^{\mathrm{mgf}}(t)$ is not less than that of $\hat{\beta}^{\mathrm{ridge}}(\lambda^*)$. But the Bayes risks of MGF (20) and ridge (22) do not depend on the likelihood and the prior (only on their first two moments), thus we prove the lower bound must be hold in general. For the upper bound, set $t = \sqrt{2\alpha}$, and denote the ith summand in (20) and in (22) by a_i and b_i, respectively. By Lemma 7, we have

$$\begin{aligned}
a_i &= \alpha H^2\left(s_i, \sqrt{2\alpha}\right) + \left[1 - H(s_i, \sqrt{2\alpha})\right]^2 / s_i \\
&< 1.035\left[1/\alpha(1 + s_i)\right] = 1.035 b_i
\end{aligned}$$

The proof of the Bayes in-sample prediction risk is similar to Theorem 1 and we omit it here.

Since $H(S, \sqrt{2\alpha})$ and S are diagonal matrices, the inequality in Lemma 7 can be extended to matrix operations, i.e. $\alpha H^2\left(S, \sqrt{2\alpha}\right) + S^{-1}\left(I - H(S, \sqrt{2\alpha})\right)^2 \prec 1.035\alpha(I + S)^{-1}$. Then for the Bayes out-of-sample prediction risk, we have

$$\begin{aligned}
&\alpha H^2(S, \sqrt{2\alpha})\Sigma + S^{-1}(I - H(S, \sqrt{2\alpha}))^2\Sigma \\
&= \left[\alpha H^2(S, \sqrt{2\alpha}) + S^{-1}(I - H(S, \sqrt{2\alpha}))^2\right]\Sigma \\
&\prec 1.035\left[\alpha(I + S)^{-1}\right]\Sigma.
\end{aligned}$$

□

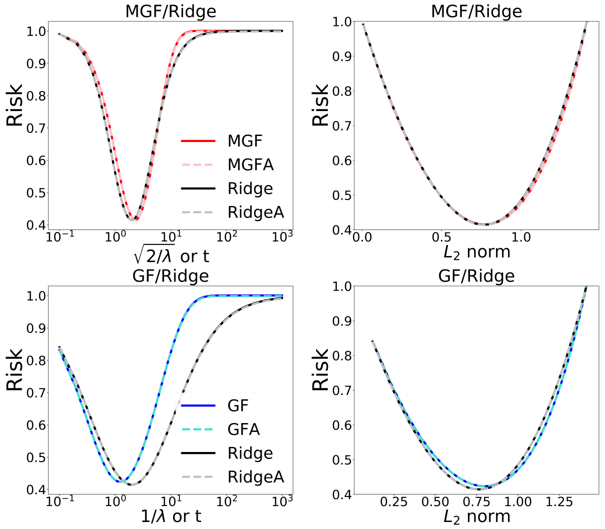

Figure 2: Comparison of the Bayes risks and asymptotic risk for MGF, GF and ridge. MGFA, GFA, and RidgeA stand for the asymptotic risk of MGF, GF and Ridge, respectively. We generate features via $X = \Sigma^{1/2}Z$, for a matrix Z with i.i.d. entries from a standard Gaussian and set $\Sigma = I, n = 1000, p = 500, \sigma^2 = r^2 = 1$ and $D(\mu) = \text{diag}(\mu_i) = \text{diag}(2\sqrt{s_i} + 10^{-3})(i = 1, \cdots, p)$.

In the first column of Figure 2, we plot MGF versus ridge (calibrated according to $\lambda = 2/t^2$) and GF (gradient flow, which is the continuous-time form of GD) versus ridge (calibrated according to $\lambda = 1/t$ (Ali, Kolter, and Tibshirani 2019)) and its asymptotic risk expressions which can be found in *supplement S.11*. It shows that there is a fairly strong agreement between the risk curves, and MGF is much closer to ridge than GF over the entire path; the maximum ratio of the Bayes risk of MGF to ridge is 1.1097 and the maximum ratio of the optima is 1.0208 which are lower than that of GF to ridge, which are 1.3663 and 1.0914, respectively; in addition, it shows that MGF converges to ridge faster than GD, which is compatible with the theoretical results (the tuning parameter λ of ridge is proportional to $\mathcal{O}(1/t^2)$ in MGF and GF requires $\mathcal{O}(1/t)$). The second column shows the remarkable agreement of the risk over the whole path when parameterized by the ℓ_2 norm of the underlying estimator (more details can be found *Supplement S.12*). And MGF is closer to ridge than GD, too. Moreover, the four plots show that the finite-sample and asymptotic risk curves are identical, which implies that the convergence is rapid. The results for other settings (the results are grossly similar) can be found in *Supplement S.12*.

Dynamical Analysis of the Depth-N Model

Is MGD still close to ridge as the model depth increases? If not, what solution would MGD prefer? We will answer this question by a simple but non-trivial depth-N linear neural networks. A natural strategy is to give the closed-form expression of MGF as for $N = 1$. Unfortunately, it is extremely difficult for the depth-N model even if for the linear setting. So, to explicitly characterize how the depth N affects the implicit regularization of MGF, we embed N into the evolution of the components of the weight vector.

The Depth-N Linear Model

We will consider the depth-N diagonal linear neural networks (i.e. where the weight matrices have diagonal structure). The input is $x \in \mathbb{R}^d$, $y \in \mathbb{R}$ and the predictor is linear model with positive weights, such that

$$f(w, x) = \langle w, x \rangle, \qquad w \in \mathbb{R}^d > 0 \tag{36}$$

The diagonal linear networks has d units, with each unit connected to just a single input unit and the output. By parameterizing w using $\beta \in \mathbb{R}^d > 0$, we introduce the depth into our model

$$w = \beta_N \odot \beta_{N-1} \odot \cdots \odot \beta_1, \tag{37}$$

where $\odot$ represents the Hadamard product. (Woodworth et al. 2020) shows that if the input and output weights for each hidden unit are initialized to have the same magnitude, then their magnitudes will remain equal and their signs will not flip throughout training. Based on (37), we equivalently parametrize the model in terms of a single shared input and output weight β for each hidden unit and get the model

$$f(\beta, x) = \langle \beta^N, x \rangle. \tag{38}$$

We will study the dynamics of this model for general N under the squared loss

$$L_N(\beta) = \frac{1}{2N}(\hat{y} - y)^2 = \frac{1}{2N}(\langle \beta^N, x \rangle - y)^2. \tag{39}$$

By running MGF over $L_N(\beta)$ (39), for any i, we can get the dynamics of β_i

$$\ddot{\beta}_i(t) = -\mu\dot{\beta}_i - \beta_i^{N-1}x_i(\hat{y} - y), \tag{40}$$

where $\beta_i \in \mathbb{R}$ and $x_i \in \mathbb{R}$ are the i-th components of β and x, respectively. From the definition of w in (37) and the dynamics of β_i, we can get the dynamics of w_i

$$\begin{aligned} \ddot{w}_i(t) =& \frac{N-1}{N} w_i^{-1}(t)\dot{w}_i^2(t) - \mu\dot{w}_i(t) \\ & - N[w_i(t)]^{\frac{2(N-1)}{N}} x_i(\hat{y} - y). \end{aligned} \tag{41}$$

Implicit Regularization Towards the Sparse Solutions

Let $i, j \in \{1, \cdots, d\}$. By (41), we can give the relationship of the evolution between $w_i(t)$ and $w_j(t)$

$$\begin{aligned} &\left[\ddot{w}_i(t) - \frac{N-1}{N} w_i^{-1}(t)\dot{w}_i^2(t) + \mu\dot{w}_i(t)\right] \cdot \\ &\left(-\frac{1}{N}(w_i(t))^{-\frac{2(N-1)}{N}} \frac{1}{x_i}\right) = \\ &\left[\ddot{w}_j(t) - \frac{N-1}{N} w_j^{-1}(t)\dot{w}_j^2(t) + \mu\dot{w}_j(t)\right] \cdot \\ &\left(-\frac{1}{N}(w_j(t))^{-\frac{2(N-1)}{N}} \frac{1}{x_j}\right). \end{aligned} \tag{42}$$

The following theorem states that $w_i(t)$ can be expressed as a function of $w_j(t)$.

Theorem 3. *For the depth-N diagonal linear model described in (38) and $w_i > 0$, there exist the following relationship between $w_i(t)$ and $w_j(t)$ in the dynamics of MGF (41) over the squared loss (39)*

$$w_i(t) = \begin{cases} \dfrac{x_i}{x_j} w_j(t) + C \cdot \exp(-ut) + C, & N = 1; \\ w_j(t)^{\frac{x_i}{x_j}} \exp\left(H(t) + C \cdot \exp(-\mu t) + C\right), & N = 2; \\ \left[\dfrac{x_i}{x_j} w_j(t)^{-\frac{N-2}{N}} - \dfrac{(N-1)(N-2)}{N^2} G(t) \right. & \\ \left. + C \cdot \exp(-\mu t) + C\right]^{-\frac{N}{N-2}}, & N \geq 3. \end{cases}$$

where $G(t)$ and $H(t)$ are the functions of t, x_i and x_j are the i-th and j-th components of input, respectively. And C is a constant independent of t.

Proof. The proof is complicated and can be found in *Supplement S.13*. □

Remark 5. *If the depth $N = 1$, the $w_i(t)$ will grow linearly with $w_j(t)$, which is consistent with the performance of MGF for the least squares (the component of label is rescaled by a linear factor). If $N \geq 2$, the evolution gap between $w_i(t)$ and $w_j(t)$ will increase rapidly. In details, as the depth increasing the components of weight vector learned at different evolution rates. We note that the larger components of weight vector are subject to an enhancement effect of the increase of the depth, while the smaller ones are subject to attenuation. This tendency will incline to sparse solutions and intensifies with the depth. Furthermore, if we apply the early-stopping for MGD with sufficiently small initialization and step size over the depth-N models, the larger components of weight vector will be learned and the smaller ones fail to be learned so that this implicit sparse regularization effect is more likely to take place.*

Figure 3 present the empirical demonstration of our conclusion from Theorem 3. It shows that for the depth-N diagonal linear neural networks, under MGD with the sufficiently small step size and initialization, the components of weight vector are learned at different evolution rates (i.e. the smaller ones converge slower and the larger ones converge faster), and this evolution gap increases according to the depth. Moreover, we note that the early stopping is crucial for MGD to converge to the sparse model. In the case of $N = 3$, if we perform the early-stopping for MGD when the time threshold is 400, the three largest values ($w_5 - w_3$) have converged, while the two smallest (w_1 and w_2) progress very slowly (close to 0), which will cause the model to be sparse.

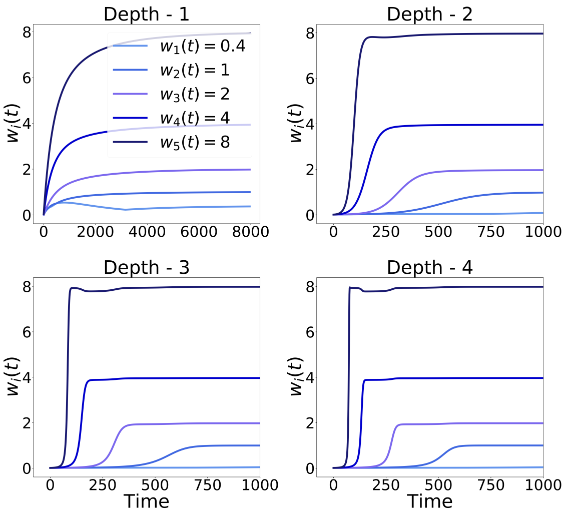

Figure 3: Dynamics of MGD over the deep regression model. Fix the same sufficiently small initialization $w_0 = 0.01$ and step size $\epsilon = 0.001$, we plot the path of w for different $N = 1, 2, 3, 4$. The five curves show the evolution of the five largest values of w separately. For depth $N = 1$, we can see that the evolution rate of the components of weight vector are similar, and converge nearly at the same time threshold. For depth $N \geq 2$, w progress very slowly after initialization (when close to zero); then, the largest one ($w_5 = 8$) converges first, followed by $w_4 = 4$, while the smallest one barely changes even for the large time thresholds. As the depth increases, we can find that w will converge at the smaller time threshold, and the movement of w will be sharper. We note that as the depth increases, the evolution gap of the components of weight vector becomes more significant. This is an implicity regularization towards sparse solutions, which intensifies with the depth.

Conclusion

The present paper studied the implicit regularization of MGF for the depth-N models. For the depth-1 model (the least squares), we characterized the close connections between MGF and ridge. In details, we proved that the risk of MGF is no more than 1.54 times that of ridge under the calibration $t = \sqrt{2/\lambda}$ and the relative Bayes risk of MGF to ridge is between 1 and 1.035 under the optimal tuning. For the depth-N model, we showed that the components of the weight vector are learned at different evolution rates, and this evolution gap increases with depth N. This tendency implies an implicit bias towards sparse solutions. The numerical experiments strongly support our theoretical results.

There are some worthwhile directions for the further work. For example, it would be interesting to explore how hyperparameters (e.g. momentum parameters and learning rate) affect the generalization performance of MGD (or other acceleration optimization algorithms, e.g. Nesterov) for the deep models. It would also be interesting to explain why there is a much tighter coupling of MGF to ridge and under ℓ_2 norms calibration in theory. Moreover, we try to theoretically characterize the impact of the depth, early-stopping and initialization on the implicit regularization of optimization. Furthermore, we carry out the experiments for the matrix factorization which can be found in *Supplement S.14*. The results implied that as the depth increasing, MGD inclines to the low-rank solutions. In our future work, we will try to give the theoretical explanations.

Acknowledgments

This work was supported by NSFC-61872076, Natural Science Foundation of Jilin Province 20200201161JC and National Key R&D Program of China (No. 2022YFA1003701).

References

Ali, A.; Dobriban, E.; and Tibshirani, R. J. 2020. The Implicit Regularization of Stochastic Gradient Flow for Least Squares. In *The 37th International Conference on Machine Learning*.

Ali, A.; Kolter, J. Z.; and Tibshirani, R. J. 2019. A Continuous-Time View of Early Stopping for Least Squares Regression. In *Proceedings of the 22th International Conference on Artificial Intelligence and Statistics*, volume 89, 1370–1378. PMLR.

Arora, S.; Cohen, N.; and Hazan, E. 2018. On the optimization of deep networks: Implicit acceleration by overparameterization. In *International Conference on Machine Learning*, 244–253. PMLR.

Arora, S.; Cohen, N.; Hu, W.; and Luo, Y. 2019. Implicit regularization in deep matrix factorization. *Advances in Neural Information Processing Systems*, 32.

Cyrus, S.; Hu, B.; Scoy, B. V.; and Lessard, L. 2018. A Robust Accelerated Optimization Algorithm for Strongly Convex Functions. *2018 Annual American Control Conference (ACC)*, 1376–1381.

Dicker; and Lee, H. 2016. Ridge regression and asymptotic minimax estimation over spheres of growing dimension. *Bernoulli*, 22(1): 1–37.

Even, M.; Berthier, R.; Bach, F. R.; Flammarion, N.; Gaillard, P.; Hendrikx, H.; Massouli'e, L.; and Taylor, A. B. 2021. A Continuized View on Nesterov Acceleration for Stochastic Gradient Descent and Randomized Gossip. *ArXiv*, abs/2106.07644.

Gidel, G.; Bach, F. R.; and Lacoste-Julien, S. 2019. Implicit Regularization of Discrete Gradient Dynamics in Deep Linear Neural Networks. In *In Advances in Neural Information Processing Systems*.

Gissin, D.; Shalev-Shwartz, S.; and Daniely, A. 2019. The implicit bias of depth: How incremental learning drives generalization. *arXiv preprint arXiv:1909.12051*.

Gunasekar, S.; Woodworth, B. E.; Bhojanapalli, S.; Neyshabur, B.; and Srebro, N. 2018. Implicit Regularization in Matrix Factorization. *2018 Information Theory and Applications Workshop (ITA)*, 1–10.

Li, J.; Nguyen, T.; Hegde, C.; and Wong, K. W. 2021. Implicit sparse regularization: The impact of depth and early stopping. *Advances in Neural Information Processing Systems*, 34: 28298–28309.

Lin, Z.; Li, H.; and Fang, C. 2020. Accelerated Optimization for Machine Learning: First-Order Algorithms. *Accelerated Optimization for Machine Learning*.

Polyak, B. 1964. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5): 1–17.

Sheng, Y.; and Ali, A. 2022. Accelerated Gradient Flow: Risk, Stability, and Implicit Regularization. *ArXiv*, abs/2201.08311.

Smith, S. L.; Dherin, B. R. U.; Barrett, D. G. T.; and De, S. 2021. Stochastic Gradient Descent. *Machine Learning with Neural Networks*.

Soudry, D.; Hoffer, E.; Nacson, M. S.; Gunasekar, S.; and Srebro, N. 2018. The Implicit Bias of Gradient Descent on Separable Data. *Journal of Machine Learning Research*, 19(70): 1–57.

Steinerberger, S. 2021. On the regularization effect of stochastic gradient descent applied to least-squares. *ETNA - Electronic Transactions on Numerical Analysis*.

Suggala, A.; Prasad, A.; and Ravikumar, P. K. 2018. Connecting Optimization and Regularization Paths. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Vardi, G.; and Shamir, O. 2021. Implicit Regularization in ReLU Networks with the Square Loss. *ArXiv*, abs/2012.05156.

Vaskevicius, T.; Kanade, V.; and Rebeschini, P. 2019. Implicit Regularization for Optimal Sparse Recovery. In *In Advances in Neural Information Processing Systems*.

Woodworth, B.; Gunasekar, S.; Lee, J. D.; Moroshko, E.; Savarese, P.; Golan, I.; Soudry, D.; and Srebro, N. 2020. Kernel and rich regimes in overparametrized models. In *Conference on Learning Theory*, 3635–3673. PMLR.

Zhang, C.; Bengio, S.; Hardt, M.; Recht, B.; and Vinyals, O. 2016. Understanding Deep Learning Requires Rethinking Generalization. *arXiv preprint arXiv:1611.03530*.

Zhang, C.; Bengio, S.; Hardt, M.; Recht, B.; and Vinyals, O. 2021. Understanding Deep Learning (Still) Requires Rethinking Generalization. *Commun. ACM*, 64(3): 107–115.

Zou, D.; Wu, J.; Braverman, V.; Gu, Q.; Foster, D. P.; and Kakade, S. M. 2021. The Benefits of Implicit Regularization from SGD in Least Squares Problems. *ArXiv*, abs/2108.04552.