The Implicit Regularization of Stochastic Gradient Flow for Least Squares

Alnur Ali [1] **Edgar Dobriban** [2] **Ryan J. Tibshirani** [3]

Abstract

We study the implicit regularization of mini-batch stochastic gradient descent, when applied to the fundamental problem of least squares regression. We leverage a continuous-time stochastic differential equation having the same moments as stochastic gradient descent, which we call *stochastic gradient flow*. We give a bound on the excess risk of stochastic gradient flow at time t, over ridge regression with tuning parameter $\lambda = 1/t$. The bound may be computed from explicit constants (e.g., the mini-batch size, step size, number of iterations), revealing precisely how these quantities drive the excess risk. Numerical examples show the bound can be small, indicating a tight relationship between the two estimators. We give a similar result relating the coefficients of stochastic gradient flow and ridge. These results hold under no conditions on the data matrix X, and across the entire optimization path (not just at convergence).

1. Introduction

Stochastic gradient descent (SGD) is one of the most widely used optimization algorithms—given the sizes of modern data sets, its scalability and ease-of-implementation means that it is usually preferred to other methods, including gradient descent (Bottou, 1998; 2003; Zhang, 2004; Bousquet & Bottou, 2008; Bottou, 2010; Bottou et al., 2016).

A recent line of work (Nacson et al., 2018; Gunasekar et al., 2018a; Soudry et al., 2018; Suggala et al., 2018; Ali et al., 2018; Poggio et al., 2019; Ji & Telgarsky, 2019) has shown that the iterates generated by gradient descent, when applied to a loss without any explicit regularizer, possess a kind of implicit ℓ_2 regularity. Implicit regularization is useful because it suggests a computational shortcut: the iterates generated by sequential optimization algorithms may serve as cheap approximations to the more expensive solution paths associated with explictly regularized problems. While a lot of the interest in implicit regularization is new, its origins can be traced back at least a couple of decades, with several authors noting the apparent connection between early-stopped gradient descent and ℓ_2 regularization (Strand, 1974; Morgan & Bourlard, 1989; Friedman & Popescu, 2004; Ramsay, 2005; Yao et al., 2007).

Thinking of SGD as a computationally cheap but noisy version of gradient descent, it is natural to ask: do the iterates generated by SGD also possess a kind of ℓ_2 regularity? Of course, the connection here may not be as clear as with gradient descent, since there should be a price to pay for the computational savings.

In this paper, we study the implicit regularization performed by mini-batch stochastic gradient descent with a constant step size, when applied to the fundamental problem of least squares regression. We defer a proper review of related work until later on, but for now mention that constant step sizes are frequently analyzed (Bach & Moulines, 2013; Défossez & Bach, 2014; Dieuleveut et al., 2017a; Jain et al., 2017; Babichev & Bach, 2018), and popular in practice, because of their simplicity. We adopt a continuous-time point-of-view, following Ali et al. (2018), and study a stochastic differential equation that we call *stochastic gradient flow*. A strength of the continuous-time perspective is that it facilitates a direct and precise comparison to ℓ_2 regularization, across the entire optimization path—not just at convergence, as is done in much of the current work on implicit regularization.

Summary of Contributions. A summary of our contributions in this paper is as follows.

- We give a bound on the excess risk of stochastic gradient flow at time t, over ridge regression with tuning parameter $\lambda = 1/t$, for all $t \geq 0$. The bound decomposes into three terms. The first term is the (scaled) variance of ridge. The second and third terms both stem from the variance due to mini-batching, and may be made smaller by, e.g., increasing the mini-batch size and/or decreasing the step size. The second term may be interpreted as the "price of stochasticity": it is nonnegative, but vanishes as time grows. The third term

[*]Equal contribution [1]Stanford University [2]University of Pennsylvania [3]Carnegie Mellon University. Correspondence to: Alnur Ali <alnurali@stanford.edu>, Edgar Dobriban <dobriban@wharton.upenn.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020.

is tied to the limiting optimization error of stochastic gradient flow: it is zero in the overparametrized (interpolating) regime (Bassily et al., 2018), but is positive otherwise, reflecting the fact that stochastic gradient flow with a constant step size fluctuates around the least squares solution as time grows. The bound holds with no conditions on the data matrix X. Numerically, the bound can be small, indicating a tight relationship between the two estimators.

- Using the bound, we show through numerical examples that stochastic gradient flow, when stopped at a time that (optimally) balances its bias and variance, yields a solution attaining risk that is 1.0032 times that of the (optimally-stopped) ridge solution, in less time—indicating that stochastic gradient flow strikes a favorable computational-statistical trade-off.
- We give a similar bound on the distance between the coefficients of stochastic gradient flow at time t, and those of ridge regression with tuning parameter $\lambda = 1/t$, which is also seen to be tight.

Outline. Next, we review related work. Section 2 covers notation, and further motivates the continuous-time approach. In Section 3, we present our bound on the excess risk of stochastic gradient flow over ridge regression. In Section 4, we present a bound relating the coefficients of the two estimators. Section 5 gives numerical examples supporting our theory. In Section 6, we conclude.

Related Work. *Stochastic Gradient Descent.* The statistical and computational properties of SGD have been studied intensely over the years, with work tracing back to Robbins & Monro (1951); Fabian (1968); Ruppert (1988); Kushner & Yin (2003); Polyak & Juditsky (1992); Nemirovski et al. (2009). On the statistical side, a lot of the work has focused on delivering optimal error rates for SGD and its many variants, e.g., with averaging, either asymptotically (Robbins & Monro, 1951; Fabian, 1968; Ruppert, 1988; Kushner & Yin, 2003; Polyak & Juditsky, 1992; Moulines & Bach, 2011; Toulis & Airoldi, 2017; Nemirovski et al., 2009), or in finite samples (Cesa-Bianchi et al., 1996; Zhang, 2004; Ying & Pontil, 2008; Cesa-Bianchi & Lugosi, 2006; Pillaud-Vivien et al., 2018; Jain et al., 2018; Mücke et al., 2019).

Notably, Bach & Moulines (2013); Défossez & Bach (2014); Dieuleveut et al. (2017a); Jain et al. (2017); Babichev & Bach (2018) studied SGD with a constant step size for least squares regression with averaging (obtaining optimal rates, which is not our focus). Good references on inference and computation include Fabian (1968); Ruppert (1988); Polyak & Juditsky (1992); Moulines & Bach (2011); Chen et al. (2016); Toulis & Airoldi (2017) and Recht et al. (2011); Duchi et al. (2015), respectively. Mandt et al. (2015); Duvenaud et al. (2016) interpreted SGD with a constant step size as doing Bayesian inference. Many works have empirically investigated the generalization properties of SGD, mainly in the context of non-convex optimization (Jastrzebski et al., 2017; Kleinberg et al., 2018; Zhang et al., 2018; Jin et al., 2019; Nakkiran et al., 2019; Saxe et al., 2019).

Implicit Regularization. Nearly all of the work in implicit regularization thus far has examined the convergence points of gradient descent, and not the whole path, for specific convex (Nacson et al., 2018; Gunasekar et al., 2018a; Soudry et al., 2018; Vaskevicius et al., 2019) and non-convex (Li et al., 2017; Wilson et al., 2017; Gunasekar et al., 2017; 2018b) problems. Notable exceptions include Rosasco & Villa (2015); Lin et al. (2016); Lin & Rosasco (2017); Neu & Rosasco (2018), who studied averaged SGD with a constant step size for least squares regression, arguing that the various algorithmic parameters (i.e., the step size, mini-batch size, number of iterations, etc.) perform a kind of implicit regularization, by inspecting the corresponding error rates. A few works have investigated implicit regularization outside of optimization (Mahoney & Orecchia, 2011; Mahoney, 2012; Gleich & Mahoney, 2014; Martin & Mahoney, 2018).

Stochastic Differential Equations. Several papers have studied the same stochastic differential equation that we do (Hu et al., 2017; Feng et al., 2017; Li et al., 2019; Feng et al., 2019), but without the focus on implicit regularization and statistical learning. Along these lines, somewhat related work can be found in the literature on Langevin dynamics (Geman & Hwang, 1986; Seung et al., 1992; Neal et al., 2011; Welling & Teh, 2011; Sato & Nakagawa, 2014; Teh et al., 2016; Raginsky et al., 2017; Cheng et al., 2019).

2. Preliminaries

2.1. Least Squares, Stochastic Gradient Descent, and Stochastic Gradient Flow

Consider the usual least squares regression problem,

$$\underset{\beta \in \mathbb{R}^p}{\text{minimize}} \ \frac{1}{2n}\|y - X\beta\|_2^2, \tag{1}$$

where $y \in \mathbb{R}^n$ is the response and $X \in \mathbb{R}^{n \times p}$ is the data matrix. Mini-batch SGD applied to (1) is the iteration

$$\begin{aligned} \beta^{(k)} &= \beta^{(k-1)} + \frac{\epsilon}{m} \cdot \sum_{i \in \mathcal{I}_k} (y_i - x_i^T \beta^{(k-1)}) x_i \\ &= \beta^{(k-1)} + \frac{\epsilon}{m} \cdot X_{\mathcal{I}_k}^T (y_{\mathcal{I}_k} - X_{\mathcal{I}_k} \beta^{(k-1)}), \end{aligned} \tag{2}$$

for $k = 1, 2, 3, \ldots$, where $\epsilon > 0$ is a fixed step size, m is the mini-batch size, and $\mathcal{I}_k \subseteq \{1, \ldots, n\}$ denotes the mini-batch on iteration k with $|\mathcal{I}_k| = m$, for all k. For simplicity, we assume the mini-batches are sampled with replacement; our results hold with minor modifications under sampling without replacement. We assume the initialization $\beta^{(0)} = 0$.

Now, adding and subtracting the negative gradient of the loss in (2) yields

$$\beta^{(k)} = \beta^{(k-1)} + \frac{\epsilon}{n} \cdot X^T(y - X\beta^{(k-1)}) \tag{3}$$
$$+ \epsilon \cdot \left(\frac{1}{m} X_{\mathcal{I}_k}^T (y_{\mathcal{I}_k} - X_{\mathcal{I}_k}\beta^{(k-1)}) - \frac{1}{n} X^T(y - X\beta^{(k-1)}) \right).$$

This may be recognized as gradient descent, plus the deviation between the sample average of m i.i.d. random variables and their mean, which motivates the continuous-time dynamics (stochastic differential equation)

$$d\beta(t) = \frac{1}{n} X^T(y - X\beta(t))\, dt + Q_\epsilon(\beta(t))^{1/2}\, dW(t), \tag{4}$$

with $\beta(0) = 0$. Here, $W(t)$ is standard p-dimensional Brownian motion. We denote the diffusion coefficient

$$Q_\epsilon(\beta) = \epsilon \cdot \mathrm{Cov}_{\mathcal{I}} \left(\frac{1}{m} X_{\mathcal{I}}^T (y_{\mathcal{I}} - X_{\mathcal{I}}\beta) \right), \tag{5}$$

where the randomness is due to $\mathcal{I} \subseteq \{1, \ldots, n\}$. We call the diffusion process (4) *stochastic gradient flow*.

At this point, it helps to recall the related work of Ali et al. (2018), who studied *gradient flow*,

$$\dot{\beta}(t) = \frac{1}{n} X^T(y - X\beta(t))dt, \quad \beta(0) = 0, \tag{6}$$

which is gradient descent for (1) with infinitesimal step sizes. In what follows, we frequently use the solution to (6),

$$\hat{\beta}^{\mathrm{gf}}(t) = (X^TX)^+ \big(I - \exp(-tX^TX/n)\big) X^Ty, \tag{7}$$

where $\exp(A)$ and A^+ denote the matrix exponential and the Moore-Penrose pseudo-inverse of A, respectively.

Unlike gradient flow, the continuous-time flow (4) does not arise by taking limits of the discrete-time dynamics (2), and should instead be interpreted as an approximation to (2). To see this, consider the Euler discretization of (4),

$$\tilde{\beta}^{(k)} = \tilde{\beta}^{(k-1)} + \frac{\epsilon}{n} \cdot X^T(y - X\tilde{\beta}^{(k-1)}) \tag{8}$$
$$+ \epsilon \cdot \mathrm{Cov}_{\mathcal{I}}^{1/2} \left(\frac{1}{m} X_{\mathcal{I}}^T (y_{\mathcal{I}} - X_{\mathcal{I}}\tilde{\beta}^{(k-1)}) \right) z_k,$$

where $z_k \sim N(0, I)$ and $\tilde{\beta}^{(0)} = 0$, i.e., (8) approximates (3) with a Gaussian process. Note that the noise in (8) is on the right scale, which also explains the presence of ϵ in (5).

Figure 1 presents a small numerical example, where we see a striking resemblance between the paths for SGD, the Euler discretization of stochastic gradient flow, and ridge regression with tuning parameter $\lambda = 1/t$.

2.2. Basic Properties of Stochastic Gradient Flow

We begin with an important lemma further motivating the differential equation (4); its proof, as with many of the results in this paper, may be found in the supplement. The result shows that both the first and second moments of the Euler discretization of (4) match those of the underlying discrete-time SGD iteration. This means that any deviation between the first two moments of the continuous-time flow (4) and discrete-time SGD must be due to discretization.

Lemma 1. *Fix y, X, $\epsilon > 0$, and $k \geq 1$. Write $\tilde{\beta}^{(k)}$ for the Euler discretization* (8) *of stochastic gradient flow, and $\beta^{(k)}$ for SGD (both using ϵ). Then, the first and second moments of $\tilde{\beta}^{(k)}$ match those of $\beta^{(k)}$, i.e., we have that both*

- $\mathbb{E}_{\tilde{Z}} \tilde{\beta}^{(k)} = \mathbb{E}_{\mathcal{I}_1, \ldots, \mathcal{I}_k} \beta^{(k)}$, *and*
- $\mathrm{Cov}_{\tilde{Z}} \tilde{\beta}^{(k)} = \mathrm{Cov}_{\mathcal{I}_1, \ldots, \mathcal{I}_k} \beta^{(k)}$.

Here, we let $\tilde{Z}$ denote the randomness inherent to $\tilde{\beta}^{\mathrm{sgf}}(t)$.

Remark 1. The result also implies that both the estimation and out-of-sample risks of $\tilde{\beta}^{(k)}$ match those of $\beta^{(k)}$; we defer a more thorough treatment of this point to Section 3.

Remark 2. Discretization, i.e., showing that (8) and (2) are close in a precise sense, turns out to be non-trivial, and is left to future work.

Next, with the above motivation in mind, we present a lemma establishing that the solution to (4) exists and is unique. The result also gives a more explicit expression for the solution to (4), which plays a key role in many of the results to come.

Lemma 2. *Fix y, X, and $\epsilon > 0$. Let $t \geq 0$. Then*

$$\hat{\beta}^{\mathrm{sgf}}(t) = \hat{\beta}^{\mathrm{gf}}(t) \tag{9}$$
$$+ \exp(-t\hat{\Sigma}) \cdot \int_0^t \exp(\tau\hat{\Sigma}) Q_\epsilon(\hat{\beta}^{\mathrm{sgf}}(\tau))^{1/2} dW(\tau)$$

is the unique solution to the differential equation (4).

Remark 3. The result actually holds for any Lipschitz continuous diffusion coefficient $Q_\epsilon(\beta(t))$, e.g., $Q_\epsilon(\beta(t)) = I$, as well as the time-homogeneous covariance $Q_\epsilon(\beta(t)) = (\epsilon/m) \cdot \hat{\Sigma}$ (Mandt et al., 2017; Wang, 2017; Dieuleveut et al., 2017b; Fan et al., 2018). In the former case, (4) reduces to (rescaled) Langevin dynamics.

2.3. Constant vs. Non-Constant Covariances

The differential equation (4) has been considered previously (Hu et al., 2017; Feng et al., 2017; Li et al., 2019; Feng et al., 2019), but several works (Mandt et al., 2017; Wang, 2017; Dieuleveut et al., 2017b; Fan et al., 2018) have found it convenient to work with the simplification

$$d\beta(t) = \frac{1}{n} X^T(y - X\beta(t))\, dt + \Big(\frac{\epsilon}{m} \cdot \hat{\Sigma} \Big)^{1/2} dW(t), \tag{10}$$

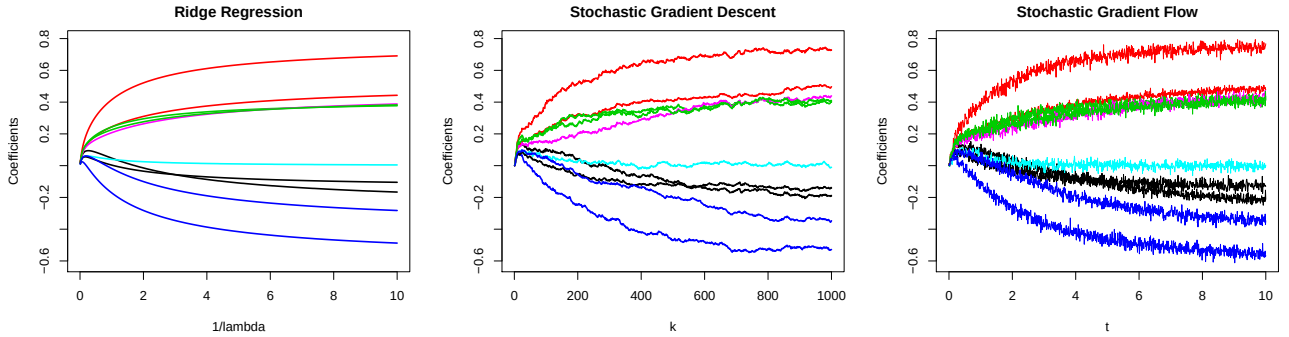

Figure 1. Solution and optimization paths for ridge regression (left panel), SGD (middle panel), and the Euler discretization of stochastic gradient flow (right panel) on a small example, where $n = 50$, $p = 10$, $m = 10$, and $\epsilon = 0.01$.

where $\beta(0) = 0$. Here, $Q_\epsilon(\beta(t)) = (\epsilon/m) \cdot \hat{\Sigma}$. However, we present a simple but telling example revealing that these two processes, i.e., the non-constant covariance process in (4), and the constant covariance process in (10), need not be close in general.

Consider the univariate responseless least squares problem,

$$\underset{\beta \in \mathbb{R}}{\text{minimize}} \ \frac{1}{2n} \sum_{i=1}^{n} (x_i \beta)^2.$$

Let $G_k = (1/m) \sum_{i \in \mathcal{I}_k} x_i^2$, for $k = 1, 2, 3, \ldots$. Then SGD for the above problem may be expressed as

$$\beta^{(k)} = \beta^{(k-1)} - \epsilon \cdot G_k \beta^{(k-1)} = \prod_{j=1}^{k-1} \left(1 - \epsilon \cdot G_j\right) \beta^{(0)}.$$

Assume the initial point is a nonzero constant, the x_i follow a continuous distribution, and ϵ is sufficiently small. Letting $t > 0$ be arbitrary, the basic estimate $1 - x \leq \exp(-x)$ combined with Markov's inequality shows that

$$\Pr(\beta^{(k)} > t) \leq \mathbb{E}\left[\exp\left(-\epsilon \cdot \sum_{j=1}^{k-1} G_j\right)\right] \beta^{(0)}/t.$$

Summing the right-hand side over $k = 1, \ldots, \infty$, we conclude that $\beta^{(k)}$ converges to zero with probability one, by the first Borel-Cantelli lemma.

Now let $G = (1/n) \sum_{i=1}^{n} x_i^2$. We may calculate for the non-constant process that $Q_\epsilon(\beta(t))^{1/2} = \theta \beta(t)$, where $\theta = (\epsilon/m \cdot G)^{1/2}$, meaning the non-constant process follows the dynamics (the sign of $Q_\epsilon^{1/2}$ may be chosen arbitrarily)

$$d\beta(t) = -G\beta(t) + \theta \beta(t) dW(t),$$

which may be recognized as a geometric Brownian motion. It can be checked that both the mean and variance of the geometric Brownian motion tend to zero as time grows, provided that $\theta^2 < 2G$, which certainly holds when $\epsilon < 1$.

On the other hand, the constant process is an Ornstein-Uhlenbeck process,

$$d\beta(t) = -G\beta(t) + \theta dW(t).$$

Again, it may be checked (e.g., Chapter 5 in Øksendal (2003)) that the process mean goes down to zero, whereas the variance tends to the constant $\epsilon/(2m)$. In other words, the limiting dynamics of the constant process exhibit constant-order fluctuations, whereas those of the non-constant process do not. Therefore, for this problem, the latter dynamics more accurately reflect those of discrete-time SGD. See Figure 2 for an example with a standard least squares regression problem.

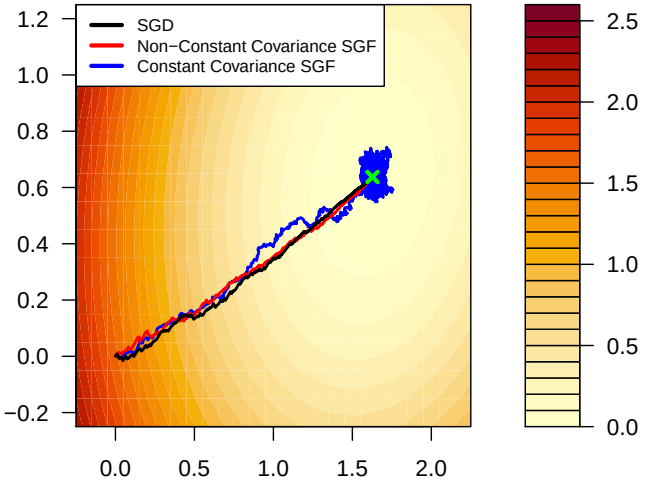

Figure 2. Trajectories for the non-constant and constant covariance processes, as well as discrete-time SGD, on a simple least squares problem, where $n = 3$, $p = 2$, $m = 2$, and $\epsilon = 0.01$. Warmer colors denote larger values of the least squares loss function, and the green X denotes the least squares solution.

We close this section with a simple result bounding the deviation between solutions to the non-constant and constant processes, in expectation. The result indicates that the two

processes can be close when the non-constant process dynamics are close to the underlying coefficients. A thorough comparison of the two processes is left to future work.

Lemma 3. *Fix y, X, and $\epsilon > 0$. Let $t \geq 0$. Write $\tilde{\beta}^{\mathrm{sgf}}(t)$ for the solution to the constant process. Then*

$$\mathbb{E}_{Z,\tilde{Z}} \|\hat{\beta}^{\mathrm{sgf}}(t) - \tilde{\beta}^{\mathrm{sgf}}(t)\|_2^2 \leq 4Lp^3\epsilon/m \times \int_0^t \mathbb{E}_Z\Big[\sum_{i=1}^n \big|(y_i - x_i^T\hat{\beta}^{\mathrm{sgf}}(\tau))^2 - 1\big|\Big] d\tau.$$

Here, we let Z, $\tilde{Z}$ denote the randomness inherent to $\hat{\beta}^{\mathrm{sgf}}(t)$, $\tilde{\beta}^{\mathrm{sgf}}(t)$, respectively, and write $L = \lambda_{\max}(\hat{\Sigma})$.

3. Statistical Risk Bounds

3.1. Measures of Risk and Notation

Here and throughout, we let the predictor matrix X be arbitrary and fixed, and assume the response y follows a standard regression model,

$$y = X\beta_0 + \eta,$$

for some fixed underlying coefficients $\beta_0 \in \mathbb{R}^p$, and noise $\eta \sim (0, \sigma^2 I)$. We consider the statistical (estimation) risk of an estimator $\hat{\beta} \in \mathbb{R}^p$,

$$\mathrm{Risk}(\hat{\beta}; \beta_0) = \mathbb{E}_{\eta,Z}\|\hat{\beta} - \beta_0\|_2^2.$$

Here Z denotes any potential randomness inherent to $\hat{\beta}$ (e.g., due to mini-batching). We also consider in-sample risk,

$$\mathrm{Risk}^{\mathrm{in}}(\hat{\beta}; \beta_0) = \frac{1}{n}\mathbb{E}_{\eta,Z}\|X\hat{\beta} - X\beta_0\|_2^2.$$

We let $\hat{\Sigma} = X^TX/n$ denote the sample covariance matrix with eigenvalues s_i and eigenvectors v_i, for $i = 1, \ldots, p$, and let $\mu = \min_i s_i$ and $L = \max_i s_i$ denote the smallest nonzero and largest eigenvalues of $\hat{\Sigma}$, respectively.

3.2. Risk Bounds

Recall the bias-variance decomposition for risk,

$$\begin{aligned}&\mathrm{Risk}(\hat{\beta}^{\mathrm{sgf}}(t); \beta_0)\\ &= \|\mathbb{E}_{\eta,Z}(\hat{\beta}^{\mathrm{sgf}}(t)) - \beta_0\|_2^2 + \mathrm{tr}\,\mathrm{Cov}_{\eta,Z}(\hat{\beta}^{\mathrm{sgf}}(t))\\ &= \mathrm{Bias}^2(\hat{\beta}^{\mathrm{sgf}}(t); \beta_0) + \mathrm{Var}_{\eta,Z}(\hat{\beta}^{\mathrm{sgf}}(t)).\end{aligned}$$

A straightforward calculation using the law of total variance shows (see the proof of Theorem 2 for details)

$$\mathrm{Bias}^2(\hat{\beta}^{\mathrm{sgf}}(t); \beta_0) = \mathrm{Bias}^2(\hat{\beta}^{\mathrm{gf}}; \beta_0)$$
$$\mathrm{Var}_{\eta,Z}(\hat{\beta}^{\mathrm{sgf}}(t)) = \mathrm{tr}\,\mathbb{E}_\eta\big[\mathrm{Cov}_Z(\hat{\beta}^{\mathrm{sgf}}(t) \,|\, \eta)\big] + \mathrm{Var}_\eta(\hat{\beta}^{\mathrm{gf}}(t)).$$

Therefore, for stochastic gradient flow, the randomness due to mini-batching contributes to the estimation variance. Hence, a tight bound on the variance due to mini-batching, $\mathrm{tr}\,\mathbb{E}_\eta\big[\mathrm{Cov}_Z(\hat{\beta}^{\mathrm{sgf}}(t) \,|\, \eta)\big]$, leads to a tight bound on the risk. The following result, which we see as one of the main technical contributions of this paper, delivers such a bound.

Lemma 4. *Fix y, X, and $\epsilon > 0$. Let $t > 0$. Then*

$$\begin{aligned}&\mathrm{tr}\,\mathrm{Cov}_Z(\hat{\beta}^{\mathrm{sgf}}(t))\\ &\quad\leq \frac{2n\epsilon}{m} \cdot \int_0^t f(\hat{\beta}^{\mathrm{sgf}}(\tau))\mathrm{tr}\,[\hat{\Sigma}\exp(2(\tau - t)\hat{\Sigma})]d\tau,\end{aligned}$$

where $f(\hat{\beta}^{\mathrm{sgf}}(\tau)) = \mathbb{E}_Z\Big[(2n)^{-1}\|y - X\hat{\beta}^{\mathrm{sgf}}(\tau)\|_2^2\Big]$.

Remark 4. The proof of the result depends critically on the special covariance structure of the diffusion coefficient, $Q_\epsilon(\beta(\tau))$, arising in the context of least squares regression. To be more specific, for a fixed β, let $h(\beta) = (y_1 - x_1^T\beta, \ldots, y_n - x_n^T\beta)$ denote the residuals at β, $F(\beta) = \mathrm{diag}(h(\beta))^2$, and $\tilde{F}(\beta) = n^{-1}h(\beta)h(\beta)^T$. Then, another calculation shows (cf. Hoffer et al. (2017); Zhang et al. (2017); Hu et al. (2017))

$$\begin{aligned}Q_\epsilon(\beta) &= \mathrm{Cov}_{\mathcal{I}}\left(\frac{1}{m}X_{\mathcal{I}}^T(y_{\mathcal{I}} - X_{\mathcal{I}}\beta)\right)\\ &= \frac{1}{nm}X^T(F(\beta) - \tilde{F}(\beta))X\\ &\preceq \frac{1}{nm}X^TF(\beta)X,\end{aligned}$$

which may be manipulated to obtain the result given in the lemma (see the supplement for details).

Remark 5. As we discuss later in Section 4, the bound on the variance due to mini-batching given in Lemma 4, turns out to be central: it may also be used to give a tight bound on the coefficient error, $\mathbb{E}_{\eta,Z}\|\hat{\beta}^{\mathrm{sgf}}(t) - \hat{\beta}^{\mathrm{ridge}}(1/t)\|_2^2$.

Inspecting the bound in Lemma 4, we see that the variance due to mini-batching, $\mathrm{tr}\,\mathbb{E}_\eta\big[\mathrm{Cov}_Z(\hat{\beta}^{\mathrm{sgf}}(t)) \,|\, \eta)\big]$, depends on the expected loss of stochastic gradient flow, $f(\hat{\beta}^{\mathrm{sgf}}(t))$. It is reasonable to expect that stochastic gradient flow converges linearly, by analogy to the results that are available for SGD (Karimi et al., 2016; Vaswani et al., 2018; Bassily et al., 2018). The following lemma gives the details.

Lemma 5. *Fix y and X. Let $t \geq 0$.*

- *If $n > p$, define*

$$u = 2\mu - 2n(n\epsilon/m)^2 \cdot \Big(\max_{i=1,\ldots,p}\big[\mathrm{diag}(\hat{\Sigma}^2)\big]_i\Big)$$
$$v = \Big[(n\epsilon/m)^2 q \cdot \max_{i=1,\ldots,p}\big[\mathrm{diag}(\hat{\Sigma}^2)\big]_i\Big]/u$$
$$w = \log\Big(\|y\|_2^2/(2n) - q/(2n)\Big).$$

Here, $q = \|P_{\mathrm{null}(X^T)}y\|_2^2$ denotes the squared norm of the projection of y onto the orthocomplement of the column space of X.

- *If $p \geq n$, define*

$$u = 2\mu - n(n\epsilon/m)^2 \cdot \Big(\max_{i=1,\ldots,p} \big[\mathrm{diag}(\hat{\Sigma}^2)\big]_i \Big)$$
$$v = 0$$
$$w = \log\Big(\|y\|_2^2/(2n) \Big).$$

In either case, set ϵ small enough so that $u > 0$. Then,

$$f(\hat{\beta}^{\mathrm{sgf}}(t)) \leq \exp(-ut + w) + v.$$

Remark 6. Lemma 5 can be seen as the continuous-time analog of, e.g., Theorem 2 in Bassily et al. (2018), and may be of independent interest.

Now define $\tilde{w} = \mathbb{E}_\eta[\exp(w)]$, $\tilde{v} = \mathbb{E}_\eta(v)$, as well as the effective variance due to mini-batching terms,

$$\nu_i(t) = \frac{\exp(w)s_i}{s_i - u/2}\big(\exp(-ut) - \exp(-2ts_i)\big) + v\big(1 - \exp(-2ts_i)\big),\ i = 1, \ldots, p. \tag{11}$$

We recall a result from Ali et al. (2018), paraphrased below.

Theorem 1 (Theorem 1 in Ali et al. (2018))**.** *Fix X. Let $t \geq 0$. Write $\hat{\beta}^{\mathrm{ridge}}(\lambda) = (X^TX + n\lambda I)^{-1}X^Ty$, for the ridge regression estimate with tuning parameter $\lambda \geq 0$. Then, $\mathrm{Bias}^2(\hat{\beta}^{\mathrm{gf}}(t); \beta_0) \leq \mathrm{Bias}^2(\hat{\beta}^{\mathrm{ridge}}(1/t); \beta_0)$, and $\mathrm{Var}(\hat{\beta}^{\mathrm{gf}}(t)) \leq 1.6862 \cdot \mathrm{Var}(\hat{\beta}^{\mathrm{ridge}}(1/t))$, so that $\mathrm{Risk}(\hat{\beta}^{\mathrm{gf}}(t); \beta_0) \leq 1.6862 \cdot \mathrm{Risk}(\hat{\beta}^{\mathrm{ridge}}(1/t); \beta_0)$.*

Putting Lemmas 4 and 5 together with Theorem 1 yields the following result, relating the risk of stochastic gradient flow to that of gradient flow and ridge regression.

Theorem 2. *Fix X. Set ϵ according to Lemma 5. Let $t > 0$.*

- *Then, relative to gradient flow,*

$$\mathrm{Risk}(\hat{\beta}^{\mathrm{sgf}}(t); \beta_0) \leq \mathrm{Bias}^2(\hat{\beta}^{\mathrm{gf}}(t); \beta_0) + \mathrm{Var}_\eta(\hat{\beta}^{\mathrm{gf}}(t)) + \epsilon \cdot \frac{n}{m}\sum_{i=1}^p \mathbb{E}_\eta \nu_i(t). \tag{12}$$

- *Relative to ridge regression,*

$$\mathrm{Risk}(\hat{\beta}^{\mathrm{sgf}}(t); \beta_0) \leq \mathrm{Bias}^2(\hat{\beta}^{\mathrm{ridge}}(1/t); \beta_0) + 1.6862 \cdot \mathrm{Var}_\eta(\hat{\beta}^{\mathrm{ridge}}(1/t)) + \epsilon \cdot \frac{n}{m}\sum_{i=1}^p \mathbb{E}_\eta \nu_i(t). \tag{13}$$

The analogous results for in-sample risk are similar, and deferred to the supplement for space reasons.

Proof. From Lemma 2, we have

$$\hat{\beta}^{\mathrm{sgf}}(t) = \hat{\beta}^{\mathrm{gf}}(t) + \int_0^t \exp\big((\tau - t)\hat{\Sigma}\big) Q_\epsilon(\beta(\tau))^{1/2} dW(\tau).$$

The law of total expectation coupled with standard properties of Brownian motion (e.g., Theorem 3.2.1 in Øksendal (2003)) implies $\mathbb{E}_{\eta,Z}(\hat{\beta}^{\mathrm{sgf}}(t)) = \mathbb{E}_\eta\big[\mathbb{E}_Z(\hat{\beta}^{\mathrm{sgf}}(t) \,|\, \eta)\big] = \mathbb{E}_\eta(\hat{\beta}^{\mathrm{gf}}(t))$. Therefore,

$$\mathrm{Bias}^2(\hat{\beta}^{\mathrm{sgf}}(t); \beta_0) = \mathrm{Bias}^2(\hat{\beta}^{\mathrm{gf}}(t); \beta_0). \tag{14}$$

Turning to the variance, the law of total variance and the above calculation implies

$$\begin{aligned}
&\mathrm{tr}\,\mathrm{Cov}_{\eta,Z}(\hat{\beta}^{\mathrm{sgf}}(t)) \\
&= \mathrm{tr}\Big(\mathbb{E}_\eta\big[\mathrm{Cov}_Z(\hat{\beta}^{\mathrm{sgf}}(t) \,|\, \eta)\big] + \mathrm{Cov}_\eta\big(\mathbb{E}_Z(\hat{\beta}^{\mathrm{sgf}}(t) \,|\, \eta)\big)\Big) \\
&= \mathrm{tr}\Big(\mathbb{E}_\eta\big[\mathrm{Cov}_Z(\hat{\beta}^{\mathrm{sgf}}(t) \,|\, \eta)\big] + \mathrm{Cov}_\eta(\hat{\beta}^{\mathrm{gf}}(t))\Big) \\
&= \mathrm{tr}\,\mathbb{E}_\eta\big[\mathrm{Cov}_Z(\hat{\beta}^{\mathrm{sgf}}(t) \,|\, \eta)\big] + \mathrm{Var}_\eta(\hat{\beta}^{\mathrm{gf}}(t)).
\end{aligned} \tag{15}$$

As for the trace appearing in (15), we have

$$\begin{aligned}
&\mathrm{tr}\,\mathbb{E}_\eta\big[\mathrm{Cov}_Z(\hat{\beta}^{\mathrm{sgf}}(t) \,|\, \eta)\big] = \mathbb{E}_\eta\big[\mathrm{tr}\,\mathrm{Cov}_Z(\hat{\beta}^{\mathrm{sgf}}(t) \,|\, \eta)\big] \\
&\leq \mathbb{E}_\eta\left[\frac{2n\epsilon}{m} \cdot \int_0^t f(\hat{\beta}^{\mathrm{sgf}}(\tau))\mathrm{tr}\,[\hat{\Sigma}\exp(2(\tau - t)\hat{\Sigma})]d\tau\right] \\
&= \frac{2n\epsilon}{m} \cdot \int_0^t \mathbb{E}_\eta[f(\hat{\beta}^{\mathrm{sgf}}(\tau))]\mathrm{tr}\,[\hat{\Sigma}\exp(2(\tau - t)\hat{\Sigma})]d\tau
\end{aligned} \tag{16}$$

$$\leq \epsilon \cdot \frac{n}{m}\sum_{i=1}^p \Big(\tilde{v}\big(1 - \exp(-2ts_i)\big) + \frac{\tilde{w}s_i}{s_i - u/2}\big(\exp(-ut) - \exp(-2ts_i)\big)\Big). \tag{17}$$

Here, the second line followed from Lemma 4. The third followed from Fubini's theorem. The fourth followed by integrating, using the eigendecomposition $\hat{\Sigma} = VSV^T$ and Lemma 5, along with one final application of Fubini's theorem. This shows the claim for gradient flow. The claim for ridge follows by applying Theorem 1. □

Remark 7. Putting (13) together with Theorem 3 in Ali et al. (2018) gives a lower bound under oracle tuning,

$$\inf_{\lambda \geq 0} \mathrm{Risk}(\hat{\beta}^{\mathrm{ridge}}(\lambda)) \leq \inf_{t \geq 0} \mathrm{Risk}(\hat{\beta}^{\mathrm{sgf}}(t)).$$

Now, subtracting the ridge risk from both sides of (13) immediately gives our main result, a bound on the excess risk of stochastic gradient flow over ridge.

Corollary 1. *Fix X. Set ϵ as in Lemma 5. Let $t > 0$. Then,*

$$\mathrm{Risk}(\hat{\beta}^{\mathrm{sgf}}(t); \beta_0) - \mathrm{Risk}(\hat{\beta}^{\mathrm{ridge}}(1/t); \beta_0) \leq 0.6862 \cdot \mathrm{Var}_\eta(\hat{\beta}^{\mathrm{ridge}}(1/t)) + \epsilon \cdot \frac{n}{m}\sum_{i=1}^p \mathbb{E}_\eta \nu_i(t). \tag{18}$$

Remark 8. For space reasons, we compare the excess risk bound (18) to the analogous bound for the time-homogeneous process (10) in the supplement.

We can understand the influence of the effective variance terms on the risks (12), (13), (18) as follows. As stochastic gradient flow moves away from initialization, the stochastic gradients become smaller, and so their variance decreases, which is captured by the first term in (11), as it goes down with time. As stochastic gradient flow approaches the least squares solution, there are two possibilities, depending on whether the solution is interpolating or not. If the solution is interpolating, then stochastic gradient flow can fit the data perfectly, and hence $v = 0$ in (11). Otherwise, stochastic gradient flow fluctuates around the solution, which is captured by the second term in (11), as it grows with time.

It is also interesting to note that the bounds (12), (13), (18) depend linearly on ϵ/m, corroborating recent empirical work (Krizhevsky, 2014; Goyal et al., 2017; Smith et al., 2017; You et al., 2017; Shallue et al., 2019).

Finally, reflecting on (18), (11), we see that the first (variance) term in (11) goes down with time, as we would expect from the bias. Interestingly, the next result shows that the risk of stochastic gradient flow may be seen as the ridge bias raised to a power strictly less than 1, plus a time-dependent scaling of the ridge variance—which is quite different from the situation with gradient flow (cf. Theorem 1).

Lemma 6. *Fix X. Set ϵ as in Lemma 5. Let $t > 0$. Define*

$$\alpha = p\tilde{w}\epsilon \cdot \frac{n\mu}{m(\mu - u/2)},$$
$$\gamma(t) = 1 + 2.164\epsilon \cdot \frac{\tilde{v}n^2 \max(1/t, L)}{m\sigma^2},$$

$\kappa = L/\mu$, and $\delta = \alpha/\|\beta_0\|_2{}^{1/\kappa}$.

- *Then, for gradient flow,*

$$\mathrm{Risk}(\hat{\beta}^{\mathrm{sgf}}(t);\beta_0) \le \mathrm{Bias}^2(\hat{\beta}^{\mathrm{gf}}(t);\beta_0) + \delta \cdot |\mathrm{Bias}(\hat{\beta}^{\mathrm{gf}}(t);\beta_0)|^{1/\kappa} + \gamma(t) \cdot \mathrm{Var}(\hat{\beta}^{\mathrm{gf}}(t)).$$

- *For ridge regression,*

$$\mathrm{Risk}(\hat{\beta}^{\mathrm{sgf}}(t);\beta_0) \le \mathrm{Bias}^2(\hat{\beta}^{\mathrm{ridge}}(1/t);\beta_0) + \delta \cdot |\mathrm{Bias}(\hat{\beta}^{\mathrm{ridge}}(1/t);\beta_0)|^{1/\kappa} + 1.6862\gamma(t) \cdot \mathrm{Var}(\hat{\beta}^{\mathrm{ridge}}(1/t)).$$

4. Coefficient Bounds

The coefficients of stochastic gradient flow and ridge regression may be close, even though the risks are not. Therefore, here, we pursue bounds on the coefficient error, $\mathbb{E}_{\eta,Z}\|\hat{\beta}^{\mathrm{sgf}}(t) - \hat{\beta}^{\mathrm{ridge}}(1/t)\|_2^2$. We start by giving a tight bound on the distance between the coefficents of gradient flow and ridge regression.

Lemma 7. *Fix X. Let $t \ge 0$. Define*

$$g(t) = \begin{cases} \frac{(1-\exp(-Lt))(1+Lt)}{Lt}, & t \le \frac{1.7933}{L} \\ \frac{(1-\exp(-\mu t))(1+\mu t)}{\mu t}, & t \ge \frac{1.7933}{\mu} \\ 1.2985, & \frac{1.7933}{L} < t < \frac{1.7933}{\mu} \end{cases}.$$

Then,

$$\mathbb{E}_\eta\|\hat{\beta}^{\mathrm{gf}}(t) - \hat{\beta}^{\mathrm{ridge}}(1/t)\|_2^2 \le (g(t)-1)^2 \cdot \mathbb{E}_\eta\|\hat{\beta}^{\mathrm{ridge}}(1/t)\|_2^2.$$

Figure 3 plots the function $g(t)$, defined in the lemma. We see that $g(t)$ has a maximum of 1.2985, and tends to 1 as either $t \to 0$ or $t \to \infty$. The behavior makes sense, as both $\hat{\beta}^{\mathrm{gf}}(t)$ and $\hat{\beta}^{\mathrm{ridge}}(1/t)$ tend to the null model as $t \to 0$, and the min-norm solution as $t \to \infty$.

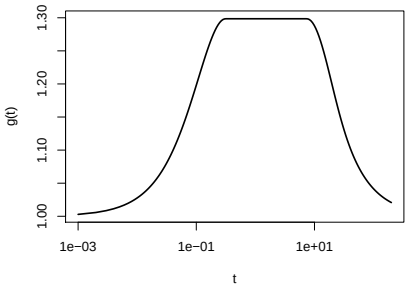

Figure 3. The function $g(t)$, defined in Lemma 7.

Our main result now follows easily, by putting Lemma 7 together with Lemma 4 from Section 3.

Theorem 3. *Fix X. Set ϵ as in Lemma 5. Let $t > 0$. Then,*

$$\mathbb{E}_{\eta,Z}\|\hat{\beta}^{\mathrm{sgf}}(t) - \hat{\beta}^{\mathrm{ridge}}(1/t)\|_2^2 \le (g(t)-1)^2 \cdot \mathbb{E}_\eta\|\hat{\beta}^{\mathrm{ridge}}(1/t)\|_2^2 + \epsilon \cdot \frac{n}{m}\sum_{i=1}^{p}\nu_i(t).$$

Proof. Expanding $\mathbb{E}_{\eta,Z}\|\hat{\beta}^{\mathrm{sgf}}(t) - \hat{\beta}^{\mathrm{ridge}}(1/t)\|_2^2$, adding and subtracting $\|\mathbb{E}_{\eta,Z}(\hat{\beta}^{\mathrm{sgf}}(t))\|_2^2$, and rearranging yields

$$\mathbb{E}_{\eta,Z}\|\hat{\beta}^{\mathrm{sgf}}(t) - \hat{\beta}^{\mathrm{ridge}}(1/t)\|_2^2 = \mathbb{E}_\eta\|\mathbb{E}_{\eta,Z}(\hat{\beta}^{\mathrm{sgf}}(t)) - \hat{\beta}^{\mathrm{ridge}}(1/t)\|_2^2 + \mathrm{tr}\,\mathbb{E}_\eta\big[\mathrm{Cov}_Z(\hat{\beta}^{\mathrm{sgf}}(t))\big].$$

As $Q_\epsilon(\beta(t))^{1/2}$ is continuous, it follows from standard properties of Brownian motion (e.g., Theorem 3.2.1 in Øksendal (2003)) that $\mathbb{E}_Z(\hat{\beta}^{\mathrm{sgf}}(t)) = \hat{\beta}^{\mathrm{gf}}(t)$. Therefore, we have

$$\mathbb{E}_{\eta,Z}\|\hat{\beta}^{\mathrm{sgf}}(t) - \hat{\beta}^{\mathrm{ridge}}(\lambda)\|_2^2 = \mathbb{E}_\eta\|\hat{\beta}^{\mathrm{gf}}(t) - \hat{\beta}^{\mathrm{ridge}}(1/t)\|_2^2 + \mathrm{tr}\,\mathbb{E}_\eta\big[\mathrm{Cov}_Z(\hat{\beta}^{\mathrm{sgf}}(t))\big].$$

Lemma 7 gives a bound on the first term in the preceding display. Lemma 4 and the same arguments used in the proof of Theorem 2 give a bound on the second term. Putting the pieces together yields the result. □

Remark 9. A bound on the coefficient error, $\mathbb{E}_{\eta,Z}\|\hat{\beta}^{\mathrm{sgf}}(t) - \hat{\beta}^{\mathrm{ridge}}(1/t)\|_2^2$, is in some sense fundamental, since the risks are close when the coefficients are. Nonetheless, obtaining risk bounds directly (as was done in Section 3) is still interesting, as these can be sharper.

5. Numerical Examples

We give numerical examples supporting our theoretical findings. We generated the data matrix according to $X = \Sigma^{1/2}W$, where the entries of W were i.i.d. following a normal distribution. We allow for correlations between the features, setting the diagonal entries of the predictor covariance Σ to 1, and the off-diagonals to 0.5. Below, we present results for $n = 100$, $p = 500$, and $m = 20$. The supplement gives additional examples with different problem sizes and data models (Student-t and Bernoulli data); the results are similar. We set $\epsilon = 2.2548\text{e-}4$, following Lemma 5.

Figure 4 plots the risk of ridge regression, discrete-time SGD (2), and Theorem 2. For ridge, we used a range of 200 tuning parameters λ, equally spaced on a log scale from 2^{-15} to 2^{15}. The expression for the risk of ridge is well-known. For Theorem 2, we set $t = 1/\lambda$. For SGD, we computed its effective time, using $t = k\epsilon$ and $t = 1/\lambda$. As for its risk, following the decomposition given in Section 3, we first computed the bias and variance of discrete-time gradient descent, using Lemma 3 in Ali et al. (2018), and then added in the variance given by Lemma 1. As a comparison, Figure 4 also plots the risks of gradient flow (7), coming from Lemma 5 in Ali et al. (2018), and discrete-time gradient descent (as was just discussed).

Though the risks look similar, there are subtle differences (the supplement gives examples with larger step sizes and smaller mini-batch sizes, where the differences are more pronounced). We also see that Theorem 2 tracks the risk of SGD closely. In fact, the maximum ratio, across the entire path, of the risk of stochastic gradient flow to that of ridge is 2.5614, whereas the same ratio for SGD to ridge is 1.7214. Figure 4 also shows the (optimal) time where each method balances its bias and variance. Choosing a tuning parameter by balancing bias and variance is common in nonparametric regression, and doing so here implies that stochastic gradient flow stops earlier than gradient flow, because the effective variance terms (11) are nonnegative. We find the optimal stopping times chosen by balancing bias and variance vs. directly minimizing risk are generally similar. Moreover, the ratio of the (optimal) risks at these times is 1.0032, indicating that stochastic gradient flow strikes a favorable computational-statistical trade-off.

Turning to Theorem 3, we consider the same experimental setup as before, now plotting the bound of Theorem 3, and the actual coefficient error $\mathbb{E}_{\eta,Z}\|\hat{\beta}^{\mathrm{sgf}}(t) - \hat{\beta}^{\mathrm{ridge}}(1/t)\|_2^2$, averaged over 30 draws of y (the underlying coefficients were drawn from a normal distribution, and scaled so the signal-to-noise ratio was roughly 1). We see the bound tracks the underlying error closely, and is quite small—indicating a tight relationship between stochastic gradient flow and ridge. For larger t, some looseness in the bound is evident, arising from the constants appearing in Lemma 5; giving sharper constants is an important problem for future work.

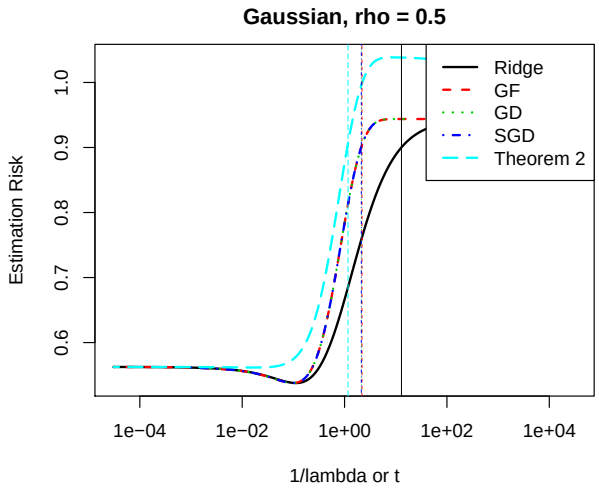

Figure 4. Risks for ridge, SGD, stochastic gradient flow, and gradient descent/flow. The excess risk of stochastic gradient flow over ridge is the distance between the cyan and black curves. The vertical lines show the stopping times that balance bias and variance.

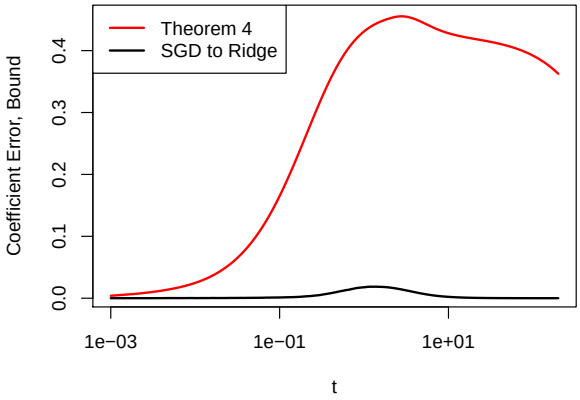

Figure 5. Comparison between Theorem 3, and the actual distance between the coefficients of SGD and ridge.

6. Discussion

We studied the implicit regularization of stochastic gradient flow, giving theoretical and empirical support for the claim that the method is closely related to ℓ_2 regularization. There are a number of important directions for future work, e.g., establishing that stochastic gradient flow and SGD are in

fact close, in a precise sense; considering general convex losses; and analyzing adaptive stochastic gradient methods.

Acknowledgements

We thank a number of people for helpful discussions, including Misha Belkin, Quanquan Gu, J. Zico Kolter, Jason Lee, Yi-An Ma, Jascha Sohl-Dickstein, Daniel Soudry, and Matus Telgarsky. We thank Dushyant Sahoo for his careful proof-reading of the paper. ED was supported in part by NSF BIGDATA grant IIS 1837992 and NSF TRIPODS award 1934960. Part of this work was completed while ED was visiting the Simons Institute.

References

Ali, A., Kolter, J. Z., and Tibshirani, R. J. A continuous-time view of early stopping for least squares regression. *arXiv preprint arXiv:1810.10082*, 2018.

Babichev, D. and Bach, F. Constant step size stochastic gradient descent for probabilistic modeling. *arXiv preprint arXiv:1804.05567*, 2018.

Bach, F. and Moulines, E. Non-strongly-convex smooth stochastic approximation with convergence rate o (1/n). In *Advances in neural information processing systems*, pp. 773–781, 2013.

Bassily, R., Belkin, M., and Ma, S. On exponential convergence of sgd in non-convex over-parametrized learning. *arXiv preprint arXiv:1811.02564*, 2018.

Bottou, L. Online learning and stochastic approximations. *On-line learning in neural networks*, 17(9):142, 1998.

Bottou, L. Stochastic learning. In *Summer School on Machine Learning*, pp. 146–168. Springer, 2003.

Bottou, L. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pp. 177–186. Springer, 2010.

Bottou, L., Curtis, F. E., and Nocedal, J. Optimization methods for large-scale machine learning. *arXiv preprint arXiv:1606.04838*, 2016.

Bousquet, O. and Bottou, L. The tradeoffs of large scale learning. In *Advances in neural information processing systems*, pp. 161–168, 2008.

Cesa-Bianchi, N. and Lugosi, G. *Prediction, learning, and games*. Cambridge university press, 2006.

Cesa-Bianchi, N., Long, P. M., and Warmuth, M. K. Worst-case quadratic loss bounds for prediction using linear functions and gradient descent. *IEEE Transactions on Neural Networks*, 7(3):604–619, 1996.

Chen, X., Lee, J. D., Tong, X. T., and Zhang, Y. Statistical inference for model parameters in stochastic gradient descent. *arXiv preprint arXiv:1610.08637, to appear in the Annals of Statistics*, 2016.

Cheng, X., Bartlett, P. L., and Jordan, M. I. Quantitative w_1 convergence of langevin-like stochastic processes with non-convex potential state-dependent noise. *arXiv preprint arXiv:1907.03215*, 2019.

Défossez, A. and Bach, F. Constant step size least-mean-square: Bias-variance trade-offs and optimal sampling distributions. *arXiv preprint arXiv:1412.0156*, 2014.

Dieuleveut, A., Durmus, A., and Bach, F. Bridging the gap between constant step size stochastic gradient descent and markov chains. *arXiv preprint arXiv:1707.06386*, 2017a.

Dieuleveut, A., Flammarion, N., and Bach, F. Harder, better, faster, stronger convergence rates for least-squares regression. *The Journal of Machine Learning Research*, 18(1): 3520–3570, 2017b.

Duchi, J. C., Chaturapruek, S., and Ré, C. Asynchronous stochastic convex optimization. *arXiv preprint arXiv:1508.00882*, 2015.

Duvenaud, D., Maclaurin, D., and Adams, R. Early stopping as nonparametric variational inference. In *Artificial Intelligence and Statistics*, pp. 1070–1077, 2016.

Fabian, V. On asymptotic normality in stochastic approximation. *The Annals of Mathematical Statistics*, 39(4): 1327–1332, 1968.

Fan, J., Gong, W., Li, C. J., and Sun, Q. Statistical sparse online regression: A diffusion approximation perspective. In *International Conference on Artificial Intelligence and Statistics*, pp. 1017–1026, 2018.

Feng, Y., Li, L., and Liu, J.-G. Semi-groups of stochastic gradient descent and online principal component analysis: properties and diffusion approximations. *arXiv preprint arXiv:1712.06509*, 2017.

Feng, Y., Gao, T., Li, L., Liu, J.-G., and Lu, Y. Uniform-in-time weak error analysis for stochastic gradient descent algorithms via diffusion approximation. *arXiv preprint arXiv:1902.00635*, 2019.

Friedman, J. and Popescu, B. Gradient directed regularization. Working paper, 2004.

Geman, S. and Hwang, C.-R. Diffusions for global optimization. *SIAM Journal on Control and Optimization*, 24 (5):1031–1043, 1986.

Gleich, D. and Mahoney, M. Anti-differentiating approximation algorithms: A case study with min-cuts, spectral, and flow. In *International Conference on Machine Learning*, pp. 1018–1025, 2014.

Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., and He, K. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

Gunasekar, S., Woodworth, B. E., Bhojanapalli, S., Neyshabur, B., and Srebro, N. Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems*, pp. 6151–6159, 2017.

Gunasekar, S., Lee, J., Soudry, D., and Srebro, N. Characterizing implicit bias in terms of optimization geometry. *arXiv preprint arXiv:1802.08246*, 2018a.

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

Hoffer, E., Hubara, I., and Soudry, D. Train longer, generalize better: closing the generalization gap in large batch training of neural networks. In *Advances in Neural Information Processing Systems*, pp. 1731–1741, 2017.

Hu, W., Li, C. J., Li, L., and Liu, J.-G. On the diffusion approximation of nonconvex stochastic gradient descent. *arXiv preprint arXiv:1705.07562*, 2017.

Jain, P., Kakade, S. M., Kidambi, R., Netrapalli, P., Pillutla, V. K., and Sidford, A. A markov chain theory approach to characterizing the minimax optimality of stochastic gradient descent (for least squares). *arXiv preprint arXiv:1710.09430*, 2017.

Jain, P., Kakade, S. M., Kidambi, R., Netrapalli, P., and Sidford, A. Parallelizing stochastic gradient descent for least squares regression: mini-batching, averaging, and model misspecification. *Journal of Machine Learning Research*, 18(223):1–42, 2018.

Jastrzebski, S., Kenton, Z., Arpit, D., Ballas, N., Fischer, A., Bengio, Y., and Storkey, A. Three factors influencing minima in sgd. *arXiv preprint arXiv:1711.04623*, 2017.

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

Jin, C., Netrapalli, P., Ge, R., Kakade, S. M., and Jordan, M. I. Stochastic gradient descent escapes saddle points efficiently. *arXiv preprint arXiv:1902.04811*, 2019.

Karimi, H., Nutini, J., and Schmidt, M. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 795–811. Springer, 2016.

Kleinberg, R., Li, Y., and Yuan, Y. An alternative view: When does sgd escape local minima? *arXiv preprint arXiv:1802.06175*, 2018.

Krizhevsky, A. One weird trick for parallelizing convolutional neural networks. *arXiv preprint arXiv:1404.5997*, 2014.

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

Li, Q., Tai, C., and Weinan, E. Stochastic modified equations and dynamics of stochastic gradient algorithms i: Mathematical foundations. *Journal of Machine Learning Research*, 20(40):1–40, 2019.

Li, Y., Ma, T., and Zhang, H. Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. *arXiv preprint arXiv:1712.09203*, 2017.

Lin, J. and Rosasco, L. Optimal rates for multi-pass stochastic gradient methods. *The Journal of Machine Learning Research*, 18(1):3375–3421, 2017.

Lin, J., Camoriano, R., and Rosasco, L. Generalization properties and implicit regularization for multiple passes sgm. In *International Conference on Machine Learning*, pp. 2340–2348, 2016.

Mahoney, M. W. Approximate computation and implicit regularization for very large-scale data analysis. In *Proceedings of the 31st ACM SIGMOD-SIGACT-SIGAI symposium on Principles of Database Systems*, pp. 143–154. ACM, 2012.

Mahoney, M. W. and Orecchia, L. Implementing regularization implicitly via approximate eigenvector computation. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pp. 121–128. Omnipress, 2011.

Mandt, S., Hoffman, M. D., and Blei, D. M. Continuous-time limit of stochastic gradient descent revisited. *NIPS-2015*, 2015.

Mandt, S., Hoffman, M. D., and Blei, D. M. Stochastic gradient descent as approximate bayesian inference. *The Journal of Machine Learning Research*, 18(1):4873–4907, 2017.

Martin, C. H. and Mahoney, M. W. Implicit self-regularization in deep neural networks: Evidence from random matrix theory and implications for learning. *arXiv preprint arXiv:1810.01075*, 2018.

Morgan, N. and Bourlard, H. Generalization and parameter estimation in feedforward nets: Some experiments. In *Advances in Neural Information Processing Systems*, 1989.

Moulines, E. and Bach, F. R. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems*, pp. 451–459, 2011.

Mücke, N., Neu, G., and Rosasco, L. Beating sgd saturation with tail-averaging and minibatching. In *Advances in Neural Information Processing Systems*, pp. 12568–12577, 2019.

Nacson, M. S., Srebro, N., and Soudry, D. Stochastic gradient descent on separable data: Exact convergence with a fixed learning rate. *arXiv preprint arXiv:1806.01796*, 2018.

Nakkiran, P., Kaplun, G., Kalimeris, D., Yang, T., Edelman, B. L., Zhang, F., and Barak, B. Sgd on neural networks learns functions of increasing complexity. *arXiv preprint arXiv:1905.11604*, 2019.

Neal, R. M. et al. Mcmc using hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2(11):2, 2011.

Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

Neu, G. and Rosasco, L. Iterate averaging as regularization for stochastic gradient descent. *arXiv preprint arXiv:1802.08009*, 2018.

Øksendal, B. *Stochastic differential equations*. Springer, 2003.

Pillaud-Vivien, L., Rudi, A., and Bach, F. Statistical optimality of stochastic gradient descent on hard learning problems through multiple passes. In *Advances in Neural Information Processing Systems*, pp. 8114–8124, 2018.

Poggio, T., Banburski, A., and Liao, Q. Theoretical issues in deep networks: Approximation, optimization and generalization. *arXiv preprint arXiv:1908.09375*, 2019.

Polyak, B. T. and Juditsky, A. B. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855, 1992.

Raginsky, M., Rakhlin, A., and Telgarsky, M. Nonconvex learning via stochastic gradient langevin dynamics: a nonasymptotic analysis. *arXiv preprint arXiv:1702.03849*, 2017.

Ramsay, J. Parameter flows. Working paper, 2005.

Recht, B., Re, C., Wright, S., and Niu, F. Hogwild: A lock-free approach to parallelizing stochastic gradient descent. In *Advances in neural information processing systems*, pp. 693–701, 2011.

Robbins, H. and Monro, S. A stochastic approximation method. *The annals of mathematical statistics*, pp. 400–407, 1951.

Rosasco, L. and Villa, S. Learning with incremental iterative regularization. In *Advances in Neural Information Processing Systems*, pp. 1630–1638, 2015.

Ruppert, D. Efficient estimations from a slowly convergent robbins-monro process. Technical report, Cornell University Operations Research and Industrial Engineering, 1988.

Sato, I. and Nakagawa, H. Approximation analysis of stochastic gradient langevin dynamics by using fokker-planck equation and ito process. In *International Conference on Machine Learning*, pp. 982–990, 2014.

Saxe, A. M., McClelland, J. L., and Ganguli, S. A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences*, 116(23):11537–11546, 2019.

Seung, H. S., Sompolinsky, H., and Tishby, N. Statistical mechanics of learning from examples. *Physical review A*, 45(8):6056, 1992.

Shallue, C. J., Lee, J., Antognini, J., Sohl-Dickstein, J., Frostig, R., and Dahl, G. E. Measuring the effects of data parallelism on neural network training. *Journal of Machine Learning Research*, 20(112):1–49, 2019.

Smith, S. L., Kindermans, P.-J., Ying, C., and Le, Q. V. Don't decay the learning rate, increase the batch size. *arXiv preprint arXiv:1711.00489*, 2017.

Soudry, D., Hoffer, E., Nacson, M. S., Gunasekar, S., and Srebro, N. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878, 2018.

Strand, O. N. Theory and methods related to the singular-function expansion and landweber's iteration for integral equations of the first kind. *SIAM Journal on Numerical Analysis*, 11(4):798–825, 1974.

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

Teh, Y. W., Thiery, A. H., and Vollmer, S. J. Consistency and fluctuations for stochastic gradient langevin dynamics. *The Journal of Machine Learning Research*, 17(1):193–225, 2016.

Toulis, P. and Airoldi, E. M. Asymptotic and finite-sample properties of estimators based on stochastic gradients. *The Annals of Statistics*, 45(4):1694–1727, 2017.

Vaskevicius, T., Kanade, V., and Rebeschini, P. Implicit regularization for optimal sparse recovery. In *Advances in Neural Information Processing Systems*, pp. 2968–2979, 2019.

Vaswani, S., Bach, F., and Schmidt, M. Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. *arXiv preprint arXiv:1810.07288*, 2018.

Wang, Y. Asymptotic analysis via stochastic differential equations of gradient descent algorithms in statistical and computational paradigms. *arXiv preprint arXiv:1711.09514*, 2017.

Welling, M. and Teh, Y. W. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pp. 681–688, 2011.

Wilson, A., Roelofs, R., Stern, M., Srebro, N., and Recht, B. The marginal value of adaptive gradient methods in machine learning. In *Advances in Neural Information Processing Systems*, 2017.

Yao, Y., Rosasco, L., and Caponnetto, A. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, 2007.

Ying, Y. and Pontil, M. Online gradient descent learning algorithms. *Foundations of Computational Mathematics*, 8(5):561–596, 2008.

You, Y., Gitman, I., and Ginsburg, B. Scaling sgd batch size to 32k for imagenet training. *arXiv preprint arXiv:1708.03888*, 6, 2017.

Zhang, C., Kjellstrom, H., and Mandt, S. Determinantal point processes for mini-batch diversification. *arXiv preprint arXiv:1705.00607*, 2017.

Zhang, C., Liao, Q., Rakhlin, A., Miranda, B., Golowich, N., and Poggio, T. Theory of deep learning iib: Optimization properties of sgd. *arXiv preprint arXiv:1801.02254*, 2018.

Zhang, T. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the twenty-first international conference on Machine learning*, pp. 116. ACM, 2004.