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-implemention 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 $\ell_2$ 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 explicitly 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 $\ell_2$ 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 $\ell_2$ 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 $\ell_2$ 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 non-negative, but vanishes as time grows. The third term
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—including 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); Défосsez & 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); Duneaud 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; Nakniran 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,

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

(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

$$\beta^{(k)} = \beta^{(k-1)} + \frac{\epsilon}{m} \sum_{i \in I_k} (y_i - x_i^T \beta^{(k-1)}) x_i$$

$$= \beta^{(k-1)} + \frac{\epsilon}{m} \cdot X_k^T (y_k - X_k \beta^{(k-1)}),$$

(2)

for $k = 1, 2, 3, \ldots$, where $\epsilon > 0$ is a fixed step size, $m$ is the mini-batch size, and $I_k \subseteq \{1, \ldots, n\}$ denotes the mini-batch on iteration $k$ with $|I_k| = m$, 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
\[
\beta^{(k)} = \beta^{(k-1)} + \frac{\epsilon}{n} \cdot X^T (y - X \beta^{(k-1)}) + \epsilon \cdot \left( \frac{1}{m} X^T I_k (y I_k - X I_k \beta^{(k-1)}) - \frac{1}{n} X^T (y - X \beta^{(k-1)}) \right).
\] (3)
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
\[
Q_\epsilon(\beta) = \epsilon \cdot \text{Cov}_I \left( \frac{1}{m} X^T I (y I - X I \beta) \right), \tag{5}
\]
where the randomness is due to \( I \subseteq \{1, \ldots, n\} \). 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),
\[
\tilde{\beta}^{\text{gf}}(t) = (X^T X)^+ (I - \exp(-t X^T / n)) X^T y, \tag{7}
\]
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),
\[
\beta^{(k)} = \tilde{\beta}^{(k-1)} + \frac{\epsilon}{n} \cdot X^T (y - X \tilde{\beta}^{(k-1)}) + \epsilon \cdot \text{Cov}^{1/2}_I \left( \frac{1}{m} X^T I (y I - X I \tilde{\beta}^{(k-1)}) \right) z_k, \tag{8}
\]
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 \( \epsilon \) 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 \( \epsilon \)). Then, the first and second moments of \( \tilde{\beta}^{(k)} \) match those of \( \beta^{(k)} \), i.e., we have that both
\[
\mathbb{E}_Z \tilde{\beta}^{(k)} = \mathbb{E}_Z \tilde{\beta}^{(k)}, \quad \text{and} \quad \text{Cov}_Z \tilde{\beta}^{(k)} = \text{Cov}_Z \beta^{(k)}.
\]
Here, we let \( Z \) denote the randomness inherent to \( \tilde{\beta}^{\text{gf}}(t) \).

**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 \( y, X, \text{ and } \epsilon > 0 \). Let \( t \geq 0 \). Then
\[
\tilde{\beta}^{\text{gf}}(t) = \tilde{\beta}^{\text{gf}}(0) + \int_0^t \exp(-t \hat{\Sigma}) \cdot \text{Cov}(\beta^{\text{gf}}(\tau))^{1/2} \, dW(\tau),
\]
where \( \hat{\Sigma} \) 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 \( Q_\epsilon(\beta(t)) = (\epsilon/m) \cdot \Sigma \) (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
\[
d\beta(t) = \frac{1}{n} X^T (y - X \beta(t)) \, dt + \left( \frac{\epsilon}{m} \cdot \hat{\Sigma} \right)^{1/2} \, dW(t).
\] (10)
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,

\[
\min_{\beta \in \mathbb{R}} \frac{1}{2n} \sum_{i=1}^{n} (x_i \beta)^2.
\]

Let \( G_k = (1/m) \sum_{i=1}^{n} x_i^2 \), 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 \( \epsilon \) 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 \( G = (1/m) \sum_{i=1}^{n} x_i^2 \). We may calculate for the non-constant process that

\[
Q_\epsilon(\beta(t))^{1/2} = \theta \beta(t), \text{ where } \theta = (\epsilon/m \cdot G)^{1/2},
\]

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.

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

**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 \( \epsilon = 0.01 \).

**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.
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 $\hat{\beta} \in \mathbb{R}^p$,

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

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

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

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,

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

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

$$\text{Var}_{\eta,Z}^m(\hat{\beta}_{\text{sgf}}(t)) = \text{tr} \mathbb{E}_\eta[\text{Cov}_Z(\hat{\beta}_{\text{sgf}}(t) | \eta)] + \text{Var}_\eta(\hat{\beta}_{\text{sgf}}(t)).$$

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, $\text{tr} \mathbb{E}_\eta[\text{Cov}_Z(\hat{\beta}_{\text{sgf}}(t) | \eta)]$, 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

$$\text{tr} \mathbb{E}_\eta[\text{Cov}_Z(\hat{\beta}_{\text{sgf}}(t) | \eta)] \leq \frac{2ne}{m} \cdot \int_0^t f(\hat{\beta}_{\text{sgf}}(\tau)) \text{tr} [\Sigma \exp(2(\tau - t)\hat{\Sigma})] d\tau,$$

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

**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 $\beta$, let $h(\beta) = (y_1 - x_1^T \beta, \ldots, y_n - x_n^T \beta)$ denote the residuals at $\beta$, $F(\beta) = \text{diag}(h(\beta))^2$, 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))

$$Q_\epsilon(\beta) = \mathbb{E}_Z \left[ \frac{1}{m} X^T (y_I - X \hat{\beta}) \right] = \frac{1}{nm} X^T (F(\beta) - \tilde{F}(\beta)) X \leq \frac{1}{nm} X^T F(\beta) X,$$

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, $\mathbb{E}_{\eta,Z} \| \hat{\beta}_{\text{sgf}}(t) - \hat{\beta}_{\text{ridge}}(1/t) \|^2_2$.

Inspecting the bound in Lemma 4, we see that the variance due to mini-batching, $\text{tr} \mathbb{E}_\eta[\text{Cov}_Z(\hat{\beta}_{\text{sgf}}(t) | \eta)]$, depends on the expected loss of stochastic gradient flow, $f(\hat{\beta}_{\text{sgf}}(t))$. 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

$$u = 2\mu - 2n(\epsilon/m)^2 \cdot \left( \max_{i=1,...,p} \left[ \text{diag}(\hat{\Sigma}^2) \right]_{ii} \right),$$

$$v = \left( \frac{(\epsilon/m)^2 q \cdot \max_{i=1,...,p} \left[ \text{diag}(\hat{\Sigma}^2) \right]_{ii}}{u} \right),$$

$$w = \log \left( \| y \|^2_2 / (2m - q / 2n) \right).$$

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

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

**Theorem 2.** Fix (Theorem 1 in Ali et al. (2018))

Let $\beta$ deferred to the supplement for space reasons.

Putting Lemmas 4 and 5 together with Theorem 1 yields the analog of, e.g., Theorem 2 in Bassily et al. (2018), and may be of independent interest.

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

$$
\nu_i(t) = \frac{\exp(s_i)}{s_i - u/2} \left( \exp(-ut) - \exp(-2ts_i) \right)
$$

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 $\epsilon$ according to Lemma 5. Let $t > 0$.

- Then, relative to gradient flow,
  $$
  \text{Risk}(\hat{\beta}^{\text{sgf}}(t); \beta_0) \leq \text{Bias}^2(\hat{\beta}^{\text{gf}}(t); \beta_0) + \text{Var}_n(\hat{\beta}^{\text{gf}}(t)) + \epsilon \cdot \frac{n}{m} \sum_{i=1}^{p} \mathbb{E}_n \nu_i(t).
  $$

- Relative to ridge regression,
  $$
  \text{Risk}(\hat{\beta}^{\text{gf}}(t); \beta_0) \leq \text{Bias}^2(\hat{\beta}^{\text{ridge}}(1/t); \beta_0) + 1.6862 \cdot \text{Var}_n(\hat{\beta}^{\text{ridge}}(1/t)) + \epsilon \cdot \frac{n}{m} \sum_{i=1}^{p} \mathbb{E}_n \nu_i(t).
  $$

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

**Proof.** From Lemma 2, we have

$$
\hat{\beta}^{\text{sgf}}(t) = \hat{\beta}^{\text{gf}}(t) + \int_0^t \exp\left((\tau - t)\hat{\Sigma}\right)Q_x(\beta(\tau))^{1/2}dW(\tau).
$$

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

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

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

$$
\text{tr Cov}_n(Z(\hat{\beta}^{\text{sgf}}(t))) = \text{tr} \left( \mathbb{E}_n[\text{Cov}_Z(\hat{\beta}^{\text{sgf}}(t) | \eta)] + \text{Cov}_n(\mathbb{E}_Z(\hat{\beta}^{\text{sgf}}(t) | \eta)) \right)
$$

$$
= \text{tr} \left( \mathbb{E}_n[\text{Cov}_Z(\hat{\beta}^{\text{sgf}}(t) | \eta)] + \text{Cov}_n(\hat{\beta}^{\text{gf}}(t)) \right)
$$

$$
= \text{tr} \mathbb{E}_n[\text{Cov}_Z(\hat{\beta}^{\text{sgf}}(t) | \eta)] + \text{Var}_n(\hat{\beta}^{\text{gf}}(t)).
$$

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

$$
\text{tr} \mathbb{E}_n[\text{Cov}_Z(\hat{\beta}^{\text{sgf}}(t) | \eta)] = \mathbb{E}_n[\text{tr Cov}_Z(\hat{\beta}^{\text{sgf}}(t) | \eta)]
$$

$$
\leq \mathbb{E}_n \left[ 2\epsilon \cdot \int_0^t f(\hat{\beta}^{\text{gf}}(\tau)) \text{tr} [\hat{\Sigma} \exp(2(\tau - t)\hat{\Sigma})]d\tau \right]
$$

$$
= \frac{2\epsilon}{m} \cdot \int_0^t \mathbb{E}_n[f(\hat{\beta}^{\text{gf}}(\tau))] \text{tr} [\hat{\Sigma} \exp(2(\tau - t)\hat{\Sigma})]d\tau
$$

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

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,

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

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 $\epsilon$ as in Lemma 5. Let $t > 0$. Then,

$$
\text{Risk}(\hat{\beta}^{\text{sgf}}(t); \beta_0) - \text{Risk}(\hat{\beta}^{\text{ridge}}(1/t); \beta_0)
$$

$$
\leq 0.6862 \cdot \text{Var}_n(\hat{\beta}^{\text{ridge}}(1/t)) + \epsilon \cdot \frac{n}{m} \sum_{i=1}^{p} \mathbb{E}_n \nu_i(t).
$$
The Implicit Regularization of Stochastic Gradient Flow for Least Squares

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 $e/m$, 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 $e$ as in Lemma 5. Let $t > 0$. Define

$$\alpha = \frac{p\mu_e}{m(\mu - u/2)},$$

$$\gamma(t) = 1 + 2.164\epsilon \cdot \frac{v^2 \max(1/t, L)}{m\sigma^2},$$

$$\kappa = L/\mu, \text{and } \delta = \alpha/\|\beta_0\|^2.1/\kappa.$$

- Then, for gradient flow,\n
  Risk($\hat{\beta}^{\text{gf}}(t); \beta_0$) $\leq$ Bias$^2(\hat{\beta}^{\text{gf}}(t); \beta_0)$
  $$+ \delta \cdot |\text{Bias}(\hat{\beta}^{\text{gf}}(t); \beta_0)|^{1/\kappa} + \gamma(t) \cdot \text{Var}(\hat{\beta}^{\text{gf}}(t)).$$

- For ridge regression, Risk($\hat{\beta}^{\text{rg}}(t); \beta_0$) $\leq$ Bias$^2(\hat{\beta}^{\text{ridge}}(1/t); \beta_0)$
  $$+ \delta \cdot |\text{Bias}(\hat{\beta}^{\text{ridge}}(1/t); \beta_0)|^{1/\kappa} + 1.6862\gamma(t) \cdot \text{Var}(\hat{\beta}^{\text{ridge}}(1/t)).$$

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, $E_{\eta, Z} \|\hat{\beta}^{\text{rg}}(t) - \bar{\beta}^{\text{ridge}}(1/t)\|^2$. We start by giving a tight bound on the distance between the coefficients of gradient flow and ridge regression.

Lemma 7. Fix $X$. Let $t \geq 0$. Define

$$g(t) = \begin{cases} 
\frac{(1-\exp(-Lt))(1+Lt)}{(1-\exp(-t))(1+\mu t)}, & t \leq \frac{1.7933}{L} \\
\frac{\mu^t}{\frac{1.7933}{L} < t < \frac{1.7933}{\mu}}, & t \geq \frac{1.7933}{\mu} 
\end{cases}$$

Then,

$$E_{\eta} \|\hat{\beta}^{\text{gf}}(t) - \hat{\beta}^{\text{ridge}}(1/t)\|^2 \leq (g(t) - 1)^2 \cdot E_{\eta} \|\hat{\beta}^{\text{ridge}}(1/t)\|^2.$$  

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 $\hat{\beta}^{\text{gf}}(t)$ and $\hat{\beta}^{\text{ridge}}(1/t)$ 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.](image)

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

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

$$E_{\eta, Z} \|\hat{\beta}^{\text{rg}}(t) - \bar{\beta}^{\text{ridge}}(1/t)\|^2 \leq (g(t) - 1)^2 \cdot E_{\eta} \|\hat{\beta}^{\text{ridge}}(1/t)\|^2 + \epsilon \cdot \frac{n}{m} \sum_{i=1}^{P} \nu_i(t).$$

Proof. Expanding $E_{\eta, Z} \|\hat{\beta}^{\text{rg}}(t) - \bar{\beta}^{\text{ridge}}(1/t)\|^2$, adding and subtracting $E_{\eta, Z} \|\hat{\beta}^{\text{rg}}(t)\|^2$, and rearranging yields

$$E_{\eta, Z} \|\hat{\beta}^{\text{rg}}(t) - \bar{\beta}^{\text{ridge}}(1/t)\|^2 = E_{\eta} E_{\eta, Z} \|\hat{\beta}^{\text{rg}}(t)\|^2 + \text{tr} E_{\eta} \left[\text{Cov} \left(\hat{\beta}^{\text{gf}}(t)\right)\right] + \epsilon \cdot \frac{n}{m} \sum_{i=1}^{P} \nu_i(t).$$

As $Q_e(\beta(t))^{1/2}$ is continuous, it follows from standard properties of Brownian motion (e.g., Theorem 3.2.1 in Øksendal (2003)) that $E_{\eta} \|\hat{\beta}^{\text{gf}}(t)\|^2 = \beta^{\text{rg}}(t)$. Therefore, we have

$$E_{\eta, Z} \|\hat{\beta}^{\text{rg}}(t) - \bar{\beta}^{\text{ridge}}(1/t)\|^2 = E_{\eta} \|\hat{\beta}^{\text{rg}}(t) - \bar{\beta}^{\text{ridge}}(1/t)\|^2 + \text{tr} E_{\eta} \left[\text{Cov} \left(\hat{\beta}^{\text{rg}}(t)\right)\right].$$

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. \[\square\]
Remark 9. A bound on the coefficient error, \( \mathbb{E}_{\eta,Z} \| \hat{\beta}_{sgf}(t) - \beta_{ridge}(1/t) \|^2_2 \), 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 \( \Sigma \) 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 \( \epsilon = 2.2548e-4 \), 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 \( \lambda \) 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 \( \mathbb{E}_{\eta,Z} \| \hat{\beta}_{sgf}(t) - \beta_{ridge}(1/t) \|^2_2 \), 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.](image)

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 \( \ell_2 \) 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


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


