
Accelerated Gradient Flow: Risk, Stability, and Implicit Regularization

Yue Sheng*
University of Pennsylvania

Alnur Ali*
Stanford University

Abstract

Acceleration and momentum are the de facto standard in modern applications of machine learning and optimization, yet the bulk of the work on implicit regularization focuses instead on unaccelerated methods. In this paper, we study the statistical risk of the iterates generated by Nesterov’s accelerated gradient method and Polyak’s heavy ball method, when applied to least squares regression, drawing several connections to explicit penalization. We carry out our analyses in continuous-time, allowing us to make sharper statements than in prior work, and revealing complex interactions between early stopping, stability, and the curvature of the loss function.

1 INTRODUCTION

Acceleration (Nesterov, 1983, 2005, 1988, 2007; Tseng, 2008; Beck and Teboulle, 2009) and momentum (Polyak, 1964, 1987) are enormously popular tools for convex and non-convex optimization alike, playing central roles in many modern applications of machine learning and statistics. As is telling, a number of recent optimization algorithms commonly used to fit deep neural networks, e.g., Adam (Kingma and Ba, 2015), AdaGrad (Duchi et al., 2011), and RMSProp (Hinton et al., 2012), either leverage momentum directly, or are routinely modified in practice to incorporate it (Sutskever et al., 2013; Yang et al., 2016; Dozat, 2016; Wilson et al., 2017; Zou et al., 2018, 2019; Défossez et al., 2020).

One plausible explanation for (at least some of) the surprising recent successes of deep neural networks is that these optimization algorithms perform *implicit regularization*, i.e., the iterates generated by these algorithms possess a kind of statistical regularizer even without the use of any explicit regularizer (Nacson et al., 2018; Gunasekar et al., 2018; Soudry et al., 2018; Suggala et al., 2018; Ali et al., 2019; Poggio et al., 2019). Implicit regularization has undoubtedly seen an explosion of interest over the last few years, but most of the analyses focus on *unaccelerated* methods, e.g., the standard gradient descent iteration (Lee et al., 2016; Gunasekar et al., 2017, 2018; Nacson et al., 2018; Jacot et al., 2018; Soudry et al., 2018; Suggala et al., 2018; Pagliana and Rosasco, 2019; Du et al., 2019b,a; Ali et al., 2019; Hastie et al., 2019; Amari et al., 2020; Vaskevicius et al., 2020; Bartlett et al., 2021), or (mini-batch) stochastic gradient descent (Nacson et al., 2018; Gunasekar et al., 2018; Jain et al., 2018; Ali et al., 2020; Wu et al., 2020a), despite the prevalence of accelerated methods in practice. The reason for this focus is probably simplicity, as acceleration can be somewhat difficult to study formally.

In this paper, we exactly characterize the risk of both Nesterov’s accelerated gradient method, as well as Polyak’s heavy ball method, across the entire optimization path, when applied to the fundamental problem of least squares regression (i.e., without regularization); as a result, we draw a number of connections to ridge regression (Hoerl and Kennard, 1970), i.e., to explicit penalization. A key feature of our approach is that we carry out the analyses in continuous-time, simplifying many of the arguments and allowing us to make sharper statements than have been made in prior work.

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

- We derive exact expressions for the estima-

*These authors contributed equally.

tion risk of the iterates generated by Nesterov’s accelerated gradient method and Polyak’s heavy ball method, holding across the entire continuous-time optimization path, i.e., for any $t \geq 0$. To do so, we study two second-order differential equations that we call *accelerated gradient flow* and *heavy ball flow*, respectively.

- We show that, under optimal tuning, the risk associated with Nesterov’s method is at most 1.5991 times the risk of ridge regression.
- We demonstrate that, in general, it is impossible to give a tight coupling between Nesterov’s method (or the heavy ball method) and ridge across the *entire* path, due to instability, i.e., because the variance of accelerated methods can grow quickly and without bound, depending on the spectrum of the sample covariance matrix. Moreover, because accelerated methods are not descent methods in general, their risks tend to oscillate. As a whole, we provide a more refined picture of the “stability-convergence speed trade-off” than prior work, which is qualitatively different from that for gradient descent.
- Nonetheless, we give a tight bound on the *relative* parameter error, i.e., the ℓ_2 norm of the difference between the accelerated vs. ridge coefficients, normalized by the length of the ridge coefficients, holding across the entire path.
- We provide numerical experiments supporting our general theory, showing that under idealized conditions accelerated gradient methods can indeed reach low-risk solutions faster than standard gradient methods can, but that early-stopping should be used with care in general.

Outline. Here is an outline for the rest of this paper. In the next section, we give some background on acceleration and implicit regularization, and review related work. In Section 3 that follows, we introduce our continuous-time framework, and present our main results for Nesterov’s method. In Section 4, we present our results for the heavy ball method. We give numerical evidence for our findings in Section 5, and wrap up with a short discussion in Section 6.

2 BACKGROUND

2.1 Least Squares, Ridge Regression, Gradient Descent, and Gradient Flow

Given a fixed design matrix $X \in \mathbb{R}^{n \times p}$, and response points $y \in \mathbb{R}^n$ arising from a canonical linear model

$$y = X\beta_0 + \varepsilon, \quad (1)$$

with underlying coefficients $\beta_0 \sim (0, (r^2/p)I)$ and noise $\varepsilon \sim (0, \sigma^2 I)$, for some $\sigma > 0$, the usual least squares regression estimate is given by solving

$$\underset{\beta \in \mathbb{R}^p}{\text{minimize}} \quad \frac{1}{2n} \|y - X\beta\|_2^2. \quad (2)$$

Applying the standard (discrete-time) gradient descent iteration to least squares regression in (2) gives

$$\beta^{(k)} = \beta^{(k-1)} + \frac{\epsilon}{n} \cdot X^T(y - X\beta^{(k-1)}), \quad \beta^{(0)} = 0, \quad (3)$$

where $k \geq 0$ is an iteration counter, and $\epsilon > 0$ is a fixed step size. Taking infinitesimally small steps in (3), i.e., sending the step size $\epsilon \rightarrow 0$, yields the ordinary differential equation called gradient flow,

$$\dot{\beta}(t) = X^T(y - X\beta(t)), \quad \beta(0) = 0, \quad t \geq 0. \quad (4)$$

Here, $\dot{\beta}(t)$ is the time derivative of $\beta : \mathbb{R}_+ \rightarrow \mathbb{R}^p$.

Also, recall that the ridge regression estimate (Hoerl and Kennard, 1970), for any $\lambda \geq 0$, is simply

$$\hat{\beta}^{\text{ridge}}(\lambda) = \arg \min_{\beta \in \mathbb{R}^p} \frac{1}{2n} \|y - X\beta\|_2^2 + \frac{\lambda}{2} \|\beta\|_2^2. \quad (5)$$

2.2 Acceleration and Momentum

Now let $0 \leq s_1 \leq \dots \leq s_p$ denote the singular values of the sample covariance matrix $\hat{\Sigma} = X^T X/n$, and write $\mu = s_1$ and $L = s_p$. It is a standard fact that gradient descent, as in (3), with step size $\epsilon \leq 1/L$ converges to a solution of (2) in $O(1/k)$ iterations, which is suboptimal for first-order methods. On the other hand, Nesterov’s accelerated gradient method attains the optimal $O(1/k^2)$ rate, and works as follows. Nesterov’s method composes a gradient step with a momentum adjustment, i.e.,

$$\begin{aligned} \beta_{\text{Nest}}^{(k)} &= \beta_{\text{Nest}}^{(k-1)} + \frac{\epsilon}{n} \cdot X^T(y - X\theta^{(k-1)}) \\ \theta^{(k)} &= \beta_{\text{Nest}}^{(k)} + \frac{k-1}{k+2} \cdot (\beta_{\text{Nest}}^{(k)} - \beta_{\text{Nest}}^{(k-1)}). \end{aligned} \quad (6)$$

Nesterov’s method (6) is itself a refinement of the heavy ball method, introduced by Polyak, which simply allows the previous iterations to carry some momentum, i.e., we perform the update

$$\begin{aligned} \beta_{\text{hb}}^{(k)} &= \beta_{\text{hb}}^{(k-1)} + \frac{\epsilon}{n} \cdot X^T(y - X\beta_{\text{hb}}^{(k-1)}) \\ &\quad + \eta \cdot (\beta_{\text{hb}}^{(k)} - \beta_{\text{hb}}^{(k-1)}), \end{aligned} \quad (7)$$

where $\eta \geq 0$ is the momentum parameter. We initialize the iterations (6), (7) at zero.

Unlike Nesterov’s method (6), which is (globally) convergent for smooth convex objectives, i.e., those having Lipschitz continuous gradients, the heavy ball method (7) need not converge even for strongly convex objectives, though it is globally convergent for (2), provided that $\hat{\Sigma}$ is positive definite. Some works have demonstrated instances of divergence (Lessard et al., 2016), and others have shown that stronger conditions are in general required to establish convergence (Zavriev and Kostyuk, 1993; Ghadimi et al., 2014; Ochs et al., 2014).

2.3 Accelerated Continuous-Time Dynamics

Polyak’s method (7) was developed with a physical interpretation in mind, so it is natural to study the iteration (7) in continuous-time. In fact, (7) can be seen as the discretization of the second-order ordinary differential equation

$$\ddot{\beta}(t) + 2\mu^{1/2}\dot{\beta}(t) = \frac{1}{n}X^T(y - X\beta(t)), \quad (8)$$

with the initialization $\beta(0) = \dot{\beta}(0) = 0$, which we refer to as *heavy ball flow*.

On the other hand, Nesterov’s method (6) has remained a bit mysterious, despite its popularity, over the years. We discuss a few different interpretations for the iteration (6) in the next section, but for our purposes the interpretation of Su et al. (2014) turns out to be most useful. Su et al. (2014) view (6) as the discretization of another closely related differential equation, in analogy to what was done with (7), (8). We refer to this differential equation, i.e.,

$$\ddot{\beta}(t) + \frac{3}{t}\dot{\beta}(t) = \frac{1}{n}X^T(y - X\beta(t)), \quad (9)$$

with the initialization $\beta(0) = \dot{\beta}(0) = 0$, as *accelerated gradient flow*. In what follows, the differential equations (8) and (9) both play a key role in understanding the statistical properties of (7) and (6).

2.4 Related Work

Before turning to our main results in this paper, we first give a brief survey of related work.

Accelerated methods. Momentum was introduced by Polyak (Polyak, 1964, 1987). Nesterov refined the idea, and furthermore showed that acceleration is optimal among first-order methods (Nesterov, 1983, 2005, 1988, 2007). In direct analogy to Polyak (1964, 1987), and building on ideas found

in O’Donoghue and Candes (2015), Su et al. (2014) interpret the iteration (6) as the discretization of (9). Krichene et al. (2015); Wibisono et al. (2016); Wilson et al. (2016); Betancourt et al. (2018) follow up, and show that a broad class of first-order methods, including Nesterov’s method, may be seen as discretizations of a “master” differential equation, called the Bregman Lagrangian. Shi et al. (2021) study a slightly different family of differential equations than (8), (9), which more accurately reflect the different underlying dynamics of Nesterov’s method and the heavy ball method, when applied to strongly convex objectives. Allen-Zhu and Orecchia (2017) show that (6) may be seen as a kind of combination of gradient descent and mirror descent. Bubeck et al. (2015) give a geometric interpretation of the iteration (6), based on localization-type ideas.

Wilson et al. (2017); Levy and Duchi (2019) carry out interesting empirical and theoretical analyses, characterizing conditions under which accelerated methods may outperform their adaptive gradient (i.e., variable metric) counterparts, and vice-versa.

Finally, and especially relevant to the current paper, Pagliana and Rosasco (2019) give excess error bounds for Nesterov’s method, showing that it attains parametric rates of convergence under suitable smoothness conditions. To prove these results, the authors interpret Nesterov’s method as a spectral shrinkage map and consider a bias-variance decomposition, showing that the accelerated bias drops faster than the unaccelerated bias as a by-product and explaining the instability commonly seen with accelerated methods. As we will see in Section 3, these techniques are reminiscent of our own.

Additionally, Chen et al. (2018) show the sum of the worst-case optimization error and the algorithmic stability (in the sense of Bousquet and Elisseeff (2002); Hardt et al. (2016)), associated with any iterative algorithm including Nesterov’s method and the heavy ball method, over a chosen loss function class, is lower bounded by the minimax excess risk over that same class. The authors claim this exhibits a so-called stability-convergence trade-off, i.e. a large minimax risk implies a slow convergence rate (and conversely), again shedding light on instability.

Implicit regularization. Nearly all of the work on implicit regularization so far has looked at unaccelerated first-order methods, as mentioned in the introduction. The literature here is massive, so we cannot give a complete coverage, but some key references include Lee et al. (2016); Gunasekar et al.

(2017, 2018); Nacson et al. (2018); Jacot et al. (2018); Soudry et al. (2018); Suggala et al. (2018); Pagliana and Rosasco (2019); Du et al. (2019b,a); Ali et al. (2019, 2020); Wu et al. (2020b); Amari et al. (2020); Vaskevicius et al. (2020); Wu et al. (2020a).

In particular, we point out Suggala et al. (2018); Du et al. (2019a); Ali et al. (2019, 2020); Wu et al. (2020b) as very relevant works, performing their respective analyses in continuous-time, which is similar to the spirit of the current paper.

3 NESTEROV ACCELERATION

We begin by considering Nesterov's method.

3.1 An Exact Risk Expression

Fix $\beta_0 \in \mathbb{R}^p$. For any estimator $\hat{\beta} \in \mathbb{R}^p$, we write

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

Denoting any solution to accelerated gradient flow (9) as $\hat{\beta}^{\text{Nest}}(t)$, for $t \geq 0$, our goal is to characterize $\text{Risk}(\hat{\beta}^{\text{Nest}}(t); \beta_0)$, and relate it to $\text{Risk}(\beta^{\text{ridge}}(\lambda); \beta_0)$, under an appropriate correspondence between the tuning parameters t and λ .

A natural strategy, pursued by prior work for gradient descent (Suggala et al., 2018; Ali et al., 2019, 2020), is to write down a closed-form expression for $\hat{\beta}^{\text{Nest}}(t)$, and hope that it is amenable to analysis. Unfortunately, from looking back at (9), this is difficult. Therefore, we proceed slightly indirectly.

Write $X = n^{1/2}US^{1/2}V^T$ for a singular value decomposition, so that $\hat{\Sigma} = VSV^T$ with $S \in \mathbb{R}^{p \times p}$ containing s_i , $i = 1, \dots, p$. Now consider a rotation by V^T in (9), i.e., multiply both sides of (9) on the left by V^T . Then, defining $\alpha(t) = V^T\beta(t)$ and $\alpha_0 = V^T\beta_0$, we obtain the decoupled system

$$\ddot{\alpha}(t) + \frac{3}{t}\dot{\alpha}(t) + S\alpha(t) = n^{-1/2}(S^{1/2})^T U^T y, \quad (10)$$

where $\alpha(0) = \dot{\alpha}(0) = 0$.

When $s_i = 0$, $i = 1, \dots, p$, (10) becomes

$$\ddot{\alpha}_i(t) + \frac{3}{t}\dot{\alpha}_i(t) = 0,$$

and it can be checked that $\alpha_i(t) = 0$ is the unique solution. On the other hand, when $s_i > 0$, some tedious calculations (cf. Su et al. (2014)) show that

$$\alpha_i(t) = \frac{v_i}{s_i} \left(1 - 2 \frac{J_1(t\sqrt{s_i})}{t\sqrt{s_i}} \right)$$

uniquely solves (10), where $v_i \in \mathbb{R}^p$ denotes a column of V (eigenvector of $\hat{\Sigma}$), and J_1 denotes the Bessel function of the first kind of order one.

Therefore, putting together the pieces, we may express a solution to (9) relatively compactly as

$$\hat{\beta}^{\text{Nest}}(t) = (X^T X)^+ (I - Vg^{\text{Nest}}(S, t)V^T)X^T y, \quad (11)$$

for any $t \geq 0$. Here, A^+ denotes the Moore-Penrose pseudo-inverse of A , and importantly $g^{\text{Nest}} : \mathbb{R}_+^{p \times p} \times \mathbb{R}_+ \rightarrow \mathbb{R}_+^{p \times p}$ is a shrinkage map, i.e.,

$$g_{ij}^{\text{Nest}}(S, t) = \begin{cases} 2 \frac{J_1(t\sqrt{s_i})}{t\sqrt{s_i}}, & i = j \\ 0, & i \neq j. \end{cases} \quad (12)$$

The following result leverages the special structure in (11), (12) to characterize $\text{Risk}(\hat{\beta}^{\text{Nest}}(t); \beta_0)$.

Lemma 1 (Exact risk of accelerated gradient flow). *Assume the data model (1). Fix $\beta_0 \in \mathbb{R}^p$. Then, for $t \geq 0$, the risk of accelerated gradient flow (11) is*

$$\begin{aligned} \text{Risk}(\hat{\beta}^{\text{Nest}}(t); \beta_0) &= \sum_{i=1}^p \left(\underbrace{4(v_i^T \beta_0)^2 \frac{(J_1(t\sqrt{s_i}))^2}{t^2 s_i}}_{\text{Bias}} + \underbrace{\frac{\sigma^2}{n} \frac{\left(1 - 2 \frac{J_1(t\sqrt{s_i})}{t\sqrt{s_i}}\right)^2}{s_i}}_{\text{Variance}} \right). \end{aligned} \quad (13)$$

Here and below, we take by convention $(J_1(t\sqrt{x}))^2/(t^2 x) = 1/4$ and $(1 - 2J_1(t\sqrt{x})/(t\sqrt{x}))^2/x = 0$, when $x = 0$.

Proof. We prove a slightly more general result. Fix $t \geq 0$. Consider an estimator $\hat{\beta}(t) \in \mathbb{R}^p$ given by

$$\hat{\beta}(t) = (X^T X)^+ (I - Vg(S, t)V^T)X^T y,$$

where $g(S, t) \in \mathbb{R}_+^{p \times p}$ is a diagonal shrinkage map. Now recall the usual bias-decomposition, i.e.,

$$\text{Risk}(\hat{\beta}(t); \beta_0) = \text{Bias}^2(\hat{\beta}(t); \beta_0) + \text{Var}(\hat{\beta}(t)),$$

where we write

$$\text{Bias}^2(\hat{\beta}(t); \beta_0) = \|\mathbb{E}[\hat{\beta}(t)] - \beta_0\|_2^2,$$

$$\text{Var}(\hat{\beta}(t)) = \text{tr}(\text{Cov}(\hat{\beta}(t))).$$

For the bias, a few elementary calculations show:

$$\begin{aligned} \text{Bias}^2(\hat{\beta}(t); \beta_0) &= \|\mathbb{E}[(X^T X)^+ (I - Vg(S, t)V^T)X^T y] - \beta_0\|_2^2 \\ &= \sum_{i=1}^p (v_i^T \beta_0)^2 g_{ii}^2. \end{aligned}$$

Similarly, for the variance, we have that $\text{Var}(\hat{\beta}(t)) = \sum_{i=1}^p \sigma^2 (1 - g_{ii})^2 / (n s_i)$. Substituting $\hat{\beta}^{\text{Nest}}(t)$ and g^{Nest} in for $\hat{\beta}(t)$ and g gives the result. \square

3.2 Risk Inflation Under Oracle Tuning

Now recall the exact risk of the ridge estimator (5) (cf. Hsu et al. (2012); Dobriban and Wager (2018); Ali et al. (2019)), i.e., for any $\lambda \geq 0$, we have

$$\begin{aligned} \text{Risk}(\hat{\beta}^{\text{ridge}}(\lambda); \beta_0) &= \sum_{i=1}^p \left(\underbrace{(v_i^T \beta_0)^2}_{\text{Bias}} \frac{\lambda^2}{(s_i + \lambda)^2} + \underbrace{\frac{\sigma^2}{n} \frac{s_i}{(s_i + \lambda)^2}}_{\text{Variance}} \right). \end{aligned} \quad (14)$$

Prior work (Theorem 3 in Ali et al. (2019)) has shown that the optimal ridge risk using the oracle tuning parameter $\lambda^* = \sigma^2 p / (r^2 n)$ admits a tight coupling with that of early-stopped gradient flow (4) for the (unregularized) least squares problem (2), at time $t = \tau / \lambda^*$ with $\tau = 1$. As it turns out, there is really nothing special about the value $\tau = 1$, which the following result demonstrates. The result strengthens Theorem 3 from Ali et al. (2019), and is useful in the current paper for giving a tight coupling between accelerated gradient flow and ridge under oracle tuning, which we do next. We write $\text{Risk}(\hat{\beta}) = \mathbb{E} \|\hat{\beta} - \beta_0\|_2^2$ for the Bayes risk of $\hat{\beta} \in \mathbb{R}^p$.

Lemma 2 (Risk inflation for gradient flow). *Assume the data model (1). For any $t \geq 0$, denote the unique solution to (4) by*

$$\hat{\beta}^{\text{gf}}(t) = (X^T X)^+ (I - V g^{\text{gf}}(S, t) V^T) X^T y, \quad (15)$$

with the diagonal shrinkage map

$$g_{ij}^{\text{gf}}(S, t) = \begin{cases} \exp(-ts_i), & i = j \\ 0, & i \neq j. \end{cases} \quad (16)$$

Then, it holds for the gradient flow estimator (15) that

$$1 \leq \frac{\inf_{t \geq 0} \text{Risk}(\hat{\beta}^{\text{gf}}(t))}{\inf_{\lambda \geq 0} \text{Risk}(\hat{\beta}^{\text{ridge}}(\lambda))} \leq 1.0786.$$

Proof. Let $\alpha = r^2 n / (\sigma^2 p)$. The Bayes risk of gradient flow was given in Lemma 5 of Ali et al. (2019):

$$\begin{aligned} \text{Risk}(\hat{\beta}^{\text{gf}}(t)) &= \frac{\sigma^2}{n} \sum_{i=1}^p \left(\alpha \exp(-2ts_i) + \frac{(1 - \exp(-ts_i))^2}{s_i} \right), \end{aligned} \quad (17)$$

for any $t \geq 0$. Meanwhile, the optimal Bayes risk for ridge is well-known, i.e.,

$$\inf_{\lambda \geq 0} \text{Risk}(\hat{\beta}^{\text{ridge}}(\lambda)) = \frac{\sigma^2}{n} \sum_{i=1}^p \frac{\alpha}{\alpha s_i + 1}, \quad (18)$$

where $\lambda^* = \sigma^2 p / (r^2 n)$ is the optimal regularization strength. Now the lower bound follows because $\hat{\beta}^{\text{ridge}}(\lambda^*)$ is the Bayes estimator under (1), i.e., (18) must be smaller than (17), for any $t \geq 0$ (see, e.g., Theorem 3 in Ali et al. (2019) for a discussion).

As for the upper bound, we compare, for $s \geq 0$,

$$\alpha \exp(-2ts) + \frac{(1 - \exp(-ts))^2}{s} \quad \text{and} \quad \frac{\alpha}{\alpha s + 1}.$$

Let $\tau \geq 0$ be arbitrary, and plug $t = \tau / \lambda$ into the first term. Changing variables and dividing the two terms together gives rise to the following min-max problem, which clearly upper bounds the risk ratio:

$$\min_{\tau \geq 0} \max_{x \geq 0} \left\{ (1+x) \exp(-2\tau x) + \frac{(1+x)(1 - \exp(-\tau x))^2}{x} \right\}.$$

We can numerically compute the optimal value over a finite set of τ values, giving the result. \square

Assuming optimal tuning, Lemma 2 says that there is really no difference between the (explicitly regularized) ridge and (implicitly regularized) gradient flow estimators (5), (4), respectively, at least as far as estimation error goes. Moreover, the lemma strengthens Theorem 3 from Ali et al. (2019), by delivering the sharper constant 1.0786 (cf. 1.2147, from the theorem); by construction, this is the sharpest possible constant available, at least with the strategy used to prove these two results.

Finally, with the above results in hand, we turn to establishing the analog of Lemma 2 for accelerated gradient flow; the following result gives the details.

Theorem 1 (Risk inflation for accelerated gradient flow). *Assume the data model (1). Then, it holds for the accelerated gradient flow estimator (11) that*

$$1 \leq \frac{\inf_{t \geq 0} \text{Risk}(\hat{\beta}^{\text{Nest}}(t))}{\inf_{\lambda \geq 0} \text{Risk}(\hat{\beta}^{\text{ridge}}(\lambda))} \leq 1.5991.$$

Proof. The strategy is roughly the same as that for the proof of Lemma 2. Let $\alpha = r^2 n / (\sigma^2 p)$. Simply taking expectations in (13) with respect to β_0 shows that the Bayes risk of $\hat{\beta}^{\text{Nest}}(t)$ satisfies, for any $t \geq 0$,

$$\begin{aligned} \text{Risk}(\hat{\beta}^{\text{Nest}}(t)) &= \frac{\sigma^2}{n} \sum_{i=1}^p \left(\alpha \frac{(2J_1(t\sqrt{s_i}))^2}{t^2 s_i} + \frac{\left(1 - 2 \frac{J_1(t\sqrt{s_i})}{t\sqrt{s_i}}\right)^2}{s_i} \right). \end{aligned}$$

The lower bound follows just as in the proof of Lemma 2. For the upper bound, we now compare

$$\alpha \frac{(2J_1(t\sqrt{x}))^2}{t^2x} + \frac{\left(1 - 2\frac{J_1(t\sqrt{x})}{t\sqrt{x}}\right)^2}{x} \quad \text{and} \quad \frac{\alpha}{\alpha x + 1},$$

for $x \geq 0$. Letting $\tau \geq 0$, plugging $t = \tau\lambda^{-1/2}$ into the first term, and dividing, we get

$$\min_{\tau \geq 0} \max_{x \geq 0} \left\{ \frac{4(1+x)J_1^2(\tau\sqrt{x})}{\tau^2x} + \frac{(1+x)\left(1 - \frac{2J_1(\tau\sqrt{x})}{\tau\sqrt{x}}\right)^2}{x} \right\}.$$

The result follows by numerical maximization. \square

3.3 Bias, Variance, and Stability

Though Theorem 1 shows the optimal ridge and accelerated risks are similar, we might hope for a bound holding *uniformly* over $t \geq 0$, assuming a suitable relation between t , λ . Unfortunately, such a result is not possible in general, as we now show.

It is natural to assume the relation $t = \lambda^{-1/2}$. Now from looking back at the bias-variance decompositions that were given in (14), (13), we would like to establish the existence of two universal constants $C_1, C_2 > 0$, such that

$$\begin{aligned} \frac{(2J_1(\sqrt{x}))^2}{x} &\leq C_1 \frac{1}{(1+x)^2} \\ \left(1 - \frac{2J_1(\sqrt{x})}{\sqrt{x}}\right)^2 &\leq C_2 \frac{x^2}{(1+x)^2}, \end{aligned}$$

where $x = t^2s$, in order to prove the required bound. However, as $J_1(x) = O(x^{-1/2})$, we see $4J_1^2(\sqrt{x})/x = O(x^{-3/2})$, and therefore $O(x^2 \cdot x^{-3/2})$ grows without bound, i.e., finding such a C_1 is impossible.

Simple empirical examples demonstrate and illuminate the issue. In Figure 2, we plot the bias, variance, and risk of accelerated gradient flow (13), heavy ball flow (21), standard gradient flow (see, e.g., Lemma 5 in Ali et al. (2019)), and ridge regression (14), for data arising from the response model (1) with $n = 500$, $p = 100$, and sample covariance matrices $\hat{\Sigma}$ having singular values satisfying $s_i = C/i^\nu$, $i = 1, \dots, p$, with $C = 1$ and $\nu \in \{0.1, 0.5, 1, 2\}$. We generated (dense) underlying coefficients β_0 , such that the signal-to-noise ratio $\|\beta_0\|_2^2/\sigma^2 = 1$. Below, we make a few observations.

- When $\nu \in \{0.1, 0.5\}$ is small, i.e., there are many sizable eigenvalues, then even with small t (as in the first row of the figure), we see that

the ridge bias drops faster than the accelerated bias, reflecting the situation described earlier.

- On the other hand, when t is large, even though we may have $\nu \in \{1, 2\}$, i.e., many small eigenvalues (as in the last row of the figure), we see the roles of the two biases reverse, so that now the accelerated bias drops faster.

This second situation, i.e., when the design matrix is ill-conditioned, is interesting to inspect further. First of all, a simple Taylor expansion of the bias terms in (13), (17) confirms the situation: when $s \approx 0$ and t is large enough, then we have $2J_1(t\sqrt{s})/(t\sqrt{s}) \ll \exp(-ts)$. But a quickly diminishing bias implies a quickly growing variance (e.g., the last two rows), revealing a kind of instability due to acceleration that is more subtle than has been described previously (Chen et al., 2018; Pagliana and Rosasco, 2019).

- We comment on one obvious aspect of the plots. The fact that accelerated methods are not descent methods implies their bias and variance are not monotone, and therefore that the accelerated risk can (and does) oscillate. These oscillations are especially pronounced in the ill-conditioned regime (the last two rows), indicating that early stopping should be used carefully.

How do we reconcile this last observation with the good performance of acceleration seen in practice (e.g., Wilson et al. (2017)), when data is high-dimensional? Well, in these settings, an interpolating solution is generally sought (Hastie et al., 2019), circumventing some of the issues described above. But when early stopping is employed with acceleration in practice, the stopping point is usually where the hold-out set error curve attains its minimum, i.e., roughly what we expect from the oracle tuning parameter (recall Theorem 1).

As a whole, the story here appears to be more complex and subtle than with standard gradient descent.

3.4 Parameter Error Bound

Although we cannot give a pointwise risk coupling between accelerated gradient flow and ridge (as was just discussed), we can still give a *relative* bound on the error $\mathbb{E}\|\hat{\beta}^{\text{Nest}}(t) - \hat{\beta}^{\text{ridge}}(\lambda)\|_2^2$, assuming the natural correspondence $t = \lambda^{-1/2}$. Our final result for this section (below) delivers such a bound.

Theorem 2 (Parameter error bound). *Assume the*

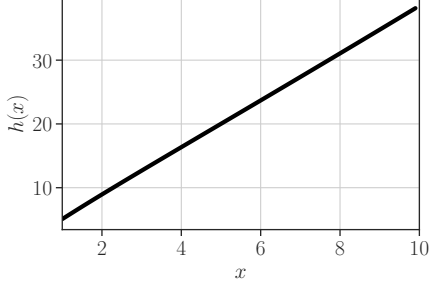


Figure 1: The function $h(x)$ from Theorem 4.

data model (1). For any $t \geq 0$, it holds that

$$\mathbb{E}\|\hat{\beta}^{\text{Nest}}(t) - \hat{\beta}^{\text{ridge}}(1/t^2)\|_2^2 \leq 0.7657 \cdot \mathbb{E}\|\hat{\beta}^{\text{ridge}}(1/t^2)\|_2^2.$$

The proof of the theorem is in the appendix. The bound tells us the parameter error is well-controlled, as the right-hand side is small when t is small and tends quickly to a constant value as t grows.

4 HEAVY BALL METHOD

Now we turn to analyzing the heavy ball method. In what follows, we assume that $\mu > 0$.

In analogy to both (11) and Lemma 1, our first result gives a closed-form solution to the heavy ball differential equation, from (8), along with a precise characterization of its estimation risk. The proof of the result is straightforward and in the appendix.

Lemma 3 (Solution and risk of heavy ball flow). *Assume the data model (1). Then, for any $t \geq 0$:*

- the heavy ball flow estimator, defined as

$$\hat{\beta}^{\text{hb}}(t) = (X^T X)^+ (I - V g^{\text{hb}}(S, t)) V^T X^T y, \quad (19)$$

where the shrinkage map has entries

$$g_{ij}^{\text{hb}}(S, t) = \exp(-\mu^{1/2} t) \times \left(\cos(t(s_i - \mu)^{1/2}) + \frac{\mu^{1/2} \sin(t(s_i - \mu)^{1/2})}{(s_i - \mu)^{1/2}} \right), \quad (20)$$

for $i = j$, and zeros everywhere else, solves (8);

- the risk of the heavy ball flow estimator (19) is

$$\text{Risk}(\hat{\beta}^{\text{hb}}(t); \beta_0) = \sum_{i=1}^p \left(\underbrace{(v_i^T \beta_0)^2 (1 - g_{ii}^{\text{hb}}(S, t))^2}_{\text{Bias}} + \underbrace{\frac{\sigma^2 (g_{ii}^{\text{hb}}(S, t))^2}{n s_i}}_{\text{Variance}} \right). \quad (21)$$

The next result bounds the risk inflation of the heavy ball flow estimator (19) over ridge (5), albeit by a worse absolute constant than in Theorem 1. The shrinkage map (20) is considerably more complex than, say, (12), which makes the analysis difficult. We suspect the constant is suboptimal and due to our proof technique, as numerical examples (see, e.g., Figure 2) indicate (19), (11) behave similarly.

Lemma 4 (Risk inflation for heavy ball flow). *Assume the data model (1). Let $\kappa = L/\mu$, and define*

$$h(x) = 8x^{2/3} + \left(1 + \left(\frac{x^{1/3} + \sqrt{5x^{2/3} - 4}}{2} \right)^2 \right) \times \left(\frac{\sqrt{5x^{2/3} - 4}}{2x^{1/3}} + \frac{3}{2} \right)^2 \exp \left(-\frac{x^{1/3} + \sqrt{5x^{2/3} - 4}}{x^{1/3}} \right). \quad (22)$$

Then, it holds for heavy ball flow (19) that

$$1 \leq \frac{\inf_{t \geq 0} \text{Risk}(\hat{\beta}^{\text{hb}}(t))}{\inf_{\lambda \geq 0} \text{Risk}(\hat{\beta}^{\text{ridge}}(\lambda))} \leq h(\kappa).$$

The proof is lengthy; see the appendix. To get a sense of the result, we plot $h(x)$ in Figure 1.

Our last result delivers a parameter error bound, similar to Theorem 2; its proof is in the appendix.

Theorem 3 (Parameter error bound for heavy ball). *Assume the data model (1). For $t \geq 0$, it holds that*

$$\mathbb{E}\|\hat{\beta}^{\text{hb}}(t) - \hat{\beta}^{\text{ridge}}(1/t^2)\|_2^2 \leq 25 \cdot \mathbb{E}\|\hat{\beta}^{\text{ridge}}(1/t^2)\|_2^2.$$

5 ADDITIONAL EXAMPLES

Due to space limitations, we present additional numerical examples with Gaussian, Student-t, and orthogonal data matrices, in the appendix; the take-away message is similar to that for Figure 2.

6 CONCLUSION

We studied the statistical properties of the iterates generated by Nesterov's accelerated gradient method and Polyak's heavy ball method, for least squares regression, making several connections to explicit penalization. There are a number of directions that would be interesting to pursue, e.g., extending the analysis to cover the high-resolution differential equations of Shi et al. (2021), as well as studying adaptive gradient methods and conjugate gradients through the same lens.

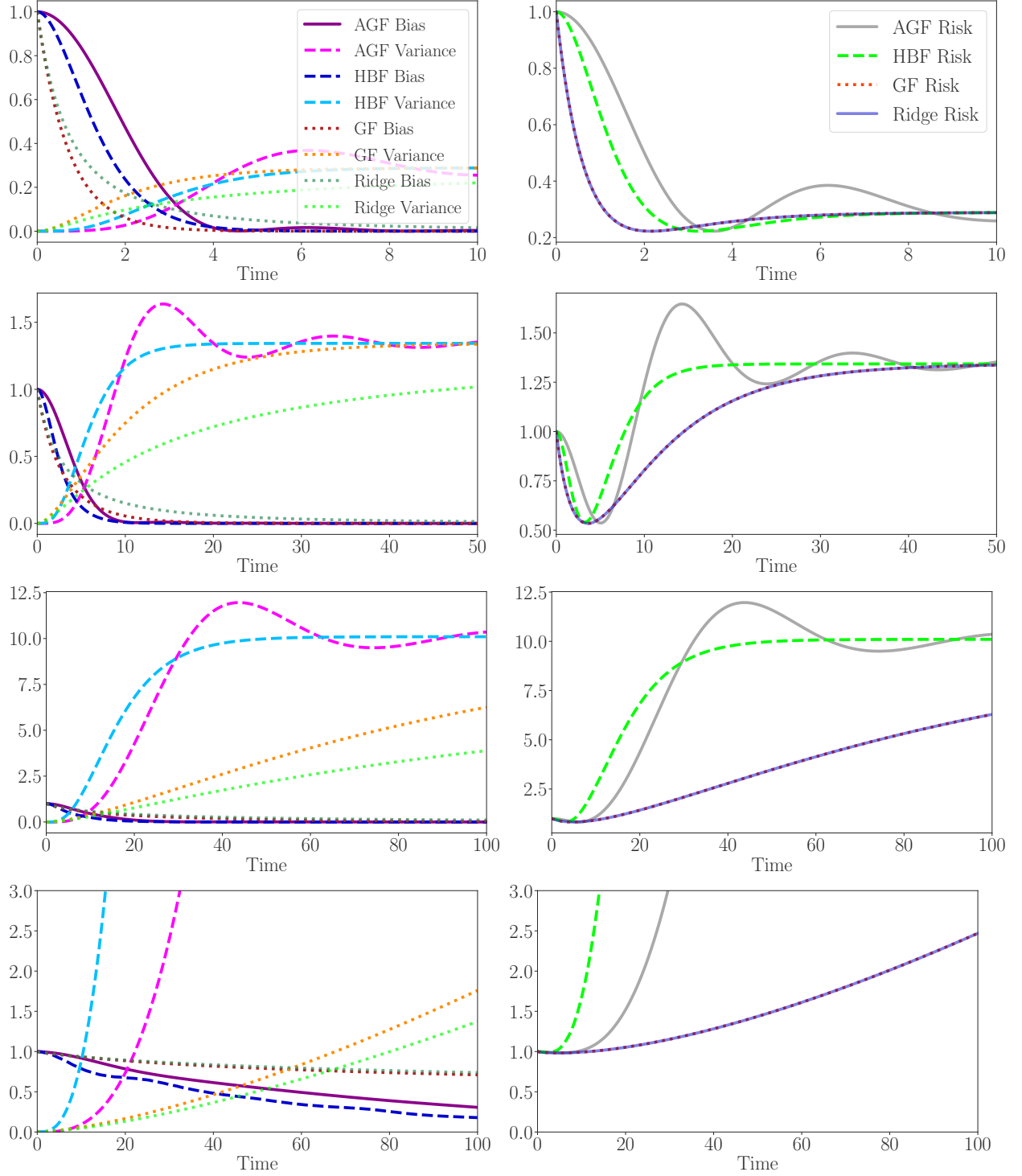


Figure 2: Bias, variance (left column), and risk (right column) for accelerated gradient flow (13), heavy ball flow (21), standard gradient flow (see, e.g., Lemma 5 in Ali et al. (2019)), and ridge regression (14). Each row corresponds to a different sample covariance matrix with its eigenvalues set according to $s_i = C/i^\nu$, for $i = 1, \dots, p$, with $C = 1$ and $\nu \in \{0.1, 0.5, 1, 2\}$, from top to bottom, respectively.

7 APPENDIX

7.1 Proof of Theorem 2

Fix ε . Using the correspondence $t = \lambda^{-1/2}$ and making a few simplifications gives

$$\|\hat{\beta}^{\text{Nest}}(t) - \hat{\beta}^{\text{ridge}}(1/t^2)\|_2^2 = \left\| n^{-1/2} t V \left((t^2 S)^+ (I - g^{\text{Nest}}(S, t)) - (t^2 S + I)^{-1} \right) ((t^2 S)^{1/2})^T U^T y \right\|_2^2.$$

So if we could find a finite numerical constant $C > 0$ satisfying

$$((t^2 S)^+ (I - g^{\text{Nest}}(S, t)) - (t^2 S + I)^{-1})^2 \preceq (C \cdot (t^2 S + I)^{-1})^2,$$

or equivalently

$$((t^2 S)^+ (I - g^{\text{Nest}}(S, t)) (t^2 S + I)^{-1} - I)^2 \preceq C^2 \cdot I,$$

then we would be done.

Noting that each of the matrices in the last display is diagonal, we define

$$f(x) = \left(1 - 2 \frac{J_1(x)}{x} \right) \frac{x^2 + 1}{x^2}.$$

Numerically maximizing $(f(x) - 1)^2$ over $x \geq 0$ gives $(f(x) - 1)^2 \leq 49/64$, which serves as such a constant. This shows the result. \square

7.2 Proof of Lemma 3

We follow a strategy similar to the one leading up to (10) for Nesterov's method, i.e., multiplying (8) on the left by V^T gives rise to the decoupled system

$$\ddot{\alpha}(t) + 2\mu^{1/2} \dot{\alpha}(t) - S\alpha_0 - n^{-1/2} S^{1/2} U^T \varepsilon + S\alpha(t) = 0, \quad (23)$$

with $\alpha(0) = \dot{\alpha}(0) = 0$. After some tedious (but standard) calculations, we obtain that a solution to (23) has the form

$$\alpha_i(t) = \frac{v_i}{s_i} - \frac{\exp(-t\mu^{1/2}) \cos(t(s_i - \mu)^{1/2}) v_i}{s_i} - \frac{\exp(-t\mu^{1/2}) \sin(t(s_i - \mu)^{1/2}) v_i \mu^{1/2}}{s_i (s_i - \mu)^{1/2}}, \quad i = 1, \dots, p. \quad (24)$$

It follows that we can write a solution to (8) as

$$\hat{\beta}^{\text{hb}}(t) = (X^T X)^+ (I - V g^{\text{hb}}(S, t)) V^T X^T y,$$

with the diagonal shrinkage map

$$g_{ij}^{\text{hb}}(S, t) = \begin{cases} \exp(-\mu^{1/2} t) \cdot \left(\cos(t(s_i - \mu)^{1/2}) + \frac{\mu^{1/2} \sin(t(s_i - \mu)^{1/2})}{(s_i - \mu)^{1/2}} \right), & i = j \\ 0, & i \neq j, \end{cases}$$

just as claimed. Invoking the risk expansion given in the proof of Lemma 1 immediately proves the second part of the result. \square

7.3 Proof of Lemma 4

Looking back at the proofs of Lemmas 1 and 2 in the main paper, we can see that proving the required result boils down to establishing the inequality

$$\alpha g^{\text{hb}}(s, t)^2 + \frac{(1 - g^{\text{hb}}(s, t))^2}{s} \leq C \cdot \frac{\alpha}{\alpha s + 1}, \quad (25)$$

for fixed $s, t > 0$ and some finite numerical constant $C > 0$.

Let $\tau > 0$ be a constant to be specified later. Consider the change of variables $t = \tau\alpha^{1/2}$, $x = (\alpha s)^{1/2}$, $a = (\alpha\mu)^{1/2}$, and $b = (\alpha(s - \mu))^{1/2}$. Plugging these and the definition of the heavy ball shrinkage map (20) into (25), and rearranging, we see that showing (25) is equivalent to showing

$$(1 + x^2) \left(e^{-\tau a} \left(\cos \tau b + \frac{a}{b} \sin \tau b \right) \right)^2 + \frac{1 + x^2}{x^2} \left(1 - e^{-\tau a} \left(\cos \tau b + \frac{a}{b} \sin \tau b \right) \right)^2 \leq C. \quad (26)$$

We proceed by bounding each of the two terms (i.e., the bias and the variance) on the left-hand side separately.

Bounding the bias. The helper Lemma 5 appearing below gives us a bound for the first (bias) term in (26), i.e., we know that

$$\left(e^{-\tau a} \left(\cos \tau b + \frac{a}{b} \sin \tau b \right) \right)^2 \leq (\tau x / \sqrt{\kappa} + 1)^2 e^{-2\tau x / \sqrt{\kappa}}.$$

To further bound the right-hand side, consider the function

$$\tilde{h}(x) = (1 + x^2)(\tau x / \sqrt{\kappa} + 1)^2 e^{-2\tau x / \sqrt{\kappa}}, \quad x \geq 0,$$

Differentiating \tilde{h} with respect to x gives

$$\frac{\partial \tilde{h}(x)}{\partial x} = -\frac{2x}{z^3} (x + z) \left(x - \frac{z - \sqrt{5z^2 - 4}}{2} \right) \left(x - \frac{z + \sqrt{5z^2 - 4}}{2} \right) e^{-2x/z},$$

where we defined $z = \kappa^{1/2} / \tau$. Some inspection reveals that there are three cases to consider.

- $0 < z \leq \frac{2}{\sqrt{5}}$: in this case, it follows that $\tilde{h}(x)$ attains its maximum over $x \geq 0$ at 0, with $\tilde{h}(0) = 1$.
- $\frac{2}{\sqrt{5}} < z < 1$: in this case, it follows that both 0 and $x^* = (z + \sqrt{5z^2 - 4})/2$ are maxima, and it can be checked numerically that $\tilde{h}(0) > \tilde{h}(x^*)$ for $z < 0.907$, but $\tilde{h}(0) \leq \tilde{h}(x^*)$ for $z \geq 0.907$.
- $z \geq 1$: in this case, it follows that $\tilde{h}(x)$ again attains its maximum at x^* .

To summarize: when $z < 0.907$, we have that $\tilde{h}(x) \leq \tilde{h}(0) = 1$ for $x \geq 0$, but when $z \geq 0.907$, we have the bound $\tilde{h}(x) \leq \tilde{h}(x^*)$. As $\kappa \geq 1 > (4/5)^{3/2}$, we may conclude that

$$\tilde{h}(x) \leq \tilde{h}(x^*) = h(x^*) - 8(x^*)^{2/3}, \quad x \geq 0, \quad (27)$$

which yields a bound on the bias term in (26).

Bounding the variance. As for the second (variance) term in (26), the helper Lemma 6 appearing below tells us that there are two cases to consider:

- for $x \leq 1$, it follows that

$$\begin{aligned} \frac{1 + x^2}{x^2} \left(1 - e^{-\tau a} \left(\cos \tau b + \frac{a}{b} \sin \tau b \right) \right)^2 &\leq \frac{1 + x^2}{x^2} \cdot 4(\tau x)^4 \\ &\leq 8\tau^4; \end{aligned} \quad (28)$$

- for $x > 1$: it follows that

$$\begin{aligned} \frac{1 + x^2}{x^2} \left(1 - e^{-\tau a} \left(\cos \tau b + \frac{a}{b} \sin \tau b \right) \right)^2 &\leq \frac{1 + x^2}{x^2} \left(1 + (\tau x / \sqrt{\kappa} + 1) e^{-\tau x / \sqrt{\kappa}} \right)^2 \\ &\leq 2 \left(1 + (\tau / \sqrt{\kappa} + 1) e^{-\tau / \sqrt{\kappa}} \right)^2. \end{aligned} \quad (29)$$

We now let $\tau = \kappa^{1/6}$, i.e., $z = \kappa^{1/3}$. Taking the maximum of (28), (29) gives

$$\begin{aligned} \max \left\{ 8\tau^4, 2 \left(1 + (\tau/\sqrt{\kappa} + 1)e^{-\tau/\sqrt{\kappa}} \right)^2 \right\} &= 8\tau^4 \\ &= 8\kappa^{2/3}, \end{aligned} \quad (30)$$

giving a bound on the variance term in (26).

Putting it all together. Adding (27), (30) together and using $z = \kappa^{1/3}$ shows that we may take $C = h(\kappa)$ in (26), completing the proof. \square

7.4 Statement and proof of helper Lemma 5

Lemma 5. *Let $s, t > 0$. Define $a = \mu^{1/2}t$, $b = t(s - \mu)^{1/2}$, and $x = ts^{1/2}$. Then, it holds that*

$$\left(\left(\cos b + \frac{a}{b} \sin b \right) e^{-a} \right)^2 \leq \left(\frac{x}{\sqrt{\kappa}} + 1 \right)^2 e^{-2x/\sqrt{\kappa}}.$$

Proof. Observe that $a \geq x/\sqrt{\kappa}$ and also that

$$\frac{a^2}{x^2 - a^2} = \frac{a^2}{b^2} = \frac{\mu t^2}{(s - \mu)t^2} = \frac{\mu}{s - \mu} = \frac{1}{s/\mu - 1} \geq \frac{1}{\kappa - 1}. \quad (31)$$

These together imply

$$\begin{aligned} \left(\left(\cos b + \frac{a}{b} \sin b \right) e^{-a} \right)^2 &= \left(\cos^2 b + \frac{2a}{b} \sin b \cos b + \frac{a^2 \sin^2 b}{b^2} \right) e^{-2a} \\ &\leq (1 + 2a + a^2) e^{-2a} \\ &\leq \left(\frac{x}{\sqrt{\kappa}} + 1 \right)^2 e^{-2x/\sqrt{\kappa}}. \end{aligned}$$

The first line follows by simply expanding the square, the second by using $\sin x \leq x$ for $x \geq 0$, and the third by using $a \geq x/\sqrt{\kappa}$. This completes the proof. \square

7.5 Statement and proof of helper Lemma 6

Lemma 6. *Let $s, t > 0$. Define $a = \mu^{1/2}t$, $b = t(s - \mu)^{1/2}$, and $x = ts^{1/2}$. Then, it holds that*

$$\begin{aligned} \left(1 - \left(\cos b + \frac{a}{b} \sin b \right) e^{-a} \right)^2 &\leq 4x^4, \quad x \leq 1; \\ \left(1 - \left(\cos b + \frac{a}{b} \sin b \right) e^{-a} \right)^2 &\leq \left(1 + (x/\sqrt{\kappa} + 1)e^{-x/\sqrt{\kappa}} \right)^2, \quad x > 1. \end{aligned}$$

Proof. To see the first inequality, we rewrite $1 - (\cos b + a/b \cdot \sin b) \exp(-a)$ as

$$(1 - e^{-a} - a) + \left(a - a \frac{\sin b}{b} \right) + \left(e^{-a} \sin^2 \frac{b}{2} \right) + \left(a(1 - e^{-a}) \frac{\sin b}{b} \right).$$

Now define two auxiliary functions $f_1(x) = x^2/2 + 1 - x - e^{-x}$ and $f_2(x) = x^2/2 - x + \sin x$. Since $f_1''(x) = 1 - e^{-x} \geq 0$ on $x \in [0, \infty)$, we have $f_1'(x) = x - 1 + e^{-x} \geq f_1'(0) = 0$ on $x \in [0, \infty)$, which means $f_1(x)$ is non-decreasing, $f_1(x) \geq f_1(0) = 0$. Thus, we have $|1 - e^{-a} - a| \leq a^2/2 \leq x^2/2$. Similarly, we can consider the derivatives of $f_2(x)$

$$f_1'(x) = x - 1 + \cos x, \quad f_2''(x) = 1 - \sin x \geq 0.$$

So $f'_2(x) \geq f'_2(0) = 0$, which means $f_2(x) \geq 0$ on $x \in [0, \infty)$, i.e. $x - \sin x \leq x^2/2$. Therefore, we have

$$\left| a - a \frac{\sin b}{b} \right| = a \left| \frac{b - \sin b}{b} \right| \leq a \left| \frac{b^2/2}{b} \right| = \frac{ab}{2} \leq \frac{a^2 + b^2}{4} = \frac{x^2}{4}.$$

The following bounds are straightforward:

$$e^{-a} \sin^2 \frac{b}{2} \leq \frac{x^2}{4}, \quad \left| a(1 - e^{-a}) \frac{\sin b}{b} \right| \leq x^2,$$

where we used $|\sin x| \leq |x|$ and $1 - \exp(-x) \leq x$ on $x \in [0, \infty)$. Putting all these together, we have

$$\left(1 - \left(\cos b + \frac{a}{b} \sin b \right) e^{-a} \right)^2 \leq \left(\frac{x^2}{2} + \frac{x^2}{4} + \frac{x^2}{4} + x^2 \right)^2 = 4x^4,$$

as claimed.

To see the second inequality, following arguments similar to those used above shows that

$$\left| \left(\cos b + \frac{a}{b} \sin b \right) e^{-a} \right| \leq (a+1)e^{-a} \leq (x/\sqrt{\kappa} + 1)e^{-x/\sqrt{\kappa}},$$

which implies

$$\left(1 - \left(\cos b + \frac{a}{b} \sin b \right) e^{-a} \right)^2 \leq \left(1 + (x/\sqrt{\kappa} + 1)e^{-x/\sqrt{\kappa}} \right)^2,$$

completing the proof. \square

7.6 Proof of Theorem 3

Fix ε . Similar to what was done in the proof of Theorem 2, putting the correspondence $\lambda = t^{-1/2}$ together with some basic manipulations, we obtain

$$\|\hat{\beta}^{\text{hb}}(t) - \hat{\beta}^{\text{ridge}}(1/t^2)\|_2^2 = \left\| n^{-1/2} t V \left((t^2 S)^+ (I - g^{\text{hb}}(S, t)) - (t^2 S + I)^{-1} \right) ((t^2 S)^{1/2})^T U^T y \right\|_2^2.$$

So (again) we need to find a constant $C > 0$ satisfying

$$\left((t^2 S)^+ (I - g^{\text{hb}}(S, t)) - (t^2 S + I)^{-1} \right)^2 \preceq (C \cdot (t^2 S + I)^{-1})^2. \quad (32)$$

Let

$$f(x) = \left(1 - \left(\cos b + \frac{a}{b} \sin b \right) e^{-a} \right) \frac{x^2 + 1}{x^2}.$$

Plugging (20) into (32), we see that showing (32) is equivalent to showing $f(x) \leq C$, for $x \geq 0$ and some finite numerical constant $C > 0$. Following arguments similar to those used when controlling the variance term in the proof of Lemma 4, we see that $f(x)^2 \leq 16$ for $x \geq 0$, so that $(f(x) - 1)^2 \leq 25$, completing the proof. \square

7.7 Additional Numerical Examples

Here we present some additional numerics for a broader range of feature covariance structures. We consider sample covariance matrices $\hat{\Sigma}$ having i.i.d. Gaussian and Student-t entries, as well as those satisfying the orthogonality condition $X^T X/n = s \cdot I$, for $s \in \{0.1, 1\}$. In Figure 3, we plot the bias, variance, and risk of accelerated gradient flow (13), heavy ball flow (21), standard gradient flow (see Lemma 5 in Ali et al. (2019)), and ridge regression (14), for each of these covariance structures and response points arising from the canonical model (1). We follow the same generic experimental setup here as in Section 5 of the main paper, i.e., we set $n = 500$, $p = 100$, and generate coefficients β_0 as well as set the noise variance σ^2 such that $\|\beta_0\|_2^2/\sigma^2 = 1$.

We include the simulations with Gaussian and Student-t data to give a sense of how accelerated methods behave under the “usual” experimental conditions. The results, plotted in the first two rows of Figure 3, roughly show most of the methods behaving similarly.

The story changes a bit when we adjust the scale of the eigenvalues of $\hat{\Sigma}$, as in the last two rows of the figure. Here we see trends similar to those in Figure 2, i.e., when $s = 0.1$ is (relatively) small, and t is large enough, then we can see the accelerated bias drop faster than the standard gradient flow bias, and some instability arises. When $s = 1$ is relatively large, the situation reverses.

References

- Alnur Ali, J. Zico Kolter, and Ryan J. Tibshirani. A continuous-time view of early stopping for least squares regression. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, pages 1370–1378, 2019.
- Alnur Ali, Edgar Dobriban, and Ryan J. Tibshirani. The implicit regularization of stochastic gradient flow for least squares. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.
- Zeyuan Allen-Zhu and Lorenzo Orecchia. Linear coupling: An ultimate unification of gradient and mirror descent. In *Proceedings of the 8th Innovations in Theoretical Computer Science*, ITCS ’17, 2017.
- Shun-ichi Amari, Jimmy Ba, Roger Grosse, Xuechen Li, Atsushi Nitanda, Taiji Suzuki, Denny Wu, and Ji Xu. When does preconditioning help or hurt generalization? 2020.
- Peter L. Bartlett, Andrea Montanari, and Alexander Rakhlin. Deep learning: a statistical viewpoint. *arXiv:2103.09177 [math.ST]*, 2021.
- Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal of Imaging Sciences*, 2(1):183–202, 2009.
- Michael Betancourt, Michael I. Jordan, and Ashia C. Wilson. On symplectic optimization. *arXiv:1802.03653 [stat.CO]*, 2018.
- O. Bousquet and A. Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.
- Sébastien Bubeck, Yin Tat Lee, and Mohit Singh. A geometric alternative to Nesterov’s accelerated gradient descent. *arXiv:1506.08187 [math.AG]*, 2015.
- Yuansi Chen, Chi Jin, and Bin Yu. Stability and convergence trade-off of iterative optimization algorithms. *arXiv:1804.01619 [stat.ML]*, 2018.
- Alexandre Défossez, Léon Bottou, Francis Bach, and Nicolas Usunier. A simple convergence proof of Adam and Adagrad. *arXiv:2003.02395 [stat.ML]*, 2020.
- Edgar Dobriban and Stefan Wager. High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics*, 46(1):247–279, 2018.
- Timothy Dozat. Incorporating Nesterov momentum into Adam. 2016.
- Simon Du, Xiyu Zhai, Barnabas Póczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *Proceedings of the Seventh International Conference on Learning Representations*, 2019a.
- Simon S. Du, Jason D. Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks, 2019b.
- John C. Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159, 2011.
- Euhanna Ghadimi, Hamid Reza Feyzmahdavian, and Mikael Johansson. Global convergence of the Heavy-ball method for convex optimization. *arXiv:1412.7457 [math.OC]*, 2014.
- Suriya Gunasekar, Blake Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nathan Srebro. Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems 30*, 2017.

- Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.
- Moritz Hardt, Benjamin Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *Proceedings of the 33rd International Conference on Machine Learning*, 2016.
- Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan Tibshirani. Surprises in high-dimensional ridgeless linear least squares interpolation. *arXiv:1903.08560 [math.ST]*, 2019.
- Geoff Hinton, Nitesh Srivastava, and Kevin Swersky. Overview of mini-batch gradient descent. Lectures slides for University of Toronto Course CSC321, 2012.
- Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.
- Daniel Hsu, Sham M. Kakade, and Tong Zhang. Random design analysis of ridge regression. In *Proceedings of the Twenty Fifth Annual Conference on Computational Learning Theory*, 2012.
- Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems 31*, 2018.
- Prateek Jain, Sham M. Kakade, Rahul Kidambi, Praneeth Netrapalli, and Aaron Sidford. Parallelizing stochastic gradient descent for least squares regression: mini-batching, averaging, and model misspecification. *Journal of Machine Learning Research*, 18(223):1–42, 2018.
- Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the Third International Conference on Learning Representations*, 2015.
- Walid Krichene, Alexandre Bayen, and Peter L Bartlett. Accelerated mirror descent in continuous and discrete time. In *Advances in Neural Information Processing Systems 28*, 2015.
- Jason D. Lee, Max Simchowitz, Michael I. Jordan, and Benjamin Recht. Gradient descent converges to minimizers. In *Proceedings of the Twenty Ninth Annual Conference on Computational Learning Theory*, 2016.
- Laurent Lessard, Benjamin Recht, and Andrew Packard. Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM Journal on Optimization*, 26(1):57–95, 2016.
- Daniel Levy and John C. Duchi. Necessary and sufficient geometries for gradient methods. In *Advances in Neural Information Processing Systems 32*, 2019.
- Mor Shpigel Nacson, Nathan Srebro, and Daniel Soudry. Stochastic gradient descent on separable data: Exact convergence with a fixed learning rate. *arXiv:1806.01796 [stat.ML]*, 2018.
- Y. Nesterov. Smooth minimization of nonsmooth functions. *Mathematical Programming, Series A*, 103: 127–152, 2005.
- Y. Nesterov. Gradient methods for minimizing composite objective function. Technical Report 76, Center for Operations Research and Econometrics (CORE), Catholic University of Louvain (UCL), 2007.
- Yurii Nesterov. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. *Soviet Mathematics Doklady*, 27(2):372–376, 1983.
- Yurii Nesterov. On an approach to the construction of optimal methods of minimization of smooth convex functions. *Ekonomika i Matematicheskie Metody*, 24(3):509–517, 1988.
- Peter Ochs, Yunjin Chen, Thomas Brox, and Thomas Pock. iPiano: Inertial proximal algorithm for nonconvex optimization. *SIAM Journal on Imaging Sciences*, 7(2):1388–1419, 2014.
- Brendan O’Donoghue and Emmanuel Candes. Adaptive restart for accelerated gradient schemes. *Foundations of Computational Mathematics*, 15(3):715–732, 2015.
- Nicolò Pagliana and Lorenzo Rosasco. Implicit regularization of accelerated methods in Hilbert spaces. *arXiv:1905.13000 [cs.LG]*, 2019.
- Tomaso Poggio, Andrzej Banburski, and Qianli Liao. Theoretical issues in deep networks: Approximation, optimization and generalization. *arXiv:1908.09375 [cs.LG]*, 2019.
- B. T. Polyak. *Introduction to Optimization*. Optimization Software, Inc., 1987.

- Boris T Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.
- Bin Shi, Simon S. Du, Michael I. Jordan, and Weijie J. Su. Understanding the acceleration phenomenon via high-resolution differential equations. *Mathematical Programming*, pages 1–70, 2021.
- Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(18):1–57, 2018.
- Weijie Su, Stephen Boyd, and Emmanuel Candes. A differential equation for modeling Nesterov’s accelerated gradient method: Theory and insights. In *Advances in Neural Information Processing Systems 27*, pages 2510–2518, 2014.
- Arun Suggala, Adarsh Prasad, and Pradeep K Ravikumar. Connecting optimization and regularization paths. In *Advances in Neural Information Processing Systems 31*, 2018.
- Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *Proceedings of the 30th International Conference on Machine Learning*, pages 1139–1147, 2013.
- P. Tseng. On accelerated proximal gradient methods for convex-concave optimization. 2008. URL <http://www.mit.edu/~dimitrib/PTseng/papers/apgm.pdf>.
- Tomas Vaskevicius, Varun Kanade, and Patrick Rebeschini. The statistical complexity of early-stopped mirror descent. In *Advances in Neural Information Processing Systems 33*, 2020.
- Andre Wibisono, Ashia Wilson, and Michael I. Jordan. A variational perspective on accelerated methods in optimization. *Proceedings of the National Academy of Sciences*, 113(47):7351–7358, 2016.
- Ashia C. Wilson, Benjamin Recht, and Michael I. Jordan. A lyapunov analysis of momentum methods in optimization. *arXiv:1611.02635 [math.OC]*, 2016.
- Ashia C Wilson, Rebecca Roelofs, Mitchell Stern, Nati Srebro, and Benjamin Recht. The marginal value of adaptive gradient methods in machine learning. In *Advances in Neural Information Processing Systems 30*, 2017.
- Jingfeng Wu, Difan Zou, Vladimir Braverman, and Quanquan Gu. Direction matters: On the implicit bias of stochastic gradient descent with moderate learning rate. *arXiv:2011.02538 [cs.LG]*, 2020a.
- Xiaoxia Wu, Edgar Dobriban, Tongzheng Ren, Shanshan Wu, Zhiyuan Li, Suriya Gunasekar, Rachel Ward, and Qiang Liu. Implicit regularization and convergence for weight normalization. In *Advances in Neural Information Processing Systems 33*, 2020b.
- Tianbao Yang, Qihang Lin, and Zhe Li. Unified convergence analysis of stochastic momentum methods for convex and non-convex optimization. *arXiv:1604.03257 [math.OC]*, 2016.
- SK Zavriev and FV Kostyuk. Heavy-ball method in nonconvex optimization problems. *Computational Mathematics and Modeling*, 4(4):336–341, 1993.
- Fangyu Zou, Li Shen, Zequn Jie, Ju Sun, and Wei Liu. Weighted adagrad with unified momentum. *arXiv:1808.03408 [cs.LG]*, 2018.
- Fangyu Zou, Li Shen, Zequn Jie, Weizhong Zhang, and Wei Liu. A sufficient condition for convergences of Adam and RMSProp. In *Proceedings of the 29th IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

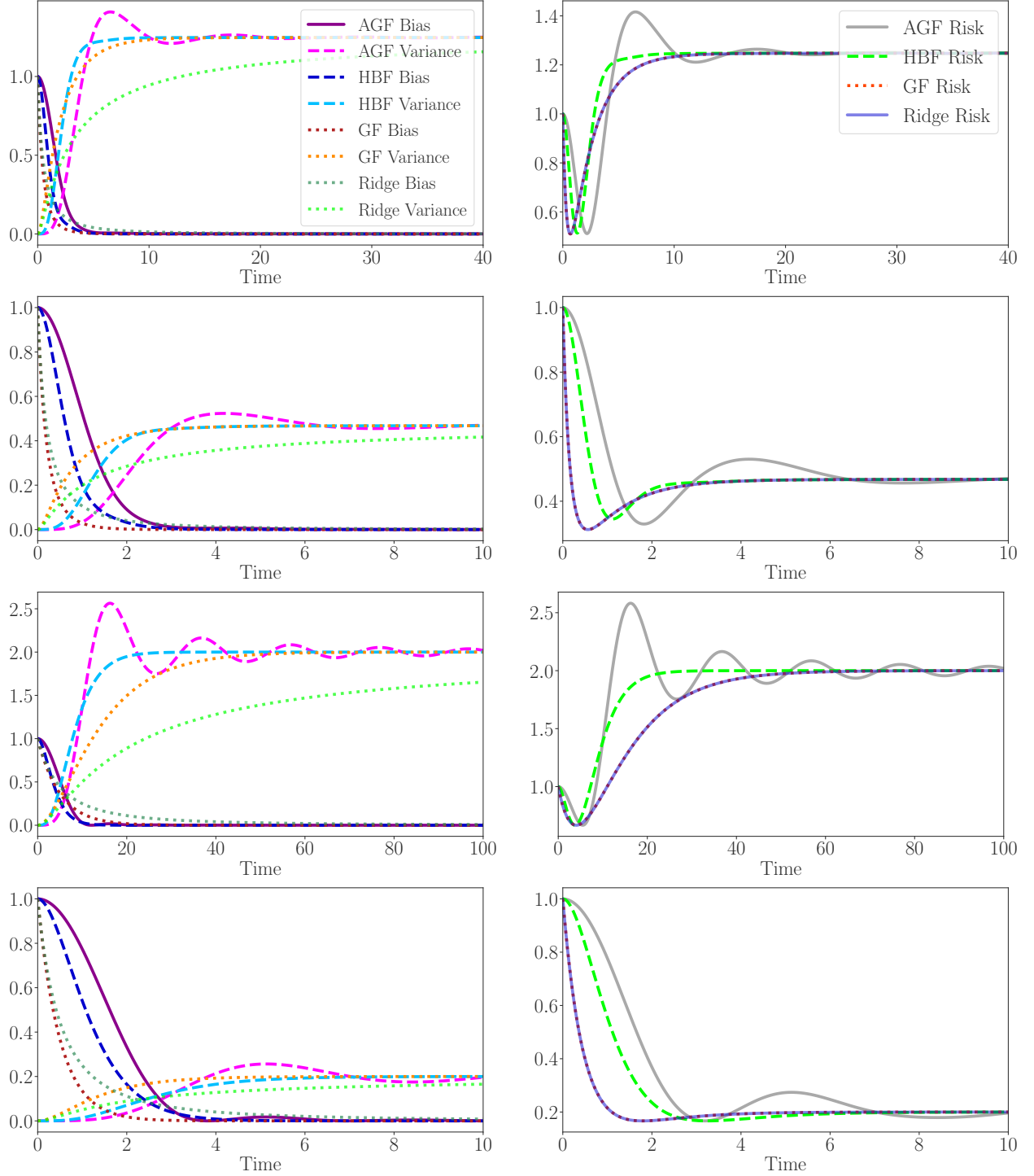


Figure 3: Bias, variance (left column), and risk (right column) for accelerated gradient flow (13), heavy ball flow (21), standard gradient flow (see Lemma 5 in Ali et al. (2019)), and ridge regression (14). The first and second row correspond to design matrices with i.i.d. standard normal entries and Student-t entries, respectively. The third and fourth rows correspond to scaled orthogonal design matrices, i.e., those satisfying $X^T X/n = s \cdot I$, for $s \in \{0.1, 1\}$, respectively.