A Continuous-Time View of Early Stopping for Least Squares

Alnur Ali
Carnegie Mellon University

J. Zico Kolter
Carnegie Mellon University

Ryan J. Tibshirani
Carnegie Mellon University

Abstract

We study the statistical properties of the iterates generated by gradient descent, applied to the fundamental problem of least squares regression. We take a continuous-time view, i.e., consider infinitesimal step sizes in gradient descent, in which case the iterates form a trajectory called *gradient flow*. Our primary focus is to compare the risk of gradient flow to that of ridge regression. Under the calibration $t = 1/\lambda$—where t is the time parameter in gradient flow, and λ the tuning parameter in ridge regression—we prove that the risk of gradient flow is no more than 1.69 times that of ridge, along the entire path (for all $t \geq 0$). This holds in finite samples with very weak assumptions on the data model (in particular, with no assumptions on the features X). We prove that the same relative risk bound holds for prediction risk, in an average sense over the underlying signal β_0. Finally, we examine limiting risk expressions (under standard Marchenko-Pastur asymptotics), and give supporting numerical experiments.

1 INTRODUCTION

Given the sizes of modern data sets, there is a growing preference towards simple estimators that have a small computational footprint and are easy to implement. Additionally, beyond efficiency and tractability considerations, there is mounting evidence that many simple and popular estimation methods perform a kind of *implicit regularization*, meaning that they appear to produce estimates exhibiting a kind of regularity, even though they do not employ an explicit regularizer.

Research interest in implicit regularization is growing, but the foundations of the idea date back at least 30 years in machine learning, where early-stopped gradient descent was found to be effective in training neural networks (Morgan and Bourlard, 1989), and at least 40 years in applied mathematics, where the same idea (here known as early-stopped Landweber iterations) was found ill-posed linear inverse problems (Strand, 1974). After a wave of research on boosting with early stopping (Buhlmann and Yu, 2003; Rosset et al., 2004; Zhang and Yu, 2005; Yao et al., 2007), more recent work focuses on the regularity properties of particular algorithms for underdetermined problems in matrix factorization, regression, and classification (Gunasekar et al., 2017; Wilson et al., 2017; Gunasekar et al., 2018). More broadly, algorithmic regularization plays a key role in training deep neural networks, via batch normalization, dropout, and other techniques.

In this paper, we focus on early stopping in gradient descent, when applied specifically to least squares regression. This is a basic problem and we are of course not the only authors to consider it; there is now a large literature on this topic (see references above, and more to come when we discuss related work shortly). However, our perspective differs from existing work in a few important ways: first, we study gradient descent in continuous-time (i.e., with infinitesimal step sizes), leading to a path of iterates known as *gradient flow*; second, we examine the regularity properties along *the entire path*, not just its convergence point (as is the focus in most of the work on implicit regularization); and third, we focus on analyzing and comparing the *risk* of gradient flow directly, which is arguably what we care about the most, in many applications.

A strength of the continuous-time perspective is that it facilitates the comparison between early stopping and ℓ_2 regularization. While the connection between these two mechanisms has been studied by many authors (and from many angles), our paper provides some of the strongest evidence for this connection to date.

Summary of Contributions. Our contributions in this paper are as follows.

- We prove that, in finite samples, under very weak

Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

assumptions on the data model (and with no assumptions on the feature matrix X), the estimation risk of gradient flow at time t is no more than 1.69 that of ridge regression at tuning parameter $\lambda = 1/t$, for all $t \geq 0$.

- We show that the same result holds for in-sample prediction risk.
- We show that the same result is also true for out-of-sample prediction risk, but now in an average (Bayes) sense, with respect to a spherical prior on the underlying signal β_0.
- For Bayes risk, under optimal tuning, our results on estimation, in-sample prediction, and out-of-sample prediction risks can all be tightened. We prove that the relative risk (measured in any of these three ways) of optimally-tuned gradient flow to optimally-tuned ridge is in between 1 and 1.22.
- We derive exact limiting formulae for the risk of gradient flow, in a Marchenko-Pastur asymptotic model where p/n (the ratio of the feature dimension to sample size) converges to a positive constant. We compare these to known limiting formulae for ridge regression.
- We support our theoretical results with numerical simulations that show the coupling between gradient flow and ridge can be extremely tight in practice (even tighter than suggested by theory).

Related Work. Various authors have made connections between ℓ_2 regularization and the iterates generated by gradient descent (when applied to different loss functions of interest): Friedman and Popescu (2004) were among the first make this explicit, and gave supporting numerical experiments, followed by Ramsay (2005), who adopted a continuous-time (gradient flow) view, as we do. Yao et al. (2007) point out that early stopped gradient descent is a spectral filter, just like ℓ_2 regularization. Subsequent work in nonparametric data models (specifically, reproducing kernel Hilbert space models), studied early-stopped gradient descent from the perspective of risk bounds, where it is shown to perform comparably to explicit ℓ_2 regularization, when each method is optimally tuned (Bauer et al., 2007; Lo Gerfo et al., 2008; Raskutti et al., 2014; Wei et al., 2017). Other works have focused on the bias-variance trade-off in early-stopped gradient boosting (Buhlmann and Yu, 2003; Zhang and Yu, 2005).

After completing this work, we became aware of the interesting recent paper by Suggala et al. (2018), who gave deterministic bounds between gradient flow and ridge regularized estimates, for problems in which the loss function is strongly convex. Their results are very different from ours: they apply to a much wider variety of problem settings (not just least squares problems), and are driven entirely by properties associated with strong convexity; our analysis, specific to least squares regression, is much more precise, and covers the important high-dimensional case (in which the strong convexity assumption is violated).

There is also a lot of related work on theory for ridge regression. Recently, Dobriban and Wager (2018) studied ridge regression (and regularized discriminant analysis) in a Marchenko-Pastur asymptotics model, deriving limiting risk expressions, and the precise form of the limiting optimal tuning parameter. Dicker (2016) gave a similar asymptotic analysis for ridge, but under a somewhat different problem setup. Hsu et al. (2012) established finite-sample concentration bounds for ridge risk. Low-dimensional theory for ridge dates back much further, see Goldenshluger and Tsybakov (2001) and others. Lastly, we point out an interesting risk inflation result in that is vaguely related to ours: Dhillon et al. (2013) showed that risk of principal components regression is at most four times that of ridge, under a natural calibration between these two estimator paths (coupling the eigenvalue threshold for the sample covariance matrix with the ridge tuning parameter).

Outline. Here is an outline for the rest of the paper. Section 2 covers preliminary material, on the problem and estimators to be considered. Section 3 gives basic results on gradient flow, and its relationship to ridge regression. Section 4 derives expressions for the estimation risk and prediction risk of gradient flow and ridge. Section 5 presents our main results on relative risk bounds (of gradient flow to ridge). Section 6 studies the limiting risk of gradient flow under standard Marchenko-Pastur asymptotics. Section 7 presents numerical examples that support our theoretical results, and Section 8 concludes with a short discussion.

2 PRELIMINARIES

2.1 Least Squares, Gradient Flow, and Ridge

Let $y \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n \times p}$ be a response vector and a matrix of predictors or features, respectively. Consider the standard (linear) least squares problem

$$\operatorname*{minimize}_{\beta \in \mathbb{R}^p} \; \frac{1}{2n} \|y - X\beta\|_2^2. \tag{1}$$

Consider gradient descent applied to (1), with a constant step size $\epsilon > 0$, and initialized at $\beta^{(0)} = 0$, which repeats the iterations

$$\beta^{(k)} = \beta^{(k-1)} + \epsilon \cdot \frac{X^T}{n} (y - X\beta^{(k-1)}), \tag{2}$$

for $k = 1, 2, 3, \ldots$. Letting $\epsilon \to 0$, we get a continuous-time ordinary differential equation

$$\dot{\beta}(t) = \frac{X^T}{n}(y - X\beta(t)), \tag{3}$$

over time $t \geq 0$, subject to an initial condition $\beta(0) = 0$. We call (3) the *gradient flow* differential equation for the least squares problem (1).

To see the connection between (2) and (3), we simply rearrange (2) to find that

$$\frac{\beta^{(k)} - \beta^{(k-1)}}{\epsilon} = \frac{X^T}{n}(y - X\beta^{(k-1)}),$$

and setting $\beta(t) = \beta^{(k)}$ at time $t = k\epsilon$, we recognize the left-hand side above as the discrete derivative of $\beta(t)$ at time t, which approaches its continuous-time derivative as $\epsilon \to 0$.

In fact, starting from the differential equation (3), we can view gradient descent (2) as one of the most basic numerical analysis techniques—the *forward Euler method*—for discretely approximating the solution (3).

Now consider the ℓ_2 regularized version of (1), called ridge regression (Hoerl and Kennard, 1976):

$$\underset{\beta \in \mathbb{R}^p}{\text{minimize}} \;\; \frac{1}{n}\|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2, \tag{4}$$

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

$$\hat{\beta}^{\text{ridge}}(\lambda) = (X^TX + n\lambda I)^{-1}X^Ty. \tag{5}$$

Though apparently unrelated, the ridge regression solution path and gradient flow path share striking similarities, and their relationship is our central focus.

2.2 The Exact Gradient Flow Solution Path

Thanks to our focus on least squares, the gradient flow differential equation in (3) is a rather special one: it is a continuous-time linear dynamical system, and has a well-known exact solution.

Lemma 1. *Fix a response y and predictor matrix X. Then the gradient flow problem (3), subject to $\beta(0) = 0$, admits the exact solution*

$$\hat{\beta}^{\text{gf}}(t) = (X^TX)^+(I - \exp(-tX^TX/n))X^Ty, \tag{6}$$

for all $t \geq 0$. Here A^+ is the Moore-Penrose generalized inverse of a matrix A, and $\exp(A) = I + A + A^2/2! + A^3/3! + \cdots$ is the matrix exponential.

Proof. This can be verified by differentiating (6) and using basic properties of the matrix exponential. □

In continuous-time, early stopping corresponds to taking the estimator $\hat{\beta}^{\text{gf}}(t)$ in (6) for any finite value of $t \geq 0$, with smaller t leading to greater regularization. We can already see that (6), like (5), applies a type of shrinkage to the least squares solution; their similarities will become more evident when we express both in spectral form, as we will do shortly in Section 3.1.

2.3 Discretization Error

In what follows, we will focus on (continuous-time) gradient flow rather than (discrete-time) gradient descent. Standard results from numerical analysis give uniform bounds between discretizations like the forward Euler method (gradient descent) and the differential equation path (gradient flow). In particular, the next result is a direct application of Theorem 212A in Butcher (2016).

Lemma 2. *For least squares, consider gradient descent (2) initialized at $\beta^{(0)} = 0$, and gradient flow (6), subject to $\beta(0) = 0$. For any step size $\epsilon < 1/s_{\max}$ where $s_{\max}$ is the largest eigenvalue of X^TX/n, and any $K \geq 1$,*

$$\max_{k=1,\ldots,k} |\beta^{(k)} - \hat{\beta}^{\text{gf}}(k\epsilon)| \leq \frac{\epsilon\|X^Ty\|_2}{2n}(\exp(K\epsilon s_{\max}) - 1).$$

The results to come can therefore be translated to the discrete-time setting, by taking a small enough ϵ and invoking Lemma 2, but we omit details for brevity.

3 BASIC COMPARISONS

3.1 Spectral Shrinkage Comparison

To compare the ridge (5) and gradient flow (6) paths, it helps to rewrite them in terms of the singular value decomposition of X. Let $X = \sqrt{n}US^{1/2}V^T$ be a singular value decomposition, so that $X^TX/n = VSV^T$ is an eigendecomposition. Then straightforward algebra brings (5), (6), on the scale of fitted values, to

$$X\hat{\beta}^{\text{ridge}}(\lambda) = US(S + \lambda I)^{-1}U^Ty, \tag{7}$$

$$X\hat{\beta}^{\text{gf}}(t) = U(I - \exp(-tS))U^Ty. \tag{8}$$

Letting s_i, $i = 1, \ldots, p$ denote the diagonal entries of S, and $u_i \in \mathbb{R}^n$, $i = 1, \ldots, p$ denote the columns of U, we see that (7), (8) are both linear smoothers (linear functions of y) of the form

$$\sum_{i=1}^p g(s_i, \kappa) \cdot u_iu_i^Ty,$$

for a spectral shrinkage map $g(\cdot, \kappa) : [0, \infty) \to [0, \infty)$ and parameter κ. This map is $g^{\text{ridge}}(s, \lambda) = s/(s + \lambda)$ for ridge, and $g^{\text{gf}}(s, t) = 1 - \exp(-ts)$ for gradient flow. We see both apply more shrinkage for smaller values of s, i.e., lower-variance directions of X^TX/n, but do so in apparently different ways.

While these shrinkage maps agree at the extreme ends (i.e., set $\lambda = 0$ and $t = \infty$, or set $\lambda = \infty$ and $t = 0$), there is no single parametrization for λ as a function of t, say $\phi(t)$, that equates $g^{\mathrm{ridge}}(\cdot, \phi(t))$ with $g^{\mathrm{gf}}(\cdot, t)$, for all $t \geq 0$. But the parametrization $\phi(t) = 1/t$ gives the two shrinkage maps grossly similar behaviors: see Figure 1 for a visualization. Moreover, as we will show later in Sections 5–7, the two shrinkage maps (under the calibration $\phi(t) = 1/t$) lead to similar risk curves for ridge and gradient flow.

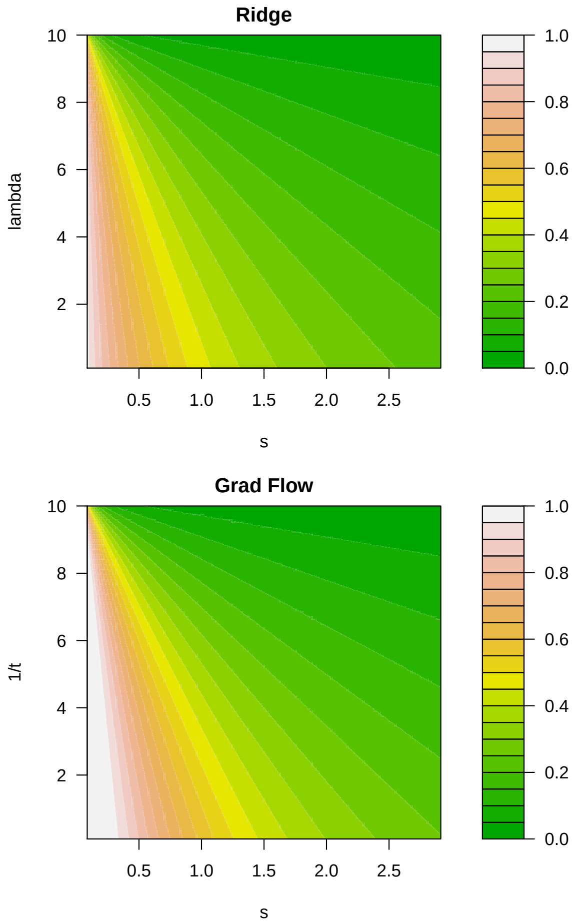

Figure 1: *Comparison of ridge and gradient flow spectral shrinkage maps, plotted as heatmaps over (s, λ) (ridge) and (s, t) (gradient flow) with the calibration $\lambda = 1/t$.*

3.2 Underlying Regularization Problems

Given our general interest in the connections between gradient descent and ridge regression, it is natural to wonder if gradient descent iterates can also be expressed as solutions to a sequence of regularized least squares problems. The following two simple lemmas certify that this is in fact the case, in both discrete- and continuous-time; their proofs may be found in the supplement.

Lemma 3. *Fix y, X, and let $X^TX/n = VSV^T$ be an eigendecomposition. Assume that we initialize $\beta^{(0)} = 0$, and we take the step size in gradient descent to satisfy $\epsilon < 1/s_{\max}$, with $s_{\max}$ denoting the largest eigenvalue of X^TX/n. Then, for each $k = 1, 2, 3, \ldots$, the iterate $\beta^{(k)}$ from step k in gradient descent (2) uniquely solves the optimization problem*

$$\underset{\beta \in \mathbb{R}^p}{\mathrm{minimize}} \ \frac{1}{n}\|y - X\beta\|_2^2 + \beta^T Q_k \beta,$$

where $Q_k = VS((I - \epsilon S)^{-k} - I)^{-1}V^T$.

Lemma 4. *Fix y, X, and let $X^TX/n = VSV^T$ be an eigendecomposition. Under the initial condition $\beta(0) = 0$, for all $t > 0$, the solution $\beta(t)$ of the gradient flow problem (3) uniquely solves the optimization problem*

$$\underset{\beta \in \mathbb{R}^p}{\mathrm{minimize}} \ \frac{1}{n}\|y - X\beta\|_2^2 + \beta^T Q_t \beta,$$

where $Q_t = VS(\exp(tS) - I)^{-1}V^T$.

Remark 1. The optimization problems that underlie gradient descent and gradient flow, in Lemmas 3 and 4, respectively, are both quadratically regularized least squares problems. In agreement with the intuition from the last subsection, we see that in both problems the regularizers penalize the lower-variance directions of X^TX/n more strongly, and this is relaxed as t or k grow. The proof of the continuous-time is nearly immediate from (8); the proof of the discrete-time result requires a bit more work. To see the link between the two results, set $t = k\epsilon$, and note that as $k \to \infty$:

$$((1 - ts/k)^{-k} - 1)^{-1} \to (\exp(ts) - 1)^{-1}.$$

4 MEASURES OF RISK

4.1 Estimation Risk

We take the feature matrix $X \in \mathbb{R}^{n \times p}$ to be fixed and arbitrary, and consider a generic response model,

$$y|\beta_0 \sim (X\beta_0, \sigma^2 I), \tag{9}$$

which we write to mean $\mathbb{E}(y|\beta_0) = X\beta_0$, $\mathrm{Cov}(y|\beta_0) = \sigma^2 I$, for an underlying coefficient vector $\beta_0 \in \mathbb{R}^p$ and error variance $\sigma^2 > 0$. We consider a spherical prior,

$$\beta_0 \sim (0, (r^2/p)I) \tag{10}$$

for some signal strength $r^2 = \mathbb{E}\|\beta_0\|_2^2 > 0$.

For an estimator $\hat{\beta}$ (i.e., measurable function of X, y), we define its estimation risk (or simply, risk) as

$$\mathrm{Risk}(\hat{\beta}; \beta_0) = \mathbb{E}\big[\|\hat{\beta} - \beta_0\|_2^2 \,\big|\, \beta_0\big].$$

We also define its Bayes risk as $\mathrm{Risk}(\hat{\beta}) = \mathbb{E}\|\hat{\beta} - \beta_0\|_2^2$.

Next we give expressions for the risk and Bayes risk of gradient flow; the derivations are straightforward and found in the supplement. We denote by s_i, $i = 1, \ldots, p$ and v_i, $i = 1, \ldots, p$ the eigenvalues and eigenvectors, respectively, of $X^T X/n$.

Lemma 5. *Under the data model (9), for any $t \geq 0$, the risk of the gradient flow estimator (6) is*

$$\mathrm{Risk}(\hat{\beta}^{\mathrm{gf}}(t); \beta_0) = \sum_{i=1}^{p} \bigg(|v_i^T \beta_0|^2 \exp(-2ts_i) + \frac{\sigma^2}{n} \frac{(1 - \exp(-ts_i))^2}{s_i} \bigg), \tag{11}$$

and under the prior (10), the Bayes risk is

$$\mathrm{Risk}(\hat{\beta}^{\mathrm{gf}}(t)) = \frac{\sigma^2}{n} \sum_{i=1}^{p} \bigg(\alpha \exp(-2ts_i) + \frac{(1 - \exp(-ts_i))^2}{s_i} \bigg), \tag{12}$$

where $\alpha = r^2 n/(\sigma^2 p)$. Here and henceforth, we take by convention $(1 - e^{-x})^2/x = 0$ when $x = 0$.

Remark 2. Compare (11) to the risk of ridge regression,

$$\mathrm{Risk}(\hat{\beta}^{\mathrm{ridge}}(\lambda); \beta_0) = \sum_{i=1}^{p} \bigg(|v_i^T \beta_0|^2 \frac{\lambda^2}{(s_i + \lambda)^2} + \frac{\sigma^2}{n} \frac{s_i}{(s_i + \lambda)^2} \bigg). \tag{13}$$

and compare (12) to the Bayes risk of ridge,

$$\mathrm{Risk}(\hat{\beta}^{\mathrm{ridge}}(\lambda)) = \frac{\sigma^2}{n} \sum_{i=1}^{p} \frac{\alpha\lambda^2 + s_i}{(s_i + \lambda)^2}, \tag{14}$$

where $\alpha = r^2 n/(\sigma^2 p)$. These ridge results follow from standard calculations, found in many other papers; for completeness, we give details in the supplement.

Remark 3. For ridge regression, the Bayes risk (14) is minimized at $\lambda^* = 1/\alpha$. There are (at least) two easy proofs of this fact. For the first, we note the Bayes risk of ridge does not depend on the distributions of $y|\beta_0$ and β_0 in (9) and (10) (just on the first two moments); in the special case that both distributions are normal, we know that $\hat{\beta}^{\mathrm{ridge}}(\lambda^*)$ is the Bayes estimator, which achieves the optimal Bayes risk (hence certainly the lowest Bayes risk over the whole ridge family). For the second proof, following Dicker (2016), we rewrite each summand in (14) as

$$\frac{\alpha\lambda^2 + s_i}{(s_i + \lambda)^2} = \frac{\alpha}{s_i + \alpha} + \frac{s(\lambda\alpha - 1)^2}{(s_i + \lambda)^2 (s_i + \alpha)},$$

and observe that this is clearly minimized at $\lambda^* = 1/\alpha$.

Remark 4. As far as we can tell, deriving the tuning parameter value t^* minimizing the gradient flow Bayes risk (12) is difficult. Nevertheless, as we will show in Section 5.3, we can still obtain interesting bounds on the optimal risk itself, $\mathrm{Risk}(\hat{\beta}^{\mathrm{gf}}(t^*))$.

4.2 Prediction Risk

We now define two predictive notions of risk. Let

$$x_0 \sim (0, \Sigma) \tag{15}$$

for a positive semidefinite matrix $\Sigma \in \mathbb{R}^{p \times p}$, and assume x_0 is independent of $y|\beta_0$. We define in-sample prediction risk and out-of-sample prediction risk (or simply, prediction risk) as, respectively,

$$\mathrm{Risk}^{\mathrm{in}}(\hat{\beta}; \beta_0) = \frac{1}{n} \mathbb{E}\big[\|X\hat{\beta} - X\beta_0\|_2^2 \,\big|\, \beta_0\big],$$
$$\mathrm{Risk}^{\mathrm{out}}(\hat{\beta}; \beta_0) = \mathbb{E}\big[(x_0^T \hat{\beta} - x_0^T \beta_0)^2 \,\big|\, \beta_0\big],$$

and their Bayes versions as, respectively, $\mathrm{Risk}^{\mathrm{in}}(\hat{\beta}) = (1/n)\mathbb{E}\|X\hat{\beta} - X\beta_0\|_2^2$, $\mathrm{Risk}^{\mathrm{out}}(\hat{\beta}) = \mathbb{E}[(x_0^T\hat{\beta} - x_0^T\beta_0)^2]$.

For space reasons, in the remainder, we will focus on out-of-sample prediction risk, and defer detailed discussion of in-sample prediction risk to the supplement. The next lemma, proved in the supplement, gives expressions for the prediction risk and Bayes prediction risk of gradient flow. We denote $\hat{\Sigma} = X^T X/n$.

Lemma 6. *Under (9), (15), the prediction risk of the gradient flow estimator (6) is*

$$\mathrm{Risk}^{\mathrm{out}}(\hat{\beta}^{\mathrm{gf}}(t); \beta_0) = \beta_0^T \exp(-t\hat{\Sigma}) \Sigma \exp(-t\hat{\Sigma}) \beta_0 + \frac{\sigma^2}{n} \mathrm{tr}\big[\hat{\Sigma}^+ (I - \exp(-t\hat{\Sigma}))^2 \Sigma\big], \tag{16}$$

and under (10), the Bayes prediction risk is

$$\mathrm{Risk}^{\mathrm{out}}(\hat{\beta}^{\mathrm{gf}}(t)) = \frac{\sigma^2}{n} \mathrm{tr}\big[\alpha \exp(-2t\hat{\Sigma})\Sigma + \hat{\Sigma}^+ (I - \exp(-t\hat{\Sigma}))^2 \Sigma\big]. \tag{17}$$

Remark 5. Compare (16) and (17) to the prediction risk and Bayes prediction risk of ridge, respectively,

$$\mathrm{Risk}^{\mathrm{out}}(\hat{\beta}^{\mathrm{ridge}}(\lambda); \beta_0) = \lambda^2 \beta_0^T (\hat{\Sigma} + \lambda I)^{-1} \Sigma (\hat{\Sigma} + \lambda I)^{-1} \beta_0 + \frac{\sigma^2}{n} \mathrm{tr}\big[\hat{\Sigma}(\hat{\Sigma} + \lambda I)^{-2} \Sigma\big], \tag{18}$$

$$\mathrm{Risk}^{\mathrm{out}}(\hat{\beta}^{\mathrm{ridge}}(\lambda)) = \frac{\sigma^2}{n} \mathrm{tr}\big[\lambda^2 \alpha (\hat{\Sigma} + \lambda I)^{-2} \Sigma + \hat{\Sigma}(\hat{\Sigma} + \lambda I)^{-2} \Sigma\big]. \tag{19}$$

These ridge results are standard, and details are given in the supplement.

Remark 6. The Bayes prediction risk of ridge (19) is again minimized at $\lambda^* = 1/\alpha$. This is not at all clear analytically, but it can be established by specializing to a normal-normal likelihood-prior pair, where (for fixed x_0) we know that $x_0^T \hat{\beta}^{\text{ridge}}(\lambda^*)$ is the Bayes estimator for the parameter $x_0^T \beta_0$ (similar to the arguments in Remark 3 for the Bayes estimation risk).

5 RELATIVE RISK BOUNDS

5.1 Relative Estimation Risk

We start with a simple but key lemma.

Lemma 7. *For all $x \geq 0$, we have (a) $e^{-x} \leq 1/(1+x)$ and (b) $1 - e^{-x} \leq 1.2985\, x/(1+x)$.*

Proof. Fact (a) can by shown via Taylor series and (b) by numerically maximizing $x \mapsto (1-e^{-x})(1+x)/x$. □

A bound on the relative risk of gradient flow to ridge, under the calibration $\lambda = 1/t$, follows immediately.

Theorem 1. *Consider the data model (9).*

(a) For all $\beta_0 \in \mathbb{R}^p$, and all $t \geq 0$, $\text{Risk}(\hat{\beta}^{\text{gf}}(t); \beta_0) \leq 1.6862 \cdot \text{Risk}(\hat{\beta}^{\text{ridge}}(1/t); \beta_0)$.

(b) The inequality in part (a) holds for the Bayes risk with respect to any prior on β_0.

(c) The results in parts (a), (b) also hold for in-sample prediction risk.

Proof. For part (a), set $\lambda = 1/t$ and compare the ith summand in (11), call it a_i, to that in (13), call it b_i. Then

$$\begin{aligned} a_i &= |v_i^T \beta_0|^2 \exp(-2ts_i) + \frac{\sigma^2}{n} \frac{(1-\exp(-ts_i))^2}{s_i} \\ &\leq |v_i^T \beta_0|^2 \frac{1}{(1+ts_i)^2} + \frac{\sigma^2}{n} 1.2985^2 \frac{t^2 s_i}{(1+ts_i)^2} \\ &\leq 1.6862 \bigg(|v_i^T \beta_0|^2 \frac{(1/t)^2}{(1/t+s_i)^2} + \frac{\sigma^2}{n} \frac{s_i}{(1/t+s_i)^2} \bigg) \\ &= 1.6862\, b_i, \end{aligned}$$

where in the second line, we used Lemma 7. Summing over $i = 1, \ldots, p$ gives the desired result.

Part (b) follows by taking an expectation on each side of the inequality in part (a). Part (c) follows similarly, with details given in the supplement. □

Remark 7. For any $t > 0$, gradient flow is in fact a unique Bayes estimator, corresponding to a normal likelihood in (9) and normal prior $\beta_0 \sim N(0, (\sigma^2/n) Q_t^{-1})$, where Q_t is as in Lemma 4. It is therefore admissible. This means the result in part (a) in the theorem (and part (b), for the same reason) cannot be true for any universal constant strictly less than 1.

5.2 Relative Prediction Risk

We extend the two simple inequalities in Lemma 7 to matrix exponentials. We use $\preceq$ to denote the Loewner ordering on positive semidefinite matrices, i.e., we use $A \preceq B$ to mean that $B - A$ is positive semidefinite.

Lemma 8. *For all $X \succeq 0$, we have (a) $\exp(-2X) \preceq (I+X)^{-2}$ and (b) $X^+(I - \exp(-X))^2 \preceq 1.6862\, X(I+X)^{-2}$.*

Proof. All matrices in question are simultaneously diagonalizable, so the claims reduce to ones about eigenvalues, i.e., reduce to checking that $e^{-2x} \leq 1/(1+x)^2$ and $(1-e^{-x})^2/x \leq 1.6862\, x/(1+x)^2$, for $x \geq 0$, and these follow by manipulating the facts in Lemma 7. □

With just a bit more work, we can bound the relative Bayes prediction risk of gradient flow to ridge, again under the calibration $\lambda = 1/t$.

Theorem 2. *Consider the data model (9), prior (10), and (out-of-sample) feature distribution (15). For all $t \geq 0$, $\text{Risk}^{\text{out}}(\hat{\beta}^{\text{gf}}(t)) \leq 1.6862 \cdot \text{Risk}^{\text{out}}(\hat{\beta}^{\text{ridge}}(1/t))$.*

Proof. Consider the matrices inside the traces in (17) and (19). Applying Lemma 8, we have

$$\begin{aligned} &\alpha \exp(-2t\hat{\Sigma}) + \hat{\Sigma}^+(I - \exp(-t\hat{\Sigma}))^2 \\ &\quad \preceq \alpha(I + t\hat{\Sigma})^{-2} + 1.6862\, t^2 \hat{\Sigma}(I + t\hat{\Sigma})^{-2} \\ &\quad \preceq 1.6862 \Big(\alpha(1/t)^2 (I/t + \hat{\Sigma})^{-2} + \hat{\Sigma}(I/t + \hat{\Sigma})^{-2} \Big). \end{aligned}$$

Let A, B be the matrices on the first and last lines in the above display, respectively. As $A \preceq B$ and $\Sigma \succeq 0$, we have $\text{tr}(A\Sigma) \leq \text{tr}(B\Sigma)$, completing the proof. □

Remark 8. The Bayes perspective here is critical; the proof breaks down for prediction risk, at an arbitrary fixed β_0, and it is not clear to us whether the result is true for prediction risk in general.

5.3 Relative Risks at Optima

We present one more helpful inequality, and defer its proof to the supplement (it is more technical than the proofs of Lemmas 7 and 8, but still straightforward).

Lemma 9. *For all $X \succeq 0$, it holds that $\exp(-2X) + X^+(I - \exp(-X))^2 \preceq 1.2147\, (I+X)^{-1}$.*

We now have the following result, on the relative Bayes risk (and Bayes prediction risk), of gradient descent to ridge regression, when both are optimally tuned.

Theorem 3. *Consider the data model (9), prior (10), and (out-of-sample) feature distribution (15).*

(a) It holds that

$$1 \leq \frac{\inf_{t \geq 0} \text{Risk}(\hat{\beta}^{\text{gf}}(t))}{\inf_{\lambda \geq 0} \text{Risk}(\hat{\beta}^{\text{ridge}}(\lambda))} \leq 1.2147.$$

(b) *The same result as in part (a) holds for both in-sample and out-of-sample prediction risk.*

Proof. For part (a), recall from Remark 3 that the optimal ridge tuning parameter is $\lambda^* = 1/\alpha$ and further, in the special case of a normal-normal likelihood-prior pair, we know that $\hat\beta^{\mathrm{ridge}}(\lambda^*)$ is the Bayes estimator so the Bayes risk of $\hat\beta^{\mathrm{gf}}(t)$, for any $t \geq 0$, must be at least that of $\hat\beta^{\mathrm{ridge}}(\lambda^*)$. But because these Bayes risks (12), (14) do not depend on the form of likelihood and prior (only on their first two moments), we know that the same must be true in general, which proves the lower bound on the risk ratio. For the upper bound, we take $t = \alpha$, and compare the ith summand in (12), call it a_i, to that in (14), call it b_i. We have

$$\begin{aligned} a_i &= \alpha \exp(-2\alpha s_i) + \frac{(1-\exp(-\alpha s_i))^2}{s_i} \\ &\leq 1.2147 \frac{\alpha}{1+\alpha s_i} = 1.2147\, b_i, \end{aligned}$$

where in the second line, we applied Lemma 9 (to the case of scalar X). Summing over $i = 1, \ldots, p$ gives the desired result.

Parts (b) follows similarly, with details in the supplement. □

6 ASYMPTOTIC RISK ANALYSIS

6.1 Marchenko-Pastur Asymptotics

Notice the Bayes risk for gradient flow (12) and ridge regression (14) depend only on the predictor matrix X via the eigenvalues of the (uncentered) sample covariance $\hat\Sigma = X^T X/n$. Random matrix theory gives us a precise understanding of the behavior of these eigenvalues, in large samples. The following assumptions are standard ones in random matrix theory (e.g., Bai and Silverstein 2010). Given a symmetric matrix $A \in \mathbb{R}^{p \times p}$, recall that its *spectral distribution* is defined as $F_A(x) = (1/p)\sum_{i=1}^p \mathbb{1}(\lambda_i(A) \leq x)$, where $\lambda_i(A)$, $i = 1, \ldots, p$ are the eigenvalues of A, and $\mathbb{1}(\cdot)$ denotes the 0-1 indicator function.

Assumption A1. The predictor matrix satisfies $X = Z\Sigma^{1/2}$, for a random matrix $Z \in \mathbb{R}^{n \times p}$ of i.i.d. entries with zero mean and unit variance, and a deterministic positive semidefinite covariance $\Sigma \in \mathbb{R}^{p \times p}$.

Assumption A2. The sample size n and dimension p both diverge, i.e., $n, p \to \infty$, with $p/n \to \gamma \in (0, \infty)$.

Assumption A3. The spectral measure F_Σ of the predictor covariance Σ converges weakly as $n, p \to \infty$ to some limiting spectral measure H.

Under the above assumptions, the seminal Marchenko-Pastur theorem describes the weak limit of the spectral measure $F_{\hat\Sigma}$ of the sample covariance $\hat\Sigma$.

Theorem 4 (Marchenko and Pastur 1967; Silverstein 1995; Bai and Silverstein 2010)**.** *Assuming A1–A3, almost surely, the spectral measure $F_{\hat\Sigma}$ of $\hat\Sigma$ converges weakly to a law $F_{H,\gamma}$, called the* empirical spectral distribution*, that depends only on H, γ.*

Remark 9. In general, a closed form for the empirical spectral distribution $F_{H,\gamma}$ is not known, except in very special cases (e.g., when $\Sigma = I$ for all n, p). However, numerical methods for approximating $F_{H,\gamma}$ have been proposed (see Dobriban 2015 and references therein).

6.2 Limiting Gradient Flow Risk

The limiting Bayes risk of gradient flow is now immediate from the representation in (12).

Theorem 5. *Assume A1–A3, as well as a data model (9) and prior (10). Then as $n, p \to \infty$ with $p/n \to \gamma \in (0, \infty)$, for each $t \geq 0$, the Bayes risk (12) of gradient flow converges almost surely to*

$$\sigma^2\gamma \int \left[\alpha_0 \exp(-2ts) + \frac{(1-\exp(-ts))^2}{s}\right] dF_{H,\gamma}(s), \tag{20}$$

where $\alpha_0 = r^2/(\sigma^2\gamma)$, and $F_{H,\gamma}$ is the empirical spectral distribution from Theorem 4.

Proof. Note that we can rewrite the Bayes risk in (12) as $(\sigma^2 p)/n[\int \alpha h_1(s)\, dF_{\hat\Sigma}(s) + \int h_2(s)\, dF_{\hat\Sigma}(s)]$, where we let $h_1(s) = \exp(-2ts)$, $h_2(s) = (1-\exp(-ts))^2/s$. Weak convergence of $F_{\hat\Sigma}$ to $F_{H,\gamma}$, from Theorem 4, implies $\int h(s)\, dF_{\hat\Sigma}(s) \to \int h(s)\, dF_{H,\gamma}(s)$ for all bounded, continuous functions h, which proves the result. □

A similar result is available for the limiting Bayes in-sample prediction risk, given in the supplement. Studying the the limiting Bayes (out-of-sample) prediction risk is much more challenging, as (17) is not simply a function of eigenvalues of $\hat\Sigma$. The proof of the next result, deferred to the supplement, relies on a key fact on the Laplace transform of the map $x \mapsto \exp(xA)$, and the asymptotic limit of a certain trace functional involving $\hat\Sigma, \Sigma$, from Ledoit and Peche (2011).

Theorem 6. *Under the conditions of Theorem 5, also assume $\mathbb{E}(Z_{ij}^{12}) \leq C_1$, $\|\Sigma\|_2 \leq C_2$, for all n, p and constants $C_1, C_2 > 0$. For each $t \geq 0$, the Bayes prediction risk (17) of gradient flow converges almost surely to*

$$\sigma^2\gamma\left[\alpha_0 f(2t) + 2\int_0^t (f(u) - f(2u))\, du\right], \tag{21}$$

where f is the inverse Laplace transform of the function

$$x \mapsto \frac{1}{\gamma}\left(\frac{1}{1-\gamma+\gamma x m(F_{H,\gamma})(-x)} - 1\right),$$

and $m(F_{H,\gamma})$ is the Stieltjes transform of $F_{H,\gamma}$ (defined precisely in the supplement).

An interesting feature of the results (20), (21) is that they are asymptotically *exact* (no hidden constants). Analogous results for ridge (by direct arguments, and Dobriban and Wager 2018, respectively) are compared in the supplement, for space reasons.

7 NUMERICAL EXAMPLES

We give numerical evidence for our theoretical results: both our relative risk bounds in Section 5, and our asymptotic risk expressions in Section 6. We generated features via $X = \Sigma^{1/2}Z$, for a matrix Z with i.i.d. entries from a distribution G (with mean zero and unit variance), for three choices of G: standard Gaussian, Student t with 3 degrees of freedom, and Bernoulli with probability 0.5 (the last two distributions were standardized). We took Σ to have all diagonal entries equal to 1 and all off-diagonals equal to $\rho = 0$ (i.e., $\Sigma = I$), or $\rho = 0.5$. For the problem dimensions, we considered $n = 1000$, $p = 500$ and $n = 500$, $p = 1000$. For both gradient flow and ridge, we used a range of 200 tuning parameters equally spaced on the log scale from 2^{-10} to 2^{10}. Lastly, we set $\sigma^2 = r^2 = 1$, where σ^2 is the noise variance in (9) and r^2 is the prior radius in (10). For each configuration of G, Σ, n, p, we computed the Bayes risk and Bayes prediction risk gradient flow and ridge, as in (12), (14), (17), (19). For $\Sigma = I$, the empirical spectral distribution from Theorem 4 has a closed form, and so we computed the limiting Bayes risk for gradient flow (20) via numerical integration (and similarly for ridge, details in the supplement).

Figure 2 shows the results for Gaussian features, $\Sigma = I$, $n = 500$, and $p = 1000$; the supplement shows results for all other cases (the results are grossly similar). The top plot shows the risk curves when calibrated according to $\lambda = 1/t$ (as per our theory). Here we see fairly strong agreement between the two risk curves, especially around their minimums; the maximum ratio of gradient flow to ridge risks is 1.2164 over the entire path (cf. the upper bound of 1.6862 from Theorem 1), and the ratio of the minimums is 1.0036 (cf. the upper bound of 1.2147 from Theorem 3). The bottom plot shows the risks when parametrized by the ℓ_2 norms of the underlying estimators. We see remarkable agreement over the whole path, with a maximum ratio of 1.0050. Moreover, in both plots, we can see that the finite-sample (dotted lines) and asymptotic risk curves (solid lines) are identical, meaning that the convergence in Theorem 5 is very rapid (and similarly for ridge).

8 DISCUSSION

In this work, we studied gradient flow (i.e., gradient descent with infinitesimal step sizes) for least squares, and pointed out a number of connections to ridge regression. We showed that, under minimal assumptions on the data model, and using a calibration $t = 1/\lambda$—where t denotes the time parameter in gradient flow, and λ the tuning parameter in ridge—the risk of gradient flow is no more than 1.69 times that of ridge, for all $t \geq 0$. We also showed that the same holds for prediction risk, in an average (Bayes) sense, with respect to any spherical prior. Though we did not pursue this, it is clear that these risk couplings could be used to port risk results from the literature on ridge regression (e.g., Hsu et al. 2012; Raskutti et al. 2014; Dicker 2016; Dobriban and Wager 2018, etc.) to gradient flow.

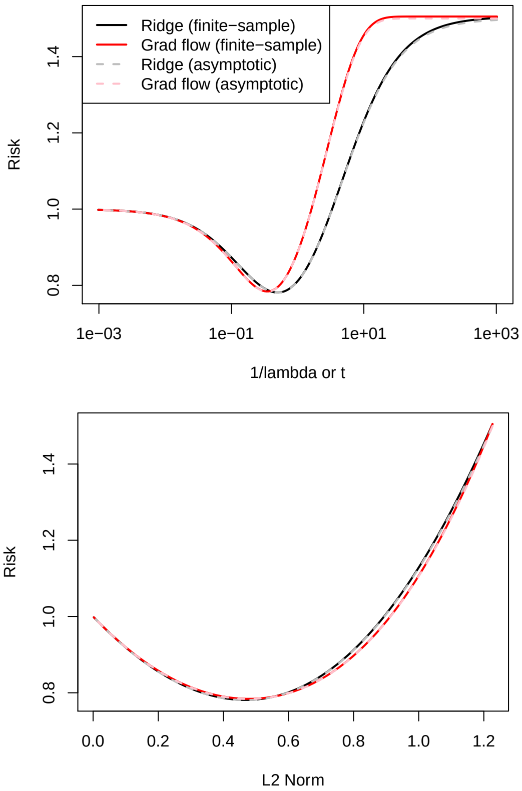

Figure 2: *Comparison of Bayes risks for gradient flow and ridge, with Gaussian features, $\Sigma = I$, $n = 500$, $p = 1000$.*

Our numerical experiments revealed that calibrating the risk curves by the underlying ℓ_2 norms of the estimators results in a much tighter coupling; developing theory to explain this phenomenon is an important challenge left to future work. Other interesting directions are to analyze the risk of a continuum version of stochastic gradient descent, or to study gradient flow beyond least squares, e.g., for logistic regression.

Acknolwedgements. We thank Veeranjaneyulu Sadhanala, whose insights led us to completely revamp the main results in our paper. AA was supported by DoE CSGF no. DE-FG02-97ER25308. ZK was supported by DARPA YFA no. N66001-17-1-4036.

References

Zhidong Bai and Jack Silverstein. *Spectral Analysis of Large Dimensional Random Matrices*. Springer, 2010.

Frank Bauer, Sergei Pereverzev, and Lorenzo Rosasco. On regularization algorithms in learning theory. *Journal of Complexity*, 23(1):52–72, 2007.

Peter Buhlmann and Bin Yu. Boosting with the ℓ_2 loss. *Journal of the American Statistical Association*, 98 (462):324–339, 2003.

John C. Butcher. *Numerical Methods for Ordinary Differential Equations*. John Wiley & Sons, 2016.

Paramveer Dhillon, Dean Foster, Sham Kakade, and Lyle Ungar. A risk comparison of ordinary least squares vs ridge regression. *The Journal of Machine Learning Research*, 14:1505–1511, 2013.

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

Edgar Dobriban. Efficient computation of limit spectra of sample covariance matrices. *Random Matrices: Theory and Applications*, 4(4):1550019, 2015.

Edgar Dobriban and Stefan Wager. High-dimensional asymptotics of prediction: ridge regression and classification. *Annals of Statistics*, 46(1):247–279, 2018.

Jerome Friedman and Bogdan Popescu. Gradient directed regularization. Working paper, 2004. URL http://www-stat.stanford.edu/~jhf/ftp/pathlite.pdf.

Alexander Goldenshluger and Alexandre Tsybakov. Adaptive prediction and estimation in linear regression with infinitely many parameters. *Annals of Statistics*, 29(6):1601–1619, 2001.

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

Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning*, 2018.

Arthur E. Hoerl and Robert W. Kennard. Ridge regression iterative estimation of the biasing parameter. *Communications in Statistics: Theory and Methods*, 5(1):77–88, 1976.

Daniel Hsu, Sham Kakade, and Tong Zhang. Random design analysis of ridge regression. In *Annual Conference on Learning Theory*, pages 9.1–9.24, 2012.

Olivier Ledoit and Sandrine Peche. Eigenvectors of some large sample covariance matrix ensembles. *Probability Theory and Related Fields*, 151(1–2):233–264, 2011.

Laura Lo Gerfo, Lorenzo Rosasco, Francesca Odone, Ernesto De Vito, and Alessandro Verri. Spectral algorithms for supervised learning. *Neural Computation*, 20(7):1873–1897, 2008.

Vladimir Alexandrovich Marchenko and Leonid Andreevich Pastur. Distribution of eigenvalues for some sets of random matrices. *Matematicheskii Sbornik*, 114(4):507–536, 1967.

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

James Ramsay. Parameter flows. Working paper, 2005.

Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Early stopping and nonparametric regression: an optimal data-dependent stopping rule. *Journal of Machine Learning Research*, 15:335–366, 2014.

Saharon Rosset, Ji Zhu, and Trevor Hastie. Boosting as a regularized path to a maximum margin classifier. *Journal of Machine Learning Research*, 5:941–973, 2004.

Jack Silverstein. Strong convergence of the empirical distribution of eigenvalues of large dimensional random matrices. *Journal of Multivariate Analysis*, 55 (2):331–339, 1995.

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

Arun S. Suggala, Adarsh Prasad, and Pradeep Ravikumar. Connecting optimization and regularization paths. In *Advances in Neural Information Processing Systems*, 2018.

Yuting Wei, Fanny Yang, and Martin J. Wainwright. Early stopping for kernel boosting algorithms: a general analysis with localized complexities. In *Advances in Neural Information Processing Systems*, 2017.

Ashia 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*, 2017.

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

Tong Zhang and Bin Yu. Boosting with early stopping: convergence and consistency. *Annals of Statistics*, 33 (4):1538–1579, 2005.