A Continuous-Time View of Early Stopping for Least Squares Regression

Alnur Ali$^1$ J. Zico Kolter$^1$ Ryan J. Tibshirani$^{1,2}$

$^1$Machine Learning Department, Carnegie Mellon University
$^2$Department of Statistics, 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. In a random matrix theory setup, which allows the number of samples $n$ and features $p$ to diverge in such a way that $p/n \to \gamma \in (0, \infty)$, we derive and analyze an asymptotic risk expression for gradient flow. In particular, we compare the asymptotic risk profile of gradient flow to that of ridge regression. When the feature covariance is spherical, we show that the optimal asymptotic gradient flow risk is between 1 and 1.25 times the optimal asymptotic ridge risk. Further, we derive a calibration between the two risk curves under which the asymptotic gradient flow risk no more than 2.25 times the asymptotic ridge risk, at all points along the path. We present a number of other results illustrating the connections between gradient flow and $\ell_2$ regularization, and numerical experiments that support our theory.

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 Landweber iteration) was found to help in ill-posed linear inverse problems (Strand, 1974). After a wave of work 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, weight decay, and other techniques (Goodfellow et al., 2016).

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 two ways. First, we study gradient descent in continuous-time, i.e., we study gradient descent with infinitesimal step sizes, leading to a path of iterates called gradient flow; and second, we focus on the regularity
properties along the entire path, not just its convergence point (as is the focus in most work on implicit regularization).

A strength of the continuous-time perspective is that it facilitates the comparison between early stopping and \( \ell_2 \) regularization. The folklore in statistics is that these two are somehow closely connected. Our paper provides, as far as we can tell, some of the most direct and rigorous evidence for this connection to date.

**Summary of Contributions.** Our specific contributions are as follows.

- We derive exact expressions for the finite-sample as well as asymptotic (out-of-sample, predictive) risk of gradient flow, at any time \( t \geq 0 \). Our setup is a standard one from random matrix theory, in which both the number of samples \( n \) and features \( p \) diverge in such a way that \( p/n \to \gamma \in (0, \infty) \). (Note that this setup covers the important high-dimensional regime, where \( \gamma \gg 1 \).)

- For a spherical feature covariance, we show that the optimal limiting risk of gradient flow (meaning optimally-tuned) is never better than the optimal limiting risk of ridge regression.

- For a spherical feature covariance, and for small enough signal sizes, we prove the optimal limiting risk of gradient flow is at most 1.25 that of ridge.

- Under the same assumptions, we show there is a mapping between the tuning parameters for gradient flow and ridge that makes the limiting risk of gradient flow at most 2.25 the limiting risk of ridge, at all points along the path.

- We backup our theoretical findings with numerical simulations that show the coupling between gradient flow and ridge to be extremely tight in practice (even tighter than suggested by theory).

**Related Work.** Various authors have made connections between \( \ell_2 \) regularization and the iterates generated by gradient descent, when applied to various loss functions of interest in modern statistical applications; e.g., Friedman and Popescu (2004) make this connection by giving supporting numerical experiments, whereas Yao et al. (2007) point out that early stopping and \( \ell_2 \) regularization are similar but generically distinct statistical mechanisms. In nonparametric data models (specifically, when the regression function is assumed to be in a reproducing kernel Hilbert space), early-stopped gradient descent has also been studied from the perspective of risk bounds, where it is shown to perform comparably to explicit \( \ell_2 \) regularization using an optimally-chosen tuning parameter (Yao et al., 2007; 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).

As far as the broader literature on implicit regularization goes, beyond the work surveyed in the introduction above, we specifically point out the work of Gunasekar et al. (2017) as relevant to ours, which studies infinitesimal gradient descent applied to a matrix factorization problem, and Gunasekar et al. (2018), which studies gradient descent for least squares regression and other problems. But in both cases (and in most current work on implicit regularization), the focus is on the regularity properties of the convergence point of the algorithm, not the whole algorithm path.

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 the same asymptotic model as we do, giving limiting risk expressions that we build upon for our analysis. Dicker (2016) gave a similar analysis for ridge, but considered a different asymptotic setup. 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 this paper. Section 2 covers preliminary material, on the problem and estimators to be considered. In Section 3, we give some basic results on gradient flow and its relationship to ridge regression. Section 4 derives the limiting risk curve of the gradient flow path, and Section 5 compares it to that from ridge. In Section, 6 we present numerical examples that support our theory. Lastly, in Section 7, we conclude and give ideas for future work.

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 least squares problem

\[
\min_{\beta \in \mathbb{R}^p} \frac{1}{2n} \| y - X\beta \|_2^2. \tag{1}
\]

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 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); standard results in numerical analysis control the approximation error of the forward Euler (gradient descent) path (e.g., Griffiths and Higham (2010)).

Now consider the \( \ell_2 \) regularized version of (1), called ridge regression (Hoerl and Kennard, 1976):

\[
\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \| y - X\beta \|_2^2 + \lambda \| \beta \|_2^2, \tag{4}
\]

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

\[
\hat{\beta}_{\text{ridge}}(\lambda) = (X^T X + n\lambda I)^{-1} X^T y. \tag{5}
\]

Though apparently unrelated, the ridge regression solution path and gradient flow path share striking similarities, and their relationship is the central focus of this paper.
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. For completeness, this is stated in the next lemma.

**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

\[
\hat{\beta}_{\text{gf}}(t) = (X^T X)^+ (I - \exp(-tX^T X/n))X^T y,
\]

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

**Proof.** This is a standard result. It can be verified by differentiating (6) and using properties of the matrix exponential. \( \square \)

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 more become clear when we express both paths in terms of the spectral decomposition of \( X \), as we will do shortly in Section 3.1.

2.3 Marchenko-Pastur Asymptotics

Much of the theory for ridge regression (and discriminant analysis, and principal component analysis) is driven by the behavior of eigenvalues of the (uncentered) sample feature covariance matrix \( \hat{\Sigma} = X^T X/n \). Random matrix theory gives us a (remarkably) precise understanding of the spread 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 \( A \in \mathbb{R}^{p \times p} \), its spectral distribution is defined as

\[
F_A(x) = \frac{1}{p} \sum_{i=1}^p \mathbb{1}(\lambda_i(A) \leq x),
\]

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 predictor dimension \( p \) both diverge, i.e., \( n, p \to \infty \), in a such a way that \( p/n \to \gamma \in (0, \infty) \).

**Assumption A3.** The spectral measure \( F_\Sigma \) of the predictor covariance \( \Sigma \) converges weakly as \( n,p \to \infty \) to some limiting spectral measure \( F \).

Under these assumptions, the seminal Marchenko-Pastur theorem describes the weak limit of the spectral measure \( F_\Sigma \) of the sample covariance \( \hat{\Sigma} = X^T X/n \).

**Theorem 1** ((Marchenko and Pastur, 1967; Silverstein, 1995; Bai and Silverstein, 2010)). Assuming A1–A3, almost surely, the spectral measure \( F_\Sigma \) of \( \Sigma \) converges weakly to a law \( F_{\Sigma, \gamma} \), called the empirical spectral distribution, that depends only on \( \Sigma, \gamma \).

When \( \Sigma = I \), the empirical spectral distribution \( F_\gamma \) is known as the Marchenko-Pastur (MP) law and has a closed form. For \( \gamma \leq 1 \), its density is

\[
\frac{dF_\gamma(s)}{ds} = \frac{1}{2\pi\gamma s} \sqrt{(b-s)(s-a)},
\]

and is supported on \([a, b]\), where \( a = (1 - \sqrt{\gamma})^2 \) and \( b = (1 + \sqrt{\gamma})^2 \). For \( \gamma > 1 \), the MP law \( F_\gamma \) additionally has a point mass at zero of probability \( 1 - 1/\gamma \).
Recently, Dobriban and Wager (2018) give a broad and insightful treatment of the risk of ridge regression using Theorem 1 and other results. As we will see in Section 4.3, we can use MP asymptotics to study the risk of gradient flow as well.

Finally, we define an important concept that we will make use of later. For a measure $G$, its Stieltjes transform $m_G$, evaluated at any $z \in \mathbb{C} \setminus \text{supp}(G)$, is

$$m_G(z) = \int_0^{\infty} \frac{1}{u-z} dG(u).$$

Conveniently, the MP law has a closed-form Stieltjes transform (e.g., Lemma 3.11 in Bai and Silverstein (2010)), for $z > 0$:

$$m_{F_\gamma}(-z) = -\frac{(1-\gamma+z) + \sqrt{(1-\gamma+z)^2 + 4\gamma z}}{2\gamma z}. \quad (8)$$

3 Basic Comparisons and Results

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 $X^TX/n = VS^TV$ is an eigendecomposition. Then straightforward algebraic manipulations bring (5), (6), on the scale of fitted values, to

$$X\widehat{\beta}_{\text{ridge}}(\lambda) = VS(S + \lambda I)^{-1}V^Ty, \quad (9)$$

$$X\widehat{\beta}_{\text{gf}}(t) = V(I - \exp(-tS))V^Ty. \quad (10)$$

Letting $s_i, i = 1, \ldots, p$ denote the diagonal entries of $S$, and $v_i \in \mathbb{R}^n, i = 1, \ldots, p$ denote the columns of $V$, we see that (9), (10) are both linear smoothers (linear functions of $y$) of the form

$$\sum_{i=1}^{p} g(s_i, \kappa) \cdot v_i v_i^T y,$$

for a spectral shrinkage map $g(\cdot, \kappa) : [0, \infty) \rightarrow [0, \infty)$ and parameter $\kappa$. This map is $g_{\text{ridge}}^r(s, \lambda) = s/(s + \lambda)$ for ridge, and $g_{\text{gf}}(s, t) = 1 - \exp(-ts)$ for gradient flow. We see both methods apply more shrinkage for smaller $s$, 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 $\lambda$ as a function of $t$, say $\phi(t)$, that equates $g_{\text{ridge}}^r(\cdot, \phi(t))$ with $g_{\text{gf}}^r(\cdot, t)$, for all $t \geq 0$. But the parametrization $\phi(t) = 1/t$ does give the two shrinkage maps grossly similar behaviors: see Figure 1 for a visualization. More importantly, as we will show in Section 5.2, the two shrinkage maps (after proper calibration) lead to similar risk curves for ridge and gradient flow, under fairly generic assumptions.
Figure 1: Comparison of ridge and gradient flow spectral shrinkage maps, for $s \in [(1 - \sqrt{\gamma})^2, (1 + \sqrt{\gamma})^2]$, $\lambda \in [0, 10]$, and $t = 1/\lambda$, when we use an aspect ratio $\gamma = 0.5$. Their behaviors are roughly similar, especially for large $\lambda$. 
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 2.** Fix $y, X$, and let $X^T X/n = VSV^T$ be an eigendecomposition. Assume that we initialize $\beta^{(0)} = 0$, and we take the step size in gradient descent to satisfy $\epsilon < 1/s_{\text{max}}$, with $s_{\text{max}}$ denoting the largest eigenvalue of $X^T X/n$. Then, for each $k = 1, 2, 3, \ldots$, the iterate $\beta^{(k)}$ from step $k$ in gradient descent (2) uniquely solves the optimization problem

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

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

**Lemma 3.** Fix $y, X$, and let $X^T X/n = VSV^T$ 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

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

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

**Remark 1.** The optimization problems that underlie gradient descent and gradient flow, in Lemmas 2 and 3, 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 $X^T X/n$ more strongly, and this is relaxed as $t$ or $k$ grow. The proof of the continuous-time is nearly immediate from (10); the proof of the discrete-time result, perhaps unexpectedly, requires a bit more care. 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 Bias, Variance, and Risk

4.1 Data Model and Setup

The data generating model that we consider consists of three components: a distribution on features, one on errors, and a prior on coefficients, which we denote by $P_x, P_\epsilon, P_\beta$, respectively. We assume the following.

- For $x \sim P_x$, we can write $x = \Sigma^{1/2} z$ for a random vector $z$ whose entries are i.i.d. from a distribution with mean zero and unit variance, and some determinisitic positive semidefinite matrix $\Sigma \in \mathbb{R}^{p \times p}$.
- For $\epsilon \sim P_\epsilon$, we have $\mathbb{E}(\epsilon) = 0, \mathbb{E}(\epsilon^2) = \sigma^2$.
- For $\beta \sim P_\beta$, we have $\mathbb{E}(\beta) = 0, \text{Cov}(\beta) = (r^2/p)I$.

The training data is generated as in

$$\beta_0 \sim P_\beta,$$

$$(x_i, \epsilon_i) \sim P_x \times P_\epsilon, \quad i = 1, \ldots, n,$$

and $y_i = x_i^T \beta_0 + \epsilon_i, \quad i = 1, \ldots, n.$
where all random draws are independent. We collect the responses in a vector $y \in \mathbb{R}^n$, and the features in a matrix $X \in \mathbb{R}^{n \times p}$ (with rows $x_i \in \mathbb{R}^p$, $i = 1, \ldots, n$).

We also consider a test point $x_0 \sim P_x$, independent of the training data. For an estimator $\hat{\beta}$ (i.e., measurable function of the training data $X,y$), we define its out-of-sample, predictive risk as

$$\text{Risk}_{n,p}(\hat{\beta}) = \mathbb{E}_{\beta_0,x_0,\epsilon}( (x_0^T \hat{\beta} - x_0^T \beta_0)^2 | X).$$

Note that our risk definition is conditional on $X$, but integrated over $\beta_0$, making it a Bayes risk of sorts. The subscripts on the expectation above emphasize what is being integrated over, but will be dropped henceforth. Lastly, we have the usual bias-variance decomposition

$$\text{Risk}_{n,p}(\hat{\beta}) = \text{Bias}_{n,p}(\hat{\beta})^2 + \text{Var}_{n,p}(\hat{\beta}),$$

where

$$\text{Bias}_{n,p}(\hat{\beta}) = \mathbb{E}[x_0^T \hat{\beta} - x_0^T \beta_0 | X], \quad \text{Var}_{n,p}(\hat{\beta}) = \mathbb{E}[x_0^T \hat{\beta} - \mathbb{E}(x_0^T \hat{\beta} | X) | X]^2.$$

### 4.2 Finite-Sample Risk Expressions

Next we give expressions for bias and variance of gradient flow; the proof is straightforward and found in the supplement.

**Lemma 4.** Assume the conditions in Section 4.1. For any $t \geq 0$, the bias and variance of the gradient flow estimator (6) are

$$\text{Bias}_{n,p}^2(\hat{\beta}_{gf}(t)) = \frac{\sigma^2}{n} \alpha_{n,p} \text{tr} \left[ \exp(-2t\hat{\Sigma})\Sigma \right],$$

$$\text{Var}_{n,p}(\hat{\beta}_{gf}(t)) = \frac{\sigma^2}{n} \text{tr} \left[ \Sigma(I - \exp(-t\hat{\Sigma})\hat{\Sigma}^+(I - \exp(-t\hat{\Sigma})) \right],$$

where $\alpha_{n,p} = r^2n/(\sigma^2p)$, and $\hat{\Sigma} = X^TX/n$.

**Remark 2.** Compare (11) and (12) to the bias and variance of ridge regression, which are also easily derived (and found in, e.g., Dobriban and Wager (2018)):

$$\text{Bias}_{n,p}^2(\hat{\beta}_{ridge}(\lambda)) = \frac{\sigma^2}{n} \alpha_{n,p} \lambda^2 \text{tr} \left[ (\hat{\Sigma} + \lambda I)^{-2}\Sigma \right],$$

$$\text{Var}_{n,p}(\hat{\beta}_{ridge}(\lambda)) = \frac{\sigma^2}{n} \text{tr} \left[ \Sigma(\hat{\Sigma} + \lambda I)^{-1}\hat{\Sigma}(\hat{\Sigma} + \lambda I)^{-1} \right].$$

Leveraging recent results in random matrix theory, Dobriban and Wager (2018) characterize the limiting behavior of the bias and variance of ridge regression, for general $\Sigma$. Their arguments rely heavily on a convergence result for a particular functional of $\hat{\Sigma}$ that is found in Ledoit and Peche (2011); while it seems these may used to characterize the limiting bias (11) for gradient flow for arbitrary $\Sigma$, the limiting variance (12) seems currently out of reach.

**Remark 3.** In the case $\Sigma = I$, the bias and variance expressions for gradient flow simplify considerably:

$$\text{Bias}_{n,p}^2(\hat{\beta}_{gf}(t)) = \frac{\sigma^2}{n} \alpha_{n,p} \sum_{i=1}^{p} \exp(-2ts_i),$$

$$\text{Var}_{n,p}(\hat{\beta}_{gf}(t)) = \frac{\sigma^2}{n} \sum_{i=1}^{r} \frac{(1 - \exp(-ts_i))^2}{s_i},$$
where \( r = \text{rank}(\Sigma) \), and \( s_1, \ldots, s_r > 0 \) are its nonzero eigenvalues. Those for ridge also simplify:

\[
\text{Bias}_{n,p}^2(\hat{\beta}_{\text{ridge}}(\lambda)) = \frac{\sigma^2}{n} \alpha_{n,p} \sum_{i=1}^{p} \frac{\lambda^2}{(s_i + \lambda)^2},
\]

(17)

\[
\text{Var}_{n,p}(\hat{\beta}_{\text{ridge}}(\lambda)) = \frac{\sigma^2}{n} \sum_{i=1}^{p} \frac{s_i}{(s_i + \lambda)^2}.
\]

(18)

We will focus on \( \Sigma = I \) in our asymptotic theory, next.

### 4.3 Limiting Risk Expressions

Working under the MP asymptotics model, and specializing to \( \Sigma = I \), our next result is an asymptotically exact expression for the risk of gradient flow.

**Lemma 5.** Assume the conditions in Section 4.1, with \( \Sigma = I \). As \( n, p \to \infty \), such that \( p/n \to \gamma \in (0, \infty) \), the risk \( \text{Risk}_{n,p}(\hat{\beta}^{gf}_{\gamma}(t)) \) of gradient flow converges, for each \( t \geq 0 \), almost surely, to

\[
\text{Risk}_{\gamma}(\hat{\beta}^{gf}_{\gamma}(t)) = \sigma^2 \gamma \int \left[ \exp(-2ts)\alpha + \frac{(1 - \exp(-ts)^2)}{s} \right] dF_{\gamma}(s),
\]

(19)

where \( \alpha = r^2/(\sigma^2\gamma) \), and \( F_{\gamma} \) is the MP law.

**Proof.** Weak convergence of \( F_{\hat{\Sigma}} \) to \( F_{\gamma} \), given by Theorem 1, is equivalent to \( \int h(s) dF_{\hat{\Sigma}}(s) \to \int h(s) dF_{\gamma}(s) \) for all bounded, continuous functions \( h \). Applying this logic to the bias (15) and variance (16) separately, then combining the integrals, proves (19). \( \square \)

**Remark 4.** Compare (19) to the analogous limiting risk for ridge, proved similarly from (17), (18) (see also in Theorem 2.1 of Dobriban and Wager (2018)):

\[
\text{Risk}_{\gamma}(\hat{\beta}_{\text{ridge}}(\lambda)) = \sigma^2 \gamma \int \frac{\lambda^2 \alpha + s}{(s + \lambda)^2} dF_{\gamma}(s).
\]

(20)

In Dobriban and Wager (2018) (see also Dicker (2016)) it is shown that the optimal tuning parameter for the limiting ridge risk (20) is given by \( \lambda^* = 1/\alpha \); plugging this into (20) and simplifying gives

\[
\text{Risk}_{\gamma}(\hat{\beta}_{\text{ridge}}(\lambda^*)) = \sigma^2 \gamma \cdot m_{F_{\gamma}}(-1/\alpha),
\]

(21)

where recall \( m_{F_{\gamma}} \) is the Stieltjes transform of the MP law, as in (8).

Somewhat remarkably, Dobriban and Wager (2018) show that \( \lambda^* = 1/\alpha \) is still the optimal ridge tuning parameter for a general feature covariance matrix \( \Sigma \), using an elegant equicontinuity argument. Note in this case, the optimal limiting ridge risk itself is no longer closed-form.

**Remark 5.** As far as we can tell, deriving the tuning parameter \( t^* \) that minimizes the gradient flow limiting risk (19) is difficult. The strategies from Dobriban and Wager (2018); Dicker (2016) for analyzing optimal tuning in ridge regression do not carry over to gradient flow for various reasons. Nevertheless, as we will show in Section 5.1, we can still obtain fairly tight (lower and upper) bounds on the optimal limiting gradient flow risk itself \( \text{Risk}_{\gamma}(\hat{\beta}^{gf}_{\gamma}(t^*)) \).

**Remark 6.** The optimal limiting ridge risk in (21) is expressed in terms of the Stieltjes transform of the MP law. It turns out that the limiting gradient flow risk (19), along its entire path \( t \geq 0 \), can be expressed in terms of another common transform on probability measures, namely the Laplace transform, denoted \( \mathcal{L}(G)(z) = \int_0^\infty \exp(-zu) dG(u) \) for a probability measure \( G \). In the supplement, we show that for all \( t \geq 0 \),

\[
\text{Risk}_{\gamma}(\hat{\beta}^{gf}_{\gamma}(t)) = \alpha \cdot \mathcal{L}(F_{\gamma})(2t) + \int_0^t \mathcal{L}(F_{\gamma})(z) dz - \int_t^{2t} \mathcal{L}(F_{\gamma})(z) dz.
\]

(22)
4.4 Convergence Rates

It is reasonable to ask how quickly the finite-sample risks approach their limiting values. The following lemma provides answers, indicating that the rates of convergence are nearly $n^{-1/2}$ for both gradient flow and ridge regression. In both cases, our analysis follows the strategy used in Theorem 1 in Dicker (2016) (who in turn invoke convergence results from Bai et al. (2003)), and the proof can be found in the supplement.

Lemma 6. Assume the conditions in Section 4.1, with $\Sigma = I$. Also assume, for $x \sim P_x$ written as $x = \Sigma^{1/2} z$, that the entries of $z$ have a finite 8th moment. Then as $n, p \to \infty$, such that $p/n \to \gamma \in (0, \infty)$, for any $\lambda > 1/(2\alpha)$, we have

$$\left| \frac{\gamma}{p/n} \cdot \text{Risk}_{n,p}(\hat{\beta}_{\text{ridge}}(\lambda)) - \text{Risk}_{n,p}(\hat{\beta}_{\text{ridge}}(\lambda)) \right| = o_P(n^{-w}),$$

where $w = 2/5 - \eta$, for any $\eta > 0$, if $\gamma \neq 1$, and $w = 1/8$ if $\gamma = 1$. Also, for any $t < 2\alpha$,

$$\left| \frac{\gamma}{p/n} \cdot \text{Risk}_{n,p}(\hat{\beta}_{gf}(t)) - \text{Risk}_{n,p}(\hat{\beta}_{gf}(t)) \right| = o_P(n^{-w}),$$

for $w$ defined in the same way.

5 Risk Comparisons

5.1 Optimal Risk Comparison

Here we relate the optimal limiting gradient flow risk $\text{Risk}_{\gamma}(\hat{\beta}_{gf}(t^*))$, where $t^* \geq 0$ denotes the (unknown) optimal tuning parameter, to that of ridge regression. We start with a lower bound.

Theorem 2. Assume the conditions from Section 4.1, with $\Sigma = I$, and $n, p \to \infty$, $p/n \to \gamma \in (0, \infty)$. Then

$$\text{Risk}_{\gamma}(\hat{\beta}_{gf}(t)) \geq \text{Risk}_{\gamma}(\hat{\beta}_{ridge}(\lambda^*)).$$

Proof. From (19), observe

$$\min_{t \geq 0} \frac{1}{\sigma^2 \gamma} \text{Risk}_{\gamma}(\hat{\beta}_{gf}(t))$$

$$= \min_{t \geq 0} \int \left[ \exp(-2ts)\alpha + \frac{(1 - \exp(-ts))^2}{s} \right] dF_{\gamma}(s)$$

$$\geq \min_{t \geq 0} \int \exp(-2ts)\alpha + \frac{(1 - \exp(-ts))^2}{s} dF_{\gamma}(s).$$

Now let us inspect the inner bracketed term in the last line, call it $f_s(t)$. Differentiating gives $f'_s(t) = 2\exp(-ts)[1 - \exp(-ts)(\alpha s + 1)]$, whose only root occurs when $\exp(-ts) = 1/(\alpha s + 1)$, and after some simplification, this gives $\min_{t \geq 0} f_s(t) = 1/(s + 1/\alpha)$. Plugging this into the integrand in the last line of the above display, then recalling the definition of the Stieltjes transform and (21), proves the result. □

Remark 7. Though its proof is simple, the result in Theorem 2 is resolute: optimal gradient flow cannot have better risk than optimal ridge regression, for any signal-to-noise ratio ($\alpha$), and a wide range of feature distributions ($P_x$ with $\Sigma = I$). This result is made less mysterious by noting that when the error and prior distributions ($P_\epsilon, P_\beta$) are both Gaussian, ridge regression with
\( \lambda^*_n, p = r^2 n / (\sigma^2 p) \) is the Bayes estimator, and no other estimator—gradient flow or otherwise—can achieve a smaller Bayes risk. As the limiting risks do not depend on the error and prior distributions (only through \( \sigma^2, r^2 \)), we should expect the same to be true in the limit. In fact, this can be made rigorous using \( \lambda^*_n, p \to \lambda^* \), and equicontinuity arguments, to serve as an alternative proof of Theorem 2.

We now turn an upper bound. The proof, deferred to the supplement, is intricate; it involves Taylor expanding the gradient flow and ridge risks, keeping infinitely many terms, and carefully coupling the results.

**Theorem 3.** Assume the same conditions as in Theorem 2. Assume further that \((1 + \sqrt{\gamma})^2 < 1/\alpha\). Then

\[
\text{Risk}_\gamma(\hat{\beta}_{df}(t^*)) \leq 1.25 \cdot \text{Risk}_\gamma(\hat{\beta}_{ridge}(\lambda^*)).
\]

**Remark 8.** The restriction that \((1 + \sqrt{\gamma})^2 < 1/\alpha\) in Theorem 3 is equivalent to \(r^2 < \sigma^2 \gamma / (1 + \sqrt{\gamma})^2\), a restriction on the signal strength \(r\). Though it is undesirable to have any restriction, an upper bound restriction on \(r\) at least allows for the signal to be arbitrarily weak, a case of significant interest. Moreover, it appears that the condition \((1 + \sqrt{\gamma})^2 < 1/\alpha\) is unavoidable (unless we completely change proof strategies); this is needed for the Stieltjes transform of the MP law to converge to a series expansion involving its moments.

**Remark 9.** The constant 1.25 in Theorem 3 is likely not as small as possible; we sought a clean-looking bound, but it could probably be made tighter by refining the coupling used in the proof.

### 5.2 Full Risk Path Comparison

We extend the results in the last subsection to cover the entire gradient flow path. The proof constructs an implicit parametrization by inverting the ridge risk curve on either side of its minimum, and is found in the supplement.

**Theorem 4.** Assume the same conditions as in Theorem 2. Assume also that \((1 + \sqrt{\gamma})^2 < 1/\alpha\). Then there is a map \(\phi : [0, \infty] \to [0, \infty]\), that satisfies \(\phi(0) = \infty, \phi(\infty) = 0\), and

\[
\text{Risk}_\gamma(\hat{\beta}_{df}(t)) \leq 2.25 \cdot \text{Risk}_\gamma(\hat{\beta}_{ridge}(\phi(t))) \quad \text{for all } t \geq 0.
\]

The map \(\phi\) is continuous on either side of \(t = \alpha\), and has a single jump discontinuity at \(t = \alpha\).

**Remark 10.** The factor of 2.25 in Theorem 4 is certainly not the smallest possible constant, and it most likely could be made tighter.

### 6 Numerical Examples

We present numerical evidence backing up the theoretical predictions from Section 5. We consider three different feature distributions in our experiments. First of all, we let the entries of \(X\) be i.i.d. standard normal, satisfying the conditions given in Section 2.3. However, we also allow for the entries of \(X\) to follow a Bernoulli distribution with mean parameter 0.4, as well as a Student-t distribution with 2 degrees of freedom; in both cases, the data does not obey the given conditions (by design), as the variance for the Student-t distribution with 2 degrees of freedom is undefined (the results for \(> 2\) degrees of freedom are actually similar, and therefore omitted). Furthermore, we consider a spherical feature covariance, but also allow for a general \(\Sigma\), by setting the correlation between each pair of features to 1/2. As far as the error distribution and prior on the coefficients go, we fix the error variance \(\sigma^2 = 1\) and signal strength \(r = 1\) throughout our experiments.

We focus here on a high-dimensional setup, where \(\gamma = 2\), by setting \(p = 1000\) and \(n = 500\) (the results for \(\gamma \leq 1\) are similar, and deferred to the supplement). Note that our choices of \(\sigma^2, \gamma\)
imply that $r$ must be less than 0.5857, in order for the conditions of Theorem 3 to hold, whereas we fixed $r = 1$ in our experiments. In order to generate risk curves for gradient flow and ridge regression, we evaluate each of the risk expressions over a discretized grid of tuning parameters $t, \lambda \in \{2^{-20}, \ldots, 2^{20}\}$.

We present the limiting as well as finite-sample risk curves of gradient flow and ridge regression, for all three feature distributions and when $\Sigma = I$, across the top row of Figure 2; the finite-sample risks for the case where $\Sigma \neq I$ are presented across the bottom row. To calibrate the risk curves, we plot them in terms of an estimator’s achieved $\ell_2$-norm (the results are similar for other measures of model complexity). Qualitatively, it is clear that the two risk curves are virtually indistinguishable across all the experimental setups, even those not satisfying the regularity conditions required by some of our results. Table 1 makes this more precise: it shows the ratio of the optimal risks, as well as the largest ratio of the respective risks; in all cases, both the ratios are nearly 1 (suggesting that there is indeed some slack in our bounds). Additionally, the asymptotics here appear to “kick in” rather fast, as the limiting and finite-sample curves essentially coincide.

![Figure 2: Comparison of risks for gradient flow and ridge, across a range of tuning parameters and feature distributions. We consider here a high-dimensional setup, where $\gamma = 2$, by setting $p = 1000$ and $n = 500$ (the results for $\gamma \leq 1$ are similar, and deferred to the supplement). Top row: the limiting as well as finite-sample risks, for a spherical feature covariance. Bottom row: the finite-sample risks, for a general feature covariance matrix.](image)

7 Discussion

We studied the statistical properties of the gradient flow (gradient descent with infinitesimal step sizes), applied to least squares regression, pointing out a number of connections to ridge regression; one of our results showed that, when the feature covariance is spherical, the optimal limiting gradient flow risk is between 1 and 1.25 times the optimal ridge risk. Although much of our theory considers
Table 1: Quantitative comparison of the risk curves from Figure 2. Below, for each feature distribution, “Min Risks” denotes the ratio of the optimal gradient flow to ridge risk, whereas “Max Risks” denotes the largest ratio of the risks along the two curves.

| Feature Distribution       | Min Risks | Max Risks |
|---------------------------|-----------|-----------|
| Gaussian (asymptotics)    | 1.0063    | 1.0552    |
| Gaussian (finite-sample)  | 1.0099    | 1.0786    |
| Student-t (asymptotics)   | 1.0063    | 1.0559    |
| Student-t (finite-sample) | 1.0098    | 1.0782    |
| Bernoulli (asymptotics)   | 1.0063    | 1.0561    |
| Bernoulli (finite-sample) | 1.0105    | 1.0825    |

a spherical feature covariance, we gave numerical results supporting the conclusion that gradient flow and ridge have similar risk profiles, even when the feature covariance was not isotropic.

Developing theory for a general feature covariance matrix is an interesting challenge that we are pursuing as part of future work. Another direction that seems worthwhile is to generalize the arguments given here to the situation beyond the squared loss.

A Supplementary Material

A.1 Proof of Lemma 2

Let $X = \sqrt{n}US^{1/2}V^T$, where $U \in \mathbb{R}^{n \times r}$, $V \in \mathbb{R}^{p \times r}$, and $S \in \mathbb{R}^{r \times r}$, so that $\hat{\Sigma} = VSV^T$. Here, $U, V$ have orthonormal columns, $S$ is diagonal, and $r = \text{rank}(\hat{\Sigma})$. We obtain for the (discrete-time) gradient descent iteration that

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

for $k = 1, 2, 3, \ldots$. Rotating by $V^T$, we get

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

i.e.,

$$
\tilde{\beta}^{(k)} = \tilde{\beta}^{(k-1)} - \epsilon S \tilde{\beta}^{(k-1)} + \tilde{y}.
$$

Unraveling the rotated gradient descent recursion in the preceding display, we find that

$$
\tilde{\beta}^{(k)} = (I - \epsilon S)^k \tilde{\beta}^{(0)} + \sum_{j=0}^{k-1} (I - \epsilon S)^j \tilde{y}.
$$

Furthermore applying the assumption that the initial point $\beta^{(0)} = 0$ yields

$$
\tilde{\beta}^{(k)} = \sum_{j=0}^{k-1} (I - \epsilon S)^j \tilde{y} = (\epsilon S)^{-1} (I - (I - \epsilon S)^k) \tilde{y}.
$$
with the second equality following after a short inductive argument. Now rotating back to the original space gives

$$V\tilde{\beta}^{(k)} = \beta^{(k)}$$

$$= V(\epsilon S)^{-1}(I - (I - \epsilon S)^k)\tilde{y},$$  \hspace{1cm} (23) with the first equality following because $VV^T = P_{\text{row}(X)}$ and $\beta^{(k)} \in \text{row}(X)$, for all $k$, where $P_A$ denotes the projection map onto the subspace $A$, and $\text{row}(X)$ denotes the row space of $X$.

A straightforward application of the matrix inversion lemma shows that

$$\langle I - (I - \epsilon S)^k \rangle^{-1} = -(I + (I - \epsilon S)^{-k} - I)^{-1};$$  \hspace{1cm} (24)

therefore, using (24), we get for (23) that

$$\tilde{y} = (\epsilon S)(I - (I - \epsilon S)^k)^{-1}V^T\beta^{(k)}$$

$$= (\epsilon S)(I + (I - \epsilon S)^{-k} - I)^{-1}V^T\beta^{(k)}.$$  \hspace{1cm} (24)

It now follows, from the preceding display, that the $k$th gradient descent iterate $\beta^{(k)}$ is given by the solution to the regularized least squares problem

$$\text{minimize}_{\beta \in \mathbb{R}^p} \frac{1}{2}\| (\epsilon S)^{-1/2}y - (\epsilon S)^{1/2}V^T\beta \|^2 + \frac{1}{2}\| (\epsilon S)^{1/2}(I - \epsilon S)^{-k} - I)^{-1/2}V^T\beta \|^2.$$  \hspace{1cm} (25)

Multiplying the loss through by $U$ and recalling that $U$ is an isometry,

$$\frac{1}{2}\| U(\epsilon S)^{-1/2}y - U(\epsilon S)^{1/2}V^T\beta \|^2 = \frac{\epsilon}{2n} \cdot \| y - X\beta \|^2,$$

meaning that the problem (25) is equivalent to

$$\text{minimize}_{\beta \in \mathbb{R}^p} \frac{\epsilon}{2n} \cdot \| y - X\beta \|^2 + \frac{\epsilon}{2} \left\| S^{1/2}(I - \epsilon S)^{-k} - I)^{-1/2}V^T\beta \|^2.$$  \hspace{1cm} (25)

The claim follows. \hfill $\square$

### A.2 Proof of Lemma 3

Recall that Lemma 1 gives the gradient flow solution at time $t$, in (6). Compare this to the solution in Lemma 3, which is

$$(X^TX + nQ_t)^{-1}X^Ty.$$  \hspace{1cm} (6)

To equate these two, we see that we must have

$$(X^TX)^+(I - \exp(-tX^TX/n)) = (X^TX + nQ_t)^{-1},$$

i.e., writing $X^TX/n = VS$ as an eigendecomposition of $X^TX/n$,

$$VS^+(I - \exp(-tS))V^T = (VS^T + Q_t)^{-1}.$$  \hspace{1cm} (6)

Inverting both sides and rearranging, we find that

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

which is as claimed in the lemma. \hfill $\square$
B Proof of Lemma 4

First of all, for any estimator \( \hat{\beta} \), observe that

\[
\text{Risk}_{n,p}(\hat{\beta}) = E_{\hat{\beta}_0, x_0, \epsilon} \left[ (x_0^T \hat{\beta} - x_0^T \beta_0)^2 | X \right] = E_{\hat{\beta}_0, \epsilon} \left[ \| \hat{\beta} - \beta_0 \|_2^2 | X \right],
\]

where \( \| z \|^2_A = z^T A z \). It follows that

\[
E_{\hat{\beta}_0, \epsilon} \left[ \| \hat{\beta} - \beta_0 \|_2^2 | X \right] = E_{\hat{\beta}_0} \left[ \| E_{\epsilon} [\hat{\beta} | X, \beta_0] - \beta_0 \|_2^2 | X \right] + \text{tr} \left( \Sigma \text{Cov}_{\hat{\beta}_0, \epsilon}(\hat{\beta} | X) \right).
\]

Now let \( X = \sqrt{n} U S^{1/2} V^T \), where \( U \in \mathbb{R}^{n \times n}, V \in \mathbb{R}^{p \times p}, \) and \( S \in \mathbb{R}^{n \times p}, \) so that \( \Sigma = V S V^T \). Here, \( U, V \) are orthogonal, and the off-diagonal entries of \( S \) are set to zero. Recalling (6), we have for any \( t \geq 0 \) that

\[
E_{\epsilon}[\hat{\beta}^{gf}(t) | X, \beta_0] = V (I - \exp(-tS)) V^T \beta_0.
\]

Therefore,

\[
E_{\epsilon}[\hat{\beta}^{gf}(t) | X, \beta_0] - \beta_0 = -\exp(-t\Sigma)\beta_0,
\]

and so

\[
\| E_{\epsilon}[\hat{\beta}^{gf}(t) | X, \beta_0] - \beta_0 \|^2_\Sigma = \text{tr} \left( \exp(-t\Sigma) \Sigma \exp(-t\Sigma) \beta_0 \beta_0^T \right),
\]

implying that

\[
E_{\hat{\beta}_0} \left[ \| E_{\epsilon}[\hat{\beta} | X, \beta_0] - \beta_0 \|^2_\Sigma | X \right] = \frac{t^2}{p} \text{tr} \left( \exp(-t\Sigma) \Sigma \exp(-t\Sigma) \right) = \frac{t^2}{p} \text{tr} \left( \exp(-2t\Sigma) \right),
\]

where we used the linearity of expectation along with properties of the trace. Finally, when \( \Sigma = I \), we obtain

\[
E_{\hat{\beta}_0, x_0} \text{Bias}_{n,p}^2(\hat{\beta}^{gf}(t)) = \frac{t^2}{p} \sum_{i=1}^p \exp(-2ts_i),
\]

where \( s_i, i = 1, \ldots, p \), are the diagonal elements of \( S \). The claim for the bias follows.

Turning now to the variance, we similarly have

\[
\text{tr} \left( \Sigma \text{Cov}_{\hat{\beta}_0, \epsilon}(\hat{\beta}^{gf}(t) | X) \right) = \text{tr} \left( \Sigma (X^T X)^+ (I - \exp(-t\Sigma)) X^T \text{Cov}_{\hat{\beta}_0, \epsilon}(\epsilon | X) X (I - \exp(-t\Sigma)) (X^T X)^+ \right)
\]

\[
= \sigma^2 \text{tr} \left( \Sigma (X^T X)^+ (I - \exp(-t\Sigma)) X^T X (I - \exp(-t\Sigma)) (X^T X)^+ \right)
\]

\[
= \sigma^2 \text{tr} \left( \Sigma \Sigma^+ (I - \exp(-t\Sigma)) \Sigma (I - \exp(-t\Sigma)) \Sigma^+ \right)
\]

\[
= \frac{\sigma^2}{n} \text{tr} \left( \Sigma (I - \exp(-t\Sigma)) \Sigma^+ (I - \exp(-t\Sigma)) \right),
\]

with the fourth line following using properties of the pseudo-inverse. Again, when \( \Sigma = I \), we obtain

\[
\text{tr} \left( \Sigma \text{Cov}_{\hat{\beta}_0, \epsilon}(\hat{\beta}^{gf}(t) | X) \right) = \frac{\sigma^2}{n} \text{tr} \left( S^+ (I - \exp(-tS))^2 \right)
\]

\[
= \frac{\sigma^2}{n} \sum_{i=1}^r \frac{(1 - \exp(-ts_i))^2}{s_i},
\]

where \( r = \text{rank}(\Sigma) \), as claimed (\( r \) here should not be confused for the signal strength).
Alternatively, the variance may be written as
\[
\text{tr}\left( \Sigma \text{Cov}_{\beta_0, \epsilon} (\hat{\beta}^{\text{gf}}(t) \mid X) \right) = \frac{\sigma^2}{n} \sum_{i=1}^{p} \left\{ \frac{(1-\exp(-ts_i))^2}{s_i}, \quad s_i > 0 \right. \\
\left. \frac{1}{t^2 s_i}, \quad s_i = 0 \right\},
\]
which is a smooth function in \((s, t)\).

\[\square\]

**B.1 Proof of Equation 22**

First of all, note that for the limiting bias of gradient flow,
\[
\text{Bias}_\gamma (\hat{\beta}^{\text{gf}}(t)) = \alpha \cdot \mathcal{L}(F_\gamma)(2t). \tag{26}
\]

Therefore, consider the limiting variance of gradient flow, i.e.,
\[
\text{Var}_\gamma (\hat{\beta}^{\text{gf}}(t)) = \int \left[ \frac{1}{s} + \frac{\exp(-2ts)}{s} - \frac{2\exp(-ts)}{s} \right] dF_\gamma(s). \tag{27}
\]

Let \( G \) be a probability measure, which we assume for simplicity is supported on \( \mathbb{R}^+ \). Then, for any \( t \geq 0 \), we have
\[
\frac{\partial}{\partial t} \left[ \int \frac{\exp(-ts)}{s} dG(s) \right] = -\mathcal{L}(G)(t),
\]
where the integration above is done over the support of \( G \); note that interchanging the order of the derivative and integral above is justified because the integrand, as well as its partial derivative with respect to \( t \), are both continuous in \((s, t)\) as long as \( s > 0 \). The fundamental theorem of calculus, then, implies
\[
\int \frac{\exp(-ts)}{s} dG(s) = \int \frac{1}{s} dG(s) - \int_0^t \mathcal{L}(G)(z)dz. \tag{28}
\]

Now applying the result in (28) to (27), for \( \gamma < 1 \), we obtain
\[
\text{Var}_\gamma (\hat{\beta}^{\text{gf}}(t)) = \int \frac{1}{s} dF_\gamma(s) + \left[ \int \frac{1}{s} dF_\gamma(s) - \int_0^{2t} \mathcal{L}(F_\gamma)(z)dz \right] - 2 \left[ \int \frac{1}{s} dF_\gamma(s) - \int_0^t \mathcal{L}(F_\gamma)(z)dz \right]
\]
\[
= 2 \int_0^t \mathcal{L}(F_\gamma)(z)dz - \int_0^{2t} \mathcal{L}(F_\gamma)(z)dz
\]
\[
= 2 \int_0^t \mathcal{L}(F_\gamma)(z)dz - \int_0^t \mathcal{L}(F_\gamma)(z)dz - \int_t^{2t} \mathcal{L}(F_\gamma)(z)dz
\]
\[
= \int_0^t \mathcal{L}(F_\gamma)(z)dz - \int_0^{2t} \mathcal{L}(F_\gamma)(z)dz,
\]
which yields the claim.

It is not too hard to see that the preceding argument still works, even when \( \gamma \geq 1 \): in this case, we may redefine the integrand in (27) to take the value \( t^2s \) when \( s = 0 \), so that the integrand, as well as its partial derivative with respect to \( t \), are now continuous in \((s, t)\), even when \( s = 0 \).
B.2 Proof of Lemma 6

For any ridge regression tuning parameter $\lambda > 0$, observe that

\[
\frac{\gamma}{p/n} \cdot \text{Risk}_{n,p}(\hat{\beta}_{\text{ridge}}(\lambda)) = \sigma^2 \gamma \sum_{i=1}^{p} \frac{\lambda^2 \alpha + s_i}{(s_i + \lambda)^2} = \sigma^2 \gamma \int_0^\infty \frac{\lambda^2 \alpha + s}{(s + \lambda)^2} d\hat{F}_{\gamma,n,p}(s)
\]

where $\hat{F}_{\gamma,n,p}$ denotes the empirical distribution function of the singular values of $\hat{\Sigma}$. We point out a slight abuse of notation in the above, where we use $\alpha$ instead of $\alpha_{n,p}$ in expanding the definition of $\text{Risk}_{n,p}(\hat{\beta}_{\text{ridge}}(\lambda))$; we will employ the same abuse of notation with $\text{Risk}_{n,p}(\hat{\beta}_{\text{ridge}}(\lambda))$, later on in the proof. Continuing, an application of integration by parts (Billingsley, 2008, page 237) to the previous display yields

\[
\int_0^\infty \frac{\lambda^2 \alpha + s}{(s + \lambda)^2} d\hat{F}_{\gamma,n,p}(s) = \left. \frac{\lambda^2 \alpha + s}{(s + \lambda)^2} \hat{F}_{\gamma,n,p}(s) \right|_0^\infty - \int_0^\infty \frac{\lambda - 2\alpha \lambda^2 - s}{(s + \lambda)^3} \hat{F}_{\gamma,n,p}(s) ds
\]

\[
= -\alpha \hat{F}_{\gamma,n,p}(0) - \int_0^\infty \frac{\lambda - 2\alpha \lambda^2 - s}{(s + \lambda)^3} \hat{F}_{\gamma,n,p}(s) ds,
\]

meaning that

\[
\frac{\gamma}{p/n} \cdot \text{Risk}_{n,p}(\hat{\beta}_{\text{ridge}}(\lambda)) = -\sigma^2 \gamma \left[ \alpha \hat{F}_{\gamma,n,p}(0) + \int_0^\infty \frac{\lambda - 2\alpha \lambda^2 - s}{(s + \lambda)^3} \hat{F}_{\gamma,n,p}(s) ds \right]. \tag{29}
\]

Using the same line of argument, we get for $\text{Risk}_\gamma(\hat{\beta}_{\text{ridge}}(\lambda))$ that

\[
\text{Risk}_\gamma(\hat{\beta}_{\text{ridge}}(\lambda)) = -\sigma^2 \gamma \left[ \alpha F_\gamma(0) + \int_0^\infty \frac{\lambda - 2\alpha \lambda^2 - s}{(s + \lambda)^3} F_\gamma(s) ds \right]. \tag{30}
\]

Combining (30), (29), and using the triangle inequality as well as Holder’s inequality, we have

\[
\frac{\gamma}{p/n} \cdot \text{Risk}_{n,p}(\hat{\beta}_{\text{ridge}}(\lambda)) - \text{Risk}_\gamma(\hat{\beta}_{\text{ridge}}(\lambda))
\]

\[
= \sigma^2 \gamma \left[ \alpha (F_\gamma(0) - \hat{F}_{\gamma,n,p}(0)) + \int_0^\infty \frac{\lambda - 2\alpha \lambda^2 - s}{(s + \lambda)^3} (F_\gamma(s) - \hat{F}_{\gamma,n,p}(s)) ds \right]
\]

\[
\leq \sigma^2 \gamma \sup_{s \in \mathbb{R}_+} |F_\gamma(s) - \hat{F}_{\gamma,n,p}(s)| + \sigma^2 \gamma \int_0^\infty \left| \frac{\lambda - 2\alpha \lambda^2 - s}{(s + \lambda)^3} \right| ds \sup_{s \in \mathbb{R}_+} |F_\gamma(s) - \hat{F}_{\gamma,n,p}(s)|. \tag{31}
\]

Now note that

\[
\int_0^\infty \left| \frac{\lambda - 2\alpha \lambda^2 - s}{(s + \lambda)^3} \right| ds = \alpha, \tag{32}
\]

provided that $\lambda > 1/(2\alpha)$. Additionally, from Theorem 2.1 in Bai et al. (2003), we have the following Berry-Esseen style result:

\[
\sup_{s \in \mathbb{R}_+} |F_\gamma(s) - \hat{F}_{\gamma,n,p}(s)| = o_\mathbb{P}(n^{-w}), \tag{33}
\]

where

\[
w = \begin{cases} 
2/5 - \eta, & \text{if } \gamma \neq 1, \\
1/8, & \text{if } \gamma = 1.
\end{cases}
\]

17
for any $\eta > 0$. The claim for ridge now follows, by putting (33), (32) together with (31).

Turning to gradient flow, similar arguments yield, for any $t \geq 0$, that

$$\gamma \cdot \frac{p/n}{\gamma} \cdot \text{Risk}_{\gamma} (\hat{\beta}_{gf}^{\gamma}(t)) = -\sigma^2 \gamma \left[ \alpha \hat{F}_{\gamma,n,p}(0) + \lim_{c \to 0^+} \int_0^c -1 - \exp(-2st) + 2 \exp(-st) - 2 \exp(-2st)st + 2 \exp(-st)st - 2 \alpha \exp(-2st)s^2 t \hat{F}_{\gamma,n,p}(s)ds \right],$$

and

$$\text{Risk}_{\gamma} (\hat{\beta}_{gf}^{\gamma}(t)) = -\sigma^2 \gamma \left[ \alpha F_{\gamma}(0) + \lim_{c \to 0^+} \int_0^c -1 - \exp(-2st) + 2 \exp(-st) - 2 \exp(-2st)st + 2 \exp(-st)st - 2 \alpha \exp(-2st)s^2 t F_{\gamma}(s)ds \right].$$

Noting that

$$\lim_{c \to 0^+} \int_0^c \left| -1 - \exp(-2st) + 2 \exp(-st) - 2 \exp(-2st)st + 2 \exp(-st)st - 2 \alpha \exp(-2st)s^2 t \right| ds = \alpha,$$

provided that $t < 2\alpha$, yields the claim for gradient flow now, in the same way as before for ridge.

**B.3 Proof of Theorem 3**

We start by inspecting the optimal limiting risk of ridge regression. Taking an exact Taylor expansion around $s = 0$ of the expression for the limiting risk, we have that

$$\frac{1}{\sigma^2} \cdot \text{Risk}_{\gamma} (\hat{\beta}_{ridge}(\lambda^*)) = \int \left[ \frac{(\lambda^*)^2 \alpha + s}{(s + \lambda^*)^2} \right] dF_{\gamma}(s)$$

$$= \int \left[ \sum_{k=0}^{\infty} (-1)^k \alpha^{k+1} s^k \right] dF_{\gamma}(s).$$

The series in the preceding display converges, provided that $b < 1/\alpha$; as long as this is the case, the helper Lemma 9 appearing below tells us that we may interchange the integral and sum, so that

$$\frac{1}{\sigma^2} \cdot \text{Risk}_{\gamma} (\hat{\beta}_{ridge}(\lambda^*)) = \sum_{k=0}^{\infty} (-1)^k \alpha^{k+1} \int s^k dF_{\gamma}(s)$$

$$= \sum_{k=0}^{\infty} (-1)^k \alpha^{k+1} M_k,$$

where we wrote $M_k = \int s^k dF_{\gamma}(s)$ to stand for the $k$th (noncentral) moment of the MP law. Rearranging, we have that

$$\frac{1}{\alpha \sigma^2} \cdot \text{Risk}_{\gamma} (\hat{\beta}_{ridge}(\lambda^*)) = \sum_{k=0}^{\infty} (-1)^k \alpha^k M_k. \quad (34)$$

For convenience in what follows, define the remainder of the series appearing in (34) as

$$R_\ell = \sum_{k=\ell}^{\infty} (-1)^k \alpha^k M_k,$$
so that

\[ R_0 = \frac{1}{\alpha \sigma^2 \gamma} \cdot \text{Risk}_\gamma(\hat{\beta}^{\text{ridge}}(\lambda^*)) \]

Turning now to the optimal limiting risk for gradient flow (5), we begin by decomposing the risk in the following way. Write

\[
\frac{1}{\sigma^2 \gamma} \cdot \text{Risk}_\gamma(\hat{\beta}^{gf}(t^*)) = \int \left[ \exp(-2t^* s)\alpha + \frac{(1 - \exp(-t^* s))^2}{s} \right] dF_\gamma(s)
\]

\[
= \int \left[ \exp(-2t^* s)\alpha + \frac{1}{s} \left( \frac{1}{T_1} + \exp(-2t^* s) - 2\exp(-t^* s) \right) \right] dF_\gamma(s).
\]

Taylor expanding the terms \( T_1, T_3, T_4 \) all around \( s = 0 \) yields

\[
T_1 = \sum_{k=0}^\infty (-1)^k \frac{(2k)!}{k!} \alpha^{k+1} s^k
\]

\[
T_3 = \sum_{k=0}^\infty (-1)^k \frac{(2k)!}{k!} \alpha^k s^k
\]

\[
T_4 = \sum_{k=0}^\infty (-1)^k \frac{(2k)!}{k!} \alpha^k s^k,
\]

where we employed the usual convention that \( 0! = 1 \). Note that the preceding series all converge for \( s \in \mathbb{R}_+ \). Combining terms, we have that

\[
\frac{1}{s} (T_2 + T_3 - T_4) = \frac{1}{s} \left[ \sum_{k=2}^\infty \frac{(-2)^k - 2(-1)^k}{k!} \alpha^k s^k \right] - \sum_{k=2}^\infty \frac{((-2)^k - 2(-1)^k)}{k!} \alpha^k s^{k-1} + \sum_{k=1}^\infty \frac{((2)^k+1 - 2(-1)^{k+1})}{(k+1)!} \alpha^{k+1} s^k,
\]

and therefore

\[
T_1 + \frac{1}{s} (T_2 + T_3 - T_4) = \sum_{k=0}^\infty (-1)^k \frac{(2k)!}{k!} \alpha^{k+1} s^k + \sum_{k=1}^\infty \frac{((-2)^{k+1} - 2(-1)^{k+1})}{(k+1)!} \alpha^{k+1} s^k
\]

\[
= \alpha + \sum_{k=1}^\infty \left[ (-2)^k \frac{k!}{k!} \alpha^{k+1} + \frac{((-2)^{k+1} - 2(-1)^{k+1})}{(k+1)!} \alpha^{k+1} s^k \right]
\]

Putting all the pieces together, we have that

\[
\frac{1}{\sigma^2 \gamma} \cdot \text{Risk}_\gamma(\hat{\beta}^{gf}(t^*)) = \int \sum_{k=0}^\infty \left[ \frac{2k! - 2k^{k+1} + 2k(k+1)!}{k!(k+1)!} (-1)^k \alpha^{k+1} s^k \right] dF_\gamma(s),
\]
Again, interchanging the integral and series above is justified by Lemma 9.

The first inequality above followed from the helper Lemmas 10 and 7 appearing below, while the second inequality followed from Lemma 8.

Now note that

\[
\frac{1}{\alpha \sigma^2 \gamma} \cdot \text{Risk}_\gamma(\hat{\beta}^\text{gf}(t^*)) = \sum_{k=0}^{\infty} a_k (-1)^k \alpha^k M_k,
\]

where we defined

\[
a_k = \frac{2k! - 2^{1+k}k! + 2^k(1+k)!}{k!(1+k)!} = \frac{2}{(k+1)!} + \frac{2^k}{k!} - \frac{2^{k+1}}{(k+1)!}.
\]

Again, interchanging the integral and series above is justified by Lemma 9.

As \(a_0 = a_1 = a_2 = 1\), we obtain

\[
\frac{1}{\alpha \sigma^2 \gamma} \cdot \text{Risk}_\gamma(\hat{\beta}^\text{gf}(t^*)) = (R_0 - R_3) + \sum_{k=3}^{\infty} a_k (-1)^k \alpha^k M_k
\]

\[
= (R_0 - R_3) + \left( a_3 R_3 + (a_4 - a_3)R_4 + (a_5 - a_4)R_5 + (a_6 - a_5)R_6 + (a_7 - a_6)R_7 + \cdots \right)
\]

\[
\leq R_0 - \left( \frac{1}{4} \right) R_3 - \left( \frac{1}{3} \right) R_4 + \left( a_5 - a_4 \right) R_5 + \left( a_7 - a_6 \right) R_7 + \cdots
\]

\[
= R_0 - \left( \frac{1}{4} \right) R_3 - \left( \frac{1}{3} \right) R_4 + \sum_{k=5,7,9,\ldots} (a_k - a_{k-1}) R_k
\]

\[
\leq R_0 - \left( \frac{1}{4} \right) R_3 - \left( \frac{1}{3} \right) R_4 + \sum_{k=5,7,9,\ldots} (a_{k-1} - a_k) |R_5|.
\] (35)

The first inequality above followed from the helper Lemmas 10 and 7 appearing below, while the second inequality followed from Lemma 8.

Now note that

\[
\sum_{k=5,7,9,\ldots} (a_{k-1} - a_k) = \sum_{k=5,7,9,\ldots} \left[ \left( \frac{2}{k!} + \frac{2^{k-1}}{(k-1)!} - \frac{2^k}{k!} \right) \right] - \left[ \left( \frac{2}{(k+1)!} + \frac{2^k}{k!} - \frac{2^{k+1}}{(k+1)!} \right) \right]
\]

\[
= \sum_{k=5,7,9,\ldots} \left[ \frac{2}{k!} - \frac{2}{(k+1)!} \right] + 2 \cdot \sum_{k=5,7,9,\ldots} \left[ \frac{2^{k+1}}{(k+1)!} - \frac{2^k}{k!} \right] + \frac{2^4}{4!}
\]

\[
= 2 \cdot \left[ \sum_{k=0}^{\infty} (-1)^k \frac{2^k}{k!} - \sum_{k=0}^{3} (-1)^k \frac{2^k}{k!} \right] - \frac{2^4}{4!} - \left[ \sum_{k=0}^{\infty} (-1)^k \frac{2^k}{k!} - \sum_{k=0}^{3} (-1)^k \frac{2^k}{k!} \right]
\]

\[
= 2 \cdot \left[ \frac{1}{e^2} + \frac{1}{3} \right] - \frac{2}{3} - \frac{2}{e^3} - \frac{3}{4}
\]

\[
\leq 0.285,
\]

by direct calculation. Therefore, returning to (35), we have that

\[
\frac{1}{\alpha \sigma^2 \gamma} \cdot \text{Risk}_\gamma(\hat{\beta}^\text{gf}(t^*)) \leq R_0 - \left( \frac{1}{4} \right) R_3 - \left( \frac{1}{3} \right) R_4 + \sum_{k=5,7,9,\ldots} (a_{k-1} - a_k) |R_5|
\]

\[
\leq R_0 - \left( \frac{1}{4} \right) R_3 - \left( \frac{1}{3} \right) R_4 + 0.285 \cdot |R_5|
\]

\[
\leq R_0 - \left( \frac{1}{4} \right) R_3 - \left( \frac{1}{3} \right) R_4 + 0.285 \cdot R_4
\]

\[
\leq R_0 - \left( \frac{1}{4} \right) R_3
\]

\[
= R_0 + \left( \frac{1}{4} \right) |R_3|
\]

\[
\leq 1.25 \cdot R_0,
\]

20
where the third as well as the final inequalities followed from Lemma 8, and the fourth inequality followed from Lemma 7. In other words, we have that

\[ \text{Risk}_\gamma(\hat{\beta}\text{مف}(t^*)) \leq 1.25 \cdot \text{Risk}_\gamma(\hat{\beta}\text{ridge}(\lambda^*)) \]

as claimed. \hfill \Box

### B.3.1 Proofs of Helper Lemmas

Below, we prove the helper lemmas used in the proof of Theorem 3, above. Throughout, we refer to the remainder terms

\[ R_\ell = \sum_{k \geq \ell} (-1)^k \alpha^k M_k, \quad \ell = 0, 1, 2, \ldots, \]

where \( M_k = \int s^k dF_\gamma(s) \) denotes the \( k \)th (noncentral) moment of the MP law. Additionally, we define the sequence

\[ a_k = \frac{2}{(k+1)!} + \frac{2^k}{k!} - \frac{2^{k+1}}{(k+1)!}, \quad k = 0, 1, 2, \ldots, \]

where \( 0! = 1 \).

**Lemma 7.** Fix some \( \sigma, \gamma, \) and \( r \) such that \( b < 1/\alpha \). Then, it holds that \( (-1)^\ell R_\ell \geq 0 \), for any \( \ell \geq 0 \).

**Proof.** For all \( \ell \geq 0 \), we can write

\[
R_\ell = \sum_{k \geq \ell} (-1)^k \alpha^k M_k
\]

\[
= \int \left[ \sum_{k \geq \ell} (-1)^k \alpha^k s^k \right] dF_\gamma(s)
\]

\[
= \int R_\ell^s dF_\gamma(s)
\]

\[
= \mathbb{E}[R_\ell^s],
\]

where we defined \( R_\ell^s = \sum_{k \geq \ell} (-1)^k \alpha^k s^k \). Above, interchanging the integral and series is justified by Lemma 9. For convenience, we also introduce the notation that \( b_k = \alpha^k s^k \); for \( k \geq 0 \).

As \( b < 1/\alpha \), for any \( k \geq 0 \), we have that (i) \( b_k \in (0, 1) \), (ii) the \( b_k \) are decreasing, and (iii) \( \lim_{k \to \infty} b_k = 0 \). (Note that we may assume that \( s \) is bounded away from zero, since when \( s = 0 \) it contributes nothing to the integral in the definition of \( R_\ell \); above; hence \( b_k > 0 \), as written in (i).)

Now let \( \ell \geq 0 \) be even. We may write \( R_\ell^s = b_\ell + R_{\ell+1}^s \). When \( R_{\ell+1}^s \geq 0 \), we have \( R_\ell^s > 0 \), because \( b_\ell > 0 \). On the other hand, when \( R_{\ell+1}^s < 0 \), we still have that \( R_\ell^s > 0 \), because \( |R_{\ell+1}^s| \leq b_{\ell+1} < b_\ell \). The reasoning for \( \ell \geq 1 \) being odd is similar.

Therefore, \( (-1)^\ell R_\ell^s \geq 0 \), for any \( \ell \geq 0 \) and \( s \) satisfying the conditions of the lemma. Consequently,

\[ \mathbb{E}[(-1)^\ell R_\ell^s] = (-1)^\ell R_\ell \geq 0, \]

as claimed. \hfill \Box

**Lemma 8.** Fix some \( \sigma, \gamma, \) and \( r \) such that \( b < 1/\alpha \). Then, for any \( \ell \geq 0 \), it holds that \( |R_\ell| \geq |R_{\ell+1}| \).

**Proof.** Just as in the proof of Lemma 7, for all \( \ell \geq 0 \), write \( R_\ell = \mathbb{E}[R_\ell^s] \), where we defined \( R_\ell^s = \sum_{k \geq \ell} (-1)^k \alpha^k s^k \). Also, define \( b_k = \alpha^k s^k \); for \( k \geq 0 \), so that

\[ R_\ell = \int \left[ \sum_{k \geq \ell} (-1)^k b_k \right] dF_\gamma(s). \]
Now note that, for even \( \ell \geq 0 \) (and any \( s \)), from Lemma 7, we have that \( R^s_\ell \geq 0 \), and so
\[
|R_\ell| = |E[R^s_\ell]| = E|R^s_\ell| = \mathbb{E}[|R^s_\ell|].
\]
Similarly, for odd \( \ell \geq 0 \), we have that \( R^s_\ell \leq 0 \), and so
\[
|R_\ell| = |E[R^s_\ell]| = |1| \cdot |E[R^s_\ell]| = |E[R^s_\ell]| = |E((-1)R^s_\ell)| = E((-1)R^s_\ell) = E[|R^s_\ell|].
\]

Therefore, if we could show that \( |R^s_\ell| \geq |R^s_{\ell+1}| \), for all \( \ell \geq 0 \), this would imply that \( E[|R^s_\ell|] \geq E[|R^s_{\ell+1}|] \) (for all \( \ell \geq 0 \)), and therefore that \( |R_\ell| \geq |R_{\ell+1}| \) (for all \( \ell \geq 0 \)), and we would be done.

To get a contradiction, suppose there exists some \( \ell^* \) such that \( |R^s_{\ell^*}| < |R^s_{\ell^*+1}| \). But a basic result on remainder terms of alternating series (Calabrese, 1962) tells us (for any \( \ell \geq 0 \), including \( \ell^* \)) that we have the bounds:
\[
\frac{b_{\ell^*+1}}{2} < |R^s_{\ell^*+1}| < \frac{b_{\ell^*}}{2} \quad \text{and} \quad \frac{b_{\ell^*}}{2} < |R^s_{\ell^*}| < \frac{b_{\ell^*-1}}{2},
\]
meaning that there exists some \( \ell^* \geq 0 \) such that
\[
\frac{b_{\ell^*}}{2} < |R^s_{\ell^*+1}| < \frac{b_{\ell^*}}{2},
\]
which is a contradiction. Therefore, \( |R^s_\ell| \geq |R^s_{\ell+1}| \), for all \( \ell \geq 0 \), and the claim follows.

Lastly, we need to verify the conditions required in order to apply the result on remainder terms from above. Using the properties (i), (ii), and (iii) established in the proof of Lemma 7, we have that
\[
0 \leq b_k - b_{k+1} = \alpha^k s^k - \alpha^{k+1} s^{k+1} = \alpha^k s^k \cdot (1 - \alpha s),
\]
whereas
\[
b_{k+1} - b_{k+2} = \alpha^{k+1} s^{k+1} - \alpha^{k+2} s^{k+2} = \alpha s \cdot \alpha^k s^k \cdot (1 - \alpha s)
\]
\[
= \alpha s \cdot (b_k - b_{k+1})
\]
\[
< b_k - b_{k+1}.
\]

Taken together, these findings mean that the sequence of differences \( (b_k - b_{k+1}) \) is decreasing down to zero, satisfying the conditions required for the result on remainder terms.

**Lemma 9.** Fix some \( \sigma, \gamma, \) and \( r \) such that \( b < 1/\alpha \). Then, it holds that both
\[
\frac{1}{\sigma^2 \gamma} \cdot \text{Risk}_\gamma(\hat{\beta}^{\text{ridge}}(\lambda^*)) = \int \left[ \sum_{k=0}^{\infty} (-1)^k \alpha^{k+1} s^k \right] dF_\gamma(s) = \sum_{k=0}^{\infty} (-1)^k \alpha^{k+1} \int s^k dF_\gamma(s), \tag{37}
\]
and
\[
\frac{1}{\sigma^2 \gamma} \cdot \text{Risk}_r(\hat{\beta}^{\ell}(t^*)) = \int \left[ \sum_{k=0}^{\infty} (-1)^k a_k \alpha^{k+1} s^k \right] dF_r(s) = \sum_{k=0}^{\infty} (-1)^k a_k \alpha^k \int s^k dF_r(s). \tag{38}
\]

**Proof.** For convenience, write \( b_k = (-1)^k \alpha^k s^k \), so that the integral after the first equals sign in (37) can be written as \( \alpha \cdot \int \left[ \sum_{k=0}^{\infty} b_k \right] dF_\gamma(s) \). Observe that \( |b_k| \leq c_k \), for all \( s, k \geq 0 \), where \( c_k = \alpha^k u^k \).

Moreover, as \( b < 1/\alpha \), we have that \( c_k \in (0,1) \), which implies that \( \sum_{k=0}^{\infty} c_k = 1/(1-\alpha u) \). Therefore, by the Weierstrass M test, the series \( \sum_{k=0}^{\infty} b_k \) converges uniformly, which means we may swap the order of the integral and the sum in (37). As for (38), a similar line of argument works, after noting that \( \max_k a_k = 1 \), as per Lemma 10. \( \square \)
Lemma 10. It holds that the sequence \((a_k)\) is nonincreasing. Moreover, we have that \(\max_k a_k = 1\).

Proof. Observe that \(a_0 = a_1 = a_2 = 1\). Therefore, let \(k \geq 3\) be arbitrary. We will show that \(a_k \geq a_{k+1}\), which will establish the first claim in the statement of the lemma. Observe that, as \(k \geq 3\),

\[
\frac{2k + 2}{2^{k+1}} + \frac{1}{2}(k + 2)(k + 1) + 2 > 2(k + 2)
\]

\[
\iff (2k + 2) + 2^k(k + 2)(k + 1) + 2^{k+2} > 2 \cdot 2^{k+1}(k + 2)
\]

\[
\iff 2(k + 2) + 2^k(k + 2)(k + 1) - 2^{k+1}(k + 2) > 2 + 2^{k+1}(k + 2) - 2^{k+2}
\]

\[
\iff \frac{2}{(k + 1)!} + \frac{2^k}{k!} > \frac{2}{(k + 2)!} + \frac{2^{k+1}}{(k + 1)!} - \frac{2^{k+2}}{(k + 2)!}
\]

\[
\iff a_k > a_{k+1},
\]

as claimed. To be clear, the second line followed by multiplying the first line through by \(2^{k+1}\). The third line followed after a bit of algebra. The fourth line followed by dividing the third line through by \((k + 2)!\).

The second claim in the statement of the lemma follows from the observation that \(a_0 = 1\) (as mentioned earlier), put together with the fact that the \(a_k\) are nonincreasing. \(\square\)

B.4 Proof of Theorem 4

Let \(f_0(\lambda) = \text{Risk}_\gamma(\hat{\beta}^{\text{ridge}}(\lambda))\) denote the risk of the ridge estimator at tuning parameter \(\lambda\), as in (20). From Dobriban and Wager (2018), we have the explicit form

\[ f_0(\lambda) = 1 + \gamma m(-\lambda) + \lambda(\lambda r^2/\sigma^2 - \gamma)m'(-\lambda), \]

where we are abbreviating here \(m = m_{F, \gamma}\) for the Stieltjes transform of the MP law, as in (8). Differentiation with respect to \(\lambda\) yields, after some simplification,

\[
f_0'(\lambda) = \frac{2\lambda r^2/\sigma^2 - 2\gamma}{(\gamma - (1 - \lambda)^2 + 2\lambda)^{3/2}}
\]

The minimizer of the ridge risk, recall, is simply \(\lambda = 1/\alpha^2 = \sigma^2\gamma/r^2\). Thus, as the denominator above is always positive, it is easy to see that \(f_0'(\lambda) < 0\) for \(\lambda < \lambda^*\) and \(f_0'(\lambda) > 0\) for \(\lambda > \lambda^*\), i.e., \(f_0\) is decreasing to the left of its minimizer, and increasing to the right of its minimizer.

Now let \(f_1(t) = f_0(1/t)\), and let \(f_2(t) = \text{Risk}_\gamma(\hat{\beta}^{g}(t))\) denote the risk of the gradient flow estimator at time \(t\), as given in (19). We collect several facts about \(f_1, f_2\):

1. \(f_1\) is minimized at \(t^* = \alpha\), and is monotone on each side of this minimizer;
2. \(f_1(0) = f_2(0)\) and \(f_1(\infty) = f_2(\infty)\);
3. \(\min_{t \in [0, \alpha]} f_2(t) \geq f_1(\alpha)\);
4. \(\max_{t \in [0, \alpha]} f_2(t) \leq f_1(0) + 1.25f_1(\alpha) \leq 2.25f_1(0)\);
5. \(\min_{t \in [\alpha, \infty]} f_2(t) \geq f_1(\alpha)\);
6. \(\max_{t \in [\alpha, \infty]} f_2(t) \leq f_1(\infty) + 1.25f_1(\alpha) \leq 2.25f_1(\infty)\);

Fact 1 is simply a restatement of the fact established at the start of this proof. Fact 2 is immediate, either by inspection of (19), (20), or by noting that both gradient flow and ridge match at the extreme
ends of shrinkage (they perform least squares with no shrinkage, and return the null estimator with infinite shrinkage). Fact 3 is a consequence of Theorem 2. Fact 4 holds as
\[
\max_{t \geq 0} f_2(t) = \max_{t \geq 0} \sigma^2 \gamma \int \left[ \exp(-2ts)\alpha + \frac{(1 - \exp(-ts))^2}{s} \right] dF_\gamma(s)
\leq \max_{t \geq 0} \sigma^2 \gamma \int \exp(-2ts)\alpha dF_\gamma(s) + \max_{t \geq 0} \sigma^2 \gamma \int \frac{(1 - \exp(-ts))^2}{s} dF_\gamma(s)
= f_2(0) + f_2(\alpha)
\leq f_1(0) + 1.25 f_1(\alpha)
\leq 2.25 f_1(0).
\]
where in the third line we use monotonicity of each of the (squared) bias and variance terms with respect to \(t\), in the fourth we used Fact 2 and Theorem 3, and in the fifth we used Fact 1. Facts 5 and 6 hold similarly.

Putting all these facts to use, let us now define a calibration map \(\phi : [0, \infty) \rightarrow [0, \infty)\)
\[
\phi(t) = \begin{cases} (f_1|_{[0, \alpha]}^{-1}(f_2(t) \land f_1(0)) & \text{for } t \in [0, \alpha], \\ (f_1|_{[\alpha, \infty]}^{-1}(f_2(t) \land f_1(\infty)) & \text{for } t \in [\alpha, \infty], \end{cases}
\]
where \(f_1|_{[0, \alpha]}\) denotes the restriction of \(f_1\) to \([0, \alpha]\) which we argued above is invertible, and similarly for \(f_1|_{[\alpha, \infty]}\). Also, here we abbreviate \(a \land b = \max\{a, b\}\). Note that by Facts 3 and 5, the above calibration map is well-defined, because we are only evaluating the inverse of \(f_1|_{[0, \alpha]}\) on a subset of \(f_1([0, \alpha])\), and similarly for \(f_1|_{[\alpha, \infty]}\).

We claim that, for all \(t \geq 0\), we have \(f_2(t)/f_1(\phi(t)) \leq 2.25\), which implies the conclusion in the theorem. To see the claim, observe that for \(t \in [0, \alpha]\),
\[
f_2(t)/f_1(\phi(t)) = \begin{cases} 1 & \text{for } f_2(t) \leq f_1(0), \\ f_2(t)/f_1(0) & \text{for } f_2(t) > f_1(0), \end{cases}
\]
and for any \(t \in [0, \alpha]\), we have \(f_2(t)/f_1(0) \leq 2.25\) by Fact 4. For \(t \in [\alpha, \infty]\),
\[
f_2(t)/f_1(\phi(t)) = \begin{cases} 1 & \text{for } f_2(t) \leq f_1(\infty), \\ f_2(t)/f_1(\infty) & \text{for } f_2(t) > f_1(\infty), \end{cases}
\]
and again for any \(t \in [\alpha, \infty]\), we have \(f_2(t)/f_1(\infty) \leq 2.25\) by Fact 6. This establishes the claim and completes the proof.

B.5 Additional Numerical Experiments

Figure 3 presents some additional numerical results for gradient flow and ridge regression. We follow the exact same experimental setup as in Section 6 of the main paper, except we consider here a low-dimensional regime, where \(\gamma = 1/2\), by setting \(p = 500\) and \(n = 1000\). The results are similar to those presented in the main paper.
Figure 3: Comparison of risks for gradient flow and ridge, across a range of tuning parameters and feature distributions. We consider here a low-dimensional setup, where $\gamma = 1/2$, by setting $p = 500$ and $n = 1000$. Top row: the limiting as well as finite-sample risks, for a spherical feature covariance. Bottom row: the finite-sample risks, for a general feature covariance matrix.
References

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

Zhidong Bai, Baiqi Miao, and Jian-Feng Yao. Convergence rates of spectral distributions of large sample covariance matrices. *SIAM Journal on Matrix Analysis and Applications*, 25(1):105–127, 2003.

Patrick Billingsley. *Probability and Measure*. John Wiley & Sons, 2008.

Peter Buhlmann and Bin Yu. Boosting with the $\ell_2$ loss. *Journal of the American Statistical Association*, 98(462):324–339, 2003.

Philip Calabrese. A note on alternating series. *The American Mathematical Monthly*, 69(3):215–217, 1962.

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(1):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 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. 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*, pages 1601–1619, 2001.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

David F. Griffiths and Desmond J. Higham. *Numerical Methods for Ordinary Differential Equations*. Springer, 2010.

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 *Proceedings of the 35th 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.

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

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

Nelson Morgan and Herve Bourlard. Generalization and parameter estimation in feedforward nets: Some experiments. *Advances in Neural Information Processing Systems*, 2, 1989.
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(1):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.

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.