The Annals of Statistics
2018, Vol. 46, No. 1, 247–279
https://doi.org/10.1214/17-AOS1549

HIGH-DIMENSIONAL ASYMPTOTICS OF PREDICTION: RIDGE REGRESSION AND CLASSIFICATION

BY EDGAR DOBRIBAN[1] AND STEFAN WAGER

University of Pennsylvania and Stanford University

We provide a unified analysis of the predictive risk of ridge regression and regularized discriminant analysis in a dense random effects model. We work in a high-dimensional asymptotic regime where $p, n \to \infty$ and $p/n \to \gamma > 0$, and allow for arbitrary covariance among the features. For both methods, we provide an explicit and efficiently computable expression for the limiting predictive risk, which depends only on the spectrum of the feature-covariance matrix, the signal strength and the aspect ratio γ. Especially in the case of regularized discriminant analysis, we find that predictive accuracy has a nuanced dependence on the eigenvalue distribution of the covariance matrix, suggesting that analyses based on the operator norm of the covariance matrix may not be sharp. Our results also uncover an exact *inverse* relation between the limiting predictive risk and the limiting estimation risk in high-dimensional linear models. The analysis builds on recent advances in random matrix theory.

1. Introduction. Suppose a statistician observes n training examples $(x_i, y_i) \in \mathbb{R}^p \times \mathcal{Y}$ drawn independently from an unknown distribution $\mathcal{D}$, and wants to find a rule for predicting y on future unlabeled draws x from $\mathcal{D}$. In other words, the statistician seeks a function $h : \mathbb{R}^p \to \mathcal{Y}$, $h(x) = g(w^\top x)$ for which $\mathbb{E}_{\mathcal{D}}[\ell(y, h(x))]$ is small, where $\ell(\cdot, \cdot)$ is a loss function; in regression $\mathcal{Y} = \mathbb{R}$ and ℓ is the squared error loss, while in classification $\mathcal{Y} = \{0, 1\}$ and ℓ is the 0–1 loss. Such prediction problems lie at the heart over several scientific and industrial endeavors in fields ranging from genetics [Wray, Goddard and Visscher (2007)] and computer vision [Russakovsky et al. (2014)] to Medicare resource allocation [Kleinberg et al. (2015)].

When the number of features p is large, accurate prediction is not always possible. Thus, in order to guarantee good results, the statistician needs to invoke some "enabling hypothesis" that encodes domain-specific knowledge about the problem and guides model fitting. Popular options include the "sparsity hypothesis," that is, that there is a good predictive rule depending only on $w^\top x$ for some sparse weight vector w [Donoho et al. (1992), Hastie, Tibshirani and Wainwright

Received December 2015; revised November 2016.

[1]Supported in part by NSF Grants DMS-1418362 and DMS-1407813, and by an HHMI International Student Research Fellowship.

MSC2010 subject classifications. Primary 62H99; secondary 62J05, 62H30.

Key words and phrases. High-dimensional asymptotics, ridge regression, regularized discriminant analysis, prediction error, random matrix theory.

(2015), Candès and Tao (2007)], the "manifold hypothesis" positing that the x_i have useful low-dimensional geometric structure [Rifai et al. (2011), Simard et al. (2000)], and several variants of an "independence hypothesis" that rely on independence assumptions for the feature distribution [Bickel and Levina (2004), Ng and Jordan (2001)]. The choice of enabling hypothesis is important from a practical perspective, as it helps choose which predictive method to use, for example, the lasso with sparsity, neighborhood-based methods under the manifold hypothesis or naive Bayes given independent features.

In many applications, however, the above enabling hypotheses are not known to apply, yet practitioners still achieve accurate high-dimensional prediction using dense, that is, nonsparse, ridge-regularized linear methods trained on highly correlated features. One striking example is the case of document classification with dictionary-based features of the form "how many times does the jth word in the dictionary appear in the current document." Even though $p \gg n$, dense ridge-regularized methods reliably work well across a wide range of problem settings [Sutton and McCallum (2006), Toutanova et al. (2003)], and sometimes even achieve state-of-the-art performance on important engineering tasks [Wang and Manning (2012)]. Another example is bioinformatics, where in a recent test of prediction algorithms [Bernau et al. (2014)], ridge regression—and a method that was previously proposed by those same authors—performed best, better than lasso regression and boosting.

The goal of this paper is to gain better understanding of the performance of ridge-regularized linear prediction methods, which are so widely in use. We focus on a random-effects hypothesis positing that the effect size of each feature is drawn independently at random, and show how working under this hypothesis enables us to get precise accuracy guarantees for ridge-regularized methods. Heuristically, results obtained under this random-effects hypothesis can be viewed as average-case analyses over dense parameters. The random-effects hypothesis is of course very strong; however, it yields a qualitatively different theory for high-dimensional prediction than popular approaches, and thus may motivate future developments.

HYPOTHESIS (Random effects). *Each predictor has a small, independent random effect on the outcome.*

This hypothesis is fruitful both theoretically and methodologically. Using random-matrix theoretic techniques [see, e.g., Bai and Silverstein (2010)], we derive closed-form expressions for the limiting out-of-sample predictive risk of ridge regularization in two settings: regression and discriminant analysis. We allow for the features x to have a general covariance structure Σ. The resulting formulas are pleasingly simple and depend on Σ through the Stieltjes transform of the limiting empirical spectral distribution. More prosaically, Σ only enters into our formulas through the almost-sure limits of $p^{-1}\operatorname{tr}((\widehat{\Sigma}+\lambda I_p)^{-1})$ and $p^{-1}\operatorname{tr}((\widehat{\Sigma}+\lambda I_p)^{-2})$,

where $\widehat{\Sigma}$ is the sample covariance and $\lambda > 0$ the ridge-regularization parameter. Notably, the same mathematical tools can describe the two settings.

From a practical perspective, we identify several high-dimensional regimes where mildly regularized discriminant analysis performs strikingly well. We hope that further work motivated by generalizations of the random-effects hypothesis could yield a new theoretical underpinning for dense high-dimensional prediction.

1.1. *Overview of results.* In the first part of our paper, we study the predictive risk of ridge regression in a high-dimensional asymptotic regime where $n, p \to \infty$ and p/n converges to a limiting aspect ratio $p/n \to \gamma > 0$. The spectral distribution, that is, the cumulative distribution function of the eigenvalues of the feature covariance matrix Σ converges weakly to a limiting spectral measure supported on $[0, \infty)$, which allows Σ to be general. After establishing formulas for the limiting predictive risk under a suitable random effects hypothesis (Theorem 2.1), we use them to gain qualitative insights about the behavior of ridge regression.

We show that, when the signal-to-noise ratio is high, the accuracy of ridge regression has a sharp phase transition at $\gamma = 1$ regardless of Σ, essentially validating a conjecture of Liang and Srebro (2010) on the "regimes of learning" problem (Section 2.1). Theorem 2.1 also implies a general *inaccuracy principle* for high-dimensional linear models, whereby there are no correlation structures Σ for which prediction and estimation are both easy (Section 2.2).

In the second part of the paper, we study regularized linear discriminant analysis (RDA) [Friedman (1989), Serdobolskii (1983)] in the two-class Gaussian problem

$$y \sim \{\pm 1\} \quad \text{with} \quad \mathbb{P}[y = \pm 1] = \pi_{\pm 1}, \quad \text{and} \quad x \sim \mathcal{N}(\mu_y, \Sigma), \tag{1}$$

where $\pi_{\pm 1}$ are known and $\mu_{\pm 1}$ and Σ are unknown. We show that the out-of-sample classification error of RDA converges to an almost-sure limit (Theorem 3.1). This limit depends on the angle between the oracle separating hyperplane and the estimated hyperplane, as well as on the limiting Bayes error (Section 3.2). This result is generalized to unequal class sample sizes in Theorem 3.2.

We can again use our result to derive qualitative insights about the behavior of RDA. We find that the limiting angle between the estimated and oracle hyperplanes converges to a nontrivial quantity as the signal strength α^2 goes to infinity (formally, $\alpha^2 = 4\mathbb{E}[\|\mu_{+1} - \mu_{-1}\|^2]$), implying that our analysis is helpful in understanding the asymptotics of RDA even in a very high signal-to-noise regime (Corollary 3.4). Finally, by studying the limits as the "regularization strength" in RDA becomes small or large, we can recover known results about Fisher's linear discriminant analysis and naive Bayes methods going back to Bickel and Levina (2004), Raudys (1967), Saranadasa (1993), and even to early ideas discussed by Kolmogorov (Section 3.4).

Mathematically, our results build on recent advances in random matrix theory. A main difficulty here is finding explicit limits of certain trace functionals involving both the sample and the population covariance matrix. For instance,

the well-known Stieltjes transform m of the limiting empirical spectral distribution (i.e., the limit distribution of the eigenvalues $\widehat{\Sigma}$, which exists in our case, as explained later) satisfies $m(-\lambda) = \lim_{p\to\infty} p^{-1}\operatorname{tr}((\widehat{\Sigma} + \lambda I_p)^{-1})$. However, standard random matrix theory does not provide simple expressions for the limits of functionals that come up in our analysis, such as $p^{-1}\operatorname{tr}(\Sigma(\widehat{\Sigma} + \lambda I_p)^{-1})$ or $p^{-1}\operatorname{tr}([\Sigma(\widehat{\Sigma} + \lambda I_p)]^{-2})$, which involve both Σ and $\widehat{\Sigma}$. For this, we leverage and build on recent results, including the work of Chen et al. (2011), Hachem, Loubaton and Najim (2007), and Ledoit and Péché (2011). Our contributions include some new explicit formulas, for which we refer to the proofs. These formulas may prove useful for the analysis of other statistical methods under high-dimensional asymptotics, such as principal component regression.

1.2. *A first example.* A key contribution of our theory is a precise understanding of the effect of correlations between the features on regularized discriminant analysis, given our random-effects model. Correlated features have a nontrivial effect on predictive accuracy that cannot be summarized using standard notions such as the condition number of Σ or the classification margin; rather, the full eigenvalue spectrum of Σ matters. A similar phenomenon also holds for PCA [Dobriban (2016)]. This observation is at odds with popular analyses of high-dimensional classification methods, in which the error bounds often depend on the operator norm $\|\Sigma\|_2$ [see, e.g., Fan, Fan and Wu (2011)], thus suggesting that existing analyses of many classification methods are not sharp.

Consider the following examples: First, Σ has eigenvalues corresponding to evenly-spaced quantiles of the standard `Exponential` distribution; Second, Σ has a depth-d `BinaryTree` covariance structure, used in genetics to model the correlations between populations with an evolutionary history described by a balanced binary tree [Pickrell and Pritchard (2012)]. In the second case, the eigenvalue spectrum equals $H_p = \sum_{i=1}^{d} 2^{-i}\delta_{2^i} + 2^{-d}\delta_{2^d}$, where δ_c is the point mass at c. In both cases, we set the class means μ_y such as to keep the Bayes error constant across experiments. Figure 1 plots our formulas for the asymptotic error rate along with empirical realizations of the classification error.

Both covariance structures are far from the identity, and have similar condition numbers. However, the `Exponential` problem is vastly more difficult for RDA than the `BinaryTree` problem. This example shows that classical notions like the classification margin or the condition number of Σ cannot satisfactorily explain the high-dimensional predictive performance of RDA; meanwhile, our asymptotic formulas are accurate even in moderate sample sizes. Our computational results are reproducible using software available from https://github.com/dobriban/high-dim-risk-experiments/.

1.3. *Related work.* Random matrix theoretic approaches have been used to study regression and classification in high-dimensional statistics [Serdobolskii (2007), Yao, Bai and Zheng (2015)], as well as in wireless communications

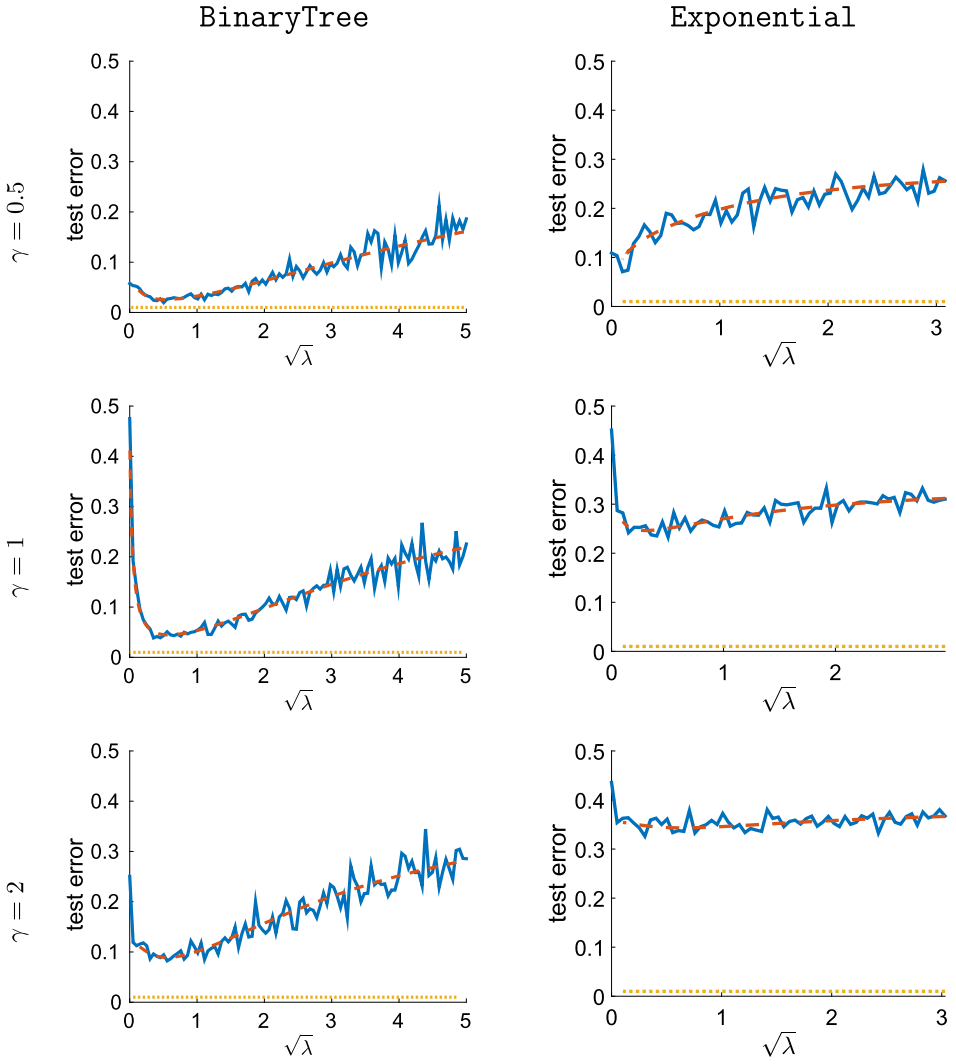

FIG. 1. *Classification error of RDA in the* `BinaryTree` *and* `Exponential` *models. The theoretical formula* (*red*, *dashed*) *is overlaid with the results from simulations* (*blue*, *solid*); *we also display the oracle error* (*yellow*, *dotted*). *The class means are drawn from* $\mu_{\pm 1} \sim \mathcal{N}(0, \alpha^2 p^{-1} I_p)$, *where* α *is calibrated such that the oracle classifier always has an error rate of* 1%. *For* `BinaryTree`, *we train on* $n = \gamma^{-1} p$ *samples*, *where* $p = 1024$; *for* `Exponential`, *we use* $n = 500$ *samples. The sizes of the two classes are equal. On the left*, p *is fixed*, *but on the right* n *is fixed* (*and* p *changes*). *We test the trained model on* 10,000 *new samples*, *and report the average classification error. Our asymptotically-motivated theoretical formulas appear to be accurate here*, *even though we only have a moderate problem size. The parameter* λ, *defined in Section* 3, *quantifies the strength of the regularization.*

[Couillet and Debbah (2011), Tulino and Verdú (2004)]. Various regression and M-estimation problems have been studied in high dimensions using approximate message passing [Bayati and Montanari (2012), Donoho and Montanari (2015)] as well as using methods inspired by random matrix theory [Bean et al. (2013)]. A remarkably early random matrix theoretic analysis of regularized discriminant analysis is due to Serdobolskii (1983). Kolmogorov and researchers around him were interested in the area even earlier (Section 3.4). In wireless communications, estimation with ridge regression is well understood; however, its prediction error has apparently not been studied.

El Karoui and Kösters (2011) study the geometric sensitivity of random matrix results, and discuss the consequences to ridge regression and regularized discriminant analysis, under weak theoretical assumptions. In contrast, we make stronger assumptions that enable explicit formulas for the limiting risk of both methods, and allow us to uncover several qualitative phenomena. Our use of results from Ledoit and Péché (2011) simplifies the proof.

We review the literature focusing on ridge regression or RDA specifically in Sections 2.3 and 3.5, respectively. Important references include, among others, Bickel and Levina (2004), Dicker (2014), El Karoui (2013), Fujikoshi, Ulyanov and Shimizu (2011), Hsu, Kakade and Zhang (2014), Saranadasa (1993) and Zollanvari and Dougherty (2015).

1.4. *Basics and notation.* We begin by reviewing some key concepts from random matrix theory (RMT) used in our analysis. RMT lets us describe the asymptotics of the eigenvalues of large matrices [see, e.g., Bai and Silverstein (2010)]. These results are typically stated in terms of the *spectral distribution*, which for a symmetric matrix A is the cumulative distribution function of its eigenvalues: $F_A(x) = p^{-1} \sum_{i=1}^{p} \mathrm{I}(\lambda_i(A) \leq x)$. In particular, the well-known Marchenko–Pastur theorem, given below, characterizes the spectral distribution of covariance matrices. We will assume the following high-dimensional asymptotic model.

ASSUMPTION HDA (High-Dimensional Asymptotics). The following conditions hold:

1. The data $X \in \mathbb{R}^{n \times p}$ are generated as $X = Z\Sigma^{1/2}$ for an $n \times p$ matrix Z with i.i.d. entries satisfying $\mathbb{E}[Z_{ij}] = 0$ and $\mathrm{Var}[Z_{ij}] = 1$, and a deterministic $p \times p$ positive semidefinite covariance matrix Σ.
2. The sample size $n \to \infty$ while the dimensionality $p \to \infty$ as well, such that the aspect ratio $p/n \to \gamma > 0$.
3. The spectral distribution F_Σ of Σ converges to a limit probability distribution H supported on $[0, \infty)$, called the population spectral distribution (PSD).

Families of covariance matrices Σ that fit the setting of this theorem include the identity covariance, `BinaryTree`, `Exponential` and the autoregressive `AR-1`

model with $\Sigma_{ij} = \rho^{|i-j|}$ [see Grenander and Szegő (1984), for the last one]. For other examples of such covariance structures, see Raudys and Saudargiene (1998) or the Appendix of Raudys (2001).

THEOREM [Marchenko and Pastur (1967), Silverstein (1995)]. *Under assumption* HDA, *the spectral distribution* $F_{\widehat{\Sigma}}$ *of the sample covariance matrix* $\widehat{\Sigma}$ *also converges weakly, with probability* 1, *to a limiting distribution supported on* $[0, \infty)$.

The limiting distribution F is called the empirical spectral distribution (ESD), and is determined uniquely by a fixed-point equation for its *Stieltjes transform*. This is defined for any distribution G supported on $[0, \infty)$ as

$$m_G(z) = \int_{l=0}^{\infty} \frac{dG(l)}{l - z}, \qquad z \in \mathbb{C} \setminus \mathbb{R}^+.$$

Given this notation, the Stieltjes transform of the spectral measure of $\widehat{\Sigma}$ satisfies

$$m_{\widehat{\Sigma}}(z) = p^{-1} \operatorname{tr}\big((\widehat{\Sigma} - zI_p)^{-1}\big) \qquad \text{converges to } m(z) \tag{2}$$

both almost surely and in expectation, for any $z \in \mathbb{C} \setminus \mathbb{R}^+$; here, we wrote $m(z) := m_F(z)$. We also define the companion Stieltjes transform $v(z)$, which is the Stieltjes transform of the limiting spectral distribution of the $n \times n$ matrix $\underline{\widehat{\Sigma}} = n^{-1} X X^\top$. Note that $\underline{\widehat{\Sigma}}$ is $n \times n$ while $\widehat{\Sigma}$ is $p \times p$. The Stieltjes transform $v(z)$ is related to $m(z)$ by

$$\gamma\left(m(z) + \frac{1}{z}\right) = v(z) + \frac{1}{z} \qquad \text{for all } z \in \mathbb{C} \setminus \mathbb{R}^+. \tag{3}$$

In addition, we denote by $m'(-\lambda)$ the derivative of the Stieltjes transform $m(z)$ evaluated at $z = -\lambda$. We can write the derivatives as

$$m'(z) = \int_{l=0}^{\infty} \frac{dG(l)}{(l - z)^2} \quad \text{and} \quad v'(z) = \gamma\left(m'(z) - \frac{1}{z^2}\right) + \frac{1}{z^2}.$$

These derivatives can also be understood in terms of empirical quantities, through the relation

$$p^{-1} \operatorname{tr}\big((\widehat{\Sigma} + \lambda I_p)^{-2}\big) \to_{\text{a.s.}} m'(-\lambda).$$

Finally, our analysis also relies on several more recent formulas for limits of trace functionals involving both Σ and $\widehat{\Sigma}$. In particular, we use a formula due to Ledoit and Péché (2011), who in the analysis of eigenvectors of sample covariance matrices showed that, under certain moment conditions detailed in the supplement [Dobriban and Wager (2018)],

$$p^{-1} \operatorname{tr}\big(\Sigma(\widehat{\Sigma} + \lambda I_p)^{-1}\big) \to_{\text{a.s.}} \frac{1}{\gamma}\left(\frac{1}{\lambda v(-\lambda)} - 1\right) \qquad \text{as } n, p \to \infty. \tag{4}$$

2. Predictive risk of ridge regression. In the first part of the paper, we study the predictive behavior of ridge regression under high-dimensional asymptotics. Suppose that we have data drawn from a p-dimensional random-design linear model with n independent observations $y_i = x_i^\top w + \varepsilon_i$. The noise terms ε_i are independent, centered, with variance 1 and are independent of the other random quantities. The x_i are arranged as the rows of the $n \times p$ matrix X, and y_i are the entries of the $n \times 1$ vector Y. We estimate w by ridge regression: $\hat{w}_\lambda = (X^\top X + \lambda n I_p)^{-1} X^\top Y$, for some $\lambda > 0$. We make the following random weights assumption, where $\alpha^2 = \mathbb{E}[\|w\|_2^2]$ is the expected signal strength.[2]

ASSUMPTION RRC (Random Regression Coefficients). The regression coefficients $w \in \mathbb{R}^p$ are random with $\mathbb{E}[w] = 0$, and $\mathrm{Var}[w] = p^{-1}\alpha^2 I_p$.

Our result about the predictive risk of ridge regression is stated in terms of the expected predictive risk $r_\lambda(X) = \mathbb{E}[(y_0 - \hat{y}_{0,\lambda})^2 \mid X]$, where $\hat{y}_{0,\lambda} = \hat{w}_\lambda^\top x_0$ and the expectation is taken over an independent random test example (x_0, y_0) from the same distribution as the training data, and over the randomness in w, ε. Here, $r_\lambda(X)$ is conditioned on the random training data X, and is therefore a random variable. Below, we will write $\gamma_p = p/n$.

THEOREM 2.1. *Consider the linear model $Y = Xw + \varepsilon$ as above. Under Assumptions* HDA *and* RRC, *suppose moreover that $\mathbb{E}[Z_{ij}^{12}]$ and $\|\Sigma\|_2$ are uniformly bounded from above. Then the expected predictive risk $r_\lambda(X)$ converges almost surely to the limiting predictive risk $R_\lambda(H, \alpha^2, \gamma)$, where*

$$R_\lambda(H, \alpha^2, \gamma) = \frac{1}{\lambda v(-\lambda)} \left\{ 1 + \left(\frac{\lambda \alpha^2}{\gamma} - 1 \right) \left(1 - \frac{\lambda v'(-\lambda)}{v(-\lambda)} \right) \right\}.$$

This function is minimized at the asymptotically optimal regularization parameter $\lambda^* = \gamma \alpha^{-2}$, for which

$$R^*(H, \alpha^2, \gamma) = \frac{1}{\lambda^* v(-\lambda^*)}. \tag{5}$$

Finally, the finite sample risk evaluated at $\lambda_p^* = \gamma_p \alpha^{-2}$ converges almost surely to R^*.

[2]It is plausible that our results should also hold under weaker assumptions, where the signal strength $\|w\|_2$ concentrates to α and w points in a "generic" direction; and other conditions that achieve the same effect have in fact been considered in the literature. For example, in analyzing ridge regression with identity covariance $\Sigma = I$, Dicker (2014) assumes that w is drawn uniformly at random from the sphere with radius α; while El Karoui (2015) studies ridge-regularized robust regression estimators under the condition that w be "diffuse," meaning that each coordinate of w is bounded as $|w_j| \le C/p^{1/2}$ for some $C > 0$. Understanding the most general conditions under which the formulas developed here remain valid could lead to new insights; such investigations, however, remain beyond the scope of the present paper.

PROOF. We begin with some algebraic manipulation to reduce our problem into a statement about trace functionals. Thanks to our generative model (part 1 of Assumption HDA and the linear model assumption), we can write the predictive risk as

$$r_\lambda(X) = 1 + \mathbb{E}\big[(\hat{w}_\lambda - w)^\top \Sigma (\hat{w}_\lambda - w) \mid X\big].$$

Meanwhile, introducing the sample covariance matrix $\widehat{\Sigma} = n^{-1} X^\top X$, we obtain by definition of the ridge regression estimator $\hat{w}_\lambda$ that

$$\hat{w}_\lambda - w = -\lambda(\widehat{\Sigma} + \lambda I_p)^{-1} w + n^{-1}(\widehat{\Sigma} + \lambda I_p)^{-1} X^\top \varepsilon.$$

In our setting of linear models, the true regression coefficients w and the noise ε are independent. Thus, using the distributional assumptions on w (from RRC) and ε (from the linear model assumption), we find that

$$\begin{aligned} r_\lambda(X) = 1 + \lambda^2 \frac{\alpha^2}{p} \operatorname{tr}\big(\Sigma(\widehat{\Sigma} + \lambda I_p)^{-2}\big) \\ + n^{-1} \operatorname{tr}\big(\Sigma(\widehat{\Sigma} + \lambda I_p)^{-1} \widehat{\Sigma} (\widehat{\Sigma} + \lambda I_p)^{-1}\big). \end{aligned}$$

Splitting the last term in two by using the relation $\widehat{\Sigma} = (\widehat{\Sigma} + \lambda I_p) - \lambda I_p$, we finally get that $r_\lambda(X)$ equals

$$1 + \frac{\gamma_p}{p} \operatorname{tr}\big(\Sigma(\widehat{\Sigma} + \lambda I_p)^{-1}\big) + \big(\lambda\alpha^2 - \gamma_p\big) \frac{\lambda}{p} \operatorname{tr}\big(\Sigma(\widehat{\Sigma} + \lambda I_p)^{-2}\big).$$

This formula provides the starting point for an RMT analysis; specifically, we seek almost sure limits for the two functionals

$$p^{-1} \operatorname{tr}\big(\Sigma(\widehat{\Sigma} + \lambda I_p)^{-1}\big) \quad \text{and} \quad p^{-1} \operatorname{tr}\big(\Sigma(\widehat{\Sigma} + \lambda I_p)^{-2}\big).$$

The convergence of the first term follows directly from the theorem of Ledoit and Péché (2011), which requires Assumption HDA, and the additional conditions that $\|\Sigma\|_2$ and $\mathbb{E}[Z_{ij}^{12}]$ are bounded. The limit of this term is given in (4). Meanwhile, in the supplement we prove the following.

LEMMA 2.2. *Under the conditions in Theorem 2.1,*

$$p^{-1} \operatorname{tr}\big(\Sigma(\widehat{\Sigma} + \lambda I_p)^{-2}\big) \to_{\text{a.s.}} \frac{1}{\gamma} \frac{v(-\lambda) - \lambda v'(-\lambda)}{[\lambda v(-\lambda)]^2}.$$

Given these results, we see that the risk $r_\lambda(X)$ converges almost surely for each $\lambda > 0$ to the desired limit:

$$r_\lambda(X) \to_{\text{a.s.}} R_\lambda = \frac{1}{\lambda v(-\lambda)} \left\{ 1 + \left(\frac{\lambda\alpha^2}{\gamma} - 1 \right) \left(1 - \frac{\lambda v'(-\lambda)}{v(-\lambda)} \right) \right\}. \tag{6}$$

Finally, with $\lambda^* = \gamma\alpha^{-2}$ the second summand in the parentheses above vanishes and we recover the formula (5) for R^*. Now, to verify that λ^* is in fact optimal,

suppose that we have a sequence of problems with *Gaussian* random weights $w \sim \mathcal{N}(0, p^{-1}\alpha^2 I_p)$; this is a special case of assumption RRC. Then ridge regression with tuning parameter $\lambda_p^* = \gamma_p \alpha^{-2}$ yields the Bayes posterior mean for w, which is a Bayes optimal estimator for any quadratic loss, including ℓ_2 prediction error, so that

$$\mathbb{E}\big[r_\lambda(X)\big] \geq \mathbb{E}\big[r_{\lambda_p^*}(X)\big] \qquad \text{for all } \lambda > 0.$$

Now we note that $r_\lambda(X)$ is uniformly bounded in X, and so the convergence statement (6) also holds in expectation, by the bounded convergence theorem. In Lemma B.1 of the supplement, we show that $r_{\lambda_p^*}(X) \to R^*$ almost surely and in expectation. Thus, taking the limit as $p \to \infty$ we conclude that $R_\lambda \geq R^*$ for all $\lambda > 0$, completing the proof. □

We highlight that for $\lambda = \lambda_p^*$ this result is a calculation of limiting Bayes risk—specifically, ℓ_2 predictive risk—in a Bayesian linear model, under high-dimensional asymptotics. A naive calculation of this quantity, assuming for instance that $\widehat{\Sigma}$ is very close to Σ, would lead to a different answer. RMT is used to get the precise result.

The required functionals of the limiting empirical eigenvalue distribution can be written in terms of almost-sure limits of simple quantities. For example, $r_{\lambda_p^*}(X) = \mathrm{Err}(\hat{w}_{\lambda_p^*})$ can equivalently be characterized as

$$\mathrm{Err}(\hat{w}_{\lambda_p^*}) - \left(\frac{\gamma_p^2}{\alpha^2} p^{-1} \operatorname{tr}\left[\left(\widehat{\Sigma} + \frac{\gamma_p}{\alpha^2} I_p\right)^{-1}\right] + 1 - \gamma\right)^{-1} \to_{\text{a.s.}} 0.$$

For general λ, the limiting error rate depends on the almost sure limits of both $p^{-1}\operatorname{tr}((\widehat{\Sigma} + \lambda I_p)^{-1})$ and $p^{-1}\operatorname{tr}((\widehat{\Sigma} + \lambda I_p)^{-2})$. The limit R_λ can be computed efficiently using the methods detailed in the supplement.

To verify the finite-sample accuracy of Theorem 2.1, we perform a simulation with the `BinaryTree` and `Exponential` models. We compute the limit risks using the algorithms in the supplement. The results in Figure 2 show that the formulas given in Theorem 2.1 appear to be accurate, even in small sized problems. In Figure 2, for `BinaryTree` we train on $n = \gamma^{-1} p$ samples, where $p = 2^4$; for `Exponential` on $n = 20$. We set the signal strength to $\alpha^2 = 1$ and generate w, X, and ε as Gaussian random variables with i.i.d. entries. The results are averaged over 500 simulation; we evaluate the empirical prediction error using a test set of size 100.

Intriguingly, Figure 2 shows that the prediction performance of ridge regression is very similar on the two problems. This presents a marked contrast to the RDA example given in the Introduction, where the two covariance structures led to very different classification performance. Thus, the effect of covariance structure depends on the loss function.

In the special case of identity covariance $\Sigma = I_p$, the quantity $R_\lambda - 1$ coincides with estimation error, so we recover known results described in, for example,

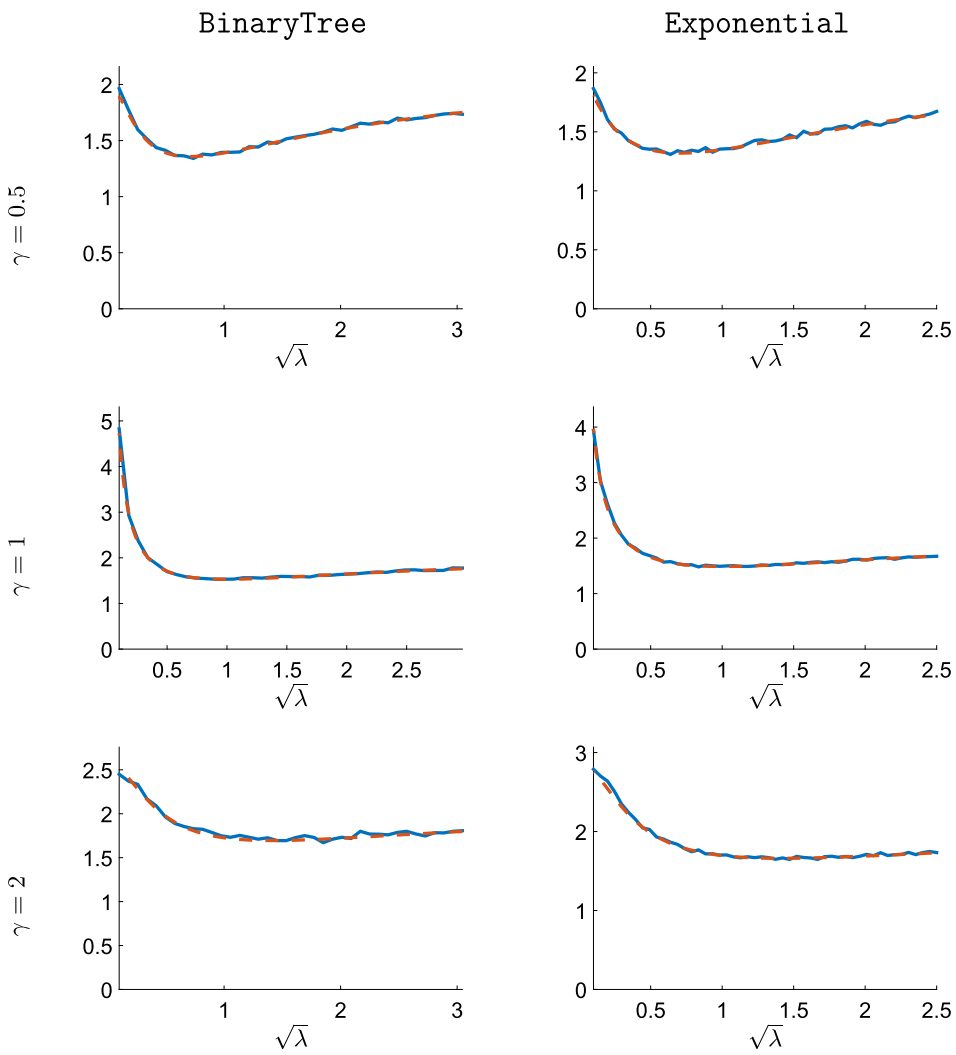

FIG. 2. *Prediction error of ridge regression in the* `BinaryTree` *and* `Exponential` *model. The theoretical formula (red, dashed) is overlaid with the results from simulations (blue, solid). The signals are drawn from* $w \sim \mathcal{N}(0, p^{-1} I_p)$. *For* `BinaryTree`, *we train on* $n = \gamma^{-1} p$ *samples, where* $p = 2^4$; *for* `Exponential` *on* $n = 20$. *We take* 100 *instances of random training data sets, and for each we test on* 500 *samples. We report the average test error over all* 50,000 *test cases.*

Tulino and Verdú (2004). The limit Stieltjes transform has an explicit expression [e.g., Bai and Silverstein (2010), page 52]:

$$m_I(-\lambda;\gamma) = \frac{-(1-\gamma+\lambda) + \sqrt{(1-\gamma+\lambda)^2 + 4\gamma\lambda}}{2\gamma\lambda}, \qquad \lambda > 0. \tag{7}$$

As shown in the supplement, Theorem 2.1 then implies that the limit predictive risk of ridge regression for general λ equals

$$R_\lambda(\alpha^2, \gamma) = 1 + \gamma m_I(-\lambda; \gamma) + \lambda(\lambda\alpha^2 - \gamma)m_I'(-\lambda; \gamma),$$

which has an explicit form. Furthermore, the optimal risk has the particularly simple form

$$R^*(\alpha^2, \gamma) = \frac{1}{2}\left[1 + \frac{\gamma - 1}{\gamma}\alpha^2 + \sqrt{\left(1 - \frac{\gamma - 1}{\gamma}\alpha^2\right)^2 + 4\alpha^2}\right].$$

2.1. *Regimes of learning.* As an application of Theorem 2.1, we study the effect of the signal strength α^2 on the prediction error of ridge regression. Equivalently, this is a study of limiting Bayes prediction risk in linear models. Liang and Srebro (2010) call this the *regimes of learning* problem and argue that, for small α^2 the Bayes error should be characterized by dimension-independent Rademacher bounds, while for large α^2 the error rate should only depend on γ. Liang and Srebro (2010) justify their claims using generalization bounds for the identity-covariance case $\Sigma = I_p$, and conjecture that similar relationships should hold in general. Using our results, we give a precise analysis of regimes of learning with general covariance Σ.

From Theorem 2.1, we know that given a signal strength α^2, the predictive risk of ridge regression with asymptotically optimal regularization converges to

$$R^*(H, \alpha^2, \gamma) = \frac{1}{\lambda^* v(-\lambda^*)} \qquad \text{with } \lambda^*(\alpha, \gamma) = \frac{\gamma}{\alpha^2}. \tag{8}$$

We now use this formula to examine the two limiting behaviors of the risk, for weak and strong signals. In order to make use of (8), we begin by computing some helpful limits involving the companion Stieltjes transform v. The proof of the following lemma is provided in the supplement; recall that, in our notation, H is the limiting *population* spectral distribution.

LEMMA 2.3. *Suppose the limit population eigenvalue distribution H has support contained in a compact set bounded away from 0. Let $v(z)$ be the companion Stieltjes transform of the ESD. Then, in the large λ limit,*

$$\lim_{\lambda\to\infty} \lambda v(-\lambda) = 1 \quad \textit{and} \quad \lim_{\lambda\to\infty} \lambda\big[1 - \lambda v(-\lambda)\big] = \gamma\mathbb{E}_H[T].$$

Meanwhile, in the small λ limit:

1. *If $\gamma < 1$, then $\lim_{\lambda\downarrow 0} \lambda v(-\lambda) = 1 - \gamma$,*
2. *If $\gamma > 1$, then $\lim_{\lambda\downarrow 0} v(-\lambda) = v(0)$, for a positive finite $v(0) > 0$ and*
3. *If $\gamma = 1$, then $\lim_{\lambda\downarrow 0} \lambda v(-\lambda)^2 = \mathbb{E}_H[T^{-1}]$.*

Here, $\mathbb{E}_H[T]$ and $\mathbb{E}_H[T^{-1}]$ denote the large-sample limits of $p^{-1}\operatorname{tr}(\Sigma)$ and $p^{-1}\operatorname{tr}(\Sigma^{-1})$, respectively.

We can now proceed to read off the behavior of ridge regression in the weak- and strong-signal limits. The weak-signal limit is relatively simple. Under the conditions of Theorem 2.1, $\lim_{\alpha^2 \to 0} R^*(H, \alpha^2, \gamma) = 1$, reflecting that for a small signal, we predict a near-zero outcome due to a large regularization. Moreover, by (8) and Lemma 2.3,

$$\lim_{\alpha^2 \to 0} \frac{R^*(H, \alpha^2, \gamma) - 1}{\alpha^2} = \lim_{\lambda \to \infty} \gamma^{-1} \lambda \left(\frac{1}{\lambda v(-\lambda)} - 1 \right) = \mathbb{E}_H[T].$$

Therefore, for small α, the difficulty of the prediction is determined to first order by the average eigenvalue, or equivalently by the average variance of the features, and does not depend on the aspect ratio $\gamma = \lim p/n$.

Conversely, the strong-signal limiting behavior of the risk depends on the aspect ratio γ, and experiences a phase transition at $\gamma = 1$. When $\gamma < 1$, Lemma 2.3 implies that the predictive risk converges to

$$\lim_{\alpha^2 \to \infty} R^*(H, \alpha^2, \gamma) = \lim_{\lambda \to 0} \frac{1}{\lambda v(-\lambda)} = \frac{1}{1 - \gamma}$$

regardless of Σ. In the Gaussian case, this quantity is known to be the $n, p \to \infty$, $p/n \to \gamma$ limit of the risk of ordinary least squares (OLS) [Dicker (2013)]. The same result for non-Gaussian data follows from the Marchenko–Pastur theorem. Thus, when $p < n$ and we have a very strong signal, ridge regression cannot outperform OLS, although of course it can do much better with a small α.

When $\gamma > 1$, the risk $R^*(H, \alpha^2, \gamma)$ can grow unboundedly large with α; and Lemma 2.3 implies that

$$\lim_{\alpha^2 \to \infty} \alpha^{-2} R^*(H, \alpha^2, \gamma) = \lim_{\lambda \to 0} \frac{1}{\gamma v(-\lambda)} = \frac{1}{\gamma v(0)} > 0. \tag{9}$$

Thus, the limiting error rate depends on the covariance matrix through $v(0)$. In general there is no closed-form expression for $v(0)$, which is instead characterized as the unique $c > 0$ for which

$$\frac{1}{\gamma} = \int_{t=0}^{\infty} \frac{tc}{1 + tc} \, dH(t).$$

In the special case $\Sigma = I_p$, however, the limiting expression simplifies to $1/[\gamma v(0)] = (\gamma - 1)/\gamma$. In other words, when $p > n$, optimally tuned ridge regression can capture a constant fraction of the signal, and its test-set fraction of explained variance tends to γ^{-1}.

Finally, in the threshold case $\gamma = 1$, the risk $R^*(H, \alpha^2, \gamma)$ scales with α:

$$\lim_{\alpha^2 \to \infty} \alpha^{-1} R^*(H, \alpha^2, \gamma) = \lim_{\lambda \to 0} \frac{1}{\lambda^{1/2} v(-\lambda)} = \frac{1}{\mathbb{E}_H[T^{-1}]^{1/2}}. \tag{10}$$

Thus, the absolute risk R^* diverges to infinity, but the normalized error $\alpha^{-2} R^*(H, \alpha^2, \gamma)$ goes to 0. This appears to be a rather unusual risk profile. In

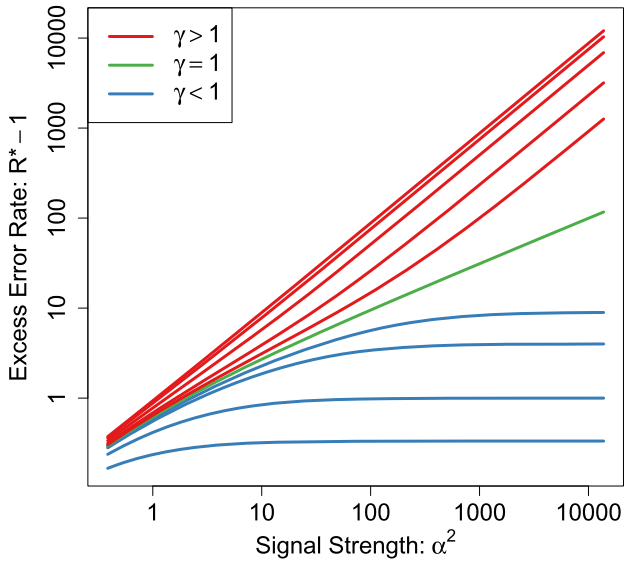

FIG. 3. *Phase transition for predictive risk of ridge regression with identity covariance* $\Sigma = I_p$. *Error rates based on our formulas are plotted for* $\gamma = 0.25, 0.5, 0.8, 0.9, 1, 1.1, 1.3, 2, 4$ *and* 8.

the case $\Sigma = I_p$, our expression simplifies further and we get the finite-α formula $R^*(\alpha^2, 1) = (\sqrt{4\alpha^2 + 1} + 1)/2$, which scales like α.

In summary, we find that for general covariance Σ, the strong-signal risk $R^*(\alpha^2, \gamma)$ scales as $\Theta(1)$ if $\gamma < 1$, as $\Theta(\alpha)$ if $\gamma = 1$, and as $\Theta(\alpha^2)$ if $\gamma > 1$. We illustrate this phenomenon in Figure 3, in the case of the identity covariance $\Sigma = I_p$. We see that when $\gamma < 1$ the error rate stabilizes, whereas when $\gamma > 1$, the error rate eventually gets a slope of 1 on the log–log scale. Finally, when $\gamma = 1$, the error rate has a log–log slope of $1/2$.

Thus, thanks to Theorem 2.1, we can derive a complete and exact answer the regimes of learning question posed by Liang and Srebro (2010) in the case of linear models. The results (9) and (10) not only show that the scalings found by Liang and Srebro (2010) with $\Sigma = I_p$ hold for arbitrary Σ, but make explicit how the slopes depend on the limiting population spectral distribution. The ease with which we were able to read off this scaling from Theorem 2.1 attests to the power of the random matrix approach.

2.2. *An inaccuracy principle for high-dimensional linear models.* Our results also reveal an intriguing inverse relationship between the prediction and estimation errors in high-dimensional linear models. Specifically, denoting the mean-squared estimation error of ridge regression as $R_{E,n}(\lambda) = \mathbb{E}[\|\hat{w}_\lambda - w^*\|^2]$, it is known that optimally tuned ridge regression satisfies, under the conditions of Theorem 2.1,

$$R_{E,n}(\lambda_p^*) \to_{\text{a.s.}} R_E := \gamma m(-\lambda^*) \qquad \text{for } \lambda^* = \gamma \alpha^{-2},$$

where m is the Stieltjes transform of the limiting empirical spectral distribution [see, e.g., Tulino and Verdú (2004), Chapter 3]. This result gives the limiting Bayes *estimation* error in high-dimensional Bayesian linear models. Combining this result with our result on *prediction*, Theorem 2.1 and with the duality relation (3), we find the following relationship between the limiting predictive risk R_P and the limiting estimation risk R_E.

COROLLARY 2.4. *In high-dimensional linear models under the conditions of Theorem 2.1, the asymptotic Bayes predictive and estimation risks are* inversely related. *For all correlation structures, that is, all limit eigenvalue distributions H of the covariance matrices Σ, one has*

$$1 - \frac{1}{R_P} = \gamma\left(1 - \frac{R_E}{\alpha^2}\right).$$

Both sides of the above equation are nonnegative: R_P cannot fall below the intrinsic noise level $\mathrm{Var}[Y \mid X] = 1$, while $R_E \le \limsup_{p\to\infty} R_{E,n}(\lambda^*) \le \limsup_{p\to\infty} R_{E,n}(0) = \alpha^2$. When $\gamma = 1$, we get the even simpler equation

$$R_E R_P = \alpha^2.$$

The product of the estimation and prediction risks equals the signal strength. Since this holds for the optimal λ^*, it also implies that for any λ we have the lower bound $R_E(\lambda) \cdot R_P(\lambda) \ge \alpha^2$; we find the explicit formula relating the two risks remarkable.

The inverse relationship may be somewhat surprising, but it has an intuitive explanation. When the features are highly correlated and v is correspondingly large, prediction is easy because y lies close to the "small" column space of the feature matrix X, but estimation of w is hard due to multicollinearity. As correlation decreases, prediction gets harder but estimation gets easier. A similar heuristic was given by Liang and Srebro (2010), without theoretical justification.

2.3. *Related work for high-dimensional ridge regression.* Random-design ridge regression in high dimensions is a thoroughly studied topic. In particular, El Karoui (2013) and Dicker (2014) study ridge regression with identity covariance $\Sigma = I_p$ in an asymptotic framework similar to ours; this special case is considerably more restrictive than a general covariance. The study of the estimation error $\mathbb{E}[\|\hat{w}_\lambda - w\|^2]$ of ridge regression has received substantial attention in the wireless communication literature; see, for example, Couillet and Debbah (2011) and Tulino and Verdú (2004) for references. To our knowledge, however, that literature has not addressed the behavior of prediction error. Finally, we also note the work of Hsu, Kakade and Zhang (2014), who provide finite-sample concentration inequalities on the prediction error of random-design ridge regression, without obtaining limiting formulas. In contrast, we give explicit limiting formulas for the prediction error.

3. Regularized discriminant analysis. In the second part of the paper, we return to regularized discriminant analysis and the two-class Gaussian discrimination problem (1). For simplicity, we will first discuss balanced populations $\pi_{+1} = \pi_{-1}$. In this case, the Bayes oracle predicts using [Anderson (2003)]

$$\hat{y}(x) = \operatorname{sign}\left(\delta^\top \Sigma^{-1}\left(x - \frac{\mu_{-1} + \mu_{+1}}{2}\right)\right) \qquad \text{with } \delta = \frac{\mu_{+1} - \mu_{-1}}{2},$$

and has an error rate $\mathrm{Err}_{\mathrm{Bayes}} = \Phi(-\Delta_{n,p})$, where $\Delta_{n,p} = \sqrt{\delta^\top \Sigma^{-1}\delta}$ is half the between-class Mahalanobis distance. The Gaussian classification problem has a rich history, going back to Fisher's pioneering work on linear discriminant analysis (LDA). When we have the same number of examples from both the positive and negative classes, that is, $n_{-1} = n_{+1} = n/2$, LDA classifies using the linear rule

$$\hat{y} = \operatorname{sign}\left(\hat{\delta}^\top \widehat{\Sigma}_c^{-1}\left(x - \frac{\hat{\mu}_{-1} + \hat{\mu}_{+1}}{2}\right)\right),$$

where

$$\hat{\delta} = \frac{\hat{\mu}_{+1} - \hat{\mu}_{-1}}{2}, \qquad \widehat{\Sigma}_c = \frac{1}{n-2}\sum_{i=1}^{n}(x_i - \hat{\mu}_{y_i})^{\otimes 2}, \quad \text{and} \quad \hat{\mu}_{\pm 1} = \frac{2}{n}\sum_{\{i:y_i=\pm 1\}} x_i.$$

Here, $\widehat{\Sigma}_c$ is the centered covariance matrix. In the low-dimensional case where n gets large while p remains fixed, LDA converges to the Bayes discrimination function [Anderson (2003), Efron (1975)]. When p is of order n, however, the matrix inverse $\widehat{\Sigma}_c^{-1}$ is unstable and the performance of LDA declines, as discussed among others by Bickel and Levina (2004). Instead, we will study regularized discriminant analysis, defined as the the linear classification rule $\hat{y} = h_{\hat{w}_\lambda}(x - \hat{\mu})$ with $\hat{\mu} = (\hat{\mu}_{-1} + \hat{\mu}_{+1})/2$, $h_w(x) = \operatorname{sign}(w^\top x)$, and $\hat{w}_\lambda = (\widehat{\Sigma}_c + \lambda I_p)^{-1}\hat{\delta}$ [Friedman (1989), Serdobolskii (1983)]. The notation was chosen to emphasize the similarities between ridge regression and RDA; the RDA weight vector $\hat{w}_\lambda$ ought not be confused with the ridge regression weight vector, also denoted $\hat{w}_\lambda$.

3.1. *High-dimensional asymptotics.* Throughout this section, we make a random-effects assumption about the class means. We denote the classification error of RDA as $\mathrm{Err}(\hat{w}_\lambda) = \mathbb{P}[y \neq \operatorname{sign}\{\hat{w}_\lambda^\top(x - \hat{\mu})\}]$. The probability is with respect to an independent test data point (x, y) from the same distribution as the training data.

ASSUMPTION RWC (Random Weights in Classification). The following conditions hold:

1. μ_{-1} and μ_{+1} are randomly generated as $\mu_{-1} = \bar{\mu} - \delta$ and $\mu_{+1} = \bar{\mu} + \delta$, where δ has i.i.d. coordinates with

$$\mathbb{E}[\delta_i] = 0, \qquad \mathrm{Var}[\delta_i] = \frac{\alpha^2}{p}, \quad \text{and} \quad \mathbb{E}\big[\delta_i^{4+\eta}\big] \leq \frac{C}{p^{2+\eta/2}}$$

for some fixed constants $\eta > 0$ and C.

2. $\bar{\mu} = (\mu_{-1} + \mu_{+1})/2$ is either fixed, or random and independent of δ, X and y, and satisfies $\limsup_{p\to\infty} \|\bar{\mu}\|_2^2 / p^{1/2-\zeta} \leq C$ almost surely for some fixed constants $\zeta > 0$ and C.

We say that the eigenvalues of Σ are uniformly bounded if $0 < b < \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq B$ for some fixed constants b and B independently of p.

THEOREM 3.1. *Consider the two-class Gaussian classification problem* (1). *Under parts* 2 *and* 3 *of Assumption* HDA, *and Assumption* RWC, *suppose moreover that the eigenvalues of* Σ *are uniformly bounded. Finally, suppose that we have equal class sizes* $n_{-1} = n_{+1}$. *Then the classification error of RDA converges almost surely*

$$\operatorname{Err}(\hat{w}_\lambda) \to_{\text{a.s.}} \Phi\big(-\Theta(\lambda)\big) \qquad \text{where } \Theta(\lambda) = \frac{\alpha^2 \tau(\lambda)}{\sqrt{\alpha^2 \eta(\lambda) + \xi(\lambda)}}$$

and τ, η *and* ξ *are determined by the limit population spectrum* H *and limit aspect ratio* γ:

$$\tau(\lambda) = \lambda m v, \qquad \eta(\lambda) = \frac{v - \lambda v'}{\gamma}, \qquad \xi(\lambda) = \frac{v'}{v^2} - 1.$$

Here, $m = m(-\lambda)$ *is the Stieltjes transform of the limit empirical spectral distribution* F *of the covariance matrix* $\widehat{\Sigma}_c$, *and* $v = v(-\lambda)$ *is the companion Stieltjes transform defined in* (3).

The proof of Theorem 3.1, provided in Section 3.6, is similar to Theorem 2.1 but more involved. The main difficulty is to evaluate the limits of certain functionals of the population and sample covariance matrices. As a part of the proof, we extend the result of Ledoit and Péché (2011), and build on technical ideas developed by Chen et al. (2011) and Hachem, Loubaton and Najim (2007).

The above result can also be extended to RDA with uneven sampling proportions. Since the limit error rates get more verbose, this is the only place where we discuss uneven sampling. Suppose that the conditions of Theorem 3.1 hold, except now our training set is comprised of $n_{\pm 1}$ samples with label $y_i = \pm 1$ such that $p/n_{\pm 1} \to \gamma_{\pm 1} > 0$. We do not assume that $n_{-1}/(n_{-1} + n_{+1}) \to \pi_-$. Consider a general regularized classifier $\operatorname{sign}(\hat{f}_\lambda(x))$, where $\hat{f}_\lambda(x) = [x - (\hat{\mu}_{+1} + \hat{\mu}_{-1})/2]^\top (\widehat{\Sigma}_c + \lambda I_p)^{-1} [\hat{\mu}_{+1} - \hat{\mu}_{-1}] + c$ for some $c \in \mathbb{R}$, where $\widehat{\Sigma}_c$, $\hat{\mu}_{\pm 1}$ are defined in the usual way. We prove in the supplement, using a similar argument to that of Theorem 3.1, that:

THEOREM 3.2. *Under the conditions of Theorem* 3.1, *and with unequal sampling, the classification error of RDA converges almost surely*:

$$\mathbb{P}\big(\operatorname{sign}\big(\hat{f}_\lambda(x)\big) \neq y\big)\big) \to_{\text{a.s.}} \pi_- \Phi(-\Theta_-) + \pi_+ \Phi(-\Theta_+), \tag{11}$$

where the effective classification margins have the form

$$\Theta_{\pm} = \mp \frac{\pm\alpha^2 m(-\lambda) + \frac{\gamma_{-1}-\gamma_{+1}}{4}\frac{1}{\gamma}(\frac{1}{\lambda v} - 1) + c}{\sqrt{Q}}, \quad \textit{and}$$

$$Q = \alpha^2 \frac{v - \lambda v'}{\gamma(\lambda v)^2} + \frac{\gamma_{-1} + \gamma_{+1}}{4}\frac{v' - v^2}{\lambda^2 v^4}.$$

It is worth mentioning that the regression and classification problems are very different statistically. In the random effects linear model, ridge regression is a linear Bayes estimator, thus the ridge regularization $\widehat{\Sigma} + \lambda I_p$ of the covariance matrix is justified statistically. However, for classification, the ridge regularization is merely a heuristic to help with the ill-conditioned sample covariance. It is thus interesting to know how much this heuristic helps improve upon unregularized LDA, and how close we get to the Bayes error. We now turn to this problem, which can be studied equivalently from a geometric perspective.

3.2. *The geometry of RDA*. The asymptotics of RDA can be understood in terms of a simple picture. The angle between the Bayes decision boundary hyperplane and the RDA discriminating hyperplane tends to an asymptotically deterministic value in the metric of the covariance matrix, and the limiting risk of RDA can be described in terms of this angle.

Recall that, in the balanced case when $n_+ = n_-$, the estimated RDA weight vector is $\hat{w}_\lambda = (\widehat{\Sigma}_c + \lambda I_p)^{-1}\hat{\delta}$, while the Bayes weight vector is $w^* = \Sigma^{-1}\delta$. In the metric induced by the inner product $\langle a, b\rangle_\Sigma = a^\top \Sigma b$, the cosine of the angle between the two is

$$\cos_\Sigma(w^*, \hat{w}_\lambda) = \hat{w}_\lambda^\top \delta / \sqrt{\hat{w}_\lambda^\top \Sigma \hat{w}_\lambda \cdot \delta^\top \Sigma^{-1}\delta}.$$

Now, as seen in the proof of Theorem 3.1,

$$\hat{w}_\lambda^\top \delta / \sqrt{\hat{w}_\lambda^\top \Sigma \hat{w}_\lambda} \to_{\text{a.s.}} \Theta(H, \gamma, \alpha^2, \lambda),$$

where $\Theta(H, \gamma, \alpha^2, \lambda)$ is the classification margin of RDA with the dependence on each parameter made explicit. Meanwhile, as discussed earlier, the Bayes error rate for the two-class Gaussian problem is $\text{Err}_{\text{Bayes}} = \Phi(-\Delta_{n,p})$, and it is easy to see that

$$\Delta_{n,p} = \sqrt{\delta^\top \Sigma^{-1}\delta} \to_{\text{a.s.}} \Delta = \alpha\sqrt{\mathbb{E}_H[T^{-1}]}.$$

Thus, it follows that our angle of interest converges

$$\cos_\Sigma(w^*, \hat{w}_\lambda) \to_{\text{a.s.}} \Gamma(H, \gamma, \alpha^2, \lambda) = \Theta(H, \gamma, \alpha^2, \lambda)/\Delta \in [0, 1],$$

and the limit of its cosine directly quantifies the inefficiency of the RDA estimator relative to the Bayes one.

We gain some insight into this angle for two special cases: when $H = \delta_1$, and by taking the limit $\alpha^2 \to \infty$. First, with $H = \delta_1$, or equivalently $\Sigma = I_p$, we curiously find that the effects of estimating the class means and the covariance matrix decouple completely, as shown in Corollary 3.3 below. The proof is provided in Section 3.7.

COROLLARY 3.3. *Under the conditions of Theorem 3.1, let $\Sigma = I_p$ for all p. Then the limiting cosine Γ of the angle between the Bayes and RDA hyperplanes is*

$$\Gamma(\delta_1, \gamma, \alpha^2, \lambda) = \frac{\alpha}{\sqrt{\alpha^2 + \gamma}} \sqrt{\frac{1 + \gamma\lambda m_I^2(-\lambda; \gamma)}{1 + \gamma m_I(-\lambda; \gamma)}},$$

where the Stieltjes transform $m_I(-\lambda; \gamma)$ for $\Sigma = I_p$ is given in (7). For $\gamma = 1$, this expression simplifies further to

$$\Gamma(\delta_1, 1, \alpha^2, \lambda) = \frac{\alpha}{\sqrt{\alpha^2 + 1}} \frac{2[\lambda(\lambda + 4)]^{1/4}}{\lambda^{1/2} + (\lambda + 4)^{1/2}}.$$

Examining $\Gamma(\delta_1, \gamma, \alpha^2, \lambda)$, we can attribute the suboptimality to two sources of noise: We need to pay a price $\alpha/\sqrt{\alpha^2 + \gamma}$ for estimating $\mu_{\pm 1}$, and a price of $([1 + \gamma\lambda m_I^2(-\lambda; \gamma)]/[1 + \gamma m_I(-\lambda; \gamma)])^{1/2}$ for estimating Σ. If we knew that $\Sigma = I_p$, we could send $\lambda \to \infty$. It is easy to verify that this would send the second term to 1, leading to a loss of efficiency $\alpha/\sqrt{\alpha^2 + \gamma}$.

In the case of a general covariance matrix Σ, we get a similar asymptotic decoupling in the strong-signal limit $\alpha^2 \to \infty$. The following claim follows immediately from Theorem 3.1.

COROLLARY 3.4. *Under the conditions of Theorem 3.1, the cosine of the angle between the optimal and learned hyperplanes has the limit as $\alpha^2 \to \infty$:*

$$\lim_{\alpha \to \infty} \Gamma(H, \gamma, \alpha^2, \lambda) = \frac{\tau(\lambda)}{\sqrt{\eta(\lambda)\mathbb{E}_H[T^{-1}]}}.$$

Thus, RDA is in general inconsistent for the Bayes hyperplane in the case of strong signals. Corollary 3.4 also implies that, in the limit $\alpha \to \infty$, the optimal λ for RDA converges to a nontrivial limit that only depends on the spectral distribution H. No such result is true for ridge regression, where $\lambda^* = \alpha^{-2}\gamma \to 0$ as $\alpha \to \infty$, regardless of Σ.

We illustrate the behavior of the cosine Γ for the `AR-1(0.9)` model in Figure 4, which displays Γ for values of α ranging from $\alpha = 0.1$ to $\alpha = 2$. We see that the Γ-curve converges to its large-α limit fairly rapidly. Moreover, somewhat unexpectedly, the optimal regularization parameter λ^*, that is, the maximizer of Γ, increases with the signal strength α^2.

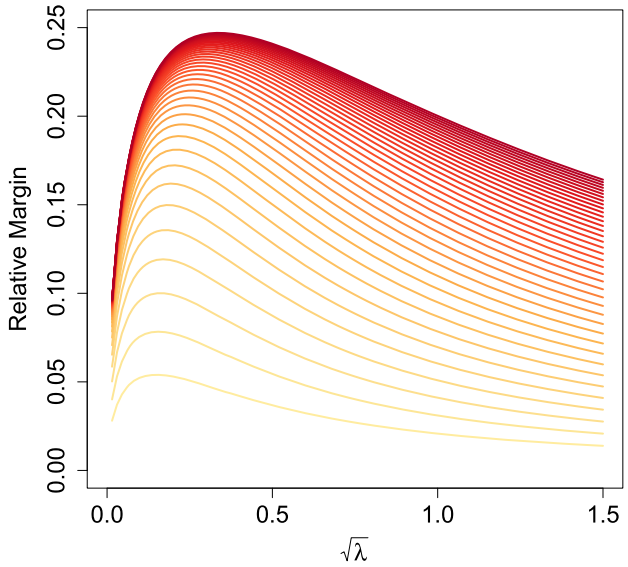

FIG. 4. *The cosine* $\Gamma(H, \gamma, \alpha^2, \lambda)$ *for the* AR-1(0.9) *model, with* $\alpha \in [0.1, 2]$. *The values of* α *used for each curve are evenly spaced, with a gap of* 0.05 *between each curve. The cosine quickly converges to a limit as* α *increases.*

Finally, we note that Efron (1975) studies the angle Γ in detail for low-dimensional asymptotics where p is fixed while $n \to \infty$; in this case, Γ converges in probability to 1, and the sampling distribution of $n(1-\Gamma)$ converges to a (scaled) χ^2_{p-1} distribution. Establishing the sampling distribution in high dimensions is interesting future work.

3.3. *Do existing theories explain the behavior of RDA?.* Theorems 3.1 and 3.2 give precise information about the error rate of RDA in our model. It is of interest to compare this to classical theories, such as Vapnik–Chervonenkis theory or Rademacher bounds, to see if they explain the behavior of RDA. In this section, we study a simulation example, and conclude that existing theory does not precisely explain the behavior of RDA.

We consider a setup with $n = p = 500$, equal class sizes, Σ an auto-regressive (AR-1) matrix such that $\Sigma_{ij} = \rho^{|i-j|}$, and $\mu_{\pm 1} \sim \mathcal{N}(0, \alpha^2 p^{-1} I_p)$. This is a natural model when the features can be ordered such that correlations decay with distance; for instance in time series and genetic data. We run experiments for different values of ρ in two settings: once with constant effect size $\alpha^2 = 1$, and once with constant oracle margin $\sqrt{\mathbb{E}[\Delta^2_{n,p}]} = 2.3$. Given $\alpha \geq 0$ and $\rho \in [0, 1)$, one can verify using the description of the limit population spectrum [Grenander and Szegő (1984)], that the limiting oracle classification margin in the AR-1 model is $\Delta = \alpha\sqrt{(1+\rho^2)/(1-\rho^2)}$; thus, with constant α^2 the oracle classifier improves as ρ increases.

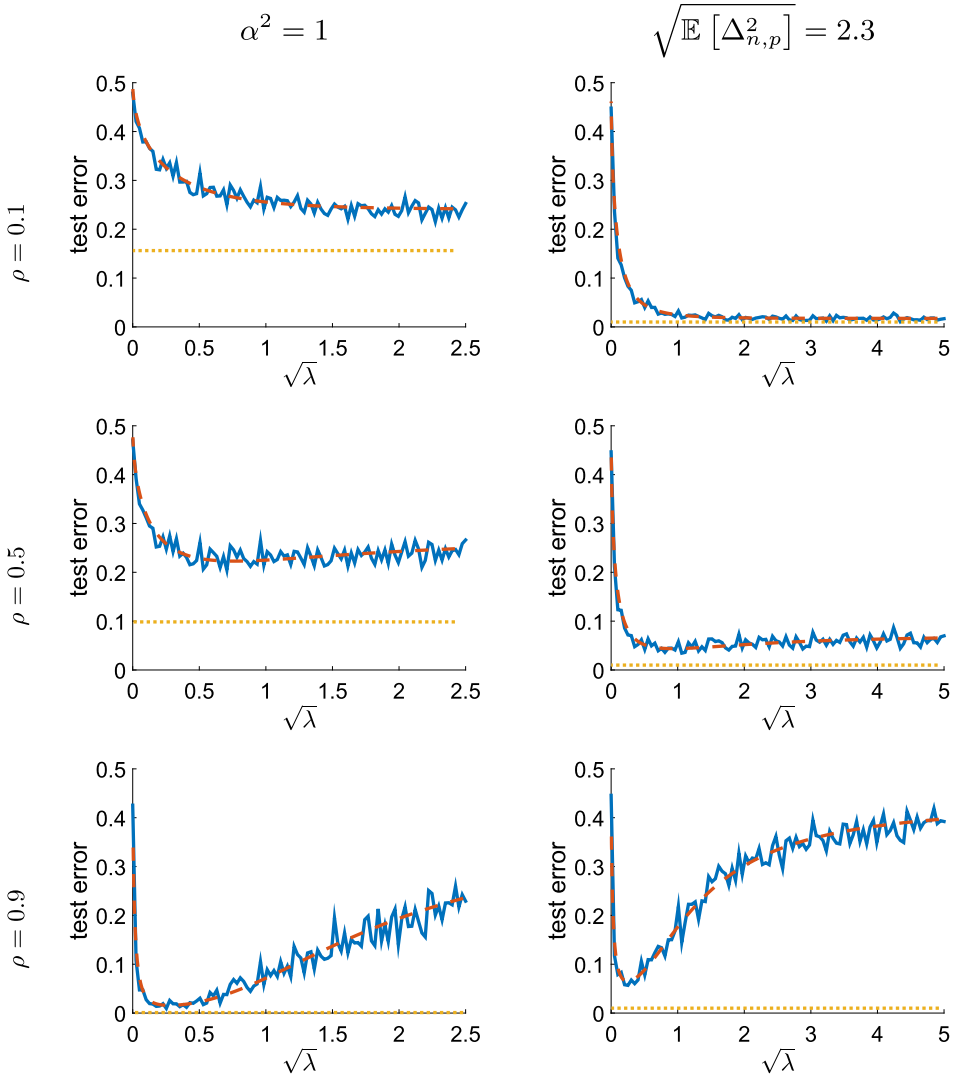

FIG. 5. *Classification error of RDA in an* AR-1 *model. The theoretical formula* (*red*, *dashed*) *is overlaid with the results from simulations* (*blue*, *solid*; *we also display the oracle error* (*yellow*, *dotted*). *In the first column*, *we keep the signal strength fixed at* $\alpha^2 = 1$, *whereas in the second column we picked* α *such as to fix the oracle error at* $\mathrm{Err}_{\mathrm{Bayes}} = 0.01$. *We test on* 10,000 *new samples*, *and report the average classification error.*

Existing results give us some intuition about what to expect. Since $n = p$, classical heuristics based on the theory of Vapnik and Chervonenkis (1971) as well as more specialized analyses [Bickel and Levina (2004), Saranadasa (1993)] predict that unregularized LDA will not work. As we will see, this matches our simulation

results. Meanwhile, Bickel and Levina (2004) study worst-case performance of the independence rule relative to the Bayes rule. In our setting, it can be verified that their results imply $\Theta_{\mathrm{IR}} \geq (1-\rho^2)/(1+\rho^2)\Delta$, where the error rate of the independence rule is $\Phi(-\Theta_{\mathrm{IR}})$. This predicts that independence rules will work better for small correlation ρ, which again will match the simulations.

The existing theory, however, is much less helpful for understanding the behavior of RDA for intermediate values of λ. A learning theoretic analysis based on Rademacher complexity suggests that the generalization performance of RDA should depend on terms that scale like $\sqrt{\|\hat{w}_\lambda\|_2^2 \operatorname{tr}\Sigma/n} \asymp \sqrt{\lambda^{-2}p/n}$ for large values of λ [e.g., Bartlett and Mendelson (2003)]. In other words, based on a classical approach, we might expect that mildly regularized RDA should not work, but using a large λ may help. Rademacher theory is not tight enough to predict what will happen for $\lambda \approx 1$.

Given this background, Figure 5 displays the performance of RDA for different values of ρ, along with our theoretically derived error from Theorem 3.1. In the $\alpha^2 = 1$ case, we find that—as predicted—unregularized LDA does poorly. However, when ρ is large, mildly regularized RDA does quite well.

Strikingly, RDA is able to benefit from the growth of the oracle classification margin with ρ, but only if we use a small positive value of λ. The analyses based on unregularized LDA or "infinitely regularized" independence rules do not cover this case. Moreover, this phenomenon is not predicted by Rademacher theory, which requires $\lambda \gg 1$ to improve over basic Vapnik–Chervonenkis bounds. Results from the constant margin case $\sqrt{\mathbb{E}[\Delta_{n,p}^2]} = 2.3$ reinforce the same interpretations. Finally, our formulas for the error rate are accurate despite the moderate sample size $n = p = 500$. In conclusion, our results describe the behavior of RDA much more precisely than existing general learning-theoretic analyses, under the random-effects models considered here.

3.4. *Linear discriminant analysis versus independence rules.* Two points along the RDA risk curve that allow for particularly simple analytic expressions occur as $\lambda \to 0$ and $\lambda \to \infty$: the former is just classical linear discriminant analysis while the latter is equivalent to an independence rule (or "naïve Bayes"). In this section, we show that by taking these limits we can recover known results about the high-dimensional asymptotics of LDA and naïve Bayes. Further, we compare these two methods over certain parameter classes.

Note that $\lambda \to \infty$ leads to a linear discriminant rule with weight vector $\hat{\delta} = \hat{\mu}_{+1} - \hat{\mu}_{-1}$. Usual independence rules take the form $\operatorname{diag}(\widehat{\Sigma}_c)^{-1}\hat{\delta}$. We will assume that all features are normalized to have equal variance, $\Sigma_{ii} = \sigma > 0$. In this case, the $\lambda \to \infty$ rule corresponds to an independence rule with oracle information about the equality of variances; which we still call "independence rule" for simplicity.

Extending our previous notation, we define the asymptotic margin of LDA and independence rules, by taking the limits of $\Theta(\lambda)$ at 0 and ∞:

$$\Theta_{\mathrm{LDA}} = \lim_{\lambda\to 0} \frac{\alpha^2\tau(\lambda)}{\sqrt{\alpha^2\eta(\lambda)+\xi(\lambda)}} \quad \text{and} \quad \Theta_{\mathrm{IR}} = \lim_{\lambda\to\infty} \frac{\alpha^2\tau(\lambda)}{\sqrt{\alpha^2\eta(\lambda)+\xi(\lambda)}}.$$

Both limits are well defined and admit simple expressions, as given below; this result is proved in Section 3.8. Let H be the limit population spectral distribution of the covariance matrices Σ; and let T be a random variable with distribution H.

THEOREM 3.5. *Under the conditions of Theorem 3.1, the margins of LDA and independence rules are equal to*

$$\Theta_{\mathrm{LDA}} = \frac{\alpha^2\sqrt{1-\gamma}\,\mathbb{E}_H[T^{-1}]}{\sqrt{\alpha^2\mathbb{E}_H[T^{-1}]+\gamma}} \quad \textit{and} \quad \Theta_{\mathrm{IR}} = \frac{\alpha^2}{\sqrt{\alpha^2\mathbb{E}_H[T]+\gamma\mathbb{E}_H[T^2]}}.$$

The formula for LDA is valid for $\gamma < 1$ while that for IR is valid for any γ.

The formulas are simpler than Theorem 3.1, as they involve the population spectral distribution H directly through its moments. For RDA, the error rate depends on H implicitly through the Stieltjes transform of the ESD F.

These formulas are equivalent to known results, some of which were obtained under slightly different parametrization. Raudys (1967) obtained the formula for IR with $H = \delta_1$, while the LDA formula was derived by Deev (1970) and Raudys (1972); see Section 3.5 for a more detailed historical account. Here, our goal was to show how these simple formulas can be recovered from the more powerful Theorem 3.1.

Saranadasa (1993) also obtains closed-form expressions for the limit risk of two classification methods, the D-criterion and the A-criterion. One can verify that these are asymptotically equivalent to LDA and IR, respectively. Our results are consistent with those of Saranadasa (1993); but they differ slightly in the modeling assumptions. In our notation, his results (as stated in his Theorem 3.2 and Corollary 3.1) are: $\Theta^S_{\mathrm{LDA}} = \alpha\sqrt{\mathbb{E}_H[T^{-1}]}\sqrt{1-\gamma}$ and $\Theta^S_{\mathrm{IR}} = \alpha/\sqrt{\mathbb{E}_H[T]}$. These results are nearly identical to Theorem 3.5, but our equations have an extra term involving γ in the denominator: γ for LDA and $\gamma\mathbb{E}_H[T^2]$ for IR. The reason is that we consider $\mu_{\pm 1}$ as random, whereas Saranadasa (1993) considers them as fixed sequences of vectors; this extra randomness yields additional variance terms.

Theorem 3.5 enables us to compare the worst-case performance of LDA and IR over suitable parameter classes of limit spectra. For $0 < k_1 \le 1 \le k_2$, we define the class

$$\mathcal{H}(k_1,k_2) = \{H : \mathbb{P}_H([k_1,k_2]) = 1, \mathbb{E}_H[T] = 1\}.$$

The bounds k_1, k_2 control the ill-conditioning of the population covariance matrix. We normalize such that the average population eigenvalue is 1, to ensure that the

scaling of the problem does not affect the answer. This parameter space is somewhat similar to the one considered by Bickel and Levina (2004). A direct comparison over these natural problem classes appears to be missing from the literature, and so we provide it below.

COROLLARY 3.6. *Under the conditions of Theorem 3.5, consider the behavior of LDA and independence rules for* $H \in \mathcal{H}(k_1, k_2)$:

1. *The worst-case margin of LDA is*

$$\bar{\Theta}_{\mathrm{LDA}}(\gamma; \alpha^2) := \inf_{H \in \mathcal{H}(k_1,k_2)} \Theta_{\mathrm{LDA}}(H, \gamma; \alpha^2) = \frac{\alpha^2\sqrt{1-\gamma}}{\sqrt{\alpha^2+\gamma}}.$$

   The least favorable distribution for LDA from the class $\mathcal{H}(k_1, k_2)$ *is the point mass at* 1: $H = \delta_1$, *that is,* $\Sigma = I_p$.
2. *The worst-case margin for independence rules is*

$$\bar{\Theta}_{\mathrm{IR}}(\mathcal{H}, \gamma; \alpha^2) := \inf_{H \in \mathcal{H}(k_1,k_2)} \Theta_{\mathrm{IR}}(H, \gamma; \alpha^2) = \frac{\alpha^2}{\sqrt{\alpha^2+\gamma(k_1+k_2-k_1k_2)}}.$$

   If $k_1 < k_2$, *the least favorable distribution is the mixture* $H = w_1\delta_{k_1} + w_2\delta_{k_2}$, *where the weights are* $w_1 = (k_2-1)/(k_2-k_1)$ *and* $w_2 = (1-k_1)/(k_2-k_1)$; *while if* $k_1 = k_2 = 1$, *it is the point mass at* 1: $H = \delta_1$.

PROOF. From Theorem 3.5, minimizing Θ_{LDA} is equivalent to minimizing $\mathbb{E}_H[T^{-1}]$ for $H \in \mathcal{H}(k_1, k_2)$. By Jensen's inequality, $\mathbb{E}_H[T^{-1}] \geq 1/\mathbb{E}_H[T] = 1$; with equality if $H = \delta_1$. This shows the first claim.

Next, again by Theorem 3.5, minimizing Θ_{IR} over $H \in \mathcal{H}(k_1, k_2)$ amounts to maximizing $\mathbb{E}_H[T^2]$ over that class. For this, note that $k_1 \leq T \leq k_2$ for a random variable T distributed according to $H \in \mathcal{H}(k_1, k_2)$. Therefore, $(T-k_1)(T-k_2) \leq 0$, and taking expectations we get the upper bound:

$$\mathbb{E}_H[T^2] \leq (k_1+k_2)\mathbb{E}_H[T] - k_1k_2 = k_1 + k_2 - k_1k_2.$$

This upper bound is achieved for any $H = w_1\delta_{k_1} + w_2\delta_{k_2}$. The weights w_i given in the corollary are required so that H has unit mean. □

This result shows a stark contrast between the worst-case behavior of LDA and independence rules: for fixed signal strength, the worst-case risk of LDA over $\mathcal{H}$ only depends on γ, and is attained with the limit of identity covariances $\Sigma = I_p$ regardless of the values of k_1, k_2. In contrast, the worst-case behavior of IR occurs for a least favorable distribution H that is as highly spread as possible. This highlights the sensitivity of IR to ill-conditioned covariance matrices. For

$0 < \gamma < 1$, we see that IR is better than LDA in the worst case over $\mathcal{H}$, that is, $\bar{\Theta}_{\mathrm{LDA}}(\gamma; \alpha^2) < \bar{\Theta}_{\mathrm{IR}}(\mathcal{H}, \gamma; \alpha^2)$, if and only if

$$\alpha^2 + 1 > (1 - \gamma)(k_1 + k_2 - k_1 k_2).$$

In particular, IR performs better than LDA for strong signals α; with weaker signals, LDA can sometimes have an edge, particularly if the covariance is poorly conditioned, quantified by a large measure of spread $k_1 + k_2 - k_1 k_2 = (k_2 - 1)(1 - k_1) + 1$.

3.5. *Literature review for high-dimensional RDA.* There has been substantial work in the former Soviet Union on high-dimensional classification; references on this work include Raudys and Young (2004), Raudys (2001), and Serdobolskii (2007). Raudys (1967) derived the $n, p \to \infty$ asymptotic error rate of independence rules in identity-covariance case $\Sigma = I_p$, while Deev (1970) and Raudys (1972) obtained the error rate of unregularized linear discriminant analysis (LDA) for general covariance Σ, again in the $n, p \to \infty$ regime. A difference is that Raudys (1972) establishes normality of the linear discriminant function, whereas Deev (1970) expands the conditional probability of misclassification.

Serdobolskii (2007) calls the framework $n, p \to \infty$, $p/n \to \gamma$ the "Kolmogorov asymptotic regime," and suggests that around 1967 Kolmogorov was interested in this area. As explained by one of our referees, Kolmogorov had suggested the problem of studying Fisher's LDA under $n, p \to \infty$ asymptotics to Y. Blagovechenskij and his PhD student, A. Deev.

For RDA, Serdobolskii (1983) [see also Chapter 5 of Serdobolskii (2007)] considered a more general setting than this paper: classification with a weight vector of the form $\Gamma(\widehat{\Sigma}_c)^{-1}\hat{\delta}$ instead of just $(\widehat{\Sigma}_c + \lambda I_p)^{-1}\hat{\delta}$, where the scalar function Γ admits the integral representation $\Gamma(x) = \int (x + t)^{-1}\, d\eta(t)$ for a suitable measure η, and is extended to matrices in the usual way. He derived a limiting formula for the error rate of this classifier under high-dimensional asymptotics. However, his results are substantially more involved and much less explicit than ours. In some cases, it is unclear to us how one could numerically compute his formulas. Furthermore, his results are proved when $\gamma < 1$, and show convergence in probability, not almost surely. We also note the work of Raudys and Skurichina (1995), who derived results about the risk of usual RDA with vanishingly small regularization $\lambda = o(1)$, and for the special case $\gamma < 1$.

In another line of work, a Japanese school [e.g., Fujikoshi, Ulyanov and Shimizu (2011), and references therein] has studied the error rates of LDA and RDA under high-dimensional asymptotics, with a focus on obtaining accurate higher-order expansions of the risk. For instance, Fujikoshi and Seo (1998) obtained asymptotic expansions for the error rate of unregularized LDA, which can be verified to be equivalent to our results in the $\lambda \to 0$ limit. More recently, Kubokawa, Hyodo and Srivastava (2013) obtained a second-order expansion of the error rate of RDA with vanishingly small regularization parameter $\lambda = O(1/n)$ in the case $\gamma < 1$.

Finally, in the signal processing and pattern recognition literature, Zollanvari, Braga-Neto and Dougherty (2011) provided asymptotic moments of estimators of the error rate of LDA, under an asymptotic framework where $n, p \to \infty$; however, this paper assumes that the covariance matrix Σ is known. More recently, Zollanvari and Dougherty (2015) provided consistent estimators for the error rate of RDA in a doubly asymptotic framework, using deterministic equivalents for random matrices. The goal of our work is rather different from theirs, in that we do not seek empirical estimators of the error rate of RDA, but instead seek simple formulas that help us understand its behavior.

3.6. *Proof of main result for RDA.* In this section we begin with an outline of the argument for Theorem 3.1 that motivates several technical lemmas, whose proof can be found in the supplementary materials. Given these technical results, at the end of this section we provide the proof of Theorem 3.1 with the details filled in.

In the Gaussian model (1), it is well known that the expected test error of an arbitrary linear classifier $h_{w,b}(x) = \text{sign}(w^\top x + b)$ is

$$\text{Err}_0(w, b) = \pi_- \Phi\left(\frac{w^\top \mu_{-1} + b}{\sqrt{w^\top \Sigma w}}\right) + \pi_+ \Phi\left(-\frac{w^\top \mu_1 + b}{\sqrt{w^\top \Sigma w}}\right). \tag{12}$$

conditional on the weight parameters w, b and the means $\mu_{\pm 1}$. The strategy is to prove that the weight parameters estimated from the training data converge to the desired limits. In the case of RDA, the relevant weight parameters are

$$\hat{w}_\lambda = (\widehat{\Sigma}_c + \lambda I_p)^{-1} \hat{\delta} \quad \text{and} \quad \hat{b}_\lambda = \hat{\delta}^\top (\widehat{\Sigma}_c + \lambda I_p)^{-1} \hat{\mu}.$$

Moreover, we can asymptotically ignore the offset term $\hat{b}_\lambda$.

LEMMA 3.7. *Under the conditions of Theorem 3.1, we have* $\hat{b}_\lambda \to_{\text{a.s.}} 0$.

This lemma suggests that we should be able to use the following simpler formula—that does not involve an offset b—in evaluating the limit of the error rate:

$$\text{Err}_1(w) = \pi_- \Phi\left(\frac{w^\top \mu_{-1}}{\sqrt{w^\top \Sigma w}}\right) + \pi_+ \Phi\left(-\frac{w^\top \mu_{+1}}{\sqrt{w^\top \Sigma w}}\right). \tag{13}$$

Recall that $\mu_{-1} = \bar{\mu} - \delta$, $\mu_{+1} = \bar{\mu} + \delta$. The second simplification we notice that we can also asymptotically ignore cross-terms of the form $\hat{w}_\lambda^\top \bar{\mu}$.

LEMMA 3.8. *Under the conditions of Theorem 3.1, we have* $\hat{w}_\lambda^\top \bar{\mu} \to_{\text{a.s.}} 0$.

Again, we may now hope to use the following even simpler formula that does not involve $\bar{\mu}$ in evaluating the limit of the error rate:

$$\text{Err}_2(w) = \Phi\left(-\frac{w^\top \delta}{\sqrt{w^\top \Sigma w}}\right). \tag{14}$$

To establish convergence of this quantity, we argue that the linear and quadratic forms involving $\hat{w}_\lambda$ concentrate around their means, and then apply random-matrix results to find the limits of those means. We start with the numerator.

LEMMA 3.9. *Under the conditions of Theorem 3.1,* $\hat{w}_\lambda^\top \delta \to_{\text{a.s.}} \alpha^2 m(-\lambda)$, *where* $m(z)$ *is the Stieltjes transform of the limit empirical eigenvalue distribution* F *of the covariance matrix* $\widehat{\Sigma}_c$.

Finding the limit of the denominator is slightly more involved. We begin by decomposing the quadratic form as

$$\hat{w}_\lambda^\top \Sigma \hat{w}_\lambda = \hat{\delta}^\top (\widehat{\Sigma}_c + \lambda I_p)^{-1} \Sigma (\widehat{\Sigma}_c + \lambda I_p)^{-1} \hat{\delta} = \tilde{A} + 2\tilde{B} + \tilde{C}, \tag{15}$$

where $M := (\widehat{\Sigma}_c + \lambda I_p)^{-1} \Sigma (\widehat{\Sigma}_c + \lambda I_p)^{-1}$ and

$$\tilde{A} := \delta^\top M \delta, \qquad \tilde{B} := \delta^\top M(\hat{\delta} - \delta), \qquad \tilde{C} := (\hat{\delta} - \delta)^\top M(\hat{\delta} - \delta).$$

One can show $\tilde{B} \to_{\text{a.s.}} 0$ similar to the analysis of the error terms in the proof of Lemmas 3.7 and 3.9; we omit the details. The two remaining terms will converge to nonzero quantities. First, we show the following.

LEMMA 3.10. *Under the conditions of Theorem 3.1,* $\tilde{A} := \delta^\top M \delta \to_{\text{a.s.}} \alpha^2 |\kappa'(\lambda)|$, *where*

$$\kappa(\lambda) = \frac{1}{\gamma}\left(\frac{1}{\lambda v(-\lambda)} - 1\right),$$

and $v(-\lambda)$ *is the companion Stieltjes transform of the ESD of the covariance matrix, defined in* (3). *Expressing the derivative explicitly, we have the limit* $\tilde{A} \to_{\text{a.s.}} (v - \lambda v')/[\gamma(\lambda v)^2]$. *The limit is strictly positive.*

The proof of the above lemma relies on the result of Ledoit and Péché (2011), and a derivative trick similar to that employed in a similar context by El Karoui and Kösters (2011), Rubio, Mestre and Palomar (2012) and Zhang et al. (2013). Finally, the last statement that we need is the following lemma, which can be proved by building on results of Hachem et al. (2008) and Chen et al. (2011).

LEMMA 3.11. *Under the conditions of Theorem 3.1,*

$$\tilde{C} := (\hat{\delta} - \delta)^\top M(\hat{\delta} - \delta) \to_{\text{a.s.}} \frac{v' - v^2}{\lambda^2 v^4},$$

where v *the companion Stieltjes transform of the ESD of the covariance matrix, defined in* (3).

With all these results, we are now ready to prove our main result.

PROOF OF THEOREM 3.1. By the decomposition (15) and Lemmas 3.10 and 3.11, we have the convergence

$$\hat{w}_\lambda^\top \Sigma \hat{w}_\lambda \to_{\text{a.s.}} \alpha^2 \frac{v - \lambda v'}{\gamma(\lambda v)^2} + \frac{v' - v^2}{\lambda^2 v^4}.$$

By Lemma 3.11, the second term is strictly positive. Therefore, combining with Lemma 3.9 and the continuous mapping theorem, we have

$$\frac{\hat{w}_\lambda^\top \delta}{\sqrt{\hat{w}_\lambda^\top \Sigma \hat{w}_\lambda}} \to_{\text{a.s.}} \frac{\alpha^2 m(-\lambda)}{[\alpha^2 \frac{v-\lambda v'}{\gamma(\lambda v)^2} + \frac{v'-v^2}{\lambda^2 v^4}]^{1/2}}. \tag{16}$$

Denote by Θ the parameter on the right-hand side. After algebraic simplification, we obtain that Θ has exactly the form stated in the theorem for the margin of RDA. To complete the proof, we show that the error rate is indeed determined by Θ. From (16) and the continuous mapping theorem, recalling the error rate $\text{Err}_2(w)$ from (14), we have $\text{Err}_2(\hat{w}_\lambda) \to_{\text{a.s.}} \Phi(-\Theta)$.

From Lemma 3.8 and the definition of the error rate $\text{Err}_1(w)$ from (13), we can move from Err_2 to Err_1:

$$\text{Err}_2(\hat{w}_\lambda) - \text{Err}_1(\hat{w}_\lambda) \to_{\text{a.s.}} 0.$$

Meanwhile, from Lemma 3.7 and the definition of the error rate $\text{Err}_0(w)$ in Equation (12), we can discard the offset $\hat{b}$, and move from Err_1 to Err_0:

$$\text{Err}_1(\hat{w}_\lambda) - \text{Err}_0(\hat{w}_\lambda, \hat{b}) \to_{\text{a.s.}} 0.$$

The last three statements imply that $\text{Err}_0(\hat{w}_\lambda, \hat{b}) \to_{\text{a.s.}} \Phi(-\Theta)$, which completes the proof of Theorem 3.1. □

3.7. *Proof of Corollary 3.3.* From the proof of Theorem 3.1, we know that the limiting error rate of RDA is $\Phi(-\Theta)$, with $\Theta = \alpha^2 m(-\lambda)/\sqrt{\alpha^2 r(\lambda) + q(\lambda)}$, and

$$r(\lambda) = \lim_{p\to\infty} \mathbb{E} p^{-1} \operatorname{tr}\big(\Sigma(\widehat{\Sigma}_c + \lambda I)^{-2}\big), \quad \text{and}$$

$$q(\lambda) = \lim_{p\to\infty} \mathbb{E} p^{-1} \operatorname{tr}\big(\big[\Sigma(\widehat{\Sigma}_c + \lambda I)^{-1}\big]^2\big).$$

Since $\Sigma = I_p$, we have that $r(\lambda) = q(\lambda)$; moreover, analogously to the argument in the proof of Lemma 3.11, this limit equals $m_I'(-\lambda; \gamma)$. Therefore, we find that

$$\Theta = \frac{\alpha^2}{\sqrt{\alpha^2 + \gamma}} \frac{m_I(-\lambda; \gamma)}{\sqrt{m_I'(-\lambda; \gamma)}}.$$

The quantity $m_I'(-\lambda; \gamma)$ can be expressed in terms of m_I by differentiating the Marchenko–Pastur equation $m_I(z; \gamma) = 1/(1 - z - \gamma - \gamma z m_I(z; \gamma))$, thus yielding $m' = m^2(1 + \gamma m)/(1 - \gamma z m^2)$ and leading to the claimed expression for Θ. The special-case formula for $\gamma = 1$ follows from elementary calculations starting from the expression (7) for $m_I(z; \gamma)$.

3.8. *Proof of Theorem 3.5.* The strategy is to compute the limits of $\tau(\lambda)$, $\eta(\lambda)$ and $\xi(\lambda)$, by first finding the limits of appropriate simpler quantities. It is helpful to represent the Stieltjes transforms and their derivatives as expectations with respect to the ESD, similar to Section 2.1. Thus, let Y be a random variable distributed according to the ESD F, and let $\underline{Y}$ be a random variable distributed according to the companion ESD $\underline{F}$. Then m, v are the Stieltjes transforms of Y and $\underline{Y}$, respectively. We note the following simple expressions, which will be used repeatedly: $m(-\lambda) = \mathbb{E}[1/(Y+\lambda)]$, $m'(-\lambda) = \mathbb{E}[1/(Y+\lambda)^2]$ and $m(-\lambda) - \lambda m'(-\lambda) = \mathbb{E}[Y/(Y+\lambda)^2]$. Furthermore, $\lambda v(-\lambda) = 1 + \gamma(\lambda m(-\lambda) - 1)$, so $\lambda v(-\lambda) = 1 - \gamma\mathbb{E}[Y/(Y+\lambda)]$.

3.8.1. *The error rate of LDA.* As is well known [e.g., Bai and Silverstein (2010)], for $\gamma < 1$, Y is supported on a compact set bounded away from 0, so $Y > c > 0$ for some c. This will allow us to take the limits as $\lambda \downarrow 0$ inside the expectation, using the dominated convergence theorem; we will not repeat this fact. Therefore, we can evaluate the limits of $\tau(\lambda)$ and $\eta(\lambda)$ by noting that $\lim_{\lambda\downarrow 0} m(-\lambda) = \mathbb{E}[Y^{-1}]$, $\lim_{\lambda\downarrow 0}[v(-\lambda) - \lambda v'(-\lambda)]\gamma^{-1} = \lim_{\lambda\downarrow 0} m(-\lambda) - \lambda m'(-\lambda) = \mathbb{E}[Y^{-1}]$, and from Lemma 2.3 we have $\lim_{\lambda\downarrow 0} \lambda v(-\lambda) = 1 - \gamma$.

To evaluate the limit of $\xi(\lambda)$, we differentiate the formula for the companion Stieltjes transform, and see $\lambda^2 v'(-\lambda) = 1 + \gamma(\lambda^2 m'(-\lambda) - 1)$. Hence, $\lim_{\lambda\downarrow 0} v'(-\lambda)/v^2(-\lambda)$ equals

$$\lim_{\lambda\downarrow 0} \frac{1+\gamma(\lambda^2 m'(-\lambda)-1)}{\lambda^2 v^2(-\lambda)} = \frac{\lim_{\lambda\downarrow 0}[1+\gamma(\lambda^2 m'(-\lambda)-1)]}{(1-\gamma)^2} = \frac{1}{1-\gamma},$$

where we have used that $\lim_{\lambda\downarrow 0} \lambda^2 m'(-\lambda) = \lim_{\lambda\downarrow 0} \mathbb{E}[\lambda^2/(Y+\lambda)^2] = 0$. We conclude that $\lim_{\lambda\downarrow 0} \xi(\lambda) = \gamma/(1-\gamma)$. Putting everything together, we find

$$\Theta_{\mathrm{LDA}} = \lim_{\lambda\downarrow 0} \frac{\alpha^2\tau}{\sqrt{\alpha^2\eta+\xi}} = \frac{\alpha^2\mathbb{E}[Y^{-1}](1-\gamma)}{\sqrt{\alpha^2\mathbb{E}[Y^{-1}]+\gamma/(1-\gamma)}}.$$

Let T be a random variable distributed according to the PSD H. By taking the limit as $z \to 0$, $z \in \mathbb{C}^+$ in the Marchenko–Pastur equation

$$m(z) = \int_{t=0}^{\infty} \frac{dH(t)}{t(1-\gamma-\gamma z m(z)) - z},$$

we find $m(0) = \int 1/[t(1-\gamma)]\,dH(t)$, or equivalently $\mathbb{E}[Y^{-1}] = \mathbb{E}[T^{-1}]/(1-\gamma)$. We see that m is well-defined, bounded away from 0 and has positive imaginary part for $z \in \mathbb{C}^+$ in a neighborhood of 0, so the limit is justified by the dominated convergence theorem. This leads to the formula for Θ_{LDA}.

3.8.2. *The error rate of IR.* Since the Stieltjes transform $\mathbb{E}[1/(Y+\lambda)]$ decays as $1/\lambda$ for $\lambda \to \infty$, we will normalize by λ. First, we evaluate the limit of $\lambda\tau(\lambda) = \lambda^2 m(-\lambda)v(-\lambda)$ as $\lambda \to \infty$. We have $\lambda m(-\lambda) = \mathbb{E}[\lambda/(Y+\lambda)]$. Since Y is

a bounded random variable, $\lim_{\lambda\to\infty} \lambda m(-\lambda) = 1$; similarly $\lim_{\lambda\to\infty} \lambda v(-\lambda) = 1$. Next, we find the limit of $\lambda^2\eta(\lambda)$ by noting

$$\lim_{\lambda\to\infty} \lambda^2\big[m(-\lambda) - \lambda m'(-\lambda)\big] = \lim_{\lambda\to\infty} \mathbb{E}\bigg[Y\bigg(\frac{\lambda}{Y+\lambda}\bigg)^2\bigg] = \mathbb{E}[Y].$$

Finally, we evaluate the limit of $\lambda^2\xi(\lambda) = \lambda^2[v'(-\lambda)/v^2(-\lambda) - 1]$. Noting that λv tends to 1, it is enough to find the limit of $\lambda^4(v'(-\lambda) - v^2(-\lambda))$. We compute

$$\begin{aligned}
\lambda^2\big(v'(-\lambda) - v^2(-\lambda)\big) &= \mathbb{E}\bigg[\bigg(\frac{\lambda}{\underline{Y}+\lambda}\bigg)^2\bigg] - \mathbb{E}\bigg[\frac{\lambda}{\underline{Y}+\lambda}\bigg]^2 \\
&= \mathbb{E}\bigg[\bigg(1 - \frac{\underline{Y}}{\underline{Y}+\lambda}\bigg)^2\bigg] - \bigg(1 - \mathbb{E}\bigg[\frac{\underline{Y}}{\underline{Y}+\lambda}\bigg]\bigg)^2 \\
&= \mathbb{E}\bigg[\bigg(\frac{\underline{Y}}{\underline{Y}+\lambda}\bigg)^2\bigg] - \mathbb{E}\bigg[\frac{\underline{Y}}{\underline{Y}+\lambda}\bigg]^2.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\lim_{\lambda\to\infty} \lambda^4\big(v'(-\lambda) - v^2(-\lambda)\big) &= \lim_{\lambda\to\infty} \bigg\{\mathbb{E}\bigg[\underline{Y}^2\bigg(\frac{\lambda}{\underline{Y}+\lambda}\bigg)^2\bigg] - \mathbb{E}\bigg[\underline{Y}\frac{\lambda}{\underline{Y}+\lambda}\bigg]^2\bigg\} \\
&= \mathbb{E}[\underline{Y}^2] - \mathbb{E}[\underline{Y}]^2.
\end{aligned}$$

Using the relationship $\underline{F} = \gamma F + (1-\gamma)\delta_0$, we can write $\mathbb{E}[\underline{Y}] = \gamma\mathbb{E}[Y]$ and $\mathbb{E}[\underline{Y}^2] = \gamma\mathbb{E}[Y^2]$. Putting everything together, we find

$$\Theta_{\mathrm{IR}} = \lim_{\lambda\to\infty} \frac{\alpha^2\lambda\tau(\lambda)}{\sqrt{\alpha^2\lambda^2\eta(\lambda) + \lambda^2\xi(\lambda)}} = \frac{\alpha^2}{\sqrt{\alpha^2\mathbb{E}[Y] + \gamma(\mathbb{E}[Y^2] - \gamma\mathbb{E}[Y]^2)}}.$$

Finally, it is known that $\mathbb{E}[Y] = \mathbb{E}[T]$ and that $\mathbb{E}[Y^2] = \mathbb{E}[T^2] + \gamma\mathbb{E}[T]^2$ [see, e.g., Lemma 2.16 in Yao, Bai and Zheng (2015)]. This leads to the claimed formula.

4. Discussion. In one of our reviews, it was pointed out that the RRC and RWC assumptions allow w to be sparse in the sense of $\|w\|_0/p \to c < 1$, by choosing a distribution with a point mass at 0 with some positive probability. In such a case, it may still make sense to construct a classifier or regressor in high dimension via thresholding. Comparing these with the dense methods could be a topic of future work.

Acknowledgments. We are grateful to the Associate Editor and two referees for their helpful comments, and for providing an account of the early developments from the 1960s on classification under $p, n \to \infty$ asymptotics as described in Section 3.5. We are especially thankful to Prof. Sarunas Raudys for a personal communication on the history of the area. We are also grateful to all our colleagues who have provided comments on earlier versions of this manuscript, including in particular, David Donoho, Jerome Friedman, Iain Johnstone, Percy Liang, Asaf Weinstein and Charles Zheng.

SUPPLEMENTARY MATERIAL

Supplement to "High-dimensional asymptotics of prediction: Ridge regression and classification" (DOI: 10.1214/17-AOS1549SUPP; .pdf). In the supplementary material, we give efficient methods to compute the risk formulas, and prove the remaining lemmas and other results.

REFERENCES

ANDERSON, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*, 3rd ed. Wiley, New York. MR1990662

BAI, Z. and SILVERSTEIN, J. W. (2010). *Spectral Analysis of Large Dimensional Random Matrices*, 2nd ed. Springer, Berlin. MR2567175

BARTLETT, P. L. and MENDELSON, S. (2003). Rademacher and Gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.* **3** 463–482.

BAYATI, M. and MONTANARI, A. (2012). The LASSO risk for Gaussian matrices. *IEEE Trans. Inform. Theory* **58** 1997–2017.

BEAN, D., BICKEL, P. J., EL KAROUI, N. and YU, B. (2013). Optimal M-estimation in high-dimensional regression. *Proc. Natl. Acad. Sci. USA* **110** 14563–14568.

BERNAU, C., RIESTER, M., BOULESTEIX, A.-L., PARMIGIANI, G., HUTTENHOWER, C., WALDRON, L. and TRIPPA, L. (2014). Cross-study validation for the assessment of prediction algorithms. *Bioinformatics* **30** i105–i112.

BICKEL, P. J. and LEVINA, E. (2004). Some theory of Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations. *Bernoulli* **10** 989–1010. MR2108040

CANDÈS, E. and TAO, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.* **35** 2313–2351. MR2382644

CHEN, L. S., PAUL, D., PRENTICE, R. L. and WANG, P. (2011). A regularized Hotelling's T^2 test for pathway analysis in proteomic studies. *J. Amer. Statist. Assoc.* **106**. 1345–1360. MR2896840

COUILLET, R. and DEBBAH, M. (2011). *Random Matrix Methods for Wireless Communications*. Cambridge Univ. Press, Cambridge.

DEEV, A. (1970). Representation of statistics of discriminant analysis and asymptotic expansion when space dimensions are comparable with sample size. *Sov. Math., Dokl.* **11** 1547–1550.

DICKER, L. (2013). Optimal equivariant prediction for high-dimensional linear models with arbitrary predictor covariance. *Electron. J. Stat.* **7** 1806–1834.

DICKER, L. (2016). Ridge regression and asymptotic minimax estimation over spheres of growing dimension. *Bernoulli* **22** 1–37.

DOBRIBAN, E. (2017). Sharp detection in PCA under correlations: All eigenvalues matter. *Ann. Statist.* **45** 1810–1833. MR3670197

DOBRIBAN, E. and WAGER, S. (2018). Supplement to "High-dimensional asymptotics of prediction: Ridge regression and classification." DOI:10.1214/17-AOS1549SUPP.

DONOHO, D. L. and MONTANARI, A. (2015). Variance breakdown of Huber (M)-estimators. $n/p \to m \in (1, \infty)$. Preprint. Available at arXiv:1503.02106.

DONOHO, D. L., JOHNSTONE, I. M., HOCH, J. C. and STERN, A. S. (1992). Maximum entropy and the nearly black object. *J. Roy. Statist. Soc. Ser. B* **54** 41–81. MR1157714

EFRON, B. (1975). The efficiency of logistic regression compared to normal discriminant analysis. *J. Amer. Statist. Assoc.* **70** 892–898.

EL KAROUI, N. (2013). Asymptotic behavior of unregularized and ridge-regularized high-dimensional robust regression estimators: Rigorous results. Preprint. Available at arXiv:1311.2445.

EL KAROUI, N. (2015). On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators. Technical report, Univ. California, Berkeley.

EL KAROUI, N. and KÖSTERS, H. (2011). Geometric sensitivity of random matrix results: Consequences for shrinkage estimators of covariance and related statistical methods. Preprint. Available at arXiv:1105.1404.

FAN, J., FAN, Y. and WU, Y. (2011). High dimensional classification. In *High-Dimensional Data Analysis* (T. Cai and X. Shen, eds.) 3–37. World Sci. Publ., Singapore. MR2848198

FRIEDMAN, J. H. (1989). Regularized discriminant analysis. *J. Amer. Statist. Assoc.* **84** 165–175.

FUJIKOSHI, Y. and SEO, T. (1998). Asymptotic approximations for EPMCs of the linear and the quadratic discriminant functions when the sample sizes and the dimension are large. *Random Oper. Stochastic Equations* **6** 269–280. MR1631003

FUJIKOSHI, Y., ULYANOV, V. V. and SHIMIZU, R. (2011). *Multivariate Statistics*: *High-Dimensional and Large-Sample Approximations*. Wiley, New York.

GRENANDER, U. and SZEGŐ, G. (1984). *Toeplitz Forms and Their Applications*, 2nd ed. Chelsea Publishing Co., New York. MR0890515

HACHEM, W., LOUBATON, P. and NAJIM, J. (2007). Deterministic equivalents for certain functionals of large random matrices. *Ann. Appl. Probab.* **17** 875–930. MR2326235

HACHEM, W., KHORUNZHIY, O., LOUBATON, P., NAJIM, J. and PASTUR, L. (2008). A new approach for mutual information analysis of large dimensional multi-antenna channels. *IEEE Trans. Inform. Theory* **54** 3987–4004.

HASTIE, T., TIBSHIRANI, R. and WAINWRIGHT, M. (2015). *Statistical Learning with Sparsity*: *The Lasso and Generalizations*. CRC Press, Boca Raton, FL.

HSU, D., KAKADE, S. M. and ZHANG, T. (2014). Random design analysis of ridge regression. *Found. Comput. Math.* **14** 569–600.

KLEINBERG, J., LUDWIG, J., MULLAINATHAN, S., OBERMEYER, Z. et al. (2015). Prediction policy problems. *Am. Econ. Rev.* **105** 491–495.

KUBOKAWA, T., HYODO, M. and SRIVASTAVA, M. S. (2013). Asymptotic expansion and estimation of EPMC for linear classification rules in high dimension. *J. Multivariate Anal.* **115** 496–515.

LEDOIT, O. and PÉCHÉ, S. (2011). Eigenvectors of some large sample covariance matrix ensembles. *Probab. Theory Related Fields* **151** 233–264.

LIANG, P. and SREBRO, N. (2010). On the interaction between norm and dimensionality: Multiple regimes in learning. In *ICML*.

MARCHENKO, V. A. and PASTUR, L. A. (1967). Distribution of eigenvalues for some sets of random matrices. *Mat. Sb.* **114** 507–536.

NG, A. and JORDAN, M. (2001). On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. In *NIPS*.

PICKRELL, J. K. and PRITCHARD, J. K. (2012). Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet.* **8** e1002967.

RAUDYS, Š. (1967). On determining training sample size of linear classifier. *Comput. Syst.* **28** 79–87 (in Russian).

RAUDYS, Š. (1972). On the amount of a priori information in designing the classification algorithm. *Technical Cybernetics* **4** 168–174 (in Russian).

RAUDYS, Š. (2001). *Statistical and Neural Classifiers*: *An Integrated Approach to Design*. Springer Science & Business Media, Berlin.

RAUDYS, Š. and SAUDARGIENE, A. (1998). Structures of the covariance matrices in the classifier design. In *Joint IAPR Intl Workshops on SPR and SSPR* 583–592. Springer, Berlin.

RAUDYS, Š. and SKURICHINA, M. (1995). Small sample properties of ridge estimate of the covariance matrix in statistical and neural net classification. *New Trends Probab. Stat.* **3** 237–245.

RAUDYS, Š. and YOUNG, D. M. (2004). Results in statistical discriminant analysis: A review of the former Soviet Union literature. *J. Multivariate Anal.* **89** 1–35.

RIFAI, S., DAUPHIN, Y., VINCENT, P., BENGIO, Y. and MULLER, X. (2011). The manifold tangent classifier. *Adv. Neural Inf. Process. Syst.* **24** 2294–2302.

RUBIO, F., MESTRE, X. and PALOMAR, D. P. (2012). Performance analysis and optimal selection of large minimum variance portfolios under estimation risk. *IEEE J. Sel. Top. Signal Process.* **6** 337–350.

RUSSAKOVSKY, O., DENG, J., SU, H., KRAUSE, J., SATHEESH, S., MA, S., HUANG, Z., KARPATHY, A., KHOSLA, A., BERNSTEIN, M. et al. (2015). ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.* **115** 211–252.

SARANADASA, H. (1993). Asymptotic expansion of the misclassification probabilities of D-and A-criteria for discrimination from two high dimensional populations using the theory of large dimensional random matrices. *J. Multivariate Anal.* **46** 154–174.

SERDOBOLSKII, V. I. (1983). On minimum error probability in discriminant analysis. *Dokl. Akad. Nauk SSSR* **27** 720–725.

SERDOBOLSKII, V. I. (2007). *Multiparametric Statistics*. Elsevier, Amsterdam.

SILVERSTEIN, J. W. (1995). Strong convergence of the empirical distribution of eigenvalues of large dimensional random matrices. *J. Multivariate Anal.* **55** 331–339.

SIMARD, P. Y., LE CUN, Y. A., DENKER, J. S. and VICTORRI, B. (2000). Transformation invariance in pattern recognition: Tangent distance and propagation. *Int. J. Imaging Syst. Technol.* **11** 181–197.

SUTTON, C. and MCCALLUM, A. (2006). An introduction to conditional random fields for relational learning. In *Introduction to Statistical Relational Learning* 93–128.

TOUTANOVA, K., KLEIN, D., MANNING, C. D. and SINGER, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *NAACL*.

TULINO, A. M. and VERDÚ, S. (2004). Random matrix theory and wireless communications. *Commun. Inf. Theory* **1** 1–182.

VAPNIK, V. N. and CHERVONENKIS, A. Y. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory Probab. Appl.* **16** 264–280.

WANG, S. and MANNING, C. D. (2012). Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the* 50*th Annual Meeting of the Association for Computational Linguistics*: *Short Papers*, *Vol.* 2 90–94. Association for Computational Linguistics, Stroudsburg PA .

WRAY, N. R., GODDARD, M. E. and VISSCHER, P. M. (2007). Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Res.* **17** 1520–1528.

YAO, J., BAI, Z. and ZHENG, S. (2015). *Large Sample Covariance Matrices and High-Dimensional Data Analysis*. Cambridge Univ. Press, Cambridge.

ZHANG, M., RUBIO, F., PALOMAR, D. P. and MESTRE, X. (2013). Finite-sample linear filter optimization in wireless communications and financial systems. *IEEE Trans. Signal Process.* **61** 5014–5025.

ZOLLANVARI, A., BRAGA-NETO, U. M. and DOUGHERTY, E. R. (2011). Analytic study of performance of error estimators for linear discriminant analysis. *IEEE Trans. Signal Process.* **59** 4238–4255.

ZOLLANVARI, A. and DOUGHERTY, E. R. (2015). Generalized consistent error estimator of linear discriminant analysis. *IEEE Trans. Signal Process.* **63** 2804–2814. MR3345588

DEPARTMENT OF STATISTICS
THE WHARTON SCHOOL
UNIVERSITY OF PENNSYLVANIA
PHILADELPHIA, PENNSYLVANIA 19104
USA
E-MAIL: dobriban@wharton.upenn.edu

GRADUATE SCHOOL OF BUSINESS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305
USA
E-MAIL: swager@stanford.edu