High-Dimensional Asymptotics of Prediction: Ridge Regression*

Edgar Dobriban

1 Introduction

- a new way to analyze and understand ridge regression in high dimensions
  - focus on out-of-sample prediction
  - in a linear regression model with dense random effects
- main result: formula for the asymptotic predictive risk
- consequences:
  - an exact inverse relationship between prediction error and estimation error (perhaps surprising?)
  - a thorough study of the "regimes of learning problem", similar in spirit to Liang and Srebro (2010) (answer some of their conjectures)
- main tool: asymptotic random matrix theory

2 A first look at ridge regression

Setup:

- Observation model: n independent observations $y_i = x_i \cdot w + \varepsilon_i$ from a p-dimensional linear model. ε_i are independent random variables with mean 0 and variance 1.
- Estimate w by ridge regression: $\hat{w}_\lambda = (X^\top X + \lambda\, n\, I_{p\times p})^{-1} X^\top Y$.
- Goal: Understand expected predictive risk $r_\lambda(X) = \mathbb{E}\left[(y - \hat{y}_\lambda)^2 \,\middle|\, X\right]$, where (x, y) is a new test example and $\hat{y}_\lambda = \hat{w}_\lambda \cdot x$.
- Assume random regression coefficients: w is random with $\mathbb{E}[w] = 0$, $\mathrm{Var}[w] = p^{-1}\alpha^2 I_{p\times p}$;
  - note: $\alpha^2 = \mathbb{E}[\|w\|_2^2]$ is the expected signal strength.

Under these conditions, with Gaussian w, ε, $\lambda_p^* = \gamma_p \alpha^{-2}$ is the Bayes-optimal regularization parameter, where $\gamma_p = p/n$. The ridge regression estimator is the posterior mean; and so it is Bayes-optimal for any quadratic loss function, including ℓ_2 estimation and prediction risk.

By elementary calculations, the finite sample expected predictive risk $r_{\lambda_p^*}(X)$ has a simple expression (where $\widehat{\Sigma} = n^{-1} X^\top X$ is the sample covariance matrix)

$$r_{\lambda_p^*}(X) = 1 + \frac{\gamma_p}{p}\,\mathrm{tr}\left(\Sigma\left(\widehat{\Sigma} + \frac{\gamma_p}{\alpha^2} I_{p\times p}\right)^{-1}\right).$$

It seems difficult to understand this quantity precisely for finite n, p. Instead, will work in an asymptotic framework where $n, p \to \infty$. Note: in classical asymptotic statistics, $n \to \infty$, while p is fixed. However, the current interest is in high-dimensional asymptotics.

*based on part of a paper with Stefan Wager, available at http://arxiv.org/abs/1507.03003

3 High-dimensional asymptotics + random matrices

- High-dimensional asymptotics
  - $n, p \to \infty$ where $p/n \to \gamma$ for some aspect ratio $0 < \gamma < \infty$
  - $n \times p$ data matrix $X = Z\Sigma^{1/2}$ for some $n \times p$ matrix Z with i.i.d. mean 0, variance 1 entries, and $p \times p$ covariance matrix Σ
  - the spectral distribution F_Σ of Σ converges weakly to a limit H, called the population spectral distribution (PSD).
    - The *spectral distribution* of a symmetric matrix A is the cumulative distribution function of its eigenvalues: $F_A(x) = p^{-1} \sum_{i=1}^{p} \mathrm{I}(\lambda_i(A) \leq x)$. (the histogram of eigenvalues)
    - In words: For large p, the histogram of eigenvalues of Σ is "close" to the probability distribution H.

- Examples:
  1. $\Sigma = I_p$ for all p.
  2. autoregressive AR(1) model with $\Sigma_{ij} = \rho^{|i-j|}$, or more generally ARMA time series models.
  3. `BinaryTree` covariance structure. Used in population genetics to model populations whose evolutionary history is described by a balanced binary tree (Pickrell and Pritchard, 2012).

The following is a basic result in the field of random matrix theory.

Theorem (Marchenko and Pastur (1967); Silverstein (1995))**.** *The spectral distribution $F_{\widehat{\Sigma}}$ of the sample covariance matrix $\widehat{\Sigma}$ converges weakly, with probability 1, to a limiting distribution F—the empirical spectral distribution (ESD)—supported on* $[0, \infty)$.

In words: for large n, p, the spectral distribution of the random matrix $\widehat{\Sigma}$ will be "close" to a certain deterministic distribution F, "with high probability".

Important consequence: Functionals depending on averages of the eigenvalues of $\widehat{\Sigma}$ will be approximately deterministic. Quantities like $\frac{\gamma_p}{p} \operatorname{tr}\left(\Sigma\left(\widehat{\Sigma} + \frac{\gamma_p}{\alpha^2} I_{p\times p}\right)^{-1}\right)$ have a deterministic limit that depends on H.

Example: Figure 1. The PSD is a mixture of point masses, while the ESD is a mixture of "bumps" centered around those point masses. Note that the ESD is different from the PSD—it is more spead out. This is true in general over every population spectrum. The difference between the population and sample spectra implies for instance that plug-in estimators of functionals of the parameter Σ or H will be in general be biased.

Closed form expressions exist only for $H = \delta_1$, in which case the ESD has the density:

$$f(x;\gamma) = \frac{\sqrt{(\gamma_+ - x)(x - \gamma_-)}}{2\pi\gamma x} I(x \in [\gamma_-, \gamma_+]), \tag{1}$$

The boundaries of the spectrum are given by $\gamma_\pm = (1 \pm \sqrt{\gamma})^2$. Note that, as γ grows from 0 to 1, the support of the spectrum increases monotonically from $[1, 1]$ to $[0, 4]$. For large $\gamma \in [0, 1]$ the sample covariance matrix is more poorly conditioned.

More generally, the limiting empirical spectral distribution F is determined uniquely by a fixed point equation for its *Stieltjes transform*, which is defined for any distribution G supported on $[0, \infty)$ as

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

With this definition, an equivalent statement to the Marchenko-Pastur theorem is the following: The Stieltjes transform of the spectral measure of $\widehat{\Sigma}$ satisfies

$$m_{\widehat{\Sigma}}(z) = \frac{1}{p} \operatorname{tr}\left(\left(\widehat{\Sigma} - z\, I_{p\times p}\right)^{-1}\right) \text{ converges to } m(z) \tag{2}$$

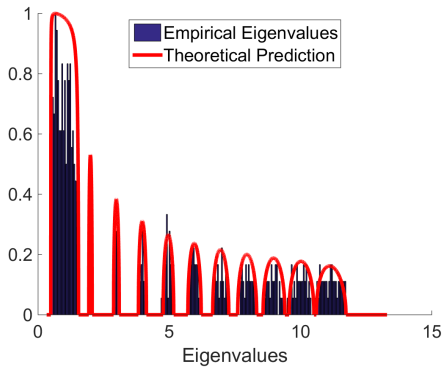

Figure 1: The density of the limit ESD, when the PSD is an equal mixture of two components: (1) a mixture of ten point masses at 2,3, . . . , 11, with weights forming an arithmetic progression with step $r = 0.005$ as follows: $0.0275, 0.0325, \ldots, 0.0725$; and (2) a uniform distribution on $[0.5, 1.5]$, with mixture weight $1/2$. $\gamma = 0.01$.

with probability one, for any $z \in \mathbb{C} \setminus \mathbb{R}^+$; here, we wrote $m(z) := m_F(z)$. In addition to $m(z)$, we also define the companion Stieltjes transform $v(z)$, which is the Stieltjes transform of the limiting spectral distribution of the matrix $\underline{\widehat{\Sigma}} = n^{-1} X X^\top$. This is related to $m(z)$ by

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

Our goal will be to express the quantities of interest (estimation error, prediction error of ridge regression etc.), in terms of the "irreducible quantities" m, v; and then interpret the results for insight.

For this, it will be useful to have expressions for more complicated trace functionals involving both Σ, and $\widehat{\Sigma}$, such as the following formula from (Ledoit and Péché, 2011):

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

Statements like this are nontrivial. While it is clear that, say, $\frac{1}{p}\operatorname{tr}\left(\Sigma^3\left(\widehat{\Sigma} + \lambda I_{p\times p}\right)^{-2}\right)$ converges, there is no general theory to tell us what the limit is. This is true more generally, as by concentration inequalities one can often show that random variables have a deterministic limit, without knowing what the limit is.

4 An asymptotic analysis of ridge regression

Return to ridge regression. Assume the high-dimensional asymptotic model, as above + some additional moment conditions.

Theorem 4.1. *The finite sample predictive risk converges almost surely to an asymptotic limit given by:*

$$r_{\lambda_p^*}(X) \to_{a.s.} R^*(H, \alpha^2, \gamma) := \frac{1}{\lambda^* v(-\lambda^*)}. \tag{5}$$

Here $\lambda^* = \gamma\alpha^{-2}$ is the limit of the optimal regularization parameters, and v is the companion Stieltjes transform of the ESD F. Another way to write this is

$$r_{\lambda_p^*}(X) - \left(\frac{\gamma^2}{\alpha^2}\frac{1}{p}\operatorname{tr}\left[\left(\widehat{\Sigma} + \frac{\gamma}{\alpha^2} I_{p\times p}\right)^{-1}\right] + 1 - \gamma\right)^{-1} \to_{a.s.} 0.$$

In words, the predictive risk of ridge regression has a limit under high-dimensional asymptotics. The limit can be expressed in terms of the signal strength α^2, the aspect ratio γ, and the limiting eigenvalue distribution of the sample covariance matrix $\widehat{\Sigma}$.

Fortunately, the result is simple enough that we can use it to gain a great deal of insight, as discussed in the next two sections.

Important special case: When $\Sigma = I_{p\times p}$, we have an explicit expression for the Stieltjes transform (e.g., Bai and Silverstein, 2010, p. 52), for $\lambda > 0$:

$$m_I(-\lambda;\gamma) = \frac{-(1-\gamma+\lambda)+\sqrt{(1-\gamma+\lambda)^2+4\gamma\lambda}}{2\gamma\lambda}.$$

This also leads to a closed form expression for the risk (with optimal λ^*).

$$R^*(\alpha^2,\gamma) = \frac{1}{2}\left[1+\frac{\gamma-1}{\gamma}\alpha^2+\sqrt{\left(1-\frac{\gamma-1}{\gamma}\alpha^2\right)^2+4\alpha^2}\right]. \tag{6}$$

Example: To verify finite-sample accuracy, we perform a simulation with the `BinaryTree` and `Exponential` models. The results in Figure 2 show that the formulas given in Theorem 4.1 appear to be accurate, even in small sized problems.

A broad claim: this is the "right" way to take high-dimensional limits to understand ridge regression. It achieves a good tradeoff between how broad it is (can capture the effect of Σ, α, γ), versus how simple the results are.

4.1 An Inaccuracy Principle for Ridge Regression

As a first consequence, the results reveal an intriguing inverse relationship between the prediction and estimation errors of ridge regression.

- mean-squared estimation error $R_{E,n}(\lambda) = \mathbb{E}\left[\|\hat{w}_\lambda - w^*\|^2\right]$
- can show
$$R_{E,n}(\lambda_p^*) = \frac{\gamma_p}{p}\,\mathrm{tr}\left(\left(\widehat{\Sigma}+\lambda_p^* I_{p\times p}\right)^{-1}\right) \tag{7}$$
- optimally tuned ridge regression satisfies, under the conditions of Theorem 4.1,
$$R_{E,n}(\lambda^*) \to_{a.s.} R_E := \gamma\, m\,(-\lambda^*) \text{ for } \lambda^* = \gamma\alpha^{-2},$$
where $m(\cdot)$ is the Stieltjes transform of the limiting empirical spectral distribution (see, e.g., Tulino and Verdú, 2004, Chapter 3). So, by (3) we find that

Corollary 4.2. *Under the conditions of Theorem 4.1, the asymptotic predictive and estimation risks of optimally-tuned ridge regression are inversely related, by the equation*

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

- This is a relation that holds for all H (i.e., for all limit spectra of Σ).
- Both sides are non-negative: R_P cannot fall below the intrinsic noise level $\mathrm{Var}\left[Y \,\middle|\, X\right] = 1$, while $R_E \le \limsup R_{E,n}(\lambda^*) \le \limsup R_{E,n}(0) = \alpha^2$.
- Not both prediction and estimation can be easy.
- Intuitively, 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 multi-collinearity. As correlation decreases, prediction gets harder but estimation gets easier. A similar heuristic was noted by Liang and Srebro (2010).

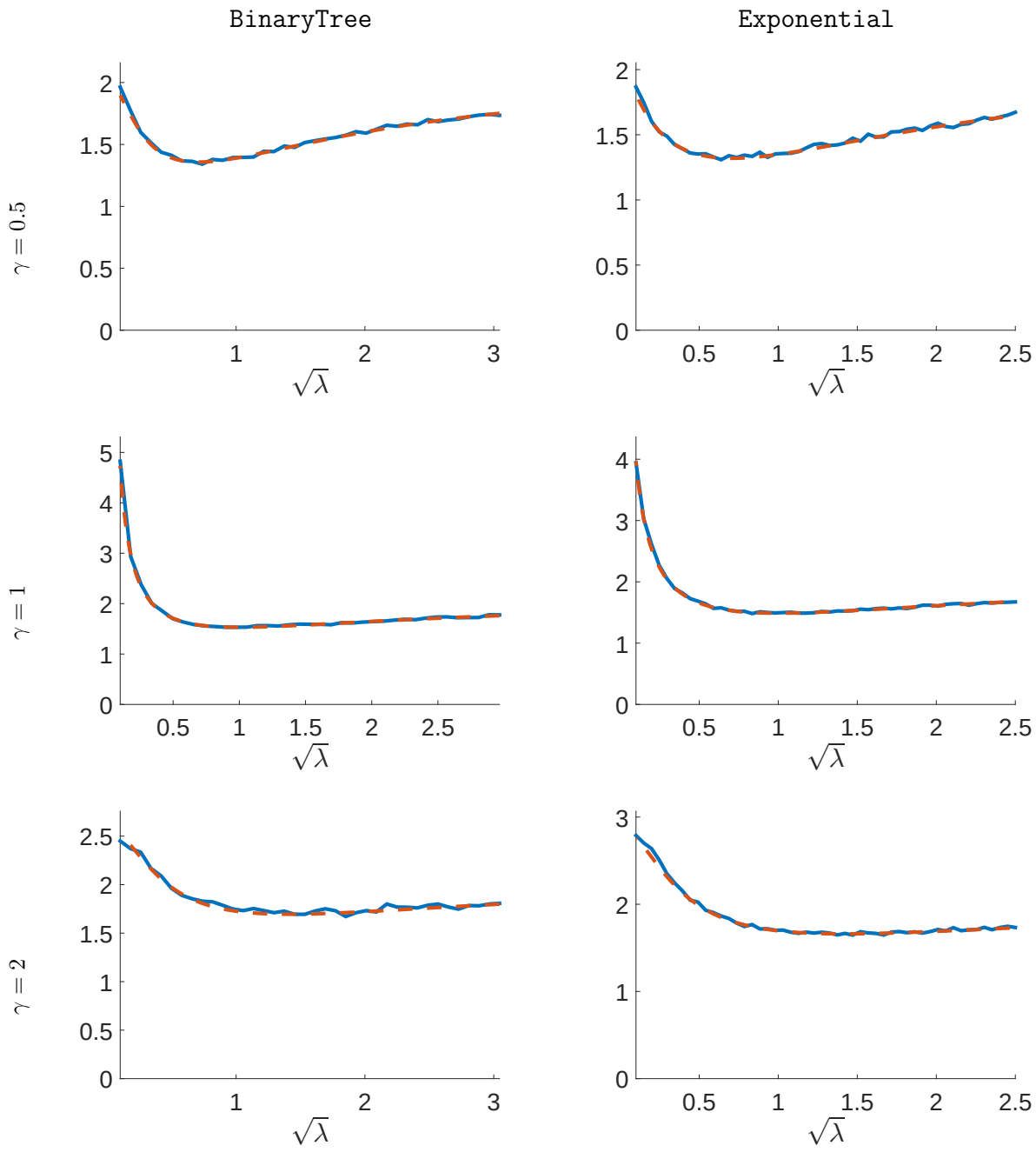

Figure 2: Prediction error of ridge regression in the `BinaryTree` and `Exponential` model. The theoretical formula (red, dashed) is overlaid with the results from simulations (blue, solid). The signals are drawn from $w \sim \mathcal{N}\left(0, p^{-1} I_{p\times p}\right)$. For `BinaryTree`, we train 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.

4.2 Regimes of Learning

- As a second consequence, we will try to understand: How does the prediction error or ridge regression depend on the signal strength α^2, the aspect ratio (dimensionality) γ, and the covariance matrix Σ?
- A more general version called the "regimes of learning" problem in Liang and Srebro (2010). For the specific case of ridge regression, they have some conjectures, which can be roughly stated as:
  - for small α^2, i.e., for weak signals, the error rate should be controlled in a dimension-independent way (e.g., by Rademacher bounds)
  - for large α^2 the error rate should strongly depend on γ, which is a proxy for the dimensionality.
- use Theorem 4.1 to examine the two limiting behaviors of the risk, for weak and strong signals.
- weak-signal limit:
  - 0th order: $\lim_{\alpha^2 \to 0} R^*(H, \alpha^2, \gamma) = 1$, reflecting that for a small signal, we predict a near-zero outcome due to a large regularization.
  - 1st order: $\lim_{\alpha^2 \to 0}(R^*(H, \alpha^2, \gamma) - 1)/\alpha^2 = \mathbb{E}_H T$, where $\mathbb{E}_H T$ is the large-sample limit of the normalized traces $p^{-1} \operatorname{tr} \Sigma$.
    - 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 size of the ratio $\gamma = p/n$. This is in line with the conjectures in Liang and Srebro (2010).

Strong-signal limit: depends the aspect ratio γ, and experiences a phase transition at $\gamma = 1$.

- When $\gamma < 1$, the predictive risk converges to

$$\lim_{\alpha^2 \to \infty} R^*(H, \alpha^2, \gamma) = \frac{1}{1 - \gamma} \tag{8}$$

  regardless of Σ. This quantity is known to be the n, $p \to \infty$, $p/n \to \gamma$ limit of the risk of ordinary least squares (OLS). 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. Moreover, we can verify that

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

  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 $\Sigma = I$, 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:

$$\lim_{\alpha^2 \to \infty} \alpha^{-1} R^*(H, \alpha^2, \gamma) = \frac{1}{\sqrt{\mathbb{E}_H\left[T^{-1}\right]}}, \tag{10}$$

  where $\mathbb{E}_H\left[T^{-1}\right]$ is the large-sample limit of $p^{-1} \operatorname{tr}\left(\Sigma^{-1}\right)$. 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.

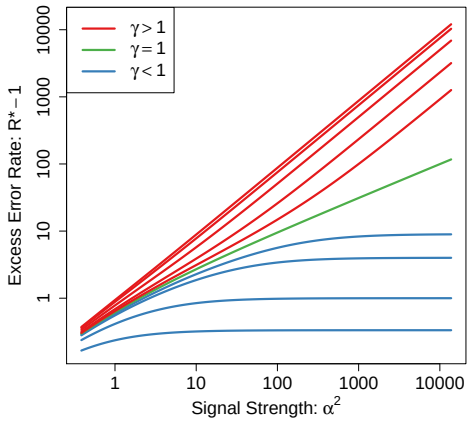

Figure 3: Phase transition for predictive risk of ridge regression with identity covariance $\Sigma = I$. Error rates are plotted for $\gamma = 0.25$, 0.5, 0.8, 0.9, 1, 1.1, 1.3, 2, 4, and 8.

- 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 $\gamma > 1$. See Figure 3.

Quote from our paper: "This leads a complete and exact answer the regimes of learning question posed by Liang and Srebro (2010) in the case of ridge regression. The results (9) and (10) not only show that the scalings found by Liang and Srebro (2010) with $\Sigma = I$ hold for arbitrary Σ, but make explicit how the slopes depend on the limiting population spectral distribution."

5 Further notes

The Stieltjes transform of a probability distribution G contains all information about G. If G has a density g, the inversion formula for Stieltjes transforms is

$$g(x) = \frac{1}{\pi} \lim_{\varepsilon \to 0^+} \text{Imag}\{m_G(x + i\varepsilon)\}. \tag{11}$$

Note that a Stieltjes transform is analytic in the upper half $\mathbb{C}^+$ of the complex plane.

Silverstein and Choi (1995) show that $v(z)$ is the unique solution with positive imaginary part of the Silverstein equation:

$$-\frac{1}{v(z)} = z - \gamma \int \frac{t\, dH(t)}{1 + tv(z)}, \; z \in \mathbb{C}^+. \tag{12}$$

This equation can be solved by a fixed-point algorithm to compute $v(z)$ for all $z \in \mathbb{C}^+$.

References

Z. Bai and J. W. Silverstein. *Spectral Analysis of Large Dimensional Random Matrices*. Springer, 2010.

O. Ledoit and S. Péché. Eigenvectors of some large sample covariance matrix ensembles. *Probab. Theory Related Fields*, 151(1-2):233–264, 2011.

P. Liang and N. Srebro. On the interaction between norm and dimensionality: Multiple regimes in learning. In *ICML*, 2010.

V. A. Marchenko and L. A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Mat. Sb.*, 114(4): 507–536, 1967.

J. K. Pickrell and J. K. Pritchard. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS genetics*, 8(11):e1002967, 2012.

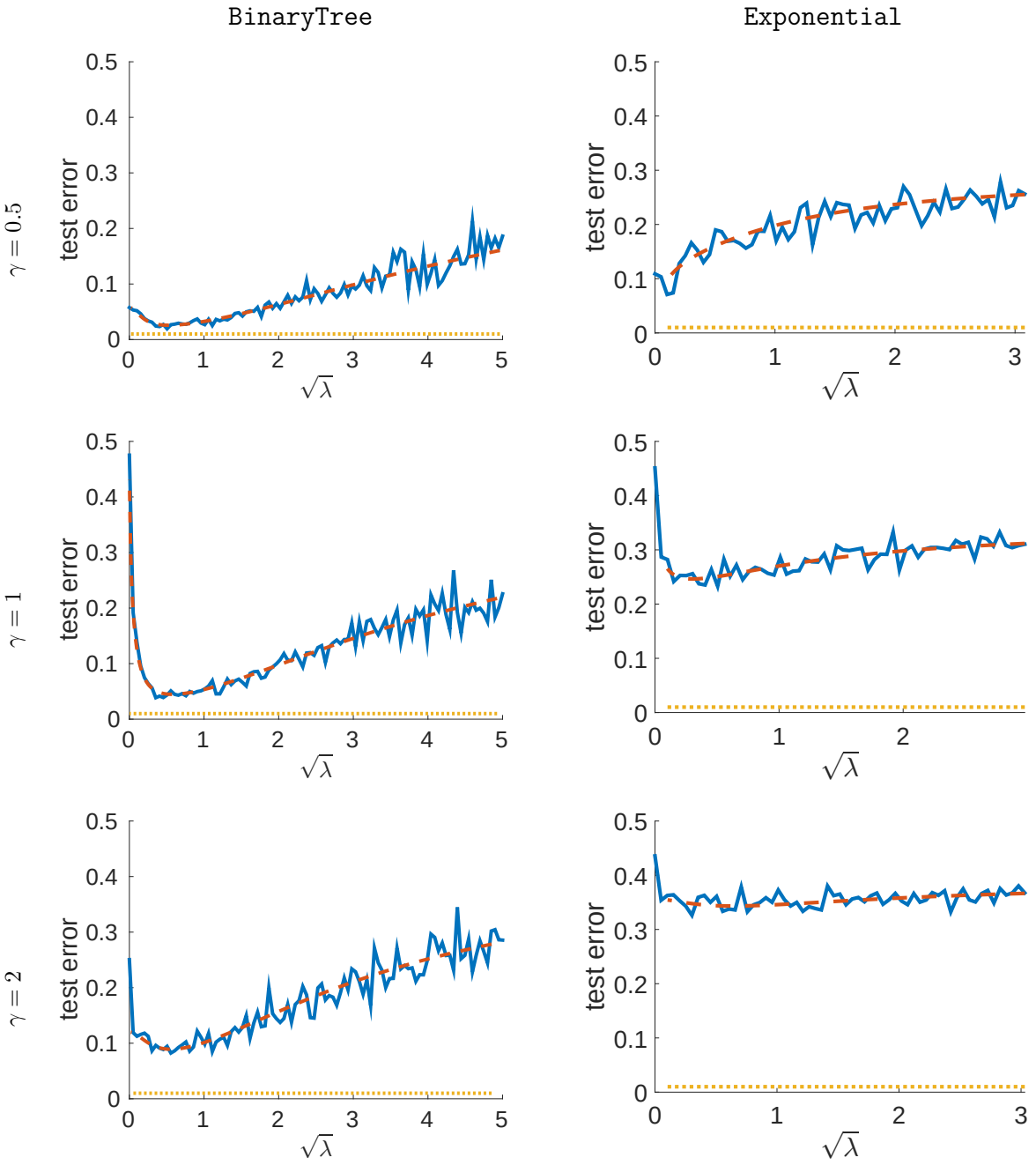

Figure 4: Classification error of RDA in the `BinaryTree` and `Exponential` models. The theoretical formula (red, dashed) is overlaid with the results from simulations (blue, solid); we also display the oracle error (yellow, dotted). The class means are drawn from $\mu_{\pm 1} \sim \mathcal{N}\left(0,\, \alpha^2\, p^{-1} I_{p\times p}\right)$, 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. 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 **??**, quantifies the strength of the regularization.

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

J. W. Silverstein and S.-I. Choi. Analysis of the limiting spectral distribution of large dimensional random matrices. *J. Multivariate Anal.*, 54(2):295–309, 1995.

A. M. Tulino and S. Verdú. Random matrix theory and wireless communications. *Communications and Information theory*, 1(1):1–182, 2004.