High-Dimensional Asymptotics of Prediction: Ridge Regression and Classification

Noah Feinberg

August 26, 2024

Problem Setting

We observe n training samples $(x_i, y_i) \in \mathbb{R}^p \times \mathcal{Y}$ drawn independently from $\mathcal{D}$. We want to find an $h(x) = g(\omega^\top x)$ that makes the following error small

$$\mathbb{E}_{\mathcal{D}}[\ell(h(x), y)]$$

Regression and classification are specific cases of this

1. Regression: $\mathcal{Y} = \mathbb{R}$, $g(x) = x$, and $\ell(x, y) = (x - y)^2$
2. Classification: $\mathcal{Y} = \{0, 1\}$, $g(x) = \text{sgn}(x)$, and ℓ is 0-1 loss

Problem Setting cont.

Hypothesis

Each predictor (ω_i) has a small, independent random effect on the outcome

As an example, in the regression setting we will assume that $\mathbb{E}[\omega] = 0$ and $\text{Var}[\omega] = p^{-1}\alpha^2 I_p$ where $\alpha^2 = \mathbb{E}\|\omega\|^2$
Then for regression X, Y are related through $X\omega = Y + \epsilon$ for an independent mean zero, unit variance ϵ. As is standard we define the ridge solution as

$$\hat{\omega}_\lambda = (X^\top X + n\lambda I_p)^{-1} X^\top Y$$

and the corresponding estimates are $\hat{y}_\lambda = \hat{\omega}_\lambda^\top x$
Furthermore, we are interested specifically in the asymptotic setting for both n, p, that is $n, p \to \infty$ and

$$\frac{p}{n} \to \gamma > 0$$

Definitions

Definition

The *spectral distribution* of a matrix $A \in \mathbb{R}^{p \times p}$ is the CDF of the eigenvalues

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

Definition

If X is a measurable space, then we say a sequence of probability measures P_i *converges weakly* to P if for all $f \in C_B(X)$ we have

$$\int_X f \, dP_i \to \int_X f \, dP$$

Marchenko Pastur

Assumptions

1. We will assume that we can factor our data matrix $X \in \mathbb{R}^{b \times p}$ as $X = Z\Sigma^{1/2}$ where Z has iid mean zero, unit variance entries, and Σ is a constant PSD covariance matrix.
2. The spectral distributions F_{Σ} converge weakly to a probability measure H called the population spectral distribution (PSD).

Theorem

Under these assumptions, then $F_{\hat{\Sigma}}$ converges weakly with probability 1 to a limiting distribution F called the empirical spectral distribution (ESD), where $\hat{\Sigma}$ is the sample covariance of X.

Stieltjes Transform

Definition

For any measure G, defined on $[0, \infty)$, it defines a function called the *Stieltjes transform* defined by

$$m_G(z) = \int_0^\infty \frac{G(t)dt}{z - t}$$

We define $m(z) := m_F(z)$ for F defined earlier, and also define the companion transform $v(z)$ as the Stieltjes transform of the limit of $\underline{\hat{\Sigma}} = \frac{1}{n} X X^\top$. These two are related by

$$\gamma(m(z) + z^{-1}) = v(z) + z^{-1}$$

For a distribution G with moments m_n, the Stieltjes transform has expansion

$$m_G(z) = \sum_{n=0}^{\infty} \frac{m_n}{z^{n+1}}$$

More Random Matrix Theory

The proof also uses the following result on random matrices and requires finite 12th moments

Theorem

$$\frac{1}{p} \operatorname{tr}\left(\Sigma(\hat{\Sigma} + \lambda I_p)^{-1}\right) \to_{a.s} \frac{1}{\gamma}\left(\frac{1}{\lambda v(-\lambda)} - 1\right)$$

Predictive Risk of Ridge Regression

Theorem

Under the previous assumptions, and the additional assumptions that $\|\Sigma\| \leq C$ *and* $\mathbb{E}[Z_{ij}^{12}] < C$ *for all the* Σ, Z*, then for all choices of* $\lambda > 0$

$$\begin{aligned} r_\lambda(X) &:= \mathbb{E}[(y - \hat{y}_\lambda)^2 \,|\, X] \\ &\to_{a.s} R(H, \alpha^2, \gamma) \\ &:= \frac{1}{\lambda v(-\lambda)} \left(1 + \left(\frac{\gamma\alpha^2}{\lambda} - 1\right)\left(1 - \frac{\lambda v'(-\lambda)}{v(-\lambda)}\right)\right) \end{aligned}$$

Cont.

Theorem cont.

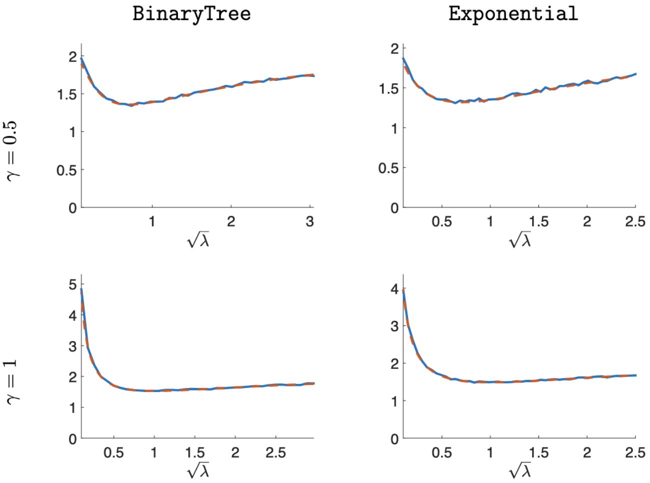
Furthermore, if we define $\gamma_p = \frac{p}{n}$ and choose the optimal ridge parameter $\lambda_p^* = \gamma_p \alpha^{-2}$ then we have

$$\begin{aligned} r_{\lambda_p^*}(X) &= 1 + \frac{\gamma_p}{p} \operatorname{tr}\left(\Sigma\left(\hat{\Sigma} + \frac{\gamma_p}{\alpha^2} I_p\right)^{-1}\right) \\ &\to_{\mathrm{a.s}} R^*(H, \alpha^2\gamma) \\ &:= \frac{1}{\lambda^* v(-\lambda^*)} \end{aligned}$$

for $\lambda^* = \gamma\alpha^{-2}$

Graph

For BinaryTree $p = 16$ and $n = p\gamma^{-1}$ while for Exponential $n = 20$ and $p = n\gamma$.



Partial Proof of Theorem

The proof for the optimal cases is shorter so we will give that

$$\begin{aligned} r_{\lambda_p^*}(X) &= 1 + \mathbb{E}[(x^\top(\omega - \hat{\omega}_{\lambda_p^*}))^2 \mid X] \\ &= 1 + \mathbb{E}[(\omega - \hat{\omega}_{\lambda_p^*})^\top (xx^\top)(\omega - \hat{\omega}_{\lambda_p^*}) \mid X] \\ &= 1 + \mathbb{E}[(\omega - \hat{\omega}_{\lambda_p^*})^\top \Sigma(\omega - \hat{\omega}_{\lambda_p^*}) \mid X] \\ &= 1 + \operatorname{tr}\left(\Sigma \mathbb{E}[(\omega - \hat{\omega}_{\lambda_p^*})(\omega - \hat{\omega}_{\lambda_p^*})^\top \mid X]\right) \end{aligned}$$

Now note that

$$\begin{aligned} \omega - \hat{\omega}_{\lambda_p^*} &= \omega - (X^\top X + n\lambda_p^* I_p)^{-1} X^\top (X\omega + X^\top \epsilon) \\ &= \omega - (X^\top X + n\lambda_p^* I_p)^{-1} (X^\top X\omega + n\lambda_p^* I_p \omega - n\lambda_p^* I_p \omega + X^\top \epsilon) \\ &= (X^\top X + n\lambda_p^* I_p)^{-1} (X^\top \epsilon - n\lambda_p^* \omega) \end{aligned}$$

Partial Proof cont.

Now we can substitute this back into where we previously were, let $A = (X^\top X + n\lambda_p^* I_p)$

$$\begin{aligned} r_{\lambda_p^*}(X) &= 1 + \operatorname{tr}\left(\Sigma \mathbb{E}[(\omega - \hat{\omega}_{\lambda_p^*})(\omega - \hat{\omega}_{\lambda_p^*})^\top \,|\, X]\right) \\ &= 1 + \operatorname{tr}\left(\Sigma A^{-1} \mathbb{E}[(X^\top \epsilon - n\lambda_p^* \omega)(X^\top \epsilon - n\lambda_p^* \omega)^\top \,|\, X] A^{-1}\right) \\ &= 1 + \operatorname{tr}\left(\Sigma A^{-1}(X^\top X + n^2 (\lambda_p^*)^2 p^{-1} \alpha^2 I_p) A^{-1}\right) \\ &= 1 + \operatorname{tr}\left(\Sigma A^{-1}\right) \\ &= 1 + \frac{\gamma_p}{p} \operatorname{tr}\left(\Sigma (\widehat{\Sigma} + \frac{\gamma_p}{\alpha^2} I_p)^{-1}\right) \end{aligned}$$

Now by the theorem of Lenoit, this converges a.s to

$$\frac{1}{\lambda^* v(-\lambda^*)}$$

Identity Covariance

In the special case of identity covariance, then the Stieltjes transform admits a simple formula

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

And from this we find that the optimal risk is equal to

$$\frac{1}{2}\left(1 + \frac{\gamma-1}{\gamma}\alpha^2 + \sqrt{\left(1 - \frac{\gamma-1}{\gamma}\alpha^2\right)^2 + 4\alpha^2}\right)$$

Learning Regimes, Small α^2

For small signal strength, the optimal regret doesn't depend on the aspect ratio γ, we can see this by using the asymptotics of the Stieltjes transform

$$\begin{aligned}\lim_{\alpha^2\to 0}\frac{1}{\lambda^* v(-\lambda^*)} &= \lim_{\alpha^2\to 0}\left(\lambda^*\sum_{n=0}^{\infty}\frac{m_n}{(\lambda^*)^{n+1}}\right)^{-1}\\ &= \lim_{\alpha^2\to 0}\left(\lambda^*\frac{m_0}{\lambda^*}\right)^{-1}\\ &= 1\end{aligned}$$

Further more, the first order behavior of this limit is

$$\lim_{\alpha^2\to 0}\frac{(\lambda^* v(-\lambda^*))^{-1}-1}{\alpha^2} = \lim_{p\to\infty} p^{-1}\operatorname{tr}(\Sigma_p)$$

Learning Regimes, Large α^2

As $\alpha^2 \to \infty$, we have the following regimes based on the aspect ratio γ

- For $\gamma < 1$

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

  which is the same as the risk for OLS

- For $\gamma > 1$ the risk may be unbounded

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

  For identity covariance this has a closed form of $\frac{\gamma-1}{\gamma}$

- Finally, when $\gamma = 1$

$$\lim_{\alpha^2 \to \infty} \alpha^{-1} R^*(H, \alpha^2\gamma) = \lim_{p \to \infty} \frac{1}{\sqrt{p^{-1}\,\mathrm{tr}(\Sigma^{-1})}}$$

Learning Regimes, Large α^2

This may be summarized by saying that for $\gamma < 1$ the risk behaves like $\Theta(1)$, for $\lambda = 1$ it behaves like $\Theta(\alpha)$, and for $\gamma > 1$ it behaves like $\Theta(\alpha^2)$

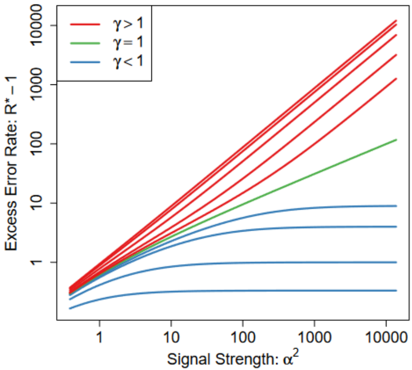


Figure 3: Phase transition for predictive risk of ridge regression with identity covariance $\Sigma = I_{p\times p}$. Error rates are plotted for $\gamma = 0.25$, 0.5, 0.8, 0.9, 1, 1.1, 1.3, 2, 4, and 8.

Inaccuracy Principle

The estimation error is defined as

$$R_{E,n}(\lambda) = \mathbb{E}\|\omega - \hat{\omega}_\lambda\|^2$$

Under the conditions of the main theorem its known that

$$\lim_{n\to\infty} R_{E,n}(\lambda^*) := R_E = \lambda^* m(-\lambda^*)$$

where m is the limiting Stieltjes transform from before. Now we can find a relation between R_E and R_P

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

In particular, for $\gamma = 1$ this simplifies to

$$R_E R_p \geq \alpha^2$$