Optimal Ridge Regularization for Out-of-Distribution Prediction

Pratik Patil[1] **Jin-Hong Du**[2] Ryan J. Tibshirani[1]

[1]Department of Statistics, University of California, Berkeley
[2]Department of Statistics & Machine Learning Department, Carnegie Mellon University

July 2024

Berkeley UNIVERSITY OF CALIFORNIA

Carnegie Mellon University

Ridge regression in high dimensions

Recent interests in high-dimensional ridge regression concern the ridge estimator:

$$\widehat{\boldsymbol{\beta}}^{\lambda} = (\boldsymbol{X}^{\top}\boldsymbol{X}/n + \lambda \boldsymbol{I}_p)^{\dagger}\boldsymbol{X}^{\top}\boldsymbol{y}/n,$$

and its prediction risk:

$$R(\widehat{\boldsymbol{\beta}}^{\lambda}) = \mathbb{E}_{\boldsymbol{x}_0, y_0}[(y_0 - \boldsymbol{x}_0^{\top}\widehat{\boldsymbol{\beta}}^{\lambda})^2 \mid \boldsymbol{X}, \boldsymbol{y}].$$

The goal is to study the behavior of its asymptotic prediction risk:

$$R(\widehat{\boldsymbol{\beta}}^{\lambda}) \to \mathscr{R}(\lambda, \phi)$$

as the feature size p and the sample size n diverge proportionally to an *aspect ratio* $p/n \to \phi \in (0, \infty)$.

Optimal ridge regression under in-distribution

For high-dimensional ridge regression, two questions for the optimal in-distribution asymptotic risk $\min_{\lambda \geq \lambda_{\min}} \mathscr{R}(\lambda, \phi)$:

(Q1) What is the behavior of the *optimal ridge penalty*, as a function of parameters such as signal-to-noise ratio, data aspect ratio, feature correlations, and signal structure?

(Q2) What is the behavior of the *optimally tuned ridge risk*, as a function of these same problem parameters?

Known results provide partial answers:

(A1) $\lambda^* = \phi/\mathrm{SNR} > 0$ in the isotropic cases when $\lambda_{\min} = 0$, while $\lambda^* < 0$ in some anisotropic cases (both signal and features) and overparameterized regimes.

(A2) $\mathscr{R}(\lambda^*, \phi)$ is monotonically increasing in ϕ.

Ridge regression under distribution shifts (motivation)

We consider two types of distribution shifts:

(i) *Covariate shift*: where $P_{\boldsymbol{x}_0} \neq P_{\boldsymbol{x}}$ but $P_{y_0|\boldsymbol{x}_0} = P_{y|\boldsymbol{x}}$.

(ii) *Regression shift*: where $P_{y_0|\boldsymbol{x}_0} \neq P_{y|\boldsymbol{x}}$ but $P_{\boldsymbol{x}_0} = P_{\boldsymbol{x}}$.

and answer two out-of-distribution problems:

(Q1′) *How does distribution shift alter optimal regularization λ^*?*

(Q2′) *How does distribution shift alter optimal risk behavior $\mathscr{R}(\lambda^*, \phi)$?*

Summary of results

Optimal regularization landscape in ridge regression.

Σ	β	Σ_0	β_0	$\phi \lessgtr 1$	$\lambda_{\min}$	**Arb. Mod.**	**Arb. SNR**	**Arb. Spec.**	**Additional Specific Data Geometry Conditions**	λ^*	**Reference**
In-distribution											
$\otimes$	$\bigcirc$	Σ	β	all	zero	✗	✓	✗		$+$	[DW, Thm. 2.1]
$\bigcirc$	$\otimes$	Σ	β	all	zero	✗	✓	✗		$+$	[HMRT, Cor. 5]
				under	neg	✗	✓	✗		$+$	[WX, Prop. 6]
				over	neg	✗	✗	✗	Strict misalignment of (Σ, β)	$+$	[WX, Thm. 4]
				over	neg	✗	✗	✗	Strict alignment of (Σ, β)	$-$	[WX, Thm. 4, Prop. 7]
$\otimes$	$\otimes$	Σ	β	over	zero	✗	✗	✗	and/or special feature model	0	[RMR, Cor. 2]
				under	neg*	✓	✓	✓		$+$	**Theorem 2** (1)
				over	neg*	✓	✓	✓	General alignment of $(\Sigma, \beta, \sigma^2)$	$-$	**Theorem 2** (2)
Out-of-distribution											
$\otimes$	$\bigcirc$	Σ_0	β	all	neg*	✓	✓	✓		$+$	**Proposition 3**
$\otimes$	$\otimes$	Σ_0	β	under	neg*	✓	✓	✓		$+$	**Theorem 4** (1)
$\otimes$	$\otimes$	I	β	over	neg*	✓	✓	✓		$+$	**Theorem 4** (2)
$\bigcirc$	$\otimes$	Σ_0	β	over	neg*	✓	✓	✓	General alignment of $(\Sigma_0, \beta, \sigma^2)$	$-$	**Theorem 4** (3)
				under	neg*	✓	✓	✓	General alignment of (Σ, β, β_0)	$-$	**Theorem 5** (1), (39)
$\otimes$	$\otimes$	Σ	β_0	under	neg*	✓	✓	✓	General misalignment of (Σ, β, β_0)	$+$	**Theorem 5** (1), (39)
				over	neg*	✓	✓	✓	General alignment of $(\Sigma, \beta, \beta_0, \sigma^2)$	$-$	**Theorem 5** (2)

Data assumptions and lower bound on negative regularization

Data assumptions:

- Covariate: Each feature vector $\boldsymbol{x}_i$ for $i \in [n]$ can be decomposed as $\boldsymbol{x}_i = \boldsymbol{\Sigma}^{1/2} \boldsymbol{z}_i$, where $\boldsymbol{z}_i \in \mathbb{R}^p$ contains i.i.d. entries z_{ij} for $j \in [p]$ with mean 0, variance 1, and bounded $4+\mu$ moments for some $\mu > 0$. (RMT structure and bounded moment)
- Response: Each response variable y_i for $i \in [n]$ has mean 0, and bounded $4+\mu$ moments. (model-free)

Lower bound on λ: Let $\mu_{\min} \in \mathbb{R}$ be the unique solution, satisfying $\mu_{\min} > -r_{\min}$, to the equation:

$$1 = \phi \, \bar{\mathrm{tr}}[\boldsymbol{\Sigma}^2 (\boldsymbol{\Sigma} + \mu_{\min} \boldsymbol{I})^{-2}],$$

and let $\lambda_{\min}(\phi)$ be given by:

$$\lambda_{\min}(\phi) = \mu_{\min} - \phi \, \bar{\mathrm{tr}}[\boldsymbol{\Sigma} (\boldsymbol{\Sigma} + \mu_{\min} \boldsymbol{I})^{-1}].$$

Out-of-distribution risk characterization

The OOD risk asymptotics read that

$$\mathscr{R}(\lambda, \phi) := \underbrace{\mathscr{B}(\lambda, \phi)}_{\text{bias}} + \underbrace{\mathscr{V}(\lambda, \phi)}_{\text{variance}} + \underbrace{\mathscr{E}(\lambda, \phi)}_{\text{extra bias}} + \underbrace{\kappa^2}_{\text{irreducible error}}, \tag{1}$$

where

$$\begin{aligned}
\mathscr{B} &= \mu^2 \cdot \boldsymbol{\beta}^\top (\boldsymbol{\Sigma} + \mu \boldsymbol{I})^{-1} (\widetilde{v}\boldsymbol{\Sigma} + \boldsymbol{\Sigma}_0)(\boldsymbol{\Sigma} + \mu \boldsymbol{I})^{-1} \boldsymbol{\beta}, \\
\mathscr{V} &= \sigma^2 \widetilde{v}, \\
\mathscr{E} &= 2\mu \cdot \boldsymbol{\beta}^\top (\boldsymbol{\Sigma} + \mu \boldsymbol{I})^{-1} \boldsymbol{\Sigma}_0 (\boldsymbol{\beta}_0 - \boldsymbol{\beta}), \\
\kappa^2 &= (\boldsymbol{\beta}_0 - \boldsymbol{\beta})^\top \boldsymbol{\Sigma}_0 (\boldsymbol{\beta}_0 - \boldsymbol{\beta}) + \sigma_0^2.
\end{aligned}$$

The optimal regularization is defined as

$$\lambda^* \in \operatorname*{argmin}_{\lambda \geq \lambda_{\min}(\phi)} \mathscr{R}(\lambda, \phi). \tag{2}$$

Optimal regularization sign characterization (IND)

Theorem (Optimal regularization sign for IND risk)

1. *(Underparameterized) When $\phi < 1$, we have $\lambda^* \geq 0$.*
2. *(Overparameterized) When $\phi > 1$, if for all $v < 1/\mu(0, \phi)$, the following general alignment holds:*

$$\frac{\bar{\mathrm{tr}}[\boldsymbol{B}\boldsymbol{\Sigma}(v\boldsymbol{\Sigma}+\boldsymbol{I})^{-2}] + \sigma^2}{\bar{\mathrm{tr}}[\boldsymbol{B}\boldsymbol{\Sigma}(v\boldsymbol{\Sigma}+\boldsymbol{I})^{-3}] + \sigma^2} > \frac{\bar{\mathrm{tr}}[\boldsymbol{\Sigma}(v\boldsymbol{\Sigma}+\boldsymbol{I})^{-2}]}{\bar{\mathrm{tr}}[\boldsymbol{\Sigma}(v\boldsymbol{\Sigma}+\boldsymbol{I})^{-3}]}, \tag{3}$$

where $\boldsymbol{B} = \boldsymbol{\beta}\boldsymbol{\beta}^\top$, then we have $\lambda^* < 0$.

- Alignment condition (3) captures how well the signal $\boldsymbol{B}$ is aligned with the feature covariance $\boldsymbol{\Sigma}$.
- λ^* could be negative in the overparameterized regime when $p > n$.

Illustration (optimal IND regularization)

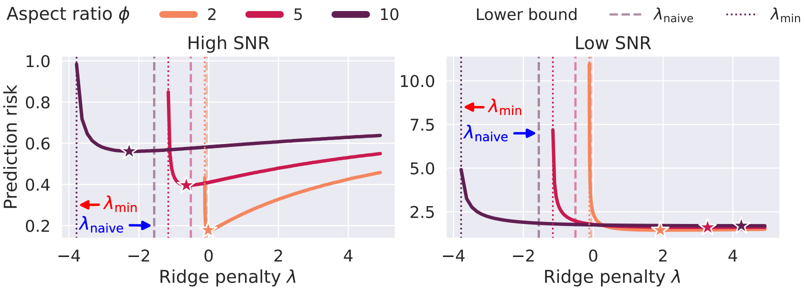

Figure: Illustration of negative or positive optimal regularization under general alignment.

- λ^* can be smaller than the previous bound.
- The more the alignment (seen as a function of SNR), the lower λ^*; the more the misalignment, the higher λ^* (seen as a function of SNR).

Optimal regularization sign characterization (OOD, covariate shift)

1. *(Underparameterized)* When $\phi < 1$, we have $\lambda^* \geq 0$.
2. *(Overparameterized)* When $\phi > 1$, if $\boldsymbol{\Sigma}_0 = \boldsymbol{I}$ (corresponding to the estimation risk), then we have $\lambda^* \geq 0$.
3. *(Overparameterized)* When $\phi > 1$, if $\boldsymbol{\Sigma} = \boldsymbol{I}$ and

$$\bar{\text{tr}}[\boldsymbol{\Sigma}_0 \boldsymbol{B}] > \bar{\text{tr}}[\boldsymbol{\Sigma}_0] \left(\bar{\text{tr}}[\boldsymbol{B}] + \frac{(1 + \mu(0, \phi))^3}{\mu(0, \phi)^3} \sigma^2 \right), \tag{4}$$

   where $\boldsymbol{B} = \boldsymbol{\beta}\boldsymbol{\beta}^\top$, then we have $\lambda^* < 0$.

- The isotropic test covariance case ($\boldsymbol{\Sigma}_0 = \boldsymbol{I}$) is similar to underparameterized cases.
- Alignment condition (4) captures how the well the signal $\boldsymbol{B}$ aligned with covariance matrix of test features $\boldsymbol{\Sigma}_0$.
- λ^* can be negative even in the isotropic train covariance case ($\boldsymbol{\Sigma} = \boldsymbol{I}$).

Optimal regularization sign characterization (OOD, label shift)

1. *(Underparameterized)* When $\phi < 1$, if $\sigma^2 = o(1)$ and for all $\mu \geq 0$, the following general alignment holds:

$$\bar{\mathrm{tr}}[\boldsymbol{B}_0 \boldsymbol{\Sigma}^2 (\boldsymbol{\Sigma} + \mu \boldsymbol{I})^{-2}] > \bar{\mathrm{tr}}[\boldsymbol{B} \boldsymbol{\Sigma}^2 (\boldsymbol{\Sigma} + \mu \boldsymbol{I})^{-2}], \tag{5}$$

   where $\boldsymbol{B} = \boldsymbol{\beta}\boldsymbol{\beta}^\top$ and $\boldsymbol{B}_0 = \boldsymbol{\beta}_0 \boldsymbol{\beta}^\top$, then we have $\lambda^* < 0$.
2. *(Overparameterized)* When $\phi > 1$, if the general alignment conditions (3) and (5) hold, then we have $\lambda^* < 0$.

- λ^* can be negative even if the design is underparameterized!

Illustration (optimal OOD regularization)

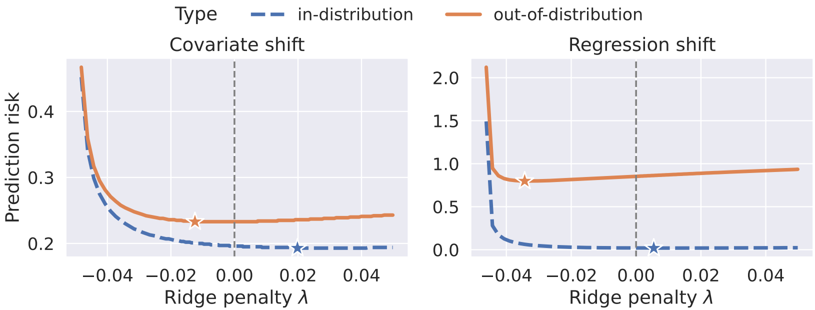

Figure: Covariate and regression shift can lead to negative optimal regularization in both underparameterized and overparameterized regimes.

The design is isotropic on the left.
The design is underparameterized on the right.

Optimal risk monotonicity

The map $\phi \mapsto \min_{\lambda \geq \lambda_{\min}(\phi)} \mathscr{R}(\lambda, \phi)$ is monotonically increasing in ϕ.

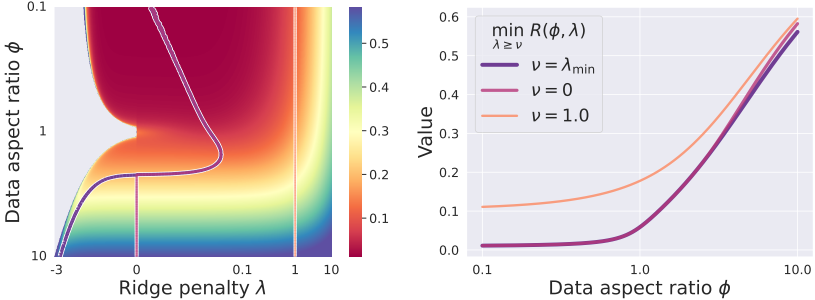

Figure: Ridge regression optimized over $\lambda \geq \nu$ for different thresholds ν has monotonic risk profile.

The previous result holds for positive λ and IND risks.
This result holds for optimizing over negative λ and for all OOD risks.