Optimal Ridge Regularization for Out-of-Distribution Prediction

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

[1]Department of Statistics, UCB [2]Department of Statistics, CMU

Ridge regression in high dimensions

Ridge estimator. 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:

$$R(\widehat{\boldsymbol{\beta}}^{\lambda}) \to \mathscr{R}(\lambda, \phi), \qquad p/n \to \phi \in (0, \infty)$$

where p is feature size, n is sample size, and ϕ is the *aspect ratio*.

Distribution shifts. We consider two types of distribution shifts:

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

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

Questions of interest. We answer two out-of-distribution problems:

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

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

Data assumptions. Feature distribution: 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^+ moments for some $\mu > 0$. Response distribution: Each response variable y_i for $i \in [n]$ has mean 0, and bounded 4^+ moments.

Lower bound on ridge regularization. Let $\mu_{\min} \in \mathbb{R}$ be the unique solution, satisfying $\mu_{\min} > -r_{\min}$, to the equation: $1 = \phi\, \bar{\operatorname{tr}}[\boldsymbol{\Sigma}^2 (\boldsymbol{\Sigma} + \mu_{\min} \boldsymbol{I})^{-2}]$, and let $\lambda_{\min}(\phi)$ be given by: $\lambda_{\min}(\phi) = \mu_{\min} - \phi\, \bar{\operatorname{tr}}[\boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \mu_{\min} \boldsymbol{I})^{-1}]$.

Summary of results

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)

Out-of-distribution risk characterization

Proposition 1 (Deterministic equivalents for OOD risk). The asymptotic OOD risk decomposes into:

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

where

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

The optimal regularization is defined as $\lambda^* \in \operatorname{argmin}_{\lambda \geq \lambda_{\min}(\phi)} \mathscr{R}(\lambda, \phi)$.

Optimal regularization sign characterization (IND)

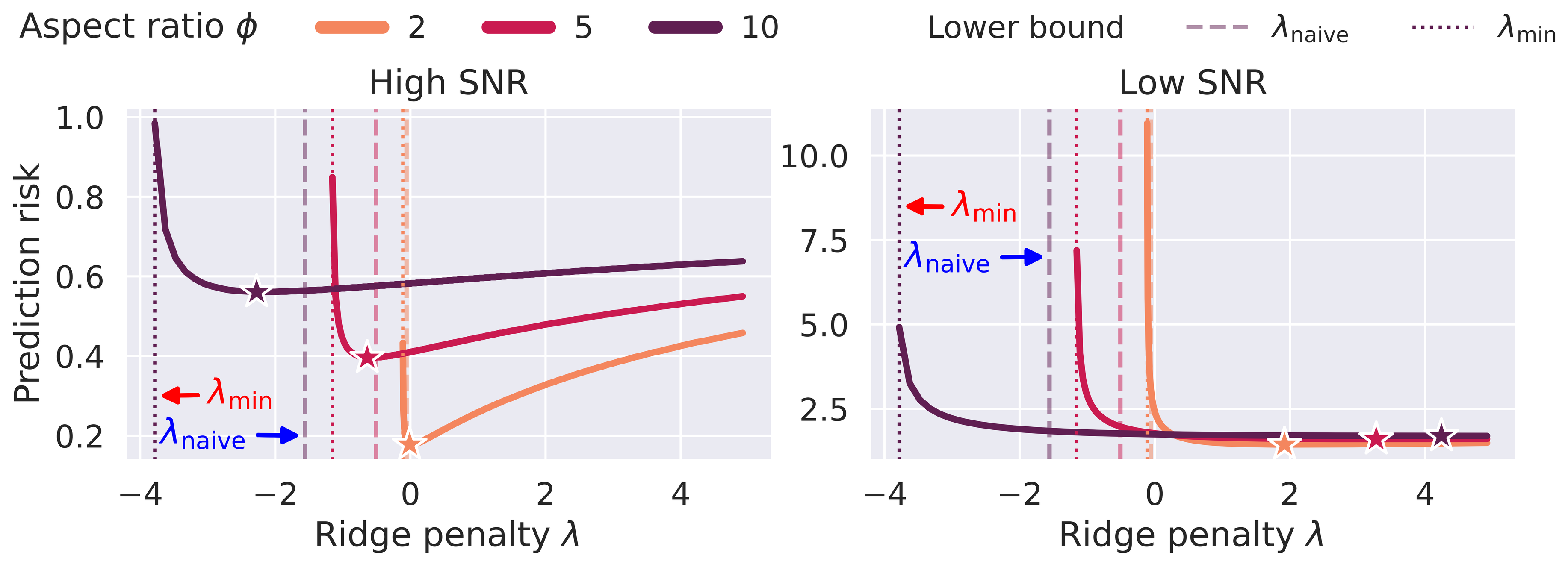


Illustration of negative or positive optimal regularization under general alignment.

Theorem 2 (Optimal regularization sign, no shift) Assume $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}$ and $\boldsymbol{\beta}_0 = \boldsymbol{\beta}$.

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:

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

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

References:

[HMRT] Trevor Hastie et al. "Surprises in high-dimensional ridgeless least squares interpolation". In: *The Annals of Statistics* 50.2 (2022), pp. 949–986

[DW] Edgar Dobriban and Stefan Wager. "High-dimensional asymptotics of prediction: Ridge regression and classification". In: *The Annals of Statistics* 46.1 (2018), pp. 247–279

[RMR] Dominic Richards, Jaouad Mourtada, and Lorenzo Rosasco. "Asymptotics of ridge (less) regression under general source condition". In: *International Conference on Artificial Intelligence and Statistics*. 2021

[WX] Denny Wu and Ji Xu. "On the Optimal Weighted ℓ_2 Regularization in Overparameterized Linear Regression". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 10112–10123

Optimal regularization sign characterization (OOD)

Theorem 4 (Optimal regularization, covariate shift). Assume $\boldsymbol{\Sigma}_0 \neq \boldsymbol{\Sigma}$ and $\boldsymbol{\beta}_0 = \boldsymbol{\beta}$.

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

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

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

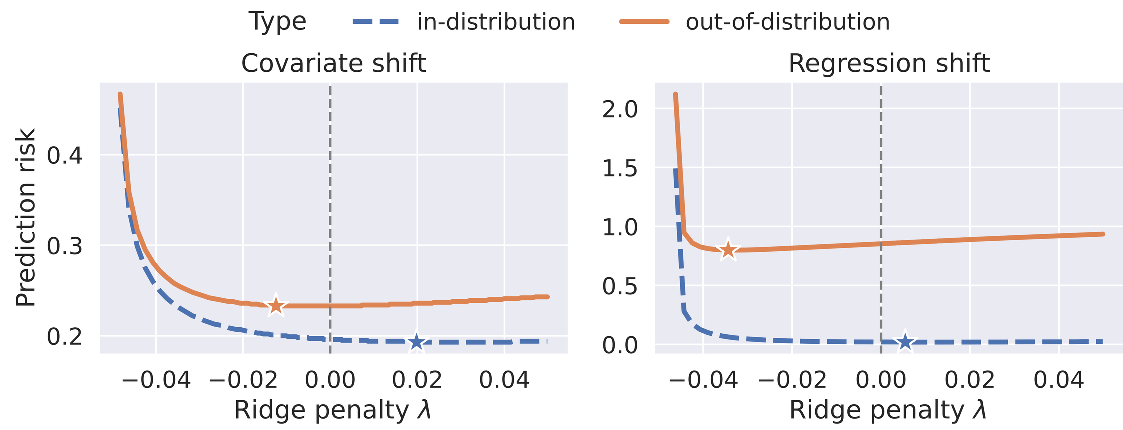
Theorem 5 (Optimal regularization, regression shift). Assume $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}$ and $\boldsymbol{\beta}_0 \neq \boldsymbol{\beta}$.

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

$$\bar{\operatorname{tr}}[\boldsymbol{B}_0 \boldsymbol{\Sigma}^2 (\boldsymbol{\Sigma} + \mu \boldsymbol{I})^{-2}] > \bar{\operatorname{tr}}[\boldsymbol{B} \boldsymbol{\Sigma}^2 (\boldsymbol{\Sigma} + \mu \boldsymbol{I})^{-2}], \tag{4}$$

where $\boldsymbol{B} = \boldsymbol{\beta}\boldsymbol{\beta}^\top$ and $\boldsymbol{B}_0 = \boldsymbol{\beta}_0 \boldsymbol{\beta}^\top$, we have $\lambda^* < 0$.

2. *(Overparameterized)* When $\phi > 1$, if conditions (2) and (4) hold, we have $\lambda^* < 0$.



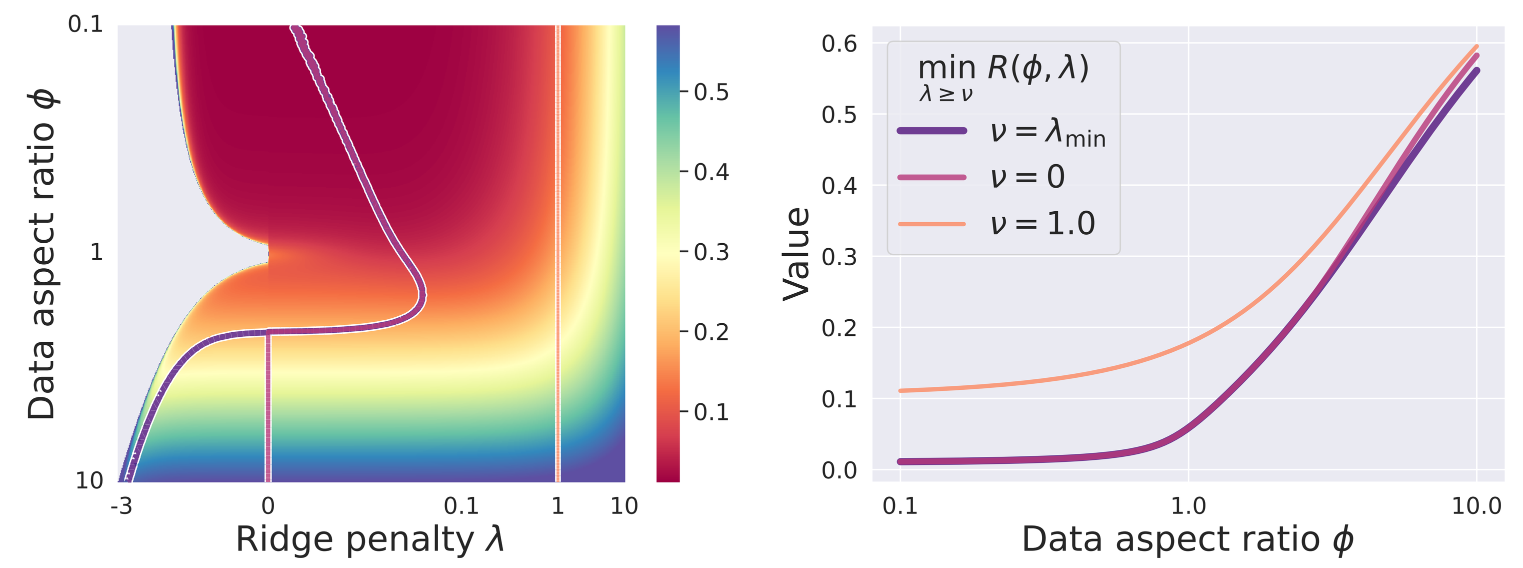
Covariate and regression shift can lead to negative optimal regularization in both overparameterized and underparameterized regimes.

Optimal risk monotonicity (both IND and OOD)

Theorem 7 (Optimal regularization, regression shift). For $\lambda \geq \lambda_{\min}(\phi)$, for all $\varepsilon > 0$ small enough, the risk of optimal ridge predictor satisfies:

$$\min_{\lambda \geq \lambda_{\min}(\phi) + \varepsilon} R(\widehat{\boldsymbol{\beta}}^{\lambda}) \simeq \min_{\lambda \geq \lambda_{\min}(\phi)} \mathscr{R}(\lambda, \phi), \tag{5}$$

and right side of (5) is monotonically increasing in ϕ if SNR and σ_0^2 are fixed. In addition, when $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ it is monotonically increasing in SNR if ϕ, σ^2, and σ_0^2 are fixed.



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