Optimal Ridge Regularization for Out-of-Distribution Prediction

Pratik Patil [1] **Jin-Hong Du** [2 3] **Ryan J. Tibshirani** [1]

Abstract

We study the behavior of optimal ridge regularization and optimal ridge risk for out-of-distribution prediction, where the test distribution deviates arbitrarily from the train distribution. We establish general conditions that determine the sign of the optimal regularization level under covariate and regression shifts. These conditions capture the alignment between the covariance and signal structures in the train and test data and reveal stark differences compared to the in-distribution setting. For example, a negative regularization level can be optimal under covariate shift or regression shift, even when the training features are isotropic or the design is underparameterized. Furthermore, we prove that the optimally tuned risk is monotonic in the data aspect ratio, even in the out-of-distribution setting and when optimizing over negative regularization levels. In general, our results do not make any modeling assumptions for the train or the test distributions, except for moment bounds, and allow for arbitrary shifts and the widest possible range of (negative) regularization levels.

1. Introduction

Regularization plays a crucial role in statistical modeling and is commonly incorporated into optimization-based models through a regularization term. Its effectiveness relies on properly scaling the regularization term, which is controlled by a penalty parameter that the data scientist needs to tune. Recent work in machine learning (precise references given shortly) has shed light on some rather surprising behavior exhibited by the optimal regularization level in overparameterized prediction models, which can be zero or even negative in certain problems with moderate signal-to-noise ratio and high dimensionality. This stands in contrast to the "typical" behavior from classical low-dimensional learning theory.

With this motivation, our paper focuses on two key questions for high-dimensional ridge regression:

(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?

To set the notation, let (x_i, y_i) for $i \in [n]$ be i.i.d. observations in $\mathbb{R}^p \times \mathbb{R}$ representing the feature vector and response. Denote the data matrix as $X = [x_1, \dots, x_n]^\top \in \mathbb{R}^{n \times p}$ and the associated response vector as $y = [y_1, \dots, y_n]^\top \in \mathbb{R}^n$. Given a ridge penalty $\lambda > 0$, recall the ridge regression fits:

$$\widehat{\beta}^\lambda = \underset{b \in \mathbb{R}^p}{\operatorname{argmin}} \; \|y - Xb\|_2^2/n + \lambda \|b\|_2^2. \tag{1}$$

Ridge regression (Hoerl & Kennard, 1970a;b) has received considerable recent attention, particularly in settings involving overparameterization, such as double descent (see, e.g., Belkin et al., 2020; Hastie et al., 2022; Muthukumar et al., 2020, and references therein) and benign overfitting (Bartlett et al., 2020; Koehler et al., 2021; Mallinar et al., 2022). This interest in ridge regression, especially its "ridgeless" limit, where $\lambda \to 0^+$, owes to its peculiar double/multiple descent risk behavior in overparameterized regimes, which (on the surface) challenges the conventional understanding of the role of regularization (Hastie, 2020). By defining the ridge estimator as:

$$\widehat{\beta}^\lambda = (X^\top X/n + \lambda I_p)^\dagger X^\top y/n, \tag{2}$$

where $A^\dagger$ denotes the Moore-Penrose pseudoinverse of A, we simultaneously accommodate the case of $\lambda > 0$, in which case (2) reduces to the unique ridge predictor obtained using (1), and the case of $\lambda = 0$, in which case (2) becomes the minimum ℓ_2-norm interpolator among many solutions to problem (1) when $p \geq \operatorname{rank}(X) = n$. Note that the above definition (2) is well defined even when $\lambda < 0$. For more background on the formulation of ridge regression with $\lambda < 0$, see Appendix B.

[1]Department of Statistics, University of California, Berkeley, CA 94720, USA. [2]Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA. [3]Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA 15213, USA.. Correspondence to: Pratik Patil <pratikpatil@berkeley.edu>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

Partial answers to questions (Q1) and (Q2) are known for the *in-distribution* squared prediction risk, defined as:

$$R(\widehat{\beta}^{\lambda}) = \mathbb{E}_{x_0,y_0}[(y_0 - x_0^\top \widehat{\beta}^{\lambda})^2 \mid X, y], \tag{3}$$

where (x_0, y_0) is a test point sampled independently from the *same* distribution $P_{x,y}$ as the training data. Note that the prediction risk (3) is conditional on the training data and is a function of (X, y) and the properties of $P_{x,y}$. Arguably, the cleanest answers to (Q1) and (Q2) are obtained under proportional asymptotics where the sample size n and the feature size p diverge proportionally to an *aspect ratio* $\phi \in (0, \infty)$. Under certain additional assumptions, the risk (3) then almost surely converges to a limit $\mathscr{R}(\lambda, \phi)$ that depends only on coarse properties of $P_{x,y}$. Analyzing the behavior of the optimal ridge penalty λ^* (which minimizes $\mathscr{R}(\lambda, \phi)$ over λ) and the optimal risk $\mathscr{R}(\lambda^*, \phi)$ then consequently allows us to answer questions (Q1) and (Q2), respectively.

For (Q1), consider a well-specified linear model $y = X\beta + \varepsilon$, where the noise $\varepsilon \sim (0_n, \sigma^2 I)$ is independent of X, and the signal is random with $\beta \sim (0_p, (\alpha^2/p)I)$. Remarkably, despite the lack of a closed-form expression for $\mathscr{R}(\lambda^*, \phi)$, Dobriban & Wager (2018) show that $\lambda^* = \phi/\textsc{snr} > 0$, where $\textsc{snr} = \alpha^2/\sigma^2$ is the signal-to-noise ratio. However, in real-data experiments, it has been observed that a negative ridge penalty can be optimal (Kobak et al., 2020). Motivated by this, Wu & Xu (2020); Richards et al. (2021) analyze the sign behavior of λ^* beyond random isotropic signals and establish sufficient conditions for when $\lambda^* < 0$ or $\lambda^* = 0$.

For (Q2), again remarkably, it follows from the results of Dobriban & Wager (2018) that for isotropic features and random isotropic signals, the risk $\mathscr{R}(\lambda^*, \phi)$ increases monotonically with the data aspect ratio ϕ. Recent work by Patil & Du (2023, Theorem 6) extends this result to anisotropic features and deterministic signals (with arbitrary response distributions of bounded moments), demonstrating that optimal ridge regression exhibits a monotonic risk profile and avoids double and multiple descents. In a certain sense, these results imply the sample-wise monotonicity of the optimal expected conditional risk (over the training data set).

We work to answer (Q1) and (Q2) more comprehensively across essentially all possible in-distribution (IND) ridge prediction problems. Furthermore, we will generalize this by also considering the *out-of-distribution* (OOD) setting, where the test point (x_0, y_0) in (3) has a different distribution P_{x_0,y_0} than the train distribution $P_{x,y} = P_x \cdot P_{y|x}$. Distribution shift occurs in many practical machine learning applications and has gained increasing attention in the learning community. We focus on two common types of distribution shifts:

(i) *Covariate shift*: where $P_{x_0} \neq P_x$ but $P_{y_0|x_0} = P_{y|x}$.
(ii) *Regression shift*: where $P_{y_0|x_0} \neq P_{y|x}$ but $P_{x_0} = P_x$.

Thus, we also study following generalizations of (Q1)–(Q2):

(Q1′) *How does distribution shift alter optimal regularization?* Specifically, under what types of shift is the optimal penalty λ^* in the OOD setting notably different (typically smaller) compared to the IND setting?

(Q2′) *How does distribution shift alter optimal risk behavior?* Specifically, is $\mathscr{R}(\lambda^*, \phi)$ still monotonic in ϕ when λ^* changes due to the distribution shift? Conversely, is optimal regularization *necessary* to ensure monotonic risk?

1.1. Summary of Results and Paper Outline

Extended risk characterization. In Section 2, we extend the scope of risk characterization for ridge regression for the out-of-distribution setting (Proposition 2.4) that: (i) does not assume any model for the train or the test distribution, apart from certain moment bounds on the train and test response distributions, (ii) does not assume that the spectrums of feature covariance or signal projections converge to fixed distributions, and (iii) allows for the widest possible range of (negative) regularization level $\lambda_{\min}$ (see Definition 2.3).

Properties of optimal regularization. In Section 3, we characterize the conditions that determine the sign of λ^* under covariate shift (Theorem 3.3) and regression shift (Theorem 3.4). These conditions capture the general alignment between the signal and the covariance spectrum, which isolates the cases where the sign of λ^* under the OOD setting changes from the IND setting. Our work subsumes and extends previously known results on optimal ridge regularization to the best of our knowledge (see Table 1 for precise comparisons).

Properties of optimal risk. In Section 4, we show that the OOD risk of the optimally tuned ridge $\mathscr{R}(\lambda^*, \phi)$ is monotonic in the data aspect ratio ϕ and $\textsc{snr}$ (Theorem 4.2). Furthermore, we establish a partial converse (Theorem 4.3) that shows risk non-monotonicity under suboptimal regularization. To prove our results, we exploit the equivalences between subsampling and ridge regression from (Patil & Du, 2023) to the OOD setting and also allow for tuning over negative regularization (Theorem 4.4).

1.2. Related Work and Comparisons

Ridge risk characterization. The asymptotic risk of ridge regression has been extensively studied in the literature under proportional asymptotics when $p/n \to \phi \in (0, \infty)$, as $n, p \to \infty$ using tools from random matrix theory and statistical physics. For well-specified linear models, expressions of risk asymptotics for the IND setting are obtained by Dobriban & Wager (2018); Hastie et al. (2022), among others. Historically, heuristic derivations of these expressions have also been derived for Gaussian process regression by Sollich (2001). Additionally, several works have explored

Table 1: **Optimal regularization landscape in ridge regression.** Here, '○' indicates either an isotropic feature or signal covariance, and '⊗' indicates anisotropic features or signal covariance. For the data aspect ratio ϕ, 'all' indicates $\phi \in (0, \infty)$, 'under' indicates $\phi \in (0, 1)$ for the underparameterized regime, and 'over' indicates $\phi \in (1, \infty)$ for the overparameterized regime. For the minimum penalty $\lambda_{\min}$, 'neg' and 'more neg' respectively indicate the naive (loose) and exact lower bound on the negative values (Definition 2.3). For the optimal penalty λ^*, green and red contrast the cases when the sign changes. 'Arb. Mod.', 'Arb. SNR.', and 'Arb. Spec.' indicate allowing for arbitrary response model, signal-to-noise ratio, and feature covariance spectrum, respectively. Please see Table 6 for the reference key.

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											
⊗	○	Σ	β	all	zero	✗	✓	✗		+	[DW, Thm. 2.1]
○	⊗	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]
⊗	⊗	Σ	β	over	zero	✗	✗	✗	and/or special feature model	0	[RMR, Cor. 2]
				under	more neg	✓	✓	✓		+	**Theorem 3.1** (1)
				over	more neg	✓	✓	✓	General alignment of $(\Sigma, \beta, \sigma^2)$	−	**Theorem 3.1** (2)
Out-of-distribution											
⊗	○	Σ_0	β	all	more neg	✓	✓	✓		+	**Proposition 3.2**
⊗	⊗	Σ_0	β	under	more neg	✓	✓	✓		+	**Theorem 3.3** (1)
⊗	⊗	I	β	over	more neg	✓	✓	✓		+	**Theorem 3.3** (2)
○	⊗	Σ_0	β	over	more neg	✓	✓	✓	General alignment of $(\Sigma_0, \beta, \sigma^2)$	−	**Theorem 3.3** (3)
				under	more neg	✓	✓	✓	General alignment of (Σ, β, β_0)	−	**Theorem 3.4** (1), (39)
⊗	⊗	Σ	β_0	under	more neg	✓	✓	✓	General misalignment of (Σ, β, β_0)	+	**Theorem 3.4** (1), (39)
				over	more neg	✓	✓	✓	General alignment of $(\Sigma, \beta, \beta_0, \sigma^2)$	−	**Theorem 3.4** (2)

risk asymptotics and its implications in different variants of ridge regression (Wei et al., 2022; Mel & Ganguli, 2021; Loureiro et al., 2021; Jacot et al., 2020; Simon et al., 2021; Zhou et al., 2023; Bach, 2024; Pesce et al., 2023). The risk asymptotics for the OOD setting are obtained by Canatar et al. (2021), D'Amour et al. (2022, Section E.5), Patil et al. (2022, Section S.6.5), Tripuraneni et al. (2021). However, these works assume either random Gaussian features or a well-specified linear model or restrict to only the positive range of regularization. Our work extends this literature by allowing for general response models and the widest possible range of negative regularization.

Behavior of optimal regularization. Under random signals with isotropic covariance, Dobriban & Wager (2018) show that the asymptotic risk over the positive range of regularization $\lambda > 0$ is minimized at $\lambda^\star = \phi/\text{SNR}$. Here $\text{SNR} = \alpha^2/\sigma^2$, where σ^2 is the noise energy and α^2 is the signal energy. Remarkably, this result is invariant of the feature covariance. Similar results under Gaussian assumptions are derived by Dicker (2016), and Han & Xu (2023) extend the result to most signals in the unit ball with high probability. However, Kobak et al. (2020) demonstrate that optimal regularization can be negative for certain signal and covariance structures in real datasets. Motivated by these curious experiments, Wu & Xu (2020) provide sufficient conditions for optimal regularization of the Bayes risk under anisotropic feature covariance and random signal, assuming a limiting spectrum distribution of the covariance matrix Σ and alignment conditions between the eigenvalues of Σ and the projections of the signal β onto the eigenspace of Σ. Furthermore, Richards et al. (2021) consider strict alignment conditions for a special feature model but do not explicitly consider negative regularization. We refer to these conditions as *strict (mis)alignment* conditions in this paper.

Our paper extends the scope of the aforementioned results for the OOD setting with general response models and allows for the widest possible range of negative regularization. The main differences to Wu & Xu (2020) are the assumptions (linear models and limiting spectrum distribution) and the hypothesis of the theorem (aligned/misaligned). Their analysis only considers high SNR regimes and analyzes the behavior of optimal regularization for bias when $\phi > 1$, although they also provide an upper bound for the noise level under special cases. We provide generic sufficient conditions. We call these *general (mis)alignment* conditions in the paper. Table 1 provides a detailed comparison summary.

Behavior of optimal risk. Under a random isotropic signal, Dobriban & Wager (2018) obtain the expression for limiting optimal risk, which can be shown to be monotonic in ϕ. Patil & Du (2023) extends these theoretical guarantees to features with an arbitrary covariance matrix and a general moment-bounded response. However, their analysis is limited to the IND setting and positive regularization. Recent works have also explored other aspects of the monotonicity of optimal risk; see, for example, Nakkiran et al. (2021); Simon et al.

(2023); Yang & Suzuki (2023).

We extend the monotonicity result in Patil & Du (2023) to the OOD setting and allow for negative regularization. In addition, we also show the monotonicity of the optimal risk in SNR. The proof technique for risk monotonicity leverages the equivalences between subsampling and ridge regularization established by Du et al. (2023); Patil & Du (2023). However, the equivalences in these works only consider cases where the ridge penalty is non-negative. We extend their analyses to accommodate negative regularization, which requires extending the properties of the parameters that appear in certain fixed-point equations under negative regularization (see Appendix F).

2. Out-of-Distribution Risk Asymptotics

Before describing the properties of λ^* in Section 3 and the behavior of optimal risk $\mathscr{R}(\widehat{\beta}^{\lambda^*})$ in Section 4, we provide the risk asymptotics in this section. For the reader's convenience, we summarize all our notation in Appendix A. We state assumptions on the train and test distributions below.

2.1. Data Assumptions

We first define a general feature and response distribution, which we will use in our subsequent assumption shortly.

Definition 2.1 (General feature and response distribution)**.** For $x \sim P_x$, it can be decomposed as $x = \Sigma^{1/2} z$, where $z \in \mathbb{R}^p$ contains i.i.d. entries with mean 0, variance 1, and $(4+\mu)$-th moment uniformly bounded for some $\mu > 0$. Here $\Sigma \in \mathbb{R}^{p \times p}$ is deterministic and symmetric with eigenvalues uniformly bounded away from 0 and $+\infty$. For $y \sim P_y$, it has mean 0 and $(4+\nu)$-th moment uniformly bounded for some $\nu > 0$. The L_2 linear projection parameter of y onto x is denoted by $\beta = \mathbb{E}[xx^\top]^{-1}\mathbb{E}[xy]$, and the variance of the conditional distribution $P_{y|x}$ is denoted by σ^2. The joint distribution $P_{x,y}$ is parameterized by $(\Sigma, \beta, \sigma^2)$.

Definition 2.1 imposes weak moment assumptions on covariates and responses, which are commonly used in random matrix theory and overparameterized risk analysis (Hastie et al., 2022; Bartlett et al., 2021). These assumptions encode a wide class of distributions over $\mathbb{R}^{p+1}$. By decomposing $\mathrm{d}P_y = \mathrm{d}P_x \cdot \mathrm{d}P_{y|x}$, we can express the response as $y = x^\top \beta + \varepsilon$, where ε is uncorrelated with x and $\mathbb{E}[\varepsilon^2] = \sigma^2$. Note that the L_2 projection parameter β minimizes the linear regression error (Györfi et al., 2006; Buja et al., 2019a;b)[1]. Also note that this formulation does not impose any specific structure on the conditional distribution $P_{y|x}$ and does not imply that (x, y) follows a linear model, as ε is also a function of x. It is possible to further relax the assumption on the feature vector x to only require an appropriate convex concentration (that proves versions of the Marchenko-Pastur law) (Louart, 2022; Cheng & Montanari, 2022) or even certain infinitesimal asymptotic freeness between the population covariance matrix Σ and the sample covariance matrix $X^\top X / n$ (LeJeune et al., 2024; Patil & LeJeune, 2024). We do not consider such relaxations here.

Under Definition 2.1, the joint distribution $P_{x,y} = P_{x,y}(\Sigma, \beta, \sigma^2)$ is parameterized by $(\Sigma, \beta, \sigma^2)$. We next state assumptions on the train and test distributions in terms of these distributions, allowing for different sets of parameters.

Assumption 2.2 (Train and test distributions)**.** Assume that $P_{x,y}$ and P_{x_0,y_0} are distributed according to Definition 2.1, parameterized by $(\Sigma, \beta, \sigma^2)$ and $(\Sigma_0, \beta_0, \sigma_0^2)$, respectively.

In this paper, we consider the following types of shifts:

(i) *Covariate shift*: where $\Sigma \neq \Sigma_0$ but $(\beta, \sigma) = (\beta_0, \sigma_0)$.
(ii) *Regression shift*: where $\Sigma = \Sigma_0$ but $(\beta, \sigma) \neq (\beta_0, \sigma_0)$.
(iii) *Joint shift*: where $\Sigma \neq \Sigma_0$ and $(\beta, \sigma) \neq (\beta_0, \sigma_0)$.

Observe that this framework also encompasses various risk notions (even for the IND setting), including the *estimation risk*, which arises when $(\Sigma_0, \beta_0, \sigma_0^2) = (I, \beta, 0)$.

2.2. Out-of-Distribution Risk Asymptotics

In this section, we obtain the asymptotic risk of ridge regression in the OOD setting. Most of the papers on ridge regression consider the range of regularization $\lambda \geq 0$. Motivated by empirical findings in Kobak et al. (2020) that negative regularization can be optimal in real datasets, some recent works consider negative regularization; see, e.g., Wu & Xu (2020); Patil et al. (2021; 2022), using a naive lower bound of $-r_{\min}(1 - \sqrt{\phi})^2$ for λ. A tighter lower bound can be obtained from Theorem 3.1 of LeJeune et al. (2024), which provides an explicit characterization of the smallest nonzero eigenvalue of Wishart-type matrices. This bound is derived by explicitly identifying the analytic continuation to the real line of a unique solution to a certain fixed-point equation over the (upper) complex half-plane (Silverstein & Choi, 1995; Dobriban, 2015). The new bound can significantly outperform the previous naive lower bound (see Figure 1 of (LeJeune et al., 2024)).

Definition 2.3 (Lower bound on negative regularization)**.** Let $\mu_{\min} \in \mathbb{R}$ be the unique solution, satisfying $\mu_{\min} > -r_{\min}$, to the equation:

$$1 = \phi \, \overline{\mathrm{tr}}[\Sigma^2 (\Sigma + \mu_{\min} I)^{-2}], \tag{4}$$

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

$$\lambda_{\min}(\phi) = \mu_{\min} - \phi \, \overline{\mathrm{tr}}[\Sigma (\Sigma + \mu_{\min} I)^{-1}]. \tag{5}$$

[1]Technically, our results can accommodate the conditional variance σ^2 depending on x with suitable regularity conditions, but for simplicity, we do not consider this variation in the current paper.

This enables feasible risk estimation over $\lambda \in (\lambda_{\min}(\phi), \infty)$. Here $\overline{\operatorname{tr}}[A]$ denotes the average trace $\operatorname{tr}[A]/p$ of a matrix $A \in \mathbb{R}^{p\times p}$. To reiterate, the difference between the bound (5) and the naive bound used in Wu & Xu (2020); Patil et al. (2021) can be significant, as seen in Figure 1.

To characterize the asymptotic out-of-distribution (OOD) risk in Proposition 2.4, we first define the non-negative constants $\mu = \mu(\lambda, \phi)$ and $\widetilde{v} = \widetilde{v}(\lambda, \phi; \Sigma_0)$ as solutions of the following fixed-point equations:

$$\mu = \lambda + \phi\,\overline{\operatorname{tr}}[\mu\Sigma(\Sigma + \mu I)^{-1}], \;\; \widetilde{v} = \frac{\phi\,\overline{\operatorname{tr}}[\Sigma_0\Sigma(\Sigma + \mu I)^{-2}]}{1 - \phi\,\overline{\operatorname{tr}}[\Sigma^2(\Sigma + \mu I)^{-2}]}.$$

One can interpret μ as the level of implicit regularization "self-induced" by the data (Bartlett et al., 2021; Misiakiewicz & Montanari, 2023). Alternatively, it is also common to parameterize the equations using its inverse $v(\lambda, \phi) = \mu^{-1}(\lambda, \phi)$, which corresponds to the Stieltjes transform of the spectrum of the sample Gram matrix in the limit. With this notation in place, we can now extend the result in Eq. (11) of Patil & Du (2023) to the OOD setting as formalized below.

Proposition 2.4 (Deterministic equivalents for OOD risk)**.** *Under Assumption 2.2, as $n, p \to \infty$ such that $p/n \to \phi \in (0, \infty)$ and $\lambda \in (\lambda_{\min}(\phi), \infty)$, the prediction risk $R(\widehat{\beta}^\lambda)$ defined in (3) admits a deterministic equivalent $R(\widehat{\beta}^\lambda) \simeq \mathscr{R}(\lambda, \phi)$, where the equivalent additively decomposes into:*

$$\mathscr{R}(\lambda, \phi) := \mathscr{B}(\lambda, \phi) + \mathscr{V}(\lambda, \phi) + \mathscr{S}(\lambda, \phi) + \kappa^2, \quad (6)$$

with the following deterministic equivalents for the bias, variance, regression shift bias, and irreducible error:

$$\begin{aligned}
\mathscr{B} &= \mu^2 \cdot \beta^\top(\Sigma + \mu I)^{-1}(\widetilde{v}\Sigma + \Sigma_0)(\Sigma + \mu I)^{-1}\beta, \\
\mathscr{V} &= \sigma^2\widetilde{v}, \\
\mathscr{S} &= 2\mu \cdot \beta^\top(\Sigma + \mu I)^{-1}\Sigma_0(\beta_0 - \beta), \\
\kappa^2 &= (\beta_0 - \beta)^\top\Sigma_0(\beta_0 - \beta) + \sigma_0^2.
\end{aligned}$$

Note that the deterministic equivalents presented in Proposition 2.4 depend not only on the regularization parameters (λ, ϕ), but also on the problem parameters $(\Sigma, \beta, \sigma^2)$ and $(\Sigma_0, \beta_0, \sigma_0^2)$, which we have omitted for notational brevity. Since the risk depends additively on σ_0^2, we focus mainly on the effect of (Σ_0, β_0) in our analysis. Extending this result to finite samples is possible by imposing additional distributional assumptions on the features and response. Techniques in Knowles & Yin (2017); Cheng & Montanari (2022); Louart (2022), among others, can be used to obtain non-asymptotic analogs of Proposition 2.4. In this paper, we will focus only on the deterministic equivalents, which capture the first-order information (akin to expectation) of interest for our goals.

3. Properties of Optimal Regularization

In this section, we focus on the optimal ridge penalty λ^* for the asymptotic out-of-distribution (OOD) risk, defined as[2]:

$$\lambda^* \in \operatorname*{argmin}_{\lambda \geq \lambda_{\min}(\phi)} \mathscr{R}(\lambda, \phi). \quad (7)$$

As discussed in Section 1.2, previous studies have explored the properties of λ^* for ridge regression summarized in Table 1. However, these studies predominantly focus on specific scenarios, such as isotropic signals or features, and do not consider the full range of negative penalty values. Furthermore, their investigations are restricted mainly to the IND setting when $(\Sigma_0, \beta_0) = (\Sigma, \beta)$. We broaden the scope of these results, considering more general scenarios, including anisotropic signals, the full range of (negative) regularization, and both IND and OOD settings.

3.1. In-Distribution Optimal Regularization

We present our initial result for the IND setting, which encompasses and extends the scope of previous works. Based on Proposition 2.4, we can characterize the properties of the optimal ridge penalty λ^* defined in (7) as follows.

Theorem 3.1 (Optimal regularization sign for IND risk)**.** *Assume the setup of Proposition 2.4 with $(\Sigma_0, \beta_0) = (\Sigma, \beta)$.*

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:*

$$\frac{\overline{\operatorname{tr}}[B\Sigma(v\Sigma + I)^{-2}] + \sigma^2}{\overline{\operatorname{tr}}[B\Sigma(v\Sigma + I)^{-3}] + \sigma^2} > \frac{\overline{\operatorname{tr}}[\Sigma(v\Sigma + I)^{-2}]}{\overline{\operatorname{tr}}[\Sigma(v\Sigma + I)^{-3}]}, \quad (8)$$

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

It is worth mentioning that although we state our results for general deterministic signals, our analysis can also incorporate random signals. In such cases, when β is random, one can simply replace B in the conclusion with its expectation $\mathbb{E}[B]$. Next, we highlight some special cases of Theorem 3.1 and compare them with previously known results.

When $\Sigma = I$ or $\mathbb{E}[B] = (\alpha^2/p)I$, it is easy to verify that the general alignment condition (8) does not hold (see Remark D.1). This corresponds to the special cases studied by Dicker (2016); Dobriban & Wager (2018), where $\lambda^* \geq 0$.

The general alignment condition (8) in Theorem 3.1 encompasses the strict alignment conditions in Wu & Xu (2020); Richards et al. (2021). Under strict alignment conditions, Wu & Xu (2020) demonstrate that the optimal ridge penalty is negative in the overparameterized and noiseless setting.

[2]Over the extended reals, there is at least one solution to (7). In case there are multiple solutions λ^* to the problem (7), the subsequent guarantees stated in the paper hold for any solution λ^*.

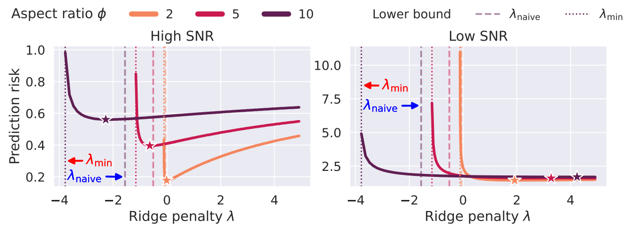

Figure 1: **Illustration of negative or positive optimal regularization under general alignment.** We plot the in-distribution risk of ridge regression against the penalty λ for varying data aspect ratios ϕ in the overparameterized regime. The left and right panels correspond to scenarios when SNR is high ($\sigma^2 = 0.01$) and low ($\sigma^2 = 1$), respectively. The data model has a covariance matrix $(\Sigma_{\text{ar1}})_{ij} := \rho_{\text{ar1}}^{|i-j|}$ with parameter $\rho_{\text{ar1}} = 0.5$, and a coefficient $\beta := \frac{1}{2}(w_{(1)} + w_{(p)})$, where $w_{(j)}$ is the jth eigenvector of Σ_{ar1}.

They assume perfect alignment or misalignment between the signal distribution and the spectrum distribution of the covariance. This also includes the strong and weak features models considered in Richards et al. (2021). When the signal is strictly aligned with the spectrum of Σ and $\sigma^2 = 0$, it can be shown that (8) holds for all $\mu > 0$ (see Proposition D.2).

The general alignment condition (8) allows for a broader range of signal and covariance structures. For example, in scenarios where the signal is the average of the largest and smallest eigenvectors of Σ, the strict alignment condition does not hold. However, these scenarios can still satisfy the general alignment conditions, as we will illustrate shortly.

In the noiseless setting, when $\sigma^2 = 0$, the alignment condition (8) can be expressed succinctly as:

$$\frac{\partial h(\mu, \Sigma_\beta)}{\partial \mu} < \frac{h(\mu, \Sigma)}{\partial \mu}, \tag{9}$$

by defining the function $h(\cdot, \Sigma_\beta)\colon \mu \mapsto \log \overline{\operatorname{tr}}[B\Sigma(\Sigma + \mu I)^{-2}]$ and $\Sigma_\beta = \Sigma\beta\beta^\top$. At a high level, these alignment conditions capture how aligned the signal vector is with the feature covariance matrix. When the alignment is strong, it indicates that the problem is effectively low-dimensional. In such cases, less regularization is needed if the signal energy in this effective direction is sufficiently large.

In the noisy setting, when $\sigma^2 \neq 0$, the condition (8) explicitly trades off the alignment of the signal and the noise level. While Wu & Xu (2020, Proposition 5) and Richards et al. (2021, Corollary 2) provide upper bounds on σ^2 for optimal negative regularization under restricted data models, (8) applies to a wider class of data models. In this sense, Theorem 3.1 extends the previous results on the occurrence of optimal negative regularization to a more general setting.

In Figure 1, we illustrate our theoretical result of Theorem 3.1. When SNR is high, we observe that the optimal ridge penalty can be negative in the overparameterized regime. In particular, for $\phi = 10$, the negative optimal ridge penalty exceeds the bound used in Wu & Xu (2020). It is worth noting that the scenario depicted in Figure 1 does not satisfy the strict alignment condition, but is incorporated by our characterization in Theorem 3.1.

3.2. Out-of-Distribution Optimal Regularization

We now investigate the behavior of the optimal ridge penalty under covariate shift or regression shift. In the IND setting, when $\Sigma_0 = \Sigma$ and the signals are isotropic ($\beta\beta^\top \simeq (\alpha^2/p)I$), previous works (Dobriban & Wager, 2018; Han & Xu, 2023) show that the optimal ridge penalty is $\lambda^* = \phi/\text{SNR} \geq 0$. Interestingly, even when allowing for negative regularization and covariate shift ($\Sigma_0 \neq \Sigma$), it is easy to check that the optimal λ remains positive in data models with isotropic signals for any $\phi \in (0, \infty)$.

Proposition 3.2 (Optimal regularization under covariate shift and random signal)**.** *When $\Sigma_0 \neq \Sigma$ and $\beta_0 = \beta$, assuming isotropic signals $\mathbb{E}[\beta\beta^\top] = (\alpha^2/p)I$, we have $\lambda^*(\phi) = \phi/\text{SNR} = \operatorname{argmin}_\lambda \mathbb{E}_\beta[\mathscr{R}(\lambda, \phi)]$. Furthermore, for $\lambda_{\min} < 0$ such that $X^\top X/n \succ \lambda_{\min} I$, even the non-asymptotic OOD risk (3) is minimized at $\lambda_p^* = \phi_p/\text{SNR}$, where $\phi_p = p/n$.*

We remark that Proposition 3.2 can also be seen as a result of a Bayes optimality argument. However, we provide a more direct argument in Appendix D.3. A result of this flavor for random features is also obtained by Tripuraneni et al. (2021, Proposition 6.1). An interesting observation from Proposition 3.2 is that the optimal ridge penalty does not depend on Σ_0. This implies that when the signal is isotropic, one does not need to worry about the covariate shift when tuning the penalty. Therefore, generalized cross-validation for IND risks (Patil et al., 2021) still yields optimal penalties even for OOD risks under isotropic signals!

However, it is important to note that this result holds specifically for random isotropic signals, which may not be realistic in practice. In scenarios where (near) random isotropic signals are not present, the question of data-dependent regularization tuning in the OOD setting is not as straightforward. We do not consider data-dependent tuning in the current paper and instead focus on the theoretical properties of the (oracle) optimal regularization (and the corresponding risk). Recently, Wang (2023) propose a method for data-dependent tuning of ridge regression under covariate shift by generating pseudo-labels with an undersmoothed ridge regression on the training data. While the method has adaptivity guarantees in the low-dimensional regime, its consistency in the proportional asymptotics regime is not clear yet, and is an interesting direction for future investigation.

Proposition 3.2 suggests that the optimal penalty λ^* remains invariant for OOD risks under isotropic signals. Although random isotropic signals make the theory more tractable, they generally are not realistic in practice. The following

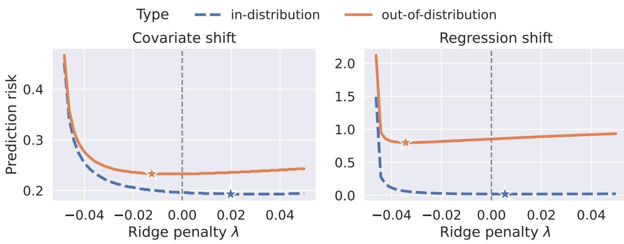

Figure 2: **Covariate and regression shift can lead to negative optimal regularization in both underparameterized and overparameterized regimes.** The plot shows the IND and OOD risks against λ in the high SNR setting ($\sigma^2 = 0.01$ and $\sigma_0^2 = 0$). The left panel shows the overparameterized regime ($\phi = 1.5$) where the optimal ridge penalty λ^* is negative under covariate shift, when $\Sigma = I$, $\Sigma_0 = \Sigma_{\text{ar1}}$, and $\beta = \beta_0 = \frac{1}{2}(w_{(1)} + w_{(p)})$. The right panel shows the underparameterized regime ($\phi = 0.5$) where the optimal ridge penalty λ^* is negative under regression shift, when $\Sigma = \Sigma_0 = \Sigma_{\text{ar1}}$, $\beta = \frac{1}{2}(w_{(1)} + w_{(p)})$, and $\beta_0 = 2\beta$.

result examines the optimal ridge penalty under deterministic signals, with similar conclusions holding for random anisotropic signals.

Theorem 3.3 (Optimal regularization under covariate shift and deterministic signal)**.** *Assume the setting of Proposition 2.4 with $\Sigma_0 \neq \Sigma$, $\beta_0 = \beta$.*

1. (Underparameterized) *When $\phi < 1$, we have $\lambda^* \geq 0$.*
2. (Overparameterized) *When $\phi > 1$, if $\Sigma_0 = I$ (corresponding to the estimation risk), then we have $\lambda^* \geq 0$.*
3. (Overparameterized) *When $\phi > 1$, if $\Sigma = I$ and*

$$\overline{\text{tr}}[\Sigma_0 B] > \overline{\text{tr}}[\Sigma_0] \left(\overline{\text{tr}}[B] + \frac{(1 + \mu(0,\phi))^3}{\mu(0,\phi)^3} \sigma^2 \right), \quad (10)$$

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

When $\phi < 1$, Part (1) of Theorem 3.3 suggests that the optimal ridge penalty λ^* remains non-negative even under covariate shift. Similarly, when $\phi > 1$, Part (2) of Theorem 3.3 (for the estimation risk) also guarantees a non-negative λ^*. However, it is quite surprising that in the overparameterized regime, even with deterministic features, the optimal ridge penalty λ^* could be negative in Part (3). In particular, in noiseless setting ($\sigma^2 = 0$), the condition (10) reduces to the strict alignment condition on (Σ_0, β); see Proposition D.2.

While Theorem 3.3 restricts $\Sigma = I$ to simplify the condition when $\phi > 1$, we provide the condition with general Σ in Equation (39). However, taking $\Sigma = I$ suffices to highlight this rather surprising sign-reversal phenomenon.

Theorem 3.4 (Optimal regularization under regression shift)**.** *Assume the setup of Proposition 2.4 with $\Sigma_0 = \Sigma$, $\beta_0 \neq \beta$.*

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

$$\overline{\text{tr}}[B_0 \Sigma^2 (\Sigma + \mu I)^{-2}] > \overline{\text{tr}}[B \Sigma^2 (\Sigma + \mu I)^{-2}], \quad (11)$$

where $B = \beta\beta^\top$ and $B_0 = \beta_0\beta^\top$, then we have $\lambda^* < 0$.
2. (Overparameterized) *When $\phi > 1$, if the general alignment conditions (8) and (11) hold, then we have $\lambda^* < 0$.*

Table 2: **Optimal ridge penalty for OOD risks on MNIST gradually becomes more negative with increasing distribution shift.** We gradually shift the test distribution by excluding samples with specific labels. For more details, please refer to Appendix G.1.2.

Excl. labels:	$\varnothing$	$\{4\}$	$\{3,4\}$	$\{2,3,4\}$	$\{1,2,3,4\}$
λ^*	1.03	0.48	0.00	-0.48	1.44

Similarly to Part (3) of Theorem 3.3, Part (1) of Theorem 3.4 is rather surprising. As shown in Table 1, λ^* is always positive for the IND setting when $\phi < 1$. However, Theorem 3.4 suggests that λ^* can be negative, even for isotropic signals, when there is some alignment or misalignment between $\beta^\top\Sigma$ and $(\beta - \beta_0)^\top\Sigma$. In fact, when $\Sigma = \Sigma_0 = I$ and $\langle \beta, \beta_0 \rangle \geq \|\beta\|_2^2$, the alignment condition $\beta^\top\Sigma^2(\Sigma + \mu I)^{-2}(\beta_0 - \beta) \geq 0$ in Theorem 3.4 always holds for all $\mu > 0$. It is worth noting that we assume $\sigma^2 = o(1)$ in Theorem 3.4 for simplicity, but a more general balance condition that holds for any $\sigma^2 > 0$ is provided in (39).

The numerical illustrations in Figure 2 demonstrate the results of Theorems 3.3 and 3.4. As shown, while the optimal ridge penalties λ^* for the IND prediction risks are positive, the OOD prediction risk can be negative and approach its lower limit. Similar observations also occur in real-world MNIST datasets (see Table 2 and Appendix G.1.2 for the experimental details). This phenomenon arises due to distribution shift, which effectively aligns β_0 and Σ_0. In some cases, it provides a possible explanation for the success of interpolators in practice, as the optimal ridge penalty λ^* can become negative under distribution shift, e.g., with random features regression (Tripuraneni et al., 2021).

Intuitively, negative regularization may be optimal in certain cases due to the implicit bias of the overparameterized ridge estimator, even when $\lambda = 0$. When the signal energy is sufficiently high, it can be beneficial to "subtract" some of this bias at the expense of increased variance. Negative regularization effectively reduces this bias and can, therefore, be the optimal choice. In a broader context, when there is implicit regularization, such as self-regularization resulting from the data structure, and we desire the overall regularization to be smaller than this inherent amount, negative "external" regularization can help counterbalance it.

4. Properties of Optimal Risk

Under the general covariance structure Σ, Theorem 6 in Patil & Du (2023) shows that optimal ridge regression exhibits a monotonic risk profile in the IND setting ($\Sigma_0 = \Sigma$) and effectively avoids the phenomena of double and multiple descents. This observation motivates a broader investigation into the monotonic behavior of regularization optimization in the out-of-distribution setting. In this section, we investigate the monotonicity of the optimal OOD risk and converse for the suboptimal OOD risk.

4.1. Optimal Risk Monotonicity

To begin with, we examine the case of isotropic signals under covariate shift. A direct consequence of Proposition 3.2 is the monotonicity property of the risk in this scenario.

Proposition 4.1 (Optimal risk under isotropic signals)**.** *When $\Sigma_0 \neq \Sigma$ and $\beta = \beta_0$, assuming isotropic signals $\mathbb{E}[\beta\beta^\top] = (\alpha^2/p)I$ the optimal risk obtained at $\lambda^*(\phi) = \phi/\mathrm{SNR}$ is given by:*

$$\mathbb{E}_\beta[\mathscr{R}(\lambda^*, \phi)] = \alpha^2\mu^* \,\overline{\mathrm{tr}}[\Sigma_0(\Sigma + \mu^* I)^{-1}] + \sigma_0^2, \quad (12)$$

where $\mu^* = \mu(\lambda^*, \phi)$. Furthermore, the left side of (12) is strictly increasing in ϕ if $\mathrm{SNR} \in (0, \infty)$ and σ_0^2 are fixed and strictly increasing in SNR if ϕ, σ^2, and σ_0^2 are fixed.

Proposition 4.1 shows that the optimal OOD risk is a monotonic function of ϕ and SNR. This is intuitive because one would expect that having more data (smaller ϕ) or larger SNR would result in a lower prediction risk. In contrast, the ridge or ridgeless predictor computed on the full data does not exhibit this property (Hastie et al., 2022, Figure 2). The optimal penalty λ^* is also monotonically increasing in the data aspect ratio ϕ when SNR is kept fixed.

However, under anisotropic signals, the optimal regularization penalty generally depends on the specific OOD risk being considered. In such cases, it is difficult to obtain analytical formulas for the optimal ridge penalty and the optimal risk. Nonetheless, we can still show that the optimal ODD risk monotonically increases in ϕ and SNR. We formalize this in the following result, which generalizes the aforementioned IND result in Patil & Du (2023, Theorem 6) to the OOD setting, allowing risk optimization over the possible range of negative regularization (with the lower limit as given in Definition 2.3).

Theorem 4.2 (Monotonicity of optimally tuned OOD risk)**.** *For $\lambda \geq \lambda_{\min}(\phi)$ where $\lambda_{\min}(\phi)$ is as in (5), for all $\epsilon > 0$ small enough, the risk of optimal ridge predictor satisfies:*

$$\min_{\lambda \geq \lambda_{\min}(\phi)+\epsilon} R(\widehat{\beta}^\lambda) \;\simeq\; \min_{\lambda \geq \lambda_{\min}(\phi)} \mathscr{R}(\lambda, \phi), \quad (13)$$

and right side of (13) is monotonically increasing in ϕ if SNR and σ_0^2 are fixed. In addition, when $\beta = \beta_0$ it is monotonically increasing in SNR if ϕ, σ^2, and σ_0^2 are fixed.

We find the monotonicity of the optimal risk in ϕ remarkable because it even holds under *arbitrary* covariate and regression shift! The monotonicity in SNR under regression shift can be similarly analyzed but requires fixing more parameters. When considering the IND prediction risk, Patil & Du (2023, Theorem 6) demonstrate that the optimally tuned ridge over $\lambda \in (0, \infty)$ exhibits a monotonic risk profile in the data aspect ratio ϕ. It is somewhat surprising that optimizing over both $(0, \infty)$ and $(\lambda_{\min}(\phi), \infty)$ yields monotonic behavior, as numerically verified in Figure 3.

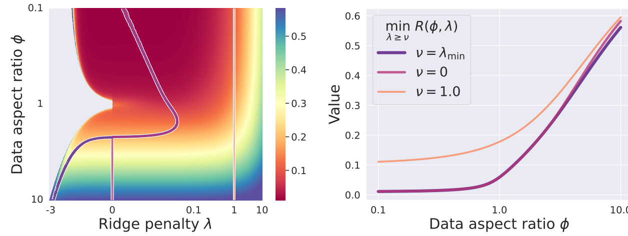

Figure 3: **Ridge regression optimized over $\lambda \geq \nu$ for different thresholds ν has monotonic risk profile.** We showcase the prediction risk of optimal ridge regression under the same data model as in Figure 1, with $\sigma^2 = 0.01$. The left panel shows the heatmap of the risks $\mathscr{R}(\lambda, \phi)$ of ridge regression for different ridge penalties λ and data aspect ratios ϕ. The lines indicate the optimized ridge risks $\min_{\lambda \geq \nu} \mathscr{R}(\lambda, \phi)$ at different thresholds ν. The right panel shows the optimized risk $\min_{\lambda \geq \nu} \mathscr{R}(\lambda, \phi)$ as a function of ϕ.

We also illustrate the optimal risk monotonicity behavior on MNIST in Figure G12. The minimum risk over $(\lambda_{\min}(\phi), \infty)$ can be significantly lower than the one over $(0, \infty)$, particularly in the overparameterized regime. Although most software packages only consider positive regularization for tuning, Figure 3 suggests that allowing negative regularization can lead to significant improvements.

Finally, we consider the converse question: is optimal regularization *necessary* for achieving monotonic risk? When considering the excess IND prediction risk under isotropic design (i.e., $\Sigma = I$), the following theorem demonstrates that the prediction risk of ridge regression is generally non-monotonic in the data aspect ratio $\phi \in (0, \infty)$.

Theorem 4.3 (Non-monotonicity of suboptimally tuned risk)**.** *When $(\Sigma_0, \beta_0) = (\Sigma, \beta)$ and $\Sigma = I$, the risk component equivalents defined in (6) have the following properties:*

1. (Bias component) *For all $\lambda > 0$, $\mathscr{B}(\lambda, \phi)$ is strictly increasing over $\phi \in (0, \lambda + 1)$ and strictly decreasing over $\phi \in (\lambda + 1, \infty)$.*
2. (Variance component) *For all $\lambda > 0$, $\mathscr{V}(\lambda, \phi)$ is strictly increasing over $\phi \in (0, \infty)$.*
3. (Risk) *When $\|\beta\|_2^2 > 0$, for all $\lambda > 0$ and $\epsilon > 0$, there exist $\sigma^2, \phi \in (0, \infty)$, such that $\partial\mathscr{R}(\lambda, \phi)/\partial\phi \leq -\epsilon$, i.e.,*

$$\max_{\sigma^2, \phi \in (0,\infty)} \; \min_{\lambda \geq \lambda_{\min}(\phi)} \partial\mathscr{R}(\lambda, \phi)/\partial\phi \leq -\epsilon. \quad (14)$$

When σ^2 is small, Theorem 4.3 implies that the risk at any fixed λ is non-monotonic, even when $\Sigma = I$. It also explains the lack of risk monotonicity in ridgeless regression. This result extends known results on non-monotonic behavior of variance of ridgeless regression (Yang et al., 2020).

4.2. Connection to Subsampling and Ensembling

We now briefly discuss the connection between subsampling and ridge regularization (Patil & Du, 2023), used to prove the OOD risk monotonicity results in Section 4. To incorporate negative regularization and OOD risks, we extend these equivalences using tools from LeJeune et al. (2024).

Before presenting the extended equivalence results, we introduce several quantities related to the subsampled ridge ensembles. For an index set $\mathcal{I} \subseteq [n]$ of size k, let $L_{\mathcal{I}} \in \mathbb{R}^{n\times n}$ be a diagonal matrix with i-th diagonal 1 if $i \in \mathcal{I}$ and 0 otherwise. The feature matrix and the response vector associated with a subsampled dataset $\{(x_i, y_i) : i \in \mathcal{I}\}$ are $L_{\mathcal{I}}X$ and $L_{\mathcal{I}}y$, respectively. Given a ridge penalty λ, let $\widehat{\beta}_k^{\lambda}(\mathcal{I})$ denote the ridge estimator fitted on the subsample $(L_{\mathcal{I}}X, L_{\mathcal{I}}y)$, consisting of k samples.

When we aggregate the estimators fitted on all subsampled datasets of size k, we obtain the so-called *full-ensemble* estimators $\widehat{\beta}_{k,\infty}^{\lambda}$, which is almost surely $\mathbb{E}[\widehat{\beta}_k^{\lambda}(\mathcal{I}) \mid X, y]$, if we draw $\mathcal{I}$ independently from the set of index sets of size k. As $k, n, p \to \infty$ such that $p/n \to \phi \in (0, \infty)$ and $p/k \to \psi \in [\phi, \infty]$, Lemma E.1 implies the OOD risk equivalence: $R(\widehat{\beta}_{k,\infty}^{\lambda}) \simeq \mathscr{R}(\lambda, \phi, \psi)$ as in (46). When $\psi = \phi$, the equivalent $\mathscr{R}(\lambda, \phi, \phi)$ reduces to $\mathscr{R}(\lambda, \phi)$ defined in (6) and analyzed in previous sections. We are now ready to present our main result in this section.

Theorem 4.4 (Optimal ensemble versus ridge regression under negative regularization)**.** *Let* $\mathscr{R}^{**} := \min_{\psi\ge\phi, \lambda\ge\lambda_{\min}(\phi)} \mathscr{R}(\lambda, \phi, \psi)$. *Then the following statements hold:*

1. (Underparameterized) *When* $\phi < 1$ *and* $\beta_0 = \beta$, $\lambda^* \ge 0$,
$$\mathscr{R}^{**} = \min_{\lambda\ge 0} \mathscr{R}(\lambda, \phi, \phi) = \min_{\psi\ge\phi} \mathscr{R}(0; \phi, \psi).$$

2. (Overparameterized) *When* $\phi \ge 1$, $\lambda^* \ge \lambda_{\min}(\phi)$,
$$\mathscr{R}^{**} = \min_{\lambda\ge\lambda_{\min}(\phi)} \mathscr{R}(\lambda, \phi, \phi) = \min_{\psi\ge\phi} \mathscr{R}(\lambda_{\min}(\phi); \phi, \psi).$$

Theorem 4.4 establishes the OOD risk equivalences between ridge predictors and full-ensemble ridgeless predictors. It shows that in the underparameterized regime, ridgeless ensembles are sufficient to achieve optimal IND risk, which is also supported by Du et al. (2023) when considering optimization over $\lambda \ge 0$. However, in the overparameterized regime, ridgeless ensembles alone are not enough to achieve optimal risk over $\lambda \ge \lambda_{\min}(\phi)$. Explicit negative regularization is required to obtain the optimal risk. In other words, the implicit regularization provided by subsampling is always positive, and under certain data geometries, using $\lambda < 0$ can improve predictive performance.

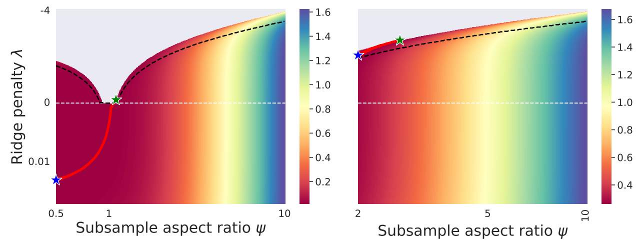

Figure 4: **Negative regularization can help achieve optimal risk in both underparameterized and overparameterized regimes.** The heatmap illustrates the prediction risks for ridge regression as a function of the ridge penalty λ and subsample aspect ratio ψ in the full ensemble. We use the same data model as Figure 3 with $\sigma^2 = 0.01$. The left and right panels show the underparameterized ($\phi = 0.5$) and overparameterized regimes ($\phi = 2$), respectively. The red paths represent the optimal risks, while the blue and green stars indicate the optimal ridge predictor and the optimal full-ensemble ridge with the largest subsample aspect ratio.

The proof of Theorem 4.4 establishes the risk equivalence between the ridge predictor and the full-ensemble ridgeless predictor. As shown in Figure 4, the risk profile of the full-ensemble ridge predictors demonstrates that negative regularization can help achieve optimal risk, particularly in the overparameterized regime when the subsample aspect ratio ψ is greater than 1.

5. Discussion

This paper investigates the optimal regularization and the optimal risk of ridge regression in the OOD setting. The analysis shows the differences between the IND and OOD settings. However, in both cases, the optimal risk is monotonic in the data aspect ratio. The main takeaway is that negative regularization can improve predictive performance in certain data geometries, even more so than the IND setting.

We close the paper by mentioning two future directions. The scope of this paper is limited to standard ridge regression. Under proportional asymptotics, similar conclusions are expected to hold for kernel ridge regression, as the gram matrix linearizes in the sense of asymptotic equivalence (Liang & Rakhlin, 2020; Bartlett et al., 2021; Sahraee-Ardakan et al., 2022; Misiakiewicz & Montanari, 2023). Beyond ridge variants, it is of interest to perform a similar analysis for the lasso. Preliminary investigations in the literature suggest that, similar to optimal ridge regression, the optimal lasso also exhibits monotonic risk in the overparameterization ratio (Patil et al., 2022). For a comparative illustration, see Figures G13 and G14. It is also of interest to understand the monotonicity properties of the optimal risk for other convex penalized M-estimators in general (Thrampoulidis et al., 2018). Empirical investigations indicate that these estimators also exhibit monotonic risk in the data aspect ratio under optimal regularization (Patil & Du, 2023). Making these observations rigorous is a promising future direction.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

Software and Data

The source code for generating all of our figures is included in the supplementary material in the folder named 'code', along with details about the computational resources used and other relevant information. The file named 'README.md' lists the organizational structure.

Acknowledgments

We thank Edgar Dobriban, Avi Feller, Qiyang Han, Dmitry Kobak, Arun Kuchibhotla, Daniel LeJeune, Jaouad Mourtada, Alessandro Rinaldo, Alexander Wei, and Yuting Wei for helpful conversations surrounding this work. We thank the anonymous reviewers for their valuable feedback and suggestions that have improved the paper (in particular, the addition of Appendix B). We also thank the computing support provided by the ACCESS allocation MTH230020 for some of the experiments performed on the Bridges2 system at the Pittsburgh Supercomputing Center. PP and RJT were supported by ONR grant N00014-20-1-2787 during the course of this work.

References

Bach, F. High-dimensional analysis of double descent for linear regression with random projections. *SIAM Journal on Mathematics of Data Science*, 6(1):26–50, 2024.

Bartlett, P. L., Long, P. M., Lugosi, G., and Tsigler, A. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

Bartlett, P. L., Montanari, A., and Rakhlin, A. Deep learning: A statistical viewpoint. *Acta Numerica*, 30:87–201, 2021.

Belkin, M., Hsu, D., and Xu, J. Two models of double descent for weak features. *SIAM Journal on Mathematics of Data Science*, 2(4):1167–1180, 2020.

Bellec, P., Du, J.-H., Koriyama, T., Patil, P., and Tan, K. Corrected generalized cross-validation for finite ensembles of penalized estimators. *arXiv preprint arXiv:2310.01374*, 2023.

Björkström, A. and Sundberg, R. A generalized view on continuum regression. *Scandinavian Journal of Statistics*, 26(1):17–30, 1999.

Buja, A., Brown, L., Berk, R., George, E., Pitkin, E., Traskin, M., Zhang, K., and Zhao, L. Models as approximations I. *Statistical Science*, 34(4):523–544, 2019a.

Buja, A., Brown, L., Kuchibhotla, A. K., Berk, R., George, E., and Zhao, L. Models as approximations II. *Statistical Science*, 34(4):545–565, 2019b.

Canatar, A., Bordelon, B., and Pehlevan, C. Out-of-distribution generalization in kernel regression. *Advances in Neural Information Processing Systems*, 2021.

Cheng, C. and Montanari, A. Dimension free ridge regression. *arXiv preprint arXiv:2210.08571*, 2022.

D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., Chen, C., Deaton, J., Eisenstein, J., Hoffman, M. D., et al. Underspecification presents challenges for credibility in modern machine learning. *The Journal of Machine Learning Research*, 23(1):10237–10297, 2022.

Davidson, J. *Stochastic Limit Theory: An Introduction for Econometricians*. Oxford University Press, 1994.

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

Dobriban, E. Efficient computation of limit spectra of sample covariance matrices. *Random Matrices: Theory and Applications*, 4(4), 2015.

Dobriban, E. and Wager, S. High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics*, 46(1):247–279, 2018.

Du, J.-H., Patil, P., and Kuchibhotla, A. K. Subsample ridge ensembles: Equivalences and generalized cross-validation. In *International Conference on Machine Learning*, 2023.

Fink, A. M. and Jodeit Jr., M. On Chebyshev's other inequality. *Lecture Notes-Monograph Series*, pp. 115–120, 1984.

Györfi, L., Kohler, M., Krzyzak, A., and Walk, H. *A Distribution-free Theory of Nonparametric Regression*. Springer, 2006.

Han, Q. and Xu, X. The distribution of ridgeless least squares interpolators. *arXiv preprint arXiv:2307.02044*, 2023.

Hastie, T. Ridge regularization: An essential concept in data science. *Technometrics*, 62(4):426–433, 2020.

Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. J. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949–986, 2022.

Hoerl, A. E. and Kennard, R. W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970a.

Hoerl, A. E. and Kennard, R. W. Ridge regression: Applications to nonorthogonal problems. *Technometrics*, 12(1): 69–82, 1970b.

Hua, T. A. and Gunst, R. F. Generalized ridge regression: a note on negative ridge parameters. *Communications in Statistics Theory and Methods*, 12(1):37–45, 1983.

Jacot, A., Simsek, B., Spadaro, F., Hongler, C., and Gabriel, F. Kernel alignment risk estimator: Risk prediction from training data. *Advances in Neural Information Processing Systems*, 2020.

Knowles, A. and Yin, J. Anisotropic local laws for random matrices. *Probability Theory and Related Fields*, 169: 257–352, 2017.

Kobak, D., Lomond, J., and Sanchez, B. The optimal ridge penalty for real-world high-dimensional data can be zero or negative due to the implicit ridge regularization. *Journal of Machine Learning Research*, 21, 2020.

Koehler, F., Zhou, L., Sutherland, D. J., and Srebro, N. Uniform convergence of interpolators: Gaussian width, norm bounds and benign overfitting. *Advances in Neural Information Processing Systems*, 2021.

LeJeune, D., Patil, P., Javadi, H., Baraniuk, R. G., and Tibshirani, R. J. Asymptotics of the sketched pseudoinverse. *SIAM Journal on Mathematics of Data Science*, 2024.

Liang, T. and Rakhlin, A. Just interpolate: Kernel "ridgeless" regression can generalize. *The Annals of Statistics*, 2020.

Louart, C. Sharp bounds for the concentration of the resolvent in convex concentration settings. *arXiv preprint arXiv:2201.00284*, 2022.

Loureiro, B., Gerbelot, C., Cui, H., Goldt, S., Krzakala, F., Mezard, M., and Zdeborova, L. Learning curves of generic features maps for realistic datasets with a teacher-student model. In *Advances in Neural Information Processing Systems*, 2021.

Mallinar, N., Simon, J. B., Abedsoltan, A., Pandit, P., Belkin, M., and Nakkiran, P. Benign, tempered, or catastrophic: A taxonomy of overfitting. *arXiv:2207.06569*, 2022.

Mel, G. and Ganguli, S. A theory of high dimensional regression with arbitrary correlations between input features and target functions: Sample complexity, multiple descent curves and a hierarchy of phase transitions. In *International Conference on Machine Learning*, 2021.

Misiakiewicz, T. and Montanari, A. Six lectures on linearized neural networks. *arXiv preprint arXiv:2308.13431*, 2023.

Muthukumar, V., Vodrahalli, K., Subramanian, V., and Sahai, A. Harmless interpolation of noisy data in regression. *IEEE Journal on Selected Areas in Information Theory*, 1 (1):67–83, 2020.

Nakkiran, P., Venkat, P., Kakade, S. M., and Ma, T. Optimal regularization can mitigate double descent. In *International Conference on Learning Representations*, 2021.

Patil, P. and Du, J.-H. Generalized equivalences between subsampling and ridge regularization. In *Conference on Neural Information Processing Systems*, 2023.

Patil, P. and LeJeune, D. Asymptotically free sketched ridge ensembles: Risks, cross-validation, and tuning. In *International Conference on Learning Representations*, 2024.

Patil, P., Wei, Y., Rinaldo, A., and Tibshirani, R. J. Uniform consistency of cross-validation estimators for high-dimensional ridge regression. In *International Conference on Artificial Intelligence and Statistics*, 2021.

Patil, P., Kuchibhotla, A. K., Wei, Y., and Rinaldo, A. Mitigating multiple descents: A model-agnostic framework for risk monotonization. *arXiv preprint arXiv:2205.12937*, 2022.

Patil, P., Du, J.-H., and Kuchibhotla, A. K. Bagging in overparameterized learning: Risk characterization and risk monotonization. *Journal of Machine Learning Research*, 24(319):1–113, 2023.

Pesce, L., Krzakala, F., Loureiro, B., and Stephan, L. Are Gaussian data all you need? Extents and limits of universality in high-dimensional generalized linear estimation. *arXiv:2302.08923*, 2023.

Richards, D., Mourtada, J., and Rosasco, L. Asymptotics of ridge (less) regression under general source condition. In *International Conference on Artificial Intelligence and Statistics*, 2021.

Ross, S. M. *Simulation*. Academic Press, 2022.

Sahraee-Ardakan, M., Emami, M., Pandit, P., Rangan, S., and Fletcher, A. K. Kernel methods and multi-layer perceptrons learn linear models in high dimensions. *arXiv preprint arXiv:2201.08082*, 2022.

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

Simon, J. B., Dickens, M., Karkada, D., and Deweese, M. The eigenlearning framework: A conservation law perspective on kernel ridge regression and wide neural networks. *arxiv:2110.03922*, 2021.

Simon, J. B., Karkada, D., Ghosh, N., and Belkin, M. More is better in modern machine learning: When infinite overparameterization is optimal and overfitting is obligatory, 2023.

Sollich, P. Gaussian process regression with mismatched models. *Advances in Neural Information Processing Systems*, 2001.

Thrampoulidis, C., Abbasi, E., and Hassibi, B. Precise error analysis of regularized M-estimators in high dimensions. *IEEE Transactions on Information Theory*, 64(8):5592–5628, 2018.

Tripuraneni, N., Adlam, B., and Pennington, J. Covariate shift in high-dimensional random feature regression. *arXiv preprint arXiv:2111.08234*, 2021.

Wang, K. Pseudo-labeling for kernel ridge regression under covariate shift. *arXiv preprint arXiv:2302.10160*, 2023.

Wei, A., Hu, W., and Steinhardt, J. More than a toy: Random matrix models predict how real-world neural representations generalize. *arXiv preprint arXiv:2203.06176*, 2022.

Wu, D. and Xu, J. On the optimal weighted ℓ_2 regularization in overparameterized linear regression. *Advances in Neural Information Processing Systems*, 33:10112–10123, 2020.

Yang, T.-L. and Suzuki, J. Dropout drops double descent. *arXiv preprint arXiv:2305.16179*, 2023.

Yang, Z., Yu, Y., You, C., Steinhardt, J., and Ma, Y. Rethinking bias-variance trade-off for generalization of neural networks. In *International Conference on Machine Learning*, pp. 10767–10777, 2020.

Zhou, L., Koehler, F., Sutherland, D. J., and Srebro, N. Optimistic rates: A unifying theory for interpolation learning and regularization in linear regression. *Journal of Data Science*, 1(1), 2023.

Appendix

This supplement serves as a companion to the paper titled "Optimal Ridge Regularization for Out-of-Distribution Prediction". In the first section, we provide an outline of the supplement in Table 3. We also include a summary of the general and specific notation used in the main paper and the supplement in Tables 4 and 5, respectively. In addition, in Table 6, we provide an explanation of and pointers to the abbreviations used in the references mentioned in Table 1.

A. Organization and Notation

A.1. Organization

Table 3: Outline of the supplement.

Section	Subsection	Purpose
Overview of supplement		
Appendix A	Appendix A.1	Organization
	Appendix A.2	General notation
	Appendix A.3	Specific notation
	Appendix A.4	Reference key
Further details in Section 1		
Appendix B	Appendix B.1	Background details on ridge regression with negative regularization
Proofs in Section 2		
Appendix C	Appendix C.1	Proof of Proposition 2.4 (out-of-distribution risk asymptotics)
Proofs in Section 3		
Appendix D	Appendix D.1	Proof of Theorem 3.1 (sign of optimal regularization for IND)
	Appendix D.2	Special cases of the general alignment condition in Theorem 3.1
	Appendix D.3	Proof of Proposition 3.2 (optimal regularization with isotropic signals)
	Appendix D.4	Proof of Theorem 3.3 (sign of optimal regularization with covariate shift)
	Appendix D.5	Proof of Theorem 3.4 (sign of optimal regularization with regression shift)
	Appendix D.6	Helper lemmas (derivatives of components of risk deterministic approximation)
Proofs in Section 4		
Appendix E	Appendix E.1	Proof of Proposition 4.1 (optimal risk under isotropic signals)
	Appendix E.2	Proof of Theorem 4.2 (risk monotonicity with optimal regularization)
	Appendix E.3	Proof of Theorem 4.3 (risk non-monotonicity with suboptimal regularization)
	Appendix E.4	Proof of Theorem 4.4 (optimal subsample versus ridge with negative regularization)
	Appendix E.5	Helper lemmas (optimal risk monotonicities, full ensemble OOD risk equivalences)
Some technical lemmas		
Appendix F	Appendix F.1	Properties of minimum limiting negative ridge penalty
	Appendix F.2	Analytic properties of certain fixed-point solutions under negative regularization
	Appendix F.3	Contour of fixed-point solutions under negative regularization
Additional experiments		
Appendix G	Appendix G.1	Additional illustrations for Section 3
	Appendix G.3	Additional illustrations for Section 5

A.2. General Notation

Table 4: A summary of general notation used in the paper and the supplement.

Notation	Description
Fonts	
Non-bold lower case	Denotes vectors (e.g., a, b, c)
Non-bold upper case	Denotes matrices (e.g., A, B, C)
Calligraphic font	Denotes sets (e.g., $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$)
Script font	Denotes certain limiting functions (e.g., $\mathscr{A}$, $\mathscr{B}$, $\mathscr{C}$)
Sets	
$\mathbb{N}$	Set of natural numbers
(a, b, c)	(Ordered) tuple of elements a, b, c
$\{a, b, c\}$	Set of elements a, b, c
$[n]$	Set $\{1, \dots, n\}$ for a natural number n
$\mathbb{R}$, $\mathbb{R}_{\geq 0}$	Set of real and non-negative real numbers
$\mathbb{C}$, $\mathbb{C}^{+}$, $\mathbb{C}^{-}$	Set of complex numbers and upper and lower complex half-planes
Operations	
$\mathrm{tr}[A]$, $\overline{\mathrm{tr}}[A]$, $\det(A)$	Trace, average trace ($\mathrm{tr}[A]/p$), and determinant of a square matrix $A \in \mathbb{R}^{p \times p}$
B^{-1}	Inverse of an invertible square matrix $B \in \mathbb{R}^{p \times p}$
$C^{\dagger}$	Moore-Penrose inverse of a general rectangular matrix $C \in \mathbb{R}^{n \times p}$
$\mathrm{rank}(C)$, $\mathrm{null}(C)$	Rank and nullity of a general rectangular matrix $C \in \mathbb{R}^{n \times p}$
$D^{1/2}$	Principal square root of a positive semidefinite matrix D
$f(D)$	Matrix obtained by applying a function $f : \mathbb{R} \to \mathbb{R}$ to a positive semidefinite matrix D
I, 1, 0	The identity matrix, the all-ones vector, the all-zeros vector
Norms	
$\Vert u \Vert_q$	The ℓ_q norm of a vector u for $q \geq 1$
$\Vert u \Vert_A$	The induced ℓ_2 norm of a vector u with respect to a positive semidefinite matrix A
$\Vert f \Vert_{L_q}$	The L_q norm of a function f for $q \geq 1$
$\Vert A \Vert_{\mathrm{op}}$	Operator (or spectral) norm of a real matrix A
$\Vert A \Vert_{\mathrm{tr}}$	Trace (or nuclear) norm of a real matrix A
$\Vert A \Vert_F$	Frobenius norm of a real matrix A
Orders	
$X = \mathcal{O}_\upsilon(Y)$	$\vert X \vert \leq C_\upsilon Y$ for some constant C_υ that may depend on the ambient parameter υ
$X \lesssim_\upsilon Y$	$X \leq C_\upsilon Y$ for some constant C_υ that may depend on the ambient parameter υ
$u \leq v$	Lexicographic ordering for vectors u and v
$A \preceq B$	Loewner ordering for symmetric matrices A and B
Asymptotics	
$\mathcal{O}_{\mathbb{P}}$, $o_{\mathbb{P}}$	Probabilistic big-O and little-o notation
$C \simeq D$	Asymptotic equivalence of matrices C and D (see Section 2 for more details)
$\xrightarrow{\text{d}}$, $\xrightarrow{\text{p}}$, $\xrightarrow{\text{a.s.}}$	Denotes convergence in distribution, probability, and almost sure convergence

Some additional conventions used throughout the supplement:

- If a proof of a statement is separated from the statement, the statement is restated (while keeping the original numbering) along with the proof for convenience.
- If no subscript is present for the norm $\Vert u \Vert$ of a vector u, it is assumed to be the ℓ_2 norm of u.
- We use C, C' to denote positive absolute constants.

A.3. Specific Notation

Table 5: A summary of specific notation used in the paper and the supplement.

Notation	Description
In-distribution	
P_x, $P_{y\|x}$, $P_{x,y}$	Distribution of the train features supported on $\mathbb{R}^p$, conditional distribution of the train response supported on $\mathbb{R}$, and the joint distribution
Σ	Covariance matrix in $\mathbb{R}^{p\times p}$ of the train feature distribution
$r_{\min}$ ($r_{\max}$)	Minimum (maximum) eigenvalue of the train covariance matrix
σ^2, α^2	Conditional variance and linearized signal energy of the train response distribution
SNR	In-distribution signal-to-noise ratio α^2/σ^2
β	Coefficients of the (population) linear projection of y onto x
$\{(x_i, y_i)\}_{i=1}^n$	Train samples of size n in $\mathbb{R}^p \times \mathbb{R}$ sampled i.i.d. from $P_{x,y}$
X, y	Train feature matrix in $\mathbb{R}^{n\times p}$ and the corresponding response vector in $\mathbb{R}^n$
$\widehat{\beta}^\lambda$	Ridge estimator fitted on train data (X, y) at regularization level λ
$\widehat{\beta}_k^\lambda(\mathcal{I})$	Ridge estimator fitted on subsampled data $(L_{\mathcal{I}}X, L_{\mathcal{I}}y)$ at the regularization level λ
$\widehat{\beta}_{k,\infty}^\lambda$	Full-ensemble ridge estimator at regularization level λ
Out-of-distribution	
P_{x_0}, $P_{y_0\|x_0}$, P_{x_0,y_0}	Distribution of the test features supported on $\mathbb{R}^p$, conditional distribution of the test response supported on $\mathbb{R}$, and the joint distribution
(x_0, y_0)	Test sample in $\mathbb{R}^p \times \mathbb{R}$ sampled from P_{x_0,y_0} independently of the train data (X, y)
Σ_0	Covariance matrix in $\mathbb{R}^{p\times p}$ of the test feature distribution
σ_0^2	Variance of the conditional test response distribution
β_0	Coefficients of the L_2 linear projection of y_0 onto x_0: $\mathbb{E}[x_0x_0^\top]^{-1}\mathbb{E}[x_0y_0]$
$R(\widehat{\beta}^\lambda)$	Squared test prediction risk of ridge estimator at level λ
Risk parameters	
ϕ, ψ	Limiting data and subsample aspect ratios
λ^*	Optimal ridge penalty
$\lambda_{\min}(\phi)$, $v_{\min}$	Minimum value of ridge regularization allowed at ϕ and the associated fixed-point parameter
$v_p(\lambda, \phi)$	A fixed-point parameter
$\mu(\lambda, \phi)$	Induced amount of implicit regularization at ridge regularization λ and aspect ratio ϕ
$\widetilde{v}_p(\lambda, \phi; \Sigma)$	Function of a fixed-point parameter
$\mathscr{R}(\lambda, \phi)$, $\mathscr{B}(\lambda, \phi)$, $\mathscr{V}(\lambda, \phi)$, $\mathscr{S}(\lambda, \phi)$	Deterministic approximation to the squared risk, the bias, the variance, and the regression shift
κ^2	Inflated out-of-distribution irreducible error
Σ_β	A certain matrix capturing the alignment of feature covariance and signal vector $\Sigma\beta\beta^\top$
$\mathscr{R}(\lambda, \phi, \psi)$	Deterministic approximation to the risk of the full-ensemble estimator

A.4. Reference Abbreviation Key

Table 6: Links to references mentioned in Table 1.

Abbreviation	Reference
HMRT	Hastie, Montanari, Rosset, and Tibshirani (2022)
DW	Dobriban and Wager (2018)
RMR	Richards, Mourtada, and Rosasco (2021)
WX	Wu and Xu (2020)
D	Dicker (2016)

B. Further Details in Section 1

B.1. Background Details on Ridge Regression with Negative Regularization

Recall that the ridge regression estimator $\widehat{\beta}_\lambda \in \mathbb{R}^p$, based on the training data X, y, is defined as the solution to the following regularized least squares problem:

$$\underset{\beta \in \mathbb{R}^p}{\text{minimize}} \; \underbrace{\frac{1}{n}\|y - X\beta\|_2^2}_{\mathcal{L}(\beta)} + \lambda \underbrace{\|\beta\|_2^2}_{\mathcal{R}(\beta)}. \tag{P}$$

Here, λ is a regularization parameter. For ease of reference in the following, denote the least squares loss by $\mathcal{L}$ and ℓ_2 regularization by $\mathcal{R}$. This is referred to as the primal form of ridge regression, denoted by (P). One can write a Lagrangian dual problem of (P) that optimizes weights $\theta \in \mathbb{R}^n$ over the data points as:

$$\underset{\theta \in \mathbb{R}^n}{\text{minimize}} \; \theta^\top (XX^\top/n + \lambda I_n)\theta/2 - \lambda\theta^\top y/n. \tag{D}$$

This is referred to as the dual form of ridge regression, denoted by (D). Let us denote the solution of this problem by $\widehat{\theta}_\lambda$. The estimator $\widehat{\beta}$ is linked to the dual weights as $\lambda\widehat{\beta}_\lambda = X^\top\widehat{\theta}_\lambda$.

Now, there are three cases depending on whether λ is positive, zero, or negative.

- When $\lambda > 0$, regardless of whether $n \geq p$ or $p > n$, both problems (P) and (D) are strictly convex and lead to the following unique ridge estimator (expressed in primal and dual forms, respectively):

$$\widehat{\beta}_\lambda = (X^\top X/n + \lambda I_p)^{-1}X^\top y/n = X^\top(XX^\top/n + \lambda I_n)^{-1}y/n. \tag{15}$$

- When $\lambda = 0$ (more precisely, defined through analytic continuation as $\lambda \to 0^+$) and $X^\top X$ is full rank (which is the case when $n \geq p$ and the design X has independent columns), the optimization problem (P) is still strictly convex and leads to the following unique solution:

$$\widehat{\beta}_0 = (X^\top X/n)^{-1}X^\top y/n.$$

- When $\lambda = 0$ (again, defined using analytic continuation as $\lambda \to 0^+$) and $XX^\top$ is full rank ($p > n$ and the design X has independent rows), the optimization problem (D) is still strictly convex and leads to the following unique solution:

$$\widehat{\beta}_0 = X^\top(XX^\top/n)^{-1}y/n.$$

- More generally, when $n \geq p$ and $\lambda \geq -s_{\min}$, where $s_{\min}$ is the smallest eigenvalue of $X^\top X/n$, the optimization problem (P) is still strictly convex and leads to the primal solution in (15).
- Similarly, when $p > n$ and $\lambda \geq -s_{\min}$, where $s_{\min}$ is the smallest eigenvalue of $XX^\top/n$, the dual to the optimization problem (D) is strictly convex and leads to the dual solution in (15).

By defining the ridge estimator as:

$$\widehat{\beta}_\lambda = (X^\top X/n + \lambda I_p)^\dagger X^\top y/n = X^\top(XX^\top/n + \lambda I_n)^\dagger y/n,$$

where $A^\dagger$ denotes the Moore-Penrose pseudoinverse of the matrix A, we can handle all the cases mentioned above simultaneously. To gain intuition, we visually compare the ridge solutions with positive and negative regularization in Figures B5 and B7 for the under- and over-parameterized regimes, respectively. See (Hua & Gunst, 1983; Björkström & Sundberg, 1999) also for some related discussion.

B.1.1. Visualizing ridge regression solutions for positive versus negative regularization (underparameterized regime)

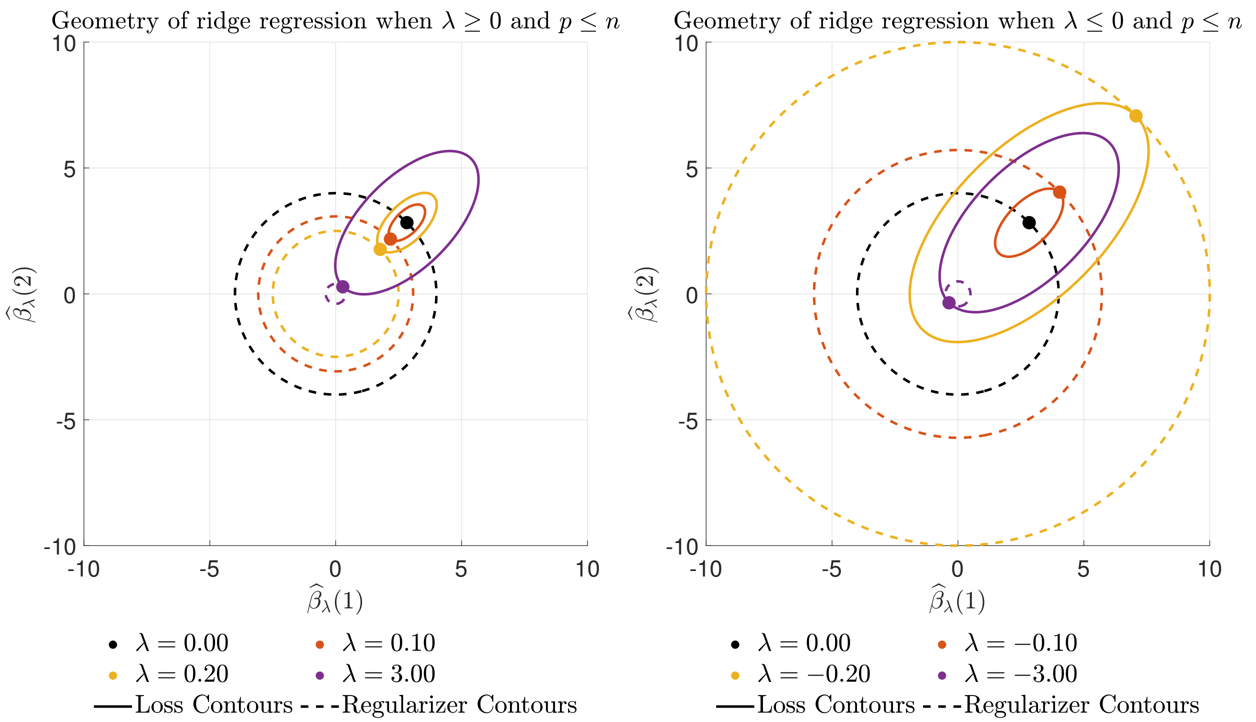

Figure B5: **Comparing the geometry of ridge solutions for positive versus negative regularization levels λ in the underparameterized regime when $p \leq n$.** We consider a two-dimensional underparameterized problem with a design matrix X having the smallest eigenvalue of $X^\top X/n$ as $s_{\min} = 4/3$ and the largest eigenvalue as $s_{\max} = 1/3$. We plot the contours of the ℓ_2 square loss $\mathcal{L}$ and the ℓ_2 regularizer $\mathcal{R}$ in the problem (P) for both positive and negative values of λ. For positive values of $\lambda \in [0, \infty]$, the loss contours touch the constraint contours from the *outside*, while for negative values of $\lambda \in (-\infty, -s_{\max}) \cup (-s_{\min}, 0)$, they touch from the *inside*. To better understand the trend, it helps to think of "tying" $+\infty$ to $-\infty$ on the real line, making it a "projective real line". The values of λ are plotted in the following order: $\boxed{0.0} \to 0.1 \to 0.2 \to 3.0 \to -3.0 \to -0.2 \to -0.1 \to \boxed{0.0}$.

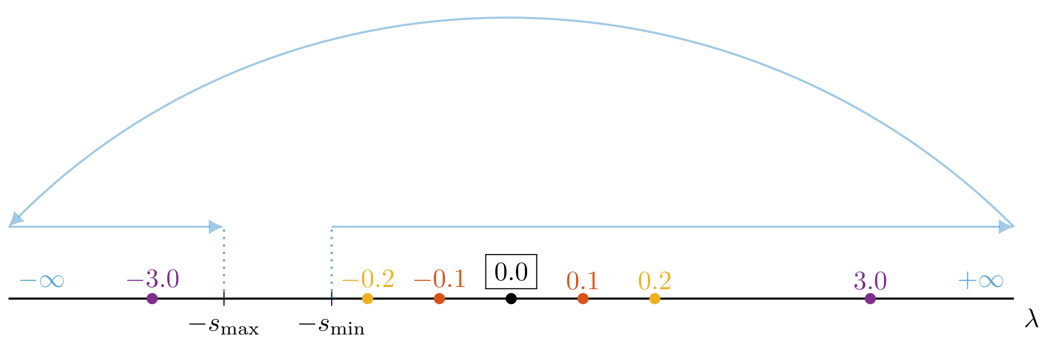

Figure B6: Illustration of the sequence of ridge regularization levels λ used in Figure B5 and the corresponding real projective line. (Note: The λ values are not necessarily drawn to scale.)

B.1.2. Visualizing ridge regression solutions for positive versus negative regularization (overparameterized regime)

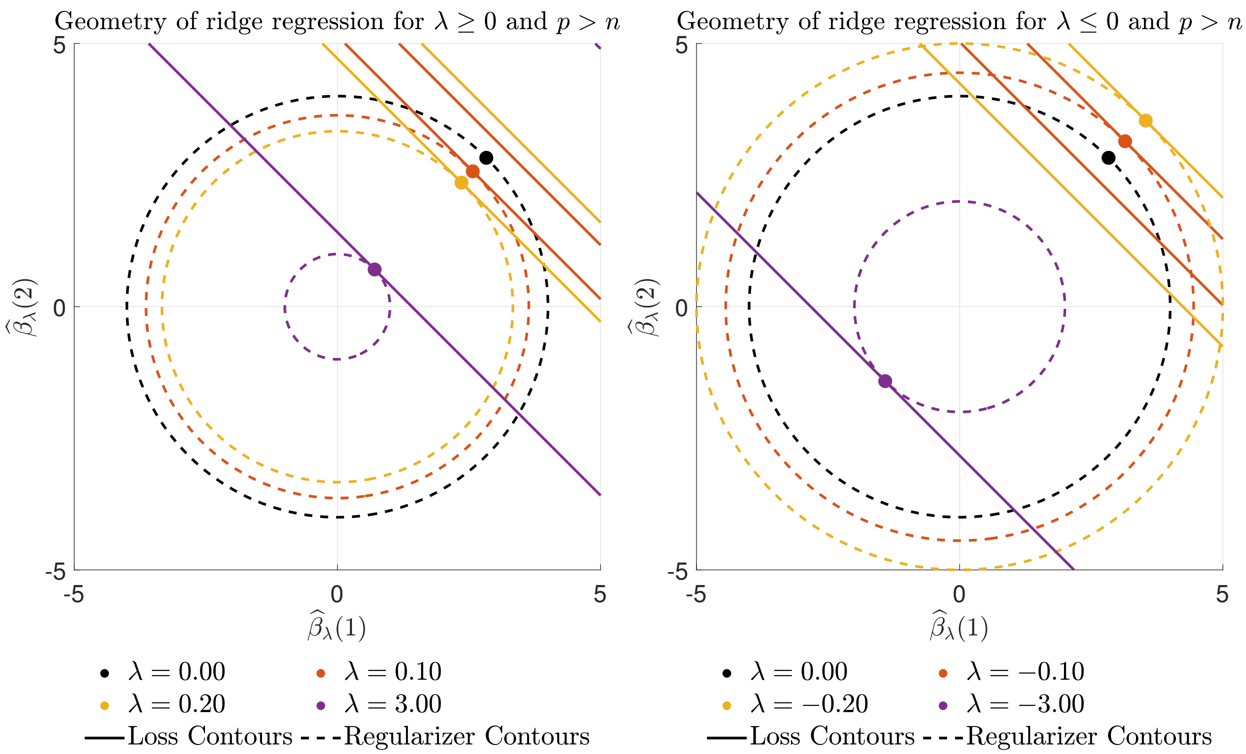

Figure B7: **Comparing the geometry of ridge solutions for positive versus negative regularization levels λ in the overparameterized regime when $p > n$.** We consider a two-dimensional overparameterized problem with the design matrix X having the smallest *non-zero* eigenvalue of $X^\top X/n$ equal to $s^+_{\min} = 1$. Similarly to Figure B5, we plot contours of the ℓ_2 squares loss $\mathcal{L}$ and the ℓ_2 regularizer $\mathcal{R}$ (from problem (P)) for both positive and negative values of λ. As before, note that for positive values of λ, the loss contours touch the constraint contours from the *outside*, while for negative values of λ, they touch from the *inside*. To reiterate, it helps to think of "tying" ∞ to $-\infty$ on the real line to better understand the trend. The values of λ are plotted in the following order: $\boxed{0.0} \to 0.1 \to 0.2 \to 3.0 \to -3.0 \to -0.2 \to -0.1 \to \boxed{0.0}$.

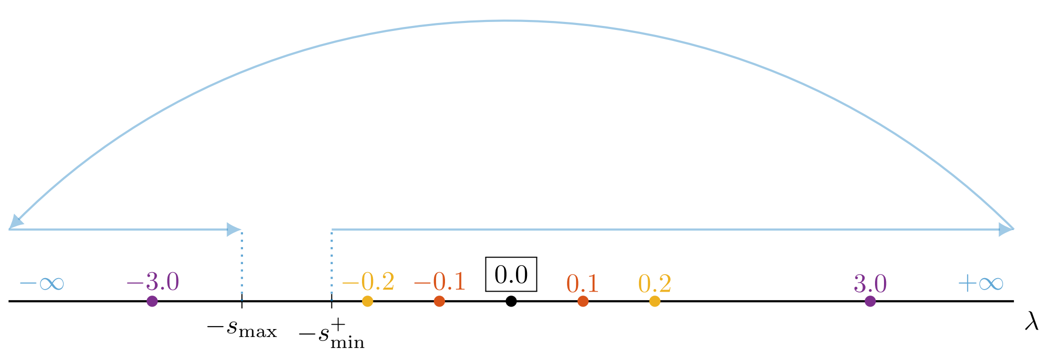

Figure B8: Illustration of the sequence of ridge regularization levels λ used in Figure B7 and the corresponding real projective line. (Note: Here $s^+_{\min}$ is the smallest *non-zero* eigenvalue of $X^\top X/n$.)

C. Proofs in Section 2

C.1. Proof of Proposition 2.4

Proposition 2.4 (Deterministic equivalents for OOD risk)**.** *Under Assumption 2.2, as $n, p \to \infty$ such that $p/n \to \phi \in (0, \infty)$ and $\lambda \in (\lambda_{\min}(\phi), \infty)$, the prediction risk $R(\widehat{\beta}^\lambda)$ defined in (3) admits a deterministic equivalent $R(\widehat{\beta}^\lambda) \simeq \mathscr{R}(\lambda, \phi)$, where the equivalent additively decomposes into:*

$$\mathscr{R}(\lambda, \phi) := \mathscr{B}(\lambda, \phi) + \mathscr{V}(\lambda, \phi) + \mathscr{S}(\lambda, \phi) + \kappa^2, \tag{6}$$

with the following deterministic equivalents for the bias, variance, regression shift bias, and irreducible error:

$$\begin{aligned}
\mathscr{B} &= \mu^2 \cdot \beta^\top (\Sigma + \mu I)^{-1} (\widetilde{v}\Sigma + \Sigma_0)(\Sigma + \mu I)^{-1}\beta, \\
\mathscr{V} &= \sigma^2 \widetilde{v}, \\
\mathscr{S} &= 2\mu \cdot \beta^\top (\Sigma + \mu I)^{-1} \Sigma_0 (\beta_0 - \beta), \\
\kappa^2 &= (\beta_0 - \beta)^\top \Sigma_0 (\beta_0 - \beta) + \sigma_0^2.
\end{aligned}$$

Proof. The main workhorse of the proof is Lemma E.1. Recall that

$$\begin{aligned}
R(\widehat{\beta}^\lambda) &= (\widehat{\beta}^\lambda - \beta_0)^\top \Sigma_0 (\widehat{\beta}^\lambda - \beta_0) + \sigma_0^2 \\
&= (\widehat{\beta}^\lambda - \beta)^\top \Sigma_0 (\widehat{\beta}^\lambda - \beta) + 2(\widehat{\beta}^\lambda - \beta)^\top \Sigma_0 (\beta - \beta_0) + \{(\beta_0 - \beta)^\top \Sigma_0 (\beta_0 - \beta) + \sigma_0^2\}.
\end{aligned} \tag{16}$$

Note that the last quadratic term of (16) is just a constant. From Lemma E.1, we know that the first term of (16) has the following deterministic equivalent:

$$(\widehat{\beta}^\lambda - \beta)^\top \Sigma_0 (\widehat{\beta}^\lambda - \beta) \simeq \mathscr{Q}_p(\lambda, \phi, \phi), \tag{17}$$

where

$$\mathscr{Q}_p(\lambda, \phi, \psi) = \widetilde{c}_p(\lambda, \phi, \psi, \Sigma_0) + \|f_{\text{NL}}\|_{L_2}^2 \widetilde{v}_p(\lambda, \phi, \psi, \Sigma_0),$$

and the non-negative constants $\widetilde{c}_p(\lambda, \phi, \psi, \Sigma_0)$ and $\widetilde{v}_p(\lambda, \phi, \psi, \Sigma_0)$ are defined through the following equations:

$$\frac{1}{v_p(\lambda, \psi)} = \lambda + \psi \int \frac{r}{1 + v_p(\lambda, \psi) r} \, \mathrm{d}H_p(r), \tag{18}$$

$$\widetilde{v}_p(\lambda, \phi, \psi, \Sigma_0) = \frac{\phi \, \overline{\operatorname{tr}}[\Sigma_0 \Sigma (v_p(\lambda, \psi)\Sigma + I)^{-2}]}{v_p(\lambda, \psi)^{-2} - \phi \displaystyle\int \frac{r^2}{(1 + v_p(\lambda, \psi) r)^2} \, \mathrm{d}H_p(r)}, \tag{19}$$

$$\widetilde{c}_p(\lambda, \phi, \psi, \Sigma_0) = \beta^\top (v_p(\lambda, \psi)\Sigma + I)^{-1} (\widetilde{v}_p(\lambda, \phi, \psi, \Sigma_0)\Sigma + \Sigma_0)(v_p(\lambda, \psi)\Sigma + I)^{-1} \beta.$$

For the second term of (16), we also have

$$(\widehat{\beta}^\lambda - \beta)^\top \Sigma_0 (\beta - \beta_0) = {f_{\text{NL}}}^\top \frac{X}{n} (\widehat{\Sigma} + \lambda I)^{-1} \Sigma_0 (\beta - \beta_0) - \beta^\top \lambda (\widehat{\Sigma} + \lambda I)^{-1} \Sigma_0 (\beta - \beta_0),$$

where f_{NL} is defined in the proof of Lemma E.1. The first term vanishes from Patil & Du (2023, Lemma D.2) because f_{NL} and X are uncorrelated, and the remaining part has operator norm tending to zero. It follows that

$$(\widehat{\beta}^\lambda - \beta)^\top \Sigma_0 (\beta - \beta_0) \simeq -\beta^\top (v_p(\lambda, \phi)\Sigma + I)^{-1} \Sigma_0 (\beta - \beta_0). \tag{20}$$

Combining (16), (17), and (20) yields

$$\begin{aligned}
R(\widehat{\beta}^\lambda) &\simeq \mathscr{R}_p(\lambda, \phi, \phi) \\
&:= \mathscr{Q}_p(\lambda, \phi, \phi) - \beta^\top (v_p(\lambda, \phi)\Sigma + I)^{-1} \Sigma_0 (\beta - \beta_0) + \{(\beta_0 - \beta)^\top \Sigma_0 (\beta_0 - \beta) + \sigma_0^2\},
\end{aligned}$$

which completes the proof. ∎

D. Proofs in Section 3

Before presenting the proof of our main results, we introduce some notation to facilitate the upcoming presentation. Define

$$q_b(\Sigma_0, B) = \mu^2 \operatorname{tr}[(\Sigma + \mu I)^{-1}\Sigma_0(\Sigma + \mu I)^{-1}B]/p \tag{21}$$

$$q_v(\Sigma_0, B) = \phi \operatorname{tr}[(\Sigma + \mu I)^{-1}\Sigma_0(\Sigma + \mu I)^{-1}B]/p \tag{22}$$

$$l(\Sigma_0, B) = \operatorname{tr}[(\Sigma + \mu I)^{-1}\Sigma_0 B]/p. \tag{23}$$

We denote the derivative with respect to μ by:

$$q_b'(\Sigma_0, B) := \frac{\partial q_b(\Sigma_0, B)}{\partial \mu},$$
$$q_v'(\Sigma_0, B) := \frac{\partial q_v(\Sigma_0, B)}{\partial \mu}.$$

We also let $B = \beta\beta^\top$ and $B_0 = \beta_0\beta^\top$.

D.1. Proof of Theorem 3.1

Theorem 3.1 (Optimal regularization sign for IND risk)**.** *Assume the setup of Proposition 2.4 with* $(\Sigma_0, \beta_0) = (\Sigma, \beta)$.

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:*

$$\frac{\overline{\operatorname{tr}}[B\Sigma(v\Sigma + I)^{-2}] + \sigma^2}{\overline{\operatorname{tr}}[B\Sigma(v\Sigma + I)^{-3}] + \sigma^2} > \frac{\overline{\operatorname{tr}}[\Sigma(v\Sigma + I)^{-2}]}{\overline{\operatorname{tr}}[\Sigma(v\Sigma + I)^{-3}]}, \tag{8}$$

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

Proof. When $(\Sigma_0, \beta_0) = (\Sigma, \beta)$, the excess bias term $\mathscr{S}(\lambda, \phi) = 0$. Therefore, only the bias and variance terms contribute to the risk. We next split the proof into two cases.

Part (1) Underparameterized regime. When $\phi < 1$, from Lemma F.2, $\mu_p(0, \phi) = 0$, and thus, we have that

$$0 = \mathscr{B}(0, \phi) \leq \min_{\lambda \in [\lambda_{\min}(\phi), 0]} \mathscr{B}(\lambda, \phi). \tag{24}$$

On the other hand, when $\phi < 1$, because $\mathscr{V}'(\lambda, \phi) < 0$, the variance is strictly decreasing over $\lambda \in (\lambda_{\min}(\phi), 0)$. We have

$$\mathscr{V}(0, \phi) < \min_{\lambda \in (\lambda_{\min}(\phi), 0)} \mathscr{V}(\lambda, \phi) \tag{25}$$

It follows that

$$\min_{\lambda \in (\lambda_{\min}(\phi), 0]} \mathscr{R}(\lambda, \phi) \geq \mathscr{R}(0, \phi)$$

and

$$\min_{\lambda \geq \lambda_{\min}} \mathscr{R}(\lambda, \phi) = \min_{\lambda \geq 0} \mathscr{R}(\lambda, \phi).$$

Equivalently, we have $\lambda^* \geq 0$.

Part (2) Overparameterized regime. When $\phi > 1$, we begin by deriving the formula of the derivative of the bias term. When $(\Sigma_0, \beta_0) = (\Sigma, \beta)$, the bias term reduces to

$$\mathscr{B}(\lambda, \phi) = \frac{q_v(\Sigma, \Sigma)}{1 - q_v(\Sigma, \Sigma)} q_b(\Sigma, B) + q_b(\Sigma, B) = \frac{q_b(\Sigma, B)}{1 - q_v(\Sigma, \Sigma)}.$$

From Lemma D.4, its derivative satisfies that

$$\begin{aligned}(1 - q_v(\Sigma, \Sigma))^2 \mathscr{B}'(\lambda, \phi) &= (1 - q_v(\Sigma, \Sigma))^2 h_2'(\mu) \\ &= q_b'(\Sigma, B)(1 - q_v(\Sigma, \Sigma)) + q_b(\Sigma, B) q_v'(\Sigma, \Sigma) \\ &= \frac{\lambda}{\mu} q_b'(\Sigma, B) + \mu q_b'(\Sigma, B) q_v(\Sigma, I) + q_b(\Sigma, B) q_v'(\Sigma, \Sigma).\end{aligned} \tag{26}$$

where the last equality is from the identity

$$1 - q_v(\Sigma, \Sigma) = \frac{\lambda}{\mu} + \mu q_v(\Sigma, I) \tag{27}$$

The in-distribution excess term vanishes, i.e., $\mathcal{E}(\lambda, \phi) = 0$. Therefore, the derivative of the risk with respect to μ satisfies that

$$\begin{aligned}&(1 - q_v(\Sigma, \Sigma))^2 \mathscr{B}'(\lambda, \phi) &(28)\\
&= \frac{\lambda}{\mu} q_b'(\Sigma, B) + \mu q_b'(\Sigma, B) q_v(\Sigma, I) + q_b(\Sigma, B) q_v'(\Sigma, \Sigma)\\
&= \frac{\lambda}{\mu} q_b'(\Sigma, B) + \mu q_b'(\Sigma, B) q_v(\Sigma, I) + q_b(\Sigma, B) q_v'(\Sigma, \Sigma)\\
&= \frac{\lambda}{\mu} \big[\mu \,\overline{\mathrm{tr}}[\Sigma_\beta(\Sigma + \mu I)^{-2}] - \mu^2 \,\overline{\mathrm{tr}}[\Sigma_\beta(\Sigma + \mu I)^{-3}]\big]\\
&\quad + 2\big[\mu \,\overline{\mathrm{tr}}[\Sigma_\beta(\Sigma + \mu I)^{-2}] - \mu^2 \,\overline{\mathrm{tr}}[\Sigma_\beta(\Sigma + \mu I)^{-3}]\big] \cdot \mu\phi \,\overline{\mathrm{tr}}[\Sigma(\Sigma + \mu I)^{-2}]\\
&\quad - 2\mu \,\overline{\mathrm{tr}}[\Sigma_\beta(\Sigma + \mu I)^{-2}] \cdot \mu\phi \,\overline{\mathrm{tr}}[\Sigma^2(\Sigma + \mu I)^{-3}]\\
&= 2\lambda\big[\,\overline{\mathrm{tr}}[\Sigma_\beta(\Sigma + \mu I)^{-3}(\Sigma + \mu I)] - \mu \,\overline{\mathrm{tr}}[\Sigma_\beta(\Sigma + \mu I)^{-3}]\big]\\
&\quad + 2\big[\mu \,\overline{\mathrm{tr}}[\Sigma_\beta(\Sigma + \mu I)^{-2}] - \mu^2 \,\overline{\mathrm{tr}}[\Sigma_\beta(\Sigma + \mu I)^{-3}]\big] \cdot \mu\phi \,\overline{\mathrm{tr}}[\Sigma(\Sigma + \mu I)^{-2}]\\
&\quad - 2\mu \,\overline{\mathrm{tr}}[\Sigma_\beta(\Sigma + \mu I)^{-2}] \cdot \mu\phi \,\overline{\mathrm{tr}}[\Sigma^2(\Sigma + \mu I)^{-3}]\\
&= 2\lambda \,\overline{\mathrm{tr}}[\Sigma_\beta \Sigma(\Sigma + \mu I)^{-3}]\\
&\quad + 2\mu^2\phi \,\overline{\mathrm{tr}}[\Sigma_\beta(\Sigma + \mu I)^{-2}] \cdot (\overline{\mathrm{tr}}[\Sigma(\Sigma + \mu I)^{-2}] - \overline{\mathrm{tr}}[\Sigma(\Sigma + \mu I)^{-3}])\\
&\quad - 2\mu^3\phi \,\overline{\mathrm{tr}}[\Sigma_\beta(\Sigma + \mu I)^{-3}] \cdot \overline{\mathrm{tr}}[\Sigma^2(\Sigma + \mu I)^{-2}]\\
&= 2\lambda \,\overline{\mathrm{tr}}[\Sigma_\beta \Sigma(\Sigma + \mu I)^{-3}]\\
&\quad + 2\mu^3\phi \,\overline{\mathrm{tr}}[\Sigma_\beta(\Sigma + \mu I)^{-2}] \cdot \overline{\mathrm{tr}}[\Sigma(\Sigma + \mu I)^{-3}]\\
&\quad - 2\mu^3\phi \,\overline{\mathrm{tr}}[\Sigma_\beta(\Sigma + \mu I)^{-3}] \cdot \overline{\mathrm{tr}}[\Sigma(\Sigma + \mu I)^{-2}]. &(29)\end{aligned}$$

When $\lambda = 0$, it follows that

$$\begin{aligned}&\mathscr{B}'(\lambda, \phi)\\ &= \frac{2\mu^3\phi}{(1 - q_v(\Sigma, \Sigma))^2}\big(\,\overline{\mathrm{tr}}[\Sigma_\beta(\Sigma + \mu I)^{-2}] \cdot \overline{\mathrm{tr}}[\Sigma(\Sigma + \mu I)^{-3}] - \overline{\mathrm{tr}}[\Sigma_\beta(\Sigma + \mu I)^{-3}] \cdot \overline{\mathrm{tr}}[\Sigma(\Sigma + \mu I)^{-2}]\big),\end{aligned}$$

and

$$\begin{aligned}&\frac{\partial \mathscr{B}(\lambda, \phi)}{\partial \lambda}\\ &= \frac{2\mu^3\phi}{(1 - q_v(\Sigma, \Sigma))^2}\big(\,\overline{\mathrm{tr}}[\Sigma_\beta(\Sigma + \mu I)^{-2}] \cdot \overline{\mathrm{tr}}[\Sigma(\Sigma + \mu I)^{-3}] - \overline{\mathrm{tr}}[\Sigma_\beta(\Sigma + \mu I)^{-3}] \cdot \overline{\mathrm{tr}}[\Sigma(\Sigma + \mu I)^{-2}]\big)\end{aligned}$$

$$\cdot \frac{\partial \mu(\lambda, \phi)}{\partial \lambda}, \tag{30}$$

where $\partial \mu(\lambda, \phi)/\partial \lambda > 0$ from Lemma F.2.

From Lemma F.2 and (30), since $\partial \mu(\lambda, \phi)/\partial \lambda > 0$, we further have that

$$\frac{\partial \mathscr{B}(\lambda, \phi)}{\partial \lambda} \propto \overline{\mathrm{tr}}[\Sigma_\beta(\Sigma + \mu I)^{-2}] \cdot \overline{\mathrm{tr}}[\Sigma(\Sigma + \mu I)^{-3}] - \overline{\mathrm{tr}}[\Sigma_\beta(\Sigma + \mu I)^{-3}] \cdot \overline{\mathrm{tr}}[\Sigma(\Sigma + \mu I)^{-2}],$$

which finishes the proof of the first conclusion.

To obtain the sign of the optimal ridge penalty, note that

$$\begin{aligned}
&(1 - q_v(\Sigma, \Sigma))^2 \mathscr{B}'(\lambda, \phi) \\
&= 2\lambda \, \overline{\mathrm{tr}}[\Sigma_\beta \Sigma(\Sigma + \mu I)^{-3}] \\
&\qquad + 2\mu^3 \phi(\overline{\mathrm{tr}}[\Sigma_\beta(\Sigma + \mu I)^{-2}] \cdot \overline{\mathrm{tr}}[\Sigma(\Sigma + \mu I)^{-3}] - \overline{\mathrm{tr}}[\Sigma_\beta(\Sigma + \mu I)^{-3}] \cdot \overline{\mathrm{tr}}[\Sigma(\Sigma + \mu I)^{-2}]) \\
&= T_1 + T_2.
\end{aligned}$$

When $\phi > 1$, because $\mu \geq 0$, we have that $T_1 < 0$ when $\lambda < 0$ and $T_1 > 0$ when $\lambda > 0$. Also, under the assumption that $\overline{\mathrm{tr}}[\Sigma_\beta(\Sigma + \mu I)^{-2}] \cdot \overline{\mathrm{tr}}[\Sigma(\Sigma + \mu I)^{-3}] - \overline{\mathrm{tr}}[\Sigma_\beta(\Sigma + \mu I)^{-3}] \cdot \overline{\mathrm{tr}}[\Sigma(\Sigma + \mu I)^{-2}] > 0$, we have that $T_2 > 0$ for all $\lambda > \lambda_{\min}(\phi)$, with equality holds only when $\lambda = 0$. Thus, $\mathscr{B}(\lambda, \phi)$ is minimized at $\lambda < 0$.

For the variance term, from Lemma D.4, we have

$$\begin{aligned}
\mathscr{V}'(\lambda, \phi) &= \frac{\sigma^2 q_v'(\Sigma, \Sigma)}{(1 - q_v(\Sigma, \Sigma))^2} \\
&= -2\sigma^2 \frac{\phi \, \overline{\mathrm{tr}}[\Sigma^2(\Sigma + \mu I)^{-3}]}{(1 - q_v(\Sigma, \Sigma))^2} \\
&= 2\phi\sigma^2 \frac{\mu \, \overline{\mathrm{tr}}[\Sigma(\Sigma + \mu I)^{-3}] - \overline{\mathrm{tr}}[\Sigma(\Sigma + \mu I)^{-2}]}{(1 - q_v(\Sigma, \Sigma))^2} \\
&:= \frac{T_3 + T_4}{(1 - q_v(\Sigma, \Sigma))^2} \\
&= \frac{-2\sigma^2 \phi \, \overline{\mathrm{tr}}[\Sigma^2(\Sigma + \mu I)^{-3}]}{(1 - q_v(\Sigma, \Sigma))^2}.
\end{aligned} \tag{31}$$

Note that $-2\sigma^2 \phi \, \overline{\mathrm{tr}}[\Sigma^2(\Sigma + \mu I)^{-3}] < 0$ and strictly increasing over $\lambda \geq 0$. Thus,

$$\begin{aligned}
&(1 - q_v(\Sigma, \Sigma))^2 \mathscr{R}'(\lambda, \phi) \\
&= (1 - q_v(\Sigma, \Sigma))^2 [\mathscr{B}'(\lambda, \phi) + \mathscr{V}'(\lambda, \phi)] \\
&= 2\lambda \, \overline{\mathrm{tr}}[\Sigma_\beta \Sigma(\Sigma + \mu I)^{-3}] \\
&\qquad + 2\mu^3 \phi\big(\overline{\mathrm{tr}}[\Sigma_\beta(\Sigma + \mu I)^{-2}] \cdot \overline{\mathrm{tr}}[\Sigma(\Sigma + \mu I)^{-3}] - \overline{\mathrm{tr}}[\Sigma_\beta(\Sigma + \mu I)^{-3}] \cdot \overline{\mathrm{tr}}[\Sigma(\Sigma + \mu I)^{-2}]\big) \\
&\qquad + 2\phi\sigma^2 \big(\mu \, \overline{\mathrm{tr}}[\Sigma(\Sigma + \mu I)^{-3}] - \overline{\mathrm{tr}}[\Sigma(\Sigma + \mu I)^{-2}]\big) \\
&= 2\lambda \, \overline{\mathrm{tr}}[\Sigma_\beta \Sigma(\Sigma + \mu I)^{-3}] \\
&\qquad + 2\phi\big(\mu^3 \, \overline{\mathrm{tr}}[\Sigma_\beta(\Sigma + \mu I)^{-2}] + \mu\sigma^2\big) \cdot \overline{\mathrm{tr}}[\Sigma(\Sigma + \mu I)^{-3}] - 2\phi\big(\mu^3 \, \overline{\mathrm{tr}}[\Sigma_\beta(\Sigma + \mu I)^{-3}] + \sigma^2\big) \cdot \overline{\mathrm{tr}}[\Sigma(\Sigma + \mu I)^{-2}].
\end{aligned}$$

Under the condition that

$$\frac{\overline{\mathrm{tr}}[\Sigma\{\mu^3(\Sigma + \mu I)^{-3}\}]}{\overline{\mathrm{tr}}[\Sigma\{\mu^2(\Sigma + \mu I)^{-2}\}]} > \frac{\overline{\mathrm{tr}}[\Sigma_\beta\{\mu^3(\Sigma + \mu I)^{-3}\}] + \sigma^2}{\overline{\mathrm{tr}}[\Sigma_\beta\{\mu^2(\Sigma + \mu I)^{-2}\}] + \sigma^2},$$

it follows that for all $\lambda \geq 0$,

$$(1 - q_v(\Sigma, \Sigma))^2 \mathscr{R}'(\lambda, \phi) > 0.$$

This implies that $\mathscr{R}(\lambda, \phi)$ is minimized at $\lambda < 0$, which finishes the proof. □

D.2. General Alignment Condition of Theorem 3.1 under Special Cases

Remark D.1 (Theorem 3.1 under isotropic features or signals)**.** When $\phi > 1$ and $\Sigma = I$, the condition above is never satisfied because

$$\frac{\overline{\operatorname{tr}}[\Sigma_\beta\{\mu^3(\Sigma+\mu I)^{-3}\}]+\sigma^2}{\overline{\operatorname{tr}}[\Sigma_\beta\{\mu^2(\Sigma+\mu I)^{-2}\}]+\sigma^2} = \frac{\left(\frac{\mu}{1+\mu}\right)^3\overline{\operatorname{tr}}[B]+\sigma^2}{\left(\frac{\mu}{1+\mu}\right)^2\overline{\operatorname{tr}}[B]+\sigma^2} > \frac{\left(\frac{\mu}{1+\mu}\right)^3}{\left(\frac{\mu}{1+\mu}\right)^2} = \frac{\overline{\operatorname{tr}}[\Sigma\{\mu^3(\Sigma+\mu I)^{-3}\}]}{\overline{\operatorname{tr}}[\Sigma\{\mu^2(\Sigma+\mu I)^{-2}\}]}.$$

Similarly, when β is isotropic, the condition above is never satisfied because

$$\frac{\overline{\operatorname{tr}}[\Sigma_\beta\{\mu^3(\Sigma+\mu I)^{-3}\}]+\sigma^2}{\overline{\operatorname{tr}}[\Sigma_\beta\{\mu^2(\Sigma+\mu I)^{-2}\}]+\sigma^2} = \frac{\overline{\operatorname{tr}}[\Sigma\{\mu^3(\Sigma+\mu I)^{-3}\}]+\sigma^2}{\overline{\operatorname{tr}}[\Sigma\{\mu^2(\Sigma+\mu I)^{-2}\}]+\sigma^2} > \frac{\overline{\operatorname{tr}}[\Sigma\{\mu^3(\Sigma+\mu I)^{-3}\}]}{\overline{\operatorname{tr}}[\Sigma\{\mu^2(\Sigma+\mu I)^{-2}\}]}$$

since $\overline{\operatorname{tr}}[\Sigma\{\mu^3(\Sigma+\mu I)^{-3}\}] < \overline{\operatorname{tr}}[\Sigma\{\mu^2(\Sigma+\mu I)^{-2}\}]$. To see this, note that $\mu > 0$ when $\phi > 1$, from Lemma F.2. Thus,

$$\begin{aligned}\overline{\operatorname{tr}}[\Sigma\{\mu^3(\Sigma+\mu I)^{-3}\}] &= \overline{\operatorname{tr}}[\Sigma\{\mu^2(\Sigma+\mu I)^{-2}\}\{\mu(\Sigma+\mu I)^{-1}\}] \\ &\leq \overline{\operatorname{tr}}[\Sigma\{\mu^2(\Sigma+\mu I)^{-2}\}]\cdot\|\mu(\Sigma+\mu I)^{-1}\|_{\mathrm{op}} \\ &\leq \overline{\operatorname{tr}}[\Sigma\{\mu^2(\Sigma+\mu I)^{-2}\}].\end{aligned}$$

Proposition D.2 (Theorem 3.1 under strict alignment conditions)**.** *Assuming random signals, zero noise under the strict alignment conditions of Wu & Xu (2020), the general alignment condition* (8) *is satisfied.*

Proof. Let h be a random variable following the empirical measure of the eigenvalues of $\Sigma = U\Lambda U$ (i.e., H_p) and g be a random variable following the empirical measure $\operatorname{diag}(U\mathbb{E}[\beta^\top\beta]U)$. When the joint distribution of (h, g) exists, there exists a function f such that g has mass $r \mapsto f(r)\,\mathrm{d}H_p(r)$. The strict alignment condition from Wu & Xu (2020) imposes f to be either strictly increasing or decreasing. This also implies that

$$\overline{\operatorname{tr}}[\Sigma_\beta] = \overline{\operatorname{tr}}[\Sigma B] \geq \overline{\operatorname{tr}}[\Sigma]\,\overline{\operatorname{tr}}[B]. \tag{32}$$

This holds because of the following Chebyshev's "other" inequality (Fink & Jodeit Jr., 1984).[3]

Fact D.3 (Positive/negative correlations of monotone functions; see, e.g., Appendix 9.9 of Ross (2022))**.** Let f and g be two real-valued functions of the same monotonicity. Let H be a probability measure on the real line. Then the following inequality holds:

$$\int f(r)g(r)\,\mathrm{d}H(r) \geq \int f(r)\,\mathrm{d}H(r)\cdot\int g(r)\,\mathrm{d}H(r). \tag{33}$$

On the other hand, if f is non-decreasing and g is non-increasing, then the inequality in (33) is reversed.

Now we will verify that this condition indeed implies (8) when $\sigma^2 = 0$. Define $f(r) = \sum_{i=1}^p(\beta^\top w_i)^2\,\mathbb{1}\{r = r_i\}$. $H_p(r) = p^{-1}\sum_{i=1}^p\mathbb{1}\{r = r_i\}$ and the transformed measure $\widetilde{H}_p(r) = r(r+\mu)^{-2}p^{-1}\sum_{i=1}^p\mathbb{1}\{r=r_i\}/\int r(r+\mu)^{-2}\,\mathrm{d}H_p(r)$.

Because $f(r)$ is increasing and $1/(r+\mu)$ is decreasing in r, it follows that

$$\int f(r)\frac{r}{r+\mu}\,\mathrm{d}\widetilde{H}_p(r) \leq \int f(r)\,\mathrm{d}\widetilde{H}_p(r)\int\frac{r}{r+\mu}\,\mathrm{d}\widetilde{H}_p(r).$$

Transforming back to $H_p(r)$ yields that

$$\int f(r)\frac{r}{(r+\mu)^3}\,\mathrm{d}H_p(r)\cdot\int\frac{r}{(r+\mu)^2}\,\mathrm{d}H_p(r) \leq \int f(r)\frac{r}{(r+\mu)^2}\,\mathrm{d}H_p(r)\cdot\int\frac{r}{(r+\mu)^3}\,\mathrm{d}H_p(r).$$

Equivalently,

$$\frac{\overline{\operatorname{tr}}[\Sigma_\beta\{\mu^2(\Sigma+\mu I)^{-2}\}]}{\overline{\operatorname{tr}}[\Sigma_\beta\{\mu^3(\Sigma+\mu I)^{-3}\}]} \geq \frac{\overline{\operatorname{tr}}[\Sigma\{\mu^2(\Sigma+\mu I)^{-2}\}]}{\overline{\operatorname{tr}}[\Sigma\{\mu^3(\Sigma+\mu I)^{-3}\}]},$$

with equality holds if and only if $f(r)$ is H_p-almost surely constant. This finishes the proof. ∎

[3]This is the second (less well-known) inequality, and not the first (more well-known) tail inequality that may come to the reader's mind!

D.3. Proof of Proposition 3.2

Proposition 3.2 (Optimal regularization under covariate shift and random signal)**.** *When $\Sigma_0 \neq \Sigma$ and $\beta_0 = \beta$, assuming isotropic signals $\mathbb{E}[\beta\beta^\top] = (\alpha^2/p)I$, we have $\lambda^*(\phi) = \phi/\mathrm{SNR} = \mathrm{argmin}_\lambda \, \mathbb{E}_\beta[\mathscr{R}(\lambda, \phi)]$. Furthermore, for $\lambda_{\min} < 0$ such that $X^\top X/n \succ \lambda_{\min} I$, even the non-asymptotic OOD risk (3) is minimized at $\lambda_p^* = \phi_p/\mathrm{SNR}$, where $\phi_p = p/n$.*

Proof. With slight abuse of notations, we use $\mathscr{R}$, $\mathscr{B}$, $\mathscr{V}$, and $\mathscr{S}$ to denote $\mathbb{E}_\beta[\mathscr{R}]$, $\mathbb{E}_\beta[\mathscr{B}]$, $\mathbb{E}_\beta[\mathscr{V}]$, and $\mathbb{E}_\beta[\mathscr{S}]$ as well. We split the proof into two parts.

Part (1) Asymptotic risk. When $\beta_0 = \beta$, the extra bias term is zero, that is, $\mathscr{S}(\lambda, \phi) = 0$. So, only the bias and variance components contribute to the risk.

We begin by deriving the formula for the derivative of the bias. For isotropic signals, from Lemma D.4 and (27), we have

$$
\begin{aligned}
&(1 - q_v(\Sigma, \Sigma))^2 \mathscr{B}'(\lambda, \phi) \\
&= \alpha^2 [q_v'(\Sigma_0, \Sigma) q_b(\Sigma, I)(1 - q_v(\Sigma, \Sigma)) \\
&\quad + q_v(\Sigma_0, \Sigma) q_b'(\Sigma, I)(1 - q_v(\Sigma, \Sigma)) + q_v(\Sigma_0, \Sigma) q_b(\Sigma, I) q_v'(\Sigma, \Sigma) \\
&\quad + q_b'(\Sigma_0, I)(1 - q_v(\Sigma, \Sigma))^2] \\
&= \alpha^2 \Bigg[q_v'(\Sigma_0, \Sigma) q_b(\Sigma, I) \left(\frac{\lambda}{\mu} + \mu q_v(\Sigma, I) \right) \\
&\quad + q_v(\Sigma_0, \Sigma) q_b'(\Sigma, I) \left(\frac{\lambda}{\mu} + \mu q_v(\Sigma, I) \right) + q_v(\Sigma_0, \Sigma) q_b(\Sigma, I) q_v'(\Sigma, \Sigma) \\
&\quad + q_b'(\Sigma_0, I) \left(\frac{\lambda}{\mu} + \mu q_v(\Sigma, I) \right)^2 \Bigg] \\
&= \frac{\alpha^2 \lambda}{\mu} \left(q_v'(\Sigma_0, \Sigma) q_b(\Sigma, I) + q_v(\Sigma_0, \Sigma) q_b'(\Sigma, I) + \frac{\lambda}{\mu} q_b'(\Sigma_0, I) + 2\mu q_b'(\Sigma_0, I) q_v(\Sigma, I) \right) \\
&\quad + \alpha^2 \Big[\mu q_v'(\Sigma_0, \Sigma) q_b(\Sigma, I) q_v(\Sigma, I) \\
&\quad + \mu q_v(\Sigma_0, \Sigma) q_b'(\Sigma, I) q_v(\Sigma, I) + q_v(\Sigma_0, \Sigma) q_b(\Sigma, I) q_v'(\Sigma, \Sigma) \\
&\quad + \mu^2 q_b'(\Sigma_0, I) q_v(\Sigma, I)^2 \Big].
\end{aligned}
$$

Notice that

$$
\begin{aligned}
q_b(\Sigma_0, I) &= \mu^2 \, \overline{\mathrm{tr}}[\Sigma_0 (\Sigma + \mu I)^{-2}] \\
&= \frac{\mu^2}{\phi} q_v(\Sigma_0, I) \\
q_b'(\Sigma_0, I) &= 2\mu \, \overline{\mathrm{tr}}[\Sigma_0 (\Sigma + \mu I)^{-2}] - 2\mu^2 \, \overline{\mathrm{tr}}[\Sigma_0 (\Sigma + \mu I)^{-3}] \\
&= 2\mu \, \overline{\mathrm{tr}}[\Sigma_0 \Sigma (\Sigma + \mu I)^{-3}] \\
&= -\frac{\mu}{\phi} q_v'(\Sigma_0, \Sigma),
\end{aligned}
$$

thus, it follows that

$$
\begin{aligned}
q_v'(\Sigma_0, \Sigma) q_b(\Sigma, I) + \mu q_b'(\Sigma_0, I) q_v(\Sigma, I) &= q_v'(\Sigma_0, \Sigma) \left[q_b(\Sigma, I) - \frac{\mu^2}{\phi} q_v(\Sigma, I) \right] = 0 \\
\mu q_b'(\Sigma, I) q_v(\Sigma, I) + q_b(\Sigma, I) q_v'(\Sigma, \Sigma) &= \frac{\mu^2}{\phi} \left[-q_v'(\Sigma, \Sigma) q_v(\Sigma, I) + q_v'(\Sigma, \Sigma) q_v(\Sigma, I) \right] = 0.
\end{aligned}
$$

Therefore, we have

$$
(1 - q_v(\Sigma, \Sigma))^2 \mathscr{B}'(\lambda, \phi)
$$

$$= \frac{\alpha^2\lambda}{\mu}\left(q_v(\Sigma_0,\Sigma)q_b'(\Sigma,I) + \frac{\lambda}{\mu}q_b'(\Sigma_0,I) + \mu q_b'(\Sigma_0,I)q_v(\Sigma,I)\right)$$
$$= -\frac{\alpha^2\lambda}{\phi}q_v(\Sigma_0,\Sigma)q_v'(\Sigma,\Sigma) - \frac{\alpha^2\lambda^2}{\mu\phi}q_v'(\Sigma_0,\Sigma) - \frac{\alpha^2\lambda\mu}{\phi}q_v'(\Sigma_0,\Sigma)q_v(\Sigma,I)$$
$$= -\frac{\alpha^2\lambda}{\phi}[q_v(\Sigma_0,\Sigma)q_v'(\Sigma,\Sigma) + \mu q_v'(\Sigma_0,\Sigma)q_v(\Sigma,I)] - \frac{\alpha^2\lambda^2}{\mu\phi}q_v'(\Sigma_0,\Sigma).$$

From (27), we further have that

$$\mathscr{B}'(\lambda,\phi) = -\frac{\alpha^2\lambda}{\phi(1-q_v(\Sigma,\Sigma))^2}[q_v(\Sigma_0,\Sigma)q_v'(\Sigma,\Sigma) - q_v'(\Sigma_0,\Sigma)q_v(\Sigma,\Sigma) + q_v'(\Sigma_0,\Sigma)]$$
$$= -\frac{\alpha^2\lambda}{\phi}\frac{\partial}{\partial\mu}\frac{q_v(\Sigma_0,\Sigma)}{1-q_v(\Sigma,\Sigma)}.$$

From Lemma D.4, we know that

$$\frac{\partial}{\partial\mu}\frac{q_v(\Sigma_0,\Sigma)}{1-q_v(\Sigma,\Sigma)} < 0$$

for all $\lambda \in (-\lambda_{\min}(\phi), +\infty)$. Furthermore, we have

$$\mathscr{R}'(\lambda,\phi) = \mathscr{B}'(\lambda,\phi) + \mathscr{V}'(\lambda,\phi)$$
$$= -\frac{\alpha^2\lambda}{\phi}\frac{\partial}{\partial\mu}\frac{q_v(\Sigma_0,\Sigma)}{1-q_v(\Sigma,\Sigma)} + \sigma^2\frac{\partial}{\partial\mu}\frac{q_v(\Sigma_0,\Sigma)}{1-q_v(\Sigma,\Sigma)}.$$

Setting the above to zero gives $\lambda^* = \phi\sigma^2/\alpha^2$. Also note that $\mathscr{R}'(\lambda,\phi)$ is negative when $\lambda < \lambda^*$ and positive when $\lambda > \lambda^*$, as well as $\partial\mu/\partial\lambda > 0$ from Lemma F.2 for all $\lambda \geq \lambda_{\min}(\phi)$. Thus, λ^* gives the optimal risk.

Part (2) Finite-sample risk. Recall the finite-sample risk is given by (3). Under Assumption 2.2 and isotropic signals, it follows that

$$R(\widehat{\beta}^\lambda) = \lambda^2\alpha^2\,\overline{\mathrm{tr}}[(\widehat{\Sigma}+\lambda I)^{-2}\Sigma_0] + \sigma^2\phi\,\overline{\mathrm{tr}}[(\widehat{\Sigma}+\lambda I)^{-2}\widehat{\Sigma}\Sigma_0].$$

Taking the derivative with respect to λ yields that

$$\frac{\partial R(\widehat{\beta}^\lambda)}{\partial\lambda} = 2\lambda\alpha^2\,\overline{\mathrm{tr}}[(\widehat{\Sigma}+\lambda I)^{-2}\Sigma_0] - 2\lambda^2\alpha^2\,\overline{\mathrm{tr}}[(\widehat{\Sigma}+\lambda I)^{-3}\Sigma_0] - 2\sigma^2\phi\,\overline{\mathrm{tr}}[(\widehat{\Sigma}+\lambda I)^{-3}\widehat{\Sigma}\Sigma_0]$$
$$= 2\lambda\alpha^2\,\overline{\mathrm{tr}}[(\widehat{\Sigma}+\lambda I)^{-3}\widehat{\Sigma}\Sigma_0] - 2\sigma^2\phi\,\overline{\mathrm{tr}}[(\widehat{\Sigma}+\lambda I)^{-3}\widehat{\Sigma}\Sigma_0]$$
$$= 2(\lambda\alpha^2 - \sigma^2\phi)\,\overline{\mathrm{tr}}[(\widehat{\Sigma}+\lambda I)^{-2}\widehat{\Sigma}\Sigma_0].$$

Setting the above to zero gives $\lambda_p^* = \phi_n/\mathrm{SNR}$. □

D.4. Proof of Theorem 3.3

Theorem 3.3 (Optimal regularization under covariate shift and deterministic signal)**.** *Assume the setting of Proposition 2.4 with* $\Sigma_0 \neq \Sigma$, $\beta_0 = \beta$.

1. (Underparameterized) *When* $\phi < 1$*, we have* $\lambda^* \geq 0$.
2. (Overparameterized) *When* $\phi > 1$*, if* $\Sigma_0 = I$ *(corresponding to the estimation risk), then we have* $\lambda^* \geq 0$.
3. (Overparameterized) *When* $\phi > 1$*, if* $\Sigma = I$ *and*

$$\overline{\mathrm{tr}}[\Sigma_0 B] > \overline{\mathrm{tr}}[\Sigma_0]\left(\overline{\mathrm{tr}}[B] + \frac{(1+\mu(0,\phi))^3}{\mu(0,\phi)^3}\sigma^2\right), \tag{10}$$

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

Proof. When $\beta_0 = \beta$, the extra bias term is zero, that is, $\mathscr{S}(\lambda,\phi) = 0$. So, only bias and variance contribute to the risk. We next split the proof into two cases.

Part (1) Underparameterized regime. When $\phi < 1$, from Lemma F.3 we have $\mu(0, \phi) = 0$ and the bias defined in Proposition 2.4 becomes zero, i.e., $\mathscr{B}(0, \phi) = 0$. From the fixed-point equation (53), we also have $\lim_{\lambda \to 0^+} \lambda/\mu(\lambda, \phi) = 0$. From Lemma D.4 and (27), we have

$$
\begin{aligned}
&(1 - q_v(\Sigma, \Sigma))^2 \mathscr{B}'(\lambda, \phi) \\
&= q_v'(\Sigma_0, \Sigma) q_b(\Sigma, B)(1 - q_v(\Sigma, \Sigma)) \\
&\quad + q_v(\Sigma_0, \Sigma) q_b'(\Sigma, B)(1 - q_v(\Sigma, \Sigma)) + q_v(\Sigma_0, \Sigma) q_b(\Sigma, B) q_v'(\Sigma, \Sigma) \\
&\quad + q_b'(\Sigma_0, B)(1 - q_v(\Sigma, \Sigma))^2 \\
&= q_v'(\Sigma_0, \Sigma) q_b(\Sigma, B) \left(\frac{\lambda}{\mu} + \mu q_v(\Sigma, I) \right) \\
&\quad + q_v(\Sigma_0, \Sigma) q_b'(\Sigma, B) \left(\frac{\lambda}{\mu} + \mu q_v(\Sigma, I) \right) + q_v(\Sigma_0, \Sigma) q_b(\Sigma, B) q_v'(\Sigma, \Sigma) \\
&\quad + q_b'(\Sigma_0, B) \left(\frac{\lambda}{\mu} + \mu q_v(\Sigma, I) \right)^2 \\
&= \frac{\lambda}{\mu} \left(q_v'(\Sigma_0, \Sigma) q_b(\Sigma, B) + q_v(\Sigma_0, \Sigma) q_b'(\Sigma, B) + \frac{\lambda}{\mu} q_b'(\Sigma_0, B) + 2\mu q_b'(\Sigma_0, B) q_v(\Sigma, I) \right) \\
&\quad + \Big[\mu q_v'(\Sigma_0, \Sigma) q_b(\Sigma, B) q_v(\Sigma, I) \\
&\quad + \mu q_v(\Sigma_0, \Sigma) q_b'(\Sigma, B) q_v(\Sigma, I) + q_v(\Sigma_0, \Sigma) q_b(\Sigma, B) q_v'(\Sigma, \Sigma) \\
&\quad + \mu^2 q_b'(\Sigma_0, B) q_v(\Sigma, I)^2 \Big].
\end{aligned}
$$

Thus, the derivative of the bias term becomes zero, i.e., $\mathscr{B}'(0, \phi) = 0$. Note that the variance term is strictly decreasing over $(\lambda_{\min}(\phi), +\infty)$ from Lemma D.4. Similarly to part (1) of the proof of Theorem 3.1, it follows that $\lambda^* \geq 0$.

Part (2) Overparameterized regime and $\Sigma = I$. When $\Sigma = I$, the above derivative in Part (1) becomes

$$
\begin{aligned}
&(1 - q_v(I, I))^2 \mathscr{B}'(\lambda, \phi) \\
&= \frac{\lambda}{\mu} \left(q_v'(\Sigma_0, I) q_b(I, B) + q_v(\Sigma_0, I) q_b'(I, B) + \frac{\lambda}{\mu} q_b'(\Sigma_0, B) + 2\mu q_b'(\Sigma_0, B) q_v(I, I) \right) \\
&\quad + \Big[\mu q_v'(\Sigma_0, I) q_b(I, B) q_v(I, I) \\
&\quad + \mu q_v(\Sigma_0, I) q_b'(I, B) q_v(I, I) + q_v(\Sigma_0, I) q_b(I, B) q_v'(I, I) \\
&\quad + \mu^2 q_b'(\Sigma_0, B) q_v(I, I)^2 \Big].
\end{aligned}
$$

Note from (21) and (22), we have that

$$
\begin{aligned}
q_b(\Sigma_0, B) &= \frac{\mu^2}{(1+\mu)^2} \,\overline{\mathrm{tr}}[\Sigma_0 B] \\
q_b'(\Sigma_0, B) &= \left(-\frac{2\mu^2}{(1+\mu)^3} + \frac{2\mu}{(1+\mu)^2} \right) \overline{\mathrm{tr}}[\Sigma_0 B] = \frac{2\mu}{(1+\mu)^3} \,\overline{\mathrm{tr}}[\Sigma_0 B] = \frac{2}{\mu(1+\mu)} q_b(\Sigma_0, B), \\
q_v(I, I) &= \frac{\phi}{(1+\mu)^2} \\
q_v'(I, I) &= -\frac{2\phi}{(1+\mu)^3} = -\frac{2}{1+\mu} q_v(I, I) \\
q_v(\Sigma_0, I) &= \frac{\phi}{(1+\mu)^2} \,\overline{\mathrm{tr}}[\Sigma_0] \\
q_v'(\Sigma_0, I) &= -\frac{2\phi}{(1+\mu)^3} \,\overline{\mathrm{tr}}[\Sigma_0] = -\frac{2}{1+\mu} q_v(\Sigma_0, I).
\end{aligned}
$$

We further have

$$
(1 - q_v(I, I))^2 \mathscr{B}'(\lambda, \phi)
$$

$$
\begin{aligned}
&= \frac{\lambda}{\mu}\left(-\frac{2}{1+\mu}q_v(\Sigma_0,I)q_b(I,B) + \frac{2}{\mu(1+\mu)}q_v(\Sigma_0,I)q_b(I,B) + \frac{2\lambda}{\mu^2(1+\mu)}q_b(\Sigma_0,B) + \frac{4}{1+\mu}q_b(\Sigma_0,B)q_v(I,I)\right) \\
&\quad + [-\frac{2\mu}{1+\mu}q_v(\Sigma_0,I)q_b(I,B)q_v(I,I) \\
&\quad + \frac{2}{1+\mu}q_v(\Sigma_0,I)q_b(I,B)q_v(I,I) - \frac{2}{1+\mu}q_v(\Sigma_0,I)q_b(I,B)q_v(I,I) \\
&\quad + \frac{2\mu}{1+\mu}q_b(\Sigma_0,B)q_v(I,I)^2] \\
&= \frac{\lambda}{\mu}\left(\frac{2\phi\mu(1-\mu)}{(1+\mu)^5}\,\overline{\mathrm{tr}}[\Sigma_0]\,\overline{\mathrm{tr}}[B] + \frac{2\lambda}{(1+\mu)^3}\,\overline{\mathrm{tr}}[\Sigma_0 B] + \frac{4\phi\mu^2}{(1+\mu)^5}\,\overline{\mathrm{tr}}[\Sigma_0 B]\right) \\
&\quad - \frac{2\mu}{1+\mu}q_v(\Sigma_0,I)q_b(I,B)q_v(I,I) \\
&\quad + \frac{2\mu}{1+\mu}q_b(\Sigma_0,B)q_v(I,I)^2 \\
&= \frac{\lambda}{\mu}\left(\frac{2\phi\mu(1-\mu)}{(1+\mu)^5}\,\overline{\mathrm{tr}}[\Sigma_0]\,\overline{\mathrm{tr}}[B] + \frac{2\lambda}{(1+\mu)^3}\,\overline{\mathrm{tr}}[\Sigma_0 B] + \frac{4\phi\mu^2}{(1+\mu)^5}\,\overline{\mathrm{tr}}[\Sigma_0 B]\right) \\
&\quad + \frac{2\phi^2\mu^3}{(1+\mu)^7}(\overline{\mathrm{tr}}[\Sigma_0 B] - \overline{\mathrm{tr}}[\Sigma_0]\,\overline{\mathrm{tr}}[B]) \\
&= \frac{\lambda}{\mu}\left(\frac{2\phi\mu(1-\mu)}{(1+\mu)^5}\,\overline{\mathrm{tr}}[\Sigma_0]\,\overline{\mathrm{tr}}[B] + \frac{2[\mu(1+\mu)^2 - \phi\mu(1-\mu)]}{(1+\mu)^5}\,\overline{\mathrm{tr}}[\Sigma_0 B]\right) \\
&\quad + \frac{2\phi^2\mu^3}{(1+\mu)^7}(\overline{\mathrm{tr}}[\Sigma_0 B] - \overline{\mathrm{tr}}[\Sigma_0]\,\overline{\mathrm{tr}}[B]) \\
&= \frac{\lambda}{\mu}\left(\frac{2\phi\mu(1-\mu)}{(1+\mu)^5}(\overline{\mathrm{tr}}[\Sigma_0]\,\overline{\mathrm{tr}}[B] - \overline{\mathrm{tr}}[\Sigma_0 B]) + \frac{2\mu}{(1+\mu)^3}\,\overline{\mathrm{tr}}[\Sigma_0 B]\right) \\
&\quad + \frac{2\phi^2\mu^3}{(1+\mu)^7}(\overline{\mathrm{tr}}[\Sigma_0 B] - \overline{\mathrm{tr}}[\Sigma_0]\,\overline{\mathrm{tr}}[B]).
\end{aligned}
$$

For the variance term, from Lemma D.4, we have

$$
\begin{aligned}
(1-q_v(I,I))^2\mathscr{V}'(\lambda,\phi) &= \sigma^2[q_v'(\Sigma_0,I)(1-q_v(I,I)) + q_v(\Sigma_0,I)q_v'(I,I)] \\
&= \sigma^2\left[q_v'(\Sigma_0,I)\left(\frac{\lambda}{\mu} + \mu q_v(I,I)\right) + q_v(\Sigma_0,I)q_v'(I,I)\right] \\
&= -\sigma^2\frac{2\lambda}{\mu(1+\mu)}q_v(\Sigma_0,I) - \sigma^2\frac{2}{1+\mu}\left[\mu q_v(\Sigma_0,I)q_v(I,I) + q_v(\Sigma_0,I)q_v(I,I)\right] \\
&= -\sigma^2\frac{2\lambda\phi}{\mu(1+\mu)^3}\,\overline{\mathrm{tr}}[\Sigma_0] - \sigma^2\frac{2\phi^2}{(1+\mu)^4}\,\overline{\mathrm{tr}}[\Sigma_0].
\end{aligned}
$$

From the fixed-point equation (53) we have $\mu(1+\mu) - \phi\mu = \lambda(1+\mu)$ and thus,

$$
\begin{aligned}
(1-q_v(I,I))^2\mathscr{V}'(\lambda,\phi) &= -\sigma^2\frac{2\phi(\mu(1+\mu)-\phi\mu)}{\mu(1+\mu)^4}\,\overline{\mathrm{tr}}[\Sigma_0] - \sigma^2\frac{2\phi^2}{(1+\mu)^4}\,\overline{\mathrm{tr}}[\Sigma_0] \\
&= -\sigma^2\frac{2\phi}{(1+\mu)^3}\,\overline{\mathrm{tr}}[\Sigma_0],
\end{aligned}
$$

which is strictly increasing in $\lambda \geq 0$.

Then we have

$$
\begin{aligned}
(1-q_v(I,I))^2\mathscr{R}'(\lambda,\phi) &= (1-q_v(\Sigma,\Sigma))^2[\mathscr{B}'(\lambda,\phi) + \mathscr{V}'(\lambda,\phi)] \\
&\geq (1-q_v(\Sigma,\Sigma))^2|_{\lambda=0}[\mathscr{B}'(0;\phi) + \mathscr{V}'(0;\phi)] \\
&= 2\lambda\left(\frac{\phi(1-\mu)}{(1+\mu)^5}(\overline{\mathrm{tr}}[\Sigma_0]\,\overline{\mathrm{tr}}[B] - \overline{\mathrm{tr}}[\Sigma_0 B]) + \frac{1}{(1+\mu)^3}\,\overline{\mathrm{tr}}[\Sigma_0 B]\right)
\end{aligned}
$$

$$
\begin{aligned}
&\quad + \frac{2\phi^2\mu^3}{(1+\mu)^7}(\overline{\operatorname{tr}}[\Sigma_0 B] - \overline{\operatorname{tr}}[\Sigma_0]\,\overline{\operatorname{tr}}[B]) - \sigma^2 \frac{2\phi}{(1+\mu)^3}\,\overline{\operatorname{tr}}[\Sigma_0] \\
&= \frac{2\lambda\phi}{(1+\mu)^5}\left(\frac{(1+\mu)^2-\phi}{\phi}\,\overline{\operatorname{tr}}[\Sigma_0 B] + \overline{\operatorname{tr}}[\Sigma_0]\,\overline{\operatorname{tr}}[B]\right) \quad (34)\\
&\quad + \left(\frac{2\lambda\phi\mu}{(1+\mu)^5} - \frac{2\lambda\phi\mu^2}{(1+\mu)^6}\right)\left(\overline{\operatorname{tr}}[\Sigma_0 B] - \overline{\operatorname{tr}}[\Sigma_0]\,\overline{\operatorname{tr}}[B]\right) \quad (35)\\
&\quad + \frac{2\phi\mu^3}{(1+\mu)^6}\left(\overline{\operatorname{tr}}[\Sigma_0 B] - \overline{\operatorname{tr}}[\Sigma_0]\left(\overline{\operatorname{tr}}[B] + \frac{(1+\mu)^3}{\mu^3}\sigma^2\right)\right). \quad (36)
\end{aligned}
$$

Note that when $\phi > 1$, $\mu(\lambda, \phi) > 0$ from Lemma F.2. Under the alignment condition, we have

$$
\overline{\operatorname{tr}}[\Sigma_0 B] > \overline{\operatorname{tr}}[\Sigma_0]\left(\overline{\operatorname{tr}}[B] + \frac{(1+\mu(0,\phi))^3}{\mu(0,\phi)^3}\sigma^2\right) \geq \overline{\operatorname{tr}}[\Sigma_0]\,\overline{\operatorname{tr}}[B].
$$

Thus, for $\lambda \geq 0$, the second term (35) is non-negative, and the third term (36) is strictly positive. When $\lambda \geq 0$, from the fixed-point equation (53) we have $\phi = (1+\mu) - \lambda(1+\mu)/\mu \leq 1+\mu$ and $(1+\mu)^2 - \phi \geq \phi(1+\mu) - \phi = \phi\mu$. Then, we know that the first term (34) is non-negative. Therefore, it follows that for all $\lambda \geq 0$,

$$
(1 - q_v(\Sigma,\Sigma))^2 \mathscr{R}'(\lambda,\phi) > 0.
$$

This implies that $\mathscr{R}(\lambda,\phi)$ is minimized at $\lambda < 0$.

Part (3) Overparameterized regime and $\Sigma_0 = I$. When $\Sigma_0 = I$, the above derivative in Part (1) becomes

$$
\begin{aligned}
&(1 - q_v(\Sigma,\Sigma))^2 \mathscr{B}'(\lambda,\phi) \\
&= \frac{\lambda}{\mu}\left(q_v'(I,\Sigma)q_b(\Sigma,B) + q_v(I,\Sigma)q_b'(\Sigma,B) + \frac{\lambda}{\mu}q_b'(I,B) + 2\mu q_b'(I,B)q_v(\Sigma,I)\right) \\
&\quad + [\mu q_v'(I,\Sigma)q_b(\Sigma,B)q_v(\Sigma,I) + \mu q_v(I,\Sigma)q_b'(\Sigma,B)q_v(\Sigma,I) \quad (37)\\
&\qquad + q_v(I,\Sigma)q_b(\Sigma,B)q_v'(\Sigma,\Sigma) + \mu^2 q_b'(I,B)q_v(\Sigma,I)^2] \quad (38)
\end{aligned}
$$

Recalling the definitions of $q_b(\cdot,\cdot)$ and $q_v(\cdot,\cdot)$ from (21) and (22), observe that

$$
\begin{aligned}
q_b(I,B) &= \mu^2\,\overline{\operatorname{tr}}[(\Sigma+\mu I)^{-2}B] \\
q_b'(I,B) &= 2\mu\,\overline{\operatorname{tr}}[(\Sigma+\mu I)^{-2}B] - 2\mu^2\,\overline{\operatorname{tr}}[(\Sigma+\mu I)^{-3}B] = \frac{2}{\mu}q_b(I,B) - 2\mu^2\,\overline{\operatorname{tr}}[(\Sigma+\mu I)^{-3}B] \\
&= 2\mu\,\overline{\operatorname{tr}}[(\Sigma+\mu I)^{-3}\Sigma B], \\
q_v(I,\Sigma) &= \phi\,\overline{\operatorname{tr}}[(\Sigma+\mu I)^{-2}\Sigma] \\
q_v'(I,\Sigma) &= -2\phi\,\overline{\operatorname{tr}}[(\Sigma+\mu I)^{-3}\Sigma].
\end{aligned}
$$

Next, we work on terms (37) and (38) that do not involve a factor of λ/μ.

$$
\begin{aligned}
&\mu q_v'(I,\Sigma)q_b(\Sigma,B)q_v(\Sigma,I) \\
&\quad + \mu q_v(I,\Sigma)q_b'(\Sigma,B)q_v(\Sigma,I) + q_v(I,\Sigma)q_b(\Sigma,B)q_v'(\Sigma,\Sigma) \\
&\quad + \mu^2 q_b'(I,B)q_v(\Sigma,I)^2 \\
&= -2\mu\phi\,\overline{\operatorname{tr}}[(\Sigma+\mu I)^{-3}\Sigma]q_b(\Sigma,B)q_v(\Sigma,I) \\
&\quad + \mu\left(\frac{2}{\mu}q_b(\Sigma,B) - 2\mu^2\,\overline{\operatorname{tr}}[(\Sigma+\mu I)^{-3}\Sigma B]\right)q_v(I,\Sigma)q_v(\Sigma,I) \\
&\quad - 2\phi\,\overline{\operatorname{tr}}[(\Sigma+\mu I)^{-3}\Sigma^2]q_v(I,\Sigma)q_b(\Sigma,B) \\
&\quad + \mu^2\left(\frac{2}{\mu}q_b(I,B) - 2\mu^2\,\overline{\operatorname{tr}}[(\Sigma+\mu I)^{-3}B]\right)q_v(\Sigma,I)^2
\end{aligned}
$$

$$
\begin{aligned}
&= 2\mu\phi\left(-\overline{\mathrm{tr}}[(\Sigma+\mu I)^{-3}\Sigma] + \frac{1}{\mu}\overline{\mathrm{tr}}[(\Sigma+\mu I)^{-2}\Sigma] - \overline{\mathrm{tr}}[(\Sigma+\mu I)^{-3}\Sigma^2]\right) q_b(\Sigma,B)q_v(\Sigma,\Sigma)\\
&\quad + 2\mu^2\left(\frac{1}{\mu}q_b(I,B) - \mu^2\,\overline{\mathrm{tr}}[(\Sigma+\mu I)^{-3}B] - \mu\,\overline{\mathrm{tr}}[(\Sigma+\mu I)^{-3}\Sigma B]\right) q_v(\Sigma,I)^2\\
&= 0.
\end{aligned}
$$

This implies that $\mathscr{B}'(0,\phi)=0$ and

$$
\begin{aligned}
&(1-q_v(\Sigma,\Sigma))^2\mathscr{B}'(\lambda,\phi)\\
&= \frac{\lambda}{\mu}\left(q_v'(I,\Sigma)q_b(\Sigma,B) + q_v(I,\Sigma)q_b'(\Sigma,B) + 2\mu q_b'(I,B)q_v(\Sigma,I)\right) + \frac{\lambda^2}{\mu^2}q_b'(I,B)\\
&= \frac{\lambda}{\mu}\left(q_v'(I,\Sigma)q_b(\Sigma,B) + q_v(I,\Sigma)q_b'(\Sigma,B) + 2\mu q_b'(I,B)q_v(\Sigma,I)\right) + \frac{\lambda^2}{\mu^2}q_b'(I,B)\\
&= \frac{\lambda}{\mu}\bigg(\frac{2}{\mu}\left(-\mu\phi\,\overline{\mathrm{tr}}[(\Sigma+\mu I)^{-3}\Sigma] + q_v(\Sigma,I)\right)q_b(\Sigma,B) - 2\mu^2\,\overline{\mathrm{tr}}[(\Sigma+\mu I)^{-3}\Sigma B]q_v(\Sigma,I)\\
&\qquad + 4\mu^2\,\overline{\mathrm{tr}}[(\Sigma+\mu I)^{-3}\Sigma B]q_v(\Sigma,I)\bigg) + \frac{\lambda^2}{\mu^2}q_b'(I,B)\\
&= \frac{\lambda}{\mu}\left(\frac{2}{\mu}\phi\,\overline{\mathrm{tr}}[(\Sigma+\mu I)^{-3}\Sigma^2]q_b(\Sigma,B) + 2\mu^2\,\overline{\mathrm{tr}}[(\Sigma+\mu I)^{-3}\Sigma B]q_v(\Sigma,I)\right) + \frac{\lambda^2}{\mu^2}q_b'(I,B)\\
&= 2\phi\lambda\left(\overline{\mathrm{tr}}[(\Sigma+\mu I)^{-3}\Sigma^2]\,\overline{\mathrm{tr}}[(\Sigma+\mu I)^{-2}\Sigma B] + \mu\,\overline{\mathrm{tr}}[(\Sigma+\mu I)^{-3}\Sigma B]\,\overline{\mathrm{tr}}[(\Sigma+\mu I)^{-2}\Sigma]\right)\\
&\qquad + 2\lambda\left(1-\phi\,\overline{\mathrm{tr}}[(\Sigma+\mu I)^{-1}\Sigma]\right)\overline{\mathrm{tr}}[(\Sigma+\mu I)^{-3}\Sigma B]\\
&= 2\phi\lambda\left(\overline{\mathrm{tr}}[(\Sigma+\mu I)^{-3}\Sigma^2]\,\overline{\mathrm{tr}}[(\Sigma+\mu I)^{-2}\Sigma B] - \overline{\mathrm{tr}}[(\Sigma+\mu I)^{-3}\Sigma B]\,\overline{\mathrm{tr}}[(\Sigma+\mu I)^{-2}\Sigma^2]\right)\\
&\qquad + 2\lambda\,\overline{\mathrm{tr}}[(\Sigma+\mu I)^{-3}\Sigma B]\\
&= 2\phi\lambda\,\overline{\mathrm{tr}}[(\Sigma+\mu I)^{-3}\Sigma^2]\,\overline{\mathrm{tr}}[(\Sigma+\mu I)^{-2}\Sigma B]\\
&\qquad + 2\lambda\,\overline{\mathrm{tr}}[(\Sigma+\mu I)^{-3}\Sigma B](1-\phi\,\overline{\mathrm{tr}}[(\Sigma+\mu I)^{-2}\Sigma^2]).
\end{aligned}
$$

From Lemma F.3 (3), we know that $1-\phi\,\overline{\mathrm{tr}}[(\Sigma+\mu I)^{-2}\Sigma^2]\geq 0$. When $\phi>1$, it then follows that $\mathscr{B}'$ is strictly negative for all $\lambda\in[\lambda_{\min}(\phi),0)$ and strictly positive for all $\lambda>0$ because $\mu(\lambda,\phi)>0$.

For the variance term, from Lemma D.4, we have

$$
\begin{aligned}
&(1-q_v(\Sigma,\Sigma))^2\mathscr{V}'(\lambda,\phi)\\
&= \sigma^2[q_v'(I,\Sigma)(1-q_v(\Sigma,\Sigma)) + q_v(I,\Sigma)q_v'(\Sigma,\Sigma)]\\
&= \sigma^2\left[q_v'(I,\Sigma)\left(\frac{\lambda}{\mu}+\mu q_v(\Sigma,I)\right) + q_v(I,\Sigma)q_v'(\Sigma,\Sigma)\right]\\
&= -2\phi\sigma^2\left[\overline{\mathrm{tr}}[(\Sigma+\mu I)^{-3}\Sigma]\left(\frac{\lambda}{\mu}+\mu q_v(\Sigma,I)\right) + \overline{\mathrm{tr}}[(\Sigma+\mu I)^{-3}\Sigma^2]q_v(I,\Sigma)\right]\\
&= -2\phi\sigma^2\left[\frac{\lambda}{\mu}\,\overline{\mathrm{tr}}[(\Sigma+\mu I)^{-3}\Sigma] + \overline{\mathrm{tr}}[(\Sigma+\mu I)^{-2}\Sigma]q_v(\Sigma,I)\right]\\
&= -2\phi\sigma^2\left[\left(1-\phi\,\overline{\mathrm{tr}}[(\Sigma+\mu I)^{-1}\Sigma]\right)\overline{\mathrm{tr}}[(\Sigma+\mu I)^{-3}\Sigma] + \overline{\mathrm{tr}}[(\Sigma+\mu I)^{-2}\Sigma]q_v(\Sigma,I)\right]\\
&= -2\phi\sigma^2\left[\overline{\mathrm{tr}}[(\Sigma+\mu I)^{-3}\Sigma] + \phi(\overline{\mathrm{tr}}[(\Sigma+\mu I)^{-2}\Sigma]^2 - \overline{\mathrm{tr}}[(\Sigma+\mu I)^{-1}\Sigma]\,\overline{\mathrm{tr}}[(\Sigma+\mu I)^{-3}\Sigma])\right],
\end{aligned}
$$

which is strictly negative for all $\lambda\geq\lambda_{\min}(\phi)$.

Combining the above two derivatives, we conclude that $\lambda^*>0$ in this case. □

D.5. Proof of Theorem 3.4

Theorem 3.4 (Optimal regularization under regression shift)**.** *Assume the setup of Proposition 2.4 with* $\Sigma_0=\Sigma$, $\beta_0\neq\beta$.

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

$$\overline{\operatorname{tr}}[B_0\Sigma^2(\Sigma + \mu I)^{-2}] > \overline{\operatorname{tr}}[B\Sigma^2(\Sigma + \mu I)^{-2}], \tag{11}$$

where $B = \beta\beta^\top$ and $B_0 = \beta_0\beta^\top$, then we have $\lambda^* < 0$.

2. (Overparameterized) *When $\phi > 1$, if the general alignment conditions* (8) *and* (11) *hold, then we have $\lambda^* < 0$.*

Proof. We split the proof into two parts.

Part (1) Underparameterized regime. From the proof of Theorem 3.1, we know that when $\Sigma_0 = \Sigma$, the bias term satisfies that

$$\min_{\lambda \in [\lambda_{\min}(\phi), 0]} \mathscr{B}(\lambda, \phi) \geq \mathscr{B}(0, \phi) = 0,$$

from (24). From Lemma D.4, the excess bias term has the following derivative:

$$\mathscr{S}'(\lambda, \phi) = -2\beta^\top\Sigma(\Sigma + \mu I)^{-2}\Sigma(\beta_0 - \beta) = -2\beta^\top\Sigma(\Sigma + \mu I)^{-2}\Sigma(\beta - \beta_0),$$

which is zero when $\lambda = 0$ because $\mu(0, \phi) = 0$ when $\phi < 1$, from Lemma F.2. Under the condition that $\beta^\top\Sigma(\Sigma + \mu I)^{-2}\Sigma(\beta - \beta_0) < 0$ for all $\mu \geq 0$, we have that $\mathscr{S}(\lambda, \phi)$ is strictly increasing in $\lambda \geq 0$.

From (31), we have

$$\mathscr{V}'(\lambda, \phi) = \frac{-2\sigma^2\phi\, \overline{\operatorname{tr}}[\Sigma^2(\Sigma + \mu I)^{-3}]}{(1 - q_v(\Sigma, \Sigma))^2}.$$

Also, since $\mathscr{B} \geq 0$ with equality holds if $\lambda = 0$. Then we have, if

$$\mathscr{S}'(\lambda, \phi) + \mathscr{V}'(\lambda, \phi) = -2\beta^\top\Sigma(\Sigma + \mu I)^{-2}\Sigma(\beta - \beta_0) + \frac{-2\sigma^2\phi\, \overline{\operatorname{tr}}[\Sigma^2(\Sigma + \mu I)^{-3}]}{(1 - q_v(\Sigma, \Sigma))^2} > 0 \tag{39}$$

for all $\mu \geq \mu(0, \phi)$, then $\lambda^* < 0$. Note that when $\lambda \to +\infty$, $\mu \to +\infty$ and the denominator of the second term tends to one, so we have $\mathscr{V}'(\lambda, \phi) \asymp \mu^{-3}$. On the other hand, the first term scales as $\mathscr{S}'(\lambda, \phi) \asymp \mu^{-2}$. Eventually, the first term dominates. Thus, the condition (39) could hold when $\mathscr{S}'(\lambda, \phi)$ is positive and large enough.

Especially, when $\sigma^2 = o(1)$ and the assumed alignment condition is met, it follows that $\mathscr{R}(\lambda, \phi) = \mathscr{B}(\lambda, \phi) + \mathscr{V}(\lambda, \phi) + \mathscr{S}(\lambda, \phi)$ is minimized over $\lambda < 0$.

Part (2) Overparameterized regime. From the proof of Theorem 3.1, we know that when (8), $\mathscr{B}(\lambda, \phi) + \mathscr{V}(\lambda, \phi)$ is minimized at $\lambda < 0$. Under the condition that $\beta^\top\Sigma(\Sigma + \mu I)^{-2}\Sigma(\beta - \beta_0) \leq 0$ for all $\mu > 0$, we have $\mathscr{S}'(\lambda, \phi) \geq 0$ over $\lambda \in (\lambda_{\min}(\phi), +\infty)$. This implies that $\mathscr{S}(\lambda, \phi)$ is increasing over $\lambda \in (\lambda_{\min}(\phi), +\infty)$. Combining the two results, we further see that the risk $\mathscr{R}(\lambda, \phi)$ is minimized at $\lambda < 0$. □

D.6. Helper Lemmas

Lemma D.4 (Out-of-distribution risk derivatives)**.** *Under the same conditions as in Proposition 2.4, we have*

$$\frac{\partial \mathscr{R}(\lambda, \phi)}{\partial \lambda} = (\mathscr{B}'(\lambda, \phi) + \mathscr{V}'(\lambda, \phi) + \mathscr{S}'(\lambda, \phi))\, \frac{\partial \mu}{\partial \lambda}, \tag{40}$$

where

$$\begin{aligned}
\mathscr{B}'(\lambda, \phi) := \frac{\partial \mathscr{B}(\lambda, \phi)}{\partial \mu} &= \frac{1}{(1 - q_v(\Sigma, \Sigma))^2}\Big[q_v'(\Sigma_0, \Sigma)q_b(\Sigma, B)(1 - q_v(\Sigma, \Sigma)) \\
&\quad + q_v(\Sigma_0, \Sigma)q_b'(\Sigma, B)(1 - q_v(\Sigma, \Sigma)) + q_v(\Sigma_0, \Sigma)q_b(\Sigma, B)q_v'(\Sigma, \Sigma) \\
&\quad + q_b'(\Sigma_0, B)(1 - q_v(\Sigma, \Sigma))^2\Big] \\
\mathscr{V}'(\lambda, \phi) := \frac{\partial \mathscr{V}(\lambda, \phi)}{\partial \mu} &= \sigma^2 \frac{q_v'(\Sigma_0, \Sigma) - q_v'(\Sigma_0, \Sigma)q_v(\Sigma, \Sigma) + q_v'(\Sigma, \Sigma)q_v(\Sigma_0, \Sigma)}{(1 - q_v(\Sigma, \Sigma))^2}
\end{aligned}$$

$$\mathscr{S}'(\lambda,\phi) := \frac{\partial \mathscr{S}(\lambda,\phi)}{\partial\mu} = -\frac{2}{\mu^2} q_b(\Sigma_0, (B-B_0)\Sigma).$$

and $\mu = 1/v_p(\lambda,\phi)$*. Furthermore, we have* $\mathscr{V}'(\lambda,\phi) < 0$ *for all* $\lambda \in (\lambda_{\min}(\phi), +\infty)$*.*

Proof. We split the proof into different parts.

Part (1) Bias term. Recall the expression for bias for out-of-distribution squared risk:

$$\mathscr{B}(\lambda,\phi) = \frac{q_v(\Sigma_0,\Sigma)}{1-q_v(\Sigma,\Sigma)} q_b(\Sigma, B) + q_b(\Sigma_0, B)$$

Define

$$h_1(\mu) = q_v(\Sigma_0,\Sigma), \qquad h_2(\mu) = \frac{q_b(\Sigma,B)}{1-q_v(\Sigma,\Sigma)}, \qquad h_3(\mu) = q_b(\Sigma_0, B). \tag{41}$$

Then we have

$$\mathscr{B}(\lambda,\phi) = h_1(\mu)h_2(\mu) + h_3(\mu),$$

and

$$\begin{aligned}
\mathscr{B}'(\lambda,\phi) &= h_1'(\mu)h_2(\mu) + h_1(\mu)h_2'(\mu) + h_3'(\mu) \\
&= q_v'(\Sigma_0,\Sigma)\frac{q_b(\Sigma,B)}{1-q_v(\Sigma,\Sigma)} + q_v(\Sigma_0,\Sigma)\frac{q_b'(\Sigma,B)(1-q_v(\Sigma,\Sigma)) + q_b(\Sigma,B)q_v'(\Sigma,\Sigma)}{(1-q_v(\Sigma,\Sigma))^2} + q_b'(\Sigma_0,B) \\
&= \frac{1}{(1-q_v(\Sigma,\Sigma))^2}\Big[q_v'(\Sigma_0,\Sigma)q_b(\Sigma,B)(1-q_v(\Sigma,\Sigma)) \\
&\quad + q_v(\Sigma_0,\Sigma)q_b'(\Sigma,B)(1-q_v(\Sigma,\Sigma)) + q_v(\Sigma_0,\Sigma)q_b(\Sigma,B)q_v'(\Sigma,\Sigma) \\
&\quad + q_b'(\Sigma_0,B)(1-q_v(\Sigma,\Sigma))^2\Big].
\end{aligned}$$

Part (2) Variance term. Recall that the variance term is given by:

$$\mathscr{V}(\lambda,\phi) = \sigma^2 \frac{q_v(\Sigma_0,\Sigma)}{1-q_v(\Sigma,\Sigma)}.$$

The derivative in μ is:

$$\mathscr{V}' = \sigma^2 \frac{1}{(1-q_v(\Sigma,\Sigma))^2}\{q_v'(\Sigma_0,\Sigma)(1-q_v(\Sigma,\Sigma)) + q_v(\Sigma_0,\Sigma)q_v'(\Sigma,\Sigma)\}.$$

Note that

$$\begin{aligned}
q_v'(\Sigma,\Sigma) &= -2\phi\,\overline{\mathrm{tr}}[\Sigma^2(\Sigma+\mu I)^{-2}] < 0 \\
q_v'(\Sigma_0,\Sigma) &= -2\phi\,\overline{\mathrm{tr}}[\Sigma_0\Sigma(\Sigma+\mu I)^{-2}] = -2\phi\,\overline{\mathrm{tr}}[(\Sigma+\mu I)^{-1}\Sigma^{1/2}\Sigma_0\Sigma^{1/2}(\Sigma+\mu I)^{-1}] < 0.
\end{aligned}$$

From Lemma F.3, we also have $q_v(\Sigma,\Sigma) > 0$ and $1 - q_v(\Sigma,\Sigma) > 0$ when $\lambda \in (\lambda_{\min}(\phi), +\infty)$. Therefore, it holds that

$$\mathscr{V}'(\lambda,\phi) < 0.$$

Part (3) Extra bias term. Recall that the extra bias term is given by

$$\mathscr{S}(\lambda,\phi) = -2\beta^\top(v\Sigma + I)^{-1}\Sigma_0(\beta-\beta_0) = -2\mu\beta^\top(\Sigma+\mu I)^{-1}\Sigma_0(\beta-\beta_0) = l(\Sigma_0,(B-B_0)\Sigma).$$

Then, the derivative is given by:

$$\mathscr{S}'(\lambda,\phi) = -2\beta^\top\Sigma(\Sigma+\mu I)^{-2}\Sigma_0(\beta-\beta_0) = -\frac{2}{\mu^2}q_b(\Sigma_0,(B-B_0)\Sigma).$$

□

E. Proofs in Section 4

E.1. Proof of Proposition 4.1

Proposition 4.1 (Optimal risk under isotropic signals)**.** *When $\Sigma_0 \neq \Sigma$ and $\beta = \beta_0$, assuming isotropic signals $\mathbb{E}[\beta\beta^\top] = (\alpha^2/p)I$ the optimal risk obtained at $\lambda^*(\phi) = \phi/\textsc{snr}$ is given by:*

$$\mathbb{E}_\beta[\mathscr{R}(\lambda^*, \phi)] = \alpha^2 \mu^* \,\overline{\mathrm{tr}}[\Sigma_0(\Sigma + \mu^* I)^{-1}] + \sigma_0^2, \tag{12}$$

where $\mu^* = \mu(\lambda^*, \phi)$. Furthermore, the left side of (12) is strictly increasing in ϕ if $\textsc{snr} \in (0, \infty)$ and σ_0^2 are fixed and strictly increasing in $\textsc{snr}$ if ϕ, σ^2, and σ_0^2 are fixed.

Proof. When $\beta = \beta_0$ and $\beta\beta^\top \simeq \alpha^2 I$, from (27), we have that

$$\begin{aligned}
\mathscr{B}(\lambda, \phi) &= \alpha^2 \frac{q_v(\Sigma_0, \Sigma)}{1 - q_v(\Sigma, \Sigma)} q_b(\Sigma, I) + \alpha^2 q_b(\Sigma_0, I) \\
&= \alpha^2 \frac{\mu^2}{\phi} \left(\frac{q_v(\Sigma_0, \Sigma)}{1 - q_v(\Sigma, \Sigma)} q_v(\Sigma, I) + q_v(\Sigma_0, I) \right) \\
&= \alpha^2 \frac{\mu}{\phi} \left(\frac{q_v(\Sigma_0, \Sigma)}{1 - q_v(\Sigma, \Sigma)} \left(1 - q_v(\Sigma, \Sigma) - \frac{\lambda}{\mu} \right) + \mu q_v(\Sigma_0, I) \right) \\
&= \alpha^2 \frac{\mu}{\phi} \left[q_v(\Sigma_0, \Sigma) + \mu q_v(\Sigma_0, I) \right] - \alpha^2 \frac{\lambda}{\phi} \frac{q_v(\Sigma_0, \Sigma)}{1 - q_v(\Sigma, \Sigma)} \\
&= \alpha^2 \mu \,\overline{\mathrm{tr}}[\Sigma_0(\Sigma + \mu I)^{-1}] - \alpha^2 \frac{\lambda}{\phi} \frac{q_v(\Sigma_0, \Sigma)}{1 - q_v(\Sigma, \Sigma)}.
\end{aligned}$$

Therefore, from Proposition 2.4 and Theorem 3.4, the optimal risk is given by

$$\begin{aligned}
R(\widehat{\beta}^{\lambda^*}) &\simeq \mathscr{B}_p(\lambda^*, \phi) + \mathscr{V}_p(\lambda^*, \phi) + \sigma_0^2 \\
&= \alpha^2 \mu^* \,\overline{\mathrm{tr}}[\Sigma_0(\Sigma + \mu^* I)^{-1}] + \left(\sigma^2 - \frac{\alpha^2 \lambda^*}{\phi} \right) \frac{q_v(\Sigma_0, \Sigma)}{1 - q_v(\Sigma, \Sigma)} + \sigma_0^2 \\
&= \alpha^2 \mu^* \,\overline{\mathrm{tr}}[\Sigma_0(\Sigma + \mu^* I)^{-1}] + \sigma_0^2 \\
&= \alpha^2 \,\overline{\mathrm{tr}}[\Sigma_0(v^* \Sigma + I)^{-1}] + \sigma_0^2,
\end{aligned}$$

where $\mu^* = \mu(\lambda^*, \phi)$ and $v^* = v(\lambda^*, \phi)$.

Note that when σ^2 and σ_0^2 are fixed, $\mathscr{R}(\lambda^*, \phi) = \phi\sigma^2 \cdot (\mu^*/\lambda^*) \cdot \overline{\mathrm{tr}}[\Sigma_0(\Sigma + \mu^* I)^{-1}] + \sigma_0^2$ is simply a function of λ^*. From Lemma F.2 (4), we have that μ^*/λ^* is strictly decreasing in λ^*. Also note that $\overline{\mathrm{tr}}[\Sigma_0(\Sigma + \mu^* I)^{-1}]$ is strictly decreasing in λ^*. Thus, we know that $\mathscr{R}(\lambda^*, \phi)$ is strictly decreasing in λ^* when ϕ, σ, σ_0 are fixed. Because λ^* is strictly decreasing in $\textsc{snr}$, we further find that $\mathscr{R}(\lambda^*, \phi)$ is strictly increasing in $\textsc{snr}$. ☐

E.2. Proof of Theorem 4.2

Theorem 4.2 (Monotonicity of optimally tuned OOD risk)**.** *For $\lambda \geq \lambda_{\min}(\phi)$ where $\lambda_{\min}(\phi)$ is as in (5), for all $\epsilon > 0$ small enough, the risk of optimal ridge predictor satisfies:*

$$\min_{\lambda \geq \lambda_{\min}(\phi) + \epsilon} R(\widehat{\beta}^\lambda) \;\simeq\; \min_{\lambda \geq \lambda_{\min}(\phi)} \mathscr{R}(\lambda, \phi), \tag{13}$$

and right side of (13) is monotonically increasing in ϕ if $\textsc{snr}$ and σ_0^2 are fixed. In addition, when $\beta = \beta_0$ it is monotonically increasing in $\textsc{snr}$ if ϕ, σ^2, and σ_0^2 are fixed.

Proof. We split the proof into different parts.

Part (1) Risk characterization and equivalence. From the proof of Theorem 4.4, we have

$$R(\widehat{\beta}^\lambda) \simeq \mathscr{R}_p(\lambda, \phi, \phi)$$

where $\mathscr{R}_p$ is defined in (50). Furthermore, for any $\overline{\psi} \in [\phi, +\infty]$, there exists a unique $\bar{\lambda} \geq \lambda_{\min}(\phi)$ defined through (44). For any pair of (λ_1, ψ_1) and (λ_2, ψ_2) on the path $\mathcal{P}(\bar{\lambda}; \phi, \overline{\psi})$ as defined in (55), we have that:

$$\mathscr{R}_p(\lambda_1; \phi, \psi_1) = \mathscr{R}_p(\lambda_2; \phi, \psi_2).$$

Part (2) Risk monotonicity. From Lemma F.3 (3), we know that the denominator of $\widetilde{v}_p$ defined in (19) is non-negative:

$$v_p(\lambda, \psi)^{-2} - \phi \int r^2 (1 + v_p(\lambda, \psi) r)^{-2} \, \mathrm{d}H_p(r) \geq v_p(\lambda, \psi)^{-2} - \psi \int r^2 (1 + v_p(\lambda, \psi) r)^{-2} \, \mathrm{d}H_p(r) \geq 0$$

for $\lambda \geq \lambda_{\min}(\phi)$. Therefore, $\mathscr{R}_p(\lambda, \phi, \psi)$ is increasing in ϕ for any fixed (λ, ψ). Furthermore, since $\mathscr{R}_p(\lambda, \phi, \psi)$ is a continuous function of ϕ and $v(\lambda; \psi)$, it follows that for $0 < \phi_1 \leq \phi_2 < \infty$,

$$\min_{\psi \geq \phi_1} \mathscr{R}_p(\lambda, \phi_1, \psi) \leq \min_{\psi \geq \phi_2} \mathscr{R}_p(\lambda, \phi_1, \psi) \leq \min_{\psi \geq \phi_2} \mathscr{R}_p(\lambda, \phi_2, \psi),$$

where the first inequality follows because $\{\psi : \psi \geq \phi_1\} \supseteq \{\psi : \psi \geq \phi_2\}$, and the second inequality follows because $\mathscr{R}_p(\lambda, \phi, \psi)$ is increasing in ϕ for a fixed ψ. Thus, $\min_{\psi \geq \phi} \mathscr{R}_p(\lambda, \phi, \psi)$ is a continuous and monotonically increasing function in ϕ.

Part (3) Optimal subsampling and regularization. Similar to the proof of Part (3) in Lemma E.1, we have

$$\min_{\psi \geq \phi} \mathscr{R}_p(0; \phi, \psi) \simeq \min_{\lambda \geq \lambda_{\min}(\phi) + \epsilon} R(\widehat{\beta}^\lambda).$$

From Part (3), we know that the former is continuous and monotonically increasing in ϕ, which finishes the proof.

Part (4) Monotonicity in signal-to-noise ratio. When $\beta_0 = \beta$, the extra bias term $\mathscr{S}$ is zero. When σ^2 is fixed, note that $\mathscr{B}$ and κ are strictly increasing in α^2 while $\mathscr{V}$ does not depend on α^2. Thus, we know that $\mathscr{R}_p$ is strictly increasing in SNR. Consequently, $\min_{\psi \geq \phi} \mathscr{R}_p(0; \phi, \psi)$ is strictly increasing in SNR. □

E.3. Proof of Theorem 4.3

Theorem 4.3 (Non-monotonicity of suboptimally tuned risk)**.** *When* $(\Sigma_0, \beta_0) = (\Sigma, \beta)$ *and* $\Sigma = I$*, the risk component equivalents defined in* (6) *have the following properties:*

1. (Bias component) *For all* $\lambda > 0$*,* $\mathscr{B}(\lambda, \phi)$ *is strictly increasing over* $\phi \in (0, \lambda + 1)$ *and strictly decreasing over* $\phi \in (\lambda + 1, \infty)$.
2. (Variance component) *For all* $\lambda > 0$*,* $\mathscr{V}(\lambda, \phi)$ *is strictly increasing over* $\phi \in (0, \infty)$.
3. (Risk) *When* $\|\beta\|_2^2 > 0$*, for all* $\lambda > 0$ *and* $\epsilon > 0$*, there exist* $\sigma^2, \phi \in (0, \infty)$*, such that* $\partial \mathscr{R}(\lambda, \phi)/\partial \phi \leq -\epsilon$*, i.e.,*

$$\max_{\sigma^2, \phi \in (0, \infty)} \min_{\lambda \geq \lambda_{\min}(\phi)} \partial \mathscr{R}(\lambda, \phi)/\partial \phi \leq -\epsilon. \tag{14}$$

Proof. For isotropic features $\Sigma = I$, analogous to the proof of Theorem 4.2, the excess prediction risk is given by

$$R(\widehat{\beta}^\lambda) \simeq R_p(\lambda, \phi, \phi)$$

where $R_p(\lambda, \phi, \psi)$ is defined in (6). When $\Sigma = I$, the non-negative constants $\widetilde{c}_p(\lambda, \psi)$ and $\widetilde{v}_p(\lambda, \phi, \psi)$ are defined through the following equations:

$$v_p(\lambda, \phi) = \frac{\sqrt{(\phi + \lambda - 1)^2 + 4\lambda} - (\phi + \lambda - 1)}{2\lambda},$$

$$\widetilde{v}_p(\lambda, \phi, \psi) = \frac{\phi \dfrac{1}{(1+v_p(\lambda, \psi))^2}}{v_p(\lambda, \psi)^{-2} - \phi \dfrac{1}{(1+v_p(\lambda, \psi))^2}},$$
$$\widetilde{c}_p(\lambda, \psi) = (v_p(\lambda, \psi)\Sigma + I)^{-2}\alpha^2.$$

From Theorem 1 of Patil & Du (2023), for all $(\lambda, \phi) \in (0, \infty)^2$, there exists $\psi = \psi(\lambda, \phi)$ such that the prediction risk (45) of the full-ensemble estimator are asymptotically equivalent:

$$R_p(\lambda, \phi, \phi) \ \simeq \ R_p(0; \phi, \psi).$$

Note that the left-hand side is simply the risk of the ridge predictor with ridge penalty λ. Furthermore, from (44), it holds that

$$\lambda + \phi \int \frac{r}{1 + v_p(0; \psi) r} \, \mathrm{d}H(r) = \psi \int \frac{r}{1 + v_p(0; \psi) r} \, \mathrm{d}H(r).$$

Taking the derivative with respect to ϕ on both sides yields that

$$\frac{\partial \psi}{\partial \phi} = 1 - \frac{\lambda v_p(0; \psi)}{1 - \lambda v_p(0; \psi)} \widetilde{v}_p(0; \phi, \psi). \tag{42}$$

We consider three cases below.

(1) $\alpha^2 = 0$ **and** $\sigma^2 > 0$**.** In this case, the excess risk equals the variance component:

$$\begin{aligned} R'_p(\lambda, \phi, \phi) &= \sigma^2 \widetilde{v}_p(\lambda, \phi, \phi) \\ &= \sigma^2 \left(\frac{1}{1 - \phi \int \left(\dfrac{v_p(\lambda, \phi) r}{1 + v_p(\lambda, \phi) r} \right)^2 \mathrm{d}H_p(r)} - 1 \right). \end{aligned}$$

Let $f(\phi) := \phi \int \left(v_p(\lambda, \phi) r / (1 + v_p(\lambda, \phi) r)\right)^2 \, \mathrm{d}H_p(r)$. Then, the monotonicity of the above display in ϕ is the same as that of f in ϕ. Note that

$$f(\phi) = \phi \left(\frac{\sqrt{(\phi + \lambda - 1)^2 + 4\lambda} - (\phi + \lambda - 1)}{\sqrt{(\phi + \lambda - 1)^2 + 4\lambda} - (\phi - \lambda - 1)} \right)^2. \tag{43}$$

Taking the derivative with respect to ϕ, we have

$$f'(\phi) = \frac{(-\phi + \lambda + 1)\left(\sqrt{(\phi + \lambda - 1)^2 + 4\lambda} - \phi - \lambda + 1\right)^2}{\sqrt{(\phi + \lambda - 1)^2 + 4\lambda}\left(\sqrt{(\phi + \lambda - 1)^2 + 4\lambda} + \lambda - \phi + 1\right)^2}.$$

Since $f'(\phi) > 0$ when $\phi \in (0, \lambda + 1)$ and $f'(\phi) < 0$ when $\phi \in (\lambda + 1, +\infty)$, we know that R_p is strictly increasing over $\phi \in (0, \lambda + 1)$ and strictly decreasing over $\phi \in (\lambda + 1, +\infty)$. Thus, the monotonicity of $R_p(\lambda, \phi, \phi)$ follows.

(2) $\alpha^2 > 0$ **and** $\sigma^2 = 0$**.** In this case, the excess risk equals the bias component:

$$\begin{aligned} R_p(\lambda, \phi, \phi) &= \widetilde{c}_p(-\lambda; \phi)(\widetilde{v}_p(\lambda, \phi, \phi) + 1) \\ &= \alpha^2 \frac{\dfrac{1}{(1 + v_p(\lambda, \phi))^2}}{1 - \phi \dfrac{1}{(1 + v_p(\lambda, \phi))^2}} \end{aligned}$$

$$= \frac{\alpha^2}{(1-\phi)v_p(\lambda,\phi)^2 + 2v_p(\lambda,\phi) + 1}.$$

Let $g(\phi) = (1-\phi)v_p(\lambda,\phi)^2 + 2v_p(\lambda,\phi) + 1$. Taking the derivative with respect to ϕ yields

$$\begin{aligned} g'(\phi) = & \frac{\left(\sqrt{(\phi+\lambda-1)^2+4\lambda} - \lambda - \phi + 1\right)}{4\lambda^2\sqrt{(\lambda+\phi-1)^2+4\lambda}} \\ & \cdot \left(\lambda + 3(\phi-1)\sqrt{\lambda^2+2\lambda(\phi+1)+\phi^2-2\phi+1} - \lambda^2 - 4\lambda(\phi+1) - 3(\phi-1)^2\right) \\ =: & \frac{\left(\sqrt{(\phi+\lambda-1)^2+4\lambda} - \lambda - \phi + 1\right)}{4\lambda^2\sqrt{(\lambda+\phi-1)^2+4\lambda}} h(\phi). \end{aligned}$$

By simple calculations, one can show that $h'(\phi) < 0$ and $h(\phi) \le h(0) < -2\lambda$ when $\lambda > 0$. Therefore, we have $g'(\phi) < 0$ and R_p is strictly increasing over $\phi \in (0,\infty)$.

(3) General cases when $\alpha^2 > 0$. Note that

$$\begin{aligned} R_p(\lambda,\phi,\phi) &= \sigma^2 \widetilde{v}_p(\lambda,\phi,\psi) + \widetilde{c}_p(-\lambda;\phi)(\widetilde{v}_p(\lambda,\phi,\phi)+1) \\ &=: f_1(\phi) + f_2(\phi), \end{aligned}$$

where f_1 first increases and then decreases in ϕ, and f_2 is a strictly increasing function. Note that only f_1 depends on σ^2. Because for any $\lambda > 0$ and $\phi \in (\lambda+1,\infty)$, $f_1'(\phi) < 0$ and its scale is proportional to σ^2, we have that for all $\epsilon > 0$, there exists $\sigma^2 > 0$ such that $-f_1'(\phi) > f_2'(\phi) + \epsilon$. This implies that

$$\max_{\sigma^2,\phi\in(0,\infty)} \min_{\lambda\ge 0} \frac{\partial \mathscr{R}(\lambda,\phi)}{\partial \phi} \le -\epsilon,$$

which completes the proof. □

E.4. Proof of Theorem 4.4

Theorem 4.4 (Optimal ensemble versus ridge regression under negative regularization)**.** *Let* $\mathscr{R}^{**} := \min_{\psi\ge\phi,\lambda\ge\lambda_{\min}(\phi)} \mathscr{R}(\lambda,\phi,\psi)$*. Then the following statements hold:*

1. (Underparameterized) *When* $\phi < 1$ *and* $\beta_0 = \beta$*,* $\lambda^* \ge 0$*,*

$$\mathscr{R}^{**} = \min_{\lambda\ge 0} \mathscr{R}(\lambda,\phi,\phi) = \min_{\psi\ge\phi} \mathscr{R}(0;\phi,\psi).$$

2. (Overparameterized) *When* $\phi \ge 1$*,* $\lambda^* \ge \lambda_{\min}(\phi)$*,*

$$\mathscr{R}^{**} = \min_{\lambda\ge\lambda_{\min}(\phi)} \mathscr{R}(\lambda,\phi,\phi) = \min_{\psi\ge\phi} \mathscr{R}(\lambda_{\min}(\phi);\phi,\psi).$$

Proof. We split the proof into different parts.

Part (1) Risk characterization. From Lemma E.2, we have $R(\widehat{\beta}^{\lambda}_{k,\infty}) \simeq \mathscr{R}_p(\lambda,\phi,\psi)$.

Part (2) Risk equivalence. From Lemma F.4, for any $\bar{\psi} \in [\phi,+\infty]$, there exists a unique $\bar{\lambda} \ge \lambda_{\min}(\phi)$ such that

$$\frac{1}{v} = \bar{\lambda} + \phi \int \frac{r}{1+vr} \,\mathrm{d}H_p(r), \quad \text{and} \quad \frac{1}{v} = \bar{\psi} \int \frac{r}{1+vr} \,\mathrm{d}H_p(r). \tag{44}$$

For any pair of (λ_1,ψ_1) and (λ_2,ψ_2) on the path $\mathcal{P}(\bar{\lambda};\phi,\bar{\psi})$ as defined in (55), we have that:

$$\mathscr{R}_p(\lambda_1;\phi,\psi_1) = \mathscr{R}_p(\lambda_2;\phi,\psi_2).$$

Part (3) Optimal risk. From (49), we have that for any $\lambda \geq \lambda_{\min}(\phi)$, there exists $\psi \geq 1$ such that $|\mathscr{R}_p(\lambda_{\min}(\phi);\phi,\psi) - \mathscr{R}_p(\lambda,\phi,\phi)| \xrightarrow{\text{a.s.}} 0$. From Lemma F.1 and Lemma F.3, $1/v_p(-\lambda;\psi) \in [-r_{\min},\infty]$ and

$$\lim_{\psi\to\phi} \mathscr{R}_p(\lambda_{\min}(\phi);\phi,\psi) = \lim_{\lambda\to\lambda_{\min}(\phi)^+} \mathscr{R}_p(\lambda,\phi,\phi) = +\infty.$$

Similar to the proof of Lemma 28 from Bellec et al. (2023), one can show that the sequence of functions $\{\mathscr{R}_p(\lambda_{\min}(\phi);\phi,\psi(\lambda)) - \mathscr{R}_p(\lambda,\phi,\phi)\}_{p\in\mathbb{N}}$ is uniformly equicontinuous on $\lambda \in \Lambda = [\lambda_{\min}(\phi)+\epsilon,\infty]$ almost surely for some small $\epsilon > 0$ such that $\mathscr{R}_p(\lambda,\phi,\psi)$ is no larger than the null risk $\mathscr{R}_p(+\infty;\phi,\phi)$ when $\lambda \in [\lambda_{\min}(\phi)+\epsilon,+\infty]$. From Theorem 21.8 of Davidson (1994), it further follows that the sequences converge to zero uniformly over Λ almost surely. This implies that

$$\begin{aligned}
0 &= \limsup_p \max_{\lambda\geq\lambda_{\min}(\phi)+\epsilon} [\mathscr{R}_p(\lambda_{\min}(\phi);\phi,\psi(\lambda)) - \mathscr{R}_p(\lambda,\phi,\phi)] \\
&\geq \min_{\psi\geq\phi} \limsup_p \max_{\lambda\geq\lambda_{\min}(\phi)+\epsilon} [\mathscr{R}_p(\lambda_{\min}(\phi);\phi,\psi) - \mathscr{R}_p(\lambda,\phi,\phi)] \\
&\geq \limsup_p \min_{\psi\geq\phi} \max_{\lambda\geq\lambda_{\min}(\phi)+\epsilon} [\mathscr{R}_p(\lambda_{\min}(\phi);\phi,\psi) - \mathscr{R}_p(\lambda,\phi,\phi)] \\
&= \limsup_p \left[\min_{\psi\geq\phi} \mathscr{R}_p(\lambda_{\min}(\phi);\phi,\psi) - \min_{\lambda\geq\lambda_{\min}(\phi)+\epsilon} \mathscr{R}_p(\lambda,\phi,\phi)\right].
\end{aligned}$$

Conversely, since for any $\psi \geq \psi(\lambda_{\min}(\phi)+\epsilon)$, there exists $\lambda \geq \lambda_{\min}(\phi)+\epsilon$ such that $|\mathscr{R}_p(\lambda_{\min}(\phi);\phi,\psi) - \mathscr{R}_p(\lambda(\psi);\phi,\phi)| \xrightarrow{\text{a.s.}} 0$. Similarly, we can show that $\{\mathscr{R}_p(\lambda_{\min}(\phi);\phi,\psi) - \mathscr{R}_p(\lambda(\psi);\phi,\phi)\}_{p\in\mathbb{N}}$ is uniformly equicontinuous on $\psi \in \Psi = [\psi(\lambda_{\min}(\phi)+\epsilon),\infty]$ almost surely. This also implies that

$$\begin{aligned}
0 &= \liminf_p \min_{\psi\geq\psi(\lambda_{\min}(\phi)+\epsilon)} [\mathscr{R}_p(\lambda_{\min}(\phi);\phi,\psi) - \mathscr{R}_p(\lambda(\psi);\phi,\phi)] \\
&\leq \max_{\lambda\geq\lambda_{\min}(\phi)} \liminf_p \min_{\psi\geq\psi(\lambda_{\min}(\phi)+\epsilon)} [\mathscr{R}_p(\lambda_{\min}(\phi);\phi,\psi) - \mathscr{R}_p(\lambda,\phi,\phi)] \\
&\leq \liminf_p \max_{\lambda\geq\lambda_{\min}(\phi)} \min_{\psi\geq\psi(\lambda_{\min}(\phi)+\epsilon)} [\mathscr{R}_p(\lambda_{\min}(\phi);\phi,\psi) - \mathscr{R}_p(\lambda,\phi,\phi)] \\
&= \liminf_p \left[\min_{\psi\geq\psi(\lambda_{\min}(\phi)+\epsilon)} \mathscr{R}_p(\lambda_{\min}(\phi);\phi,\psi) - \min_{\lambda\geq\lambda_{\min}(\phi)} \mathscr{R}_p(\lambda,\phi,\phi)\right].
\end{aligned}$$

Combining the previous two inequalities implies that

$$\min_{\lambda\geq\lambda_{\min}(\phi)+\epsilon} \mathscr{R}_p(\lambda,\phi,\phi) \simeq \min_{\psi\geq\psi(\lambda_{\min}(\phi)+\epsilon)} \mathscr{R}_p(\lambda_{\min}(\phi);\phi,\psi).$$

Since

$$\begin{aligned}
\min_{\lambda\geq\lambda_{\min}(\phi)+\epsilon} \mathscr{R}_p(\lambda,\phi,\phi) &= \min_{\lambda\geq\lambda_{\min}(\phi)} \mathscr{R}_p(\lambda,\phi,\phi) \\
\min_{\psi\geq\psi(\lambda_{\min}(\phi)+\epsilon)} \mathscr{R}_p(\lambda_{\min}(\phi);\phi,\psi) &= \min_{\psi\geq\phi} \mathscr{R}_p(\lambda_{\min}(\phi);\phi,\psi),
\end{aligned}$$

we further have

$$\min_{\lambda\geq\lambda_{\min}(\phi)} \mathscr{R}_p(\lambda,\phi,\phi) = \min_{\psi\geq\phi} \mathscr{R}_p(\lambda_{\min}(\phi);\phi,\psi),$$

which holds for $\phi \in (0,\infty)$. This finishes the proof of the second conclusion.

Part (4) Optimal risk when $\phi < 1$. When $\beta_0 = \beta$, the excess bias term $\mathscr{S} \equiv 0$. From Lemma D.4, we have that the bias component $\widetilde{c}_p(\lambda,\phi,\phi)$ of the risk equivalent is minimized at $\lambda = 0$ when $\phi < 1$. Since $\widetilde{v}_p(\lambda,\phi,\phi)$ is a strictly increasing function in $v_p(\lambda,\phi)$ and $v_p(\lambda,\phi)$ is a strictly decreasing function in λ, we see that $\widetilde{v}_p(\lambda,\phi,\phi)$ is a strictly decreasing function in λ. Thus, we have that

$$\min_{\lambda\in[\lambda_{\min},0]} \mathscr{R}_p(\lambda,\phi,\phi) \geq \mathscr{R}_p(0;\phi,\phi).$$

This implies that

$$\min_{\lambda\geq\lambda_{\min}} \mathscr{R}_p(\lambda,\phi,\phi) \geq \min_{\lambda\geq0} \mathscr{R}_p(0;\phi,\phi),$$

which finishes the proof of the first conclusion. □

E.5. Helper Lemmas

Lemma E.1 (Monotonicity of generalized prediction risk with optimal ridge regularization)**.** *Suppose Assumption 2.2 hold. Define the generalized mean squared risk for a estimator $\widehat{\beta}$ as:*

$$R(\widehat{\beta}; A, b, \beta_0) = \|L_{A,b}(\widehat{\beta} - \beta_0)\|_2^2, \tag{45}$$

where $L_{A,b}(\beta) = A\beta + b$ is a linear functional and (A, b) is independent of (X, y) such that $\|A\|_{\mathrm{op}}$ and $\|b\|_2$ are almost surely bounded. Then, as $k, n, p \to \infty$ such that $p/n \to \phi \in (0, \infty)$ and $p/k \to \psi \in [\phi, \infty]$, there exists a sequence of random variables $\{\mathscr{Q}_p(\lambda, \phi, \psi)\}_{p=1}^{\infty}$ that is asymptotically equivalent to the risk of the full-ensemble ridge predictor,

$$R(\widehat{\beta}^{\lambda}_{k,\infty}; A, b, \beta_0) \simeq \mathscr{Q}_p(\lambda, \phi, \psi) := \widetilde{c}_p(\lambda, \phi, \psi, A^\top A) + \|f_{\mathrm{NL}}\|_{L_2}^2 \widetilde{v}_p(\lambda, \phi, \psi, A^\top A), \tag{46}$$

where the non-negative constants $\widetilde{c}_p(\lambda, \phi, \psi, A^\top A)$ and $\widetilde{v}_p(\lambda, \phi, \psi, A^\top A)$ are defined through the following equations:

$$\begin{aligned}
\frac{1}{v_p(\lambda, \psi)} &= \lambda + \psi \int \frac{r}{1 + v_p(\lambda, \psi) r} \,\mathrm{d}H_p(r), \\
\widetilde{v}_p(\lambda, \phi, \psi, A^\top A) &= \frac{\phi \,\overline{\mathrm{tr}}[A^\top A \Sigma (v_p(\lambda, \psi)\Sigma + I)^{-2}]}{v_p(\lambda, \psi)^{-2} - \phi \displaystyle\int \frac{r^2}{(1 + v_p(\lambda, \psi) r)^2} \,\mathrm{d}H_p(r)}, \\
\widetilde{c}_p(\lambda, \phi, \psi, A^\top A) &= \beta_0^\top (v_p(\lambda, \psi)\Sigma + I)^{-1} (\widetilde{v}_p(\lambda, \phi, \psi, A^\top A)\Sigma + A^\top A)(v_p(\lambda, \psi)\Sigma + I)^{-1} \beta_0.
\end{aligned}$$

For the ridge predictor when $k = n$ and $\lambda_{\min}(\phi)$ defined in (5), the optimal risk equivalence $\min_{\lambda \geq \lambda_{\min}(\phi)} \mathscr{Q}_p(\lambda, \phi, \phi)$ is monotonically increasing in ϕ.

Proof. Given an observation (x, y), recall the decomposition $y = f_{\mathrm{LI}}(x) + f_{\mathrm{NL}}(x)$ explained in Section 2. For n i.i.d. samples from the same distribution as (x, y), we define analogously the vector decomposition:

$$y = f_{\mathrm{LI}} + f_{\mathrm{NL}}, \tag{47}$$

where $f_{\mathrm{LI}} = X\beta_0$ and $f_{\mathrm{NL}} = [f_{\mathrm{NL}}(x_i)]_{i \in [n]}$.

Note that

$$R(\widehat{\beta}^{\lambda}_{k,\infty}; A, b, \beta_0) = R(\widehat{\beta}^{\lambda}_{k,\infty}; A, 0, \beta_0) + 2b^\top A(\widehat{\beta}^{\lambda}_{k,\infty} - \beta_0) + \|b\|_2^2.$$

By Theorem 3 of Patil & Du (2023), the cross term vanishes, that is, $b^\top A(\widehat{\beta}^{\lambda}_{k,\infty} - \beta_0) \xrightarrow{\text{a.s.}} 0$. We then have

$$|R(\widehat{\beta}^{\lambda_1}_{k_1,\infty}; A, b, \beta_0) - R(\widehat{\beta}^{\lambda_2}_{k_2,\infty}; A, b, \beta_0)| \xrightarrow{\text{a.s.}} |R(\widehat{\beta}^{\lambda_1}_{k_1,\infty}; A, 0, \beta_0) - R(\widehat{\beta}^{\lambda_2}_{k_2,\infty}; A, 0, \beta_0)|.$$

Thus, it suffices to analyze $R(\widehat{\beta}^{\lambda}_{k,\infty}; A, 0, \beta_0)$. To simplify the notation, we define

$$R_p(\lambda, \phi, \psi) := R\big(\widehat{\beta}^{\lambda_1}_{\lfloor p/\psi \rfloor,\infty}; A, 0_p, \beta_0\big) \tag{48}$$

to indicate the dependency solely on p and (λ, ϕ, ψ). We split the proof into different parts.

Part (1) Risk characterization. Under Assumption 2.2, from Equation (11) of Patil & Du (2023), we have that for $\lambda \geq 0$,

$$R(\widehat{\beta}^{\lambda}_{k,\infty}; A, b, \beta_0) \simeq \mathscr{Q}_p(\lambda, \phi, \psi)$$

where $\mathscr{Q}_p$ is defined in (46). Note that for $\lambda \in [\lambda_{\min}, 0)$, the fixed-point solution $v_p(\lambda, \psi)$ satisfies the same fixed-point equation as the one for $\lambda \geq 0$. Since the above deterministic equivalent depends on λ only through $v_p(\lambda, \psi)$, it also applies to $\lambda \in [\lambda_{\min}, 0)$.

Part (2) Risk equivalence. From Lemma F.4, we have that, for any $\overline{\psi} \in [\phi, +\infty]$, there exists $\overline{\lambda}$ uniquely defined through (44). For any pair of (λ_1, ψ_1) and (λ_2, ψ_2) on the path $\mathcal{P}(\overline{\lambda}; \phi, \overline{\psi})$ as defined in (55), the generalized risk functionals (45) of the full-ensemble estimator are asymptotically equivalent:

$$R_p(\lambda_1; \phi, \psi_1) \simeq \mathscr{Q}_p(\lambda_1; \phi, \psi_1) = \mathscr{Q}_p(\lambda_2; \phi, \psi_2) \simeq R_p(\lambda_2; \phi, \psi_2). \tag{49}$$

Part (3) Risk monotonicity. Note that from Lemma F.10 (3) and Lemma F.11 (3) Du et al. (2023), $v_p(\lambda, \psi)^{-2} - \phi \int r^2 (1 + v_p(\lambda, \psi) r)^{-2} \, \mathrm{d}H_p(r)$ is non-negative. Then we have

$$\widetilde{v}_p(\lambda, \phi, \psi, A^\top A) = \frac{\overline{\mathrm{tr}}[A^\top A \Sigma (v_p(\lambda, \psi)\Sigma + I)^{-2}]}{\phi^{-1} v_p(\lambda, \psi)^{-2} - \displaystyle\int \frac{r^2}{(1 + v_p(\lambda, \psi) r)^2} \, \mathrm{d}H_p(r)}$$

is increasing in $\phi \in (0, \psi]$ for any fixed ψ. Thus, $\mathscr{Q}_p(\lambda, \phi, \psi)$ is increasing in ϕ for any fixed (λ, ψ). Furthermore, since $\mathscr{Q}_p(\lambda, \phi, \psi)$ is a continuous function of ϕ and $v(\lambda; \psi)$, it follows that for $0 < \phi_1 \le \phi_2 < \infty$,

$$\min_{\psi \ge \phi_1} \mathscr{Q}_p(\lambda, \phi_1, \psi) \le \min_{\psi \ge \phi_2} \mathscr{Q}_p(\lambda, \phi_1, \psi) \le \min_{\psi \ge \phi_2} \mathscr{Q}_p(\lambda, \phi_2, \psi),$$

where the first inequality follows because $\{\psi : \psi \ge \phi_1\} \supseteq \{\psi : \psi \ge \phi_2\}$, and the second inequality follows because $\mathscr{Q}_p(\lambda, \phi, \psi)$ is increasing in ϕ for a fixed ψ. Thus, $\min_{\psi \ge \phi} \mathscr{Q}_p(\lambda, \phi, \psi)$ is a continuous and monotonically increasing function in ϕ.

Part (4) Optimal subsampling and regularization. From (49), we have that for any $\lambda \ge \lambda_{\min}(\phi)$, there exists $\psi \ge 1$ such that $|\mathscr{Q}_p(\lambda_{\min}(\phi); \phi, \psi) - R_p(\lambda, \phi, \phi)| \xrightarrow{\text{a.s.}} 0$. From Lemma F.1 and Lemma F.3, $1/v_p(-\lambda; \psi) \in [-r_{\min}, \infty]$ and $\lim_{\psi \to \phi} \mathscr{Q}_p(\lambda_{\min}(\phi); \phi, \psi) = +\infty$. Similarly to the proof of Lemma 28 from Bellec et al. (2023), one can show that the sequence of functions $\{\mathscr{Q}_p(\lambda_{\min}(\phi); \phi, \psi(\lambda)) - R_p(\lambda, \phi, \phi)\}_{p \in \mathbb{N}}$ is uniformly equicontinuous on $\lambda \in \Lambda = [\lambda_{\min}(\phi) + \epsilon, \infty]$ almost surely for some small $\epsilon > 0$ such that $|\mathscr{Q}_p(\lambda_{\min}(\phi); \phi, \psi(\lambda))|$ is not greater than the null risk $\mathscr{Q}_p(+\infty; \phi, \phi)$ when $\lambda \in [\lambda_{\min}(\phi) + \epsilon, +\infty]$. From Theorem 21.8 of Davidson (1994), it further follows that the sequences converge to zero uniformly over Λ almost surely. This implies that

$$\begin{aligned}
0 &= \limsup_p \max_{\lambda \ge \lambda_{\min}(\phi) + \epsilon} [\mathscr{Q}_p(\lambda_{\min}(\phi); \phi, \psi(\lambda)) - R_p(\lambda, \phi, \phi)] \\
&\ge \min_{\psi \ge \phi} \limsup_p \max_{\lambda \ge \lambda_{\min}(\phi) + \epsilon} [\mathscr{Q}_p(\lambda_{\min}(\phi); \phi, \psi) - R_p(\lambda, \phi, \phi)] \\
&\ge \limsup_p \min_{\psi \ge \phi} \max_{\lambda \ge \lambda_{\min}(\phi) + \epsilon} [\mathscr{Q}_p(\lambda_{\min}(\phi); \phi, \psi) - R_p(\lambda, \phi, \phi)] \\
&= \limsup_p \left[\min_{\psi \ge \phi} \mathscr{Q}_p(\lambda_{\min}(\phi); \phi, \psi) - \min_{\lambda \ge \lambda_{\min}(\phi) + \epsilon} R_p(\lambda, \phi, \phi) \right].
\end{aligned}$$

Conversely, since for any $\psi \ge \psi(\lambda_{\min}(\phi) + \epsilon)$, there exists $\lambda \ge \lambda_{\min}(\phi) + \epsilon$ such that $|\mathscr{Q}_p(\lambda_{\min}(\phi); \phi, \psi) - R_p(\lambda(\psi); \phi, \phi)| \xrightarrow{\text{a.s.}} 0$. Similarly, we can show that $\{\mathscr{Q}_p(\lambda_{\min}(\phi); \phi, \psi) - R_p(\lambda(\psi); \phi, \phi)\}_{p \in \mathbb{N}}$ is uniformly equicontinuous on $\psi \in \Psi = [\psi(\lambda_{\min}(\phi) + \epsilon), \infty]$ almost surely. This also implies that

$$\begin{aligned}
0 &= \liminf_p \min_{\psi \ge \psi(\lambda_{\min}(\phi) + \epsilon)} [\mathscr{Q}_p(\lambda_{\min}(\phi); \phi, \psi) - R_p(\lambda(\psi); \phi, \phi)] \\
&\le \max_{\lambda \ge \lambda_{\min}(\phi)} \liminf_p \min_{\psi \ge \psi(\lambda_{\min}(\phi) + \epsilon)} [\mathscr{Q}_p(\lambda_{\min}(\phi); \phi, \psi) - R_p(\lambda, \phi, \phi)] \\
&\le \liminf_p \max_{\lambda \ge \lambda_{\min}(\phi)} \min_{\psi \ge \psi(\lambda_{\min}(\phi) + \epsilon)} [\mathscr{Q}_p(\lambda_{\min}(\phi); \phi, \psi) - R_p(\lambda, \phi, \phi)] \\
&= \liminf_p \left[\min_{\psi \ge \psi(\lambda_{\min}(\phi) + \epsilon)} \mathscr{Q}_p(\lambda_{\min}(\phi); \phi, \psi) - \min_{\lambda \ge \lambda_{\min}(\phi)} R_p(\lambda, \phi, \phi) \right].
\end{aligned}$$

Combining the previous two inequalities implies that

$$\min_{\lambda \ge \lambda_{\min}(\phi) + \epsilon} R_p(\lambda, \phi, \phi) \simeq \min_{\psi \ge \psi(\lambda_{\min}(\phi) + \epsilon)} \mathscr{Q}_p(\lambda_{\min}(\phi); \phi, \psi).$$

Since

$$\min_{\psi \geq \psi(\lambda_{\min}(\phi)+\epsilon)} \mathscr{Q}_p(\lambda_{\min}(\phi);\phi,\psi) = \min_{\psi \geq \psi(\lambda_{\min}(\phi))} \mathscr{Q}_p(\lambda_{\min}(\phi);\phi,\psi) = \min_{\psi\geq\phi} \mathscr{Q}_p(\lambda_{\min}(\phi);\phi,\psi),$$

we further have

$$\min_{\lambda \geq \lambda_{\min}(\phi)+\epsilon} R_p(\lambda,\phi,\phi) \simeq \min_{\psi\geq\phi} \mathscr{Q}_p(\lambda_{\min}(\phi);\phi,\psi).$$

From the second part, we know that the latter is continuous and monotonically increasing in ϕ, which finishes the proof. ☐

Lemma E.2 (Out-of-distribution full-ensemble risk asymptotics)**.** *Under Assumption 2.2, as $k, n, p \to \infty$ such that $p/n \to \phi \in (0,\infty)$ and $p/k \to \psi \in [\phi,\infty]$, for $\lambda \geq \lambda_{\min}(\phi)$, it holds that*

$$\begin{aligned} R(\widehat{\beta}^{\lambda}_{k,\infty}) &\simeq \mathscr{R}_p(\lambda,\phi,\psi) \\ &:= \mathscr{Q}_p(\lambda,\phi,\psi) - 2\beta^\top (v_p(\lambda,\phi)\Sigma + I)^{-1}\Sigma_0(\beta-\beta_0) + [(\beta_0-\beta)^\top\Sigma_0(\beta_0-\beta)+\sigma_0^2]. \end{aligned} \tag{50}$$

where

$$\mathscr{Q}_p(\lambda,\phi,\psi) := \widetilde{c}_p(\lambda,\phi,\psi,\Sigma_0) + \|f_{\rm NL}\|^2_{L_2}\widetilde{v}_p(\lambda,\phi,\psi,\Sigma_0),$$

where the non-negative constants $\widetilde{c}_p(\lambda,\phi,\psi,\Sigma_0)$ and $\widetilde{v}_p(\lambda,\phi,\psi,\Sigma_0)$ are defined through the following equations:

$$\begin{aligned} \frac{1}{v_p(\lambda,\psi)} &= \lambda + \psi\int \frac{r}{1+v_p(\lambda,\psi)r}\,\mathrm{d}H_p(r), \\ \widetilde{v}_p(\lambda,\phi,\psi,\Sigma_0) &= \frac{\phi\,\overline{\mathrm{tr}}[\Sigma_0\Sigma(v_p(\lambda,\psi)\Sigma+I)^{-2}]}{v_p(\lambda,\psi)^{-2} - \phi\displaystyle\int\frac{r^2}{(1+v_p(\lambda,\psi)r)^2}\,\mathrm{d}H_p(r)}, \\ \widetilde{c}_p(\lambda,\phi,\psi,\Sigma_0) &= \beta^\top(v_p(\lambda,\psi)\Sigma+I)^{-1}(\widetilde{v}_p(\lambda,\phi,\psi,\Sigma_0)\Sigma+\Sigma_0)(v_p(\lambda,\psi)\Sigma+I)^{-1}\beta. \end{aligned}$$

Proof. Similar to the proof of Proposition 2.4, we have the decomposition

$$\begin{aligned} R(\widehat{\beta}^{\lambda}_{k,\infty}) &= (\widehat{\beta}^{\lambda}_{k,\infty}-\beta_0)^\top\Sigma_0(\widehat{\beta}^{\lambda}_{k,\infty}-\beta_0)+\sigma_0^2 \\ &= (\widehat{\beta}^{\lambda}_{k,\infty}-\beta)^\top\Sigma_0(\widehat{\beta}^{\lambda}_{k,\infty}-\beta) + 2(\widehat{\beta}^{\lambda}_{k,\infty}-\beta)^\top\Sigma_0(\beta-\beta_0) + [(\beta_0-\beta)^\top\Sigma_0(\beta_0-\beta)+\sigma_0^2] \\ &\simeq \mathscr{Q}_p(\lambda,\phi,\psi) + 2\mathbb{E}_{I\sim\mathcal{I}_k}[(\widehat{\beta}(I)-\beta)^\top\Sigma_0(\beta-\beta_0)\mid\mathcal{D}_n] + [(\beta_0-\beta)^\top\Sigma_0(\beta_0-\beta)+\sigma_0^2] \\ &\simeq \mathscr{Q}_p(\lambda,\phi,\psi) - 2\beta^\top(v_p(\lambda,\psi)\Sigma+I)^{-1}\Sigma_0(\beta-\beta_0) + [(\beta_0-\beta)^\top\Sigma_0(\beta_0-\beta)+\sigma_0^2], \end{aligned}$$

where the first asymptotic equivalent is from Lemma E.1 and the second is from Patil & Du (2023, Lemma D.1 (1)). This finishes the proof. ☐

F. Technical Lemmas

F.1. Fixed-Point Equations for Minimum Ridge Penalty

Recall that under Assumption 2.2, the minimum ridge penalty $\lambda_{\min} = \lambda_{\min}(\phi)$ can be determined by the following equations:

$$\begin{aligned} 1 &= \phi\int\left(\frac{v_0 r}{1+v_0 r}\right)^2\mathrm{d}P(r), \qquad -r_{\min} < v_0^{-1} \leq 0, \\ \frac{1}{v_0} &= \lambda_{\min} + \phi\int\frac{r}{1+v_0 r}\,\mathrm{d}P(r), \end{aligned}$$

or equivalently

$$1 = \phi\int\left(\frac{r}{\mu_0+r}\right)^2\mathrm{d}P(r), \qquad -r_{\min} < \mu_0 \leq 0,$$

$$\mu_0 = \lambda_{\min} + \phi \int \frac{\mu_0 r}{\mu_0 + r} \, \mathrm{d}P(r).$$

We next analyze the properties of μ_0 in ϕ.

Lemma F.1 (Continuity properties with the minimum regularization parameter)**.** *Let $a > 0$ and $b < \infty$ be real numbers. Let P be a probability measure supported on $[a, b]$. Consider the function $v_0(\cdot) : \phi \mapsto \mu_0(\phi)$, over $(0, \infty)$, where $-a \le \mu_0(\phi) \le 0$ is the unique solution to the following fixed-point equation:*

$$1 = \phi \int \left(\frac{r}{\mu_0(\phi) + r} \right)^2 \mathrm{d}P(r), \tag{51}$$

Then the following properties hold:

(1) *The range of $\mu_0(\phi)$ is $[-a, \infty)$.*
(2) *$\mu_0(\phi)$ is continuous and strictly increasing over $\phi \in (0, \infty)$. In addition, $\mu_0(\phi)$ has a root at $\phi = 1$.*
(3) *$\lambda_{\min}(\phi) = \mu_0(\phi)(1 - \phi \int r/(\mu_0(\phi) + r) \, \mathrm{d}P(r))$ is non-positive over $\phi \in (0, \infty)$ with zero obtained when $\phi = 1$. Furthermore, it is strictly increasing over $\phi \in (0, 1)$ and strictly decreasing over $\phi \in (1, \infty)$.*

Proof. The existence of the solution $v_0(\phi) = 1/\mu_0(\phi)$ to the fixed-point equation

$$\frac{1}{\phi} = \int \left(\frac{v_0(\phi) r}{1 + v_0(\phi) r} \right)^2 \mathrm{d}P(r)$$

follows from Theorem 3.1 of LeJeune et al. (2024). Next, we split the proof into different parts.

Part (1). Define $h(x) = \int (xr)^2/(1 + xr)^2 \, \mathrm{d}P(r)$.

When $\phi < 1$, $h(v_0(\phi)) = 1/\phi > 1$, which implies that $1/v_0(\phi) \in [-a, 0)$. When $\phi = +\infty$, $1/v_0(\phi) = -a$ if $P(a) > 0$.

When $\phi \ge 1$, $h(v_0(\phi)) = 1/\phi \le 1$, which implies that $v_0(\phi) \in (0, \infty]$, with infinity obtained when $\phi = 1$, or equivalently, $1/v_0(\phi) \in [0, \infty)$.

Part (2). Since $g(t) = h(t^{-1})^{-1}$ is positive, strictly increasing and continuous over $t \in [-a, \infty)$, by the continuous inverse theorem, we have that $1/v_0(\phi) = g^{-1}(\phi)$ is strictly increasing and continuous over $\phi \in (0, \infty)$. From Part (1), we also have $1/v_0(1) = 0$, which finishes the proof.

Part (3). Consider the function $f(x) = 1 - \phi \int r/(x + r) \, \mathrm{d}P(r)$. When $\phi < 1$, we know that $f(\mu_0(\phi)) > 0$ because from (51),

$$1 = \phi \int \left(\frac{r}{\mu_0(\phi) + r} \right)^2 \mathrm{d}P(r) > \phi \int \frac{r}{\mu_0(\phi) + r} \, \mathrm{d}P(r),$$

which holds because $\mu_0(\phi) < 0$ from Part (2). Analogously, when $\phi > 1$, $f(\mu_0(\phi)) < 0$. Therefore, $\lambda_{\min}(\phi) = \mu_0(\phi) f(\mu_0(\phi)) \le 0$ with equality obtained when $\phi = 1$.

Because $f(x)$ is strictly decreasing in both ϕ and x, combining the sign properties, we further find that $\lambda_{\min}(\phi) = \mu_0(\phi) f(\mu_0(\phi))$ is strictly increasing over $\phi \in (0, 1)$ and strictly decreasing over $\phi \in (1, \infty)$. □

F.2. Properties of Fixed-Point Equations under Negative Regularization

Our analysis involves $v(\lambda, \phi)$ as a unique solution to the fixed-point equation in

$$\frac{1}{v(\lambda, \phi)} = \lambda + \phi \int \frac{r}{1 + v(\lambda, \phi) r} \, \mathrm{d}H_p(r). \tag{52}$$

Define $\mu(\lambda, \phi) = 1/v(\lambda, \phi)$. Equivalently, we have that $\mu(\lambda, \phi)$ is a unique solution to the following fixed-point equation:

$$\mu(\lambda, \phi) = \lambda + \phi \int \frac{\mu(\lambda, \phi) r}{\mu(\lambda, \phi) + r} \, \mathrm{d}H_p(r). \tag{53}$$

The analytic properties of the function $\lambda \mapsto v(\lambda, \phi)$ on $(\lambda_{\min}(\phi), +\infty)$ for $\phi \in (1, \infty)$ are detailed in Lemma F.2.

Lemma F.2 (Analytic properties in the regularization parameter)**.** *Let $0 < a \le b < \infty$ be real numbers. Let P be a probability measure supported on $[a, b]$. Let $\phi > 0$ be a real number. For $\lambda > \lambda_{\min}(\phi)$, let $v(\lambda, \phi) > 0$ denote the solution to the fixed-point equation*

$$\mu(\lambda, \phi) = \lambda + \phi \int \frac{\mu(\lambda, \phi) r}{r + \mu(\lambda, \phi)} \, \mathrm{d}P(r).$$

Then the following properties hold:

(1) (Monotonicity) *For $\phi \in (0, \infty)$, the function $\lambda \mapsto \mu(\lambda, \phi)$ is strictly increasing in $\lambda \in (\lambda_{\min}(\phi), \infty)$.*
(2) (Range) *For $\phi \in (0, 1]$, $\lim_{\lambda \to \lambda_{\min}(\phi)^-} \mu(\lambda, \phi) = -\infty$ and $\mu(0, \phi) = 0$. For $\phi \in (1, \infty)$, $\lim_{\lambda \to \lambda_{\min}(\phi)^-} \mu(\lambda, \phi) \in (0, \infty)$. For $\phi \in (0, \infty)$, $\lim_{\lambda \to +\infty} \mu(\lambda, \phi) = +\infty$.*
(3) (Differentiability) *For $\phi \in (0, \infty)$, the function $\lambda \mapsto \mu(\lambda, \phi)$ is differentiable over Λ.*
(4) *The function $\lambda \mapsto \lambda / \mu_p(\lambda, \phi)$ is strictly increasing in λ with $\lim_{\lambda \to 0} \lambda / \mu_p(\lambda, \phi) = 0$ and $\lim_{\lambda \to 0} \lambda / \mu_p(\lambda, \phi) = 1$.*

Proof. Note that

$$\lambda = \mu(\lambda, \phi) - \phi \int \frac{\mu(\lambda, \phi) r}{r + \mu(\lambda, \phi)} \, \mathrm{d}P(r).$$

Define a function f by

$$f(x) = x - \phi \int \frac{xr}{x + r} \, \mathrm{d}P(r).$$

Observe that $\mu(\lambda, \phi) = f^{-1}(\lambda)$. The claim of differentiability of the function $\lambda \mapsto \mu(\lambda, \phi)$ follows from the differentiability and strict monotonicity of f, similar to Patil et al. (2022, Lemma S.6.14).

For the last property, from the definition of the fixed-point equation, we have

$$1 = \frac{\lambda}{\mu_p(\lambda, \phi)} + \phi \int \frac{r}{r + \mu(\lambda, \phi)} \, \mathrm{d}P(r).$$

Because $\phi \int \frac{r}{r + \mu(\lambda, \phi)} \, \mathrm{d}P(r)$ is strictly decreasing in λ, we have that $\lambda / \mu_p(\lambda, \phi)$ is strictly increasing in λ. Because $\lim_{\lambda \to +\infty} \mu_p(\lambda, \phi) = +\infty$, we know that

$$\lim_{\lambda \to \infty} \phi \int \frac{r}{r + \mu(\lambda, \phi)} \, \mathrm{d}P(r) = 0.$$

and thus,

$$\lim_{\lambda \to \infty} \frac{\lambda}{\mu_p(\lambda, \phi)} = 1.$$

On the other hand, because $\lim_{\lambda \to 0} \mu_p(\lambda, \phi) = 0$ for $\phi \in (0, 1]$ and $\lim_{\lambda \to 0} \mu_p(\lambda, \phi) < \infty$, it directly implies that

$$\lim_{\lambda \to 0} \frac{\lambda}{\mu_p(\lambda, \phi)} = 0.$$

Consequently, the conclusion follows. □

Lemma F.3 (Continuity properties in the aspect ratio for ridge regression, adapted from Patil et al. (2023))**.** *Let $a > 0$, $b < \infty$ and $\lambda \in \mathbb{R}$ be real numbers. Let P be a probability measure supported on $[a, b]$. For $\lambda \in \mathbb{R}$, let $\Phi(\lambda) = \{\phi \in (0, \infty) \mid \lambda_{\min}(\phi) < \lambda\}$. Consider the function $\mu(\lambda, \cdot) : \phi \mapsto \mu(\lambda, \phi)$, over $\phi \in \Phi(\lambda)$ such that $\lambda_{\min}(\phi) \le \lambda$, where $\mu(\lambda, \phi) \ge -a$ is the unique solution to the following fixed-point equation:*

$$\mu(\lambda, \phi) = \lambda + \phi \int \frac{\mu(\lambda, \phi) r}{\mu(\lambda, \phi) + r} \, \mathrm{d}P(r). \tag{54}$$

Then the following properties hold:

(1) *The range of the function $\mu(-\lambda; \cdot)$ is a subset of $[\lambda, \infty]$ when $\lambda \ge 0$ and $[-a, \infty]$ when $\lambda < 0$.*
(2) *The function $\mu(-\lambda; \cdot)$ is continuous and strictly increasing over $\Phi(\lambda)$. Furthermore, $\lim_{\phi \to \infty} \mu(\lambda, \phi) = +\infty$, $\lim_{\phi \to 0^+} \mu(\lambda, \phi) = \lambda$ when $\lambda \ge 0$, and $\lim_{\phi \to 0^+} \mu(\lambda, \phi) = -\infty$ when $\lambda < 0$.*

(3) *The function* $\widetilde{v}_v(-\lambda; \cdot) : \phi \mapsto \widetilde{v}_v(\lambda, \phi)$*, where*

$$\widetilde{v}_v(\lambda, \phi) = \left(\mu(\lambda, \phi)^2 - \phi \int (\mu(\lambda, \phi) r)^2 (\mu(\lambda, \phi) + r)^{-2} \, \mathrm{d}P(r)\right)^{-1},$$

is positive and continuous over $\Phi(\lambda)$*.*

(4) *The function* $\widetilde{v}_b(-\lambda; \cdot) : \phi \mapsto \widetilde{v}_b(\lambda, \phi)$*, where*

$$\widetilde{v}_b(\lambda, \phi) = \widetilde{v}_v(\lambda, \phi) \int \phi (\mu(\lambda, \phi) r)^2 (\mu(\lambda, \phi) + r)^{-2} \, \mathrm{d}P(r),$$

is positive and continuous over $\Phi(\lambda)$*.*

Proof. Properties (1)-(4) follow similarly to Patil et al. (2022, Lemma S.6.15). □

F.3. Contours of Fixed-Point Solutions under Negative Regularization

Lemma F.4 (Contours of fixed-point solutions)**.** *As* $n, p \to \infty$ *such that* $p/n \to \phi \in (0, \infty)$*, let* $\lambda_{\min} : \phi \mapsto \lambda_{\min}(\phi)$ *as defined in* (5)*. For any* $\overline{\psi} \in [\phi, +\infty]$*, there exists a unique value* $\overline{\lambda} \geq \lambda_{\min}(\overline{\psi})$ *(or conversely for* $\overline{\lambda} \in [\lambda_{\min}(\phi), \infty]$*, there exists a unique value* $\overline{\psi} \in [\phi, \infty]$*) such that for all* (λ, ψ) *on the path (as in Figure F9)*

$$\mathcal{P} = \{(1 - \theta) \cdot (\overline{\lambda}, \phi) + \theta \cdot (\lambda_{\min}(\overline{\psi}), \overline{\psi}) \mid \theta \in [0, 1]\}, \tag{55}$$

it holds that

$$\mu(\lambda, \psi) = \mu(\overline{\lambda}, \phi) = \mu(\lambda_{\min}(\overline{\psi}), \overline{\psi}),$$

where $\mu(\lambda, \psi)$ *is as defined in* (53)*.*

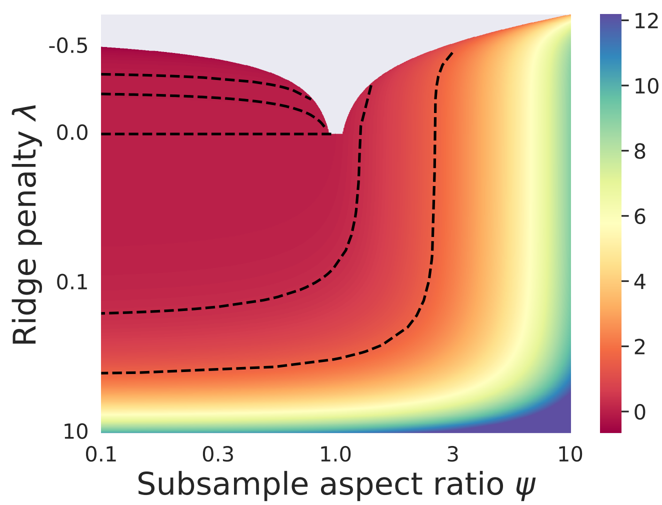

Figure F9: Heatmap of $\mu_p(\lambda, \psi)$ for isotropic covariance matrix $\Sigma = I$ in the symmetric log-log scale (where the logarithmic scale is applied symmetrically to both positive and negative values on the x-axis and y-axis). The 5 black dashed lines indicate 5 different equivalence paths. The boundary of negative ridge penalties is given by $-(1 - \sqrt{\psi})^2$.

Proof. Note that $\mu(\lambda_{\min}(\phi), \phi) = \mu_0(\phi)$ which is the solution to the fixed-point equation (51). From Lemma F.1 (2), we see that the function $\psi \mapsto \mu(\lambda_{\min}(\psi), \psi)$ is strictly increasing over $\psi \in [\phi, \infty]$ with the range

$$\lim_{\psi \to \phi} \mu(\lambda_{\min}(\psi), \psi) = \mu_0(\phi), \quad \lim_{\psi \to +\infty} \mu(\lambda_{\min}(\psi), \psi) = +\infty.$$

From Lemma F.3, the function $\lambda \mapsto \mu(\lambda, \phi)$ is strictly increasing over $\lambda \in [\lambda_{\min}, \infty]$ with range

$$\lim_{\lambda \to \lambda_{\min}} \mu(\lambda, \phi) = \mu_0(\phi), \quad \lim_{\lambda \to +\infty} \mu(\lambda, \phi) = +\infty.$$

For $\bar{\psi} \in [\phi, \infty]$, by the intermediate value theorem, there exists a unique $\bar{\lambda} \in [\lambda_{\min}(\phi), \infty]$ such that $v(\bar{\lambda}, \phi) = v(-\lambda_{\min}(\bar{\psi}); \bar{\psi})$. Conversely, for $\bar{\lambda} \in [\lambda_{\min}(\phi), \infty]$, there also exists a unique $\bar{\psi} \in [\phi, \infty]$ such that $v(\bar{\lambda}, \phi) = v(-\lambda_{\min}(\bar{\psi}); \bar{\psi})$.

Based on the definition of fixed-point solutions, it follows that

$$\mu(\bar{\lambda}, \phi) = \bar{\lambda} + \phi \int \frac{\mu(\bar{\lambda}, \phi) r}{\mu(\bar{\lambda}, \phi) + r} \, \mathrm{d}H_p(r) = \lambda_{\min}(\phi) + \bar{\psi} \int \frac{\mu(-\lambda_{\min}; \bar{\psi}) r}{\mu(-\lambda_{\min}; \bar{\psi}) + r} \, \mathrm{d}H_p(r) = \mu(-\lambda_{\min}; \bar{\psi}).$$

Then, for any $(\lambda, \psi) = (1-\theta)(\bar{\lambda}, \phi) + \theta(\lambda_{\min}(\bar{\psi}), \bar{\psi})$ on the path $\mathcal{P}$, we have

$$\begin{aligned}
\mu(\bar{\lambda}, \phi) &= (1-\theta)\mu(\bar{\lambda}, \phi) + \theta \mu(\lambda_{\min}(\bar{\psi}), \bar{\psi}) \\
&= (1-\theta)\bar{\lambda} + (1-\theta)\phi \int \frac{\mu(\bar{\lambda}, \phi) r}{\mu(\bar{\lambda}, \phi) + r} \, \mathrm{d}H_p(r) + \theta \lambda_{\min}(\bar{\psi}) + \theta \bar{\psi} \int \frac{\mu(-\lambda_{\min}(\bar{\psi}); \bar{\psi}) r}{\mu(-\lambda_{\min}(\bar{\psi}); \bar{\psi}) + r} \, \mathrm{d}H_p(r) \\
&= \lambda + \psi \int \frac{\mu(\bar{\lambda}, \phi) r}{\mu(\bar{\lambda}, \phi) + r} \, \mathrm{d}H_p(r).
\end{aligned}$$

Because $\mu(\lambda, \psi)$ is the unique solution to the fixed-point equation:

$$\mu(\lambda, \psi) = \lambda + \psi \int \frac{\mu(\lambda, \psi) r}{\mu(\lambda, \psi) + r} \, \mathrm{d}H_p(r),$$

it then follows that $\mu(\lambda, \psi) = \mu(\bar{\lambda}, \phi) = \mu(\lambda_{\min}(\bar{\psi}), \bar{\psi})$. This completes the proof. □

G. Experimental Details and Additional Numerical Illustrations

The source code for reproducing the results of this paper can be found at the following location: `https://github.com/jaydu1/ood-ridge`.

G.1. Additional Illustrations for Section 3

G.1.1. Numerical verification of (8) under generic alignment scenarios

Under Assumption 2.2, suppose the data model has a covariance matrix Σ to be $(\Sigma_{\mathrm{ar1}})_{ij} := \rho_{\mathrm{ar1}}^{|i-j|}$ with parameter $\rho_{\mathrm{ar1}} \in (0,1)$, and a coefficient $\beta := \frac{1}{2}(w_{(1)} + w_{(p)})$, where $w_{(j)}$ is the jth eigenvector of Σ_{ar1}.

$$\overline{\mathrm{tr}}[\Sigma B] = \frac{1}{4p}(w_{(1)} + w_{(p)})^\top \Sigma (w_{(1)} + w_{(p)}) = \frac{1}{4p} w_{(1)}^\top \Sigma w_{(1)} + \frac{1}{4p} w_{(p)}^\top \Sigma w_{(p)} = \frac{1}{4}(r_{(1)} + r_{(p)}).$$

On the other hand, we also have

$$\overline{\mathrm{tr}}[\Sigma]\,\overline{\mathrm{tr}}[B] = \sum_{j=1}^{p} \frac{r_{(j)}}{p} \cdot \frac{\|\beta\|_2^2}{p} = \frac{1}{2p^2} \sum_{j=1}^{p} r_{(j)}$$

One can numerically verify that when $p = 500$ and $\rho_{\mathrm{ar1}} = 0.5$,

$$r_{(1)} + r_{(p)} \approx 3.33 > 2 = \frac{2}{p} \sum_{j=1}^{p} r_{(j)}$$

which contradicts the implication (32) of the strict alignment condition.

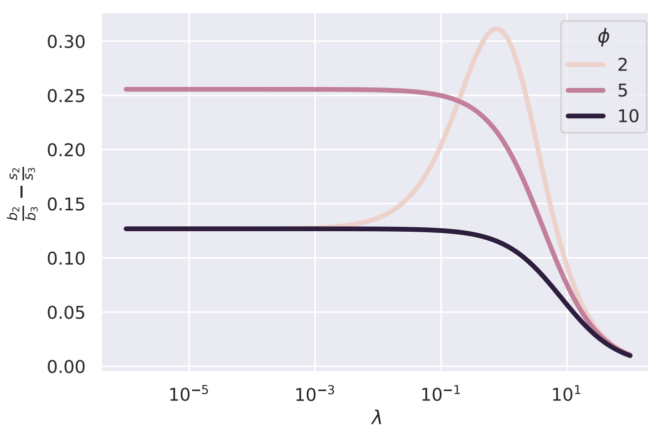

Figure G10: Numerical evaluation of $b_2/b_3 - s_2/s_3$ under the same data model in Figure 1, where $s_k = \overline{\mathrm{tr}}[\Sigma\{\mu^k(\Sigma+\mu I)^{-k}\}]$ and $b_k = \overline{\mathrm{tr}}[\Sigma_\beta\{\mu^k(\Sigma+\mu I)^{-k}\}]$ for $k = 2, 3$.

On the other hand, from Figure G10, we see that the general alignment condition

$$\frac{\overline{\mathrm{tr}}[\Sigma_\beta\{\mu^2(\Sigma+\mu I)^{-2}\}]}{\overline{\mathrm{tr}}[\Sigma_\beta\{\mu^3(\Sigma+\mu I)^{-3}\}]} = \frac{\overline{\mathrm{tr}}[\Sigma\{\mu^2(\Sigma+\mu I)^{-2}\}]}{\overline{\mathrm{tr}}[\Sigma\{\mu^3(\Sigma+\mu I)^{-3}\}]}$$

holds for various data aspect ratios in the noiseless setting.

G.1.2. Real Data Illustration

Following the approach of Kobak et al. (2020), we consider a similar setup using random Fourier features on MNIST.

Feature generation. The pixel values are normalized to the range $[-1, 1]$. The features are then mapped from the original dimension of 28×28 to 1000 random Fourier features. This is achieved by multiplying the features with a 784×500 random matrix W, where the elements are independently drawn from a normal distribution with mean 0 and standard deviation 0.2. The real and imaginary parts of $\exp(-iXW)$ are taken as separate features, where $X \in \mathbb{R}^{n \times 784}$.

Training details. For training, we randomly select $n = 64$ images, and for testing, we hold out 10,000 images. The response variable y represents the digit value ranging from 0 to 9. Our model includes an intercept term, which is not subject to penalization. To estimate the expected out-of-distribution risk, we average the risks across 100 random samples from the training set.

Distribution shift. To generate distribution shift, we gradually exclude samples with given labels in the test set:

- `Type 1`: $\varnothing$
- `Type 2`: $\{4\}$
- `Type 3`: $\{3, 4\}$
- `Type 4`: $\{2, 3, 4\}$
- `Type 5`: $\{1, 2, 3, 4\}$

In other words, `Type` 1 represents no covariate shift, while `Type` 5 represents the case with potentially the most severe covariate shift. The results are summarized in Figure G11. We observe a clear pattern where the optimal ridge penalty shifts toward negative values, suggesting that Theorem 3.3 may occur in real-world datasets. However, for `Type` 5, the optimal ridge penalty becomes positive again. This could be due to the removal of Class 1, which reduces the degree of alignment. Also, observe that the minimum risks increase from `Type 1` to `5`.

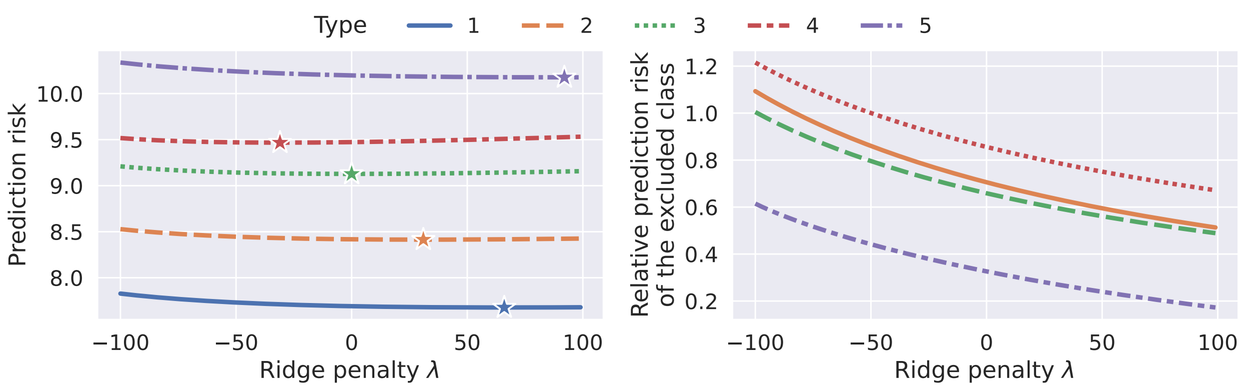

Figure G11: **Effect of distribution shift on the optimal ridge penalty on MNIST.** (a) The left panel illustrates the risk profile (against the regularization penalty) of ridge regression on the MNIST dataset when subjected to different types of distribution shifts. Different colors represent different types of shift from less severe (`Type` 1) to more severe (`Type` 5). The y-axis represents the out-of-distribution prediction risk for the task of accurately predicting the digit value for unseen images. The figure shows a clear pattern where the optimal ridge penalty shifts towards negative values, in the spirit of Theorem 3.4. The only exception seems to be `Type` 5 for which the optimal ridge penalty becomes positive again. It is likely due to the removal of Class 1, which reduces the degree of alignment. (b) The right panel shows the relative OOD prediction risk computed only on the excluded class, compared to the IND prediction risk of the ridge predictor fitted only on the training data of the same class. The reason we compare the relative prediction risk is to compensate for the differences in the conditional variances in the different classes. Observe that Class 1 (`Type` 5) has the lowest relative prediction risk, which partially explains the increase in optimal regularization of `Type` 5 in the left panel.

G.2. Additional Illustrations for Section 4

Experiment details. Similar to the setup in Appendix G.1.2, we further vary the number of training sample size n from 25 to 200, and inspect the OOD prediction risk of the optimal ridge predictor. The results are summarized in Figure G12.

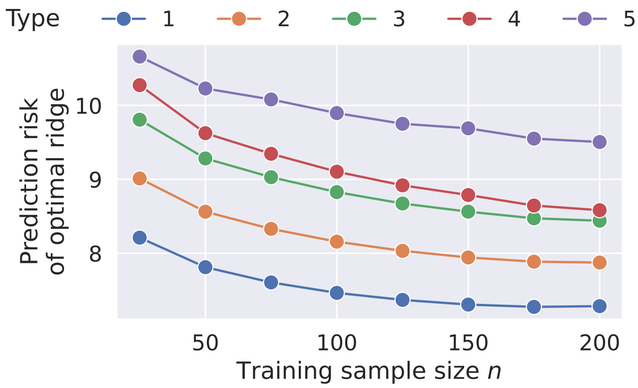

Figure G12: **Effect of distribution shift on the risk monotonicity behavior of optimal ridge on MNIST.** The figure illustrates the risk profile (against the training sample size) of optimal ridge regression on the MNIST dataset when subjected to different types of distribution shifts. We follow the same setup as for Table 2 (see Appendix G.1.2 for more details) and vary the number of training sample size n from 25 to 200, and inspect the OOD prediction risk of the optimal ridge predictor. Different colors represent different types of shift from less severe (`Type` 1) to more severe (`Type` 5). The y-axis represents the out-of-distribution prediction risk for the task of accurately predicting the digit value for unseen images. The figure shows a clear pattern where the optimal ridge exhibits a monotonically decreasing risk in the training sample size n.

G.3. Additional Illustrations for Section 5 (Ridge versus Lasso Monotonicities)

G.3.1. Subsampled ridgeless versus full ridge

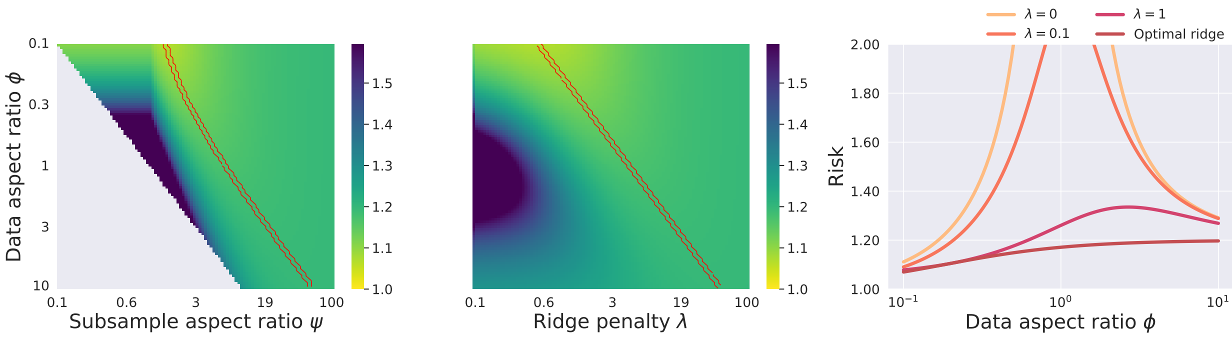

Figure G13: **Ridge optimal risk monotonicity and suboptimal risk non-monotonicity.** Optimal ridge has a monotonic risk in the data aspect ratio, but the risk for fixed ridge penalty λ may not be monotonic in the data aspect ratio. The data model has $\Sigma = \Sigma_{\mathrm{ar1}}$ with $\rho_{\mathrm{ar1}} = 0.25$ (the covariance matrix of an auto-regressive process of order 1 (AR(1)) is given by Σ_{ar1}, where $(\Sigma_{\mathrm{ar1}})_{ij} = \rho_{\mathrm{ar1}}^{|i-j|}$ for some parameter $\rho_{\mathrm{ar1}} \in (0, 1)$), β being the leading eigenvector of Σ and $\sigma = 0.5$. The leftmost panel shows the limiting risk of the full-ensemble ridgeless regression at various data and subsample aspect ratios (ϕ, ψ). The middle panel shows the limiting risk of the ridge predictor (on the full data) at various data aspect ratios and regularization penalties (ϕ, λ). We highlight the optimal risks at a given data aspect ratio for the leftmost and middle panels using slender red lines. Observe that the optimal risk in both cases is increasing as a function of ϕ.

G.3.2. Subsampled lassoless versus full lasso

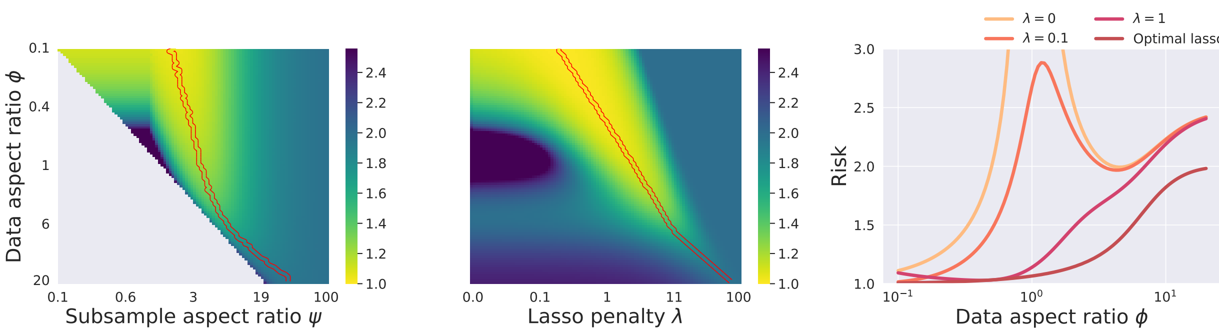

Figure G14: **Lasso optimal risk monotonicity and suboptimal risk non-monotonicity.** Similar to ridge regression, optimal lasso has a monotonic risk in the over-parameterization ratio, but the risk for fixed lasso penalty λ may not be monotonic in the data aspect ratio. The data model has $\Sigma = I$ and $\beta_j \overset{\text{i.i.d.}}{\sim} \epsilon P_{1/\sqrt{\phi\epsilon}} + (1-\epsilon)P_0$ where the sparsity level is $\epsilon = 0.01$ and $\sigma^2 = 1$, such that $\mathrm{SNR} = 1$. As for ridge regression in Figure G13, optimal risks for each data aspect ratio are highlighted using slender red lines in the left and middle panels. Similarly to the ridge curves in Figure G13, observe that the optimal risk in both cases is increasing as a function of ϕ.