On the Optimal Weighted ℓ_2 Regularization in Overparameterized Linear Regression

Denny Wu[*]
University of Toronto and Vector Institute
dennywu@cs.toronto.edu

Ji Xu[*]
Columbia University
jixu@cs.columbia.edu

Abstract

We consider the linear model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_\star + \boldsymbol{\epsilon}$ with $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ in the overparameterized regime $p > n$. We estimate $\boldsymbol{\beta}_\star$ via generalized (weighted) ridge regression: $\hat{\boldsymbol{\beta}}_\lambda = (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{\Sigma}_w)^\dagger \boldsymbol{X}^\top \boldsymbol{y}$, where $\boldsymbol{\Sigma}_w$ is the weighting matrix. Under a random design setting with general data covariance $\boldsymbol{\Sigma}_x$ and general prior on the true coefficients $\mathbb{E}\boldsymbol{\beta}_\star \boldsymbol{\beta}_\star^\top = \boldsymbol{\Sigma}_\beta$, we provide an exact characterization of the prediction risk $\mathbb{E}(y - \boldsymbol{x}^\top \hat{\boldsymbol{\beta}}_\lambda)^2$ in the proportional asymptotic limit $p/n \to \gamma \in (1, \infty)$. Our general setup leads to a number of interesting findings. We outline precise conditions that decide the sign of the optimal choice λ_{opt} of the ridge parameter λ, based on the *alignment* between $\boldsymbol{\Sigma}_x$ and $\boldsymbol{\Sigma}_\beta$; this rigorously justifies the surprising empirical observation that λ_{opt} can be *negative* in the overparameterized regime. We also discuss the risk monotonicity of optimally tuned ridge regression, and confirm the double descent phenomenon for principal component regression (PCR) under anisotropic $\boldsymbol{X}$ and $\boldsymbol{\beta}_\star$. Finally, we determine the optimal $\boldsymbol{\Sigma}_w$ for both the ridgeless ($\lambda \to 0$) and optimally regularized ($\lambda = \lambda_{\text{opt}}$) case, and demonstrate the advantage of the weighted objective over standard ridge regression and PCR.

1 Introduction

In this work we consider learning the target signal $\boldsymbol{\beta}_\star$ in the following linear regression model:

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}_\star + \epsilon_i, \quad i = 1, 2, \ldots, n$$

where each feature vector $\boldsymbol{x}_i \in \mathbb{R}^p$ and noise $\epsilon_i \in \mathbb{R}$ are drawn i.i.d. from the two independent random variables $\tilde{\boldsymbol{x}}$ and $\tilde{\epsilon}$ satisfying $\mathbb{E}\tilde{\epsilon} = 0$, $\mathbb{E}\tilde{\epsilon}^2 = \tilde{\sigma}^2$, $\tilde{\boldsymbol{x}} = \boldsymbol{\Sigma}_x^{1/2}\boldsymbol{z}/\sqrt{n}$, and the components of $\boldsymbol{z}$ are i.i.d. random variables with zero mean, unit variance, and bounded 12th absolute central moment. To estimate $\boldsymbol{\beta}_\star$ from $(\boldsymbol{x}_i, y_i)$, we consider the following generalized ridge regression estimator:

$$\hat{\boldsymbol{\beta}}_\lambda \quad = \quad (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{\Sigma}_w)^\dagger \boldsymbol{X}^\top \boldsymbol{y}, \tag{1.1}$$

in which $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ is the feature matrix, $\boldsymbol{y}$ is vector of the observations, $\boldsymbol{\Sigma}_w$ is a positive definite weighting matrix, and the symbol † denotes the Moore-Penrose pseudo-inverse. When $\lambda \geq 0$, $\hat{\boldsymbol{\beta}}_\lambda$ minimizes the squared loss plus a weighted ℓ_2 regularization: $\min_{\boldsymbol{\beta}} \sum_{i=1}^n (y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2 + \lambda \boldsymbol{\beta}^\top \boldsymbol{\Sigma}_w \boldsymbol{\beta}$. Note that $\boldsymbol{\Sigma}_w = \boldsymbol{I}_d$ reduces the objective to standard ridge regression.

While the standard ridge regression estimator is relatively well-understood in the data-abundant regime ($n > p$), several interesting properties have been recently discovered in high dimensions, especially when $p > n$. For instance, the double descent phenomenon suggests that overparameterization may not result in overfitting due to the *implicit regularization* of the least squares estimator [HMRT19, BLLT19]. This implicit regularization also relates to the surprising empirical finding that the optimal ridge parameter λ can be negative in the overparameterized regime [KLS20].

[*]Equal contribution; alphabetical ordering.

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

Motivated by the observations above, we analyze the estimator $\hat{\boldsymbol{\beta}}_\lambda$ in the proportional limit: $p/n \to \gamma \in (1, \infty)$[2] as $n, p \to \infty$. We place a general prior on the true parameters (independent of $\tilde{\boldsymbol{x}}$ and $\tilde{\epsilon}$): $\mathbb{E}\boldsymbol{\beta}_\star\boldsymbol{\beta}_\star^\top = \boldsymbol{\Sigma}_{\boldsymbol{\beta}}$, which covers both *random* and *deterministic* $\boldsymbol{\beta}_*$. Our goal is to study the prediction risk of $\hat{\boldsymbol{\beta}}_\lambda$: $\mathbb{E}_{\tilde{x},\tilde{\epsilon},\boldsymbol{\beta}_\star}(\tilde{y} - \tilde{\boldsymbol{x}}^\top\hat{\boldsymbol{\beta}}_\lambda)^2$, where $\tilde{y} = \tilde{\boldsymbol{x}}^\top\boldsymbol{\beta}_\star + \tilde{\epsilon}$[3]. Compared to previous high-dimensional analysis of ridge regression [DW18], our setup is generalized in two important aspects:

Anisotropic $\boldsymbol{\Sigma}_x$ and $\boldsymbol{\Sigma}_\beta$. Our analysis handles general prior $\boldsymbol{\Sigma}_\beta$ and data covariance $\boldsymbol{\Sigma}_x$, in contrast to previous works which assume either isotropic features or signal (e.g., [DW18, HMRT19]). Note that the isotropic assumption on the signal or features implies that each component is roughly of the same magnitude, which may not hold true in practice. For instance, the optimal ridge penalty is *provably* non-negative when either the signal $\boldsymbol{\Sigma}_\beta$ [DW18, Theorem 2.1] or the features $\boldsymbol{\Sigma}_x$ [HMRT19, Theorem 5] is isotropic. On the other hand, it has been empirically demonstrated that the optimal ridge for real-world data can be negative [KLS20]. While this observation cannot be captured by previous works, our less restrictive assumptions lead to a concise description of this phenomenon.

Weighted ℓ_2 Regularization. We consider generalized ridge regression instead of simple isotropic shrinkage. While the generalized formulation has also been studied (e.g., [HK70, Cas80]), to the best of our knowledge, no existing work computes the exact risk in the overparameterized proportional limit and decides the corresponding optimal $\boldsymbol{\Sigma}_w$. Our setting is also inspired by recent observations in deep learning that weighted ℓ_2 regularization often achieves better generalization compare to isotropic weight decay [ZWXG18]. Our analysis illustrates the benefit of weighted ℓ_2 regularization.

Under the general setup (1.1), the contributions of this work can be summarized as (see Figure 1):

- **Exact Asymptotic Risk.** In Section 4 we derive the prediction risk $R(\lambda)$ of our estimator (1.1) in its bias-variance decomposition (see Figure 2). We also characterize principal component regression (PCR) and confirm the double descent phenomenon under more general setting than [XH19].
- **"Negative Ridge" Phenomenon.** In Section 5, we analyze the optimal regularization strength λ_{opt} under different $\boldsymbol{\Sigma}_w$, and provide precise conditions under which the optimal λ_{opt} is negative in the overparameterized regime. In brief, we show that λ_{opt} is negative when SNR is large and the large directions of $\boldsymbol{\Sigma}_x$ and $\boldsymbol{\beta}_\star$ are *aligned* (see Figure 4), and vice versa. On the other hand, we show that the optimal ridge penalty is always non-negative in the underparameterized regime ($p < n$); this implies an implicit ℓ_2 regularization effect of overparameterization for certain cases.
- **Optimal Weighting Matrix $\boldsymbol{\Sigma}_w$.** In Section 6, we decide the optimal $\boldsymbol{\Sigma}_w$ for both the optimally regularized ridge estimator ($\lambda = \lambda_{\text{opt}}$) and the ridgeless limit ($\lambda \to 0$). In the ridgeless limit, based on the bias-variance decomposition, we show that the optimal $\boldsymbol{\Sigma}_w$ should interpolate between $\boldsymbol{\Sigma}_x$, which minimizes the variance, and $\boldsymbol{\Sigma}_\beta^{-1}$, which minimizes the bias. Whereas for the optimally regularized case, in many settings the optimal $\boldsymbol{\Sigma}_w$ is simply $\boldsymbol{\Sigma}_\beta^{-1}$ (Figure 6) (for more general result see Theorem 10). We demonstrate the benefit of weighted regularization over standard ridge regression and PCR, and also propose a heuristic choice of $\boldsymbol{\Sigma}_w$ when information of $\boldsymbol{\beta}_\star$ is not present.

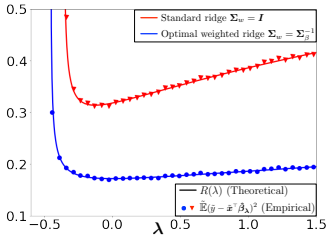


Figure 1: Illustration of the "negative ridge" phenomenon and the advantage of weighted ℓ_2 regularization under "aligned" $\boldsymbol{\Sigma}_x$ and $\boldsymbol{\Sigma}_\beta$. We set $\gamma = 2$, $\tilde{\sigma}^2 = 0$. Red: standard ridge regression ($\boldsymbol{\Sigma}_w = \boldsymbol{I}$); note that the lowest prediction risk is achieved when $\lambda < 0$. Blue: optimally weighted ridge regression ($\boldsymbol{\Sigma}_w = \boldsymbol{\Sigma}_\beta^{-1}$), which achieves lower risk compared to the isotropic shrinkage.

Notations: We denote $\tilde{\mathbb{E}}$ as taking expectation over $\boldsymbol{\beta}_\star, \tilde{x}, \tilde{\epsilon}$. Let $\boldsymbol{d}_x$, $\boldsymbol{d}_\beta$, $\boldsymbol{d}_w$ be the vectors of the eigenvalues of $\boldsymbol{\Sigma}_x, \boldsymbol{\Sigma}_\beta$ and $\boldsymbol{\Sigma}_w$ respectively. We use $\mathbb{I}_{\mathcal{S}}$ as the indicator function of set $\mathcal{S}$. We write $\xi = \mathbb{E}(\tilde{\boldsymbol{x}}^\top\boldsymbol{\beta}_\star)^2/(\gamma\tilde{\sigma}^2)$ as the signal-to-noise ratio (SNR) of the problem.

2 Related Works

Asymptotics of Ridge Regression. The prediction risk of standard ridge regression ($\boldsymbol{\Sigma}_w = \boldsymbol{I}_d$) in the proportional asymptotics has been widely studied. When the data is isotropic, precise characterization can be obtained from random matrix theory [Kar13, Dic16, HMRT19], approximate message

[2] Some of our results also apply to the underparameterized case, as we explicitly highlight in the sequel.

[3] When $\boldsymbol{\beta}_\star$ is deterministic, $\mathbb{E}_{\tilde{x},\tilde{\epsilon},\boldsymbol{\beta}_\star}(\tilde{y} - \tilde{\boldsymbol{x}}^\top\hat{\boldsymbol{\beta}}_\lambda)^2$ reduces to the prediction risk for one fixed $\boldsymbol{\beta}_\star$.

passing algorithm [DM16], or the convex Gaussian min-max theorem[4][TAH18]. Under general data covariance, closely related to our work is [DW18], which considered a random effects model with isotropic prior on the target coefficients ($\boldsymbol{\Sigma}_\beta = \boldsymbol{I}_d$). Our risk characterization is built upon the general random matrix result of [RM11, LP11]. Similar tools have been applied in the analysis of sketching [LD19] and the connection between ridge regression and early stopping [AKT19, Lol20].

Weighted Regularization. The formulation (1.1) was first introduced in [HK70], and many choices of weighting matrix have been proposed [Str78, Cas80, MS05, MS18]; yet since these estimators are usually derived in the $n > p$ setup, their effectiveness in the high-dimensional and overparameterized regime is largely unknown. In semi-supervised linear regression, it is known that weighted matrix estimated from unlabeled data can improve the model performance [RC15, TCG20]. In deep learning, anisotropic Gaussian prior on the parameters enjoyed empirical success [LW17, ZTSG19]. Additionally, decoupled weight decay [LH17] and elastic weight consolidation [KPR+17] can both be interpreted as an ℓ_2 regularization weighted by an approximate Fisher information matrix [ZWXG18], which relates to the Fisher-Rao norm [LPRS17]. Finally, beyond the ℓ_2 penalty, weighted regularization is also effective in LASSO regression [Zou06, CWB08, BVDBS+15].

Benefit of Overparamterization. Our overparameterized setting is partially motivated by the double descent phenomenon [BHMM18], which can be theoretically explained in linear regression [AS17, HMRT19, BLLT19], random features regression [MM19, dRBK20, DL20], and max-margin classification [MRSY19, DKT19, HMX20], although translation to neural networks can be more nuanced [BES+20]. For least squares regression, it has been shown in special cases that overparameterization induces an implicit ℓ_2 regularization [KLS20, DLM19], which agrees with the absence of overfitting. This observation also leads to the speculation that the optimal ridge penalty in the overparameterized regime may be negative, to partially cancel out the implicit regularization. While the possibility of negative ridge parameter has been noted in [HG83, BS99], theoretical understanding of its benefit is largely missing, expect for heuristic argument (and empirical evidence) in [KLS20]. We provide a rigorous characterization of this "negative ridge" phenomenon.

Concurrent Works. Independent to our work, [RMR20] computed the asymptotic prediction risk under a similar extension of the isotropic assumption on $\boldsymbol{\Sigma}_\beta$, but did not consider the sign of λ_{opt}. We note that their result requires codiagonalizability of the covariances and certain functional relations between the eigenvalues, which is more restrictive than our setting. [TB20] provided a non-asymptotic analysis of ridge regression and constructed a specific spike model[5] in which negative regularization may lead to better generalization bound than interpolation ($\lambda = 0$). In a companion work [ABG+20], we connect properties of the ridgeless limit of the generalized ridge regression estimator to the implicit bias of preconditioned gradient descent (e.g., natural gradient descent), which allows us to decide the optimal preconditioner (for generalization) in the interpolation setting.

3 Setup and Assumptions

In addition to the prediction risk of the weighted ridge estimator $\hat{\boldsymbol{\beta}}_\lambda = (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{\Sigma}_w)^\dagger \boldsymbol{X}^\top \boldsymbol{y}$, the setup of which we outlined in Section 1, we also analyze the principal component regression (PCR) estimator: for $\theta \in [0, 1]$, the PCR estimator is given as $\hat{\boldsymbol{\beta}}_\theta = (\boldsymbol{X}_\theta^\top \boldsymbol{X}_\theta)^\dagger \boldsymbol{X}_\theta^\top \boldsymbol{y}$, where $\boldsymbol{X}_\theta = \boldsymbol{X}\boldsymbol{U}_\theta$ and the columns of $\boldsymbol{U}_\theta \in \mathbb{R}^{p \times \theta p}$ are the leading θp eigenvectors of $\boldsymbol{\Sigma}_x$.

Under the setup on $(\tilde{\boldsymbol{x}}, \boldsymbol{\beta}_\star, \tilde{\epsilon})$ described in Section 1, the prediction risk of (1.1) can be simplified as

$$\begin{aligned}\tilde{\mathbb{E}}\Big(\tilde{y} - \tilde{x}^\top \hat{\boldsymbol{\beta}}_\lambda\Big)^2 \;=\; & \underbrace{\tilde{\sigma}^2\bigg(1 + \frac{1}{n}\operatorname{tr}\Big(\boldsymbol{\Sigma}_{x/w}\Big(\boldsymbol{X}_{/w}^\top \boldsymbol{X}_{/w} + \lambda \boldsymbol{I}\Big)^{-1} - \lambda \boldsymbol{\Sigma}_{x/w}\Big(\boldsymbol{X}_{/w}^\top \boldsymbol{X}_{/w} + \lambda \boldsymbol{I}\Big)^{-2}\Big)\bigg)}_{\text{Part 1, Variance}} \\ & + \underbrace{\frac{\lambda^2}{n}\operatorname{tr}\Big(\boldsymbol{\Sigma}_{x/w}\Big(\boldsymbol{X}_{/w}^\top \boldsymbol{X}_{/w} + \lambda \boldsymbol{I}\Big)^{-1} \boldsymbol{\Sigma}_{w\beta}\Big(\boldsymbol{X}_{/w}^\top \boldsymbol{X}_{/w} + \lambda \boldsymbol{I}\Big)^{-1}\Big)}_{\text{Part 2, Bias}}, \end{aligned} \tag{3.1}$$

where $\boldsymbol{X}_{/w} = \boldsymbol{X}\boldsymbol{\Sigma}_w^{-1/2}$, $\boldsymbol{\Sigma}_{x/w} = \boldsymbol{\Sigma}_w^{-1/2}\boldsymbol{\Sigma}_x\boldsymbol{\Sigma}_w^{-1/2}$, $\boldsymbol{\Sigma}_{w\beta} = \boldsymbol{\Sigma}_w^{1/2}\boldsymbol{\Sigma}_\beta\boldsymbol{\Sigma}_w^{1/2}$. Note that the variance term does not depend on the true signal, and the bias is independent of the noise level. Let $\boldsymbol{d}_{x/w}$ be

[4] We remark that convergence and uniqueness of AMP and CGMT can be challenging to establish for $\lambda < 0$. Also, to our knowledge the current AMP framework does not handle the *joint* relation between $\boldsymbol{\Sigma}_x$ and $\boldsymbol{\Sigma}_\beta$.

[5] Our negative ridge construction relies on the general notion of *alignment*, which subsumes the spike model.

the eigenvalues of $\boldsymbol{\Sigma}_{x/w}$ and $\boldsymbol{\Sigma}_{x/w} = \boldsymbol{U}_{x/w}\boldsymbol{D}_{x/w}\boldsymbol{U}_{x/w}^\top$ be the eigendecomposition of $\boldsymbol{\Sigma}_{x/w}$, where $\boldsymbol{U}_{x/w}$ is the eigenvector matrix and $\boldsymbol{D}_{x/w} = \operatorname{diag}(\boldsymbol{d}_{x/w})$. Let $\boldsymbol{d}_{w\beta} \triangleq \operatorname{diag}\left(\boldsymbol{U}_{x/w}^\top \boldsymbol{\Sigma}_{w\beta} \boldsymbol{U}_{x/w}\right)$. When $\boldsymbol{\Sigma}_w = \boldsymbol{I}$, $\boldsymbol{d}_{w\beta}$ characterizes the strength of the signal $\boldsymbol{\beta}_\star$ along the directions of the eigenvectors of feature covariance $\boldsymbol{\Sigma}_x$. To simplify the RHS of (3.1), we make the following assumption:

Assumption 1. *Let $d_{x/w,i}$ and $d_{w\beta,i}$ be the ith entry of $\boldsymbol{d}_{x/w}$ and $\boldsymbol{d}_{w\beta}$, respectively. The empirical distribution of $(d_{x/w,i}, d_{w\beta,i})$ jointly converges to non-negative random variables (h, g). Further, $\min_i d_{x/w,i} \geq c_l$, $\max_i (d_{x/w,i}, d_{w\beta,i})) \leq c_u$ and $\|\boldsymbol{\Sigma}_{w\beta}\| \leq c_u$ for some $c_l, c_u > 0$ independent of p.*

One can check that $\boldsymbol{\Sigma}_x$ and $\boldsymbol{\Sigma}_\beta$ studied in [DW18, HMRT19, XH19] (with $\boldsymbol{\Sigma}_w = \boldsymbol{I}$) are special cases of Assumption 1 with either h or g being a point mass. Our Assumption 1 thus covers much more general settings of $\boldsymbol{\Sigma}_x$ and $\boldsymbol{\Sigma}_\beta$, which allows us to precisely analyze the *negative ridge* phenomenon.

4 Risk Characterization

With the aforementioned assumptions, we now present our characterization of the prediction risk.

Theorem 1. *Under Assumption 1, the asymptotic prediction risk is given as*

$$\tilde{\mathbb{E}}\left(\tilde{y} - \tilde{x}^\top \hat{\boldsymbol{\beta}}_\lambda\right)^2 \xrightarrow{p} \frac{m'(-\lambda)}{m^2(-\lambda)} \cdot \left(\gamma \mathbb{E} \frac{gh}{(h \cdot m(-\lambda) + 1)^2} + \tilde{\sigma}^2\right) := R(\lambda),\ \forall \lambda > -c_0 \quad (4.1)$$

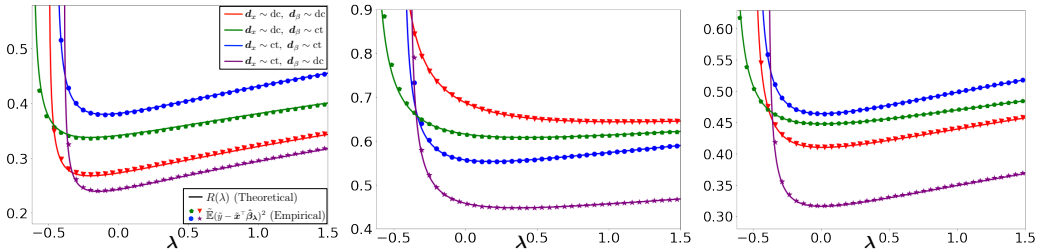
where $c_0 = (\sqrt{\gamma} - 1)^2 c_l$, and $m(z)$ is the Stieltjes transform of the limiting distribution of the eigenvalues of $\boldsymbol{X}_{/w}\boldsymbol{X}_{/w}^\top$. Additionally, $m(-\lambda), m'(-\lambda) > 0$ satisfy the self-consistent equations:

$$\lambda = \frac{1}{m(-\lambda)} - \gamma \mathbb{E} \frac{h}{1 + h \cdot m(-\lambda)} \quad (4.2)$$

$$1 = \left(\frac{1}{m^2(-\lambda)} - \gamma \mathbb{E} \frac{h^2}{(h \cdot m(-\lambda) + 1)^2}\right) m'(-\lambda). \quad (4.3)$$

Note that the condition $\lambda > -c_0$ ensures both $m(-\lambda)$ and $m'(-\lambda)$ exist and are positive. Furthermore, it can be shown from prior works [DW18, XH19] that the variance term (part 1) in (3.1), converges to $\tilde{\sigma}^2 \frac{m'(-\lambda)}{m^2(-\lambda)}$. Our main contribution is to characterize the bias term, Part 2, under significantly less restrictive assumption on $(\boldsymbol{\Sigma}_x, \boldsymbol{\Sigma}_\beta, \boldsymbol{\Sigma}_w)$. In particular, we show that

$$\text{Part 2} \xrightarrow{\text{p}} \frac{m'(-\lambda)}{m^2(-\lambda)} \cdot \gamma \mathbb{E} \frac{gh}{(h \cdot m(-\lambda) + 1)^2}, \quad \forall \lambda > -c_0.$$



(a) Aligned, noiseless (b) Misaligned, noiseless (c) Random, noiseless

Figure 2: Finite sample prediction risk $\tilde{\mathbb{E}}(\tilde{y} - \tilde{\boldsymbol{x}}^\top \boldsymbol{\beta}_\star)^2$ (experiment) and the asymptotic risk $R(\lambda)$ (theory) against λ for standard ridge regression ($\boldsymbol{\Sigma}_w = \boldsymbol{I}_d$). We set $\gamma = 2$ and $(n, p) = (300, 600)$. 'dc' and 'ct' stand for for discrete and continuous distribution, respectively. We write 'aligned' if $\boldsymbol{d}_x$ and $\boldsymbol{d}_\beta$ have the same order, 'misaligned' for the reverse, and 'random' for random order. Colors indicate different combinations of $\boldsymbol{d}_x$ and $\boldsymbol{d}_\beta$. Note that our derived risk $R(\lambda)$ matches the experimental values, and in the aligned and noiseless case, the optimal risk is achieved when $\lambda < 0$ (predicted by Theorem 4). The noisy case is presented in Appendix D.

We illustrate the results of Theorem 1 in Figure 2 (noiseless case) and Figure 8 (noisy case) for both discrete and continuous design for $\boldsymbol{d}_x$ and $\boldsymbol{d}_\beta$ with $\boldsymbol{\Sigma}_x = \operatorname{diag}(\boldsymbol{d}_x), \boldsymbol{\Sigma}_\beta = \operatorname{diag}(\boldsymbol{d}_\beta)$ and $\boldsymbol{\Sigma}_w = \boldsymbol{I}$ (see design details in Appendix D). Note that Assumption 1 specifies a joint relation between $\boldsymbol{d}_x (= \boldsymbol{d}_{x/w})$ and $\boldsymbol{d}_\beta (= \boldsymbol{d}_{w\beta})$. In the following section, we mainly consider the three following relations, which allow us to precisely determine the sign of λ_{opt}.

Definition 2. *For two vectors $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^p$, we say $\boldsymbol{a}$ is aligned (misaligned) with $\boldsymbol{b}$ if the order of $\boldsymbol{a}$ is the same as (reverse of) the order of $\boldsymbol{b}$, i.e., $a_i \geq a_j$ iff $b_i \geq (\leq)\ b_j$ for all i, j. Additionally, we say $\boldsymbol{a}$ and $\boldsymbol{b}$ have random relation if given the order of one vector, the order of the other is uniformly permuted at random.*

Intuitively, aligned $\boldsymbol{d}_x$ and $\boldsymbol{d}_\beta$ implies that when one component in $\boldsymbol{d}_x$ has large magnitude, then so does the corresponding component in $\boldsymbol{d}_\beta$, and vice versa (see Figure 3). In Figure 2, we plot the prediction risk of all three joint relations defined above (see Appendix D for details).

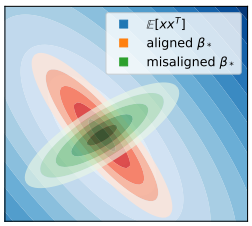


Figure 3: Alignment between $\boldsymbol{x}$ and $\boldsymbol{\beta}_*$ in 2D.

Theorem 1 allows us to compute the risk of the generalized ridge estimator $\hat{\boldsymbol{\beta}}_\lambda$ and also its ridgeless limit, i.e., the minimum $\|\hat{\boldsymbol{\beta}}\|_{\boldsymbol{\Sigma}_w}$ norm interpolant (taking $\boldsymbol{\Sigma}_w = \boldsymbol{I}$ yields the min ℓ_2 norm solution).

Connection to PCR estimator. Note that the principal component regression (PCR) estimator is closely related to the ridgeless estimator in the following sense: picking the leading θp eigenvectors of $\boldsymbol{\Sigma}_x$ (for some $\theta \in [0, 1]$) is equivalent to setting the remaining $(1-\theta)p$ eigenvalues of $\boldsymbol{\Sigma}_w$ to be infinity [HG83]. The following corollary characterizes the prediction risk of the PCR estimator $\hat{\beta}_\theta$:

Corollary 3. *Given Assumption 1 and $\boldsymbol{\Sigma}_w = \boldsymbol{I}$, and h has continuous and strictly increasing quantile function Q_h. Then for all $\theta \in (0, 1]$, as $n, p \to \infty$,*

$$\tilde{\mathbb{E}}\Big(\tilde{y} - \tilde{x}\hat{\boldsymbol{\beta}}_\theta\Big)^2 \xrightarrow{p} \begin{cases} \dfrac{m_\theta'(0)}{m_\theta^2(0)} \cdot \left(\gamma \mathbb{E}\dfrac{gh}{(h_\theta \cdot m_\theta(0) + 1)^2} + \tilde{\sigma}^2\right), & \theta\gamma > 1 \\ \left(\gamma \mathbb{E}[gh \cdot \mathbb{I}_{h < Q_h(1-\theta)}] + \tilde{\sigma}^2\right) \dfrac{1}{1 - \theta\gamma}, & \theta\gamma < 1 \end{cases} \tag{4.4}$$

where $h_\theta = h \cdot \mathbb{I}_{h \geq Q_h(1-\theta)}$ and $m_\theta(z)$ satisfies $-z = m_\theta^{-1}(z) - \gamma \mathbb{E} h_\theta \cdot (1 + h_\theta \cdot m_\theta(z))^{-1}$. In addition, if $\mathbb{E}[g|h]$ is a decreasing function of h, and h has continuous p.d.f., then the asymptotic prediction risk of $\hat{\boldsymbol{\beta}}_\theta$ is a decreasing function of θ when $\theta\gamma > 1$.

Corollary 3 confirms double descent under more general settings of $(\boldsymbol{\Sigma}_x, \boldsymbol{\Sigma}_\beta)$ than [XH19], i.e. the risk exhibits a spike as $\theta\gamma \to 1^-$, and then decreases as we further overparameterize by increasing θ. In Section 6 we compare the PCR estimator $\hat{\boldsymbol{\beta}}_\theta$ with the minimum $\|\hat{\boldsymbol{\beta}}\|_{\boldsymbol{\Sigma}_w}$ norm solution.

Remark. *The PCR estimator [XH19] and the ridgeless regression estimator (considered in [HMRT19]) are fundamentally different in the following way: in ridgeless regression, increasing the model size corresponds to changing γ, which also alters the dimensions of $\boldsymbol{\beta}_\star$; in contrast, in PCR, increasing θ does not change the data generating process (which is a more natural setting).*

In terms of the risk curve, Figure 9(a) shows that the ridgeless regression estimator can exhibit "multiple descent" as γ increases, whereas Corollary 3 and Figure 9(b) demonstrate that in the misaligned case, the PCR risk is monotonically decreasing in the overparameterized regime $\theta\gamma > 1$.

5 Analysis of Optimal λ_{opt}

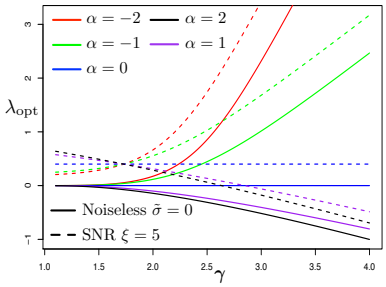
In this section, we focus on the optimal weighted ridge estimator and determine the sign of the optimal regularization parameter λ_{opt}. Taking the derivatives of (4.1) yields

$$R'(\lambda) \cdot \frac{m^3(-\lambda)}{2\gamma(m'(-\lambda))^2} = \underbrace{\left(-\tilde{\sigma}^2 \frac{\mathbb{E}\frac{\zeta^2}{(1+\zeta)^3}}{1 - \gamma\mathbb{E}\frac{\zeta^2}{(1+\zeta)^2}}\right)}_{\text{Part 3}} + \underbrace{\left(\mathbb{E}\frac{gh\zeta}{(1+\zeta)^3} - \frac{\gamma\mathbb{E}\frac{\zeta^2}{(1+\zeta)^3}\mathbb{E}\frac{gh}{(1+\zeta)^2}}{1 - \gamma\mathbb{E}\frac{\zeta^2}{(1+\zeta)^2}}\right)}_{\text{Part 4}}, \tag{5.1}$$

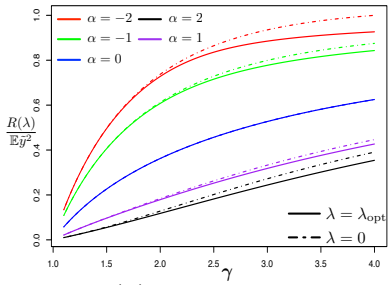
where $\zeta = h \cdot m(-\lambda)$. For certain special cases, we obtain a closed form solution for λ_{opt} (see details in Appendix B.1) and recover the result from [HMRT19, DW18][6] and beyond:

- When $h \overset{\text{a.s.}}{=} c$ (i.e., isotropic features [HMRT19]), the optimal λ_{opt} is achieved at c/ξ.
- When $g \overset{\text{a.s.}}{=} c$ (i.e., isotropic signals [DW18]), the optimal λ_{opt} is achieved at $\tilde{\sigma}^2/c$.
- When $\mathbb{E}[g|h] \overset{\text{a.s.}}{=} \mathbb{E}[g]$ (e.g., random order), the optimal λ_{opt} is achieved at $\tilde{\sigma}^2/\mathbb{E}[g]$.

[6] In [HMRT19], $h \overset{\text{a.s.}}{=} 1$. In [DW18], $\tilde{\sigma}^2 = 1$ and their signal strength α^2 is equivalent to $c\gamma$ in our setting.



Left: Optimal Ridge λ_{opt}.



Right: $R(\lambda)$: Optimal vs. ridgeless.

Figure 4: We set $\boldsymbol{\Sigma}_w = \boldsymbol{I}$ and $\boldsymbol{\Sigma}_\beta = \boldsymbol{\Sigma}_x^\alpha$ where $\boldsymbol{d}_x$ has two point masses on 1 and 5 with probability $3/4$ and $1/4$ respectively. **Left**: optimal λ; solid lines represents the noiseless case $\tilde{\sigma} = 0$ and dashed lines represents SNR $\xi = 5$. **Right**: prediction risk of the ridgeless ($R(\lambda_{\text{opt}})$, dashed lines) and optimally regularized ($R(\lambda_{\text{opt}})$, solid lines) estimator in the noiseless case. We normalize the prediction risk as $\mathbb{E}\tilde{y}^2 = \mathbb{E}(\tilde{\boldsymbol{x}}^\top \boldsymbol{\beta}_\star)^2$.

Although λ_{opt} may not have a tractable form in general, we may infer the sign of λ_{opt}. Note that in (5.1), Part 3 is due to the variance term (Part 1) and Part 4 from the bias term (Part 2) in (3.1). We therefore consider the sign of Part 3 and Part 4 separately in the following theorem.

Theorem 4. *Under Assumption 1, we have*

- *Part 3 (derivative of variance) is negative for all $\lambda > -c_0$.*
- *If $\mathbb{E}[g|h]$ is an increasing function of h on its support, then Part 4 (derivative of bias) is positive for all $\lambda > 0$. At $\lambda = 0$, Part 4 is non-negative and achieves 0 only if $\mathbb{E}[g|h] \overset{a.s.}{=} \mathbb{E}[g]$.*
- *If $\mathbb{E}[g|h]$ is a decreasing function of h on its support, then Part 4 is negative for all $\lambda \in (-c_0, 0)$. At $\lambda = 0$, Part 4 is non-positive and achieves 0 only if $\mathbb{E}[g|h] \overset{a.s.}{=} \mathbb{E}[g]$.*

The first point in Theorem 4 is consistent with the well-understood variance reduction property of ridge regularization. On the other hand, when the prediction risk is dominated by the bias term (i.e., $\tilde{\sigma}^2 = o(1)$) and both $\boldsymbol{d}_{x/w}$ and $\boldsymbol{d}_{w\beta}$ converge to non-trivial distributions, the second and third point of Theorem 4 reveal the following surprising phenomena (see Figure 2 (a) and (b)):

M1 $\lambda_{\text{opt}} < 0$ when $\boldsymbol{d}_{x/w}$ aligns with $\boldsymbol{d}_{w\beta}$, or in general, $\mathbb{E}[g|h]$ is a strictly increasing function of h. In the context of standard ridge regression, it means that shrinkage regularization only *increases the bias* in the overparameterized regime when features are informative, i.e., the projection of the signal is large in the directions where the feature variance is large.

M2 $\lambda_{\text{opt}} > 0$ when $\boldsymbol{d}_{x/w}$ is misaligned with $\boldsymbol{d}_{w\beta}$, or in general, $\mathbb{E}[g|h]$ is a strictly decreasing function of h. This is to say, in standard ridge regression, when features are not informative, i.e., the projection of the signal is small in the directions of large feature variance, shrinkage is beneficial even in the *absence of label noise* (the variance term is zero).

M1 and M2, together with aforementioned special case when g and h have random relation, provide a precise characterization of the sign of λ_{opt}. In particular, M1 confirms the "negative ridge" phenomenon empirically observed in [KLS20] and outlines concise conditions under which it occurs. We emphasize that neither M1 nor M2 would be observed when one of $\boldsymbol{\Sigma}_{x/w}$ and $\boldsymbol{\Sigma}_{w\beta}$ is identity. In other words, these observations arise from our more general assumption on $(\boldsymbol{\Sigma}_{x/w}, \boldsymbol{\Sigma}_{w\beta})$.

Implicit regularization of overparameterization. Taking both the bias and variance into account, Theorem 4 suggests a bias-variance tradeoff between Part 3 and Part 4, and λ_{opt} will eventually become positive as $\tilde{\sigma}^2$ increases (i.e., risk is dominated by variance, for which positive λ is beneficial). For certain special cases, we can provide a lower bound for the transition from $\lambda_{\text{opt}} < 0$ to $\lambda_{\text{opt}} > 0$.

Proposition 5. *Given Assumption 1, let $(h, g) = (1, 1)$ with probability $1 - q$ and $(h, g) = (h_1, g_1)$ with probability q, where $h_1 > 1$ and $g_1 > 1$. Denote $\bar{\gamma} = \gamma - 1$. Then $\lambda_{opt} < 0$ if*

$$\tilde{\sigma}^2 < (h_1 - 1)(g_1 - 1)h_1 \cdot \max\left(\frac{(\gamma q - 1)^3 \bar{\gamma}^3 (1 - q)}{(1 - q)\gamma^2(\bar{\gamma}^3 q^2 + (\gamma q - 1)^3 h_1^2)}, \frac{\gamma q (1 - q)\bar{\gamma}^3}{(1 - q)(h_1 + \bar{\gamma})^3 + q h_1^2 \gamma^3} \right).$$

As q approaches 0 or 1, the above upper bound goes to 0 because $\boldsymbol{\Sigma}_x$, $\boldsymbol{\Sigma}_\beta$ becomes closer to $\boldsymbol{I}$. Otherwise, when $\gamma q > 1$, the upper bound suggests $\tilde{\sigma}^2 = O(g_1 \gamma)$ which implies the SNR $\xi = \Omega(h_1/\gamma)$.

Hence, as γ increases, λ_{opt} remains negative for a lower SNR, which coincides with the intuition that overparameterization has an implicit effect of ℓ_2 regularization (Figure 4 Left). Indeed, the following proposition suggests such implicit regularization is only present in the overparameterized regime:

Proposition 6. *When $\gamma < 1$, λ_{opt} on $(-c_0, \infty)$ is always non-negative under Assumption 1.*

In Figure 4 we confirm our findings in Theorem 4 (for additional results on different distributions see Figure 10). Specifically, we set $\boldsymbol{\Sigma}_w = \boldsymbol{I}, \boldsymbol{\Sigma}_x = \text{diag}(\boldsymbol{d}_x)$ and $\boldsymbol{\Sigma}_\beta = \boldsymbol{\Sigma}_x^\alpha$. As we increase α from negative to positive, the relation between $\boldsymbol{d}_x$ and $\boldsymbol{d}_\beta$ transitions from misaligned to aligned. The left panel shows that the sign of λ_{opt} is the exact opposite to the sign of α in the noiseless case (i.e. the variance is 0), which is consistent with M1 and M2. Moreover, when $\boldsymbol{d}_x$ aligns with $\boldsymbol{d}_\beta$, λ_{opt} decreases as γ becomes larger, which agrees with our observation on the implicit ℓ_2 regularization of overparameterization. Last but not least, in Figure 4 (Right) we see that the optimal ridge regression estimator leads to considerable improvement over the ridgeless estimator. We comment that this improvement becomes more significant as γ or condition number of $\boldsymbol{\Sigma}_x$ and $\boldsymbol{\Sigma}_\beta$ increases.

Risk monotonicity of optimal ridge regression. [DS20, Proposition 6] showed that for isotropic data ($\boldsymbol{\Sigma}_x = \boldsymbol{I}$), the asymptotic prediction risk of optimally-tuned ridge regression monotonically increases with γ. This is to say, under proper regularization, more training data always helps the test performance [KH92]. Here we extend this result to data with general covariance and isotropic $\boldsymbol{\beta}_*$.

Proposition 7. *Given $\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top] = \boldsymbol{\Sigma}_x$ satisfying Assumption 1 and $\mathbb{E}[\boldsymbol{\beta}_*\boldsymbol{\beta}_*^\top] = \frac{c}{p}\boldsymbol{I}$[7], the asymptotic prediction risk of the optimally-tuned ridge regression estimator (i.e., $\boldsymbol{\Sigma}_w = \boldsymbol{I}$) with $\lambda_{\text{opt}} = \gamma\tilde{\sigma}^2/c$ is an increasing function of $\gamma \in (0, \infty)$.*

As shown in Figure 5 (where $\boldsymbol{d}_x$ has 3 point masses and $\boldsymbol{d}_\beta = 1$), ℓ_2 regularization can suppress "multiple descent", and the risk of the optimally-tuned ridge estimator (purple) is monotone w.r.t. γ. We remark that establishing such characterization under general orientation of $\boldsymbol{\beta}_*$ (anisotropic $\boldsymbol{\Sigma}_\beta$) can be challenging, because the optimal regularization λ_{opt} may not have a convenient closed-form. We leave the analysis for general $\boldsymbol{\Sigma}_\beta$ as future work.

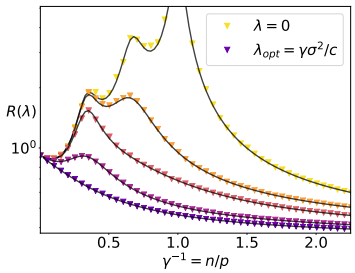


Figure 5: Impact of ridge regularization on the risk curve (SNR=3). Darker color corresponds to larger λ.

6 Optimal Weighting Matrix

Having characterized the optimal regularization strength, we now turn to the optimal weighting matrix $\boldsymbol{\Sigma}_w$. Toward this goal, we additionally require the following assumptions on $(\boldsymbol{\Sigma}_x, \boldsymbol{\Sigma}_\beta, \boldsymbol{\Sigma}_w)$:

Assumption 2. *The covariance matrix $\boldsymbol{\Sigma}_x$ and the weighting matrix $\boldsymbol{\Sigma}_w$ share the same set of eigenvectors, i.e., we have the following eigendecompositions: $\boldsymbol{\Sigma}_x = \boldsymbol{U}\boldsymbol{D}_x\boldsymbol{U}^\top$ and $\boldsymbol{\Sigma}_w = \boldsymbol{U}\boldsymbol{D}_w\boldsymbol{U}^\top$, where $\boldsymbol{U} \in \mathbb{R}^{p\times p}$ is orthogonal, and $\boldsymbol{D}_x = \text{diag}(\boldsymbol{d}_x), \boldsymbol{D}_w = \text{diag}(\boldsymbol{d}_w)$.*

Define $\bar{\boldsymbol{d}}_\beta = \text{diag}(\boldsymbol{U}^\top\boldsymbol{\Sigma}_\beta\boldsymbol{U})$. Note that when $\boldsymbol{\Sigma}_\beta$ also shares the same eigenvector matrix $\boldsymbol{U}$, then $\bar{\boldsymbol{d}}_\beta = \boldsymbol{d}_\beta$, which is simply the eigenvalues of $\boldsymbol{\Sigma}_\beta$.

Assumption 3. *Let $d_{x,i}, \bar{d}_{\beta,i}, d_{w,i}$ be the ith element of $\boldsymbol{d}_x, \bar{\boldsymbol{d}}_\beta, \boldsymbol{d}_w$ respectively. We assume that the empirical distribution of $(d_{x,i}, \bar{d}_{\beta,i}, d_{w,i})$ jointly converges to $(s, v, s/r)$, where s, v, r are non-negative random variables. Further, there exists constants $c_l, c_u > 0$ independent of n and p such that $\min_i(\min(d_{x,i}, \bar{d}_{\beta,i}, d_{w,i})) \geq c_l$, $\max_i(\max(d_{x,i}, \bar{d}_{\beta,i}, d_{w,i})) \leq c_u$ and $\|\boldsymbol{\Sigma}_\beta\| \leq c_u$.*

For notational convenience, we define $\mathcal{H}_w$ and $\mathcal{H}_r$ to be the sets of all $\boldsymbol{\Sigma}_w$ and r, respectively, that satisfy Assumption 2 and Assumption 3. Additionally, let $\mathcal{S}_w$ and $\mathcal{S}_r$ be the subset of $\mathcal{H}_w$ and $\mathcal{H}_r$ such that $r = f(s)$ for some function f (this represents $\boldsymbol{\Sigma}_w \in \mathcal{H}_w$ that only depends on $\boldsymbol{\Sigma}_x$ but not $\boldsymbol{\Sigma}_\beta$). By Assumption 2 and 3, the empirical distribution of $(d_{x/w,i}, d_{w\beta,i})$ jointly converges to $(r, sv/r)$ and satisfies the boundedness requirement in Assumption 1. Thus by Theorem 1 we have:

$$R(r, \lambda) \triangleq \frac{m_r'(-\lambda)}{m_r^2(-\lambda)} \cdot \left(\gamma\mathbb{E}\frac{sv}{(r \cdot m_r(-\lambda) + 1)^2} + \tilde{\sigma}^2\right), \tag{6.1}$$

where $m_r(-\lambda)$ satisfies the equation $\lambda = m_r^{-1}(-\lambda) - \gamma\mathbb{E}(1 + r \cdot m_r(-\lambda))^{-1}r$. It is clear that when $r \overset{\text{a.s.}}{=} s$, (6.1) reduces to the standard ridge regression with $\boldsymbol{\Sigma}_w = \boldsymbol{I}$, and for $r \overset{\text{a.s.}}{=} 1$, the equation

[7] Note that the parameter scaling differs from the previous setting by γ to be consistent with that of [DS20].

reduces to the cases of isotropic features ($\boldsymbol{\Sigma}_w = \boldsymbol{\Sigma}_x$). Note that (6.1) indicates that the impact of $\boldsymbol{\Sigma}_\beta$ on the risk is fully captured by $\bar{\boldsymbol{d}}_\beta$. Hence we define $\bar{\boldsymbol{\Sigma}}_\beta = \boldsymbol{U}\,\mathrm{diag}(\bar{\boldsymbol{d}}_\beta)\boldsymbol{U}^\top$, which corresponds to $r \overset{\text{a.s.}}{=} sv$, and is equivalent to $\boldsymbol{\Sigma}_\beta$ when $\boldsymbol{\Sigma}_\beta$ also shares the same eigenvector matrix $\boldsymbol{U}$. In the following subsections, we discuss the optimal $\boldsymbol{\Sigma}_w$ for two types of estimator: the minimum $\|\hat{\beta}\|_{\boldsymbol{\Sigma}_w}$ solution (taking $\lambda \to 0$), and the optimally weighted ridge estimator ($\lambda = \lambda_{\text{opt}}$). Note that the risk for both estimators is scale-invariant over $\boldsymbol{\Sigma}_w$ and r. Hence, when we define a specific choice of $(\boldsymbol{\Sigma}_w, r)$, we simultaneously consider all pairs $(c\boldsymbol{\Sigma}_w, r/c)$ for $c > 0$. Finally, we note that the choice of $r \overset{\text{a.s.}}{=} s \cdot \mathbb{E}[v|s] \in \mathcal{S}_r$ plays a key role in our analysis, and its corresponding choice of $\boldsymbol{\Sigma}_w$ is given as $\boldsymbol{\Sigma}_w = (f_v(\boldsymbol{\Sigma}_x))^{-1}$, where $f_v(s) \triangleq \mathbb{E}[v|s]$ and f_v applies to the eigenvalues of $\boldsymbol{\Sigma}_x$.

6.1 Minimum $\|\hat{\beta}\|_{\boldsymbol{\Sigma}_w}$ solution

Taking the ridgeless limit leads to the following bias-variance decomposition of the prediction risk,

$$\text{Bias:}\quad \mathrm{R_b}(r) \triangleq \frac{m_r'(0)}{m_r^2(0)} \cdot \gamma \mathbb{E}\frac{sv}{(r \cdot m_r(0) + 1)^2} \qquad \text{Variance:}\quad \mathrm{R_v}(r) \triangleq \frac{m_r'(0)}{m_r^2(0)} \cdot \tilde{\sigma}^2.$$

In the previous sections we observe a bias-variance tradeoff in choosing the optimal λ. Interesting, the following theorem illustrates a similar bias-variance tradeoff in choosing the optimal $\boldsymbol{\Sigma}_w$:

Theorem 8. *Given Assumptions 2 and 3,*

- $r \overset{\text{a.s.}}{=} sv$ *(i.e.,* $\boldsymbol{\Sigma}_w = \bar{\boldsymbol{\Sigma}}_\beta^{-1}$*) is the optimal choice in* $\mathcal{H}_r$ *that minimizes the bias function* $R_b(r)$*. Additionally,* $r \overset{\text{a.s.}}{=} \mathbb{E}[v|s] \cdot s$ *(i.e.,* $\boldsymbol{\Sigma}_w = (f_v(\boldsymbol{\Sigma}_x))^{-1}$*) is the optimal in* $\mathcal{S}_r$ *that minimizes* $R_b(r)$*.*
- $r \overset{\text{a.s.}}{=} 1$ *(i.e.,* $\boldsymbol{\Sigma}_w = \boldsymbol{\Sigma}_x$*) is optimal in both* $\mathcal{S}_r$ *and* $\mathcal{H}_r$ *that minimizes the variance function* $R_v(r)$*.*

Theorem 8 implies that the variance is minimized when $\boldsymbol{\Sigma}_w = \boldsymbol{\Sigma}_x$. Since the variance term does not depend on $\boldsymbol{\beta}_\star$, it is not surprising that the optimal $\boldsymbol{\Sigma}_w$ is also independent of $\boldsymbol{\Sigma}_\beta$. Furthermore, this result is consistent with the intuition that to minimize the variance, $\hat{\beta}_\lambda$ should be penalized more in the higher variance directions of $\boldsymbol{\Sigma}_x$, and vice versa. On the other hand, Theorem 8 also implies that the bias is minimized when $\boldsymbol{d}_w = 1/\bar{\boldsymbol{d}}_\beta$ which does not depend on $\boldsymbol{d}_x$. While this characterization may not be intuitive, when $\bar{\boldsymbol{d}}_\beta = \boldsymbol{d}_\beta$ (i.e., $\boldsymbol{\Sigma}_\beta$ also shares the same eigenvector matrix $\boldsymbol{U}$), one analogy is that since the quadratic regularization corresponds to the a Gaussian prior $\mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_w^{-1})$, it is reasonable to match $\boldsymbol{\Sigma}_w^{-1}$ with the covariance of $\boldsymbol{\beta}_\star$, which gives the maximum a posteriori (MAP) estimate. In general, the optimal $\boldsymbol{\Sigma}_w$ admits a bias-variance tradeoff (i.e., the bias and variance are optimal under different $\boldsymbol{\Sigma}_w$) except for the special case of $\boldsymbol{\Sigma}_x\bar{\boldsymbol{\Sigma}}_\beta = \boldsymbol{I}$.

Additionally, the following proposition demonstrates the advantage of the minimum $\|\hat{\boldsymbol{\beta}}\|_{\boldsymbol{\Sigma}_w}$ solution over the PCR estimator in the noiseless case.

Proposition 9. *Given Assumption 2 and 3 and* $\tilde{\sigma} = 0$*, suppose* s *and* $\mathbb{E}[v|s] \cdot s$ *both have continuous and strictly increasing quantile functions. Then the minimum* $\|\hat{\boldsymbol{\beta}}\|_{\boldsymbol{\Sigma}_w}$ *solution outperforms the PCR estimator for all* $\theta \in [0, 1)$ *when* $\boldsymbol{\Sigma}_w = \bar{\boldsymbol{\Sigma}}_\beta^{-1} \in \mathcal{H}_w$*, or when* $\boldsymbol{\Sigma}_w = (f_v(\boldsymbol{\Sigma}_x))^{-1} \in \mathcal{S}_w$*.*

6.2 Optimal weighted ridge estimator

Finally, we consider the optimally-tuned weighted shrinkage and discuss the optimal choice of $\boldsymbol{\Sigma}_w$.

Theorem 10. *Suppose Assumptions 2 and 3 hold. Then* $r \overset{\text{a.s.}}{=} sv$ *(i.e.,* $\boldsymbol{\Sigma}_w = \bar{\boldsymbol{\Sigma}}_\beta^{-1}$*) is the optimal solution in* $\mathcal{H}_r$ *that minimizes* $\min_\lambda R(r, \lambda)$*. Additionally,* $r \overset{\text{a.s.}}{=} \mathbb{E}[v|s] \cdot s$ *(i.e.,* $\boldsymbol{\Sigma}_w = (f_v(\boldsymbol{\Sigma}_x))^{-1}$*) is the optimal solution in* $\mathcal{S}_r$ *that minimizes* $\min_\lambda R(r, \lambda)$*.*

In contrast to the ridgeless setting in Theorem 8, the optimal $\boldsymbol{d}_w$ for general λ_{opt} does not depend on the noise level but only on $\bar{\boldsymbol{d}}_\beta$, the strength of the signal in the directions of the eigenvectors of $\boldsymbol{\Sigma}_x$. We conjecture that this is because in the optimally weighted estimator, λ_{opt} is capable of balancing the bias-variance tradeoff; therefore the weighting matrix may not need to adjust to the label noise and can be chosen solely based on the signal $\boldsymbol{\beta}_\star$. Indeed, as previously discussed, $\boldsymbol{\Sigma}_w = \boldsymbol{\Sigma}_\beta^{-1}$ is a preferable choice of prior under the Bayesian perspective when $\boldsymbol{d}_\beta = \bar{\boldsymbol{d}}_\beta$.

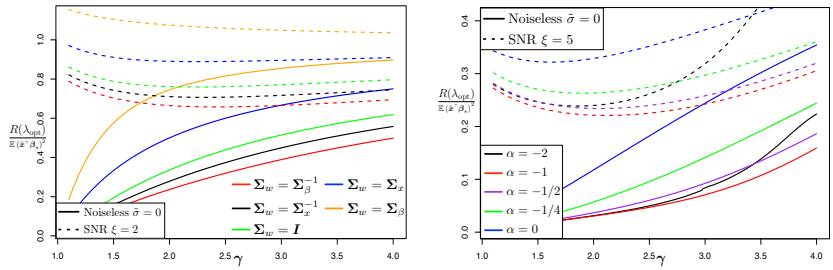


Figure 6: $R(\lambda_{\text{opt}})/\mathbb{E}(\tilde{\boldsymbol{x}}^\top \boldsymbol{\beta}_\star)^2$ against γ for various weighting matrix $\boldsymbol{\Sigma}_w$. Solid lines represent the noiseless case $\tilde{\sigma} = 0$ and the dashed lines represent the noisy case with fixed SNR ξ. We set $\boldsymbol{d}_x$ to be aligned with $\boldsymbol{d}_\beta$ and **Left**: $\boldsymbol{d}_x$ to have 4 point masses $(1, 2, 3, 4)$ with equal probabilities and $\boldsymbol{d}_\beta$ with 2 point masses on 1 and 5 with probabilities $3/4$ and $1/4$, respectively; **Right**: $\boldsymbol{d}_x$ has 2 point masses on 1 and 5 with probabilities $3/4$ and $1/4$, respectively, and $\boldsymbol{\Sigma}_\beta = \boldsymbol{\Sigma}_x^2$; we set $\boldsymbol{\Sigma}_w = \boldsymbol{\Sigma}_\beta^\alpha$.

Theorem 10 is supported by Figure 6, where we plot the prediction risk of the generalized ridge regression estimator under different $\boldsymbol{\Sigma}_w$ and optimally tuned λ_{opt}. We consider a simple discrete construction for aligned $\boldsymbol{d}_x$ and $\boldsymbol{d}_\beta (= \bar{\boldsymbol{d}}_\beta)$. On the left panel, we enumerate a few standard choices of $\boldsymbol{\Sigma}_w$: $\boldsymbol{\Sigma}_x, \boldsymbol{\Sigma}_\beta, \boldsymbol{I}, \boldsymbol{\Sigma}_x^{-1}$ and the optimal choice $\boldsymbol{\Sigma}_\beta^{-1}$ (red). On the right, we take $\boldsymbol{\Sigma}_w$ to be powers of $\boldsymbol{\Sigma}_\beta$ around the optimal $\boldsymbol{\Sigma}_\beta^{-1}$. In both setups, we confirm that $\boldsymbol{\Sigma}_\beta^{-1}$ achieves the lowest risk uniformly over γ, as predicted by Theorem 10.

Note that our main results require knowledge of $\boldsymbol{\Sigma}_x$ and $\bar{\boldsymbol{\Sigma}}_\beta$. While $\boldsymbol{\Sigma}_x$ can be obtained in a semi-supervised setting using unlabeled data (e.g., [RC15, TCG20]), it is typically difficult to estimate $\bar{\boldsymbol{\Sigma}}_\beta$ directly from data. Without prior knowledge on $\bar{\boldsymbol{\Sigma}}_\beta$, Theorem 10 suggests that $r \overset{\text{a.s.}}{=} \mathbb{E}[v|s] \cdot s$ is the optimal r that only depends on s. That is, $\boldsymbol{\Sigma}_w = (f_v(\boldsymbol{\Sigma}_x))^{-1}$ is the optimal $\boldsymbol{\Sigma}_w$ that only depends on $\boldsymbol{\Sigma}_x$. In the special case of $\mathbb{E}[v|s] = \mathbb{E}[v]$, standard ridge regression ($\boldsymbol{\Sigma}_w = \boldsymbol{I}$) is optimal in $\mathcal{S}_w$. When the exact form of $f_v(s)$ is also not known, we may use a polynomial or power function of s to approximate either $f_v(s)$ or $1/f_v(s)$, whose coefficients can be considered as hyper-parameters to be cross-validated. We demonstrate the effectiveness of this heuristic in Figure 7: although our proposed $\boldsymbol{\Sigma}_w = f_v(\boldsymbol{\Sigma}_x)^{-1}$ (blue) is worse than the actual optimal (red) $\boldsymbol{\Sigma}_w = \boldsymbol{\Sigma}_\beta^{-1}$ (same as $\bar{\boldsymbol{\Sigma}}_\beta^{-1}$ due to diagonal design), it is the best choice among weighting matrices that only depend on $\boldsymbol{\Sigma}_x$. In addition, we seek the best approximation of $f_v(s)$ by applying a power transformation on $\boldsymbol{\Sigma}_x$, and we observe that certain powers of $\boldsymbol{\Sigma}_x$ also outperform the standard isotropic regularization.

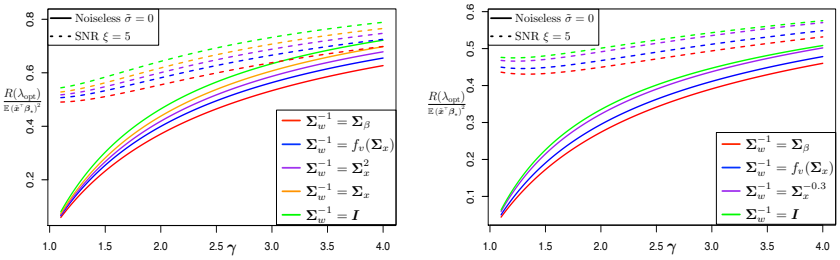


Figure 7: $R(\lambda_{\text{opt}})/\mathbb{E}(\tilde{\boldsymbol{x}}^\top \boldsymbol{\beta}_\star)^2$ against γ for various weighting matrix $\boldsymbol{\Sigma}_w$ under noiseless $\tilde{\sigma} = 0$ (solid lines) and noisy setting with fixed SNR ξ (dashed lines). **Left**: We set $f_v(s)$ as an increasing function of s on its support; **Right**: We set $f_v(s)$ as a decreasing function of s on its support. Note that the heuristically chosen weighting matrices often outperform the standard ridge regression estimator (green).

7 Conclusion

We provide a precise asymptotic characterization of the prediction risk of generalized ridge regression in the overparameterized regime. Our result greatly generalizes previous high-dimensional analysis of ridge regression, which enables us to discover and theoretically justify various interesting findings, including the negative ridge phenomenon, the implicit regularization of overparameterization, and a concise description of the optimal weighted shrinkage. Future works include extending our analysis to border settings, such as more general eigenvalue conditions [XH19] or the random features regression model [MM19]. Another important direction is to construct weighting matrix $\boldsymbol{\Sigma}_w$ solely from training data that outperforms isotropic shrinkage in the overparameterized regime.

8 Broader Impact

This work does not present any foreseeable direct societal consequence.

Acknowledgement

The authors would like to thank Murat A. Erdogdu, Daniel Hsu and Taiji Suzuki for comments and suggestions, and also anonymous NeurIPS reviewers 1 and 3 for helpful feedback. DW was partially funded by CIFAR, NSERC and LG Electronics. JX was supported by a Cheung-Kong Graduate School of Business Fellowship.

References

[ABG+20] Shun-ichi Amari, Jimmy Ba, Roger Grosse, Xuechen Li, Atsushi Nitanda, Taiji Suzuki, Denny Wu, and Ji Xu, *When does preconditioning help or hurt generalization?*, arXiv preprint arXiv:2006.10732 (2020).

[AKT19] Alnur Ali, J Zico Kolter, and Ryan J Tibshirani, *A continuous-time view of early stopping for least squares*, International Conference on Artificial Intelligence and Statistics, vol. 22, 2019.

[AS17] Madhu S Advani and Andrew M Saxe, *High-dimensional dynamics of generalization error in neural networks*, arXiv preprint arXiv:1710.03667 (2017).

[BES+20] Jimmy Ba, Murat Erdogdu, Taiji Suzuki, Denny Wu, and Tianzong Zhang, *Generalization of two-layer neural networks: An asymptotic viewpoint*, International Conference on Learning Representations, 2020.

[BHMM18] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal, *Reconciling modern machine learning and the bias-variance trade-off*, arXiv preprint arXiv:1812.11118 (2018).

[BLLT19] Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler, *Benign overfitting in linear regression*, arXiv preprint arXiv:1906.11300 (2019).

[BS99] Anders Björkström and Rolf Sundberg, *A generalized view on continuum regression*, Scandinavian Journal of Statistics **26** (1999), no. 1, 17–30.

[BVDBS+15] Małgorzata Bogdan, Ewout Van Den Berg, Chiara Sabatti, Weijie Su, and Emmanuel J Candès, *Slope—adaptive variable selection via convex optimization*, The annals of applied statistics **9** (2015), no. 3, 1103.

[Cas80] George Casella, *Minimax ridge regression estimation*, The Annals of Statistics (1980), 1036–1056.

[CWB08] Emmanuel J Candes, Michael B Wakin, and Stephen P Boyd, *Enhancing sparsity by reweighted ℓ_1 minimization*, Journal of Fourier analysis and applications **14** (2008), no. 5-6, 877–905.

[Dic16] Lee H Dicker, *Ridge regression and asymptotic minimax estimation over spheres of growing dimension*, Bernoulli **22** (2016), no. 1, 1–37.

[DKT19] Zeyu Deng, Abla Kammoun, and Christos Thrampoulidis, *A model of double descent for high-dimensional binary linear classification*, arXiv preprint arXiv:1911.05822 (2019).

[DL20] Oussama Dhifallah and Yue M Lu, *A precise performance analysis of learning with random features*, arXiv preprint arXiv:2008.11904 (2020).

[DLM19] Michał Dereziński, Feynman Liang, and Michael W Mahoney, *Exact expressions for double descent and implicit regularization via surrogate random design*, arXiv preprint arXiv:1912.04533 (2019).

[DM16] David Donoho and Andrea Montanari, *High dimensional robust m-estimation: Asymptotic variance via approximate message passing*, Probability Theory and Related Fields **166** (2016), no. 3-4, 935–969.

[dRBK20] Stéphane d'Ascoli, Maria Refinetti, Giulio Biroli, and Florent Krzakala, *Double trouble in double descent: Bias and variance (s) in the lazy regime*, arXiv preprint arXiv:2003.01054 (2020).

[DS20] Edgar Dobriban and Yue Sheng, *Wonder: Weighted one-shot distributed ridge regression in high dimensions.*, Journal of Machine Learning Research **21** (2020), no. 66, 1–52.

[DW18] Edgar Dobriban and Stefan Wager, *High-dimensional asymptotics of prediction: Ridge regression and classification*, The Annals of Statistics **46** (2018), no. 1, 247–279.

[HG83] Tsushung A Hua and Richard F Gunst, *Generalized ridge regression: a note on negative ridge parameters*, Communications in Statistics-Theory and Methods **12** (1983), no. 1, 37–45.

[HK70] Arthur E Hoerl and Robert W Kennard, *Ridge regression: Biased estimation for nonorthogonal problems*, Technometrics **12** (1970), no. 1, 55–67.

[HMRT19] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani, *Surprises in high-dimensional ridgeless least squares interpolation*, arXiv preprint arXiv:1903.08560 (2019).

[HMX20] Daniel Hsu, Vidya Muthukumar, and Ji Xu, *On the proliferation of support vectors in high dimensions*, arXiv preprint arXiv:2009.10670 (2020).

[Kar13] Noureddine El Karoui, *Asymptotic behavior of unregularized and ridge-regularized high-dimensional robust regression estimators: rigorous results*, arXiv preprint arXiv:1311.2445 (2013).

[KH92] Anders Krogh and John A Hertz, *A simple weight decay can improve generalization*, Advances in neural information processing systems, 1992, pp. 950–957.

[KLS20] Dmitry Kobak, Jonathan Lomond, and Benoit Sanchez, *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), no. 169, 1–16.

[KPR+17] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al., *Overcoming catastrophic forgetting in neural networks*, Proceedings of the national academy of sciences **114** (2017), no. 13, 3521–3526.

[LD19] Sifan Liu and Edgar Dobriban, *Ridge regression: Structure, cross-validation, and sketching*, arXiv preprint arXiv:1910.02373 (2019).

[LH17] Ilya Loshchilov and Frank Hutter, *Decoupled weight decay regularization*, arXiv preprint arXiv:1711.05101 (2017).

[Lol20] Panagiotis Lolas, *Regularization in high-dimensional regression and classification via random matrix theory*, arXiv preprint arXiv:2003.13723 (2020).

[LP11] Olivier Ledoit and Sandrine Péché, *Eigenvectors of some large sample covariance matrix ensembles*, Probability Theory and Related Fields **151** (2011), no. 1-2, 233–264.

[LPRS17] Tengyuan Liang, Tomaso Poggio, Alexander Rakhlin, and James Stokes, *Fisher-rao metric, geometry, and complexity of neural networks*, arXiv preprint arXiv:1711.01530 (2017).

[LW17] Christos Louizos and Max Welling, *Multiplicative normalizing flows for variational bayesian neural networks*, Proceedings of the 34th International Conference on Machine Learning-Volume 70, JMLR. org, 2017, pp. 2218–2227.

[MM19] Song Mei and Andrea Montanari, *The generalization error of random features regression: Precise asymptotics and double descent curve*, arXiv preprint arXiv:1908.05355 (2019).

[MRSY19] Andrea Montanari, Feng Ruan, Youngtak Sohn, and Jun Yan, *The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime*, arXiv preprint arXiv:1911.01544 (2019).

[MS05] Yuzo Maruyama and William E Strawderman, *A new class of generalized bayes minimax ridge regression estimators*, The Annals of Statistics **33** (2005), no. 4, 1753–1770.

[MS18] Yuichi Mori and Taiji Suzuki, *Generalized ridge estimator and model selection criteria in multivariate linear regression*, Journal of Multivariate Analysis **165** (2018), 243–261.

[RC15] Kenneth Joseph Ryan and Mark Vere Culp, *On semi-supervised linear regression in covariate shift problems*, The Journal of Machine Learning Research **16** (2015), no. 1, 3183–3217.

[RM11] Francisco Rubio and Xavier Mestre, *Spectral convergence for a general class of random matrices*, Statistics & probability letters **81** (2011), no. 5, 592–602.

[RMR20] Dominic Richards, Jaouad Mourtada, and Lorenzo Rosasco, *Asymptotics of ridge (less) regression under general source condition*, arXiv preprint arXiv:2006.06386 (2020).

[SC95] Jack W Silverstein and Sang-Il Choi, *Analysis of the limiting spectral distribution of large dimensional random matrices*, Journal of Multivariate Analysis **54** (1995), no. 2, 295–309.

[Str78] William E Strawderman, *Minimax adaptive generalized ridge regression estimators*, Journal of the American Statistical Association **73** (1978), no. 363, 623–627.

[TAH18] Christos Thrampoulidis, Ehsan Abbasi, and Babak Hassibi, *Precise error analysis of regularized m-estimators in high dimensions*, IEEE Transactions on Information Theory **64** (2018), no. 8, 5592–5628.

[TB20] Alexander Tsigler and Peter L Bartlett, *Benign overfitting in ridge regression*, arXiv preprint arXiv:2009.14286 (2020).

[TCG20] T Tony Cai and Zijian Guo, *Semisupervised inference for explained variance in high dimensional linear regression and its applications*, Journal of the Royal Statistical Society: Series B (Statistical Methodology) (2020).

[XH19] Ji Xu and Daniel J Hsu, *On the number of variables to use in principal component regression*, Advances in Neural Information Processing Systems, 2019, pp. 5095–5104.

[Zou06] Hui Zou, *The adaptive lasso and its oracle properties*, Journal of the American statistical association **101** (2006), no. 476, 1418–1429.

[ZTSG19] Han Zhao, Yao-Hung Hubert Tsai, Russ R Salakhutdinov, and Geoffrey J Gordon, *Learning neural networks with adaptive regularization*, Advances in Neural Information Processing Systems, 2019, pp. 11389–11400.

[ZWXG18] Guodong Zhang, Chaoqi Wang, Bowen Xu, and Roger Grosse, *Three mechanisms of weight decay regularization*, arXiv preprint arXiv:1810.12281 (2018).