# Memorize to Generalize: on the Necessity of Interpolation in High Dimensional Linear Regression

Chen Cheng[1] John Duchi[1,2] Rohith Kuditipudi[3]

Departments of [1]Statistics, [2]Electrical Engineering, and [3]Computer Science
Stanford University

February 2022; Revised June 2022

**Abstract**

We examine the necessity of interpolation in overparameterized models, that is, when achieving optimal predictive risk in machine learning problems requires (nearly) interpolating the training data. In particular, we consider simple overparameterized linear regression $y = X\theta + w$ with random design $X \in \mathbb{R}^{n\times d}$ under the proportional asymptotics $d/n \to \gamma \in (1, \infty)$. We precisely characterize how prediction (test) error necessarily scales with training error in this setting. An implication of this characterization is that as the label noise variance $\sigma^2 \to 0$, any estimator that incurs at least $\mathsf{c}\sigma^4$ training error for some constant $\mathsf{c}$ is necessarily suboptimal and will suffer growth in excess prediction error at least linear in the training error. Thus, optimal performance requires fitting training data to substantially higher accuracy than the inherent noise floor of the problem.

## 1 Introduction

Conventional machine learning wisdom [e.g. 25] posits that the size of a model's training data must be large relative to its effective capacity—for which parameter count often serves as a proxy—in order for the model to have good generalization. Yet despite the fact that many common families of modern machine learning models (e.g., deep neural networks) are overparameterized in the sense that they are demonstrably able to interpolate arbitrary relabelings of their training data, they tend to generalize remarkably well in practice even after optimizing the empirical risk to zero [26].

This *benign overfitting* phenomenon has spurred considerable recent interest and effort within the learning theory community toward understanding learning in the overparameterized regime, where the empirical risk minimizer is underdetermined [6, 7, 8, 17, 21, 5, 9, 19, 20]. Yet while overparameterized interpolating models evidently generalize well, both in theory and practice, there nonetheless remains at least some reason to be skeptical of the notion that interpolation is necessarily "benign." Indeed, numerous desiderata beyond prediction risk—for example, privacy and security concerns—motivate an explicit preference for models that do not interpolate, or in particular, memorize, their training data. An alternative and perhaps less auspicious explanation for benign overfitting is that many of the crowdsourced benchmarks the machine learning community uses to evaluate models, such as ImageNet [13], have limited label uncertainty: examples with high annotator disagreement are in many cases explicitly withheld [13, 22], mitigating the danger of overfitting to label noise.

Thus, while interpolation may *suffice* to learn models with strong generalization, it is natural to wonder whether interpolation—or more evocatively, memorization—is *necessary* for learning in the overparameterized regime. Here we take a phenomenological approach, developing a simple model to explicate and predict behavior of statistical learning procedures, and motivated by the question of the necessity of memorization, we precisely characterize how prediction risk must scale with empirical risk. Considering a simple linear model $y = x^\top\theta + w$,

we define memorization in terms of the empirical risk, and formulate the cost of not fitting the training data as an optimization problem over a class of estimators $\mathcal{H}$,

$$\begin{aligned} \underset{\widehat{\theta}\in\mathcal{H}}{\text{minimize}} \quad & \mathsf{Pred}\left(\widehat{\theta}\right) := \mathbb{E}\left[(x^\top\widehat{\theta} - y)^2 \mid X\right] \\ \text{subject to} \quad & \mathsf{Train}\left(\widehat{\theta}\right) := \frac{1}{n}\mathbb{E}\left[\left\|X\widehat{\theta} - Y\right\|_2^2 \mid X\right] \geq \epsilon^2 \,, \end{aligned} \tag{1}$$

where the expectations in $\mathsf{Pred}$ and $\mathsf{Train}$ are taken conditional on the over the training data $Y$ defining $\widehat{\theta}$ conditional on $X$, as well as the future data point $(x, y)$, so that $\mathsf{Pred}\,(\cdot)$ and $\mathsf{Train}\,(\cdot)$ denote the expected prediction and training error given a prior over the true model parameter $\theta$ (respectively).

We take as inspiration the recent line of work [14, 11], which gives scenarios in which certain formal notions of memorization are necessary for a model to generalize well. We build on this by studying the extent to which memorization remains necessary even in the simplest settings: random design linear regression with independent noise. For our initial analysis, we assume the estimator $\widehat{\theta}$ is linear in $y$, which includes least-norm interpolants and ridge regression as special cases. Here, we obtain a tight asymptotic characterization of the optimal solution to the problem (1) (see Theorems 1 and 3). Key to our analysis is to show that, even though problem (1) is non-convex, strong duality obtains, and then leverage tools from random matrix theory to obtain analytic formulae for the optimal prediction risk by integrating over the spectrum of the empirical data covariance. We find that memorization of label noise is in fact necessary for generalization even in the simple case of linear regression; in particular, the threshold $\epsilon^2$ above which the optimal prediction risk is no longer achievable tends to zero asymptotically faster than the variance of the label noise—so we must fit linear regression models to (training) accuracy substantially better than the intrinsic noise floor of the problem. Beyond this threshold the excess prediction risk grows linearly with the empirical risk. Finally, assuming Gaussian noise $w$ and a Gaussian prior over $\theta$ in problem (1), we extend our analysis to hold not only for linear estimators, but for general $\mathcal{H}$ comprised of all square-integrable estimators (see Theorem 4), meaning that our characterization holds for (essentially) any estimator.

### 1.1 Related work

Neither interpolation nor memorization of training data is a new phenomenon in machine learning. Classical algorithms, such as $k$-nearest neighbors and (kernel) support vector machines, explicitly encode the training data into the learned model. Some explicitly interpolate training data and still enjoy performance guarantees; for example, the 1-nearest neighbor algorithm interpolates its training data and has classification risk at most twice the Bayes' error [12].

Nonetheless, the success of deep learning has spurred renewed interest in interpolating models. Recent work has sought to develop an understanding of "implicit regularization": whereas most minimizers of the empirical risk may generalize poorly, standard learning algorithms used in practice such as (stochastic) gradient descent tend to converge to solutions that do generalize well, even in the absence of explicit regularization terms in the training objective [15, 24, 16, 2, 3, 18]. In the particular case of overparameterized linear regression, gradient descent initialized at the origin trivially recovers the ordinary least-squares (OLS) estimator, which in overparameterized settings is the minimum norm interpolant. Most relevant to our work, Hastie et al. [17] give formulae for the asymptotic error of ridge-type estimators, including the minimum norm interpolant, as the number of features $d$ and

training observations $n$ tend to infinity in the proportional regime where $d/n \to \gamma > 1$ for both isotropic and anisotropic features. Muthukumar et al. [21] give corresponding non-asymptotic lower bounds, with matching upper bounds for certain particular feature distributions, on the minimal error achievable among all interpolating solutions. Bartlett et al. [5] consider regression over general Hilbert spaces, showing that the minimum norm interpolant achieves optimal error assuming certain conditions on the effective rank of the feature covariance. Our results complement this line of work: not only can overparameterized interpolating models generalize well, but in fact interpolation is *necessary* to achieve good generalization.

Our work pursues a line of inquiry Feldman [14] originates, which studies memorization in the setting of multi-class classification, where the data distribution is a heavy-tailed mixture over a finite set of subpopulations. He defines memorization in terms of the sensitivity of a model's predictions to the inclusion or exclusion of a particular observation in its training data, and under the assumption that the class labelings of distinct subpopulations are essentially independent—i.e., an observation drawn from one subpopulation yields limited to no information about the labels of the other subpopulations—proves that memorization is necessary to achieve optimal generalization. Brown et al. [11] extend these results, which are specific to label memorization, to incorporate an information-theoretic notion of memorizing the input observations in carefully constructed combinatorial settings, including next-symbol prediction and clustering on the hypercube. In contrast, we attempt a simpler tack: ordinary linear regression with standard distributional assumptions, construing memorization strictly in terms of training error.

## 2 Problem formulation

Given a design matrix $X = \mathbb{R}^{n\times d}$ $(d \geq n)$, an unknown signal $\theta \in \mathbb{R}^d$ and a noise vector $w$ such that $\mathbb{E}[w] = 0$ and $\mathsf{Var}(w) = \sigma^2 I_n$, consider the standard linear model

$$y = X\theta + w.$$

We assume that $X$ has i.i.d. mean zero rows $x_1^\top, \cdots, x_n^\top$ with covariance $\Sigma \in \mathbb{R}^{d\times d}$. The training error of an estimator $\widehat{\theta} = \widehat{\theta}(X, y)$, a function of $X$ and the responses $y$ whose dependence on both we typically leave implicit, is $\mathsf{Train}_{X,\theta}(\widehat{\theta}) = \frac{1}{n}\mathbb{E}_w[\|X\widehat{\theta} - y\|_2^2 \mid X, \theta]$, while the prediction (generalization) error is $\mathsf{Pred}_{X,\theta}(\widehat{\theta}) = \mathbb{E}_{x,w}[(x^\top\theta - x^\top\widehat{\theta})^2 \mid X, \theta]$, where $x$ is an independent copy from the input distribution. We consider a Bayesian formulation where the ground truth $\theta$ has a prior distribution independent of the data and the noise, and the posterior training and generalization errors are $\mathsf{Train}_X(\widehat{\theta}) = \mathbb{E}_\theta[\mathsf{Train}_{X,\theta}(\widehat{\theta})]$ and $\mathsf{Pred}_X(\widehat{\theta}) = \mathbb{E}_\theta[\mathsf{Pred}_{X,\theta}(\widehat{\theta})]$.

Given a constraint on the training error $\epsilon \in [0, \infty)$, we can then formalize the cost of not fitting the training data via the following optimization problem over a hypothesis class of estimators $\mathcal{H}$.

$$\begin{aligned} &\underset{\widehat{\theta}\in\mathcal{H}}{\text{minimize}}\ \mathsf{Pred}_X\left(\widehat{\theta}\right) \\ &\text{subject to}\ \mathsf{Train}_X\left(\widehat{\theta}\right) \geq \epsilon^2 \end{aligned} \tag{2}$$

Here, the constraint is on the average training error (over $y$); any estimator that on *each* input $y$ has prescribed error $\epsilon^2$ immediately satisfies the constraints (2). We mainly study the *cost of not fitting*

$$\mathsf{Cost}_X(\epsilon) := \min_{\widehat{\theta}\in\mathcal{H}(\epsilon)} \mathsf{Pred}_X\left(\widehat{\theta}\right) - \min_{\widehat{\theta}\in\mathcal{H}(0)} \mathsf{Pred}_X\left(\widehat{\theta}\right), \tag{3}$$

where for a given $\mathcal{H}$ we define the set $\mathcal{H}(\epsilon) := \{\widehat{\theta} \in \mathcal{H} \mid \mathsf{Train}_X(\widehat{\theta}) \geq \epsilon^2\} \subset \mathcal{H}$.

Noting that $\mathcal{H}(t)$ is a decreasing set in $t$, we always have $\mathsf{Cost}_X(\epsilon) \geq 0$. Of course, the best estimator need not necessarily memorize the entire dataset—as we shall see, some amount of regularization can help—and so we also specifically consider the *cost of not interpolating* with respect to the minimum norm interpolating solution $\widehat{\theta}_{\mathsf{ols}} := X^\top (XX^\top)^{-1} y$, defining

$$\overline{\mathsf{Cost}}_X(\epsilon) := \min_{\widehat{\theta} \in \mathcal{H}(\epsilon)} \mathsf{Pred}_X\left(\widehat{\theta}\right) - \mathsf{Pred}_X\left(\widehat{\theta}_{\mathsf{ols}}\right). \tag{4}$$

We study problem (2), in particular through the lens of the quantities (3) and (4), under the following assumptions.

**Assumption A1** (Proportional asymptotics and spherical prior)**.** The dimension $d := d(n)$ satisfies $d/n \to \gamma \in (1, \infty)$. The data matrix $X = [x_1 \; x_2 \; \cdots \; x_n]^\top \in \mathbb{R}^{n \times d}$, where $X := X(n) = (x_{ij}(n))_{i \in [n], j \in [d]}$ forms a triangular array of random variables with independent rows. There is a deterministic sequence of symmetric positive definite matrices $\Sigma := \Sigma(n) \in \mathbb{R}^{d \times d}$ such that $X = Z\Sigma^{\frac{1}{2}}$, where $Z = (z_{ij})_{i \in [n], j \in [d]}$ and $z_{ij}$ are i.i.d. random variables with distribution independent of $n$ such that $\mathbb{E}[z_{ij}] = 0$, $\mathsf{Var}(z_{ij}) = 1$, and $\mathbb{E}[z_{ij}^4] \leq M$ for a universal constant $M$. In addition, we assume $\theta$ has prior independent of the data $X, y$, with zero mean and variance $\mathsf{Var}(\theta) = I_d/d$.

Under Assumption A1, for each $n$, $x_1(n), \cdots, x_n(n)$ are i.i.d. random vectors such that

$$\mathbb{E}[x_i(n)] = 0, \qquad \mathsf{Var}\left(x_i(n)\right) = \Sigma(n).$$

Meanwhile, examples of priors satisfying the assumption include the uniform prior on the unit sphere $\mathbb{S}^{d-1}$ and the Gaussian prior $\mathsf{N}(0, I_d/d)$, where note that $\mathbb{E}[\|\theta\|_2^2] = 1$. We assume $\gamma > 1$, and hence, as the model is overparameterized, zero training error is attainable.

While at first blush appearing restrictive, our main results characterize the cost of not fitting for linear estimators.

**Assumption A2** (Linear estimators)**.** The hypothesis class consists of all linear estimators, i.e.,

$$\mathcal{H} = \left\{\widehat{\theta}(X, y) = Ay \,, A := A(X) \in \mathbb{R}^{d \times n}\right\},$$

where $A$ may depend on the features $X$ but not the labels $y$.

Notably, the hypothesis class of linear estimators contains the popular ridge estimator $\widehat{\theta}_\lambda := (X^\top X + \lambda I)^{-1} X^\top y$ and minimum norm interpolant $\widehat{\theta}_{\mathsf{ols}} := (X^\top X)^\dagger X^\top y$. Because we seek exact optimality results for more general estimators, we follow standard practice in minimax and asymptotic statistics to choose a prior on the "true" parameter $\theta$. In classical linear regression, the prior of choice is a Gaussian, so that Anderson's theorem (1955) guarantees the posterior mean is minimax for any symmetric loss, and so the optimal estimator is linear. In our case, a similar result holds, though it is more subtle because of the nonconvex constraint (2) on training error; Theorem 4 to come guarantees that when the prior and noise are both Gaussian, the optimal estimator solving problem (2) belongs to the collection of linear estimators. Thus, our main results extend immediately to the general class of all square integrable estimators:

**Assumption A2′** (Estimators with Gaussian prior)**.** The parameter $\theta \sim \mathsf{N}(0, I_d/d)$ and the noise $w \sim \mathsf{N}(0, \sigma^2 I_n)$. The hypothesis class consists of measurable, square integrable $\widehat{\theta} : \mathbb{R}^{n\times d+n} \to \mathbb{R}^d$, i.e.,

$$\mathcal{H} = \left\{ \widehat{\theta} = \widehat{\theta}(X, y) \mid \mathbb{E}_y[\|\widehat{\theta}(X, y)\|_2^2 \mid X] < \infty \right\}.$$

We return to more discussion in Section 3.3.

# 3 Main results

## 3.1 The isotropic case

We first consider the isotropic setting where $\Sigma = I$ for all $n$, and thus $x_{ij}$ are i.i.d. random variables with zero mean and unit variance. Before stating the main theorem regarding the quantity $\mathsf{Cost}_X(\epsilon)$, we first characterize the optimal solution to the cost of not fitting problem (2) via strong duality, illustrating the role random matrix theory plays in computing the optimal solution value. We postpone most of the technical details to Section 4.

When $\mathcal{H}$ consists of linear estimators $\widehat{\theta} = Ay$, we define the shorthand $\mathcal{P}(A) := \mathsf{Pred}_X(\widehat{\theta})$ and $\mathcal{T}(A) := \mathsf{Train}_X(\widehat{\theta})$, with which we express the cost of not fitting problem (2) as

$$\begin{array}{ll} \underset{A\in\mathbb{R}^{d\times n}}{\text{minimize}} & \mathcal{P}(A) = \dfrac{1}{d}\|AX - I\|_F^2 + \sigma^2 \|A\|_F^2 \\ \text{subject to} & \mathcal{T}(A) = \dfrac{1}{nd}\|XAX - X\|_F^2 + \dfrac{\sigma^2}{n}\|XA - I\|_F^2 \geq \epsilon^2 . \end{array} \tag{5}$$

The problem—while nonconvex—has quadratic objective and a single quadratic constraint. Thus we may leverage strong duality [10, Appendix B.1], writing a Lagrangian and solving, to conclude that for some $\rho_n := \rho_n(\epsilon)$ such that $I - \frac{\rho_n}{d} X^\top X \succ 0$, the optimal $A$ for the problem (2) is

$$A(\rho_n) = \left( I - \rho_n \sigma^2 \left( I - \frac{\rho_n}{d} X^\top X \right)^{-1} \right) (X^\top X + d\sigma^2 I)^{-1} X^\top,$$

where $\rho_n$ is the dual optimal value of the Lagrange multiplier associated with the constraint $\mathcal{T}(A) \geq \epsilon^2$. When $\rho_n = 0$, the constraint is inactive, so $A(0)$ is the global minimizer of the unconstrained problem and evidently corresponds to a ridge regression estimate; we have $\mathsf{Cost}_X(\epsilon) = \mathcal{P}(A(\rho_n)) - \mathcal{P}(A(0))$ and $\mathcal{T}(A(\rho_n)) = \epsilon^2$. Substituting $A = A(\rho)$ into $\mathcal{P}(A)$ and $\mathcal{T}(A)$, we obtain

$$\begin{aligned} \mathcal{P}(A(\rho)) - \mathcal{P}(A(0)) &= \frac{\rho^2\sigma^4}{d} \mathsf{Tr}\left( \left( I - \frac{\rho}{d} X^\top X \right)^{-2} \frac{X^\top X}{d} \left( \frac{X^\top X}{d} + \sigma^2 I \right)^{-1} \right), \\ \mathcal{T}(A(\rho)) &= \frac{\sigma^4}{n} \mathsf{Tr}\left( \left( I - \frac{\rho}{d} X^\top X \right)^{-2} \left( \frac{X^\top X}{d} + \sigma^2 I \right)^{-1} \right). \end{aligned}$$

We may now leverage high-dimensional random matrix theory and asymptotics. Let $X$ have singular values $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. Denoting the empirical spectral distribution of $\frac{1}{d} XX^\top$ via its c.d.f. $H_n(s) := \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\lambda_i^2/d \leq s}$, we equivalently have

$$\mathcal{P}(A(\rho)) - \mathcal{P}(A(0)) = \frac{\rho^2 n}{d} \int \frac{\sigma^4 s}{(1-\rho s)^2 (s + \sigma^2)} dH_n(s),$$

$$\mathcal{T}(A(\rho)) = \int \frac{\sigma^4}{(1-\rho s)^2\,(s+\sigma^2)} dH_n(s)\,.$$

By standard results in random matrix theory (see Lemma A.1), $H_n$ converges weakly to the Marchenko-Pastur c.d.f. $H$, which has support $[\lambda_-, \lambda_+]$ ro $\lambda_\pm := \left(1 \pm 1/\sqrt{\gamma}\right)^2$, and density

$$dH(s) = \frac{\gamma}{2\pi} \frac{\sqrt{(\lambda_+ - s)(s-\lambda_-)}}{s} \mathbb{1}_{s\in[\lambda_-,\lambda_+]} ds\,. \tag{6}$$

Therefore for any fixed $0 \le \rho < \frac{1}{1+\sqrt{\gamma}}$,

$$\lim_{n\to\infty} \left(\mathcal{P}(A(\rho)) - \mathcal{P}(A(0))\right) = \frac{\rho^2}{\gamma} \int \frac{\sigma^4 s}{(1-\rho s)^2 (s+\sigma^2)} dH(s)\,,$$
$$\lim_{n\to\infty} \mathcal{T}(A(\rho)) = \int \frac{\sigma^4}{(1-\rho s)^2\,(s+\sigma^2)} dH(s)\,.$$

Setting $\rho = 0$ corresponds to making the constraint (5) inactive, so we therefore define the memorization threshold

$$\epsilon_\sigma^2 := \int \frac{\sigma^4}{s+\sigma^2} dH(s), \tag{7}$$

and observe that for any $\epsilon^2 \ge \epsilon_\sigma^2$, there exists a $\rho \ge 0$ such that $\lim_{n\to\infty} \mathcal{T}(A(\rho)) = \epsilon^2$. Given that $\mathcal{T}(A(\rho_n)) = \epsilon^2$, we expect that $\lim_{n\to\infty} \rho_n = \rho$ and therefore should have

$$\lim_{n\to\infty} \mathsf{Cost}_X(\epsilon) = \lim_{n\to\infty} \left(\mathcal{P}(A(\rho)) - \mathcal{P}(A(0))\right) = \frac{\rho^2}{\gamma} \int \frac{\sigma^4 s}{(1-\rho s)^2 (s+\sigma^2)} dH(s).$$

We can make each of these steps rigorous (see Section 4), yielding the following theorem.

**Theorem 1.** *Let Assumption A1 and either Assumption A2 or A2′ hold. Then as $n \to \infty$,*

*(i) (**threshold value**) for $\epsilon_\sigma$ defined in Eq. (7), $\epsilon_\sigma^2 = \frac{\sigma^4}{\sigma^2+1-1/\gamma} + o(\sigma^4)$.*

*(ii) (**no cost below threshold**) if $\epsilon < \epsilon_\sigma$, then with probability one $\lim_{n\to\infty} \mathsf{Cost}_X(\epsilon) = 0$. In addition, for the ridge estimator $\widehat{\theta}_{d\sigma^2} = (X^\top X + d\sigma^2 I)^{-1} X^\top y$, we have*

$$\lim_{n\to\infty} \left( \min_{\widehat{\theta}\in\mathcal{H}(\epsilon)} \mathsf{Pred}_X\left(\widehat{\theta}\right) - \mathsf{Pred}_X\left(\widehat{\theta}_{d\sigma^2}\right) \right) = 0\,.$$

*(iii) (**cost of not fitting**) if $\epsilon \ge \epsilon_\sigma$, there exists a scalar $\rho := \rho(\epsilon) \in \left[0, \lambda_+^{-1}\right)$ that uniquely solves*

$$\int \frac{\sigma^4}{(1-\rho s)^2\,(s+\sigma^2)} dH(s) = \epsilon^2, \tag{8}$$

*and with probability one*

$$\lim_{n\to\infty} \mathsf{Cost}_X(\epsilon) = \frac{\rho^2}{\gamma} \int \frac{\sigma^4 s}{(1-\rho s)^2 (s+\sigma^2)} dH(s). \tag{9}$$

*For the constants $\mathsf{c} := \frac{2}{\lambda_-^2+\sigma^2}$ and $\mathsf{C} := \frac{(1-1/\sqrt{2})^2\lambda_-}{\lambda_+^2\gamma}$, we have $\lim_{n\to\infty} \mathsf{Cost}_X(\epsilon) \ge \mathsf{C}\epsilon^2$ whenever $\epsilon^2 \ge \mathsf{c}\sigma^4$.*

Part (i) of Theorem 1 characterizes the threshold for the constraint on training error above which no linear estimator can achieve optimal generalization; from part (ii), so long as the constraint is below this threshold, optimal generalization remains attainable. Together, parts (i) and (iii) of the theorem imply that for an estimator to achieve optimal generalization, the estimator must incur $O(\sigma^4)$ training error as the label noise variance $\sigma^2$ tends to zero. When $\sigma^2$ is small, this is quadratically smaller than the inherent noise floor in the problem. Moreover, part (iii) implies eventually for sufficiently large $\epsilon$ that $\mathsf{Cost}_X(\epsilon)$ grows linearly in terms of the constraint on training error $\mathsf{Train}_X(\widehat{\theta}) = \epsilon^2$—by not memorizing, we are essentially paying the same additional amount of error in generalization in terms of training error up to a constant factor. We conclude that memorization for high dimensional linear regression—training to accuracy quadratically smaller than the inherent noise floor in the problem—is necessary, and with the "necessity" increasing as the signal-to-noise ratio grows.

We now turn to look specifically at the cost of exact interpolation; instead of comparing against the best linear estimator, we characterize $\overline{\mathsf{Cost}}_X(\epsilon)$ (see Eq. (4)), the prediction error of $\widehat{\theta} \in \mathcal{H}(\epsilon)$ to the minimum norm interpolant $\widehat{\theta}_{\mathsf{ols}}$. We provide a proof of the following theorem in Appendix C.

**Theorem 2.** *Let Assumption A1 and either Assumption A2 or A2′ hold. Then*

*(i) (**interpolation cost**) for any $\epsilon \geq 0$, $\mathsf{Cost}_X(\epsilon) - \overline{\mathsf{Cost}}_X(\epsilon) = \mathsf{Pred}_X(\widehat{\theta}_{\mathsf{ols}}) - \mathsf{Pred}_X(\widehat{\theta}(0))$, and with probability one*

$$\lim_{n\to\infty} \left(\mathsf{Pred}_X(\widehat{\theta}_{\mathsf{ols}}) - \mathsf{Pred}_X(\widehat{\theta}(0))\right) = \frac{\sigma^4}{\gamma}\int \frac{1}{s(s+\sigma^2)} dH(s) = \frac{\sigma^4}{\gamma\,(1-1/\gamma)^3} + o(\sigma^4).$$

*(ii) (**interpolation threshold**) for any $\sigma > 0$, there exists a $\rho = \rho_{\mathrm{ols}} \in (0, \lambda_+^{-1})$ that uniquely solves*

$$\rho^2 \int \frac{s}{(1-\rho s)^2\,(s+\sigma^2)} dH(s) = \int \frac{1}{s(s+\sigma^2)} dH(s), \tag{10}$$

*where for the threshold $\epsilon^2_{\sigma,\mathrm{ols}} := \int \frac{\sigma^4}{(1-\rho_{\mathrm{ols}} s)^2 (s+\sigma^2)} dH(s)$ we have*

$$\lim_{n\to\infty} \overline{\mathsf{Cost}}_X(\epsilon) \begin{cases} < 0 & \text{if } \epsilon < \epsilon_{\sigma,\mathrm{ols}} \\ = 0 & \text{if } \epsilon = \epsilon_{\sigma,\mathrm{ols}} \\ > 0 & \text{if } \epsilon > \epsilon_{\sigma,\mathrm{ols}}. \end{cases}$$

*In comparison to the threshold $\epsilon_\sigma$ in Eq. (7) and Theorem 1, we have $\epsilon_\sigma < \epsilon_{\sigma,\mathrm{ols}} \leq \frac{2\lambda_+}{\lambda_-}\epsilon_\sigma$.*

Part (i) shows that the minimum norm interpolant is nearly optimal, at least as $\sigma^2 \to 0$: its prediction error over the best (linear) estimator scales asymptotically as $O(\sigma^4/\gamma)$, and as the aspect ratio $\gamma$ increases it becomes closer and closer to optimal. Part (ii) complements this result, showing that if the constraint $\epsilon$ on the training error of an estimator is at most $\epsilon^2 \leq \epsilon^2_{\sigma,\mathrm{ols}} = O(\sigma^4)$, there are better estimators than the minimum norm interpolant; one concrete example here is the optimal ridge estimator $\widehat{\theta}_{d\sigma^2}$, which has asymptotic training error, as we see from Theorem 1.

### 3.2 Features with general covariance

In this section, we develop analogous results to those for the identity covariance in Sec. 3.1, showing that the results are not merely some fragile and magical consequences of isotropy. Here, we make the following assumption about the covariance matrix $\Sigma$.

**Assumption A3.** The population covariance $\Sigma$ has eigenvalues $t_1 \geq t_2 \geq \cdots \geq t_d \geq 0$, where $t_1 = 1$ and there exists $\kappa < \infty$ such that $t_d \geq 1/\kappa$. The empirical spectral distribution $T_n(s) := \frac{1}{d}\sum_{i=1}^{d} \mathbb{1}_{t_i \leq s}$ of $\Sigma$ converges weakly to a c.d.f. $T$.

Under this assumption, the empirical distribution for the eigenvalues of $\frac{1}{d}XX^\top$ converges weakly to a distribution with deformed Marchenko-Pastur c.d.f. $G$. (See Lemma A.3 for the precise definition.) With the limit $G$ and recalling the Marchenko-Pastur c.d.f. $H$, we may characterize $\mathsf{Cost}_X(\epsilon)$ for general covariances $\Sigma$. The result is analogous to Theorem 1, modulo the condition number $\kappa$ and the alternative limit $G$. To that end, define the deformed threshold

$$\epsilon^2_{\sigma,\text{def}} := \int \frac{\sigma^4}{s+\sigma^2} dG(s), \tag{11}$$

comparing to the definition (7) of $\epsilon_\sigma = \int \frac{\sigma^4}{s+\sigma^2} dH(s)$. We then have the following theorem, whose proof we provide in Appendix D.

**Theorem 3.** *Let Assumptions A1 and A3 hold, $\sigma > 0$, and let $G$ be the deformed Marchenko-Pastur c.d.f. in Lemma A.3. If either Assumption A2 or A2′ holds, then as $n \to \infty$,*

*(i) (**threshold value**) for $\epsilon_{\sigma,\text{def}}$ defined in Eq. (11), $\epsilon^2_{\sigma,\text{def}} \leq \epsilon^2_{\sqrt{\kappa}\sigma}/\kappa = \frac{\kappa\sigma^4}{\kappa\sigma^2+1-1/\gamma} + o(\sigma^4)$.*

*(ii) (**no cost below threshold**) if $\epsilon < \epsilon_{\sigma,\text{def}}$, then with probability one $\lim_{n\to\infty} \mathsf{Cost}_X(\epsilon) = 0$. In addition, define the ridge estimator $\widehat{\theta}_{d\sigma^2} = (X^\top X + d\sigma^2 I)^{-1} X^\top y$, we have*

$$\lim_{n\to\infty} \left( \min_{\widehat{\theta} \in \mathcal{H}(\epsilon)} \mathsf{Pred}_X\left(\widehat{\theta}\right) - \mathsf{Pred}_X\left(\widehat{\theta}_{d\sigma^2}\right) \right) = 0.$$

*(iii) (**cost of not fitting**) If $\epsilon \geq \epsilon_{\sigma,\text{def}}$, there exists $\rho_{\text{def}} = \rho_{\text{def}}(\epsilon) \in [0, 1/\lambda_+)$ that uniquely solves*

$$\kappa\sigma^4 \cdot \left( \int \frac{1}{(1-\rho s)^2 (s+\kappa\sigma^2)} dH(s) - \int \frac{1}{s+\kappa\sigma^2} dH(s) \right) = \epsilon^2 - \epsilon^2_{\sigma,\text{def}}, \tag{12}$$

*where $H$ is the Marchenko-Pastur c.d.f. (6). Further, with probability one*

$$\liminf_{n\to\infty} \mathsf{Cost}_X(\epsilon) \geq \frac{\rho^2_{\text{def}}}{\gamma} \int \frac{\sigma^4 s}{(1-\rho_{\text{def}} s)^2 (s+\sigma^2)} dH(s).$$

*For the constants $\mathsf{c} := \frac{2\kappa}{\lambda_- + \kappa\sigma^2}$ and $\mathsf{C} = \frac{\lambda_- (1-1/\sqrt{2})^2}{\kappa\lambda_+^2\gamma}$, we have $\liminf_{n\to\infty} \mathsf{Cost}_X(\epsilon) \geq \mathsf{C}\epsilon^2$ whenever $\epsilon^2 \geq \mathsf{c}\sigma^4$.*

### 3.3 Optimality of general estimators in Gaussian case

While, as we discuss before Assumption A2′, the lower bounds in Theorems 1, 2, and 3 apply over the class of linear estimators, which allows our exact predictive risk characterizations, these results hold for all estimators satisfying mild regularity conditions under a Gaussianity assumption on the data distribution. Our main insight here is that when the prior and noise distributions are Gaussian, for all $\epsilon \geq 0$, the linear estimator class contains the optimal estimator among the broader class of all square integrable estimators with training error at least $\epsilon^2$. Of course, this is trivial when $\epsilon = 0$, as given $(X, y)$ in such a model, the posterior on $\theta$ is Gaussian. That the result holds for $\epsilon > 0$ is a bit more subtle. Specifically, we have the following theorem, whose proof we provide in Appendix E.

**Theorem 4.** *Let Assumptions A1 and A2′ hold. Let $\mathcal{H}_{\text{lin}}$ and $\mathcal{H}_{\text{sq}}$ denote the classes of linear and square integrable estimators in Assumptions A2 and A2′, respectively. Then for all $\epsilon \geq 0$, $\inf_{\widehat{\theta} \in \mathcal{H}_{\text{sq}}(\epsilon)} \mathsf{Pred}_X(\widehat{\theta}) = \min_{\widehat{\theta} \in \mathcal{H}_{\text{lin}}(\epsilon)} \mathsf{Pred}_X(\widehat{\theta})$.*

Observing in the Gaussian case that the posterior over $\theta \mid y$ is has mean linear in $y$ and covariance independent of $y$, the main idea underlying the proof is to factor the prediction and training error over the marginal distribution of $y$, as

$$\mathsf{Pred}_X\left(\widehat{\theta}\right) = \mathbb{E}_y\left[\mathbb{E}_{\theta|y}\left[\left\|\Sigma^{\frac{1}{2}}\left(\widehat{\theta}(X,y) - \theta\right)\right\|_2^2 \,\middle|\, y\right] \,\middle|\, X\right]$$
$$\mathsf{Train}_X\left(\widehat{\theta}\right) = \mathbb{E}_y\left[\left\|X\widehat{\theta}(X,y) - y\right\|_2^2 \,\middle|\, X\right].$$

Thus the cost of not fitting problem (2) is a functional (infinite-dimensional) optimization problem over $\mathcal{H}_{\text{sq}}$, with a quadratic objective and a single quadratic constraint, for which we show that strong duality still obtains. Applying the appropriate Karush-Kuhn-Tucker conditions, we can then recover that the optimal estimator is linear, and in particular is

$$\widehat{\theta}(X,y) = \left(I - \rho(\epsilon)\sigma^2\left(\Sigma - \frac{\rho(\epsilon)}{d}X^\top X\right)^{-1}\right)(X^\top X + d\sigma^2 I)^{-1}X^\top y.$$

Here $\rho(\epsilon)$ is the dual optimal value of the Lagrange multiplier for the constraint on training error, and it is identical to that in Theorems 1 and 3. See Section 4.1 for the details.

## 4 Proof of Theorem 1

### 4.1 Reduction by strong duality

We first provide some technical lemmas to reduce the nonconvex problem (2). The lemmas will be useful in both the isotropic case and the general covariance case, and in particular the key ingredient that allows for this reduction is strong duality in quadratic optimization.

The first lemma gives an equivalent formulation of the cost of not fitting problem (2) using the closed forms of $\mathsf{Pred}_X(\widehat{\theta})$ and $\mathsf{Train}_X(\widehat{\theta})$. We defer the proof to Appendix B.1.

**Lemma 4.1.** *Let Assumption A2 hold and assume $X = Z\Sigma^{\frac{1}{2}}$. Then for any $\widehat{\theta} \in \mathcal{H}$ the following is an equivalent formulation of problem (2):*

$$\begin{aligned} \underset{A \in \mathbb{R}^{d\times n}}{\text{minimize}} \quad & \mathcal{P}(A;\Sigma) := \frac{1}{d}\left\|\Sigma^{\frac{1}{2}}(AX - I)\right\|_F^2 + \sigma^2\left\|\Sigma^{\frac{1}{2}}A\right\|_F^2 \\ \text{subject to} \quad & \mathcal{T}(A;\Sigma) := \frac{1}{nd}\|XAX - X\|_F^2 + \frac{\sigma^2}{n}\|XA - I\|_F^2 \geq \epsilon^2. \end{aligned} \tag{13}$$

As strong duality holds for this problem [cf. 10, Appendix B.1], we derive in Lemma 4.2 the optimality criteria via studying the dual. We postpone the proof details to Appendix B.2.

**Lemma 4.2.** *There exists a* $\rho_n := \rho_n(\epsilon, \Sigma) \geq 0$ *such that* $\Sigma - \frac{\rho_n}{d} X^\top X \succ 0$ *and the optimal solution of problem* (13) *is* $A_n := A(\rho_n, \Sigma)$*, where*

$$A(\rho, \Sigma) = \left(I - \rho\sigma^2 \left(\Sigma - \frac{\rho}{d} X^\top X\right)^{-1}\right) X^\top (XX^\top + d\sigma^2 I)^{-1} \tag{14a}$$

$$= \left(I - \rho\sigma^2 \left(\Sigma - \frac{\rho}{d} X^\top X\right)^{-1}\right) (X^\top X + d\sigma^2 I)^{-1} X^\top. \tag{14b}$$

$A(\rho, \Sigma)$ *is defined for* $\rho \in D$*, where* $D$ *is the interval for all* $\rho \geq 0$ *such that* $\Sigma - \frac{\rho}{d} X^\top X \succ 0$.

We suppress the dependence of $A, \rho$ on the data matrix $X$ for simplicity.

In the next lemma we derive the exact forms of the constraint $\mathcal{T}(A(\rho, \Sigma); \Sigma)$ and the growth of the objective $\mathcal{P}(A(\rho, \Sigma); \Sigma) - \mathcal{P}(A(0, \Sigma); \Sigma)$. We defer the proof to Appendix B.3.

**Lemma 4.3.** *Let the conditions of Lemma 4.2 hold, and assume* $XX^\top$ *is non-singular. Then for any* $\rho \in D$ *we have*

$$\begin{aligned}
&\mathcal{P}(A(\rho, \Sigma); \Sigma) - \mathcal{P}(A(0, \Sigma); \Sigma) \\
&= \frac{\rho^2 \sigma^4}{d} \mathsf{Tr}\left(\left(\Sigma - \frac{\rho}{d} X^\top X\right)^{-1} \Sigma \left(\Sigma - \frac{\rho}{d} X^\top X\right)^{-1} X^\top X \left(X^\top X + d\sigma^2 I\right)^{-1}\right),
\end{aligned}$$

*and*

$$\begin{aligned}
&\mathcal{T}(A(\rho, \Sigma); \Sigma) \\
&= \frac{d\sigma^4}{n} \mathsf{Tr}\left(\Sigma \left(\Sigma - \frac{\rho}{d} X^\top X\right)^{-1} X^\top X \left(\Sigma - \frac{\rho}{d} X^\top X\right)^{-1} \Sigma \left(X^\top X\right)^\dagger \left(X^\top X + d\sigma^2 I\right)^{-1}\right).
\end{aligned}$$

### 4.2 Main proof of Theorem 1

By Theorem 4, we only need to prove under Assumption A2 with the linear hypothesis class $\mathcal{H} = \{\widehat{\theta} : \widehat{\theta} = A(X)y\}$.

**Part I: Memorization threshold.** From Eq. (7), we can directly write out

$$\epsilon_\sigma^2 = \int \frac{\sigma^4}{s + \sigma^2} dH(s) = \sigma^4 \cdot \lim_{y \to 0^+} m_H(-\sigma^2 + iy)\,, \tag{15}$$

where $m_H : \mathbb{C}_+ \to \mathbb{C}_+$ is the Stieltjes transform (cf. (18)) of the Marchenko-Pastur law.

**Lemma 4.4.** *For any* $\sigma^2 > 0$*,*

$$\lim_{y \to 0^+} m_H(-\sigma^2 + iy) = \frac{\sigma^2 + o(\sigma^2)}{\sigma^2 \cdot (1 - 1/\gamma + \sigma^2)}.$$

We defer the proof to Appendix B.4. We conclude the proof of (i) by applying Lemma 4.4 to Eq. (15),

$$\epsilon_\sigma^2 = \sigma^4 \cdot \frac{\sigma^2 + o(\sigma^2)}{\sigma^2 \cdot (1 - 1/\gamma + \sigma^2)} = \frac{\sigma^4}{\sigma^2 + 1 - 1/\gamma} + o(\sigma^4)\,.$$

**Part II: No cost below threshold.** Invoke Lemma 4.2 and set $\rho = 0$ (when the constraint is not active) to obtain the global minimizer for the unconstrained problem

$$A(0, I) = X^\top (XX^\top + d\sigma^2 I)^{-1} = (X^\top X + d\sigma^2 I)^{-1} X^\top,$$

so the ridge estimator $\widehat{\theta}_{d\sigma^2}$ is optimal in $\mathcal{H}(0)$. Thus we must prove that $\widehat{\theta}_{d\sigma^2} \in \mathcal{H}(\epsilon)$ eventually, for which it suffices to show

$$\liminf_{n\to\infty} \mathsf{Train}_X \left(\widehat{\theta}_{d\sigma^2}\right) = \liminf_{n\to\infty} \mathcal{T}(A(0, I); I) > \epsilon^2 \,,$$

where $\mathcal{T}(A; \Sigma)$ is defined in Eq. (13). When $\Sigma = I$, we can compute the exact limits in Lemma 4.3 when $n \to \infty$.

**Lemma 4.5.** *Fix $0 \le \rho < \lambda_+^{-1}$. Then with probability one*

$$\lim_{n\to\infty} \left(\mathcal{P}(A(\rho, I); I) - \mathcal{P}(A(0, I); I)\right) = \frac{\rho^2}{\gamma} \int \frac{\sigma^4 s}{(1-\rho s)^2 (s+\sigma^2)} dH(s),$$
$$\lim_{n\to\infty} \mathcal{T}(A(\rho, I); I) = \int \frac{\sigma^4}{(1-\rho s)^2 (s+\sigma^2)} dH(s).$$

Invoke Lemma 4.5 above for $\rho = 0$ to conclude that with probability one

$$\lim_{n\to\infty} \mathcal{T}(A(0, I); I) = \lim_{n\to\infty} \int \frac{\sigma^4}{s+\sigma^2} dH_n(s) = \int \frac{\sigma^4}{s+\sigma^2} dH(s) = \epsilon_\sigma^2 > \epsilon^2.$$

**Part III: Cost of not-fitting above threshold.** First we show for any $\epsilon \ge \epsilon_\sigma$ there exists a unique $\rho = \rho(\epsilon) \in \left[0, \lambda_+^{-1}\right)$ that solves the fixed point (8), i.e.

$$\int \frac{\sigma^4}{(1-\rho s)^2 \, (s+\sigma^2)} dH(s) = \epsilon^2.$$

As the left hand side is increasing in $\rho$ and when $\rho \downarrow 0$, the integral approaches $\epsilon_\sigma^2 = \int \frac{\sigma^4}{s+\sigma^2} dH(s)$. On the other hand, by substituting in the exact formula of $dH(s)$ in Eq. (6), we see as $s \uparrow \lambda_+$,

$$\frac{\sigma^4}{(1-\lambda_+^{-1} s)^2 \, (s+\sigma^2)} dH(s) = (1 + o(1)) \frac{\gamma \lambda_+ \sigma^4 \sqrt{\lambda_+ - \lambda_-}}{2\pi \, (\lambda_+ + \sigma^2)} \cdot (\lambda_+ - s)^{-\frac{3}{2}} \, ds, \tag{16}$$

so that the improper integral diverges when $\rho = \lambda_+^{-1}$. Monotone convergence then implies that the integral approaches $\infty$ as $\rho \uparrow \lambda_+^{-1}$.

It remains to show the limiting statement (9) in part (iii) of the theorem and the growth lower bounds. To do so, we leverage the duality calculations in Lemma 4.2 to transfer between the training error $\epsilon$ and the Lagrange multiplier $\rho$, using that to construct upper and lower bounds on $\mathsf{Cost}_X(\epsilon)$. By Lemma 4.2, the estimator

$$\widehat{\theta}(\rho) := A(\rho, I) y$$

is the optimal solution to problem (13) when $\epsilon^2 = \mathcal{T}(A(\overline{\rho}, I); I)$, that is, $A(\rho, I)$ solves

$$\underset{A \in \mathbb{R}^{d\times n}}{\text{minimize}} \;\; \mathcal{P}(A; I) \quad \text{subject to} \quad \mathcal{T}(A; I) \ge \mathcal{T}(A(\rho, I); I).$$

Thus, whenever $\mathcal{T}(A(\rho, I); I) < \epsilon^2$ it holds that

$$\mathsf{Cost}_X(\epsilon) \geq \mathcal{P}(A(\rho, I); I) - \mathcal{P}(A(0, I); I) \tag{17a}$$

while when $\mathcal{T}(A(\rho, I); I) > \epsilon^2$, it holds that

$$\mathsf{Cost}_X(\epsilon) \leq \mathcal{P}(A(\rho, I); I) - \mathcal{P}(A(0, I); I). \tag{17b}$$

We will give matching upper and lower bounds to the quantities (17) to show the limit (9).

Let $\rho(\epsilon) \in (0, \lambda_+^{-1})$ be the $\rho$ satisfying the fixed point (8), where $\rho(\epsilon) > 0$ as $\epsilon^2 > \epsilon_\sigma$ by assumption (as otherwise $\lim_n \mathsf{Cost}_X(\epsilon) = 0$ by part (ii) of the theorem). For any $\rho \in \left[0, \lambda_+^{-1}\right)$, Lemma 4.5 implies

$$\lim_{n\to\infty} \mathcal{T}(A(\rho, I); I) = \int \frac{\sigma^4}{(1-\rho s)^2(s+\sigma^2)} dH(s).$$

Then $\rho > \rho(\epsilon)$ implies that $\lim_n \mathcal{T}(A(\rho, I); I) > \epsilon^2$, while $\rho < \rho(\epsilon)$ implies that $\lim_n \mathcal{T}(A(\rho, I); I) < \epsilon^2$. In particular, the inequalities (17) and these limits on $\mathcal{T}$ combine to give that

$$\limsup_{n\to\infty} \mathsf{Cost}_X(\epsilon) \leq \liminf_{n\to\infty} \left[\mathcal{P}(A(\rho, I); I) - \mathcal{P}(A(0, I); I)\right]$$

whenever $\rho > \rho(\epsilon)$, while if $\rho < \rho(\epsilon)$ we have

$$\liminf_{n\to\infty} \mathsf{Cost}_X(\epsilon) \geq \limsup_{n\to\infty} \left[\mathcal{P}(A(\rho, I); I) - \mathcal{P}(A(0, I); I)\right].$$

We can now apply the limiting expansion of $\mathcal{P}(A(\rho)) - \mathcal{P}(A(0))$ in Lemma 4.5, which yields that for any $0 \leq \rho_0 < \rho(\epsilon) < \rho_1 < \lambda_+^{-1}$, we have

$$\begin{aligned}
\frac{\rho_0^2}{\gamma} \int \frac{\sigma^4 s}{(1-\rho_0 s)^2(s+\sigma^2)} dH(s) &= \lim_{n\to\infty} \left[\mathcal{P}(A(\rho_0, I); I) - \mathcal{P}(A(0, I); I)\right] \\
&\leq \liminf_{n\to\infty} \mathsf{Cost}_X(\epsilon) \leq \limsup_{n\to\infty} \mathsf{Cost}_X(\epsilon) \\
&\leq \lim_{n\to\infty} \left[\mathcal{P}(A(\rho_1, I); I) - \mathcal{P}(A(0, I); I)\right] = \frac{\rho_1^2}{\gamma} \int \frac{\sigma^4 s}{(1-\rho_1 s)^2(s+\sigma^2)} dH(s)
\end{aligned}$$

Take $\rho_1 \downarrow \rho(\epsilon)$ and $\rho_0 \uparrow \rho(\epsilon)$ to obtain the limit (9).

We complete the proof of part (iii) of the theorem via the following final lemma, which provides a linear lower bound for $\lim_{n\to\infty} \mathsf{Cost}_X(\epsilon)$.

**Lemma 4.6.** *Let* $\mathsf{c} = \frac{2}{\lambda_-^2 + \sigma^2}$. *If* $\epsilon^2 \geq \mathsf{c}\sigma^4$, *then*

$$\lim_{n\to\infty} \mathsf{Cost}_X(\epsilon) \geq \frac{(1-1/\sqrt{2})^2}{\lambda_+^2} \frac{\lambda_-}{\gamma} \cdot \epsilon^2.$$

*Proof.* Taking $\rho$ to solve the fixed point (8), the limit (9) yields

$$\lim_n \mathsf{Cost}_X(\epsilon) \overset{(9)}{=} \frac{\rho^2\sigma^4}{\gamma} \int \frac{s}{(1-\rho s)^2(s+\sigma^2)} dH(s) \geq \frac{\rho^2\lambda_-}{\gamma} \int \frac{\sigma^4}{(1-\rho s)^2(s+\sigma^2)} dH(s) \overset{(8)}{=} \frac{\rho^2\lambda_-}{\gamma} \epsilon^2,$$

Thus it suffices to show that $\rho \geq \frac{1}{\lambda_+}(1 - 1/\sqrt{2})$. To see this, we leverage the following inequalities:

$$\frac{1}{(1-\rho\lambda_+)^2} \geq \int \frac{1}{(1-\rho s)^2} dH(s) \geq \int \frac{\lambda_- + \sigma^2}{(1-\rho s)^2(s+\sigma^2)} dH(s) = \frac{\lambda_- + \sigma^2}{\sigma^4} \epsilon^2 \geq 2,$$

the last inequality holding for $\epsilon^2 \geq \frac{2\sigma^4}{\sigma^2 + \lambda_-}$. Rearranging $(1-\rho\lambda_+)^2 \leq \frac{1}{2}$ yields $\rho \geq \frac{1}{\lambda_+}(1 - 1/\sqrt{2})$, which implies the claimed result. ∎

## 5 Discussion

By characterizing the excess prediction error in linear regression models as a function of constraints on training error, this paper gives insights into the necessity—in achieving optimal prediction risk—of memorization for learning. Our results support the natural conclusion that interpolation is particularly beneficial in settings with low label noise, which as we note earlier, may include some of the most widely-used existing benchmarks for deep learning. Even more, they suggest that—at least when the noise is low—memorization may simply be necessary, so that a deeper understanding of the generalization of modern machine learning algorithms may require a careful look at more precise noise properties of the prediction problems at hand.

In the anisotropic setting, our lower bounds on prediction error depend on the condition number of the data covariance, and thus our bounds not apply, i.e., are vacuous, in settings such as sparse covariance or kernel regression. Extending our results to these settings is an interesting direction for future work. Furthermore, our analysis relies heavily on the fact that both the prediction and empirical risk are quadratic in the case of least-squares regression, and thus strong duality obtains. Proving similar results in settings such as linear binary classification, where the optimal unconstrained estimator, i.e., margin maximizing solution, is nonlinear and the risk no longer quadratic, is an exciting open problem.

## References


[1] T. W. Anderson. The integral of a symmetric unimodal function over a symmetric convex set and some probability inequalities. *Proceedings of the American Mathematical Society*, 6(2):170–176, 1955.

[2] S. Arora, N. Cohen, W. Hu, and Y. Luo. Implicit regularization in deep matrix factorization. In *Advances in Neural Information Processing Systems 32*, 2019.

[3] S. Arora, S. Du, W. Hu, Z. Li, and R. Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

[4] Z. Bai and J. W. Silverstein. *Spectral analysis of large dimensional random matrices*, volume 20 of *Springer Series in Statistics*. Springer, 2010.

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

[6] M. Belkin, D. Hsu, and P. Mitra. Overfitting or perfect fitting? Risk bounds for classification and regression rules that interpolate. In *Advances in Neural Information Processing Systems 31*, pages 2300–2311. Curran Associates, Inc., 2018.

[7] M. Belkin, S. Ma, and S. Mandal. To understand deep learning we need to understand kernel learning. In *Proceedings of the 35th International Conference on Machine Learning*, pages 541–549, 2018.

[8] M. Belkin, A. Rakhlin, and A. B. Tsybakov. Does data interpolation contradict statistical optimality? In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, pages 1611–1619, 2019.

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

[10] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[11] G. Brown, M. Bun, V. Feldman, A. Smith, and K. Talwar. When is memorization of irrelevant training data necessary for high-accuracy learning? *arXiv:2012.06421 [cs.LG]*, 2021.

[12] T. M. Cover and P. E. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13:21–27, 1967.

[13] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei. ImageNet: a large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[14] V. Feldman. Does learning require memorization? A short tale about a long tail. In *Proceedings of the Fifty-Second Annual ACM Symposium on the Theory of Computing*, pages 954–959, 2020.

[15] S. Gunasekar, B. Woodworth, S. Bhojanapalli, B. Neyshabur, and N. Srebro. Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems 30*, 2017.

[16] S. Gunasekar, J. Lee, D. Soudry, and N. Srebro. Characterizing implicit bias in terms of optimization geometry. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

[17] T. Hastie, A. Montanari, S. Rosset, and R. Tibshirani. Surprises in high-dimensional ridgeless linear least squares interpolation. *arXiv:1903.08560 [math.ST]*, 2019.

[18] Z. Ji and M. Telgarsky. Gradient descent aligns the layers of deep linear networks. In *Proceedings of the Seventh International Conference on Learning Representations*, 2019.

[19] T. Liang and A. Rakhlin. Just interpolate: Kernel "ridgeless" regression can generalize. *Annals of Statistics*, 48:1329–1347, 2020.

[20] S. Mei and A. Montanari. The generalization error of random features regression: Precise asymptotics and double descent curve. *Communications on Pure and Applied Mathematics*, 2021.

[21] V. Muthukumar, K. Vodrahalli, and A. Sahai. Harmless interpolation of noisy data in regression. In *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, 2019.

[22] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar. Do ImageNet classifiers generalize to ImageNet? In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

[23] J. W. Silverstein. Strong convergence of the empirical distribution of eigenvalues of large dimensional random matrices. *Journal of Multivariate Analysis*, 55(2):331–339, 1995.

[24] D. Soudry, E. Hoffer, M. S. Nacson, S. Gunasekar, and N. Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(18):1–57, 2018.

[25] V. N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, XVI(2):264–280, 1971.

[26] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. In *Proceedings of the Fifth International Conference on Learning Representations*, 2017.

# A Asymptotics of random matrices

In this appendix, we review the classical results regarding singular values of random matrices we require. Consider a triangular array of independent and identically distributed random variables $(z_{ij}(n))_{i\in[n],j\in[d]}$ for $n = 1, 2, \cdots$ and $d := d(n)$. We write $Z := Z(n) = (z_{ij}(n)) \in \mathbb{R}^{n\times d}$. Throughout we assume the proportional asymptotics $d/n \to \gamma \in (1, \infty)$, so the matrices $Z$ have rank at most $n$. We assume throughout that the entries $z_{ij}$ satisfy $\mathbb{E}[z_{ij}] = 0$ and $\mathbb{E}[z_{ij}^2] = 1$. We have the following standard Marchenko-Pastur and Bai-Yin laws.

**Lemma A.1** (Marchenko-Pastur law, Bai and Silverstein [4], Thm. 3.4)**.** *Let $Z$ have singular values $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$, and let $\frac{1}{d}ZZ^\top$ have spectral distribution with c.d.f.*

$$H_n(s) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\lambda_i^2/d \leq s}.$$

*Then with probability one $H_n$ converges weakly to the c.d.f. $H$ supported on $[\lambda_-, \lambda_+]$, with*

$$\lambda_+ := \left(1 + \frac{1}{\sqrt{\gamma}}\right)^2 \quad \text{and} \quad \lambda_- := \left(1 - \frac{1}{\sqrt{\gamma}}\right)^2,$$

*and $H$ has density*

$$dH(s) = \frac{\gamma}{2\pi}\frac{\sqrt{(\lambda_+ - s)(s - \lambda_-)}}{s}\mathbb{1}_{s\in[\lambda_-,\lambda_+]}ds.$$

**Lemma A.2** (Bai-Yin law, Bai and Silverstein [4], Thm. 5.10)**.** *Let the conditions of Lemma A.1 hold, and assume additionally that $\sup_{ij}\mathbb{E}[z_{ij}^4] < \infty$. Then the largest and smallest singular values $\lambda_1 = \lambda_1(Z)$ and $\lambda_n = \lambda_n(Z)$ of $Z$ satisfy*

$$\frac{\lambda_1^2}{d} \overset{a.s.}{\to} \lambda_+ = \left(1 + \frac{1}{\sqrt{\gamma}}\right)^2, \qquad \frac{\lambda_n^2}{d} \overset{a.s.}{\to} \lambda_- = \left(1 - \frac{1}{\sqrt{\gamma}}\right)^2.$$

We also consider random matrices whose rows have non-identity covariance. In these cases, we assume a deterministic sequence of symmetric positive definite matrices $\Sigma := \Sigma(n) \in \mathbb{R}^{d\times d}$. We let $t_1 \geq t_2 \geq \cdots \geq t_d > 0$ denote the eigenvalues of $\Sigma$ and let $T_n$ denote the associated c.d.f.

$$T_n(s) := \frac{1}{d}\sum_{i=1}^{d} \mathbb{1}_{t_i \leq s},$$

assuming that $T_n$ converges weakly to some c.d.f. $T$ on $\mathbb{R}_+$. With this, we can state a limiting law for the spectral distribution of $\frac{1}{d}Z\Sigma Z^\top$. In the statement of the lemma, we require the *Stieltjes transform* of a measure. Letting $\mathbb{C}_+ := \{z \in \mathbb{C} \mid \mathrm{Im}(z) > 0\}$ be those elements of $\mathbb{C}$ with positive imaginary part, recall that for a measure on $\mathbb{R}$ with c.d.f. $F$, the Stieltjes transform of $m_F : \mathbb{C}_+ \to \mathbb{C}_+$ of $F$ is

$$m_F(z) := \int \frac{1}{s - z} dF(s). \tag{18}$$

Then we have the following

**Lemma A.3** (Deformed Marchenko-Pastur law, Silverstein [23]). *Let the conditions of Lemma A.1 and those on the spectral distribution $T_n$ of $\Sigma$ above hold. Let $\frac{1}{d}Z\Sigma Z^\top$ have spectral distribution with c.d.f.*

$$G_n(s) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\lambda_i^2/d \le s}.$$

*Then with probability one, $G_n$ converges weakly to the c.d.f. $G$ whose Stieltjes transform $m_G$ satisfies the fixed point equation*

$$m_G(z) = -\left(z - \int \frac{\tau}{1+\tau m_G(z)/\gamma} dT(\tau)\right)^{-1}.$$

Lemma A.3 is slightly different from the result of Silverstein [23, Thm. 1.1], whose original theorem holds for the empirical spectral distributions of $\frac{1}{n}Z\Sigma Z^\top$. Lemma A.3 follows from the change of variables $n = \frac{d}{\gamma}(1+o(1))$.

# B Proofs of identities in Theorem 1

## B.1 Proof of Lemma 4.1

This is essentially trivial: by definition, we can write

$$\begin{aligned}
\mathsf{Pred}_X(\widehat{\theta}) &= \mathbb{E}_\theta\left[\mathsf{Pred}_{X,\theta}(\widehat{\theta})\right] = \mathbb{E}_{\theta,w}\left[\|(AX-I)\theta + Aw\|_\Sigma^2 \mid X\right] \\
&= \mathsf{Tr}\left(\mathbb{E}_{\theta,w}\left[((AX-I)\theta+Aw)^\top \Sigma\left((AX-I)\theta+Aw\right) \mid X\right]\right) \\
&= \mathsf{Tr}\left(\mathbb{E}_\theta\left[\Sigma(AX-I)\theta\theta^\top(AX-I)^\top \mid X\right]\right) + \sigma^2\mathsf{Tr}\left(A^\top\Sigma A\right) \\
&= \frac{1}{d}\left\|\Sigma^{\frac{1}{2}}\left(AX-I\right)\right\|_F^2 + \sigma^2\left\|\Sigma^{\frac{1}{2}}A\right\|_F^2,
\end{aligned}$$

where in the last line we use $\mathbb{E}[\theta\theta^\top] = I_d/d$. Similarly

$$\begin{aligned}
\mathsf{Train}_X(\widehat{\theta}) &= \mathbb{E}_\theta\left[\mathsf{Train}_{X,\theta}(\widehat{\theta})\right] = \frac{1}{n}\mathbb{E}_{\theta,w}\left[\|(XA-I)\left(X\theta+w\right)\|_2^2 \mid X\right] \\
&= \frac{1}{n}\mathsf{Tr}\left(\mathbb{E}_{\theta,w}\left[\left(X\theta+w\right)^\top(XA-I)^\top(XA-I)\left(X\theta+w\right) \mid X\right]\right) \\
&= \frac{1}{n}\mathsf{Tr}\left((XA-I)X\theta\theta^\top X^\top(XA-I)^\top\right) + \frac{\sigma^2}{n}\mathsf{Tr}\left((XA-I)(XA-I)^\top\right) \\
&= \frac{1}{nd}\|XAX-X\|_F^2 + \frac{\sigma^2}{n}\|XA-I\|_F^2 .
\end{aligned}$$

## B.2 Proof of Lemma 4.2

While problem (13) is non-convex, it consists of a quadratic objective and quadratic constraint, and taking $A \to \infty$ shows that there certainly exist feasible points in the interior of the set of $A$ satisfying $\mathcal{T}(A;\Sigma) \ge \epsilon^2$. Thus, strong duality holds [10, Appendix B.1]. We therefore consider the Lagrangian dual problem, introducing the dual multplier $\lambda \ge 0$ for the constraint and writing the Lagrangian

$$\mathcal{L}(A,\lambda) = \mathcal{P}(A;\Sigma) + \lambda(\epsilon^2 - \mathcal{T}(A;\Sigma))$$

$$= \frac{1}{d}\left\|\Sigma^{\frac{1}{2}}\left(AX - I\right)\right\|_F^2 + \sigma^2\left\|\Sigma^{\frac{1}{2}}A\right\|_F^2 - \lambda\left(\frac{1}{nd}\left\|XAX - X\right\|_F^2 + \frac{\sigma^2}{n}\left\|XA - I\right\|_F^2\right) + \lambda\epsilon^2.$$

Using $\mathcal{L}$, we begin by demonstrating the first claim of the lemma, that is, that if $\Sigma - \frac{\lambda}{n}X^TX \not\succ 0$, then we have $\inf_A \mathcal{L}(A,\lambda) = -\infty$. To see this, first let $V \in \mathbb{R}^{d\times r}$ be an orthogonal basis for $X$'s row space and $V^\perp$ its orthogonal complement. Then $XV^\perp = 0$ and $\Sigma - \frac{\lambda}{n}X^\top X$ failing to be positive definite is equivalent to

$$\begin{bmatrix}V & V^\perp\end{bmatrix}^\top\left(\Sigma - \frac{\lambda}{n}X^\top X\right)\begin{bmatrix}V & V^\perp\end{bmatrix} = \begin{bmatrix}V^\top\left(\Sigma - \frac{\lambda}{n}X^\top X\right)V & 0\\ 0 & (V^\perp)^\top\Sigma V^\perp\end{bmatrix}$$

failing to be positive definite. Then as $(V^\perp)^\top\Sigma V^\perp \succ 0$ by assumption that $\Sigma \succ 0$, it must thus be the case that $V^\top(\Sigma - \frac{\lambda}{n}X^\top X)V \not\succ 0$. We leverage this indefiniteness to observe that, as $V$ spans the row space of $X$, there exists a unit vector $\nu \in \mathbb{R}^d$, $\|\nu\| = 1$, and vector $\mu \in \mathbb{R}^n$ satisfying $\nu = X^\top\mu \in \mathbb{R}^d$ and

$$\alpha := \nu^\top\left(\Sigma - \frac{\lambda}{n}X^\top X\right)\nu \le 0. \tag{19}$$

To show that the non-positivity (19) entails $\inf_A \mathcal{L}(A,\lambda) = -\infty$ requires a few additional steps. We detour by taking the gradient of the Lagrangian with respect to $A$ (this will be useful later),

$$\begin{aligned}
&\frac{\partial}{\partial A}\mathcal{L}(A,\lambda)\\
&= \frac{1}{d}\left(\Sigma AXX^\top - \Sigma X^\top\right) + \sigma^2\Sigma A - \frac{\lambda}{n}\left\{\frac{1}{d}\left(X^\top XAXX^\top - X^\top XX^\top\right) + \sigma^2\left(X^\top XA - X^\top\right)\right\}\\
&= \frac{1}{d}\left(\Sigma - \frac{\lambda}{n}X^\top X\right)AXX^\top + \sigma^2\left(\Sigma - \frac{\lambda}{n}X^\top X\right)A - \frac{1}{d}\left(\Sigma - \frac{\lambda}{n}X^\top X - \frac{\lambda d\sigma^2}{n}I\right)X^\top\\
&= \frac{1}{d}\left(\Sigma - \frac{\lambda}{n}X^\top X\right)A\left(XX^\top + d\sigma^2 I\right) - \frac{1}{d}\left(\Sigma - \frac{\lambda}{n}X^\top X - \frac{\lambda d\sigma^2}{n}I\right)X^\top.
\end{aligned} \tag{20}$$

Using the $\mu$ defining $\nu = X^\top\mu$ in Eq. (19), let $t \in \mathbb{R}$ be unspecified and take $A = t\nu\mu^\top$. Define the function $L(t) = \mathcal{L}(t\nu\mu^\top,\lambda)$, for which we have

$$\begin{aligned}
\frac{d}{dt}L(t) &= \mathsf{Tr}\left(\frac{\partial}{\partial A}\mathcal{L}(t\nu\mu^\top,\lambda)(\nu\mu^\top)^\top\right)\\
&= \frac{t}{d}\nu^\top\left(\Sigma - \frac{\lambda}{n}X^\top X\right)\nu\mu^\top\left(XX^\top + d\sigma^2 I\right)\mu - \frac{1}{d}\nu^\top\left(\Sigma - \frac{\lambda}{n}X^\top X - \frac{\lambda d\sigma^2}{n}I\right)X^\top\mu\\
&\overset{(i)}{=} \frac{t}{d}\alpha\cdot\left(\|\nu\|_2^2 + d\sigma^2\|\mu\|_2^2\right) - \frac{1}{d}\nu^\top\left(\Sigma - \frac{\lambda}{n}X^\top X - \frac{\lambda d\sigma^2}{n}I\right)\nu\\
&\overset{(ii)}{=} \frac{t\alpha}{d}\cdot\left(1 + d\sigma^2\|\mu\|_2^2\right) - \frac{\alpha}{d} + \frac{\lambda\sigma^2}{n}
\end{aligned}$$

where step $(i)$ substitutes the definition (19) of $\alpha$ and that $X^\top\mu = \nu$, while step $(ii)$ similarly uses the definition of $\alpha$ and that $\|\nu\|_2 = 1$ by assumption. We consider two cases: if $\alpha < 0$, then taking $t \to \infty$ yields $L'(t) \to -\infty$, so that $L(t) \to -\infty$ and $\inf_A \mathcal{L}(A,\lambda) = -\infty$. If $\alpha = 0$, then $L'(t) = \frac{\lambda\sigma^2}{n} > 0$, and so taking $t \to -\infty$ yields $\mathcal{L}(A,\lambda) \to -\infty$ as well. As such, the optimal $\lambda \ge 0$ must satisfy $\Sigma - \frac{\lambda}{n}X^\top X \succ 0$, as we desired to show.

Having verified that $\Sigma - \frac{\lambda}{n} X^\top X \succ 0$, we can use the derivative (20) and solve for the $A$ satisfying the stationary condition $\frac{\partial}{\partial A}\mathcal{L}(A, \lambda) = 0$, obtaining

$$\frac{1}{d}\left(\Sigma - \frac{\lambda}{n} X^\top X\right) A \left(XX^\top + d\sigma^2 I\right) - \frac{1}{d}\left(\Sigma - \frac{\lambda}{n} X^\top X - \frac{\lambda d\sigma^2}{n} I\right) X^\top = 0.$$

Solving this equation yields

$$\begin{aligned} A &= \left(\Sigma - \frac{\lambda}{n} X^\top X\right)^{-1} \left(\Sigma - \frac{\lambda}{n} X^\top X - \frac{\lambda d\sigma^2}{n} I\right) X^\top \left(XX^\top + d\sigma^2 I\right)^{-1} \\ &= \left(I - \frac{\lambda d\sigma^2}{n}\left(\Sigma - \frac{\lambda}{n} X^\top X\right)^{-1}\right) X^\top (XX^\top + d\sigma^2 I)^{-1} \\ &= \left(I - \frac{\lambda d\sigma^2}{n}\left(\Sigma - \frac{\lambda}{n} X^\top X\right)^{-1}\right) (X^\top X + d\sigma^2 I)^{-1} X^\top . \end{aligned}$$

In the last equation we use the matrix identity $X^\top (XX^\top + d\sigma^2 I)^{-1} = (X^\top X + d\sigma^2 I)^{-1} X^\top$, which follows directly via the SVD of $X$. We complete the proof by identifying $\rho_n := \frac{\lambda n}{d}$.

### B.3 Proof of Lemma 4.3

The proof is essentially pure calculations. For reference, we divide the proof into three parts.

I. We compute formulas for $A(\rho; \Sigma)X - I$ and $XA(\rho; \Sigma) - I$.

II. Derive the expansion for $\mathcal{P}(A(\rho, \Sigma); \Sigma) - \mathcal{P}(A(0, \Sigma); \Sigma)$.

III. Derive the expansion for for $\mathcal{T}(A(\rho, \Sigma); \Sigma)$.

Throughout we write $A(\rho) = A(\rho; \Sigma)$ for simplicity.

**Part I: Computing $A(\rho)X - I$ and $XA(\rho) - I$.** We first substitute expression (14a) for $A(\rho)$ into the difference $A(\rho)X - I$ to obtain

$$\begin{aligned} &A(\rho)X - I \\ &= \left(I - \rho\sigma^2\left(\Sigma - \frac{\rho}{d} X^\top X\right)^{-1}\right)(X^\top X + d\sigma^2 I)^{-1} X^\top X - I \\ &= -\rho\sigma^2\left(\Sigma - \frac{\rho}{d} X^\top X\right)^{-1}(X^\top X + d\sigma^2 I)^{-1} X^\top X + (X^\top X + d\sigma^2 I)^{-1}\left(X^\top X - X^\top X - d\sigma^2 I\right) \\ &\overset{(i)}{=} -\rho\sigma^2\left(\Sigma - \frac{\rho}{d} X^\top X\right)^{-1} X^\top X (X^\top X + d\sigma^2 I)^{-1} - d\sigma^2 (X^\top X + d\sigma^2 I)^{-1} \\ &= \left\{-\rho\sigma^2\left(\Sigma - \frac{\rho}{d} X^\top X\right)^{-1} X^\top X - d\sigma^2\left(\Sigma - \frac{\rho}{d} X^\top X\right)^{-1}\left(\Sigma - \frac{\rho}{d} X^\top X\right)\right\}(X^\top X + d\sigma^2 I)^{-1} \\ &= -\sigma^2\left(\Sigma - \frac{\rho}{d} X^\top X\right)^{-1}\left\{\rho X^\top X + d\left(\Sigma - \frac{\rho}{d} X^\top X\right)\right\}(X^\top X + d\sigma^2 I)^{-1} \\ &= -d\sigma^2\left(\Sigma - \frac{\rho}{d} X^\top X\right)^{-1} \Sigma (X^\top X + d\sigma^2 I)^{-1}, \end{aligned} \tag{21}$$

where in step $(i)$ we use that $X^\top X$ and $(X^\top X + d\sigma^2 I)^{-1}$ commute. Similarly, we can compute $XA(\rho) - I$ by using the alternative formulation (14b) for $A(\rho)$, substituting to obtain

$$XA(\rho) - I$$

$$
\begin{aligned}
&= X\left(I - \rho\sigma^2\left(\Sigma - \frac{\rho}{d}X^\top X\right)^{-1}\right)X^\top(XX^\top + d\sigma^2 I)^{-1} - I \\
&= -\rho\sigma^2 X\left(\Sigma - \frac{\rho}{d}X^\top X\right)^{-1}X^\top(XX^\top + d\sigma^2 I)^{-1} + \left(XX^\top - XX^\top - d\sigma^2 I\right)(XX^\top + d\sigma^2 I)^{-1} \\
&= -\rho\sigma^2 X\left(\Sigma - \frac{\rho}{d}X^\top X\right)^{-1}X^\top(XX^\top + d\sigma^2 I)^{-1} - d\sigma^2(XX^\top + d\sigma^2 I)^{-1} \\
&= -d\sigma^2\left\{\frac{\rho}{d}X\left(\Sigma - \frac{\rho}{d}X^\top X\right)^{-1}X^\top + I\right\}\left(XX^\top + d\sigma^2 I\right)^{-1}.
\end{aligned}
$$

As $X$ is wide and $XX^\top$ is non-singular by assumption, $\lim_{\lambda\downarrow 0} X(X^\top X + \lambda I)^{-1}X^\top = I$ and therefore

$$
\begin{aligned}
&XA(\rho) - I \\
&= -d\sigma^2\left\{\frac{\rho}{d}X\left(\Sigma - \frac{\rho}{d}X^\top X\right)^{-1}X^\top + \lim_{\lambda\downarrow 0} X(X^\top X + \lambda I)^{-1}X^\top\right\}\left(XX^\top + d\sigma^2 I\right)^{-1} \\
&= -\lim_{\lambda\downarrow 0} d\sigma^2 X\left\{\left(\frac{d}{\rho}\Sigma - X^\top X\right)^{-1} + \left(X^\top X + \lambda I\right)^{-1}\right\}\cdot X^\top\left(XX^\top + d\sigma^2 I\right)^{-1} \\
&= -\lim_{\lambda\downarrow 0} d\sigma^2 X\cdot\left\{\left(\frac{d}{\rho}\Sigma - X^\top X\right)^{-1}\left(\lambda I + \frac{d}{\rho}\Sigma\right)\left(X^\top X + \lambda I\right)^{-1}\right\}X^\top\cdot\left(XX^\top + d\sigma^2 I\right)^{-1} \\
&\overset{(i)}{=} -\lim_{\lambda\downarrow 0} d\sigma^2 X\cdot\left\{\left(\frac{d}{\rho}\Sigma - X^\top X\right)^{-1}\left(\lambda I + \frac{d}{\rho}\Sigma\right)X^\top\left(XX^\top + \lambda I\right)^{-1}\right\}\cdot\left(XX^\top + d\sigma^2 I\right)^{-1} \\
&= -d\sigma^2 X\left(\Sigma - \frac{\rho}{d}X^\top X\right)^{-1}\Sigma X^\top\left(XX^\top\right)^{-1}\left(XX^\top + d\sigma^2 I\right)^{-1}, \qquad (22)
\end{aligned}
$$

where in step $(i)$ we use that $\left(X^\top X + \lambda I\right)^{-1}X^\top = X^\top\left(XX^\top + \lambda I\right)^{-1}$.

**Part II: Computing $\mathcal{P}(A(\rho,\Sigma);\Sigma) - \mathcal{P}(A(0,\Sigma);\Sigma)$.** As

$$
\begin{aligned}
&\mathcal{P}(A(\rho,\Sigma);\Sigma) - \mathcal{P}(A(0,\Sigma);\Sigma) \qquad (23) \\
&= \frac{1}{d}\left\|\Sigma^{\frac{1}{2}}\left(A(\rho)X - I\right)\right\|_F^2 + \sigma^2\left\|\Sigma^{\frac{1}{2}}A(\rho)\right\|_F^2 - \frac{1}{d}\left\|\Sigma^{\frac{1}{2}}\left(A(0)X - I\right)\right\|_F^2 - \sigma^2\left\|\Sigma^{\frac{1}{2}}A(0)\right\|_F^2 \\
&= \underbrace{\frac{1}{d}\left(\left\|\Sigma^{\frac{1}{2}}\left(A(\rho)X - I\right)\right\|_F^2 - \left\|\Sigma^{\frac{1}{2}}\left(A(0)X - I\right)\right\|_F^2\right)}_{\text{(I)}} + \underbrace{\sigma^2\left(\left\|\Sigma^{\frac{1}{2}}A(\rho)\right\|_F^2 - \left\|\Sigma^{\frac{1}{2}}A(0)\right\|_F^2\right)}_{\text{(II)}},
\end{aligned}
$$

we compute terms (I) and (II) separately. For (I) we substitute in the explicit form (21) of $A(\rho)X - I$ to obtain

$$
\text{(I)} = d\sigma^4\mathsf{Tr}\left(\Sigma\left(\Sigma - \frac{\rho}{d}X^\top X\right)^{-1}\Sigma\left(X^\top X + d\sigma^2 I\right)^{-2}\Sigma\left(\Sigma - \frac{\rho}{d}X^\top X\right)^{-1}\right) - d\sigma^4\mathsf{Tr}\left(\Sigma\left(X^\top X + d\sigma^2 I\right)^{-2}\right).
$$

We then use the identity

$$
\left(\Sigma - \frac{\rho}{d}X^\top X\right)^{-1}\Sigma = I + \left(\Sigma - \frac{\rho}{d}X^\top X\right)^{-1}\cdot\frac{\rho}{d}X^\top X
$$

to obtain further that

$$
\text{(I)} = d\sigma^4\mathsf{Tr}\left(\Sigma\left(I + \left(\Sigma - \frac{\rho}{d}X^\top X\right)^{-1}\cdot\frac{\rho}{d}X^\top X\right)\left(X^\top X + d\sigma^2 I\right)^{-2}\left(I + \frac{\rho}{d}X^\top X\cdot\left(\Sigma - \frac{\rho}{d}X^\top X\right)^{-1}\right)\right)
$$

$$
\begin{aligned}
&\quad - d\sigma^4 \mathsf{Tr}\left(\Sigma\left(X^\top X + d\sigma^2 I\right)^{-2}\right) \\
&= d\sigma^4 \mathsf{Tr}\left(\Sigma\left(\Sigma - \frac{\rho}{d}X^\top X\right)^{-1} \cdot \frac{\rho}{d}X^\top X\left(X^\top X + d\sigma^2 I\right)^{-2}\right) \\
&\quad + d\sigma^4 \mathsf{Tr}\left(\Sigma\left(X^\top X + d\sigma^2 I\right)^{-2}\frac{\rho}{d}X^\top X \cdot \left(\Sigma - \frac{\rho}{d}X^\top X\right)^{-1}\right) \\
&\quad + d\sigma^4 \mathsf{Tr}\left(\Sigma\left(\Sigma - \frac{\rho}{d}X^\top X\right)^{-1} \cdot \frac{\rho}{d}X^\top X\left(X^\top X + d\sigma^2 I\right)^{-2}\frac{\rho}{d}X^\top X \cdot \left(\Sigma - \frac{\rho}{d}X^\top X\right)^{-1}\right) \\
&= \rho\sigma^4 \mathsf{Tr}\left(\Sigma\left(\Sigma - \frac{\rho}{d}X^\top X\right)^{-1} X^\top X\left(X^\top X + d\sigma^2 I\right)^{-2}\right) \\
&\quad + \rho\sigma^4 \mathsf{Tr}\left(\Sigma\left(X^\top X + d\sigma^2 I\right)^{-2} X^\top X\left(\Sigma - \frac{\rho}{d}X^\top X\right)^{-1}\right) \\
&\quad + \frac{\rho^2\sigma^4}{d}\mathsf{Tr}\left(\Sigma\left(\Sigma - \frac{\rho}{d}X^\top X\right)^{-1} X^\top X\left(X^\top X + d\sigma^2 I\right)^{-2} X^\top X\left(\Sigma - \frac{\rho}{d}X^\top X\right)^{-1}\right).
\end{aligned} \tag{24}
$$

For term (II), we substitute in formula (14b) for $A(\rho)$ and use that $X^\top X$ and $(X^\top X + d\sigma^2 I)^{-1}$ commute, yielding that

$$
\begin{aligned}
\text{(II)} &= \sigma^2 \mathsf{Tr}\left(\Sigma\left(I - \rho\sigma^2\left(\Sigma - \frac{\rho}{d}X^\top X\right)^{-1}\right) X^\top X (X^\top X + d\sigma^2 I)^{-2}\left(I - \rho\sigma^2\left(\Sigma - \frac{\rho}{d}X^\top X\right)^{-1}\right)\right) \\
&\quad - \sigma^2 \mathsf{Tr}\left(\Sigma X^\top X (X^\top X + d\sigma^2 I)^{-2}\right) \\
&= -\rho\sigma^4 \mathsf{Tr}\left(\Sigma\left(\Sigma - \frac{\rho}{d}X^\top X\right)^{-1} X^\top X\left(X^\top X + d\sigma^2 I\right)^{-2}\right) \\
&\quad - \rho\sigma^4 \mathsf{Tr}\left(\Sigma\left(X^\top X + d\sigma^2 I\right)^{-2} X^\top X\left(\Sigma - \frac{\rho}{d}X^\top X\right)^{-1}\right) \\
&\quad + \rho^2\sigma^6 \mathsf{Tr}\left(\Sigma\left(\Sigma - \frac{\rho}{d}X^\top X\right)^{-1} X^\top X\left(X^\top X + d\sigma^2 I\right)^{-2}\left(\Sigma - \frac{\rho}{d}X^\top X\right)^{-1}\right).
\end{aligned}
$$

Substituting the equality (24) for term (I) and the above identity for term (II) back into the expansion (23) of $\mathcal{P}(A(\rho,\Sigma);\Sigma) - \mathcal{P}(A(0,\Sigma);\Sigma)$, we get our desired expansion:

$$
\begin{aligned}
&\mathcal{P}(A(\rho,\Sigma);\Sigma) - \mathcal{P}(A(0,\Sigma);\Sigma) \\
&= \frac{\rho^2\sigma^4}{d}\mathsf{Tr}\left(\Sigma\left(\Sigma - \frac{\rho}{d}X^\top X\right)^{-1} X^\top X\left(X^\top X + d\sigma^2 I\right)^{-2} X^\top X\left(\Sigma - \frac{\rho}{d}X^\top X\right)^{-1}\right) \\
&\quad + \rho^2\sigma^6 \mathsf{Tr}\left(\Sigma\left(\Sigma - \frac{\rho}{d}X^\top X\right)^{-1} X^\top X\left(X^\top X + d\sigma^2 I\right)^{-2}\left(\Sigma - \frac{\rho}{d}X^\top X\right)^{-1}\right) \\
&= \frac{\rho^2\sigma^4}{d}\mathsf{Tr}\left(\Sigma\left(\Sigma - \frac{\rho}{d}X^\top X\right)^{-1} X^\top X\left(X^\top X + d\sigma^2 I\right)^{-2}\left(X^\top X + d\sigma^2 I\right)\left(\Sigma - \frac{\rho}{d}X^\top X\right)^{-1}\right) \\
&= \frac{\rho^2\sigma^4}{d}\mathsf{Tr}\left(\left(\Sigma - \frac{\rho}{d}X^\top X\right)^{-1}\Sigma\left(\Sigma - \frac{\rho}{d}X^\top X\right)^{-1} X^\top X\left(X^\top X + d\sigma^2 I\right)^{-1}\right).
\end{aligned}
$$

**Part III: Computing $\mathcal{T}(A(\rho,\Sigma);\Sigma)$.** Leveraging the expansion

$$
\mathcal{T}(A(\rho,\Sigma);\Sigma) = \frac{1}{nd}\left\|XA(\rho)X - X\right\|_F^2 + \frac{\sigma^2}{n}\left\|XA(\rho) - I\right\|_F^2
$$

$$= \frac{1}{nd}\mathsf{Tr}\left((XA(\rho)-I)XX^\top(XA(\rho)-I)^\top\right) + \frac{\sigma^2}{n}\mathsf{Tr}\left((XA(\rho)-I)(XA(\rho)-I)^\top\right)$$
$$= \frac{1}{nd}\mathsf{Tr}\left((XA(\rho)-I)\left(XX^\top + d\sigma^2 I\right)(XA(\rho)-I)^\top\right),$$

we can substitute the expression (22) for $XA(\rho)-I$ to obtain

$$\mathcal{T}(A(\rho,\Sigma);\Sigma) =$$
$$\frac{d\sigma^4}{n}\mathsf{Tr}\left(X\left(\Sigma - \frac{\rho}{d}X^\top X\right)^{-1}\Sigma X^\top\left(XX^\top\right)^{-1}\left(XX^\top + d\sigma^2 I\right)^{-1}\left(XX^\top\right)^{-1}X\Sigma\left(\Sigma - \frac{\rho}{d}X^\top X\right)^{-1}X^\top\right).$$

Leveraging the identity $X^\top(XX^\top + d\sigma^2 I)^{-1} = (X^\top X + d\sigma^2 I)^{-1}X^\top$ and that $(XX^\top)^{-1}$ and $(XX^\top + \lambda I)^{-1}$ commute, we have

$$\begin{aligned} X^\top(XX^\top)^{-1}(XX^\top + d\sigma^2 I)^{-1}(XX^\top)^{-1}X &= X^\top(XX^\top)^{-2}X(X^\top X + d\sigma^2 I)^{-1} \\ &= (X^\top X)^\dagger(X^\top X + d\sigma^2 I)^{-1}. \end{aligned}$$

Substituting this into the preceding display gives

$$\begin{aligned} &\mathcal{T}(A(\rho,\Sigma);\Sigma) \\ &= \frac{d\sigma^4}{n}\mathsf{Tr}\left(X\left(\Sigma - \frac{\rho}{d}X^\top X\right)^{-1}\Sigma(X^\top X)^\dagger(X^\top X + d\sigma^2 I)^{-1}\Sigma\left(\Sigma - \frac{\rho}{d}X^\top X\right)^{-1}X^\top\right) \\ &= \frac{d\sigma^4}{n}\mathsf{Tr}\left(\Sigma\left(\Sigma - \frac{\rho}{d}X^\top X\right)^{-1}X^\top X\left(\Sigma - \frac{\rho}{d}X^\top X\right)^{-1}\Sigma\left(X^\top X\right)^\dagger\left(X^\top X + d\sigma^2 I\right)^{-1}\right) \end{aligned}$$

by the cyclic property of the trace, as desired.

### B.4 Proof of Lemma 4.4

By Bai and Silverstein [4, Lemma 3.11] we can exactly compute

$$\begin{aligned} \lim_{y\to 0^+} m_H(-\sigma^2 + iy) &= \frac{1 - 1/\gamma + \sigma^2 - \sqrt{(1+1/\gamma+\sigma^2)^2 - 4/\gamma}}{-2\sigma^2/\gamma} \\ &= \frac{\sqrt{(1-1/\gamma+\sigma^2)^2 + 4\sigma^2/\gamma} - (1 - 1/\gamma + \sigma^2)}{2\sigma^2/\gamma} \\ &= \frac{2\sigma^2/\gamma + o(\sigma^2/\gamma)}{2\sigma^2/\gamma \cdot (1 - 1/\gamma + \sigma^2)}, \end{aligned}$$

completing the proof.

### B.5 Proof of Lemma 4.5

As the Bai-Yin law (Lemma A.2) guarantees the convergence of the smallest eigenvalue of $\frac{1}{n}XX^\top$ and $XX^\top$ is eventually non-singular, we can invoke the identities on the prediction and training error in Lemma 4.3. Therefore

$$\mathcal{P}(A(\rho,I);I) - \mathcal{P}(A(0,I);I) = \frac{\rho^2\sigma^4}{d}\mathsf{Tr}\left(\left(I - \frac{\rho}{d}X^\top X\right)^{-2}X^\top X\left(X^\top X + d\sigma^2 I\right)^{-1}\right)$$

$$= \frac{\rho^2\sigma^4}{d/n} \cdot \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\left(1-\rho\lambda_i^2/d\right)^2} \cdot \frac{\lambda_i^2}{d} \cdot \frac{1}{\lambda_i^2/d+\sigma^2}$$
$$= \frac{\rho^2}{d/n}\int \frac{\sigma^4 s}{(1-\rho s)^2(s+\sigma^2)} dH_n(s).$$

By the assumption that $\rho < \lambda_+^{-1}$, the Bai-Yin law (Lemma A.2) guarantees that $I - \frac{\rho}{d}XX^\top$ is eventually positive definite and with probability one $\lambda_1^2/d \to \lambda_+$. The function $s \mapsto \frac{\sigma^4 s}{(1-\rho s)^2(s+\sigma^2)}$ is thus eventually bounded on the support of $H_n$. Applying the Marchenko-Pastur law, we deduce

$$\lim_{n\to\infty} \left(\mathcal{P}(A(\rho, I); I) - \mathcal{P}(A(0, I); I)\right) = \frac{\rho^2}{\gamma}\int \frac{\sigma^4 s}{(1-\rho s)^2(s+\sigma^2)} dH(s).$$

For the second limit in Lemma 4.5, we can again leverage $\Sigma = I$ in Lemma 4.3 to compute

$$\begin{aligned}
\mathcal{T}(A(\rho, I); I) &= \frac{d\sigma^4}{n}\mathsf{Tr}\left(\left(I-\frac{\rho}{d}X^\top X\right)^{-1} X^\top X \left(I-\frac{\rho}{d}X^\top X\right)^{-1}\left(X^\top X\right)^\dagger \left(X^\top X + d\sigma^2 I\right)^{-1}\right) \\
&= \frac{\sigma^4}{n}\mathsf{Tr}\left(\left(I-\frac{\rho}{d}X^\top X\right)^{-1} \frac{X^\top X}{d} \left(I-\frac{\rho}{d}X^\top X\right)^{-1}\left(\frac{X^\top X}{d}\right)^\dagger \left(\frac{X^\top X}{d} + \sigma^2 I\right)^{-1}\right) \\
&= \sigma^4 \cdot \frac{1}{n}\sum_{i=1}^{n} \frac{1}{1-\rho\lambda_i^2/d} \cdot \frac{\lambda_i^2}{d} \cdot \frac{1}{1-\rho\lambda_i^2/d} \cdot \frac{1}{\lambda_i^2/d} \cdot \frac{1}{\lambda_i^2/d+\sigma^2} \\
&= \int \frac{\sigma^4}{(1-\rho s)^2(s+\sigma^2)} dH_n(s).
\end{aligned}$$

Applying the Marchenko-Pastur law gives the desired limit.

# C Proof of Theorem 2

We only need to prove under Assumption A2 thanks to Theorem 4. First, we recall our standard notation that $X$ has singular values $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$ and empirical spectral c.d.f. $H_n(s) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\lambda_i^2/d \leq s}$. We first prove (most of) part (i) of the theorem, which we state as a lemma. It is immediate by the definitions (3) and (4) of $\mathsf{Cost}$ and $\overline{\mathsf{Cost}}$ that $\mathsf{Cost}_X(\epsilon) - \overline{\mathsf{Cost}}_X(\epsilon) = \mathsf{Pred}_X(\widehat{\theta}_{\mathsf{ols}}) - \mathsf{Pred}_X(\widehat{\theta}(0))$, so we focus on the latter quantity.

**Lemma C.1.** *With probability 1*

$$\lim_{n\to\infty}\left(\mathsf{Pred}_X\left(\widehat{\theta}_{\mathsf{ols}}\right) - \mathsf{Pred}_X\left(\widehat{\theta}(0)\right)\right) = \frac{\sigma^2}{\gamma}\left(\int \frac{1}{s} dH(s) - \int \frac{1}{s+\sigma^2} dH(s)\right).$$

*Proof.* By the Bai-Yin law (Lemma A.2) we may assume that $XX^\top \succ 0$, as this eventually holds with probability 1. Let $\widehat{\theta}(0) = A(0, I)y$ for $A(0, I) = (X^\top X + d\sigma^2 I)^{-1}X^\top$ be the optimal unconstrained estimator (recall Lemma 4.2) and $\widehat{\theta}_{\mathsf{ols}} = A_{\mathsf{ols}}y$ for $A_{\mathsf{ols}} = X^\top(XX^\top)^{-1} = X^\dagger$. Then

$$\mathsf{Pred}_X\left(\widehat{\theta}_{\mathsf{OLS}}\right) - \mathsf{Pred}_X\left(\widehat{\theta}(0)\right) = \mathcal{P}(A_{\mathsf{ols}}; I) - \mathcal{P}(A(0, I); I). \tag{25}$$

We expand each of the prediction errors above in turn.

For the first, we have the identity

$$\mathcal{P}(A_{\mathsf{ols}}; I) = \frac{1}{d}\left\|A_{\mathsf{ols}}X - I\right\|_F^2 + \sigma^2\left\|A_{\mathsf{ols}}\right\|_F^2$$

$$= \frac{1}{d}\mathsf{Tr}\left(\left(X^\top(XX^\top)^{-1}X - I_d\right)^2\right) + \sigma^2\mathsf{Tr}\left(X^\top(XX^\top)^{-2}X\right) = \frac{d-n}{d} + \sigma^2\mathsf{Tr}\left((XX^\top)^{-1}\right),$$

where we have used that $X^\top(XX^\top)^{-1}X - I_d$ is a projection matrix of rank $d-n$. For the second,

$$\begin{aligned}
&\mathcal{P}(A(0,I);I) = \frac{1}{d}\left\|A(0,I)X - I\right\|_F^2 + \sigma^2\left\|A(0,I)\right\|_F^2 \\
&= \frac{1}{d}\mathsf{Tr}\left(\left(X^\top(XX^\top + d\sigma^2 I)^{-1}X - I\right)^2\right) + \sigma^2\mathsf{Tr}\left(X^\top\left(XX^\top + d\sigma^2 I\right)^{-2}X\right) \\
&\overset{(i)}{=} 1 + \frac{1}{d}\mathsf{Tr}\left((XX^\top)^2(XX^\top + d\sigma^2 I)^{-2} - 2XX^\top(XX^\top + d\sigma^2 I)^{-1}\right) + \sigma^2\mathsf{Tr}\left(XX^\top(XX^\top + d\sigma^2 I)^{-2}\right) \\
&= 1 + \frac{1}{d}\mathsf{Tr}\left(XX^\top\left(XX^\top - 2(XX^\top + d\sigma^2 I) + d\sigma^2 I\right)\left(XX^\top + d\sigma^2 I\right)^{-2}\right) \\
&= 1 + \frac{1}{d}\mathsf{Tr}\left(XX^\top\left(XX^\top + d\sigma^2 I\right)^{-1}\right),
\end{aligned}$$

where in step $(i)$ we use that $XX^\top$ and $(XX^\top + d\sigma^2 I)^{-1}$ commute and the cyclic property of the trace. Substituting these equalities into expression (25) yields

$$\mathsf{Pred}_X\left(\widehat{\theta}_{\mathsf{OLS}}\right) - \mathsf{Pred}_X\left(\widehat{\theta}(0)\right) = -\frac{n}{d} + \sigma^2\mathsf{Tr}\left((XX^\top)^{-1}\right) + \frac{1}{d}\mathsf{Tr}\left(XX^\top\left(XX^\top + d\sigma^2 I\right)^{-1}\right).$$

From this point, we expand the traces in terms of the empirical spectral distributions $H_n$, so multiplying and dividing $XX^\top$ by $d$ and normalizing the traces by $n$, we obtain

$$\mathsf{Pred}_X\left(\widehat{\theta}_{\mathsf{OLS}}\right) - \mathsf{Pred}_X\left(\widehat{\theta}(0)\right) = -\frac{n}{d} + \frac{\sigma^2 n}{d}\int\frac{1}{s}dH_n(s) + \frac{n}{d}\int\frac{s}{s+\sigma^2}dH_n(s).$$

We may apply the Bai-Yin law (Lemma A.2) and the Marchenko-Pastur law (Lemma A.1), so $\lambda_{\min}(XX^\top/d)$ converges with probability 1, and thus almost surely

$$\lim_{n\to\infty}\left(\mathsf{Pred}_X\left(\widehat{\theta}_{\mathsf{OLS}}\right) - \mathsf{Pred}_X\left(\widehat{\theta}(0)\right)\right) = -\frac{1}{\gamma} + \frac{\sigma^2}{\gamma}\int\frac{1}{s}dH(s) + \frac{1}{\gamma}\int\frac{s}{s+\sigma^2}dH(s).$$

An algebraic manipulation gives the lemma. □

Noting that $\frac{1}{s} - \frac{1}{s+\sigma^2} = \frac{\sigma^2}{s(s+\sigma^2)}$ gives the first equality of part (i) of the theorem. We divide the remainder of the proof into two parts. In the first, we perform an asymptotic expansion of the integral in Lemma C.1 to finalize part (i). In the second, we prove part (ii), including the existence of the threshold $\rho$ and the limiting values of $\overline{\mathsf{Cost}}_X(\epsilon)$.

**Finalizing Theorem 2 (i): The cost of minimum norm interpolation.** As in our derivation of Eq. (15), we can apply Bai and Silverstein [4, Lemma 3.11] to the integral form of Lemma C.1. Recalling Bai and Silverstein's result, we have

$$\int\frac{1}{s+\sigma^2}dH(s) = \frac{1 - 1/\gamma + \sigma^2 - \sqrt{\left(1 - 1/\gamma + \sigma^2\right)^2 + 4\sigma^2/\gamma}}{-2\sigma^2/\gamma}. \tag{26}$$

As

$$\left[\left(1-\frac{1}{\gamma}+\sigma^2\right)-\sqrt{\left(1-\frac{1}{\gamma}+\sigma^2\right)^2+\frac{4\sigma^2}{\gamma}}\right]\left[\left(1-\frac{1}{\gamma}+\sigma^2\right)+\sqrt{\left(1-\frac{1}{\gamma}+\sigma^2\right)^2+\frac{4\sigma^2}{\gamma}}\right]=\frac{4\sigma^2}{\gamma},$$

we then use that $H$ has support bounded away from zero to immediately obtain

$$\begin{aligned}\int\frac{1}{s}dH(s)&=\lim_{\sigma\downarrow 0}\frac{1-1/\gamma+\sigma^2-\sqrt{(1-1/\gamma+\sigma^2)^2+4\sigma^2/\gamma}}{-2\sigma^2/\gamma}\\&=\lim_{\sigma\downarrow 0}\frac{2}{\sqrt{(1-1/\gamma+\sigma^2)^2+4\sigma^2/\gamma}+(1-1/\gamma+\sigma^2)}=\frac{1}{1-1/\gamma}.\end{aligned}$$

As $\frac{\sigma^2}{s(s+\sigma^2)}=\frac{1}{s}-\frac{1}{s+\sigma^2}$, we then again use identity (26) and Lemma C.1 to see that

$$\begin{aligned}&\frac{\sigma^2}{\gamma}\cdot\left(\int\frac{1}{s}dH(s)-\int\frac{1}{s+\sigma^2}dH(s)\right)\\&=\frac{\sigma^2}{\gamma}\cdot\left(\frac{1}{1-1/\gamma}-\frac{2}{\sqrt{(1-1/\gamma+\sigma^2)^2+4\sigma^2/\gamma}+(1-1/\gamma+\sigma^2)}\right)\\&=\frac{\sigma^2}{\gamma}\cdot\left(\frac{1}{1-1/\gamma}-\frac{2}{(1-1/\gamma+\sigma^2)+\frac{4\sigma^2/\gamma}{2(1-1/\gamma+\sigma^2)}+(1-1/\gamma+\sigma^2)+o(\sigma^2)}\right)\\&=\frac{\sigma^2}{\gamma}\cdot\left(\frac{1}{1-1/\gamma}-\frac{1}{1-1/\gamma+\frac{\sigma^2}{1-1/\gamma}+o(\sigma^2)}\right)\\&=\frac{\sigma^4}{\gamma\,(1-1/\gamma)^3}+o(\sigma^4)\,,\end{aligned}$$

where we use the Taylor expansions $\sqrt{x^2+t}=x+\frac{t}{2x}+o(t^2)$ and $\frac{1}{x+t}=\frac{1}{x}-\frac{t}{x^2}+o(t^2)$, valid for any fixed $x>0$.

**Proving Theorem 2 (ii): interpolation threshold.** To obtain the threshold value $\rho_{\text{ols}}$, we derive the limit $\lim_{n\to\infty}\overline{\mathsf{Cost}}_X(\epsilon)$ for any $\epsilon>0$. As Lemma C.1 shows,

$$\lim_{n\to\infty}\left(\mathsf{Cost}_X(\epsilon)-\overline{\mathsf{Cost}}_X(\epsilon)\right)=\frac{\sigma^4}{\gamma}\int\frac{1}{s(s+\sigma^2)}dH(s).$$

Applying Theorem 1 for the limiting value of $\mathsf{Cost}_X(\epsilon)$, we recall the definition (7) of $\epsilon_\sigma^2=\int\frac{\sigma^4}{s+\sigma^2}dH(s)$. Choose $\rho=\rho(\epsilon)$ to be $\rho(\epsilon)=0$ if $\epsilon<\epsilon_\sigma$ and to satisfy $\epsilon^2=\int\frac{\sigma^4}{(1-\rho s)^2(s+\sigma^2)}dH(s)$ when $\epsilon\geq\epsilon_\sigma$, as in Eq. (8) in Theorem 1, which decreases continuously to $\rho(\epsilon_\sigma)=0$. The theorem then implies

$$\lim_{n\to\infty}\mathsf{Cost}_X(\epsilon)=\frac{\rho^2}{\gamma}\int\frac{\sigma^4 s}{(1-\rho s)^2(s+\sigma^2)}dH(s).$$

Adding and subtracting $\mathsf{Cost}_X(\epsilon)$, we therefore have with probability 1 that

$$\lim_{n\to\infty}\overline{\mathsf{Cost}}_X(\epsilon)=\lim_{n\to\infty}\mathsf{Cost}_X(\epsilon)-\lim_{n\to\infty}\left(\mathsf{Cost}_X(\epsilon)-\overline{\mathsf{Cost}}_X(\epsilon)\right)$$

$$= \frac{\rho^2}{\gamma} \int \frac{\sigma^4 s}{(1-\rho s)^2(s+\sigma^2)} dH(s) - \frac{\sigma^4}{\gamma} \int \frac{1}{s(s+\sigma^2)} dH(s) \tag{27}$$

(compare with Eq. (10)). Notably, $\rho = \rho(\epsilon)$ satisfies $\rho = 0$ whenever $\epsilon < \epsilon_\sigma$, so that

$$\lim_{n\to\infty} \overline{\mathsf{Cost}}_X(\epsilon) = -\frac{\sigma^4}{\gamma} \int \frac{1}{s(s+\sigma^2)} dH(s) < 0$$

for $\epsilon < \epsilon_\sigma$.

Now, consider the $\rho_{\mathrm{ols}}$ solving identity (10) and the associated value $\epsilon_{\sigma,\mathrm{ols}}$, where it is evident that $\rho_{\mathrm{ols}} > 0$. Then the preceding calculations yield immediately that

$$\lim_{n\to\infty} \overline{\mathsf{Cost}}_X(\epsilon_{\sigma,\mathrm{ols}}) = \frac{\sigma^4}{\gamma} \cdot \left( \rho_{\mathrm{ols}}^2 \int \frac{s}{(1-\rho_{\mathrm{ols}} s)^2 (s+\sigma^2)} dH(s) - \int \frac{1}{s(s+\sigma^2)} dH(s) \right) = 0.$$

Because the value $\rho = \rho(\epsilon)$ solving the identity (8) is increasing in $\epsilon \geq \epsilon_\sigma$, we conclude that $\rho(\epsilon) > \rho_{\mathrm{ols}}$ for $\epsilon > \epsilon_{\sigma,\mathrm{ols}}$ and $\epsilon_{\sigma,\mathrm{ols}} > \epsilon_\sigma$. Combining everything to this point and the limit (27), we see that

$$\lim_{n\to\infty} \overline{\mathsf{Cost}}_X(\epsilon) \begin{cases} > 0 & \text{if } \epsilon > \epsilon_{\sigma,\mathrm{ols}} \\ < 0 & \text{if } \epsilon < \epsilon_{\sigma,\mathrm{ols}}. \end{cases}$$

Lastly, we provide the concrete claimed bounds on $\epsilon_{\sigma,\mathrm{ols}}$ in terms of $\epsilon_\sigma$. We have already seen that $\epsilon_{\sigma,\mathrm{ols}} > \epsilon_\sigma$, and so the claimed upper bound revolves around lower bounding $\rho_{\mathrm{ols}}$ so that we may provide an upper bound on $\epsilon_{\sigma,\mathrm{ols}} = \int \frac{\sigma^4}{(1-\rho_{\mathrm{ols}} s)^2 (s+\sigma^2)} dH(s)$. To that end, note that identity (10) gives a lower bound for $\rho_{\mathrm{ols}}$: as

$$\int \left( \frac{\rho_{\mathrm{ols}}^2 s^2}{(1-\rho_{\mathrm{ols}} s)^2} - 1 \right) \cdot \frac{1}{s\,(s+\sigma^2)} dH(s) = 0\,,$$

we must have

$$\sup_{s\in[\lambda_-,\lambda_+]} \frac{\rho_{\mathrm{ols}}^2 s^2}{(1-\rho_{\mathrm{ols}} s)^2} - 1 \geq 0, \quad \text{so} \quad \rho_{\mathrm{ols}} \geq \frac{1}{2\lambda_+}.$$

Invoking the lower bound $\rho_{\mathrm{ols}} \cdot 2\lambda_+ \geq 1$ and that $s/\lambda_- \geq 1$ on the support of $H$, we have

$$\begin{aligned} \epsilon_{\sigma,\mathrm{ols}}^2 &= \int \frac{\sigma^4}{(1-\rho_{\mathrm{ols}} s)^2 (s+\sigma^2)} dH(s) \\ &\leq \frac{4\lambda_+^2}{\lambda_-} \cdot \sigma^4 \cdot \rho_{\mathrm{ols}}^2 \int \frac{s}{(1-\rho_{\mathrm{ols}} s)^2 (s+\sigma^2)} dH(s) = \frac{4\lambda_+^2 \sigma^4}{\lambda_-} \int \frac{1}{s(s+\sigma^2)} dH(s), \end{aligned}$$

where we used the identity (10). Noting that $\frac{1}{s} \leq \frac{1}{\lambda_-}$ and using the definition (7) of $\epsilon_\sigma = \int \frac{\sigma^4}{s+\sigma^2} dH(s)$ gives the final bound that $\epsilon_{\sigma,\mathrm{ols}}^2 \leq \frac{4\lambda_+^2}{\lambda_-^2} \epsilon_\sigma^2$, as desired.

# D Proof of Theorem 3

The proof follows a similar approach to that we use in the proof of Theorem 1 in Section 4: we compute formulae for the training and prediction errors conditional on the data matrices $X$, then use these to provide the bounds on the memorization threshold and costs for fitting to accuracy worse than that threshold. While in the proof of Theorem 1, we could develop explicit spectral limits for the error measures of interest, here exact forms are difficult, but we

can obtain tight enough bounds (mitigated by the condition number $\kappa$ of the covariance $\Sigma$ of the data vectors $x$) to give the desired results. With that in mind, we note that Lemmas 4.1, 4.2, and 4.3 all continue to hold, so that the reduction via strong duality applies. In particular, the optimal linear estimator $A$ in the form $\widehat{\theta} = Ay$ continues to take the form $A(\rho, \Sigma)$ in (14).

Throughout the proof, we let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$ denote the singular values of $X$ and $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_n \geq 0$ those of $Z$, and so the empirical spectral c.d.f.s of $\frac{1}{d}XX^\top$ and $\frac{1}{d}ZZ^\top$ are (respectively)

$$G_n(s) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\lambda_i^2/d \leq s} \quad \text{and} \quad H_n(s) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\mu_i^2/d \leq s}.$$

By the Marchenko-Pastur and deformed Marchenko-Pastur laws (Lemmas A.1 and A.3), $G_n$ and $H_n$ converge weakly (almost surely) to c.d.f.s $G$ and $H$, respectively. Again, we only need to prove under Assumption A2 by applying Theorem 4.

**Part I: Memorization threshold.** We begin with the expansion of $\epsilon_{\sigma,\mathrm{def}}$ and the bound $\epsilon^2_{\sigma,\mathrm{def}} \leq \epsilon^2_{\sqrt{\kappa}\sigma}/\kappa$. Rewriting $\epsilon_\sigma$ and $\epsilon_{\sigma,\mathrm{def}}$ in terms of the limits arising from their respective Marchenko-Pastur laws, we have

$$\epsilon^2_{\sigma,\mathrm{def}} = \int \frac{\sigma^4}{s+\sigma^2} dG(s) = \lim_{n\to\infty} \int \frac{\sigma^4}{s+\sigma^2} dG_n(s) = \lim_{n\to\infty} \frac{d\sigma^4}{n} \mathsf{Tr}\left((XX^\top + d\sigma^2 I)^{-1}\right),$$
$$\epsilon^2_{\sqrt{\kappa}\sigma}/\kappa = \int \frac{\kappa\sigma^4}{s+\kappa\sigma^2} dH(s) = \lim_{n\to\infty} \int \frac{\kappa\sigma^4}{s+\kappa\sigma^2} dH_n(s) = \lim_{n\to\infty} \frac{d\sigma^4}{n} \mathsf{Tr}\left((ZZ^\top/\kappa + d\sigma^2 I)^{-1}\right).$$

As $XX^\top = Z\Sigma Z^\top \succeq ZZ^\top/\kappa$, we have $\mathsf{Tr}(ZZ^\top/\kappa + d\sigma^2 I)^{-1} \geq \mathsf{Tr}(XX^\top + d\sigma^2 I)^{-1}$ and thus $\epsilon^2_{\sigma,\mathrm{def}} \leq \epsilon^2_{\sqrt{\kappa}\sigma}/\kappa$.

**Part II: No cost below threshold.** It is immediate via Lemma 4.2 that the global minimizer for the unconstrained problem (2) (with $\epsilon = 0$) is $A(0, \Sigma)$, that is, $\rho = 0$ as the constraint is inactive and

$$A(0,\Sigma) = X^\top(XX^\top + d\sigma^2 I)^{-1} = (X^\top X + d\sigma^2 I)^{-1} X^\top.$$

Then as usual $\inf_{\widehat{\theta}\in\mathcal{H}(0)} \mathsf{Pred}_X(\widehat{\theta}) = \mathsf{Pred}_X(\widehat{\theta}_{d\sigma^2})$, where we recall $\widehat{\theta}_{d\sigma^2}$ is the ridge estimator. To prove that $\lim_{n\to\infty} \mathsf{Cost}_X(\epsilon) = 0$ when $\epsilon < \epsilon_{\sigma,\mathrm{def}}$, it is thus sufficient to show that $\widehat{\theta}_{d\sigma^2}$ is contained in $\mathcal{H}(\epsilon)$ eventually, which amounts to proving

$$\liminf_{n\to\infty} \mathsf{Train}_X\left(\widehat{\theta}_{d\sigma^2}\right) = \liminf_{n\to\infty} \mathcal{T}(A(0,\Sigma);\Sigma) > \epsilon^2.$$

Invoking the expansion of $\mathcal{T}(A(\rho,\Sigma);\Sigma)$ in Lemma 4.3 and setting $\rho = 0$, we obtain

$$\mathcal{T}(A(0,\Sigma);\Sigma) = \frac{d\sigma^4}{n} \mathsf{Tr}\left(X^\top X \left(X^\top X\right)^\dagger \left(X^\top X + d\sigma^2 I\right)^{-1}\right) = \int \frac{\sigma^4}{s+\sigma^2} dG_n(s).$$

By weak convergence,

$$\lim_{n\to\infty} \mathcal{T}(A(0,\Sigma);\Sigma) = \lim_{n\to\infty} \int \frac{\sigma^4}{s+\sigma^2} dG_n(s) = \int \frac{\sigma^4}{s+\sigma^2} dG(s) = \epsilon^2_{\sigma,\mathrm{def}} > \epsilon^2,$$

so indeed we have $\widehat{\theta}_{d\sigma^2} \in \mathcal{H}(\epsilon)$ as desired.

**Part III: Cost of not-fitting above threshold.** Our starting point is to demonstrate the existence and uniqueness of $\rho_{\text{def}} \in \left[0, \lambda_+^{-1}\right)$ solving the identity (12). For this, we note that the difference

$$\Delta_H(\rho) := \int \left[ \frac{1}{(1-\rho s)^2 (s + \kappa\sigma^2)} - \frac{1}{s+\sigma^2} \right] dH(s)$$

is monotone increasing in $\rho$, and $\Delta_H(0) = 0$. That $\Delta_H(\rho) \to \infty$ as $\rho \uparrow \lambda_+^{-1}$ is then an immediate consequence of the expansion (16) of the left integrand above.

We turn to the second claim in part (iii): the lower bound on $\mathsf{Cost}_X(\epsilon)$. We (roughly) reduce the general covariance case to the isotropic case, then apply our previous results and techniques. To do so, we require the following lemma, which upper-bounds the training error growth and lower-bounds the prediction error growth. The proof is essentially tedious algebraic manipulations, so we defer it to Appendix D.1.

**Lemma D.1.** *Let the same conditions of Lemma 4.3 hold and assume $\rho\lambda_1^2/d < 1$. Then*

$$\mathcal{P}(A(\rho,\Sigma);\Sigma) - \mathcal{P}(A(0,\Sigma);\Sigma) \geq \frac{\rho^2\sigma^4}{d} \mathsf{Tr}\left( \left(I - \frac{\rho}{d} ZZ^\top\right)^{-2} \frac{ZZ^\top}{d} \cdot \left( \frac{ZZ^\top}{d} + \sigma^2 I \right)^{-1} \right),$$

$$\mathcal{T}(A(\rho,\Sigma);\Sigma) - \mathcal{T}(A(0,\Sigma);\Sigma) \leq \frac{\kappa\sigma^4}{n} \mathsf{Tr}\left[ \left( \left(I - \frac{\rho}{d} ZZ^\top\right)^{-2} - I \right) \left( \frac{1}{d} ZZ^\top + \kappa\sigma^2 I \right)^{-1} \right].$$

We use the upper and lower bounds in Lemma D.1, coupled with the strong duality guarantees in Lemma 4.2 (and the identities (14)), to prove the desired growth of the $\mathsf{Cost}_X(\epsilon)$. Consider any $0 \leq \rho < \rho_{\text{def}}$, where $\rho_{\text{def}}$ satisfies the identity (12). By construction and duality, $A(\rho,\Sigma)$ is the optimal solution to the problem

$$\begin{aligned} \underset{A \in \mathbb{R}^{d\times n}}{\text{minimize}} \quad & \mathcal{P}(A;\Sigma) \\ \text{subject to} \quad & \mathcal{T}(A(\rho,\Sigma);\Sigma) - \mathcal{T}(A;\Sigma) \leq 0. \end{aligned}$$

Thus, whenever $\mathcal{T}(A(\rho,\Sigma);\Sigma) < \epsilon^2$ it holds that

$$\mathsf{Cost}_X(\epsilon) \geq \mathcal{P}(A(\rho,\Sigma);\Sigma) - \mathcal{P}(A(0,\Sigma);\Sigma). \tag{28}$$

Therefore, to prove that $\mathsf{Cost}_X(\epsilon)$ grows it is sufficient to show that eventually $\mathcal{T}(A(\rho,\Sigma);\Sigma) < \epsilon$ for our chosen $\rho$ and provide lower bounds on the difference $\mathcal{P}(A(\rho,\Sigma);\Sigma) - \mathcal{P}(A(0,\Sigma);\Sigma)$.

To that end, let us take limits of $\mathcal{T}$. Applying the upper bound in Lemma D.1, we have

$$\begin{aligned} &\limsup_{n\to\infty} \mathcal{T}(A(\rho,\Sigma);\Sigma) \\ &\leq \limsup_{n\to\infty} \mathcal{T}(A(0,\Sigma);\Sigma) + \limsup_{n\to\infty} \frac{\kappa\sigma^4}{n} \mathsf{Tr}\left[ \left( \left(I - \frac{\rho}{d} ZZ^\top\right)^{-2} - I \right) \left( \frac{1}{d} ZZ^\top + \kappa\sigma^2 I \right)^{-1} \right] \\ &= \epsilon_{\sigma,\text{def}}^2 + \limsup_{n\to\infty} \kappa\sigma^4 \int \frac{\rho s(2-\rho s)}{(1-\rho s)^2 (s+\kappa\sigma^2)} dH_n(s) \end{aligned}$$

with probability 1. As $\rho < \rho_{\text{def}} < \lambda_+^{-1}$, the quantity $\frac{s(2-\rho s)}{(1-\rho s)^2(s+\kappa\sigma^2)}$ is eventually bounded on the support $[\lambda_-, \lambda_+] + o(1)$ of $H_n$ by the Bai-Yin law (Lemma A.2), and so with probability one

$$\kappa\sigma^4 \int \frac{\rho s(2-\rho s)}{(1-\rho s)^2 (s+\kappa\sigma^2)} dH_n(s) \to \kappa\sigma^4 \int \frac{\rho s(2-\rho s)}{(1-\rho s)^2 (s+\kappa\sigma^2)} dH(s)$$

$$
\begin{aligned}
&= \kappa\sigma^4 \int \left( \frac{1}{(1-\rho s)^2 (s+\kappa\sigma^2)} - \frac{1}{s+\kappa\sigma^2} \right) dH(s) \\
&< \kappa\sigma^4 \int \left( \frac{1}{(1-\rho_{\mathrm{def}} s)^2 (s+\kappa\sigma^2)} - \frac{1}{s+\kappa\sigma^2} \right) dH(s) \\
&= \epsilon^2 - \epsilon^2_{\sigma,\mathrm{def}},
\end{aligned}
$$

where the last line follows from the definition (12) of $\rho_{\mathrm{def}}$. In particular, with probability 1 we have

$$
\limsup_{n\to\infty} \mathcal{T}(A(\rho,\Sigma);\Sigma) < \epsilon^2_{\sigma,\mathrm{def}} + \epsilon^2 - \epsilon^2_{\sigma,\mathrm{def}} = \epsilon^2,
$$

and therefore inequality (28) implies that with probability 1,

$$
\liminf_{n\to\infty} \mathsf{Cost}_X(\epsilon) \geq \liminf_{n\to\infty} \left[ \mathcal{P}(A(\rho,\Sigma);\Sigma) - \mathcal{P}(A(0,\Sigma);\Sigma) \right].
$$

We now apply Lemma D.1 again, invoking the lower bound on the prediction errors to obtain

$$
\begin{aligned}
\liminf_{n\to\infty} \mathsf{Cost}_X(\epsilon) &\geq \lim_{n\to\infty} \frac{\rho^2\sigma^4}{d} \mathsf{Tr}\left( \left( I - \frac{\rho}{d} ZZ^\top \right)^{-2} \frac{ZZ^\top}{d} \cdot \left( \frac{ZZ^\top}{d} + \sigma^2 I \right)^{-1} \right) \\
&= \lim_{n\to\infty} \rho^2\sigma^4 \cdot \frac{n}{d} \cdot \int \frac{s}{(1-\rho s)^2 (s+\sigma^2)} dH_n(s) \\
&= \frac{\rho^2}{\gamma} \int \frac{\sigma^4 s}{(1-\rho s)^2 (s+\sigma^2)} dH(s).
\end{aligned}
$$

Taking $\rho \uparrow \rho_{\mathrm{def}}$ yields the second claim of part (iii).

Our last step is to prove a concrete lower bound showing that $\mathsf{Cost}_X(\epsilon)$ grows linearly in $\epsilon^2$ provided that $\epsilon^2 \geq \frac{2\kappa\sigma^4}{\lambda_- + \kappa\sigma^2}$, in parallel to the result in Lemma 4.6. We state a small integral inequality:

**Lemma D.2.** *Let $\rho = \rho_{\mathrm{def}}$ solve the fixed point (12). Then*

$$
\int \frac{\kappa\sigma^4}{(1-\rho s)^2 (s+\kappa\sigma^2)} dH(s) \geq \epsilon^2.
$$

*Proof.* The identity (12) shows that the integral in the statement of the lemma equals $\int \frac{\kappa\sigma^4}{s+\kappa\sigma^2} dH(s) + \epsilon^2 - \epsilon^2_{\sigma,\mathrm{def}}$. Recall that by part (i) of Theorem 3, we have $\epsilon^2_{\sigma,\mathrm{def}} \leq \epsilon^2_{\sqrt{\kappa}\sigma}/\kappa = \int \frac{\kappa\sigma^4}{s+\kappa\sigma^2} dH(s)$. ☐

Taking $\rho = \rho_{\mathrm{def}}$ to solve the fixed point (12), we apply the second claim in part (iii) to see that

$$
\begin{aligned}
\liminf_{n\to\infty} \mathsf{Cost}_X(\epsilon) &\geq \frac{\rho^2}{\gamma} \int \frac{\sigma^4 s}{(1-\rho s)^2 (s+\sigma^2)} dH(s) \geq \frac{\rho^2\lambda_-}{\gamma} \int \frac{\sigma^4}{(1-\rho s)^2 (s+\sigma^2)} dH(s) \\
&\geq \frac{\rho^2\lambda_-}{\kappa\gamma} \int \frac{\kappa\sigma^4}{(1-\rho s)^2 (s+\kappa\sigma^2)} dH(s) \geq \frac{\rho^2\lambda_-}{\kappa\gamma} \epsilon^2 \qquad (29)
\end{aligned}
$$

by Lemma D.2. It remains to lower bound $\rho = \rho_{\mathrm{def}} < \lambda_+^{-1}$. For this, we observe that

$$
\frac{1}{(1-\rho\lambda_+)^2} \geq \int \frac{1}{(1-\rho s)^2} dH(s) \geq \frac{\lambda_- + \kappa\sigma^2}{\kappa\sigma^4} \int \frac{\kappa\sigma^4}{(1-\rho s)^2 (s+\kappa\sigma^2)} dH(s) \geq \frac{\lambda_- + \kappa\sigma^2}{\kappa\sigma^4} \cdot \epsilon^2,
$$

again applying Lemma D.2. In particular, whenever $\frac{\lambda_+ + \kappa\sigma^2}{\kappa\sigma^4}\epsilon^2 \geq 2$, we obtain $(1-\rho\lambda_+)^{-2} \geq 2$, or $\rho_{\mathrm{def}} \geq \frac{1}{\lambda_+}(1 - 1/\sqrt{2})$. Substituting in inequality (29) gives the lower bound on $\liminf_n \mathsf{Cost}_X(\epsilon)$.

## D.1 Proof of Lemma D.1

We prove each claim of the lemma in turn. For the first, we use the shorthand $\Delta_{\mathcal{P}}(\rho) := \mathcal{P}(A(\rho,\Sigma);\Sigma) - \mathcal{P}(A(0,\Sigma);\Sigma)$. Then applying Lemma 4.3, we have

$$\Delta_{\mathcal{P}}(\rho) = \frac{\rho^2\sigma^4}{d}\mathsf{Tr}\left(\left(\Sigma - \frac{\rho}{d}X^\top X\right)^{-1}\Sigma\left(\Sigma - \frac{\rho}{d}X^\top X\right)^{-1}X^\top\left(XX^\top + d\sigma^2 I\right)^{-1}X\right),$$

and making the substitution $X = Z\Sigma^{\frac{1}{2}}$ immediately yields

$$\begin{aligned}
&\Delta_{\mathcal{P}}(\rho)\\
&= \frac{\rho^2\sigma^4}{d}\mathsf{Tr}\left(\left(\Sigma - \frac{\rho}{d}\Sigma^{\frac{1}{2}}Z^\top Z\Sigma^{\frac{1}{2}}\right)^{-1}\Sigma\left(\Sigma - \frac{\rho}{d}\Sigma^{\frac{1}{2}}Z^\top Z\Sigma^{\frac{1}{2}}\right)^{-1}\Sigma^{\frac{1}{2}}Z^\top\left(Z\Sigma Z^\top + d\sigma^2 I\right)^{-1}Z\Sigma^{\frac{1}{2}}\right)\\
&= \frac{\rho^2\sigma^4}{d}\mathsf{Tr}\left(Z\left(I - \frac{\rho}{d}Z^\top Z\right)^{-2}Z^\top\cdot\left(Z\Sigma Z^\top + d\sigma^2 I\right)^{-1}\right).
\end{aligned}$$

As $Z(I - \frac{\rho}{d}Z^\top Z)^{-2}Z^\top \succeq 0$ and $Z\Sigma Z^\top + d\sigma^2 I \preceq ZZ^\top + d\sigma^2 I$ as $\Sigma \preceq I$ by assumption, we can leverage that the mapping $A \mapsto \mathsf{Tr}(AC)$ is increasing in the positive definite order for $C \succeq 0$ to obtain that

$$\begin{aligned}
\Delta_{\mathcal{P}}(\rho) &\geq \frac{\rho^2\sigma^4}{d}\mathsf{Tr}\left(Z\left(I - \frac{\rho}{d}Z^\top Z\right)^{-2}Z^\top\cdot\left(ZZ^\top + d\sigma^2 I\right)^{-1}\right)\\
&= \frac{\rho^2\sigma^4}{d}\mathsf{Tr}\left(\left(I - \frac{\rho}{d}ZZ^\top\right)^{-2}\frac{ZZ^\top}{d}\cdot\left(\frac{ZZ^\top}{d} + \sigma^2 I\right)^{-1}\right),
\end{aligned}$$

where in the last line we used the identity $Z(I - \frac{\rho}{d}Z^\top Z)^{-1} = (I - \frac{\rho}{d}ZZ^\top)^{-1}Z$. This gives the first claim of Lemma D.1.

We turn to the upper bound on the training error, for which we use the shorthand $\Delta_{\mathcal{T}}(\rho) := \mathcal{T}(A(\rho,\Sigma);\Sigma) - \mathcal{T}(A(0,\Sigma);\Sigma)$. Beginning from the expansion of $\mathcal{T}$ in Lemma 4.3, we have

$$\begin{aligned}
\frac{n}{d\sigma^4}\Delta_{\mathcal{T}}(\rho) = \mathsf{Tr}&\left[\Sigma\left(\Sigma - \frac{\rho}{d}X^\top X\right)^{-1}X^\top X\left(\Sigma - \frac{\rho}{d}X^\top X\right)^{-1}\Sigma\left(X^\top X\right)^\dagger\left(X^\top X + d\sigma^2 I\right)^{-1}\right]\\
&- \mathsf{Tr}\left[X^\top X(X^\top X)^\dagger(X^\top X + d\sigma^2 I)^{-1}\right].
\end{aligned} \tag{30}$$

Leveraging the identities $X = Z\Sigma^{\frac{1}{2}}$ and that

$$\begin{aligned}
(X^\top X)^\dagger(X^\top X + d\sigma^2 I)^{-1} &= X^\top(XX^\top)^{-2}(XX^\top + d\sigma^2 I)^{-1}X\\
&= \Sigma^{\frac{1}{2}}Z^\top(Z\Sigma Z^\top)^{-2}(Z\Sigma Z^\top + d\sigma^2 I)^{-1}Z\Sigma^{\frac{1}{2}},
\end{aligned}$$

the right hand side of the expansion (30) becomes

$$\begin{aligned}
&\mathsf{Tr}\left[\left(\Sigma^{\frac{1}{2}}\left(I - \frac{\rho}{d}Z^\top Z\right)^{-1}Z^\top Z\left(I - \frac{\rho}{d}Z^\top Z\right)^{-1}\Sigma^{\frac{1}{2}} - \Sigma^{\frac{1}{2}}Z^\top Z\Sigma^{\frac{1}{2}}\right)X^\top(XX^\top)^{-2}(XX^\top + d\sigma^2 I)^{-1}X\right]\\
&= \mathsf{Tr}\left[\Sigma^{\frac{1}{2}}\left(\left(I - \frac{\rho}{d}Z^\top Z\right)^{-1}Z^\top Z\left(I - \frac{\rho}{d}Z^\top Z\right)^{-1} - Z^\top Z\right)\Sigma Z^\top(Z\Sigma Z^\top)^{-2}(Z\Sigma Z^\top + d\sigma^2 I)^{-1}Z\Sigma^{\frac{1}{2}}\right]\\
&= \mathsf{Tr}\left[\Sigma^{\frac{1}{2}}\left(\left(I - \frac{\rho}{d}Z^\top Z\right)^{-2} - I\right)Z^\top(Z\Sigma Z^\top)^{-1}\left(Z\Sigma Z^\top + d\sigma^2 I\right)^{-1}Z\Sigma^{\frac{1}{2}}\right],
\end{aligned}$$

where we have used that $(I - \frac{\rho}{d}Z^\top Z)^{-1}$ and $Z^\top Z$ commute and eliminated one inverse of $Z\Sigma Z^\top$. The singular value decomposition gives the equality $(I - \frac{\rho}{d}Z^\top Z)^{-2}Z^\top = Z^\top(I - \frac{\rho}{d}ZZ^\top)^{-2}$, where $I$ is an identity matrix of appropriate size. The cyclic property of the trace and that $(Z\Sigma Z^\top)^{-1}$ and $(Z\Sigma Z^\top + d\sigma^2 I)^{-1}$ comute then allows us to substitute into the identity (30) to obtain

$$\Delta_{\mathcal{T}}(\rho) = \frac{d\sigma^4}{n}\mathsf{Tr}\left[\left(\left(I - \frac{\rho}{d}ZZ^\top\right)^{-2} - I\right)\left(Z\Sigma Z^\top + d\sigma^2 I\right)^{-1}\right].$$

Lastly, we again use the monotonicity of $A \mapsto \mathsf{Tr}(AC)$ for $C \succeq 0$ and that $Z\Sigma Z^\top + d\sigma^2 I \succeq ZZ^\top/\kappa + d\sigma^2 I$ to get claimed upper bound in the lemma.

# E Proof of Theorem 4

We provide the proof conditional on $X$, implicitly conditioning throughout. As $\mathcal{H}_{\mathrm{lin}}(\epsilon) \subset \mathcal{H}_{\mathrm{sq}}(\epsilon)$, we only need to show

$$\inf_{\widehat{\theta} \in \mathcal{H}_{\mathrm{sq}}(\epsilon)} \mathsf{Pred}_X\left(\widehat{\theta}\right) \geq \min_{\widehat{\theta} \in \mathcal{H}_{\mathrm{lin}}(\epsilon)} \mathsf{Pred}_X\left(\widehat{\theta}\right).$$

First we note that in the Gaussian setting that $\theta \sim \mathsf{N}(0, \frac{1}{d}I)$, we have $y = X\theta + \varepsilon \sim \mathsf{N}(0, \frac{XX^\top}{d} + \sigma^2 I)$. By a standard calculation, the conditional distribution of $\theta$ given $y$ is

$$\theta \mid y \sim \mathsf{N}\left(\left(X^\top X + d\sigma^2 I\right)^{-1} X^\top y, \sigma^2\left(X^\top X + d\sigma^2 I\right)^{-1}\right),$$

and therefore for any $\widehat{\theta}(X, y) \in \mathcal{H}_{\mathrm{sq}}$,

$$\begin{aligned}
\mathsf{Pred}_X\left(\widehat{\theta}\right) &= \mathbb{E}_y\left[\mathbb{E}_{\theta|y}\left[\left\|\Sigma^{\frac{1}{2}}\left(\widehat{\theta} - \theta\right)\right\|_2^2 \mid y\right]\right] \\
&= \mathbb{E}_y\left[\left\|\Sigma^{\frac{1}{2}}\left(\widehat{\theta} - \left(X^\top X + d\sigma^2 I\right)^{-1} X^\top y\right)\right\|_2^2 + \sigma^2 \mathsf{Tr}\left(\Sigma\left(X^\top X + d\sigma^2 I\right)^{-1}\right)\right].
\end{aligned}$$

Notably, the posterior mean $\mathbb{E}[\theta \mid y]$ always minimizes the prediction risk. By Lemma 4.2 we know there is a $\rho$ such that $\widehat{\theta}(\rho) := A(\rho, \Sigma)y$ is optimal for problem (2) where

$$A(\rho, \Sigma) = \left(I - \rho\sigma^2\left(\Sigma - \frac{\rho}{d}X^\top X\right)^{-1}\right)(X^\top X + d\sigma^2 I)^{-1} X^\top.$$

We consider two cases, depending on whether the value of the dual variable $\rho = 0$ or $\rho > 0$.

**Case I:** $\rho = 0$**.** In this case $\widehat{\theta}(0) = \left(X^\top X + d\sigma^2 I\right)^{-1} X^\top y \in \mathcal{H}_{\mathrm{lin}}(\epsilon) \subset \mathcal{H}_{\mathrm{sq}}(\epsilon)$. But this is the posterior mean, that is, $\widehat{\theta}(0) = \mathbb{E}[\theta \mid y]$, which is thus optimal.

**Case II:** $\rho > 0$**.** As $\mathsf{Pred}_X(\widehat{\theta}(\rho))$ is continuous in $\rho$, if we can prove for any $\widehat{\theta} \in \mathcal{H}_{\mathrm{sq}}(\epsilon)$ and any $0 \leq \overline{\rho} < \rho$ that

$$\mathsf{Pred}_X\left(\widehat{\theta}\right) \geq \mathsf{Pred}_X\left(\widehat{\theta}(\overline{\rho})\right), \tag{31}$$

taking $\overline{\rho} \uparrow \rho$ completes the proof. (Note that $\rho$ is the optimal dual variable for problem (2), and so $\widehat{\theta}(\overline{\rho}) \in \mathcal{F}(\epsilon)$.)

To show claim (31), let $\mu = \mathsf{N}(0, \frac{1}{d} XX^\top + \sigma^2 I)$ be the marginal distribution over $y$. We construct a sequence of random measures $\mu_1, \mu_2, \cdots$, by sampling $y_i \overset{\text{iid}}{\sim} \mu$ and constructing the empirical measure

$$\mu_m = \frac{1}{m} \sum_{i=1}^{m} \delta_{y_i}.$$

In this case the optimization problem

$$\begin{aligned} \underset{\widehat{\theta}(X,y_i) \in \mathbb{R}^d, 1 \le i \le m}{\text{minimize}} \quad & \int \left\| \Sigma^{\frac{1}{2}} \left( \widehat{\theta} - \left( X^\top X + d\sigma^2 I \right)^{-1} X^\top y \right) \right\|_2^2 d\mu_m \\ \text{subject to} \quad & \int \left\| X\widehat{\theta} - y \right\|_2^2 d\mu_m \ge \int \left\| X\widehat{\theta}(\overline{\rho}) - y \right\|_2^2 d\mu_m \end{aligned}$$

is a finite dimensional optimization problem with (strongly convex) quadratic objective and a single quadratic constraint. Then strong duality obtains [10, Appendix B.1], so we can write the stationary condition that for some $\lambda \ge 0$,

$$\Sigma \left( \widehat{\theta}(X, y_i) - \left( X^\top X + d\sigma^2 I \right)^{-1} X^\top y_i \right) - \lambda X^\top (X\widehat{\theta}(X, y_i) - y_i) = 0$$

simultaneously for $i = 1, \ldots, m$. Rewriting gives

$$\left( \Sigma - \lambda X^\top X \right) \widehat{\theta}(X, y_i) = \left( \Sigma \left( X^\top X + d\sigma^2 I \right)^{-1} - \lambda I \right) X^\top y_i, \quad \text{for } i = 1, \ldots, m.$$

By an identical argument to that we use to prove Lemma 4.2 in Appendix B.2, it must be the case that $\Sigma - \lambda X^\top X \succ 0$ and thus for each $i = 1, \ldots, m$,

$$\begin{aligned} \widehat{\theta}(X, y_i) &= \left( \Sigma - \lambda X^\top X \right)^{-1} \left( \Sigma - \lambda X^\top X - \lambda d\sigma^2 I \right) \left( X^\top X + d\sigma^2 I \right)^{-1} X^\top y_i \\ &= \left( I - \lambda d\sigma^2 \left( \Sigma - \lambda X^\top X \right)^{-1} \right) \left( X^\top X + d\sigma^2 I \right)^{-1} X^\top y_i. \end{aligned}$$

By inspection, this estimator is linear in $y$, and for the choice $\lambda = \frac{\overline{\rho}}{d}$ takes identical values at $y_1, \ldots, y_m$ as $\widehat{\theta}(\overline{\rho})$. The constraints of the problem (32) are satisfied and the KKT conditions hold, so (an) optimal solution is $\widehat{\theta}(\overline{\rho})$.

For any $\widehat{\theta} \in \mathcal{H}_{\text{sq}}(\epsilon)$, whenever the training errors satisfy

$$\int \left\| X\widehat{\theta} - y \right\|_2^2 d\mu_m \ge \int \left\| X\widehat{\theta}(\overline{\rho}) - y \right\|_2^2 d\mu_m,$$

we must have

$$\int \left\| \Sigma^{\frac{1}{2}} \left( \widehat{\theta} - \mathbb{E}[\theta \mid y] \right) \right\|_2^2 d\mu_m \ge \int \left\| \Sigma^{\frac{1}{2}} \left( \widehat{\theta}(\overline{\rho}) - \mathbb{E}[\theta \mid y] \right) \right\|_2^2 d\mu_m. \tag{32}$$

By the law of large numbers, if $\widehat{\theta}$ is square integrable, then with probability one

$$\lim_{m \to \infty} \int \left\| X\widehat{\theta} - y \right\|_2^2 d\mu_m = \int \left\| X\widehat{\theta} - y \right\|_2^2 d\mu \ge \epsilon^2 \overset{(\star)}{>} \int \left\| X\widehat{\theta}(\overline{\rho}) - y \right\|_2^2 d\mu = \lim_{m \to \infty} \int \left\| X\widehat{\theta} - y \right\|_2^2 d\mu_m,$$

where inequality $(\star)$ holds by the assumption that $\overline{\rho} < \rho = \rho(\epsilon)$, yielding the difference in training errors. Thus Eq. (32) holds eventually for all large $m$. Again applying the law of large numbers and taking $m \to \infty$, we establish the desired prediction error gap (31).