Spurious Correlations in High Dimensional Regression: The Roles of Regularization, Simplicity Bias and Over-Parameterization

Anonymous authors
Paper under double-blind review

Abstract

Learning models have been shown to rely on spurious correlations between non-predictive features and the associated labels in the training data, with negative implications on robustness, bias and fairness. In this work, we provide a statistical characterization of this phenomenon for high-dimensional regression, when the data contains a predictive *core* feature x and a *spurious* feature y. Specifically, we quantify the amount of spurious correlations $\mathcal{C}$ learned via linear regression, in terms of the data covariance and the strength λ of the ridge regularization. As a consequence, we first capture the simplicity of y through the spectrum of its covariance, and its correlation with x through the Schur complement of the full data covariance. Next, we prove a trade-off between $\mathcal{C}$ and the in-distribution test loss $\mathcal{L}$, by showing that the value of λ that minimizes $\mathcal{L}$ lies in an interval where $\mathcal{C}$ is increasing. Finally, we investigate the effects of over-parameterization via the random features model, by showing its equivalence to regularized linear regression. Our theoretical results are supported by numerical experiments on Gaussian, Color-MNIST, and CIFAR-10 datasets.

1 Introduction

Machine learning systems have been shown to learn from patterns that are statistically correlated with the intended task, despite not being causally predictive Geirhos et al. (2020); Xiao et al. (2021). As a concrete example, a blue background in a picture might be positively correlated with the presence of a boat in the foreground, and while not being a predictive feature per se, a trained deep learning model could use this information to bias its prediction. In the literature, this statistical (but non causal) connection is referred to as a *spurious correlation* between a feature and the learning task. A recent and extensive line of research has investigated the extent to which deep learning models manifest this behavior Geirhos et al. (2019); Xiao et al. (2021) and has proposed different mitigation approaches Sagawa et al. (2020a); Liu et al. (2021), given its implications to robustness, bias, and fairness Zliobaite (2015); Zhou et al. (2021). The phenomenon, also referred to as shortcut learning, is often attributed to the relative "simplicity" of spurious features Geirhos et al. (2020); Shah et al. (2020); Hermann & Lampinen (2020) and to the implicit bias of over-parameterized models toward learning simpler patterns Belkin et al. (2019); Rahaman et al. (2019); Kalimeris et al. (2019). Consequently, the *core features* that are informative about the task (e.g., the boat in the foreground) may be neglected, as *spurious features* (e.g., the blue background) provide an easier shortcut to minimize the loss function.

Figure 1: *Left two panels:* pictorial representation of the core (spurious) feature x (y) and an independent core feature $\tilde{x}$, taken from an image of a boat and a truck in the CIFAR-10 dataset. *Right two panels:* examples from a binary Color-MNIST dataset, where the labels correspond to the number shapes, and the zeros (ones) are colored in blue (red) with probability $(1+\alpha)/2$.

Prior work has attempted to formalize the *simplicity bias* relying on boolean functions Qiu et al. (2024), model-specific biases Morwani et al. (2023), one-dimensional features Shah et al. (2020) and their pairwise interactions Pezeshki et al. (2021). However, when considering high-dimensional natural data (e.g., the boat and its background in Figure 1), it remains unclear, based on these notions, what exactly makes the features easy or difficult to learn, and to what extent a trained model relies on spurious correlations. Furthermore, while Sagawa et al. (2020b) show that over-parameterization can exacerbate spurious correlations when re-weighting the objective on minority groups (e.g., boats with a green background), its effect on models trained via empirical risk minimization (ERM) is less understood. This is a critical point when additional group membership annotations are too expensive to obtain, and ERM is a key part of training Liu et al. (2021); Ahmed et al. (2021).

Our work tackles these issues: we provide a rigorous characterization of the statistical mechanisms behind learning spurious correlations in high-dimensional data, focusing on the solution obtained via ERM. Formally, we model the input sample z as composed by two distinct features, *i.e.*, $z = [x, y]$, where $x \in \mathbb{R}^d$ is the core feature and $y \in \mathbb{R}^d$ the spurious one. The first panel of Figure 1 provides an illustration with a boat in the foreground (x) and its blue background (y). Then, we quantify spurious correlations via the covariance $\mathcal{C}$ (see (2.3)) between the label g ("boat") and the model output given $\tilde{z} = [\tilde{x}, y]$ as input. Here, $\tilde{x}$ (a truck in the foreground) is a new core feature independent of everything else, see the second panel of Figure 1. Now, if $\mathcal{C}$ is positive, it means that the model is biased towards g only because of y, since $\tilde{x}$ is independent from x and g. More precisely, we provide a sharp, non-asymptotic characterization of $\mathcal{C}$ for linear regression (Theorem 1). Armed with such a characterization, we then:

• Interpret $\mathcal{C}$ via upper bounds on its magnitude (Proposition 4.1). This highlights the role of the regularization strength and of the data covariance via *(i)* its Schur complement with respect to the covariance of the core feature x, and *(ii)* the covariance of the spurious feature y. Specifically, we link the smallest eigenvalue of the Schur complement to the strength of the correlation between y and x, and the largest eigenvalue of the spurious covariance to the simplicity of y.

• Prove a trade-off between $\mathcal{C}$ and the test loss (Proposition 4.3), which implies that spurious correlations can be beneficial to performance when learning in-distribution. Specifically, we show that the optimal regularization minimizing the test loss lies in an interval where $\mathcal{C}$ is positive and monotonically increasing.

• Investigate the role of over-parameterization via a *random features* (RF) model. Specifically, we show that the RF model is equivalent to linear regression with an effective regularization that depends on the over-parameterization (Theorem 2). This allows to leverage the earlier analysis on regularized linear regression to quantify spurious correlations in over-parameterized, non-linear models.

Throughout the paper, the theoretical results are supported by numerical experiments on Gaussian data, Color-MNIST, and CIFAR-10, which validates our analysis even in settings not strictly following the modeling choices. Additional discussion about the related work is deferred to Appendix B.

2 PRELIMINARIES

Notation. Given a vector v, we denote by $\|v\|_2$ its Euclidean norm. Given a matrix A, we denote by $\mathrm{tr}(A)$ and $\|A\|_{\mathrm{op}}$ its trace and operator (spectral) norm. Given a symmetric matrix A, we denote by $\lambda_{\min}(A)$ ($\lambda_{\max}(A)$) its smallest (largest) eigenvalue. All complexity notations $\Omega(\cdot)$, $\mathcal{O}(\cdot)$, $\omega(\cdot)$, $o(\cdot)$ and $\Theta(\cdot)$ are understood for large data size n, input dimension d, and number of parameters p. We indicate with $C, c > 0$ numerical constants, independent of n, d, p, whose value may change from line to line.

Setting. We consider supervised learning with n training samples $\{(z_1, g_1), \dots, (z_n, g_n)\}$ and labels defined by a (not necessarily deterministic) function of the inputs $g_i = f^*(z_i)$, where $z_i \in \mathbb{R}^{2d}$ denotes the i-th training input and $g_i \in \mathbb{R}$ the corresponding label. Input samples are composed by two distinct parts (or *features*), *i.e.*, $z_i^\top = [x_i^\top, y_i^\top]$, with $x_i, y_i \in \mathbb{R}^d$, and they are sampled i.i.d. from the distribution $\mathcal{P}_{XY}$. We further denote with $\mathcal{P}_X$ ($\mathcal{P}_Y$) the marginal distribution of the x_i-s (y_i-s). The features x and y have the same dimension d to ease the presentation.

We focus on the setting where the labels g_i depend only on x_i, *i.e.*, $g_i = f^*(z_i) = f^*_x(x_i)$ for some (not necessarily deterministic) function f^*_x. Hence, y_i is independent from g_i, after conditioning on x_i. We highlight that the independence between y_i and g_i is conditional to x_i, as the covariance between y_i and x_i is in general non-zero. We refer to y_i as the *spurious feature* of the i-th sample, and to x_i as its *core feature*. As an example, x_i may represent the main object in an image and y_i the (not necessarily independent) background, see Figure 1.

In this setup, the training data is used to learn $f^*(z)$ through a parametric model $f(\theta, z)$ via regularized empirical risk minimization (ERM). Specifically, we perform the following optimization in parameter space:

$$\hat{\theta} = \arg\min_\theta \left(\frac{1}{n} \sum_{i=1}^{n} \ell\left(f(\theta, z_i), g_i\right) + \lambda \left\|\theta\right\|_2^2 \right), \tag{2.1}$$

for some regularization term $\lambda \geq 0$, where ℓ is a loss function[1]. We define the (in-distribution) test loss associated to the model $f(\hat{\theta}, \cdot)$ as

$$\mathcal{L}(\hat{\theta}) = \mathbb{E}_{z \sim \mathcal{P}_{XY},\, g = f^*(z)} \left[\ell\left(f(\hat{\theta}, z), g \right) \right]. \tag{2.2}$$

[1]In general, existence and uniqueness of $\hat{\theta}$ depend on the choice of the model $f(\theta, z)$, the loss function ℓ and the regularization term λ. For the purposes of our work, we will precisely define $\hat{\theta}$ for linear regression (Section 3) and for random features (Section 5).

Spurious correlations. We express the extent to which a model $f(\hat{\theta}, \cdot)$ learns spurious correlations between the spurious feature y and the label g as

$$\mathcal{C}(\hat{\theta}) = \operatorname{Cov}\left(f\left(\hat{\theta}, [\tilde{x}^\top, y^\top]^\top\right), g\right), \tag{2.3}$$

where the covariance is computed on the probability space of $[x^\top, y^\top]^\top \sim \mathcal{P}_{XY}$, $g = f_x^*(x)$ and of the independent core feature $\tilde{x} \sim \mathcal{P}_X$. In words, $\mathcal{C}(\hat{\theta})$ expresses how the output of the model $f(\hat{\theta}, \cdot)$ evaluated on an out-of-distribution sample $[\tilde{x}^\top, y^\top]^\top$ (where the two features are sampled independently from the marginal distributions $\mathcal{P}_X$ and $\mathcal{P}_Y$) correlates to the label associated to the in-distribution sample $g = f^*(z) = f_x^*(x)$. We highlight that, if the model $f(\hat{\theta}, \cdot)$ *does not rely* on the spurious feature y, then $\mathcal{C}(\hat{\theta}) = 0$ as x and $\tilde{x}$ are independent. We formally connect (2.3) to the out-of-distribution test loss in Appendix E.

3 Precise analysis for linear regression

To study $\mathcal{C}(\cdot)$ as defined in (2.3), we focus on a high-dimensional *linear regression* model, *i.e.*,

$$f_{\text{LR}}(\theta, z) = z^\top \theta, \tag{3.1}$$

where $\theta \in \mathbb{R}^{2d}$. The data also follows a linear model, *i.e.*, $g_i = z_i^\top \theta^* + \epsilon_i = x_i^\top \theta_x^* + \epsilon_i$, where $\theta^* \in \mathbb{R}^{2d}$, $\theta_x^* \in \mathbb{R}^d$, and ϵ_i is label noise. Notice that this implies that $\theta^* = [\theta_x^{*\top}, \mathbf{0}_d^\top]^\top$, where $\theta_x^*, \mathbf{0}_d \in \mathbb{R}^d$ and each entry of $\mathbf{0}_d$ is 0. We set $\|\theta^*\|_2 = \|\theta_x^*\|_2 = 1$ and let the ϵ_i-s be i.i.d. (and independent from the z_i-s), mean-0, sub-Gaussian, with variance $\sigma^2 > 0$. We introduce the shorthands $Z = [z_1^\top, \dots, z_n^\top]^\top \in \mathbb{R}^{n \times 2d}$, $G = [g_1, \dots, g_n]^\top \in \mathbb{R}^n$, and $\mathcal{E} = [\epsilon_1, \dots, \epsilon_n]^\top \in \mathbb{R}^n$ to indicate the data matrix, the labels, and the noise vector respectively. Then, using a quadratic loss, (2.1) reads

$$\hat{\theta}_{\text{LR}}(\lambda) = \arg\min_\theta \left(\frac{1}{n}\|Z\theta - G\|_2^2 + \lambda\|\theta\|_2^2\right) = \left(Z^\top Z + n\lambda I\right)^{-1} Z^\top G, \tag{3.2}$$

where the second step holds for $\lambda > 0$ and, if $Z^\top Z$ is invertible, also for $\lambda = 0$.

Assumption 1 (Data distribution)**.** *$\{z_i\}_{i=1}^n$ are n i.i.d. samples from a mean-0, Gaussian distribution $\mathcal{P}_{XY}$, such that its covariance $\Sigma := \mathbb{E}\left[zz^\top\right] \in \mathbb{R}^{2d \times 2d}$ is invertible, with $\lambda_{\max}(\Sigma) = \mathcal{O}(1)$, $\lambda_{\min}(\Sigma) = \Omega(1)$, and $\operatorname{tr}(\Sigma) = 2d$.*

This requirement could be relaxed to having sub-Gaussian data. We focus on the Gaussian case for simplicity, deferring the discussion on the generalization to Appendix C.2.

Warm-up: no regularization ($\lambda = 0$). Our first result concerns the un-regularized setting.

Proposition 3.1. *Let $\lambda = 0$ and $Z^\top Z \in \mathbb{R}^{2d \times 2d}$ be invertible*[2]*. Let $\mathcal{C}(\hat{\theta}_{\text{LR}}(0))$ be the amount of spurious correlations learned by the model $f_{\text{LR}}(\hat{\theta}_{\text{LR}}(0))$. Then, we have that $\mathbb{E}_{\mathcal{E}}[\mathcal{C}(\hat{\theta}_{\text{LR}}(0))] = 0$. Furthermore, if Assumption 1 holds and $n = \omega(d)$, $|\mathcal{C}(\hat{\theta}_{\text{LR}}(0))| = \mathcal{O}(\log d/\sqrt{d})$, with probability at least $1 - 2\exp(-c\log^2 d)$ over Z and $\mathcal{E}$, where c is an absolute constant.*

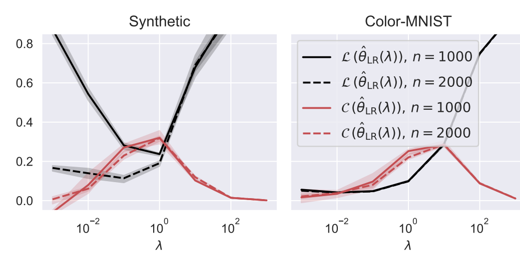

Figure 2: Test loss $\mathcal{L}(\hat{\theta}_{\text{LR}}(\lambda))$ (black) and spurious correlations $\mathcal{C}(\hat{\theta}_{\text{LR}}(\lambda)$ (red) as a function of λ for two values of the number of samples n. *Left:* synthetic Gaussian dataset; *right:* binary Color-MNIST dataset (additional details in Appendix F).

In words, $f_{\text{LR}}(\hat{\theta}_{\text{LR}}(0))$ *does not learn* any spurious correlation between the spurious feature y and the label g. This is also clear from Figure 2, where we report in red the value of $\mathcal{C}(\hat{\theta}_{\text{LR}}(\lambda))$, which approaches 0 as λ becomes small. The idea of the argument is to write explicitly the solution $\hat{\theta}_{\text{LR}}(0) = \left(Z^\top Z\right)^{-1} Z^\top G = \theta^* + \left(Z^\top Z\right)^{-1} Z^\top \mathcal{E}$, where in the second step we separate the ground truth θ^* (which does not capture any dependence on y) from a term only depending on the label noise, which is mean-0 and independent from y. This directly gives the first result, while the second bound is obtained via standard concentration results on $\lambda_{\min}\left(Z^\top Z\right)$. The details are in Appendix C.

General case with regularization ($\lambda > 0$). Setting a regularizer $\lambda > 0$ often reduces the test loss, see the black curve in Figure 2. However, it also leads to non-trivial spurious correlations, and our main result provides a non-asymptotic characterization of this phenomenon.

[2]Under Assumption 1, this holds with probability 1 for $n \geq 2d$.

Theorem 1. *Let Assumption 1 hold, $n = \Theta(d)$ and $\mathcal{C}(\hat{\theta}_{\mathrm{LR}}(\lambda))$ be the amount of spurious correlations learned by the model $f_{\mathrm{LR}}(\hat{\theta}_{\mathrm{LR}}(\lambda))$ for $\lambda > 0$. Denote by $P_y \in \mathbb{R}^{2d \times 2d}$ the projector on the last d elements of the canonical basis in $\mathbb{R}^{2d}$, and set*

$$\mathcal{C}^{\Sigma}(\lambda) := {\theta^*}^{\top} \Sigma \left(\Sigma + \tau(\lambda) I\right)^{-1} P_y \Sigma \theta^*, \tag{3.3}$$

where $\tau := \tau(\lambda)$ is implicitly defined as the unique positive solution of

$$1 - \frac{\lambda}{\tau} = \frac{1}{n} \operatorname{tr}\left(\left(\Sigma + \tau I\right)^{-1} \Sigma\right). \tag{3.4}$$

Then, for every $t \in (0, 1/2)$, $\mathbb{P}_{Z,\mathcal{E}}\left(\left|\mathcal{C}(\hat{\theta}_{\mathrm{LR}}(\lambda)) - \mathcal{C}^{\Sigma}(\lambda)\right| \geq t\right) \leq C d \exp\left(-dt^4/C\right)$, where C is an absolute constant.

In words, Theorem 1 guarantees that $|\mathcal{C}(\hat{\theta}_{\mathrm{LR}}(\lambda)) - \mathcal{C}^{\Sigma}(\lambda)| = o(1)$ with high probability (e.g., setting $t = d^{-1/5}$). Thus, for large d, n, we can estimate $\mathcal{C}(\hat{\theta}_{\mathrm{LR}}(\lambda))$ via the deterministic quantity $\mathcal{C}^{\Sigma}(\lambda)$, which depends on the true parameter θ^*, the covariance of the data Σ, and the regularization λ via the parameter $\tau(\lambda)$ introduced in (3.4). Note that, since $\hat{\theta}_{\mathrm{LR}}(\lambda)$ is given by (3.2), when $\lambda > 0$ it cannot be decomposed as $\theta^* + \left(Z^{\top} Z\right)^{-1} Z^{\top} \mathcal{E}$ (as in the proof of Proposition 3.1 for $\lambda = 0$). Thus, we rely on the non-asymptotic characterization of $\hat{\theta}_{\mathrm{LR}}(\lambda)$ recently provided by Han & Xu (2023). In particular, in the proportional regime $n = \Theta(d)$, their analysis allows to provide concentration bounds on a certain family of low-dimensional functions of $\hat{\theta}_{\mathrm{LR}}(\lambda)$, which includes $\mathcal{C}$ as defined in (2.3). The details are in Appendix C.

4 Roles of regularization and simplicity bias

We now interpret $\mathcal{C}^{\Sigma}(\lambda)$, which characterizes the spurious correlations via Theorem 1, in terms of the data covariance Σ and the regularization λ. To do so, we introduce the following notation

$$\Sigma =: \left(\begin{array}{c|c} \Sigma_{xx} & \Sigma_{xy} \\ \hline \Sigma_{yx} & \Sigma_{yy} \end{array}\right), \qquad S_x^{\Sigma} := \Sigma_{yy} - \Sigma_{yx} \Sigma_{xx}^{-1} \Sigma_{xy} \tag{4.1}$$

where the block $\Sigma_{xx} = \mathbb{E}_{x \sim \mathcal{P}_X}\left[x x^{\top}\right] \in \mathbb{R}^{d \times d}$ ($\Sigma_{yy} = \mathbb{E}_{y \sim \mathcal{P}_Y}\left[y y^{\top}\right] \in \mathbb{R}^{d \times d}$) denotes the covariance of the core (spurious) feature sampled from its marginal distribution. The off-diagonal blocks are $\Sigma_{xy} = \Sigma_{yx}^{\top} = \mathbb{E}_{[x^{\top}, y^{\top}]^{\top} \sim \mathcal{P}_{XY}}\left[x y^{\top}\right] \in \mathbb{R}^{d \times d}$. S_x^{Σ} denotes the Schur complement of Σ with respect to the top-left $d \times d$ block Σ_{xx}. In our setting, S_x^{Σ} offers a helpful statistical interpretation. In fact, for multivariate Gaussian data, it corresponds to the conditional covariance of y given x, *i.e.*, $S_x^{\Sigma} = \operatorname{Cov}(y | x = \bar{x}) = \mathbb{E}_{y|x=\bar{x}}[(y - \mathbb{E}_{y|x=\bar{x}}[y])(y - \mathbb{E}_{y|x=\bar{x}}[y])^{\top}]$. Therefore, the spectrum of S_x^{Σ} describes the degree of dependence between y and x: on the one hand, if its eigenvalues are small, the feature y is close to be determined by the knowledge of the feature x (*i.e.*, y is highly correlated with x); on the other hand, if its eigenvalues are large, the two features tend to be independent. We provide an intuitive example based on the Color-MNIST dataset to better visualize the Schur complement in a low dimensional setting in Appendix F. At this point, leveraging the decomposition of Σ in (4.1) and the Schur complement S_x^{Σ}, we provide the following bounds on $\mathcal{C}^{\Sigma}(\lambda)$, which are proved in Appendix C.

Proposition 4.1. *Let $\mathcal{C}^{\Sigma}(\lambda)$ and S_x^{Σ} be defined in* (3.3) *and* (4.1)*, respectively. Then,*

$$\left|\mathcal{C}^{\Sigma}(\lambda)\right| \leq \min\left(\left\|\Sigma_{yx}\right\|_{\mathrm{op}}, \frac{\lambda_{\max}\left(\Sigma\right)^2}{\tau(\lambda)}, \tau(\lambda) \sqrt{\operatorname{Var}(g) - \sigma^2} \frac{\lambda_{\max}\left(\Sigma_{yy}\right) - \lambda_{\min}\left(S_x^{\Sigma}\right)}{\lambda_{\min}\left(S_x^{\Sigma}\right) \sqrt{\lambda_{\min}\left(\Sigma_{xx}\right)}}\right). \tag{4.2}$$

We discuss the three upper bounds in (4.2) below.

(i): $|\mathcal{C}^{\Sigma}(\lambda)| \leq \|\Sigma_{yx}\|_{\mathrm{op}}$. The off-diagonal blocks $\Sigma_{yx} = \mathbb{E}[y x^{\top}]$ and $\Sigma_{xy} = \Sigma_{yx}^{\top}$ describe the correlation between y and x. In the limit case $\|\Sigma_{yx}\|_{\mathrm{op}} = 0$, we have that x and y are uncorrelated and, therefore, $\mathcal{C}^{\Sigma}(\lambda) = 0$, as there is no spurious correlation that the model can learn.

(ii): $|\mathcal{C}^{\Sigma}(\lambda)| \leq \lambda_{\max}\left(\Sigma\right)^2 / \tau(\lambda)$. From (3.4), one obtains that $\tau(\lambda) \to \infty$ as $\lambda \to \infty$. Thus, the bound implies that $\mathcal{C}^{\Sigma}(\lambda)$ approaches 0 as λ grows large. This captures the intuition that, when the regularization λ is large, the minimization in (3.2) is biased towards solutions with small norm and, therefore, the output of the model is small, which drives to 0 the spurious correlations as defined in (2.3). The behavior is confirmed by Figure 2: $|\mathcal{C}(\hat{\theta}_{\mathrm{LR}}(\lambda))|$ is decreasing for large values of λ and it eventually vanishes; at the same time, large values of λ make the output of the model small, which in turn increases the test loss $\mathcal{L}(\hat{\theta}_{\mathrm{LR}}(\lambda))$.

(iii): The third bound in (4.2), after isolating the term depending on the covariance of the core feature x ($\sqrt{\lambda_{\min}\left(\Sigma_{xx}\right)}$) and on the scaling of the labels ($\sqrt{\mathrm{Var}(g)-\sigma^2}$), depends on *(a)* $\tau(\lambda)$, *(b)* $\lambda_{\min}\left(S_x^{\Sigma}\right)$, and *(c)* $\lambda_{\max}\left(\Sigma_{yy}\right)$. As for *(a)*, we note that $\mathcal{C}^{\Sigma}(\lambda)$ approaches 0 for small values of λ. In fact, the RHS of (3.4) is smaller or equal to $2d/n$; thus, if we consider $2d<n$, we also get $\tau \leq \lambda\left(1-2d/n\right)^{-1}$, which implies $\tau(\lambda)\to 0$ as $\lambda \to 0$. This is in agreement with Proposition 3.1, which handles the case without regularization, and also with the numerical experiments of Figure 2. As for *(b)*, we note that the bound is decreasing with $\lambda_{\min}\left(S_x^{\Sigma}\right)$. This is in agreement with the earlier discussion on how the spectrum of the Schur complement S_x^{Σ} measures the degree of independence between the spurious feature y and the core feature x. Finally, as for *(c)*, we note that the bound is increasing with $\lambda_{\max}\left(\Sigma_{yy}\right)$, which is connected below to the *simplicity* of the spurious feature y. The increasing (decreasing) trend of $\mathcal{C}^{\Sigma}(\lambda)$ w.r.t. $\lambda_{\max}\left(\Sigma_{yy}\right)$ ($\lambda_{\min}\left(S_x^{\Sigma}\right)$) is clearly displayed in Figure 5 for Gaussian data, available in Appendix F.

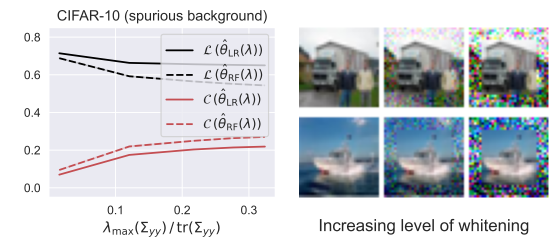

Figure 3: Test loss $\mathcal{L}(\hat{\theta}_{\mathrm{LR/RF}}(\lambda))$ (black) and spurious correlations $\mathcal{C}(\hat{\theta}_{\mathrm{LR/RF}}(\lambda)$ (red) as a function of $\lambda_{\max}\left(\Sigma_{yy}\right)/\,\mathrm{tr}(\Sigma_{yy})$ on a CIFAR-10 dataset for different levels of whitening (additional details in Appendix F).

The connection between $\lambda_{\max}\left(\Sigma_{yy}\right)$ and the *simplicity bias* of ERM can be illustrated via our initial image recognition example. The (spurious) background feature is intuitively an easy pattern to learn from the model: the pixels corresponding to the spurious feature behave consistently across the training data. This in turn skews the spectrum of Σ_{yy}, which has few dominant directions with eigenvalues much larger than the others. Note that this interpretation is similar to the model-dependent definition of simplicity in Morwani et al. (2023). An empirical verification is provided in Figure 3, where we consider the CIFAR-10 dataset, restricted to the "boat" and "truck" classes. Before training a regression model, we whiten up to some level the background feature (as defined in Figure 1) to make it harder to learn, see the right side of Figure 3. Then, for different levels of whitening, we report $\mathcal{C}(\hat{\theta}_{\mathrm{LR}}(\lambda))$ as a function of $\lambda_{\max}\left(\Sigma_{yy}\right)$. We normalize $\lambda_{\max}\left(\Sigma_{yy}\right)$ by the trace $\mathrm{tr}(\Sigma_{yy})$ to exclude the size of the pattern from our experiment[3]. The red curve shows an increasing trend: small values of $\lambda_{\max}\left(\Sigma_{yy}\right)$ correspond to significant whitening and, hence, to small spurious correlations, as predicted by Proposition 4.1.

Trade-off between $\mathcal{L}(\hat{\theta}_{\mathbf{LR}}(\lambda))$ and $\mathcal{C}(\hat{\theta}_{\mathbf{LR}}(\lambda))$. Figure 2 shows that there is an interval of values for the regularization ($\lambda \sim 10^{-1}$) where the test loss $\mathcal{L}(\hat{\theta}_{\mathrm{LR}}(\lambda))$ is decreasing in λ, while the spurious correlations $\mathcal{C}(\hat{\theta}_{\mathrm{LR}}(\lambda))$ are increasing. This evidence suggests a natural trade-off between these two quantities, mediated by λ. To theoretically capture such trade-off, we first provide a non-asymptotic concentration bound for $\mathcal{L}(\hat{\theta}_{\mathrm{LR}}(\lambda))$.

Proposition 4.2. *Let Assumption 1 hold, $n=\Theta(d)$ and $\mathcal{L}(\hat{\theta}_{\mathrm{LR}}(\lambda))$ be defined according to* (2.2). *Set*

$$\mathcal{L}^{\Sigma}(\lambda) := \left(\sigma^2+\tau(\lambda)^2\left\|\left(\Sigma+\tau(\lambda)I\right)^{-1}\Sigma^{1/2}\theta^*\right\|_2^2\right) / \left(1-\mathrm{tr}\left(\left(\Sigma+\tau(\lambda)I\right)^{-2}\Sigma^2\right)/n\right), \tag{4.3}$$

where $\tau(\lambda)$ is defined via (3.4). *Then, for every $t\in(0,1/2)$, $\mathbb{P}_{Z,\mathcal{E}}\left(\left|\mathcal{L}(\hat{\theta}_{\mathrm{LR}}(\lambda))-\mathcal{L}^{\Sigma}(\lambda)\right|\geq t\right)\leq Cd\exp\left(-dt^4/C\right)$.*

In words, Proposition 4.2 guarantees that $|\mathcal{L}(\hat{\theta}_{\mathrm{LR}}(\lambda))-\mathcal{L}^{\Sigma}(\lambda)|=o(1)$ with high probability. Its proof is an adaptation of Theorem 3.1 in Han & Xu (2023), and the details are in Appendix C. Armed with the non-asymptotic bounds of Theorem 1 and Proposition 4.2, we characterize the trade-off between $\mathcal{L}(\hat{\theta}_{\mathrm{LR}}(\lambda))$ and $\mathcal{C}(\hat{\theta}_{\mathrm{LR}}(\lambda))$ by studying the monotonicity of $\mathcal{L}^{\Sigma}(\lambda)$ and $\mathcal{C}^{\Sigma}(\lambda)$.

Proposition 4.3. *Let $\mathcal{C}^{\Sigma}(\lambda)$ and $\mathcal{L}^{\Sigma}(\lambda)$ be defined as in* (3.3) *and* (4.3). *Then, if $2d<n$, we have that $\mathcal{L}^{\Sigma}(\lambda)$ is monotonically decreasing in a right neighborhood of $\lambda=0$, and there exists $\lambda_{\mathcal{L}}>0$ such that $\mathcal{L}^{\Sigma}(\lambda)$ is monotonically increasing for $\lambda\geq\lambda_{\mathcal{L}}$. Furthermore, if $\Sigma_{xx}=I$, then $\mathcal{C}^{\Sigma}(\lambda)$ is non-negative and there exists $\lambda_{\mathcal{C}}$ such that $\mathcal{C}^{\Sigma}(\lambda)$ is monotonically increasing for $\lambda\leq\lambda_{\mathcal{C}}$. Finally, as long as*

$$\frac{2d}{n}\leq\frac{\lambda_{\min}(\Sigma)}{4}\min\left(1,\frac{2\lambda_{\max}(\Sigma)/\sigma^2}{\left(\lambda_{\max}(\Sigma)/\lambda_{\min}(\Sigma)+1\right)^2}\right), \tag{4.4}$$

we have that $\lambda_{\mathcal{C}}\geq\lambda_{\mathcal{L}}$.

[3] If y has 0-mean, then $\mathbb{E}\|y\|_2^2=\mathrm{tr}(\Sigma_{yy})$, i.e., the trace captures the size of the pattern.

In words, Proposition 4.3 shows that $\mathcal{C}^{\Sigma}(\lambda)$ grows with λ at least until the regularization equals a value $\lambda_{\mathcal{C}}$. For example, in Figure 2, $\lambda_{\mathcal{C}} \sim 1$ for a Gaussian data and $\lambda_{\mathcal{C}} \sim 10$ for Color-MNIST. Furthermore, in this interval, $\mathcal{L}^{\Sigma}(\lambda)$ is initially decreasing and then increasing as $\lambda \geq \lambda_{\mathcal{L}}$. These trends in turn imply that the optimal value $\lambda^*_{\mathcal{L}}$ that minimizes the test loss is s.t. $\lambda^*_{\mathcal{L}} \in (0, \lambda_{\mathcal{C}}]$ – an interval where the spurious correlations are strictly positive and increasing. The proof of Proposition 4.3 (whose details are in Appendix C) relies on the monotonicity of $\tau(\lambda)$ in λ, and the last statement follows from showing that $\tau(\lambda_{\mathcal{C}}) \geq \lambda_{\min}\left(S_x^{\Sigma}\right) \geq \lambda_{\min}(\Sigma) \geq \tau(\lambda_{\mathcal{L}})$. The upper bound on $2d/n$ in (4.4) is required to prove that $\lambda_{\min}(\Sigma) \geq \tau(\lambda_{\mathcal{L}})$ and, due to Assumption 1, it is implied by taking $n = \omega(d)$. We note that the latter scaling holds in standard datasets, e.g., MNIST ($n = 6 \cdot 10^4$, $2d \approx 2 \cdot 10^3$ when considering the 3 color channels) and CIFAR-10 ($n = 5 \cdot 10^4$, $2d \approx 3 \cdot 10^3$).

5 Role of over-parameterization

Our analysis has so far focused on linear regression, highlighting the role of data covariance and regularization. However, moving to complex predictive models, such as neural networks, may lead to differences in the degree to which spurious correlations are learned. As an example, in the left panel of Figure 6 in Appendix F, we train an over-parameterized two-layer neural network on the binary Color-MNIST and CIFAR-10 datasets, for different values of the regularizer λ. While for high values of λ the results are qualitatively similar to the ones in Figure 2, a striking difference is that spurious correlations remain significant even when there is little to no regularization (i.e., $\lambda \approx 0$), in sharp contrast with Proposition 3.1. We also note that the phenomenon is in line with previous empirical work Sagawa et al. (2020a). We bridge the gap between linear regression and over-parameterized models by focusing on *random features*:

$$f_{\mathrm{RF}}(z, \theta) = \phi(Vz)^{\top}\theta, \tag{5.1}$$

where V is a $p \times 2d$ matrix s.t. $V_{i,j} \sim_{\text{i.i.d.}} \mathcal{N}(0, 1/(2d))$, and ϕ is an activation applied component-wise. We consider ϕ to be L Lipschitz, and odd, such that its 1st Hermite coefficient $\mu_1 \neq 0$. The number of parameters of this model is p, as V is a fixed random matrix and $\theta \in \mathbb{R}^p$ contains trainable parameters. The scaling of input data ($\mathrm{tr}(\Sigma) = 2d$) and the variance of the entries of V guarantee that the pre-activations of the model (*i.e.*, the entries of the vector $Vz \in \mathbb{R}^p$) are of constant order. We consider the ERM in (2.1) with a quadratic loss $\hat{\theta}_{\mathrm{RF}}(\lambda) = \arg\min_{\theta}(\frac{1}{n}\left\|\Phi\theta - G\right\|_2^2 + \lambda\left\|\theta\right\|_2^2)$, where we set $\Phi := [\phi(Vz_1), \ldots, \phi(Vz_n)]^{\top} \in \mathbb{R}^{n \times p}$. When $\lambda = 0$, if $\Phi\Phi^{\top}$ is invertible, the minimization above does not necessarily have a unique solution. In that case, we set $\hat{\theta}_{\mathrm{RF}}(0)$ to be the solution obtained via gradient descent with 0 initialization, which corresponds to the min-norm interpolator (see equation (33) in Bartlett et al. (2021)). Then[4], we can write, for $\lambda \geq 0$, $\hat{\theta}_{\mathrm{RF}}(\lambda) = \Phi^{\top}\left(\Phi\Phi^{\top} + n\lambda I\right)^{-1} G$.

Theorem 2. *Let Assumptions 1 hold, $n = \Theta(d)$, $p = \omega(n \log^4 n)$, $\log p = \Theta(\log n)$, and $z \in \mathbb{R}^{2d}$ be sampled from a distribution satisfying Assumption 1, not necessarily with the same covariance as $\mathcal{P}_{XY}$, independent from everything else. Let $f_{\mathrm{RF}}(z, \hat{\theta}_{\mathrm{RF}}(\lambda))$ be the RF model defined in* (5.1)*, and $f_{\mathrm{LR}}(z, \hat{\theta}_{\mathrm{LR}}(\tilde{\lambda}))$ be the linear regression model defined in* (3.1) *with $\hat{\theta}_{\mathrm{LR}}(\tilde{\lambda})$ given by* (3.2)*. Then, for $\lambda \geq 0$,*

$$\left|f_{\mathrm{RF}}(z, \hat{\theta}_{\mathrm{RF}}(\lambda)) - f_{\mathrm{LR}}(z, \hat{\theta}_{\mathrm{LR}}(\tilde{\lambda}))\right| = \mathcal{O}\left(\frac{d^{1/4}\log d}{p^{1/4}} + \frac{\log^{3/2} d}{d^{1/8}}\right) = o(1), \tag{5.2}$$

with probability at least $1 - C\sqrt{d}\log^2 d/\sqrt{p} - C\log^3 d/d^{1/4}$, where the effective regularization *$\tilde{\lambda}$ is given by*

$$\tilde{\lambda} = \frac{2\tilde{\mu}^2 d}{\mu_1^2 n} + \frac{2d}{\mu_1^2 p}\lambda, \tag{5.3}$$

and $\tilde{\mu}^2 = \sum_{k\geq 2}\mu_k^2$, with μ_k denoting the k-th Hermite coefficient of ϕ.

In words, Theorem 2 shows that the over-parameterized RF model, when evaluated on a new test sample (not necessarily from the same distribution as the input data), is asymptotically equivalent to linear regression with regularization $\tilde{\lambda}$, given by (5.3). In particular, even in the ridgeless case ($\lambda = 0$), the RF model is equivalent to linear regression with strictly positive regularization.

Thus, we expect the presence of spurious correlations, just like in Figure 6, since $\mathcal{C}(\hat{\theta}_{\mathrm{RF}}(0))$ approaches $\mathcal{C}^{\Sigma}(\tilde{\lambda})$ with $\tilde{\lambda} > 0$. Notably, the effective regularization $\tilde{\lambda}$ depends on the activation ϕ via its Hermite coefficients, and it increases with the ratio $\tilde{\mu}^2/\mu_1^2$. This is also verified in Figure 6 via experiments on Gaussian data, as discussed in Appendix F. We finally remark that Theorem 2 holds in more generality than Assumption 1. In Appendix D, we provide the full argument using the less stringent Assumption 6.

[4] $\Phi\Phi^{\top}$ is proved to be invertible with high probability in Lemma D.3.

REFERENCES

Faruk Ahmed, Yoshua Bengio, Harm van Seijen, and Aaron Courville. Systematic generalisation with group invariant predictions. In *International Conference on Learning Representations*, 2021.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2020.

Jimmy Ba, Murat A Erdogdu, Taiji Suzuki, Zhichao Wang, Denny Wu, and Greg Yang. High-dimensional asymptotics of feature learning: How one gradient step improves the representation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Peter L Bartlett, Andrea Montanari, and Alexander Rakhlin. Deep learning: a statistical viewpoint. *Acta numerica*, 30: 87–201, 2021.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

Simone Bombari and Marco Mondelli. Privacy for free in the over-parameterized regime. *arXiv preprint arXiv:2410.14787*, 2024.

Simone Bombari, Mohammad Hossein Amani, and Marco Mondelli. Memorization and optimization in deep neural networks with minimum over-parameterization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Simone Bombari, Shayan Kiyani, and Marco Mondelli. Beyond the universal law of robustness: Sharper laws for random features and neural tangent kernels. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Sebastien Bubeck and Mark Sellke. A universal law of robustness via isoperimetry. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

C. Chang, G. Adam, and A. Goldenberg. Towards robust classification model by counterfactual and invariant data generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021a.

Xiangyu Chang, Yingcong Li, Samet Oymak, and Christos Thrampoulidis. Provable benefits of overparameterization in model compression: From double descent to pruning neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 6974–6983, 2021b.

Chen Cheng and Andrea Montanari. Dimension free ridge regression. *The Annals of Statistics*, 52(6):2879 – 2912, 2024. doi: 10.1214/24-AOS2449.

Alexandru Damian, Jason Lee, and Mahdi Soltanolkotabi. Neural networks can learn representations with gradient descent. In *Conference on Learning Theory (COLT)*, 2022.

Elvis Dohmatob and Alberto Bietti. On the (non-) robustness of two-layer neural networks in different learning regimes. *arXiv preprint arXiv:2203.11864*, 2022.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations (ICLR)*, 2019.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. ISSN 2522-5839.

Sebastian Goldt, Marc Mézard, Florent Krzakala, and Lenka Zdeborová. Modeling the influence of data structure on learning in neural networks: The hidden manifold model. *Physical Review X*, 10(4):041044, 2020.

Sebastian Goldt, Bruno Loureiro, Galen Reeves, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. The gaussian equivalence of generative models for learning with shallow neural networks. In *Mathematical and Scientific Machine Learning*, pp. 426–471. PMLR, 2022.

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

Hamed Hassani and Adel Javanmard. The curse of overparametrization in adversarial training: Precise analysis of robust generalization for random features regression. *The Annals of Statistics*, 52(2):441 – 465, 2024.

Trevor J. Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *Annals of statistics*, 50 2:949–986, 2019.

Katherine Hermann and Andrew Lampinen. What shapes feature representations? Exploring datasets, architectures, and training. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Katherine Hermann, Hossein Mobahi, Thomas FEL, and Michael Curtis Mozer. On the foundations of shortcut learning. In *The Twelfth International Conference on Learning Representations*, 2024.

Hong Hu and Yue M. Lu. Universality laws for high-dimensional learning with random features. *IEEE Transactions on Information Theory*, 69(3):1932–1964, 2023.

Pavel Izmailov, Polina Kirichenko, Nate Gruver, and Andrew Gordon Wilson. On feature learning in the presence of spurious correlations. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022.

Dimitris Kalimeris, Gal Kaplun, Preetum Nakkiran, Benjamin Edelman, Tristan Yang, Boaz Barak, and Haofeng Zhang. Sgd on neural networks learns functions of increasing complexity. In *Advances in Neural Information Processing Systems*, 2019.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. In *The Eleventh International Conference on Learning Representations*, 2023.

Donghwan Lee, Behrad Moniri, Xinmeng Huang, Edgar Dobriban, and Hamed Hassani. Demystifying disagreement-on-the-line in high dimensions. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

Neil Rohit Mallinar, Austin Zane, Spencer Frei, and Bin Yu. Minimum-norm interpolation under covariate shift. In *Forty-first International Conference on Machine Learning*, 2024.

Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75(4):667–766, 2022.

Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Generalization error of random feature and kernel methods: Hypercontractivity and kernel matrix concentration. *Applied and Computational Harmonic Analysis*, 59:3–84, 2022. ISSN 1063-5203. Special Issue on Harmonic Analysis and Machine Learning.

Mazda Moayeri, Sahil Singla, and Soheil Feizi. Hard imagenet: Segmentations for objects with strong spurious cues. In *Advances in Neural Information Processing Systems*, 2022.

Behrad Moniri, Donghwan Lee, Hamed Hassani, and Edgar Dobriban. A theory of non-linear feature learning with one gradient step in two-layer neural networks. In *International Conference on Machine Learning (ICML)*, 2024.

Andrea Montanari and Basil N Saeed. Universality of empirical risk minimization. In *Conference on Learning Theory*, pp. 4310–4312. PMLR, 2022.

Andrea Montanari, Feng Ruan, Youngtak Sohn, and Jun Yan. The generalization error of max-margin linear classifiers: Benign overfitting and high dimensional asymptotics in the overparametrized regime. *arXiv preprint arXiv:1911.01544*, 2019.

Depen Morwani, jatin batra, Prateek Jain, and Praneeth Netrapalli. Simplicity bias in 1-hidden layer neural networks. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Quynh Nguyen, Marco Mondelli, and Guido Montufar. Tight bounds on the smallest eigenvalue of the neural tangent kernel for deep ReLU networks. In *International Conference on Machine Learning (ICML)*, 2021.

Ryan O'Donnell. *Analysis of Boolean Functions*. Cambridge University Press, 2014.

Mohammad Pezeshki, Sékou-Oumar Kaba, Yoshua Bengio, Aaron Courville, Doina Precup, and Guillaume Lajoie. Gradient starvation: A learning proclivity in neural networks. In *Advances in Neural Information Processing Systems*, 2021.

Gregory Plumb, Marco Tulio Ribeiro, and Ameet Talwalkar. Finding and fixing spurious patterns with explanations. *Transactions on Machine Learning Research*, 2022.

GuanWen Qiu, Da Kuang, and Surbhi Goel. Complexity matters: Dynamics of feature learning in the presence of spurious correlations. In *International Conference on Machine Learning (ICML)*, 2024.

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, 2007.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*, 2020a.

Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8346–8356, 2020b.

Seonguk Seo, Joon-Young Lee, and Bohyung Han. Information-theoretic bias reduction via causal view of spurious correlation. In *AAAI Conference on Artificial Intelligence*, 2022.

Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. The pitfalls of simplicity bias in neural networks. In *Advances in Neural Information Processing Systems*, 2020.

Sahil Singla and Soheil Feizi. Salient imagenet: How to discover spurious features in deep learning? In *International Conference on Learning Representations*, 2022.

Yanke Song, Sohom Bhattacharya, and Pragya Sur. Generalization error of min-norm interpolators in transfer learning. *arXiv preprint arXiv:2406.13944*, 2024.

Christos Thrampoulidis, Samet Oymak, and Babak Hassibi. Regularized linear regression: A precise analysis of the estimation error. In *Conference on Learning Theory*, pp. 1683–1709. PMLR, 2015.

Rishabh Tiwari and Pradeep Shenoy. Overcoming simplicity bias in deep networks using a feature sieve. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Nilesh Tripuraneni, Ben Adlam, and Jeffrey Pennington. Overparameterization improves robustness to covariate shift in high dimensions. In *Advances in Neural Information Processing Systems*, 2021.

Victor Veitch, Alexander D'Amour, Steve Yadlowsky, and Jacob Eisenstein. Counterfactual invariance to spurious correlations in text classification. In *Advances in Neural Information Processing Systems*, 2021.

Roman Vershynin. *Introduction to the non-asymptotic analysis of random matrices*, pp. 210–268. Cambridge University Press, 2012.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*. Cambridge university press, 2018.

Kai Yuanqing Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. In *International Conference on Learning Representations (ICLR)*, 2021.

Fan Yang, Hongyang R. Zhang, Sen Wu, Christopher Ré, and Weijie J. Su. Precise high-dimensional asymptotics for quantifying heterogeneous transfers. *arXiv preprint arXiv:2010.11750*, 2023.

Wenqian Ye, Guangtao Zheng, Xu Cao, Yunsheng Ma, and Aidong Zhang. Spurious correlations in machine learning: A survey. *arXiv preprint arXiv:2402.12715*, 2024.

Jingzhao Zhang, Aditya Krishna Menon, Andreas Veit, Srinadh Bhojanapalli, Sanjiv Kumar, and Suvrit Sra. Coping with label shift via distributionally robust optimisation. In *International Conference on Learning Representations*, 2021.

Chunting Zhou, Xuezhe Ma, Paul Michel, and Graham Neubig. Examining and combating spurious features under distribution shift. In *International Conference on Machine Learning (ICML)*, 2021.

Indre Zliobaite. On the relation between accuracy and fairness in binary classification. In *2nd ICML Workshop on Fairness, Accountability, and Transparency in Machine Learning (FATML)*, 2015.

A Additional notation

We define a sub-Gaussian random variable according to Proposition 2.5.2 in Vershynin (2018), and $\|X\|_{\psi_2} := \inf\{t > 0 \,:\, \mathbb{E}\left[\exp(X^2/t^2)\right] \leq 2\}$. If $X \in \mathbb{R}^n$ is a random vector, then $\|X\|_{\psi_2} := \sup_{\|u\|_2=1} \left\|u^\top X\right\|_{\psi_2}$. When we state that a random variable or vector X is sub-Gaussian, we implicitly mean $\|X\|_{\psi_2} = \mathcal{O}(1)$, *i.e.* its sub-Gaussian norm does not increase with the scalings of the problem.

We say that X respects the Lipschitz concentration property if, for all 1-Lipschitz continuous functions φ, we have $\|\varphi(X) - \mathbb{E}[\varphi(X)]\|_{\psi_2} = \mathcal{O}(1)$. Notice that then, if X is Lipschitz concentrated, then $X - \mathbb{E}[X]$ is sub-Gaussian.

Given two symmetric matrices A, B, we use the notation $A \succeq B$ if $A - B$ is p.s.d. Notice that if $A \succeq B \succ 0$, then we also have $B^{-1} \succeq A^{-1}$. We denote with $\|A\|_F$ the Frobenius norm of A, and with $\ker(A)$ its kernel space. If A is a square matrix, we use the notation $\mathrm{diag}(A)$ to denote a matrix identical to A on the diagonal, and 0 everywhere else. We let $A \circ B$ denote the Hadamard (component-wise) product between matrices, and $A^{\circ k}$ denote $A \circ A \circ ... \circ A$, where A appears k times.

B Related work

Spurious correlations. Learning from spurious correlations in a training dataset is rather common Geirhos et al. (2019); Arjovsky et al. (2020); Geirhos et al. (2020); Sagawa et al. (2020a); Xiao et al. (2021); Singla & Feizi (2022) and it has unwanted consequences, e.g., lack of robustness towards domain shift, prediction bias and compromised algorithmic fairness Zliobaite (2015); Geirhos et al. (2019); Zhou et al. (2021); Veitch et al. (2021); Seo et al. (2022). Thus, multiple mitigation approaches have been proposed, with Sagawa et al. (2020a); Zhang et al. (2021) or without Liu et al. (2021); Ahmed et al. (2021) available annotations. Specifically, Tiwari & Shenoy (2023) exploit the difference in the features learned at different layers of a deep neural network; Izmailov et al. (2022); Kirichenko et al. (2023) re-train the last layer of the ERM solution to adapt the features to the distribution shift; and Chang et al. (2021a); Plumb et al. (2022) mitigate the problem via data augmentation.

Simplicity bias. Recent work has shown that deep learning models have a bias towards learning from "easier" patterns Belkin et al. (2019); Rahaman et al. (2019); Kalimeris et al. (2019). In shortcut learning, this property is formalized in different ways across the literature. The difficulty of a feature is defined in terms of the minimum complexity of a network that learns it by Hermann & Lampinen (2020) and in terms of the smallest amount of linear segments that separate different classes by Shah et al. (2020). Moayeri et al. (2022) connect the simplicity to the position and size of the features in an image. Morwani et al. (2023) define the simplicity bias in 1-hidden layer neural networks via the rank of a projection operator that does not alter them substantially, and they focus on a dataset generated via an independent features model learned via the NTK. The NTK is also used to analyze gradient starvation Pezeshki et al. (2021) and feature availability Hermann et al. (2024), regarded as explanations of the simplicity bias. Qiu et al. (2024) focus on parity functions and staircases, analyzing the learning dynamics of features having different complexity.

High-dimensional regression. The test loss of linear regression when the input dimension d scales proportionally with the sample size n has been characterized precisely both in-distribution (Hastie et al., 2019; Cheng & Montanari, 2024) and under covariate shift Yang et al. (2023); Mallinar et al. (2024); Song et al. (2024). Furthermore, Montanari et al. (2019); Chang et al. (2021b); Han & Xu (2023) have studied the distribution of the ERM solution via the convex Gaussian min-max Theorem Thrampoulidis et al. (2015). Specifically, our work builds on the non-asymptotic characterization provided by Han & Xu (2023).

In contrast with linear regression where the number of parameters equals the input dimension, random features models Rahimi & Recht (2007) capture the effects of over-parameterization, as the number of parameters is independently of d and n. Mei & Montanari (2022) have characterized the test loss of random features, showing that it displays a double descent Belkin et al. (2019). Furthermore, the RF model has been used to understand a wide family of phenomena such as feature learning Ba et al. (2022); Damian et al. (2022); Moniri et al. (2024), robustness under adversarial attacks Dohmatob & Bietti (2022); Bombari et al. (2023); Hassani & Javanmard (2024), and distribution shift Tripuraneni et al. (2021); Lee et al. (2023). The equivalence between an over-parameterized RF model and a regularized linear one has also been studied in detail Goldt et al. (2022; 2020); Hu & Lu (2023); Montanari & Saeed (2022). However, existing rigorous results show the equivalence at the level of training and test error. In contrast, we are interested in the covariance defined in (2.3) and, for this reason, we prove an equivalence at the level of the predictor (Theorem 2).

C PROOFS FOR LINEAR REGRESSION

Proof of Proposition 3.1. Note that

$$\hat{\theta}_{\mathrm{LR}}(0) = \left(Z^\top Z\right)^{-1} Z^\top G. \tag{C.1}$$

Since we have $g_i = z_i^\top \theta^* + \epsilon_i$, (C.1) reads

$$\hat{\theta}_{\mathrm{LR}}(0) = \left(Z^\top Z\right)^{-1} Z^\top \left(Z\theta^* + \epsilon\right) = \theta^* + \left(Z^\top Z\right)^{-1} Z^\top \mathcal{E}. \tag{C.2}$$

Then, we can plug this result in the definition of $\mathcal{C}(\hat{\theta})$ in (2.3) to obtain

$$\begin{aligned}
\mathbb{E}_{\mathcal{E}}\left[\mathcal{C}(\hat{\theta}_{\mathrm{LR}}(0))\right] &= \mathbb{E}_{\mathcal{E}}\left[\mathrm{Cov}_{[x^\top, y^\top]^\top \sim P_{XY},\, g=f_x^*(x),\, \tilde{x}\sim P_X}\left(f_{\mathrm{LR}}\left(\hat{\theta}_{\mathrm{LR}}(0), [\tilde{x}^\top, y^\top]^\top\right), g\right)\right] \\
&= \mathbb{E}_{\mathcal{E}}\left[\mathrm{Cov}_{[x^\top, y^\top]^\top \sim P_{XY},\, \tilde{x}\sim P_X}\left([\tilde{x}^\top, y^\top]\hat{\theta}_{\mathrm{LR}}(0), x^\top \theta_x^*\right)\right] \\
&= \mathbb{E}_{\mathcal{E}}\left[\mathrm{Cov}_{[x,y]\sim P_{XY},\, \tilde{x}\sim P_X}\left(\tilde{x}^\top \theta_x^* + [\tilde{x}^\top, y^\top]\left(Z^\top Z\right)^{-1} Z^\top \mathcal{E}, x^\top \theta_x^*\right)\right] \\
&= \mathrm{Cov}_{[x,y]\sim P_{XY},\, \tilde{x}\sim P_X}\left(\tilde{x}^\top \theta_x^*, x^\top \theta_x^*\right) \\
&= 0,
\end{aligned} \tag{C.3}$$

where in the second line we used that $\mathcal{E}$ is independent from everything else, in fourth line we used $\mathbb{E}\left[\mathcal{E}\right] = 0$, and that $\mathcal{E}$ is independent from all the other random variables, and the last step holds since $\tilde{x}$ is independent from x.

For the second part of the statement we have that

$$\begin{aligned}
\mathcal{C}(\hat{\theta}_{\mathrm{LR}}(0)) &= \mathrm{Cov}_{[x^\top, y^\top]^\top \sim P_{XY},\, \tilde{x}\sim P_X}\left([\tilde{x}^\top, y^\top]\left(Z^\top Z\right)^{-1} Z^\top \mathcal{E}, x^\top \theta_x^*\right) \\
&= \mathrm{Cov}_{[x^\top, y^\top]^\top \sim P_{XY},\, \tilde{x}\sim P_X}\left(\mathcal{E}^\top Z\left(Z^\top Z\right)^{-1} P_y [x^\top, y^\top]^\top, [x^\top, y^\top]\theta^*\right) \\
&= \mathcal{E}^\top Z\left(Z^\top Z\right)^{-1} P_y \Sigma \theta^*,
\end{aligned} \tag{C.4}$$

where in the second line we introduced $P_y \in \mathbb{R}^{2d\times 2d}$, defined as the projector on the last d elements of the canonical basis in $\mathbb{R}^{2d}$. Then, since $\mathcal{E}$ is a sub-Gaussian vector (the entries are mean-0, i.i.d. sub-Gaussian) independent from everything else, we have that, with probability at least $1 - 2\exp\left(-c_1 \log^2 d\right)$,

$$\left|\mathcal{C}(\hat{\theta}_{\mathrm{LR}}(0))\right| \le \log d \left\|Z\left(Z^\top Z\right)^{-1} P_y \Sigma \theta^*\right\|_2 \le \log d \left\|Z\left(Z^\top Z\right)^{-1}\right\|_{\mathrm{op}} \left\|P_y\right\|_{\mathrm{op}} \left\|\Sigma\right\|_{\mathrm{op}} \left\|\theta^*\right\|_2 \le \frac{\log d \left\|\Sigma\right\|_{\mathrm{op}}}{\sqrt{\lambda_{\min}\left(Z^\top Z\right)}}, \tag{C.5}$$

where we used $\left\|P_y\right\|_{\mathrm{op}} = 1$ and $\left\|\theta^*\right\|_2 = 1$. Since Z is a $n \times 2d$ matrix with independent rows having second moment Σ, by Theorem 5.39 in Vershynin (2012) (see Remark 5.40), we have that

$$\left\|\frac{Z^\top Z}{n} - \Sigma\right\|_{\mathrm{op}} = \mathcal{O}\left(\sqrt{\frac{d}{n}}\right) = o(1), \tag{C.6}$$

with probability at least $1-2\exp\left(-c_2 d\right)$. Hence, with this probability, by Weyl's inequality, we also have

$$\lambda_{\min}\left(Z^\top Z\right) \geq n\lambda_{\min}\left(\Sigma\right) - \left\|Z^\top Z - n\Sigma\right\|_{\text{op}} = \Theta(n), \tag{C.7}$$

where the last step holds because of Assumption 1. Thus, we have that (C.5) reads

$$\left|\mathcal{C}(\hat{\theta}_{\text{LR}}(0))\right| \leq \frac{\log d \left\|\Sigma\right\|_{\text{op}}}{\sqrt{\lambda_{\min}\left(Z^\top Z\right)}} = \mathcal{O}\left(\frac{\log d}{\sqrt{n}}\right), \tag{C.8}$$

with probability at least $1-2\exp\left(-c_3\log^2 d\right)$ over Z and $\mathcal{E}$, which gives the desired result. □

Proof of Theorem 1. As in Han & Xu (2023), we define the Gaussian sequence model $\hat{\theta}^\rho \in \mathbb{R}^{2d}$ as

$$\hat{\theta}^\rho = \left(\Sigma + \tau(\lambda)I\right)^{-1}\Sigma^{1/2}\left(\Sigma^{1/2}\theta^* + \frac{\gamma\rho}{\sqrt{2d}}\right), \tag{C.9}$$

where ρ is a standard Gaussian vector in $\mathbb{R}^{2d}$. In the equation above, $\gamma > 0$ is implicitly defined via

$$\frac{n\gamma^2}{2d} = \sigma^2 + \mathbb{E}_\rho\left[\left\|\Sigma^{1/2}\left(\hat{\theta}^\rho - \theta^*\right)\right\|_2^2\right] \tag{C.10}$$

On the other hand, following a similar argument as the one in (C.3), we have that, for every $\theta \in \mathbb{R}^{2d}$,

$$\begin{aligned}\mathcal{C}(\theta) &= \text{Cov}_{[x^\top, y^\top]^\top \sim P_{XY},\, \tilde{x}\sim P_X}\left([\tilde{x}^\top, y^\top]\theta, x^\top\theta_x^*\right)\\ &= \theta^\top\mathbb{E}_{[x^\top, y^\top]^\top \sim P_{XY},\, \tilde{x}\sim P_X}\left[[\tilde{x}^\top, y^\top]^\top x^\top\right]\theta_x^*\\ &= \theta^\top\mathbb{E}_{[x^\top, y^\top]^\top \sim P_{XY}}\left[[\mathbf{0}^\top, y^\top]^\top[x^\top, \mathbf{0}^\top]^\top\right]\theta^*\\ &= \theta^\top P_y\mathbb{E}_{[x^\top, y^\top]^\top \sim P_{XY}}\left[[x^\top, y^\top]^\top[x^\top, y^\top]^\top\right]\theta^*\\ &= \theta^\top P_y\Sigma\theta^*,\end{aligned} \tag{C.11}$$

where the third line holds since $\tilde{x}$ has 0 mean and is independent with x and y, and by definition of θ_x^*, and the fourth line holds because $P_y[x^\top, y^\top]^\top = [\mathbf{0}^\top, y^\top]^\top$ and because the last d entries of θ^* are 0 (*i.e.*, $P_y\theta^* = 0$). Thus, since we have that $\left\|P_y\Sigma\theta^*\right\|_2 \leq \left\|P_y\right\|_{\text{op}}\left\|\Sigma\right\|_{\text{op}}\left\|\theta^*\right\|_2 \leq \left\|\Sigma\right\|_{\text{op}}$ because of Assumption 1, we have that $\mathcal{C}(\cdot): \mathbb{R}^{2d} \to \mathbb{R}$ is a $\left\|\Sigma\right\|_{\text{op}}$-Lipschitz function.

Now, since $\mathcal{P}_{XY}$ is multivariate Gaussian, Theorem 2.3 of Han & Xu (2023) gives that, for any 1-Lipschitz function $\varphi: \mathbb{R}^{2d} \to \mathbb{R}$, and any $t \in (0, 1/2)$,

$$\mathbb{P}_{Z,G}\left(\left|\varphi(\hat{\theta}_{\text{LR}}(\lambda)) - \mathbb{E}_\rho\left[\varphi\left(\hat{\theta}^\rho\right)\right]\right| \geq t\right) \leq C_1 d\exp\left(-dt^4/C_1\right), \tag{C.12}$$

where C_1 is a constant depending on $\lambda_{\min}\left(\Sigma\right)$, $\left\|\Sigma\right\|_{\text{op}}$, σ^2, and $n/d = \Theta(1)$. Since $\mathcal{C}(\cdot)$ is linear, notice that we have

$$\mathbb{E}_\rho\left[\mathcal{C}\left(\hat{\theta}^\rho\right)\right] = \mathcal{C}\left(\mathbb{E}_\rho\left[\hat{\theta}^\rho\right]\right) = \mathcal{C}\left(\left(\Sigma + \tau(\lambda)I\right)^{-1}\Sigma\theta^*\right) = {\theta^*}^\top\Sigma\left(\Sigma + \tau(\lambda)I\right)^{-1}P_y\Sigma\theta^* = \mathcal{C}^\Sigma(\lambda), \tag{C.13}$$

where we used (C.11) in the third step, and the definition of $\mathcal{C}^\Sigma(\lambda)$ in (3.3) in the last one. Thus, setting $\varphi(\cdot)$ to be $\mathcal{C}(\cdot)/\left\|\Sigma\right\|_{\text{op}}$, and plugging (C.13) in (C.12) we obtain

$$\mathbb{P}_{Z,G}\left(\frac{\left|\mathcal{C}(\hat{\theta}_{\text{LR}}(\lambda)) - \mathcal{C}^\Sigma(\lambda)\right|}{\left\|\Sigma\right\|_{\text{op}}} \geq t\right) \leq C_1 d\exp\left(-dt^4/C_1\right), \tag{C.14}$$

which gives the thesis after absorbing the constant $\left\|\Sigma\right\|_{\text{op}}$ in t, and noticing that the bound is still true for $t \in (0, 1/2)$ since $\left\|\Sigma\right\|_{\text{op}} \geq \text{tr}(\Sigma)/2d = 1$ by Assumption 1. □

Proposition C.1. *Let $\mathcal{C}^\Sigma(\lambda)$ be defined in* (3.3)*, and let $S_x^{\Sigma+\tau(\lambda)I}$ be the Schur complement of $\Sigma + \tau(\lambda)I$ with respect to the top-left $d \times d$ block. Then, we have that*

$$\mathcal{C}^\Sigma(\lambda) = \tau(\lambda)\,{\theta_x^*}^\top\left(\Sigma_{xx} + \tau(\lambda)I\right)^{-1}\Sigma_{xy}\left(S_x^{\Sigma+\tau(\lambda)I}\right)^{-1}\Sigma_{yx}\theta_x^*. \tag{C.15}$$

Proof. During the proof, to ease the notation, we will often leave implicit the dependence of τ on λ. Then, we can write

$$\begin{aligned}
\mathcal{C}^{\Sigma}(\lambda) &= {\theta^*}^\top \Sigma \left(\Sigma + \tau I\right)^{-1} P_y \Sigma \theta^* \\
&= {\theta^*}^\top \left(\Sigma + \tau I - \tau I\right) \left(\Sigma + \tau I\right)^{-1} P_y \Sigma \theta^* \\
&= -\tau {\theta^*}^\top \left(\Sigma + \tau I\right)^{-1} P_y \Sigma \theta^* + {\theta^*}^\top P_y \Sigma \theta^* \\
&= -\tau {\theta^*}^\top \left(\Sigma + \tau I\right)^{-1} P_y \Sigma \theta^*,
\end{aligned} \tag{C.16}$$

where the last step holds since $P_y \theta^* = 0$. This expression can be further manipulated using the notation introduced in (4.1). We also introduce the following notation

$$\left(\Sigma + \tau I\right)^{-1} = \left(\begin{array}{c|c} \left[\left(\Sigma + \tau I\right)^{-1}\right]_{xx} & \left[\left(\Sigma + \tau I\right)^{-1}\right]_{xy} \\ \hline \left[\left(\Sigma + \tau I\right)^{-1}\right]_{yx} & \left[\left(\Sigma + \tau I\right)^{-1}\right]_{yy}, \end{array}\right) \tag{C.17}$$

where we divided $\left(\Sigma + \tau I\right)^{-1}$ in four $d \times d$ blocks. Notice that, the expression in (C.16) only depends on $\left[\left(\Sigma + \tau I\right)^{-1}\right]_{xy}$, *i.e.*,

$$\mathcal{C}^{\Sigma}(\lambda) = -\tau {\theta^*}^\top P_x \left(\Sigma + \tau I\right)^{-1} P_y \Sigma \theta^* = -\tau {\theta^*_x}^\top \left[\left(\Sigma + \tau I\right)^{-1}\right]_{xy} \Sigma_{yx} \theta^*_x, \tag{C.18}$$

where we denoted the projector on the first d elements of the canonical basis of $\mathbb{R}^{2d}$ as $P_x \in \mathbb{R}^{2d \times 2d}$. Exploiting the Schur complement $S_x^{\Sigma + \tau I}$, it holds that

$$\left[\left(\Sigma + \tau I\right)^{-1}\right]_{xy} = -\left(\Sigma_{xx} + \tau I\right)^{-1} \Sigma_{xy} \left(S_x^{\Sigma + \tau I}\right)^{-1}, \tag{C.19}$$

which combined with (C.18) proves (C.15). □

Proof of Proposition 4.1. During the proof, to ease the notation, we will often leave implicit the dependence of τ on λ. Then, according to (3.3), we have that

$$\left|\mathcal{C}^{\Sigma}(\lambda)\right| = \left|{\theta^*}^\top \Sigma \left(\Sigma + \tau I\right)^{-1} P_y \Sigma P_x \theta^*\right| \leq \left\|\theta^*\right\|_2^2 \left\|\Sigma \left(\Sigma + \tau I\right)^{-1}\right\|_{\text{op}} \left\|P_y \Sigma P_x\right\|_{\text{op}} \leq \left\|\Sigma_{yx}\right\|_{\text{op}}, \tag{C.20}$$

and

$$\left|\mathcal{C}^{\Sigma}(\lambda)\right| = \left|{\theta^*}^\top \Sigma \left(\Sigma + \tau I\right)^{-1} P_y \Sigma \theta^*\right| \leq \left\|\theta^*\right\|_2^2 \left\|\Sigma\right\|_{\text{op}}^2 \frac{1}{\lambda_{\min}\left(\Sigma\right) + \tau} \leq \frac{\lambda_{\max}\left(\Sigma\right)^2}{\tau}. \tag{C.21}$$

Then, using (C.15), we get

$$
\begin{aligned}
\mathcal{C}^{\Sigma}(\lambda) &= \tau\, \theta_x^{*\top} \left(\Sigma_{xx}+\tau I\right)^{-1} \Sigma_{xy} \left(S_x^{\Sigma+\tau I}\right)^{-1} \Sigma_{yx}\theta_x^* \\
&= \tau\, \theta_x^{*\top} \left(\Sigma_{xx}+\tau I\right)^{-1/2} \left(\Sigma_{xx}+\tau I\right)^{-1/2} \Sigma_{xy} \left(S_x^{\Sigma+\tau I}\right)^{-1} \Sigma_{yx} \left(\Sigma_{xx}+\tau I\right)^{-1/2} \left(\Sigma_{xx}+\tau I\right)^{1/2} \theta_x^* \\
&\le \tau \left\| \left(\Sigma_{xx}+\tau I\right)^{-1/2} \theta_x^* \right\|_2 \left\| \left(\Sigma_{xx}+\tau I\right)^{1/2} \theta_x^* \right\|_2 \left\| \left(\Sigma_{xx}+\tau I\right)^{-1/2} \Sigma_{xy} \left(S_x^{\Sigma+\tau I}\right)^{-1} \Sigma_{yx} \left(\Sigma_{xx}+\tau I\right)^{-1/2} \right\|_{\text{op}} \\
&\le \tau \frac{1}{\sqrt{\lambda_{\min}\left(\Sigma_{xx}\right)+\tau}} \sqrt{\theta_x^{*\top}\Sigma_{xx}\theta_x^* + \tau} \left\| \left(\Sigma_{xx}+\tau I\right)^{-1/2} \Sigma_{xy} \left(S_x^{\Sigma+\tau I}\right)^{-1} \Sigma_{yx} \left(\Sigma_{xx}+\tau I\right)^{-1/2} \right\|_{\text{op}} \\
&\le \tau \frac{\sqrt{\mathbb{E}_{x\sim \mathcal{P}_X}\left[\left(x^\top \theta_x^*\right)^2\right]+\tau}}{\sqrt{\lambda_{\min}\left(\Sigma_{xx}\right)+\tau}} \frac{\left\| \left(\Sigma_{xx}+\tau I\right)^{-1/2} \Sigma_{xy} \right\|_{\text{op}}^2}{\lambda_{\min}\left(S_x^{\Sigma+\tau I}\right)} \\
&= \tau \frac{\sqrt{\mathbb{E}_{g=x^\top\theta_x^*+\epsilon}\left[g^2\right]-\sigma^2+\tau}}{\sqrt{\lambda_{\min}\left(\Sigma_{xx}\right)+\tau}} \frac{\lambda_{\max}\left(\Sigma_{yx}\left(\Sigma_{xx}+\tau I\right)^{-1}\Sigma_{xy}\right)}{\lambda_{\min}\left(\Sigma_{yy}+\tau I-\Sigma_{yx}\left(\Sigma_{xx}+\tau I\right)^{-1}\Sigma_{xy}\right)} \\
&\le \tau \frac{\sqrt{\mathbb{E}_g\left[g^2\right]-\sigma^2+\tau}}{\sqrt{\lambda_{\min}\left(\Sigma_{xx}\right)+\tau}} \frac{\lambda_{\max}\left(\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}\right)}{\lambda_{\min}\left(\Sigma_{yy}+\tau I-\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}\right)} \\
&= \tau \frac{\sqrt{\mathbb{E}_g\left[g^2\right]-\sigma^2+\tau}}{\sqrt{\lambda_{\min}\left(\Sigma_{xx}\right)+\tau}} \frac{\lambda_{\max}\left(\Sigma_{yy}\right)-\lambda_{\min}\left(S_x^{\Sigma}\right)}{\lambda_{\min}\left(S_x^{\Sigma}\right)+\tau} \\
&\le \tau \sqrt{\mathrm{Var}(g)-\sigma^2}\, \frac{\lambda_{\max}\left(\Sigma_{yy}\right)-\lambda_{\min}\left(S_x^{\Sigma}\right)}{\lambda_{\min}\left(S_x^{\Sigma}\right)\sqrt{\lambda_{\min}\left(\Sigma_{xx}\right)}},
\end{aligned}
\tag{C.22}
$$

where in the fifth line we denoted with $\mathcal{P}_X$ the marginal distribution of the core feature x, and $\mathbb{E}_g[\cdot]$ from the sixth line on denotes an expectation with respect to g distributed as the labels of the model. The last step simplifies the expression with respect to τ, and it holds since $\mathrm{Var}(g)-\sigma^2 = \theta_x^{*\top}\Sigma_{xx}\theta_x^* \ge \lambda_{\min}\left(\Sigma_{xx}\right)$. This, together with (C.20) and (C.21) gives the desired result. □

Proof of Proposition 4.2. During the proof, to ease the notation, we will often leave implicit the dependence of τ on λ. Then, as in Han & Xu (2023) and in the proof of Theorem 1, we define the Gaussian sequence model $\hat{\theta}^\rho \in \mathbb{R}^{2d}$ as (C.9) where ρ is a standard Gaussian vector in $\mathbb{R}^{2d}$ and $\gamma > 0$ is implicitly defined via

$$
\frac{n\gamma^2}{2d} = \sigma^2 + \mathbb{E}_\rho\left[\left\|\Sigma^{1/2}\left(\hat{\theta}^\rho-\theta^*\right)\right\|_2^2\right] = \sigma^2 + \left\|\Sigma^{1/2}\left(\left(\Sigma+\tau I\right)^{-1}\Sigma - I\right)\theta^*\right\|_2^2 + \frac{\gamma^2}{2d}\,\mathrm{tr}\left(\left(\Sigma+\tau I\right)^{-2}\Sigma^2\right), \tag{C.23}
$$

which also reads

$$
\frac{n\gamma^2}{2d} = \frac{\sigma^2 + \tau^2\left\|\left(\Sigma+\tau I\right)^{-1}\Sigma^{1/2}\theta^*\right\|_2^2}{1-\frac{\mathrm{tr}\left(\left(\Sigma+\tau I\right)^{-2}\Sigma^2\right)}{n}}. \tag{C.24}
$$

Then, due to Theorem 3.1 of Han & Xu (2023) on the prediction risk, since $\mathcal{P}_{XY}$ is a multivariate Gaussian due to Assumption 1, we have that, for any $t \in (0, 1/2)$,

$$
\mathbb{P}_{Z,G}\left(\left|\mathcal{L}(\hat{\theta}_{\text{LR}}(\lambda)) - \mathcal{L}^{\Sigma}(\lambda)\right| \ge t\right) \le Cd\exp\left(-dt^4/C\right), \tag{C.25}
$$

where C is a positive constant depending on $\lambda_{\min}\left(\Sigma\right)$, $\|\Sigma\|_{\text{op}}$, σ^2, and $n/d = \Theta(1)$. □

Proof of Proposition 4.3 As $\tau(\lambda)$ is an increasing function of λ, all the statements on the monotonicity of $\mathcal{L}^{\Sigma}(\lambda)$ and $\mathcal{C}^{\Sigma}(\lambda)$ can be proved by showing monotonicity w.r.t. τ (whose dependence w.r.t. λ is left implicit throughout the argument). In particular, we have

$$\frac{\mathrm{d}\mathcal{L}^{\Sigma}(\lambda)}{\mathrm{d}\tau} = \frac{\frac{\mathrm{d}}{\mathrm{d}\tau}\left(\tau^2\left\|(\Sigma+\tau I)^{-1}\Sigma^{1/2}\theta^*\right\|_2^2\right)\left(1-\frac{\mathrm{tr}\left((\Sigma+\tau I)^{-2}\Sigma^2\right)}{n}\right)}{\left(1-\frac{\mathrm{tr}\left((\Sigma+\tau I)^{-2}\Sigma^2\right)}{n}\right)^2} - \frac{\left(\sigma^2+\tau^2\left\|(\Sigma+\tau I)^{-1}\Sigma^{1/2}\theta^*\right\|_2^2\right)\frac{\mathrm{d}}{\mathrm{d}\tau}\left(1-\frac{\mathrm{tr}\left((\Sigma+\tau I)^{-2}\Sigma^2\right)}{n}\right)}{\left(1-\frac{\mathrm{tr}\left((\Sigma+\tau I)^{-2}\Sigma^2\right)}{n}\right)^2}. \tag{C.26}$$

To study the sign of the above expression, it suffices to focus on the numerators, as the denominator is always positive.

Note that the RHS of (3.4) is smaller or equal to $2d/n$; thus, as $2d < n$, we also get $\tau \leq \lambda(1-2d/n)^{-1}$, which implies $\tau(\lambda) \to 0$ as $\lambda \to 0$. Hence, to show that $\mathcal{L}^{\Sigma}(\lambda)$ is monotonically decreasing in a right neighborhood of $\lambda = 0$, it suffices to show that (C.26) evaluated in $\tau = 0$ is strictly negative. For $\tau = 0$, the first factor in the numerator of the first term in (C.26) is 0, as the following chain of equalities holds:

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\tau}\left(\tau^2\left\|(\Sigma+\tau I)^{-1}\Sigma^{1/2}\theta^*\right\|_2^2\right) &= \frac{\mathrm{d}}{\mathrm{d}\tau}\left(\left\|(\Sigma/\tau+I)^{-1}\Sigma^{1/2}\theta^*\right\|_2^2\right)\\
&= \theta^{*\top}\frac{\mathrm{d}}{\mathrm{d}\tau}(\Sigma/\tau+I)^{-2}\Sigma\theta^*\\
&= -\theta^{*\top}(\Sigma/\tau+I)^{-2}\left(\frac{\mathrm{d}}{\mathrm{d}\tau}(\Sigma/\tau+I)^2\right)(\Sigma/\tau+I)^{-2}\Sigma\theta^*\\
&= -\theta^{*\top}(\Sigma/\tau+I)^{-2}\left(-\frac{2\Sigma}{\tau^2}(\Sigma/\tau+I)\right)(\Sigma/\tau+I)^{-2}\Sigma\theta^*\\
&= 2\tau\theta^{*\top}(\Sigma+\tau I)^{-3}\Sigma^2\theta^*.
\end{aligned} \tag{C.27}$$

Furthermore, the second term gives

$$-\sigma^2\frac{\mathrm{d}}{\mathrm{d}\tau}\left(1-\frac{\mathrm{tr}\left((\Sigma+\tau I)^{-2}\Sigma^2\right)}{n}\right) = \frac{\sigma^2}{n}\frac{\mathrm{d}}{\mathrm{d}\tau}\left(\sum_{k=1}^{2d}\frac{\lambda_k^2}{(\lambda_k+\tau)^2}\right) = -\frac{2\sigma^2}{n}\sum_{k=1}^{2d}\frac{\lambda_k^2}{(\lambda_k+\tau)^3} < 0, \tag{C.28}$$

where λ_k denotes the k-th eigenvalue of Σ. This gives the first claim.

To show that there exists $\lambda_{\mathcal{L}} > 0$ such that $\mathcal{L}^{\Sigma}(\lambda)$ is monotonically increasing for $\lambda \geq \lambda_{\mathcal{L}}$ we will show that the derivative of $\mathcal{L}^{\Sigma}(\lambda)$ with respect to τ is positive for all $\tau \geq \tau_{\mathcal{L}} := \tau(\lambda_{\mathcal{L}})$. For simplicity, in the rest of the argument we use the notation $\lambda_{\max}$ and $\lambda_{\min}$ to indicate the largest and smallest eigenvalues of Σ, respectively. Instead, the notation $\lambda_{\min}(\cdot)$ still represents the smallest eigenvalue of its argument. For the first factor of the first term of (C.26), continuing from (C.27), we have

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\left(\tau^2\left\|(\Sigma+\tau I)^{-1}\Sigma^{1/2}\theta^*\right\|_2^2\right) \geq 2\frac{1}{\lambda_{\max}(\Sigma/\tau+I)^3}\lambda_{\min}(\Sigma/\tau)^2 = \frac{2\lambda_{\min}^2}{\tau^2(\lambda_{\max}/\tau+1)^3}. \tag{C.29}$$

For the second factor of the first term of (C.26), we have

$$1-\frac{\mathrm{tr}\left((\Sigma+\tau I)^{-2}\Sigma^2\right)}{n} = 1-\frac{1}{n}\sum_{k=1}^{2d}\frac{\lambda_k^2}{(\lambda_k+\tau)^2} \geq 1-\frac{2d\lambda_{\max}^2}{n\tau^2}. \tag{C.30}$$

For the first factor of the second term of (C.26), we have

$$\tau^2\left\|(\Sigma+\tau I)^{-1}\Sigma^{1/2}\theta^*\right\|_2^2 \leq \tau^2\frac{\lambda_{\max}}{(\lambda_{\min}+\tau)^2} \tag{C.31}$$

For the second factor of the second term of (C.26), we have

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\left(1-\frac{\operatorname{tr}\left((\Sigma+\tau I)^{-2}\Sigma^2\right)}{n}\right)=-\frac{1}{n}\frac{\mathrm{d}}{\mathrm{d}\tau}\left(\sum_{k=1}^{2d}\frac{\lambda_k^2}{(\lambda_k+\tau)^2}\right)=\frac{2}{n}\sum_{k=1}^{2d}\frac{\lambda_k^2}{(\lambda_k+\tau)^3}\leq\frac{4d\lambda_{\max}^2}{n\tau^3}. \tag{C.32}$$

Thus, putting together (C.26), (C.29), (C.30), (C.31), and (C.32), the monotonicity of $\mathcal{L}^{\Sigma}(\lambda)$ is implied by

$$\frac{2\lambda_{\min}^2}{\tau^2\left(\lambda_{\max}/\tau+1\right)^3}\left(1-\frac{2d\lambda_{\max}^2}{n\tau^2}\right)\overset{?}{\geq}\left(\sigma^2+\tau^2\frac{\lambda_{\max}}{(\lambda_{\min}+\tau)^2}\right)\frac{4d\lambda_{\max}^2}{n\tau^3}. \tag{C.33}$$

Now, we have that the above inequality holds for sufficiently large τ: the LHS is $\Theta(1/\tau^2)$ (considering fixed the other quantities), while the RHS is $\Theta(1/\tau^3)$; and the desired statement is therefore proved.

Next, we set $\tau_{\mathcal{C}} := \tau(\lambda_{\mathcal{C}}) = \sqrt{\lambda_{\min}\left(S_x^{\Sigma}\right)}$ and show that $\mathcal{C}^{\Sigma}(\lambda)$ is monotonically increasing for $\tau\in[0,\tau_{\mathcal{C}}]$. Plugging $\Sigma_{xx}=I$ in (C.15) we get

$$\mathcal{C}^{\Sigma}(\lambda)=\tau\,\theta_x^{*\top}\left(\Sigma_x+\tau I\right)^{-1}\Sigma_{xy}\left(S_x^{\Sigma+\tau I}\right)^{-1}\Sigma_{yx}\theta_x^*=\frac{\tau}{1+\tau}\,\theta_x^{*\top}\Sigma_{xy}\left(\Sigma_{yy}+\tau I-\frac{\Sigma_{yx}\Sigma_{xy}}{1+\tau}\right)^{-1}\Sigma_{yx}\theta_x^*. \tag{C.34}$$

By the product rule, and introducing the shorthand $A(\tau)=\Sigma_{yy}+\tau I-\frac{\Sigma_{xy}^\top\Sigma_{xy}}{1+\tau}$, we have

$$\begin{aligned}
\frac{\mathrm{d}C^{\Sigma}(\lambda)}{\mathrm{d}\tau}&=\left(\frac{\mathrm{d}}{\mathrm{d}\tau}\left(\frac{\tau}{1+\tau}\right)\right)\left(\theta_x^{*\top}\Sigma_{xy}A(\tau)^{-1}\Sigma_{xy}^\top\theta_x^*\right)+\left(\frac{\tau}{1+\tau}\right)\left(\frac{\mathrm{d}}{\mathrm{d}\tau}\left(\theta_x^{*\top}\Sigma_{xy}A(\tau)^{-1}\Sigma_{xy}^\top\theta_x^*\right)\right)\\
&=\frac{1}{(1+\tau)^2}\,\theta_x^{*\top}\Sigma_{xy}A(\tau)^{-1}\Sigma_{xy}^\top\theta_x^*+\left(\frac{\tau}{1+\tau}\right)\left(\theta_x^{*\top}\Sigma_{xy}A(\tau)^{-1}\left(-\frac{\mathrm{d}}{\mathrm{d}\tau}A(\tau)\right)A(\tau)^{-1}\Sigma_{xy}^\top\theta_x^*\right)\\
&=\frac{1}{(1+\tau)^2}\theta_x^{*\top}\Sigma_{xy}A(\tau)^{-1}\Sigma_{xy}^\top\theta_x^*-\frac{\tau}{1+\tau}\left(\theta_x^{*\top}\Sigma_{xy}A(\tau)^{-1}\left(I+\frac{\Sigma_{xy}^\top\Sigma_{xy}}{(1+\tau)^2}\right)A(\tau)^{-1}\Sigma_{xy}^\top\theta_x^*\right)\\
&=\frac{1}{(1+\tau)}\theta_x^{*\top}\Sigma_{xy}A(\tau)^{-1}\left(\frac{A(\tau)}{1+\tau}-\tau\left(I+\frac{\Sigma_{xy}^\top\Sigma_{xy}}{(1+\tau)^2}\right)\right)A(\tau)^{-1}\Sigma_{xy}^\top\theta_x^*\\
&=\frac{1}{(1+\tau)}\theta_x^{*\top}\Sigma_{xy}A(\tau)^{-1}\left(\frac{\Sigma_{yy}+\tau I}{1+\tau}-\frac{\Sigma_{xy}^\top\Sigma_{xy}}{(1+\tau)^2}-\tau I-\tau\frac{\Sigma_{xy}^\top\Sigma_{xy}}{(1+\tau)^2}\right)A(\tau)^{-1}\Sigma_{xy}^\top\theta_x^*\\
&=\frac{1}{(1+\tau)^2}\theta_x^{*\top}\Sigma_{xy}A(\tau)^{-1}\left(\Sigma_{yy}-\Sigma_{xy}^\top\Sigma_{xy}-\tau^2 I\right)A(\tau)^{-1}\Sigma_{xy}^\top\theta_x^*\\
&=\frac{1}{(1+\tau)^2}\theta_x^{*\top}\Sigma_{xy}A(\tau)^{-1}\left(S_x^{\Sigma}-\tau^2 I\right)A(\tau)^{-1}\Sigma_{xy}^\top\theta_x^*,
\end{aligned} \tag{C.35}$$

where in the second line we used the identity $\frac{\mathrm{d}}{\mathrm{d}\tau}\left(A(\tau)^{-1}\right)=A(\tau)^{-1}\left(-\frac{\mathrm{d}}{\mathrm{d}\tau}A(\tau)\right)A(\tau)^{-1}$. Then, if $\tau\leq\sqrt{\lambda_{\min}\left(S_x^{\Sigma}\right)}=\tau_{\mathcal{C}}$, we have that $\left(S_x^{\Sigma}-\tau^2 I\right)$ is p.s.d., which in turn implies $\frac{\mathrm{d}C^{\Sigma}(\lambda)}{\mathrm{d}\tau}\geq 0$, thus giving the desired claim. The non-negativity of $\mathcal{C}^{\Sigma}(\lambda)$ readily follows from (C.34).

For the last statement, setting $\tau_{\mathcal{L}}=\lambda_{\min}\left(\Sigma\right)$ we show that $\mathcal{L}^{\Sigma}(\lambda)$ is monotonically increasing for all $\tau\in[\tau_{\mathcal{L}},+\infty)$ as long as the additional bound on $2d/n$ holds. As $\lambda_{\min}\left(S_x^{\Sigma}\right)\leq\lambda_{\min}\left(\Sigma_{yy}\right)\leq\operatorname{tr}\left(\Sigma_{yy}\right)/d=\operatorname{tr}\left(\Sigma-\Sigma_{xx}\right)/d=1$, we also have

$$\tau_{\mathcal{C}}=\sqrt{\lambda_{\min}\left(S_x^{\Sigma}\right)}\geq\lambda_{\min}\left(S_x^{\Sigma}\right)\geq\lambda_{\min}\left(\Sigma\right)=\tau_{\mathcal{L}}, \tag{C.36}$$

where the second inequality follows from Lemma C.3. Thus, from the monotonicity of $\tau(\lambda)$ in λ, the final result readily follows.

It remains to prove the monotonicity of $\mathcal{L}^{\Sigma}(\lambda)$ in $[\lambda_{\min}\left(\Sigma\right),+\infty)$. To do so, we again study the sign of (C.26).

For the first factor of the first term of (C.26), we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\tau}\left(\tau^2\left\|\left(\Sigma+\tau I\right)^{-1}\Sigma^{1/2}\theta^*\right\|_2^2\right) &= 2\theta^{*\top}\left(\Sigma/\tau+I\right)^{-3}\left(\Sigma/\tau\right)^2\theta^* \\ &= 2\theta^{*\top}\Sigma^{1/2}\left(\Sigma+\tau I\right)^{-1}\left(\Sigma+\tau I\right)^{-1}\tau\Sigma\left(\Sigma+\tau I\right)^{-1}\Sigma^{1/2}\theta^* \\ &\geq \frac{2}{\tau}\lambda_{\min}\left(\Sigma\left(\Sigma+\tau\right)^{-1}\right)\tau^2\left\|\left(\Sigma+\tau I\right)^{-1}\Sigma^{1/2}\theta^*\right\|_2^2 \\ &= \frac{2}{\tau}\frac{\lambda_{\min}}{\lambda_{\min}+\tau}\tau^2\left\|\left(\Sigma+\tau I\right)^{-1}\Sigma^{1/2}\theta^*\right\|_2^2.\end{aligned} \tag{C.37}$$

For the second factor of the first term of (C.26), we have

$$1-\frac{\operatorname{tr}\left(\left(\Sigma+\tau I\right)^{-2}\Sigma^2\right)}{n} = 1-\frac{1}{n}\sum_{k=1}^{2d}\frac{\lambda_k^2}{\left(\lambda_k+\tau\right)^2} \geq 1-\frac{2d}{n}. \tag{C.38}$$

For the second factor of the second term of (C.26), we have

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\left(1-\frac{\operatorname{tr}\left(\left(\Sigma+\tau I\right)^{-2}\Sigma^2\right)}{n}\right) = \frac{2}{n}\sum_{k=1}^{2d}\frac{\lambda_k^2}{\left(\lambda_k+\tau\right)^3} \leq \frac{2}{n}\frac{1}{\tau\left(\lambda_{\min}+\tau\right)}\sum_{k=1}^{2d}\frac{\lambda_k^2}{\left(\lambda_k+\tau\right)} \leq \frac{4d}{n\tau\left(\lambda_{\min}+\tau\right)}. \tag{C.39}$$

Thus, putting together (C.26), (C.37), (C.38), and (C.39), the monotonicity of $\mathcal{L}^{\Sigma}(\lambda)$ is implied by

$$\frac{2}{\tau}\frac{\lambda_{\min}}{\lambda_{\min}+\tau}\tau^2\left\|\left(\Sigma+\tau I\right)^{-1}\Sigma^{1/2}\theta^*\right\|_2^2\left(1-\frac{2d}{n}\right) \overset{?}{\geq} \left(\sigma^2+\tau^2\left\|\left(\Sigma+\tau I\right)^{-1}\Sigma^{1/2}\theta^*\right\|_2^2\right)\frac{4d}{n\tau\left(\lambda_{\min}+\tau\right)} \tag{C.40}$$

Since we assumed that $2d/n \leq \lambda_{\min}/4 \leq 1/4$, we have

$$\frac{2}{\tau}\frac{\lambda_{\min}}{\lambda_{\min}+\tau}\left(1-\frac{2d}{n}\right)-\frac{4d}{n\tau\left(\lambda_{\min}+\tau\right)} = \frac{2}{\tau}\frac{\lambda_{\min}}{\lambda_{\min}+\tau}\left(1-\frac{2d}{n}-\frac{2d}{n\lambda_{\min}}\right) \geq \frac{1}{\tau}\frac{\lambda_{\min}}{\lambda_{\min}+\tau}, \tag{C.41}$$

and

$$\begin{aligned}\tau^2\left\|\left(\Sigma+\tau I\right)^{-1}\Sigma^{1/2}\theta^*\right\|_2^2 &\geq \lambda_{\min}\left(\tau^2\Sigma\left(\Sigma+\tau I\right)^{-2}\right) \\ &= \min_k\frac{\tau^2\lambda_k}{\left(\lambda_k+\tau\right)^2} \\ &= \min_k\frac{\lambda_k}{\left(\lambda_k/\tau+1\right)^2} \\ &\geq \min_k\frac{\lambda_k}{\left(\lambda_k/\lambda_{\min}+1\right)^2} \\ &= \lambda_{\min}\min_k\frac{\lambda_k/\lambda_{\min}}{\left(\lambda_k/\lambda_{\min}+1\right)^2} \\ &= \frac{\lambda_{\max}}{\left(\lambda_{\max}/\lambda_{\min}+1\right)^2},\end{aligned} \tag{C.42}$$

where in the fourth line we used that $\tau \geq \lambda_{\min}$, and in the last step we used that $f(x) := x/(x+1)^2$ is decreasing for $x \geq 1$.

Thus, using (C.41) and (C.42) gives that (C.40) is implied by

$$\frac{\lambda_{\max}}{\left(\lambda_{\max}/\lambda_{\min}+1\right)^2}\frac{1}{\tau}\frac{\lambda_{\min}}{\lambda_{\min}+\tau} \overset{?}{\geq} \sigma^2\frac{4d}{n\tau\left(\lambda_{\min}+\tau\right)}, \tag{C.43}$$

which holds since we assumed

$$\frac{2d}{n} \leq \frac{1}{2\sigma^2}\frac{\lambda_{\max}\lambda_{\min}}{\left(\lambda_{\max}/\lambda_{\min}+1\right)^2}. \tag{C.44}$$

□

C.1 Proofs on S_x^Σ

For completeness, in this section we prove two known results about S_x^Σ.

Lemma C.2. *Let $z = \left[x^\top, y^\top\right]^\top \sim \mathcal{P}_{XY}$ be distributed according to a mean-0, multivariate Gaussian distribution with covariance Σ, such that Σ is invertible. Then, the Schur complement S_x^Σ of Σ with respect to the top left block Σ_{xx} (see* (4.1)*) corresponds to the conditional covariance of y given x,* i.e.*,*

$$S_x^\Sigma = \mathrm{Cov}\left(y|x=\bar{x}\right) = \mathbb{E}_{y|x=\bar{x}}\left[\left(y - \mathbb{E}_{y|x=\bar{x}}[y]\right)\left(y - \mathbb{E}_{y|x=\bar{x}}[y]\right)^\top\right]. \tag{C.45}$$

Proof. Consider the expression $z^\top \Sigma^{-1} z$. According to the notation in (4.1) and in (C.17), we have

$$z^\top \Sigma^{-1} z = x^\top \left[\Sigma^{-1}\right]_{xx} x + y^\top \left[\Sigma^{-1}\right]_{yy} y + x^\top \left[\Sigma^{-1}\right]_{xy} y + y^\top \left[\Sigma^{-1}\right]_{yx} x. \tag{C.46}$$

Then, the formulas for the inverse of a block matrix give

$$\begin{aligned} &z^\top \Sigma^{-1} z \\ &= x^\top \left(\Sigma_{xx}^{-1} + \Sigma_{xx}^{-1}\Sigma_{xy} S_x^{\Sigma^{-1}} \Sigma_{yx}\Sigma_{xx}^{-1}\right) x + y^\top S_x^{\Sigma^{-1}} y + x^\top \left(-\Sigma_{xx}^{-1}\Sigma_{xy} S_x^{\Sigma^{-1}}\right) y + y^\top \left(-S_x^{\Sigma^{-1}}\Sigma_{yx}\Sigma_{xx}^{-1}\right) x \\ &= x^\top \Sigma_{xx}^{-1} x + \left(y - \Sigma_{yx}\Sigma_{xx}^{-1} x\right)^\top S_x^{\Sigma^{-1}} \left(y - \Sigma_{yx}\Sigma_{xx}^{-1} x\right). \end{aligned} \tag{C.47}$$

Then, denoting with $p(x,y)$ and $p(x)$ the probability density functions of $z = \left[x^\top, y^\top\right]^\top$ and x respectively, we get that the probability density function of y conditioned on x takes the form

$$\begin{aligned} p(y|x) &= \frac{p(x,y)}{p(x)} \\ &= \frac{\sqrt{(2\pi)^d \det\left(\Sigma_{xx}\right)}}{\sqrt{(2\pi)^{2d} \det\left(\Sigma\right)}} \frac{\exp\left(-\left[x^\top, y^\top\right]\Sigma^{-1}\left[x^\top, y^\top\right]^\top/2\right)}{\exp\left(-x^\top \Sigma_{xx}^{-1} x/2\right)} \\ &= \frac{1}{\sqrt{(2\pi)^d \det\left(S_x^\Sigma\right)}} \exp\left(-\left[x^\top, y^\top\right]\Sigma^{-1}\left[x^\top, y^\top\right]^\top/2 + x^\top \Sigma_{xx}^{-1} x/2\right) \\ &= \frac{\exp\left(-\left(y - \Sigma_{yx}\Sigma_{xx}^{-1} x\right)^\top S_x^{\Sigma^{-1}}\left(y - \Sigma_{yx}\Sigma_{xx}^{-1} x\right)/2\right)}{\sqrt{(2\pi)^d \det\left(S_x^\Sigma\right)}}, \end{aligned} \tag{C.48}$$

where we used Schur formula for the determinants in the third line, and (C.47) in the last step. Thus, we have that $p(y|x)$ describes the density of a multivariate Gaussian random variable, with covariance S_x^Σ. □

Lemma C.3. *Let $\Sigma \in \mathbb{R}^{2d\times 2d}$ be a p.s.d., invertible matrix. Then, the Schur complement $S_x^\Sigma \in \mathbb{R}^{d\times d}$ of Σ with respect to the top left block Σ_{xx} (see* (4.1)*) is such that*

$$\lambda_{\min}\left(S_x^\Sigma\right) \geq \lambda_{\min}\left(\Sigma\right). \tag{C.49}$$

Proof. Let $\Gamma \in \mathbb{R}^{2d\times d}$ be the rank-d matrix defined as

$$\Gamma = \left(\frac{\Sigma_{xx}^{1/2}}{\Sigma_{yx}\Sigma_{xx}^{-1/2}}\right), \tag{C.50}$$

and $S \in \mathbb{R}^{2d\times 2d}$ as the matrix containing S_x^Σ in its bottom-right $d \times d$ block, and 0 everywhere else. Then, we have that

$$\Sigma = S + \Gamma\Gamma^\top, \tag{C.51}$$

where both S and $\Gamma\Gamma^\top$ are rank-d p.s.d. matrices.

Denoting by $\lambda_k(S)$ the k-th largest eigenvalue of S, by the Courant–Fischer–Weyl min-max principle, we can write

$$\lambda_k(S) = \max\nolimits_{W,\,\dim(W)=k} \min\nolimits_{u\in W,\,\|u\|_2=1}\left(u^\top S u\right), \tag{C.52}$$

where with W we denote a generic k-dimensional subspace of $\mathbb{R}^{2d}$. Thus, the desired result follows from

$$\begin{aligned}\lambda_{\min}(\Sigma) &= \lambda_{\min}\left(S+\Gamma\Gamma^\top\right)\\ &= \min_{\|u\|_2=1} u^\top\left(S+\Gamma\Gamma^\top\right)u\\ &\leq \min_{u\in\ker(\Gamma\Gamma^\top),\,\|u\|_2=1} u^\top\left(S+\Gamma\Gamma^\top\right)u\\ &= \min_{u\in\ker(\Gamma\Gamma^\top),\,\|u\|_2=1} u^\top S u\\ &\leq \max_{W,\,\dim(W)=d}\min_{u\in W,\,\|u\|_2=1} u^\top S u\\ &= \lambda_d(S)\\ &= \lambda_{\min}\left(S_x^\Sigma\right),\end{aligned} \tag{C.53}$$

where the last step holds since the d smallest eigenvalues of S are equal to 0, and the d largest correspond to the ones of S_x^Σ. □

C.2 Remarks on Assumption 1

Our results on linear regression rely on Assumption 1, and in particular on the training samples to be normally distributed. This assumption is made for technical convenience, as the concentration results in Theorem 1 and Proposition 4.2 still hold under the following milder requirement.

Assumption 2 (Data distribution)**.** *The input samples $\{z_i\}_{i=1}^n$ are n i.i.d. samples from a mean-0, sub-Gaussian distribution $\mathcal{P}_{XY}$, such that*

1. *its covariance $\Sigma \in \mathbb{R}^{2d\times 2d}$ is invertible, with $\lambda_{\max}(\Sigma) = \mathcal{O}(1)$, $\lambda_{\min}(\Sigma) = \Omega(1)$, and $\mathrm{tr}(\Sigma) = 2d$;*
2. *for $z \sim P_Z$, the random variable $\Sigma^{-1/2}z$ has independent, mean-0, unit variance, sub-Gaussian entries.*

This assumption resembles the requirements A-B in Section 2.2 in Han & Xu (2023), where we also included the scaling of the trace. To formally state the equivalent of Theorem 1 and Proposition 4.2, one also has to enforce the following technical condition on the true parameter θ^*.

Assumption 3. *Let $\delta = 1/72$, then we assume that*

$$\theta^* \text{ s.t. } \left\|\Sigma^{1/2}\tau(\Sigma+\tau I)^{-1}\theta^*\right\|_\infty \leq C d^{\delta-1/2}. \tag{C.54}$$

In Proposition 10.3 in Han & Xu (2023), it is shown that this condition excludes a negligible fraction ($Ce^{-n^{2\delta}/C}$) of the θ^* on the unit ball. Since we set $\delta = 1/72$, following the same arguments of the proofs of Theorem 1 and Proposition 4.2, we have that Theorems 2.4 and 3.1 in Han & Xu (2023) imply the results below.

Theorem 3. *Let Assumptions 6 and 3 hold, and let $n = \Theta(d)$. Let $\hat{\theta}_{\mathrm{LR}}(\lambda)$ be defined as in* (3.2)*, and let $\mathcal{C}(\hat{\theta}_{\mathrm{LR}}(\lambda))$ be the amount of spurious correlations learned by the model $f_{\mathrm{LR}}\left(\hat{\theta}_{\mathrm{LR}}(\lambda)\right)$ as defined in* (2.3)*. Then, for any $\lambda > 0$, we have that, for every $t \in (0, 1/2)$,*

$$\mathbb{P}_{Z,G}\left(\left|\mathcal{C}(\hat{\theta}_{\mathrm{LR}}(\lambda)) - \mathcal{C}^\Sigma(\lambda)\right| \geq t\right) \leq C t^{-13} d^{-1/8}, \tag{C.55}$$

where $\mathcal{C}^\Sigma(\lambda)$ is defined in (3.3)*, and C is a an absolute constant.*

Proposition C.4. *Let Assumptions 6 and 3 hold, and let $n = \Theta(d)$. Let $\hat{\theta}_{\mathrm{LR}}(\lambda)$ be defined as in* (3.2)*, and let $\mathcal{L}(\hat{\theta}_{\mathrm{LR}}(\lambda))$ be the in-distribution test loss of the model $f_{\mathrm{LR}}(\hat{\theta}_{\mathrm{LR}}(\lambda))$ as defined in* (3.1)*. Then, , for any $\lambda > 0$, we have that, for every $t \in (0, 1/2)$,*

$$\mathbb{P}_{Z,G}\left(\left|\mathcal{L}(\hat{\theta}_{\mathrm{LR}}(\lambda)) - \mathcal{L}^\Sigma(\lambda)\right| \geq t\right) \leq C t^{-c} d^{-1/6.5}, \tag{C.56}$$

where $\mathcal{L}^\Sigma(\lambda)$ is defined in (4.3)*, and C and c are positive absolute constants.*

D Proofs for Random Features

Assumption 4 (Activation function)**.** *The activation $\phi : \mathbb{R} \to \mathbb{R}$ is a non-linear, odd, Lipschitz function, such that its first Hermite coefficient $\mu_1 \neq 0$.*

This choice is motivated by theoretical convenience and is similar to the one considered in Hu & Lu (2023). We believe that our result can be extended to a more general setting, as the ones in Mei & Montanari (2022); Mei et al. (2022), with a more involved analysis. We refer to O'Donnell (2014) for background on Hermite coefficients.

Assumption 5 (Over-parameterization)**.** *We let p grow s.t. $p = \omega\left(n \log^4 n\right)$ and $\log p = \Theta(\log n)$.*

This requires the width of the model (and, hence, its number of parameters) to grow faster (by at least a poly-log factor) than the number of training samples.

Finally, our requirements on the data are less restrictive than those coming from Assumption 1.

Assumption 6 (Data distribution, less restrictive)**.** *$\{z_i\}_{i=1}^n$ are n i.i.d. samples from a mean-0, Lipschitz concentrated distribution $\mathcal{P}_{XY}$, with covariance Σ s.t. $\operatorname{tr}(\Sigma) = 2d$. Furthermore, the labels g_i are i.i.d. sub-Gaussian random variables.*

Note that the labels g_i are not required to follow a linear model $g_i = z_i^\top \theta^* + \epsilon_i$. The Lipschitz concentration property (see Appendix A for details) corresponds to data having well-behaved tails, it includes the distributions considered in Assumption 1, as well as the uniform distribution on the sphere or the hypercube (Vershynin, 2018), and it is a common requirement in the related literature (Nguyen et al., 2021; Bubeck & Sellke, 2021; Bombari et al., 2022).

Lemma D.1. *We have that*

$$\left\| V \right\|_{\text{op}} = \mathcal{O}\left(\sqrt{\frac{p}{d}}\right), \tag{D.1}$$

$$\left\| Z \right\|_{\text{op}} = \mathcal{O}\left(\sqrt{d}\right), \tag{D.2}$$

with probability at least $1 - 2\exp\left(-cd\right)$ over V and Z, where c is an absolute constant. Furthermore, for every $i \in [n]$, we have

$$\left\| \left\| z_i \right\|_2 - \sqrt{2d} \right\|_{\psi_2} = \mathcal{O}\left(1\right). \tag{D.3}$$

Proof. V has independent, mean-0, unit variance, sub-Gaussian entries. Then, the first statement is a direct consequences of Theorem 4.4.5 of Vershynin (2018) and of the scaling $d = o(p)$.

By Assumption 6, we have that Z has i.i.d. mean-0, Lipschitz concentrated rows. This property also implies that the rows are i.i.d. sub-Gaussian. Thus, by Remark 5.40 in Vershynin (2012), we have that

$$\left\| Z^\top Z - n\Sigma \right\|_{\text{op}} = \mathcal{O}\left(n \frac{d}{n}\right) = \mathcal{O}\left(d\right), \tag{D.4}$$

with probability at least $1 - 2\exp\left(-c_1 d\right)$. Then, conditioning on this high probability event, by Weyl's inequality, we have

$$\left\| Z^\top Z \right\|_{\text{op}} \leq \left\| n\Sigma \right\|_{\text{op}} + \left\| Z^\top Z - n\Sigma \right\|_{\text{op}} = \mathcal{O}\left(d\right), \tag{D.5}$$

where the last step follows from the argument used to prove $\left\| \Sigma \right\|_{\text{op}} = \mathcal{O}\left(1\right)$ in Lemma C.1 in Bombari & Mondelli (2024).

For the last statement, we have

$$2d = \operatorname{tr}(\Sigma) = \operatorname{tr}\left(\mathbb{E}\left[zz^\top\right]\right) = \mathbb{E}\left[\operatorname{tr}\left(zz^\top\right)\right] = \mathbb{E}\left[\operatorname{tr}\left(z^\top z\right)\right] = \mathbb{E}\left[\left\| z \right\|_2^2\right], \tag{D.6}$$

where we used the cyclic property of the trace. Furthermore, we have

$$\left\| \left\| z \right\|_2 - \mathbb{E}\left[\left\| z \right\|_2\right] \right\|_{\psi_2} = \mathcal{O}\left(1\right), \tag{D.7}$$

since z is Lipschitz concentrated. Then,

$$0 \leq 2d - \mathbb{E}\left[\left\| z \right\|_2\right]^2 = \mathbb{E}\left[\left(\left\| z \right\|_2 - \mathbb{E}\left[\left\| z \right\|_2\right]\right)^2\right] \leq C_1, \tag{D.8}$$

for some absolute constant C_1. Thus, as $\sqrt{1-x} \geq 1 - x$ for $x \in [0, 1]$, we obtain

$$1 - \frac{C_1}{2d} \leq \sqrt{1 - \frac{C_1}{2d}} \leq \frac{\mathbb{E}\left[\left\| z \right\|_2\right]}{\sqrt{2d}} \leq 1. \tag{D.9}$$

Plugging this last result in (D.7) gives the desired claim. □

Lemma D.2. *We have that, denoting with $\tilde{\mu}^2 = \sum_{k\geq 2} \mu_k^2$, with μ_k denoting the k-th Hermite coefficient of ϕ,*

$$\left\| \mathbb{E}_V \left[\Phi\Phi^\top \right] - p \left(\mu_1^2 \frac{ZZ^\top}{2d} + \tilde{\mu}^2 I \right) \right\|_{\text{op}} = \mathcal{O}\left(\frac{p \log^3 d}{\sqrt{d}} \right), \tag{D.10}$$

with probability at least $1 - 2\exp\left(-c \log^2 d\right)$ over Z, where c is an absolute constant.

Proof. For all $i \in [n]$, we define the functions $\phi^{(i)} : \mathbb{R} \to \mathbb{R}$ as $\phi^{(i)}(\cdot) = \phi(\|z_i\|_2 \cdot / \sqrt{2d})$. Note that $\phi^{(i)}$ is odd, since ϕ is odd by Assumption 4. Thus, denoting with $\mu_k^{(i)}$ the k-th Hermite coefficient of $\phi^{(i)}$, for every $i \in [n]$, we have that $\mu_k^{(i)} = 0$ for all even k. This implies that, by denoting with v a random vector distributed as the rows of V, *i.e.*, $\sqrt{2d}\, v$ is a standard Gaussian vector, we have

$$\begin{aligned}
\left[\mathbb{E}_V \left[\Phi\Phi^\top\right]\right]_{ij} &= p\,\mathbb{E}_v \left[\phi(z_i^\top v)\phi(z_j^\top v)\right] \\
&= p\,\mathbb{E}_v \left[\phi^{(i)}\left(\frac{z_i^\top}{\|z_i\|_2} \sqrt{2d} v \right) \phi^{(j)}\left(\frac{z_j^\top}{\|z_j\|_2} \sqrt{2d} v \right) \right] \\
&= p \sum_{k=0}^{+\infty} \mu_k^{(i)} \mu_k^{(j)} \left(\frac{z_i^\top z_j}{\|z_i\|_2 \|z_j\|_2} \right)^k \\
&= p \mu_1^{(i)} \mu_1^{(j)} \frac{z_i^\top z_j}{\|z_i\|_2 \|z_j\|_2} + p \sum_{k\geq 3} \mu_k^{(i)} \mu_k^{(j)} \left(\frac{z_i^\top z_j}{\|z_i\|_2 \|z_j\|_2} \right)^k .
\end{aligned} \tag{D.11}$$

Then, denoting with $D_k \in \mathbb{R}^{n\times n}$ the diagonal matrix containing $\mu_k^{(i)} / \|z_i\|_2^k$ in its i-th entry, we can write

$$\mathbb{E}_V \left[\Phi\Phi^\top\right] = p D_1 Z Z^\top D_1 + p \sum_{k\geq 3} D_k \left(ZZ^\top\right)^{\circ k} D_k. \tag{D.12}$$

Notice that, due to the last statement in Lemma D.1, we have that, jointly for all $i \in [n]$,

$$\left| \frac{\|z_i\|_2}{\sqrt{2d}} - 1 \right| = \mathcal{O}\left(\frac{\log d}{\sqrt{d}} \right), \tag{D.13}$$

with probability at least $1 - 2\exp(-c_1 \log^2 d)$. Then, conditioning on such high probability event and denoting with ρ a standard Gaussian random variable, for all $i \in [n]$ we have

$$\begin{aligned}
\left| \mu_1^{(i)} - \mu_1 \right| &= \left| \mathbb{E}_\rho \left[\rho \phi^{(i)}(\rho) \right] - \mathbb{E}_\rho \left[\rho\phi(\rho) \right] \right| \\
&= \left| \mathbb{E}_\rho \left[\rho \left(\phi\left(\frac{\|z_i\|_2}{\sqrt{2d}} \rho \right) - \phi(\rho) \right) \right] \right| \\
&= \left| \mathbb{E}_\rho \left[\rho \left(\phi\left(\left(\frac{\|z_i\|_2}{\sqrt{2d}} - 1 \right) \rho + \rho \right) - \phi(\rho) \right) \right] \right| \\
&\leq \mathbb{E}_\rho \left[|\rho| \left| \phi\left(\left(\frac{\|z_i\|_2}{\sqrt{2d}} - 1 \right) \rho + \rho \right) - \phi(\rho) \right| \right] \\
&\leq L \mathbb{E}_\rho \left[|\rho| \left| \frac{\|z_i\|_2}{\sqrt{2d}} - 1 \right| |\rho| \right] \\
&= L \left| \frac{\|z_i\|_2}{\sqrt{2d}} - 1 \right| \mathbb{E}_\rho \left[\rho^2 \right] \\
&= \mathcal{O}\left(\frac{\log d}{\sqrt{d}} \right),
\end{aligned} \tag{D.14}$$

where we used Jensen's inequality in the fourth line, the L-Lipschitzness of ϕ in the fifth line, and (D.13) in the last step. With a similar approach, denoting with $\|\cdot\|_{L^2}$ the L^2 norm with respect to the Gaussian measure, we have that for all $i \in [n]$

$$
\begin{aligned}
\left|\left\|\phi^{(i)}\right\|_{L^2} - \|\phi\|_{L^2}\right| &\leq \left\|\phi^{(i)} - \phi\right\|_{L^2} \\
&= \mathbb{E}_\rho\left[\left(\phi^{(i)}(\rho) - \phi(\rho)\right)^2\right]^{1/2} \\
&= \mathbb{E}_\rho\left[\left(\phi\left(\left(\frac{\|z_i\|_2}{\sqrt{2d}} - 1\right)\rho + \rho\right) - \phi(\rho)\right)^2\right]^{1/2} \\
&\leq L\left|\frac{\|z_i\|_2}{\sqrt{2d}} - 1\right| \mathbb{E}_\rho\left[\rho^2\right]^{1/2} \\
&= \mathcal{O}\left(\frac{\log d}{\sqrt{d}}\right),
\end{aligned}
\tag{D.15}
$$

which directly implies that, for all $i \in [n]$, $\left\|\phi^{(i)}\right\|_{L^2} = \sum_{k\geq 0}\left(\mu_k^{(i)}\right)^2 = \Theta(1)$, and that

$$
\left|\sum_{k\geq 3}\left(\mu_k^{(i)}\right)^2 - \sum_{k\geq 3}\mu_k^2\right| \leq \left|\left\|\phi^{(i)}\right\|_{L^2}^2 - \|\phi\|_{L^2}^2\right| + \left|\left(\mu_1^{(i)}\right)^2 - \mu_1^2\right| = \mathcal{O}\left(\frac{\log d}{\sqrt{d}}\right). \tag{D.16}
$$

Thus, we are ready to estimate the operator norm of the off-diagonal part of the second term on the RHS of (D.12), specifically

$$
\begin{aligned}
&\left\|\sum_{k\geq 3} D_k \left(ZZ^\top\right)^{\circ k} D_k - \operatorname{diag}\left(\sum_{k\geq 3} D_k \left(ZZ^\top\right)^{\circ k} D_k\right)\right\|_{\text{op}} \\
&\quad\leq \sum_{k\geq 3}\left\|D_k \left(ZZ^\top\right)^{\circ k} D_k - \operatorname{diag}\left(D_k \left(ZZ^\top\right)^{\circ k} D_k\right)\right\|_F \\
&\quad\leq \sum_{k\geq 3}\max\nolimits_{i\neq j}\left(\frac{\left|z_i^\top z_j\right|}{\|z_i\|_2\|z_j\|_2}\right)^k\left(\sum_{i\in[n], j\in[n]}\left(\mu_k^{(i)}\mu_k^{(j)}\right)^2\right)^{1/2} \\
&\quad\leq \max\nolimits_{i\neq j}\left(\frac{\left|z_i^\top z_j\right|}{\|z_i\|_2\|z_j\|_2}\right)^3\sum_{i=0}^{n}\sum_{k\geq 3}\left(\mu_k^{(i)}\right)^2 \\
&\quad= \mathcal{O}\left(\frac{1}{d^{3/2}}\log^3 d\, n\right) = \mathcal{O}\left(\frac{\log^3 d}{\sqrt{d}}\right),
\end{aligned}
\tag{D.17}
$$

where in the first step we replaced the operator norm with the Frobenius norm, and used triangle inequality; in the fifth line we used that $\|z_i\|_2 = \Theta(\sqrt{d})$ for all $i \in [n]$ (true because of (D.13)), and that jointly for all $i \neq j$ we have $\left|z_i^\top z_j\right| / \|z_j\|_2 = \mathcal{O}(\log d)$ with probability at least $1 - 2\exp(-c_2 \log^2 d)$ since the z_i-s are independent sub-Gaussian vectors (since they are mean-0 and Lipschitz concentrated). The diagonal part of the second term on the RHS of (D.12) respects

$$
\left\|\operatorname{diag}\left(\sum_{k\geq 3} D_k \left(ZZ^\top\right)^{\circ k} D_k\right) - \tilde{\mu}^2 I\right\|_{\text{op}} = \max\nolimits_{i\in[n]}\left|\sum_{k\geq 3}\left(\mu_k^{(i)}\right)^2 - \sum_{k\geq 3}\mu_k^2\right| = \mathcal{O}\left(\frac{\log d}{\sqrt{d}}\right), \tag{D.18}
$$

because of (D.16). Lastly, notice that,

$$\begin{aligned}
\left\| D_1 Z Z^\top D_1 - \mu_1^2 \frac{ZZ^\top}{2d} \right\|_{\text{op}} &= \sup_{\|u\|_2=1} \left| u^\top D_1 Z Z^\top D_1 u - \mu_1^2 u^\top \frac{ZZ^\top}{2d} u \right| \\
&= \sup_{\|u\|_2=1} \left| \left\| Z^\top D_1 u \right\|_2^2 - \mu_1^2 \left\| \frac{Z^\top}{\sqrt{2d}} u \right\|_2^2 \right| \\
&\leq \sup_{\|u\|_2=1} \left(\left\| Z^\top D_1 u \right\|_2 + \mu_1 \left\| \frac{Z^\top}{\sqrt{2d}} u \right\|_2 \right) \sup_{\|u\|_2=1} \left(\left\| Z^\top D_1 u - \mu_1 \frac{Z^\top}{\sqrt{2d}} u \right\|_2 \right) \qquad \text{(D.19)} \\
&\leq \left(\left\| Z^\top D_1 \right\|_{\text{op}} + \mu_1 \left\| \frac{Z^\top}{\sqrt{2d}} \right\|_{\text{op}} \right) \left\| Z^\top D_1 - \mu_1 \frac{Z^\top}{\sqrt{2d}} \right\|_{\text{op}} \\
&\leq \left(\|Z\|_{\text{op}} \|D_1\|_{\text{op}} + \mu_1 \frac{\|Z\|_{\text{op}}}{\sqrt{2d}} \right) \|Z\|_{\text{op}} \left\| D_1 - \frac{\mu_1}{\sqrt{2d}} \right\|_{\text{op}} .
\end{aligned}$$

By Lemma D.1, we have that $\|Z\|_{\text{op}} = \mathcal{O}\left(\sqrt{d}\right)$ with probability at least $1 - 2\exp(-c_2 d)$, and since $\|z_i\|_2 = \Theta(\sqrt{d})$ and $\mu_1^{(i)} = \mathcal{O}(1)$ for all $i \in [n]$ (true because of (D.13) and (D.14) respectively), we have that $\|D_1\|_{\text{op}} = \mathcal{O}\left(1/\sqrt{d}\right)$. Furthermore, we have

$$\begin{aligned}
\left\| D_1 - \frac{\mu_1}{\sqrt{2d}} \right\|_{\text{op}} &= \max_i \left| \frac{\mu_1^{(i)}}{\|z_i\|_2} - \frac{\mu_1}{\sqrt{2d}} \right| \\
&\leq \max_i \frac{1}{\|z_i\|_2} \left(\left| \mu_1^{(i)} - \mu_1 \right| + \mu_1 \left| 1 - \frac{\|z_i\|_2}{\sqrt{2d}} \right| \right) \qquad \text{(D.20)} \\
&= \mathcal{O}\left(\frac{\log d}{d} \right),
\end{aligned}$$

where the last step is a consequence of (D.13) and (D.14). Then, we have that (D.19) reads

$$\left\| D_1 Z Z^\top D_1 - \mu_1^2 \frac{ZZ^\top}{2d} \right\|_{\text{op}} = \mathcal{O}\left(\frac{\log d}{\sqrt{d}} \right). \tag{D.21}$$

A standard application of the triangle inequality to (D.17), (D.18) and (D.21) gives

$$\left\| \left(D_1 Z Z^\top D_1 + \sum_{k \geq 3} D_k \left(ZZ^\top \right)^{\circ k} D_k \right) - \left(\mu_1^2 \frac{ZZ^\top}{2d} + \tilde{\mu}^2 I \right) \right\|_{\text{op}} = \mathcal{O}\left(\frac{\log^3 d}{\sqrt{d}} \right), \tag{D.22}$$

with probability at least $1 - 2\exp\left(-c_3 \log^2 d\right)$ over Z (where we used $\mu_2 = 0$ since ϕ is odd), which readily gives the thesis when plugged in (D.12). □

Lemma D.3. *We have that*

$$\|\Phi\|_{\text{op}} = \mathcal{O}\left(\sqrt{p}\right), \tag{D.23}$$

$$\left\| \Phi\Phi^\top - \mathbb{E}_V\left[\Phi\Phi^\top\right] \right\|_{\text{op}} = \mathcal{O}\left(\sqrt{pd}\right), \tag{D.24}$$

$$\lambda_{\min}\left(\Phi\Phi^\top\right) = \Omega(p), \tag{D.25}$$

with probability at least $1 - 2\exp\left(-c \log^2 d\right)$ over Z and V, where c is an absolute constant.

Proof. $\Phi^\top$ is a matrix with i.i.d. rows in the probability space of V. In particular, its i-th row takes the form

$$\left[\Phi^\top\right]_{i:} = \phi\left(ZV_{i:}\right) = \phi\left(ZV_{i:}\right) - \mathbb{E}_V\left[\phi\left(ZV_{i:}\right)\right], \tag{D.26}$$

where the last step holds since the (Gaussian) distribution of $V_{i:}$ is symmetric and ϕ is an odd function by Assumption 4. Then, since $\sqrt{2d}\,V_{i:}$ is a standard Gaussian (and hence Lipschitz concentrated) random vector, and ϕ is a Lipschitz continuous function, we have that

$$\left\| [\Phi^\top]_{i:} \right\|_{\psi_2} = \mathcal{O}\left(\frac{\|Z\|_{\text{op}}}{\sqrt{d}} \right) = \mathcal{O}\left(1\right), \tag{D.27}$$

where the $\|\cdot\|_{\psi_2}$ is meant on the probability space of V, and the second step holds with probability at least $1 - 2\exp\left(-c_1 d\right)$ over Z due to Lemma D.1. Conditioning on this high probability event, $\Phi^\top$ is a $p \times n$ matrix whose rows are i.i.d. mean-0 sub-Gaussian random vectors in $\mathbb{R}^n$. Then, by Lemma B.7 in Bombari et al. (2022), we have

$$\left\|\Phi^\top\right\|_{\text{op}} = \mathcal{O}\left(\sqrt{n} + \sqrt{p}\right) = \mathcal{O}\left(\sqrt{p}\right), \tag{D.28}$$

with probability at least $1 - 2\exp(-c_2 n)$ over V, where the second step holds because $n = o(p)$.

For the second part of the proof, we again follow the argument in Lemma B.7 in Bombari et al. (2022), which in turn exploits the discussion in Remark 5.40 in Vershynin (2012), and conclude that

$$\left\|\Phi\Phi^\top - \mathbb{E}_V\left[\Phi\Phi^\top\right]\right\|_{\text{op}} = \mathcal{O}\left(p\sqrt{\frac{n}{p}}\right) = \mathcal{O}\left(\sqrt{pn}\right) = \mathcal{O}\left(\sqrt{pd}\right), \tag{D.29}$$

with probability at least $1 - 2\exp(-c_3 n)$ over Z and V.

For the last statement, Lemma D.2 and Weyl's inequality imply that, with probability at least $1 - 2\exp\left(-c_2 \log^2 d\right)$ over Z we have

$$\begin{aligned}\lambda_{\min}\left(\Phi\Phi^\top\right) &\geq p\,\lambda_{\min}\left(\mu_1^2 \frac{ZZ^\top}{2d} + \tilde{\mu}^2 I\right) - \left\|\mathbb{E}_V\left[\Phi\Phi^\top\right] - p\left(\mu_1^2 \frac{ZZ^\top}{2d} + \tilde{\mu}^2 I\right)\right\|_{\text{op}} - \left\|\Phi\Phi^\top - \mathbb{E}_V\left[\Phi\Phi^\top\right]\right\|_{\text{op}} \\ &\geq p\tilde{\mu}^2 - \mathcal{O}\left(\frac{p\log^3 d}{\sqrt{d}}\right) - \mathcal{O}\left(\sqrt{pd}\right) = \Omega(p),\end{aligned} \tag{D.30}$$

where the last step is true since $\tilde{\mu} \neq 0$, as ϕ is non-linear by Assumption 4. □

Lemma D.4. *Let $\tilde{\phi} : \mathbb{R} \to \mathbb{R}$ be defined as $\tilde{\phi}(\cdot) := \phi(\cdot) - \mu_1(\cdot)$, and set*

$$n' = \min\left(\left\lfloor \frac{p}{\log^4 p} \right\rfloor, \left\lfloor \frac{d^{3/2}}{\log^3 d} \right\rfloor\right). \tag{D.31}$$

Let $\{\hat{z}_i\}_{i=1}^{n'}$ be n' i.i.d. random variables sampled from a distribution respecting Assumption 6, not necessarily with the same covariance as $\mathcal{P}_{XY}$, and independent from V. Then, if $\tilde{\Phi}_{n'} \in \mathbb{R}^{n' \times p}$ is defined as the matrix containing $\tilde{\phi}(V\hat{z}_i)$ in its i-th row, we have that

$$\left\|\tilde{\Phi}_{n'}\right\|_{\text{op}} = \mathcal{O}\left(\sqrt{p}\right), \tag{D.32}$$

with probability at least $1 - 2\exp\left(-c\log^2 d\right)$ over $\{\hat{z}_i\}_{i=1}^{n'}$ and V, where c is an absolute constant.

Proof. The proof follows the same strategy as Lemma C.8 in Bombari & Mondelli (2024), with the only difference that they work under their Assumption 1.2, *i.e.* that the data is normalized as $\|\hat{z}_i\|_2 = \sqrt{2d}$ (in our notation). This difference, however, does not affect the result. We can in fact condition on the high probability event that all $\hat{z}_i$ are such that $\|\hat{z}_i\|_2 = \Theta(\sqrt{d})$, which holds with probability at least $1 - 2\exp\left(-cd\right)$ by Lemma D.1, and proceed in the same way (as their Equation (C.78) now holds) until their Equation (C.81), which requires their Lemma C.7, *i.e.* that

$$\left\|\mathbb{E}_V\left[\tilde{\Phi}_{n'}\tilde{\Phi}_{n'}^\top\right]\right\|_{\text{op}} = \mathcal{O}\left(p\right). \tag{D.33}$$

This holds also in our case, as it can be proven following the argument in (D.17) and (D.18), where now the n in the last line of (D.17) has to be replaced with n', making the RHS there being $\mathcal{O}\left(1\right)$, as $n' = \mathcal{O}\left(\frac{d^{3/2}}{\log^3 d}\right)$ by definition. Lastly, the normalization of the data is used one more time in their Equation (C.91), but it is not critical to obtain the result, as $\|\hat{z}_i\|_2 = \Theta(\sqrt{d})$ is sufficient. We remark that Assumption 1.2 in Bombari & Mondelli (2024) also requires the covariance of the distribution to be well-conditioned, which however is not required for the purposes of the above mentioned lemmas. □

Lemma D.5. *Let $\tilde{\phi}:\mathbb{R}\to\mathbb{R}$ be defined as $\tilde{\phi}(\cdot):=\phi(\cdot)-\mu_1(\cdot)$, and let $z\in\mathbb{R}^{2d}$ be sampled from a distribution respecting Assumption 6, not necessarily with the same covariance as $\mathcal{P}_{XY}$, and independent from V. Then we have that*

$$\left\|\mathbb{E}_z\left[\tilde{\phi}(Vz)\tilde{\phi}(Vz)^\top\right]\right\|_{\text{op}}=\mathcal{O}\left(\log^4 d+\frac{p\log^3 d}{d^{3/2}}\right), \tag{D.34}$$

with probability at least $1-2p^2\exp\left(-c\log^2 d\right)$ over V, where c is an absolute constant.

Proof. The proof follows a similar path as the one in Lemma C.15 in Bombari & Mondelli (2024). In particular, set

$$n'=\min\left(\left\lfloor\frac{p}{\log^4 p}\right\rfloor,\left\lfloor\frac{d^{3/2}}{\log^3 d}\right\rfloor\right),\qquad N=p^2 n', \tag{D.35}$$

and let $\tilde{\Phi}_N\in\mathbb{R}^{N\times p}$ be a matrix containing $\tilde{\phi}(V\hat{z}_i)$ in its i-th row, where every $\{\hat{z}_i\}_{i=1}^N$ is sampled independently from the same distribution of z. Thus, $\tilde{\Phi}_N$ can be seen as the vertical stacking of p^2 matrices with size $n'\times p$. All these matrices respect the hypotheses of Lemma D.4, and hence have their operator norm bounded by $\mathcal{O}\left(\sqrt{p}\right)$ with probability at least $1-2\exp\left(-c_1\log^2 d\right)$. Thus, performing a union bound over these p^2 matrices, we get

$$\left\|\tilde{\Phi}_N^\top\tilde{\Phi}_N\right\|_{\text{op}}=\mathcal{O}\left(p^2 p\right)=\mathcal{O}\left(\frac{Np}{n'}\right)=\mathcal{O}\left(N\log^4 p+\frac{Np\log^3 d}{d^{3/2}}\right), \tag{D.36}$$

with probability at least $1-2p^2\exp\left(-c_1\log^2 d\right)$ over V and $\{\hat{z}_i\}_{i=1}^N$.

Via the same argument used for the last statement of Lemma D.1, denoting with $v_k\in\mathbb{R}^{2d}$ the k-th row of V, we have that $\|v_k\|_2=\mathcal{O}(1)$ uniformly for every k with probability at least $1-2p\exp(-c_2 d)$. Conditioning on such event, we have that each entry of $\tilde{\phi}(V\hat{z}_1)$ is sub-Gaussian (with uniformly bounded sub-Gaussian norm), since $\hat{z}_1$ is sub-Gaussian (as it is mean-0 and Lipschitz concentrated) and $\tilde{\phi}$ is a Lipschitz function. Thus, we have that each entry of $\mathbb{E}_{\hat{z}_1}\left[\tilde{\phi}(V\hat{z}_1)\right]$ is $\mathcal{O}(1)$ (see Proposition 2.5.2 in Vershynin (2018)), and therefore that $\left\|\mathbb{E}_{z_1}\left[\tilde{\phi}(V\hat{z}_1)\right]\right\|_{\psi_2}=\mathcal{O}\left(\left\|\mathbb{E}_{z_1}\left[\tilde{\phi}(V\hat{z}_1)\right]\right\|_2\right)=\mathcal{O}\left(\sqrt{p}\right)$. Then, conditioning on the high probability event $\|V\|_{\text{op}}=\mathcal{O}\left(\sqrt{p/d}\right)$ given by Lemma D.1, we have

$$\left\|\tilde{\phi}(V\hat{z}_1)\right\|_{\psi_2}\le\left\|\tilde{\phi}(V\hat{z}_1)-\mathbb{E}_{z_1}\left[\tilde{\phi}(V\hat{z}_1)\right]\right\|_{\psi_2}+\left\|\mathbb{E}_{z_1}\left[\tilde{\phi}(V\hat{z}_1)\right]\right\|_{\psi_2}=\mathcal{O}\left(\sqrt{\frac{p}{d}}+\sqrt{p}\right)=\mathcal{O}\left(\sqrt{p}\right), \tag{D.37}$$

where the second step holds because $\hat{z}_1$ is Lipschitz concentrated and $\tilde{\phi}$ is Lipschitz. Since the rows of $\tilde{\Phi}_N$ are identically distributed, this also holds jointly for all other $\hat{z}_i$-s, for $i\in[N]$. Then, $\tilde{\Phi}_N/\sqrt{p}$ is a matrix with independent sub-Gaussian rows, and by Theorem 5.39 in Vershynin (2012) (see their Remark 5.40 and Equation (5.25)), we have that

$$\frac{1}{p}\left\|\frac{\tilde{\Phi}_N^\top\tilde{\Phi}_N}{N}-\mathbb{E}_z\left[\tilde{\phi}(Vz)\tilde{\phi}(Vz)^\top\right]\right\|_{\text{op}}=\mathcal{O}\left(\sqrt{\frac{p}{N}}\right), \tag{D.38}$$

with probability at least $1-2\exp\left(-c_3 p\right)$ over $\{\hat{z}_i\}_{i=1}^N$. Then, we have

$$\begin{aligned}\left\|\mathbb{E}_z\left[\tilde{\phi}(Vz)\tilde{\phi}(Vz)^\top\right]\right\|_{\text{op}}&\le\left\|\frac{\tilde{\Phi}_N^\top\tilde{\Phi}_N}{N}-\mathbb{E}_z\left[\tilde{\phi}(Vz)\tilde{\phi}(Vz)^\top\right]\right\|_{\text{op}}+\frac{\left\|\tilde{\Phi}_N^\top\tilde{\Phi}_N\right\|_{\text{op}}}{N}\\&=\mathcal{O}\left(p\sqrt{\frac{p}{N}}\right)+\mathcal{O}\left(\log^4 p+\frac{p\log^3 d}{d^{3/2}}\right)\\&=\mathcal{O}\left(\sqrt{p}\sqrt{\frac{\log^4 p}{p}}+\sqrt{p}\sqrt{\frac{\log^3 d}{d^{3/2}}}\right)+\mathcal{O}\left(\log^4 p+\frac{p\log^3 d}{d^{3/2}}\right)\\&=\mathcal{O}\left(\log^4 p+\frac{p\log^3 d}{d^{3/2}}\right),\end{aligned} \tag{D.39}$$

where the first step follows from the triangle inequality, the second step is a consequence of (D.38) and (D.36), and the third step follows from the definition of N.

Taking the intersection between the high probability events in (D.36), (D.37) and (D.38), the previous equation then holds with probability at least $1 - 2p^2 \exp\left(-c_4 \log^2 d\right)$ over V and $\{\hat{z}_i\}_{i=1}^N$. Also note that its LHS does not depend on $\{\hat{z}_i\}_{i=1}^N$, which were introduced as auxiliary random variables. Thus, the high probability bound holds restricted to the probability space of V, and the desired result follows. □

Lemma D.6. *Let $\tilde{\phi} : \mathbb{R} \to \mathbb{R}$ be defined as $\tilde{\phi}(\cdot) := \phi(\cdot) - \mu_1(\cdot)$ and $\tilde{\Phi} \in \mathbb{R}^{n \times p}$ as the matrix containing $\tilde{\phi}(Vz_i)$ in its i-th row. Then, we have*

$$\left\|\tilde{\Phi} V\right\|_{\text{op}} = \mathcal{O}\left(\sqrt{p}\log d + \frac{p \log d}{d}\right), \tag{D.40}$$

with probability at least $1 - 2\exp\left(-c\log^2 d\right)$ over Z and V, where c is an absolute constant.

Proof. Note that $\tilde{\phi}$ is Lipschitz (since ϕ is Lipschitz by Assumption 4). During all the proof, we condition on the event $\|Z\|_{\text{op}} = \mathcal{O}\left(\sqrt{d}\right)$ and $\|z_i\|_2 = \Theta\left(\sqrt{d}\right)$ for all $i \in [n]$, which holds with probability at least $1 - 2\exp\left(-c_1 d\right)$ by Lemma D.1. During the proof we also use the shorthand $v \in \mathbb{R}^{2d}$ to denote a random vector such that $\sqrt{2d}\, v$ is a standard Gaussian vector, *i.e.*, it has the same distribution as the rows of V. This implies

$$\mathbb{E}_v\left[\left\|\tilde{\phi}\left(Zv\right)\right\|_2\right] = \mathcal{O}\left(\sqrt{n}\right), \qquad \left\|\left\|\tilde{\phi}\left(Zv\right)\right\|_2 - \mathbb{E}_v\left[\left\|\tilde{\phi}\left(Zv\right)\right\|_2\right]\right\|_{\psi_2} = \mathcal{O}\left(1\right), \tag{D.41}$$

and

$$\mathbb{E}_v\left[\|v\|_2\right] = \mathcal{O}\left(1\right), \qquad \left\|\|v\|_2 - \mathbb{E}_v\left[\|v\|_2\right]\right\|_{\psi_2} = \mathcal{O}\left(\frac{1}{\sqrt{d}}\right), \tag{D.42}$$

where both sub-Gaussian norms are meant on the probability space of v, and where the very first equation follows from the discussion in Lemma C.3 in Bombari et al. (2022). Then, there exists an absolute constant C_1 such that we jointly have

$$\left\|\tilde{\phi}\left(Zv\right)\right\|_2 \leq C_1\sqrt{d}, \qquad \|v\|_2 \leq C_1 \tag{D.43}$$

with probability at least $1 - 2\exp\left(-c_2 d\right)$ over v.

Let E_k be the indicator defined on the high probability event above with respect to the random variable $v_k := V_{:k}$ (the k-th row of V), *i.e.*

$$E_k := \mathbf{1}\left(\|v_k\|_2 \leq C_1 \text{ and } \left\|\tilde{\phi}\left(Zv_k\right)\right\|_2 \leq C_1\sqrt{d}\right), \tag{D.44}$$

and we define $E \in \mathbb{R}^{p \times p}$ as the diagonal matrix containing E_k in its k-th entry. Notice that we have $\|I - E\|_{\text{op}} = 0$ with probability at least $1 - 2p\exp\left(-c_2 d\right)$, and $\mathbb{E}_V\left[\|I - E\|_{\text{op}}\right] \leq 2p\exp\left(-c_2 d\right)$.

Thus, we have

$$\begin{aligned}
\left\|\mathbb{E}_V\left[\tilde{\Phi}\left(I - E\right)V\right]\right\|_{\text{op}} &\leq \mathbb{E}_V\left[\left\|\tilde{\Phi}\right\|_{\text{op}}\|1 - E\|_{\text{op}}\|V\|_{\text{op}}\right] \\
&\leq \mathbb{E}_V\left[\left\|\tilde{\Phi}\right\|_{\text{op}}^2\|V\|_{\text{op}}^2\right]^{1/2}\mathbb{E}_V\left[\|I - E\|_{\text{op}}^2\right]^{1/2} \\
&\leq \mathbb{E}_V\left[\left\|\tilde{\Phi}\right\|_{\text{op}}^4\right]^{1/4}\mathbb{E}_V\left[\|V\|_{\text{op}}^4\right]^{1/4}\left(2p\exp\left(-c_2 d\right)\right)^{1/2} \\
&\leq \mathbb{E}_V\left[\left\|\tilde{\Phi}\right\|_F^4\right]^{1/4}\mathbb{E}_V\left[\|V\|_F^4\right]^{1/4}\left(2p\exp\left(-c_2 d\right)\right)^{1/2} \\
&= o(1),
\end{aligned} \tag{D.45}$$

where the last step holds because of our initial conditioning on Z: the first two terms are the sum of finite powers of sub-Gaussian random variables (the entries of $\tilde{\Phi}$ and V), and thus (see Proposition 2.5.2 in Vershynin (2018)) the first two factors in the third line of the previous equation will be $\mathcal{O}(p^\alpha)$ for some finite α, which gives the last line due to Assumption 5.

As in Lemma D.2, we introduce the notation (for all $i \in [n]$) $\tilde{\phi}^{(i)} : \mathbb{R} \to \mathbb{R}$ such that $\tilde{\phi}^{(i)}(\cdot) = \tilde{\phi}(\|z_i\|_2 \cdot /\sqrt{2d})$. Thus, denoting with $v \in \mathbb{R}^{2d}$ a random vector such that $\sqrt{2d}v$ is standard Gaussian (*i.e.*, distributed as the rows of V), we can write

$$\left[\mathbb{E}_V\left[\tilde{\Phi}V\right]\right]_{ij} = p\left[\mathbb{E}_v\left[\tilde{\phi}(Zv)v^\top\right]\right]_{ij} = \frac{p}{\sqrt{2d}}\mathbb{E}_v\left[\tilde{\phi}^{(i)}\left(\frac{z_i^\top}{\|z_i\|_2}\sqrt{2d}v\right)\left(e_j^\top\left(\sqrt{2d}v\right)\right)\right] = \frac{p}{\sqrt{2d}}\frac{\tilde{\mu}_1^{(i)} z_i^\top e_j}{\|z_i\|_2}, \quad \text{(D.46)}$$

where $\tilde{\mu}_1^{(i)}$ is the first Hermite coefficient of $\tilde{\phi}^{(i)}$. Then, denoting with $\tilde{D} \in \mathbb{R}^{n\times n}$ the diagonal matrix containing $\tilde{\mu}_1^{(i)}/\|z_i\|_2$ in its i-th entry, we can write

$$\left\|\mathbb{E}_V\left[\tilde{\Phi}V\right]\right\|_{\text{op}} = \frac{p}{\sqrt{2d}}\left\|\tilde{D}Z\right\|_{\text{op}} \le \frac{p}{\sqrt{2d}}\left\|\tilde{D}\right\|_{\text{op}}\|Z\|_{\text{op}}. \quad \text{(D.47)}$$

Then, since we conditioned on $\|z_i\|_2 = \Theta\left(\sqrt{d}\right)$ for all $i \in [n]$, following the same argument as in (D.14), and since the first Hermite coefficient of $\tilde{\phi}$ is 0 by definition, we have

$$\left\|\mathbb{E}_V\left[\tilde{\Phi}V\right]\right\|_{\text{op}} = \mathcal{O}\left(\frac{p}{\sqrt{d}}\frac{\log d}{d}\sqrt{d}\right) = \mathcal{O}\left(\frac{p\log d}{d}\right), \quad \text{(D.48)}$$

with probability at least $1 - 2\exp\left(-c_3\log^2 d\right)$ over Z. A standard application of the triangle inequality to this last equation and (D.45) then gives

$$\left\|\mathbb{E}_V\left[\tilde{\Phi}EV\right]\right\|_{\text{op}} \le \left\|\mathbb{E}_V\left[\tilde{\Phi}V\right]\right\|_{\text{op}} + \left\|\mathbb{E}_V\left[\tilde{\Phi}\left(I-E\right)V\right]\right\|_{\text{op}} = \mathcal{O}\left(\frac{p\log d}{d}\right), \quad \text{(D.49)}$$

with probability at least $1 - 2\exp\left(-c_4\log^2 d\right)$ over Z.

Let's now look at

$$\tilde{\Phi}EV - \mathbb{E}_V\left[\tilde{\Phi}EV\right] = \sum_{k=1}^{p}\tilde{\phi}(Zv_k)E_k v_k^\top - \mathbb{E}_{v_k}\left[\tilde{\phi}(Zv_k)E_k v_k^\top\right] =: \sum_{k=1}^{p} W_k, \quad \text{(D.50)}$$

where we defined the shorthand $W_k = \tilde{\phi}(Zv_k)E_k v_k^\top - \mathbb{E}_{v_k}\left[\tilde{\phi}(Zv_k)E_k v_k^\top\right]$. (D.50) is the sum of p i.i.d. mean-0 random matrices W_k (in the probability space of V), such that

$$\begin{aligned}\sup_{v_k}\left\|\tilde{\phi}(Zv_k)E_k v_k^\top - \mathbb{E}_{v_k}\left[\tilde{\phi}(Zv_k)E_k v_k^\top\right]\right\|_{\text{op}} &\le 2\sup_{v_k}\left\|\tilde{\phi}(Zv_k)E_k v_k^\top\right\|_{\text{op}} \\ &= 2\sup_{v_k}\left\|\tilde{\phi}(Zv_k)\right\|_2\|v_k\|_2 E_k \\ &\le 2C_1^2\sqrt{d},\end{aligned} \quad \text{(D.51)}$$

because of (D.44). Then, by matrix Bernstein's inequality for rectangular matrices (see Exercise 5.4.15 in Vershynin (2018)), we have that

$$\mathbb{P}_V\left(\left\|\tilde{\Phi}EV - \mathbb{E}_V\left[\tilde{\Phi}EV\right]\right\|_{\text{op}} \ge t\right) \le (n+d)\exp\left(-\frac{t^2/2}{\sigma^2 + 2C_1^2\sqrt{d}t/3}\right), \quad \text{(D.52)}$$

where σ^2 is defined as

$$\sigma^2 = p\max\left(\left\|\mathbb{E}_{v_k}\left[W_k W_k^\top\right]\right\|_{\text{op}}, \left\|\mathbb{E}_{v_k}\left[W_k^\top W_k\right]\right\|_{\text{op}}\right). \quad \text{(D.53)}$$

For every matrix A, we have $\mathbb{E}\left[\left(A-\mathbb{E}[A]\right)\left(A-\mathbb{E}[A]\right)^\top\right] = \mathbb{E}\left[AA^\top\right] - \mathbb{E}[A]\mathbb{E}[A]^\top \preceq \mathbb{E}\left[AA^\top\right]$. Thus,

$$\begin{aligned}\left\|\mathbb{E}_{v_k}\left[W_k W_k^\top\right]\right\|_{\text{op}} &\le \left\|\mathbb{E}_{v_k}\left[\tilde{\phi}(Zv_k)E_k v_k^\top v_k E_k \tilde{\phi}(Zv_k)^\top\right]\right\|_{\text{op}} \\ &\le \left\|\mathbb{E}_{v_k}\left[\tilde{\phi}(Zv_k)\tilde{\phi}(Zv_k)^\top\right]\right\|_{\text{op}}\sup_{v_k}\left(E_k\|v_k\|_2^2\right) \\ &\le C_1^2\left\|\mathbb{E}_{v_k}\left[\tilde{\phi}(Zv_k)\tilde{\phi}(Zv_k)^\top\right]\right\|_{\text{op}} \\ &= \mathcal{O}(1),\end{aligned} \quad \text{(D.54)}$$

where the last step is a direct consequence of Lemma D.2, applied to $\tilde{\Phi}$ instead of to Φ, and holds with probability at least $1-2\exp\left(-c_5\log^2 d\right)$ over Z. For the other argument in the $\max$ in (D.53) we similarly have

$$
\begin{aligned}
\left\|\mathbb{E}\left[W_k^\top W_k\right]\right\|_{\text{op}} &\leq \left\|\mathbb{E}_{v_k}\left[v_k E_k \tilde{\phi}(Zv_k)^\top \tilde{\phi}(Zv_k) E_k v_k^\top\right]\right\|_{\text{op}} \\
&\leq \left\|\mathbb{E}_{v_k}\left[v_k v_k^\top\right]\right\|_{\text{op}} \sup_{v_k}\left(E_k\left\|\tilde{\phi}(Zv_k)\right\|_2^2\right) \\
&\leq \frac{1}{d} C_1^2 d \\
&= \mathcal{O}(1).
\end{aligned}
\tag{D.55}
$$

Then, plugging these last two equations in (D.52) we get

$$
\mathbb{P}_V\left(\left\|\tilde{\Phi}EV-\mathbb{E}_V\left[\tilde{\Phi}EV\right]\right\|_{\text{op}} \geq \sqrt{p}\log d\right) \leq (n+d)\exp\left(-\frac{p\log^2 d/2}{C_2 p + 2C_1^2\sqrt{d}\sqrt{p}\log d/3}\right) \leq 2\exp\left(-c_6\log^2 d\right),
\tag{D.56}
$$

where we used Assumption 5. Then, applying a triangle inequality and using (D.49) and (D.56), we get

$$
\left\|\tilde{\Phi}EV\right\|_{\text{op}} = \mathcal{O}\left(\sqrt{p}\log d + \frac{p\log d}{d}\right),
\tag{D.57}
$$

with probability at least $1-2\exp\left(-c_7\log^2 d\right)$ over Z, V. Then, since $E=I$ with probability at least $1-2p\exp\left(-c_3 d\right)$, using Assumption 5 we get the desired result.

□

Lemma D.7. *Let $z \in \mathbb{R}^{2d}$ be sampled from a distribution respecting Assumption 6, not necessarily with the same covariance as $\mathcal{P}_{XY}$, independent from everything else, and let $f_{\text{RF}}(z, \hat{\theta}_{\text{RF}}(\lambda))$ be the RF model defined in (5.1).Then, we have that*

$$
\left|f_{\text{RF}}(z, \hat{\theta}_{\text{RF}}(\lambda)) - \mu_1^2 p \frac{z^\top Z^\top}{2d}\left(\Phi\Phi^\top + n\lambda I\right)^{-1} G\right| = \mathcal{O}\left(\frac{d^{1/4}\log d}{p^{1/4}} + \frac{\log^{3/2} d}{d^{1/8}}\right) = o(1),
\tag{D.58}
$$

with probability $1 - C\sqrt{d}\log^2 d/\sqrt{p} - C\log^3 d/d^{1/4}$ over Z, G, V and z, where C is an absolute constant

Proof. Let $\tilde{\phi}:\mathbb{R}\to\mathbb{R}$ be defined as $\tilde{\phi}(\cdot) := \phi(\cdot) - \mu_1(\cdot)$, and let $\tilde{\Phi} \in \mathbb{R}^{n\times p}$ be defined as the matrix containing $\tilde{\phi}(Vz_i)$ in its i-th row. Then, introducing the shorthand $\hat{G} = \left(\Phi\Phi^\top + n\lambda I\right)^{-1} G$, we can write

$$
\begin{aligned}
f_{\text{RF}}(z, \hat{\theta}_{\text{RF}}(\lambda)) &= \left(\mu_1 Vz + \tilde{\phi}(Vz)\right)^\top \left(\mu_1 VZ^\top + \tilde{\Phi}^\top\right)\hat{G} \\
&= \mu_1^2 p \frac{z^\top Z^\top}{2d}\hat{G} + \mu_1^2 z^\top\left(V^\top V - \frac{p}{2d} I\right) Z^\top \hat{G} + \mu_1 z^\top V^\top \tilde{\Phi}^\top \hat{G} + \tilde{\phi}(Vz)^\top \Phi\hat{G}.
\end{aligned}
\tag{D.59}
$$

Notice that since every entry of G is sub-Gaussian and independent by Assumption 6, Theorem 3.1.1 in Vershynin (2018) readily gives $\|G\|_2 = \mathcal{O}(\sqrt{n}) = \mathcal{O}\left(\sqrt{d}\right)$ with probability at least $1-2\exp(-c_1 d)$ over G. Then, conditioning on the high probability event described by Lemma D.3, we get

$$
\left\|\hat{G}\right\|_2 \leq \left(\lambda_{\min}\left(\Phi\Phi^\top + n\lambda I\right)\right)^{-1}\|G\|_2 \leq \left(\lambda_{\min}\left(\Phi\Phi^\top\right)\right)^{-1}\|G\|_2 = \mathcal{O}\left(\frac{\sqrt{d}}{p}\right),
\tag{D.60}
$$

with probability at least $1-2\exp\left(-c_2\log^2 d\right)$ over V, Z and G. We will condition on this high probability event until the end of the proof. Let's then investigate the last 3 terms on the RHS of (D.59) separately:

(i) A direct application of Theorem 5.39 of Vershynin (2012) (see their Equation 5.23) gives

$$
\left\|\frac{2d}{p}V^\top V - I\right\|_{\text{op}} = \mathcal{O}\left(\sqrt{\frac{d}{p}}\right),
\tag{D.61}
$$

with probability at least $1 - 2\exp(-c_3 d)$ over V. Then, since z is sub-Gaussian and independent from everything else, with probability $1 - 2\exp\left(-c_4 \log^2 d\right)$ over itself we have

$$\begin{aligned} \left|\mu_1^2 z^\top \left(V^\top V - \frac{p}{2d} I\right) Z^\top \hat{G}\right| &\le \log d \left\|\left(V^\top V - \frac{p}{2d} I\right) Z^\top \hat{G}\right\|_2 \\ &\le \log d \left\|V^\top V - \frac{p}{2d} I\right\|_{\text{op}} \|Z\|_{\text{op}} \left\|\hat{G}\right\|_2 \\ &= \mathcal{O}\left(\log d \sqrt{\frac{p}{d}}\sqrt{d}\,\frac{\sqrt{d}}{p}\right) = \mathcal{O}\left(\frac{\sqrt{d}\log d}{\sqrt{p}}\right), \end{aligned} \tag{D.62}$$

where the third step holds with probability at least $1 - 2\exp(-c_5 d)$ over Z due to Lemma D.1.

(ii) As before, since z is sub-Gaussian and independent from everything else, with probability $1 - 2\exp\left(-c_4 \log^2 d\right)$ we have

$$\begin{aligned} \left|\mu_1 z^\top V^\top \tilde{\Phi}^\top \hat{G}\right| &\le \log d \left\|V^\top \tilde{\Phi}^\top\right\|_{\text{op}} \left\|\hat{G}\right\|_2 \\ &= \mathcal{O}\left(\log d \left(\sqrt{p}\log d + \frac{p\log d}{d}\right)\frac{\sqrt{d}}{p}\right) = \mathcal{O}\left(\log^2 d \left(\sqrt{\frac{d}{p}} + \frac{1}{\sqrt{d}}\right)\right), \end{aligned} \tag{D.63}$$

where the second step holds because of Lemma D.6, and holds with probability at least $1 - 2\exp\left(-c_6 \log^2 d\right)$ over Z and V.

(iii) For the last term of the RHS of (D.59), its second moment in the probability space of z reads

$$\begin{aligned} \mathbb{E}_z\left[\hat{G}^\top \Phi^\top \tilde{\phi}(Vz)\,\tilde{\phi}(Vz)^\top \Phi \hat{G}\right] &\le \left\|\mathbb{E}_z\left[\tilde{\phi}(Vz)\,\tilde{\phi}(Vz)^\top\right]\right\|_{\text{op}} \|\Phi\|_{\text{op}}^2 \left\|\hat{G}\right\|_2^2 \\ &= \mathcal{O}\left(\left(\log^4 d + \frac{p\log^3 d}{d^{3/2}}\right)\sqrt{d}\frac{\sqrt{d}}{p}\right) \\ &= \mathcal{O}\left(\frac{d\log^4 d}{p} + \frac{\log^3 d}{\sqrt{d}}\right), \end{aligned} \tag{D.64}$$

where the second step follows from Lemmas D.5 and D.3, and holds with probability at least $1 - 2\exp\left(-c_7 \log^2 d\right)$ over Z and V. Then, by Markov inequality, we have that there exists a constant C_1 such that

$$\left(\tilde{\phi}(Vz)^\top \Phi \hat{G}\right)^2 < C_1 \left(\frac{d\log^4 d}{p} + \frac{\log^3 d}{\sqrt{d}}\right) t, \tag{D.65}$$

with probability at least $1 - 1/t$ over z. Setting

$$t = \min\left(\frac{\sqrt{p}}{\sqrt{d}\log^2 d}, d^{1/4}\right) = \omega(1), \tag{D.66}$$

since $p = \omega\left(d\log^4 d\right)$ by Assumption 5, we have

$$\left|\tilde{\phi}(Vz)^\top \Phi \hat{G}\right| = \mathcal{O}\left(\frac{d^{1/4}\log d}{p^{1/4}} + \frac{\log^{3/2} d}{d^{1/8}}\right), \tag{D.67}$$

with probability at least $1 - \sqrt{d}\log^2 d/\sqrt{p} - \log^3 d/d^{1/4}$.

Then, plugging (i), (ii) and (iii) in (D.59) gives

$$\left|f_{\text{RF}}(z, \hat{\theta}_{\text{RF}}(\lambda)) - \mu_1^2 p \frac{z^\top Z^\top}{2d}\hat{G}\right| = \mathcal{O}\left(\frac{d^{1/4}\log d}{p^{1/4}} + \frac{\log^{3/2} d}{d^{1/8}}\right), \tag{D.68}$$

with probability at least $1 - c_8 \frac{\sqrt{d}\log^2 d}{\sqrt{p}} - c_8 \frac{\log^3 d}{d^{1/4}}$, which gives the desired result. □

Proof of Theorem 2 Let $E \in \mathbb{R}^{n\times n}$ be the matrix defined as

$$E = \Phi\Phi^\top - p\left(\mu_1^2 \frac{ZZ^\top}{2d} + \tilde{\mu}^2 I\right). \tag{D.69}$$

Note that

$$\|E\|_{\text{op}} \le \left\|\Phi\Phi^\top - \mathbb{E}_V\left[\Phi\Phi^\top\right]\right\|_{\text{op}} + \left\|\mathbb{E}_V\left[\Phi\Phi^\top\right] - p\left(\mu_1^2 \frac{ZZ^\top}{2d} + \tilde{\mu}^2 I\right)\right\|_{\text{op}} = \mathcal{O}\left(p\left(\sqrt{\frac{d}{p}} + \frac{\log^3 d}{\sqrt{d}}\right)\right), \tag{D.70}$$

with probability at least $1 - 2\exp\left(-c_1 \log^2 d\right)$ over Z, V due to Lemmas D.2 and D.3. By the Woodbury matrix identity (or Hua's identity), we have

$$\begin{aligned}
\left(\Phi\Phi^\top + n\lambda I\right)^{-1} &= \left(p\left(\mu_1^2 \frac{ZZ^\top}{2d} + \tilde{\mu}^2 I\right) + E + n\lambda I\right)^{-1} \\
&= \left(\mu_1^2 p \frac{ZZ^\top}{2d} + \left(\tilde{\mu}^2 p + n\lambda\right) I + E\right)^{-1} \\
&= \left(\mu_1^2 p \frac{ZZ^\top}{2d} + \left(\tilde{\mu}^2 p + n\lambda\right) I\right)^{-1} \\
&\quad - \left(\mu_1^2 p \frac{ZZ^\top}{2d} + \left(\tilde{\mu}^2 p + n\lambda\right) I\right)^{-1} E \left(\mu_1^2 p \frac{ZZ^\top}{2d} + \left(\tilde{\mu}^2 p + n\lambda\right) I + E\right)^{-1},
\end{aligned} \tag{D.71}$$

which gives

$$\begin{aligned}
&\left|\mu_1^2 p \frac{z^\top Z^\top}{2d}\left(\Phi\Phi^\top + n\lambda I\right)^{-1} G - \mu_1^2 p \frac{z^\top Z^\top}{2d}\left(\mu_1^2 p \frac{ZZ^\top}{2d} + \left(\tilde{\mu}^2 p + n\lambda\right) I\right)^{-1} G\right| \\
&\le \left|\mu_1^2 p \frac{z^\top Z^\top}{2d}\left(\mu_1^2 p \frac{ZZ^\top}{2d} + \left(\tilde{\mu}^2 p + n\lambda\right) I\right)^{-1} E \left(\mu_1^2 p \frac{ZZ^\top}{2d} + \left(\tilde{\mu}^2 p + n\lambda\right) I + E\right)^{-1} G\right| \\
&\le \log d \left\|\frac{p}{2d} Z^\top \left(\mu_1^2 p \frac{ZZ^\top}{2d} + \left(\tilde{\mu}^2 p + n\lambda\right) I\right)^{-1} E \left(\mu_1^2 p \frac{ZZ^\top}{2d} + \left(\tilde{\mu}^2 p + n\lambda\right) I + E\right)^{-1} G\right\|_2 \\
&\le \log d \, \frac{p}{2d} \|Z\|_{\text{op}} \frac{1}{\tilde{\mu}^2 p} \|E\|_{\text{op}} \frac{1}{\lambda_{\min}\left(\Phi\Phi^\top\right)} \|G\|_2 \\
&= \mathcal{O}\left(\log d \, \frac{p}{d} \sqrt{d} \, \frac{1}{p} \, p\left(\sqrt{\frac{d}{p}} + \frac{\log^3 d}{\sqrt{d}}\right) \frac{1}{p} \sqrt{d}\right) \\
&= \mathcal{O}\left(\log d \sqrt{\frac{d}{p}} + \frac{\log^4 d}{\sqrt{d}}\right).
\end{aligned} \tag{D.72}$$

Here, the second step holds with probability at least $1 - 2\exp\left(-c_2 \log^2 d\right)$ since z is sub-Gaussian and independent from everything else; the fourth step is a consequence of Lemma D.1, (D.70), Lemma D.3, and $\|G\|_2 = \mathcal{O}\left(\sqrt{d}\right)$ (see the argument prior to (D.60)), and as a whole holds with probability $1 - 2\exp\left(-c_3 \log^2 d\right)$ over Z, G, and V.

Note that the second term in the LHS of (D.72) can be written as

$$\begin{aligned}
\mu_1^2 p \frac{z^\top Z^\top}{2d}\left(\mu_1^2 p \frac{ZZ^\top}{2d} + \left(\tilde{\mu}^2 p + n\lambda\right) I\right)^{-1} G &= z^\top Z^\top \left(ZZ^\top + n\left(\frac{2\tilde{\mu}^2 d}{\mu_1^2 n} + \frac{2d}{\mu_1^2 p}\lambda\right) I\right)^{-1} G \\
&= z^\top \left(Z^\top Z + n\left(\frac{2\tilde{\mu}^2 d}{\mu_1^2 n} + \frac{2d}{\mu_1^2 p}\lambda\right) I\right)^{-1} Z^\top G \\
&= f_{\text{LR}}(z, \hat{\theta}_{\text{LR}}(\tilde{\lambda})),
\end{aligned} \tag{D.73}$$

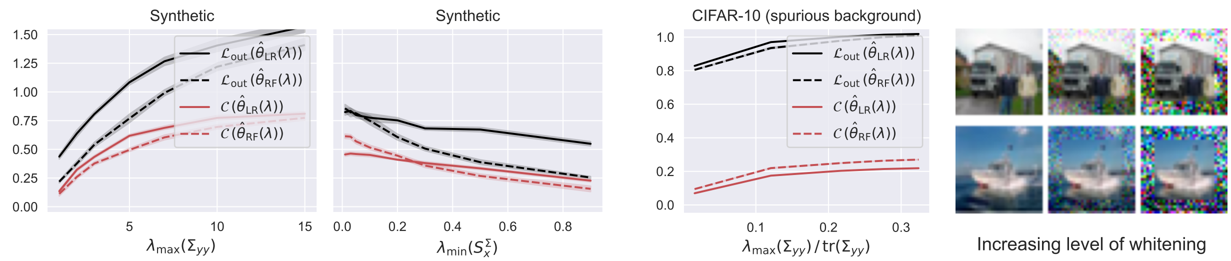

Figure 4: Out-of-distribution test loss $\mathcal{L}(\hat\theta_{\rm LR/RF}(\lambda))$ (black) and spurious correlations $\mathcal{C}(\hat\theta_{\rm LR/RF}(\lambda)$ (red) as a function of $\lambda_{\max}\left(\Sigma_{yy}\right)$ (first panel) and $\lambda_{\min}\left(S_x^\Sigma\right)$ (second panel) on a Gaussian synthetic dataset, and for the CIFAR-10 experiment (third panel). We consider the same set-up as Figures 5 and 3, for Gaussian and CIFAR-10 data, respectively.

where the second line is due to the classical identity $A^\top\left(AA^\top+\kappa I\right)^{-1}=\left(A^\top A+\kappa I\right)^{-1}A^\top$, and the third line uses the definition in (3.1), with $\hat\theta_{\rm LR}(\tilde\lambda)$ defined in (3.2) and

$$\tilde\lambda=\frac{2\tilde\mu^2 d}{\mu_1^2 n}+\frac{2d}{\mu_1^2 p}\lambda. \tag{D.74}$$

Furthermore, the first term of the LHS of (D.72) satisfies

$$\left|\mu_1^2 p\frac{z^\top Z^\top}{2d}\left(\Phi\Phi^\top+n\lambda I\right)^{-1}G-f_{\rm RF}(z,\hat\theta_{\rm RF}(\lambda))\right|=\mathcal{O}\left(\frac{d^{1/4}\log d}{p^{1/4}}+\frac{\log^{3/2}d}{d^{1/8}}\right), \tag{D.75}$$

with probability $1-C_1\sqrt{d}\log^2 d/\sqrt{p}-C_1\log^3 d/d^{1/4}$ over Z,G,V and z, due to Lemma D.7. Thus, an application of the triangle inequality together with (D.72), (D.73) and (D.75) gives the desired result. □

E Connection with out-of-distribution loss

Our definition of $\mathcal{C}(\hat\theta)$ in (2.3) formalizes to the regression setting the definition in the survey Ye et al. (2024) and the fairness metric in Zliobaite (2015) (when interpreting y as the protected variable). Furthermore, in the context of classification, it can connected to the worst group accuracy Sagawa et al. (2020a;b). In fact, our definition (2.3) is also related to the *out-of-distribution* test loss. To show this, let $\tilde x$ and $[x^\top,y^\top]^\top$ be sampled independently from $\mathcal{P}_X$ and $\mathcal{P}_{XY}$ respectively. For simplicity, assume that $\mathbb{E}\left[f(\hat\theta,[\tilde x^\top,y])^2\right]=\mathbb{E}\left[f^*_x(\tilde x)^2\right]=1$ and $\mathbb{E}\left[f(\hat\theta,[\tilde x^\top,y])\right]=\mathbb{E}\left[f^*_x(\tilde x)\right]=0$. Thus, for the quadratic loss, we readily get

$$\mathbb{E}_{\tilde x,y}\left[\left(f(\hat\theta,[\tilde x^\top,y])-f^*_x(\tilde x)\right)^2\right]=2-2\mathbb{E}_{\tilde x,y}\left[f(\hat\theta,[\tilde x^\top,y])f^*_x(\tilde x)\right]=2-2\mathrm{Cov}\left(f(\hat\theta,[\tilde x^\top,y]),f^*_x(\tilde x)\right). \tag{E.1}$$

Denoting with S the covariance matrix of the three random variables $f(\hat\theta,[\tilde x^\top,y])$, $f^*_x(\tilde x)$, and $f^*_x(x)$, we have

$$S=\begin{pmatrix}1 & \rho & \mathcal{C}\\ \rho & 1 & 0\\ \mathcal{C} & 0 & 1\end{pmatrix}, \tag{E.2}$$

where we introduced the shorthands $\mathcal{C}=\mathrm{Cov}\left(f(\hat\theta,[\tilde x^\top,y]),f^*_x(x)\right)$ and $\rho=\mathrm{Cov}\left(f(\hat\theta,[\tilde x^\top,y]),f^*_x(\tilde x)\right)$. Since S is p.s.d., its determinant has to be non-negative, hence

$$1-\rho^2-\mathcal{C}^2\geq 0, \tag{E.3}$$

which, when plugged in (E.1), gives

$$\mathbb{E}_{\tilde x,y}\left[\left(f(\hat\theta,[\tilde x^\top,y])-f^*_x(\tilde x)\right)^2\right]\geq 2-2\sqrt{1-\mathcal{C}(\hat\theta)^2}. \tag{E.4}$$

This implies that an increase in $\mathcal{C}(\hat\theta)$ hurts the performance of the model when core and spurious features are sampled independently (and, thus, the model is tested out-of-distribution). This bound suggests the close connection between

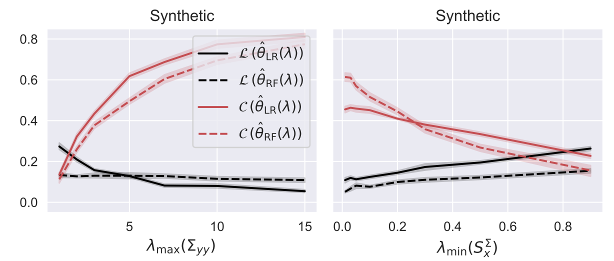

Figure 5: Test loss $\mathcal{L}(\hat{\theta}_{\rm LR/RF}(\lambda))$ (black) and spurious correlations $\mathcal{C}(\hat{\theta}_{\rm LR/RF}(\lambda)$ (red) as a function of $\lambda_{\max}\left(\Sigma_{yy}\right)$ (left) and $\lambda_{\min}\left(S_x^\Sigma\right)$ (right) on a synthetic Gaussian dataset, for both linear regression and random features, with $\lambda = 1$ (additional details in Appendix F).

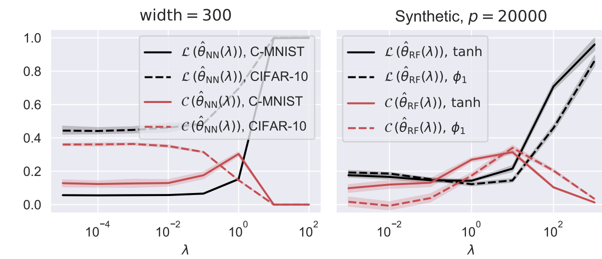

Figure 6: Test loss $\mathcal{L}(\hat{\theta}_{\rm NN/RF}(\lambda))$ (black) and spurious correlations $\mathcal{C}(\hat{\theta}_{\rm NN/RF}(\lambda)$ (red) as a function of λ. *Left:* 2-layer fully connected ReLU network, trained on the binary color(C)-MNIST and CIFAR-10 (boats and trucks). *Right:* RF model with $\tanh$ and $\phi_1 = h_1 + 0.1\, h_3$ activation.

$\mathcal{C}$ and the out-of-distribution test loss. In Figure 4, we repeat the same experiments of Figures 5-3 connecting $\mathcal{C}$ to $\lambda_{\min}\left(S_x^\Sigma\right)$ and $\lambda_{\max}\left(\Sigma_{yy}\right)$, and report in black the out-of-distribution test loss. The plots clearly show that $\mathcal{C}$ and the out-of-distribution test loss follow a similar trend, for both linear regression and random features.

We conclude the section by noting that learning spurious correlations can be beneficial to minimize the (in-distribution) test loss. In fact, the spurious features in y are effectively correlated with the labels, due to their correlation with the core feature x, and hence they can be helpful at prediction time. This phenomenon is numerically supported by Figures 5 and 3, where for a fixed value of λ, easier spurious features (or higher correlations) generate both higher values of $\mathcal{C}(\hat{\theta}_{\rm LR}(\lambda))$ and lower values of $\mathcal{L}(\hat{\theta}_{\rm LR}(\lambda))$. In words, while a blue background cannot strictly predict the label "boat", it is a useful feature in prediction as long as the boats in the test data tend to have a blue background.

F Experimental details

All the plots in the figures report the average over 10 independent trials, with a shaded area describing a confidence interval of 1 standard deviation. For the Gaussian and Color-MNIST datasets, every iteration involves re-generating (or re-coloring) the data, while for the CIFAR-10 dataset the randomness comes from the model and the training algorithm.

Effect of over-parameterization and activation. The right panel of Figure 6 presents the test loss $\mathcal{L}(\hat{\theta}_{\rm RF}(\lambda))$ (in black) and the spurious correlations $\mathcal{C}(\hat{\theta}_{\rm RF}(\lambda))$ (in red) for two activation functions: $\tanh$ and $\phi_1 = h_1 + 0.1 h_3$, where h_1 and h_3 denote the first and third Hermite polynomials, respectively. Notice that this gives $\tilde{\mu}^2/\mu_1^2 \sim 0.1$ for $\tanh$, $\tilde{\mu}^2/\mu_1^2 \sim 0.01$ for ϕ_1, and we take $d = 400$ and $n = 2000$. As expected, $\mathcal{C}(\hat{\theta}_{\rm RF}(0)) > 0$ for the $\tanh$ activation function, since $\tilde{\lambda} \sim 0.05$ (which matches the corresponding value in Figure 2). On the other hand, $\mathcal{C}(\hat{\theta}_{\rm RF}(0)) \sim 0$ for the activation ϕ_1, since $\tilde{\lambda} \sim 0.005$. As λ grows, $\mathcal{C}(\hat{\theta}_{\rm RF}(\lambda))$ goes to 0 faster for the $\tanh$ activation function (which has higher $\tilde{\lambda}$), as predicted by the second upper bound in Proposition 4.1.

Synthetic Gaussian data generation. This follows the same model across all the numerical experiments presented in the paper. In particular, we fix $d = 400$ and set $\Sigma_{xx} = I$. Σ_{yy} is a diagonal matrix, such that its first entry equals $\lambda_{\max}\left(\Sigma_{yy}\right)$ and all the other entries equal $\left(d - \lambda_{\max}\left(\Sigma_{yy}\right)\right)/(d-1)$. In this way, $\mathrm{tr}(\Sigma) = 2d$. Then, we set the off-diagonal blocks Σ_{xy} and Σ_{yx} to the same diagonal matrix, so that

$$\Sigma_{xy} = \Sigma_{yx} = (\Sigma_{yy} - \beta I)^{1/2}, \tag{F.1}$$

which implies that the Schur complement $S_x^\Sigma = \beta I$ and, therefore, $\lambda_{\min}\left(S_x^\Sigma\right) = \beta$. To conclude, we set the ground truth $\theta_x^* = e_1$, *i.e.*, the first element of the canonical basis in $\mathbb{R}^d$. This design choice is motivated by our interest in capturing the role of $\lambda_{\max}\left(\Sigma_{yy}\right)$ and to have an easy control on the Schur complement S_x^Σ (which is therefore chosen to be proportional to the identity).

Unless differently stated in the figure, $n = 2000$, $\lambda_{\max}\left(\Sigma_{yy}\right) = 2$, $\beta = 0.5$, and $\lambda = 1$. Furthermore, to generate the labels, we add an independent noise with variance $\sigma^2 = 0.25$, and we subtract this quantity from the test loss, so that the optimal predictor θ^* has a test loss equal to 0.

When we use an RF model, on every dataset, unless differently stated in the figure, we use $\tanh$ as activation function, with $p = 20000$ neurons.

Binary color MNIST. This dataset is graphically shown in Figure 1. To generate it, we take a subset of the MNIST training dataset ($n = 1000$ samples as default, unless differently specified) made only of zeros and ones. Then, for every training image, we color the white portion in blue (red) with probability $(1+\alpha)/2$ if the digit is a zero (one), and red (blue) otherwise. For the test set, we proceed in the same way, but setting $\alpha = 0$, to make the core feature (the digit) effectively independent from the spurious one (the color). For all the experiments, we set $\beta = 1 - \alpha^2 = 0.25$.

CIFAR-10. For the experiments on CIFAR-10, we implicitly suppose that the middle 22×22 square contains the core, predictive feature x. Thus, we sum to all the channels of the outer region white noise with increasing variance, and we later clamp the pixels to ensure their value is between 0 and 1. Increasing the variance of the noise, this progressively makes the outer portion being dominated by random noise, thus reducing its value of $\lambda_{\max}(\Sigma_{yy}) / \operatorname{tr}(\Sigma_{yy})$ when estimating the covariance on the perturbed training set. At test time, we take the images from the CIFAR-10 test set, and we add the same level of noise. To compute $\mathcal{C}$, we create an out-of-distribution dataset where the core features are randomly permuted across different backgrounds. We always consider the subset of boats and trucks, which contains $n = 10000$ images.

2-layer neural network. In the experiments shown in Figure 6, we consider a 2-layer neural network trained with gradient descent and quadratic loss on the Color-MNIST and CIFAR-10 datasets. For both datasets, we train for 1000 epochs, with learning rate 0.003, and batch size 1000.