

Surprising benefits of ridge regularization for noiseless regression

Anonymous Authors¹

Abstract

Numerous recent works show that overparameterization implicitly reduces variance for minimum-norm interpolators, suggesting vanishing benefits for ridge regularization in high dimensions. However, empirical findings suggest that this narrative may not hold true for robust generalization. In this paper we reveal that for overparameterized linear regression, the robust risk is minimized for a positive regularization coefficient even when the training data is noiseless. Hence, we effectively provide, to the best of our knowledge, the first theoretical analysis on the phenomenon of robust overfitting.

1. Introduction

Conventional statistical wisdom suggests that optimizing the data fit, by minimizing a loss $\mathcal{L}(\theta)$ with respect to parameters θ , should be exercised with care: if a global minimizer achieves zero or near-zero training loss (i.e., it *interpolates*), we run the risk of overfitting (i.e., high variance) and thus sub-optimal prediction performance. For regression on linearized models, *ridge regularization* is commonly used to reduce the effect of noise and to obtain an estimator with better generalization. In this paper, we study the ridge estimate for linear models that minimizes $\mathcal{L}(\theta) + \lambda \|\theta\|_2^2$, reducing model complexity by paying the price of worse data fit.

This classical rationale is challenged by recent observations on overparameterized models, often referred to as double descent: Neural networks, for example, do not sacrifice generalization performance on i.i.d. samples when they are very wide and trained until convergence to achieve vanishing $\mathcal{L}(\theta)$ (Nakkiran et al., 2020), i.e. they *interpolate*.

For linear regression without additional prior knowledge, a natural interpolator to study is the *minimum-norm interpolator*: not only is it the ridge estimate with $\lambda \rightarrow 0$ but it also

corresponds to the solution of gradient descent initialized at zero. A plethora of recent papers explicitly study generalization properties of min-norm interpolators (Dobriban & Wager, 2018; Ghorbani et al., 2021; Hastie et al., 2019; Bartlett et al., 2020; Mei & Montanari, 2019; Muthukumar et al., 2020a;b) and show that the variance decreases as the overparameterization ratio increases. Motivated by the double descent phenomenon (Belkin et al., 2018), most of them focus on settings where the optimal regularization parameter satisfies $\lambda_{\text{opt}} \leq 0$ (Kobak et al., 2020; Wu & Xu, 2020; Richards et al., 2021), implying that it is redundant or even detrimental to explicitly regularize with $\lambda > 0$.

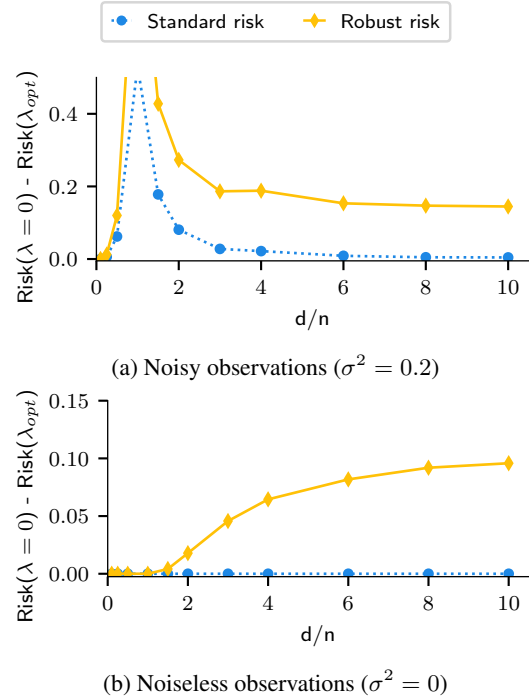


Figure 1. Decrease of population risks for linear regression when using the ridge estimate $\lambda_{\text{opt}} > 0$ as opposed to the minimum-norm interpolant ($\lambda = 0$) (detailed settings as in Fig. 2). While the standard risk only benefits from ridge regularization in the noisy case, the robust risk decreases even in the noiseless and heavily overparameterized setting $d/n \gg 1$.

¹Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

that analyze the *standard* test risk. Since this measure of performance assumes identically distributed training and test data however, it does not reflect the *robust risk* of models when the test data has a shifted distribution, is attacked by adversaries, or primarily contains samples from minority groups. In fact, mounting empirical evidence suggests that regularization is indeed helpful for robust generalization, even when it *does not* benefit the standard risk (Rice et al., 2020; Sagawa et al., 2020a;b). This phenomenon is sometimes referred to as *robust overfitting*.

In the presence of noise, the following intuition holds true: since the robust risk amplifies estimation errors, its variance is larger and hence regularization, e.g. via early stopping, can be beneficial for generalization (Sanyal et al., 2021). However, we observe that even when the training data is entirely *noiseless*, robust overfitting persists! In particular, we observe that for truly high-dimensional feature models, the robust risk of linear minimum ℓ_2 norm interpolators (i.e. $\lambda \rightarrow 0$) is larger compared to an explicit ridge estimate with $\lambda > 0$. Figure 1 shows the difference between the risk of the optimal regularized estimator ($\lambda = \lambda_{\text{opt}}$) and the unregularized estimator ($\lambda = 0$). We can see that in contrast to the standard risk, the robust risk benefits from regularization in the overparameterized regime $d \gg n$ and even for noiseless data.

To date, prior work does not predict nor explain our observations in Figure 1 that contradict intuition: if ℓ_2 -norm minimization is yielding a good solution for noiseless data, why would it help to deteriorate data fit and increase weight on the ridge penalty? For linear ridge regression with isotropic Gaussian covariates, we show that asymptotically as $d, n \rightarrow \infty$ and $d/n \rightarrow \gamma$, a strictly positive ℓ_2 penalty leads to a systematic improvement in robust generalization. Our results provide the first rigorous explanation for robust overfitting even in the absence of noise.

2. Risk minimization framework

In this section, we describe the setup for our theoretical analysis of linear regression. We define the data generating process, the standard and robust risks, and formally introduce the estimators that we consider.

2.1. Problem setting

We consider an observation model with covariates $x \in \mathbb{R}^d$ drawn from a standard normal distribution with zero mean, i.e., $x \sim \mathcal{N}(0, I_d)$, and with the target variable $y \in \mathbb{R}$ defined as a noisy observation of a linear function $y = \langle \theta^*, x \rangle + \xi$ with $\xi \sim \mathcal{N}(0, \sigma^2)$. Here, θ^* denotes the ground truth vector which we set to have unit ℓ_2 -norm. As mentioned before, we focus on the noiseless case of $\sigma = 0$ since overfitting is not expected for that case and include $\sigma > 0$ only for completeness.

Throughout this paper, we assume \mathbb{P} to be an isotropic Gaussian distribution, and study the high dimensional asymptotic regime where $d/n \rightarrow \gamma$ as both the dimensionality d and the number of samples n tend to infinity. This regime (Bühlmann & Van De Geer, 2011; Wainwright, 2019) is widely studied as it yields precise predictions for many real world problems where both the input dimension and the data set size are large. It is also the predominant setting considered in previous theoretical papers that discuss overparameterized linear models (Ali et al., 2020; Deng et al., 2021; Dobriban & Wager, 2018; Hastie et al., 2019; Javanmard et al., 2020; Javanmard & Soltanolkotabi, 2020; Sur & Candès, 2019).

2.2. Standard and robust risk

We now introduce the standard and robust evaluation metrics for regression. We define the standard risk of an estimator $\hat{\theta}$ to be the population mean squared error

$$\mathbf{R}(\hat{\theta}) := \mathbb{E}_{X \sim \mathbb{P}} \left(\langle \hat{\theta}, X \rangle - \langle \theta^*, X \rangle \right)^2, \quad (1)$$

where the expectation is taken over the marginal feature distribution \mathbb{P} . Note that conditioned on the training data, this risk is fixed, and our asymptotic bounds hold almost surely over draws of the training set.

The broad application of ML models in real-world decision-making processes increases requirements on their robustness. One example that is studied extensively in the literature, is robustness against adversarial ϵ -bounded ℓ_p -perturbations with $p = 2$ or $p = \infty$. Importantly, the ground truth is not guaranteed to achieve zero robust risk if we measure robustness with respect to any perturbation in the ϵ -ball around a test point x (see e.g. Raghunathan et al. (2020)). To make the problem well-behaved, we consider *consistent* perturbations, that are constrained to an ϵ -ball around x and, in addition, need to be orthogonal to θ^* . Therefore, the ground truth θ^* is indeed robust against such attacks and is the unique vector achieving zero robust risk. We can write the adversarially robust risk with respect to consistent ℓ_p perturbations as

$$\mathbf{R}_\epsilon(\hat{\theta}) := \mathbb{E}_{X \sim \mathbb{P}} \max_{\delta \in \mathcal{U}_p(\epsilon)} \left(\langle \hat{\theta}, X + \delta \rangle - \langle \theta^*, X \rangle \right)^2, \quad (2)$$

with $\mathcal{U}_p(\epsilon) := \{\delta \in \mathbb{R}^d : \|\delta\|_p \leq \epsilon \text{ and } \langle \theta^*, \delta \rangle = 0\}$. The following lemma provides a closed-form expression of the robust risk. A similar result for inconsistent attacks has already been shown before (Lemma 3.1. in (Javanmard et al., 2020)).

Lemma 2.1. *Assume that \mathbb{P}_X is the isotropic Gaussian distribution, and let Π_\perp be the orthonormal projection onto the groundtruth vector θ^* . Then, the robust risk (2) with*

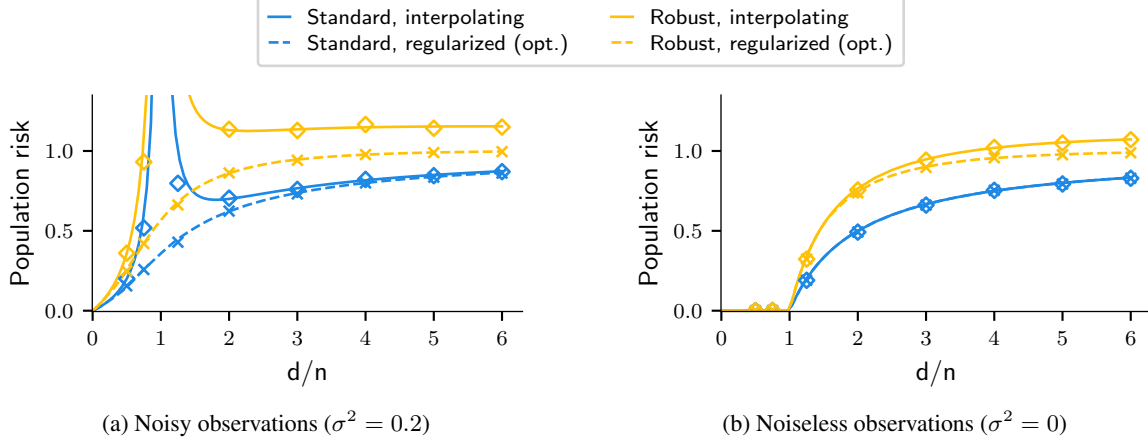


Figure 2. Asymptotic theoretical predictions for $d, n \rightarrow \infty$ (curves) and experimental results with finite d, n (markers) for the robust (yellow, $\epsilon = 0.4$) and standard (blue) risk of the min-norm solution (solid, *interpolating*) and the ridge regression estimate with optimal λ (dashed, *regularized*) for noisy (left) and noiseless data (right). The experiments use $n = 1000$ samples of the model described in Section 2.1. We observe that the gap between the robust risk of the interpolating and optimally regularized estimator persists even in the noiseless case.

respect to ℓ_2 perturbations is given by

$$\mathbf{R}_\epsilon(\theta) = \|\theta^* - \theta\|_2^2 + 2\epsilon\sqrt{2/\pi}\|\Pi_\perp \theta\|_2\|\theta^* - \theta\|_2 + \epsilon^2\|\Pi_\perp \theta\|_2^2. \quad (3)$$

In many applications, security concerns go beyond robustness to adversarial ℓ_p -perturbations and one may expect estimators to also be robust against small distribution shifts. As earlier work (Sinha et al., 2018) points out, distribution shift robustness and adversarial robustness are equivalent for losses that are convex in the parameter θ . In particular, for consistent ℓ_p perturbations, the adversarially robust risk upper bounds the distributionally robust risk in a neighborhood of mean-shifted distributions (see Appendix A.2). Hence, in our setting, adversarial robustness implies distributional robustness.

2.3. Interpolating and regularized estimator

We study standard linear ridge regression:

$$\hat{\theta}_\lambda = \arg \min_{\theta} \frac{1}{n} \sum_{i=0}^n (y_i - \langle \theta, x_i \rangle)^2 + \lambda \|\theta\|_2^2. \quad (4)$$

For $\lambda > 0$, we obtain a regularized predictor $\hat{\theta}_\lambda$. In addition, we also consider the minimum norm interpolator, namely the unregularized estimator that corresponds to $\lambda \rightarrow 0$ and can be written as follows:

$$\hat{\theta}_0 = \arg \min_{\theta} \|\theta\|_2 \text{ such that } \langle \theta, x_i \rangle = y_i \text{ for all } i. \quad (5)$$

For linear regression on isotropic features there exists a well-known correspondence between the optimization path of zero-initialized gradient descent and the regularization path

of the ridge regression estimator (see for example (Ali et al., 2019; 2020)). Hence, the results presented in this paper for ridge regularization directly translate to early stopped gradient descent on the mean squared loss.

3. Main results for ridge regression

In this section we prove that preventing interpolation on noiseless samples via ridge regularization (4) with $\lambda > 0$ improves the robust risk relative to an un-regularized, interpolating predictor. We further provide an explanation for why regularization is beneficial for the robust risk even when it is not beneficial for improving the standard risk.

3.1. Main result

The following theorem provides a precise asymptotic description of the consistent ℓ_2 robust risk for the ridge regression estimate in (4) as $d/n \rightarrow \gamma$ and $d, n \rightarrow \infty$. The proof uses techniques from (Hastie et al., 2019; Knowles & Yin, 2014) and can be found in Appendix B.

Theorem 3.1. Assume isotropic Gaussian covariates, i.e., $\mathbb{P}_X = \mathcal{N}(0, I_d)$. Then, for $d, n \rightarrow \infty$ with $d/n \rightarrow \gamma$ and for any $\lambda \geq 0$, the robust risk (2) with respect to consistent ℓ_2 perturbations of the estimators $\hat{\theta}_\lambda$ defined in Equations (4), (5) converges to

$$\mathbf{R}_\epsilon(\hat{\theta}_\lambda) \xrightarrow{a.s.} \mathcal{B} + \mathcal{V} + \epsilon^2 \mathcal{P} + \sqrt{\frac{8\epsilon^2}{\pi} \mathcal{P}(\mathcal{B} + \mathcal{V})}, \quad (6)$$

where $\mathcal{P} = \mathcal{B} + \mathcal{V} - \lambda^2(m(-\lambda))^2$ and $\mathcal{B} = \lambda^2 m'(-\lambda)$, $\mathcal{V} = \sigma^2 \gamma(m(-\lambda) - \lambda m'(-\lambda))$ are the asymptotic bias and variance with $\mathbf{R}(\hat{\theta}_\lambda) \xrightarrow{a.s.} \mathcal{B} + \mathcal{V}$. The function $m(z)$ is

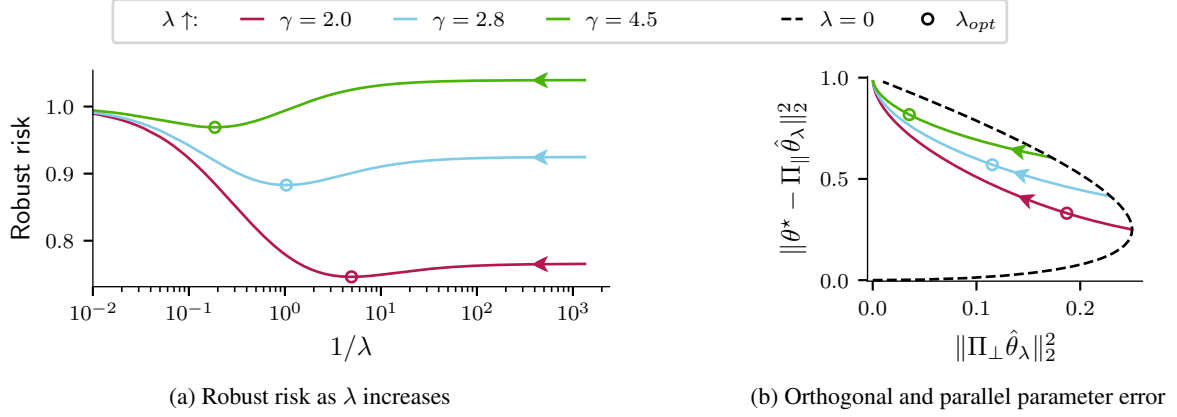


Figure 3. Theoretical curves for the robust risk with $\epsilon = 0.4$ (a) and decomposed terms (b) as λ increases (arrow direction) for different choices of the overparameterization ratio $d/n \rightarrow \gamma$. In (b) we observe for large $\gamma > 1$ that, as λ increases, the orthogonal error $\|\Pi_{\perp} \hat{\theta}_{\lambda}\|_2$ decreases, whereas the parallel error $\|\theta^* - \Pi_{\parallel} \hat{\theta}_{\lambda}\|_2$ increases. For $\epsilon > 0$, the optimal λ is large enough to prevent interpolation.

given by $m(z) = \frac{1-\gamma-z-\sqrt{(1-\gamma-z)^2-4\gamma z}}{2\gamma z}$ and m' is the derivative of m .

Using the precise asymptotics derived in this theorem, we plot the asymptotic standard and robust risk of the ridge estimate with the optimal λ parameter and the min-norm interpolator in Figure 2.¹

Firstly, Figure 2a reveals that ridge regularization reduces the robust risk even for $d > n$ well beyond the interpolation threshold where previous works show that variance is negligible. For the robust risk we use $\epsilon = 0.4$.

Moreover, Figure 2b shows that the beneficial effect of ridge regularization persists even for noiseless data, that is when $\mathcal{V} = 0$. This supports our statement that regularization not only helps reduce variance, but also reduces the part of the robust risk that is unaffected by noise in the overparameterized regime. Furthermore, we show that experiments (depicted by the markers in Figure 2) run with finite d and n closely match the predictions in Theorem 3.1 for $d, n \rightarrow \infty$ and $d/n \rightarrow \gamma$. This indicates that the asymptotic regime that we adopt in this paper indeed characterizes the high-dimensional non-asymptotic regime well. Finally, even though Theorem 3.1 assumes isotropic Gaussian covariates $\Sigma_d = I_d$, we can extend the result to more general covariance matrices following the same argument as in (Hastie et al., 2019) based on random matrix theory (Knowles & Yin, 2014).

3.2. Intuitive explanations and discussion

We now shed light on the phenomena revealed by Theorem 3.1 and Figure 2. In particular we discuss why regular-

¹Here we choose λ using the population risk oracle, in practice one would resort to standard tools such as cross-validation techniques (Patil et al., 2021).

ization can reduce the robust risk even in a noiseless setting and why the effect is not noticeable for the standard risk.

For this purpose, we examine the robust risk as a function of λ , depicted in Figure 3a for different overparameterization ratios $\gamma > 1$ and $\epsilon = 0.4$. In order to understand the overfitting phenomenon, we decompose the ridge estimate $\hat{\theta}_{\lambda}$ into its projection Π_{\parallel} on the ground-truth direction θ^* and its projection Π_{\perp} onto its orthogonal complement, i.e., $\hat{\theta}_{\lambda} = \Pi_{\parallel} \hat{\theta}_{\lambda} + \Pi_{\perp} \hat{\theta}_{\lambda}$. For the noiseless setting ($\sigma^2 = 0$), the closed-form expression of the consistent robust risk involves the parallel error $\|\theta^* - \Pi_{\parallel} \hat{\theta}_{\lambda}\|_2^2$ and the orthogonal error $\|\Pi_{\perp} \hat{\theta}_{\lambda}\|_2^2$ as follows:

$$\mathbf{R}_{\epsilon}(\hat{\theta}_{\lambda}) = \|\theta^* - \Pi_{\parallel} \hat{\theta}_{\lambda}\|_2^2 + (1 + \epsilon^2) \|\Pi_{\perp} \hat{\theta}_{\lambda}\|_2^2 + \sqrt{\frac{8\epsilon^2}{\pi} \|\Pi_{\perp} \hat{\theta}_{\lambda}\|_2^2 (\|\theta^* - \Pi_{\parallel} \hat{\theta}_{\lambda}\|_2^2 + \|\Pi_{\perp} \hat{\theta}_{\lambda}\|_2^2)} \quad (7)$$

The proof can be found in Appendix A.1. As λ increases, Figure 3b shows that the parallel error increases faster than the orthogonal error decreases. When the perturbation strength ϵ is large enough, we see in expression (7) that the orthogonal error is weighted more heavily. Hence, for the robust risk, some nonzero λ that prevents interpolation yields the best trade-off. On the other hand the standard risk with $\epsilon = 0$ weighs both errors equally, resulting in an optimum at $\lambda = 0$.

4. Conclusions

We show that high dimensional linear regression can overfit with respect to the robust risk even when there is no noise in the training data set. Our results put into perspective the modern narrative that interpolating overparameterized models yield good performance without explicit regularization and motivate the use of ridge regularization and early stopping for improved robust generalization.

References

- Ali, A., Kolter, J. Z., and Tibshirani, R. J. A continuous-time view of early stopping for least squares regression. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 89, pp. 1370–1378, 16–18 Apr 2019.
- Ali, A., Dobriban, E., and Tibshirani, R. The implicit regularization of stochastic gradient flow for least squares. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 119, pp. 233–244, 13–18 Jul 2020.
- Bai, Z. D. and Yin, Y. Q. Limit of the Smallest Eigenvalue of a Large Dimensional Sample Covariance Matrix. *The Annals of Probability*, 21(3):1275 – 1294, 1993.
- Bartlett, P. L., Long, P. M., Lugosi, G., and Tsigler, A. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.
- Belkin, M., Ma, S., and Mandal, S. To understand deep learning we need to understand kernel learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 80, pp. 541–549, 2018.
- Bühlmann, P. and Van De Geer, S. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.
- Bühlmann, P. et al. Invariance, causality and robustness. *Statistical Science*, 35(3):404–426, 2020.
- Chen, Y. and Bühlmann, P. Domain adaptation under structural causal models. *arXiv preprint arXiv:2010.15764*, 2020.
- Deng, Z., Kammoun, A., and Thrampoulidis, C. A model of double descent for high-dimensional binary linear classification. *Information and Inference: A Journal of the IMA*, 04 2021.
- Dobriban, E. and Wager, S. High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics*, 46(1):247 – 279, 2018.
- Ghorbani, B., Mei, S., Misiakiewicz, T., and Montanari, A. Linearized two-layers neural networks in high dimension. *The Annals of Statistics*, 49(2):1029 – 1054, 2021.
- Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. J. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019.
- Javanmard, A. and Soltanolkotabi, M. Precise statistical analysis of classification accuracies for adversarial training. *arXiv preprint arXiv:2010.11213*, 2020.
- Javanmard, A., Soltanolkotabi, M., and Hassani, H. Precise tradeoffs in adversarial training for linear regression. In *Proceedings of the Conference on Learning Theory (COLT)*, pp. 2034–2078, 2020.
- Knowles, A. and Yin, J. Anisotropic local laws for random matrices. *Probability Theory and Related Fields*, 169: 257–352, 2014.
- Kobak, D., Lomond, J., and Sanchez, B. The optimal ridge penalty for real-world high-dimensional data can be zero or negative due to the implicit ridge regularization. *Journal of Machine Learning Research*, 21(169):1–16, 2020.
- Mei, S. and Montanari, A. The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv preprint arXiv:1908.05355*, 2019.
- Muthukumar, V., Narang, A., Subramanian, V., Belkin, M., Hsu, D., and Sahai, A. Classification vs regression in overparameterized regimes: Does the loss function matter? *arXiv preprint arXiv:2005.08054*, 2020a.
- Muthukumar, V., Vodrahalli, K., Subramanian, V., and Sahai, A. Harmless interpolation of noisy data in regression. *IEEE Journal on Selected Areas in Information Theory*, 1 (1):67–83, 2020b.
- Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., and Sutskever, I. Deep double descent: Where bigger models and more data hurt. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.
- Patil, P., Wei, Y., Rinaldo, A., and Tibshirani, R. Uniform consistency of cross-validation estimators for high-dimensional ridge regression. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 3178–3186, 2021.
- Raghuathan, A., Xie, S. M., Yang, F., Duchi, J., and Liang, P. Understanding and mitigating the tradeoff between robustness and accuracy. In *International Conference on Machine Learning*, pp. 7909–7919. PMLR, 2020.
- Rice, L., Wong, E., and Kolter, Z. Overfitting in adversarially robust deep learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 8093–8104, 2020.
- Richards, D., Mourtada, J., and Rosasco, L. Asymptotics of ridge (less) regression under general source condition. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 3889–3897. PMLR, 2021.
- Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020a.

- Sagawa, S., Raghunathan, A., Koh, P. W., and Liang, P. An investigation of why overparameterization exacerbates spurious correlations. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 119, pp. 8346–8356, 13–18 Jul 2020b.
- Sanyal, A., Dokania, P. K., Kanade, V., and Torr, P. How benign is benign overfitting? In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.
- Sinha, A., Namkoong, H., and Duchi, J. Certifying some distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018.
- Sur, P. and Candès, E. J. A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116(29):14516–14525, 2019.
- Wainwright, M. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019. ISBN 9781108498029.
- Wu, D. and Xu, J. On the optimal weighted l_2 regularization in overparameterized linear regression. *arXiv preprint arXiv:2006.05800*, 2020.

A. Setting

In Section A.1 we proof Lemma 2.1 which gives a closed form expression for the robust risk. Furthermore, Section A.2 discusses that the robust risk (2) upper-bounds the worst case risk under distributional mean shifts.

A.1. Proof of Lemma 2.1

Define $\tilde{y}_i = y_i - \langle x_i, \theta \rangle$, and note that using similar arguments as in Section 6.2. (Javanmard et al., 2020)

$$\begin{aligned} \max_{\delta_i \in \mathcal{U}_2(\epsilon)} (\tilde{y}_i - \langle \delta_i, \theta \rangle)^2 &= \left(\max_{\delta_i \in \mathcal{U}_2(\epsilon)} |\tilde{y}_i - \langle \delta_i, \theta \rangle| \right)^2 \\ &= (|\tilde{y}_i| + \max_{\|\delta_i\|_2 \leq \epsilon, \delta_i \perp \theta^*} |\langle \delta_i, \theta \rangle|)^2 \\ &= (|\tilde{y}_i| + \epsilon \|\Pi_\perp \theta\|_2)^2. \end{aligned}$$

With this characterization, we can derive a convenient expression for the robust risk:

$$\begin{aligned} \mathbf{R}_\epsilon(\theta) &= \mathbb{E}_X (|\langle X, \theta^* - \theta \rangle| + \epsilon \|\Pi_\perp \theta\|_2)^2 \\ &= \mathbb{E}_X (\langle X, \theta^* - \theta \rangle)^2 + 2\epsilon \mathbb{E}_X |\langle X, \theta^* - \theta \rangle| \|\Pi_\perp \theta\|_2 + \epsilon^2 \|\Pi_\perp \theta\|_2^2. \end{aligned} \quad (8)$$

Since we assume isotropic Gaussian features, that is $\mathbb{P}_X = \mathcal{N}(0, I)$, we can further simplify

$$\mathbf{R}_\epsilon(\theta) = \|\theta - \theta^*\|_2^2 + 2\epsilon \sqrt{2/\pi} \|\Pi_\perp \theta\|_2 \|\theta - \theta^*\|_2 + \epsilon^2 \|\Pi_\perp \theta\|_2^2$$

which concludes the proof.

A.2. Distribution shift robustness and consistent adversarial robustness

In this section we rigorously introduce distribution shift robustness and show the relation to consistent ℓ_p adversarial robustness for certain types of distribution shifts.

When learned models are deployed in the wild, the i.i.d. assumption does not always hold. That is, the test loss might be evaluated on samples from a slightly different distribution. Shifts in the mean of the covariate distribution is a standard intervention studied in the invariant causal prediction literature (Bühlmann et al., 2020; Chen & Bühlmann, 2020). For mean shifts in the null space of the ground truth θ^* we define an alternative evaluation metric that we refer to as the *distributionally robust risk* defined as follows:

$$\begin{aligned} \tilde{\mathbf{R}}_\epsilon(\theta) &:= \max_{\mathbb{Q} \in \mathcal{V}_\epsilon(\mathbb{P})} \mathbb{E}_{X \sim \mathbb{Q}} \ell_{\text{test}}(\langle \theta, X + \delta \rangle, \langle \theta^*, X \rangle), \text{ with} \\ \mathcal{V}_\epsilon(\mathbb{P}) &:= \{\mathbb{Q} \in \mathcal{P} : \|\mu_\mathbb{P} - \mu_\mathbb{Q}\|_p \leq \epsilon \text{ and } \langle \mu_\mathbb{P} - \mu_\mathbb{Q}, \theta^* \rangle = 0\} \end{aligned}$$

where \mathcal{V}_ϵ is the neighborhood of mean shifted probability distributions and ℓ_{test} is a convex loss function.

A duality between distribution shift robustness and adversarial robustness has been established in earlier work such as (Sinha et al., 2018) for general convex, continuous losses ℓ_{test} . For our setting, the following lemma holds.

Lemma A.1. For any $\epsilon \geq 0$ and θ , we have $\tilde{\mathbf{R}}_\epsilon(\theta) \leq \mathbf{R}_\epsilon(\theta)$.

Proof. The proof follows directly from the definition and consistency of the perturbations $\mathcal{U}_p(\epsilon)$ and orthogonality of the mean shifts for the neighborhood \mathcal{V}_p . By defining the random variable $w = x - \mu_\mathbb{P}$ for $x \sim \mathbb{P}$ we have the distributional equivalence

$$x' = \mu_\mathbb{P} + \delta + w \stackrel{d}{=} x + \delta$$

for $x' \sim \mathbb{Q}$ and $x \sim \mathbb{P}$ with $\mu_\mathbb{Q} - \mu_\mathbb{P} = \delta$ and hence

$$\begin{aligned} \tilde{\mathbf{R}}_\epsilon(\theta) &= \max_{\mathbb{Q} \in \mathcal{V}_\epsilon(\mathbb{P})} \mathbb{E}_{x \sim \mathbb{Q}} \ell_{\text{test}}(\langle \theta, x \rangle, \langle \theta^*, x \rangle) = \max_{\|\delta\|_p \leq \epsilon, \delta \perp \theta^*} \mathbb{E}_{x \sim \mathbb{P}} \ell_{\text{test}}(\langle \theta, x + \delta \rangle, \langle \theta^*, x \rangle) \\ &\leq \mathbb{E}_{x \sim \mathbb{P}} \max_{\|\delta\|_p \leq \epsilon, \delta \perp \theta^*} \ell_{\text{test}}(\langle \theta, x + \delta \rangle, \langle \theta^*, x \rangle) = \mathbf{R}_\epsilon(\theta) \end{aligned}$$

where the first line follows from orthogonality of the mean-shift to θ^* . \square

B. Proof of Theorem 3.1

In this section, we provide a proof of Theorem 3.1, which characterizes the asymptotic risk of the linear regression estimator $\hat{\theta}_\lambda$ defined in Equation (4).

We first introduce some notation and give the standard closed form solution for the ridge regression estimate $\hat{\theta}_\lambda$. Denoting the input data matrix by $X \in \mathbb{R}^{d \times n}$, the observation vector $y \in \mathbb{R}^n$ reads $y = X^\top \theta^* + \xi$ with $\xi \sim \mathcal{N}(0, I)$ the noise vector. the noise vector containing i.i.d. zero-mean σ^2 -variance Gaussian noise as entries. Defining the empirical covariance matrix as $\hat{\Sigma} = \frac{1}{n} X^\top X$ yields the ridge estimate

$$\begin{aligned} \hat{\theta}_\lambda &= \frac{1}{n} (\lambda I_d + \hat{\Sigma})^{-1} X^\top y \\ &= (\lambda I_d + \hat{\Sigma})^{-1} \hat{\Sigma} \theta^* + \frac{1}{n} (\lambda I_d + \hat{\Sigma})^{-1} X^\top \xi. \end{aligned} \quad (9)$$

For $\lambda \rightarrow 0$, we obtain the min-norm interpolator

$$\hat{\theta}_0 = \lim_{\lambda \rightarrow 0} \hat{\theta}_\lambda = (\lambda I_d + \hat{\Sigma})^{-1} X^\top y = \hat{\Sigma}^\dagger X^\top y$$

where $\hat{\Sigma}^\dagger$ denotes the Moore-Penrose pseudo inverse.

We now compute the adversarial risk of this estimator. By Equation (3), the adversarial risk depends on the estimator only via the two terms $\|\hat{\theta}_\lambda - \theta^*\|_2$ and $\|\Pi_\perp \hat{\theta}_\lambda\|_2$. To characterize the asymptotic risk, we hence separately derive asymptotic expressions for each of both terms. The following convergence results hold almost surely with respect to the draws of the train dataset, with input features X and observations y , as $n, d \rightarrow \infty$.

Step 1: Characterizing $\|\hat{\theta}_\lambda - \theta^*\|_2^2$. Here, we show that

$$\|\hat{\theta}_\lambda - \theta^*\|_2^2 \rightarrow \mathcal{B} + \mathcal{V}. \quad (10)$$

where $\mathcal{B} = \lambda^2 m'(-\lambda)$, $\mathcal{V} = \sigma^2 \gamma(m(-\lambda) - \lambda m'(-\lambda))$ are the asymptotic bias and variance as in the Theorem and Theorem 5 of (Hastie et al., 2019), where the authors shows that $\mathbb{E}_\xi \|\hat{\theta}_\lambda - \theta^*\|_2^2 \rightarrow \mathcal{B} + \mathcal{V}$ and the expectation is taken over the observation noise ξ in the train dataset. In this paper, we define the population risks without the expectation over the noise. Hence, in a first step, the goal is to extend Theorem 5 (Hastie et al., 2019) for the standard risk $\mathbf{R}(\hat{\theta}_\lambda) = \|\hat{\theta}_\lambda - \theta^*\|_2^2$ such that (10) holds almost surely over the draws of the training data.

Using Equation (9) we can rewrite

$$\begin{aligned} \|\hat{\theta}_\lambda - \theta^*\|_2^2 &= \left\| \left(I_d - (\lambda I_d + \hat{\Sigma})^{-1} \hat{\Sigma} \right) \theta^* + \frac{1}{n} (\lambda I_d + \hat{\Sigma})^{-1} X^\top \xi \right\|_2^2 \\ &= \underbrace{\left\| \left(I_d - (\lambda I_d + \hat{\Sigma})^{-1} \hat{\Sigma} \right) \theta^* \right\|_2^2}_{T_1} + \underbrace{\left\langle \frac{\xi}{\sqrt{n}}, (\lambda I_d + \hat{\Sigma})^{-2} \hat{\Sigma} \frac{\xi}{\sqrt{n}} \right\rangle}_{T_2} \\ &\quad + \underbrace{\left\langle \frac{X^\top}{\sqrt{n}} (\lambda I_d + \hat{\Sigma})^{-1} \left(I_d - (\lambda I_d + \hat{\Sigma})^{-1} \hat{\Sigma} \right) \theta^*, \frac{\xi}{\sqrt{n}} \right\rangle}_{T_3}, \end{aligned}$$

where we used for the second term that $\left\langle \frac{\xi}{\sqrt{n}}, \frac{X}{\sqrt{n}} (\lambda I_d + \hat{\Sigma})^{-2} \frac{X^\top}{\sqrt{n}} \frac{\xi}{\sqrt{n}} \right\rangle = \left\langle \frac{\xi}{\sqrt{n}}, (\lambda I_d + \hat{\Sigma})^{-2} \hat{\Sigma} \frac{\xi}{\sqrt{n}} \right\rangle$.

The first term $T_1 \rightarrow \mathcal{B}$ directly via Theorem 5 (Hastie et al., 2019). We next show that $T_2 \rightarrow \mathcal{V}$ and $T_3 \rightarrow 0$ almost surely, which establishes Equation 10.

Proof that $T_2 \rightarrow \mathcal{V}$: While Theorem 5 (Hastie et al., 2019) also shows that $\mathbb{E}_\xi \text{tr} \left(\frac{1}{n} \xi \xi^\top \hat{\Sigma} (\lambda I_d + \hat{\Sigma})^{-2} \right) \rightarrow \mathcal{V}$, we require the convergence almost surely over a single draw of ξ . In fact, this directly follows from the same argument as used for the proof of Theorem 5 (Hastie et al., 2019) and the fact that $\left\| \frac{\xi}{\sqrt{n}} \right\|_2^2 \rightarrow \sigma^2$. Hence $\left\langle \frac{\xi}{\sqrt{n}}, (\lambda I_d + \hat{\Sigma})^{-2} \hat{\Sigma} \frac{\xi}{\sqrt{n}} \right\rangle \rightarrow \mathcal{V}$ almost surely over the draws of ξ .

Proof that $T_3 \rightarrow 0$: This follows straight forwardly from sub-Gaussian concentration inequalities and from the fact that

$$\left\| \frac{X}{\sqrt{n}} (\lambda I_d + \widehat{\Sigma})^{-1} \left(I_d - (\lambda I_d + \widehat{\Sigma})^{-1} \widehat{\Sigma} \right) \theta^* \right\|_2 = O(1),$$

which is a direct consequence of the Bai-Yin theorem (Bai & Yin, 1993), stating that for sufficiently large n , the non zero eigenvalues of $\widehat{\Sigma}$ can be almost surely bounded by $(1 + \sqrt{\gamma})^2 \geq \lambda_{\max}(\widehat{\Sigma}) \geq \lambda_{\min}(\widehat{\Sigma}) \geq (1 - \sqrt{\gamma})^2$. Hence we can conclude the first part of the proof.

Step 2: Characterizing $\|\Pi_{\perp} \hat{\theta}_{\lambda}\|_2$. Here, we show that

$$\|\Pi_{\perp} \hat{\theta}_{\lambda}\|_2^2 \rightarrow \mathcal{B} + \mathcal{V} - \lambda^2(m(-\lambda))^2 =: \mathcal{P}. \quad (11)$$

We assume without loss of generality that $\|\theta^*\|_2 = 1$ and hence $\Pi_{\perp} = I_d - \theta^*(\theta^*)^{\top}$. It follows that

$$\begin{aligned} \|\Pi_{\perp} \hat{\theta}_{\lambda}\|_2^2 &= \|\hat{\theta}_{\lambda}\|_2^2 - \left(\langle \hat{\theta}_{\lambda}, \theta^* \rangle \right)^2 \\ &= \|\theta^* - \hat{\theta}_{\lambda} - \theta^*\|_2^2 - \left(1 - \langle \theta^* - \hat{\theta}_{\lambda}, \theta^* \rangle \right)^2 \\ &= \|\theta^* - \hat{\theta}_{\lambda}\|_2^2 - 2\langle \theta^* - \hat{\theta}_{\lambda}, \theta^* \rangle + 1 - \left(1 - \langle \theta^* - \hat{\theta}_{\lambda}, \theta^* \rangle \right)^2 \\ &= \|\theta^* - \hat{\theta}_{\lambda}\|_2^2 - \left(\langle \theta^* - \hat{\theta}_{\lambda}, \theta^* \rangle \right)^2. \end{aligned}$$

The convergence of the first term is already known from step 1. Hence, it is only left to find an asymptotic expression for $\langle \theta^* - \hat{\theta}_{\lambda}, \theta^* \rangle$. Inserting the closed form expression from Equation (9), we obtain:

$$\langle \theta^* - \hat{\theta}_{\lambda}, \theta^* \rangle = \langle I_d - (\lambda I_d + \widehat{\Sigma})^{-1} \widehat{\Sigma} \rangle \theta^*, \theta^* \rangle - \langle (\lambda I_d + \widehat{\Sigma})^{-1} \frac{X^{\top} \xi}{n}, \theta^* \rangle. \quad (12)$$

Note that $\langle (\lambda I_d + \widehat{\Sigma})^{-1} \frac{X^{\top} \xi}{n}, \theta^* \rangle$ vanishes almost surely over the draws of ξ using the same reasoning as in the first step. Hence, we only need to find an expression for the first term on the RHS of Equation (12). Note that we can use Woodbury's matrix identity to write:

$$\langle I_d - (\lambda I_d + \widehat{\Sigma})^{-1} \widehat{\Sigma} \rangle \theta^*, \theta^* \rangle = \lambda \langle (\lambda I_d + \widehat{\Sigma})^{-1} \theta^*, \theta^* \rangle.$$

However, the expression on the RHS appears exactly in the proof of Theorem 1 (Hastie et al., 2019) (Equation 116), which shows that $\lambda \langle (\lambda I_d + \widehat{\Sigma})^{-1} \theta^*, \theta^* \rangle \rightarrow \lambda m(-\lambda)$ with $m(z)$ as in Theorem 3.1. Hence the proof of almost sure convergence (11) of $\|\Pi_{\perp} \hat{\theta}_{\lambda}\|_2$ is complete.

Substituting Equations (10) and (11) into robust risk (3) expression yields:

$$\mathbf{R}_{\epsilon}(\hat{\theta}_{\lambda}) \xrightarrow{\text{a.s.}} \mathcal{B} + \mathcal{V} + \epsilon^2 \mathcal{P} + \sqrt{\frac{8\epsilon^2}{\pi} \mathcal{P}(\mathcal{B} + \mathcal{V})},$$

and the proof is complete.