Interpolation can hurt robust generalization even when there is no noise

Anonymous Author(s)
Affiliation
Address
email

Abstract

Numerous recent works show that overparameterization implicitly reduces variance for minimum-norm interpolators and max-margin classifiers, suggesting vanishing benefits for ridge regularization in high dimensions. Putting this narrative into perspective, our paper reveals the unexpected benefits of a non-zero ridge penalty even when there is no noise: we prove that for both overparameterized linear regression and classification, avoiding interpolation can in fact significantly improve generalization. The benefits are particularly prominent for distributionally and adversarially robust risks and we hence effectively provide, to the best of our knowledge, the first theoretical analysis on the phenomenon of robust overfitting.

1 Introduction

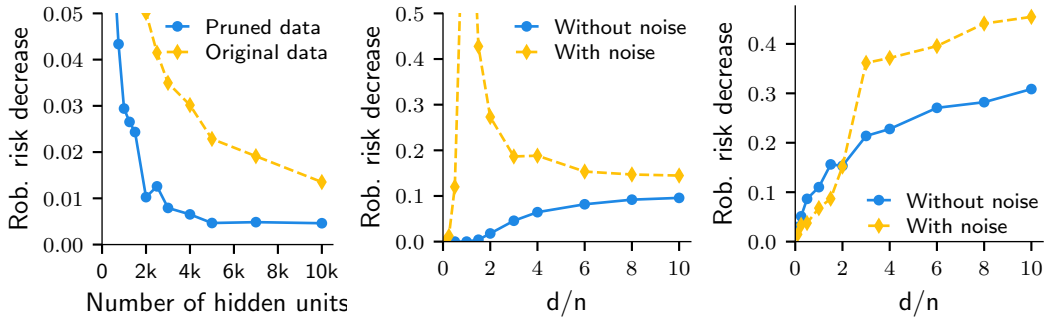
Conventional statistical wisdom suggests that optimizing the data fit by minimizing a loss $\mathcal{L}(\theta)$ with respect to parameters θ should be exercised with care: if a global minimizer achieves zero or near-zero training loss (i.e. it *interpolates*), we run the risk of overfitting (i.e. high variance) and thus sub-optimal prediction performance. *Regularization* is commonly used to reduce the effect of noise and to obtain an estimator with better generalization. Specifically, it induces worse data fit by limiting model complexity, for example via an explicit penalty term $R(\theta)$, which in turn imposes certain structural properties on the minimizer of the penalized loss $\mathcal{L}(\theta) + \lambda R(\theta)$. This classical rationale, however, does not apply to overparameterized models: Large neural networks, for example, have good generalization performance on i.i.d. samples when $\mathcal{L}(\theta)$ vanishes and label noise is present [29].

Since interpolators are not unique in the overparameterized regime, it is crucial to study the *implicit bias* of specific interpolating estimators. A large body of recent work analyzes the properties of the solution found via gradient descent at convergence for common choices of $\mathcal{L}(\theta)$ (see e.g. [34, 8, 9, 16, 21, 22, 40]). For example, for linear and logistic regression, it is well-known that gradient descent converges to the minimum-ℓ_2-norm and maximum-ℓ_2-margin solutions [16, 22, 25, 40], respectively, that also correspond to the minimizers of the loss $\mathcal{L}(\theta) + \lambda\|\theta\|_2^2$ in the limit of $\lambda \to 0$ [34].

How well do these specific interpolators generalize? A plethora of recent papers explicitly study generalization properties on min-norm interpolators [11, 13, 17, 4, 26, 27, 28] and max-margin solutions [10, 27, 37] and show that the variance decreases as overparameterization ratio increases. Motivated by the double descent phenomenon [5], most of them focus on settings where the optimal regularization parameter $\lambda_{\text{opt}} \leq 0$ [24, 47, 33], implying that it is redundant or even detrimental to explicitly regularize with $\lambda > 0$.

The narrative that regularization is obsolete for large overparameterized models is drawn from theoretical and experimental findings 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

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.



(a) Neural networks (b) Linear regression (c) Linear classification

Figure 1: In this plot we exemplify the benefits of early stopping (a) and nonzero ℓ_2 regularization (b),(c) for the robust risk (robust risk decrease). The decrease is strictly positive across most overparameterization ratios d/n even in the absence of noise. (a) Neural networks fit on (binary 1-3) MNIST that arguably has minimal noise compared to other image datasets and that we further sanitize. (b) Linear regression on noiseless observations for $n = 1k$. (c) Adversarial training on logistic regression for $n = 1k$. Note that for small d/n, adversarial training does not interpolate which leads to better robust accuracy (see Appendix D.3 for a more detailed discussion). Experimental details can be found in Appendix B. See Sections 3 and 4 for the precise settings for (b) and (c).

regularization is indeed helpful for robust generalization, even when it *does not* benefit the standard risk [32, 35, 36]. 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 [38]. However, we observe that even when estimating entirely noiseless signals, robust overfitting persists! In particular, we observe that for truly high-dimensional feature models, the robust risk of minimum ℓ_2 norm/max-margin interpolators (i.e. $\lambda \to 0$) suffers from overfitting when compared to explicit ridge regularization with $\lambda > 0$ that prevents interpolation. In Figure 1 we provide three examples including neural networks on sanitized image data linear models (see Section B for experimental details).

To date, prior work does not predict nor explain our observations. Further, they contradict a possible intuition for the noiseless case: if the min ℓ_2-norm and max-ℓ_2-margin interpolators have high risks as $\lambda \to 0$, the ℓ_2-norm induced bias is potentially suboptimal and hence increasing the penalty weight $\lambda > 0$ should only degrade performance. In this paper, we prove that this intuition in fact does not hold true in general in the high dimensional overparameterized regime. As noise is known to have negligible effects in the overparameterized regime, we focus on consistent training losses and noiseless observations – precisely the scenarios that are expected to benefit the least from a non-zero ℓ_2 penalty. In particular, we study point-wise losses averaged over the original data points or consistent adversarial perturbation sets around the training samples. For the square and logistic loss with isotropic Gaussian covariates, we then show that 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.

In Section 2 we formally define the evaluation metrics and training losses used to obtain interpolating and regularized estimators. We then present our precise asymptotic results for the robust risk of the minimum norm interpolator in Section 3 and the max-margin solution of the robust loss for logistic regression in Section 4. In each section we also provide additional insights to build intuition for the phenomenon of robust overfitting occurring even in the noiseless and linear case.

2 Risk minimization framework

In this section, we describe the setup for our theoretical analysis of linear and logistic regression. We define the data generating process, the standard and robust risks, and formally introduce the estimators that we consider. We present general concepts that apply for both regression and classification and follow up with more specific details in the respective sections.

We consider the supervised learning problem of estimating a mapping from d-dimensional real-valued features $x \in \mathbb{R}^d$ to a target $y \in \mathcal{Y} \subseteq \mathbb{R}$, given a set of input-output pairs $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$. The data is generated by drawing n i.i.d. feature vectors x from the marginal distribution of the inputs $\mathbb{P}$ and observing targets y that follow a conditional distribution depending on the linear ground truth function $\langle \theta^\star, x \rangle$. The exact expressions for regression and classification are given in Sections 3 and 4, respectively.

Throughout this paper, we assume $\mathbb{P}$ to be an isotropic Gaussian distribution, and study the high dimensional asymptotic regime where $d/n \to \gamma$ as both the dimensionality d and the number of samples n tend to infinity. This regime [6, 45] 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 [2, 10, 11, 17, 20, 19, 42].

2.1 Standard and robust risk

We now introduce the standard and robust evaluation metrics for regression and classification that we study in this paper. For a given test loss $\ell_{\text{test}} \colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, we define the standard risk of an estimator $\hat{\theta}$ as

$$\mathbf{R}(\hat{\theta}) := \mathbb{E}_{X \sim \mathbb{P}} \ell_{\text{test}} \left(\langle \hat{\theta}, X \rangle, \langle \theta^\star, X \rangle \right), \tag{1}$$

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. For regression tasks we choose $\ell_{\text{test}}(u, v) = (u - v)^2$ which yields the population mean squared error, and for classification we choose the 0-1 loss $\ell_{\text{test}}(u, v) = \mathbb{1}_{\text{sgn}(u) \neq \text{sgn}(v)}$ which yields the probability of making a classification error.

As more ML models are productionalized, there are increasing requirements on their robustness. One popular notion is robustness with respect to ϵ-bounded ℓ_p perturbations with $p = 2$ or $p = \infty$. We assume that $\theta^\star$ is indeed robust to begin with, that is $\theta^\star$ minimizes the population robust risk (see also for example [31]). The perturbations then have to be *consistent* in the sense that they do not change the ground truth label. Consequently, we can write the adversarially robust risk with respect to ℓ_p perturbations as

$$\mathbf{R}_\epsilon(\hat{\theta}) := \mathbb{E}_{X \sim \mathbb{P}} \max_{\delta \in \mathcal{U}_p(\epsilon)} \ell_{\text{test}}(\langle \hat{\theta}, X + \delta \rangle, \langle \theta^\star, X \rangle), \tag{2}$$

with $\mathcal{U}_p(\epsilon) := \{\delta \in \mathbb{R}^d : \|\delta\|_p \leq \epsilon \text{ and } \langle \theta^\star, \delta \rangle = 0\}$. In many application scenarios, robustness against adversarial ℓ_p-perturbations is not a main security concern. Instead, the estimator should be robust against small distribution shifts of the training and test data. As earlier work [39] pointed 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.3). Hence, a small adversarially robust risk implies a small distributionally robust risk.

2.2 Interpolation and regularization

In order to find an estimator $\hat{\theta}$ with small population risk, we minimize an average training loss $\ell_{\text{train}}(\langle \theta, x \rangle, y)$. Even though our results include the general setting for noisy observations, the fact that regularization helps generalization is much more surprising in the noiseless setting. Furthermore, the effect of noise is diminishing in the overparameterized regime. Hence, throughout this paper, we focus on the setting where all labels y in the loss follow the ground truth function and the signal is noiseless.

This also includes the adversarially robust loss (2) where the average is taken over worst-case bounded ℓ_p perturbations around the training points. We choose consistent perturbations orthogonal to the ground truth $\theta^\star$ for the same reason that we focus on the noiseless setting. Even though the orthogonal subspace constraint is not standard, inconsistent perturbations add noisy signal to the training loss — which is what we want to disentangle from finite-sample effects. Yet, our results are not limited to consistent attacks; we empirically show in Appendix D.1 that they also hold when training with inconsistent attacks.

In the noiseless overparameterized setting where $d > n$, many *interpolating estimators* $\hat{\theta}_0(\epsilon)$ with vanishing average loss $\frac{1}{n}\sum_{i=1}^{n} \max_{\delta \in \mathcal{U}_p(\epsilon)} \ell_{\text{train}}(\langle \hat{\theta}_0(\epsilon), x\rangle, y)$ may exist. In order to resolve ambiguity, we choose the solution of gradient descent on the average loss as our implicit bias. For the logistic and square loss ℓ_{train} the resulting estimator is equivalent to the minimizer of *ridge-regularized* objective

$$\mathcal{L}_{\epsilon,\lambda}(\theta) := \frac{1}{n}\sum_{i=1}^{n} \max_{\delta \in \mathcal{U}_p(\epsilon)} \ell_{\text{train}}(\langle \theta, x_i + \delta\rangle, y_i) + \lambda \|\theta\|_2^2 \tag{3}$$

as $\lambda \to 0$ [34]. Our goal is to compare the interpolating estimators with *regularized* estimators that minimize $\mathcal{L}_{\epsilon,\lambda}$ with $\lambda > 0$ and do not achieve vanishing training loss. Finally, we remark that for linear regression on isotropic Gaussian features, it is well known that ridge regularization corresponds to early stopping [1, 2, 41].

3 Regularization for linear regression

In this section we prove that preventing interpolation on noiseless samples via ridge regularization (3) 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 when it is not beneficial for improving the standard risk.

3.1 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 $y = \langle \theta^\star, x\rangle + \xi$ with $\xi \sim \mathcal{N}(0, \sigma^2)$. Here, $\theta^\star$ 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. We illustrate the phenomenon of robust overfitting for consistent ℓ_2 perturbations sets (2). By Lemma A.2 in the appendix, this is equivalent to minimizing the distributionally robust risk for ℓ_2-bounded mean shifts. Due to rotation invariance of the problem our results hold for sparse and dense vectors alike. We expect robust overfitting also for ℓ_∞ perturbations and different choices of $\mathcal{L}_{\epsilon,\lambda}$ and $\theta^\star$ but leave a detailed discussion for future work.

3.2 Interpolating and regularized estimator

We study the ridge penalized loss for regression (3) that corresponds to $\epsilon = 0$ and choosing ℓ_{train} as the squared loss which yields

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

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 $\lambda \to 0$ and can be written as follows:

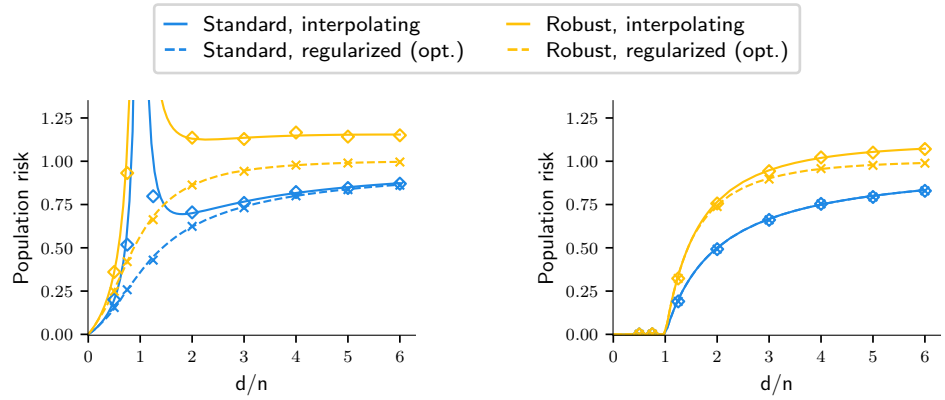
$$\hat{\theta}_0 = \arg\min_\theta \|\theta\|_2 \ \text{ such that } \ \langle \theta, x_i\rangle = y_i \ \text{ for all } i. \tag{5}$$

The minimum norm interpolator is the solution of gradient descent whereas ridge regression with $\lambda > 0$ corresponds to early stopping path [1, 2].

The attentive reader might wonder why we choose $\epsilon = 0$, corresponding to standard training although we study the robust risk. For this we need to note that consistent adversarial training with $\epsilon > 0$ simply recovers the ground truth as it reveals full information during training. On the other hand, adversarial training with inconsistent perturbations has a similar effect to ℓ_2 regularization and in particular prevents the estimator from interpolating (see more detailed discussion in C.1).

3.3 Main result and experiments

The following theorem provides a precise asymptotic description of the consistent ℓ_2 robust risk for the ridge regression estimate in (4) as $d/n \to \gamma$ and $d, n \to \infty$. The proof uses techniques from [17, 23] and can be found in Appendix E.1.



(a) Noisy observations

(b) Noiseless observations

Figure 2: We plot the asymptotic theoretical predictions for $d, n \to \infty$ (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 the optimal λ (dashed, *regularized*) for noisy data with $\sigma^2 = 0.2$ (left) and noiseless data (right). For the experiments we use $n = 1000$ and the model described in Section 3.1. We observe that the gap between the robust risk of the interpolating and optimally regularized estimator remains even when $\sigma^2 = 0$.

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

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

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* $\boldsymbol{R}(\hat{\theta}_\lambda) \xrightarrow{a.s.} \mathcal{B} + \mathcal{V}$ *as shown in Corollary 5 of [17]. The function* $m(z)$ *is 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*. Finally,*

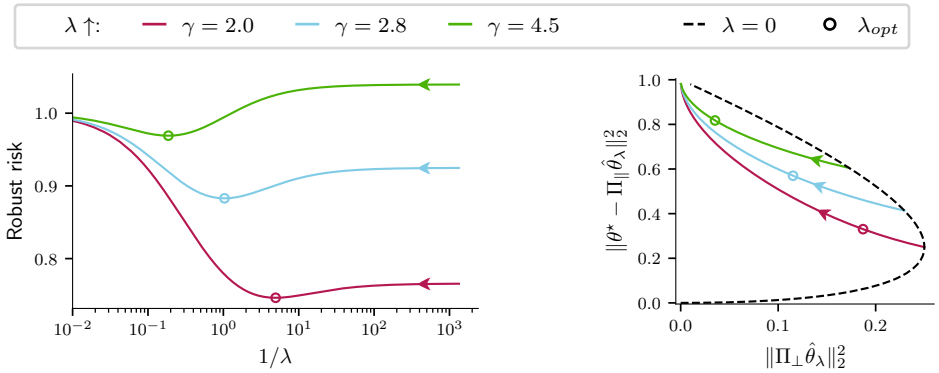
Using the precise asymptotics derived in this theorem, we plot the asymptotic standard and robust risk of the ridge estimate with the optimal λ parameter and the min-norm interpolator in Figure 2.[1] Firstly, Figure 2a reveals that ridge regularization reduces the robust risk even for $d > n$ well beyond the interpolation threshold where previous works show that variance is negligible. 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 $d, n \to \infty$ and $d/n \to \gamma$. 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 matricies following the same argument as in [17] based on random matrix theory [23].

3.4 Intuitive explanations and discussion

We now shed light on the phenomena revealed by Theorem 3.1 and Figure 2. In particular we discuss why regularization 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 $\Pi_\parallel$ on the ground-truth direction $\theta^\star$ and

[1] Here we choose λ using the population risk oracle, in practice one would resort to standard tools such as cross-validation techniques [30]



(a) Robust risk as λ increases

(b) Orthogonal and parallel parameter error

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 overparamterization ratio $d/n \to \gamma$. We see in (b) that for large $\gamma > 1$, as λ increases, the orthogonal error $\|\Pi_\perp \hat{\theta}_\lambda\|_2$ decreases whereas the parallel error $\|\theta^\star - \Pi_\| \hat{\theta}_\lambda\|_2$ increases. For $\epsilon > 0$, the best λ is chosen at a large λ that prevents interpolation.

its projection $\Pi_\perp$ onto its orthogonal complement, i.e., $\hat{\theta}_\lambda = \Pi_\| \hat{\theta}_\lambda + \Pi_\perp \hat{\theta}_\lambda$. For the noiseless setup ($\sigma^2 = 0$), the closed-form expression of the consistent robust risk involves the parallel error $\|\theta^\star - \Pi_\| \hat{\theta}_\lambda\|_2^2$ and the orthogonal error $\|\Pi_\perp \hat{\theta}_\lambda\|_2^2$ as follows:

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

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 Regularization for logistic regression

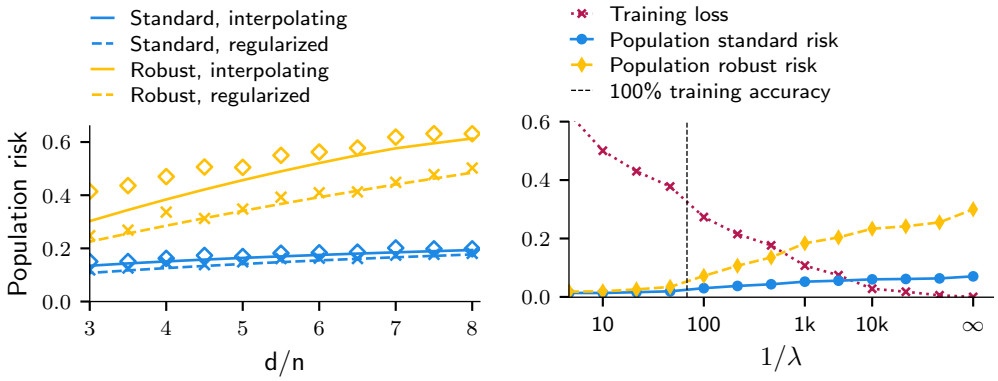
In this section, we show that the max-margin solution (corresponding to $\lambda \to 0$ in $\mathcal{L}_{\epsilon,\lambda}$ (2) for the logistic loss with ℓ_∞-attacks) yields worse generalization than when choosing $\lambda > 0$. We derive precise asymptotic predictions that show a systematic benefit for the standard and robust risk under ℓ_2 regularization when $d, n \to \infty$ and $d/n \to \gamma$. We explain this phenomenon by arguing that unexpectedly, for the ℓ_∞ robust logistic loss, a large ℓ_2 penalty in fact enhances the bias of the estimator towards approximately sparse solutions. As a consequence, for a sparse ground truth, both the robust and standard risk benefit from a non-zero regularization parameter.

4.1 Problem setting

We consider a discriminative data-generating model for classification with isotropic Gaussian features $x \sim \mathcal{N}(0, I_d)$. Since the focus of this paper is on interpolation with noiseless data, we observe deterministic ground truth labels $\operatorname{sgn}\langle \theta^\star, x\rangle$ for all covariates $x \in \mathbb{R}^d$ and fixed vector $\theta^\star$ in this section. We refer to Appendix D.3, F for results under observation noise.

We illustrate the phenomenon of robust overfitting for consistent ℓ_∞ perturbation sets $\mathcal{U}_\infty(\epsilon)$, a standard choice in the experimental and theoretical classification literature [14, 19, 32, 38]. Similar to linear regression, the adversarially robust risk computed with such perturbation sets can be shown to upper bound the distributionally robust risk with respect to mean shifts (see Appendix A.3) so that adversarial robustness implies distributional robustness.

We consider the scenario where the ground truth has inherent low-dimensional structure in the form of sparsity $\theta^\star = (1, 0, \ldots, 0)^\top$. As we discuss later in this section, ℓ_∞ robustness and sparsity via its



(a) Fixed n, d vs asymptotic predicitons

(b) Risks of $\hat{\theta}_\lambda(\epsilon)$

Figure 4: (a) The markers show the robust and standard risks for empirical training using $n = 1000$ and $\epsilon = 0.05$ when interpolating ($\lambda = 0$) and regularizing ($\lambda = 1$). Additionally, the lines display the corresponding asymptotic theoretical predictions from Theorem 4.1. (b) The training losses, robust risks and standard risks of empirical simulations with $\epsilon = 1$ and $n = 1000$, plottet with respect to increasing $1/\lambda$. Both figures demonstrate that regularization improves the robust and standard risk. See B for experimental details.

convex surrogate, the ℓ_1 norm, are related objectives. Nonetheless, we show that the inductive bias of the ℓ_∞ robust logistic loss with $\lambda \to 0$ (that is the robust max-margin solution) is imperfect and that alternatively choosing a non-zero ℓ_2 penalty surprisingly promotes better sparsity and generalization.

4.2 Interpolating and regularized estimators

For training, we consider the common ℓ_2-regularized robust logistic regression estimate that corresponds to Equation (3) by choosing ℓ_{train} to be the logistic loss:

$$\hat{\theta}_\lambda(\epsilon) := \arg\min_\theta \frac{1}{n} \sum_{i=1}^{n} \max_{\delta \in \mathcal{U}_\infty(\epsilon)} \log(1 + e^{-\langle \theta, x_i + \delta \rangle y_i}) + \lambda \|\theta\|_2^2. \tag{8}$$

As $\lambda \to 0$, we obtain the max-margin estimator [34] on consistent perturbation sets of the training data

$$\hat{\theta}_0(\epsilon) := \arg\min_\theta \|\theta\|_2 \ \text{ such that } \ y_i \langle \theta, x_i \rangle - \epsilon \|\Pi_\perp \theta\|_1 \geq 1 \ \text{ for all } i, \tag{9}$$

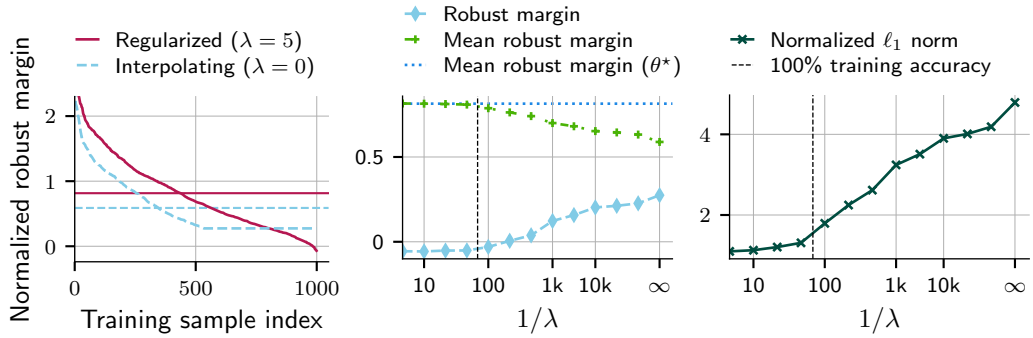
where $\Pi_\perp$ is the projection onto the orthogonal complement of $\theta^\star$. Similar to linear regression, this estimator is of particular interest since it corresponds to the gradient descent solution of the unregularized loss with $\lambda = 0$ (see for example [22, 40], [19]).

We recall that we focus on noiseless training to disentangle the effect of noise from finite-sample effects. This also motivates why we not only evaluate with consistent attacks but also train with consistent adversarial perturbation sets: by definition, for points close to the true decision boundary (which are likely in our data model), ℓ_p perturbed points parallel to $\theta^\star$ cross the true decision boundary. Hence, the ground truth does not achieve 100% robust accuracy on the training dataset. Nonetheless, we also include a discussion on inconsistent adversarial training in Appendix D.1 for completeness and show that it behaves similarly to training with consistent perturbation sets.

4.3 Main results and experiments

We now present an informal statement describing the asymptotic standard and robust risk for linear classification with the logistic loss as $d, n \to \infty$. We refer to Appendix F for the precise statement and the proof. The latter is inspired by the works [19, 37] and uses the Convex Gaussian Minimax Theorem (CGMT) [15, 43].

Theorem 4.1 (Informal)**.** *Assume that $\epsilon = \epsilon_0/\sqrt{d}$ for some constant ϵ_0 independent of n, d and $\theta^\star = (1, 0, \cdots, 0)$. Then, the robust and standard risks of the ridge regularized estimator $\hat{\theta}_\lambda(\epsilon)$ with*



(a) Robust margin decay of $\hat{\theta}_\lambda(\epsilon)$ (b) Robust margin of $\hat{\theta}_\lambda(\epsilon)$ (c) Sparsity of $\hat{\theta}_\lambda(\epsilon)$

Figure 5: (a) The ordered sample-wise penalized margin $\langle x_i, \theta\rangle - \epsilon\|\Pi_\perp \theta\|_1$ of the normalized estimator $\hat{\theta}/\|\hat{\theta}\|_2$ when interpolating and regularizing with $\lambda = 5$. For larger λ, the minimum margin is decreases while the average margin (horizontal lines) decreases. This trendis also observed in (b) where the minimum penalized margin (robust margin) and the average penalized margin as in (a) are plotted with respect to $1/\lambda$. The horizontal line is the corresponding average penalized margin of $\theta^\star$. (c) The ℓ_1 norm of the normalized estimator increasing with $1/\lambda$. See B for experimental details.

$\lambda \geq 0$ *converge in probability as* $d, n \to \infty$, $d/n \to \gamma$

$$\boldsymbol{R}(\hat{\theta}_\lambda(\epsilon)) \to \frac{1}{\pi}\arccos\left(\frac{\nu_\parallel^\star}{\nu^\star}\right) \;,\; \boldsymbol{R}_\epsilon(\hat{\theta}_\lambda(\epsilon)) \to \frac{1}{\pi}\arccos\left(\frac{\nu_\parallel^\star}{\nu^\star}\right) + \frac{1}{2}\mathrm{erf}\left(\frac{\epsilon_0\delta^\star}{\sqrt{2}\nu^\star}\right) + I\left(\frac{\epsilon_0\delta^\star}{\nu^\star}, \frac{\nu_\parallel^\star}{\nu^\star}\right)$$

with erf *the error function,*

$$I(t,u) := \int_0^t \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)\mathrm{erf}\left(\frac{xu}{\sqrt{2(1-u^2)}}\right)dx$$

and where $\nu^\star = \sqrt{(\nu_\perp^\star)^2 + (\nu_\parallel^\star)^2}$, *and* $\nu_\perp^\star, \nu_\parallel^\star, \delta^\star$ *are the unique solution of a scalar optimization problem specified in Appendix F that depends on* $\theta^\star, \gamma, \epsilon_0$ *and* λ.

We illustrate the consequences of this theorem in Figure 4a where we plot the standard and robust risk obtained by the asymptotic theoretical predictions for $d, n \to \infty$ and simulations for $n = 1000$ and $\epsilon = 0.1$. The theoretical curves are obtained by solving the scalar optimization problem specified in Appendix F; the empirical setup is described in Appendix B. We observe that the precise asymptotics indeed predict benefits of regularization for adversarial logistic regression. This trend is further supported by our simulation results closely following the theoretical curves. Even though the standard risk also benefits from regularization in our setting, the gains for the robust risk are more significant and increase with the overparameterization ratio.

Finally, in Figure 4b we observe for a fixed overparameterization ratio d/n how optimizing the average adversarial logistic loss beyond 100% training accuracy substantially hurts generalization. In particular, this challenges the common belief that training overparameterized classifiers even beyond the point of achieving 100% training accuracy improves generalization. Even though the theory and simulations in this section are for varying λ, we show in Appendix D.2 that the same trends as for $1/\lambda \to \infty$ can be observed epoch-wise when training with gradient descent as the epochs $t \to \infty$. We leave the proof for the benefits of early stopping as interesting future work.

4.4 Why regularization improves regularization

While Theorem 4.1 specifies the precise asymptotic behavior of the standard and robust risks and predicts the benefits of regularization for generalization, it is hard to extract intuitive insights directly from the proof. Instead, in this section we provide intuition for why a non-zero ridge penalty induces a sparser $\hat{\theta}_\lambda(\epsilon)$ than the ℓ_1 penalized max-margin solution $\hat{\theta}_0(\epsilon)$ and use simulations to support our claims.

As shown in Appendix F, the unregularized and consistent adversarial robust loss $\mathcal{L}_{\epsilon,\lambda}$ with ℓ_∞ perturbations can be written as $\mathcal{L}_\epsilon(\theta) = \frac{1}{n}\sum_{i=1}^n \log(1 + e^{-y_i\langle\theta, x_i\rangle + \epsilon\|\Pi_\perp\theta\|_1})$. Hence, adversarial

training with ℓ_∞ perturbations penalize solutions with large ℓ_1-norm, which in turn is well known to encourage sparse solutions (see [44] and references therein). However, in contrast to an explicit additive ℓ_1 penalty, the ℓ_1 term appears inside the loss function. As illustrated in Figure 5c, this leads to the effect that additionally regularizing with an ℓ_2 penalty leads to a smaller normalized ℓ_1 norm and hence a sparser solution.

To gain intuition for this phenomenon, we consider two cases: large λ inducing small $\|\hat{\theta}\|_2$ and small λ inducing large $\|\hat{\theta}\|_2$ solutions. As we argue below and illustrate in Figure 5a, the former corresponds to large-average margin solutions whereas the latter approaches a large-min margin solution for small enough λ. Since the ground truth $\theta^\star$ is sparse and has a small min-margin and large-average margin (horizontal dotted line in Figure 5b) itself, the large-average margin solution is likely to align better with the ground truth. Hence, we expect it to be sparse and have a low standard and robust risk. The improved sparsity is indeed observed in Figure 5c and the corresponding standard risk decrease is plotted in Figure 4b.

We now reason why large $\|\hat{\theta}\|_2$ leads to the max-min margin solution. Due to the logistic loss, most training points have large values $y_i\langle\hat{\theta}, x_i\rangle$ for large $\|\hat{\theta}\|_2$ and do not contribute to the loss. In contrast, small $\|\hat{\theta}\|_2$ lead to a large average-margin solution since *all training points* contribute to the loss, most easily seen via the Taylor series expansion $\mathcal{L}_\epsilon(\theta) \approx -\frac{1}{n}\sum_{i=1}^n y_i\langle x_i, \theta\rangle + \epsilon\|\Pi_\perp\theta\|_1$.

5 Related work

Understanding robust overfitting The current literature attempting to explain robust overfitting [32] primarily focuses on the effect of noise and the smoothness of decision boundaries learned by neural networks trained to convergence [12, 38, 46]. A slightly different line of work [35] shows that regularization is needed in order for overparameterized models to achieve good classification accuracy on all subpopulations. However, a theoretical understanding of the role of regularization is still missing.

Theory for adversarial robustness Recent works [20, 19] provide a precise description of the robust risks for logistic and linear regression when trained with adversarial attacks based on the Convex Gaussian Minimax Theorem [15, 43]. The analysis focuses on inconsistent attacks for both evaluation and training. For linear regression, the authors observe that adversarial training with ℓ_2 perturbations mitigates the peak in the double descent curve around $d = n$ and hence acts similarly to ridge regularization as explicitly studied in the paper [17].

6 Conclusions

We show that overparameterized linear models can overfit with respect to the robust risk even when there is no noise in the training dataset. 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. Perhaps surprisingly, we further observe that ridge regularization enhances the bias of logistic regression trained with adversarial ℓ_∞ attacks towards sparser solutions. Thus, ridge regularization can have a positive effect on generalization for high-dimensional linear models going beyond variance reduction. While our theoretical results hold for the ridge penalty, we experimentally observe the same behavior for early stopping gradient descent. While our theoretical results apply for ridge regularization, we also observe the same behavior for early stopping. A formal proof could reveal important insights to better understand the phenomenon and we leave this as future work. Furthermore, it would be interesting to study whether our observations also transfer to random feature and kernel regression.

References

[1] A. Ali, J. Z. Kolter, and R. J. Tibshirani. A continuous-time view of early stopping for least squares, 2019.

[2] A. Ali, E. Dobriban, and R. J. Tibshirani. The implicit regularization of stochastic gradient flow for least squares, 2020.

[3] M. ApS. *MOSEK Optimizer API for Python 9.2.45*, 2021. URL `https://docs.mosek.com/9.2/pythonapi/index.html`.

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

[5] M. Belkin, S. Ma, and S. Mandal. To understand deep learning we need to understand kernel learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 80, pages 541–549, 2018.

[6] P. Bühlmann and S. Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

[7] Y. Chen and P. Bühlmann. Domain adaptation under structural causal models, 2020.

[8] L. Chizat and F. Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.

[9] L. Chizat and F. Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss, 2020.

[10] Z. Deng, A. Kammoun, and C. Thrampoulidis. A model of double descent for high-dimensional binary linear classification, 2020.

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

[12] C. Dong, L. Liu, and J. Shang. Data profiling for adversarial training: On the ruin of problematic data, 2021.

[13] B. Ghorbani, S. Mei, T. Misiakiewicz, and A. Montanari. Linearized two-layers neural networks in high dimension. *arXiv preprint arXiv:1904.12191*, 2019.

[14] I. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

[15] Y. Gordon. On milman's inequality and random subspaces which escape through a mesh in rn. In *Geometric aspects of functional analysis*, pages 84–106. Springer, 1988.

[16] S. Gunasekar, J. D. Lee, D. Soudry, and N. Srebro. Implicit bias of gradient descent on linear convolutional networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.

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

[18] A. Javanmard and M. Soltanolkotabi. Precise statistical analysis of classification accuracies for adversarial training, 2020.

[19] A. Javanmard and M. Soltanolkotabi. Precise statistical analysis of classification accuracies for adversarial training, 2020.

[20] A. Javanmard, M. Soltanolkotabi, and H. Hassani. Precise tradeoffs in adversarial training for linear regression. In *Proceedings of the Conference on Learning Theory (COLT)*, pages 2034–2078, 2020.

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

[22] Z. Ji and M. Telgarsky. Risk and parameter convergence of logistic regression, 2019.

[23] A. Knowles and J. Yin. Anisotropic local laws for random matrices, 2016.

[24] D. Kobak, J. Lomond, and B. Sanchez. The optimal ridge penalty for real-world high-dimensional data can be zero or negative due to the implicit ridge regularization. *Journal of Machine Learning Research*, 21(169):1–16, 2020.

[25] K. Lyu and J. Li. Gradient descent maximizes the margin of homogeneous neural networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.

[26] S. Mei and A. Montanari. The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv preprint arXiv:1908.05355*, 2019.

[27] V. Muthukumar, A. Narang, V. Subramanian, M. Belkin, D. Hsu, and A. Sahai. Classification vs regression in overparameterized regimes: Does the loss function matter? *arXiv preprint arXiv:2005.08054*, 2020.

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

[29] P. Nakkiran, G. Kaplun, Y. Bansal, T. Yang, B. Barak, and I. Sutskever. Deep double descent: Where bigger models and more data hurt. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.

[30] P. Patil, Y. Wei, A. Rinaldo, and R. Tibshirani. Uniform consistency of cross-validation estimators for high-dimensional ridge regression. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 3178–3186, 2021.

[31] A. Raghunathan, S. M. Xie, F. Yang, J. C. Duchi, and P. Liang. Adversarial training can hurt generalization. *arXiv preprint arXiv:1906.06032*, 2019.

[32] L. Rice, E. Wong, and J. Z. Kolter. Overfitting in adversarially robust deep learning, 2020.

[33] D. Richards, J. Mourtada, and L. Rosasco. Asymptotics of ridge (less) regression under general source condition. In *International Conference on Artificial Intelligence and Statistics*, pages 3889–3897. PMLR, 2021.

[34] S. Rosset, J. Zhu, and T. Hastie. Margin maximizing loss functions. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems*, volume 16. MIT Press, 2004.

[35] S. Sagawa*, P. W. Koh*, T. B. Hashimoto, and P. Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=ryxGuJrFvS`.

[36] S. Sagawa, A. Raghunathan, P. W. Koh, and P. 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*, pages 8346–8356, 13–18 Jul 2020.

[37] F. Salehi, E. Abbasi, and B. Hassibi. The impact of regularization on high-dimensional logistic regression, 2019.

[38] A. Sanyal, P. K. Dokania, V. Kanade, and P. Torr. How benign is benign overfitting? In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[39] A. Sinha, H. Namkoong, and J. Duchi. Certifying some distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018.

[40] D. Soudry, E. Hoffer, M. S. Nacson, S. Gunasekar, and N. Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70):1–57, 2018.

[41] A. Suggala, A. Prasad, and P. K. Ravikumar. Connecting optimization and regularization paths. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.

[42] P. Sur and E. J. Candès. A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116(29):14516–14525, 2019.

[43] C. Thrampoulidis, S. Oymak, and B. Hassibi. Regularized linear regression: A precise analysis of the estimation error. In *Proceedings of the Conference on Learning Theory (COLT)*, pages 1683–1709. PMLR, 2015.

[44] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society (Series B)*, 58:267–288, 1996.

[45] M. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019. ISBN 9781108498029.

[46] B. Wu, J. Chen, D. Cai, X. He, and Q. Gu. Do wider neural networks really help adversarial robustness?, 2021.

[47] D. Wu and J. Xu. On the optimal weighted ℓ_2 regularization in overparameterized linear regression. *arXiv preprint arXiv:2006.05800*, 2020.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] We specified the models for which our results hold.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] See Section E,F
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See Section B.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section B.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section B.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes] See Section B.
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]