The Implicit Bias of Benign Overfitting

Document Version:
Publisher's PDF, also known as Version of record

Citation for published version:
Shamir, O 2023, 'The Implicit Bias of Benign Overfitting', *Journal of Machine Learning Research*, vol. 24.
<https://arxiv.org/abs/2201.11489>

Total number of authors:
1

Published In:
Journal of Machine Learning Research

How does open access to this work benefit you?
Let us know @ library@weizmann.ac.il

(article begins on next page)

The Implicit Bias of Benign Overfitting

Ohad Shamir   OHAD.SHAMIR@WEIZMANN.AC.IL
Department of Computer Science and Applied Mathematics
Weizmann Institute of Science
Rehovot, Israel

Editor: Daniel Hsu

Abstract

The phenomenon of benign overfitting, where a predictor perfectly fits noisy training data while attaining near-optimal expected loss, has received much attention in recent years, but still remains not fully understood beyond well-specified linear regression setups. In this paper, we provide several new results on when one can or cannot expect benign overfitting to occur, for both regression and classification tasks. We consider a prototypical and rather generic data model for benign overfitting of linear predictors, where an arbitrary input distribution of some fixed dimension k is concatenated with a high-dimensional distribution. For linear regression which is not necessarily well-specified, we show that the minimum-norm interpolating predictor (that standard training methods converge to) is biased towards an inconsistent solution in general, hence benign overfitting will generally *not* occur. Moreover, we show how this can be extended beyond standard linear regression, by an argument proving how the existence of benign overfitting on some regression problems precludes its existence on other regression problems. We then turn to classification problems, and show that the situation there is much more favorable. Specifically, we prove that the max-margin predictor (to which standard training methods are known to converge in direction) is asymptotically biased towards minimizing a weighted *squared hinge loss*. This allows us to reduce the question of benign overfitting in classification to the simpler question of whether this loss is a good surrogate for the misclassification error, and use it to show benign overfitting in some new settings.

Keywords: benign overfitting, interpolating predictors, implicit bias, surrogate losses

1. Introduction

The ability of learning algorithms to succeed despite overfitting is a curious phenomenon in statistical learning, which has received much interest in the past few years. A particular version of it, commonly denoted as "benign overfitting" (Bartlett et al., 2020), refers to situations which combine the following: (1) The trained predictor interpolates the data, in the sense that it achieves perfect prediction accuracy on the training data; (2) No predictor in the relevant hypothesis class can achieve perfect accuracy w.r.t. the underlying data distribution; yet (3) The trained predictor has near-optimal accuracy w.r.t. the underlying data distribution. The combination of (1) and (2) implies that the predictor overfits (in the sense that the training error is significantly smaller than the test error), and (3) implies that this overfitting is "benign" (in the simple sense that the predictor still performs well). This phenomenon is intriguing, because some version of it appears to occur frequently in large-scale learning problems, yet it cannot be easily explained using standard learning theoretic

tools such as uniform convergence (which requires the performance on the training data and the underlying distribution to be similar). This has led to a flurry of papers in the past few years, attempting to understand why and when benign overfitting occurs, and whether uniform convergence can or cannot explain its occurence (see discussion of related work below).

So far, most of the theoretical work on benign overfitting has focused on linear (or kernel) regression problems using the square loss, with some works extending this to classification problems. The relatively most well-understood situation is plain linear regression in a well-specified setting, where we are training a linear predictor $\mathbf{x} \mapsto \mathbf{x}^\top \mathbf{w}$, $\mathbf{w} \in \mathbb{R}^d$ with respect to the square loss, and when the outputs y satisfy $y = \mathbf{x}^\top \mathbf{w}^* + \xi$ for some $\mathbf{w}^* \in \mathbb{R}^d$, where ξ is zero-mean noise. In this setting, we know that benign overfitting occurs (roughly speaking) whenever the dimension of the input $\mathbf{x}$ is sufficiently large, and the distribution has many directions of small (but non-zero) variance that the trained predictor can utilize to perfectly fit the training data, without significantly affecting the distribution of the predictions on new samples. A helpful feature of this setting is that there is a closed-form expression for the predictor returned by standard gradient-based methods, when trained to convergence on the average square loss – namely, the minimum-norm interpolating predictor. For classification, the situation is more complicated, because the predictor that standard gradient-based methods converge to with appropriate losses (namely, the max-margin predictor) does not have a closed-form expression in general. Thus, most recent works on classification focused on more specific setups (as discussed in the related work section below).

In this paper, we provide several new results on benign overfitting, which can be used to analyze when benign overfitting may – or may not – occur in settings beyond those studied so far in the literature. We focus on a prototypical data model for benign overfitting, where for some fixed integer k, the input distribution in $\mathbb{R}^d$ is composed of an arbitrary distribution on the first k coordinates, and a *high-dimensional* distribution on the last $d - k$ coordinates (in the sense that i.i.d. samples from that distribution are approximately orthogonal). As we discuss later on, such a setting is essentially necessary for benign overfitting to occur, even for well-specified linear regression. In this setting, we study benign overfitting by considering the asymptotic consistency of the predictor learned by standard gradient-based methods, as both d and the sample size m diverge to infinity, and d increases sufficiently faster than m. Our results can be informally summarized as follows:

- Beginning with linear regression, we show that once the distribution is not well-specified, the minimum-norm interpolating predictor (a.k.a. min-norm predictor) returned by standard training methods will generally *not* be consistent, and hence benign overfitting will generally not occur. A bit more concretely, in our data model a consistent predictor $\hat{\mathbf{w}}$ should be such that asymptotically,

$$\hat{\mathbf{w}}_{|k} \approx \left(\mathbb{E}\left[\mathbf{x}\mathbf{x}^\top \right]^{-1} \right)_k \mathbb{E}[y\mathbf{x}] \,, \tag{1}$$

where $\mathbf{v}_{|k}$ of a vector $\mathbf{v}$ denote its first k coordinates, and $\left(\mathbb{E}\left[\mathbf{x}\mathbf{x}^\top \right]^{-1} \right)_k$ refers to the first k rows of the inverse covariance matrix. In other words, we expect the first k coordinates

to converge to the first k coordinates of the least-squares solution w.r.t. the underlying distribution. In contrast, we prove that the min-norm predictor asymptotically satisfies

$$\hat{\mathbf{w}}_{|k} \approx \mathbb{E}\left[\frac{\mathbf{x}_{|k}\mathbf{x}_{|k}^\top}{\|\mathbf{x}_{|d-k}\|^2}\right]^{-1} \mathbb{E}\left[\frac{y\mathbf{x}_{|k}}{\|\mathbf{x}_{|d-k}\|^2}\right] , \tag{2}$$

where $\mathbf{x}_{|d-k}$ are the last $d-k$ coordinates of $\mathbf{x}$. Clearly, the expressions in Eq. (1) and Eq. (2) are generally not the same, except in some special cases (an important one, as we shall see, being a well-specified setting). Thus, we argue that one should not expect benign overfitting to generally occur in linear regression with the square loss.

- We show that benign overfitting will generally not occur in some natural extensions of linear regression with the square loss, even in a well-specified setting. These include (1) regression with generalized linear models (or equivalently, with a single neuron predictor); and (2) regression with losses other than the square loss. To do so, we present a new observation that may be of independent interest. Roughly speaking, we argue that any interpolating predictor can be seen as returning the optimum of the average loss, but simultaneously, it is also the optimum of many other types of average loss objectives, which reflect rather different learning problems with different optimal predictors. The trained predictor does not "know" which of these learning problems it is actually solving, so benign overfitting (with the trained predictor achieving low expected loss) can only occur in some of them. As a result, benign overfitting is implicitly "biased" towards certain learning problems, and its occurence in one problem precludes its occurence in another. We note that this is somewhat reminiscent of the paper Muthukumar et al. (2021), which pointed out that interpolating predictors can be insensitive to the type of loss function used for training, but we take this in a rather different direction.

- Having discussed regression problems, we turn to binary classification problems (where our goal is to minimize misclassification error), and show that the situation there is rather different. Concretely, we consider linear predictors under the same data model as before, and the max-margin predictor $\hat{\mathbf{w}}$ (to which gradient-based methods are known to asymptotically converge in direction). Perhaps surprisingly, we prove that $\hat{\mathbf{w}}$ has a rather clean asymptotic characterization: Its first k coordinates asymptotically minimize the expectation of a (weighted) *squared hinge loss* on those coordinates,

$$\hat{\mathbf{w}}_{|k} \approx \arg\min_{\mathbf{v}\in\mathbb{R}^k} \mathbb{E}\left[\frac{[1-y\mathbf{x}_{|k}^\top\mathbf{v}]_+^2}{\|\mathbf{x}_{|d-k}\|^2}\right] ,$$

(where $[z]_+ := \max\{0, z\}$), and the last $d-k$ coordinates of $\hat{\mathbf{w}}$ are asymptotically immaterial. Thus, we get that the existence of benign overfitting in our model is reduced to a simpler question: Whether the data distribution is such that minimizing the expectation of this weighted squared hinge loss is a good surrogate for minimizing misclassification error. Although not true in the worst case, it is reasonable to assume that it will be true in many cases, because this loss still encourages the sign of $\mathbf{x}_{|k}^\top\hat{\mathbf{w}}_k$ to accord with the output y. This is in contrast to regression, in which our results suggest that benign overfitting is more brittle. Also, we note that unlike many previous works on benign overfitting in

classification, our result does not require that the distribution is such that the min-norm and max-margin predictors coincide. Based on this result, we study more specifically the case of linearly separable distributions with label noise, and provide a few positive results: For example, for just about any choice of distribution on the first k coordinates, we will have benign overfitting at least for some positive amount of label noise. Moreover, under some stronger assumptions on the input distribution (e.g., a mixture of symmetric distributions), we will have benign overfitting for any non-trivial level of label noise.

Overall, we hope that the results provided in this paper will allow us to understand the phenomenon of benign overfitting beyond the settings studied so far.

The paper is structured as follows: After discussing related work below, we study linear regression with the square loss in Section 3, and other regression settings in Section 4. We then turn to classification problems in Section 5, and conclude with a discussion in Section 6. The formal proofs of all our results appear in Appendix A.

1.1 Related Work

Papers such as (Zhang et al., 2017) popularized the notion that modern learning systems (such as deep learning) tend to perfectly fit the training data, while still performing well on test data. The literature on the theory of this phenomenon is by now very large, and we will only discuss here the papers most relevant to our work (see for example Belkin (2021) for a more comprehensive survey).

A line of works (e.g., Belkin et al. (2018a,b,b, 2019b); Mei and Montanari (2019); Liang and Rakhlin (2020); Belkin et al. (2019a)) showed that this phenomenon is not reserved to deep learning, and occurs already in linear and kernel learning. More recently, papers such as Bartlett et al. (2020); Hastie et al. (2019); Belkin et al. (2020) studied conditions for benign overfitting in linear regression with the square loss in a well-specified setting. In particular, Bartlett et al. (2020) considered general distributions, and showed how the occurrence of benign overfitting can be characterized in terms of the eigenvalues of the input covariance matrix, and how having many low-variance directions is in some sense necessary for benign overfitting to occur. Koehler et al. (2021); Zhou et al. (2021) showed how these results can be recovered and extended from a uniform convergence perspective, again for well-specified linear regression. Other works which study benign overfitting and its relationship to classical learning theory include Nagarajan and Kolter (2019); Negrea et al. (2020); Yang et al. (2021); Bartlett and Long (2021); Bachmann et al. (2021); Koehler et al. (2021); Zhou et al. (2020); Muthukumar et al. (2021).

Understanding benign overfitting in classification problems has been more challenging, since the max-margin predictor to which gradient-based methods are known to converge to (in direction) does not have a closed-form solution. Many of the existing works focus on settings where the max-margin predictor and the (closed-form) least squares predictor coincide (as originally argued in Muthukumar et al. (2021)). Wang and Thrampoulidis (2021) and Cao et al. (2021) use this to study a setting where the two classes are a symmetric mixture of Gaussian (or subgaussian) distributions, without label noise. Chatterji and Long (2021) studies a setting where the two classes are a mixture of two product distributions, and with label noise, by studying the trajectory of gradient descent on the training data. Montanari et al. (2019) considers classification problem where the inputs are Gaussian, and

the labels are generated according to a logistic link function, and derives a formula for the asymptotic prediction error of the max-margin classifier, in a setting where the ratio of the dimension and the sample size converges to some fixed positive limit. Recently, Frei et al. (2022) managed to prove the existence of benign overfitting in two-layer neural networks with smoothed leaky ReLU activations, assuming the data comes from a mixture of two well-separated distributions. Other works studying benign overfitting and classification include Liang and Recht (2021); McRae et al. (2021); Poggio and Liao (2019); Thrampoulidis (2020); Hu et al. (2021); Wang et al. (2021).

2. Preliminaries

Notation. We use bold-faced letter to denote vectors, and assume that they are in column form unless specified otherwise. Given a vector $\mathbf{w} \in \mathbb{R}^d$, we let w_i denote its i-th coordinate, $\mathbf{w}_{|k} \in \mathbb{R}^k$ to denote its first k coordinates, and $\mathbf{w}_{|d-k} \in \mathbb{R}^{d-k}$ to denote its last $d-k$ coordinates. We also use this notation when the vector already has a subscript for a different purpose, e.g. $\mathbf{x}_{i|k}$ refers to the first k coordinates of $\mathbf{x}_i$. Given two vectors $\mathbf{u}, \mathbf{v}$ of the same size, $\mathbf{u} \succeq \mathbf{v}$ means that $u_i \geq v_i$ for all i. We use $[\cdot]_+$ to denote the ReLU function $z \mapsto \max\{0, z\}$. Given a vector $\mathbf{v}$, $[\mathbf{v}]_+$ refers to applying the ReLU function entry-wise. $\mathbf{1}(\cdot)$ to denote the indicator function. $[m]$ is shorthand for $\{1, \ldots, m\}$. I denotes the identity matrix (whose size should be clear from context). Given a matrix A, $A_{i,j}$ refers to its entry at row i column j, $\lambda_{\min}(A)$ refers to its minimal eigenvalue, $\text{Tr}(A)$ to its trace, $\|A\|$ refers to its spectral norm, and $\|A\|_F$ refers to the Frobenius norm. It is well known that $\|A\| \leq \|A\|_F$.

Convergence in probability and the law of large numbers. For the asymptotic results in our paper, we will often consider a sequence of real-valued random variables $\{X_d\}$ indexed by d, and say that they converge in probability to some fixed number a (or $X_d \xrightarrow{P} a$) if $\Pr(|X_d - a| \geq \epsilon) \xrightarrow{d \to \infty} 0$ for all $\epsilon \geq 0$. Note that this slightly extends the usual notion of convergence in probability, in that we do not require the random variables to share the same probability space. In addition, we will often utilize the following version of the (weak) law of large numbers for such sequences:

Lemma 1 *If $\sup_d \mathbb{E}[X_d^2] < \infty$, then for any sequence of positive integers $\{m_d\}_{d=1}^\infty$ diverging to ∞, if we let $\{X_{d,i}\}_{i=1}^{m_d}$ be m_d i.i.d. copies of X_d, then $\frac{1}{m_d} \sum_{i=1}^{m_d} X_{d,i} - \mathbb{E}[X_d] \xrightarrow{P} 0$.*

The proof immediately follows from Chebyshev's inequality, which implies that since $\mathbb{E}[X_d^2] < \sigma^2$ for some finite σ^2 and any d, then $\Pr\left(\left|\frac{1}{m_d} \sum_{i=1}^{m_d} X_{d,i} - \mathbb{E}[X_d]\right| \geq \epsilon\right) \leq \frac{\sigma^2}{m_d^2 \epsilon^2} \xrightarrow{d \to \infty} 0$ for all $\epsilon > 0$. This law of large numbers trivially extends to vector-valued and matrix-valued random variables of some fixed dimension k.

The min-norm predictor. In linear regression problems with the square loss, we attempt to find a predictor $\mathbf{w}$ minimizing the expected squared loss $\mathbb{E}_{(\mathbf{x},y)}[(\mathbf{x}^\top \mathbf{w} - y)^2]$. It is well-known that if we attempt this by running standard gradient-based methods on the empirical risk objective $\frac{1}{m} \sum_{i=1}^m (\mathbf{x}_i^\top \mathbf{w} - y_i)^2$ (for some training sample $\{(\mathbf{x}_i, y_i)\}_{i=1}^m$), and the dimension is at least m, these methods will generally converge to the (unique) point

$$\hat{\mathbf{w}} := \arg \min_{\mathbf{w} \in \mathbb{R}^d} \|\mathbf{w}\| : \frac{1}{m} \sum_{i=1}^m (\mathbf{x}_i^\top \mathbf{w} - y_i)^2 = 0$$

(e.g., Zhang et al. (2021)). This predictor is also referred to as the minimum-norm interpolating predictor, or simply the min-norm predictor. For our purposes, we will need a somewhat more general result. Concretely, instead of the square loss specifically, suppose we have some loss function $\ell(\mathbf{x}^\top \mathbf{w}; y)$, which for any y is minimized at some unique prediction value $\mathbf{x}^\top \mathbf{w}$. As before, suppose that we attempt to minimize the expected loss by running a standard gradient-based method over the empirical risk objective, which is now $\frac{1}{m} \sum_{i=1}^m \ell(\mathbf{x}_i^\top \mathbf{w}; y_i)$, assuming we converge to a globally minimal solution $\hat{\mathbf{w}}$. The following proposition shows that under a mild assumption, $\hat{\mathbf{w}}$ is still the minimum-norm interpolating predictor:

Proposition 2 *Fix a function $L(\mathbf{w}) = \frac{1}{m} \sum_{i=1}^m \ell(\mathbf{x}_i^\top \mathbf{w}; y_i)$, where each $\ell(\cdot; y_i)$ is a non-negative continuous function which equals 0 at some unique point denoted as $\ell_{y_i}^{-1}(0)$. Suppose we run an arbitrary iterative training method, that converges to a point $\hat{\mathbf{w}}$ such that $L(\hat{\mathbf{w}}) = 0$ and $\hat{\mathbf{w}} \in span\{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$. Then $\hat{\mathbf{w}}$ is the unique point in $\arg\min_{\mathbf{w}} \|\mathbf{w}\| : L(\mathbf{w}) = 0$.*

Since gradient-based methods rely on iterative updates along the gradient of $L(\cdot)$ (or gradients of single loss functions $\ell(\mathbf{x}_i^\top \mathbf{w}; y_i)$), they generally remain in $span\{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ assuming we initialize at $\mathbf{0}$, and thus the theorem implies that such methods will converge to minimum-norm solutions that minimize the empirical risk.

Benign overfitting for regression. For linear prediction problems, benign overfitting is inherently a high-dimensional phenomenon (since when the dimension is fixed, uniform convergence generally occurs). Thus, the most appropriate way to study benign overfitting is to consider a *sequence* of input distributions over $\mathbb{R}^d$ (indexed by d), and study the performance of the learned predictors as both d and the training set size diverge to infinity. Concretely, for the setting studied in the previous paragraph, a standard way to formally define benign overfitting is as follows:

Definition 3 (Benign Overfitting for the min-norm predictor) *Given a non-negative function $\ell(p; y)$ on $\mathbb{R}^2$, a sequence of distributions $\{\mathcal{D}_d\}_{d=k+1}^\infty$ on $\mathbb{R}^d \times \mathbb{R}$ satisfies* benign overfitting, *if there exists a monotonically increasing sequence of positive integers $\{m_d\}_{d=k+1}^\infty$ such that the following holds:*

- *For any sufficiently large d, if we sample m_d samples $\{(\mathbf{x}_i, y_i)\}_{i=1}^{m_d}$ i.i.d. from $\mathcal{D}_d$, then with probability approaching 1, $\frac{1}{m_d} \sum_{i=1}^{m_d} \ell(\mathbf{x}_i^\top \mathbf{w}; y_i) = 0$ for some $\mathbf{w} \in \mathbb{R}^d$.*

- *Picking $\hat{\mathbf{w}}_d = \arg\min_{\mathbf{w}} \|\mathbf{w}\| : \frac{1}{m} \sum_{i=1}^{m_d} \ell(\mathbf{x}_i^\top \mathbf{w}; y_i) = 0$ to be the min-norm predictor, and defining $R_d(\mathbf{w}) := \mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}_d} \left[\ell(\mathbf{x}^\top \mathbf{w}; y) \right]$, it holds that $\inf_d \inf_{\mathbf{w} \in \mathbb{R}^d} R_d(\mathbf{w}) > 0$ as well as $R_d(\hat{\mathbf{w}}_d) - \inf_{\mathbf{w} \in \mathbb{R}^d} R_d \xrightarrow{P} 0$.*

In other words, the min-norm predictor $\hat{\mathbf{w}}_d$ is asymptotically optimal, in the sense that its expected loss converges to the best possible expected loss among linear predictors, as the sample size and d diverge to infinity at an appropriate rate. Moreover, this holds despite overfitting, in the sense that the training error (which is 0) does not converge to the (strictly positive) expected loss.

The max-margin predictor, and benign overfitting for classification. In linear binary classification problems, we consider distributions where the examples $(\mathbf{x}, y)$ are such that $y \in \{-1, +1\}$, and the predictor (specified by a vector $\mathbf{w}$) is $\mathbf{x} \mapsto \text{sign}(\mathbf{x}^\top \mathbf{w})$. In this

case, we generally care only about the direction of the predictor $\mathbf{w}$, and its expected misclassification error, namely $\Pr_{(\mathbf{x},y)}(y\mathbf{x}^\top\mathbf{w} \leq 0)$. Whereas in regression, standard methods converge to the minimum-norm interpolating predictor, the characterization in classification is a bit different. Concretely, using standard convex classification losses with exponential tails (such as the logistic or cross-entropy loss), it is by now well-known that gradient-based methods ran on the average loss w.r.t. a given dataset $\{\mathbf{x}_i, y_i\}_{i=1}^m$ converge in direction to the max-margin predictor

$$\hat{\mathbf{w}} \;=\; \arg\min_{\mathbf{w}\in\mathbb{R}^d} \|\mathbf{w}\| \;\; : \;\; \min_{i\in[m]} y_i\mathbf{x}_i^\top\mathbf{w} \geq 1$$

(Soudry et al., 2018; Ji and Telgarsky, 2020), which by definition achieves zero misclassification error on the dataset. In this setup, we can define benign overfitting as follows: Similar to the case of regression, we need to consider a sequence of distributions $\{\mathcal{D}_d\}_{d=k+1}^\infty$ (this time on $\mathbb{R}^d \times \{-1, +1\}$) and sample sizes $\{m_d\}_{d=k+1}^\infty$, which induce a sequence of max-margin predictors $\{\hat{\mathbf{w}}_d\}_{d=k+1}^\infty$ (as defined above) when trained on samples $\{\mathbf{x}_i, y_i\}_{i=1}^{m_d}$. Letting $R_d(\mathbf{w}) = \Pr_{(\mathbf{x},y)\sim\mathcal{D}_d}(y\mathbf{x}^\top\mathbf{w} \leq 0)$, we say that the sequence $\{\mathcal{D}_d\}_{d=k+1}^\infty$ satisfies benign overfitting, if $\hat{\mathbf{w}}_d$ exists with probability approaching 1 (as $d \to \infty$), and

$$\inf_d \inf_{\mathbf{w}\in\mathbb{R}^d} R_d(\mathbf{w}) > 0 \quad \text{as well as} \quad R_d(\hat{\mathbf{w}}_d) - \inf_{\mathbf{w}\in\mathbb{R}^d} R_d(\mathbf{w}) \xrightarrow{P} 0 \;. \tag{3}$$

Note that this definition is similar to the one we had for regression (Definition 3), except that $R_d(\cdot)$ is defined with respect to misclassification error, and $\hat{\mathbf{w}}_d$ is now defined as the max-margin predictor.

3. Linear Regression with the Square Loss

We begin by considering the setting of linear regression with the square loss, where our goal is to minimize $\mathbb{E}[(\mathbf{x}^\top\mathbf{w} - y)^2]$ with respect to an underlying distribution over $(\mathbf{x}, y)$. As discussed in the previous section, taking an i.i.d. sample of m training examples, and running standard training methods to convergence, generally results in the min-norm predictor interpolating the data, $\hat{\mathbf{w}} = \arg\min \|\mathbf{w}\| \;:\; \forall i \in [m], \mathbf{x}_i^\top\mathbf{w} = y_i$. The question we ask here is whether this predictor is asymptotically consistent and satisfies benign overfitting.

To motivate our approach, let us first consider the well-specified setting, where $\mathbb{E}[y|\mathbf{x}] = \mathbf{x}^\top\mathbf{w}^*$ for some fixed $\mathbf{w}^*$, and $\mathbf{x}$ is zero-mean. This setting was studied in detail in Bartlett et al. (2020); Hastie et al. (2019); Belkin et al. (2020). In particular, an important corollary of the results of Bartlett et al. (2020) is that for benign overfitting to occur, it is *necessary* that for some k, the smallest $d - k$ eigenvalues of the covariance matrix Σ, namely $\lambda_{k+1}, \ldots, \lambda_d$, satisfy $m \cdot \frac{\sum_{i>k}\lambda_i^2}{(\sum_{i>k}\lambda_i)^2} \to 0$. Letting $\Sigma = \Sigma_k \oplus \Sigma_{d-k}$ be the decomposition of Σ with respect to the first k and last $d - k$ eigenvalues, this is equivalent to requiring $m \cdot \frac{\|\Sigma_{d-k}\|_F^2}{\mathrm{Tr}^2(\Sigma_{d-k})} \to 0$. The following simple lemma implies that this in turn is equivalent to requiring the distribution of $\mathbf{x}_{|d-k}$ to be high-dimensional, in the sense that inner products of independent copies of $\mathbf{x}_{|d-k}$ will be very small compared to the "typical" size of $\mathbf{x}_{|d-k}$:

Lemma 4 *Let $\mathbf{z}, \mathbf{z}'$ be two i.i.d. random vectors in $\mathbb{R}^n$ such that $\mathbb{E}[\mathbf{z}\mathbf{z}^\top] = \Sigma$. Then $\mathbb{E}[(\mathbf{z}^\top\mathbf{z}')^2] = \|\Sigma\|_F^2$. Thus, if $m \cdot \frac{\|\Sigma\|_F^2}{Tr^2(\Sigma)} \to 0$, then $m \cdot \frac{\mathbb{E}[(\mathbf{z}^\top\mathbf{z}')^2]}{\mathbb{E}^2[\|\mathbf{z}\|^2]} \to 0$ as well.*

7

Proof The proof follows from the observation that $\text{Tr}(\Sigma) = \mathbb{E}[\|\mathbf{z}\|^2]$, and that $\mathbb{E}[(\mathbf{z}^\top \mathbf{z}')^2]$ equals

$$\mathbb{E}\left[\left(\sum_{i=1}^n z_i z_i'\right)^2\right] \;=\; \sum_{i,j=1}^n \mathbb{E}[z_i z_j z_i' z_j'] \;=\; \sum_{i,j=1}^n \mathbb{E}[z_i z_j] \cdot \mathbb{E}[z_i' z_j'] \;=\; \sum_{i,j=1}^n \Sigma_{i,j}^2 \;=\; \|\Sigma\|_F^2 \;.$$

∎

Thus, we see that to get benign overfitting even in a well-specified setting, we generally need to consider distributions which have a very high-dimensional component "spread" in many directions. Intuitively, this component will be (approximately) mutually orthogonal across different samples, thus allowing the linear predictor to fit the training data, but without significantly affecting the prediction on new samples. A prototypical example to keep in mind is any spherically symmetric distribution in $\mathbb{R}^{d-k}$, in which case it is not difficult to show that

$$m \cdot \frac{\mathbb{E}[(\mathbf{z}^\top \mathbf{z}')^2]}{\mathbb{E}^2[\|\mathbf{z}\|^2]} \;=\; m \cdot \frac{\|\mathbb{E}[\mathbf{z}\mathbf{z}^\top]\|_F^2}{\mathbb{E}^2[\|\mathbf{z}\|^2]} \;=\; \frac{m}{d-k} \;,$$

which goes to zero as $d \to \infty$ sufficiently faster than m. Moreover, assuming $\|\mathbf{z}\|$ does not fluctuate too wildly, the empirical quantity $m \cdot \left(\frac{\mathbf{z}^\top \mathbf{z}'}{\|\mathbf{z}\|^2}\right)^2$ will strongly concentrate around 0 as $d \to \infty$.

Motivated by this, we focus on distributions where the inputs vectors $\mathbf{x} \in \mathbb{R}^d$ are such that the first k coordinates $\mathbf{x}_{|k}$ (for some fixed k independent[1] of d) have some arbitrary distribution, whereas the last $d-k$ coordinates $\mathbf{x}_{|d-k}$ form a high-dimensional distribution, in the sense that given an i.i.d. sample $\{(\mathbf{x}_i, y_i)\}_{i=1}^m$, it holds with high probability that $\frac{\sup_{i \neq j} |\mathbf{x}_{i|d-k}^\top \mathbf{x}_{j|d-k}|}{\min_i \|\mathbf{x}_{i|d-k}\|^2}$ converges to 0 (as the dimension d increases sufficiently fast compared to m). We note that assuming such a split to the first k and last $d-k$ coordinates is mostly to simplify the presentation, and is without much loss of generality: Indeed, by a suitable rotation of the coordinate system, all our results extend to distributions which can be split into a high-dimensional distribution in some $d-k$-dimensional subspace, and an arbitrary distribution in the complementary k-dimensional subspace. Moreover, variants of this assumption, or related assumptions, are very common in the literature on benign overfitting (e.g., the "junk features" model of Zhou et al. (2020), or the "weak features" model of Muthukumar et al. (2021)), although these tend to assume some particular (e.g., Gaussian) distribution on the first k coordinates, whereas we allow that distribution to be rather generic. Finally, we note that we generally do not assume how y depends on $\mathbf{x}$ (unless specified otherwise).

3.1 A Deterministic Perturbation Bound

Focusing on data samples from such distributions, let us now turn to analyze what form the min-norm predictor takes, as a function of some training data $\{(\mathbf{x}_i, y_i)\}_{i=1}^m$. Intuitively, the analysis rests on viewing the high-dimensional components of the data, $\{\mathbf{x}_{i|d-k}\}_{i=1}^m$, as

---

1. Our proof techniques readily extend to the case of k growing with d at a sufficiently slow rate, but we choose to consider k fixed for simplicity.

a perturbation of perfectly orthogonal vectors. To quantify this, we introduce the perturbation matrix $E \in \mathbb{R}^{m \times m}$, defined as

$$E_{i,j} = \mathbf{x}_{i|d-k}^{\top} \mathbf{x}_{j|d-k} \cdot \mathbf{1}(i \neq j) \quad \forall i, j \in [m] . \tag{4}$$

Note that if $\{\mathbf{x}_{i|d-k}\}_{i=1}^m$ are perfectly orthogonal, then E is the zero matrix. Also, we will use $\hat{\mathbb{E}}[f(\mathbf{x}, y)]$ as shorthand for $\frac{1}{m} \sum_{i=1}^m f(\mathbf{x}_i, y_i)$ for any function f. The key technical result we will need is the following deterministic perturbation bound, which bounds the distance of $\hat{\mathbf{w}}_{|k}$ (the first k coordinates) from a certain expression, as well as bounding the norm of the last $d - k$ coordinates $\hat{\mathbf{w}}_{|d-k}$.

Theorem 5 *Fix some $k \in [d-1]$ and $\{(\mathbf{x}_i, y_i)\}_{i=1}^m \subseteq \mathbb{R}^d \times \mathbb{R}$, such that $\{\mathbf{x}_i\}_{i=1}^m$ are linearly independent, $\frac{\|E\|}{\min_{i \in [m]} \|\mathbf{x}_{i|d-k}\|^2} \leq \frac{1}{2}$, and $\frac{1}{m} \leq \frac{1}{2}\lambda_{\min}\left(\hat{\mathbb{E}}\left[\frac{\mathbf{x}_{|k}\mathbf{x}_{|k}^{\top}}{\|\mathbf{x}_{|d-k}\|^2}\right]\right)$. Then the min-norm predictor $\hat{\mathbf{w}}$ exists and satisfies*

$$\left\| \hat{\mathbf{w}}_{|k} - \left(\hat{\mathbb{E}}\left[\frac{\mathbf{x}_{|k}\mathbf{x}_{|k}^{\top}}{\|\mathbf{x}_{|d-k}\|^2} \right] \right)^{-1} \hat{\mathbb{E}}\left[\frac{y\mathbf{x}_{|k}}{\|\mathbf{x}_{|d-k}\|^2} \right] \right\|$$

$$\leq \frac{2 \left\| \hat{\mathbb{E}}\left[\frac{y\mathbf{x}_{|k}}{\|\mathbf{x}_{|d-k}\|^2} \right] \right\|}{\lambda_{\min}\left(\hat{\mathbb{E}}\left[\frac{\mathbf{x}_{|k}\mathbf{x}_{|k}^{\top}}{\|\mathbf{x}_{|d-k}\|^2} \right] \right)^2} \cdot \frac{1}{m} + \frac{2\sqrt{\hat{\mathbb{E}}[\|\mathbf{x}_{|k}\|^2] \cdot \hat{\mathbb{E}}[y^2]} \cdot}{\min_{i \in [m]} \|\mathbf{x}_{i|d-k}\|^4} \cdot m \cdot \|E\| ,$$

as well as

$$\|\hat{\mathbf{w}}_{|d-k}\| \leq \frac{\sqrt{\hat{\mathbb{E}}[\|\mathbf{x}_{|d-k}\|^2] \cdot \hat{\mathbb{E}}[y^2]}}{\min_{i \in [m]} \|\mathbf{x}_{i|d-k}\|^2} \cdot \left(1 + \frac{2\|E\|}{\min_{i \in [m]} \|\mathbf{x}_{i|d-k}\|^2} \right) \cdot m .$$

The key take-away from this theorem is as follows: Assuming various ratios and empirical moments of the dataset are bounded, then

$$\left\| \hat{\mathbf{w}}_{|k} - \left(\hat{\mathbb{E}}\left[\frac{\mathbf{x}_{|k}\mathbf{x}_{|k}^{\top}}{\|\mathbf{x}_{|d-k}\|^2} \right] \right)^{-1} \hat{\mathbb{E}}\left[\frac{y\mathbf{x}_{|k}}{\|\mathbf{x}_{|d-k}\|^2} \right] \right\| \leq \mathcal{O}\left(\frac{1}{m} + m \cdot \|E\| \right) ,$$

where $E \to 0$ as the inner products between pairs of vectors in $\{\mathbf{x}_{i|d-k}\}_{i=1}^m$ go to zero. Assuming this convergence to zero is sufficiently fast compared to m, that $m \to \infty$, and that the law of large numbers hold, we get that

$$\hat{\mathbf{w}}_{|k} \to \left(\mathbb{E}\left[\frac{\mathbf{x}_{|k}\mathbf{x}_{|k}^{\top}}{\|\mathbf{x}_{|d-k}\|^2} \right] \right)^{-1} \mathbb{E}\left[\frac{y\mathbf{x}_{|k}}{\|\mathbf{x}_{|d-k}\|^2} \right] .$$

As to $\hat{\mathbf{w}}_{|d-k}$, we effectively bound its norm by $\mathcal{O}(m)$, which scales with m but not with the dimension. If the input distribution on the last $d - k$ coordinates is sufficiently high-dimensional, this implies that given a new sample $\mathbf{x}$, the contribution of $\mathbf{x}_{|d-k}^{\top}\hat{\mathbf{w}}_{|d-k}$ to the predicted value (namely $\mathbf{x}^{\top}\hat{\mathbf{w}} = \mathbf{x}_{|k}^{\top}\hat{\mathbf{w}}_{|k} + \mathbf{x}_{|d-k}^{\top}\hat{\mathbf{w}}_{|d-k}$) is asymptotically negligible. Thus, the asymptotic expression of $\hat{\mathbf{w}}_{|k}$ eventually determines the behavior of the learned predictor.

Before continuing, let us provide an informal and partial proof sketch, explaining where the approximate expression for $\hat{\mathbf{w}}_{|k}$ in Theorem 5 comes from. To that end, let X be a matrix whose i-th row is $\mathbf{x}_i$, and $\mathbf{y} = (y_1, \ldots, y_m)$. By standard results, $\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \|\mathbf{w}\|$: $X\mathbf{w} = \mathbf{y}$ has the closed-form expression $X^\top (XX^\top)^{-1} \mathbf{y}$. Letting $X_{|k}$ be the first k columns of X, it follows that $\hat{\mathbf{w}}_{|k} = X_{|k}^\top ((XX^\top)^{-1} \mathbf{y}) = X_{|k}^\top (X_{|k} X_{|k}^\top + X_{|d-k} X_{|d-k}^\top) \mathbf{y}$. Suppose for simplicity that $\{\mathbf{x}_{i|d-k}\}_{i=1}^m$ are precisely orthogonal (so that $E = 0$ in the theorem above, and $X_{|d-k} X_{|d-k}^\top$ equals a diagonal matrix D). As a result, we get $\hat{\mathbf{w}}_{|k} = X_{|k}^\top \left(X_{|k} X_{|k}^\top + D \right)^{-1} \mathbf{y}$. By the Woodbury matrix identity and some algebraic manipulations, this equals $\left(I + X_{|k}^\top D^{-1} X_{|k} \right)^{-1} X_{|k}^\top D^{-1} \mathbf{y}$, or equivalently,

$$\left(\frac{1}{m} I + \frac{1}{m} X_{|k}^\top D^{-1} X_{|k} \right)^{-1} \left(\frac{1}{m} X_{|k}^\top D^{-1} \mathbf{y} \right) \;=\; \left(\frac{1}{m} I + \hat{\mathbb{E}}\left[\frac{\mathbf{x}_{|k} \mathbf{x}_{|k}^\top}{\|\mathbf{x}_{|d-k}\|^2} \right] \right)^{-1} \cdot \hat{\mathbb{E}}\left[\frac{y \mathbf{x}_{|k}}{\|\mathbf{x}_{|d-k}\|^2} \right] \;,$$

which approaches $\left(\hat{\mathbb{E}}\left[\frac{\mathbf{x}_{|k} \mathbf{x}_{|k}^\top}{\|\mathbf{x}_{|d-k}\|^2} \right] \right)^{-1} \cdot \hat{\mathbb{E}}\left[\frac{y \mathbf{x}_{|k}}{\|\mathbf{x}_{|d-k}\|^2} \right]$ as m increases.

3.2 Asymptotic Characterization of the min-norm predictor

Let us now turn to show how Theorem 5 can lead to a formal asymptotic characterization of the min-norm predictor $\hat{\mathbf{w}}$, in a statistical setting where the training data is sampled from some underlying distribution. To do so, we will need to impose assumptions on the distribution, which ensure that the perturbation matrix E from Theorem 5 indeed converges to 0, and that the various quantities in the bounds are well-behaved. One such set of sufficient conditions is the following:

Assumption 1 *Suppose $\{\mathcal{D}_d\}_{d=k+1}^\infty$ is a sequence of distributions on $\mathbb{R}^d \times \mathbb{R}$, and $\{m_d\}_{d=k+1}^\infty$ a monotonically increasing sequence of positive integers diverging to ∞, such that the following hold:*

1. *Letting $\mathbb{E}_d$ be shorthand for $\mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}_d}$, it holds that*

$$\sup_d \max \left\{ \mathbb{E}_d[\|\mathbf{x}\|^4] \;,\; \mathbb{E}_d[y^4] \;,\; \mathbb{E}_d\left[\frac{\|y\mathbf{x}_{|k}\|^2}{\|\mathbf{x}_{|d-k}\|^4} \right] \;,\; \mathbb{E}_d\left[\frac{\|\mathbf{x}_{|k}\|^4}{\|\mathbf{x}_{|d-k}\|^4} \right] \right\} < \infty \;,$$

and $\inf_d \lambda_{\min} \left(\mathbb{E}_d\left[\frac{\mathbf{x}_{|k} \mathbf{x}_{|k}^\top}{\|\mathbf{x}_{|d-k}\|^2} \right] \right) > 0$.

2. *If we sample m_d i.i.d. samples $\{\mathbf{x}_i, y_i\}_{i=1}^{m_d}$ from $\mathcal{D}_d$, then with probability approaching 1, $\{\mathbf{x}_i\}_{i=1}^{m_d}$ are linearly independent, and $\min_{i \in [m_d]} \|\mathbf{x}_{i|d-k}\|$ is at least some $c > 0$ independent of d.*

3. *$m_d \cdot \|E\| \xrightarrow{P} 0$, where E is as defined in Eq. (4).*

4. *$m_d^2 \cdot \|\mathbb{E}_d[\mathbf{x}_{|d-k} \mathbf{x}_{|d-k}^\top]\| \to 0$.*

Since the assumptions are rather technical, let us provide one simple example to keep in mind, which satisfies the above:

Example 1 *Suppose that* $\lim_{d\to\infty} \frac{m_d^3 \log(d)}{d} = 0$, *and that* $\mathcal{D}_d$ *is defined as follows:* $(\mathbf{x}_{|k}, y)$ *has some fixed distribution (independent of d), with bounded moments up to order* 4 *and such that* $\mathbb{E}[\mathbf{x}_{|k}\mathbf{x}_{|k}^\top]$ *is positive definite; And* $\mathbf{x}_{|d-k}$ *is an independent zero-mean Gaussian with covariance matrix* $\frac{1}{d-k} \cdot I$.
In this case, $\|\mathbf{x}_{|d-k}\|$ *strongly concentrates around* 1 *as* $d \to \infty$ *increases, and* $\{\mathbf{x}_{i|d-k}\}_{i=1}^{m_d}$ *are linearly independent with probability* 1, *hence conditions* 1 *and* 2 *in the assumption clearly holds. As to conditions* 3 *and* 4, *letting* $Z \in \mathbb{R}^{m\times(d-k)}$ *be the matrix whose rows are* $\mathbf{x}_{i|d_k}$, *and* D *the diagonal of* $ZZ^\top$, *we have* $\|E\| = \|ZZ^\top - D\| \le \|ZZ^\top - I\| + \|I - D\|$, *which is at most* $\mathcal{O}(\sqrt{m_d \log(d)/d})$ *with probability converging to* 1 *as* $d \to \infty$ *(see for example (Zhu, 2012)). Also, clearly* $\|\mathbb{E}[\mathbf{x}_{|d-k}\mathbf{x}_{|d-k}^\top]\| = \frac{1}{d-k}$. *Combined with the assumption* $\lim_{d\to\infty} \frac{m_d^3 \log(d)}{d} = 0$, *conditions 3,4 follow.*

We remark that in the example, we require $\lim_{d\to\infty} \frac{m_d^3 \log(d)}{d} = 0$, which is a stronger assumption on the scaling of d vs. m_d compared to previous work on linear regression (which usually consider $m_d/d \to 0$ under similar distributional assumptions). On the flip side, the proof technique allows us to analyze more general settings which go beyond well-specified linear regression.

In any case, we emphasize that Assumption 1 applies far more broadly than Example 1: For instance, it generally applies to any spherically-symmetric distribution of $\mathbf{x}_{|d-k}$ (possibly dependent on $\mathbf{x}_{|k}, y$) so that $\|\mathbf{x}_{|d-k}\|$ is bounded (or at least concentrated) in some fixed interval bounded away from 0. Also, Assumption 1 itself is not the most general possible, in the sense that it focuses on situations where $\|\mathbf{x}\|, y$ and $\frac{1}{\|\mathbf{x}_{|d-k}\|}$ are scaled so that they are essentially bounded independent of d. Moreover, using Theorem 5 above one can analyze even more general situations: For example, that the data norm scales with d, while only bounding various ratios between relevant quantities.

Under Assumption 1, let us now proceed to formally state our asymptotic characterization of the min-norm predictor:

Theorem 6 *Suppose* $\{\mathcal{D}_d\}_{d=k+1}^\infty$ *and* $\{m_d\}_{d=k+1}^\infty$ *satisfy Assumption 1. For any d, let* $\hat{\mathbf{w}}_d = \arg\min \|\mathbf{w}\| : \forall i \in [m_d],\ \mathbf{x}_i^\top \mathbf{w} = y_i$ *be the min-norm predictor w.r.t. a training set of size* m_d *sampled i.i.d. from* $\mathcal{D}_d$. *Then as* $d \to \infty$, $\hat{\mathbf{w}}_d$ *exists with probability approaching* 1, *and satisfies*

$$\left\| \hat{\mathbf{w}}_{d|k} - \left(\mathbb{E}_d \left[\frac{\mathbf{x}_{|k}\mathbf{x}_{|k}^\top}{\|\mathbf{x}_{|d-k}\|^2} \right] \right)^{-1} \mathbb{E}_d \left[\frac{y\mathbf{x}_{|k}}{\|\mathbf{x}_{|d-k}\|^2} \right] \right\| \xrightarrow{P} 0 \quad and \quad \mathbb{E}_d\left[(\mathbf{x}^\top \hat{\mathbf{w}}_d - \mathbf{x}_{|k}^\top \hat{\mathbf{w}}_{d|k})^2 \right] \xrightarrow{P} 0 .$$

3.3 Implications for benign overfitting in regression

Having established this asymptotic characterization of $\hat{\mathbf{w}}$, we now turn to discuss its implications to benign overfitting in linear regression. Our bottom-line message is that in general, $\hat{\mathbf{w}}_d$ is asymptotically *not* an optimal predictor, and hence benign overfitting will not occur.

To see this, consider any sequence of distributions $\{\mathcal{D}_d\}$ as in Theorem 6. For any $\mathcal{D}_d$, the expected loss has the form

$$R_d(\mathbf{w}) := \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_d}[(\mathbf{x}^\top\mathbf{w} - y)^2] = \mathbf{w}^\top \mathbb{E}_d[\mathbf{x}\mathbf{x}^\top]\mathbf{w} - 2\mathbb{E}_d[y\mathbf{x}^\top]\mathbf{w} + \mathbb{E}_d[y^2] .$$

Assuming $\mathbb{E}_d[\mathbf{x}\mathbf{x}^\top]$ is positive definite, it follows that the unique minimizer equals $\mathbf{w} = \mathbb{E}[\mathbf{x}\mathbf{x}^\top]^{-1}\mathbb{E}[y\mathbf{x}]$, and in particular

$$\mathbf{w}_{|k} = \left(\mathbb{E}_d[\mathbf{x}\mathbf{x}^\top]^{-1}\right)_k \mathbb{E}_d[y\mathbf{x}] \ ,$$

where $\left(\mathbb{E}_d[\mathbf{x}\mathbf{x}^\top]^{-1}\right)_k$ refers to the first k rows of the inverse covariance matrix. Thus, for benign overfitting to occur, we need that the min-norm predictor $\hat{\mathbf{w}}_d$ will be such that $\hat{\mathbf{w}}_{d|k}$ equals this expression, at least asymptotically. However, Theorem 6 implies that $\hat{\mathbf{w}}_{d|k}$ is asymptotically biased towards a different expression, namely $\left(\mathbb{E}_d\left[\frac{\mathbf{x}_{|k}\mathbf{x}_{|k}^\top}{\|\mathbf{x}_{|d-k}\|^2}\right]\right)^{-1}\mathbb{E}_d\left[\frac{y\mathbf{x}_{|k}}{\|\mathbf{x}_{|d-k}\|^2}\right]$. Thus, unless the two expressions somehow exactly coincide, there is no reason to believe that benign overfitting to occur, even though the covariance structure of the inputs $\mathbf{x}$ can be a textbook case of amenability to benign overfitting (in terms of the conditions of Theorem 6 or previous papers on benign overfitting in regression). The following example illustrates this:

Example 2 *In the setting of Theorem 6, suppose $k = 1$, x_1 (the first coordinate of $\mathbf{x}$) is uniform on the interval $[-a, a]$ for some arbitrary $a > 0$, $y = \exp(x_1)$, and for all $j \in \{2, \ldots, d\}$, $x_j = \sqrt{\frac{y}{d-1}} \cdot r_j$, where each r_j is an independent standard Gaussian random variable. Then $\mathbb{E}[x_1 x_j] = \mathbb{E}[y x_j] = 0$ for all $j > 1$, and therefore*

$$R_d(\mathbf{w}) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_d}[(\mathbf{x}^\top\mathbf{w} - y)^2] = \mathbf{w}^\top\mathbb{E}[\mathbf{x}\mathbf{x}^\top]\mathbf{w} - 2\mathbb{E}[y\mathbf{x}]^\top\mathbf{w} + \mathbb{E}[y^2]$$

$$= \mathbb{E}[x_1^2] \cdot w_1^2 + \frac{\mathbb{E}[y]}{d-1} \cdot \sum_{j=2}^{d} w_j^2 - 2\mathbb{E}[yx_1] \cdot w_1 + \mathbb{E}[y^2] \ .$$

By differentiating the above w.r.t. $\mathbf{w}$, it is easily verified that $R_d(\cdot)$ achieves a minimal value only when $w_j = 0$ for all $j > 1$, and $w_1 = \frac{\mathbb{E}[yx_1]}{\mathbb{E}[x_1^2]} = \frac{\mathbb{E}[\exp(x_1)x_1]}{\mathbb{E}[x_1^2]}$, which is a strictly positive number dependent only on a. However, by Theorem 6 and standard concentration results for the Gaussian distribution, the first coordinate of $\hat{\mathbf{w}}_d$ converges in probability to the different value $\mathbb{E}\left[\frac{yx_1}{\exp(x_1)}\right] / \mathbb{E}\left[\frac{x_1^2}{\exp(x_1)}\right] = \frac{\mathbb{E}[x_1]}{\mathbb{E}[x_1^2\exp(-x_1)]} = 0$. It follows that $R_d(\hat{\mathbf{w}}_d) - \inf_{\mathbf{w}} R_d(\mathbf{w})$ is lower bounded by a positive number independent of d, and therefore we do not have benign overfitting.

The reader familiar with previous literature might wonder how this can possibly accord with previous results (such as Bartlett et al. (2020)), which show that benign overfitting *does* occur for linear regression with the square loss, under the kind of input distributions we study here. The reason is that these results assume a well-specified setting, where $\mathbb{E}[y|\mathbf{x}] = \mathbf{x}^\top\mathbf{w}^*$ for some fixed $\mathbf{w}^*$ (see for example Assumption 4 in Definition 1 of Bartlett et al. (2020)). In the example above, this does not hold, since $\mathbb{E}[y|\mathbf{x}] = \exp(x_1)$ is not a linear function of $\mathbf{x}$. Had we been in a well-specified setting with $\mathbb{E}[y|\mathbf{x}] = \mathbf{x}_{|k}^\top\mathbf{w}_{|k}^*$ for some $\mathbf{w}^*$, benign overfitting *would* generally occur, because then we have that

$$\left(\mathbb{E}\left[\frac{\mathbf{x}_{|k}\mathbf{x}_{|k}^\top}{\|\mathbf{x}_{|d-k}\|^2}\right]\right)^{-1} \cdot \mathbb{E}\left[\frac{y\mathbf{x}_{|k}}{\|\mathbf{x}_{|d-k}\|^2}\right] = \left(\mathbb{E}\left[\frac{\mathbf{x}_{|k}\mathbf{x}_{|k}^\top}{\|\mathbf{x}_{|d-k}\|^2}\right]\right)^{-1} \cdot \mathbb{E}\left[\frac{\mathbf{x}_{|k}\mathbf{x}_{|k}^\top}{\|\mathbf{x}_{|d-k}\|^2}\right]\mathbf{w}_{|k}^* = \mathbf{w}_{|k}^* \ ,$$

which now coincides with the optimal solution on the first k coordinates. However, a well-specified distribution is the exception rather than the rule in practice.

4. Linear Regression Beyond the Square Loss

In the previous section, we studied linear regression with the square loss, with our main conclusion being that benign overfitting should not be expected in general, beyond well-specified distributions. In this section, we study what happens if we do focus on well-specified distributions, but consider more general regression problems (beyond linear regression with the square loss). We will see that here again, benign overfitting can generally fail to hold.

Concretely, suppose that instead of the square loss, we have some non-negative loss function $\ell(\mathbf{x}^\top\mathbf{w}; y)$, which is relevant for regression in the sense that for any y, it equals 0 at some unique value $\ell_y^{-1}(0)$. As discussed in Section 2 and Proposition 2, we still expect standard gradient-based training methods to converge to a minimum-norm interpolating predictor, assuming they manage to drive the average loss on the training set to 0: Namely, given a dataset $\{\mathbf{x}_i, y_i\}_{i=1}^m$, the unique point in $\arg\min_{\mathbf{w}} \|\mathbf{w}\| : \frac{1}{m}\sum_{i=1}^m \ell(\mathbf{x}_i^\top\mathbf{w}; y_i) = 0$. The question now is whether this predictor enjoys benign overfitting.

To answer this question, one option is to try and repeat the analysis from the previous section, depending on the choice of $\ell(\cdot)$. However, we will take a different approach, which allows us to study this using the results we already developed for linear regression with the square loss. Our crucial observation can be phrased as the following, very simple lemma:

Lemma 7 *Given a dataset $\{\mathbf{x}_i, y_i\}_{i=1}^m$ and non-negative loss $\ell(\cdot)$ as above,*

$$\arg\min_{\mathbf{w}} \|\mathbf{w}\| : \frac{1}{m}\sum_{i=1}^m \ell(\mathbf{x}_i^\top\mathbf{w}; y_i) = 0 \quad equals \quad \arg\min_{\mathbf{w}} \|\mathbf{w}\| : \frac{1}{m}\sum_{i=1}^m (\mathbf{x}_i^\top\mathbf{w} - \ell_{y_i}^{-1}(0))^2 = 0 \ .$$

The proof of the lemma trivially follows from the observation that if $\frac{1}{m}\sum_{i=1}^m \ell(\mathbf{x}_i^\top\mathbf{w}; y_i) = 0$, then $\ell(\mathbf{x}_i^\top\mathbf{w}; y_i) = 0$ for all i, hence by assumption, $(\mathbf{x}_i^\top\mathbf{w} - \ell_{y_i}^{-1}(0))^2 = 0$ for all i.

The lemma implies that the same method that converges to the minimum-norm minimizer of the average loss w.r.t. $\ell(\cdot)$, also simultaneously converges to the minimum-norm minimizer of the average loss w.r.t. the square loss with target values $\ell^{-1}(y_i)$. However, assuming that the training set $\{(\mathbf{x}_i, y_i)\}_{i=1}^m$ is sampled i.i.d. from some underlying distribution, it is evident that they represent empirical risk minimization of two distinct statistical learning problems: One being minimizing $\mathbb{E}[\ell(\mathbf{x}^\top\mathbf{w}; y)]$, and the other minimizing $\mathbb{E}[(\mathbf{x}^\top\mathbf{w} - \ell_y^{-1}(0))^2]$. In general, these are different learning problems, with distinct optima with respect to the underlying data distribution, yet the returned $\hat{\mathbf{w}}$ is exactly the same one. Thus, if we have benign overfitting in one problem (with the trained predictor $\hat{\mathbf{w}}$ having near-minimal expected loss), we generally *cannot* expect to have benign overfitting in the other problem. Thus, *the very fact that we can show benign overfitting in settings such as well-specified linear regression with the square loss, precludes the possibility of having benign overfitting in other learning problems.*

In what follows, we exemplify this observation on two types of well-specified regression problems. The first setting we study is a generalized linear model. Concretely, we consider predictors of the form $\mathbf{x} \mapsto \sigma(\mathbf{x}^\top\mathbf{w})$, where $\mathbf{w}$ is the parameter vector and $\sigma(\cdot)$ is some strictly monotonic non-linear function, and assume that $\mathbb{E}[y|\mathbf{x}] = \sigma(\mathbf{x}^\top\mathbf{w}^*)$ for some $\mathbf{w}^*$. In the context of neural networks, this can also be viewed as training a single neuron using some nonlinear activation function $\sigma(\cdot)$. In this setting, standard gradient-based methods trained on the average square loss (i.e., $\min_{\mathbf{w}} \frac{1}{m}\sum_{i=1}^m (\sigma(\mathbf{x}_i^\top\mathbf{w}) - y_i)^2$) will indeed generally converge

to the min-norm predictor, namely $\arg\min_{\mathbf{w}} \|\mathbf{w}\|$: $\frac{1}{m_d}\sum_{i=1}^{m_d}(\sigma(\mathbf{x}_i^\top\mathbf{w}) - y_i)^2 = 0$ (see for example the proof of Yehudai and Shamir (2020, Thm. 3.2), combined with Proposition 2). The following corollary of our previous results implies that for just about any choice of input distribution on the first k coordinates, and just about any choice of a strictly monotonic non-linear $\sigma(\cdot)$, we generally *cannot* expect benign overfitting to occur, even if the model is well-specified:

Corollary 8 *Suppose that $\sigma : \mathbb{R} \to \mathbb{R}$ is a function whose inverse $\sigma^{-1}(\cdot)$ exists and is Lipschitz continuous. Consider any sequence of distributions $\{\mathcal{D}_d\}_{d=k+1}^\infty$ and integers $\{m_d\}_{d=k+1}^\infty$ satisfying Assumption 1, such that for any d and $(\mathbf{x}, y) \sim \mathcal{D}_d$, $y = \sigma(\mathbf{x}_{|k}^\top\mathbf{w}^*) + \xi$ for some fixed $\mathbf{w}^* \in \mathbb{R}^k$ and random variable ξ. Then the min-norm predictor $\hat{\mathbf{w}}_d$ satisfies*

$$\left\| \hat{\mathbf{w}}_{d|k} - \left(\mathbb{E}_d\left[\frac{\mathbf{x}_{|k}\mathbf{x}_{|k}^\top}{\|\mathbf{x}_{|d-k}\|^2}\right]\right)^{-1} \cdot \mathbb{E}_d\left[\frac{\sigma^{-1}\left(\sigma(\mathbf{x}_{|k}^\top\mathbf{w}^*) + \xi\right)\mathbf{x}_{|k}}{\|\mathbf{x}_{|d-k}\|^2}\right] \right\| \xrightarrow{P} 0$$

and $\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_d}\left[(\mathbf{x}^\top\hat{\mathbf{w}}_d - \mathbf{x}_{|k}^\top\hat{\mathbf{w}}_{d|k})^2\right] \xrightarrow{P} 0$

The corollary follows immediately from the observation that since σ is invertible, $\hat{\mathbf{w}}_d$ is also the minimum-norm minimizer of $\frac{1}{m_d}\sum_{i=1}^{m_d}(\mathbf{x}_i^\top\mathbf{w} - \sigma^{-1}(y_i))^2 = 0$, and that the moment conditions in Assumption 1 are still satisfied if we replace y by $\sigma^{-1}(y)$ (since $|\sigma^{-1}(y)| \leq c_\sigma(1 + |y|)$ for some $c_\sigma > 0$ dependent only on σ). Applying Theorem 6 on $\hat{\mathbf{w}}_d$ with y replaced by $\sigma^{-1}(y)$, we get that

$$\left\| \hat{\mathbf{w}}_{d|k} - \left(\mathbb{E}_d\left[\frac{\mathbf{x}_{|k}\mathbf{x}_{|k}^\top}{\|\mathbf{x}_{|d-k}\|^2}\right]\right)^{-1} \mathbb{E}_d\left[\frac{\sigma^{-1}(y)\mathbf{x}_{|k}}{\|\mathbf{x}_{|d-k}\|^2}\right] \right\| \xrightarrow{P} 0 \,,$$

and plugging in $y = \sigma(\mathbf{x}_{|k}^\top) + \xi$ results in the theorem.

When ξ is independent zero-mean noise, and $\sigma(\cdot)$ (and hence $\sigma^{-1}(\cdot)$) is a linear function, then the asymptotic expression for $\hat{\mathbf{w}}_{d|k}$ in the theorem above reduces to $\mathbf{w}^*$, which is indeed the optimal vector we would hope to converge to. However, when $\sigma(\cdot)$ is nonlinear, the expression is not $\mathbf{w}^*$ in general, and hence we do not get asymptotic consistency. To give just one simple example, suppose that $\|\mathbf{x}_{|d-k}\| = 1$ with probability 1, $\sigma(0) = 0$, $\mathbf{w}^* = \mathbf{0}$ and $\mathbb{E}[\mathbf{x}_{|k}] \neq \mathbf{0}$. In this case the asymptotic expression for $\hat{\mathbf{w}}_{d|k}$ in the theorem reduces to, $\left(\mathbb{E}[\mathbf{x}_{|k}\mathbf{x}_{|k}^\top]\right)^{-1}\mathbb{E}[\mathbf{x}_{|k}]\cdot\mathbb{E}[\sigma^{-1}(\xi)]$. For this to equal $\mathbf{w}^*$ (namely $\mathbf{0}$), we need that $\mathbb{E}[\sigma^{-1}(\xi)] = 0$. However, since $\sigma(\cdot)$ (and hence $\sigma^{-1}(\cdot)$) is non-linear, the equation above will not hold for "most" zero-mean distributions of ξ. In other words, even if we fix the input distribution, then just by playing around with the distribution of the noise term ξ, we can easily encounter situations where benign overfitting does not hold. Concretely, the following lemma (whose proof is in the appendix) shows that *no* nonlinear $\sigma(\cdot)$ can possibly satisfy $\mathbb{E}[\sigma^{-1}(\xi)] = 0$ for all zero-mean distributions:

Lemma 9 *Suppose that $\sigma^{-1}(\cdot)$ is a function on $\mathbb{R}$ such that $\mathbb{E}[\sigma^{-1}(\xi)] = 0$ for all zero-mean random variables ξ with support of size at most 2. Then $\sigma^{-1}(\cdot)$ (and hence $\sigma(\cdot)$) must be a homogeneous linear function (that is, $\exists c \in \mathbb{R}$ s.t. $\forall z \in \mathbb{R}$, $\sigma^{-1}(z) = cz$).*

Next, we go back to linear regression, but now assume that we use some convex loss which is not necessarily the square loss (say, the absolute loss). Here again, standard gradient methods trained on the average loss will generally converge to a min-norm interpolating predictor (thanks to convexity and Proposition 2). However, we cannot expect benign overfitting to occur in general for this predictor:

Corollary 10 *Consider any sequence of distributions $\{\mathcal{D}_d\}_{d=k+1}^{\infty}$ and integers $\{m_d\}_{d=k+1}^{\infty}$ satisfying Assumption 1. Suppose we use the loss function $\ell(\mathbf{x}^{\top}\mathbf{w}; y) = f(\mathbf{x}^{\top}\mathbf{w} - y)$ for some non-negative function f which has a unique root at 0. Then the min-norm predictor $\hat{\mathbf{w}}$ satisfies $\mathbb{E}_{(\mathbf{x},y) \sim \mathcal{D}_d}\left[(\mathbf{x}^{\top}\hat{\mathbf{w}}_d - \mathbf{x}_{|k}^{\top}\hat{\mathbf{w}}_{d|k})^2\right] \xrightarrow{P} 0$ and*

$$\left\| \hat{\mathbf{w}}_{d|k} - \left(\mathbb{E}_d \left[\frac{\mathbf{x}_{|k}\mathbf{x}_{|k}^{\top}}{\|\mathbf{x}_{|d-k}\|^2} \right] \right)^{-1} \cdot \mathbb{E}_d \left[\frac{y\mathbf{x}_{|k}}{\|\mathbf{x}_{|d-k}\|^2} \right] \right\| \xrightarrow{P} 0 .$$

The proof is immediate from observing that $\hat{\mathbf{w}}_d$ is also $\arg\min \|\mathbf{w}\| : \frac{1}{m_d} \sum_{i=1}^{m_d} (\mathbf{x}_i^{\top}\mathbf{w} - y_i)^2 = 0$, and applying Theorem 6 on this related linear regression problem. Crucially, note that $\hat{\mathbf{w}}_{d|k}$ has the same asymptotic characterization as if we have used the square loss, and there is no reason to believe that this is also an optimal solution (w.r.t. the first k coordinates) of $\mathbb{E}[f(\mathbf{x}^{\top}\mathbf{w} - y)]$ when $f(\cdot)$ is not the square loss. Let us illustrate this with a simple example:

Example 3 *In the setting of Corollary 10, suppose $f(z) = |z|$ is the absolute loss, $k = 1$, $x_1 = 1$ with probability 1, and $y = x_1 + \xi$ for some independent random variable ξ. Then the first coordinate of $\hat{\mathbf{w}}_d$ converges in probability to $\mathbb{E}[yx_1]/\mathbb{E}[x_1^2] = 1 + \mathbb{E}[\xi]$. However, the expected absolute loss is $R_d(\mathbf{w}) = \mathbb{E}[|\mathbf{x}^{\top}\mathbf{w} - y|] = \mathbb{E}\left[\left|w_1 + \sum_{j=2}^{d} x_j w_j - (1 + \xi)\right|\right]$, which is easily verified to be minimized only when $w_1 = 1 + \text{med}(\xi)$ (where $\text{med}(\xi)$ is the median of ξ). Thus, whenever $\text{med}(\xi) \neq \mathbb{E}[\xi]$ (which generally occurs when ξ has a non-symmetric distribution), $R_d(\hat{\mathbf{w}}_d) - \inf_{\mathbf{w}} R_d(\mathbf{w})$ does not converge to 0, and we do not have benign overfitting.*

Remark 11 (Implicit bias towards a weighted square loss problem) *In Theorem 6, for linear regression with the square loss, we saw that $\hat{\mathbf{w}}_{d|k}$ asymptotically equals*

$$\left(\mathbb{E}_d \left[\frac{\mathbf{x}_{|k}\mathbf{x}_{|k}^{\top}}{\|\mathbf{x}_{|d-k}\|^2} \right] \right)^{-1} \mathbb{E}_d \left[\frac{y\mathbf{x}_{|k}}{\|\mathbf{x}_{|d-k}\|^2} \right] .$$

This can be equivalently seen as the minimum-norm optimum of the objective function

$$\mathbb{E}_d \left[\left(\frac{\mathbf{x}^{\top}}{\|\mathbf{x}_{|d-k}\|}\mathbf{w} - \frac{y}{\|\mathbf{x}_{|d-k}\|} \right)^2 \right] .$$

In other words, even though $\hat{\mathbf{w}}$ minimizes $\frac{1}{m} \sum_{i=1}^{m} (\mathbf{x}_i^{\top}\mathbf{w} - y_i)^2$, and should asymptotically minimize $\mathbb{E}[(\mathbf{x}^{\top}\mathbf{w} - y)^2]$ for benign overfitting to occur, its first k coordinates actually optimize a weighted version of this problem, where both $\mathbf{x}, y$ are scaled down by $\|\mathbf{x}_{|d-k}\|$.

This can be explained via the approach developed in this section: It trivially holds that $\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \|\mathbf{w}\| : \frac{1}{m} \sum_{i=1}^{m} (\mathbf{x}_i^\top \mathbf{w} - y_i)^2 = 0$ also equals

$$\arg\min_{\mathbf{w}} \|\mathbf{w}\| : \frac{1}{m} \sum_{i=1}^{m} \left(\frac{\mathbf{x}_i^\top}{\|\mathbf{x}_{i|d-k}\|} \mathbf{w} - \frac{y_i}{\|\mathbf{x}_{i|d-k}\|} \right)^2 = 0 \ ,$$

and in terms of asymptotic behavior, it turns out that $\hat{\mathbf{w}}$ is actually "consistent" with respect to the statistical problem associated with the latter, weighted loss function, and not the former unweighted one.

5. Linear Binary Classification

The results in the previous section suggest that many natural extensions of well-specified linear regression with the square loss will generally not satisfy benign overfitting. These were all regression problems, where to get low loss the prediction value must be close to some optimal value.

In this section, we turn to consider binary linear classification setups, where we only care about the sign of $\mathbf{x}^\top \mathbf{w}$ rather than its exact value, and see that the situation there is much more favorable. As in the case of regression, we will focus on input distributions which can be decomposed to some arbitrary distribution on the first k coordinates, and a high-dimensional distribution on the last $d - k$ coordinates (for example, a spherically symmetric distribution).

As discussed in Section 2, whereas for regression we care about the minimum-norm interpolating predictor, for classification we care about the max-margin predictor, $\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} : \min_{i \in [m]} y_i \mathbf{x}_i^\top \mathbf{w} \geq 1$. To study benign overfitting for such problems, we need an asymptotic characterization of $\hat{\mathbf{w}}$, similar to what we have done for regression. However, this might seem difficult, since unlike the min-norm predictor, the max-margin predictor does not have a closed-form expression. In fact, many previous analyses of benign overfitting in classification resorted to additional assumptions which make the max-margin predictor coincide with the min-norm solution, $\arg\min_{\mathbf{w}} \|\mathbf{w}\| : \frac{1}{m_d} \sum_{i=1}^{m_d} (\mathbf{x}_i^\top \mathbf{w} - y_i)^2 = 0$. We take a different route, which applies even when the max-margin and min-norm solutions do not coincide: We show that at least for distributions falling within our framework, the first k coordinates of $\hat{\mathbf{w}}_d$ asymptotically minimize the expectation of a (weighted) *squared hinge loss* on those coordinates, namely $\arg\min_{\mathbf{v} \in \mathbb{R}^k} \mathbb{E} \left[\frac{[1 - y\mathbf{x}_{|k}^\top \mathbf{v}]_+^2}{\|\mathbf{x}_{|d-k}\|^2} \right]$. As in the case of regression, we will first show how to derive this using a deterministic perturbation bound (depending on the extent to which the high-dimensional components in the data are far from being perfectly orthogonal), followed by a probabilistic asymptotic characterization of the max-margin predictor, and finally discuss its implications.

5.1 A deterministic perturbation bound

Recall that our analysis for regression relied on the assumption that the input distribution on the last $d - k$ coordinates is high-dimensional, in the sense that $\mathbf{x}_{i|d-k}^\top \mathbf{x}_{j|d-k} \approx 0$ for $i \neq j$. With this scenario in mind, we present the following deterministic perturbation bound, which characterizes $\hat{\mathbf{w}}_{|k}$ and the norm of $\hat{\mathbf{w}}_{|d-k}$ when such inner products are small.

As in the case of regression, we utilize a perturbation matrix $E \in \mathbb{R}^{m \times m}$, which is now defined as

$$E_{i,j} \;=\; y_i y_j \mathbf{x}_{i|d-k}^\top \mathbf{x}_{j|d-k} \cdot \mathbf{1}(i \neq j) \quad \forall i, j \in [m] \;, \tag{5}$$

and use $\hat{\mathbb{E}}[f(\mathbf{x}, y)]$ as shorthand for $\frac{1}{m} \sum_{i=1}^m f(\mathbf{x}_i, y_i)$ for any function f.

Theorem 12 *Fix some* $k \in [d-1]$ *and* $\{(\mathbf{x}_i, y_i)\}_{i=1}^m \subseteq \mathbb{R}^d \times \{-1, +1\}$. *Suppose that* $\epsilon_0 := \frac{2\|E\| \cdot \max_{i \in [m]} \|\mathbf{x}_{i|d-k}\|^2}{\min_{i \in [m]} \|\mathbf{x}_{i|d-k}\|^4} \leq \frac{1}{2}$. *Then the max-margin predictor* $\hat{\mathbf{w}} = \arg\min \|\mathbf{w}\| : \forall i \in [m], y_i \mathbf{x}_i^\top \mathbf{w} \geq 1$ *(assuming it exists) satisfies the following:*

$$\hat{\mathbf{w}}_{|k} \;=\; \arg\min_{\mathbf{v} \in \mathbb{R}^k} \; (1+\epsilon_{\mathbf{v}}) \cdot \hat{\mathbb{E}}\left[\frac{[1 - y\mathbf{x}_{|k}^\top \mathbf{v}]_+^2}{\|\mathbf{x}_{|d-k}\|^2}\right] + \frac{\|\mathbf{v}\|^2}{m} \quad and \quad \|\hat{\mathbf{w}}_{|d-k}\|^2 \;\leq\; \frac{5m}{\min_{i \in [m]} \|\mathbf{x}_{i|d-k}\|^2} \;,$$

where $\epsilon_{\mathbf{v}}$ *satisfies* $\sup_{\mathbf{v} \in \mathbb{R}^k} |\epsilon_{\mathbf{v}}| \leq \epsilon_0$, *and.*

Thus, we see that $\hat{\mathbf{w}}_{|k}$ is essentially the minimizer of the empirical average of the (weighted) squared hinge loss discussed earlier, plus a certain regularization term which decays with the data size m. This is modified by a $(1 + \epsilon_{\mathbf{v}})$ multiplicative parameter, where $|\epsilon_{\mathbf{v}}| \leq \epsilon_0$ converges uniformly to 0 as $\|E\| \to 0$. As to $\hat{\mathbf{w}}_{|d-k}$, as in the case of regression, we bound its norm by an expression generally scaling with m but not with d, which implies that its contribution to the prediction (assuming a high-dimensional distribution on these coordinates) is negligible as $d \to \infty$.

Before continuing, let us informally explain how this squared hinge loss arises in our analysis (with the formal proof deferred as usual to the appendix). To simplify matters, let us suppose that $\{\mathbf{x}_{i|d-k}\}_{i=1}^m$ are precisely orthogonal (so that $E = 0$ in the theorem above). In that case, the max-margin predictor $\hat{\mathbf{w}}$ can be equivalently written as

$$\arg\min_{\mathbf{w} \in \mathbb{R}^d} \|\mathbf{w}_{|k}\|^2 + \|\mathbf{w}_{|d-k}\|^2 \quad : \quad \forall i \in [m], \; y_i \mathbf{x}_{i|k}^\top \mathbf{w}_{|k} + y_i \mathbf{x}_{i|d-k}^\top \mathbf{w}_{|d-k} \geq 1 \;. \tag{6}$$

For any fixed $\mathbf{w}_{|k}$, we therefore wish to make $\|\mathbf{w}_{|d-k}\|^2$ as small as possible, while satisfying the constraints, which can also be written as $\forall i \in [m], \; y_i \mathbf{x}_{i|d-k}^\top \mathbf{w}_{|d-k} \geq 1 - y_i \mathbf{x}_{i|k}^\top \mathbf{w}_{|k}$. Since $\{y_i \mathbf{x}_{i|d-k}\}_{i=1}^m$ are orthogonal, it is easy to see that we should pick $\hat{\mathbf{w}}_{|d-k}$ as follows: If $1 - y_i \mathbf{x}_{i|k}^\top \mathbf{w}_{|k} \leq 0$, we should make $y_i \mathbf{x}_{i|d-k}^\top \mathbf{w}_{|d-k} = 0$, and if $1 - y_i \mathbf{x}_{i|k}^\top \mathbf{w}_{|k} > 0$, we should make $y_i \mathbf{x}_{i|d-k}^\top \mathbf{w}_{|d-k} = 1 - y_i \mathbf{x}_{i|k}^\top \mathbf{w}_{|k}$. By orthogonality of the $\{\mathbf{x}_{i|d-k}\}_{i=1}^m$ vectors, it follows that the optimal $\hat{\mathbf{w}}_{|d-k}$ equals

$$\sum_{i:1-y_i \mathbf{x}_{i|k}^\top \mathbf{w}_{|k} > 0} \left(1 - y_i \mathbf{x}_{i|k}^\top \mathbf{w}_{|k}\right) \cdot \frac{y_i \mathbf{x}_{i|d-k}}{\|\mathbf{x}_{i|d-k}\|^2} \;=\; \sum_i \left[1 - y_i \mathbf{x}_{i|k}^\top \mathbf{w}_{|k}\right]_+ \cdot \frac{y_i \mathbf{x}_{i|d-k}}{\|\mathbf{x}_{i|d-k}\|^2} \;.$$

Again by orthogonality, it follows that $\|\mathbf{w}_{|d-k}\|^2 = \sum_i \frac{[1 - y_i \mathbf{x}_{i|k}^\top \mathbf{w}_{|k}]_+^2}{\|\mathbf{x}_{i|d-k}\|^2}$. Plugging this into Eq. (6), we get that $\hat{\mathbf{w}}_{|k}$ equals

$$\arg\min_{\mathbf{w}_{|k} \in \mathbb{R}^k} \|\mathbf{w}_{|k}\|^2 + \sum_{i=1}^m \frac{[1 - y_i \mathbf{x}_{i|k}^\top \mathbf{w}_{|k}]_+^2}{\|\mathbf{x}_{i|d-k}\|^2} \;=\; \arg\min_{\mathbf{w}_{|k} \in \mathbb{R}^k} \frac{\|\mathbf{w}_{|k}\|^2}{m} + \frac{1}{m}\sum_{i=1}^m \frac{[1 - y_i \mathbf{x}_{i|k}^\top \mathbf{w}_{|k}]_+^2}{\|\mathbf{x}_{i|d-k}\|^2} \;.$$

Substituting $\mathbf{v}$ instead of $\mathbf{w}_{|k}$ results in the expression for $\hat{\mathbf{w}}_{|k}$ appearing in the theorem (with $\epsilon_{\mathbf{v}} = 0$). The proof of Theorem 12 essentially generalizes this argument to the case where $\{\mathbf{x}_{i|d-k}\}_{i=1}^m$ are only approximately orthogonal, and also provides a bound for $\|\hat{\mathbf{w}}_{|d-k}\|$.

5.2 Asymptotic characterization of the max-margin predictor

Having established the perturbation bound in the previous section, let us now show how this can lead to an asymptotic characterization of the max-margin predictor, under suitable distributional assumptions. As in the case of regression, we present a set of sufficient conditions (which are not the most general possible):

Assumption 2 *Suppose* $\{\mathcal{D}_d\}_{d=k+1}^{\infty}$ *is a sequence of distributions on* $\mathbb{R}^d \times \mathbb{R}$, *and* $\{m_d\}_{d=k+1}^{\infty}$ *a monotonically increasing sequence of positive integers diverging to* ∞, *such that the following hold:*

1. *Letting* $\mathbb{E}_d$ *be shorthand for* $\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_d}$, *it holds that* $\sup_d \mathbb{E}_d\left[\frac{1+\|y\mathbf{x}_k\|^4}{\|\mathbf{x}_{|d-k}\|^4}\right] < \infty$.

2. *With probability approaching 1 over sampling* m_d *samples from* $\mathcal{D}_d$, *the max-margin predictor* $\hat{\mathbf{w}}_d$ *exists, and* $\max_{i\in[m_d]} \max\left\{\frac{1}{\|\mathbf{x}_{i|d-k}\|}, \|\mathbf{x}_{i|d-k}\|\right\} \leq c$ *for some constant* $c > 0$ *independent of* d.

3. $m_d \cdot \|E\| \xrightarrow{P} 0$, *where* E *is as defined in Eq. (5) w.r.t. an i.i.d. sample of size* m_d *from* $\mathcal{D}_d$.

4. $m_d \cdot \|\mathbb{E}_d[\mathbf{x}_{|d-k}\mathbf{x}_{|d-k}^{\top}]\| \to 0$.

5. *There exists some constant* $c' > 0$ *independent of* d, *such that with probability approaching 1, the function* $\hat{g}_d(\mathbf{v}) := \hat{\mathbb{E}}_d\left[\frac{[1-y\mathbf{x}_{|k}^{\top}\mathbf{v}]_+^2}{\|\mathbf{x}_{|d-k}\|^2}\right] = \frac{1}{m_d}\sum_{i=1}^{m_d}\frac{[1-y_i\mathbf{x}_{i|k}^{\top}\mathbf{v}]_+^2}{\|\mathbf{x}_{i|d-k}\|^2}$ *has a minimizer with norm at most* c'.

6. *Letting*

$$g_d(\mathbf{v}) := \mathbb{E}_d\left[\frac{[1-y\mathbf{x}_{|k}^{\top}\mathbf{v}]_+^2}{\|\mathbf{x}_{|d-k}\|^2}\right] \ ,$$

   it holds that $\inf_{\mathbf{v}} \limsup_d (g_d(\mathbf{v}) - \inf_{\mathbf{u}} g_d(\mathbf{u})) = 0$.

The first four conditions are similar to the condition in Assumption 1 for regression, with the only differences being that conditions 1 and 2 require slightly different functions of the data to be bounded. The rest of the assumptions are very mild when we think of the distribution of $y, \mathbf{x}_{|k}, \|\mathbf{x}_{|d-k}\|$ as converging to some fixed distribution as $d \to \infty$ (for instance, in the setting of Example 1). In that case, we would expect the minimum-norm minimizers of $\hat{\mathbb{E}}_d\left[\frac{[1-y\mathbf{x}_{|k}^{\top}\mathbf{v}]_+^2}{\|\mathbf{x}_{|d-k}\|^2}\right]$ to converge to some fixed limit (as $d, m_d \to \infty$), and therefore condition 5 should automatically hold. Similarly, condition 6 should also hold, as it is equivalent to requiring that the set of near-minimizers of the functions $\{g_d(\cdot)\}$ asymptotically overlap.

With these conditions at hand, we can now state our asymptotic characterization of the max-margin predictor, in terms of the expected (weighted) squared hinge loss function $g_d(\cdot)$ defined above:

Theorem 13 *Suppose* $\{\mathcal{D}_d\}_{d=k+1}^{\infty}$ *and* $\{m_d\}_{d=k+1}^{\infty}$ *satisfy Assumption 2. Then the max-margin predictor* $\hat{\mathbf{w}}_d = \arg\min \|\mathbf{w}\| : \forall i \in [m_d],\ y_i\mathbf{x}_i^\top\mathbf{w} \geq 1$ *satisfies*

$$g_d(\hat{\mathbf{w}}_{d|k}) - \inf_{\mathbf{v}} g_d(\mathbf{v}) \xrightarrow{P} 0 \quad and \quad \mathbb{E}_d\left[(\mathbf{x}^\top\hat{\mathbf{w}}_d - \mathbf{x}_{|k}^\top\hat{\mathbf{w}}_{d|k})^2\right] \xrightarrow{P} 0 \ .$$

It is interesting to note that the max-margin predictor is asymptotically characterized in terms of a squared hinge loss, even though this loss does not appear explicitly in its definition (and moreover, the max-margin predictor itself arises from training gradient-based methods on losses which are definitely not the squared hinge loss). Instead, the loss naturally arises from our analysis. We note that this loss achieves the same value as the square loss for examples $(\mathbf{x}, y)$ where $\mathbf{x}^\top\hat{\mathbf{w}}_d = y$, but is otherwise distinct. Thus, there is no contradiction with previous results on benign overfitting in classification that focused on situations where the max-margin and min-norm predictors coincide.

5.3 Implications for benign overfitting in classification

Theorem 13 implies that in our binary classification model, the max-margin predictor $\hat{\mathbf{w}}$ is such that its last $d - k$ coordinates are asymptotically immaterial, whereas the first k coordinates asymptotically minimize the function $g_d(\mathbf{v}) = \mathbb{E}_d\left[\frac{[1-y\mathbf{x}_{|k}^\top\mathbf{v}]_+^2}{\|\mathbf{x}_{|d-k}\|^2}\right]$. Thus, the next natural step is to understand whether the minimizers of this function on the first k coordinates result in good predictors with respect to the misclassification error $\Pr(y\mathbf{x}^\top\mathbf{w} \leq 0)$. We note that since this loss is not identical to misclassification error, we cannot hope this to hold in the worst-case: Indeed, see Appendix B for an explicit example. . However, it is also not an unreasonable requirement, since predictors that attempt to minimize $\frac{[1-y\mathbf{x}_{|k}^\top\mathbf{v}]_+^2}{\|\mathbf{x}_{|d-k}\|^2}$ over $\mathbf{v}$ will also tend to make $y\mathbf{x}_{|k}^\top\mathbf{v}$ positive, and hence (since the last $d - k$ coordinates have negligible effect) make the expected misclassification error $\Pr(y\mathbf{x}^\top\mathbf{w} \leq 0)$ small. A different way to phrase this question is whether the data distribution is such that the weighted squared hinge loss function $\mathbf{v} \mapsto \frac{[1-y\mathbf{x}_{|k}^\top\mathbf{v}]_+^2}{\|\mathbf{x}_{|d-k}\|^2}$ is a good *surrogate loss function* for the misclassification error. We note that this question has already been studied for other surrogate losses (see Frei et al. (2021); Ji et al. (2022) for some recent examples).

Our results imply that asymptotically, the max-margin predictor depends only on the joint distribution of $\mathbf{x}_{|k}, \|\mathbf{x}_{|d-k}\|, y$. Thus, to simplify the discussion in the remainder of this section, we will assume this distribution is fixed for all d, and satisfies some mild conditions:

Assumption 3 *The distribution sequence* $\{\mathcal{D}_d\}_{d=k+1}^{\infty}$ *on* $(\mathbf{x}, y) \in \mathbb{R}^d \times \{-1, +1\}$ *is such that the joint distribution of* $(\mathbf{x}_{|k}, \|\mathbf{x}_{|d-k}\|, y)$ *is the same under any* d. *Moreover, this distribution* $\mathcal{D}$ *is such that* $\mathbb{E}[\mathbf{x}_{|k}\mathbf{x}_{|k}^\top]$ *is positive definite, and* $\Pr(\|\mathbf{x}_{|d-k}\| \in [l, u]) = 1$ *for some* $[l, u] \subset (0, \infty)$.

To give a simple example, consider the case where $\mathcal{D}_d$ is defined as some fixed distribution over $(\mathbf{x}_{|k}, y)$, and the marginal distribution of $\mathbf{x}_{|d-k}$ is uniform over some origin-centered
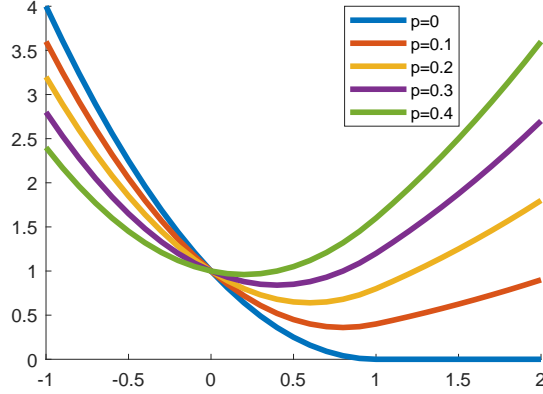
Figure 1: Graphical illustration of the function $\ell_p(\cdot)$ from Eq. (7) for different values of p.

sphere[2]. Under this assumption, the function g_d can be rewritten as

$$g(\mathbf{v}) \; := \; \mathbb{E}_{\mathbf{x}_{|k}, z, y}\left[z \cdot [1 - y\mathbf{x}_{|k}^\top \mathbf{v}]_+^2\right] \quad , \quad \mathbf{v} \in \mathbb{R}^k$$

for some fixed distribution over $\mathbf{x}_{|k} \in \mathbb{R}^k$, $y \in \{-1, +1\}$, and $z \in (0, \infty)$.

Our goal now will be to illustrate how our characterization allows us to prove that benign overfitting does occur in some classification setups, which to the best of our knowledge have not been explicitly studied before.

For benign overfitting to occur, we need situations where no predictor attains zero error w.r.t. the underlying data distribution. In binary classification setups, the simplest (and most well-studied) case where this occurs is when we have an underlying distribution $\mathcal{D}_{\text{clean}}$ which is *linearly separable* w.r.t. $(\mathbf{x}_{|k}, y)$ (i.e., there is some unit vector $\mathbf{w}^* \in \mathbb{R}^k$ such that $\text{Pr}_{\mathcal{D}_{\text{clean}}}(y\mathbf{x}_{|k}^\top \mathbf{w}^* < \gamma) = 0$ for some margin parameter $\gamma > 0$), but where there is random label noise (with each y flipped to $-y$ with some probability $p > 0$), resulting in a final distribution $\mathcal{D}$. In such a distribution, $\mathbf{w}^*$ is still an optimal predictor, but now necessarily its expected misclassification error equals p. To model this setting, it will be convenient to assume that $(\mathbf{x}_{|k}, y)$ is still distributed as $\mathcal{D}_{\text{clean}}$, and that we wish to find a predictor $\mathbf{w}$ satisfying $\text{Pr}_{\mathcal{D}\text{clean}}(y\mathbf{x}_{|k}^\top \mathbf{w} \le 0) = 0$. However, the predictor $\hat{\mathbf{w}}$ we learn is with respect to the "noisy" labels, and is characterized by the weighted squared hinge loss: Namely, by Theorem 13, $\hat{\mathbf{w}}_{d|k}$ is asymptotically the minimizer of

$$L_p(\mathbf{w}) \; := \; \mathbb{E}_{(\mathbf{x}_{|k}, z, y) \sim \mathcal{D}_{\text{clean}}}\left[(1 - p) \cdot z \cdot [1 - y\mathbf{x}_{|k}^\top \mathbf{w}]_+^2 + p \cdot z \cdot [1 + y\mathbf{x}_{|k}^\top \mathbf{w}]_+^2\right] \tag{7}$$

$$= \; \mathbb{E}_{(\mathbf{x}_{|k}, z, y) \sim \mathcal{D}_{\text{clean}}} \mathbb{E}[z \cdot \ell_p(y\mathbf{x}_{|k}^\top \mathbf{w})] \quad \text{where} \quad \ell_p(\beta) := (1 - p) \cdot [1 - \beta]_+^2 + p \cdot [1 + \beta]_+^2$$

(see Figure 1 for a graphical illustration). It is easily verified that for any $p \in (0, \frac{1}{2})$, ℓ_p is a strongly convex function, and therefore $L_p(\cdot)$ is a strongly convex function, as long as

---

2. One can also consider the case of $\mathbf{x}_{|d-k}$ being a zero-mean Gaussian with covariance matrix $\frac{\sigma^2}{d-k}I$ for some $\sigma^2 > 0$, in which case $\|\mathbf{x}_{|d-k}\|$ will concentrate around σ. In the assumption, we slightly simplify this by assuming $\|\mathbf{x}_{|d-k}\|$ already has such a limit distribution.

$\mathbb{E}[z \cdot \mathbf{x}_{|k}\mathbf{x}_{|k}^{\top}]$ is positive definite (see Lemma 19 in the appendix for a formal definition of strong convexity and a proof). Therefore, $L_p(\cdot)$ has a unique minimizer $\mathbf{w}_p^*$, to which $\hat{\mathbf{w}}_{d|k}$ converges to. Overall, we get that for benign overfitting, it is sufficient that $\mathbf{w}_p^*$ is an optimal predictor in terms of misclassification error on the "clean" labels. This is formalized in the following theorem:

Theorem 14 *Under the conditions of Theorem 13 and Assumption 3, define $R_d(\mathbf{w})$ to equal $\mathrm{Pr}_{(\mathbf{x},y)\sim\mathcal{D}_d}(y\mathbf{x}^{\top}\mathbf{w} \leq 0)$. Also, suppose that the distribution of $(\mathbf{x}_{|k}, y)$ corresponds to some linearly separable distribution $\mathcal{D}_{clean}$ with labels flipped with some probability $p \in (0, \frac{1}{2})$.*

Then benign overfitting (as defined in Eq. (3)) holds under the following condition: The (unique) minimizer $\mathbf{w}_p^*$ of $L_p(\mathbf{w})$ satisfies $\mathrm{Pr}_{(\mathbf{x}_{|k},y)\sim\mathcal{D}_{clean}}(y\mathbf{x}_{|k}^{\top}\mathbf{w}_p^* \leq 0) = 0$.

Focusing on such linearly-separable-with-label-noise distributions, we now turn to study some cases where the condition on $\mathbf{w}_p^*$ in Theorem 14 indeed holds. For example, the following theorem implies that under mild assumptions, just about *any* choice of distribution $\mathcal{D}_{\text{clean}}$ satisfies benign overfitting, for some non-trivial (distribution-dependent) regime of label noise. As far as we can surmise, this is not at all obvious from the original characterization of the max-margin predictor, where the data points appear as constraints and where introducing label noise changes these constraints in possibly complicated ways. However, using our characterization and properties of the squared hinge loss, the result follows from a rather straightforward continuity argument.

Theorem 15 *Fix any distribution $\mathcal{D}_{clean}$ over $(\mathbf{x}_{|k}, z, y)$ satisfying Assumption 3, which is linearly separable w.r.t. $(\mathbf{x}_{|k}, y)$, and where $\mathbf{x}_{|k}$ has bounded support. Then there exists some $a \in (0, \frac{1}{2})$ (dependent on $\mathcal{D}_{clean}$), such that for all $p \in (0, a)$, the minimizer $\mathbf{w}_p^*$ of $L_p(\cdot)$ satisfies $\mathrm{Pr}_{(\mathbf{x}_{|k},y)\sim\mathcal{D}_{clean}}(y\mathbf{x}_{|k}^{\top}\mathbf{w}_p^* \leq 0) = 0$.*

Intuitively, the proof proceeds by arguing that $\mathbf{w}_p^*$ is continuous in $p \in (0, \frac{1}{2})$, and as $p \to 0$, necessarily converges to some predictor which separates the data with positive margin. Hence, small perturbations of the predictor will maintain linear separability, and therefore $\mathbf{w}_p^*$ will remain a linear separator for small positive values of p.

This result holds for generic linearly separable distributions, but does not specify the amount of label noise under which benign overfitting occurs. In the following theorem, we identify one simple class of distributions where benign overfitting occurs with any amount of label noise up to $\frac{1}{2}$:

Theorem 16 *Fix any distribution $\mathcal{D}_{clean}$ on $(\mathbf{x}_{|k}, z, y)$ satisfying Assumption 3, such that the distribution of $(\mathbf{x}_{|k}, y)$ is linearly separable. Moreover, suppose that for some unit vector $\mathbf{u}$, and conditioned on any value of y, $\mathbf{u}^{\top}\mathbf{x}$ and $(I - \mathbf{u}\mathbf{u}^{\top})\mathbf{x}$ are mutually independent, and the distributions of $(I - \mathbf{u}\mathbf{u}^{\top})\mathbf{x}$ and $-(I - \mathbf{u}\mathbf{u}^{\top})\mathbf{x}$ are identical. Then for all $p \in (0, \frac{1}{2})$, the minimizer $\mathbf{w}_p^*$ of $L_p(\cdot)$ satisfies $\mathrm{Pr}_{\mathcal{D}_{clean}}(y\mathbf{x}_{|k}^{\top}\mathbf{w}_p^* \leq 0) = 0$.*

The conditions in the theorem refer to a situation where there is some distinguished direction $\mathbf{u}$, such that conditioned on y, $\mathbf{x}_{|k} = \beta\mathbf{u} + \mathbf{s}$ for some mutually independent random variables $\beta, \mathbf{s}$ where $\mathbf{s}$ is orthogonal to $\mathbf{u}$ and has a symmetric distribution. Examples where this occurs include any one-dimensional distribution, and a mixture of any two symmetric

distributions with means in span($\mathbf{u}$) (one for $y = 1$ and one for $y = -1$, and assuming linear separability). Note that unlike most previous results on benign overfitting in classification, the distributions do not need to be identical nor satisfy any additional structural properties.

6. Discussion

In this paper, we presented several new results on benign overfitting, for both regression and classification. For linear regression with the square loss, we argued that benign overfitting should not be expected to hold in general, once we go beyond well-specified distributions. Moreover, we showed how this can be extended beyond linear regression with the square loss, by an argument proving how the existence of benign overfitting on some regression problems precludes its existence on other regression problems. On the more positive side, for classification problems, we showed that the max-margin is implicitly biased towards minimizing a weighted squared hinge loss w.r.t. the underlying distribution (at least in a model where an arbitrary k-dimensional distribution is concatenated with a high-dimensional distribution). We use it to show benign overfitting in various settings, by considering cases where this squared hinge loss is a good surrogate for the misclassification error.

Overall, we hope that our observations here will allow us to understand benign overfitting beyond the settings studied so far in the literature. For example, it would be interesting to identify other settings where the structure of the squared hinge loss means that the max-margin predictor will have benign overfitting properties. Moreover, our results focused on input distributions with a clean separation between a few "important" coordinates, and a high dimensional distribution on the other coordinates. Although this is a prototypical setting for benign overfitting, our insights can potentially be extended to other input distributions, and identifying them can be an interesting direction for future research.

Another, more technical issue is that our asymptotic characterization of the min-norm or max-margin predictor require the high-dimensional distribution to be sufficiently "spread", which ultimately requires the dimension d to scale sufficiently faster than the sample size m_d. For instance, as discussed in Example 1, if we consider a distribution on $\mathbf{x} \in \mathbb{R}^d$ such that $\mathbf{x}_{|d-k}$ is zero-mean Gaussian with covariance matrix $\frac{1}{d-k} \cdot I$, then we need $m_d^3 \log(d)/d \to 0$ for the perturbation matrix E to decay sufficiently fast, and for Theorem 6 to hold. A similar requirement also applies to our classification results. For the purposes of our paper, this is not a major issue, since our goal was to understand when benign overfitting might or might not occur assuming the dimension is sufficiently large, and indeed we did not attempt to optimize this condition. Nevertheless, since other papers usually assume a milder scaling of d vs. m_d, it would be interesting to understand whether results similar to ours can be obtained under such conditions.

Acknowledgments

Appendix A. Proofs

A.1 Proof of Theorem 5

We will utilize the following matrix inverse perturbation result:

Lemma 17 *Let A be some positive definite matrix with minimal eigenvalue $\lambda_{\min}(A) > 0$. Then for any symmetric matrix E of the same size such that $\|E\| \leq \frac{\lambda_{\min}(A)}{2}$, it holds that $A + E$ is invertible and*

$$\|(A+E)^{-1} - A^{-1}\| \leq \frac{2}{\lambda_{\min}(A)^2}\|E\|.$$

Proof By Weyl's inequality, $\lambda_{\min}(A + E) \geq \lambda_{\min}(A) - \|E\| \geq \frac{\lambda_{\min}(A)}{2} > 0$, hence $A + E$ is invertible. Moreover, by the Woodbury matrix identity,

$$(A+E)^{-1} = A^{-1} - A^{-1}E(I + A^{-1}E)^{-1}A^{-1},$$

which implies

$$\|(A+E)^{-1} - A^{-1}\| \leq \|A^{-1}\|^2 \cdot \|E\| \cdot \|(I + A^{-1}E)^{-1}\| \leq \frac{\|A^{-1}\|^2 \cdot \|E\|}{1 - \|A^{-1}\| \cdot \|E\|}.$$

Noting that $\|A^{-1}\| = \frac{1}{\lambda_{\min}(A)}$ and $\|E\| \leq \frac{\lambda_{\min}(A)}{2}$, it follows that the above is at most

$$\frac{\frac{1}{\lambda_{\min}(A)^2}\|E\|}{1 - \frac{1}{\lambda_{\min}(A)} \cdot \frac{\lambda_{\min}(A)}{2}} = \frac{2}{\lambda_{\min}(A)^2}\|E\|. \qquad \blacksquare$$

We now turn to the proof itself. Let $X \in \mathbb{R}^{m \times d}$ be the matrix whose i-th row is $\mathbf{x}_i$, and let $X_{|k}, X_{|d-k}$ be its first k and last $d - k$ columns respectively (so that the i-th row of $X_{|k}$ is $\mathbf{x}_{i|k}$, and that of $X_{|d-k}$ is $\mathbf{x}_{i|d-k}$). Also, let $\mathbf{y} = (y_1, \ldots, y_m)$. Since $\{\mathbf{x}_i\}_{i=1}^m$ are linearly independent, X has full row rank. Therefore, $XX^\top$ is invertible, and $\hat{\mathbf{w}} = \arg\min \|\mathbf{w}\| : X\mathbf{w} = \mathbf{y}$ can be written in closed form as

$$\hat{\mathbf{w}} = X^\top(XX^\top)^{-1}\mathbf{y} = X^\top \left(X_{|k}X_{|k}^\top + X_{|d-k}X_{|d-k}^\top\right)^{-1}\mathbf{y} = X^\top \left(D + X_{|k}X_{|k}^\top + E\right)^{-1}\mathbf{y},$$

where E is as defined in the theorem statement (namely, the off-diagonal entries of $X_{|d-k}X_{|d-k}^\top$), and

$$D := X_{|d-k}X_{|d-k}^\top - E = \mathrm{diag}(\|\mathbf{x}_{i|d-k}\|^2, \ldots, \|\mathbf{x}_{m|d-k}\|^2)$$

is a diagonal matrix. In what follows, it will be useful to note that $\lambda_{\min}(D) = \min_{i \in [m]} \|\mathbf{x}_{m|d-k}\|^2$.

Continuing, we have by the representation of $\hat{\mathbf{w}}$ above that

$$\hat{\mathbf{w}} = X^\top(D + X_{|k}X_{|k}^\top)^{-1}\mathbf{y} + X^\top E'\mathbf{y} \quad \text{where} \quad E' = (D + X_{|k}X_{|k}^\top + E)^{-1} - (D + X_{|k}X_{|k}^\top)^{-1}. \tag{8}$$

Considering the first k and last $d - k$ coordinates separately, it follows that

$$\hat{\mathbf{w}}_{|k} = X_{|k}^\top(D + X_{|k}X_{|k}^\top)^{-1}\mathbf{y} + X_{|k}^\top E'\mathbf{y} \quad \text{and} \quad \hat{\mathbf{w}}_{|d-k} = X_{|d-k}^\top(D + X_{|k}X_{|k}^\top)^{-1}\mathbf{y} + X_{|d-k}^\top E'\mathbf{y}, \tag{9}$$

23

where E' is as defined in Eq. (8).

Applying Lemma 17 with $A = D + X_{|k}X_{|k}^\top$ (noting that $\lambda_{\min}(D + X_{|k}X_{|k}^\top) \geq \lambda_{\min}(D) = \min_{i \in [m]} \|\mathbf{x}_{i|d-k}\|^2 > 0$ and that we assume $\|E\| \leq \frac{\lambda_{\min}(D)}{2}$), it follows that $\|E'\| \leq \frac{2}{\lambda_{\min}(D)^2}\|E\|$. As a result, and using the fact that the spectral norm is upper bounded by the Frobenius norm, we have

$$
\begin{aligned}
\|X_{|k}^\top E' \mathbf{y}\| &\leq \|X_{|k}\| \cdot \|\mathbf{y}\| \cdot \|E'\| \leq m \cdot \sqrt{\frac{1}{m}\|X_{|k}\|_F^2} \cdot \sqrt{\frac{1}{m}\|\mathbf{y}\|^2} \cdot \frac{2}{\lambda_{\min}(D)^2}\|E\| \\
&= \frac{2m\|E\|}{\lambda_{\min}(D)^2} \cdot \sqrt{\hat{\mathbb{E}}[\|\mathbf{x}_{|k}\|^2] \cdot \hat{\mathbb{E}}[y^2]} \, .
\end{aligned}
$$

An identical calculation implies that $\|X_{|d-k}^\top E' \mathbf{y}\| \leq \frac{2m\|E\|}{\lambda_{\min}(D)^2} \cdot \sqrt{\hat{\mathbb{E}}[\|\mathbf{x}_{|d-k}\|^2] \cdot \hat{\mathbb{E}}[y^2]}$. Plugging these bounds back into Eq. (9), it follows that

$$
\left\| \hat{\mathbf{w}}_{|k} - X_{|k}^\top(D + X_{|k}X_{|k}^\top)^{-1}\mathbf{y} \right\| \leq \frac{2m\|E\|}{\lambda_{\min}(D)^2} \cdot \sqrt{\hat{\mathbb{E}}[\|\mathbf{x}_{|k}\|^2] \cdot \hat{\mathbb{E}}[y^2]} \tag{10}
$$

and

$$
\left\| \hat{\mathbf{w}}_{|d-k} - X_{|d-k}^\top(D + X_{|k}X_{|k}^\top)^{-1}\mathbf{y} \right\| \leq \frac{2m\|E\|}{\lambda_{\min}(D)^2} \cdot \sqrt{\hat{\mathbb{E}}[\|\mathbf{x}_{|d-k}\|^2] \cdot \hat{\mathbb{E}}[y^2]} \, . \tag{11}
$$

We now turn to analyze $X_{|k}^\top(D + X_{|k}X_{|k}^\top)^{-1}\mathbf{y}$ and $X_{|d-k}^\top(D + X_{|k}X_{|k}^\top)^{-1}\mathbf{y}$, starting with the first expression. Using the Woodbury matrix identity, we have

$$
\begin{aligned}
(D + X_{|k}X_{|k}^\top)^{-1} &= D^{-1} - D^{-1}X_{|k}(I + X_{|k}^\top D^{-1}X_{|k})^{-1}X_{|k}^\top D^{-1} \\
&= D^{-1}\left(I - X_{|k}(I + X_{|k}^\top D^{-1}X_{|k})^{-1}X_{|k}^\top D^{-1}\right)
\end{aligned}
$$

(note that $I + X_{|k}^\top D^{-1}X_{|k}$ is indeed invertible, since it is the sum of the identity matrix and a positive semidefinite matrix). This implies that

$$
\begin{aligned}
X_{|k}^\top\left(D + X_{|k}X_{|k}^\top\right)^{-1}\mathbf{y} &= X_{|k}^\top D^{-1}\left(I - X_{|k}(I + X_{|k}^\top D^{-1}X_{|k})^{-1}X_{|k}^\top D^{-1}\right)\mathbf{y} \\
&= \left(I - X_{|k}^\top D^{-1}X_{|k}\left(I + X_{|k}^\top D^{-1}X_{|k}\right)^{-1}\right)X_{|k}^\top D^{-1}\mathbf{y} \\
&= \left(\left(I + X_{|k}^\top D^{-1}X_{|k}\right) - X_{|k}^\top D^{-1}X_{|k}\right)\left(I + X_{|k}^\top D^{-1}X_{|k}\right)^{-1}X_{|k}^\top D^{-1}\mathbf{y} \\
&= \left(I + X_{|k}^\top D^{-1}X_{|k}\right)^{-1}X_{|k}^\top D^{-1}\mathbf{y} \\
&= \left(\frac{1}{m}I + \frac{1}{m}X_{|k}^\top D^{-1}X_{|k}\right)^{-1}\left(\frac{1}{m}X_{|k}^\top D^{-1}\mathbf{y}\right) \\
&= \left(\frac{1}{m}I + \hat{\mathbb{E}}\left[\frac{\mathbf{x}_{|k}\mathbf{x}_{|k}^\top}{\|\mathbf{x}_{|d-k}\|^2}\right]\right)^{-1} \cdot \hat{\mathbb{E}}\left[\frac{y\mathbf{x}_{|k}}{\|\mathbf{x}_{|d-k}\|^2}\right] \, . \tag{12}
\end{aligned}
$$

Using Lemma 17 and the assumption that $\frac{1}{m} \leq \frac{1}{2}\lambda_{\min}\left(\hat{\mathbb{E}}\left[\frac{\mathbf{x}_{|k}\mathbf{x}_{|k}^\top}{\|\mathbf{x}_{|d-k}\|^2}\right]\right)$, it follows that

$$\left\|\left(\frac{1}{m}I + \hat{\mathbb{E}}\left[\frac{\mathbf{x}_{|k}\mathbf{x}_{|k}^\top}{\|\mathbf{x}_{|d-k}\|^2}\right]\right)^{-1} - \left(\hat{\mathbb{E}}\left[\frac{\mathbf{x}_{|k}\mathbf{x}_{|k}^\top}{\|\mathbf{x}_{|d-k}\|^2}\right]\right)^{-1}\right\| \leq \frac{2}{\lambda_{\min}\left(\hat{\mathbb{E}}\left[\frac{\mathbf{x}_{|k}\mathbf{x}_{|k}^\top}{\|\mathbf{x}_{|d-k}\|^2}\right]\right)^2 \cdot m}.$$

Combining this with Eq. (12) and the Cauchy-Schwarz inequality, it follows that

$$\left\|X_{|k}^\top\left(D + X_{|k}X_{|k}^\top\right)^{-1}\mathbf{y} - \left(\hat{\mathbb{E}}\left[\frac{\mathbf{x}_{|k}\mathbf{x}_{|k}^\top}{\|\mathbf{x}_{|d-k}\|^2}\right]\right)^{-1}\cdot\hat{\mathbb{E}}\left[\frac{y\mathbf{x}_{|k}}{\|\mathbf{x}_{|d-k}\|^2}\right]\right\|$$

$$= \left\|\left(\left(\frac{1}{m}I + \hat{\mathbb{E}}\left[\frac{\mathbf{x}_{|k}\mathbf{x}_{|k}^\top}{\|\mathbf{x}_{|d-k}\|^2}\right]\right)^{-1} - \left(\hat{\mathbb{E}}\left[\frac{\mathbf{x}_{|k}\mathbf{x}_{|k}^\top}{\|\mathbf{x}_{|d-k}\|^2}\right]\right)^{-1}\right)\cdot\hat{\mathbb{E}}\left[\frac{y\mathbf{x}_{|k}}{\|\mathbf{x}_{|d-k}\|^2}\right]\right\|$$

$$\leq \left\|\left(\frac{1}{m}I + \hat{\mathbb{E}}\left[\frac{\mathbf{x}_{|k}\mathbf{x}_{|k}^\top}{\|\mathbf{x}_{|d-k}\|^2}\right]\right)^{-1} - \left(\hat{\mathbb{E}}\left[\frac{\mathbf{x}_{|k}\mathbf{x}_{|k}^\top}{\|\mathbf{x}_{|d-k}\|^2}\right]\right)^{-1}\right\|\cdot\left\|\hat{\mathbb{E}}\left[\frac{y\mathbf{x}_{|k}}{\|\mathbf{x}_{|d-k}\|^2}\right]\right\|$$

$$\leq \frac{2}{\lambda_{\min}\left(\hat{\mathbb{E}}\left[\frac{\mathbf{x}_{|k}\mathbf{x}_{|k}^\top}{\|\mathbf{x}_{|d-k}\|^2}\right]\right)^2 \cdot m}\cdot\left\|\hat{\mathbb{E}}\left[\frac{y\mathbf{x}_{|k}}{\|\mathbf{x}_{|d-k}\|^2}\right]\right\|.$$

Combining the above with Eq. (10) using a triangle inequality, it follows that

$$\left\|\hat{\mathbf{w}}_{|k} - \left(\hat{\mathbb{E}}\left[\frac{\mathbf{x}_{|k}\mathbf{x}_{|k}^\top}{\|\mathbf{x}_{|d-k}\|^2}\right]\right)^{-1}\cdot\hat{\mathbb{E}}\left[\frac{y\mathbf{x}_{|k}}{\|\mathbf{x}_{|d-k}\|^2}\right]\right\|$$

$$\leq \frac{2\left\|\hat{\mathbb{E}}\left[\frac{y\mathbf{x}_{|k}}{\|\mathbf{x}_{|d-k}\|^2}\right]\right\|}{\lambda_{\min}\left(\hat{\mathbb{E}}\left[\frac{\mathbf{x}_{|k}\mathbf{x}_{|k}^\top}{\|\mathbf{x}_{|d-k}\|^2}\right]\right)^2 \cdot m} + \frac{2m\|E\|}{\lambda_{\min}(D)^2}\cdot\sqrt{\hat{\mathbb{E}}[\|\mathbf{x}_{|k}\|^2]\cdot\hat{\mathbb{E}}[y^2]}.$$

Recalling that $\lambda_{\min}(D) = \min_{i\in[m]}\|\mathbf{x}_{i|d-k}\|^2$, the first bound in the theorem follows.

As to bound on $\|\mathbf{w}_{|d-k}\|$ in the theorem, recalling Eq. (11), we need to analyze the expression $X_{|d-k}^\top(D + X_{|k}X_{|k}^\top)^{-1}\mathbf{y}$. By Cauchy-Schwarz, its norm is at most

$$\|X_{|d-k}\|\cdot\left\|(D + X_{|k}X_{|k}^\top)^{-1}\right\|\cdot\|\mathbf{y}\| = \frac{\|X_{|d-k}\|\cdot\|\mathbf{y}\|}{\lambda_{\min}(D + X_{|k}X_{|k}^\top)} \leq \frac{\|X_{|d-k}\|_F\cdot\|\mathbf{y}\|}{\lambda_{\min}(D)}$$

$$= \frac{m\cdot\sqrt{\hat{\mathbb{E}}[\|\mathbf{x}_{|d-k}\|^2]\cdot\hat{\mathbb{E}}[y^2]}}{\lambda_{\min}(D)}.$$

Combining this with Eq. (11), it follows that

$$\|\hat{\mathbf{w}}_{|d-k}\| \leq \frac{m\cdot\sqrt{\hat{\mathbb{E}}[\|\mathbf{x}_{|d-k}\|^2]\cdot\hat{\mathbb{E}}[y^2]}}{\lambda_{\min}(D)}\cdot\left(1 + \frac{2\|E\|}{\lambda_{\min}(D)}\right).$$

Recalling that $\lambda_{\min}(D) = \min_{i\in[m]}\|\mathbf{x}_{i|d-k}\|^2$, the second bound in the theorem follows.

A.2 Proof of Theorem 6

Assumption 1 implies that the law of large numbers holds with respect to the random variables $\|\mathbf{x}_{|k}\|^2$, $\|\mathbf{x}_{|d-k}\|^2$, $\frac{y\mathbf{x}_{|k}}{\|\mathbf{x}_{|d-k}\|^2}$ and $\frac{\mathbf{x}_{|k}\mathbf{x}_{|k}^\top}{\|\mathbf{x}_{|d-k}\|^2}$, so their empirical average over m_d i.i.d. instances (where $m_d \to \infty$) converges in probability to their (finite) expectations. Moreover, by assumption 2, $\min_{i \in [m_d]} \cdot \|\mathbf{x}_{i|d-k}\|$ is at least some positive constant with probability approaching 1, and $m_d\|E\|$ becomes arbitrarily small by assumption 3. All this implies that as $d, m_d \to \infty$, then with probability approaching 1, the conditions of Theorem 5 hold, and the upper bound in its first displayed equation is arbitrarily small. Hence, $\left\| \hat{\mathbf{w}}_{d|k} - \left(\hat{\mathbb{E}}_d \left[\frac{\mathbf{x}_{|k}\mathbf{x}_{|k}^\top}{\|\mathbf{x}_{|d-k}\|^2} \right] \right)^{-1} \hat{\mathbb{E}}_d \left[\frac{y\mathbf{x}_{|k}}{\|\mathbf{x}_{|d-k}\|^2} \right] \right\| \xrightarrow{P} 0$, where $\hat{\mathbb{E}}_d[\cdot]$ is the uniform distribution over $\{\mathbf{x}_i, y_i\}_{i=1}^{m_d}$ sampled i.i.d. from $\mathcal{D}_d$. To prove the first part of the theorem, it remains to show that the expression $\left(\hat{\mathbb{E}}_d \left[\frac{\mathbf{x}_{|k}\mathbf{x}_{|k}^\top}{\|\mathbf{x}_{|d-k}\|^2} \right] \right)^{-1} \hat{\mathbb{E}}_d \left[\frac{y\mathbf{x}_{|k}}{\|\mathbf{x}_{|d-k}\|^2} \right]$ converges in probability to $\left(\mathbb{E}_d \left[\frac{\mathbf{x}_{|k}\mathbf{x}_{|k}^\top}{\|\mathbf{x}_{|d-k}\|^2} \right] \right)^{-1} \mathbb{E}_d \left[\frac{y\mathbf{x}_{|k}}{\|\mathbf{x}_{|d-k}\|^2} \right]$. This holds, because again by the law of large numbers and our assumptions,

$$\left\| \hat{\mathbb{E}}_d \left[\frac{y\mathbf{x}_{|k}}{\|\mathbf{x}_{|d-k}\|^2} \right] - \mathbb{E}_d \left[\frac{y\mathbf{x}_{|k}}{\|\mathbf{x}_{|d-k}\|^2} \right] \right\| \xrightarrow{P} 0 \quad \text{and} \quad \left\| \hat{\mathbb{E}}_d \left[\frac{\mathbf{x}_{|k}\mathbf{x}_{|k}^\top}{\|\mathbf{x}_{|d-k}\|^2} \right] - \mathbb{E}_d \left[\frac{\mathbf{x}_{|k}\mathbf{x}_{|k}^\top}{\|\mathbf{x}_{|d-k}\|^2} \right] \right\| \xrightarrow{P} 0,$$

and moreover $\mathbb{E}_d \left[\frac{\mathbf{x}_{|k}\mathbf{x}_{|k}^\top}{\|\mathbf{x}_{|d-k}\|^2} \right]$ is positive definite by assumption 1, hence the convergence also holds with respect to the inverse of the matrices.

As to the second part, recalling that $\mathbb{E}_d$ is shorthand for $\mathbb{E}_{(\mathbf{x},y) \sim \mathcal{D}_d}$, and fixing some $\hat{\mathbf{w}}_d$, we have

$$\mathbb{E}_d \left[\left(\mathbf{x}^\top \hat{\mathbf{w}}_d - \hat{\mathbf{x}}_{|k}^\top \hat{\mathbf{w}}_{d|k} \right)^2 \right] = \mathbb{E}_d \left[\left(\mathbf{x}_{|d-k}^\top \hat{\mathbf{w}}_{d|d-k} \right)^2 \right] = \hat{\mathbf{w}}_{d|d-k}^\top \mathbb{E}_d[\mathbf{x}_{d-k}\mathbf{x}_{|d-k}^\top]\hat{\mathbf{w}}_{d|d-k}$$
$$\leq \|\hat{\mathbf{w}}_{d|d-k}\|^2 \cdot \|\mathbb{E}_d[\mathbf{x}_{d-k}\mathbf{x}_{|d-k}^\top]\| .$$

Assumptions $1, 2, 3$ and Theorem 5 imply that with probability approaching 1, $\|\hat{\mathbf{w}}_{d|d-k}\| \leq c' \cdot m_d$, where $c' > 0$ is a constant independent of d. Plugging into the displayed equation above, and using assumption 4, implies that the bound in the displayed equation converges to 0. This holds with probability approaching 1 over the choice of $\hat{\mathbf{w}}_d$, hence overall $\mathbb{E}_d \left[\left(\mathbf{x}^\top \hat{\mathbf{w}}_d - \hat{\mathbf{x}}_{|k}^\top \hat{\mathbf{w}}_{d|k} \right)^2 \right] \xrightarrow{P} 0$.

A.3 Proof of Lemma 9

Considering ξ which equals 0 almost surely, we clearly have $\sigma^{-1}(0) = 0$. More generally, fix some $c \in \mathbb{R}$ and $z > 0$, and consider the random variable

$$\xi = \begin{cases} -c & \text{w.p. } \frac{z}{z+1} \\ cz & \text{w.p. } \frac{1}{z+1} , \end{cases}$$

which is easily verified to be zero mean. The assumption $\mathbb{E}[\sigma^{-1}(\xi)] = 0$ translates to

$$\frac{z}{z+1} \cdot \sigma^{-1}(-c) + \frac{1}{z+1} \cdot \sigma^{-1}(cz) = 0 \quad \Longrightarrow \quad \sigma^{-1}(cz) = -\sigma^{-1}(-c) \cdot z \ .$$

Fixing $c = 1$ and studying this equation as a function of $z > 0$, we see that $\sigma^{-1}(\cdot)$ is necessarily linear over $[0, \infty)$ (with slope $-\sigma^{-1}(-1)$). Similarly, fixing $c = -1$, we get that σ^{-1} is necessarily linear over $(-\infty, 0]$ (with slope $\sigma^{-1}(1)$). Thus, it only remains to show that $-\sigma^{-1}(-1) = \sigma^{-1}(1)$, which follows by considering the random variable ξ uniformly distributed on $\{-1, 1\}$, and noting that $\mathbb{E}[\sigma^{-1}(\xi)] = 0$ implies $\sigma^{-1}(-1) + \sigma^{-1}(1) = 0$ in this case.

A.4 Proof of Theorem 12

Consider the optimization problem

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^m} \boldsymbol{\alpha}^\top D \boldsymbol{\alpha} \ : \ D\boldsymbol{\alpha} \succeq \mathbf{r} \quad \Longleftrightarrow \quad \min_{\boldsymbol{\alpha} \in \mathbb{R}^m} \sum_{i=1}^m D_{i,i} \alpha_i^2 \ : \ \forall i \in [m], \ D_{i,i}\alpha_i \geq r_i$$

for some diagonal matrix $D \in \mathbb{R}^{m \times m}$ with positive entries $D_{i,i}$ on the diagonal, and a vector $\mathbf{r} \in \mathbb{R}^m$. Note that for this problem, it is easily verified that the optimum satisfies $\alpha_i = \frac{[r_i]_+}{D_{i,i}}$ for all $i \in [m]$, hence the optimal value is $\|[D^{-1/2}\mathbf{r}]_+\|^2 = \sum_{i=1}^d \frac{[r_i]_+^2}{D_{i,i}}$. The following key technical lemma quantifies by how much the optimal value changes, if we perturb D be some symmetric matrix E:

Lemma 18 *Fix some integer $m \geq 1$, some $\mathbf{r} \in \mathbb{R}^m$, a diagonal matrix $D \in \mathbb{R}^{m \times m}$ with positive diagonal entries, and an $m \times m$ symmetric matrix E such that $\|E\| < \frac{\lambda_{\min}(D)}{2}$. Then the set $\{\boldsymbol{\alpha} \in \mathbb{R}^m : (D+E)\boldsymbol{\alpha} \succeq \mathbf{r}\}$ is not empty. Moreover, letting*

$$a^* = \min_{\boldsymbol{\alpha} \in \mathbb{R}^m} \boldsymbol{\alpha}^\top (D+E)\boldsymbol{\alpha} \ : \ (D+E)\boldsymbol{\alpha} \succeq \mathbf{r} \ , \tag{13}$$

we have

$$\left| a^* - \|[D^{-1/2}\mathbf{r}]_+\|^2 \right| \leq \frac{2\|E\| \cdot \|D\|}{\lambda_{\min}(D)^2} \cdot \|[D^{-1/2}\mathbf{r}]_+\|^2 \ .$$

Proof The fact that $\{\boldsymbol{\alpha} \in \mathbb{R}^m : (D+E)\boldsymbol{\alpha} \succeq \mathbf{r}\}$ is not empty follows from the observation that $D + E$ is positive definite and hence invertible (since we assume $\|E\| \leq \frac{\lambda_{\min}(D)}{2}$). Thus, the set contains for instance the vector $(D+E)^{-1}\mathbf{r}$. Also, note that the minimum a^* is indeed attained, as we are minimizing a strongly convex function with (feasible) linear constraints.

To continue, let us perform the variable change $\boldsymbol{\beta} = (D+E)\boldsymbol{\alpha}$ (which is valid since $D + E$ is invertible), so $\boldsymbol{\alpha} = (D+E)^{-1}\boldsymbol{\beta}$, and

$$a^* = \min_{\boldsymbol{\beta} \in \mathbb{R}^m} \boldsymbol{\beta}^\top (D+E)^{-1}\boldsymbol{\beta} \ : \ \boldsymbol{\beta} \succeq \mathbf{r} \ . \tag{14}$$

By Lemma 17 and the assumption $\|E\| \leq \frac{\lambda_{\min}(D)}{2}$, it follows that

$$\|(D+E)^{-1} - D^{-1}\| \leq \frac{2\|E\|}{\lambda_{\min}(D)^2} \ .$$

This implies that

$$D^{-1} + \frac{2\|E\|}{\lambda_{\min}(D)^2}I \ \succeq \ (D+E)^{-1} \ \succeq \ D^{-1} - \frac{2\|E\|}{\lambda_{\min}(D)^2}I,$$

where $A \succeq B$ for symmetric matrices A, B implies that $A - B$ is positive semidefinite. Plugging this back into Eq. (14), it follows that

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^m:\boldsymbol{\beta}\succeq\mathbf{r}} \boldsymbol{\beta}^\top \left(D^{-1} + \frac{2\|E\|}{\lambda_{\min}(D)^2}I\right)\boldsymbol{\beta} \ \geq \ a^* \ \geq \ \min_{\boldsymbol{\beta}\in\mathbb{R}^m:\boldsymbol{\beta}\succeq\mathbf{r}} \boldsymbol{\beta}^\top \left(D^{-1} - \frac{2\|E\|}{\lambda_{\min}(D)^2}I\right)\boldsymbol{\beta} \ .$$

Now, it is easily verified that for a diagonal matrix $A \in \mathbb{R}^{m\times m}$ with non-negative diagonal entries,

$$\min_{\boldsymbol{\beta}:\boldsymbol{\beta}\succeq\mathbf{r}} \boldsymbol{\beta}^\top A\boldsymbol{\beta} = \min_{\boldsymbol{\beta}:\forall i, \beta_i \geq r_i} \sum_{i=1}^m A_{i,i}\beta_i^2 \ = \ \sum_{i=1}^m A_{i,i}[r_i]_+^2.$$

Plugging this into the previous displayed equation, it follows that

$$\sum_{i=1}^m \left(\frac{1}{D_{i,i}} + \frac{2\|E\|}{\lambda_{\min}(D)^2}\right)[r_i]_+^2 \ \geq \ a^* \ \geq \ \sum_{i=1}^m \left(\frac{1}{D_{i,i}} - \frac{2\|E\|}{\lambda_{\min}(D)^2}\right)[r_i]_+^2 \ .$$

Therefore,

$$\left|a^* - \sum_{i=1}^m \frac{[r_i]_+^2}{D_{i,i}}\right| \ \leq \ \sum_{i=1}^m [r_i]_+^2 \cdot \frac{2\|E\|}{\lambda_{\min}(D)^2} \ = \ \sum_{i=1}^m \frac{[r_i]_+^2}{D_{i,i}}\cdot\frac{2\|E\|D_{i,i}}{\lambda_{\min}(D)^2} \ \leq \ \sum_{i=1}^m \frac{[r_i]_+^2}{D_{i,i}}\cdot\frac{2\|E\|\cdot\|D\|}{\lambda_{\min}(D)^2} \ .$$

Noting that $\sum_{i=1}^m \frac{[r_i]_+^2}{D_{i,i}} = \|[D^{-1/2}\mathbf{r}]_+\|^2$ and plugging in the displayed equation above, the bound in the lemma follows. ∎

With this lemma in hand, we can now turn to prove the theorem. $\hat{\mathbf{w}}$ can be equivalently written as

$$\hat{\mathbf{w}} \ = \ \arg\min_{\mathbf{w}\in\mathbb{R}^d} \frac{\|\mathbf{w}\|^2}{m} \ : \ \forall i \in [m], \ y_i\mathbf{x}_i^\top\mathbf{w} \geq 1 \ .$$

writing $\mathbf{w} = (\mathbf{v}, \mathbf{u})$ where $\mathbf{v} \in \mathbb{R}^k, \mathbf{u} \in \mathbb{R}^{d-k}$, the above is equivalent to

$$\arg\min_{\mathbf{v}\in\mathbb{R}^k, \mathbf{u}\in\mathbb{R}^{d-k}} \frac{\|\mathbf{v}\|^2}{m} + \frac{\|\mathbf{u}\|^2}{m} \ : \ \forall i \in [m], \ y_i\mathbf{x}_{i|d-k}^\top\mathbf{u} \geq 1 - y_i\mathbf{x}_{i|k}^\top\mathbf{v} \ . \tag{15}$$

This in turn is equivalent to

$$\arg\min_{\mathbf{v}\in\mathbb{R}^k} \frac{\|\mathbf{v}\|^2}{m} + f_m(\mathbf{v}) \tag{16}$$

where

$$f_m(\mathbf{v}) \ = \ \min_{\mathbf{u}\in\mathbb{R}^{d-k}} \frac{\|\mathbf{u}^2\|}{m} \ : \ \forall i \in [m], \ y_i\mathbf{x}_{i|d-k}^\top\mathbf{u} \geq 1 - y_i\mathbf{x}_{i|k}^\top\mathbf{v} \ .$$

Let $Z_{|k}, Z_{|d-k}$ be $m \times (d-k)$ matrices whose i-th rows are (respectively) $y_i\mathbf{x}_i$ and $y_i\mathbf{x}_{i|d-k}$. Also, let $\mathbf{1}$ be the all-ones vector in $\mathbb{R}^m$. Thus, we can write

$$f_m(\mathbf{v}) \ = \ \min_{\mathbf{u}\in\mathbb{R}^{d-k}} \frac{\|\mathbf{u}\|^2}{m} \ : \ Z_{|d-k}\mathbf{u} \succeq \mathbf{1} - Z_{|k}\mathbf{v} \ .$$

Clearly, the optimal $\mathbf{u}$ must lie in the row span of $Z_{|d-k}$ (otherwise, we can further reduce $\|\mathbf{u}\|^2$ by projecting to that subspace, without violating the constraints). Thus, any optimal $\mathbf{u}$ can be written as $Z_{|d-k}^\top \boldsymbol{\alpha}$ for some $\boldsymbol{\alpha} \in \mathbb{R}^m$, so we can rewrite the displayed equation above as

$$f_m(\mathbf{v}) \;=\; \min_{\boldsymbol{\alpha}\in\mathbb{R}^m} \frac{1}{m}\boldsymbol{\alpha}^\top \left(Z_{|d-k} Z_{|d-k}^\top \right)\boldsymbol{\alpha} \;\;:\;\; Z_{|d-k} Z_{|d-k}^\top \boldsymbol{\alpha} \succeq \mathbf{1} - Z_{|k}\mathbf{v} \;.$$

Letting

$$Z_{|d-k} Z_{|d-k}^\top \;=\; D + E \;,$$

where E is as defined in the theorem statement, and D being a diagonal matrix consisting of the diagonal of $Z_{|d-k} Z_{|d-k}^\top$ (namely $(\|\mathbf{x}_{1|d-k}\|^2,\ldots,\|\mathbf{x}_{m|d-k}\|^2)$), we can write the above as

$$f_m(\mathbf{v}) \;=\; \min_{\boldsymbol{\alpha}\in\mathbb{R}^m} \frac{1}{m}\boldsymbol{\alpha}^\top (D+E)\boldsymbol{\alpha} \;\;:\;\; (D+E)\boldsymbol{\alpha} \succeq \mathbf{1} - Z_{|k}\mathbf{v} \;.$$

Applying Lemma 18 on the equation above (and noting that $\|E\| < \frac{\lambda_{\min}(D)}{2}$, which follows from the assumption $\frac{2\|E\|\cdot\|D\|}{\lambda_{\min}(D)^2} = \frac{2\|E\|\cdot\max_{i\in[m]}\|\mathbf{x}_{i|d-k}\|^2}{\min i\in[m]\|\mathbf{x}_{i|d-k}\|^4} \le \frac{1}{2} < 1$ and the fact that $1 \le \frac{\|D\|}{\lambda_{\min}(D)}$), we get that

$$\left| f_m(\mathbf{v}) - \frac{1}{m}\|[D^{-1/2}(\mathbf{1}-Z_{|k}\mathbf{v})]_+\|^2 \right| \;\le\; \frac{2\|E\|\cdot\|D\|}{\lambda_{\min}(D)^2}\cdot\frac{1}{m}\|[D^{-1/2}(\mathbf{1}-Z_{|k}\mathbf{v})]_+\|^2 \;. \quad (17)$$

Plugging this back into Eq. (16), and plugging in $\frac{1}{m}\|[D^{-1/2}(\mathbf{1}-Z_{|k}\mathbf{v})]_+\|^2 = \frac{1}{m}\sum_{i=1}^m \frac{[1-y_i\mathbf{x}_{i|k}^\top\mathbf{v}]_+^2}{\|\mathbf{x}_{i|d-k}\|^2}$ (which holds by definition of $D, Z_{|k}$), we get that

$$\hat{\mathbf{w}}_{|k} \;=\; \arg\min_{\mathbf{v}} \frac{\|\mathbf{v}\|^2}{m} + (1+\epsilon_\mathbf{v})\cdot\frac{1}{m}\sum_{i=1}^m \frac{[1-y_i\mathbf{x}_{i|k}^\top\mathbf{v}]_+^2}{\|\mathbf{x}_{i|d-k}\|^2} \;, \quad (18)$$

where $\epsilon_\mathbf{v}\in\mathbb{R}$ satisfies

$$\sup_{\mathbf{v}\in\mathbb{R}^k} |\epsilon_\mathbf{v}| \;\le\; \frac{2\|E\|\cdot\|D\|}{\lambda_{\min}(D)^2} = \epsilon_0 \;,$$

proving the first bound in the theorem.

To get the second bound, note that since $\hat{\mathbf{w}}_{|k}$ minimizes the expression in Eq. (18), which for $\mathbf{v}=0$ equals $\frac{1+\epsilon_\mathbf{v}}{m}\sum_{i=1}^m \frac{1}{\|\mathbf{x}_{i|d-k}\|^2} \le \frac{1+\epsilon_0}{\min_i\|\mathbf{x}_{i|d-k}\|^2} = \frac{1+\epsilon_0}{\lambda_{\min}(D)}$, it must hold that

$$\frac{\|\hat{\mathbf{w}}_{|k}\|^2}{m} + (1+\epsilon_{\hat{\mathbf{w}}_{|k}})\cdot\frac{1}{m}\sum_{i=1}^m \frac{[1-y_i\mathbf{x}_{i|k}^\top\hat{\mathbf{w}}_{|k}]_+^2}{\|\mathbf{x}_{i|d-k}\|^2} \;\le\; \frac{1+\epsilon_0}{\lambda_{\min}(D)} \;,$$

and since $1+\epsilon_{\hat{\mathbf{w}}_{|k}} \ge 1-\epsilon_0 > 0$, it follows that

$$\frac{1}{m}\sum_{i=1}^m \frac{[1-y_i\mathbf{x}_{i|k}^\top\hat{\mathbf{w}}_{|k}]_+^2}{\|\mathbf{x}_{i|d-k}\|^2} \;\le\; \frac{1+\epsilon_0}{(1-\epsilon_0)\lambda_{\min}(D)} \;. \quad (19)$$

Next, recall from Eq. (16) and the fact that $(\hat{\mathbf{w}}_{|k}, \hat{\mathbf{w}}_{|d-k})$ jointly optimize Eq. (15) that

$$f_m(\hat{\mathbf{w}}_{|k}) \;=\; \frac{\|\hat{\mathbf{w}}_{|d-k}\|^2}{m} \; .$$

Combining with Eq. (17) (with $\mathbf{v} = \hat{\mathbf{w}}_{|k}$), it follows that

$$\frac{\|\hat{\mathbf{w}}_{|d-k}\|^2}{m} \leq \left(1 + \frac{2\|E\|\cdot\|D\|}{\lambda_{\min}(D)^2}\right)\cdot\frac{1}{m}\|[D^{-1/2}\left(\mathbf{1} - Z_{|k}\hat{\mathbf{w}}_{|k}\right)]_+\|^2 = (1+\epsilon_0)\cdot\frac{1}{m}\sum_{i=1}^{m}\frac{[1 - y_i\mathbf{x}_{i|k}^\top\hat{\mathbf{w}}_{|k}]_+^2}{\|\mathbf{x}_{i|d-k}\|^2} \; ,$$

where we recall that $\epsilon_0 = \frac{2\|E\|\cdot\|D\|}{\lambda_{\min}(D)^2}$. Combining this with Eq. (19), we get

$$\frac{\|\hat{\mathbf{w}}_{|d-k}\|^2}{m} \;\leq\; \frac{(1+\epsilon_0)^2}{(1-\epsilon_0)\lambda_{\min}(D)} \; ,$$

which is less than $5/\lambda_{\min}(D)$ (since $\epsilon_0 \in [0, \frac{1}{2}]$). Multiplying both sides by m, and plugging in $\lambda_{\min}(D) = \min_{i\in[m]} \|\mathbf{x}_{i|d-k}\|^2$, results in the second bound in the theorem.

A.5 Proof of Theorem 13

Assumptions 2 and 3 imply that with probability approaching 1 as d increases, $\hat{\mathbf{w}}_d$ exists and the parameter ϵ_0 from Theorem 12 is arbitrarily small. Under that event, Theorem 12 applies, and $\|\hat{\mathbf{w}}_{d|d-k}\|^2 \leq c_0 \cdot m_d$ for some $c_0 > 0$ independent of d. Therefore, with probability approaching 1,

$$\mathbb{E}_d\left[\left(\mathbf{x}^\top\hat{\mathbf{w}}_d - \mathbf{x}_{|k}^\top\hat{\mathbf{w}}_{d|k}\right)^2\right] \;=\; \mathbb{E}_d\left[\left(\mathbf{x}_{|d-k}^\top\hat{\mathbf{w}}_{d|d-k}\right)^2\right] \;=\; \hat{\mathbf{w}}_{d|d-k}^\top\mathbb{E}_d[\mathbf{x}_{d-k}\mathbf{x}_{|d-k}^\top]\hat{\mathbf{w}}_{d|d-k}$$

$$\leq \; \|\hat{\mathbf{w}}_{d|d-k}\|^2 \cdot \|\mathbb{E}_d[\mathbf{x}_{d-k}\mathbf{x}_{|d-k}^\top]\| \;\leq\; c_0 \cdot m_d \cdot \|\mathbb{E}_d[\mathbf{x}_{d-k}\mathbf{x}_{|d-k}^\top]\| \; ,$$

which by assumption 4 converges to 0. All this implies that $\mathbb{E}_d\left[\left(\mathbf{x}^\top\hat{\mathbf{w}}_d - \mathbf{x}_{|k}^\top\hat{\mathbf{w}}_{d|k}\right)^2\right] \xrightarrow{P} 0$.

We now turn to analyze $\hat{\mathbf{w}}_{d|k}$, which by Theorem 12 and assumptions 2,3 satisfies

$$\hat{\mathbf{w}}_{d|k} \;=\; \arg\min_{\mathbf{v}\in\mathbb{R}^k} \; (1 + \epsilon_{d,\mathbf{v}}) \cdot \hat{g}_d(\mathbf{v}) + \frac{\|\mathbf{v}\|^2}{m_d} \tag{20}$$

with probability approaching 1, where $\hat{g}_d(\mathbf{v}) := \hat{\mathbb{E}}_d\left[\frac{[1-y\mathbf{x}_{|k}^\top\mathbf{v}]_+^2}{\|\mathbf{x}_{|d-k}\|^2}\right] = \frac{1}{m_d}\sum_{i=1}^{m_d}\frac{[1-y_i\mathbf{x}_{i|k}^\top\mathbf{v}]_+^2}{\|\mathbf{x}_{i|d-k}\|^2}$ and $\sup_{\mathbf{v}}|\epsilon_{d,\mathbf{v}}| \leq \epsilon_d$ for some random variable ϵ_d satisfying $m_d\epsilon_d \xrightarrow{P} 0$ as $d \to \infty$.

First, recalling that $g_d(\mathbf{v}) = \mathbb{E}_d\left[\frac{[1-y\mathbf{x}_{|k}^\top\mathbf{v}]_+^2}{\|\mathbf{x}_{|d-k}\|^2}\right]$, we have by assumption 1 and the law of large numbers that $\hat{g}_d(\mathbf{v}) - g_d(\mathbf{v}) \xrightarrow{P} 0$ for any fixed $\mathbf{v}$. Moreover, by assumption 2, both $g_d(\cdot)$ and $\hat{g}_d(\cdot)$ are $c_{\mathcal{V}}$-Lipschitz on any fixed compact set $\mathcal{V}$ in $\mathbb{R}^k$ (where $c_{\mathcal{V}}$ depends on $\mathcal{V}$ but not on d, and with probability approaching 1 for $\hat{g}_d$). Therefore, by a standard covering number argument, it follows that

$$\sup_{\mathbf{v}\in\mathcal{V}} |\hat{g}_d(\mathbf{v}) - g_d(\mathbf{v})| \xrightarrow{P} 0$$

for any compact set $\mathcal{V} \subset \mathbb{R}^k$.

We now wish to argue that with probability approaching 1, $\hat{\mathbf{w}}_{d|k}$ lies in some compact set $\{\mathbf{v} \in \mathbb{R}^k : \|\mathbf{v}\| \leq c''\}$ where c'' is a constant independent of d. Applying the displayed equation above, this would imply that

$$\hat{g}_d(\hat{\mathbf{w}}_{d|k}) - g_d(\hat{\mathbf{w}}_{d|k}) \xrightarrow{P} 0 \ . \tag{21}$$

To justify this, let us compare $\hat{\mathbf{w}}_{d|k}$ to $\hat{\mathbf{v}}_d$, which we define as the minimum-norm minimizer of[3] $\hat{g}_d(\cdot)$. By Eq. (20), we have with probability approaching 1 for any large enough d that

$$(1 - \epsilon_d)\hat{g}_d(\hat{\mathbf{v}}_d) + \frac{\|\hat{\mathbf{w}}_{d|k}\|^2}{m_d} \ \leq \ (1 + \epsilon_{d,\hat{\mathbf{w}}_{d|k}})\hat{g}_d(\hat{\mathbf{w}}_{d|k}) + \frac{\|\hat{\mathbf{w}}_{d|k}\|^2}{m_d}$$

$$\leq \ (1 + \epsilon_{d,\hat{\mathbf{v}}_d})\hat{g}_d(\hat{\mathbf{v}}_d) + \frac{\|\hat{\mathbf{v}}_d\|^2}{m_d} \ \leq \ (1 + \epsilon_d)\hat{g}_d(\hat{\mathbf{v}}_d) + \frac{\|\hat{\mathbf{v}}_d\|^2}{m_d} \ .$$

Multiplying both sides by m_d and switching sides, it follows that

$$\|\hat{\mathbf{w}}_{d|k}\|^2 \ \leq \ 2m_d\epsilon_d \cdot g_d(\hat{\mathbf{v}}_d) + \|\hat{\mathbf{v}}_d\|^2 \ \leq \ 2c^2 m_d\epsilon_d + \|\hat{\mathbf{v}}_d\|^2 \ ,$$

where the last transition follows from $\hat{\mathbf{v}}_d$ being a minimizer of $\hat{g}_d(\cdot)$, hence $\hat{g}_d(\hat{\mathbf{v}}_d) \leq \hat{g}_d(\mathbf{0}) = \hat{\mathbb{E}}_d[\frac{1}{\|\mathbf{x}_{|d-k}\|^2}] \leq c^2$ by assumption 2. Recalling that $m_d\epsilon_d \xrightarrow{P} 0$, it follows in particular $\|\hat{\mathbf{w}}_{d|k}\|^2 \leq 1 + \|\hat{\mathbf{v}}_d\|^2$ with probability approaching 1. By assumption 5 in the theorem, $\|\hat{\mathbf{v}}_d\| \leq c'$ for some constant c' with probability approaching 1. Overall, we get that with probability approaching 1, $\hat{\mathbf{w}}_{d|k}$ lies in some compact set $\{\mathbf{v} \in \mathbb{R}^k : \|\mathbf{v}\| \leq c''\}$ with c'' independent of d, hence justifying Eq. (21) as discussed earlier.

Next, let us fix some reference vector $\mathbf{v}_0 \in \mathbb{R}^k$. By assumption 3 and the law of large numbers, $\hat{g}_d(\mathbf{v}_0) - g_d(\mathbf{v}_0) \xrightarrow{P} 0$. Combined with Eq. (21), it follows that

$$\delta_d := \big(g_d(\hat{\mathbf{w}}_{d|k}) - g_d(\mathbf{v}_0)\big) - \big(\hat{g}_d(\hat{\mathbf{w}}_{d|k}) - \hat{g}_d(\mathbf{v}_0)\big) \ \xrightarrow{P} \ 0$$

Thus, we have the following:

$$
\begin{aligned}
g_d(\hat{\mathbf{w}}_{d|k}) - g_d(\mathbf{v}_0) &= \hat{g}_d(\hat{\mathbf{w}}_{d|k}) - \hat{g}_d(\mathbf{v}_0) + \delta_d \\
&= (1 + \epsilon_{d,\hat{\mathbf{w}}_{d|k}}) \cdot \hat{g}_d(\hat{\mathbf{w}}_{d|k}) - \hat{g}_d(\mathbf{v}_0) + \delta_d - \epsilon_{d,\hat{\mathbf{w}}_{d|k}} \cdot \hat{g}_d(\hat{\mathbf{w}}_{d|k}) \\
&\leq (1 + \epsilon_{d,\hat{\mathbf{w}}_{d|k}}) \cdot \hat{g}_d(\hat{\mathbf{w}}_{d|k}) + \frac{\|\hat{\mathbf{w}}_{d|k}\|^2}{m_d} - \hat{g}_d(\mathbf{v}_0) + \delta_d - \epsilon_{d,\hat{\mathbf{w}}_{d|k}} \cdot \hat{g}_d(\hat{\mathbf{w}}_{d|k}) \\
&\overset{Eq.\ (20)}{\leq} (1 + \epsilon_{d,\mathbf{v}_0}) \cdot \hat{g}_d(\mathbf{v}_0) + \frac{\|\mathbf{v}_0\|^2}{m_d} - \hat{g}_d(\mathbf{v}_0) + \delta_d - \epsilon_{d,\hat{\mathbf{w}}_{d|k}} \cdot \hat{g}_d(\hat{\mathbf{w}}_{d|k}) \\
&= \epsilon_{d,\mathbf{v}_0} \cdot \hat{g}_d(\mathbf{v}_0) + \frac{\|\mathbf{v}_0\|^2}{m_d} + \delta_d - \epsilon_{d,\hat{\mathbf{w}}_{d|k}} \cdot \hat{g}_d(\hat{\mathbf{w}}_{d|k}) \ .
\end{aligned}
$$

We now argue that all the terms in the bound above converge (deterministically or in probability) to 0: As to the first term, we know that $\sup_{\mathbf{v}} |\epsilon_{d,\mathbf{v}}| \xrightarrow{P} 0$, and $\hat{g}_d(\mathbf{v}_0)$ is at most

---

3. A minimizer of $\hat{g}_d(\cdot)$ always exists, since it is convex piecewise-quadratic with finitely many pieces. The minimum-norm minimizer is unique, since if there were two minimizers of equal minimal norm, their average would also be a minimizer by convexity of $\hat{g}_d(\cdot)$, and with a smaller norm which is a contradiction.

some fixed value independent of d with probability approaching 1, by assumption 2 and definition of $\hat{g}_d(\cdot)$. As to the second term, it converges to 0 since $m_d \to \infty$. The third term converges in probability to 0 as discussed above. As to the last term, we know that $\sup_{\mathbf{v}} |\epsilon_{d,\mathbf{v}}| \xrightarrow{P} 0$, and $\hat{g}_d(\hat{\mathbf{w}}_{d|k})$ is bounded by some constant independent of d with probability approaching 1 (since $\hat{\mathbf{w}}_{d_k}$ lies in some fixed compact set with probability approaching 1 as discussed earlier, and the values of $\hat{g}_d(\cdot)$ are bounded independent of d on any fixed compact set with probability approaching 1, by assumption 2).

The displayed equation above and the following discussion implies that for any $\beta > 0$, $\Pr(g_d(\hat{\mathbf{w}}_{d|k}) - g_d(\mathbf{v}_0) \geq \beta) \to 0$. This holds for any fixed $\mathbf{v}_0$. Combined with the assumption $\inf_{\mathbf{v}} \limsup_d (g_d(\mathbf{v}) - \inf_{\mathbf{u}} g_d(\mathbf{u})) \leq 0$ (which implies that $\limsup_d (g_d(\mathbf{v}_0) - \inf_{\mathbf{v}} g_d(\mathbf{v}))$ can be made arbitrarily small by choosing $\mathbf{v}_0$ appropriately), it follows that $\Pr(g_d(\hat{\mathbf{w}}_{d|k}) - \inf_{\mathbf{v}} g_d(\mathbf{v}) \geq \beta) \to 0$ for all $\beta > 0$. But since $g_d(\hat{\mathbf{w}}_{d|k}) - \inf_{\mathbf{v}} g_d(\mathbf{v})$ is necessarily non-negative, we get that $g_d(\hat{\mathbf{w}}_{d|k}) - \inf_{\mathbf{v}} g_d(\mathbf{v}) \xrightarrow{P} 0$.

A.6 Proof of Theorem 14

Since the labels are flipped with some probability $p > 0$, we have $\inf_d \inf_{\mathbf{w} \in \mathbb{R}^d} R_d(\mathbf{w}) > 0$. Thus, it remains to prove that under the condition stated in the theorem, $\Pr_{(\mathbf{x},y) \sim \mathcal{D}_d}(y\mathbf{x}^\top \hat{\mathbf{w}}_d)$ converges in probability to p.

As discussed before the theorem, $L_p(\cdot)$ has a unique minimizer $\mathbf{w}_p^*$. Therefore, by Theorem 13, $\hat{\mathbf{w}}_{d|k} \xrightarrow{P} \mathbf{w}_p^*$. Since we assume $\Pr_{(\mathbf{x}_{|k},y) \sim \mathcal{D}_{\text{clean}}}(y\mathbf{x}_{|k}^\top \mathbf{w}_p^* \leq 0) = 0$, we argue that

$$\Pr_{(\mathbf{x}_{|k},y) \sim \mathcal{D}_{\text{clean}}}(y\mathbf{x}_{|k}^\top \hat{\mathbf{w}}_{d|k} \leq 0) \xrightarrow{P} 0 \ . \tag{22}$$

We note that formally proving this requires some care, as $\Pr(y\mathbf{x}_{|k}^\top \mathbf{w} \leq 0)$ is not necessarily continuous in $\mathbf{w}$ (otherwise Eq. (22) would follow immediately by continuity). To show Eq. (22) formally, define for all $\gamma > 0$ the set $\mathcal{U}_\gamma := \{\mathbf{u} \in \mathbb{R}^k : \mathbf{u}^\top \mathbf{w}_p^* > \gamma, \|\mathbf{u}\| \leq \frac{1}{\gamma}\}$. Clearly, since $\Pr_{(\mathbf{x}_{|k},y) \sim \mathcal{D}_{\text{clean}}}(y\mathbf{x}_{|k}^\top \mathbf{w}_p^* \leq 0) = 0$, we have

$$\Pr_{(\mathbf{x}_{|k},y) \sim \mathcal{D}_{\text{clean}}}(y\mathbf{x}_{|k} \in \mathcal{U}_\gamma) \xrightarrow{\gamma \to 0} 1 \ . \tag{23}$$

Moreover, since $\hat{\mathbf{w}}_{d|k} \xrightarrow{P} \mathbf{w}_p^*$, it holds that $y\mathbf{x}_{|k}^\top \hat{\mathbf{w}}_{d|k} \xrightarrow{P} y\mathbf{x}_{|k}^\top \hat{\mathbf{w}}_p^*$ simultaneously for all vectors $y\mathbf{x}_{|k}$ of some bounded norm. Therefore, for any fixed γ,

$$\Pr_{(\mathbf{x}_{|k},y) \sim \mathcal{D}_{\text{clean}}}(y\mathbf{x}_{|k}^\top \hat{\mathbf{w}}_{d|k} \leq 0 \mid y\mathbf{x}_{|k} \in \mathcal{U}_\gamma)$$

$$= \Pr_{(\mathbf{x}_{|k},y) \sim \mathcal{D}_{\text{clean}}} \left(y\mathbf{x}_{|k}^\top \hat{\mathbf{w}}_{d|k} \leq 0 \mid y\mathbf{x}_{|k}^\top \mathbf{w}_p^* > \gamma, \|y\mathbf{x}_{|k}\| \leq \frac{1}{\gamma} \right) \xrightarrow{P} 0 \ . \tag{24}$$

Recalling that for any two events E, A over some probability space,

$$\Pr(A) = \Pr(A|E) \cdot \Pr(E) + \Pr(A|\neg E) \cdot \Pr(\neg E) \leq \Pr(A|E) + \Pr(\neg E) \ ,$$

it follows that for any fixed γ,

$$\Pr_{(\mathbf{x}_{|k},y) \sim \mathcal{D}_{\text{clean}}} \left(y\mathbf{x}_{|k}^\top \hat{\mathbf{w}}_{d|k} \leq 0 \right) \leq \Pr_{(\mathbf{x}_{|k},y) \sim \mathcal{D}_{\text{clean}}}(y\mathbf{x}_{|k}^\top \hat{\mathbf{w}}_{d|k} \leq 0 \mid y\mathbf{x}_{|k} \in \mathcal{U}_\gamma) + \Pr(y\mathbf{x}_{|k} \notin \mathcal{U}_\gamma).$$

Combined with Eq. (23) and Eq. (24), it follows that by picking γ sufficiently small, we can make
$\Pr_{(\mathbf{x}_{|k},y)\sim\mathcal{D}_{\text{clean}}}\left(y\mathbf{x}_{|k}^{\top}\hat{\mathbf{w}}_{d|k}\leq 0\right)$ asymptotically smaller than any positive number with arbitrarily high probability, from which Eq. (22) follows.

From Eq. (22) and the definition of $\mathcal{D}_{\text{clean}}$, it follows that

$$\Pr_{(\mathbf{x},y)\sim\mathcal{D}_d}(y\mathbf{x}_{|k}^{\top}\hat{\mathbf{w}}_{d|k}\leq 0)\xrightarrow{P}p\ .$$

By Theorem 13, we also have that

$$\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_d}\left[(\mathbf{x}^{\top}\hat{\mathbf{w}}_d-\mathbf{x}_{|k}^{\top}\hat{\mathbf{w}}_{d|k})^2\right]\xrightarrow{P}0\ . \tag{25}$$

Combining the last two observations, we can use similar arguments as above to prove that

$$\Pr_{(\mathbf{x},y)\sim\mathcal{D}_d}(y\mathbf{x}^{\top}\hat{\mathbf{w}}_d\leq 0)\xrightarrow{P}p\ , \tag{26}$$

which as discussed earlier implies the theorem statement. Formally, let $\tilde{\mathcal{D}}_d$ refer to $\mathcal{D}_d$, where y is distributed according to the "clean" distribution $\mathcal{D}_{\text{clean}}$. Also, let $\mathcal{U}_{\gamma}^d=\{\mathbf{u}\in\mathbb{R}^d:\mathbf{u}_{|k}^{\top}\mathbf{w}_p^*>\gamma,\|\mathbf{u}\|\leq\frac{1}{\gamma}\}$. Similar to before, we have

$$\Pr_{(\mathbf{x},y)\sim\tilde{\mathcal{D}}_d}(y\mathbf{x}\in\mathcal{U}_{\gamma}^d)\xrightarrow{\gamma\to 0}1\ .$$

By applying Markov's inequality on Eq. (25), it follows that for any $\gamma>0$, the measure of points $y\mathbf{x}\in\mathcal{U}_{\gamma}^d$ such that $|y\mathbf{x}^{\top}\hat{\mathbf{w}}_d-y\mathbf{x}_{|k}^{\top}\hat{\mathbf{w}}_{d|k}|>\frac{\gamma}{2}$ goes to 0 in probability. For all other points in $\mathcal{U}_{\gamma}^d$, we have $y\mathbf{x}_{|k}^{\top}\hat{\mathbf{w}}_{d|k}>\gamma\ \Rightarrow\ y\mathbf{x}^{\top}\hat{\mathbf{w}}_d>\frac{\gamma}{2}$. Recalling that $y\mathbf{x}_{|k}^{\top}\hat{\mathbf{w}}_{d|k}\xrightarrow{P}y\mathbf{x}_{|k}^{\top}\mathbf{w}_p^*$ (which is $>\gamma$) uniformly for all $y\mathbf{x}\in\mathcal{U}_{\gamma}^d$, we get that

$$\Pr_{(\mathbf{x},y)\sim\tilde{\mathcal{D}}_d}(y\mathbf{x}^{\top}\hat{\mathbf{w}}_d\leq 0\mid y\mathbf{x}\in\mathcal{U}_{\gamma}^d)\xrightarrow{P}0\ .$$

Combining the two displayed equation above, and using the same arguments made in the context of Eq. (23) and Eq. (24), it follows that by choosing γ small enough, the probability $\Pr_{(\mathbf{x},y)\sim\tilde{\mathcal{D}}_d}(y\mathbf{x}^{\top}\hat{\mathbf{w}}_d\leq 0)$ can be made asymptotically smaller than any positive numbers with arbitrarily high probability, from which it follows that

$$\Pr_{(\mathbf{x},y)\sim\tilde{\mathcal{D}}_d}(y\mathbf{x}^{\top}\hat{\mathbf{w}}_d\leq 0)\xrightarrow{P}0\ .$$

Switching from $\tilde{\mathcal{D}}_d$ to $\mathcal{D}_d$ (which involves flipping y randomly with probability p), Eq. (26) follows.

A.7 Proof of Theorem 15

Recall that a function $f:\mathbb{R}^k\to\mathbb{R}$ is λ-strongly convex, if for any $\mathbf{u},\mathbf{v}\in\mathbb{R}^k$ and $\alpha\in[0,1]$,

$$f(\alpha\mathbf{u}+(1-\alpha)\mathbf{v})\ \leq\ \alpha\cdot f(\mathbf{u})+(1-\alpha)f(\mathbf{v})-\alpha(1-\alpha)\cdot\frac{\lambda}{2}\|\mathbf{u}-\mathbf{v}\|^2\ . \tag{27}$$

SHAMIR

Note that any λ-strongly convex function is also λ' strongly convex for any $\lambda' \in [0, \lambda]$. Also, it is well-known that any convex function is 0-strongly convex, that if f is λ-strongly convex, then $c \cdot f$ is $c \cdot \lambda$-strongly convex, and that a sum of a λ-strongly convex function and a λ'-strongly convex function is $(\lambda + \lambda')$-strongly convex. Moreover, if f is λ-strongly convex, it always has a finite unique minimizer $\mathbf{w}^*$, and $f(\mathbf{w}) - f(\mathbf{w}^*) \geq \frac{\lambda}{2}\|\mathbf{w} - \mathbf{w}^*\|^2$ for any $\mathbf{w}$.

We start with the following auxiliary lemma, which implies that $L_p(\cdot)$ is strongly convex under Assumption 3 (since positive definiteness of $\mathbb{E}[\mathbf{x}_{|k}\mathbf{x}_{|k}^\top]$ implies positive definiteness of $\mathbb{E}[z \cdot \mathbf{x}_{|k}\mathbf{x}_{|k}^\top]$ for z bounded in a positive interval):

Lemma 19 *If $\mathbb{E}[z \cdot \mathbf{x}_{|k}\mathbf{x}_{|k}^\top]$ is positive definite (with minimal eigenvalue $\lambda_{\min} > 0$), then for any $p \in (0, \frac{1}{2}]$, $L_p(\cdot)$ (as defined in Eq. (7) is $2p\lambda_{\min}$-strongly convex.*

Proof We have $L_p(\mathbf{w}) = \mathbb{E}[z \cdot \ell_p(y\mathbf{x}_{|k}^\top\mathbf{w})]$, where $\ell_p(\beta) = (1 - p)[1 - \beta]_+^2 + p[1 + \beta]_+^2$. We first argue that ℓ_p is $2p$-strongly convex: Indeed, it can be easily verified that $\ell_p(\beta) = p\beta^2 + (1 - 2p)[1 - \beta]_+^2 + h(\beta)$, where $h(\beta)$ is a convex function that equals $-2\beta + 1$ on $(-\infty, -1]$, $\beta^2 + 2$ on $[-1, +1]$, and $2\beta + 1$ on $[1, \infty)$. Therefore, ℓ_p is the sum of the $2p$-strongly convex function $p\beta^2$ and convex functions, hence is $2p$-strongly convex itself.

Continuing, note that by the theorem's assumptions,

$$\mathbb{E}[z \cdot (\mathbf{x}_{|k}^\top\mathbf{u} - \mathbf{x}_{|k}^\top\mathbf{v})^2] = (\mathbf{u} - \mathbf{v})^\top\mathbb{E}[z \cdot \mathbf{x}_{|k}\mathbf{x}_{|k}^\top] \cdot (\mathbf{u} - \mathbf{v}) \geq \lambda_{\min}\|\mathbf{u} - \mathbf{v}\|^2 .$$

Combining this with the strong convexity of ℓ_p, we get that for any $\mathbf{u}\mathbf{v} \in \mathbb{R}^k$ and $\alpha \in [0, 1]$,

$$
\begin{aligned}
L_p(\alpha\mathbf{u} + (1 - \alpha)\mathbf{v}) &= \mathbb{E}\left[z \cdot \ell_p\left(\alpha\mathbf{x}_{|k}^\top\mathbf{u} + (1 - \alpha)\mathbf{x}_{|k}^\top\mathbf{v}\right)\right] \\
&\leq \mathbb{E}\left[z\alpha\ell_p(\mathbf{x}_{|k}^\top\mathbf{u}) + z(1 - \alpha)\ell_p(\mathbf{x}_{|k}^\top\mathbf{v}) - p\alpha(1 - \alpha)z(\mathbf{x}_{|k}^\top\mathbf{u} - \mathbf{x}_{|k}^\top\mathbf{v})^2\right] \\
&\leq \alpha L_p(\mathbf{u}) + (1 - \alpha)L_p(\mathbf{v}) - p\alpha(1 - \alpha)\lambda_{\min}\|\mathbf{u} - \mathbf{v}\|^2
\end{aligned}
$$

Which by Eq. (27), implies that L_p is $2p\lambda_{\min}$-strongly convex. ∎

We now continue with the proof of the theorem. Recall that

$$L_p(\mathbf{w}) = \mathbb{E}\left[(1 - p) \cdot z \cdot [1 - y\mathbf{x}_{|k}^\top\mathbf{w}]_+^2 + p \cdot z \cdot [1 + y\mathbf{x}_{|k}^\top\mathbf{w}]_+^2\right] ,$$

and in particular, $L_0(\mathbf{w}) = \mathbb{E}[z \cdot [1 - y\mathbf{x}_{|k}^\top\mathbf{w}]_+^2]$, which by the linear separability assumption achieves a minimal value of 0 at some $\mathbf{w}^*$. For any $p \in [0, \frac{1}{2})$, let

$$\tilde{L}_p(\mathbf{w}) = \mathbb{E}[z \cdot [1 - y\mathbf{x}_{|k}^\top\mathbf{w}]_+^2] + \frac{p}{1 - 2p} \cdot g(\mathbf{w}) \quad \text{where} \quad g(\mathbf{w}) := \mathbb{E}[z \cdot [1 - y\mathbf{x}_{|k}^\top\mathbf{w}]_+^2 + z \cdot [1 + y\mathbf{x}_{|k}^\top\mathbf{w}]_+^2] .$$

It is easy to check that $(1 - 2p) \cdot \tilde{L}_p(\mathbf{w}) = L_p(\mathbf{w})$ for all $\mathbf{w}$, hence a minimizer $\mathbf{w}_p^*$ of $L_p(\cdot)$ is also a minimizer of $\tilde{L}_p(\cdot)$, and $\mathbf{w}^*$ is a minimizer of $\tilde{L}_0(\cdot)$. Moreover, by Lemma 19, $\tilde{L}_p(\mathbf{w})$ is $\frac{2p\lambda_{\min}}{1 - 2p}$-strongly convex (which would also imply that its minimizer $\mathbf{w}_p^*$ always exists and is unique).

34

Next, we argue that $\mathbf{w}_p^*$ is continuous as a function of p in $(0, \frac{1}{2})$: Otherwise, there is some $p_0 \in (0, \frac{1}{2})$ and a sequence of values $p_1, p_2, \ldots$ converging to p_0, such that $\mathbf{w}_{p_j}^*$ remains bounded away from $\mathbf{w}_{p_0}^*$, say by some minimal distance $\delta > 0$. Let us see why that is not possible: By $\frac{2p\lambda_{\min}}{1-2p}$-strong convexity of $\tilde{L}_p(\cdot)$ and the fact that $\mathbf{w}_p^*$ is a minimizer, it would imply

$$\tilde{L}_p(\mathbf{w}_{p_0}^*) - \tilde{L}_p(\mathbf{w}_p^*) \geq \frac{p\lambda_{\min}}{(1-2p)}\|\mathbf{w}_{p_0}^* - \mathbf{w}_p^*\|^2 \geq \frac{p\lambda_{\min}\delta^2}{1-2p}$$

if $p = p_j$ for some j. Similarly, by $\frac{2p_0\lambda_{\min}}{1-2p_0}$-strong convexity of $\tilde{L}_{p_0}(\cdot)$, and the fact that $\mathbf{w}_{p_0}^*$ is a minimizer, we would have

$$\tilde{L}_{p_0}(\mathbf{w}_p^*) - \tilde{L}_{p_0}(\mathbf{w}_{p_0}^*) \geq \frac{p_0\lambda_{\min}}{1-2p_0}\|\mathbf{w}_p^* - \mathbf{w}_{p_0}^*\|^2 \geq \frac{p_0\lambda_{\min}\delta^2}{1-2p_0}$$

for any $p = p_j$. Summing the last two displayed equations for any $p = p_j$, it follows that $(\tilde{L}_{p_j}(\mathbf{w}_{p_0}^*) - \tilde{L}_{p_0}(\mathbf{w}_{p_0}^*)) + (\tilde{L}_{p_0}(\mathbf{w}_{p_j}^*) - \tilde{L}_{p_j}(\mathbf{w}_{p_j}^*))$ is bounded away from 0 as $j \to \infty$, but this contradicts the fact that $\tilde{L}_{p_j}(\mathbf{w}_{p_0}^*) - \tilde{L}_{p_0}(\mathbf{w}_{p_0}^*) \overset{j\to\infty}{\longrightarrow} 0$ and $\tilde{L}_{p_0}(\mathbf{w}_{p_j}^*) - \tilde{L}_{p_j}(\mathbf{w}_{p_j}^*) \overset{j\to\infty}{\longrightarrow} 0$.

Now, since $\mathbf{w}_p^*$ is a continuous function of p in $(0, \frac{1}{2})$, it must have a limit point $\hat{\mathbf{w}}$ as $p \to 0$. Since $\lim_{p\to 0} \tilde{L}_p(\mathbf{w}^*) = \tilde{L}_0(\mathbf{w}^*)$ and $\tilde{L}_0(\mathbf{w}^*) \leq \tilde{L}_0(\mathbf{w}_p^*) \leq \tilde{L}_p(\mathbf{w}_p^*) \leq \tilde{L}_p(\mathbf{w}^*)$, we must have $\lim_{p\to 0} \tilde{L}_p(\mathbf{w}_p^*) = \tilde{L}_0(\mathbf{w}^*)$. But since $\mathbf{w}_p^* \overset{p\to 0}{\longrightarrow} \hat{\mathbf{w}}$ and $\tilde{L}_p$ is Lipschitz in any fixed neighborhood of $\hat{\mathbf{w}}$ (with a uniform upper bound on the Lipschitz constant), it follows that $\tilde{L}_p(\hat{\mathbf{w}}) \overset{p\to 0}{\longrightarrow} \tilde{L}_0(\mathbf{w}^*)$. Recalling that $\tilde{L}_0(\mathbf{w}^*) = L_0(\mathbf{w}^*) = 0$, it follows that

$$\lim_{p\to 0} \tilde{L}_p(\hat{\mathbf{w}}) = 0 .$$

Combined with the fact that $\tilde{L}_p(\hat{\mathbf{w}}) \geq \mathbb{E}[z \cdot [1 - y\mathbf{x}_{|k}^\top \hat{\mathbf{w}}]_+^2] \geq \mathbb{E}[\frac{z}{4}\mathbf{1}_{y\mathbf{x}_{|k}^\top \hat{\mathbf{w}} < \frac{1}{2}}] \geq \frac{l}{4}\Pr(y\mathbf{x}_{|k}^\top \hat{\mathbf{w}} < \frac{1}{2})$ regardless of p, it follows that

$$\Pr\left(y\mathbf{x}_{|k}^\top \hat{\mathbf{w}} < \frac{1}{2}\right) = 0 .$$

Now, let B be such that $\Pr(\|y\mathbf{x}_{|k}\| \leq B) = 1$ (such a B exists by assumption). Note that since $\mathbf{w}_p^* \overset{p\to 0}{\longrightarrow} \hat{\mathbf{w}}$, then for any $p > 0$ sufficiently small, we must have $\|\mathbf{w}_p^* - \hat{\mathbf{w}}\| \leq \frac{1}{4B}$. For any $y, \mathbf{x}_{|k}$ such that $\|y\mathbf{x}_{|k}\| \leq B$, the event $y\mathbf{x}_{|k}^\top \mathbf{w}_p^* < \frac{1}{4}$ implies

$$y\mathbf{x}_{|k}^\top \hat{\mathbf{w}} = y\mathbf{x}_{|k}^\top \mathbf{w}_p^* + y\mathbf{x}_{|k}^\top(\hat{\mathbf{w}} - \mathbf{w}_p^*) < \frac{1}{4} + \|y\mathbf{x}_{|k}\|\|\hat{\mathbf{w}} - \mathbf{w}_p^*\| \leq \frac{1}{4} + \frac{B}{4B} = \frac{1}{2}.$$

But since we showed that $y\mathbf{x}_{|k}^\top \hat{\mathbf{w}} < \frac{1}{2}$ occurs with probability 0, it follows that the event $y\mathbf{x}_{|k}^\top \mathbf{w}_p^* < \frac{1}{4}$ also occurs with probability 0, namely

$$\Pr\left(y\mathbf{x}_{|k}^\top \mathbf{w}_p^* < \frac{1}{4}\right) = 0 .$$

In particular, we get that for all sufficiently small p, the misclassification error probability of $\mathbf{w}_p^*$ is 0.

A.8 Proof of Theorem 16

We first argue that for all $p \in (0, \frac{1}{2}]$, $\mathbf{w}_p^*$ must be in span$(\mathbf{u})$. Otherwise, suppose that $\mathbf{w}_p^* = \alpha \mathbf{u} + \mathbf{r}$ for some $\alpha \in \mathbb{R}$ and non-zero vector $\mathbf{r}$ orthogonal to $\mathbf{u}$. Then we argue that $\alpha \mathbf{u} - \mathbf{r}$ (which is distinct from $\mathbf{w}_p^*$) must also be a minimizer of L_p, because conditioned on any y, z, the distribution of

$$\mathbf{x}_{|k}^\top \mathbf{w}_p^* = \mathbf{x}_{|k}^\top (\alpha \mathbf{u} + \mathbf{r}) = \alpha \mathbf{x}_{|k}^\top \mathbf{u} + \mathbf{r}^\top (I - \mathbf{u}\mathbf{u}^\top) \mathbf{x}_{|k}$$

is the same as

$$\mathbf{x}_{|k}^\top (\alpha \mathbf{u} - \mathbf{r}) = \alpha \mathbf{x}_{|k}^\top \mathbf{u} - \mathbf{r}^\top (I - \mathbf{u}\mathbf{u}^\top) \mathbf{x}_{|k} \ ,$$

and the the value of $L_p(\cdot)$ conditioned on any y, z depends just on these quantities. But since $L_p(\cdot)$ is strongly convex, its minimizer must be unique, which is a contradiction.

Next, let $\mathbf{w}^* \in \mathbb{R}^k$ such that $L_0(\mathbf{w}^*) = \mathbb{E}[z \cdot [1 - y\mathbf{x}_{|k}^\top \mathbf{w}^*]_+^2] = 0$ (such a vector exists by the linear separability assumption). We argue that we can assume $\mathbf{w}^* \in$ span$(\mathbf{u})$ without loss of generality: If not, and it equals $\alpha\mathbf{u} + \mathbf{r}$ with $\mathbf{r} \neq 0$ orthogonal to $\mathbf{u}$, then by the same arguments as above, $\alpha\mathbf{u} - \mathbf{r}$ also minimizes $L_0(\cdot)$. But $L_0(\cdot)$ is convex, so the average of the two points (which is $\alpha\mathbf{u}$) is a minimizer of $L_0(\cdot)$, and we can take $\mathbf{w}^*$ to be that minimizer.

Finally, we argue that if we write $\mathbf{w}_p^*$ as $\alpha_p \mathbf{u}$, and $\mathbf{w}^*$ as $\alpha \mathbf{u}$, then the sign of α_p and α must be the same. Indeed, suppose without loss of generality that $\alpha > 0$ (otherwise, flip $\mathbf{u}$ to $-\mathbf{u}$, and note that α cannot be zero, since then $\mathbf{w}^* = 0$ and it cannot possibly satisfy the theorem assumptions). Since $L_0(\mathbf{w}^*) = \mathbb{E}[z \cdot [1 - y\mathbf{x}_{|k}^\top \alpha\mathbf{u}]_+^2] = 0$ and $z \geq l > 0$ with probability 1, it follows that $zy\mathbf{x}_{|k}^\top \mathbf{u} > 0$ with probability 1. Therefore,

$$\frac{d}{d\beta} L_p(\beta \mathbf{u}) \mid_{\beta=0} \ = \ -2(1 - 2p)\mathbb{E}[zy\mathbf{x}_{|k}^\top \mathbf{u}] < 0$$

for any $p \in (0, \frac{1}{2})$, which by convexity of $\beta \mapsto L_p(\beta\mathbf{u})$ implies that the (unique) minimizer $\mathbf{w}_p^* = \alpha_p \mathbf{u}$ of $L_p(\cdot)$ must satisfy $\alpha_p > 0$. Overall, we have

$$\Pr(y\mathbf{x}_{|k}^\top \mathbf{w}_p^* \leq 0) \ = \ \Pr(\alpha_p y\mathbf{x}_{|k}^\top \mathbf{u} \leq 0) \ = \ \Pr(\alpha y\mathbf{x}_{|k}^\top \mathbf{u} \leq 0) \ = \ \Pr(y\mathbf{x}_{|k}^\top \mathbf{w}^* \leq 0) \ = \ 0 \ .$$

as required.

Appendix B. Minimizers of the Squared Hinge Loss Can Lead to Large Misclassification Error

Fix some distribution $\mathcal{D}$ over examples $(\mathbf{x}, y) \in \mathbb{R}^d \times \{-1, +1\}$. If the distribution is linearly separable on the first k coordinates, it is easy to see that a minimizer of the expectation of the weighted squared hinge loss, $\mathbb{E}\left[\frac{[1 - y\mathbf{x}_{|k}^\top \mathbf{w}]_+^2}{\|\mathbf{x}_{|d-k}\|} \right]$ will also minimize the expected misclassification error (probability that $y\mathbf{x}_{|k}^\top \mathbf{w} \leq 0$), since it will return a point $\mathbf{w}$ such that $1 - y\mathbf{x}_{|k}^\top \mathbf{w} \leq 0$ with probability 1 (and such a point exists by the linear separability assumption). However, this can badly break down when there isn't linear separability. Concretely, suppose that $\|\mathbf{x}_{|d-k}\| = 1$ with probability 1, and that we introduce label noise, so that the sign of y is flipped with some probability p. In this case, the expected loss can be written as

$$L_p(\mathbf{w}) \ = \ \mathbb{E}_{(\mathbf{x}_{|k}, y)} \left[(1 - p) \cdot [1 - y\mathbf{x}_{|k}^\top \mathbf{w}]_+^2 + p \cdot [1 + y\mathbf{x}_{|k}^\top \mathbf{w}]_+^2 \right] \ ,$$

where the expectation is with respect to the "clean" labels. In this case, the minimizer of the above might have an expected misclassification error of $1/2$ (even if p is arbitrarily small). To see this, it is enough to produce some finite linearly-separable dataset, such that 50% of the points will be misclassified by the minimizer of $L_p(\cdot)$ (and then random label flipping will keep the error rate at 50%). The existence of such a dataset was essentially shown for a more general setting in Long and Servedio (2010), and below we instantiate their analysis for our setting with more explicit guarantees:

Proposition 20 *For any $p \in (0, \frac{1}{12})$, there exists a dataset $\{\mathbf{x}_i, y_i\}_{i=1}^4 \subseteq \mathbb{R}^2 \times \{-1, +1\}$, where $\max_i \|\mathbf{x}\|_i \leq 1$, such that:*

- *There exists a unit vector $\mathbf{w}^*$ for which $\min_i y_i \mathbf{x}_i^\top \mathbf{w}^* \geq p$*

- *If $\hat{\mathbf{w}}$ is a minimizer of $L_p(\mathbf{w}) = \frac{1}{4} \sum_{i=1}^4 \left((1-p) \cdot [1 - y_i \mathbf{x}_i^\top \mathbf{w}]_+^2 + p \cdot [1 + y_i \mathbf{x}_i^\top \mathbf{w}]_+^2 \right)$, then $\hat{\mathbf{w}}$ misclassifies two of the four points.*

Proof

Let $y_1 = y_2 = y_3 = y_4 = 1$, and

$$\mathbf{x}_1 = \mathbf{x}_2 = \begin{pmatrix} p \\ -p \end{pmatrix} \, , \, \mathbf{x}_3 = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \, , \, \mathbf{x}_4 = \begin{pmatrix} p \\ 5p \end{pmatrix} \, .$$

It is easily verified that $\mathbf{w}^* = (p, 0)$ satisfies $\min_i y_i \mathbf{x}_i^\top \mathbf{w}^* \geq p$. Also, by Lemma 19, it is easily verified that $L_p(\cdot)$ is strongly convex. Therefore, the minimizer is unique, and we claim that for any small enough $p > 0$, it equals

$$\hat{\mathbf{w}} = \left(\frac{5 - 16p}{3 + 8p} \, , \, \frac{1-p}{3p(3 + 8p)} \right) \, .$$

In that case, for the two points $\mathbf{x}_1 = \mathbf{x}_2 = (p, -p)$,

$$\mathbf{x}_1^\top \hat{\mathbf{w}} = \mathbf{x}_2^\top \hat{\mathbf{w}} = -\frac{1 - 16p + 48p^2}{9 + 24p} \, ,$$

which is negative for any small enough $p \in (0, \frac{1}{12})$, hence two of the four points are misclassified. To verify that $\hat{\mathbf{w}}$ above is indeed the minimizer, let $\ell_p(z) := (1-p)[1 - z]_+^2 + p[1 + z]_+^2$ (so that $L_p(\mathbf{w}) = \mathbb{E}_{(\mathbf{x},y)}[\ell_p(y\mathbf{x}^\top \mathbf{w})]$), and note that

$$4 \cdot \nabla L_p(\hat{\mathbf{w}}) = 4 \cdot \nabla L_p(\hat{w}_1, \hat{w}_2) = 2p\ell_p'(p(\hat{w}_1 - \hat{w}_2)) \begin{pmatrix} 1 \\ -1 \end{pmatrix} + \ell_p'(w_1) \begin{pmatrix} 1 \\ 0 \end{pmatrix} + p\ell_p'(p(\hat{w}_1 + 5\hat{w}_2)) \begin{pmatrix} 1 \\ 5 \end{pmatrix},$$

$$\tag{28}$$

where

$$\ell_p'(z) = -2(1-p)[1 - z]_+ + 2p[1 + z]_+ \, .$$

A tedious but routine calculation shows that for any $p \in (0, \frac{1}{12})$, it holds that $p(\hat{w}_1 - \hat{w}_2) \in [-1, 0)$, $p(\hat{w}_1 + 5\hat{w}_2) \in [0, 1]$, and $\hat{w}_1 > 1$. Plugging in the corresponding expressions for $\ell_p'(z)$ into Eq. (28), we get the $\mathbf{0}$ vector. Hence, $\nabla L_p(\hat{\mathbf{w}})$, and since $L_p(\cdot)$ is convex, it follows that $\hat{\mathbf{w}}$ is indeed its minimizer. ∎

References

Gregor Bachmann, Seyed-Mohsen Moosavi-Dezfooli, and Thomas Hofmann. Uniform convergence, adversarial spheres and a simple remedy. *arXiv preprint arXiv:2105.03491*, 2021.

Peter L Bartlett and Philip M Long. Failures of model-dependent generalization bounds for least-norm interpolation. *Journal of Machine Learning Research*, 22(204):1–15, 2021.

Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

Mikhail Belkin. Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation. *arXiv preprint arXiv:2105.14368*, 2021.

Mikhail Belkin, Daniel J Hsu, and Partha Mitra. Overfitting or perfect fitting? risk bounds for classification and regression rules that interpolate. *Advances in Neural Information Processing Systems*, 31:2300–2311, 2018a.

Mikhail Belkin, Siyuan Ma, and Soumik Mandal. To understand deep learning we need to understand kernel learning. In *International Conference on Machine Learning*, pages 541–549. PMLR, 2018b.

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, 2019a.

Mikhail Belkin, Alexander Rakhlin, and Alexandre B Tsybakov. Does data interpolation contradict statistical optimality? In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1611–1619. PMLR, 2019b.

Mikhail Belkin, Daniel Hsu, and Ji Xu. Two models of double descent for weak features. *SIAM Journal on Mathematics of Data Science*, 2(4):1167–1180, 2020.

Yuan Cao, Quanquan Gu, and Mikhail Belkin. Risk bounds for over-parameterized maximum margin classification on sub-gaussian mixtures. *arXiv preprint arXiv:2104.13628*, 2021.

Niladri S Chatterji and Philip M Long. Finite-sample analysis of interpolating linear classifiers in the overparameterized regime. *Journal of Machine Learning Research*, 22(129): 1–30, 2021.

Spencer Frei, Yuan Cao, and Quanquan Gu. Agnostic learning of halfspaces with gradient descent via soft margins. In *International Conference on Machine Learning*, pages 3417–3426. PMLR, 2021.

Spencer Frei, Niladri S Chatterji, and Peter Bartlett. Benign overfitting without linearity: Neural network classifiers trained by gradient descent for noisy linear data. In *Conference on Learning Theory*, pages 2668–2703. PMLR, 2022.

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

Tianyang Hu, Jun Wang, Wenjia Wang, and Zhenguo Li. Understanding square loss in training overparametrized neural network classifiers. *arXiv preprint arXiv:2112.03657*, 2021.

Ziwei Ji and Matus Telgarsky. Directional convergence and alignment in deep learning. *Advances in Neural Information Processing Systems*, 33, 2020.

Ziwei Ji, Kwangjun Ahn, Pranjal Awasthi, Satyen Kale, and Stefani Karp. Agnostic learnability of halfspaces via logistic loss. In *International Conference on Machine Learning*, pages 10068–10103. PMLR, 2022.

Frederic Koehler, Lijia Zhou, Danica J Sutherland, and Nathan Srebro. Uniform convergence of interpolators: Gaussian width, norm bounds, and benign overfitting. *arXiv preprint arXiv:2106.09276*, 2021.

Tengyuan Liang and Alexander Rakhlin. Just interpolate: Kernel "ridgeless" regression can generalize. *The Annals of Statistics*, 48(3):1329–1347, 2020.

Tengyuan Liang and Benjamin Recht. Interpolating classifiers make few mistakes. *arXiv preprint arXiv:2101.11815*, 2021.

Philip M Long and Rocco A Servedio. Random classification noise defeats all convex potential boosters. *Machine learning*, 78(3):287–304, 2010.

Andrew D McRae, Santhosh Karnik, Mark A Davenport, and Vidya Muthukumar. Harmless interpolation in regression and classification with structured features. *arXiv preprint arXiv:2111.05198*, 2021.

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*, 2019.

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

Vidya Muthukumar, Adhyyan Narang, Vignesh Subramanian, Mikhail Belkin, Daniel Hsu, and Anant Sahai. Classification vs regression in overparameterized regimes: Does the loss function matter? *Journal of Machine Learning Research*, 22(222):1–69, 2021.

Vaishnavh Nagarajan and J Zico Kolter. Uniform convergence may be unable to explain generalization in deep learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Jeffrey Negrea, Gintare Karolina Dziugaite, and Daniel Roy. In defense of uniform convergence: Generalization via derandomization with an application to interpolating predictors. In *International Conference on Machine Learning*, pages 7263–7272. PMLR, 2020.

Tomaso Poggio and Qianli Liao. Generalization in deep network classifiers trained with the square loss. Technical report, CBMM Memo No, 2019.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878, 2018.

Christos Thrampoulidis. Theoretical insights into multiclass classification: A high-dimensional asymptotic view. *Neural Information Processing Systems (NeuRIPS 2020)*, 2020.

Ke Wang and Christos Thrampoulidis. Benign overfitting in binary classification of gaussian mixtures. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4030–4034. IEEE, 2021.

Ke Wang, Vidya Muthukumar, and Christos Thrampoulidis. Benign overfitting in multiclass classification: All roads lead to interpolation. *arXiv preprint arXiv:2106.10865*, 2021.

Zitong Yang, Yu Bai, and Song Mei. Exact gap between generalization error and uniform convergence in random feature models. *arXiv preprint arXiv:2103.04554*, 2021.

Gilad Yehudai and Ohad Shamir. Learning a single neuron with gradient methods. In *Conference on Learning Theory*, pages 3756–3786. PMLR, 2020.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *5th International Conference on Learning Representations, ICLR*, 2017.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

Lijia Zhou, DJ Sutherland, and Nati Srebro. On uniform convergence and low-norm interpolation learning. *Advances in Neural Information Processing Systems*, 33, 2020.

Lijia Zhou, Frederic Koehler, Danica J Sutherland, and Nathan Srebro. Optimistic rates: A unifying theory for interpolation learning and regularization in linear regression. *arXiv preprint arXiv:2112.04470*, 2021.

Shenghuo Zhu. A short note on the tail bound of wishart distribution. *arXiv preprint arXiv:1212.5860*, 2012.