The Implicit Bias of Benign Overfitting

Ohad Shamir OHAD.SHAMIR@WEIZMANN.AC.IL
Weizmann Institute of Science

Editors: Po-Ling Loh and Maxim Raginsky

Abstract

The phenomenon of benign overfitting, where a predictor perfectly fits noisy training data while attaining low expected loss, has received much attention in recent years, but still remains not fully understood beyond simple linear regression setups. In this paper, we show that for regression, benign overfitting is "biased" towards certain types of problems, in the sense that its existence on one learning problem precludes its existence on other learning problems. On the negative side, we use this to argue that one should not expect benign overfitting to occur in general, for several natural extensions of the plain linear regression problems studied so far. We then turn to classification problems, and show that the situation there is much more favorable. Specifically, we consider a model where an arbitrary input distribution of some fixed dimension k is concatenated with a high-dimensional distribution, and prove that the max-margin predictor (to which gradient-based methods are known to converge in direction) is asymptotically biased towards minimizing the expected *squared hinge loss* w.r.t. the k-dimensional distribution. 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

Benign overfitting is an intriguing phenomenon in statistical learning, which has received much interest in the past few years, and appears to occur frequently in large-scale learning problems (such as deep learning). It refers to situations which combine the following: (1) The trained predictor achieves essentially perfect prediction accuracy on the training data; (2) No predictor in the relevant hypothesis class can achieve perfect prediction accuracy w.r.t. the underlying data distribution (e.g., the Bayes error is strictly positive); yet (3) The trained predictor has good prediction accuracy w.r.t. the underlying data distribution. This phenomenon is intriguing, because 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 $y = \mathbf{x}^\top \mathbf{w}^* + \xi$, where ξ is zero-mean noise. In this setting, we know that benign overfitting occurs (roughly speaking) whenever the distribution of the input $\mathbf{x}$ has a high-dimensional distribution, with 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 examples. A helpful feature of this setting is that there is a closed-form expression for the predictor returned by gradient-based methods, when trained to convergence on the average square loss. For classification, the situation is more complicated, because the predictor that 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 propose a new perspective on benign overfitting, which can potentially be used to analyze when benign overfitting may – or may not – occur in settings beyond those studied so far in the literature. Roughly speaking, we argue that any predictor that perfectly fits the training data 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 (w.r.t. the underlying data distribution). The trained predictor doesn't "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.

To make the argument a bit more concrete, let us consider a linear prediction setup, where we have some non-negative loss function $\ell(p; y)$ so that the loss of a predictor $\mathbf{w} \in \mathbb{R}^d$ on an example $(\mathbf{x}, y)$ equals $\ell(\mathbf{x}^\top \mathbf{w}; y)$. Thus, our goal is to minimize $\mathbb{E}_{(\mathbf{x},y)}[\ell(\mathbf{x}^\top \mathbf{w}; y)]$ over $\mathbf{w}$. With an eye towards benign overfitting, suppose that we attempt this by running a gradient-based method over the empirical risk objective $\frac{1}{m}\sum_{i=1}^m \ell(\mathbf{x}_i^\top \mathbf{w}; y_i)$ (with respect to an i.i.d. training set $\{\mathbf{x}_i, y_i\}_{i=1}^m$, and without any regularization), assuming we achieve a globally minimal solution $\hat{\mathbf{w}}$. In regression, $\ell(\cdot)$ is commonly such that for any value y, there is some unique value p such that $\ell(p; y) = 0$. In this case, we understand where gradient-based methods converge to, as shown in the following theorem (the proof, like most proofs in the paper, appears in the appendix):

Theorem 1 *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 tend to converge to minimum-norm solutions that minimize the empirical risk. In itself, the theorem is not surprising, and is based on well-known ideas (see for example Zhang et al. (2021) for a derivation of a somewhat more specific version). However, our crucial observation is the following rather immediate corollary:

Corollary 2 *The point $\hat{\mathbf{w}}$ defined in Thm. 1 is also the (unique) point that satisfies $\tilde{L}(\hat{\mathbf{w}}) = 0$ and $\hat{\mathbf{w}} \in$ span$\{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$, where $\tilde{L}(\mathbf{w}) := \frac{1}{m}\sum_{i=1}^m (\mathbf{x}_i^\top \mathbf{w} - \ell_{y_i}^{-1}(0))^2$.*

The corollary follows simply by applying the theorem on the "loss" function $\tilde{\ell}(\mathbf{x}_i^\top \mathbf{w}; y_i) := (\mathbf{x}_i^\top \mathbf{w} - \ell_{y_i}^{-1}(0))^2$, which is non-negative and equals zero when $\mathbf{x}_i^\top \mathbf{w} = \ell_{y_i}^{-1}(0)$. Hence, the same

method that converges to the minimum-norm root of $L(\cdot)$, also simultaneously converges to the minimum-norm root of $\tilde{L}(\cdot)$.

Comparing $L(\cdot)$ from the theorem and $\tilde{L}(\cdot)$ from the corollary, we see that they are both the empirical average of a certain loss function over m training examples. 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. However, 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 the paper, we build on this simple observation to provide several new results, both positive and negative, about the existence of benign overfitting in various regression and classification problems. We focus on a prototypical setting for benign overfitting, where the input distribution in $\mathbb{R}^d$ is composed of an arbitrary fixed input distribution on the first k coordinates, and a high-dimensional Gaussian distribution on the other $d-k$ coordinates (which facilitates overfitting the training data). In this setting, we study consistency of the predictor learned by gradient-based methods, as both d, the sample size and their ratio diverge to infinity. Informally, our results are as follows:

- Beginning with regression, we provide evidence that benign overfitting generally *does not* occur in several natural extensions of the linear regression settings studied so far. These include (1) linear regression with the square loss in an agnostic or misspecified setting (where $\mathbb{E}[y|\mathbf{x}]$ does not necessary equal $\mathbf{x}^\top \mathbf{w}^*$ for some fixed $\mathbf{w}^*$, see Thm. 6); (2) Regression with generalized linear models (or equivalently, with a single neuron predictor, see Thm. 7); and (3) Regression with losses other than the square loss (see Thm. 9). In line with our observations above, these negative results hold exactly *because* benign overfitting can occur in high-dimensional well-specified linear regression problems.

- The negative results above depend on the fact that in regression, we generally require the predictor's output to exactly equal some target value, which is a rather stringent requirement. In contrast, classification problems are easier, in the sense that we only need the confidence in predicting one class to be larger than the confidence in another. Indeed, papers such as Muthukumar et al. (2021) pointed out that in natural setups, benign overfitting in classification can occur even when benign overfitting in regression fails. Inspired by our previous observations, we investigate the behavior of the learned predictor for binary classification, under the same input distribution as earlier (where an arbitrary distribution on the first k coordinates is concatenated with a high-dimensional distribution). Perhaps unexpectedly, we show that in this model, the max-margin predictor (to which gradient-based methods are known to asymptotically converge in direction) is implicitly biased towards minimizing the expected *squared hinge loss* w.r.t. the underlying data distribution (see Thm. 10). Thus, the consistency of the learned predictor (and hence benign overfitting) in our model is reduced to a simpler question: Whether the data distribution is such that minimizing the squared hinge loss is a good surrogate for minimizing misclassification error. We note that unlike many previous works on benign overfitting in classification, we do not require that the distribution is such that the least-squares 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 (see Thm. 12). Moreover, under some stronger distributional assumptions, we will have benign overfitting for label noise arbitrarily close to $1/2$ (see Thm. 13).

Overall, we hope that the perspective suggested in this paper will allow us to understand benign overfitting beyond the settings studied so far. For example, it would be interesting to identify other settings where the structure of the square hinge loss means that the max-margin predictor will have benign overfitting properties, as well as other distributions which lead to similar implicit biases.

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 occurence 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. 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. 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).

2. Preliminaries

Notation. We use bold-faced letter to denote vectors, and assume that they are in column form unless specified otherwise. Let $\mathbf{e}_i \in \mathbb{R}^d$ be the i-th standard basis vector. 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 $z \mapsto \max\{0, z\}$. $[m]$ is shorthand for $\{1, \ldots, m\}$. I_d is the $d \times d$ identity matrix. We use standard asymptotic notation $\mathcal{O}(\cdot), \Omega(\cdot)$ to hide constants (generally with respect to a dimension parameter $d \to \infty$). We use $\xrightarrow{a.s.}$ and $\xrightarrow{P}$ to denote convergence almost surely and in probability, respectively (of a sequence of random variables to some fixed limit).

Benign Overfitting. 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 diverges to infinity. For the setting studied in Thm. 1, a common way to define benign overfitting as follows:

Definition 3 (Benign Overfitting for minimum-norm interpolators) *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 is a monotonically increasing sequence of 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 almost surely, there exists some $\mathbf{w} \in \mathbb{R}^d$ such that $\frac{1}{m_d}\sum_{i=1}^{m_d} \ell(\mathbf{x}_i^\top \mathbf{w}; y_i)) = 0$.*
- *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 minimum-norm minimizer of the average loss over the dataset, 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 $\inf_{d>k} \inf_{\mathbf{w}\in\mathbb{R}^d} R_d(\mathbf{w}) > 0$ as well as $R_d(\hat{\mathbf{w}}_d) - \inf_{\mathbf{w}\in\mathbb{R}^d} R_d(\mathbf{w}) \xrightarrow{P} 0$.*

In other words, $\hat{\mathbf{w}}_d$ 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 ∞ at an appropriate rate. Defining benign overfitting in classification is a bit different, and is left to Sec. 4.

3. Regression

We begin by considering a regression setup, where the loss of a predictor $\mathbf{w}$ on an example $(\mathbf{x}, y)$ is denoted as $\ell(\mathbf{x}^\top \mathbf{w}; y)$, and is minimized w.r.t. the first argument at some unique point. A prototypical case where we might hope to have benign overfitting is when the inputs are composed or a "signal" component, concatenated with a high-dimensional random component which can be used to fit the training data, but without materially affecting the performance on new examples. Formally, we focus on the following particularly nice instantiation of this idea (in the context of benign overfitting as defined in the previous section):

Assumption 1 *For any distribution $\mathcal{D}_d$, $d > k$ on $(\mathbf{x}, y) \in \mathbb{R}^d \times \mathbb{R}$, the distribution of y and $\mathbf{x} = (\mathbf{x}_{|k}, \mathbf{x}_{|d-k})$ (where we separate the first k and last $d-k$ coordinates) satisfies the following:*

- $\mathbf{x}_{|k}$ *and* y *have a fixed distribution independent of* d*, such that* $\mathbb{E}[\|\mathbf{x}_{|k}\|^2]$, $\mathbb{E}[y^2]$, $\mathbb{E}[\|y\mathbf{x}_{|k}\|^2]$ *are all finite, and* $\mathbb{E}[\mathbf{x}_{|k}\mathbf{x}_{|k}^\top]$ *is positive definite.*
- $\mathbf{x}_{|d-k}$ *is independent of* $\mathbf{x}_{|k}$*, and is zero-mean Gaussian with covariance matrix* $\frac{1}{d-k} \cdot I_{d-k}$.

Remark 4 *In what follows, it will be convenient to embed a sequence of random variables (each being a function of samples from a different* $\mathcal{D}_d$*) in one common probability space, so that notions such as almost sure convergence make sense. Formally, we can do this as follows: Consider a doubly-infinite matrix* X*, where the first* k *columns correspond to an infinite i.i.d. sequence of input vectors in* $\mathbb{R}^k$*, distributed as the first* k *coordinates of* $\mathcal{D}_d$ *(which is independent of* d*), and the rest of the entries are i.i.d. standard Gaussian. Also, we let* $\mathbf{y}$ *be an infinite i.i.d. sequence of output values distributed as in* $\mathcal{D}_d$ *(which is independent of* d*). Then, for each* d *and integer* m_d*, we let* $\{\mathbf{x}_i\}_{i=1}^{m_d}$ *be the rows of the* $m_d \times d$ *top-left submatrix of* X*, with the last* $d-k$ *coordinates of each* $\mathbf{x}_i$ *multiplied by* $\frac{1}{\sqrt{d-k}}$*. Also, we let* $\{y_i\}_{i=1}^{m_d}$ *be the first* m_d *entries of* $\mathbf{y}$*. By Assumption 1,* $\{(\mathbf{x}_i, y_i)\}_{i=1}^{m_d}$ *has the same distribution as an i.i.d. sample from* $\mathcal{D}_d$*, for all* d*, while sharing a common probability space.*

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. We note that we scale the covariance matrix by $\frac{1}{d-k}$, to ensure that $\|\mathbf{x}\|$ is almost surely bounded as $d \to \infty$. Moreover, it means that the values we choose for the predictor $\mathbf{w}$ in the last $d-k$ coordinates are asymptotically immaterial (as we will prove formally in our theorems), and they will serve merely to fit the training data. Also, we note that the choice of a Gaussian distribution on the last $d-k$ coordinates is merely for simplicity: Essentially, our analysis only really requires a light-tailed distribution which is "spread", in the sense that independent samples are asymptotically orthogonal (as $d \to \infty$), and that the minimum-norm interpolator exists with arbitrarily high probability for large enough d. The choice of the distribution will only affect technical details such as how fast d needs to grow compared to the sample size m_d. For our choice, we will need the following assumption:

Assumption 2 *The sequence of positive integers* $\{m_d\}_{d=k+1}^{\infty}$ *is monotonically increasing, diverges to* ∞*, satisfies* $m_d \leq d-k$ *for all* d*, and* $\lim_{d\to\infty} \frac{m_d^3 \log(d)}{d} = 0$.

We now present a basic theorem on the asymptotic behavior of the minimum-norm interpolator:

Theorem 5 *For any sequence of distributions* $\{\mathcal{D}_d\}_{d=k+1}^{\infty}$ *and integers* $\{m_d\}_{d=k+1}^{\infty}$ *satisfying Assumptions 1 and 2, it holds that if* $\hat{\mathbf{w}}_d := \arg\min_{\mathbf{w}\in\mathbb{R}^d} \|\mathbf{w}\| : \frac{1}{m_d}\sum_{i=1}^{m_d}(\mathbf{x}_i^\top \mathbf{w} - y_i)^2 = 0$ *, where* $\{\mathbf{x}_i, y_i\}_{i=1}^{m_d}$ *are sampled i.i.d. from* $\mathcal{D}_d$*, then the sequence* $\{\hat{\mathbf{w}}_d\}_{d=k+1}^{\infty}$ *satisfies*

$$\hat{\mathbf{w}}_{d|k} \xrightarrow{a.s.} \mathbb{E}\left[\mathbf{x}_{|k}\mathbf{x}_{|k}^\top\right]^{-1} \cdot \mathbb{E}\left[y\mathbf{x}_{|k}\right] \quad \text{and} \quad \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_d}\left[(\mathbf{x}^\top\hat{\mathbf{w}}_d - \mathbf{x}_{|k}^\top\hat{\mathbf{w}}_{d|k})^2\right] \xrightarrow{a.s.} 0\,.$$

In other words, the first k coordinates of $\hat{\mathbf{w}}_d$ converges almost surely to the (unique) minimum-norm minimizer of $\mathbb{E}[(\mathbf{x}_{|k}^\top\mathbf{w} - y)^2]$, and the last $d-k$ coordinates are asymptotically irrelevant. Since the behavior of any other fixed predictor $\mathbf{w}^*$ will also asymptotically depend just on its first k coordinates, we see that the expected loss of $\hat{\mathbf{w}}_d$ converges to a minimal value over all linear

predictors. Assuming no predictor makes the expected square loss precisely zero, we therefore get benign overfitting (as defined in Definition 3). .

In itself, this result is not surprising: We deliberately focus on a model which is well-known to be particularly amenable to benign overfitting. However, inspired by the observations discussed in the introduction, let us show how a slight tweak of this setup easily makes benign overfitting *fail* to occur in general. In fact, this already occurs for linear regression with the square loss, as formalized in the following theorem:

Theorem 6 *Consider any sequence of distributions $\{\mathcal{D}_d\}_{d=k+1}^{\infty}$ and integers $\{m_d\}_{d=k+1}^{\infty}$ satisfying Assumptions 1 and 2, except that conditioned on $\mathbf{x}_{|k}$, $\mathbf{x}_{|d-k}$ is zero-mean Gaussian with a covariance matrix $\frac{g(\mathbf{x}_{|k})}{d-k} \cdot I_{d-k}$ (where $g : \mathbb{R}^k \mapsto \mathbb{R}$ is any measurable function such that $\Pr\left(g(\mathbf{x}_{|k}) \in (l,u)\right) = 1$ for some positive $l, u \in \mathbb{R}$). If we let $\hat{\mathbf{w}}_d := \arg\min_{\mathbf{w}\in\mathbb{R}^d} \|\mathbf{w}\| : \frac{1}{m_d}\sum_{i=1}^{m_d}(\mathbf{x}_i^\top \mathbf{w} - y_i)^2 = 0$ where $\{\mathbf{x}_i, y_i\}_{i=1}^{m_d}$ is an i.i.d. sample from $\mathcal{D}_d$, then*

$$\hat{\mathbf{w}}_{d|k} \xrightarrow{a.s.} \left(\mathbb{E}\left[\frac{\mathbf{x}_{|k}\mathbf{x}_{|k}^\top}{g(\mathbf{x}_{|k})}\right]\right)^{-1} \cdot \mathbb{E}\left[\frac{y\mathbf{x}_{|k}}{g(\mathbf{x}_{|k})}\right] \quad \text{and} \quad \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_d}\left[(\mathbf{x}^\top\hat{\mathbf{w}}_d - \mathbf{x}_{|k}^\top\hat{\mathbf{w}}_{d|k})^2\right] \xrightarrow{a.s.} 0\,.$$

Proof $\hat{\mathbf{w}}_d$ can be equivalently written as $\arg\min_{\mathbf{w}\in\mathbb{R}^d} \|\mathbf{w}\| : \frac{1}{m_d}\sum_{i=1}^{m_d}\left(\frac{\mathbf{x}_i^\top}{\sqrt{g(\mathbf{x}_{i|k})}}\mathbf{w} - \frac{y_i}{\sqrt{g(\mathbf{x}_{i|k})}}\right)^2 = 0$. Thus, $\hat{\mathbf{w}}_d$ can be seen as a minimal norm interpolator of a related linear regression problem, where we scale both $\mathbf{x}_i$ and y_i by $\sqrt{g(\mathbf{x}_{|k})}$. We note that now this falls exactly in the framework of Assumption 1, with the last $d-k$ coordinates having a marginal zero-mean Gaussian distribution with covariance matrix $\frac{1}{d-k}\cdot I_{d-k}$; where the first k coordinates have the distribution $\frac{1}{\sqrt{g(\mathbf{x}_{|k})}}\mathbf{x}_{|k}$; and where the target values have distribution $\frac{y}{\sqrt{g(\mathbf{x}_{|k})}}$. Plugging this into Thm. 5 and using the assumptions on $g(\cdot)$, the result easily follows. ■

The important point to notice here is that now the predictor we converge to is such that its last $d-k$ coordinates are still asymptotically negligible, whereas the first k coordinates *do not* look like the standard minimum-norm optimal predictor we would expect, which is still $\mathbb{E}[\mathbf{x}_{|k}\mathbf{x}_{|k}^\top]^{-1}\mathbb{E}[y\mathbf{x}_{|k}]$ in the first k coordinates. Thus, unless the two somehow exactly coincide, we should not expect benign overfitting to occur, even though the covariance structure of the inputs $\mathbf{x}$ is a textbook case of amenability to benign overfitting (in the sense of having many small positive eigenvalues). The following example illustrates this:

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

$$R_d(\mathbf{w}) := \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_d}[(\mathbf{x}^\top\mathbf{w} - y)^2] = \mathbb{E}[x_1^2]\cdot w_1^2 + \frac{\mathbb{E}[g(x_1)]}{d-1}\cdot\sum_{j=2}^{d} w_j^2 - 2\mathbb{E}[yx_1]\cdot w_1 + \mathbb{E}[y^2]\,.$$

$R_d(\cdot)$ 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 Thm. 6, the first coordinate

of $\hat{\mathbf{w}}_d$ *converges almost surely to the different value* $\mathbb{E}\left[\frac{yx_1}{g(x_1)}\right] / \mathbb{E}\left[\frac{x_1^2}{g(x_1)}\right] = \frac{\mathbb{E}[x_1]}{\mathbb{E}[x_1^2 \exp(-x_1)]} = 0$. *Thus, we get 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 (under assumption 1, 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 $\mathbb{E}\left[\frac{\mathbf{x}_{|k}\mathbf{x}_{|k}^\top}{g(\mathbf{x}_{|k})}\right]^{-1} \cdot \mathbb{E}\left[\frac{y\mathbf{x}_{|k}}{g(\mathbf{x}_{|k})}\right]$ equals $\mathbb{E}\left[\frac{\mathbf{x}_{|k}\mathbf{x}_{|k}^\top}{g(\mathbf{x}_{|k})}\right]^{-1} \cdot \mathbb{E}\left[\frac{\mathbf{x}_{|k}\mathbf{x}_{|k}^\top}{g(\mathbf{x}_{|k})}\right] \mathbf{w}_{|k}^* = \mathbf{w}_{|k}^*$, which now coincides with the optimal solution on the first k coordinates (which equals $\mathbb{E}[\mathbf{x}_{|k}\mathbf{x}_{|k}^\top]^{-1} \cdot \mathbb{E}\left[y\mathbf{x}_{|k}\right] = \mathbb{E}[\mathbf{x}_{|k}\mathbf{x}_{|k}^\top]^{-1} \cdot \mathbb{E}[\mathbf{x}_{|k}\mathbf{x}_{|k}^\top]\mathbf{w}_{|k}^* = \mathbf{w}_{|k}^*$).

Next, we turn to study two important settings beyond linear regression with the square loss, and in both cases, show that we should not expect benign overfitting to occur in general, exactly because it generally occurs for linear regression with the square loss under Assumption 1.

The first model we study is when we wish to fit a generalized linear model, namely a predictor of the form $\mathbf{x} \mapsto \sigma(\mathbf{x}^\top \mathbf{w})$, where $\mathbf{w}$ is the parameter vector and $\sigma(\cdot)$ is some fixed non-linear function. In the context of neural networks, this can also be seen as training a single neuron using some nonlinear activation function $\sigma(\cdot)$. In this setting, it is not difficult to show that 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 Thm. 1). The following theorem 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 in a well-specified setting where $\mathbb{E}[y|\mathbf{x}] = \sigma(\mathbf{x}^\top \mathbf{w}^*)$ for some $\mathbf{w}^*$:

Theorem 7 *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 Assumptions 1 and 2, 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*. Given a training set* $\{\mathbf{x}_i, y_i\}_{i=1}^{m_d}$ *sampled i.i.d. from* $\mathcal{D}_d$*, let* $\hat{\mathbf{w}}_d = \arg\min_{\mathbf{w}} \|\mathbf{w}\| : \frac{1}{m_d}\sum_{i=1}^{m_d}(\sigma(\mathbf{x}_i^\top \mathbf{w}) - y_i)^2 = 0$*. Then* $\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_d}\left[(\mathbf{x}^\top \hat{\mathbf{w}}_d - \mathbf{x}_{|k}^\top \hat{\mathbf{w}}_{d|k})^2\right] \xrightarrow{a.s.} 0$ *and*

$$\hat{\mathbf{w}}_{d|k} \xrightarrow{a.s.} \left(\mathbb{E}[\mathbf{x}_{|k}\mathbf{x}_{|k}^\top]\right)^{-1} \cdot \mathbb{E}\left[\sigma^{-1}\left(\sigma(\mathbf{x}_{|k}^\top \mathbf{w}^*) + \xi\right)\mathbf{x}_{|k}\right] .$$

The theorem follows immediately from the observation that $\hat{\mathbf{w}}_d$ 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 $|\sigma^{-1}(y)| \le c_\sigma(1 + |y|)$ for some $c_\sigma > 0$ dependent only on σ). Hence, by Thm. 5, $\hat{\mathbf{w}}_{d|k}$ converges almost surely to $\left(\mathbb{E}[\mathbf{x}_{|k}\mathbf{x}_{|k}^\top]\right)^{-1}\mathbb{E}[\sigma^{-1}(y)\mathbf{x}_{|k}]$, which equals the expression stated in the theorem above.

When ξ is independent zero-mean noise, and $\sigma(\cdot)$ (and hence $\sigma^{-1}(\cdot)$) are linear functions, then the asymptotic expression for $\hat{\mathbf{w}}_{d|k}$ in the theorem easily reduces to $\mathbf{w}^*$, which is indeed the optimal predictor we would hope to converge to. However, this generally breaks when $\sigma(\cdot)$ is nonlinear. To give just one simple example, suppose that $\sigma(0) = 0$, $\mathbf{w}^* = \mathbf{0}$ and $\mathbb{E}[\mathbf{x}_{|k}] \neq \mathbf{0}$, in which case the asymptotic expression 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. 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 fitting 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 8 *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 the loss is not the square loss (say, the absolute loss). Here again, we cannot expect benign overfitting to occur in general:

Theorem 9 *Consider any sequence of distributions $\{\mathcal{D}_d\}_{d=k+1}^{\infty}$ and integers $\{m_d\}_{d=k+1}^{\infty}$ satisfying Assumptions 1 and 2. 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. Let $\hat{\mathbf{w}}_d := \arg\min_{\mathbf{w}} \|\mathbf{w}\| : \frac{1}{m_d}\sum_{i=1}^{m_d} \ell(\mathbf{x}_i^\top\mathbf{w}; y_i) = 0$. Then $\hat{\mathbf{w}}_{d|k} \xrightarrow{a.s.} \mathbb{E}\left[\mathbf{x}_{|k}\mathbf{x}_{|k}^\top\right]^{-1} \cdot \mathbb{E}\left[y\mathbf{x}_{|k}\right]$ and $\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_d}\left[(\mathbf{x}^\top\hat{\mathbf{w}}_d - \mathbf{x}_{|k}^\top\hat{\mathbf{w}}_{d|k})^2\right] \xrightarrow{a.s.} 0$.*

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 Thm. 5 on this related linear regression problem. Crucially, note that $\hat{\mathbf{w}}_{d|k}$ converges almost surely to the unique minimum-norm minimizer of $\mathbb{E}[(\mathbf{x}_{|k}^\top\mathbf{w} - y)^2]$, 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 2 *In the setting of Thm. 9, suppose $f(z) = |z|$ is the absolute loss, $k = 1$, $x_1 = 1$ with probability 1, and $y = x_1 + \xi$ for some independent zero-mean noise term ξ. Then the first coordinate of $\hat{\mathbf{w}}_d$ converges almost surely to $\mathbb{E}[yx_1]/\mathbb{E}[x_1^2] = 1$. 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 $w_1 = 1 + med(\xi)$ (where $med(\xi)$ is the median of ξ). Thus, whenever $med(\xi) \neq 0 = \mathbb{E}[\xi]$ (which occurs whenever 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.*

4. Classification

The results in the previous section suggest that many natural extensions of well-specified linear regression will generally not satisfy benign overfitting in our model. These were all regression problems, where to get (asymptotically) optimal accuracy required the predictions to converge to some single 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. More specifically, we will now 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 direction of the predictor $\mathbf{w}$, and its expected misclassification rate, namely $\Pr_{(\mathbf{x},y)}(y\mathbf{x}^\top \mathbf{w} \leq 0)$.

In regression, we saw that gradient-based methods ran on the empirical risk function generally converge to the minimum-norm predictor which zeroes the empirical risk, assuming such a predictor exists (see Thm. 1). For linear classification, the characterization is a bit different: Using a convex classification loss with exponential tails (such as the logistic 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. Our goal now is to understand when can we hope to have benign overfitting for $\hat{\mathbf{w}}$. 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 for any large enough d, $\hat{\mathbf{w}}_d$ exists almost surely, and

$$\inf_d \inf_{\mathbf{w} \in \mathbb{R}^d} R_d(\mathbf{w}) > 0 \quad \text{as well as} \quad \lim_{d \to \infty} \mathbb{E}\left[R_d(\hat{\mathbf{w}}_d) - \inf_{\mathbf{w} \in \mathbb{R}^d} R_d(\mathbf{w})\right] = 0\,. \tag{1}$$

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.

As in the case of regression, we will focus on distributions which satisfy Assumption 1, namely where some arbitrary distribution on the first k coordinates is concatenated with a bounded-norm Gaussian distribution on the last $d-k$ coordinates. Again, this is a prototypical model for which we might hope for benign overfitting to occur, and the Gaussianity assumption can be relaxed to other light-tailed and "spread" distributions.

In regression, we were able to characterize the asymptotic behavior of $\hat{\mathbf{w}}_d$ as the least-square solution w.r.t. the data distribution on the first k coordinates. To understand the case of classification, we will need a similar characterization of what $\hat{\mathbf{w}}_d$ converges to. At first glance, this might seem difficult, as the max-margin predictor on a given dataset does not have a closed-form expression (unlike the case of the least squares solution), and in fact, many previous analyses of benign overfitting in classification resorted to additional assumptions which make the max-margin predictor coincides with the least-squares 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 least-squares solutions do not coincide: We show that at least for distributions satisfying Assumption 1, the first k coordinates of $\hat{\mathbf{w}}_d$ asymptotically minimize the expected *squared hinge loss*, $\ell(y\mathbf{x}^\top \mathbf{w}) = [1 - y\mathbf{x}^\top \mathbf{w}]_+^2 = (\max\{0, 1 - y\mathbf{x}^\top \mathbf{w}\})^2$. This is formalized in the following theorem:

Theorem 10 *Consider any sequence of distributions $\{\mathcal{D}_d\}_{d=k+1}^\infty$ satisfying Assumptions 1 and 2, and where $y \in \{-1, +1\}$. Let $\hat{\mathbf{w}}_d = \arg\min_{\mathbf{w} \in \mathbb{R}^d} \|\mathbf{w}\| : \min_{i \in [m_d]} y_i \mathbf{x}_i^\top \mathbf{w} \geq 1$ (where $\{\mathbf{x}_i, y_i\}_{i=1}^{m_d}$ are i.i.d. from $\mathcal{D}_d$). Furthermore, suppose that for some large enough d_0, $\sup_{d \geq d_0} \|\hat{\mathbf{v}}_d\|$*

is almost surely bounded, where $\hat{\mathbf{v}}_d$ *is the minimum-norm minimizer*[1] *of the function* $g_d(\mathbf{v}) := \frac{1}{m_d}\sum_{i=1}^{m_d}[1 - y_i\mathbf{x}_{i|k}^\top \mathbf{v}]_+^2$ *on* $\mathbb{R}^k$*. Then letting*

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

it holds that

$$g(\hat{\mathbf{w}}_{d|k}) \xrightarrow{a.s.} \inf_{\mathbf{v}\in\mathbb{R}^k} g(\mathbf{v}) \;\; \textit{and} \;\; \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_d}\left[(\mathbf{x}^\top \hat{\mathbf{w}}_d - \mathbf{x}_{|k}^\top \hat{\mathbf{w}}_{d|k})^2\right] \xrightarrow{a.s.} 0 \;.$$

The fact that the characterization is in terms of the squared hinge loss might be surprising at first, as this loss does not appear explicitly in the definition of the max-margin predictor (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 least-squares predictors coincide.

Before continuing, let us informally explain how the squared hinge loss arises in our analysis. Given a dataset $\{(\mathbf{x}_i, y_i)\}_{i=1}^{m_d}$, recall that $\mathbf{x}_{i|k}$ refers to the first k coordinates of $\mathbf{x}_i$, and $\mathbf{x}_{i|d-k}$ refers to the last $d-k$ coordinates (which have a high-dimensional isotropic Gaussian distribution). To simplify matters, let us suppose that $\{\mathbf{x}_{i|d-k}\}_{i=1}^{m_d}$ are precisely unit norm and orthogonal. In that case, the max-margin predictor on the training set can be equivalently written as $\arg\min_{\mathbf{w}\in\mathbb{R}^d} \|\mathbf{w}_{|k}\|^2 + \|\mathbf{w}_{|d-k}\|^2 : \forall i,\; y_i\mathbf{x}_{i|k}^\top \mathbf{w}_{|k} + y_i\mathbf{x}_{i|d-k}^\top \mathbf{w}_{|d-k} \geq 1$. 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 $\forall i,\; y_i\mathbf{x}_{i|d-k}^\top \mathbf{w}_{|d-k} \geq 1 - y_i\mathbf{x}_{i|k}^\top \mathbf{w}_{|k}$. Since $\{y_i\mathbf{x}_{i|d-k}\}_{i=1}^{m_d}$ 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, it follows that the optimal $\hat{\mathbf{w}}_{|d-k}$ is such that $\|\mathbf{w}_{|d-k}\|^2 = \|\sum_{i=1}^{m_d} y_i\mathbf{x}_{i|d-k}^\top \mathbf{w}_{|d-k}\|^2$, and that this equals $\sum_{i=1}^{m_d}(y_i\mathbf{x}_{i|d-k}^\top \mathbf{w}_{|d-k})^2 = \sum_{i=1}^{m_d}[1 - y_i\mathbf{x}_{i|k}^\top \mathbf{w}_{|k}]_+^2$. Plugging into the above, we get $\arg\min_{\mathbf{w}_{|k}\in\mathbb{R}^k} \|\mathbf{w}_{|k}\|^2 + \sum_{i=1}^{m_d}[1 - y_i\mathbf{x}_{i|k}^\top \mathbf{w}_{|k}]_+^2 \;=\; \arg\min_{\mathbf{w}_{|k}\in\mathbb{R}^k} \frac{\|\mathbf{w}_{|k}\|^2}{m_d} + \frac{1}{m_d}\sum_{i=1}^{m_d}[1 - y_i\mathbf{x}_{i|k}^\top \mathbf{w}_{|k}]_+^2$. As $d \to \infty$, we have $m_d \to \infty$. Hence, the first term becomes negligible, and the second term converges pointwise to $\mathbb{E}[[1 - y\mathbf{x}_{|k}^\top \mathbf{w}_{|k}]_+^2]$, namely the expected squared hinge loss w.r.t. the first k coordinates. The formal proof of Thm. 10 is more complicated, but follows a similar idea.

Overall, we see that benign overfitting in our binary classification model boils down to the optimality (w.r.t. expected misclassification error) of the minimizer of the expected squared hinge loss on the first k coordinates. Since the squared hinge loss is not the same as misclassification error, we cannot hope this to always hold: Indeed, we prove in Appendix B that there exist distributions where minimizing the squared hinge loss can lead to trivial misclassification error, and therefore benign overfitting will not occur in these cases. However, for general distributions, it is not unreasonable to assume that minimizing the squared hinge loss *will* lead to low misclassification error. This is because predictors that attempt to minimize the squared hinge loss will also tend to make

1. A minimizer always exists, since $g_d(\mathbf{v})$ 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 $g_d(\cdot)$, and with a smaller norm which is a contradiction.

$y\mathbf{x}_{|k}^\top \mathbf{w}_{|k}$ 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. 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. Since we are now only concerned with the behavior on the first k coordinates, we will focus from now on solely on examples $(\mathbf{x}, y)$ coming from some fixed distribution, where $\mathbf{x} \in \mathbb{R}^k$.

Recall that for benign overfitting, 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 *linearly separable* distribution $\mathcal{D}_{\text{clean}}$ (with some unit vector $\mathbf{w}^*$ such that $\Pr(y\mathbf{x}^\top \mathbf{w}^* < \gamma) = 0$ 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 $(\mathbf{x}, y)$ is still distributed as $\mathcal{D}_{\text{clean}}$, and that we wish to find a predictor $\mathbf{w}$ satisfying $\Pr_{(\mathbf{x},y)\sim\mathcal{D}_{\text{clean}}}(y\mathbf{x}^\top \mathbf{w} \leq 0) = 0$. However, the predictor we learn is with respect to the "noisy" labels, where the expected squared hinge loss can be written as

$$\begin{aligned} L_p(\mathbf{w}) &:= \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_{\text{clean}}}\left[(1-p)\cdot[1-y\mathbf{x}^\top\mathbf{w}]_+^2 + p\cdot[1+y\mathbf{x}^\top\mathbf{w}]_+^2\right] \\ &= \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_{\text{clean}}}\mathbb{E}[\ell_p(y\mathbf{x}^\top\mathbf{w})] \quad \text{where} \quad \ell_p(z) := (1-p)\cdot[1-z]_+^2 + p\cdot[1+z]_+^2 \,. \end{aligned} \tag{2}$$

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 $\mathbb{E}[\mathbf{x}\mathbf{x}^\top]$ is positive definite (see Lemma 17 in the appendix for a formal definition of strong convexity and a proof). Therefore, $L_p(\cdot)$ has a unique minimizer $\mathbf{w}_p^*$. In that case, Thm. 10 implies that $\hat{\mathbf{w}}_{d|k}$ converges parameterically to $\mathbf{w}_p^*$. Thus, for benign overfitting, it is sufficient that $\mathbf{w}_p^*$ achieves zero error with respect to the "clean" labels. This is formalized in the following theorem:

Theorem 11 *Under the conditions of Thm. 10, let $R_d(\mathbf{w}) = \Pr_{(\mathbf{x},y)\sim\mathcal{D}_d}(y\mathbf{x}^\top\mathbf{w} \leq 0)$. Then the benign overfitting property specified in Eq. (1) holds under the following condition: The first k coordinates of $\mathcal{D}_d$ corresponds to some linearly separable distribution $\mathcal{D}_{clean}$ with labels flipped with some probability $p \in (0, \frac{1}{2})$, and the minimizer $\mathbf{w}_p^*$ of $L_p(\mathbf{w})$ satisfies $\Pr_{(\mathbf{x},y)\sim\mathcal{D}_{clean}}(y\mathbf{x}^\top\mathbf{w}_p^* \leq 0) = 0$.*

Focusing on such linearly-separable-with-label-noise distributions, we now turn to study some cases where the condition on $\mathbf{w}_p^*$ in Thm. 11 indeed holds. For example, the following theorem implies that under mild assumptions, just about *any* choice of distribution on the first k coordinates 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 12 *Fix any distribution $\mathcal{D}_{clean}$ on $(\mathbf{x}, y) \in \mathbb{R}^k \times \{-1, +1\}$, such that there exists some $\mathbf{w}^* \in \mathbb{R}^k$ so that $\mathbb{E}[[1-y\mathbf{x}^\top\mathbf{w}^*]_+^2] = 0$. Suppose $\mathbb{E}[\mathbf{x}\mathbf{x}^\top]$ is positive definite and $\mathbf{x}$ 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 $\Pr_{(\mathbf{x},y)\sim\mathcal{D}_{clean}}(y\mathbf{x}^\top\mathbf{w}_p^* \leq 0) = 0$.*

We note that the assumption on $\mathbf{w}^*$ is equivalent to having linear separability w.r.t. $\mathcal{D}_{\text{clean}}$ with margin $\frac{1}{\|\mathbf{w}^*\|}$ (the unit vector $\mathbf{w} = \frac{\mathbf{w}^*}{\|\mathbf{w}\|_*}$ satisfies $\Pr(y\mathbf{x}^\top\mathbf{w} < \frac{1}{\|\mathbf{w}^*\|}) = 0$).

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 13 *Fix any distribution $\mathcal{D}_{clean}$ on $(\mathbf{x}, y) \in \mathbb{R}^k \times \{-1, +1\}$, such that there exists some $\mathbf{w}^* \in \mathbb{R}^k$ so that $\mathbb{E}[[1 - y\mathbf{x}^\top\mathbf{w}^*]_+^2] = 0$. Suppose that for some unit vector $\mathbf{u}$, and conditioned on 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 $\Pr_{\mathcal{D}_{clean}}(y\mathbf{x}^\top\mathbf{w}_p^* \leq 0) = 0$.*

The conditions in the theorem refer to a situation where there is some distinguished direction $\mathbf{u}$, such that conditioned on y, the component of the input distribution orthogonal to $\mathbf{u}$ is independent of the distribution along $\mathbf{u}$, and symmetric around the origin. Examples where this occurs include any one-dimensional distribution, and a mixture of any two symmetric distributions with means in $\text{span}(\mathbf{u})$ (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.

5. Discussion

In this paper, we proposed and studied a new perspective on the phenomenon of benign overfitting. We argue that when interpolating the training data, the learned predictor simultaneously optimizes the average loss function of many different types of problems, and the existence of benign overfitting on one problem precludes its existence on another. On the negative side, we argue that this makes benign overfitting difficult to establish for regression settings beyond the well-specified linear ones studied so far. On the positive side, for classification problems, we identify an implicit bias of the learned max-margin predictor to minimize the expected squared hinge loss with respect to the underlying distribution (at least in a simple model where an arbitrary low-dimensional distribution is concatenated with a high-dimensional one). We use it to show benign overfitting in various settings, by considering cases where the 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 square 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 large number of small "unimportant" Gaussian coordinates, a setting which is prototypical for benign overfitting. Nevertheless, since our results did not crucially rely on any special properties of the Gaussian distribution, we believe our insights should be extendable to more general input distributions, and identifying them can be an interesting direction for future research.

Acknowledgements

This research is supported in part by European Research Council (ERC) grant 754705. We thank Gilad Yehudai and the anonymous COLT reviewers for some very helpful comments.

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.

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.

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.

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.

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

Appendix A. Proofs

A.1. Proof of Thm. 1

Define

$$\mathcal{W} = \{\mathbf{w} : L(\mathbf{w}) = 0, \mathbf{w} \in \text{span}\{\mathbf{x}_1, \ldots, \mathbf{x}_m\}\} \,.$$

By definition, we have $\hat{\mathbf{w}} \in \mathcal{W}$.

Next, we argue that if $\mathbf{w}^*$ is any minimum norm solution satisfying $L(\mathbf{w}^*) = 0$ (a non-empty set since by assumption, $L(\hat{\mathbf{w}}) = 0$ and $L(\cdot)$ is continuous), then $\mathbf{w}^* \in \text{span}\{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$, so $\mathbf{w}^* \in \mathcal{W}$ as well. The reason is that by definition of L, $L(\mathbf{w}^*)$ depends on $\mathbf{w}^*$ only via the values of $\mathbf{x}_1^\top \mathbf{w}^*, \ldots, \mathbf{x}_m^\top \mathbf{w}^*$. Thus, if $\mathbf{w}^*$ is not in $\text{span}\{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$, we could have found a smaller-norm solution $\mathbf{w}$ such that $L(\mathbf{w}) = 0$ (which would be a contradiction), by projecting $\mathbf{w}^*$ on this linear span.

Now, we argue that all $\mathbf{w} \in \mathcal{W}$ have the same norm, so in particular, $\|\hat{\mathbf{w}}\| = \|\mathbf{w}^*\|$, which implies that

$$\hat{\mathbf{w}} \in \{\arg\min_{\mathbf{w}} \|\mathbf{w}\| \quad : \quad L(\mathbf{w}) = 0\} \,. \tag{3}$$

To see this, fix any $\mathbf{w} \in \mathcal{W}$, and let X be a matrix whose i-th column is $\mathbf{x}_i$. By definition of $\mathcal{W}$, there exists a vector $\boldsymbol{\alpha}$ such that $\mathbf{w} = X\boldsymbol{\alpha}$, and moreover, $X^\top \mathbf{w} = \mathbf{z}$, where $\mathbf{z}$ is the vector such that $z_i = \ell_{y_i}^{-1}(0)$. Combining, we get that

$$X^\top X \boldsymbol{\alpha} = X^\top \mathbf{w} = \mathbf{z} \,,$$

which implies that

$$\boldsymbol{\alpha} \in \mathcal{A} := \{\mathbf{u} + \mathbf{v} : (X^\top X)\mathbf{v} = \mathbf{0}\},$$

with $\mathbf{u}$ being some fixed vector (independent of $\boldsymbol{\alpha}$) satisfying $X^\top X \mathbf{u} = \mathbf{z}$ (we note that such a $\mathbf{u}$ must exist, since by the fact that $\mathbf{w}^* \in \mathcal{W}$, it follows that $\mathbf{w}^* = X\boldsymbol{\alpha}^*$ for some $\boldsymbol{\alpha}^*$, hence $X^\top X \boldsymbol{\alpha}^* = X^\top \mathbf{w}^* = \mathbf{z}$). Note that for any $\boldsymbol{\alpha} \in \mathcal{A}$, $X\boldsymbol{\alpha}$ has the same norm:

$$\|X\boldsymbol{\alpha}\|^2 = \boldsymbol{\alpha}^\top X^\top X \boldsymbol{\alpha} = \mathbf{u}^\top (X^\top X)(X^\top X)\mathbf{u} \,,$$

and since $\mathbf{w} = X\boldsymbol{\alpha}$, it follows that any $\mathbf{w} \in \mathcal{W}$ has the same norm, from which Eq. (3) follows.

Finally, we need to prove that $\hat{\mathbf{w}}$ is the *unique* minimal-norm point achieving $L(\hat{\mathbf{w}}) = 0$. By definition, $\mathbf{w}^*$ is such a minimal-norm point. Assume by contradiction that $\hat{\mathbf{w}} \neq \mathbf{w}^*$, and consider their average, $\mathbf{w}_0 := \frac{1}{2}(\hat{\mathbf{w}} + \mathbf{w}^*)$. Since $X\mathbf{w}^* = X\hat{\mathbf{w}} = X\mathbf{w}_0$, it follows that $L(\mathbf{w}_0) = L(\hat{\mathbf{w}}) = L(\mathbf{w}^*) = 0$, yet

$$\|\mathbf{w}_0\|^2 \;=\; \frac{\|\hat{\mathbf{w}}\|^2 + \|\mathbf{w}^*\|^2 + 2\hat{\mathbf{w}}^\top \mathbf{w}^*}{4} \;<\; \frac{\|\hat{\mathbf{w}}\|^2 + \|\mathbf{w}^*\|^2 + 2\|\hat{\mathbf{w}}\| \cdot \|\mathbf{w}^*\|}{4} \;=\; \frac{4\|\mathbf{w}^*\|^2}{4} \;=\; \|\mathbf{w}^*\|^2$$

(since we showed $\|\hat{\mathbf{w}}\| = \|\mathbf{w}^*\|$, and $\hat{\mathbf{w}}^\top \mathbf{w}^* \leq \|\hat{\mathbf{w}}\| \cdot \|\mathbf{w}^*\|$ with equality only when the vectors are equal). Overall, we get that $L(\mathbf{w}_0) = 0$ and $\|\mathbf{w}_0\| < \|\mathbf{w}^*\|$, contradicting the definition of $\mathbf{w}^*$.

A.2. Proof of Thm. 5

Fix some d, and let $X \in \mathbb{R}^{m_d \times d}$ be the matrix whose rows are $\mathbf{x}_1, \ldots, \mathbf{x}_{m_d}$. Let $X_{|k} \in \mathbb{R}^{m_d \times k}$ denote its first k columns, and $X_{|d-k}$ the rest of the columns. Since $m_d \leq d - k$, the rows of X are almost surely linearly independent (since the m_d rows of $X_{|d-k}$ are i.i.d. Gaussian and hence almost surely independent). Thus, $XX^\top$ is invertible, and $\hat{\mathbf{w}}_d = \arg\min \|\mathbf{w}\| : X\mathbf{w} = \mathbf{y}$ can be written in closed form as

$$\hat{\mathbf{w}}_d \;=\; X^\top (XX^\top)^{-1}\mathbf{y} \;=\; X^\top \left(X_{|k}X_{|k}^\top + X_{|d-k}X_{|d-k}^\top\right)^{-1} \mathbf{y}\,. \tag{4}$$

Note that $X_{|d-k}$ is an $m_d \times (d-k)$ matrix composed of i.i.d. Gaussian entries of variance $\frac{1}{d-k}$. By a tail bound for Wishart matrices (see for example Zhu (2012)), it is easily verified that with probability at least $1 - \frac{1}{d^2}$,

$$\|X_{|d-k}X_{|d-k}^\top - I\| \;\leq\; \mathcal{O}\left(\sqrt{\frac{m_d \log(d)}{d-k}}\right) \;=\; \mathcal{O}\left(\sqrt{\frac{m_d \log(d)}{d}}\right)\,,$$

where $I = I_{d-k}$ is the identity matrix. Thus, we can write

$$X_{|k}X_{|k}^\top + X_{|d-k}X_{|d-k}^\top = A + E \text{ where } A = I + X_{|k}X_{|k}^\top\,,\; \Pr\left(\|E\| > \mathcal{O}\left(\sqrt{\frac{m_d \log(d)}{d}}\right)\right) \leq \frac{1}{d^2}\,. \tag{5}$$

Note that A is a positive definite matrix with minimal eigenvalue ≥ 1, hence A^{-1} exists and $\|A^{-1}\| \leq 1$ regardless of d.

By taking a union bound over all sufficiently large d, and using the assumption that $\frac{m_d^3 \log(d)}{d} \to 0$, it follows that $m_d \cdot E$ (and definitely E) almost surely converges to 0. Therefore, for any large enough d, $I + A^{-1}E$ is invertible with arbitrarily high probability, and by the Woodbury matrix identity we get that

$$\left(X_{|k}X_{|k}^\top + X_{|d-k}X_{|d-k}^\top\right)^{-1} \;=\; (A+E)^{-1} \;=\; A^{-1} - A^{-1}E(I + A^{-1}E)^{-1}A^{-1}\,.$$

Thus,

$$\|(A+E)^{-1} - A^{-1}\| \;\leq\; \|A^{-1}\|^2 \cdot \|E\| \cdot \|(I + A^{-1}E)^{-1}\| \;\leq\; \frac{\|A^{-1}\|^2 \cdot \|E\|}{1 - \|A^{-1}\| \cdot \|E\|}\,, \tag{6}$$

which even multiplied by m_d, converges with d almost surely to 0 . Again by the Woodbury matrix identity,

$$A^{-1} \;=\; (I + X_{|k}X_{|k}^\top)^{-1} \;=\; I - X_{|k}(I + X_{|k}^\top X_{|k})^{-1}X_{|k}^\top \,.$$

Collecting the top two displayed equations, it follows that

$$(X_{|k}X_{|k}^\top + X_{|d-k}X_{|d-k}^\top)^{-1} \;=\; (A+E)^{-1} \;=\; I - X_{|k}(I + X_{|k}^\top X_{|k})^{-1}X_{|k}^\top + E',$$

where E' is some matrix such that $m_d \cdot E'$ converges almost surely to 0. Plugging this back into Eq. (4) and focusing on the first k coordinates of $\hat{\mathbf{w}}_d$, it follows that

$$\begin{aligned}
\hat{\mathbf{w}}_{d|k} \;&=\; X_{|k}^\top \left(X_{|k}X_{|k}^\top + X_{|d-k}X_{|d-k}^\top\right)^{-1}\mathbf{y} \\
&=\; \left(I - X_{|k}^\top X_{|k}(I + X_{|k}^\top X_{|k})^{-1}\right)X_{|k}^\top\mathbf{y} + X_{|k}^\top E'\mathbf{y} \\
&=\; \left((I + X_{|k}^\top X_{|k}) - X_{|k}^\top X_{|k}\right)(I + X_{|k}^\top X_{|k})^{-1}X_{|k}^\top\mathbf{y} + X_{|k}^\top E'\mathbf{y} \\
&=\; (I + X_{|k}^\top X_{|k})^{-1}X_{|k}^\top\mathbf{y} + X_{|k}^\top E'\mathbf{y} \\
&=\; \left(\frac{1}{m_d}I + \frac{1}{m_d}X_{|k}^\top X_{|k}\right)^{-1}\left(\frac{1}{m_d}X_{|k}^\top\mathbf{y}\right) + X_{|k}^\top E'\mathbf{y}\,. \qquad (7)
\end{aligned}$$

Let us now understand how this expression behaves as $d \to \infty$. First, since $m_d \to \infty$, we have $\frac{1}{m_d}I \to 0$. Second, we have $\frac{1}{m_d}X_{|k}^\top X_{|k} = \frac{1}{m_d}\sum_{i=1}^{m_d}\mathbf{x}_{i|k}\mathbf{x}_{i|k}^\top$, a $k \times k$ matrix, which by Assumption 1 and the law of large numbers[2] converges almost surely to the (positive definite) matrix $\mathbb{E}[\mathbf{x}_{|k}\mathbf{x}_{|k}^\top]$. Third, we have $\frac{1}{m_d}X_{|k}^\top\mathbf{y} = \frac{1}{m_d}\sum_{i=1}^{m_d} y_i\mathbf{x}_{i|k}$, which again by by Assumption 1 and the law of large numbers converges almost surely to the vector $\mathbb{E}[y\mathbf{x}_{|k}]$. Fourth, we have

$$\|X_{|k}^\top E'\mathbf{y}\| \le \|X_{|k}\| \cdot \|\mathbf{y}\| \cdot \|E'\| \;\le\; m_d \cdot \sqrt{\frac{1}{m_d}\|X_{|k}\|_F^2} \cdot \sqrt{\frac{1}{m_d}\|\mathbf{y}\|^2} \cdot \|E'\| \,.$$

Again by the law of large numbers, both $\frac{1}{m_d}\|X_{|k}\|^2 = \frac{1}{m_d}\sum_{i=1}^{m_d}\|\mathbf{x}_{i|k}\|^2$ and $\frac{1}{m_d}\|\mathbf{y}\|^2 = \frac{1}{m_d}\sum_{i=1}^{m_d} y_i^2$ converge almost surely to their (finite) expectations, hence they are almost surely bounded, and since $m_d \cdot E'$ converges almost surely to 0, it follows that the bound in the displayed equation converges almost surely to 0. Plugging these observations back into Eq. (7), we get that $\hat{\mathbf{w}}_{d|k}$ converges almost surely to $\mathbb{E}[\mathbf{x}_{|k}\mathbf{x}_{|k}^\top]^{-1} \cdot \mathbb{E}[y\mathbf{x}_{|k}]$ as stated in the theorem.

We will now show the second assertion in the theorem, which is equivalent to proving

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

By Eq. (4), Eq. (5) and Eq. (6),

$$\|\hat{\mathbf{w}}_{d|d-k}\| \;=\; \left\|X_{|d-k}^\top\left(X_{|k}X_{|k}^\top + X_{|d-k}X_{|d-k}^\top\right)^{-1}\mathbf{y}\right\| \;=\; \left\|X_{|d-k}^\top\left((I + X_{|k}X_{|k}^\top)^{-1} + E'\right)\mathbf{y}\right\| \,,$$

2. Throughout the proofs, we use the following standard version of the strong law of large numbers: If $\{r_d\}_{d=1}^\infty$ are a sequence of i.i.d. random variables on $\mathbb{R}$ such that $\mathbb{E}[|r_1|] < \infty$, then $\frac{1}{d}\sum_{i=1}^d r_i$ converges almost surely to $\mathbb{E}[r_1]$. This trivially extends to vector-valued random variables in $\mathbb{R}^k$ where k is fixed.

where E' is some matrix which converges almost surely to 0. The expression above is at most

$$\|X_{|d-k}\| \cdot \left(\|(I + X_{|k} X_{|k}^\top)^{-1}\| + \|E'\| \right) \cdot \|\mathbf{y}\| \;\leq\; \|X_{|d-k}\| \cdot \|\mathbf{y}\| \cdot (1 + \|E'\|) \,,$$

where we used the fact that $I + X_{|k} X_{|k}^\top$ is a positive definite matrix with all eigenvalues being at least 1. Overall, this implies that

$$\frac{\|\hat{\mathbf{w}}_{d|d-k}\|}{m_d} \;\leq\; \left(\sqrt{\frac{1}{m_d} \|X_{|d-k}\|^2} \right) \left(\sqrt{\frac{1}{m_d} \|\mathbf{y}\|^2} \right) (1 + \|E'\|) \,.$$

By Assumption 1 and the law of large numbers, both $\frac{1}{m_d}\|\mathbf{y}\|^2 = \frac{1}{m_d}\sum_{i=1}^{m_d} y_i^2$ and $\frac{1}{m_d}\|X_{|d-k}\|^2 \leq \frac{1}{m_d}\|X_{|d-k}\|_F^2 = \frac{1}{m_d}\sum_{i=1}^{m_d} \|\mathbf{x}_{i|d-k}\|^2$ converge almost surely to their (finite) expectations, and we already stated that $\|E'\|$ converges almost surely to 0. Thus, almost surely, $\limsup_{d\to\infty} \frac{\|\hat{\mathbf{w}}_{d|d-k}\|}{m_d}$ has some finite value. Hence, if $\mathbf{z} \in \mathbb{R}^{d-k}$ is zero-mean Gaussian with covariance matrix $\frac{1}{d-k} I$ (that is, the same distribution as $\mathbf{x}_{|d-k}$), we have

$$\mathbb{E}_{\mathbf{z}}[(\mathbf{z}^\top \hat{\mathbf{w}}_{d|d-k})^2] \;=\; \frac{1}{d-k} \cdot \|\hat{\mathbf{w}}_{d|d-k}\|^2 \;=\; \frac{m_d^2}{d-k} \cdot \left(\frac{\|\hat{\mathbf{w}}_{d|d-k}\|}{m_d} \right)^2 \,,$$

and since $\frac{m_d^2}{d-k} \to 0$ with d, it follows that $\mathbb{E}_{\mathbf{z}}[(\mathbf{z}^\top \hat{\mathbf{w}}_{d|d-k})^2]$ converges almost surely to 0 as required.

A.3. Proof of Lemma 8

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 Thm. 10

We note that as in the proof of Thm. 5, our assumptions imply that $\hat{\mathbf{w}}_d$ exists almost surely for all $d > k$.

Since the proof is a bit lengthy, we split it into three subsections: First we state and prove an auxiliary lemma about the optimal value of a certain optimization problem. Then, we analyze the $\hat{\mathbf{w}}_{d|k}$ term, and finally, we analyze the $\hat{\mathbf{w}}_{d|d-k}$ term.

A.4.1. AN AUXILIARY LEMMA

Lemma 14 *Fix some integer $m \geq 1$, some $\mathbf{r} \in \mathbb{R}^m$, and an $m \times m$ symmetric matrix E such that $\|E\| < \frac{1}{2}$. Then the set $\{\boldsymbol{\alpha} \in \mathbb{R}^m : (I+E)\boldsymbol{\alpha} \succeq \mathbf{r}\}$ is not empty. Moreover, letting*

$$a^* \;=\; \min_{\boldsymbol{\alpha} \in \mathbb{R}^m} \|\boldsymbol{\alpha}\|^2 + \boldsymbol{\alpha}^\top E \boldsymbol{\alpha} \;\;:\;\; (I+E)\boldsymbol{\alpha} \succeq \mathbf{r}\,, \tag{8}$$

we have

$$\left|a^* - \|[\mathbf{r}]_+\|^2\right| \;\leq\; 2\|E\| \cdot \|[\mathbf{r}]_+\|^2\,,$$

where $[\mathbf{r}]_+$ denotes applying $[\cdot]_+$ element-wise.

To prove the lemma, we will first state and prove the following helper lemma:

Lemma 15 *For any symmetric matrix E such that $\|E\| \leq \frac{1}{2}$, it holds that $I+E$ is invertible and*

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

Proof By Weyl's inequality, $\lambda_{\min}(I+E) \geq \lambda_{\min}(I) - \|E\| \geq \frac{1}{2} > 0$, hence $I+E$ is invertible. Moreover, by the Woodbury matrix identity,

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

which implies

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

Noting that $\|E\| \leq \frac{1}{2}$, it follows that the above is at most $\frac{\|E\|}{1/2} = 2\|E\|$. ■

Proof [Proof of Lemma 14] The fact that $\{\boldsymbol{\alpha} \in \mathbb{R}^m : (I+E)\boldsymbol{\alpha} \succeq \mathbf{r}\}$ is not empty follows from the observation that $I+E$ is positive definite and hence invertible (since we assume $\|E\| < \frac{1}{2}$). Thus, the set contains for instance the vector $(I+E)^{-1}\mathbf{r}$. 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 $\boldsymbol{\beta} = (I+E)\boldsymbol{\alpha}$ (which is valid since $I+E$ is invertible), so $\boldsymbol{\alpha} = (I+E)^{-1}\boldsymbol{\beta}$. Noting that $a^* = \min_{\boldsymbol{\alpha}:(I+E)\boldsymbol{\alpha} \succeq \mathbf{r}} \boldsymbol{\alpha}^\top (I+E)\boldsymbol{\alpha}$, it follows that

$$a^* \;=\; \min_{\boldsymbol{\beta} \in \mathbb{R}^m : \boldsymbol{\beta} \succeq \mathbf{r}} \boldsymbol{\beta}^\top (I+E)^{-1} \boldsymbol{\beta}\,. \tag{9}$$

By the assumption $\|E\| \leq \frac{1}{2}$ and Lemma 15, it follows that

$$\|(I+E)^{-1} - I\| \;\leq\; 2\|E\|\,.$$

This implies that

$$(1+2\|E\|)I \;\succeq\; (I+E)^{-1} \;\succeq\; (1-2\|E\|)I,$$

where $A \succeq B$ for symmetric matrices A, B implies that $A - B$ is positive semidefinite. Plugging this back into Eq. (9), it follows that

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^m : \boldsymbol{\beta} \succeq \mathbf{r}} (1+2\|E\|) \cdot \|\boldsymbol{\beta}\|^2 \;\geq\; a^* \;\geq\; \min_{\boldsymbol{\beta} \in \mathbb{R}^m : \boldsymbol{\beta} \succeq \mathbf{r}} (1-2\|E\|) \cdot \|\boldsymbol{\beta}\|^2\,.$$

Now, it is easily verified that $\min_{\boldsymbol{\beta}:\boldsymbol{\beta}\succeq\mathbf{r}} \|\boldsymbol{\beta}\|^2 = \sum_{i=1}^m [r_i]_+^2$. Plugging this in the previous displayed equation, it follows that

$$(1+2\|E\|)\sum_{i=1}^m [r_i]_+^2 \;\geq\; a^* \;\geq\; (1-2\|E\|)\sum_{i=1}^m [r_i]_+^2 \,.$$

Therefore,

$$\left| a^* - \sum_{i=1}^m [r_i]_+^2 \right| \;\leq\; \sum_{i=1}^m [r_i]_+^2 \cdot 2\|E\| \,,$$

from which the bound in the lemma follows. ■

A.4.2. ANALYSIS OF $\hat{\mathbf{w}}_{d|k}$

$\hat{\mathbf{w}}_d$ can be equivalently written as

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

Let us write the first k coordinates of $\mathbf{w}$ as $\mathbf{v}$, the last $d-k$ coordinates of $\mathbf{w}$ as $\mathbf{u}$, and the last $d-k$ coordinates of $\mathbf{x}_i$ as $\mathbf{z}_i$. Then the above can be equivalently written as

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

or again equivalently as

$$\arg\min_{\mathbf{v}\in\mathbb{R}^k} \frac{\|\mathbf{v}\|^2}{m_d} + f_{m_d}(\mathbf{v}) \tag{10}$$

where

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

(we note that since $d-k > m_d$, and the coordinates of $\mathbf{z}_i$ are i.i.d. Gaussian, then almost surely, the set $\{y_i\mathbf{z}_i\}_{i=1}^{m_d} \subset \mathbb{R}^{d-k}$ is linearly independent, and the constraints in the displayed equation above are feasible).

Our goal will now be to argue that as d, m_d go to infinity, $f_{m_d}(\mathbf{v})$ converges to a simple closed-form expression independent of $\{(\mathbf{x}_i, y_i)\}_{i=1}^{m_d}$. To do so, let Z, X be $m_d \times (d-k)$ matrices whose i-th rows are $y_i\mathbf{z}_i^\top$ and $y_i\mathbf{x}_{i|k}^\top$ respectively, for $i\in[m_d]$. Also, let $\mathbf{1}$ be the all-ones vector in $\mathbb{R}^{m_d}$. Thus, we can write

$$f_{m_d}(\mathbf{v}) \;=\; \min_{\mathbf{u}\in\mathbb{R}^{d-k}} \frac{\|\mathbf{u}\|^2}{m_d} \;\;:\;\; Z\mathbf{u}\succeq\mathbf{1}-X\mathbf{v} \,.$$

Clearly, the optimal $\mathbf{u}$ must lie in the row span of Z (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 $Z^\top\boldsymbol{\alpha}$ for some $\boldsymbol{\alpha}\in\mathbb{R}^{m_d}$, so we can rewrite the displayed equation above as

$$f_{m_d}(\mathbf{v}) \;=\; \min_{\boldsymbol{\alpha}\in\mathbb{R}^{m_d}} \frac{1}{m_d}\boldsymbol{\alpha}^\top\left(ZZ^\top\right)\boldsymbol{\alpha} \;\;:\;\; ZZ^\top\boldsymbol{\alpha}\succeq\mathbf{1}-X\mathbf{v} \,.$$

Letting

$$E = ZZ^\top - I\,,$$

we can write the above as

$$f_{m_d}(\mathbf{v}) \;=\; \min_{\boldsymbol{\alpha}\in\mathbb{R}^{m_d}} \frac{1}{m_d}\left(\|\boldsymbol{\alpha}\|^2 + \boldsymbol{\alpha}^\top E\boldsymbol{\alpha}\right) \;\;:\;\; (I+E)\boldsymbol{\alpha} \succeq \mathbf{1} - X\mathbf{v}\,. \tag{11}$$

Similar to the proof of Thm. 5, we can use a tail bound (e.g., Zhu (2012)) for Wishart matrices to argue that with probability at least $1-\frac{1}{d^2}$,

$$\|E\| \;\leq\; \mathcal{O}\left(\sqrt{\frac{m_d\log(d)}{d}}\right)\,. \tag{12}$$

Let us now sketch the rest of the proof for analyzing $\hat{\mathbf{w}}_{d|k}$, followed by a more rigorous analysis. Informally, as d increases and E goes to zero, Eq. (11) "converges" to the optimization problem

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^{m_d}} \frac{\|\boldsymbol{\alpha}\|^2}{m_d} \;\;:\;\; \boldsymbol{\alpha} \succeq \mathbf{1} - X\mathbf{v}\,,$$

whose solution is $\boldsymbol{\alpha} = [1-X\mathbf{v}]_+$ (with $[\cdot]_+$ applied element-wise), which leads to an optimal objective value of $\frac{1}{m_d}[1-X\mathbf{v}]_+^2 = \frac{1}{m_d}\sum_{i=1}^{m_d}[1-y_i\mathbf{x}_{i|k}^\top\mathbf{v}]_+^2$. Since $y_i, \mathbf{x}_{i|k}$ are i.i.d., this in turn converges to $\mathbb{E}[(1-y\mathbf{x}_{|k}^\top\mathbf{v})]$ almost surely. Plugging this back into Eq. (10), and noting that $\frac{1}{m_d}\|\mathbf{v}\|^2 \to 0$ for any fixed $\mathbf{v}$, we get that asymptotically we are looking for an optimum of $\min_{\mathbf{v}\in\mathbb{R}^k}\mathbb{E}[1-y\mathbf{x}_{|k}^\top\mathbf{v}]_+^2]$, hence $\hat{\mathbf{w}}_{d|k}$ is asymptotically a minimizer of this function as stated in the theorem.

To formally justify this informal argument, we apply Lemma 14 on Eq. (11) with $\mathbf{r} = 1 - X\mathbf{v}$, $m = m_d$, recalling that $\|E\| = \mathcal{O}\left(\sqrt{\frac{m_d\log(d)}{d}}\right)$ (which is less than $\frac{1}{2}$ for any large enough d) with probability at least $1-\frac{1}{d^2}$. Thus, we get that with probability at least $1-\frac{1}{d^2}$, it holds simultaneously for any $\mathbf{v}$ that

$$\left|f_{m_d}(\mathbf{v}) - \frac{1}{m_d}\|[\mathbf{1}-X\mathbf{v}]_+\|^2\right| \;\leq\; \frac{2\|E\|}{m_d}\cdot\|[\mathbf{1}-X\mathbf{v}]_+\|^2 \;\leq\; \mathcal{O}\left(\sqrt{\frac{m_d\log(d)}{d}}\right)\cdot\frac{1}{m_d}\|[\mathbf{1}-X\mathbf{v}]_+\|^2\,. \tag{13}$$

Plugging this back into Eq. (10), and plugging in $\frac{1}{m_d}\|[\mathbf{1}-X\mathbf{v}]_+\|^2 = \frac{1}{m_d}\sum_{i=1}^{m_d}[1-y_i\mathbf{x}_{i|k}^\top\mathbf{v}]_+^2$, we get that with probability at least $1-\frac{1}{d^2}$,

$$\begin{aligned}\hat{\mathbf{w}}_{d|k} \;&=\; \arg\min_{\mathbf{v}\in\mathbb{R}^k} \frac{\|\mathbf{v}\|^2}{m_d} + (1+\epsilon_{d,\mathbf{v}})\cdot\frac{1}{m_d}\sum_{i=1}^{m_d}[1-y_i\mathbf{x}_{i|k}^\top\mathbf{v}]_+^2 \\ &=\; \arg\min_{\mathbf{v}\in\mathbb{R}^k} (1+\epsilon_{d,\mathbf{v}})\cdot g_d(\mathbf{v}) + \frac{\|\mathbf{v}\|^2}{m_d}\,,\end{aligned} \tag{14}$$

where $g_d(\mathbf{v}) := \frac{1}{m_d}\sum_{i=1}^{m_d}[1-y_i\mathbf{x}_{i|k}^\top\mathbf{v}]_+^2$ as defined in the theorem statement, and $\epsilon_{d,\mathbf{v}}$ satisfies

$$\sup_{\mathbf{v}\in\mathbb{R}^k}|\epsilon_{d,\mathbf{v}}| \;\leq\; \epsilon_d := \mathcal{O}\left(\sqrt{\frac{m_d\log(d)}{d}}\right)\,.$$

Let us now consider the sequence $\{\hat{\mathbf{w}}_{d|k}\}_{d=d_0}^{\infty}$, where we pick d_0 sufficiently large so that Eq. (14) and the bound on $\epsilon_{d,\mathbf{v}}$ holds for any $d \geq d_0$ with probability at least $1 - \frac{1}{d^2}$. By a union bound, it follows that these hold simultaneously for all $\{\hat{\mathbf{w}}_{d|k}\}_{d=d_0}^{\infty}$, with probability at least $1 - o_{d_0}(1)$, where $o_{d_0}(1)$ signifies an expression that goes to 0 with d_0. Assuming this event holds, let us compare this sequence to the sequence $\{\hat{\mathbf{v}}_d\}_{d=d_0}^{\infty}$ of random vectors (defined over the same probability space), where $\hat{\mathbf{v}}_d$ is the minimum-norm minimizer of $g_d(\mathbf{v})$. We want to argue that for any d, if Eq. (14) holds, then $\|\hat{\mathbf{w}}_{d|k}\|$ is not much larger than $\|\hat{\mathbf{v}}_d\|$. In the theorem, we assume that $\sup_d \|\hat{\mathbf{v}}_d\|$ is almost surely bounded, so the above would imply that $\sup_{d \geq d_0} \|\hat{\mathbf{w}}_{d|k}\|$ is also bounded by some finite number with probability at least $1 - o_{d_0}(1)$. To justify this, note that by Eq. (14), definition of $\hat{\mathbf{v}}_d$, and the fact that $\sup_{\mathbf{v}} |\epsilon_{d,\mathbf{v}}| \leq \epsilon_d$, we have with probability $1 - o_{d_0}(1)$ that for all $d \geq d_0$,

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

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

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

where the last transition follows from $\hat{\mathbf{v}}_d$ being a minimizer of $g_d(\cdot)$, hence $g_d(\hat{\mathbf{v}}_d) \leq g_d(\mathbf{0}) = 1$. Recalling that $m_d\epsilon_d = \mathcal{O}\left(\sqrt{\frac{m_d^3\log(d)}{d}}\right) \overset{d\to\infty}{\longrightarrow} 0$ by Assumption 2, it follows in particular that with probability at least $1-o_{d_0}(1)$, $\|\hat{\mathbf{w}}_{d|k}\|^2 \leq 1+\|\hat{\mathbf{v}}_d\|^2$ for all large enough d. Recalling that $\sup_d \|\hat{\mathbf{v}}_d\|$ is bounded almost surely, we overall get the following: For any d_0, there exists some finite B_{d_0} such than with probability at least $1 - o_{d_0}(1)$,

$$\{\hat{\mathbf{w}}_{d|k}\}_{d=d_0}^{\infty} \subset \mathcal{V}_{d_0} := \{\mathbf{v} \in \mathbb{R}^k : \|\mathbf{v}\| \leq B_{d_0}\}\,,$$

and each $\hat{\mathbf{w}}_{d|k}$ satisfies Eq. (14).

We now reach the final stage of the analysis of $\hat{\mathbf{w}}_{d|k}$. Fix the function

$$g(\mathbf{v}) := \mathbb{E}[[1 - y\mathbf{x}_{|k}\mathbf{v}]_+^2]\,,$$

as defined in the theorem. We will the require the following observation, which we state as a lemma:

Lemma 16 *For any d_0, the sequence of functions $\{g_d(\cdot)\}_{d=d_0}^{\infty}$ converges to $g(\cdot)$ almost surely, uniformly on the compact set $\mathcal{V}_{d_0}$.*

Proof It is easy to see that by the law of large numbers, for any fixed $\mathbf{v} \in \mathcal{V}_{d_0}$, $g_d(\mathbf{v})$ converges almost surely to $g(\mathbf{v})$. Thus, $\{g_d(\cdot)\}_{d=d_0}^{\infty}$ almost surely converges to $g(\cdot)$ pointwise. To show that this pointwise convergence implies uniform convergence, it is enough to prove that almost surely, $\{g_d(\cdot)\}_{d=d_0}^{\infty}$ is equicontinuous, a sufficient condition for which is that this infinite sequence of functions on $\mathcal{V}_{d_0}$ has a uniformly bounded Lipschitz parameter (uniformly over d). Since each $g_d(\cdot)$ is differentiable, it is enough to prove a finite upper bound on $\sup_{d \geq d_0} \sup_{\mathbf{v} \in \mathcal{V}_{d_0}} \|\nabla g_d(\mathbf{v})\|$:

Indeed, using Jensen's inequality and the Cauchy-Schwartz inequality, we have

$$\begin{aligned}\|\nabla g_d(\mathbf{v})\| &= \left\|-\frac{2}{m_d}\sum_{i=1}^{m_d}[1-y_i\mathbf{x}_{i|k}^\top\mathbf{v}]_+\cdot y_i\mathbf{x}_{i|k}\right\| \le \frac{2}{m_d}\sum_{i=1}^{m}\left\|[1-y_i\mathbf{x}_{i|k}^\top\mathbf{v}]_+\cdot y_i\mathbf{x}_{i|k}\right\| \\ &\le \frac{2}{m_d}\sum_{i=1}^{m}(1+|y_i\mathbf{x}_{i|k}^\top\mathbf{v}|)\cdot\|y_i\mathbf{x}_{i|k}\| \le \frac{2}{m_d}\sum_{i=1}^{m_d}\left(\|y_i\mathbf{x}_{i|k}\|+\|y_i\mathbf{x}_{i|k}\|^2\|\mathbf{v}\|\right) \\ &\le 2\left(\sqrt{\frac{1}{m_d}\sum_{i=1}^{m_d}\|y_i\mathbf{x}_{i|k}\|^2}\right)+2\left(\frac{1}{m_d}\sum_{i=1}^{m_d}\|y_i\mathbf{x}_{i|k}\|^2\right)\|\mathbf{v}\|. \qquad (15)\end{aligned}$$

Assumption 1 implies that $\|y\mathbf{x}_{|k}\|^2$ has bounded expectation. Therefore, by the law of large numbers, $\frac{1}{m_d}\sum_{i=1}^{m_d}\|y_i\mathbf{x}_{i|k}\|^2$ converges almost surely to $\mathbb{E}[\|y\mathbf{x}_{|k}\|^2]$. Moreover, $\|\mathbf{v}\|$ is also bounded, since $\mathbf{v}\in\mathcal{V}_{d_0}$ and $\mathcal{V}_{d_0}$ is a compact set. In view of Eq. (15), we get that $\|\nabla g_d(\mathbf{v})\|$ is almost surely bounded uniformly for all $\mathbf{v}\in\mathcal{V}_{d_0}$ and $d\ge d_0$. As discussed previously, this implies equicontinuity and hence uniform convergence. ■

Now, fix some reference vector $\mathbf{v}_0\in\mathbb{R}^k$. We make the following observations:

1. Since $\hat{\mathbf{w}}_{d|k}$ is a minimizer of the expression in Eq. (14) (with probability $1-o_{d_0}(1)$ for all $d\ge d_0$), it follows that with the same probability, for any fixed reference vector $\mathbf{v}_0$,

$$\begin{aligned}(1-\epsilon_d)\cdot g_d(\hat{\mathbf{w}}_{d|k}) &\le (1+\epsilon_{d,\hat{\mathbf{w}}_{d|k}})\cdot g_d(\hat{\mathbf{w}}_{d|k})+\frac{\|\hat{\mathbf{w}}_{d|k}\|^2}{m_d} \le (1+\epsilon_{d,\mathbf{v}_0})\cdot g_d(\mathbf{v}_0)+\frac{\|\mathbf{v}_0\|^2}{m_d} \\ &\le (1+\epsilon_d)\cdot g_d(\mathbf{v}_0)+\frac{\|\mathbf{v}_0\|^2}{m_d}\,,\end{aligned}$$

   and therefore

$$g_d(\hat{\mathbf{w}}_{d|k}) \le \frac{1+\epsilon_d}{1-\epsilon_d}\cdot g_d(\mathbf{v}_0)+\frac{\|\mathbf{v}_0\|^2}{(1-\epsilon_d)m_d}\,.$$

2. By Lemma 16 above, and the fact that $\{\hat{\mathbf{w}}_{d|k}\}_{d=d_0}^{\infty}\subset\mathcal{V}_{d_0}$ with probability $1-o_{d_0}(1)$, it follows that with the same probability,

$$|g_d(\hat{\mathbf{w}}_{d|k})-g(\hat{\mathbf{w}}_{d|k})|\overset{d\to\infty}{\longrightarrow}0\,.$$

   Moreover, by the law of large numbers, it holds with probability 1 that

$$|g_d(\mathbf{v}_0)-g(\mathbf{v}_0)|\overset{d\to\infty}{\longrightarrow}0\,.$$

3. Since $|\epsilon_{d,\mathbf{v}}|\le\epsilon_d\overset{d\to\infty}{\longrightarrow}0$, we have that $\frac{1+\epsilon_d}{1-\epsilon_d}$ and $1-\epsilon_d$ converge to 1.

Combining these three observations, and the fact that $\frac{\|\mathbf{v}_0\|^2}{m_d}\overset{d\to\infty}{\longrightarrow}0$, it follows that with probability at least $1-o_{d_0}(1)$,

$$\limsup_{d\to\infty} g(\hat{\mathbf{w}}_{d|k})-g(\mathbf{v}_0)\le 0\,.$$

This holds for any $\mathbf{v}_0$, hence $g(\hat{\mathbf{w}}_{d|k})$ must converge to $\inf_{\mathbf{v}}g(\mathbf{v})$ with probability at least $1-o_{d_0}(1)$. Taking d_0 to infinity, we get that $g(\hat{\mathbf{w}}_{d|k})$ must asymptotically converge to $\inf_{\mathbf{v}}g(\mathbf{v})$ with probability 1, as stated in the theorem.

A.4.3. ANALYSIS OF $\hat{\mathbf{w}}_{d|d-k}$

We now turn to analyze the last $d-k$ coordinates of $\hat{\mathbf{w}}_d$, namely $\hat{\mathbf{w}}_{d|d-k}$. First, we note that since $\hat{\mathbf{w}}_{d|k}$ minimizes the expression in Eq. (14), which equals $1+\epsilon_{d,\mathbf{0}}$ when $\mathbf{v}=0$, it must hold that

$$\frac{\|\hat{\mathbf{w}}_{d|k}\|^2}{m_d}+\frac{1+\epsilon_{d,\hat{\mathbf{w}}_{d|k}}}{m_d}\sum_{i=1}^{m_d}[1-y_i\mathbf{x}_{i|k}^\top\hat{\mathbf{w}}_{d|k}]_+^2 \;\leq\; 1+\epsilon_{d,\mathbf{0}}\,.$$

Recalling that $\sup_{\mathbf{v}}|\epsilon_{d,\mathbf{v}}|\leq\epsilon_d\overset{d\to\infty}{\longrightarrow}0$, it follows in particular that $\frac{1}{m_d}\sum_{i=1}^{m_d}[1-y_i\mathbf{x}_{i|k}^\top\hat{\mathbf{w}}_{d|k}]_+^2\;\leq\;2$ (all this for large enough d and with probability at least $1-\frac{1}{d^2}$, as specified there).

Next, note that by Eq. (10) and the fact that $\hat{\mathbf{w}}_{d|k},\hat{\mathbf{w}}_{d|d-k}$ jointly optimize it over $(\mathbf{v},\mathbf{u})$, we have

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

Combined with Eq. (13) (using $\mathbf{v}=\hat{\mathbf{w}}_{d|k}$), it follows that with probability at least $1-\frac{1}{d^2}$,

$$\frac{\|\hat{\mathbf{w}}_{d|d-k}\|^2}{m_d}\;\leq\;\left(1+\mathcal{O}\left(\sqrt{\frac{m_d\log(d)}{d}}\right)\right)\cdot\frac{1}{m_d}\|[\mathbf{1}-X\hat{\mathbf{w}}_{d|k}]_+\|^2\,,$$

and since we showed that $\frac{1}{m_d}\|[\mathbf{1}-X\hat{\mathbf{w}}_{d|k}]_+\|^2=\frac{1}{m_d}\sum_{i=1}^{m_d}[1-y_i\mathbf{x}_{i|k}^\top\hat{\mathbf{w}}_{d|k}]_+^2\;\leq\;2$ under the same event, it follows that

$$\frac{\|\hat{\mathbf{w}}_{d|d-k}\|^2}{m_d}\;\leq\;2+\mathcal{O}\left(\sqrt{\frac{m_d^2\log(d)}{d}}\right)$$

with probability at least $1-\frac{1}{d^2}$. By a union bound, it follows that this holds simultaneously for all sufficiently large d, with arbitrarily high probability, hence almost surely, $\limsup_{d\to\infty}\frac{\|\hat{\mathbf{w}}_{d|d-k}\|^2}{m_d}\leq 2$.

As a result, if $\mathbf{z}\in\mathbb{R}^{d-k}$ is zero-mean Gaussian with covariance matrix $\frac{1}{d-k}I$, we have

$$\mathbb{E}_{\mathbf{z}}[(\mathbf{z}^\top\hat{\mathbf{w}}_{d|d-k})^2]\;=\;\frac{1}{d-k}\cdot\|\hat{\mathbf{w}}_{d|d-k}\|^2\;=\;\frac{m_d}{d-k}\cdot\frac{\|\hat{\mathbf{w}}_{d|d-k}\|^2}{m_d}\,,$$

where $\frac{m_d}{d-k}\overset{d\to\infty}{\longrightarrow}0$ by assumption, and $\limsup_d\frac{\|\hat{\mathbf{w}}_{d|d-k}\|^2}{m_d}\leq 2$ almost surely, hence $\mathbb{E}_{\mathbf{z}}[(\mathbf{z}^\top\hat{\mathbf{w}}_{d|d-k})^2]$ converges almost surely to 0. Recalling that by Assumption 1, $\mathbf{z}$ has the same distribution as $\mathbf{x}_{|d-k}$, it follows that

$$\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_d}\left[(\mathbf{x}^\top\hat{\mathbf{w}}_d-\mathbf{x}_{|k}^\top\hat{\mathbf{w}}_{d|k})^2\right]\;=\;\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_d}\left[(\mathbf{x}_{|d-k}^\top\hat{\mathbf{w}}_{d|d-k})\right]\;\overset{a.s.}{\longrightarrow}\;0\,.$$

A.5. Proof of Thm. 11

Since we consider labels flipped with some positive probability p, we trivially 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 almost surely to p.

As discussed before the theorem, $L_p(\cdot)$ has a unique minimizer $\mathbf{w}_p^*$. Therefore, by Thm. 10, $\hat{\mathbf{w}}_{d|k} \xrightarrow{a.s.} \mathbf{w}_p^*$. Since we assume $\Pr_{(\mathbf{x},y)\sim\mathcal{D}_{\text{clean}}}(y\mathbf{x}^\top \mathbf{w}_p^* \le 0) = 0$, we argue that

$$\Pr_{(\mathbf{x},y)\sim\mathcal{D}_{\text{clean}}}(y\mathbf{x}^\top \hat{\mathbf{w}}_{d|k} \le 0) \xrightarrow{a.s.} 0\,. \tag{16}$$

We note that formally proving this requires some care, as $\Pr(y\mathbf{x}^\top \mathbf{w} \le 0)$ is not necessarily continuous in $\mathbf{w}$ (otherwise Eq. (16) would follow immediately by continuity). To show Eq. (16) formally, define for all $\gamma > 0$ the set $\mathcal{Z}_\gamma := \{\mathbf{z} \in \mathbb{R}^k : \mathbf{z}^\top \mathbf{w}_p^* > \gamma, \|\mathbf{z}\| \le \frac{1}{\gamma}\}$. Clearly,

$$\Pr_{(\mathbf{x},y)\sim\mathcal{D}_{\text{clean}}}(y\mathbf{x} \in \mathcal{Z}_\gamma) \xrightarrow{\gamma\to 0} 1\,. \tag{17}$$

Moreover, since $\hat{\mathbf{w}}_{d|k} \xrightarrow{a.s.} \mathbf{w}_p^*$, it holds with probability 1 that $y\mathbf{x}^\top \hat{\mathbf{w}}_{d|k} \longrightarrow y\mathbf{x}^\top \hat{\mathbf{w}}_p^*$ simultaneously for all vectors $y\mathbf{x}$ of some bounded norm. Therefore, for any fixed γ,

$$\Pr_{(\mathbf{x},y)\sim\mathcal{D}_{\text{clean}}}(y\mathbf{x}^\top \hat{\mathbf{w}}_{d|k} \le 0 \mid y\mathbf{x} \in \mathcal{Z}_\gamma) = \Pr_{(\mathbf{x},y)\sim\mathcal{D}_{\text{clean}}}\left(y\mathbf{x}^\top \hat{\mathbf{w}}_{d|k} \le 0 \mid y\mathbf{x}^\top \mathbf{w}_p^* > \gamma, \|y\mathbf{x}\| \le \frac{1}{\gamma}\right) \xrightarrow{a.s.} 0\,. \tag{18}$$

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

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

it follows that for any fixed γ,

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

Combined with Eq. (17) and Eq. (18), it follows that by picking γ small enough, then almost surely, we can make $\Pr_{(\mathbf{x},y)\sim\mathcal{D}_{\text{clean}}}\left(y\mathbf{x}^\top \hat{\mathbf{w}}_{d|k} \le 0\right)$ asymptotically smaller than arbitrarily small positive numbers, from which Eq. (16) follows.

From Eq. (16), it follows that $\Pr_{(\mathbf{x},y)\sim\mathcal{D}_d}(y\mathbf{x}_{|k}^\top \hat{\mathbf{w}}_{d|k} \le 0) \xrightarrow{a.s.} p$. By Thm. 10, we also have that

$$\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_d}\left[(\mathbf{x}^\top \hat{\mathbf{w}}_d - \mathbf{x}_{|k}^\top \hat{\mathbf{w}}_{d|k})^2\right] \xrightarrow{a.s.} 0\,. \tag{19}$$

Thus, we can use similar arguments as above, to prove that

$$\Pr_{(\mathbf{x},y)\sim\mathcal{D}_d}(y\mathbf{x}^\top \hat{\mathbf{w}}_d \le 0) \xrightarrow{a.s.} p\,, \tag{20}$$

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 $\mathcal{Z}_\gamma^d = \{\mathbf{z} \in \mathbb{R}^d : \mathbf{z}_{|k}^\top \mathbf{w}_p^* > \gamma, \|\mathbf{z}\| \le \frac{1}{\gamma}\}$. Similar to before, we have

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

By applying Markov's inequality on Eq. (19), it follows that for any $\gamma > 0$, the measure of points $y\mathbf{x} \in \mathcal{Z}_\gamma^d$ 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 almost surely. For all other points, 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 $y\mathbf{x}_{|k}^\top \hat{\mathbf{w}}_{d|k} \xrightarrow{a.s.} y\mathbf{x}_{|k}^\top \mathbf{w}_p^*$ (which is $> \gamma$) uniformly for all $y\mathbf{x} \in \mathcal{Z}_\gamma^d$, we get that

$$\Pr_{(\mathbf{x},y)\sim\tilde{\mathcal{D}}_d}\left(y\mathbf{x}^\top \hat{\mathbf{w}}_d \le 0 \mid y\mathbf{x} \in \mathcal{Z}_\gamma^d\right) \xrightarrow{a.s.} 0\,.$$

Combining the two displayed equation above, and using the same arguments as we made in the context of Eq. (17) and Eq. (18), it follows that almost surely, $\Pr_{(\mathbf{x},y)\sim\tilde{\mathcal{D}}_d}(y\mathbf{x}^\top \hat{\mathbf{w}}_d \le 0)$ can be made asymptotically smaller than arbitrarily small positive numbers, from which it follows that

$$\Pr_{(\mathbf{x},y)\sim\tilde{\mathcal{D}}_d}\left(y\mathbf{x}^\top \hat{\mathbf{w}}_d \le 0\right) \xrightarrow{a.s.} 0\,.$$

Switching from $\tilde{\mathcal{D}}_d$ to $\mathcal{D}_d$ (which involves flipping y randomly with probability p), Eq. (20) follows.

A.6. Proof of Thm. 12

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 $\alpha \in [0,1]$,

$$f(\alpha\mathbf{u} + (1-\alpha)\mathbf{v}) \;\le\; \alpha \cdot f(\mathbf{u}) + (1-\alpha)f(\mathbf{v}) - \alpha(1-\alpha)\cdot\frac{\lambda}{2}\|\mathbf{u}-\mathbf{v}\|^2\,. \tag{21}$$

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 $f(\mathbf{w}) - f(\mathbf{w}^*) \ge \frac{\lambda}{2}\|\mathbf{w}-\mathbf{w}^*\|^2$ for any $\mathbf{w}$.

We start with the following auxiliary lemma, which implies that $L_p(\cdot)$ is strongly convex:

Lemma 17 *If $\mathbb{E}[\mathbf{x}\mathbf{x}^\top]$ 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. (2)) is $2p\lambda_{\min}$-strongly convex.*

Proof We have $L_p(\mathbf{w}) = \mathbb{E}[\ell_p(y\mathbf{x}_{|k}^\top \mathbf{w})]$, 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 $(-\infty,-1]$, β^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.

$$\mathbb{E}[(\mathbf{x}^\top\mathbf{u} - \mathbf{x}^\top\mathbf{v})^2] \;=\; (\mathbf{u}-\mathbf{v})^\top\mathbb{E}[\mathbf{x}\mathbf{x}^\top](\mathbf{u}-\mathbf{v}) \ge \lambda_{\min}\|\mathbf{u}-\mathbf{v}\|^2\,.$$

Combining, we get that

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

Which by Eq. (21), implies that L_p is $2p\lambda_{\min}$-strongly convex. ■

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

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

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

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

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 17, $\tilde{L}_p(\mathbf{w})$ is $\frac{2p\lambda}{1-2p}$-strongly convex for some $\lambda > 0$ (which would also imply that its minimizer $\mathbf{w}_p^*$ always exists and is unique).

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 $\frac{2p\lambda}{1-2p}$-strong convexity of $\tilde{L}_p(\cdot)$ and the fact that $\mathbf{w}_p^*$ is a minimizer, it would imply

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

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

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

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 $\tilde{L}_{p_j}(\mathbf{w}_{p_0}^*) - \tilde{L}_{p_0}(\mathbf{w}_{p_0}^*) \stackrel{j\to\infty}{\longrightarrow} 0$ and $\tilde{L}_{p_0}(\mathbf{w}_{p_j}^*) - \tilde{L}_{p_j}(\mathbf{w}_{p_j}^*) \stackrel{j\to\infty}{\longrightarrow} 0$.

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 $\mathbf{w}_p^* \stackrel{p\to 0}{\longrightarrow} \hat{\mathbf{w}}$ 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 $\tilde{L}_p(\hat{\mathbf{w}}) \stackrel{p\to 0}{\longrightarrow} \tilde{L}_0(\mathbf{w}^*)$. Recalling that $\tilde{L}_0(\mathbf{w}^*) = L_0(\mathbf{w}^*) = 0$, it follows that

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

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

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

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

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

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

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

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

A.7. Proof of Thm. 13

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 the distribution of

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

is the same as

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

and the the value of $L_p(\cdot)$ depends just on the distribution of $y\mathbf{x}^\top\mathbf{w}_p^*$. But since $L_p(\cdot)$ is strongly convex, its minimizer must be unique, which is a contradiction.

Next, we argue that we can assume $\mathbf{w}^*$ (which minimizes $L_0(\mathbf{w}) = \mathbb{E}[[1 - y\mathbf{x}^\top\mathbf{w}]_+^2]$) to be 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 $L_0(\mathbf{w}^*) = \mathbb{E}[[1 - y\mathbf{x}^\top\alpha\mathbf{u}]_+^2] = 0$, it follows that $y\mathbf{x}^\top\mathbf{u} > 0$ with probability 1. Therefore,

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

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

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

as required.

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

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

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

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:

Theorem 18 *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 $L_p(\mathbf{w}) \;=\; \frac{1}{4}\sum_{i=1}^4 \left((1-p) \cdot [1 - y\mathbf{x}^\top \mathbf{w}]_+^2 + p \cdot [1 + y\mathbf{x}^\top \mathbf{w}]_+^2 \right)$, 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 17, 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

$$4{\cdot}\nabla L_p(\hat{\mathbf{w}}) = 4{\cdot}\nabla L_p(\hat{w}_1, \hat{w}_2) \;=\; 2p\ell_p'(p(\hat{w}_1{-}\hat{w}_2)) \begin{pmatrix} 1 \\ -1 \end{pmatrix} {+} \ell_p'(w_1) \begin{pmatrix} 1 \\ 0 \end{pmatrix} {+} p\ell_p'(p(\hat{w}_1{+}5\hat{w}_2)) \begin{pmatrix} 1 \\ 5 \end{pmatrix}, \tag{22}$$

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. (22), 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. ■