Optimal Excess Risk Bounds for Empirical Risk Minimization on p-Norm Linear Regression

Ayoub El Hanchi[*] Murat A. Erdogdu[†]

September 26, 2023

Abstract

We study the performance of empirical risk minimization on the p-norm linear regression problem for $p \in (1, \infty)$. We show that, in the realizable case, under no moment assumptions, and up to a distribution-dependent constant, $O(d)$ samples are enough to exactly recover the target. Otherwise, for $p \in [2, \infty)$, and under weak moment assumptions on the target and the covariates, we prove a high probability excess risk bound on the empirical risk minimizer whose leading term matches, up to constants that depend only on p, the asymptotically optimal rate. We extend this result to the case $p \in (1, 2)$ under mild assumptions that guarantee the existence of the Hessian of the risk at its minimizer.

1 Introduction

Real-valued linear prediction is a fundamental problem in machine learning. Traditionally, the square loss has been the default choice for this problem. The performance of empirical risk minimization (ERM) on linear regression under the square loss, as measured by the excess risk, has been studied extensively both from an asymptotic [Whi82; Vaa98; LC06] and a non-asymptotic point of view [AC11; HKZ12; Oli16; LM16; Sau18; Mou22]. A major achievement of the last decade has been the development of non-asymptotic excess risk bounds for ERM on this problem under weak assumptions, and which match, up to constant factors, the asymptotically optimal rate.

In this paper, we consider the more general family of p-th power losses $t \mapsto |t|^p$ for a user-chosen $p \in (1, \infty)$. Under mild assumptions, the classical asymptotic theory can still be applied to ERM under these losses, yielding the asymptotic distribution of the excess risk. However, to the best of our knowledge, the problem of deriving non-asymptotic excess risk bounds for ERM for $p \in (1, \infty) \setminus \{2\}$ remains open, and, as we discuss below, resists the application of standard tools from the literature.

Our motivation for extending the case $p = 2$ to $p \in (1, \infty)$ is twofold. Firstly, the freedom in the choice of p allows us to better capture our prediction goals. For example, we might only care about how accurate our prediction is on average, in which case, the choice $p = 1$ is appropriate. At the other extreme, we might insist that we do as well as possible on a subset of inputs of probability 1, in which case the choice $p = \infty$ is best. A choice of $p \in (1, \infty)$ therefore allows us to interpolate between these two extremes, with the case $p = 2$ offering a balanced choice. Secondly, different choices of p have complementary qualities. On the one hand, small values of p allow us to operate with weak assumptions, making them applicable in more general cases. On the other, larger values of p yield predictions whose optimality is less sensitive to changes in the underlying distribution: for $p = \infty$, the best predictor depends only on the support of this distribution.

To sharpen our discussion, let us briefly formalize our problem. There is an input random vector $X \in \mathbb{R}^d$ and and output random variable $Y \in \mathbb{R}$, and we are provided with n i.i.d. samples $(X_i, Y_i)_{i=1}^n$. We select our set of predictors to be the class of linear functions $\left\{x \mapsto \langle w, x\rangle \mid w \in \mathbb{R}^d\right\}$, and choose a value $p \in (1, \infty)$ with the corresponding loss $\ell_p(t) := |t|^p/[p(p-1)]$.[1] Using this loss, we define the associated risk and empirical

[*] Department of Computer Science at University of Toronto, and Vector Institute, aelhan@cs.toronto.edu

[†] Department of Computer Science at University of Toronto, and Vector Institute, erdogdu@cs.toronto.edu

[1] The rescaling of the loss by $1/[p(p-1)]$ is motivated by aesthetic reasons we encounter later.

risk by

$$R_p(w) := \mathrm{E}[\ell_p(\langle w, X\rangle - Y)], \qquad\qquad R_{p,n}(w) := \frac{1}{n}\sum_{i=1}^{n} \ell_p(\langle w, X_i\rangle - Y_i).$$

We perform empirical risk minimization $\hat{w}_p \in \operatorname{argmin}_{w\in\mathbb{R}^d} R_{p,n}(w)$, and our goal is to derive high probability bounds on the excess risk $R_p(\hat{w}_p) - R_p(w_p^*)$, where w_p^* is the risk minimizer. For issues related to the computation of an empirical risk minimizer $\hat{w}_p$, we refer the reader to the rich recent literature dealing with this problem [Bub+18; Adi+19; APS19; JLS22].

To see why the problem we are considering is difficult, let us briefly review some of the recent literature. Most closely related to our problem are the results of [AC11; HKZ12; Oli16; LM16], who derive high probability non-asymptotic excess risk bounds for the case $p = 2$. The best such bounds are found in [Oli16] and [LM16], who both operate under weak assumptions on (X, Y), requiring at most the existence of fourth moments of Y and the components X^j of X for $j \in [d]$. Unfortunately, the analysis in [Oli16] relies on the closed form expression of the empirical risk minimizer $\hat{w}_2$, and therefore cannot be extended to other values of p. Similarly, the analysis in Lecué and Mendelson [LM16] relies on an exact decomposition of the excess loss $\ell_2(\langle w, X\rangle - Y) - \ell_2(\langle w_p^*, X\rangle - Y)$ in terms of "quadratic" and "multiplier" components, which also does not extend to other values of p.

To address these limitations, the work of Mendelson [Men18] extends the ideas of Mendelson [Men14] and Lecué and Mendelson [LM16] to work for loss functions more general than the square loss. Roughly speaking, the main result of Mendelson [Men18] states that as long as the loss is strongly convex and smooth in a neighbourhood of 0, then the techniques developed by Mendelson [Men14] can still be applied to obtain high probability excess risk bounds. Unfortunately, the loss functions $\ell_p(t)$ are particularly ill-behaved in precisely this sense, as $\ell_p''(t) \to 0$ when $t \to 0$ for $p > 2$, and $|\ell_p''(t)| \to \infty$ as $t \to 0$ for $p \in (1, 2)$. This makes the analysis of the excess risk of ERM in the case $p \in (1, \infty) \setminus \{2\}$ particularly challenging using well-established methods.

Contrary to the non-asymptotic regime, the asymptotic properties of the excess risk of ERM under the losses ℓ_p are better understood [Ron84; BRW92; Nie92; Arc96; HS96; LL05], and can be derived from the more general classical asymptotic theory of M-estimators [VW96; Vaa98; LC06] under mild regularity conditions. In particular, these asymptotic results imply that the excess risk of ERM with n samples satisfies

$$\mathrm{E}[R_p(\hat{w}_p)] - R_p(w_p^*) = \frac{\mathrm{E}\Big[\|\nabla \ell_p(\langle w_p^*, X\rangle - Y)\|_{H_p^{-1}}^2\Big]}{2n} + o\bigg(\frac{1}{n}\bigg) \quad \text{as} \quad n \to \infty, \tag{1.1}$$

where $H_p := \nabla^2 R_p(w_p^*)$ is the Hessian of the risk at its minimizer. We refer the reader to the discussions in Ostrovskii and Bach [OB21] and Mourtada and Gaïffas [MG22] for more details. As we demonstrate in Theorem 2, the rate of convergence of ERM for the square loss derived in Oliveira [Oli16] and Lecué and Mendelson [LM16] matches the asymptotic rate (1.1) up to constant factors. Ideally, we would like our high probability excess risk bounds for the cases $p \in (1, \infty) \setminus \{2\}$ to also match the optimal rate (1.1), although it is not yet clear how to derive any meaningful such bounds.

In this paper, we prove the first high probability excess risk bounds for ERM under the p-th power loss $\ell_p(t)$ for any $p \in (1, \infty) \setminus \{2\}$. Our assumptions on (X, Y) are weak, arise naturally from the analysis, and reduce to the standard ones for the case $p = 2$. Furthermore, the rate we derive matches, up to a constant that depends only on p, the asymptotically optimal rate (1.1).

We split the analysis in three cases. The first is when the problem is realizable, i.e. $Y = \langle w^*, X\rangle$ for some $w^* \in \mathbb{R}^d$. This edge case is not problematic for the analysis of the case $p = 2$, but as discussed above, the ℓ_p losses are ill-behaved around 0 for $p \in (1, \infty) \setminus \{2\}$, requiring us to treat this case separately. The second case is when the problem is not realizable and $p \in (2, \infty)$. The final case is when the problem is not realizable and $p \in (1, 2)$, which turns out to be the most technically challenging. In Section 2 we present our main results and in Section 3 we provide their proofs.

Notation. We denote the components of the random vector $X \in \mathbb{R}^d$ by X^j for $j \in [d]$. We assume the support of X is not contained in any hyperplane, i.e. $\mathrm{P}(\langle w, X\rangle = 0) = 1$ only if $w = 0$. This is without loss of generality as argued in Oliveira [Oli16] and Mourtada [Mou22]. For a positive semi-definite matrix A, we denote the bilinear form it induces on $\mathbb{R}^d$ by $\langle \cdot, \cdot\rangle_A$, and define $\|\cdot\|_A = \sqrt{\langle \cdot, \cdot\rangle_A}$.

2 Main results

In this section we state our main results. We start in Section 2.1 where we introduce constants that help us formulate our theorems. In Section 2.2, we state the best known results for both the case $p = 2$ and the realizable case where $Y = \langle w^*, X\rangle$. Finally, in Section 2.3, we state our theorems.

2.1 Norm equivalence and small ball constants

To state our results, we will need to define two types of quantities first. The first kind are related to norms and their equivalence constants, which we will use in the analysis of the non-realizable case. The second are small ball probabilities, and which we will use for the realizable case.

We start by introducing the following functions on our space of coefficients $\mathbb{R}^d$. For $p, q \in [1, \infty)$, define, with the convention $\infty^{1/p} = \infty$ for all $p \in [1, \infty)$,

$$\|w\|_{L^p} := \mathrm{E}[|\langle w, X\rangle|^p]^{1/p}, \qquad \|w\|_{L^q,p} := \mathrm{E}[\|w\|^q_{\nabla^2 \ell_p(\langle w^*_p, X\rangle - Y)}]^{1/q}. \tag{2.1}$$

As suggested by the notation, under appropriate assumptions, these are indeed norms on $\mathbb{R}^d$. In that case, we will be interested in norm equivalence constants between them

$$C_{a\to b} := \sup_{w\in\mathbb{R}^d\setminus\{0\}} \frac{\|w\|_a}{\|w\|_b}, \qquad \sigma_p^2 := C^4_{(L^4,p)\to(L^2,p)}, \tag{2.2}$$

where a and b stand for one of L^p or (L^q, p). Let us note that since we work in a finite dimensional vector space, all norms are equivalent, so that as soon as the quantities defined in (2.1) are indeed norms, the constants defined in (2.2) are finite. Furthermore, as suggested by the notation, σ_p^2 may be viewed as the maximum second moment of the random variables $\|w\|^2_{\nabla^2\ell_p(\langle w^*_p, X\rangle - Y)}$ over the unit sphere of $\|\cdot\|_{L^2,p}$. Finally, we record the following identities for future use

$$\|w\|_{L^2,p} = \|w\|_{H_p}, \qquad \|w\|_{L^q,2} = \|w\|_{L^q}, \qquad \sigma_2^2 = C^4_{L^4,L^2}. \tag{2.3}$$

The first identity holds by linearity, and the second by noticing that $\nabla^2\ell_2(\langle w, X\rangle - Y) = XX^T$.

We now turn to small ball probabilities. We define the following functions on $\mathbb{R}^d$, for $q \in [1, \infty)$,

$$\rho_0(w) := \mathrm{P}(\langle w, X\rangle = 0), \qquad \rho_q(w, \kappa) := \mathrm{P}(|\langle w, X\rangle| > \kappa\|w\|_{L^q}). \tag{2.4}$$

Assumptions on the functions ρ_0 and ρ_2 have been used extensively in the recent literature, see e.g. [Men14; KM15; LM17a; LM17b; Men18; LM18; Mou22]. In particular, a standard assumption postulates the existence of strictly positive constants β_0, and (β_2, κ_2) such that $\rho_0(w) \leq 1 - \beta_0$ and $\rho_2(w, \kappa_2) \geq \beta_2$ for all $w \in \mathbb{R}^d$. Conditions of this type are usually referred to as small ball conditions. Efforts have been made to understand when these conditions hold [Men14; RV15; LM17b] as well as reveal the dimension dependence of the constants with which they do [Sau18]. Here we prove that such conditions always hold for finite dimensional spaces. We leave the proof of Lemma 1 to the Appendix to not distract from our main development.

Lemma 1. *ρ_0 is upper semi-continuous. Furthermore, if for some $q \in [1, \infty)$, $\mathrm{E}[|X_j|^q] < \infty$ for all $j \in [d]$, then $\rho_q(\cdot, \kappa)$ is lower semi-continuous for any $\kappa \geq 0$. Moreover, for all $\kappa \in [0, 1)$*

$$\rho := \sup_{w\in\mathbb{R}^d\setminus\{0\}} \rho_0(w) < 1, \qquad \inf_{w\in\mathbb{R}^d\setminus\{0\}} \rho_q(w, \kappa) > 0.$$

2.2 Background

To better contextualize our results, we start by stating the best known high probability bound on ERM for the square loss, which we deduce from Oliveira [Oli16] and Lecué and Mendelson [LM16].

Theorem 2 (Oliveira [Oli16] and Lecué and Mendelson [LM16])**.** *Assume that $\mathrm{E}[Y^2] < \infty$ and $\mathrm{E}[(X^j)^4] < \infty$ for all $j \in [d]$, and let $\delta \in (0, 1]$. If*

$$n \geq 196\sigma_2^2(d + 2\log(4/\delta)),$$

then, with probability at least $1 - \delta$

$$R_2(\hat{w}_2) - R_2(w_2^*) \leq \frac{16\,\mathrm{E}[\|\nabla \ell_2(\langle w_2^*, X\rangle - Y)\|^2_{H_2^{-1}}]}{n\delta}.$$

Up to a constant factor and the dependence on δ, Theorem 2 recovers the optimal bound (1.1). Let us briefly comment on the differences between Theorem 2 and the comparable statements in the original papers. First, the finiteness of σ_2^2 is deduced from the finiteness of the fourth moments of the components of X, instead of being assumed as in Oliveira [Oli16] (see the discussion in Section 3.1 in Oliveira [Oli16]). Second we combine Theorem 3.1 from [Oli16] with the proof technique of Lecué and Mendelson [LM16] to achieve a slightly better bound that the one achieved by the proof technique used in the proof of Theorem 4.2 in Oliveira [Oli16], while avoiding the dependence on the small ball-constant present in the bound of Theorem 1.3 in Lecué and Mendelson [LM16], which is known to incur additional dimension dependence in some cases [Sau18].

We now move to the realizable case, where $Y = \langle w^*, X\rangle$ so that $w_p^* = w^*$ for all $p \in (1, \infty)$. We immediately note that Theorem 2 is still applicable in this case, and ensures that we recover w^* exactly with no more than $n = O(\sigma_2^2 d)$ samples. However, we can do much better, while getting rid of all the moment assumptions in Theorem 2. Indeed, it is not hard to see that $\hat{w}_p \neq w^*$ only if for some $w \in \mathbb{R}^d \setminus \{0\}$, $\langle w, X_i\rangle = 0$ for all $i \in [n]$ (taking $w = \hat{w}_p - w_p^*$ works). The implicit argument in Theorem 2 then uses the pointwise bound (see Lemma B.2 in Oliveira [Oli16])

$$\mathrm{P}(\cap_{i=1}^n \{\langle w, X_i\rangle = 0\}) \leq \exp\left(-\frac{n}{2\sigma_2^2}\right)$$

and uniformizes it over the L^2 unit sphere in $\mathbb{R}^d$, where the L^2 norm is as defined in (2.1). However, we can use the much tighter bound ρ^n where ρ is as defined in Lemma 1. To the best of our knowledge, the realizable case has not been studied explicitly before in the literature. However, with the above considerations in mind, we can deduce the following result from Lecué and Mendelson [LM17b], which uniformizes the pointwise bound we just discussed using a VC dimension argument.

Theorem 3 (Corollary 2.5, Lecué and Mendelson [LM17b])**.** *Assume that there exists* $w^* \in \mathbb{R}^d$ *such that* $Y = \langle w^*, X\rangle$*. Let* $\delta \in (0, 1]$*. If*

$$n \geq O\left(\frac{d + \log(1/\delta)}{(1-\rho)^2}\right)$$

then for any $p \in (1, \infty)$*,* $\hat{w}_p = w^*$ *with probability at least* $1 - \delta$*.*

2.3 Results

We are now in position to state our main results. As discussed in Section 1, the ℓ_p losses have degenerate second derivatives as $t \to 0$. When the problem is realizable, the risk is not twice differentiable at its minimizer for the cases $p \in (1, 2)$, and is degenerate for the cases $p \in (2, \infty)$. If we want bounds of the form (1.1), we must exclude this case from our analysis. This is in part what motivates us to study the realizable case separately. Our first main result is a strengthening of Theorem 3, and relies on a combinatorial argument to uniformize the pointwise estimate discussed in Section 2.2.

Theorem 4. *Assume that there exists* $w^* \in \mathbb{R}^d$ *such that* $\langle w^*, X\rangle = Y$*. Then for all* $n \geq d$*, and for all* $p \in (1, \infty)$*, we have*

$$\mathrm{P}(\hat{w}_p \neq w^*) \leq \binom{n}{d-1}\rho^{n-d+1}.$$

Furthermore, if

$$n \geq \begin{cases} O(d + \log(1/\delta)/\log(1/\rho)) & \textit{if} \quad 0 \leq \rho < e^{-1} \\ O\left(\dfrac{d + \log(1/\delta)}{1-\rho}\right) & \textit{if} \quad e^{-1} \leq \rho < e^{-1/e} \\ O\left(\dfrac{d\log(1/(1-\rho)) + \log(1/\delta)}{1-\rho}\right) & \textit{if} \quad e^{-1/e} \leq \rho < 1, \end{cases}$$

then with probability at least $1 - \delta$*,* $\hat{w}_p = w^*$*.*

Comparing Theorem 3 and Theorem 4, we see that the bound on the number of samples required to reach a confidence level δ in Theorem 4 is uniformly smaller than the one in Theorem 3. The proof of Theorem 4 can be found in the Appendix.

We now move to the more common non-realizable case. Our first theorem here gives a non-asymptotic bound for the excess risk of ERM under a p-th power loss for $p \in (2, \infty)$. To the best of our knowledge, no such result is known in the literature.

Theorem 5. *Let $p \in (2, \infty)$ and $\delta \in (0, 1]$. Assume that no $w \in \mathbb{R}^d$ satisfies $Y = \langle w, X \rangle$. Further, assume that $\mathrm{E}[|Y|^p] < \infty$, $\mathrm{E}[|X^j|^p] < \infty$, and $\mathrm{E}[|\langle w_p^*, X\rangle - Y|^{2(p-2)}(X^j)^4] < \infty$ for all $j \in [d]$. If*

$$n \geq 196\sigma_p^2(d + 2\log(4/\delta)),$$

then with probability at least $1 - \delta$

$$R_p(\hat{w}_p) - R_p(w_p^*) \leq \frac{2048p^2\,\mathrm{E}\Big[\|\nabla \ell_p(\langle w_p^*, X\rangle - Y)\|_{H_p^{-1}}^2\Big]}{n\delta} + \left(\frac{512p^4c_p^2\,\mathrm{E}\Big[\|\nabla \ell_p(\langle w_p^*, X\rangle - Y)\|_{H_p^{-1}}^2\Big]}{n\delta}\right)^{p/2},$$

where we used c_p to denote $C_{L^p \to (L^2, p)}$ as defined in (2.2).

Up to a constant factor that depends only on p and the dependence on δ, the bound of Theorem 5 is precisely of the form of the optimal bound (1.1). Indeed, as $p > 2$, the second term is $o(1/n)$. At the level of assumptions, the finiteness of the p-th moment of Y and the components of X is necessary to ensure that the risk R_p is finite for all $w \in \mathbb{R}^d$. The last assumption $\mathrm{E}[|Y - \langle w_p^*, X\rangle|^{2(p-2)}(X^j)^4] < \infty$ is a natural extension of the fourth moment assumption in Theorem 2. In fact, all three assumptions in Theorem 5 reduce to those of Theorem 2 as $p \to 2$. It is worth noting that the constant c_p has the alternative expression $\sup_{w \in \mathbb{R}^d \setminus \{0\}}\{\|w\|_{L^p}/\|w\|_{H_p}\}$ by (2.3), i.e. it is the norm equivalence constant between the L^p norm and the norm induced by H_p. Using again (2.3), we see that $c_p \to 1$ as $p \to 2$. As $p \to \infty$, c_p grows, and we suspect in a dimension dependent way. However, this does not affect the asymptotic optimality of our rate as c_p only enters an $o(1/n)$ term in our bound.

We now turn to the case of $p \in (1, 2)$. Here we will need a slightly stronger version of non-realizability to ensure that the risk is twice differentiable at its minimizer. Our main result follows.

Theorem 6. *Let $p \in (1, 2)$ and $\delta \in (0, 1]$. Assume that $\mathrm{P}(|\langle w_p^*, X\rangle - Y|^{2-p} > 0) = 1$ and $\mathrm{E}[|\langle w_p^*, X\rangle - Y|^{2(p-2)}] < \infty$. Further, assume that $\mathrm{E}[|Y|^p] < \infty$, $\mathrm{E}[(X^j)^2] < \infty$, $\mathrm{E}[|\langle w_p^*, X\rangle - Y|^{2(p-2)}(X^j)^4] < \infty$ for all $j \in [d]$. If*

$$n \geq 196\sigma_p^2(d + 2\log(4/\delta)),$$

then, with probability at least $1 - \delta$

$$R_p(\hat{w}_p) - R_p(w_p^*) \leq \frac{8192}{p-1}\frac{\mathrm{E}\Big[\|\nabla \ell_p(\langle w_p^*, X\rangle - Y)\|_{H_p^{-1}}^2\Big]}{n\delta} + \frac{1}{p-1}\left(\frac{524288\,\mathrm{E}\Big[\|\nabla \ell_p(\langle w_p^*, X\rangle - Y)\|_{H_p^{-1}}^2\Big]\sigma_p^{6-2p}d^{(2-p)}c_p^{2-p}c_p^*}{n\delta}\right)^{1/(p-1)}$$

where we used c_p^* to denote $\mathrm{E}[|Y - \langle w_p^*, X\rangle|^{2(p-2)}]$ and c_p to denote $C_{L^2 \to (L^2, p)}^2$.

Just as in Theorems 2 and 5, Theorem 6 is asymptotically optimal up to a constant factor that depends on p. Indeed, since $1 < p < 2$, $1/(p-1) > 1$, and the second term is $o(1/n)$. From the point of view of assumptions, we have two additional assumptions compared to Theorem 5. First, we require a stronger version of non-realizability by assuming $\langle w_p^*, X\rangle \neq Y$ almost surely. This assumption is necessary to prove the twice differentiability of the risk at its minimizer using the standard result that allows the exchange of differentiation and expectation, see e.g. Theorem 2.27 in [Fol13]. It is also known that for the case $p \in [1, 2)$, there are situations where the asymptotic bound (1.1) does not hold, as the limiting distribution of the coefficients $\hat{w}_p$ as $n \to \infty$ does not necessarily converge to a Gaussian, and depends heavily on the distribution of $\langle w_p^*, X\rangle - Y$, see e.g. Lai and Lee [LL05] and Knight [Kni98]. Overall, we suspect that perhaps a slightly weaker version of our assumptions is necessary for a fast rate like (1.1) to hold.

The second additional assumption we require is the existence of the $2(2-p)$ negative moment of $|\langle w_p^*, X\rangle - Y|$. In the majority of applications, one adds an intercept to the original covariates, so that this negative moment assumption is already implied by the standard assumption $\mathrm{E}[|Y - \langle w_p^*, X\rangle|^{2(p-2)}(X^j)^4] < \infty$. In the rare case where an intercept is not included, any negative moment assumption on $|\langle w_p^*, X\rangle - Y|$ can be used instead, at the cost of a larger factor in the $o(1/n)$ term. Finally, similar to how the constant c_p of Theorem 5 deteriorates as $p \to \infty$, the constant c_p^* of Theorem 6 gets worse as $p \to 1$. It is unclear to us however if it acquires dimension dependence.

3 Proofs

3.1 Proof of Theorem 2

Here we give a detailed proof of Theorem 2. While the core technical result can be deduced by combining results from [Oli16] and [LM16], here we frame the proof in a way that makes it easy to extend to the cases $p \in (1, \infty)$, and differently from either paper. We split the proof in three steps. First notice that as the loss is a quadratic function of w, we can express it exactly using a second order Taylor expansion around the minimizer w_2^*

$$\ell_2(\langle w, X\rangle - Y) - \ell_2(\langle w_2^*, X\rangle - Y) = \langle \nabla \ell_2(\langle w_2^*, X\rangle - Y), w - w_2^*\rangle + \frac{1}{2}\|w - w_2^*\|^2_{\nabla^2 \ell_2(\langle w_2^*, X\rangle - Y)}.$$

Taking empirical averages and expectations of both sides respectively shows that the excess empirical risk and excess risk also admit such a decomposition

$$\begin{aligned} R_{2,n}(w) - R_{2,n}(w_2^*) &= \langle \nabla R_{2,n}(w_2^*), w - w_2^*\rangle + \frac{1}{2}\|w - w_2^*\|^2_{H_{2,n}}, \\ R_2(w) - R_2(w_2^*) &= \frac{1}{2}\|w - w_2^*\|^2_{H_2}, \end{aligned} \tag{3.1}$$

where in the second equality we used that the gradient of the risk vanishes at the minimizer w_2^*. Therefore, to bound the excess risk, it is sufficient to bound the norm $\|w - w_2^*\|_{H_2}$. This is the goal of the second step, where we use two ideas. First, by definition, the excess empirical risk of the empirical risk minimizer satisfies the upper bound

$$R_{2,n}(\hat{w}_2) - R_{2,n}(w_2^*) \leq 0. \tag{3.2}$$

Second, we use the Cauchy-Schwartz inequality to lower bound the excess empirical risk by

$$R_{2,n}(\hat{w}_2) - R_{2,n}(w_2^*) \geq -\|\nabla R_{2,n}(w_2^*)\|_{H_2^{-1}} \|\hat{w}_2 - w_2^*\|_{H_2} + \frac{1}{2}\|\hat{w}_2 - w_2^*\|^2_{H_{2,n}}, \tag{3.3}$$

and we further lower bound it by deriving high probability bounds on the two random terms $\|\nabla R_{2,n}(w_2^*)\|_{H_2^{-1}}$ and $\|\hat{w}_2 - w_2^*\|^2_{H_{2,n}}$. The first can easily be bounded using Chebyshev's inequality and the elementary fact that the variance of the average of n i.i.d. random variables is $1/n$ the variance of the original random variable. Here we state a slightly more general result. The straightforward proof is relegated to the Appendix.

Lemma 7. *Let $p \in (1, \infty)$. If $p \in (1, 2)$, let the assumptions of Theorem 6 hold. Then with probability at least $1 - \delta/2$*

$$\|\nabla R_{p,n}(w_p^*)\|_{H_p^{-1}} \leq \sqrt{2\,\mathrm{E}\left[\|\nabla \ell_p(\langle w_p^*, X\rangle - Y)\|^2_{H_p^{-1}}\right]/(n\delta)}.$$

For the second random term $\|\hat{w}_2 - w_2^*\|^2_{H_{2,n}}$, we use Theorem 3.1 of Oliveira [Oli16], which we restate here, emphasizing that the existence of fourth moments of the components of the random vector is enough to ensure the existence of the needed norm equivalence constant.

Proposition 8 (Theorem 3.1, Oliveira [Oli16])**.** *Let $Z \in \mathbb{R}^d$ be a random vector satisfying $\mathrm{E}[Z_j^4] < \infty$ for all $j \in [d]$ and assume that $\mathrm{P}(\langle v, Z\rangle = 0) = 1$ only if $v = 0$. For $p \in [1, \infty)$ and $v \in \mathbb{R}^d$, define*

$$\|v\|_{L^p} := \mathrm{E}[(\langle v, Z\rangle)^p]^{1/p}, \qquad \sigma^2 := \left(\sup_{v \in \mathbb{R}^d \setminus \{0\}} \|v\|_{L^4}/\|v\|_{L^2}\right)^4.$$

Let $(Z_i)_{i=1}^n$ be i.i.d. samples of Z. Then, with probability at least $1-\delta$, for all $v \in \mathbb{R}^d$,

$$\frac{1}{n}\sum_{i=1}^n \langle v, Z_i\rangle^2 \geq \left(1 - 7\sigma\sqrt{\frac{d + 2\log(2/\delta)}{n}}\right)\|v\|_{L^2}^2.$$

Using this result we can immediately deduce the required high-probability lower bound on the second random term $\|\hat{w}_2 - w_2^*\|_{H_{2,n}}^2$, we leave the proof to the Appendix.

Corollary 9. *Under the assumptions of Theorem 2, if $n \geq 196\sigma_2^2(d + 2\log(4/\delta))$, then with probability at least $1-\delta/2$, for all $w \in \mathbb{R}^d$,*

$$\|w - w_2^*\|_{H_{2,n}}^2 \geq \frac{1}{2}\|w - w_2^*\|_{H_2}^2.$$

Combining Lemma 7, Corollary 9, and (3.3) yields that with probability at least $1-\delta$

$$R_{2,n}(\hat{w}_2) - R_{2,n}(w_2^*) \geq -\sqrt{2\,\mathrm{E}\left[\|\nabla \ell_p(\langle w_p^*, X\rangle - Y)\|_{H_p^{-1}}^2\right]/(n\delta)}\,\|\hat{w}_2 - w_2^*\|_{H_2} + \frac{1}{4}\|\hat{w}_2 - w_2^*\|_{H_2}^2. \tag{3.4}$$

The final step is to combine the upper bound (3.2) and the lower bound (3.4). This gives that with probability at least $1-\delta$

$$\|\hat{w}_2 - w_2^*\|_{H_2} \leq 4\sqrt{2\,\mathrm{E}\left[\|\nabla \ell_p(\langle w_p^*, X\rangle - Y)\|_{H_p^{-1}}^2\right]/(n\delta)}.$$

Replacing in (3.1) finishes the proof. □

3.2 Proof of Theorem 5

The main challenge in moving from the case $p = 2$ to the case $p \in (2, \infty)$ is that the second order Taylor expansion of the loss is no longer exact. The standard way to deal with this problem is to assume that the loss is upper and lower bounded by quadratic functions, i.e. that it is smooth and strongly convex. Unfortunately, as discussed in Section 1, the ℓ_p loss is not strongly convex for any $p > 2$, so we need to find another way to deal with this issue. Once this has been resolved however, the strategy we used in the proof of Theorem 2 can be applied almost verbatim to yield the result. Remarkably, a result of [Adi+22] allows us to upper and lower bound the p-th power loss for $p \in (2, \infty)$ by its second order Taylor expansion around a point, up to some residual terms. An application of this result yields the following Lemma.

Lemma 10. *Let $p \in [2, \infty)$. Then:*

$$R_{p,n}(w) - R_{p,n}(w_p^*) \geq \frac{1}{8(p-1)}\|w - w_p^*\|_{H_{p,n}}^2 + \langle \nabla R_{p,n}(w_p^*), w - w_p^*\rangle, \tag{3.5}$$

$$R_p(w) - R_p(w_p^*) \leq \frac{2p}{(p-1)}\|w - w_p^*\|_{H_p}^2 + p^p\|w - w_p^*\|_{L^p}^p. \tag{3.6}$$

Up to constant factors that depend only on p and an L^p norm residual term, Lemma 10 gives matching upper and lower bounds on the excess risk and excess empirical risks in terms of their second order Taylor expansions around the minimizer. We can thus use the approach taken in the proof of Theorem 2 to derive our result. The only additional challenge is the control of the term $\|\hat{w}_p - w_p^*\|_{L^p}$, which we achieve by reducing it to an $\|\hat{w}_p - w_p^*\|_{H_p}$ term using norm equivalence. We leave the details to the Appendix.

3.3 Proof of Theorem 6

The most technically challenging case is when $p \in (1, 2)$. Indeed as seen in the proof of Theorem 2, the most involved step is lower bounding the excess empirical risk with high probability. For the case $p \in [2, \infty)$, we achieved this by having access to a pointwise quadratic lower bound, which is not too surprising. Indeed, at small scales, we expect the second order Taylor expansion to be accurate, while at large scales, we expect the p-th power loss to grow at least quadratically for $p \in [2, \infty)$.

In the case of $p \in (1, 2)$, we are faced with a harder problem. Indeed, as $p \to 1$, the ℓ_p losses behave almost linearly at large scales. This means that we cannot expect to obtain a global quadratic lower bound as for the case $p \in [2, \infty)$, so we will need a different proof technique. Motivated by closely related concerns, Bubeck et al. [Bub+18] introduced the following approximation to the p-th power function

$$\gamma_p(t, x) := \begin{cases} \dfrac{p}{2} t^{p-2} x^2 & \text{if} \quad x \leq t \\ x^p - \left(1 - \dfrac{p}{2}\right) t^p & \text{if} \quad x > t, \end{cases}$$

for $t, x \in [0, \infty)$ and with $\gamma_p(0, 0) = 0$. This function was further studied by [Adi+19], whose results we use to derive the following Lemma.

Lemma 11. *Let $p \in (1, 2)$. Under the assumptions of Theorem 6, we have*

$$R_{p,n}(w) - R_{p,n}(w_p^*) \geq \frac{1}{4p^2} \frac{1}{n} \sum_{i=1}^{n} \gamma_p\big(|\langle w_p^*, X_i \rangle - Y_i|, |\langle w - w_p^*, X_i \rangle|\big) + \langle \nabla R_{p,n}(w_p^*), w - w_p^* \rangle, \tag{3.7}$$

$$R_p(w) - R_p(w_p^*) \leq \frac{4}{(p-1)} \|w - w_p^*\|_{H_p}^2. \tag{3.8}$$

As expected, while we do have the desired quadratic upper bound, the lower bound is much more cumbersome, and is only comparable to the second order Taylor expansion when $|\langle w - w_p^*, X_i \rangle| \leq |\langle w_p^*, X_i \rangle - Y_i|$. What we need for the proof to go through is a high probability lower bound of order $\Omega(\|w - w^*\|_{H_p}^2)$ on the first term in the lower bound (3.7). We obtain this in the following result. The rest of the proof of Theorem 6 is left to the Appendix.

Proposition 12. *Let $\delta \in (0, 1]$. Under the assumptions of Theorem 6, if $n \geq 196\sigma_p^2(d + 2\log(4/\delta))$, then with probability at least $1 - \delta/2$, for all $w \in \mathbb{R}^d$,*

$$\frac{1}{n} \sum_{i=1}^{n} \gamma_p\big(|\langle w_p^*, X_i \rangle - Y_i|, |\langle w - w_p^*, X_i \rangle|\big) \geq \frac{1}{8} \min\Big\{\|w - w_p^*\|_{H_p}^2, \varepsilon^{2-p}\|w - w_p^*\|_{H_p}^p\Big\},$$

where $\varepsilon^{p-2} := 8\sigma_p^{3-p}(dc_p)^{(2-p)/2}\sqrt{c_p^*}$, and c_p and c_p^* are as defined in Theorem 6.

Proof. Let $\varepsilon > 0$ and let $T \in (0, \infty)$ be a truncation parameter we will set later. Define the truncated vector

$$\tilde{X} := X \cdot \mathbb{1}_{[0,T]}(\|X\|_{H_p^{-1}}),$$

and the constant $\beta := T\varepsilon$. By Lemma 3.3 in [Adi+19], we have that $\gamma_p(t, \lambda x) \geq \min\{\lambda^2, \lambda^p\}\gamma_p(t, x)$ for all $\lambda \geq 0$. Furthermore, it is straightforward to verify that $\gamma_p(t, x)$ is decreasing in t and increasing in x. Therefore, we have, for any $w \in \mathbb{R}^d$,

$$\begin{aligned} &\frac{1}{n} \sum_{i=1}^{n} \gamma_p\big(|\langle w_p^*, X_i \rangle - Y_i|, |\langle w - w_p^*, X_i \rangle|\big) \\ &\geq \min\Big\{\varepsilon^{-2}\|w - w_p\|_{H_p}^2, \varepsilon^{-p}\|w - w_p\|_{H_p}^p\Big\} \frac{1}{n} \sum_{i=1}^{n} \gamma_p\left(|\langle w_p^*, X_i \rangle - Y_i|, \left|\left\langle \frac{\varepsilon(w - w_p^*)}{\|w - w_p^*\|_{H_p}}, X_i \right\rangle\right|\right) \\ &\geq \min\Big\{\varepsilon^{-2}\|w - w_p\|_{H_p}^2, \varepsilon^{-p}\|w - w_p\|_{H_p}^p\Big\} \cdot \inf_{\|w\|_{H_p} = \varepsilon} \frac{1}{n} \sum_{i=1}^{n} \gamma_p\big(|\langle w_p^*, X_i \rangle - Y_i|, |\langle w, X_i \rangle|\big). \end{aligned} \tag{3.9}$$

The key idea to control this last infimum is to truncate $\langle w, X_i \rangle$ from above by using the truncated vector $\tilde{X}$, and $|\langle w_p^*, X_i\rangle - Y_i|$ from below by forcing it to be greater than β. By the monotonicity properties of γ_p discussed above, we get that the infimum in (3.9) is lower bounded by

$$\begin{aligned}&\inf_{\|w\|_{H_p}=\varepsilon} \frac{1}{n}\sum_{i=1}^{n} \gamma_p(\max\{|\langle w_p^*, X_i\rangle - Y_i|, \beta\}, |\langle w, \tilde{X}_i\rangle|) \\ &= \frac{\varepsilon^2 p}{2} \inf_{\|w\|_{H_p}=1} \frac{1}{n}\sum_{i=1}^{n} \max\{|\langle w_p^*, X_i\rangle - Y_i|, \beta\}^{p-2} |\langle w, \tilde{X}_i\rangle|^2,\end{aligned}$$

where the equality follows by the fact that with the chosen truncations, the second argument of γ_p is less than or equal to the first. Define the random vector

$$Z = \max\{|\langle w_p^*, X\rangle - Y|, \beta\}^{(p-2)/2} \tilde{X}.$$

Then, by removing the truncations, we see that the components of Z have finite fourth moments by assumption. Using Proposition 8, and under our constraint on n, we get that with probability at least $1 - \delta/2$,

$$\begin{aligned}&\inf_{\|w\|_{H_p}=1} \frac{1}{n}\sum_{i=1}^{n} \max\{|\langle w_p^*, X_i\rangle - Y_i|, \beta\}^{p-2} |\langle w, \tilde{X}_i\rangle|^2 = \inf_{\|w\|_{H_p}=1} \frac{1}{n}\sum_{i=1}^{n} \langle w, Z_i\rangle^2 \\ &\geq \frac{1}{2} \inf_{\|w\|_{H_p}=1} \mathrm{E}\left[\max\{|\langle w_p^*, X\rangle - Y|, \beta\}^{(p-2)} \langle w, \tilde{X}\rangle^2\right] \\ &\geq \frac{1}{2}\left(1 - \sup_{\|w\|_{H_p}=1} \mathrm{E}\left[|\langle w_p^*, X\rangle - Y|^{p-2}\langle w, X\rangle^2 \left(\mathbb{1}_{[0,\beta)}(|\langle w_p^*, X\rangle - Y|) + \mathbb{1}_{(T,\infty)}(\|X\|_{H_p^{-1}})\right)\right]\right)\end{aligned}$$

Finally, we make use of Holder's inequality and our moment assumptions to bound this last supremum. Optimizing over the choice of T and ε yields the result. Details of this last step are in the Appendix. □

References

[AC11] J.-Y. Audibert and O. Catoni. "Robust Linear Least Squares Regression". In: *The Annals of Statistics* 39.5 (Oct. 2011), pp. 2766–2794. ISSN: 0090-5364, 2168-8966. DOI: 10.1214/11-AOS918. (Visited on 01/17/2023).

[Adi+19] D. Adil et al. "Iterative Refinement for P-Norm Regression". In: *Proceedings of the 2019 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*. Proceedings. Society for Industrial and Applied Mathematics, Jan. 2019, pp. 1405–1424. DOI: 10.1137/1.9781611975482.86. (Visited on 05/17/2023).

[Adi+22] D. Adil et al. *Fast Algorithms for ℓ_p-Regression*. Nov. 2022. DOI: 10.48550/arXiv.2211.03963. arXiv: 2211.03963 [cs, math]. (Visited on 05/17/2023).

[APS19] D. Adil, R. Peng, and S. Sachdeva. "Fast, Provably Convergent IRLS Algorithm for p-Norm Linear Regression". In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc., 2019. (Visited on 05/17/2023).

[Arc96] M. A. Arcones. "The Bahadur-Kiefer Representation of Lp Regression Estimators". In: *Econometric Theory* 12.2 (1996), pp. 257–283. ISSN: 0266-4666. JSTOR: 3532831. (Visited on 07/10/2023).

[BRW92] Z. D. Bai, C. R. Rao, and Y. Wu. "M-Estimation of Multivariate Linear Regression Parameters Under a Convex Discrepancy Function". In: *Statistica Sinica* 2.1 (1992), pp. 237–254. ISSN: 1017-0405. JSTOR: 24304129. (Visited on 07/10/2023).

[Bub+18] S. Bubeck et al. "An Homotopy Method for Lp Regression Provably beyond Self-Concordance and in Input-Sparsity Time". In: *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*. STOC 2018. New York, NY, USA: Association for Computing Machinery, June 2018, pp. 1130–1137. ISBN: 978-1-4503-5559-9. DOI: 10.1145/3188745.3188776. (Visited on 05/17/2023).

[Fol13] G. B. Folland. *Real Analysis: Modern Techniques and Their Applications*. John Wiley & Sons, June 2013. ISBN: 978-1-118-62639-9.

[HKZ12] D. Hsu, S. M. Kakade, and T. Zhang. "Random Design Analysis of Ridge Regression". In: *Proceedings of the 25th Annual Conference on Learning Theory*. JMLR Workshop and Conference Proceedings, June 2012, pp. 9.1–9.24. (Visited on 01/17/2023).

[HS96] X. He and Q.-M. Shao. "A General Bahadur Representation of M-estimators and Its Application to Linear Regression with Nonstochastic Designs". In: *The Annals of Statistics* 24.6 (Dec. 1996), pp. 2608–2630. ISSN: 0090-5364, 2168-8966. DOI: 10.1214/aos/1032181172. (Visited on 07/10/2023).

[JLS22] A. Jambulapati, Y. P. Liu, and A. Sidford. "Improved Iteration Complexities for Overconstrained P-Norm Regression". In: *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*. STOC 2022. New York, NY, USA: Association for Computing Machinery, June 2022, pp. 529–542. ISBN: 978-1-4503-9264-8. DOI: 10.1145/3519935.3519971. (Visited on 05/24/2023).

[KM15] V. Koltchinskii and S. Mendelson. "Bounding the Smallest Singular Value of a Random Matrix Without Concentration". In: *International Mathematics Research Notices* 2015.23 (2015), pp. 12991–13008. ISSN: 1687-0247. DOI: 10.1093/imrn/rnv096.

[Kni98] K. Knight. "Limiting Distributions for $L\sb 1$ Regression Estimators under General Conditions". In: *The Annals of Statistics* 26.2 (Apr. 1998), pp. 755–770. ISSN: 0090-5364, 2168-8966. DOI: 10.1214/aos/1028144858. (Visited on 05/17/2023).

[LC06] E. L. Lehmann and G. Casella. *Theory of Point Estimation*. Springer Science & Business Media, May 2006. ISBN: 978-0-387-22728-3.

[LL05] P. Y. Lai and S. M. S. Lee. "An Overview of Asymptotic Properties of Lp Regression under General Classes of Error Distributions". In: *Journal of the American Statistical Association* 100.470 (2005), pp. 446–458. ISSN: 0162-1459. JSTOR: 27590567. (Visited on 05/16/2023).

[LM16] G. Lecué and S. Mendelson. "Performance of Empirical Risk Minimization in Linear Aggregation". In: *Bernoulli* 22.3 (Aug. 2016), pp. 1520–1534. ISSN: 1350-7265. DOI: 10.3150/15-BEJ701. (Visited on 01/17/2023).

[LM17a] G. Lecué and S. Mendelson. "Regularization and the Small-Ball Method II: Complexity Dependent Error Rates". In: *Journal of Machine Learning Research* 18.146 (2017), pp. 1–48. ISSN: 1533-7928. (Visited on 05/04/2023).

[LM17b] G. Lecué and S. Mendelson. "Sparse Recovery under Weak Moment Assumptions". In: *Journal of the European Mathematical Society* 19.3 (Feb. 2017), pp. 881–904. ISSN: 1435-9855. DOI: 10.4171/jems/682. (Visited on 05/04/2023).

[LM18] G. Lecué and S. Mendelson. "Regularization and the Small-Ball Method I: Sparse Recovery". In: *The Annals of Statistics* 46.2 (Apr. 2018), pp. 611–641. ISSN: 0090-5364, 2168-8966. DOI: 10.1214/17-AOS1562. (Visited on 05/02/2023).

[Men14] S. Mendelson. "Learning without Concentration". In: *Proceedings of The 27th Conference on Learning Theory*. PMLR, May 2014, pp. 25–39. (Visited on 01/17/2023).

[Men18] S. Mendelson. "Learning without Concentration for General Loss Functions". In: *Probability Theory and Related Fields* 171.1 (June 2018), pp. 459–502. ISSN: 1432-2064. DOI: 10.1007/s00440-017-0784-y. (Visited on 04/18/2023).

[MG22] J. Mourtada and S. Gaïffas. "An Improper Estimator with Optimal Excess Risk in Misspecified Density Estimation and Logistic Regression". In: *Journal of Machine Learning Research* 23.31 (2022), pp. 1–49. ISSN: 1533-7928. (Visited on 05/16/2023).

[Mou22] J. Mourtada. "Exact Minimax Risk for Linear Least Squares, and the Lower Tail of Sample Covariance Matrices". In: *The Annals of Statistics* 50.4 (Aug. 2022), pp. 2157–2178. ISSN: 0090-5364, 2168-8966. DOI: 10.1214/22-AOS2181. (Visited on 01/17/2023).

[Nie92] W. Niemiro. "Asymptotics for M-Estimators Defined by Convex Minimization". In: *The Annals of Statistics* 20.3 (Sept. 1992), pp. 1514–1533. ISSN: 0090-5364, 2168-8966. DOI: 10.1214/aos/1176348782. (Visited on 07/10/2023).

[OB21] D. M. Ostrovskii and F. Bach. “Finite-Sample Analysis of M-Estimators Using Self-Concordance”. In: *Electronic Journal of Statistics* 15.1 (Jan. 2021), pp. 326–391. ISSN: 1935-7524, 1935-7524. DOI: 10.1214/20-EJS1780. (Visited on 05/16/2023).

[Oli16] R. I. Oliveira. “The Lower Tail of Random Quadratic Forms with Applications to Ordinary Least Squares”. In: *Probability Theory and Related Fields* 166.3 (Dec. 2016), pp. 1175–1194. ISSN: 1432-2064. DOI: 10.1007/s00440-016-0738-9. (Visited on 01/17/2023).

[Ron84] A. E. Ronner. “Asymptotic Normality of P-Norm Estimators in Multiple Regression”. In: *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* 66.4 (Sept. 1984), pp. 613–620. ISSN: 1432-2064. DOI: 10.1007/BF00531893. (Visited on 05/16/2023).

[RV15] M. Rudelson and R. Vershynin. “Small Ball Probabilities for Linear Images of High-Dimensional Distributions”. In: *International Mathematics Research Notices* 2015.19 (2015), pp. 9594–9617. ISSN: 1687-0247. DOI: 10.1093/imrn/rnu243.

[Sau18] A. Saumard. “On Optimality of Empirical Risk Minimization in Linear Aggregation”. In: *Bernoulli* 24.3 (Aug. 2018), pp. 2176–2203. ISSN: 1350-7265. DOI: 10.3150/17-BEJ925. (Visited on 04/18/2023).

[Vaa98] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press, 1998. ISBN: 978-0-521-78450-4. DOI: 10.1017/CBO9780511802256. (Visited on 05/15/2023).

[VW96] A. W. Van Der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer Series in Statistics. New York, NY: Springer, 1996. ISBN: 978-1-4757-2547-6 978-1-4757-2545-2. DOI: 10.1007/978-1-4757-2545-2. (Visited on 07/10/2023).

[Whi82] H. White. “Maximum Likelihood Estimation of Misspecified Models”. In: *Econometrica* 50.1 (1982), pp. 1–25. ISSN: 0012-9682. DOI: 10.2307/1912526. JSTOR: 1912526. (Visited on 05/15/2023).

A Preliminary results

In this section we provide the proof of some basic results we used in the main paper.

A.1 Proof of Lemma 7

Proof. We compute the expectation:

$$\begin{aligned}
\mathrm{E}\Big[\|\nabla R_{p,n}(w_p^*)\|_{H_p^{-1}}^2\Big] &= \mathrm{E}\Big[\|n^{-1}\nabla \ell_p(\langle w_p^*, X_i\rangle - Y_i)\|_{H_p^{-1}}^2\Big] \\
&= n^{-2}\sum_{i=1}^{n} \mathrm{E}\Big[\|\nabla \ell_p(\langle w_p^*, X_i\rangle - Y_i)\|_{H_p^{-1}}^2\Big] \\
&\quad + 2n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{i-1}\langle \mathrm{E}[\nabla \ell_p(\langle w_p^*, X_i\rangle - Y_i)], \mathrm{E}[\nabla \ell_p(\langle w_p^*, X_j\rangle - Y_j)]\rangle_{H_p^{-1}} \\
&= n^{-1}\,\mathrm{E}\Big[\|\nabla \ell_p(\langle w_p^*, X\rangle - Y)\|_{H_p^{-1}}^2\Big]
\end{aligned}$$

where in the second line we expanded the inner product of the sums into its n^2 terms, used linearity of expectation, and used the independence of the samples to take the expectation inside the inner product. In the last line, we used the fact that the samples are identically distributed to simplify the first term. For the second term, we used the fact that the expectation of the gradient of the loss at the risk minimizer vanishes. Applying Markov's inequality finishes the proof. □

A.2 Proof of Corollary 9

Proof. We have

$$\begin{aligned}
\|w - w_2^*\|_{H_{2,n}}^2 &= (w - w_2^*)^T H_{2,n}(w - w_2^*) \\
&= \frac{1}{n}\sum_{i=1}^{n}(w - w_2^*)^T \nabla^2 \ell_p(\langle w_2^*, X_i\rangle - Y_i)(w - w_2^*) \\
&= \frac{1}{n}\sum_{i=1}^{n}\langle w - w_2^*, X_i\rangle^2.
\end{aligned}$$

Now by assumption, the components of the vector X have finite fourth moment so that applying Proposition 8 and using the condition on n yields the result. □

A.3 Proof of Lemma 10.

Proof. By Lemma 2.5 in [Adi+22], we have for all $t, s \in \mathbb{R}$

$$\ell_p(t) - \ell_p(s) - \ell_p'(s)(t - s) \geq \frac{1}{8(p-1)}\ell_p''(s)(t - s)^2.$$

Recall that by the chain rule

$$\nabla \ell_p(\langle w, X\rangle - Y) = \ell_p'(\langle w, X\rangle - Y)X \qquad \nabla^2 \ell_p(\langle w, X\rangle - Y) = \ell_p''(\langle w, X\rangle - Y)XX^T.$$

Replacing t and s by $\langle w, X_i\rangle - Y_i$ and $\langle w_p^*, X_i\rangle - Y_i$ respectively, and using the formulas for the gradient and Hessian we arrive at

$$\begin{aligned}
\ell_p(\langle w, X_i\rangle - Y_i) - \ell_p(\langle w_p^*, X_i\rangle - Y_i) \geq \frac{1}{8(p-1)}&(w - w_p^*)^T \nabla^2 \ell_p(\langle w_p^*, X_i\rangle - Y_i)(w - w_p^*) \\
&+ \langle \nabla \ell_p(\langle w_p^*, X_i\rangle - Y_i), w - w_p^*\rangle
\end{aligned}$$

Averaging over $i \in [n]$ yields the first inequality. The proof of the second inequality proceeds in the same way and uses instead the upper bound of Lemma 2.5 in [Adi+22]. We omit it here. □

A.4 Proof of Lemma 11

Proof. Both inequalities follow from Lemma 4.5 in Adil et al. [Adi+19]. (3.7) follows from a straightforward calculation using the lower bound of Lemma 4.5 in Adil et al. [Adi+19]; we omit it here. The upper bound requires a bit more work. We have by the quoted Lemma

$$\ell_p(t) - \ell_p(s) - \ell_p'(s)(t-s) \leq \frac{4}{p(p-1)}\gamma_p(|s|, |t-s|).$$

Now assume that $|s| > 0$. If $|t-s| \leq |s|$, we have

$$\gamma_p(|s|, |t-s|) = \frac{p}{2}|s|^{p-2}(t-s)^2 \leq |s|^{p-2}(t-s)^2 = \ell_p''(s)(t-s)^2.$$

Otherwise, if $|t-s| > |s|$, then we have

$$\gamma_p(|s|, |t-s|) = |t-s|^p - (1-p/2)|s|^p \leq (t-s)^2|t-s|^{p-2} \leq |s|^{p-2}(t-s)^2 = \ell_p''(s)(t-s)^2.$$

Therefore in both cases we have $\gamma_p(|s|, |t-s|) \leq \ell_p''(s)(t-s)^2$ as long as $|s| > 0$. Replacing t and s by $\langle w, X\rangle - Y$ and $\langle w_p^*, X\rangle - Y$ respectively we get, on the event that $\langle w_p^*, X\rangle - Y \neq 0$

$$\ell_p(\langle w, X\rangle - Y) - \ell_p(\langle w_p^*, X\rangle - Y) - \langle \nabla \ell_p(\langle w_p^*, X\rangle - Y), w - w_p^*\rangle \leq \frac{4}{p(p-1)}\|w - w_p^*\|_{\nabla^2 \ell_p(\langle w_p^*, X\rangle - Y)}$$

Recalling that by assumption $\mathrm{P}\big(\langle w_p^*, X\rangle - Y \neq 0\big) = 1$, taking expectation of both sides, and bounding $1/p \leq 1$ finishes the proof of (3.8). □

B Differentiability of the risk

In this section, we study the differentiability properties of the risk. We start by showing that under a subset of our assumptions, the risk is differentiable everywhere on $\mathbb{R}^d$.

Lemma 13. *Let $p \in (1, \infty)$ and assume that $\mathrm{E}[|Y|^p] < \infty$ and $\mathrm{E}[|X_j|^p] < \infty$ for all $j \in [d]$. Then R_p is differentiable on $\mathbb{R}^d$, and*

$$\nabla R_p(w) = \mathrm{E}[\nabla \ell_p(\langle w, X\rangle - Y)].$$

Proof. Let $w \in \mathbb{R}^d$. We want to show that

$$\lim_{\Delta \to 0} \frac{|R_p(w + \Delta) - R_p(w) - \mathrm{E}[\langle \nabla \ell_p(\langle w, X\rangle - Y), \Delta\rangle]|}{\|\Delta\|} = 0,$$

where, for convenience, we take the norm $\|\cdot\|$ to be the Euclidean norm. Define the function $\phi(w, X, Y) := \ell_p(\langle w, X\rangle - Y)$ and note that by the chain rule ϕ is differentiable as a function of w on all of $\mathbb{R}^d$. Now let $(\Delta_k)_{k=1}^n$ be a sequence in $\mathbb{R}^d$ such that $\lim_{k\to\infty}\|\Delta_k\| = 0$. Then

$$\begin{aligned}
&\lim_{k\to\infty} \frac{|R_p(w + \Delta_k) - R_p(w) - \mathrm{E}[\langle \nabla \phi(w, X, Y), \Delta_k\rangle]|}{\|\Delta_k\|} \\
&= \lim_{k\to\infty} \frac{|\mathrm{E}[\phi(w + \Delta_k, X, Y) - \phi(w, X, Y) - \langle \nabla \phi(w, X, Y), \Delta_k\rangle]|}{\|\Delta_k\|} \\
&\leq \lim_{k\to\infty} \mathrm{E}\left[\frac{|\phi(w + \Delta_k, X, Y) - \phi(w, X, Y) - \langle \nabla \phi(w, X, Y), \Delta_k\rangle|}{\|\Delta_k\|}\right].
\end{aligned} \tag{B.1}$$

Our goal is to interchange the limit and expectation. For that, we will use the dominated convergence theorem. We construct our dominating function as follows. Let $R := \sup_{k\in\mathbb{N}}\|\Delta_k\|$, and note that $R < \infty$

since $\|\Delta_k\| \to 0$ as $k \to \infty$. Then we have

$$\begin{aligned}
&\frac{|\phi(w+\Delta_k, X, Y) - \phi(w, X, Y) - \langle \nabla\phi(w, X, Y), \Delta_k\rangle|}{\|\Delta_k\|} \\
&\leq \frac{|\phi(w+\Delta_k, X, Y) - \phi(w, X, Y)|}{\|\Delta_k\|} + \frac{|\langle \nabla\phi(w, X, Y), \Delta_k\rangle|}{\|\Delta_k\|} \\
&\leq \frac{\left\langle \int_0^1 \nabla\phi(w + t\Delta_k, X, Y)dt, \Delta_k \right\rangle}{\|\Delta_k\|} + \|\nabla\phi(w, X, Y)\| \\
&\leq \left\| \int_0^1 \nabla\phi(w + t\Delta_k, X, Y)dt \right\| + \|\nabla\phi(w, X, Y)\| \\
&\leq \int_0^1 \|\nabla\phi(w + t\Delta_k, X, Y)\|dt + \|\nabla\phi(w, X, Y)\| \\
&\leq 2 \sup_{\Delta \in B(0,R)} \|\nabla\phi(w + \Delta, X, Y)\| \\
&\leq \frac{2}{p-1}\|X\| \sup_{\Delta \in B(0,R)} |\langle w + \Delta, X\rangle - Y|^{p-1} \\
&\leq \frac{2}{p-1}\|X\| \sup_{\Delta \in B(0,R)} \max\{2^{p-1}, 1\}\Big(|\langle w, X\rangle - Y|^{p-1} + |\langle \Delta, X\rangle|^{p-1}\Big) \\
&= \frac{2^p}{p-1}\Big\{|\langle w, X\rangle - Y|^{p-1}\|X\| + R^{p-1}\|X\|^p\Big\} =: g(X, Y),
\end{aligned}$$

where the second line follows by triangle inequality, the third from the fundamental theorem of calculus applied component-wise, the fourth by Cauchy-Schwartz inequality, the fifth by Jensen's inequality and the convexity of the norm, and the eighth by the inequality $|a + b|^q \leq \max\{2^{q-1}, 1\}(|a|^q + |b|^q)$ valid for $q > 0$. It remains to show that $g(X, Y)$ is integrable. We have

$$\begin{aligned}
\mathrm{E}[g(X, Y)] &= \frac{2^p}{p-1}\,\mathrm{E}\Big[|\langle w, X\rangle - Y|^{p-1}\|X\| + R^{p-1}\|X\|^p\Big] \\
&= \frac{2^p}{p-1}\left\{\sum_{j=1}^d \mathrm{E}\Big[|\langle w, X\rangle - Y|^{p-1}|X^j|\Big] + R^{p-1}\,\mathrm{E}\left[\left(\sum_{j=1}^d |X^j|\right)^p\right]\right\} \\
&\leq \frac{2^p}{p-1}\left\{\sum_{j=1}^d \mathrm{E}[|\langle w, X\rangle - Y|^p]^{\frac{p-1}{p}}\,\mathrm{E}[|X_j|^p]^{1/p} + R^{p-1}d^p \sum_{j=1}^d \mathrm{E}\big[|X^j|^p\big]\right\} \\
&< \infty,
\end{aligned}$$

where in the second line we used that the Euclidean norm is bounded by the 1-norm, in the third we used Holder's inequality, and the last line follows from our assumptions. Applying the dominated convergence theorem, we interchange the limit and the expectation in (B.1). Recalling that ϕ is differentiable finishes the proof. □

We now turn to the twice differentiability of the risk. We start with the easy case $p \in [2, \infty)$. The proof is very similar to that of Lemma 13 and we omit it here.

Lemma 14. *Let $p \in [2, \infty)$ and assume that $\mathrm{E}[|Y|^p] < \infty$ and $\mathrm{E}[|X_j|^p] < \infty$ for all $j \in [d]$. Then R_p is twice differentiable on $\mathbb{R}^d$, and*

$$\nabla^2 R_p(w) = \mathrm{E}[\nabla^2 \ell_p(\langle w, X\rangle - Y)].$$

The case $p \in (1, 2)$ is more complicated. The following lemma establishes the twice differentiability of the risk at its minimizer under a subset of the assumptions of Theorem 6.

Lemma 15. *Let $p \in (1, 2)$. Assume that $\mathrm{P}\big(|\langle w_p^*, X\rangle - Y| = 0\big) = 0$ and $\mathrm{E}[|\langle w_p^*, X\rangle - Y|^{p-2}(X^j)^2] < \infty$ for all $j \in [d]$. Then R_p is twice differentiable at w_p^* and*

$$\nabla^2 R_p(w_p^*) = \mathrm{E}[\nabla^2 \ell_p(\langle w_p^*, X\rangle - Y)]$$

Proof. The difficulty in the proof compared to Lemma 13 and Lemma 14 stems from the fact that the loss is not twice differentiable at zero. We still rely on the dominated convergence theorem, but the construction of the dominating function is slightly more intricate. Using the setup of the proof of Lemma 13, and following the same line of arguments, we arrive at

$$\begin{aligned}&\lim_{k\to\infty} \frac{\|\nabla R_p(w_p^* + \Delta_k) - \nabla R_p(w_p^*) - \mathrm{E}\big[\nabla^2 \phi(w_p^*, X, Y)\Delta_k\big]\|}{\|\Delta_k\|} \\ &\leq \lim_{k\to\infty} \mathrm{E}\left[\frac{\|\nabla \phi(w_p^* + \Delta_k, X, Y) - \nabla \phi(w_p^*, X, Y) - \nabla^2 \phi(w_p^*, X, Y)\Delta_k\|}{\|\Delta_k\|}\right], \end{aligned} \tag{B.2}$$

where we have used the fact that since $\mathrm{P}\big(|\langle w_p^*, X\rangle - Y| = 0\big) = 0$, $\phi(w, X, Y)$ is almost surely twice differentiable at w_p^*. To finish the proof, it remains to construct a dominating function for the above sequence to justify the interchange of the limit and expectation. We consider two cases.

Case 1: $\|\Delta_k\| \geq \big|\langle w_p^*, X\rangle - Y\big|/(2\|X\|) =: R(X, Y)$. Then we have

$$\begin{aligned}&\frac{\|\nabla \phi(w_p^* + \Delta_k, X, Y) - \nabla \phi(w_p^*, X, Y) - \nabla^2 \phi(w_p^*, X, Y)\Delta_k\|}{\|\Delta_k\|} \\ &\leq \frac{\|\nabla \phi(w_p^* + \Delta_k, X, Y)\| + \|\nabla \phi(w_p^*, X, Y)\| + \|\nabla^2 \phi(w_p^*, X, Y)\Delta_k\|}{\|\Delta_k\|} \\ &\leq \frac{\Big(\big|\langle w_p^* + \Delta, X\rangle - Y\big|^{p-1} + \big|\langle w_p^*, X\rangle - Y\big|^{p-1}\Big)\|X\|}{(p-1)\|\Delta_k\|} + \|\nabla^2 \phi(w_p^*, X, Y)\|_{op} \\ &\leq \frac{2\big|\langle w_p^*, X\rangle - Y\big|^{p-1}\|X\|}{(p-1)\|\Delta_k\|} + \big|\langle w_p^*, X\rangle - Y\big|^{p-2}\|X\|^2 + \frac{|\langle \Delta_k/\|\Delta_k\|, X\rangle|^{p-1}\|X\|}{(p-1)\|\Delta_k\|^{2-p}} \\ &\leq \frac{4\big|\langle w_p^*, X\rangle - Y\big|^{p-2}\|X\|^2}{(p-1)} + \big|\langle w_p^*, X\rangle - Y\big|^{p-2}\|X\|^2 + \frac{\|X\|^p}{(p-1)\|\Delta_k\|^{2-p}} \\ &\leq \frac{7\big|\langle w_p^*, X\rangle - Y\big|^{p-2}\|X\|^2}{(p-1)}\end{aligned}$$

where the second line follows by triangle inequality, the third by definition of the operator norm, the fourth by $|a + b|^q \leq |a|^q + |b|^q$ valid for $q \in (0, 1)$, and the fifth and sixth by Cauchy-Schwartz inequality and the assumed lower bound on $\|\Delta_k\|$.

Case 2: $\|\Delta_k\| < R(X, Y)$. We starting by noting that, for all $\Delta \in B(0, R(X, Y)) := \big\{x \in \mathbb{R}^d \mid \|x\| < R(X, Y)\big\}$, we have

$$\big|\langle w_p^* + \Delta, X\rangle - Y\big| \geq \big|\langle w_p^*, X\rangle - Y\big| - |\langle \Delta, X\rangle| \geq \big|\langle w_p^*, X\rangle - Y\big| - \|\Delta\|\|X\| > \big|\langle w_p^*, X\rangle - Y\big|/2 > 0.$$

Therefore $\phi(w, X, Y)$ is twice differentiable on $B(0, R(X, Y))$. Now

$$\begin{aligned}
&\frac{\|\nabla\phi(w_p^* + \Delta_k, X, Y) - \nabla\phi(w_p^*, X, Y) - \nabla^2\phi(w_p^*, X, Y)\Delta_k\|}{\|\Delta_k\|} \\
&\leq \frac{\|\nabla\phi(w_p^* + \Delta_k, X, Y) - \nabla\phi(w_p^*, X, Y)\| + \|\nabla^2\phi(w_p^*, X, Y)\Delta_k\|}{\|\Delta_k\|} \\
&\leq \frac{\left\|\left(\int_0^1 \nabla^2\phi(w_p^* + t\Delta_k, X, Y)dt\right)\Delta_k\right\|}{\|\Delta_k\|} + \|\nabla^2\phi(w_p^*, X, Y)\|_{op} \\
&\leq \left\|\int_0^1 \nabla^2\phi(w + t\Delta_k, X, Y)dt\right\|_{op} + \|\nabla^2\phi(w_p^*, X, Y)\|_{op} \\
&\leq \int_0^1 \|\nabla^2\phi(w + t\Delta_k, X, Y)\|_{op}dt + \|\nabla^2\phi(w_p^*, X, Y)\|_{op} \\
&\leq 2 \sup_{\Delta\in B(0,R(X,Y))} \|\nabla^2\phi(w_p^* + \Delta, X, Y)\|_{op} \\
&\leq 2\|X\|_2^2 \sup_{\Delta\in B(0,R(X,Y))} \left|\langle w_p^* + \Delta, X\rangle - Y\right|^{p-2} \\
&\leq 4\left|\langle w_p^*, X\rangle - Y\right|^{p-2}\|X\|_2^2
\end{aligned}$$

where the second line follows from the triangle inequality, the third follows from the twice differentiability of ϕ on $B(0, R(X, Y))$ and the fundamental theorem of calculus applied component-wise, the fifth by Jensen's inequality, and the last by definition of $R(X, Y)$ and the above lower bound. We therefore define our dominating function by

$$g(X, Y) := 8\left|\langle w_p^*, X\rangle - Y\right|^{p-2}\|X\|_2^2.$$

It is then immediate from our assumptions that $g(X, Y)$ is integrable. Interchanging the limit and the expectation in (B.2) and recalling that ϕ is almost surely twice differentiable at w_p^* finishes the proof.

□

C Proof of Lemma 1

We start with the claim that ρ_0 is upper-semicontinuous. We want to show that for any $w \in \mathbb{R}^d$ and any sequence $(w_k)_{k=1}^\infty$ converging to w (in the norm topology)

$$\limsup_{k\to\infty} \rho_0(w_k) \leq \rho_0(w).$$

Fix a $w \in \mathbb{R}^d$ and let $(w_k)_{k=1}^\infty$ be a sequence in $\mathbb{R}^d$ satisfying $\lim_{k\to\infty}\|w - w_k\| = 0$, where for convenience we take $\|\cdot\|$ to be the Euclidean norm on $\mathbb{R}^d$. Then we have by (reverse) Fatou's Lemma

$$\limsup_{k\to\infty} \rho_0(w_k) = \limsup_{k\to\infty} \mathrm{E}\left[\mathbb{1}_{\{0\}}(\langle w_k, X\rangle)\right] \leq \mathrm{E}\left[\limsup_{k\to\infty} \mathbb{1}_{\{0\}}(\langle w_k, X\rangle)\right]. \tag{C.1}$$

Now we bound the inner limsup pointwise. We split this task in two cases. If $\langle w, X\rangle = 0$, then

$$\limsup_{k\to\infty} \mathbb{1}_{\{0\}}(\langle w_k, X\rangle) \leq 1 = \mathbb{1}_{\{0\}}(\langle w, X\rangle). \tag{C.2}$$

Otherwise we have $\delta := |\langle w, X\rangle| > 0$. But then, by the convergence of $(w_k)_{k=1}^\infty$ to w, there exists a $K \in \mathbb{N}$ such that for all $k \geq K$ we have $\|w_k - w\| < \delta/(2\|X\|)$. This implies that for all $k \geq K$

$$|\langle w_k, X\rangle| = |\langle w, X\rangle - \langle w - w_k, X\rangle| \geq |\langle w, X\rangle| - |\langle w - w_k, X\rangle| \geq \delta - \|w_k - w\|_2\|X\| \geq \delta/2 > 0.$$

We conclude that

$$\limsup_{k\to\infty} \mathbb{1}_{\{0\}}(\langle w_k, X\rangle) = \lim_{k\to\infty} \mathbb{1}_{\{0\}}(\langle w_k, X\rangle) = 0 = \mathbb{1}_{\{0\}}(\langle w, X\rangle). \tag{C.3}$$

Combining (C.1), (C.2), and (C.3) proves the upper-semicontinuity of ρ_0. Essentially the same proof shows the lower-semicontinuity of $\rho_q(\cdot, \kappa)$ for any $\kappa \geq 0$; we omit it here.

For the remaining claims, first notice that the function ρ_0 is scale invariant, i.e. for all $w \in \mathbb{R}^d$ and all $c \in \mathbb{R}$, we have $\rho_0(cw) = \rho_0(w)$. Therefore

$$\sup_{w \in \mathbb{R}^d \setminus \{0\}} \rho_0(w) = \sup_{w \in \mathbb{S}^{d-1}} \rho_0(w),$$

where $\mathbb{S}^{d-1}$ is the Euclidean unit sphere. By assumption on the random vector X, we know that $\rho_0(w) < 1$ for all $w \in \mathbb{S}^{d-1}$. Furthermore since ρ_0 is upper semicontinuous, and $\mathbb{S}^{d-1}$ is compact, ρ_0 attains its supremum on $\mathbb{S}^{d-1}$ at some point $w_0 \in \mathbb{S}^{d-1}$. From this we conclude that

$$\rho = \sup_{w \in \mathbb{R}^d \setminus \{0\}} \rho_0(w) = \rho_0(w_0) < 1.$$

Finally, we turn to the claim about ρ_q. Since $\mathrm{E}[|X^j|^q] < \infty$, the function $\|\cdot\|_{L^q}$ is a norm on $\mathbb{R}^d$, from which it follows that $\rho_q(w, \kappa)$ is also scale invariant for any κ. Therefore

$$\inf_{w \in \mathbb{R}^d \setminus \{0\}} \rho_q(w, \kappa) = \inf_{w \in S_q} \rho_q(w, \kappa),$$

where S_q is the unit sphere of the norm $\|\cdot\|_{L^q}$. Now fix $\kappa \in [0, 1)$. We claim that $\rho_q(w, \kappa) > 0$ for all $w \in S_q$. Suppose not. Then there exists a $w \in S_q$ such that $|\langle w, X \rangle| \leq \kappa$ with probability 1, but then we get the contradiction

$$1 = \|w\|_{L^q} = \mathrm{E}[|\langle w, X \rangle|^q]^{1/q} \leq \kappa < 1.$$

therefore $\rho_q(w, \kappa) > 0$ for all $w \in S_q$. Now since $\rho_q(\cdot, \kappa)$ is lower-semicontinuous, and S_q is compact, $\rho_q(\cdot, \kappa)$ attains its infimum on S_q at some point $w_q \in S_q$. From this we conclude

$$\inf_{w \in \mathbb{R}^d \setminus \{0\}} \rho_q(w, \kappa) = \rho_q(w_q, \kappa) > 0.$$

D Proof of Theorem 4

Fix $p \in (1, \infty)$, and let $\hat{w} := \hat{w}_p$. Our goal will be to upper bound the probability $\mathrm{P}(\hat{w} \neq w^*)$. By assumption, we have that $Y = \langle w^*, X \rangle$, so that $Y_i = \langle w^*, X_i \rangle$ for all $i \in [n]$. Since $\hat{w}$ minimizes the empirical risk, we must also have $\langle \hat{w}, X_i \rangle = Y_i = \langle w^*, X_i \rangle$ for all $i \in [n]$. Let $A \in \mathbb{R}^{n \times d}$ denote the matrix whose i-th row is X_i. Then we have the following implications.

$$\hat{w} \neq w^* \Rightarrow \langle \hat{w} - w^*, X_i \rangle = 0 \; \forall i \in [n] \Rightarrow \exists w \in \mathbb{R}^d \setminus \{0\} \mid Aw = 0 \Leftrightarrow \mathrm{rowrank}(A) < d. \tag{D.1}$$

Let $r := \mathrm{rowrank}(A)$. We claim the following equivalence

$$\mathrm{rowrank}(A) < d \Leftrightarrow \exists S \subset [n] \mid |S| = d - 1 \wedge \forall i \in [n] \setminus S \; X_i \in \mathrm{span}(\{X_k \mid k \in S\}). \tag{D.2}$$

Indeed the implication ($\Leftarrow$) follows by definition of the rowrank of A. For the implication ($\Rightarrow$), by definition, $\{X_i \mid i \in [n]\}$ is a spanning set for the row space of A, therefore it can be reduced to a basis of it $\{X_k \mid k \in S_1\}$ for some indices $S_1 \subset [n]$ with $|S_1| = r$. If $r = d - 1$, then the choice $S = S_1$ satisfies the right side of (D.2). Otherwise, let $S_2 \subset [n] \setminus S_1$ with $|S_2| = d - 1 - r$. Such a subset exists since by assumption $n \geq d > d - 1$. Then the set $S := S_1 \cup S_2$ satisfies the right side of (D.2). Combining (D.1) and (D.2) we arrive at:

$$\begin{aligned} \mathrm{P}(\hat{w} \neq w^*) &\leq \mathrm{P}\left(\bigcup_{\substack{S \subset [n] \\ |S| = d-1}} \{\forall i \in [n] \setminus S \; X_i \in \mathrm{span}(\{X_k \mid k \in S\})\} \right) \\ &\leq \sum_{\substack{S \subset [n] \\ |S| = d-1}} \mathrm{P}(\forall i \in [n] \setminus S \; X_i \in \mathrm{span}(\{X_k \mid k \in S\})) \end{aligned} \tag{D.3}$$

where the second inequality follows from the union bound. We now bound each of the terms of the sum. Fix $S = \{i_1, \dots, i_{d-1}\} \subset [n]$ with $|S| = d-1$. Let $Z_S = n((X_{i_j})_{j=1}^{d-1})$ be a non-zero vector orthogonal to $\operatorname{span}(\{X_k \mid k \in S\})$. Such a vector must exist since $\dim(\operatorname{span}(\{X_k \mid k \in S\})) < d$; see Lemma 16 below for an explicit construction of the function n. Denote by P_{Z_S} the distribution of Z_S and $P_{(X_i)_{i\in[n]\setminus S}} = \prod_{i=1}^{n-d-1} P_X$ the distribution of $(X_i)_{i\in[n]\setminus S}$, where P_X is the distribution of X. Note that since Z_S is a function of $(X_{i_j})_{j=1}^{d}$ only, it is independent of $(X_i)_{i\in[n]\setminus S}$. In particular, the joint distribution of $(Z_S, (X_i)_{i\in[n]\setminus S})$ is given by the product $P_{Z_S} \times P_{(X_i)_{i\in[n]\setminus S}}$. Now if $X_i \in \operatorname{span}(\{X_k \mid k \in S\})$, then by definition of Z_S, $\langle Z_S, X_i\rangle = 0$. Therefore

$$\begin{aligned}
\mathrm{P}(\forall i \in [n]\setminus S\ X_i \in \operatorname{span}(\{X_k \mid k \in S\})) &\leq \mathrm{P}(\forall i \in [n]\setminus S\ \langle Z_S, X_i\rangle = 0) \\
&= \mathrm{E}\left[\prod_{i\in[n]\setminus S} \mathbb{1}_{\{0\}}(\langle Z_S, X_i\rangle)\right] \\
&= \int \left\{\prod_{i\in[n]\setminus S} \mathbb{1}_{\{0\}}(\langle y_S, x_i\rangle)\right\} P_{Z_S}(dz_S) P_{(X_i)_{i\in[n]\setminus S}}(d(x_i)_{i\in[n]\setminus S}) \\
&= \int P_{Z_S}(dy_s) \left\{\prod_{i\in[n]\setminus S} \int \mathbb{1}_{\{0\}}(\langle y_S, x_i\rangle) P_X(dx_i)\right\} \\
&= \int \left\{\prod_{i\in[n]\setminus S} \mathrm{P}(\langle z_S, X\rangle = 0)\right\} P_{Z_S}(dy_s) \\
&= \int \left\{\prod_{i\in[n]\setminus S} \rho_0(y_s)\right\} P_{Z_S}(dy_s) \\
&\leq \rho^{n-d+1}, \qquad (D.4)
\end{aligned}$$

where in the third line we used the independence of Z_S and $(X_i)_{i\in[n]\setminus S}$, in the fourth we used the independence of the $(X_i)_{i\in\setminus S}$, in the sixth we used the definition of ρ_0 in 2.4, and in the last line we used the fact that $z_S \neq 0$ and the definition of ρ in Lemma 1. Combining the inequalities (D.3) and (D.4) yields the result.

Lemma 16. *Let $m \in \{1, \dots, d-1\}$ and let $(x_j)_{j=1}^m$ be a sequence of points in $\mathbb{R}^d$. Denote by $A \in \mathbb{R}^{m\times d}$ the matrix whose j-th row is x_j and let A^+ be its pseudo-inverse. Let $(b_i)_{i=1}^d$ be an ordered basis of $\mathbb{R}^d$, and define*

$$\begin{aligned}
k &:= \min\{i \in [n] \mid (I - A^+A)b_i \neq 0\} \\
n((x_j)_{j=1}^m) &:= (I - A^+A)b_k
\end{aligned}$$

Then $n((x_j)_{j=1}^m)$ is non-zero and is orthogonal to $\operatorname{span}(\{x_j \mid j \in [m]\})$.

Proof. We start by showing that k is well defined. First note that $I - A^+A$ is the orthogonal projector onto the kernel of A, which is non-trivial since $\dim(\ker(A)) = d - \dim(\operatorname{Im}(A)) \geq d - m \geq 1$. Now we claim that there exists an $i \in [d]$ such that $(I - A^+A)b_i \neq 0$. Suppose not, then for any $w \in \mathbb{R}^d$, we have $(I - A^+A)w = (I - A^+A)(\sum_{i=1}^d c_i b_i) = \sum_{i=1}^d c_i(I - A^+A)b_i = 0$, implying that $I - A^+A = 0$, but this contradicts the non-triviality of $\ker(A)$. This proves that k is well-defined, which in turn proves that $n((x_j)_{j=1}^m) \neq 0$. It remains to prove the orthogonality claim. Let $v \in \operatorname{span}(\{x_j \mid j \in [m]\})$. Then there exists coefficients $c \in \mathbb{R}^m$ such that $v = A^T c$. Therefore

$$\langle v, n((x_j)_{j=1}^m)\rangle = \langle A^T c, n((x_j)_{j=1}^m)\rangle = c^T A(I - AA^+)b_k = 0,$$

where the last equality holds since $(I - AA^+)b_k \in \ker(A)$. □

E Detailed proof of Theorem 5

We proceed similarly to the proof of Theorem 2. By definition of the empirical risk minimizer, we have the upper bound

$$R_{p,n}(\hat{w}_p) - R_{p,n}(w_p^*) \leq 0. \tag{E.1}$$

Using (3.5) from Lemma 10 and the Cauchy-Schwarz inequality, we obtain the pointwise lower bound

$$R_{p,n}(\hat{w}_p) - R_{p,n}(w_p^*) \geq -\|\nabla R_{p,n}(w_p^*)\|_{H_p^{-1}} \|\hat{w}_p - w_p^*\|_{H_p} + \frac{1}{8(p-1)} \|\hat{w}_p - w_p^*\|_{H_{p,n}}^2. \tag{E.2}$$

Using Lemma 7 we have that, with probability at least $1 - \delta/2$,

$$\|\nabla R_{p,n}(w_p^*)\|_{H_p^{-1}} \leq \sqrt{2\,\mathrm{E}\left[\|\nabla \ell_p(\langle w_p^*, X\rangle - Y)\|_{H_p^{-1}}^2\right]/(n\delta)}. \tag{E.3}$$

It remains to control $\|\hat{w}_p - w_p^*\|_{H_{p,n}}^2$ from below. Define the random vector

$$Z = |\langle w_p^*, X\rangle - Y|^{(p-2)/2} X$$

Then, for any $w \in \mathbb{R}^d$, we have

$$\begin{aligned}
\|w - w_p^*\|_{H_{p,n}}^2 &= (w - w_p^*)^T H_{p,n} (w - w_p^*) \\
&= \frac{1}{n} \sum_{i=1}^n (w - w_p^*)^T \nabla^2 \ell_p(\langle w_p^*, X_i\rangle - Y_i)(w - w_p^*) \\
&= \frac{1}{n} \sum_{i=1}^n \langle w - w_p^*, Z_i\rangle^2
\end{aligned}$$

By assumption, the components of the random vector Z have finite fourth moment. Applying Proposition 8, and using the condition on n assumed in the statement of Theorem 5, we get that, with probability at least $1 - \delta/2$, for all $w \in \mathbb{R}^d$,

$$\|w - w_p^*\|_{H_{p,n}}^2 \geq \frac{1}{2} \|w - w_p^*\|_{H_p}^2. \tag{E.4}$$

Combining (E.3) and (E.4) with (E.2) gives that with probability at least $1 - \delta$,

$$\begin{aligned}
R_{p,n}(\hat{w}_p) - R_{p,n}(w_p^*) \geq -\sqrt{2\,\mathrm{E}\left[\|\nabla \ell_p(\langle w_p^*, X\rangle - Y)\|_{H_p^{-1}}^2\right]/(n\delta)}\, \|\hat{w}_p - w_p^*\|_{H_p} \\
+ \frac{1}{16(p-1)} \|\hat{w}_p - w_p^*\|_{H_p}^2.
\end{aligned} \tag{E.5}$$

Further combining (E.5) with (E.1) and rearranging yields that with probability at least $1 - \delta$

$$\|\hat{w}_p - w_p^*\|_{H_p}^2 \leq \frac{512p^2\,\mathrm{E}\left[\|\nabla \ell_p(\langle w_p^*, X\rangle - Y)\|_{H_p^{-1}}^2\right]}{n\delta} \tag{E.6}$$

The last step is to bound the excess risk of the empirical risk minimizer using the bound (E.6) and (3.6) from Lemma 10. For that, we control the L^p norm term in (3.6) as follows

$$\begin{aligned}
p^p \|\hat{w}_p - w_p^*\|_{L^p}^p &= \left(p^2 \frac{\|\hat{w}_p - w_p^*\|_{L^p}^2}{\|\hat{w}_p - w_p^*\|_{H_p}^2} \|\hat{w}_p - w_p^*\|_{H_p}^2\right)^{p/2} \\
&\leq \left(p^2 \sup_{w \in \mathbb{R}^d \setminus \{0\}} \left\{\frac{\|w\|_{L^p}^2}{\|w\|_{H_p}^2}\right\} \|\hat{w}_p - w_p^*\|_{H_p}^2\right)^{p/2} \\
&= \left(p^2 c_p^2 \|\hat{w}_p - w_p^*\|_{H_p}^2\right)^{p/2}.
\end{aligned} \tag{E.7}$$

Combining (E.6), (3.6), and (E.7) yields the result.

F Detailed proof of Theorem 6

In this section, we provide a detailed proof of Theorem 6. We finish the proof of Proposition 12, and conclude by bringing all the pieces together to prove Theorem 6.

F.1 Rest of the proof of Proposition 12

We continue the proof by bounding the last remaining supremum. We have

$$
\begin{aligned}
&\sup_{\|w\|_{H_p}=1} \mathrm{E}\Big[|\langle w_p^*, X\rangle - Y|^{p-2}\langle w, X\rangle^2\Big(\mathbb{1}_{[0,\beta)}(|\langle w_p^*, X\rangle - Y|) + \mathbb{1}_{(T,\infty)}(\|X\|_{H_p^{-1}})\Big)\Big] \\
&\leq \sup_{\|w\|_{H_p}=1}\left\{\mathrm{E}\Big[|\langle w_p^*, X\rangle - Y|^{2(p-2)}\langle w, X\rangle^4\Big]^{1/2}\right\}\Big(\mathrm{P}\big(|\langle w_p^*, X\rangle - Y| < \beta\big) + \mathrm{P}\Big(\|X\|_{H_p^{-1}} > T\Big)\Big)^{1/2} \\
&= \left(\sup_{\|w\|_{H_p}=1}\|w\|_{L^4,p}^2\right)\Big(\mathrm{P}\big(|\langle w_p^*, X\rangle - Y| < \beta\big) + \mathrm{P}\Big(\|X\|_{H_p^{-1}} > T\Big)\Big)^{1/2} \\
&= \sigma_p\Big(\mathrm{P}\big(|\langle w_p^*, X\rangle - Y| < \beta\big) + \mathrm{P}\Big(\|X\|_{H_p^{-1}} > T\Big)\Big)^{1/2},
\end{aligned}
$$

where the first inequality follows from Cauchy Schwartz inequality, and the subsequent equalities by definitions of $\|\cdot\|_{L^4,p}$ and σ_p^2. It remains to bound the tail probabilities. Recall that $\beta = T\varepsilon$, so that on the one hand we have

$$
\begin{aligned}
\mathrm{P}\big(|\langle w_p^*, X\rangle - Y| < \beta\big) &= \mathrm{P}\big(|\langle w_p^*, X\rangle - Y| < T\varepsilon\big) \\
&= \mathrm{P}\big(|\langle w_p^*, X\rangle - Y|^{-1} > (T\varepsilon)^{-1}\big) \\
&= \mathrm{P}\Big(|\langle w_p^*, X\rangle - Y|^{2(p-2)} > (T\varepsilon)^{2(p-2)}\Big) \\
&\leq \mathrm{E}[|\langle w_p^*, X\rangle - Y|^{2(p-2)}](T\varepsilon)^{2(2-p)} \\
&= c_p^*(T\varepsilon)^{2(2-p)}, \qquad \text{(F.1)}
\end{aligned}
$$

where we applied Markov's inequality in the fourth line, and the last follows by definition of c_p^*. On the other hand, we have

$$
\mathrm{E}\Big[\|X\|_{H_p^{-1}}^2\Big] = \mathrm{E}\big[X^T H_p^{-1} X\big] = \mathrm{E}\big[\mathrm{Tr}\big(H_p^{-1}XX^T\big)\big] = \mathrm{Tr}\big(H_p^{-1}\Sigma\big)
$$

where $\Sigma = \mathrm{E}[XX^T]$. Define $\widetilde{H}_p := \Sigma^{-1/2}H_p\Sigma^{-1/2}$. Then by the above, we have

$$
\mathrm{E}\Big[\|X\|_{H_p^{-1}}^2\Big] = \mathrm{Tr}\Big(\widetilde{H}_p^{-1}\Big) \leq d\lambda_{\max}(\widetilde{H}_p^{-1}) = \frac{d}{\lambda_{\min}(\widetilde{H}_p)}
$$

where $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ are the largest and smallest eigenvalues of a positive definite matrix A respectively. Now, using the variational characterization of the smallest eigenvalue, we have

$$
\mathrm{E}\Big[\|X\|_{H_p^{-1}}^2\Big] \leq d \sup_{v\in\mathbb{R}^d\setminus\{0\}} \frac{\|v\|_2^2}{v^T\tilde{H}_p v} = d \sup_{w\in\mathbb{R}^d\setminus\{0\}} \frac{w^T\Sigma w}{w^T H_p w} = d \sup_{w\in\mathbb{R}^d\setminus\{0\}} \frac{\|w\|_{L^2}^2}{\|w\|_{H_p}^2} = dc_p^2
$$

where the second equality is obtained by the change of variable $w = \Sigma^{-1/2}v$, the third by the definition of $\|\cdot\|_{L^2}$ and $\|\cdot\|_{H_p}$, and the last by definition of c_p in Theorem 6. Therefore, by Markov's inequality, we get

$$
\mathrm{P}\Big(\|X\|_{H_p^{-1}} > T\Big) \leq \frac{dc_p}{T^2} \qquad \text{(F.2)}
$$

Combining the inequalities (F.1) and (F.2), we obtain

$$
\mathrm{P}\big(|\langle w_p^*, X\rangle - Y| < T\varepsilon\big) + \mathrm{P}\Big(\|X\|_{H_p^{-1}} > T\Big) \leq c_p^* T^{2(2-p)}\varepsilon^{2(2-p)} + \frac{dc_p}{T^2} \qquad \text{(F.3)}
$$

Minimizing over T we get

$$T^* := \left(\frac{dc_p}{c_p^*(2-p)}\right)^{1/(6-2p)}$$

which ensures that

$$\left(\mathrm{P}\left(|\langle w_p^*, X\rangle - Y| < T^*\varepsilon\right) + \mathrm{P}\left(\|X\|_{H_p^{-1}} > T^*\right)\right)^{1/2} \leq \sqrt{2}(c_p^*)^{1/(6-2p)}\left(\varepsilon\sqrt{d}\sqrt{c_p}\right)^{(2-p)/(3-p)}$$

Therefore, choosing

$$\varepsilon^{2-p} := \frac{1}{8\sigma_p^{3-p}\sqrt{c_p^*}(dc_p)^{(2-p)/2}},$$

ensures that the supremum we are bounding is less than 1/2. □

F.2 Proof of Theorem 6

We follow the same proof strategy as the one used in the proofs of Theorems 2 and 5. By definition of the empirical risk minimizer, we have

$$R_{p,n}(\hat{w}_p) - R_{p,n}(w_p^*) \leq 0. \tag{F.4}$$

Using (3.7) from Lemma 11 and the Cauchy-Schwarz inequality, we have the lower bound

$$\begin{aligned} R_{p,n}(\hat{w}_p) - R_{p,n}(w_p^*) \geq -\|\nabla R_{p,n}(w_p^*)\|_{H_p^{-1}}\|\hat{w}_p - w_p^*\|_{H_p} \\ + \frac{1}{4}\frac{1}{n}\sum_{i=1}^{n}\gamma_p\left(|\langle w_p^*, X_i\rangle - Y_i|, |\langle w - w_p^*, X_i\rangle|\right) \end{aligned} \tag{F.5}$$

Using Lemma 7 we have that, with probability at least $1-\delta/2$,

$$\|\nabla R_{p,n}(w_p^*)\|_{H_p^{-1}} \leq \sqrt{2\,\mathrm{E}\left[\|\nabla \ell_p(\langle w_p^*, X\rangle - Y)\|_{H_p^{-1}}^2\right]/(n\delta)}. \tag{F.6}$$

On the other hand, by Proposition 12, we have with probability $1-\delta/2$, for all $w \in \mathbb{R}^d$,

$$\frac{1}{n}\sum_{i=1}^{n}\gamma_p\left(|\langle w_p^*, X_i\rangle - Y_i|, |\langle w - w_p^*, X_i\rangle|\right) \geq \frac{1}{8}\min\left\{\|w - w_p^*\|_{H_p}^2, \varepsilon^{2-p}\|w - w_p^*\|_{H_p}^p\right\}, \tag{F.7}$$

where ε is as defined in Proposition 12. We now consider two cases. If $\|\hat{w}_p - w_p^*\|_{H_p}^2 \leq \varepsilon^{2-p}\|\hat{w}_p - w_p^*\|_{H_p}^p$, then combining (F.4), (F.5), (F.6), and (F.7) gives

$$\|\hat{w}_p - w_p^*\|_{H_p}^2 \leq \frac{2048\,\mathrm{E}\left[\|\nabla \ell_p(\langle w_p^*, X\rangle - Y)\|_{H_p^{-1}}^2\right]}{n\delta}. \tag{F.8}$$

Otherwise, $\|\hat{w}_p - w_p^*\|_{H_p}^2 > \varepsilon^{2-p}\|\hat{w}_p - w_p^*\|_{H_p}^p$, then again combining (F.4), (F.5), (F.6), and (F.7) gives

$$\|\hat{w}_p - w_p^*\|_{H_p}^2 \leq \left(\frac{2048\,\mathrm{E}\left[\|\nabla \ell_p(\langle w_p^*, X\rangle - Y)\|_{H_p^{-1}}^2\right]\varepsilon^{2(p-2)}}{n\delta}\right)^{1/(p-1)} \tag{F.9}$$

In either case, we have, using (F.8) and (F.9), with probability at least $1-\delta$,

$$\|\hat{w}_p - w_p^*\|_{H_p}^2 \leq \frac{2048\,\mathrm{E}\left[\|\nabla \ell_p(\langle w_p^*, X\rangle - Y)\|_{H_p^{-1}}^2\right]}{n\delta} + \left(\frac{2048\,\mathrm{E}\left[\|\nabla \ell_p(\langle w_p^*, X\rangle - Y)\|_{H_p^{-1}}^2\right]\varepsilon^{2(p-2)}}{n\delta}\right)^{1/(p-1)}. \tag{F.10}$$

Combining this last inequality with (3.8) from Lemma 11 finishes the proof.