Agnostic Estimation for Misspecified Phase Retrieval Models

Matey Neykov **Zhaoran Wang** **Han Liu**
Department of Operations Research and Financial Engineering
Princeton University, Princeton, NJ 08544
{mneykov, zhaoran, hanliu}@princeton.edu

Abstract

The goal of noisy high-dimensional phase retrieval is to estimate an s-sparse parameter $\boldsymbol{\beta}^* \in \mathbb{R}^d$ from n realizations of the model $Y = (\boldsymbol{X}^\top \boldsymbol{\beta}^*)^2 + \varepsilon$. Based on this model, we propose a significant semi-parametric generalization called *misspecified phase retrieval* (MPR), in which $Y = f(\boldsymbol{X}^\top \boldsymbol{\beta}^*, \varepsilon)$ with unknown f and $\mathrm{Cov}(Y, (\boldsymbol{X}^\top \boldsymbol{\beta}^*)^2) > 0$. For example, MPR encompasses $Y = h(|\boldsymbol{X}^\top \boldsymbol{\beta}^*|) + \varepsilon$ with increasing h as a special case. Despite the generality of the MPR model, it eludes the reach of most existing semi-parametric estimators. In this paper, we propose an estimation procedure, which consists of solving a cascade of two convex programs and provably recovers the direction of $\boldsymbol{\beta}^*$. Our theory is backed up by thorough numerical results.

1 Introduction

In scientific and engineering fields researchers often times face the problem of quantifying the relationship between a given outcome Y and corresponding predictor vector $\boldsymbol{X}$, based on a sample $\{(Y_i, \boldsymbol{X}_i^\top)^\top\}_{i=1}^n$ of n observations. In such situations it is common to postulate a linear "working" model, and search for a d-dimensional signal vector $\boldsymbol{\beta}^*$ satisfying the following familiar relationship:

$$Y = \boldsymbol{X}^\top \boldsymbol{\beta}^* + \varepsilon. \tag{1.1}$$

When the predictor $\boldsymbol{X}$ is high-dimensional in the sense that $d \gg n$, it is commonly assumed that the underlying signal $\boldsymbol{\beta}^*$ is s-sparse. In a certain line of applications, such as X-ray crystallography, microscopy, diffraction and array imaging[1], one can only measure the magnitude of $\boldsymbol{X}^\top \boldsymbol{\beta}^*$ but not its phase (i.e., sign in the real domain). In this case, assuming model (1.1) may not be appropriate. To cope with such applications in the high-dimensional setting, [7] proposed the thresholded Wirtinger flow (TWF), a procedure which consistently estimates the signal $\boldsymbol{\beta}^*$ in the following *real sparse noisy phase retrieval model*:

$$Y = (\boldsymbol{X}^\top \boldsymbol{\beta}^*)^2 + \varepsilon, \tag{1.2}$$

where one additionally knows that the predictors have a Gaussian random design $\boldsymbol{X} \sim \mathcal{N}(0, \mathbf{I}_d)$. In the present paper, taking an agnostic point of view, we recognize that both models (1.1) and (1.2) represent an idealized view of the data generating mechanism. In reality, the nature of the data could be better reflected through the more flexible viewpoint of a single index model (SIM):

$$Y = f(\boldsymbol{X}^\top \boldsymbol{\beta}^*, \varepsilon), \tag{1.3}$$

where f is an unknown link function, and it is assumed that $\|\boldsymbol{\beta}^*\|_2 = 1$ for identifiability. A recent line of work on high-dimensional SIMs [25, 27], showed that under Gaussian designs, one can apply ℓ_1 regularized least squares to successfully estimate the direction of $\boldsymbol{\beta}^*$ and its support. The crucial condition allowing for the above somewhat surprising application turns out to be:

$$\mathrm{Cov}(Y, \boldsymbol{X}^\top \boldsymbol{\beta}^*) \neq 0. \tag{1.4}$$

While condition (1.4) is fairly generic, encompassing cases with a binary outcome, such as logistic regression and one-bit compressive sensing [5], it fails to capture the phase retrieval model (1.2).

[1] In such applications it is typically assumed that $\boldsymbol{X} \in \mathbb{C}^d$ is a complex normal random vector. In this paper for simplicity we only consider the real case $\boldsymbol{X} \in \mathbb{R}^d$.

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

More generally, it is easy to see that when the link function f is even in its first coordinate, condition (1.4) fails to hold. The goal of the present manuscript is to formalize a class of SIMs, which includes the noisy phase retrieval model as a special case in addition to various other additive and non-additive models with even link functions, and develop a procedure that can successfully estimate the direction of $\boldsymbol{\beta}^*$ up to a global sign. Formally, we consider models (1.3) with Gaussian design that satisfy the following moment assumption:

$$\mathrm{Cov}(Y, (\boldsymbol{X}^\top \boldsymbol{\beta}^*)^2) > 0. \tag{1.5}$$

Unlike (1.4), one can immediately check that condition (1.5) is satisfied by model (1.2). In §2 we give multiple examples, both abstract and concrete, of SIMs obeying this constraint. Our second moment constraint (1.5) can be interpreted as a semi-parametric robust version of phase-retrieval. Hence, we will refer to the class of models satisfying condition (1.5) as *misspecified phase retrieval* (MPR) models. In this point of view it is worth noting that condition (1.4) relates to linear regression in a way similar to how condition (1.5) relates to the phase retrieval model. Our motivation for studying SIMs under such a constraint can ultimately be traced to the vast sufficient dimension reduction (SDR) literature. In particular, we would like to point out [22] as a source of inspiration.

Contributions. Our first contribution is to formulate a novel and easily implementable two-step procedure, which consistently estimates the direction of $\boldsymbol{\beta}^*$ in an MPR model. In the first step we solve a semidefinite program producing a unit vector $\widehat{\mathbf{v}}$, such that $|\widehat{\mathbf{v}}^\top \boldsymbol{\beta}^*|$ is sufficiently large. Once such a pilot estimate is available, we consider solving an ℓ_1 regularized least squares on the augmented outcome $\widetilde{Y}_i = (Y_i - \overline{Y})\boldsymbol{X}_i^\top \widehat{\mathbf{v}}$, where $\overline{Y}$ is the average of Y_i's, to produce a second estimate $\widehat{\mathbf{b}}$, which is then normalized to obtain the final refined estimator $\widehat{\boldsymbol{\beta}} = \widehat{\mathbf{b}}/\|\widehat{\mathbf{b}}\|_2$. In addition to being universally applicable to MPR models, our procedure has an algorithmic advantage in that it relies solely on convex optimization, and as a consequence we can obtain the corresponding global minima of the two convex programs in polynomial time.

Our second contribution is to rigorously demonstrate that the above procedure consistently estimates the direction of $\boldsymbol{\beta}^*$. We prove that for a given MPR model, with high probability, one has:

$$\min_{\eta \in \{-1,1\}} \|\widehat{\boldsymbol{\beta}} - \eta \boldsymbol{\beta}^*\|_2 \lesssim \sqrt{s \log d / n},$$

provided that the sample size n satisfies $n \gtrsim s^2 \log d$. While the same rates (with different constants) hold for the TWF algorithm of [7] in the special case of noisy phase retrieval model, our procedure provably achieves these rates over the broader class of MPR models.

Lastly, we propose an optional refinement of our algorithm, which shows improved performance in the numerical studies.

Related Work. The phase retrieval model has received considerable attention in the recent years by statistics, applied mathematics as well as signal processing communities. For the non-sparse version of (1.2), efficient algorithms have been suggested based on both semidefinite programs [8, 10] and non-convex optimization methods that extend gradient descent [9]. Additionally, a non-traditional instance of phase retrieval model (which also happens to be a special case of the MPR model) was considered by [11], where the authors suggested an estimation procedure originally proposed for the problem of mixed regression. For the noisy sparse version of model (1.2), near optimal solutions were achieved with a computationally infeasible program by [20]. Subsequently, a tractable gradient descent approach achieving minimax optimal rates was developed by [7].

Abstracting away from the phase retrieval or linear model settings, we note that inference for SIMs in the case when d is small or fixed, has been studied extensively in the literature [e.g., 18, 24, 26, 34, among many others]. In another line of research on SDR, seminal insights shedding light on condition (1.4) can be found in, e.g., [12, 21, 23]. The modified condition (1.5) traces roots to [22], where the authors designed a procedure to handle precisely situations where (1.4) fails to hold. More recently, there have been active developments for high-dimensional SIMs. [27] and later [31] demonstrated that under condition (1.4), running the least squares with ℓ_1 regularization can obtain a consistent estimate of the direction of $\boldsymbol{\beta}^*$, while [25] showed that this procedure also recovers the signed support of the direction. Excess risk bounds were derived in [14]. Very recently, [16] extended this observation to other convex loss functions under a condition corresponding to (1.4) depending implicitly on the loss function of interest. [28] proposed a non-parametric least squares with an equality ℓ_1 constraint to handle simultaneous estimation of $\boldsymbol{\beta}^*$ as well as f. [17] considered a smoothed-out U-process type of loss function with ℓ_1 regularization, and proved their approach works for a sub-class of functions satisfying condition (1.4). None of the aforementioned works on SIMs can be directly applied to tackle the MPR class (1.5). A generic procedure for estimating sparse principal eigenvectors was

developed in [37]. While in principle this procedure can be applied to estimate the direction in MPR models, it requires proper initialization, and in addition, it requires knowledge of the sparsity of the vector $\boldsymbol{\beta}^*$. We discuss this approach in more detail in §4.

Regularized procedures have also been proposed for specific choices of f and Y. For example, [36] studied consistent estimation under the model $\mathbb{P}(Y = 1|\boldsymbol{X}) = (h(\boldsymbol{X}^\top\boldsymbol{\beta}^*) + 1)/2$ with binary Y, where $h : \mathbb{R} \mapsto [-1, 1]$ is possibly unknown. Their procedure is based on taking pairs of differences in the outcome, and therefore replaces condition (1.4) with a different type of moment conditon. [35] considered the model $Y = h(\boldsymbol{X}^\top\boldsymbol{\beta}^*) + \varepsilon$ with a known continuously differentiable and monotonic h, and developed estimation and inferential procedures based on the ℓ_1 regularized quadratic loss, in a similar spirit to the TWF algorithm suggested by [7]. In conclusion, although there exists much prior related work, to the best of our knowledge, none of the available literature discusses the MPR models in the generality we attempt in the present manuscript.

Notation. We now briefly outline some commonly used notations. Other notations will be defined as needed throughout the paper. For a (sparse) vector $\mathbf{v} = (v_1, \ldots, v_p)^\top$, we let $S_{\mathbf{v}} := \mathrm{supp}(\mathbf{v}) = \{j : v_j \neq 0\}$ denote its support, $\|\mathbf{v}\|_p$ denote the ℓ_p norm (with the usual extension when $p = \infty$) and $\mathbf{v}^{\otimes 2} := \mathbf{v}\mathbf{v}^\top$ is a shorthand for the outer product. With a standard abuse of notation we will denote by $\|\mathbf{v}\|_0 = |\mathrm{supp}(\mathbf{v})|$ the cardinality of the support of $\mathbf{v}$. We often use $\mathbf{I}_d$ to denote a $d \times d$ identity matrix. For a real random variable X, define

$$\|X\|_{\psi_2} = \sup_{p\geq 1} p^{-1/2}(\mathbb{E}|X|^p)^{1/p}, \quad \|X\|_{\psi_1} = \sup_{p\geq 1} p^{-1}(\mathbb{E}|X|^p)^{1/p}.$$

Recall that a random variable is called *sub-Gaussian* if $\|X\|_{\psi_2} < \infty$ and *sub-exponential* if $\|X\|_{\psi_1} < \infty$ [e.g., 32]. For any integer $k \in \mathbb{N}$ we use the shorthand notation $[k] = \{1, \ldots, k\}$. We also use standard asymptotic notations. Given two sequences $\{a_n\}, \{b_n\}$ we write $a_n = O(b_n)$ if there exists a constant $C < \infty$ such that $a_n \leq Cb_n$, and $a_n \asymp b_n$ if there exist positive constants c and C such that $c < a_n/b_n < C$.

Organization. In §2 and §3 we introduce the MPR model class and our estimation procedure, and §3.1 is dedicated to stating the theoretical guarantees of our proposed algorithm. Simulation results are given in §4. A brief discussion is provided in §5. We defer the proofs to the appendices due to space limitations.

2 MPR Models

In this section we formally introduce MPR models. In detail, we argue that the class of such models is sufficiently rich, including numerous models of interest. Motivated by the setup in the sparse noisy phase retrieval model (1.2), we assume throughout the remainder of the paper that $\boldsymbol{X} \sim \mathcal{N}(0, \mathbf{I}_d)$. We begin our discussion with a formal definition.

Definition 2.1 (MPR Models)**.** Assume that we are given model (1.3), where $\boldsymbol{X} \sim \mathcal{N}(0, \mathbf{I}_d)$, $\varepsilon \perp\!\!\!\perp \boldsymbol{X}$ and $\boldsymbol{\beta}^* \in \mathbb{R}^d$ is an s-sparse unit vector, i.e., $\|\boldsymbol{\beta}^*\|_2 = 1$. We call such a model *misspecified phase retrieval* (MPR) model, if the link function f and noise ε further satisfy, for $Z \sim \mathcal{N}(0, 1)$ and $K > 0$,

$$c_0 := \mathrm{Cov}(f(Z, \varepsilon), Z^2) > 0, \qquad (2.1) \qquad\qquad \|Y\|_{\psi_1} \leq K. \qquad (2.2)$$

Both assumptions (2.1) and (2.2) impose moment restrictions on the random variable Y. Assumption (2.1) states that Y is positively correlated with the random variable $(\boldsymbol{X}^\top\boldsymbol{\beta}^*)^2$, while assumption (2.2) requires Y to have somewhat light-tails. Also, as mentioned in the introduction, the unit norm constraint on the vector $\boldsymbol{\beta}^*$ is required for the identifiability of model (1.3). We remark that the class of MPR models is convex in the sense that if we have two MPR models $f_1(\boldsymbol{X}^\top\boldsymbol{\beta}^*, \varepsilon)$ and $f_2(\boldsymbol{X}^\top\boldsymbol{\beta}^*, \varepsilon)$, all models generated by their convex combinations $\lambda f_1(\boldsymbol{X}^\top\boldsymbol{\beta}^*, \varepsilon) + (1-\lambda) f_2(\boldsymbol{X}^\top\boldsymbol{\beta}^*, \varepsilon)$ $(\lambda \in [0, 1])$ are also MPR models. It is worth noting the $>$ direction in (2.1) is assumed without loss of generality. If $\mathrm{Cov}(Y, (\boldsymbol{X}^\top\boldsymbol{\beta}^*)^2) < 0$ one can apply the same algorithm to $-Y$. However, the knowledge of the direction of the inequality is important.

In the following, we restate condition (2.1) in a more convenient way, enabling us to easily calculate the explicit value of the constant c_0 in several examples.

Proposition 2.2. Assume that there exists a version of the map $\varphi(z) : z \mapsto \mathbb{E}[f(Z, \varepsilon)|Z = z]$ such that $\mathbb{E}D^2\varphi(Z) > 0$, where D^2 is the second distributional derivative of φ and $Z \sim \mathcal{N}(0, 1)$. Then the SIM (1.3) satisfies assumption (2.1) with $c_0 = \mathbb{E}D^2\varphi(Z)$.

We now provide three concrete MPR models as warm up examples for the more general examples discussed in Proposition 2.3 and Remark 2.3. Consider the models:

$$Y = (\boldsymbol{X}^\top \boldsymbol{\beta}^*)^2 + \varepsilon, \quad (2.3) \qquad Y = |\boldsymbol{X}^\top \boldsymbol{\beta}^*| + \varepsilon, \quad (2.4) \qquad Y = |\boldsymbol{X}^\top \boldsymbol{\beta}^* + \varepsilon|, \quad (2.5)$$

where $\varepsilon \perp\!\!\!\perp \boldsymbol{X}$ is sub-exponential noise, i.e., $\|\varepsilon\|_{\psi_1} \leq K_\varepsilon$ for some $K_\varepsilon > 0$. Model (2.3) is the noisy phase retrieval model considered by [7], while models (2.4) and (2.5) were both discussed in [11], where the authors proposed a method to solve model (2.5) in the low-dimensional regime. Below we briefly explain why these models satisfy conditions (2.1) and (2.2). First, observe that in all three models we have a sum of two sub-exponential random variables, and hence by the triangle inequality it follows that the random variable Y is also sub-exponential, which implies (2.2). To understand why (2.1) holds, by applying Proposition 2.2 we have $c_0 = \mathbb{E}2 = 2 > 0$ for model (2.3), $c_0 = \mathbb{E}2\delta_0(Z) = 2/\sqrt{2\pi} > 0$ for model (2.4), and $c_0 = \mathbb{E}2\delta_0(Z + \varepsilon) = 2\mathbb{E}\phi(\varepsilon) > 0$ for model (2.5), where $\delta_0(\cdot)$ is the Dirac delta function centered at zero, and ϕ is the density of the standard normal distribution.

Admittedly, calculating the second distributional derivative could be a laborious task in general. In the remainder of this section we set out to find a simple to check generic sufficient condition on the link function f and error term ε, under which both (2.1) and (2.2) hold. Before giving our result note that the condition $\mathbb{E}D^2\varphi(Z) > 0$ is implied whenever φ is strictly convex and twice differentiable. However, strictly convex functions φ may violate assumption (2.2) as they can inflate the tails of Y arbitrarily (consider, e.g., $f(x, \varepsilon) = x^4 + \varepsilon$). Moreover, the functions in example (2.4) and (2.5) fail to be twice differentiable. In the following result we handle those two problems, and in addition we provide a more generic condition than convexity, which suffices to ensure the validity of (2.1).

Proposition 2.3. The following statements hold.

(i) Let the function φ defined in Proposition 2.2 be such that the map $z \mapsto \varphi(z) + \varphi(-z)$ is non-decreasing on $\mathbb{R}_0^+$ and and there exist $z_1 > z_2 > 0$ such that $\varphi(z_1) + \varphi(-z_1) > \varphi(z_2) + \varphi(-z_2)$. Then (2.1) holds.

(ii) A sufficient condition for (i) to hold, is that $z \mapsto \varphi(z)$ is convex and sub-differentiable at every point $z \in \mathbb{R}$, and there exists a point $z_0 \in \mathbb{R}_0^+$ satisfying $\varphi(z_0) + \varphi(-z_0) > 2\varphi(0)$.

(iii) Assume that there exist functions g_1, g_2 such that $f(Z, \varepsilon) \leq g_1(Z) + g_2(\varepsilon)$, and g_1 is *essentially quadratic* in the sense that there exists a closed interval $\mathcal{I} = [a, b]$ with $0 \in \mathcal{I}$, such that for all z satisfying $g_1(z) \in \mathcal{I}^c$ we have $|g_1(z)| \leq Cz^2$ for a sufficiently large constant $C > 0$, and let $g_2(\varepsilon)$ be sub-exponential. Then (2.2) holds.

Remark 2.4. Proposition 2.3 shows that the class of MPR models is sufficiently broad. By (i) and (ii) it immediately follows that the additive models

$$Y = h(\boldsymbol{X}^\top \boldsymbol{\beta}^*) + \varepsilon, \quad (2.6)$$

where the link function h is even and increasing on $\mathbb{R}_0^+$ or convex, satisfy the covariance condition (2.1) by (i) and (ii) of Proposition 2.3 respectively. If h is also essentially quadratic and ε is sub-exponentially distributed, using (iii) we can deduce that Y in (2.6) is a sub-exponential random variable, and hence under these restrictions model (2.6) is an MPR model. Both examples (2.3) and (2.4) take this form.

Additionally, Proposition 2.3 implies that the model

$$Y = h(\boldsymbol{X}^\top \boldsymbol{\beta}^* + \varepsilon) \quad (2.7)$$

satisfies (2.1), whenever the link h is a convex sub-differentiable function, such that $h(z_0)+h(-z_0) > 2h(0)$ for some $z_0 > 0$, $\mathbb{E}|h(z + \varepsilon)| < \infty$ for all $z \in \mathbb{R}$ and $\mathbb{E}|h(Z + \varepsilon)| < \infty$. This conclusion follows because under the latter conditions the function $\varphi(z) = \mathbb{E}h(z + \varepsilon)$ satisfies (ii), which is proved in Appendix C under Lemma C.1. Moreover, if it turns out that h is essentially quadratic and $h(2\varepsilon)$ is sub-exponential, then by Jensen's inequality we have $2h(Z+\varepsilon) \leq h(2Z)+h(2\varepsilon)$ and hence (iii) implies that (2.2) is also satisfied. Model (2.5) is of the type (2.7). Unlike the additive noise models (2.6), models (2.7) allow noise corruption even within the argument of the link function. On the negative side, it should be apparent that (2.1) fails to hold in cases where φ is an odd function, i.e., $\varphi(z) = -\varphi(-z)$. In many such cases (e.g. when φ is monotone or non-constant and non-positive/non-negative on $\mathbb{R}^+$), one would have $\mathrm{Cov}(Y, \boldsymbol{X}^\top \boldsymbol{\beta}^*) = \mathbb{E}[\varphi(Z)Z] \neq 0$, and hence direct application of the ℓ_1 regularized least squares algorithm is possible as we discussed in the introduction.

3 Agnostic Estimation for MPR

In this section we describe and motivate our two-step procedure, which consists of a convex relaxation and an ℓ_1 regularized least squares program, for performing estimation in the MPR class of models

described by Definition 2.1. We begin our motivation by noting that any MPR model satisfies the following inequality

$$\mathrm{Cov}((Y-\mu)\boldsymbol{X}^\top\boldsymbol{\beta}^*, \boldsymbol{X}^\top\boldsymbol{\beta}^*) = \mathbb{E}\{(Y-\mu)(\boldsymbol{X}^\top\boldsymbol{\beta}^*)^2\} = \mathrm{Cov}(f(Z,\varepsilon), Z^2) = c_0 > 0, \quad (3.1)$$

where we have denoted $\mu := \mathbb{E}Y$. This simple observation plays a major role in the motivation of our procedure. Notice that in view of condition (1.4), inequality (3.1) implies that if instead of observing Y we had observed $\check{Y} = g(\boldsymbol{X}^\top\boldsymbol{\beta}^*, \varepsilon) = (Y-\mu)\boldsymbol{X}^\top\boldsymbol{\beta}^*$. However, there is no direct way of generating the random variable $\check{Y}$, as doing so would require the knowledge of $\boldsymbol{\beta}^*$ and the mean μ. Here, we propose to roughly estimate $\boldsymbol{\beta}^*$ by a vector $\widehat{\mathbf{v}}$ first, use an empirical estimate $\overline{Y}$ of μ, and then obtain the ℓ_1 regularized least squares estimate on the augmented variable $\widetilde{Y} = (Y-\overline{Y})\boldsymbol{X}^\top\widehat{\mathbf{v}}$ to sharpen the convergence rate. At first glance it might appear counter-intuitive that introducing a noisy estimate of $\boldsymbol{\beta}^*$ can lead to consistent estimates, as the so-defined $\widetilde{Y}$ variable depends on the projection of $\boldsymbol{X}$ on $\mathrm{span}\{\boldsymbol{\beta}^*, \widehat{\mathbf{v}}\}$. Decompose

$$\widehat{\mathbf{v}} = (\widehat{\mathbf{v}}^\top\boldsymbol{\beta}^*)\boldsymbol{\beta}^* + \widehat{\boldsymbol{\beta}}^\perp, \quad (3.2)$$

where $\widehat{\boldsymbol{\beta}}^\perp \perp \boldsymbol{\beta}^*$. To better motivate this proposal, in the following we analyze the population least squares fit, based on the augmented variable $\check{Y} = (Y-\mu)\boldsymbol{X}^\top\widehat{\mathbf{v}}$ for some *fixed* unit vector $\widehat{\mathbf{v}}$ with decomposition (3.2). Writing out the population solution for least squares yields:

$$[\mathbb{E}\boldsymbol{X}^{\otimes 2}]^{-1}\mathbb{E}[\boldsymbol{X}\check{Y}] = \underbrace{\mathbb{E}[\boldsymbol{X}(Y-\mu)\boldsymbol{X}^\top(\widehat{\mathbf{v}}^\top\boldsymbol{\beta}^*)\boldsymbol{\beta}^*]}_{\boldsymbol{I}_1} + \underbrace{\mathbb{E}[\boldsymbol{X}(Y-\mu)\boldsymbol{X}^\top\widehat{\boldsymbol{\beta}}^\perp]}_{\boldsymbol{I}_2}. \quad (3.3)$$

We will now argue that left hand side of (3.3) is proportional to $\boldsymbol{\beta}^*$. First, we observe that $\boldsymbol{I}_1 = c_0(\widehat{\mathbf{v}}^\top\boldsymbol{\beta}^*)\boldsymbol{\beta}^*$, since multiplying by any vector $\mathbf{b} \perp \boldsymbol{\beta}^*$ yields $\mathbf{b}^\top\boldsymbol{I}_1 = 0$ by independence. Second, and perhaps more importantly, we have that $\boldsymbol{I}_2 = 0$. To see this, we first take a vector $\mathbf{b} \in \mathrm{span}\{\boldsymbol{\beta}^*, \widehat{\boldsymbol{\beta}}^\perp\}^\perp$. Since the three variables $\mathbf{b}^\top\boldsymbol{X}$, $Y-\mu$ and $\widehat{\boldsymbol{\beta}}^\perp\boldsymbol{X}$ are independent, we have $\mathbf{b}^\top\boldsymbol{I}_2 = 0$. Multiplying by $\boldsymbol{\beta}^*$ we have ${\boldsymbol{\beta}^*}^\top\boldsymbol{I}_2 = 0$ since ${\boldsymbol{\beta}^*}^\top\boldsymbol{X}(Y-\mu)$ is independent of $\boldsymbol{X}^\top\widehat{\boldsymbol{\beta}}^\perp$. Finally, multiplying by $\widehat{\boldsymbol{\beta}}^\perp$ yields $\boldsymbol{I}_2^\top\widehat{\boldsymbol{\beta}}^\perp = 0$, since $(\boldsymbol{X}^\top\widehat{\boldsymbol{\beta}}^\perp)^2$ is independent of $Y-\mu$.

(a) Initialization (b) Second Step

Figure 1: An illustration of the estimates $\widehat{\mathbf{v}}$ and $\widehat{\boldsymbol{\beta}}$ produced by the first and second steps of Algorithm 1. After the first step we can guarantee that the vector $\boldsymbol{\beta}^*$ belongs to one of two spherical caps which contain all vectors $\mathbf{w}$ such that $|\widehat{\mathbf{v}}^\top\mathbf{w}| \geq \kappa$ for some constant $\kappa > 0$, provided that the sample size $n \gtrsim s^2 \log d$ is sufficiently large. After the second step we can guarantee that the vector $\boldsymbol{\beta}^*$ belongs to one of two spherical caps in (b), which are shrinking with (n, s, d) at a faster rate.

It is noteworthy to mention that the above derivation crucially relies on the fact that the Y variable was centered, and the vector $\widehat{\mathbf{v}}$ was fixed. In what follows we formulate a pilot procedure which produces an estimate $\widehat{\mathbf{v}}$ such that $|\widehat{\mathbf{v}}^\top\boldsymbol{\beta}^*| \geq \kappa > 0$. A proper initialization algorithm can be achieved by using a spectral method, such as the Principal Hessian Directions (PHD) proposed by [22]. Cast into the framework of SIM, the PHD framework implies the following simple observation:

Lemma 3.1. If we have an MPR model, then $\mathrm{argmax}_{\|\mathbf{v}\|_2=1}\, \mathbf{v}^\top\mathbb{E}[Y(\boldsymbol{X}^{\otimes 2} - \mathbf{I})]\mathbf{v} = \pm\boldsymbol{\beta}^*$.

A proof of this fact can be found in Appendix C. Lemma 3.1 encourages us to look into the following sample version maximization problem

$$\mathrm{argmax}_{\|\mathbf{v}\|_2=1, \|\mathbf{v}\|_0=s} n^{-1}\mathbf{v}^\top\textstyle\sum_{i=1}^n [Y_i(\boldsymbol{X}_i^{\otimes 2} - \mathbf{I})]\mathbf{v}, \quad (3.4)$$

which targets a restricted (s-sparse) principal eigenvector. Unfortunately, solving such a problem is a computationally intensive task, and requires knowledge of s. Here we take a standard route of relaxing the above problem to a convex program, and solving it efficiently via semidefinite programming (SDP). A similar in spirit SDP relaxation for solving sparse PCA problems, was originally proposed by [13]. Instead of solving (3.4), define $\widehat{\boldsymbol{\Sigma}} = n^{-1}\sum_{i=1}^n Y_i(\boldsymbol{X}_i^{\otimes 2} - \mathbf{I})$, and solve the following convex

program:

$$\widehat{\mathbf{A}} = \mathrm{argmax}_{\mathrm{tr}(\mathbf{A})=1, \mathbf{A}\in\mathbb{S}_+^d} \mathrm{tr}(\widehat{\mathbf{\Sigma}}\mathbf{A}) - \lambda_n \sum_{i,j=1}^d |A_{ij}|, \tag{3.5}$$

where $\mathbb{S}_+^d$ is the convex cone of non-negative semidefinite matrices, and λ_n is a regularization parameter encouraging element-wise sparsity in the matrix $\mathbf{A}$. The hopes of introducing the optimization program above are that $\widehat{\mathbf{A}}$ will be a good first estimate of $\boldsymbol{\beta}^{*\otimes 2}$. In practice it could turn out that the matrix $\widehat{\mathbf{A}}$ is not rank one, hence we suggest taking $\widehat{\mathbf{v}}$ as the principal eigenvector of $\widehat{\mathbf{A}}$. In theory we show that with high probability the matrix $\widehat{\mathbf{A}}$ will indeed be of rank one. Observation (3.3), Lemma 3.1 and the SDP formulation motivate the agnostic two-step estimation procedure for misspecified phase retrieval in Algorithm 1.

Algorithm 1

input :$(Y_i, \boldsymbol{X}_i)_{i=1}^n$: data, λ_n, ν_n: tuning parameters

1. Split the sample into two approximately equal sets S_1, S_2, with $|S_1| = \lfloor n/2 \rfloor$, $|S_2| = \lceil n/2 \rceil$.
2. Let $\widehat{\mathbf{\Sigma}} := |S_1|^{-1} \sum_{i\in S_1} Y_i(\boldsymbol{X}_i^{\otimes 2} - \mathbf{I}_d)$. Solve (3.5). Let $\widehat{\mathbf{v}}$ be the first eigenvector of $\widehat{\mathbf{A}}$.
3. Let $\overline{Y} = |S_2|^{-1} \sum_{i\in S_2} Y_i$. Solve the following program:

$$\widehat{\mathbf{b}} = \mathrm{argmin}_{\mathbf{b}} (2|S_2|)^{-1} \sum_{i\in S_2} ((Y_i - \overline{Y})\boldsymbol{X}_i^\top \widehat{\mathbf{v}} - \boldsymbol{X}_i^\top \mathbf{b})^2 + \nu_n \|\mathbf{b}\|_1. \tag{3.6}$$

4. Return $\widehat{\boldsymbol{\beta}} := \widehat{\mathbf{b}}/\|\widehat{\mathbf{b}}\|_2$.

The sample split is required to ensure that after decomposition (3.2), the vector $\widehat{\boldsymbol{\beta}}^\perp$ and the value $\widehat{\mathbf{v}}^\top \boldsymbol{\beta}^*$ are independent of the remaining sample. In §3.1 we demonstrate that Algorithm 1 succeeds with optimal (in the noisy regime) ℓ_2 rate $\sqrt{s \log d/n}$, provided that $s^2 \log d \lesssim n$. The latter requirement on the sample size suffices to guarantee that the solution $\widehat{\mathbf{A}}$ of optimization program (3.5) is rank one. Figure 1 illustrates the two steps of Algorithm 1. In addition to our main procedure, we propose an optional refinement step (Algorithm 2) in which one applies steps 3. and 4. of Algorithm 1 on the full dataset using the output vector $\widehat{\boldsymbol{\beta}}$ of Algorithm 1. Doing so can potentially result in additional stability and further refinements of the rate constant.

Algorithm 2 Optional Refinement

input :$(Y_i, \boldsymbol{X}_i)_{i=1}^n$: data, ν_n': tuning parameter, output $\widehat{\boldsymbol{\beta}}$ from the Algorithm 1

5. Let $\overline{Y} = n^{-1} \sum_{i\in[n]} Y_i$. Solve the following program:

$$\widehat{\mathbf{b}} = \mathrm{argmin}_{\mathbf{b}} (2n)^{-1} \sum_{i=1}^n ((Y_i - \overline{Y})\boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}} - \boldsymbol{X}_i^\top \mathbf{b})^2 + \nu_n' \|\mathbf{b}\|_1. \tag{3.7}$$

6. Return $\widehat{\boldsymbol{\beta}}' := \widehat{\mathbf{b}}/\|\widehat{\mathbf{b}}\|_2$.

3.1 Theoretical Guarantees

In this section we present our main theoretical results, which consist of theoretical justification of our procedures, as well as lower bounds for certain types of SIM (1.3). To simplify the presentation for this section, we slightly change the notation and assume that the sample size is $2n$ and $S_1 = [n]$ and $S_2 = \{n+1, \dots, 2n\}$. Of course this abuse of notation does not restrict our analysis to only even sample size cases.

Our first result shows that the optimization program (3.5) succeeds in producing a vector $\widehat{\mathbf{v}}$ which is close to the vector $\boldsymbol{\beta}^*$.

Proposition 3.2. Assume that n is large enough so that $s\sqrt{\log d/n} < (1/6 - \kappa/4)c_0/(C_1 + C_2)$ for some small but fixed $\kappa > 0$ and constants C_1, C_2 (depending on f and ε). Then there exists a value of $\lambda_n \asymp \sqrt{\log d/n}$ such that the principal eigenvector $\widehat{\mathbf{v}}$ of $\widehat{\mathbf{A}}$, the solution of (3.5), satisfies

$$|\widehat{\mathbf{v}}^\top \boldsymbol{\beta}^*| \geq \kappa > 0,$$

with probability at least $1 - 4d^{-1} - O(n^{-1})$.

Proposition 3.2 shows that the first step of Algorithm 1 narrows down the search for the direction of $\boldsymbol{\beta}^*$ to a union of two spherical caps (i.e., the estimate $\widehat{\mathbf{v}}$ satisfies $|\widehat{\mathbf{v}}^\top \boldsymbol{\beta}^*| \geq \kappa$ for some constant $\kappa > 0$, see also Figure 1a). Our main result below, demonstrates that in combination with program (3.6) this suffices to recover the direction of $\boldsymbol{\beta}^*$ at an optimal rate with high probability.

Theorem 3.3. There exist values of $\lambda_n, \nu_n \asymp \sqrt{\log d/n}$ and a constant $R > 0$ depending on f and ε, such that if $s\sqrt{\log d/n} < R$ and $\log(d)\log^2(n)/n = o(1)$, the output of Algorithm 1 satisfies:

$$\sup_{\|\boldsymbol{\beta}^*\|_2=1, \|\boldsymbol{\beta}^*\|_0 \leq s} \mathbb{P}_{\boldsymbol{\beta}^*}\left(\min_{\eta\in\{1,-1\}} \|\widehat{\boldsymbol{\beta}} - \eta\boldsymbol{\beta}^*\|_2 > L\sqrt{\frac{s\log d}{n}}\right) \leq O(d^{-1} \vee n^{-1}), \tag{3.8}$$

where L is a constant depending solely on f and ε.

We remark that although the estimation rate is of the order $\sqrt{s\log d/n}$, our procedure still requires that $s\sqrt{\log d/n}$ is sufficiently small. This phenomenon is similar to what has been observed by [7], and it is our belief that this requirement cannot be relaxed for computationally feasible algorithms. We would further like to mention that while in bound (3.8) we control the worst case probability of failure, it is less clear whether the estimate $\widehat{\boldsymbol{\beta}}$ is universally consistent (i.e., whether the sup can be moved inside the probability in (3.8)).

4 Numerical Experiments

In this section we provide numerical experiments based on the three models (2.3), (2.4) and (2.5) where the random variable $\varepsilon \sim \mathcal{N}(0,1)$. All models are compared with the Truncated Power Method (TPM), proposed in [37]. For model (2.3) we also compare the results of our approach to the ones given by the TWF algorithm of [7]. Our setup is as follows. In all scenarios the vector $\boldsymbol{\beta}^*$ was held fixed at $\boldsymbol{\beta}^* = (\underbrace{-s^{-1/2}, s^{-1/2}, \ldots, s^{-1/2}}_{s}, \underbrace{0, \ldots 0}_{d-s})$. Since our theory requires that $n \gtrsim s^2\log d$, we have setup four different sample sizes $n \approx \theta s^2 \log d$, where $\theta \in \{4, 8, 12, 16\}$. We let s depend on the dimension d and we take $s \approx \log d$. In addition to the suggested approach in Algorithm 1, we also provide results using the refinement procedure (see Algorithm 3.7). We also provide the values of two "warm" starts of our algorithm, produced by solving program (3.5) with half and full data correspondingly. It is evident that for all scenarios the second step of Algorithms 1 and 2 outperform the warm start from SDP, except in Figure 2 (b), (c), when the sample size is simply two small to for the warm start on half of the data to be accurate. All values we report are based on an average over 100 simulations.

The SDP parameter was kept at a constant value (0.015) throughout all simulations, and we observed that varying this parameter had little influence on the final SDP solution. To select the ν_n parameter for (3.6) a pre-specified grid of parameters $\{\nu^1, \ldots, \nu^l\}$ was selected, and the following heuristic procedure based on K-fold cross-validation was used. We divide S_2 into $K = 5$ approximately equally sized non-intersecting sets $S_2 = \cup_{j\in[K]}\widetilde{S}_2^j$. For each $j \in [K]$ and $k \in [l]$ we run (3.6) on the set $\cup_{r\in[K], r\neq j}\widetilde{S}_2^r$ with a tuning parameter $\nu_n = \nu^k$ to obtain an estimate $\widehat{\boldsymbol{\beta}}_{k,-\widetilde{S}_2^j}$. Lemma 3.1 then justifies the following criteria to select the optimal index for selecting $\widehat{\nu}_n = \nu^{\widehat{l}}$ where

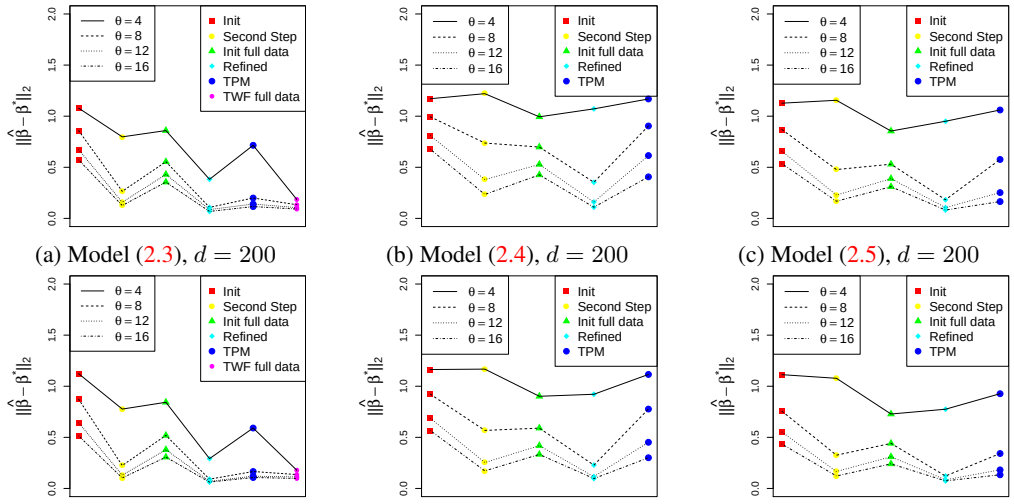
$$\widehat{l} = \operatorname*{argmax}_{k\in[l]} \sum_{j\in[K]} \sum_{i\in\widetilde{S}_2^j} Y_i(\boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}}_{k,-\widetilde{S}_2^j})^2.$$

Our experience suggests this approach works well in practice provided that the values $\{\nu^1, \ldots, \nu^l\}$ are selected within appropriate range and are of the magnitude $\sqrt{\log d/n}$.

Since the TPM algorithm requires an estimate of the sparsity s, we tuned it as suggested in Section 4.1.2 of [37]. In particular, for each scenario we considered the set of possible sparsities $K = \{s, 2s, 4s, 8s\}$. For each $k \in K$ the algorithm is ran on the first part of the data S_1, to obtain an estimate $\widehat{\boldsymbol{\beta}}_k$, and the final estimate is taken to be $\widehat{\boldsymbol{\beta}}_{\widehat{k}}$ where $\widehat{k}$ is given by

$$\widehat{k} = \operatorname*{argmax}_{k\in K} \widehat{\boldsymbol{\beta}}_k^\top |S_2|^{-1} \sum_{i\in S_2} Y_i(\boldsymbol{X}_i^{\otimes 2} - \mathbf{I}_d)\widehat{\boldsymbol{\beta}}_k.$$

The TPM is ran for 2000 iterations. In the case of phase retrieval, the TWF algorithm was also ran at a total number of 2000 iterations, using the tuning parameters originally suggested in [7]. As expected the TWF algorithm which targets the sparse phase retrieval model in particular outperforms our approach in the case when the sample size n is small, however our approach performs very comparatively to the TWF, and in fact even slightly better once we increase the sample size. It is possible that the TWF algorithm can perform better if it is ran for a longer than 2000 iterations, though in most cases it appeared to have converged to its final value. The results are visualized on Figure 2 above. The TPM algorithm, has performance comparable to that of Algorithm 1, is always



(a) Model (2.3), $d = 200$ (b) Model (2.4), $d = 200$ (c) Model (2.5), $d = 200$

(d) Model (2.3), $d = 400$ (e) Model (2.4), $d = 400$ (f) Model (2.5), $d = 400$

Figure 2: Simulation results for the three examples considered in §2, in two different settings for the dimension $d = 200, 400$. Here the parameter $\theta \approx \frac{n}{s^2 \log d}$ describes the relationship between sample size, dimension and sparsity of the signal. Algorithm 2 dominates in most settings, with exceptions when θ is too small, in which case none of the approaches provides meaningful results.

worse than the estimate produced by Algorithm 2, and it needs an initialization (for the first step of Algorithm 1 is used) and further requires a rough knowledge of the sparsity s, whereas both Algorithms 1 and 2 do not require an estimate of s.

5 Discussion

In this paper we proposed a two-step procedure for estimation of MPR models with standard Gaussian designs. We argued that the MPR models form a rich class including numerous additive SIMs (i.e., $Y = h(\boldsymbol{X}^\top \boldsymbol{\beta}^*) + \varepsilon$) with an even and increasing on $\mathbb{R}^+$ link function h. Our algorithm is based solely on convex optimization, and achieves optimal rates of estimation.

Our procedure does require that the sample size $n \gtrsim s^2 \log d$ to ensure successful initialization. The same condition has been exhibited previously, e.g., in [7] for the phase retrieval model, and in works on sparse principal components analysis [see, e.g., 3, 15, 33]. We anticipate that for a certain subclass of MPR models, the sample size requirement $n \gtrsim s^2 \log d$ is necessary for computationally efficient algorithms to exist. We conjecture that models (2.3)-(2.5) are such models. It is however certainly not true that this sample size requirement holds for all models from the MPR class. For example, the following model can be solved efficiently by applying the Lasso algorithm, without requiring the sample size scaling $n \gtrsim s^2 \log d$

$$Y = \mathrm{sign}(\boldsymbol{X}^\top \boldsymbol{\beta}^* + c),$$

where $c < 0$ is fixed. This discussion leads to the important question under what conditions of the (known) link and error distribution (f, ε) one can efficiently solve the SIM $Y = f(\boldsymbol{X}^\top \boldsymbol{\beta}^*, \varepsilon)$ with an optimal sample complexity. We would like to investigate this issue further in our future work.

Acknowledgments: The authors would like to thank the reviewers and meta-reviewers for carefully reading the manuscript and their helpful suggestions which improved the presentation. The authors would also like to thank Professor Xiaodong Li for kindly providing the code for the TWF algorithm.

References

[1] Adamczak, R. and Wolff, P. (2015). Concentration inequalities for non-Lipschitz functions with bounded derivatives of higher order. *Probability Theory and Related Fields*, **162** 531–586.

[2] Amini, A. A. and Wainwright, M. J. (2008). High-dimensional analysis of semidefinite relaxations for sparse principal components. In *IEEE International Symposium on Information Theory*.

[3] Berthet, Q. and Rigollet, P. (2013). Complexity theoretic lower bounds for sparse principal component detection. In *Conference on Learning Theory*.

[4] Bickel, P. J., Ritov, Y. and Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and dantzig selector. *The Annals of Statistics* 1705–1732.

[5] Boufounos, P. T. and Baraniuk, R. G. (2008). 1-bit compressive sensing. In *Annual Conference on Information Sciences and Systems*.

[6] Bühlmann, P. and van de Geer, S. (2011). *Statistics for high-dimensional data: Methods, theory and applications*. Springer.

[7] Cai, T. T., Li, X. and Ma, Z. (2015). Optimal rates of convergence for noisy sparse phase retrieval via thresholded Wirtinger flow. *arXiv:1506.03382*.

[8] Candès, E. J., Li, X. and Soltanolkotabi, M. (2015). Phase retrieval from coded diffraction patterns. *Applied and Computational Harmonic Analysis*, **39** 277–299.

[9] Candès, E. J., Li, X. and Soltanolkotabi, M. (2015). Phase retrieval via Wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, **61** 1985–2007.

[10] Candès, E. J., Strohmer, T. and Voroninski, V. (2013). Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming. *Communications on Pure and Applied Mathematics*, **66** 1241–1274.

[11] Chen, Y., Yi, X. and Caramanis, C. (2013). A convex formulation for mixed regression with two components: Minimax optimal rates. *arXiv:1312.7006*.

[12] Cook, R. D. and Ni, L. (2005). Sufficient dimension reduction via inverse regression. *Journal of the American Statistical Association*, **100**.

[13] d'Aspremont, A., El Ghaoui, L., Jordan, M. I. and Lanckriet, G. R. (2007). A direct formulation for sparse PCA using semidefinite programming. *SIAM review*, **49** 434–448.

[14] Ganti, R., Rao, N., Willett, R. M. and Nowak, R. (2015). Learning single index models in high dimensions. *arXiv preprint arXiv:1506.08910*.

[15] Gao, C., Ma, Z. and Zhou, H. H. (2014). Sparse CCA: Adaptive estimation and computational barriers. *arXiv:1409.8565*.

[16] Genzel, M. (2016). High-dimensional estimation of structured signals from non-linear observations with general convex loss functions. *arXiv:1602.03436*.

[17] Han, F. and Wang, H. (2015). Provable smoothing approach in high dimensional generalized regression model. *arXiv:1509.07158*.

[18] Horowitz, J. L. (2009). *Semiparametric and nonparametric methods in econometrics*. Springer.

[19] Laurent, B. and Massart, P. (2000). Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics* 1302–1338.

[20] Lecué, G. and Mendelson, S. (2013). Minimax rate of convergence and the performance of erm in phase recovery. *arXiv:1311.5024*.

[21] Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, **86** 316–327.

[22] Li, K.-C. (1992). On principal Hessian directions for data visualization and dimension reduction: Another application of Stein's lemma. *Journal of the American Statistical Association*, **87** 1025–1039.

[23] Li, K.-C. and Duan, N. (1989). Regression analysis under link violation. *The Annals of Statistics* 1009–1052.

[24] McCullagh, P. and Nelder, J. (1989). *Generalized linear models*. Chapman & Hall/CRC.

[25] Neykov, M., Liu, J. S. and Cai, T. (2016). L_1-regularized least squares for support recovery of high dimensional single index models with Gaussian designs. *Journal of Machine Learning Research*, **17** 1–37.

[26] Peng, H. and Huang, T. (2011). Penalized least squares for single index models. *Journal of Statistical Planning and Inference*, **141** 1362–1379.

[27] Plan, Y. and Vershynin, R. (2015). The generalized Lasso with non-linear observations. *IEEE Transactions on information theory*.

[28] Radchenko, P. (2015). High dimensional single index models. *Journal of Multivariate Analysis*, **139** 266–282.

[29] Raskutti, G., Wainwright, M. J. and Yu, B. (2010). Restricted eigenvalue properties for correlated Gaussian designs. *Journal of Machine Learning Research*, **11** 2241–2259.

[30] Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics* 1135–1151.

[31] Thrampoulidis, C., Abbasi, E. and Hassibi, B. (2015). Lasso with non-linear measurements is equivalent to one with linear measurements. *arXiv preprint, arXiv:1506.02181*.

[32] Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv:1011.3027*.

[33] Wang, Z., Gu, Q. and Liu, H. (2015). Sharp computational-statistical phase transitions via oracle computational model. *arXiv:1512.08861*.

[34] Xia, Y. and Li, W. (1999). On single-index coefficient regression models. *Journal of the American Statistical Association*, **94** 1275–1285.

[35] Yang, Z., Wang, Z., Liu, H., Eldar, Y. C. and Zhang, T. (2015). Sparse nonlinear regression: Parameter estimation and asymptotic inference. *arXiv; 1511:04514*.

[36] Yi, X., Wang, Z., Caramanis, C. and Liu, H. (2015). Optimal linear estimation under unknown nonlinear transform. In *Advances in Neural Information Processing Systems*.

[37] Yuan, X.-T. and Zhang, T. (2013). Truncated power method for sparse eigenvalue problems. *Journal of Machine Learning Research*, **14** 899–925.

A Notation

In addition to the notation defined in §1, throughout the appendices we use $\odot$ to denote the Hadamard (or element-wise) product, and dot product will sometimes be denoted with angle notation $\langle \cdot, \cdot \rangle$, to facilitate the display of long equations. For a matrix $\mathbf{A}$ we denote the max and ℓ_2 norms with $\|\mathbf{A}\|_{\max} = \max_{i,j} |A_{ij}|$ and $\|\mathbf{A}\|_2 = \sup_{\|\mathbf{v}\|_2=1} \|\mathbf{A}\mathbf{v}\|_2$ respectively. If $\mathbf{A}$ is symmetric we denote its spectrum ordered in decreasing manner by $\lambda_j(\mathbf{A})$.

B Auxiliary Results

Here we collect several results which we use in the later development.

Lemma B.1 (Lemma 5 [2])**.** Consider the following optimization program:

$$\widehat{\mathbf{Z}} = \underset{\mathrm{tr}(\mathbf{Z})=1, \mathbf{Z}\in\mathbb{S}_+^d}{\operatorname{argmax}} \mathrm{tr}(\mathbf{A}\mathbf{Z}) - \lambda_n \sum_{i,j=1}^{d} |Z_{ij}|, \tag{B.1}$$

where $\mathbb{S}_+^d$ is the set of all the $d \times d$ positive semi-definite matrices. Suppose there exists a matrix $\mathbf{U}$ (independent of $\widehat{\mathbf{z}}$) satisfying:

$$U_{ij} = \begin{cases} \operatorname{sign}(\widehat{z}_i)\operatorname{sign}(\widehat{z}_j), & \text{if } \widehat{z}_i\widehat{z}_j \neq 0; \\ \in [-1,1], & \text{otherwise.} \end{cases} \tag{B.2}$$

Then if $\widehat{\mathbf{z}}$ is the principal eigenvector of the matrix $\mathbf{A} - \lambda_n \mathbf{U}$, $\widehat{\mathbf{z}}\widehat{\mathbf{z}}^\top$ is the optimal solution to problem (B.1).

For convenience of the reader we briefly recall the notation and result on Gaussian concentration of non-Lipschitz functions used by [1], which we apply in Lemma E.8 below. For the set $[\ell]$, we denote with P_ℓ the set of its partitions into non-empty and non-intersecting disjoint sets. For a partition $\mathcal{J} = \{J_1, \ldots, J_k\}$, and an ℓ-indexed matrix $\mathbf{A} = (a_{\mathbf{i}})_{\mathbf{i}\in[n]^\ell}$, define the norm:

$$\|\mathbf{A}\|_{\mathcal{J}} = \sup\Big\{ \sum_{\mathbf{i}\in[n]^\ell} a_{\mathbf{i}} \prod_{l=1}^{k} x^{(l)}_{\mathbf{i}_{J_l}} : \|x^{(l)}_{\mathbf{i}_{J_l}}\|_2 \leq 1, 1 \leq l \leq k \Big\},$$

where the indexing should be understood as $\mathbf{i}_I := (i_k)_{k\in I}$. Given the convention that $\#\mathcal{J} = |\mathcal{J}|$ is the cardinality of the set $\mathcal{J}$ we restate (a shortened) version of Theorem 1.4 of [1].

Theorem B.2 (Theorem 1.4 [1])**.** Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a random vector with independent components, such that for all $i \leq n$, $\|X_i\|_{\psi_2} \leq \Upsilon$. Then for every polynomial $f : \mathbb{R}^n \mapsto \mathbb{R}$ of degree L and every $p \geq 2$ we have:

$$\|f(X) - \mathbb{E}f(X)\|_p \leq K_L \sum_{\ell=1}^{L} \Upsilon^\ell \sum_{\mathcal{J}\in P_\ell} p^{\#\mathcal{J}/2} \|\mathbb{E}\mathbf{D}^\ell f(X)\|_{\mathcal{J}}.$$

Here $\mathbf{D}^\ell$ is the ℓ^{th} derivative of f.

C Preliminary Proofs

Proof of Proposition 2.2. We have the following equality:

$$c_0 = \mathrm{Cov}(f(Z,\varepsilon), Z^2) = \mathbb{E}[\varphi(Z)Z^2] - \mathbb{E}\varphi(Z) = \mathbb{E}D^2\varphi(Z) > 0,$$

where the last (and key) equation follows by Stein's Lemma [see, e.g., Lemma 4 of 30]. □

Proof of Lemma 3.1. First of all we notice that:

$$\mathbb{E}[Y(\boldsymbol{X}^{\otimes 2} - \mathbf{I})] = \mathbb{E}[(Y-\mu)(\boldsymbol{X}^{\otimes 2} - \mathbf{I})] = \mathbb{E}(Y-\mu)\boldsymbol{X}^{\otimes 2}.$$

Hence proving Lemma 3.1 is equivalent to showing:

$$\boldsymbol{\beta}^* = \underset{\|\mathbf{v}\|_2=1}{\operatorname{argmax}} \mathbb{E}[(Y-\mu)(\mathbf{v}^\top \boldsymbol{X})^2].$$

Next, decompose:

$$\mathbf{v}^\top \boldsymbol{X} = (\mathbf{v}^\top \boldsymbol{\beta}^*)\boldsymbol{\beta}^{*\top}\boldsymbol{X} + (\mathbf{v} - (\mathbf{v}^\top\boldsymbol{\beta}^*)\boldsymbol{\beta}^{*\top})\boldsymbol{X} = (\mathbf{v}^\top\boldsymbol{\beta}^*)\boldsymbol{\beta}^{*\top}\boldsymbol{X} + \boldsymbol{\beta}^{\perp\top}\boldsymbol{X},$$

where we used the shorthand notation $\boldsymbol{\beta}^{\perp} := \mathbf{v} - (\mathbf{v}^{\top}\boldsymbol{\beta}^*)\boldsymbol{\beta}^*$, for the vector $\boldsymbol{\beta}^{\perp}$ which is orthogonal to $\boldsymbol{\beta}$. In terms of this notation, we have the following identity:

$$\begin{aligned}\mathbb{E}[(Y-\mu)(\mathbf{v}^{\top}\boldsymbol{X})^2] &= (\mathbf{v}^{\top}\boldsymbol{\beta}^*)^2\mathbb{E}[(Y-\mu)(\boldsymbol{\beta}^{*\top}\boldsymbol{X})^2] \\ &\quad + 2(\mathbf{v}^{\top}\boldsymbol{\beta}^*)\mathbb{E}[(Y-\mu)(\boldsymbol{\beta}^{*\top}\boldsymbol{X})(\boldsymbol{\beta}^{\perp\top}\boldsymbol{X})] + \mathbb{E}[(Y-\mu)(\boldsymbol{\beta}^{\perp\top}\boldsymbol{X})^2].\end{aligned}$$

We next deal with the last two terms of the above decomposition. Since $\boldsymbol{\beta}^{\perp\top}\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{\beta}^{*\top}\boldsymbol{X}$ we have that the second term:

$$\mathbb{E}[(Y-\mu)(\boldsymbol{\beta}^{*\top}\boldsymbol{X})(\boldsymbol{\beta}^{\perp\top}\boldsymbol{X})] = \mathbb{E}[(Y-\mu)(\boldsymbol{\beta}^{*\top}\boldsymbol{X})]\mathbb{E}[\boldsymbol{\beta}^{\perp\top}\boldsymbol{X}] = 0.$$

For the third term due to the same independence ($\boldsymbol{\beta}^{\perp\top}\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{\beta}^{*\top}\boldsymbol{X}$) we have:

$$\mathbb{E}[(Y-\mu)(\boldsymbol{\beta}^{\perp\top}\boldsymbol{X})^2] = \mathbb{E}[(Y-\mu)]\mathbb{E}[(\boldsymbol{\beta}^{\perp\top}\boldsymbol{X})^2] = 0.$$

Hence:

$$\mathbb{E}[(Y-\mu)(\mathbf{v}^{\top}\boldsymbol{X})^2] = (\mathbf{v}^{\top}\boldsymbol{\beta}^*)^2\mathbb{E}[(Y-\mu)(\boldsymbol{\beta}^{*\top}\boldsymbol{X})^2] = (\mathbf{v}^{\top}\boldsymbol{\beta}^*)^2 c_0.$$

Since (2.1) implies that $c_0 > 0$, by Cauchy-Schwartz the maximizer of the above expression is $\mathbf{v} = \pm\boldsymbol{\beta}^*$. □

Proof of Proposition 2.3. We prove the three statements in turn.

(i) Let Z' be an independent copy of Z. We have the following chain of equalities:

$$c_0 = \mathbb{E}(\varphi(Z) - \mathbb{E}\varphi(Z))Z^2 = \mathbb{E}(\varphi(Z) - \mathbb{E}\varphi(Z'))Z^2 = \mathbb{E}(\varphi(Z) - \varphi(Z'))Z^2.$$

By symmetry one also has: $c_0 = \mathbb{E}(\varphi(Z') - \varphi(Z))(Z')^2$. Adding the last two equations yields:

$$\begin{aligned}2c_0 &= \mathbb{E}[(\varphi(Z) - \varphi(Z'))(Z^2 - (Z')^2)] \\ &= \mathbb{E}_{X,X'\sim|\mathcal{N}(0,1)|}[(\varphi(X) + \varphi(-X) - (\varphi(X') + \varphi(-X')))(X^2 - (X')^2)]/2 \\ &> 0,\end{aligned}$$

where we used the fact that $\operatorname{sign}(\varphi(X) + \varphi(-X) - (\varphi(X') + \varphi(-X'))) = \operatorname{sign}(X^2 - (X')^2)$. The last inequality is strict since by our condition the integrand is strictly positive on the set $[0, z_2] \times [z_1, \infty) \subset \mathbb{R}^2$ which is a set of positive Lebesgue measure.

(ii) To see (ii) for any two points $x < y$, take $v_x \in \partial\varphi(x)$ and $v_y \in \partial\varphi(y)$ to be arbitrary points in the corresponding sub-differentials. Adding the following two inequalities:

$$\varphi(x) - \varphi(y) \geq v_y(x-y), \quad \varphi(y) - \varphi(x) \geq v_x(y-x),$$

we conclude that $(v_x - v_y)(x-y) \geq 0$. Notice that since φ is convex, by Jensen's inequality, $\varphi(z) + \varphi(-z) \geq 2\varphi(0)$ for any $z \geq 0$. Next take $z > z' > 0$, and consider the difference:

$$\varphi(z) + \varphi(-z) - \varphi(z') - \varphi(-z') \geq (v_{z'} - v_{-z'})(z - z') \geq 0,$$

where $v_{z'} \in \partial\varphi(z')$ and $v_{-z'} \in \partial\varphi(-z')$ are arbitrary sub-gradients, and the last inequality follows by the fact that $z' > 0$ and hence $v_{z'} \geq v_{-z'}$ as we verified before. The above inequality becomes strict whenever $z' \geq z_0$ since $v_{z'} - v_{-z'}$ is non-decreasing and by assumption:

$$z_0 v_{z_0} \geq \varphi(z_0) - \varphi(0) > \varphi(0) - \varphi(-z_0) \geq v_{-z_0} z_0,$$

and hence $v_{z_0} - v_{-z_0} > 0$. Hence we may take $z_1 = 2z_0$ and $z_2 = z_0$ in (i) to complete the proof.

(iii) Statement (iii) is an implication of the fact that we can control the tail bound of $g_1(Z)$. Notice that when $0 < t \leq \max\{|a|, |b|\}$ we trivially have $\mathbb{P}(|g_1(Z)| \geq t) \leq 1$. When $t > \max\{|a|, |b|\}$ using our assumption, by a standard normal tail bound we have:

$$\mathbb{P}(|g_1(Z)| \geq t) \leq \mathbb{P}(|Z| \geq \sqrt{t/C}) \leq 2\exp(-t/(2C)) \leq \exp(1 - t/(2C)).$$

Hence setting $K = \max\{|a|, |b|, 2C\}$ shows that in any case $\mathbb{P}(|g_1(Z)| \geq t) \leq \exp(1 - t/K)$, which shows that $\|g_1(Z)\|_{\psi_1} \leq cK < \infty$ for some absolute constant c. Finally by the triangle inequality we conclude:

$$\|g_1(Z)\|_{\psi_1} + \|g_2(\varepsilon)\|_{\psi_1} < \infty,$$

which completes the proof.

□

Lemma C.1. If h is a convex function such that $h(z_0) + h(-z_0) > 2h(0)$ for some $z_0 > 0$, $\mathbb{E}|h(z + \varepsilon)| < \infty$ for every $z \in \mathbb{R}$, and $\mathbb{E}|h(Z + \varepsilon)| < \infty$ we have $\varphi(z) = \mathbb{E}h(z + \varepsilon)$ is convex, sub-differentiable and there exists a $z_0' > 0$ such that $\varphi(z_0') + \varphi(-z_0') > 2\varphi(0)$.

Proof of Lemma C.1. Since the function h is convex and the expectation is a linear operator it follows that $\varphi(z)$ is indeed convex. The linearity of the expectation operator, coupled with the fact that the function $|h(z+\varepsilon)|$ is integrable for all z, additionally implies that $\varphi(z)$ is sub-differentiable with $\mathbb{E}\overline{\partial}_\varepsilon h(z+\varepsilon) \in \partial\varphi(z)$[2], where $\overline{\partial}_\varepsilon h(z+\varepsilon) \in \partial h(z+\varepsilon)$ is chosen so that $\varepsilon \mapsto \overline{\partial}_\varepsilon h(z+\varepsilon)$ is a function[3]. Next, notice that for any fixed ε, we have:

$$\mathbb{E}_Z[h(Z+\varepsilon)+h(-Z+\varepsilon)] > 2h(\varepsilon).$$

The last inequality is strict, since by Jensen's inequality $\mathbb{E}_Z[h(Z+\varepsilon)+h(-Z+\varepsilon)] \geq h(\mathbb{E}Z+\varepsilon)+h(-\mathbb{E}Z+\varepsilon) = 2h(\varepsilon)$, and equality can be achieved only when h is linear, which is not the case since $h(z_0)+h(-z_0) > 2h(0)$ by assumption. Take an expectation with respect to ε and exchange the expectations (by Fubini's theorem, recall that $\mathbb{E}|h(Z+\varepsilon)| < \infty$) to obtain:

$$\mathbb{E}_Z\mathbb{E}_\varepsilon[h(Z+\varepsilon)+h(-Z+\varepsilon)] > 2\mathbb{E}_\varepsilon h(\varepsilon).$$

Naturally, the above implies the existence of z_0' such that:

$$\mathbb{E}_\varepsilon[h(z_0'+\varepsilon)+h(-z_0'+\varepsilon)] > 2\mathbb{E}_\varepsilon h(\varepsilon),$$

and completes the proof. □

D Proofs for Initialization Step

Proof of Proposition 3.2. The proof follows by an application of Lemma B.1 and Lemma D.2. □

Lemma D.1. Let $\mathbf{A} = a\mathbf{v}\mathbf{v}^\top - b\mathbf{w}\mathbf{w}^\top$ be a symmetric rank two matrix, with $a > b \geq 0$ and $\|\mathbf{v}\|_2 = \|\mathbf{w}\|_2 = 1$, and let $\mathbf{N}$ be a symmetric noise matrix. Then, assuming that $\|\mathbf{N}\|_2 \leq \frac{a-b}{2}$ the principal eigenvector $\widehat{\mathbf{v}}$ of $\mathbf{A}+\mathbf{N}$ satisfies:

$$|\widehat{\mathbf{v}}^\top\mathbf{v}| \geq \Big[\frac{a-b-2\|\mathbf{N}\|_2}{a}\Big]^{1/2}$$

Proof of Lemma D.1. First off, an elementary calculation shows that the non-zero spectrum of $\mathbf{A}$ is:

$$\{\lambda_1(\mathbf{A}), \lambda_d(\mathbf{A})\} = \Big\{\frac{a-b\pm\sqrt{(a-b)^2+4ab(1-\mathbf{v}^\top\mathbf{w})}}{2}\Big\}.$$

Next we have:

$$a(\mathbf{v}^\top\widehat{\mathbf{v}})^2 + \|\mathbf{N}\|_2 \geq \widehat{\mathbf{v}}^\top(\mathbf{A}+\mathbf{N})\widehat{\mathbf{v}} \geq \lambda_1(\mathbf{A}) - \|\mathbf{N}\|_2,$$

and hence:

$$(\mathbf{v}^\top\widehat{\mathbf{v}})^2 \geq \frac{a-b+\sqrt{(a-b)^2+4ab(1-\mathbf{v}^\top\mathbf{w})}}{2a} - 2\frac{\|\mathbf{N}\|_2}{a} \tag{D.1}$$

$$\geq \frac{a-b-2\|\mathbf{N}\|_2}{a}, \tag{D.2}$$

where the last inequality follows by Cauchy-Schwartz. □

Lemma D.2. Assume that n is large enough so that $s\sqrt{\frac{\log d}{n}} < (\frac{1}{6}-\frac{\kappa}{4})\frac{c_0}{(C_1+C_2)}$ for some small but fixed $\kappa > 0$ and constants C_1, C_2 as defined in Lemmas D.6 and D.7. Put $\lambda_n = (C_1+C_2)\sqrt{\frac{\log d}{n}}$. There exists a sign matrix $\widehat{\mathbf{U}}$ with $\widehat{\mathbf{U}}_{S_{\boldsymbol{\beta}^*}S_{\boldsymbol{\beta}^*}} = \operatorname{sign}(\boldsymbol{\beta}^*_{S_{\boldsymbol{\beta}^*}})\operatorname{sign}(\boldsymbol{\beta}^*_{S_{\boldsymbol{\beta}^*}})^\top$ such that the principal eigenvector of $\widehat{\boldsymbol{\Sigma}} - \lambda\widehat{\mathbf{U}}$, $\widehat{\mathbf{v}}$ satisfies:

$$|\widehat{\mathbf{v}}^\top\boldsymbol{\beta}^*| \geq \kappa,$$

with probability at least $1-4d^{-1}-O(n^{-1})$.

Remark D.3. The proof of Lemma D.2 also shows that with high probability the vector $\widehat{\mathbf{v}}$ can be identified with a vector $\widetilde{\mathbf{v}}$ (the principal eigenvector of $\widehat{\boldsymbol{\Sigma}}_{S_{\boldsymbol{\beta}^*},S_{\boldsymbol{\beta}^*}} - \lambda_n\widehat{\mathbf{U}}_{S_{\boldsymbol{\beta}^*},S_{\boldsymbol{\beta}^*}}$ see below) which is independent of the data $\mathbf{X}_{S^c_{\boldsymbol{\beta}^*}}$ such that $\widehat{\mathbf{v}} \equiv \widetilde{\mathbf{v}}$ with high probability. This becomes evident upon realizing that the matrix $\mathbf{N}_{S_{\boldsymbol{\beta}^*}S_{\boldsymbol{\beta}^*}}$ depends solely on $\mathbf{X}_{S_{\boldsymbol{\beta}^*}}$. In addition it is evident that the support $\operatorname{supp}(\widehat{\mathbf{v}}) \subset S_{\boldsymbol{\beta}^*}$ and $\operatorname{supp}(\widetilde{\mathbf{v}}) \subset S_{\boldsymbol{\beta}^*}$.

[2]The fact that $\mathbb{E}\overline{\partial}_\varepsilon h(z+\varepsilon)$ exists is implied by $\mathbb{E}|h(z+\varepsilon)| < \infty$.

[3]Note also that $\varepsilon \mapsto \overline{\partial}_\varepsilon h(z+\varepsilon)$ is monotone and hence measurable.

Proof of Lemma D.2. Setting $\lambda_n = (C_1 + C_2)\sqrt{\frac{\log d}{n}}$ and using Lemma D.4 gives us that

$$\widehat{\mathbf{\Sigma}} - \lambda_n \widehat{\mathbf{U}} = \left[\begin{array}{c|c} \eta \boldsymbol{\beta}^*_{S_{\boldsymbol{\beta}^*}} \boldsymbol{\beta}^{*\top}_{S_{\boldsymbol{\beta}^*}} - \lambda_n \operatorname{sign}(\boldsymbol{\beta}^*_{S_{\boldsymbol{\beta}^*}}) \operatorname{sign}(\boldsymbol{\beta}^*_{S_{\boldsymbol{\beta}^*}})^\top + \mathbf{N}_{S_{\boldsymbol{\beta}^*} S_{\boldsymbol{\beta}^*}} & \mathbf{N}_{S_{\boldsymbol{\beta}^*} S^c_{\boldsymbol{\beta}^*}} - \lambda_n \widehat{\mathbf{U}}_{S_{\boldsymbol{\beta}^*} S^c_{\boldsymbol{\beta}^*}} \\ \hline \mathbf{N}_{S^c_{\boldsymbol{\beta}^*} S_{\boldsymbol{\beta}^*}} - \lambda_n \widehat{\mathbf{U}}_{S^c_{\boldsymbol{\beta}^*} S_{\boldsymbol{\beta}^*}} & \mathbf{N}_{S^c_{\boldsymbol{\beta}^*} S^c_{\boldsymbol{\beta}^*}} - \lambda_n \widehat{\mathbf{U}}_{S^c_{\boldsymbol{\beta}^*} S^c_{\boldsymbol{\beta}^*}} \end{array}\right].$$

We can select the sign matrix $\widehat{\mathbf{U}}$ such that all three terms which do not correspond to the $S_{\boldsymbol{\beta}^*} S_{\boldsymbol{\beta}^*}$ "corner" of the above visualization are $\equiv 0$, since by Lemma D.4 we have that $\|\mathbf{N}\|_{\max} \leq (C_1 + C_2)\sqrt{\frac{\log d}{n}} \leq \lambda_n$ with high probability. Recall that by our assumption on the sample size we have $\lambda_n \leq \frac{c_0}{6s}$ and hence $\lambda_n \widehat{\mathbf{U}}_{S^c_{\boldsymbol{\beta}^*} S_{\boldsymbol{\beta}^*}} \leq \frac{c_0}{6} \frac{\operatorname{sign}(\boldsymbol{\beta}^*_{S_{\boldsymbol{\beta}^*}})}{\sqrt{s}} \frac{\operatorname{sign}(\boldsymbol{\beta}^*_{S_{\boldsymbol{\beta}^*}})^\top}{\sqrt{s}}$. Using Lemmas D.1 and D.4 on the event $\|\mathbf{N}_{S_{\boldsymbol{\beta}^*} S_{\boldsymbol{\beta}^*}}\|_2 \leq (C_1 + C_2) s \sqrt{\frac{\log d}{n}}$ we have:

$$\begin{aligned} |\widehat{\mathbf{v}}^\top_{S_{\boldsymbol{\beta}^*}} \boldsymbol{\beta}^*_{S_{\boldsymbol{\beta}^*}}| &\geq \frac{\eta - \frac{c_0}{6} - 2\|\mathbf{N}_{S_{\boldsymbol{\beta}^*} S_{\boldsymbol{\beta}^*}}\|_2}{\eta} \geq \\ &\geq 1 - \frac{1}{3} - 4\|\mathbf{N}_{S_{\boldsymbol{\beta}^*} S_{\boldsymbol{\beta}^*}}\|_2 / c_0 \geq \frac{2}{3} - 4\frac{(C_1 + C_2)}{c_0} s\sqrt{\frac{\log d}{n}} \geq \kappa, \end{aligned}$$

for values of n large enough so that the above expression is positive, which concludes the proof. □

Lemma D.4. We have that:

$$\widehat{\mathbf{\Sigma}} = \eta \boldsymbol{\beta}^* \boldsymbol{\beta}^{*\top} + \mathbf{N}, \tag{D.3}$$

where $\eta > c_0/2$ and $\|\mathbf{N}_{S_{\boldsymbol{\beta}^*} S_{\boldsymbol{\beta}^*}}\|_2 \leq (C_1 + C_2) s \sqrt{\frac{\log d}{n}}$ and $\|\mathbf{N}\|_{\max} \leq (C_1 + C_2)\sqrt{\frac{\log d}{n}}$ with probability at least $1 - 4d^{-1} - O(n^{-1})$, where C_1 and C_2 are constants depending on f, ε.

Proof of Lemma D.4. First we observe that decomposition (D.3) holds with:

$$\begin{aligned} \eta \boldsymbol{\beta}^* \boldsymbol{\beta}^{*\top} &= \frac{1}{n}\sum_{i=1}^n Y_i (\boldsymbol{\beta}^{*\top} \boldsymbol{X}_i)^2 \boldsymbol{\beta}^* \boldsymbol{\beta}^{*\top} + \frac{1}{n}\sum_{i=1}^n Y_i (\mathbf{P}_{\boldsymbol{\beta}^{*\perp}} - \mathbf{I}_d), \\ \mathbf{N} &= \frac{1}{n}\sum_{i=1}^n Y_i (\boldsymbol{\beta}^{*\top} \boldsymbol{X}_i)(\boldsymbol{\beta}^* \boldsymbol{X}_i^\top \mathbf{P}_{\boldsymbol{\beta}^{*\perp}} + \mathbf{P}_{\boldsymbol{\beta}^{*\perp}} \boldsymbol{X}_i \boldsymbol{\beta}^{*\top}) + \frac{1}{n}\sum_{i=1}^n Y_i [\mathbf{P}_{\boldsymbol{\beta}^{*\perp}} (\boldsymbol{X}_i^{\otimes 2} - \mathbf{I}_d) \mathbf{P}_{\boldsymbol{\beta}^{*\perp}}], \end{aligned}$$

where $\mathbf{P}_{\boldsymbol{\beta}^{*\perp}} = (\mathbf{I}_d - \boldsymbol{\beta}^* \boldsymbol{\beta}^{*\top})$. Lemma D.5 shows that $\eta \geq c_0/2$ with probability at least $1 - O(n^{-1})$. Next, Lemma D.6 and Lemma D.7 show that:

$$\|\mathbf{N}\|_{\max} \leq (C_1 + C_2)\sqrt{\frac{\log d}{n}},$$

with probability at least $1 - 4d^{-1} - O(n^{-1})$, where the constants C_1 and C_2 depend on f, ε. Using the fact that $\|\mathbf{N}_{S_{\boldsymbol{\beta}^*} S_{\boldsymbol{\beta}^*}}\|_2 \leq \|\mathbf{N}_{S_{\boldsymbol{\beta}^*} S_{\boldsymbol{\beta}^*}}\|_1 \leq s\|\mathbf{N}\|_{\max}$ completes the proof. □

Lemma D.5. We have that η defined in (D.3) satisfies

$$\eta \geq c_0/2,$$

with probability at least $1 - \frac{4 \operatorname{Var}[f(Z,\varepsilon)(Z^2-1)]}{c_0^2} n^{-1}$.

Proof of Lemma D.5. Grouping the first two terms by Chebyshev's inequality we have that:

$$\mathbb{P}\Big(\Big|\frac{1}{n}\sum_{i=1}^n Y_i ((\boldsymbol{\beta}^{*\top} \boldsymbol{X}_i)^2 - 1) - c_0\Big| \geq t\Big) \leq \frac{\operatorname{Var}[f(Z,\varepsilon)(Z^2-1)]}{nt^2}.$$

Notice that in the last inequality we have $\operatorname{Var}[(f(Z,\varepsilon)(Z^2-1)] < \infty$ since we are assuming that $f(Z,\varepsilon)$ is sub-exponential. Setting $t = c_0/2$ brings the above probability bound to zero at a rate n^{-1}.

□

Lemma D.6. We have that:

$$\Big\|\frac{1}{n}\sum_{i=1}^n Y_i (\boldsymbol{\beta}^{*\top} \boldsymbol{X}_i)[\boldsymbol{\beta}^* \boldsymbol{X}_i^\top \mathbf{P}_{\boldsymbol{\beta}^{*\perp}} + \mathbf{P}_{\boldsymbol{\beta}^{*\perp}} \boldsymbol{X}_i \boldsymbol{\beta}^{*\top}]\Big\|_{\max} \leq C_1 \sqrt{\frac{\log d}{n}},$$

where C_1 is a constant depending on f, ε, with probability at least $1 - 2d^{-1} - \frac{\mathrm{Var}[f^2(Z,\varepsilon)Z^2]}{(\mathbb{E}[f^2(Z,\varepsilon)Z^2])^2} n^{-1}$.

Proof of Lemma D.6. We will only deal with the first term of the sum, as the second term follows by the same argument after transposition. First notice that $Y_i(\boldsymbol{\beta}^{*\top}\boldsymbol{X}_i) \perp\!\!\!\perp \boldsymbol{X}_i^\top \mathbf{P}_{\boldsymbol{\beta}^{*\perp}}$. Analyzing the first part of this term row-wise, for $j \in S_{\boldsymbol{\beta}^*}$ ($\beta_j^* \neq 0$) we have:

$$\frac{1}{n}\sum_{i=1}^n Y_i \beta_j^* (\boldsymbol{\beta}^{*\top}\boldsymbol{X}_i)\boldsymbol{X}_i^\top \mathbf{P}_{\boldsymbol{\beta}^{*\perp}} \sim \mathcal{N}\Big(0, \beta_j^{*2}\frac{1}{n^2}\sum_{i=1}^n Y_i^2 (\boldsymbol{\beta}^{*\top}\boldsymbol{X}_i)^2 \mathbf{P}_{\boldsymbol{\beta}^{*\perp}}\Big).$$

By Chebyshev's inequality we have:

$$\mathbb{P}\Big(\Big|\frac{1}{n}\sum_{i=1}^n Y_i^2(\boldsymbol{\beta}^{*\top}\boldsymbol{X}_i)^2 - \mathbb{E}[f^2(Z,\varepsilon)Z^2]\Big| \geq t\Big) \leq \frac{\mathrm{Var}[f^2(Z,\varepsilon)Z^2]}{nt^2},$$

assuming $\mathrm{Var}[f^2(Z,\varepsilon)Z^2] < \infty$. Putting $t = \mathbb{E}[f^2(Z,\varepsilon)Z^2]$ brings the above probability converge to zero at a rate n^{-1}. Hence, by a conditioning argument, a standard normal tail bound coupled with the facts that $\|\mathbf{P}_{\boldsymbol{\beta}^{*\perp}}\|_2 \leq 1, |\beta_j^*| \leq 1$ and a union bound, we obtain:

$$\mathbb{P}\Big(\max_{j\in[d]}\Big\|\frac{1}{n}\sum_{i=1}^n Y_i(\boldsymbol{\beta}^{*\top}\boldsymbol{X}_i)\beta_j^*\boldsymbol{X}_i^\top \mathbf{P}_{\boldsymbol{\beta}^{*\perp}}\Big\|_\infty > t\Big) \leq 2d^2 \exp\Big(-\frac{nt^2}{4\mathbb{E}[f^2(Z,\varepsilon)Z^2]}\Big),$$

Plugging in

$$t = 2\sqrt{3\mathbb{E}f^2(Z,\varepsilon)Z^2}\sqrt{\frac{\log d}{n}},$$

brings the probability to $2d^{-1}$. This completes the proof with $C_1 = 4\sqrt{3\mathbb{E}f^2(Z,\varepsilon)Z^2}$. □

Lemma D.7. Let $Y_i = f(\boldsymbol{X}_i^\top\boldsymbol{\beta}^*, \varepsilon)$, where $\boldsymbol{X}_i \sim \mathcal{N}(0, \mathbf{I})$. Assume that f and ε are such that $\|f(Z,\varepsilon)\|_{\psi_1} \leq K$ for $Z \sim \mathcal{N}(0,1)$ and $Z \perp\!\!\!\perp \varepsilon$, and in addition let $\log d = o(n/\log^2 n)$. Then:

$$\Big\|\frac{1}{n}\sum_{i=1}^n Y_i \mathbf{P}_{\boldsymbol{\beta}^{*\perp}}(\boldsymbol{X}_i^{\otimes 2} - \mathbf{I}_d)\mathbf{P}_{\boldsymbol{\beta}^{*\perp}}\Big\|_{\max} \leq \sqrt{\frac{C_2 \log d}{n}},$$

with probability at least $1 - 2d^{-1} - (2^{11}K^4/(\mathbb{E}f^2(Z,\varepsilon))^2 + e)n^{-1}$, for some absolute value C_2 depending on K, and large values of n.

Proof of Lemma D.7. Notice that by the properties of the multivariate normal distribution one has that $Y_i \perp\!\!\!\perp \mathbf{P}_{\boldsymbol{\beta}^{*\perp}}(\boldsymbol{X}_i^{\otimes 2} - \mathbf{I}_d)\mathbf{P}_{\boldsymbol{\beta}^{*\perp}}$. Next we have that $\boldsymbol{Z}_i := \mathbf{P}_{\boldsymbol{\beta}^{*\perp}}\boldsymbol{X}_i \sim \mathcal{N}(0, \mathbf{P}_{\boldsymbol{\beta}^{*\perp}})$, and thus, since $\|\mathbf{P}_{\boldsymbol{\beta}^{*\perp}}\|_2 \leq 1$, we have that each individual entry of $\boldsymbol{Z}_i$ is a normally distributed random variable with variance at most one. Hence we have that for any $j, k \in [d]$: $\|\boldsymbol{Z}_{ij}\boldsymbol{Z}_{ik}\|_{\psi_1} \leq 2\|\boldsymbol{Z}_{ij}\|_{\psi_2}\|\boldsymbol{Z}_{ik}\|_{\psi_2} \leq 2$, and hence conditionally on Y_i one has

$$\|\boldsymbol{Z}_{ij}\boldsymbol{Z}_{ik} - \mathbb{E}\boldsymbol{Z}_{ij}\boldsymbol{Z}_{ik}\|_{\psi_1} = \|\boldsymbol{Z}_{ij}\boldsymbol{Z}_{ik} - \mathbf{P}_{\boldsymbol{\beta}^{*\perp},jk}\|_{\psi_1} \leq 4,$$

for all $j, k \in [d]$. Next conditionally on the Y_i values and a Bernstein type of inequality (see, e.g., Proposition 5.16 of [32]) we obtain:

$$\mathbb{P}\Big(\Big\|\frac{1}{n}\sum_{i=1}^n Y_i \mathbf{P}_{\boldsymbol{\beta}^{*\perp}}(\boldsymbol{X}_i^{\otimes 2} - \mathbf{I}_d)\mathbf{P}_{\boldsymbol{\beta}^{*\perp}}\Big\|_{\max} \geq t\Big) \tag{D.4}$$

$$\leq 2d^2 \exp\Big[-c\min\Big(\frac{nt^2}{16n^{-1}\sum_{i=1}^n Y_i^2}, \frac{nt}{4\max_{i\in[n]}|Y_i|}\Big)\Big], \tag{D.5}$$

for an absolute constant $c > 0$. Using the union bound and the fact that Y_i are sub-exponential we obtain:

$$\mathbb{P}(\max|Y_i| \geq t) \leq n\exp(1 - t/(c'K)), \tag{D.6}$$

for some absolute constant. Setting $t = 2c'K\log(n)$ brings the above probability converging to zero at a rate n^{-1}. Furthermore by Chebyshev's inequality we obtain:

$$\mathbb{P}\Big(\Big|n^{-1}\sum_{i=1}^n Y_i^2 - \mathbb{E}Y^2\Big| \geq t\Big) \leq \mathrm{Var}(Y_i^2)n^{-1}t^{-2} \leq 2^9K^4n^{-1}t^{-2}, \tag{D.7}$$

and thus we can set $t = \mathbb{E}Y^2/2$ to bring the above probability to zero at a rate n^{-1}. In addition we have $\mathbb{E}Y^2/2 \leq n^{-1}\sum_{i=1}^n Y_i^2 \leq 2\mathbb{E}Y^2$ with probability at least $2^{11}K^4n^{-1}/(\mathbb{E}Y^2)^2$. Selecting

$t = \sqrt{\frac{96\mathbb{E}Y^2 \log d}{cn}}$ in (D.4) gives us that:

$$t \le \frac{\mathbb{E}Y^2}{c' K \log n} \le \frac{16 n^{-1} \sum_{i=1}^n Y_i^2}{4 \max_{i\in[n]} |Y_i|},$$

where the first inequality in the preceding display holds for large enough values of n so long as $\log d = o(n/\log^2 n)$. Hence we conclude:

$$\Big\| \frac{1}{n} \sum_{i=1}^n Y_i \mathbf{P}_{\boldsymbol{\beta}^{*\perp}} (\boldsymbol{X}_i^{\otimes 2} - \mathbf{I}_d) \mathbf{P}_{\boldsymbol{\beta}^{*\perp}} \Big\|_{\max} \le \sqrt{\frac{96\mathbb{E}Y^2 \log d}{cn}},$$

with probability at least $1 - 2d^{-1} - (2^{11}K^4/(\mathbb{E}Y^2)^2 + e)n^{-1}$. Taking into account that $\mathbb{E}Y^2 \le 4K^2$ we obtain that $C_2 = 384K^2/c$. □

E Proofs for Second Step

Remark E.1. For simplicity of presentation we will subtract n from the indexes of the set S_2 in the proofs, i.e., instead of having observations indexed in the range $S_2 = \{n+1, \dots, 2n\}$ we will pretend that our observations are in the range $\{1, \dots, n\}$.

Proof of Theorem 3.3. Take the fixed estimate $\widehat{\mathbf{v}}$ from the first step (recall that $\|\widehat{\mathbf{v}}\|_2 = 1$), and decompose it to:

$$\widehat{\mathbf{v}} = (\widehat{\mathbf{v}}^\top \boldsymbol{\beta}^*)\boldsymbol{\beta}^* + \widehat{\boldsymbol{\beta}}^\perp.$$

By the Pythagorean theorem we have $1 = \|\widehat{\mathbf{v}}\|_2^2 = (\widehat{\mathbf{v}}^\top \boldsymbol{\beta}^*)^2 \|\boldsymbol{\beta}^*\|_2^2 + \|\widehat{\boldsymbol{\beta}}^\perp\|_2^2$, which implies that

$$\|\widehat{\boldsymbol{\beta}}^\perp\|_2 = \sqrt{1 - (\widehat{\mathbf{v}}^\top \boldsymbol{\beta}^*)^2} \le 1. \tag{E.1}$$

Put $\alpha := c_0 \widehat{\mathbf{v}}^\top \boldsymbol{\beta}^*$ so by Lemma D.2 we have $|\alpha| > \kappa c_0$ with high probability. By formulation (3.6) we have:

$$\frac{1}{2n}\|\mathbf{X}(\widehat{\mathbf{b}} - \alpha\boldsymbol{\beta}^*)\|_2^2 + \lambda_n \|\widehat{\mathbf{b}}\|_1 \le \frac{1}{n}\Big\langle (\boldsymbol{Y} - \overline{\boldsymbol{Y}}) \odot \mathbf{X}\widehat{\mathbf{v}} - \alpha \mathbf{X}\boldsymbol{\beta}^*, \mathbf{X}(\widehat{\mathbf{b}} - \alpha\boldsymbol{\beta}^*)\Big\rangle + \nu_n \|\alpha\boldsymbol{\beta}^*\|_1.$$

We handle the empirical process term in Lemma E.3, which also presents the main difficulty in the analysis of the ℓ_1 regularized least squares procedure. Using this result we conclude that:

$$\frac{1}{2n}\|\mathbf{X}(\widehat{\mathbf{b}} - \alpha\boldsymbol{\beta}^*)\|_2^2 + \nu_n \|\widehat{\mathbf{b}}\|_1 \le \widetilde{C}\sqrt{\frac{\log d}{n}}\Big[\|\widehat{\mathbf{b}} - \alpha\boldsymbol{\beta}^*\|_1 + \frac{1}{\sqrt{n}}\|\mathbf{X}(\widehat{\mathbf{b}} - \alpha\boldsymbol{\beta}^*)\|_2\Big] + \nu_n\|\alpha\boldsymbol{\beta}^*\|_1, \tag{E.2}$$

with probability at least $1 - O(n^{-1} + d^{-1})$. We now distinguish two cases. First assume that $\|\mathbf{X}(\widehat{\mathbf{b}} - \alpha\boldsymbol{\beta}^*)\|_2 > 2\widetilde{C}\sqrt{\log d}$. Then (E.2) implies that:

$$\frac{1}{4n}\|\mathbf{X}(\widehat{\mathbf{b}} - \alpha\boldsymbol{\beta}^*)\|_2^2 + \nu_n \|\widehat{\mathbf{b}}\|_1 \le \widetilde{C}\sqrt{\frac{\log d}{n}}\|\widehat{\mathbf{b}} - \alpha\boldsymbol{\beta}^*\|_1 + \nu_n\|\alpha\boldsymbol{\beta}^*\|_1, \tag{E.3}$$

Next using a standard trick [see, e.g., 4, 6] we have:

$$\|\widehat{\mathbf{b}}\|_1 = \|\widehat{\mathbf{b}}_{S_{\boldsymbol{\beta}^*}}\|_1 + \|\widehat{\mathbf{b}}_{S^c_{\boldsymbol{\beta}^*}}\|_1 \ge \|\alpha\boldsymbol{\beta}^*_{S_{\boldsymbol{\beta}^*}}\|_1 - \|\widehat{\mathbf{b}}_{S_{\boldsymbol{\beta}^*}} - \alpha\boldsymbol{\beta}^*_{S_{\boldsymbol{\beta}^*}}\|_1 + \|\widehat{\mathbf{b}}_{S^c_{\boldsymbol{\beta}^*}}\|_1,$$
$$\|\widehat{\mathbf{b}} - \alpha\boldsymbol{\beta}^*\|_1 = \|\widehat{\mathbf{b}}_{S_{\boldsymbol{\beta}^*}} - \alpha\boldsymbol{\beta}^*_{S_{\boldsymbol{\beta}^*}}\|_1 + \|\widehat{\mathbf{b}}_{S^c_{\boldsymbol{\beta}^*}}\|_1.$$

Selecting $\nu_n \ge 2\widetilde{C}\sqrt{\frac{\log d}{n}}$, the above equalities in combination with (E.3) guarantee that:

$$\frac{1}{4n}\|\mathbf{X}(\widehat{\mathbf{b}} - \alpha\boldsymbol{\beta}^*)\|_2^2 + \nu_n\|\widehat{\mathbf{b}}_{S^c_{\boldsymbol{\beta}^*}} - \alpha\boldsymbol{\beta}^*_{S^c_{\boldsymbol{\beta}^*}}\|_1 \le 3\nu_n\|\widehat{\mathbf{b}}_{S_{\boldsymbol{\beta}^*}} - \alpha\boldsymbol{\beta}^*_{S_{\boldsymbol{\beta}^*}}\|_1. \tag{E.4}$$

Using Corollary 1 from [29], since clearly $\mathbf{I}_d$ satisfies the RE condition of order $2s$ with constants $(3,1)$ (i.e., $\forall S \in \binom{[d]}{2s}$ $\forall \boldsymbol{\theta} \in \{\|\boldsymbol{\theta}_{S^c}\|_1 \le 3\|\boldsymbol{\theta}_S\|_1\}$ we have $\|\boldsymbol{\theta}\|_2 \le \|\mathbf{I}\boldsymbol{\theta}\|_2$) we can further bound:

$$\frac{1}{4n}\|\mathbf{X}(\widehat{\mathbf{b}} - \alpha\boldsymbol{\beta}^*)\|_2^2 \ge \frac{1}{4\cdot 8^2}\|\widehat{\mathbf{b}} - \alpha\boldsymbol{\beta}^*\|_2^2, \tag{E.5}$$

with probability at least $1 - c'\exp(-c''n)$ if $n > c'''4^2 s\log d$ where $c', c'', c''' > 0$ are absolute constants. On the above event, (E.4) implies:

$$\frac{1}{4\cdot 8^2}\|\widehat{\mathbf{b}} - \alpha\boldsymbol{\beta}^*\|_2^2 \le 3\nu_n\|\widehat{\mathbf{b}}_{S_{\boldsymbol{\beta}^*}} - \alpha\boldsymbol{\beta}^*_{S_{\boldsymbol{\beta}^*}}\|_1 \le 3\nu_n\sqrt{2s}\|\widehat{\mathbf{b}} - \alpha\boldsymbol{\beta}^*\|_2,$$

where we used Cauchy-Schwartz and the fact that the vector $\widehat{\mathbf{b}}_{S_{\boldsymbol{\beta}^*}} - \alpha\boldsymbol{\beta}^*_{S_{\boldsymbol{\beta}^*}}$ is at most $2s$ sparse. The above inequality gives us that:

$$\|\widehat{\mathbf{b}} - \alpha\boldsymbol{\beta}^*\|_2 \leq 12 \cdot 8^2\sqrt{2s}\nu_n. \tag{E.6}$$

In the second case when $\|\mathbf{X}(\widehat{\mathbf{b}} - \alpha\boldsymbol{\beta}^*)\|_2 \leq 2\widetilde{C}\sqrt{\log d}$, on the event (E.5) we have:

$$\|\widehat{\mathbf{b}} - \alpha\boldsymbol{\beta}^*\|_2 \leq 32\widetilde{C}\sqrt{\frac{\log d}{n}},$$

and we see that in either case bound (E.6) holds. Before we complete the proof we need the following straightforward result:

Lemma E.2. Assume that n is large enough so that $12 \cdot 8^2\sqrt{2s}\nu_n \leq \kappa c_0/2$. Then with probability at least $1 - O(n^{-1} + d^{-1})$ we have:

$$\min_{\eta\in\{1,-1\}} \left\| \frac{\widehat{\mathbf{b}}}{\|\widehat{\mathbf{b}}\|_2} - \eta\boldsymbol{\beta}^* \right\|_2 \leq \frac{38 \cdot 8^2\sqrt{2s}\nu_n}{\kappa c_0}$$

Finally notice that $s\sqrt{\frac{\log d}{n}} < R$ implies that $12 \cdot 8^2\sqrt{2s}\nu_n \leq \kappa c_0/2$ when R is small enough. □

Proof of Lemma E.2. Put $r = 12 \cdot 8^2\sqrt{2s}\nu_n$ for brevity. By (E.6) and the triangle inequality we can conclude that:

$$|\alpha| - r \leq \|\widehat{\mathbf{b}}\|_2 \leq r + |\alpha|.$$

Additionally:

$$\left\| \frac{\widehat{\mathbf{b}}}{\|\widehat{\mathbf{b}}\|_2} - \operatorname{sign}(\alpha)\boldsymbol{\beta}^* \right\|_2 \leq \frac{\|\widehat{\mathbf{b}} - \alpha\boldsymbol{\beta}^*\|_2 + |\|\widehat{\mathbf{b}}\|_2 - |\alpha||}{\|\widehat{\mathbf{b}}\|_2} \leq \frac{2r}{|\alpha| - r} \leq \frac{4r}{|\alpha|} \leq \frac{4r}{\kappa c_0},$$

with the last two inequalities holding with high probability when $r < \kappa c_0/2$ ($\leq |\alpha|/2$ with high probability by Lemma D.2). This completes the proof. □

Lemma E.3. There exists a constant $\widetilde{C}$ depending on f, ε such that:

$$\frac{1}{n}\Big\langle (\boldsymbol{Y} - \overline{\boldsymbol{Y}}) \odot \mathbf{X}\widehat{\mathbf{v}} - \alpha\mathbf{X}\boldsymbol{\beta}^*, \mathbf{X}(\widehat{\mathbf{b}} - \alpha\boldsymbol{\beta}^*)\Big\rangle \leq \widetilde{C}\sqrt{\frac{\log d}{n}}\Big[\frac{1}{\sqrt{n}}\|\mathbf{X}(\widehat{\mathbf{b}} - \alpha\boldsymbol{\beta}^*)\|_2 + \|\widehat{\mathbf{b}} - \alpha\boldsymbol{\beta}^*\|_1\Big],$$

with probability at least $1 - O(n^{-1} + d^{-1})$.

Proof of Lemma E.3. Using Hölder's inequality we obtain:

$$\frac{1}{n}\Big\langle (\boldsymbol{Y} - \overline{\boldsymbol{Y}}) \odot \mathbf{X}\widehat{\mathbf{v}} - \alpha\mathbf{X}\boldsymbol{\beta}^*, \mathbf{X}(\widehat{\mathbf{b}} - \alpha\boldsymbol{\beta}^*)\Big\rangle \leq \frac{1}{n}\|\mathbf{X}^\top[(\boldsymbol{Y} - \boldsymbol{\mu}) \odot \mathbf{X}\widehat{\mathbf{v}} - \alpha\mathbf{X}\boldsymbol{\beta}^*]\|_\infty\|\widehat{\mathbf{b}} - \alpha\boldsymbol{\beta}^*\|_1 \tag{E.7}$$

$$+ \frac{1}{n}\|(\overline{\boldsymbol{Y}} - \boldsymbol{\mu}) \odot \mathbf{X}\widehat{\mathbf{v}}\|_2\|\mathbf{X}(\widehat{\mathbf{b}} - \alpha\boldsymbol{\beta}^*)\|_2$$

where we have set $\boldsymbol{\mu} := \mathbb{E}\boldsymbol{Y}$ for brevity. We first handle the second term. We have $\frac{1}{\sqrt{n}}\|(\overline{\boldsymbol{Y}} - \boldsymbol{\mu}) \odot \mathbf{X}\widehat{\mathbf{v}}\|_2 = |\overline{Y} - \mu|\frac{1}{\sqrt{n}}\|\mathbf{X}\widehat{\mathbf{v}}\|_2$. Since Y_i is assumed to be sub-exponential by a Bernstein type of inequality we have:

$$\mathbb{P}(|\overline{Y} - \mu| \geq t) \leq 2\exp(-c\min(nt^2/4K^2, nt/2K))$$

where c is an absolute constant. Thus we conclude that $|\overline{Y} - \mu| \leq \frac{2K}{\sqrt{c}}\sqrt{\frac{\log d}{n}}$ with probability at least $2d^{-1}$, for values of n such that $\sqrt{\frac{\log d}{n}} < c$. Also since we have $\mathbf{X}\widehat{\mathbf{v}} \sim \mathcal{N}(0, \mathbf{I}_n)$ we obtain that $\|\mathbf{X}\widehat{\mathbf{v}}\|_2^2 \sim \chi^2_n$. Hence by Chebyshev's inequality we obtain:

$$\mathbb{P}(|\|\mathbf{X}\widehat{\mathbf{v}}\|_2^2/n - 1| \geq t) \leq 2/(nt),$$

and thus by plugging in $t = 1$, we conclude that $\frac{1}{\sqrt{n}}\|\mathbf{X}\widehat{\mathbf{v}}\|_2 \leq \sqrt{2}$ with probability at least $1 - 2n^{-1}$. Hence $\frac{1}{\sqrt{n}}\|(\overline{\boldsymbol{Y}} - \boldsymbol{\mu}) \odot \mathbf{X}\widehat{\mathbf{v}}\|_2 \leq \widetilde{C}_1\sqrt{\frac{\log d}{n}}$ with probability at least $1 - O(n^{-1}) - 2d^{-1}$.

Next we analyze the sup norm term appearing in inequality (E.7). The first fact we observe is that by construction this term is unbiased since:

$$\begin{aligned}
&\mathbb{E}[(Y-\mu)\boldsymbol{X}^{\otimes 2}\boldsymbol{\beta}^*(\widehat{\mathbf{v}}^\top\boldsymbol{\beta}^*)-\alpha\boldsymbol{X}^{\otimes 2}\boldsymbol{\beta}^*]+\mathbb{E}[(Y-\mu)\boldsymbol{X}^{\otimes 2}\widehat{\boldsymbol{\beta}}^\perp]\\
&=\underbrace{\boldsymbol{\beta}^*\mathbb{E}[(Y-\mu)(\boldsymbol{X}^\top\boldsymbol{\beta}^*)^2(\widehat{\mathbf{v}}^\top\boldsymbol{\beta}^*)-\alpha{\boldsymbol{\beta}^*}^\top\boldsymbol{X}^{\otimes 2}\boldsymbol{\beta}^*]}_{0}\\
&+\underbrace{\mathbb{E}[(Y-\mu)\mathbf{P}_{\boldsymbol{\beta}^{*\perp}}\boldsymbol{X}^{\otimes 2}\boldsymbol{\beta}^*(\widehat{\mathbf{v}}^\top\boldsymbol{\beta}^*)]}_{0}-\underbrace{\mathbb{E}[\alpha\mathbf{P}_{\boldsymbol{\beta}^{*\perp}}\boldsymbol{X}^{\otimes 2}\boldsymbol{\beta}^*]}_{0}\\
&+\underbrace{\mathbb{E}[(Y-\mu)\mathbf{P}_{\boldsymbol{\beta}^{*\perp}}\boldsymbol{X}^{\otimes 2}\widehat{\boldsymbol{\beta}}^\perp]}_{0}+\boldsymbol{\beta}\underbrace{\mathbb{E}[(Y-\mu){\boldsymbol{\beta}^*}^\top\boldsymbol{X}^{\otimes 2}\widehat{\boldsymbol{\beta}}^\perp]}_{0}.
\end{aligned}$$

Now according to the decomposition in the preceding display, we break down the sup norm term in (E.7) into mean zero terms using the triangle inequality:

$$\begin{aligned}
n^{-1}\|\mathbf{X}^\top[(\boldsymbol{Y}-\boldsymbol{\mu})\odot\mathbf{X}\widehat{\mathbf{v}}-\alpha\mathbf{X}\boldsymbol{\beta}^*]\|_\infty &\le n^{-1}\|\mathbf{P}_{\{\boldsymbol{\beta}^*,\widehat{\boldsymbol{\beta}}^\perp\}^\perp}\mathbf{X}^\top[(\boldsymbol{Y}-\boldsymbol{\mu})\odot\mathbf{X}\widehat{\boldsymbol{\beta}}^\perp]\|_\infty\\
&+n^{-1}\|\boldsymbol{\beta}^*{\boldsymbol{\beta}^*}^\top\mathbf{X}^\top[(\boldsymbol{Y}-\boldsymbol{\mu})\odot\mathbf{X}\widehat{\boldsymbol{\beta}}^\perp]\|_\infty\\
&+\frac{n^{-1}}{\|\widehat{\boldsymbol{\beta}}^\perp\|_2^2}\|\widehat{\boldsymbol{\beta}}^\perp(\widehat{\boldsymbol{\beta}}^\perp)^\top\mathbf{X}^\top[(\boldsymbol{Y}-\boldsymbol{\mu})\odot\mathbf{X}\widehat{\boldsymbol{\beta}}^\perp]\|_\infty\\
&+n^{-1}\|\boldsymbol{\beta}^*{\boldsymbol{\beta}^*}^\top\mathbf{X}^\top[(\boldsymbol{Y}-\boldsymbol{\mu})\odot\mathbf{X}\boldsymbol{\beta}^*-c_0\mathbf{X}\boldsymbol{\beta}^*]\|_\infty\\
&+n^{-1}\|\mathbf{P}_{\boldsymbol{\beta}^{*\perp}}\mathbf{X}^\top[(\boldsymbol{Y}-\boldsymbol{\mu})\odot\mathbf{X}\boldsymbol{\beta}^*-c_0\mathbf{X}\boldsymbol{\beta}^*]\|_\infty,
\end{aligned}\tag{E.8}$$

where in the last two terms we used the fact that $|\widehat{\mathbf{v}}^\top\boldsymbol{\beta}^*|\le 1$ and $\mathbf{P}_{\{\boldsymbol{\beta}^*,\widehat{\boldsymbol{\beta}}^\perp\}}$ is the projection on the space $\mathrm{span}\{\boldsymbol{\beta}^*,\widehat{\boldsymbol{\beta}}^\perp\}^\perp$. We use Lemma E.4 to control the first term of the decomposition. Lemma E.5 handles the second term, and Lemmas E.7 and E.9 show concentration for the remaining terms. We conclude that there exists a constant $\widetilde{C}_2$ such that:

$$n^{-1}\|\mathbf{X}^\top[(\boldsymbol{Y}-\boldsymbol{\mu})\odot\mathbf{X}\widehat{\mathbf{v}}-\alpha\mathbf{X}\boldsymbol{\beta}^*]\|_\infty\le\widetilde{C}_2\sqrt{\frac{\log d}{n}},$$

with probability at least $1-O(n^{-1}+d^{-1})$, which is what we aimed to show with $\widetilde{C}=\max(\widetilde{C}_1,\widetilde{C}_2)$. □

Lemma E.4. We have that:

$$\Big\|\frac{1}{n}\sum_{i=1}^n(Y_i-\mu)\boldsymbol{X}_i^\top\widehat{\boldsymbol{\beta}}^\perp\boldsymbol{X}_i^\top\mathbf{P}_{\{\boldsymbol{\beta}^*,\widehat{\boldsymbol{\beta}}^\perp\}^\perp}\Big\|_\infty\le\|\widehat{\boldsymbol{\beta}}^\perp\|_2C_3\sqrt{\frac{\log d}{n}},$$

for an absolute constant C_3 depending on f and ε with probability at least $1-2d^{-1}-\frac{\mathrm{Var}(f(Z,\varepsilon)-\mathbb{E}f(Z,\varepsilon))^2}{[\mathbb{E}(f(Z,\varepsilon)-\mathbb{E}f(Z,\varepsilon))^2]^2}n^{-1}$.

Proof of Lemma E.4. Notice that $\boldsymbol{X}_i^\top\mathbf{P}_{\{\boldsymbol{\beta}^*,\widehat{\boldsymbol{\beta}}^\perp\}^\perp}$ is independent of $(Y_i-\mu)\boldsymbol{X}_i^\top\widehat{\boldsymbol{\beta}}^\perp$. Hence conditionally on $\boldsymbol{Y}$ and $\mathbf{X}\widehat{\boldsymbol{\beta}}^\perp$ we have

$$\frac{1}{n}\sum_{i=1}^n(Y_i-\mu)\boldsymbol{X}_i^\top\widehat{\boldsymbol{\beta}}^\perp\boldsymbol{X}_i^\top\mathbf{P}_{\{\boldsymbol{\beta}^*,\widehat{\boldsymbol{\beta}}^\perp\}^\perp}\sim\mathcal{N}\Big(0,\frac{1}{n^2}\sum_{i=1}^n(Y_i-\mu)^2(\boldsymbol{X}_i^\top\widehat{\boldsymbol{\beta}}^\perp)^2\mathbf{P}_{\{\boldsymbol{\beta}^*,\widehat{\boldsymbol{\beta}}^\perp\}^\perp}\Big).$$

Next using Chebyshev's inequality we can control the probability of spread about the mean:

$$\mathbb{P}\Big(\Big|\frac{1}{n}\sum_{i=1}^n\frac{(\boldsymbol{X}_i^\top\widehat{\boldsymbol{\beta}}^\perp)^2}{\|\widehat{\boldsymbol{\beta}}^\perp\|_2^2}(Y_i-\mu)^2-\mathbb{E}(f(Z,\varepsilon)-\mathbb{E}f(Z,\varepsilon))^2\Big|\ge t\Big)\le\frac{\mathrm{Var}[(f(Z,\varepsilon)-\mathbb{E}f(Z,\varepsilon))^2]}{nt^2},\tag{E.9}$$

by setting $t=\mathbb{E}(f(Z,\varepsilon)-\mathbb{E}f(Z,\varepsilon))^2$. Using the fact that $\|\mathbf{P}_{\{\boldsymbol{\beta}^*,\widehat{\boldsymbol{\beta}}^\perp\}^\perp}\|_2\le 1$, by a standard normal tail bound and union bound on the event $\frac{1}{n}\sum_{i=1}^n(\boldsymbol{X}_i^\top\widehat{\boldsymbol{\beta}}^\perp)^2(Y_i-\mu)^2\le 2\|\widehat{\boldsymbol{\beta}}^\perp\|_2^2\mathbb{E}(f(Z,\varepsilon)-\mathbb{E}f(Z,\varepsilon))^2$ we obtain:

$$\mathbb{P}\Big(\Big\|\frac{1}{n}\sum_{i=1}^n(Y_i-\mu)\boldsymbol{X}_i^\top\widehat{\boldsymbol{\beta}}^\perp\boldsymbol{X}_i^\top\mathbf{P}_{\{\boldsymbol{\beta}^*,\widehat{\boldsymbol{\beta}}^\perp\}^\perp}\Big\|_\infty\ge t\Big)\le 2d\exp(-nt^2/[4\|\widehat{\boldsymbol{\beta}}^\perp\|_2^2\mathbb{E}(f(Z,\varepsilon)-\mathbb{E}f(Z,\varepsilon))^2]).$$

Select $t = 2\sqrt{2\mathbb{E}(f(Z,\varepsilon) - \mathbb{E}f(Z,\varepsilon))^2}\|\widehat{\boldsymbol{\beta}}^\perp\|_2\sqrt{\frac{\log d}{n}}$ yields the desired bound with

$$C_3 = 2\sqrt{2\mathbb{E}(f(Z,\varepsilon) - \mathbb{E}f(Z,\varepsilon))^2}\|\widehat{\boldsymbol{\beta}}^\perp\|_2.$$

□

Lemma E.5. We have that:

$$n^{-1}\|\boldsymbol{\beta}^*\boldsymbol{\beta}^{*\top}\mathbf{X}^\top[(\boldsymbol{Y} - \boldsymbol{\mu}) \odot \mathbf{X}\widehat{\boldsymbol{\beta}}^\perp]\|_\infty \leq C_4\sqrt{\frac{\log d}{n}},$$

for an absolute constant C_4 depending on f and ε with probability at least $1 - 2d^{-1} - \frac{\operatorname{Var} Z^2(f(Z,\varepsilon) - \mathbb{E}f(Z,\varepsilon))^2}{[\mathbb{E}Z^2(f(Z,\varepsilon) - \mathbb{E}f(Z,\varepsilon))^2]^2}n^{-1}$.

Proof of Lemma E.5. Notice that $\|\boldsymbol{\beta}^*\|_\infty \leq \|\boldsymbol{\beta}^*\|_2 = 1$ and thus:

$$n^{-1}\|\boldsymbol{\beta}^*\boldsymbol{\beta}^{*\top}\mathbf{X}^\top[(\boldsymbol{Y} - \boldsymbol{\mu}) \odot \mathbf{X}\widehat{\boldsymbol{\beta}}^\perp]\|_\infty \leq n^{-1}|\boldsymbol{\beta}^{*\top}\mathbf{X}^\top[(\boldsymbol{Y} - \boldsymbol{\mu}) \odot \mathbf{X}\widehat{\boldsymbol{\beta}}^\perp]|.$$

Next since $((\widehat{\boldsymbol{\beta}}^\perp)^\top \boldsymbol{X}_i) \perp\!\!\!\perp (\boldsymbol{\beta}^{*\top}\boldsymbol{X}_i)(Y_i - \mu)$, conditioning on $\{(\boldsymbol{\beta}^{*\top}\boldsymbol{X}_i)(Y_i - \mu)\}_{i\in[n]}$ we obtain:

$$\frac{1}{n}\sum_{i=1}^n(\boldsymbol{\beta}^{*\top}\boldsymbol{X}_i)(Y_i - \mu)((\widehat{\boldsymbol{\beta}}^\perp)^\top\boldsymbol{X}_i) \sim \mathcal{N}\Big(0, \frac{\|\widehat{\boldsymbol{\beta}}\|_2^2}{n^2}\sum_{i=1}^n(\boldsymbol{\beta}^{*\top}\boldsymbol{X}_i)^2(Y_i - \mu)^2\Big),$$

Next,

$$\mathbb{P}\Big(\Big|\frac{1}{n}\sum_{i=1}^n(\boldsymbol{\beta}^{*\top}\boldsymbol{X}_i)^2(Y_i - \mu)^2 - \mathbb{E}Z^2(f(Z,\varepsilon) - \mathbb{E}f(Z,\varepsilon))^2\Big| \geq t\Big) \leq \frac{\operatorname{Var}[Z^2(f(Z,\varepsilon) - \mathbb{E}f(Z,\varepsilon))^2]}{nt^2}, \tag{E.10}$$

by setting $t = \mathbb{E}Z^2(f(Z,\varepsilon) - \mathbb{E}f(Z,\varepsilon))^2$ we can control the variance term above. The final bound follows after an application of a standard Gaussian tail bound, where C_4 turns out to be $C_4 = \|\widehat{\boldsymbol{\beta}}\|_2 2\sqrt{\mathbb{E}Z^2(f(Z,\varepsilon) - \mathbb{E}f(Z,\varepsilon))^2}$. □

Lemma E.6. For large enough values of n we have:

$$\frac{n^{-1}}{\|\widehat{\boldsymbol{\beta}}^\perp\|_2^2}\|\widehat{\boldsymbol{\beta}}^\perp(\widehat{\boldsymbol{\beta}}^\perp)^\top\mathbf{X}^\top[(\boldsymbol{Y} - \boldsymbol{\mu}) \odot \mathbf{X}\widehat{\boldsymbol{\beta}}^\perp]\|_\infty \leq C_5\sqrt{\frac{\log d}{n}},$$

with prob at least $1 - 2d^{-1} - O(n^{-1})$.

Proof of Lemma E.6. We have that $\|\widehat{\boldsymbol{\beta}}^\perp\|_\infty \leq \|\widehat{\boldsymbol{\beta}}^\perp\|_2$, and hence:

$$\frac{n^{-1}}{\|\widehat{\boldsymbol{\beta}}^\perp\|_2^2}\|\widehat{\boldsymbol{\beta}}^\perp(\widehat{\boldsymbol{\beta}}^\perp)^\top\mathbf{X}^\top[(\boldsymbol{Y} - \boldsymbol{\mu}) \odot \mathbf{X}\widehat{\boldsymbol{\beta}}^\perp]\|_\infty \leq \frac{n^{-1}}{\|\widehat{\boldsymbol{\beta}}^\perp\|_2}|(\widehat{\boldsymbol{\beta}}^\perp)^\top\mathbf{X}^\top[(\boldsymbol{Y} - \boldsymbol{\mu}) \odot \mathbf{X}\widehat{\boldsymbol{\beta}}^\perp]|.$$

Observe that $\boldsymbol{X}_i^\top\widehat{\boldsymbol{\beta}}^\perp$ is independent from $Y_i - \mu$, and in addition $\boldsymbol{X}_i^\top\widehat{\boldsymbol{\beta}}^\perp \sim \mathcal{N}(0, \|\widehat{\boldsymbol{\beta}}^\perp\|_2^2)$. Hence $(\boldsymbol{X}_i^\top\widehat{\boldsymbol{\beta}}^\perp)^2/\|\boldsymbol{\beta}^\perp\|_2^2 \sim \chi_1^2$. Next we make usage of the decomposition:

$$\frac{1}{n}\sum_i(Y_i - \mu)(\boldsymbol{X}_i^\top\widehat{\boldsymbol{\beta}}^\perp)^2 = \frac{\|\widehat{\boldsymbol{\beta}}^\perp\|_2^2}{n}\sum_i(Y_i - \mu)((\boldsymbol{X}_i^\top\widehat{\boldsymbol{\beta}}^\perp)^2/\|\widehat{\boldsymbol{\beta}}^\perp\|_2^2 - 1) + \frac{\|\widehat{\boldsymbol{\beta}}^\perp\|_2^2}{n}\sum_i(Y_i - \mu).$$

Since Y_i is assumed to be sub-exponential, the second concentrates about zero by Proposition 5.16 in [32]:

$$\mathbb{P}\Big(\Big|n^{-1}\sum_{i=1}^n Y_i - \mu\Big| \geq t\Big) \leq 2\exp(-c\min(nt^2/K^2, nt/K)).$$

Selecting $t = \frac{K}{\sqrt{c}}\sqrt{\frac{\log d}{n}}$ gives a bound on the probability equal to $2d^{-1}$, for values of n large enough so that $\sqrt{\frac{\log d}{n}} \leq \sqrt{c}$. For the remaining term, conditionally on $\{Y_i\}_{i\in[n]}$, by Lemma 1 of [19] we obtain:

$$\begin{aligned}\mathbb{P}\Big(\Big|n^{-1}\sum_i(Y_i - \mu)((\boldsymbol{X}_i^\top\widehat{\boldsymbol{\beta}}^\perp)^2/\|\widehat{\boldsymbol{\beta}}^\perp\|_2^2 - 1)\Big| \geq 2\sqrt{n^{-1}\sum_{i=1}^n(Y_i - \mu)^2\sqrt{t} + 2\max_{i\in[n]}|Y_i - \mu|t}\Big) \\ \leq 2\exp(-nt).\end{aligned} \tag{E.11}$$

Next, by the triangle inequality:

$$\sqrt{n^{-1}\sum_{i=1}^{n}(Y_i-\mu)^2} \leq \sqrt{n^{-1}\sum_{i=1}^{n}Y_i^2}+|\mu|, \quad \max_{i\in[n]}|Y_i-\mu| \leq \max_{i\in[n]}|Y_i|+|\mu|.$$

The inequalities in the preceding display allow us to reuse the results (D.6) and (D.7) of Lemma D.7. Thus conditioning on these events (E.11) implies:

$$\mathbb{P}\Big(\Big|n^{-1}\sum_i (Y_i-\mu)((\boldsymbol{X}_i^\top\widehat{\boldsymbol{\beta}}^\perp)^2/\|\widehat{\boldsymbol{\beta}}^\perp\|_2^2-1)\Big| \geq (2\sqrt{2\mathbb{E}Y^2}+\mu)\sqrt{t}+(4c'K\log n)t\Big) \leq 2\exp(-nt),$$

on an event failing with probability at most $\Big(\frac{\operatorname{Var}Y^2}{[\mathbb{E}Y^2]^2}+e\Big)n^{-1}$. Selecting $t=\frac{\log d}{n}$ implies that with probability at least $1-2d^{-1}-\Big(\frac{\operatorname{Var}Y^2}{[\mathbb{E}Y^2]^2}+e\Big)n^{-1}$ we have:

$$\Big|n^{-1}\sum_i (Y_i-\mu)((\boldsymbol{X}_i^\top\widehat{\boldsymbol{\beta}}^\perp)^2/\|\widehat{\boldsymbol{\beta}}^\perp\|_2^2-1)\Big| \leq [(2\sqrt{2\mathbb{E}Y^2}+\mu)+4c'K]\sqrt{\frac{\log d}{n}},$$

with the probability of failing being at most $\exp(-nt)\leq\max(2d^{-1},O(n^{-1}))$. We remind the reader that we are assuming $\log(d)=o(n/\log^2(n))$. This completes the proof with $C_5=\Big((2\sqrt{2\mathbb{E}Y^2}+\mu)+4c'K+\frac{K}{\sqrt{c}}\Big)\|\widehat{\boldsymbol{\beta}}^\perp\|_2$. □

Lemma E.7. We have that:

$$n^{-1}\|\boldsymbol{\beta}^*\boldsymbol{\beta}^{*\top}\mathbf{X}^\top[(\boldsymbol{Y}-\boldsymbol{\mu})\odot\mathbf{X}\boldsymbol{\beta}^*-c_0\mathbf{X}\boldsymbol{\beta}^*]\|_\infty \leq C_6\sqrt{\frac{\log d}{n}},$$

with probability at least $1-O(n^{-1})-3d^{-1}$.

Proof of Lemma E.7. As in Lemma E.5 we have:

$$n^{-1}\|\boldsymbol{\beta}^*\boldsymbol{\beta}^{*\top}\mathbf{X}^\top[(\boldsymbol{Y}-\boldsymbol{\mu})\odot\mathbf{X}\boldsymbol{\beta}^*-c_0\mathbf{X}\boldsymbol{\beta}^*]\|_\infty \leq n^{-1}|\boldsymbol{\beta}^{*\top}\mathbf{X}^\top[(\boldsymbol{Y}-\boldsymbol{\mu})\odot\mathbf{X}\boldsymbol{\beta}^*-c_0\mathbf{X}\boldsymbol{\beta}^*]|,$$

since $\|\boldsymbol{\beta}^*\|_\infty\leq 1$. We decompose the right hand side of the preceding display to:

$$\frac{1}{n}\sum_{i=1}^{n}[(\boldsymbol{\beta}^{*\top}\boldsymbol{X}_i)^2Y_i-(c_0+\mu)]+\frac{(c_0+\mu)}{n}\sum_{i=1}^{n}[1-(\boldsymbol{\beta}^{*\top}\boldsymbol{X}_i)^2].$$

To handle the first term one can easily use Chebyshev's inequality to obtain convergence with probability at least $(\log d)^{-1}$. However, to sharpen this rate, we work around the classic Chebyshev's inequality, by making usage of recent concentration results on polynomials of sub-Gaussian random variables proved in [1]. We have the following:

Lemma E.8. We have that:

$$\frac{1}{n}\sum_{i=1}^{n}[(\boldsymbol{\beta}^{*\top}\boldsymbol{X}_i)^2Y_i-(c_0+\mu)] \leq \widetilde{C}_6\sqrt{\frac{\log d}{n}},$$

with probability at least $1-\max(O(n^{-1}),d^{-1})$.

Usual concentration bounds on the χ^2 distribution can be used to control the second term. Using Lemma 1 of [19] we obtain:

$$\mathbb{P}\Big(\Big|\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{\beta}^{*\top}\boldsymbol{X}_i)^2-1\Big|\geq 2\sqrt{t}+2t\Big)\leq 2\exp(-nt).$$

Select $t=\sqrt{\frac{\log d}{n}}$ to complete the proof assuming that $t<1$ and setting $C_6=4(c_0+\mu)+\widetilde{C}_6$. □

Proof of Lemma E.8. First we construct the random variable $Z_i=\eta_i|\boldsymbol{\beta}^{*\top}\boldsymbol{X}_i|^{1/2}|Y_i|^{1/4}$, where η_i is a Rademacher random variable. Notice that $Z_i^4=(\boldsymbol{\beta}^{*\top}\boldsymbol{X}_i)^2Y_i$, and hence $\mathbb{E}Z_i^4=(c_0+\mu)$. We now argue that Z is a sub-Gaussian random variable. By Hölder's inequality, and the definition of ψ_2 norm we have:

$$\mathbb{E}|Z|^p\leq\sqrt{\mathbb{E}|\boldsymbol{\beta}^{*\top}\boldsymbol{X}|^p\mathbb{E}|Y|^{p/2}}\leq(p\|\boldsymbol{\beta}^{*\top}\boldsymbol{X}\|_{\psi_2}(\|Y\|_{\psi_1}/2)^{1/2})^{p/2}\leq(p(\|Y\|_{\psi_1}/2)^{1/2})^{p/2},$$

where we used that since ${\boldsymbol{\beta}^*}^\top \boldsymbol{X} \sim \mathcal{N}(0,1)$ we have $\|{\boldsymbol{\beta}^*}^\top \boldsymbol{X}\|_{\psi_2} \le 1$. Hence $\|Z\|_{\psi_2} \le (K/2)^{1/4}$, and thus Z is sub-Gaussian as claimed.

For the remaining part recall the notation preceding Theorem B.2. For $f(x) = x^4$ and $F(\boldsymbol{x}) = \sum_{i=1}^n f(x_i)$ we have $\mathbf{D}^\ell F(\boldsymbol{x}) = \mathrm{diag}_d(f^{(\ell)}(x_1), \dots, f^{(\ell)}(x_n))$ for $\ell \in [4]$. Using the definition of ψ_2 norm we can easily estimate $\mathbb{E}[|Z|^\ell] \le (\sqrt{\ell})^\ell \|Z\|_{\psi_2}^\ell$. To this end we observe the following:

$$\|\mathrm{diag}_\ell\{x_1, \dots, x_n\}\|_{\mathcal{J}} = \mathbb{1}(\#\mathcal{J} = 1)\|\boldsymbol{x}\|_2 + \mathbb{1}(\#\mathcal{J} \ge 2)\|\boldsymbol{x}\|_{\max}.$$

Hence:

$$\|\mathbb{E}\mathbf{D}^\ell F(\boldsymbol{Z})\|_{\mathcal{J}} \le [\mathbb{1}(\#\mathcal{J} = 1)\sqrt{n} + \mathbb{1}(\#\mathcal{J} \ge 2)]4!/(4-\ell)!(\sqrt{4-\ell})^{4-\ell}\|Z\|_{\psi_2}^{4-\ell},$$

for $\ell \in [4]$, where with a slight abuse of notation we understand $(\sqrt{4-\ell})^{(4-\ell)} = 1$ when $\ell = 4$. Using the moment estimate of Theorem B.2 we obtain:

$$\begin{aligned}
&\|F(\boldsymbol{Z}) - \mathbb{E}F(\boldsymbol{Z})\|_k \\
&\le K_4 \sum_{\ell \in [4]} \|Z\|_{\psi_2}^\ell \sum_{\mathcal{J} \in \mathcal{P}_\ell} k^{\#\mathcal{J}/2}[\mathbb{1}(\#\mathcal{J} = 1)\sqrt{n} + \mathbb{1}(\#\mathcal{J} \ge 2)] \frac{4!(\sqrt{4-\ell})^{4-\ell}\|Z\|_{\psi_2}^{4-\ell}}{(4-\ell)!} \\
&\le \widetilde{K}_4[\sqrt{kn} + k^2],
\end{aligned}$$

where $\mathcal{P}_\ell$ is the set of partitions of $[\ell]$, the absolute constant K_4 depends solely on the dimension four, and $\widetilde{K}_4$ on the $\|Z\|_{\psi_2}$ norm and K_4. Next by Chebyshev's inequality:

$$\mathbb{P}(n^{-1}|F(\boldsymbol{Z}) - \mathbb{E}F(\boldsymbol{Z})| \ge t) \le \frac{\widetilde{K}_4^k[\sqrt{k/n} + k^2/n]^k}{t^k}.$$

Applying this inequality with $k = \min(\lceil \log d \rceil, \lceil (n \log d)^{1/4} \rceil)$, and $t = 2e\widetilde{K}_4\sqrt{\frac{\log d}{n}}$ gives us that:

$$\frac{1}{n}\sum_{i=1}^n [({\boldsymbol{\beta}^*}^\top \boldsymbol{X}_i)^2 Y_i - (c_0 + \mu)] \le \widetilde{C}_5\sqrt{\frac{\log d}{n}},$$

with probability at least $1 - \exp(-\min(\lceil \log d \rceil, \lceil (n \log d)^{1/4} \rceil)) \ge 1 - \max(O(n^{-1}), d^{-1})$ where $\widetilde{C}_5 = 2e\widetilde{K}_4$. This is what we wanted to show. □

Lemma E.9. We have:

$$n^{-1}\|\mathbf{P}_{\boldsymbol{\beta}^{*\perp}}\mathbf{X}^\top[(\boldsymbol{Y} - \boldsymbol{\mu}) \odot \mathbf{X}\boldsymbol{\beta}^* - c_0\mathbf{X}\boldsymbol{\beta}^*]\|_\infty \le C_6\sqrt{\frac{\log d}{n}}.$$

with probability at least $1 - O(n^{-1}) - 2d^{-1}$.

Proof of Lemma E.9. We have that $\mathbf{P}_{\boldsymbol{\beta}^{*\perp}}\mathbf{X}^\top$ is independent of $(\boldsymbol{Y} - \boldsymbol{\mu}) \odot \mathbf{X}\boldsymbol{\beta}^* - c_0\mathbf{X}\boldsymbol{\beta}^*$ and thus:

$$\frac{1}{n}\mathbf{P}_{\boldsymbol{\beta}^{*\perp}}\mathbf{X}^\top[(\boldsymbol{Y} - \boldsymbol{\mu}) \odot \mathbf{X}\boldsymbol{\beta}^* - c_0\mathbf{X}\boldsymbol{\beta}^*] \sim \mathcal{N}\Big(0, \frac{1}{n^2}\sum_{i=1}^n (Y_i - \mu - c_0)^2({\boldsymbol{\beta}^*}^\top \boldsymbol{X}_i)^2 \mathbf{P}_{\boldsymbol{\beta}^{*\perp}}\Big).$$

By Chebyshev's inequality we obtain that

$$\left|\frac{1}{n}\sum_{i=1}^n (Y_i - \mu - c_0)^2({\boldsymbol{\beta}^*}^\top \boldsymbol{X}_i)^2\right| \le 2\mathbb{E}((f(Z,\varepsilon) - \mu - c_0)^2 Z^2)$$

with probability at least $\frac{\mathrm{Var}((f(Z,\varepsilon)-\mu-c_0)^2Z^2)}{[\mathbb{E}((f(Z,\varepsilon)-\mu-c_0)^2Z^2)]^2 n}$. Since $\|\mathbf{P}_{\boldsymbol{\beta}^{*\perp}}\|_2 \le 1$, by a standard Gaussian tail bound we obtain:

$$\mathbb{P}\left(\|\mathbf{P}_{\boldsymbol{\beta}^{*\perp}}\mathbf{X}^\top[(\boldsymbol{Y} - \boldsymbol{\mu}) \odot \mathbf{X}\boldsymbol{\beta}^* - c_0\mathbf{X}\boldsymbol{\beta}^*]\|_\infty \ge t\right) \le 2d\exp\Big(-\frac{nt^2}{4\mathbb{E}((f(Z,\varepsilon) - \mu - c_0)^2 Z^2)}\Big).$$

Setting $t = 2\sqrt{2\mathbb{E}((f(Z,\varepsilon) - \mu - c_0)^2 Z^2)\frac{\log d}{n}}$, completes the proof. □