Near Optimal Heteroscedastic Regression with Symbiotic Learning

Dheeraj Baby DHEERAJ@UCSB.EDU
University of California, Santa Barbara

Aniket Das KETD@GOOGLE.COM
Google Research, Bangalore

Dheeraj Nagaraj DHEERAJNAGARAJ@GOOGLE.COM
Google Research, Bangalore

Praneeth Netrapalli PNETRAPALLI@GOOGLE.COM
Google Research, Bangalore

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

We consider the classical problem of heteroscedastic linear regression where, given n i.i.d. samples $(\mathbf{x}_i, y_i)$ drawn from the model $y_i = \langle \mathbf{w}^*, \mathbf{x}_i \rangle + \epsilon_i \cdot \langle \mathbf{f}^*, \mathbf{x}_i \rangle$, $\mathbf{x}_i \sim \mathcal{N}(0, \mathbf{I})$, $\epsilon_i \sim \mathcal{N}(0, 1)$, our aim is to *estimate the regressor* $\mathbf{w}^*$ *without prior knowledge of the noise parameter* $\mathbf{f}^*$. In addition to classical applications of such models in statistics (Jobson and Fuller, 1980), econometrics (Harvey, 1976), time series analysis (Engle, 1982) etc., it is also particularly relevant in machine learning problems where data is collected from multiple sources of varying (but apriori unknown) quality, e.g., in the training of large models (Devlin et al., 2019) on web-scale data. In this work, we develop an algorithm called *SymbLearn* (short for *Symb*iotic *Learn*ing) which estimates $\mathbf{w}^*$ in squared norm upto an error of $\tilde{O}(\|\mathbf{f}^*\|^2 \cdot ({}^{1}\!/\!{}_{n} + ({}^{d}\!/\!{}_{n})^2))$, and prove that this rate is minimax optimal modulo logarithmic factors. This represents a substantial improvement upon the previous best known upper bound of $\tilde{O}(\|\mathbf{f}^*\|^2 \cdot {}^{d}\!/\!{}_{n})$. Our algorithm is essentially an alternating minimization procedure which comprises of two key subroutines 1. An adaptation of the classical weighted least squares heuristic to estimate $\mathbf{w}^*$ (dating back to at least Davidian and Carroll (1987)), for which our work presents the first non-asymptotic guarantee; 2. A novel non-convex pseudogradient descent procedure for estimating $\mathbf{f}^*$, which draws inspiration from the phase retrieval literature. As corollaries of our analysis, we obtain fast non-asymptotic rates for two important problems, linear regression with multiplicative noise, and phase retrieval with multiplicative noise, both of which could be of independent interest. Beyond this, the proof of our lower bound, which involves a novel adaptation of LeCam's two point method for handling infinite mutual information quantities (thereby preventing a direct application of standard techniques such as Fano's method), could also be of broader interest for establishing lower bounds for other heteroscedastic or heavy tailed statistical problems.

Keywords: Linear regression, heteroscedasticity, phase retrieval, alternating minimization

1. Introduction

A popular trend in machine learning (ML) in the recent years has been the shift from training models on carefully curated datasets such as ImageNet (Deng et al., 2009), PennTreeBank (Marcus et al., 1993) etc. to training on a much larger corpus of data collected from all over the web (Devlin et al., 2019; Brown et al., 2020). While this has enabled training of models using much larger amounts of data, the data, so collected from all over the web, also has large variations in quality. It is clear that one should consider the quality of different data points while training the model

– giving more importance to high quality data points (i.e., less noise) and vice versa, rather than giving equal importance to all the data points. However, the quality of a data point is not apriori known and needs to be learned from the data itself. This motivates the problem of learning with *heteroscedastic* noise, which finds applications in several domains such as regression (Davidian and Carroll, 1987), large-scale classification (Collier et al., 2022) and deep learning (Patrini et al., 2017).

In this work, we consider a prototypical version of this problem: linear regression with heteroscedastic noise, where the noise variance is a rank 1 quadratic function of the covariates. That is, given n independently and identically distributed (i.i.d) samples $(\mathbf{x}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$ such that $\mathbf{x}_i \sim \mathcal{N}(0, \mathbf{I})$, $\epsilon_i \sim \mathcal{N}(0, 1)$, sampled independently of $\mathbf{x}_i$[1], and

$$\mathbf{y}_i = \langle \mathbf{w}^*, \mathbf{x}_i \rangle + \epsilon_i \langle \mathbf{f}^*, \mathbf{x}_i \rangle \,, \tag{1}$$

Our objective is to estimate $\mathbf{w}^*$ *without apriori knowledge of the noise model* $\mathbf{f}^*$. The problem of linear regression with heteroscedastic noise has been widely studied in statistics (Jobson and Fuller, 1980; Davidian and Carroll, 1987; Carroll and Ruppert, 2017), econometrics (Goldfeld and Quandat, 1972; Harvey, 1976), and timeseries analysis (Engle, 1982; Bollerslev, 1986; Nelson, 1991). Most of these classical works analyze maximum likelihood or other estimators and obtain asymptotic rates of convergence for a large class of noise models. The main message of these works is that, *asymptotically*, it is possible to obtain dramatically improved rates of estimation of $\mathbf{w}^*$ in the heteroscedastic model (1), compared to the homogeneous setting where $y_i = \langle \mathbf{w}^*, \mathbf{x}_i \rangle + \|\mathbf{f}^*\| \cdot \mathcal{N}(0, 1)$. While there have been empirical results suggesting that the rate of estimation for heteroscedastic setting can be improved over that of homogeneous setting even for small sample sizes n (Jobson and Fuller, 1980), it has been an open problem to obtain such improved *non-asymptotic* rates of estimation. In fact, even for the interesting special case of (1), where $\mathbf{f}^*$ is parallel to $\mathbf{w}^*$, used to capture covariate uncertainty (Harvey, 1976; Xu and Shimada, 2000; Xu, 2019), no better results are known compared to the homogeneous setting.

1.1. Contributions

The main contribution of this work is to propose an algorithm – SymbLearn (short for Symbiotic Learning) – that obtains an estimate $\widehat{\mathbf{w}}$ for (1) satisfying $\|\widehat{\mathbf{w}} - \mathbf{w}^*\|^2 = \tilde{O}\big(\|\mathbf{f}^*\|^2\big(\frac{1}{n} + \big(\frac{d}{n}\big)^2\big)\big)$, and an information theoretic lower bound of $\tilde{\Omega}\big(\|\mathbf{f}^*\|^2\big(\frac{1}{n} + \big(\frac{d}{n}\big)^2\big)\big)$, which matches the upper bound up to logarithmic factors. For $n > \tilde{\Omega}\left(d\right)$, this is a strict improvement over the estimation rate for the homogeneous setting, which is $\mathcal{O}\left(\|\mathbf{f}^*\|^2 \cdot \frac{d}{n}\right)$. An informal version of our result is presented below.

Theorem 1 (Main Result (Informal)) *Consider any $\delta \in (0, \frac{1}{2})$ and let $n \geq \tilde{\Omega}(d)$. With probability at least $1-\delta$, the output $\hat{\mathbf{w}}$ of SymbLearn (Algorithm 5) satisfies $\|\hat{\mathbf{w}} - \mathbf{w}^*\|^2 = \tilde{O}\left(\|\mathbf{f}^*\|^2 \left({}^1\!/_n + ({}^d\!/_n)^2\right)\right)$ for the heteroscedastic regression problem (1). The rate achieved by SymbLearn is minimax optimal up-to poly-logarithmic factors. For the same problem, the Ordinary Least Squares estimator $\hat{\mathbf{w}}_{\mathsf{OLS}}$ exhibits a sub-optimal error rate of $\|\hat{\mathbf{w}}_{\mathsf{OLS}} - \mathbf{w}^*\|^2 = \tilde{\Theta}(\|\mathbf{f}^*\|^2 {}^d\!/_n)$.*

Conceptually, our algorithm is an iterative re-weighted least squares algorithm, where data points are weighted according to their estimated noise. This idea has a long history (Jobson and Fuller,

[1] Our results can be extended to any well-conditioned covariance matrix Σ of $\mathbf{x}_i$ and any zero-mean subGaussian random variable ϵ_i, but for ease of exposition, we only consider identity covariance and standard normal ϵ_i.

1980) with impressive practical performance. To the best of our knowledge, our work presents the first non-asymptotic analysis of such an approach. Our key technical contributions are listed below:

Upper bound : We first show that the mere presence of structure in the noise as in (1) does not automatically improve the convergence rate of Ordinary Least Squares (OLS) estimator, which is computed as $\hat{\mathbf{w}}_{\mathsf{OLS}} = \mathrm{argmin}_{\mathbf{w}} 1/n \sum_{i=1}^{n} (\langle \mathbf{w}, \mathbf{x}_i \rangle - y_i)^2 = \big(\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T\big)^{-1} \big(\sum_{i=1}^{n} \mathbf{x}_i y_i\big)$. This exhibits a rate of $\|\hat{\mathbf{w}}_{\mathsf{OLS}} - \mathbf{w}^*\|^2 = \tilde{\Theta}\big(\|\mathbf{f}^*\|^2 {}^{d}/_{n}\big)$ for this problem (Theorem 3). This motivates us to consider a weighted least squares (WLS) objective: $\hat{\mathbf{w}}_{\mathsf{WLS}} = \mathrm{argmin}_{\mathbf{w}} 1/n \sum_{i=1}^{n} \alpha_i \cdot (\langle \mathbf{w}, \mathbf{x}_i \rangle - y_i)^2$, where the weight of an example α_i is inversely proportional to the variance of noise in that example i.e., $\alpha_i \propto \langle \mathbf{f}^*, \mathbf{x}_i \rangle^{-2}$. However, this cannot be implemented since $\mathbf{f}^*$ is unknown. Our first step is to compute an estimate $\widehat{\mathbf{f}}_{\mathsf{Spec}}$ of $\mathbf{f}^*$, using a standard spectral method, which guarantees that $\|\widehat{\mathbf{f}}_{\mathsf{Spec}} - \mathbf{f}^*\|^2 = \tilde{O}\big(\|\mathbf{f}^*\|^2 \cdot \frac{d}{n}\big)$. After selecting the weights α_i appropriately using $\widehat{\mathbf{f}}_{\mathsf{Spec}}$, and accounting for the uncertainty in this estimate compared to $\mathbf{f}^*$, we first show that the resulting estimate $\hat{\mathbf{w}}_{\mathsf{WLS}}$ achieves a rate of $\|\hat{\mathbf{w}}_{\mathsf{WLS}} - \mathbf{w}^*\|^2 = \tilde{O}\big(\|\mathbf{f}^*\|^2 \big(\frac{1}{n} + \big(\frac{d}{n}\big)^{1.5}\big)\big)$, which is already significantly better than the rate achieved by OLS. To show this, we perform a fine-grained analysis of the design matrix of WLS, which involves establishing a strict spectral gap for certain heavy-tailed random matrices that have infinite expectation. This rate can be improved further by obtaining better estimators for $\mathbf{f}^*$ than the spectral method, and using it to come up with better weights α_i in WLS. Concretely, given $\hat{\mathbf{w}}_{\mathsf{WLS}}$ and $\widehat{\mathbf{f}}_{\mathsf{Spec}}$, we hope to obtain a better estimate of $\mathbf{f}^*$ with $\widehat{\mathbf{f}}_{\mathsf{WLS}} = \mathrm{argmin}_{\mathbf{f}} \frac{1}{n} \sum_{i=1}^{n} \beta_i \cdot \Big((y_i - \langle \hat{\mathbf{w}}_{\mathsf{WLS}}, \mathbf{x}_i \rangle)^2 - \langle \mathbf{f}, \mathbf{x}_i \rangle^2\Big)^2$, for an appropriately chosen β_i depending on $\widehat{\mathbf{f}}_{\mathsf{Spec}}$. We analyze a *pseudogradient* descent algorithm on this objective and show that $\|\widehat{\mathbf{f}}_{\mathsf{WLS}} - \mathbf{f}^*\|^2 = \tilde{O}\big(\|\mathbf{f}^*\|^2 \big(\frac{1}{n} + \big(\frac{d}{n}\big)^{1.5}\big)\big)$, improving over $\widehat{\mathbf{f}}_{\mathsf{Spec}}$. The overall algorithm, SymbLearn, alternates between WLS estimation of $\mathbf{w}^*$ and pseudogradient descent estimation of $\mathbf{f}^*$, achieving a convergence rate of $\tilde{O}\big(\|\mathbf{f}^*\|^2 \big(\frac{1}{n} + \big(\frac{d}{n}\big)^2\big)\big)$ for both $\|\widehat{\mathbf{w}} - \mathbf{w}^*\|^2$ and $\|\widehat{\mathbf{f}} - \mathbf{f}^*\|^2$.

Lower bound : We show that SymbLearn is near optimal by proving a minimax lower bound of $\tilde{\Omega}(\|\mathbf{f}^*\|^2({}^1/_n + {}^{d^2}/_{n^2}))$ for $\mathbb{E}\|\widehat{\mathbf{w}} - \mathbf{w}^*\|^2$, which holds even when $\mathbf{f}^*$ is known. The $\frac{\|\mathbf{f}^*\|^2}{n}$ term which arises due to uncertainty of $\mathbf{w}^*$ in the direction parallel to $\mathbf{f}^*$ is straightforward to obtain. However the $\tilde{\Omega}\left(\|\mathbf{f}^*\|^2 (d/n)^2\right)$ term arising due to uncertainty in directions perpendicular to $\mathbf{f}^*$, is challenging to obtain. The key obstacle is that that for any two instances of the heteroscedastic regression model with regressors $\mathbf{w}_1^*$ and $\mathbf{w}_2^*$ and a common noise model $\mathbf{f}^*$, the KL divergence between them can be infinite, precluding the direct application of standard techniques such as Fano's method or Assouad's lemma (Tsybakov, 2009). Via a refined version of Assoud's lemma, we exploit the symmetry in Gaussian random variables to consider lower bounds for fixed designs (i.e, the covariates $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are fixed). We then reduce this to the problem of obtaining lower bounds for 'typical' fixed designs, which is then solved via an intricate covariate-bucketing argument which bounds certain heavy tailed random variables whose expectation is infinite. Our methods might be useful in establishing lower bounds in other statistical estimation problems with heteroscedastic or heavy-tailed noise.

As a byproduct of our analysis, we also obtain improved rates of estimation for both linear regression as well as phase retrieval with *multiplicative noise* as we describe below.

Linear Regression with Multiplicative Noise : A special case of the heteroscedastic regression problem, when $\mathbf{w}^* = \mathbf{f}^*$, corresponds to linear regression with multiplicative noise. The task is to estimate $\mathbf{w}^*$ given n i.i.d samples $(\mathbf{x}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$ such that $\mathbf{x}_i \sim \mathcal{N}(0, \mathbf{I})$ and $\mathbf{y}_i = \langle \mathbf{w}^*, \mathbf{x}_i \rangle (1 + \epsilon_i)$, with $\epsilon_i \sim \mathcal{N}(0,1)$ drawn independently of $\mathbf{x}_i$. While this is a classical model for regression with covariate uncertainty (Xu and Shimada, 2000), but the best known rate of estimation in the literature is still $\tilde{\Theta}(\|\mathbf{w}^*\|^2 d/n)$. While SymbLearn improves this to $\tilde{O}(\|\mathbf{w}^*\|^2(1/n + d^2/n^2))$, we also design a simpler algorithm, called Self-SymbLearn, which achieves this improved rate. Self-SymbLearn is similar to Iteratively Reweighted Least Squares (IRLS) (Mukhoty et al., 2019), and our analysis constitutes the first *nonasymptotic* analysis of such an algorithm for this problem.

Phase Retrieval with Multiplicative Noise The estimation of $\mathbf{f}^*$ (up-to a sign) is an important sub-routine in SymbLearn. This problem can be reduced to estimating $\mathbf{f}^*$ with data of the form $(\mathbf{x}_i, \epsilon_i^2(\langle \mathbf{f}^*, \mathbf{x}_i \rangle)^2 + \delta_i)$, where $\epsilon_i \sim \mathcal{N}(0,1)$ and $|\delta_i| \le \delta$. This corresponds to phase retrieval with multiplicative noise and a (small) additive adversarial error. Phase retrieval is a well-studied non-convex optimization problem encountered in physical sciences (Candes et al., 2015; Shechtman et al., 2015) and is typically studied without noise (see Chen et al. (2019) and references therein). While additive noise has been considered in this setting (Cai et al., 2016), multiplicative noise has not been studied in the literature to the best of our knowledge. This could help model covariate uncertainty in this setting as explained in Xu and Shimada (2000).

1.2. Related Work

Our work is most closely related to Anava and Mannor (2016); Chaudhuri et al. (2017), which obtain non-asymptotic rates of estimation of $\mathbf{w}^*$ in the heteroscedastic regression model (1). While Anava and Mannor (2016) considers this problem through the lens of online learning, Chaudhuri et al. (2017) considers the active learning setting. In the offline setting considered in this paper, neither of these works improve upon the OLS rate of $\|\widehat{\mathbf{w}} - \mathbf{w}^*\|^2 \le \mathcal{O}\left(\|\mathbf{f}^*\|^2 \cdot d/n\right)$

Our work is an instance of *statistical learning with a nuisance component*, with the regressor $\mathbf{w}^*$ being the *target parameter* and the noise model $\mathbf{f}^*$ being a *nuisance parameter*.Thus, our approach bears some resemblance to well-established methods for this problem, such as Double/Debiased Machine Learning (Chernozhukov et al., 2018a,b; Fingerhut et al., 2022) and Orthogonal Statistical Learning (or OSL) (Mackey et al., 2018; Foster and Syrgkanis, 2023; Liu et al., 2022), where an estimate of the nuisance parameter can improve the estimation of the target parameter. These methods partition the data into disjoint subsets (sample splitting), using the first subset to estimate the nuisance parameter, and then estimating the target parameter using the nuisance estimate and the second subset. Similarly, SymbLearn uses separate subsets for estimating $\mathbf{w}^*$ and $\mathbf{f}^*$ in each iteration. However, our approach is cyclic and iterative, where *a rough initial estimate of the target parameter can lead to improved nuisance parameter estimation, which in turn allows refined estimation of the target parameter and so on.* This leads to our alternating minimization-based approach where each step alternates between $\mathbf{w}^*$ estimation and $\mathbf{f}^*$ estimation. Typical realizations of OSL do not use any apriori target parameter estimate for nuisance estimation and is not iterative and cyclic (e.g. Meta-Algorithm 1 of Foster and Syrgkanis (2023) estimates the nuisance as $\hat{g} = \mathrm{Alg}(\mathcal{G}, S_1)$ and target as $\hat{\theta} = \mathrm{Alg}(\mathcal{G}, S_2; \hat{g})$). In our case, such a single stage estimation achieves a sub-optimal rate of $\tilde{O}(\|\mathbf{f}^*\|^2(1/n + (d/n)^{1.5}))$ (see discussion after Theorem 5). Moreover, the OSL meta-algorithm treats $\mathrm{Alg}(\mathcal{G}, S_1)$ and $\mathrm{Alg}(\mathcal{G}, S_2; \hat{g})$ as blackbox subroutines with certain high-probability convergence guarantees, whereas we design such and analyze such subroutines for heteroscedastic regression.

1.3. Organization

In Section 2, we present the problem setting and preliminaries. In Section 3, we present algorithms to solve partial versions of the heteroscedastic problem, with additional information. By using these algorithms as subroutines, we derive our main algorithm SymbLearn for the full heteroscedastic problem in Section 4. We present our main results for this algorithm as well as a matching lower bound in Section 5 with a high level proof idea in Section 6. We present some experimental results in Section 7 and conclude in Section 8.

1.4. Notation

The boldface lower letters (e.g. $\mathbf{x}$) represent vectors in $\mathbb{R}^d$ and boldface capital letters (e.g. $\mathbf{A}$) represent matrices in $\mathbb{R}^{m\times n}$. $\mathbf{A}_i$ denotes i^{th} row of matrix $\mathbf{A}$ and x_j denotes j^{th} element of vector $\mathbf{x}$. For an indexed vector $\mathbf{x}_i$; we use $x_{i,j}$ to denote the j^{th} element of $\mathbf{x}_i$. We use $[n]$ to denote the set $\{1, 2, \dots, n\}$. The ℓ_p norm of a vector $\mathbf{v}$ is denoted using $\|\mathbf{v}\|_p$. For any matrix $\mathbf{A}$, $\|\mathbf{A}\|_2$ and $\|\mathbf{A}\|_F$ denote the spectral and Frobenius norms of $\mathbf{A}$ respectively. We let $\mathbf{I}$ denote the identity matrix, whose dimension is clear from the context. Unless otherwise specified, $\|\mathbf{v}\| = \|\mathbf{v}\|_2$ for any vector $\mathbf{v}$ and $\|\mathbf{A}\| = \|\mathbf{A}\|_2$ for any matrix $\mathbf{A}$. We use the O notation to characterize the dependence of our error bounds on the number of samples n, the covariate dimension d and the confidence level δ, suppressing numerical constants. The $\tilde{O}$ notation suppresses polylogarithmic factors in n, d and $1/\delta$. We also assume that $\delta \leq 1/2$ wherever it appears so we can write $\log(C/\delta) \leq C\log(1/\delta)$. By $\mathsf{polylog}(x)$ we refer to some fixed poly-logarithmic function evaluated at x which can be different when invoked in different bounds. We also hide $\mathsf{polylog}(dn/\delta)$ terms within the $\tilde{O}$ notation.

2. Problem Formulation and Preliminaries

Our data set consists of i.i.d samples $(\mathbf{x}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$ of $i \in [n]$. We assume that $\mathbf{x}_i \sim \mathcal{N}(0, \mathbf{I})$ and $y_i = \langle \mathbf{w}^*, \mathbf{x}_i\rangle + \epsilon_i \langle \mathbf{f}^*, \mathbf{x}_i\rangle$ where $\epsilon_i \sim \mathcal{N}(0, 1)$ and independent of $\mathbf{x}_i$, and $\mathbf{w}^*, \mathbf{f}^* \in \mathbb{R}^d$ are unknown vectors. Our task is to estimate the regressor $\mathbf{w}^*$ without prior knowledge of $\mathbf{f}^*$. For the rest of this paper, we assume that the data is sampled from this model.

Remark 2 *We can relax our assumptions to allow sub-Gaussian ϵ_i. We can also consider $\mathbf{x}_i \sim \mathcal{N}(0, \Sigma)$ and 'whiten' it by estimating $\hat{\Sigma} \approx \Sigma$ and considering $\hat{\Sigma}^{-1/2}\mathbf{x}_i$ as the co-variates. We do not consider these scenarios for the sake of simplicity.*

2.1. Estimation of $\mathbf{w}^*$ using ordinary least squares (OLS)

Algorithm 1 OLS

```
Input: $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n) \in \mathbb{R}^d \times \mathbb{R}$.
$\hat{\mathbf{w}}_{\mathsf{OLS}} = \left(\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^T\right)^{-1}\left(\sum_{i=1}^n \mathbf{x}_i y_i\right)$
Output $\hat{\mathbf{w}}_{\mathsf{OLS}}$
```

Algorithm 2 Spectral Method

```
Input: $\hat{\mathbf{w}} \in \mathbb{R}^d$, $(\mathbf{x}_i, y_i)_{i=1}^n \in \mathbb{R}^d \times \mathbb{R}$
$\hat{\mathbf{S}} = \frac{1}{n}\sum_{i=1}^n (y_i - \langle \hat{\mathbf{w}}, \mathbf{x}_i\rangle)^2 \mathbf{x}_i\mathbf{x}_i^T$
$\mathbf{u} = $ Top Eigenvector$(\hat{\mathbf{S}})$, $R = \sqrt{\frac{\|\hat{\mathbf{S}}\|}{3}}$
Output $R \cdot \mathbf{u}$.
```

A classical approach for estimating $\mathbf{w}^*$ is through ordinary least squares (OLS) (pseudocode in Algorithm 1). The following result, proved in Appendix A.4, gives the convergence rate of OLS.

Theorem 3 (Ordinary Least Squares(OLS)) *For any $\delta \in (0, \frac{1}{2})$, and $n \geq d\mathsf{polylog}(\frac{d}{\delta})$, the OLS estimator for the heteroscedastic linear regression problem satisfies $\|\hat{\mathbf{w}}_{\mathsf{OLS}} - \mathbf{w}^*\|^2 \leq \frac{d\|\mathbf{f}^*\|^2}{n}\mathsf{polylog}(\frac{nd}{\delta})$ with probability at least $1-\delta$. Furthermore, $\|\hat{\mathbf{w}}_{\mathsf{OLS}} - \mathbf{w}^*\|^2 = \Theta({}^{d\|\mathbf{f}^*\|^2}/{}_{n})$ with probability at least $1 - d/n^c$, $c \geq 1$.*

2.2. Estimation of $\mathbf{f}^*$ using spectral method

A standard approach to estimate $\mathbf{f}^*$, given an estimate $\hat{\mathbf{w}} \approx \mathbf{w}^*$, is through the spectral method (Chaudhuri et al., 2017; Chen et al., 2021), originally proposed in the context of phase retrieval (Netrapalli et al., 2013). A pseudocode is presented in Algorithm 2. The following theorem, which is proved in Appendix B, gives a performance guarantee for the spectral method.

Theorem 4 (Spectral Method) *Consider any $\delta \in (0, {}^{1}/{}_{2})$. Suppose we have $\hat{\mathbf{w}}$ satisfying $\|\hat{\mathbf{w}} - \mathbf{w}^*\|^2 \leq \epsilon$. Then, for $n \geq d\mathsf{polylog}(\frac{d}{\delta})$, the output $\hat{\mathbf{f}}_{\mathsf{S}}$ of the spectral method satisfies:*

$$\|\hat{\mathbf{f}}_{\mathsf{S}} - \mathbf{f}^*\|^2 \leq \left(\left(\|\mathbf{f}^*\|^2 + \epsilon\right)\frac{d}{n} + \frac{\epsilon^2}{\|\mathbf{f}^*\|^2}\right)\mathsf{polylog}(\tfrac{nd}{\delta}).$$

Remark: Note that $\mathbf{f}^*$ can be recovered only up to a sign i.e., both $\pm\mathbf{f}^*$ are valid solutions. For any estimate $\widehat{\mathbf{f}}$, by $\|\widehat{\mathbf{f}} - \mathbf{f}^*\|$ we mean $\min\left(\|\widehat{\mathbf{f}} - \mathbf{f}^*\|, \|\widehat{\mathbf{f}} + \mathbf{f}^*\|\right)$. Theorem 3 and 4 imply that running the spectral method with $\hat{\mathbf{w}} = \hat{\mathbf{w}}_{\mathsf{OLS}}$ gives us an estimate $\hat{\mathbf{f}}_{\mathsf{S}}$ that satisfies $\|\hat{\mathbf{f}}_{\mathsf{S}} - \mathbf{f}^*\|^2 \leq \tilde{O}(\|\mathbf{f}^*\|^2 \cdot {}^{d}/{}_{n})$.

3. Estimating with Partial Information

In this section we discuss procedures to estimate $\mathbf{w}^*$ and $\mathbf{f}^*$ when partial information is known. We then combine these procedures to develop our Algorithm SymbLearn in Section 4.

3.1. Estimate $\mathbf{w}^*$ when $\mathbf{f}^*$ is approximately known

```
Algorithm 3 Weighted Least Squares (WLS)
Input: $(\mathbf{x}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$ for $i \in [n]$, noise model $\hat{\mathbf{f}} \in \mathbb{R}^d$, reg. parameter $\lambda$
  Output $\hat{\mathbf{w}}_{\hat{\mathbf{f}},\lambda}$ given by $\hat{\mathbf{w}}_{\hat{\mathbf{f}},\lambda} = \left(\sum_{i=1}^n \frac{\mathbf{x}_i\mathbf{x}_i^T}{\langle\hat{\mathbf{f}},\mathbf{x}_i\rangle^2+\lambda}\right)^{-1}\left(\sum_{i=1}^n \frac{\mathbf{x}_i y_i}{\langle\hat{\mathbf{f}},\mathbf{x}_i\rangle^2+\lambda}\right)$.
```

Suppose $\hat{\mathbf{f}}$ is known such that $\hat{\mathbf{f}} \approx \mathbf{f}^*$. Then, the weighted least squares (WLS) estimator $\hat{\mathbf{w}}_{\hat{\mathbf{f}},\lambda}$ is given in Algorithm 3, where $\lambda \geq 0$ is a regularizer. The intuition for the weights $\left(\langle\hat{\mathbf{f}}, \mathbf{x}_i\rangle^2 + \lambda\right)^{-1}$ is that the variance of the observation y_i conditioned on $\mathbf{x}_i$ is $\langle\mathbf{f}^*, \mathbf{x}_i\rangle^2$. The role of λ is to ensure the weights don't become arbitrarily large when $\langle\mathbf{f}^*, \mathbf{x}_i\rangle^2$ is very small as well as to account for the approximation error $\|\hat{\mathbf{f}} - \mathbf{f}^*\|$. This is made precise in the following theorem.

Theorem 5 (Weighted Least Squares (WLS)) *Consider any $\delta \in (0, \frac{1}{2})$. Suppose we know $\hat{\mathbf{f}}$ such that $\left\|\hat{\mathbf{f}} - \mathbf{f}^*\right\|^2 \leq \epsilon$. If λ is chosen such that $\lambda \geq \max\{\epsilon, \|\hat{\mathbf{f}}\|^2 {}^{d^2}/{}_{n^2}\}\mathsf{polylog}(\frac{nd}{\delta})$, the weighted least squares estimator $\hat{\mathbf{w}}_{\hat{\mathbf{f}},\lambda}$ satisfies the following with probability at least $1-\delta$:*

$$\|\hat{\mathbf{w}}_{\hat{\mathbf{f}},\lambda} - \mathbf{w}^*\|^2 \leq \left[\frac{\|\mathbf{f}^*\|^2}{n} + \frac{d\|\mathbf{f}^*\|\sqrt{\lambda}}{n} + \frac{\epsilon}{n} + \frac{d\sqrt{\epsilon\lambda}}{n}\right]\mathsf{polylog}(\tfrac{nd}{\delta}).$$

The WLS estimator obtains a better estimate of $\mathbf{w}^*$ than OLS using an estimate $\hat{\mathbf{f}}$ of $\mathbf{f}^*$. In particular, for $\epsilon = 0$ (i.e., $\hat{\mathbf{f}} = \mathbf{f}^*$), WLS attains a rate of $\tilde{O}(\|\mathbf{f}^*\|^2(\frac{1}{n} + \frac{d^2}{n^2}))$. This is in contrast to OLS, which does not utilize the knowledge of $\mathbf{f}^*$, and consequently achieves a suboptimal rate of $\tilde{O}\big(\|\mathbf{f}^*\|^2 \cdot \frac{d}{n}\big)$. We refer to Appendix A.2 for a proof of Theorem 5.

Obtaining $\tilde{O}(\|\mathbf{f}^*\|^2(\frac{1}{n} + (\frac{d}{n})^{1.5}))$ Rates From Theorems 3 and 4, we know that the output of the spectral method $\hat{\mathbf{f}}_{\mathsf{S}}$ (computed using an OLS estimate) satisfies $\|\hat{\mathbf{f}}_{\mathsf{S}} - \mathbf{f}^*\|^2 = \tilde{O}(\|\mathbf{f}^*\|^2 d/n)$. Moreover, from Theorem 5 we note that the WLS estimate computed using $\hat{\mathbf{f}}_{\mathsf{S}}$ $\hat{\mathbf{w}} = \hat{\mathbf{w}}_{\hat{\mathbf{f}}_{\mathsf{S}},\lambda}$ (where $\lambda = \tilde{\Theta}(\|\hat{\mathbf{f}}_{\mathsf{S}}\| d/n)$) satisfies $\|\hat{\mathbf{w}} - \mathbf{w}^*\|^2 = \tilde{O}(\|\mathbf{f}^*\|^2(\frac{1}{n} + (\frac{d}{n})^{1.5}))$. This rate is an improvement upon the OLS estimate but strictly worse than the minimax optimal rate attained by SymbLearn . We show in Section 7, the empirical performance of this strategy is also worse than that of SymbLearn.

3.2. Estimate $\mathbf{f}^*$ when $\mathbf{w}^*$ is approximately known

Algorithm 4 `Phase Retrieval with Multiplicative Noise`

Input: samples $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$. Estimates $\hat{\mathbf{w}}, \hat{\mathbf{f}}$, relaxation parameter $\bar{\mu}$, step sizes α_0, α_1 and number of steps K.

```
Divide the data into $K$ batches of size $m = \lfloor n/K \rfloor$ as $\left\{(\mathbf{x}_1^{(t)}, y_1^{(t)}), \ldots, (\mathbf{x}_m^{(t)}, y_m^{(t)})\right\}_{t=1}^{K}$
Set $\hat{\mathbf{f}}_0 = \hat{\mathbf{f}}$, matrix valued step size $\mathbf{D} = \alpha_0 \frac{\hat{\mathbf{f}}\hat{\mathbf{f}}^T}{\|\hat{\mathbf{f}}\|^2} + \alpha_1 \left(\mathbf{I} - \frac{\hat{\mathbf{f}}\hat{\mathbf{f}}^T}{\|\hat{\mathbf{f}}\|^2}\right)$
for $t \in \{0, \ldots, K-1\}$ do
    Set pseudo stochastic gradient $\mathbf{g}_t$ as
        $$\mathbf{g}_t \leftarrow \frac{1}{m} \sum_{i=1}^{m} \langle \hat{\mathbf{f}}, \mathbf{x}_i^{(t)} \rangle \mathbf{x}_i^{(t)} \mathbb{1}\left(|\langle \hat{\mathbf{f}}, \mathbf{x}_i^{(t)} \rangle| \geq \bar{\mu}\right) \frac{\langle \mathbf{f}_k, \mathbf{x}_i^{(t)} \rangle^2 - \left(y_i^{(t)} - \langle \hat{\mathbf{w}}, \mathbf{x}_i^{(t)} \rangle\right)^2}{\langle \hat{\mathbf{f}}, \mathbf{x}_i^{(t)} \rangle^4}$$
    Perform the pseudo-SGD update $\mathbf{f}_{t+1} = \mathbf{f}_t - \mathbf{D}\mathbf{g}_t$
end for
Output $\mathbf{f}_K$
```

We now wish to refine our estimate of $\mathbf{f}^*$ using estimates $\widehat{\mathbf{w}}$ and $\widehat{\mathbf{f}}$ of $\mathbf{w}^*$ and $\mathbf{f}^*$ respectively. For this, we design a weighted phase retrieval algorithm which adapts to the quality of $\widehat{\mathbf{w}}$ and $\widehat{\mathbf{f}}$. Suppose we are given $\widehat{\mathbf{w}} = \mathbf{w}^*$, then we observe that $y_i - \langle \widehat{\mathbf{w}}, \mathbf{x}_i \rangle = \epsilon_i \cdot \langle \mathbf{f}^*, \mathbf{x}_i \rangle$. Since ϵ_i is a zero-mean symmetric random variable, this is equivalent to observing $(y_i - \langle \mathbf{w}^*, \mathbf{x}_i \rangle)^2 = \epsilon_i^2 \langle \mathbf{f}^*, \mathbf{x}_i \rangle^2 = (1 + \zeta_i) \langle \mathbf{f}^*, \mathbf{x}_i \rangle^2$, where $\zeta_i = \epsilon_i^2 - 1$ is a zero mean subexponential random variable. Thus, estimating $\pm\mathbf{f}^*$ given $(\mathbf{x}_i, (1 + \zeta_i) \langle \mathbf{f}^*, \mathbf{x}_i \rangle^2)$ is phase retrieval with multiplicative noise.
When expressed as $\epsilon_i^2 \langle \mathbf{f}^*(\mathbf{f}^*)^\intercal, \mathbf{x}_i(\mathbf{x}_i)^\intercal \rangle_{\mathsf{HS}}$ ($\langle \cdot, \cdot \rangle_{\mathsf{HS}}$ is the Hilbert-Schmidt inner product), this reduces to linear regression with multiplicative noise in the space of rank-1 matrices. Thus, we consider estimating $\mathbf{f}^*$ by (roughly) solving for the minimizer of the weighted loss:

$$\mathcal{L}^{\mathsf{mul}}(\mathbf{f}) = \frac{1}{n} \sum_{i=1}^{n} \frac{((y_i - \langle \hat{\mathbf{w}}, \mathbf{x}_i \rangle)^2 - \langle \mathbf{f}, \mathbf{x}_i \rangle^2)^2}{\langle \hat{\mathbf{f}}, \mathbf{x}_i \rangle^4} \tag{2}$$

The term $\langle \hat{\mathbf{f}}, \mathbf{x}_i \rangle^4$ appears in the denominator since $(y_i - \langle \hat{\mathbf{w}}, \mathbf{x}_i \rangle)^2$ has a variance of the order $\langle \mathbf{f}^*, \mathbf{x}_i \rangle^4$ conditioned on $\mathbf{x}_i$ and we re-weight just like in Algorithm 3. To this end, we design

Algorithm 4 to minimize $\mathcal{L}^{\mathsf{mul}}(\mathbf{f})$ via an approximate gradient descent procedure The following theorem presents the performance guarantee for Algorithm 4. Its proof is presented in Appendix C.

Theorem 6 (Phase retrieval with multiplicative noise) *In Algorithm 4, Assume $\bar{\mu} < \|\hat{\mathbf{f}}\|$, let $r := \lceil \log\left(\frac{\|\hat{\mathbf{f}}\|_2}{\bar{\mu}}\right) \rceil$, $\Delta_{\hat{\mathbf{w}}} := \mathbf{w}^* - \hat{\mathbf{w}}$, $\Delta_{\hat{\mathbf{f}}} := \mathbf{f}^* - \hat{\mathbf{f}}$ and $\delta \in (0, 1/2)$. Let the step sizes be $\alpha_0 = c\|\hat{\mathbf{f}}\|^2$ and $\alpha_1 = c\bar{\mu}\|\hat{\mathbf{f}}\|$ for some small enough constant c. Suppose for large enough constant C^{par}:*

$$\bar{\mu} > C^{\mathsf{par}} \max(\|\Delta_{\hat{\mathbf{f}}}\|, \|\Delta_{\hat{\mathbf{w}}}\|) \log \tfrac{m}{\delta}; \quad m \geq C^{\mathsf{par}} \max\left(\tfrac{\|\hat{\mathbf{f}}\|}{\bar{\mu}} \log\left(\tfrac{r}{\delta}\right), \tfrac{\|\hat{\mathbf{f}}\| d \log^4(\frac{m}{\delta})}{\bar{\mu}}, \tfrac{\|\hat{\mathbf{f}}\|^2 \log^4(\frac{m}{\delta})}{\bar{\mu}^2}\right).$$

Then, the output $\mathbf{f}_K$ of K steps of Algorithm 4 satisfies the following with probability at least $1 - K\delta$:

$$\|\mathbf{f}_K - \mathbf{f}^*\| \leq \exp(-\gamma K)\|\Delta_{\hat{\mathbf{f}}}\| + C\left[\frac{\|\hat{\mathbf{f}}\| \log\left(\frac{1}{\delta}\right)}{\sqrt{m}} + \frac{\|\Delta_{\hat{\mathbf{w}}}\|^2}{\bar{\mu}} + \log\left(\tfrac{d}{\delta}\right)\sqrt{\frac{d\bar{\mu}\|\hat{\mathbf{f}}\|}{m}}\right].$$

where $\gamma > 0$ is a universal constant. In particular, setting $K = \Theta(\log(m))$ and $\bar{\mu} = \tilde{\Theta}\left(\max\{\|\Delta_{\hat{\mathbf{f}}}\|, \|\Delta_{\hat{\mathbf{w}}}\|\}\right)$ we have:

$$\|\mathbf{f}_K - \mathbf{f}^*\|^2 \leq \tilde{O}\left(\frac{\|\mathbf{f}^*\|^2}{m} + \|\Delta_{\hat{\mathbf{w}}}\|^2 + \|\Delta_{\hat{\mathbf{f}}}\|^2 \cdot \frac{d}{m} + \frac{d}{m}\|\mathbf{f}^*\| \cdot (\|\Delta_{\hat{\mathbf{w}}}\| + \|\Delta_{\hat{\mathbf{f}}}\|)\right)$$

Quality Adaptivity of Noisy Phase Retrieval If $\hat{\mathbf{f}}$ is the spectral method estimate of $\mathbf{f}^*$, and $\hat{\mathbf{w}} = \hat{\mathbf{w}}_{\hat{\mathbf{f}},\lambda}$ (output of WLS) for appropriately chosen $\lambda = \tilde{\Theta}(\|\hat{\mathbf{f}}\|^2 d/n)$, then $\|\hat{\mathbf{w}} - \mathbf{w}^*\|^2 = \tilde{O}\left(\|\mathbf{f}^*\|^2 (1/n + (d/n)^{3/2})\right)$ is lower order compared to $\|\hat{\mathbf{f}} - \mathbf{f}^*\|^2 = \tilde{O}\left(\|\mathbf{f}^*\|^2 \cdot d/n\right)$. In this case, $\|\mathbf{f}_K - \mathbf{f}^*\|^2 \leq \tilde{O}(\|\hat{\mathbf{w}} - \mathbf{w}^*\|^2)$. This *transfers the error in estimation of $\mathbf{w}^*$ to the error in estimation of* $\mathbf{f}^*$. That is, a better estimate of $\mathbf{w}^*$ leads to a better estimate of $\mathbf{f}^*$, which in turn improves $\mathbf{w}^*$ estimation via WLS. Having developed quality-adaptive algorithms for estimating both $\mathbf{w}^*$ and $\mathbf{f}^*$, the SymbLearn algorithm naturally follows by iteratively alternating between the two.

4. Algorithm

Combining our observations in Section 3 we present the main algorithm of our work. In the general case, $\mathbf{f}^*$ is arbitrary and unknown. For the sake of simplicity, we will assume that we can divide the dataset into $2K$ disjoint parts. We illustrate our algorithm in Figure 1

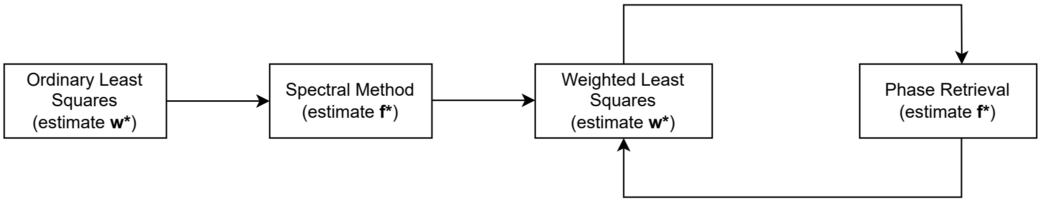

Ordinary Least Squares (estimate **w***)
Spectral Method (estimate **f***)
Weighted Least Squares (estimate **w***)
Phase Retrieval (estimate **f***)

Figure 1: An Illustration of SymbLearn

Regression with Multiplicative Noise When $\mathbf{w}^* = \pm\mathbf{f}^*$, the problem coincides with linear regression with multiplicative noise. For this special case, we design a much simpler algorithm (Self-SymbLearn , see Appendix A.3), which given an estimate of $\hat{\mathbf{w}} \approx \mathbf{w}^*$, iteratively applies the weighted least squares procedure (Algorithm 3) to get a better estimate with $\hat{\mathbf{f}} = \hat{\mathbf{w}}$.

Algorithm 5 SymbLearn

Require: $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n) \in \mathbb{R}^d \times \mathbb{R}$. Number of steps K. Number of phase retrieval steps K_p. Weights $\lambda_1, \dots, \lambda_K$, $\bar{\mu}_1, \dots, \bar{\mu}_K$ and step sizes $(\alpha_0^1, \alpha_1^1), \dots, (\alpha_0^K, \alpha_1^K)$.

```
Divide the data into $2K$ disjoint parts
Compute $\hat{\mathbf{w}}^{(1)}$ with OLS (Algorithm 1) and 1st data partition
Compute $\hat{\mathbf{f}}^{(1)}$ using the spectral method with input $\hat{\mathbf{w}}^{(1)}$ (Algorithm 2) and 2nd data partition
for $k \in \{1, \dots, K-1\}$ do
    Compute $\hat{\mathbf{w}}^{(k+1)}$ using WLS (Algorithm 3) with input parameters $\lambda_k$ and $\hat{\mathbf{f}}^k$ and data from $(2k+1)^{\text{th}}$ partition
    Compute $\hat{\mathbf{f}}^{(k+1)}$ using Phase Retrieval (Algorithm 4) for $K_p$ steps with inputs $(\hat{\mathbf{w}}^{(k)}, \hat{\mathbf{f}}^{(k)}, \bar{\mu}_k, \alpha_0^k, \alpha_1^k)$ and data from $(2k+2)^{\text{th}}$ partition
end for
Output $\hat{\mathbf{w}}^K$.
```

5. Main Results

Our first main result stated below establishes the estimation guarantee for $\mathbf{w}^*$ via SymbLearn. The proof of this theorem is presented in Appendix D.

Theorem 7 (SymbLearn) *Consider any $\delta \in (0, 1/2)$. Let $S_k = \sum_{j=0}^{k} 1/2^j$. Then, for $n \geq d\text{polylog}(d/\delta)$, $K = \lceil \log_2(n) \rceil$ and $K_p = \Theta(\log(n/K))$, if we run SymbLearn (Algorithm 5) with $\bar{\mu}_k = \|\hat{\mathbf{f}}_k\| \cdot \sqrt{\max(\frac{k}{m}, \frac{(k)d^2}{m^2}) + (d/m)^{S_k}}\,\text{polylog}(\frac{nd}{\delta})$, $\lambda_k = \|\hat{\mathbf{f}}_k\|^2 \left(\max(\frac{k}{m}, \frac{(k)d^2}{m^2}) + (d/m)^{S_k} \right) \text{polylog}(\frac{nd}{\delta})$ then the output $\hat{\mathbf{w}}_K$ satisfies the following with probability at least $1 - \delta$:*

$$\|\hat{\mathbf{w}}_K - \mathbf{w}^*\|^2 \leq \tilde{O}\left(\|\mathbf{f}^*\|^2 \left(1/n + d^2/n^2 \right) \right),$$

Our second result below establishes the minimax optimality of SymbLearn with a lower bound construction. Let $P_{\mathbf{w},\mathbf{f}}$ denote the heteroscedastic regression model parameterized by $\mathbf{w}$ and $\mathbf{f}$, i.e., $P_{\mathbf{w},\mathbf{f}}$ is a probability measure supported on $\mathbb{R}^d \times \mathbb{R}$ such that $(\mathbf{x}, y) \sim P_{\mathbf{w},\mathbf{f}}$ implies $\mathbf{x} \sim \mathcal{N}(0, \mathbf{I})$ and $y \mid \mathbf{x} \sim \mathcal{N}(\langle \mathbf{w}, \mathbf{x} \rangle, \langle \mathbf{f}, \mathbf{x} \rangle^2)$. We refer to Appendix G for the proof.

Theorem 8 (Minimax Lower Bound for Heteroscedastic Regression) *Consider any $\mathbf{f}^* \in \mathbb{R}^d$, and let $\hat{\mathbf{w}} : \left(\mathbb{R}^d \times \mathbb{R}\right)^n \to \mathbb{R}^d$ denote any arbitrary measurable function that estimates $\mathbf{w}^*$ given i.i.d samples $(\mathbf{x}_i, y_i)_{i \in [n]} \overset{\text{iid}}{\sim} P_{\mathbf{w}^*,\mathbf{f}^*}$. Furthermore, let $\mathcal{W}$ denote the family of estimators $\hat{\mathbf{w}}$. Then,*

$$\inf_{\hat{\mathbf{w}} \in \mathcal{W}} \sup_{\mathbf{w}^* \in \mathbb{R}^d} \mathbb{E}_{(\mathbf{x}_i, y_i)_{i \in [n]} \overset{\text{iid}}{\sim} P_{\mathbf{w}^*,\mathbf{f}^*}} \left[\|\hat{\mathbf{w}} - \mathbf{w}^*\|^2 \right] \geq \tilde{\Omega}\left(\frac{\|\mathbf{f}^*\|^2}{n} + \frac{\|\mathbf{f}^*\|^2 d^2}{n^2} \right).$$

6. Key Ideas behind Proofs

6.1. Weighted Least Squares (WLS) : Theorem 5

For simplicity, consider the setting where $\hat{\mathbf{f}} = \mathbf{f}^*$ and $\lambda = 0$ in Algorithm 3. The empirical covariance matrix $\mathbf{H} = \frac{1}{n} \sum_{i=1}^{n} \frac{\mathbf{x}_i \mathbf{x}_i^\intercal}{\langle \mathbf{x}_i, \mathbf{f}^* \rangle^2}$ has infinite expectation. We show via a straight forward calculation that the expected squared error $\mathbb{E}\left[\|\hat{\mathbf{w}}_{\mathsf{WLS}} - \mathbf{w}^*\|^2\right] = \frac{1}{n}\mathbb{E}\,\mathrm{Tr}\big(\mathbf{H}^{-1}\big)$. The main technical difficulty is to precisely upper bound this trace. Note that $(\mathbf{f}^*)^\intercal\mathbf{H}\mathbf{f}^* = 1$ and for $\mathbf{u} \perp \mathbf{f}^*$, we have that $\langle \mathbf{x}_i, \mathbf{u}\rangle$ is independent of $\langle \mathbf{x}_i, \mathbf{f}^*\rangle$. Note that $\mathbf{u}^\intercal\mathbf{H}\mathbf{u} = \frac{1}{n}\sum_{i=1}^n \frac{\langle \mathbf{u},\mathbf{x}_i\rangle^2}{\langle \mathbf{f}^*,\mathbf{x}_i\rangle^2}$. We show that, typically there exist $O(1)$ samples i such that $|\langle \mathbf{x}_i, \mathbf{f}^*\rangle|$ has the value $\tilde{\Theta}(\frac{1}{n})$ and the corresponding value of $\langle \mathbf{u}, \mathbf{x}_i\rangle$ is $\tilde{\Theta}(1)$. We show that such extreme values contribute most to $\frac{1}{n}\sum_{i=1}^n \frac{\langle \mathbf{u},\mathbf{x}_i\rangle^2}{\langle \mathbf{f}^*,\mathbf{x}_i\rangle^2}$, allowing us to show that $\mathbf{u}^\intercal\mathbf{H}\mathbf{u} = \tilde{\Theta}(n)$ followed by a covering argument to show that $\mathbf{H} \succeq \|\mathbf{f}^*\|^{-2}\left[\mathbf{e}\mathbf{e}^T + \frac{n}{d}\left(\mathbf{I} - \mathbf{e}\mathbf{e}^T\right)\right]$ with high probability ($\mathbf{e} := \mathbf{f}^*/\|\mathbf{f}^*\|$, the extra $1/d$ appearing due to the covering argument). This concludes the proof of Theorem 5 in this setting. Handling $\widehat{\mathbf{f}} \neq \mathbf{f}^*$ and $\lambda > 0$ are straight forward and is described in the full proof of Theorem 5 in Appendix A.2.

6.2. Phase retrieval with multiplicative noise : Theorem 6

Algorithm 4 estimates $\mathbf{f}^*$ given prior estimates of $\mathbf{f}^*$ and $\mathbf{w}^*$ via pseudogradient descent on a carefully designed nonconvex squared loss, which draws motivation from similar approaches in the noise-free phase retrieval literature (Chen et al., 2019; Netrapalli et al., 2013). Consider the hypothetical scenario in which we have apriori access to both $\mathbf{f}^*$ and $\mathbf{w}^*$, and iid samples $(\mathbf{x}_i, y_i)_{i\in[n]}$. We observe that $(y_i - \langle \mathbf{w}^*, \mathbf{x}_i\rangle)^2 = \epsilon_i^2\langle \mathbf{f}^*, \mathbf{x}_i\rangle^2$, where $\mathbb{E}\left[\epsilon_i^2\langle \mathbf{f}^*, \mathbf{x}_i\rangle^2|\mathbf{x}_i\right] = \langle \mathbf{f}^*, \mathbf{x}_i\rangle^2$ and $\mathsf{var}(\epsilon_i^2\langle \mathbf{f}^*, \mathbf{x}_i\rangle^2|\mathbf{x}_i) = 3\langle \mathbf{f}^*, \mathbf{x}_i\rangle^4$. Hence, $\mathbf{f}^*$ is the (expected) minimizer of the following nonconvex quartic loss function which cannot be computed from samples.

$$\mathcal{L}^{\mathsf{mul}}(\mathbf{f}) = \frac{1}{m}\sum_{i=1}^m \frac{\left[(y_i - \langle \mathbf{w}^*, \mathbf{x}_i\rangle)^2 - \langle \mathbf{f}, \mathbf{x}_i\rangle^2\right]^2}{\langle \mathbf{f}^*, \mathbf{x}_i\rangle^4}.$$

The choice of this loss is also motivated by the effectiveness of the WLS procedure for estimating $\mathbf{w}^*$. In fact, writing $\mathbf{A} = \mathbf{f}\mathbf{f}^\intercal$, $\mathcal{L}^{\mathsf{mul}}$ can be interpreted as a weighted least squares objective on the manifold of rank 1 matrices, equipped with the Hilbert Schmidt inner product:

$$\mathcal{L}^{\mathsf{mul}}(\mathbf{A}) = \frac{1}{m}\sum_{i=1}^m \frac{\left[(y_i - \langle \mathbf{w}^*, \mathbf{x}_i\rangle)^2 - \left\langle \mathbf{A}, \mathbf{x}_i\mathbf{x}_i^T\right\rangle_{\mathsf{HS}}\right]^2}{\left\langle \mathbf{A}, \mathbf{x}_i\mathbf{x}_i^T\right\rangle_{\mathsf{HS}}^2}.$$

In a practical setting, where we only have access to estimates $\hat{\mathbf{f}}$ and $\hat{\mathbf{w}}$, we consider the following computable approximation of $\mathcal{L}^{\mathsf{mul}}$:

$$\bar{\mathcal{L}}(\mathbf{f}) = \frac{1}{m}\sum_{i=1}^m \mathbb{1}(|\langle \hat{\mathbf{f}}, \mathbf{x}_i\rangle| \geq \bar{\mu})\frac{\left[(y_i - \langle \hat{\mathbf{w}}, \mathbf{x}_i\rangle)^2 - \langle \mathbf{f}, \mathbf{x}_i\rangle^2\right]^2}{\langle \hat{\mathbf{f}}, \mathbf{x}_i\rangle^4},$$

where the 'regularization' term $\mathbb{1}(|\langle \hat{\mathbf{f}}, \mathbf{x}_i\rangle| \geq \bar{\mu})$ ensures that the Hessian of this loss is well-defined. The *pseudogradient* $\mathcal{G}$, which is an approximation of $\nabla\bar{\mathcal{L}}(\mathbf{f})$, is then defined as follows:

$$\mathcal{G}(\mathbf{f}) := \frac{1}{m}\sum_{i=1}^m \mathbb{1}(|\langle \hat{\mathbf{f}}, \mathbf{x}_i\rangle| \geq \bar{\mu})\frac{\left[\langle \mathbf{f}, \mathbf{x}_i\rangle^2 - (y_i - \langle \hat{\mathbf{w}}, \mathbf{x}_i\rangle)^2\right]\langle \hat{\mathbf{f}}, \mathbf{x}_i\rangle\mathbf{x}_i}{\langle \hat{\mathbf{f}}, \mathbf{x}_i\rangle^4}$$

We show in Appendix C that $\mathcal{G}(\mathbf{f})$ behaves like the gradient of a strongly convex function whenever $\mathbf{f}$ is initialized sufficiently close to $\mathbf{f}^*$. As a consequence, we show that the pseudogradient descent procedure $\mathbf{f}_{t+1} = \mathbf{f}_t - \mathbf{D}\mathcal{G}(\mathbf{f}_t)$, initialized with $\mathbf{f}_0 = \hat{\mathbf{f}}$ under appropriately chosen matrix valued step-sizes $\mathbf{D}$ (where the matrix valued step-size 'pre-conditions' the gradient similar to the modified Newton method), converges exponentially fast to a local neighborhood of $\mathbf{f}^*$.

6.3. SymbLearn Guarantees : Theorem 7

In order to prove Theorem 7 (full proof in Appendix D), we use the fact that the OLS (Algorithm 1) estimate $\hat{\mathbf{w}}_0$ satisfies $\|\hat{\mathbf{w}}_0 - \mathbf{w}^*\|^2 = \tilde{O}(\|\mathbf{f}^*\|^2 \cdot d/n)$. The spectral method (Algorithm 2) then takes this as the input and estimates $\|\hat{\mathbf{f}}_0 - \mathbf{f}^*\|^2 = \tilde{O}(\|\mathbf{f}^*\|^2 \cdot d/n)$. Running WLS (Algorithm 3) with input $\hat{\mathbf{f}}_0$ and appropriately chosen λ_0 gives $\|\hat{\mathbf{w}}_1 - \mathbf{w}^*\|^2 = \tilde{O}\big(\frac{1}{n} + (\frac{d}{n})^{\frac{3}{2}}\big)$. The phase retrieval algorithm with input $\hat{\mathbf{w}}_1$ and $\hat{\mathbf{f}}_0$ then transfers this error bound to $\hat{\mathbf{f}}_1$, giving $\|\hat{\mathbf{f}}_1 - \mathbf{f}^*\|^2 = \tilde{O}\big(\frac{1}{n} + (\frac{d}{n})^{3/2}\big)$. This iterative procedure gives us $\|\hat{\mathbf{f}}_K - \mathbf{f}^*\|^2, \|\hat{\mathbf{w}}_K - \mathbf{w}^*\|^2 = \tilde{O}\left(\frac{1}{n} + (\frac{d}{n})^2\right)$ for $K = \lceil \log_2(n) \rceil$.

6.4. Lower Bound : Theorem 8

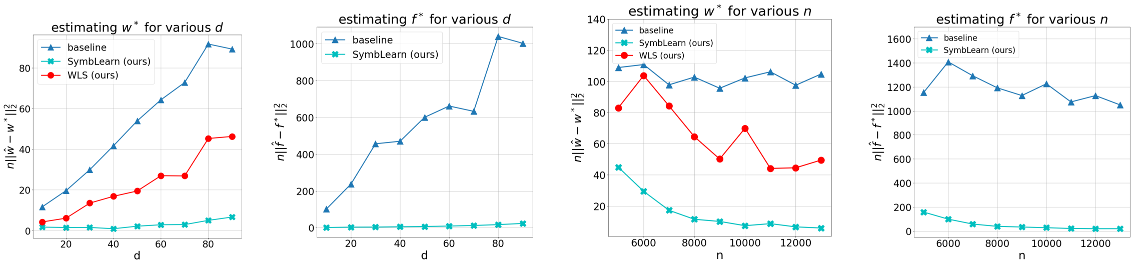
The key technical challenge in the lower bound analysis is that given any $\mathbf{f}^* \in \mathbb{R}^d$ and two instances of the heteroscedastic regression problem, $P_{\mathbf{w}_1,\mathbf{f}^*}$ and $P_{\mathbf{w}_2,\mathbf{f}^*}$, $\mathsf{KL}\left(P_{\mathbf{w}_1,\mathbf{f}^*} \middle|\middle| P_{\mathbf{w}_2,\mathbf{f}^*}\right) = \infty$ unless $\mathbf{w}_1 - \mathbf{w}_2$ is parallel to $\mathbf{f}^*$. This precludes the direct use of standard techniques such as LeCam's method or Assoud's Lemma. We first obtain a coarse lower bound by considering instances $P_{\mathbf{w}_1,\mathbf{f}^*}$ and $P_{\mathbf{w}_2,\mathbf{f}^*}$ such that $\mathbf{w}_1 - \mathbf{w}_2$ is parallel to $\mathbf{f}^*$. Here, $\mathsf{KL}\left(P_{\mathbf{w}_1,\mathbf{f}^*} \middle|\middle| P_{\mathbf{w}_2,\mathbf{f}^*}\right) = \|\mathbf{w}_1 - \mathbf{w}_2\|^2/\|\mathbf{f}^*\|^2$, and a direct application of LeCam's method obtains a lower bound of $\Omega(\|\mathbf{f}^*\|^2/n)$. To obtain the finer $\tilde{\Omega}(\|\mathbf{f}^*\|^2 d^2/n^2)$ lower bound, we develop a refined version of Assoud's lemma for heavy tailed random variables. First, we lower bound the minimax risk by the Bayes risk over the uniform spherical distribution supported over the orthogonal subspace of $\mathbf{f}^*$ as follows.

$$\sup_{\mathbf{w}\in\mathbb{R}^d} \mathbb{E}_{(\mathbf{x}_i,y_i)_{i\in[n]} \overset{\text{iid}}{\sim} P_{\mathbf{w}^*,\mathbf{f}}}\left[\|\hat{\mathbf{w}} - \mathbf{w}^*\|^2\right] \geq \mathbb{E}_{\mathbf{x}_{1:n}}\left[\mathbb{E}_{\mathbf{w}\sim S_\alpha(\mathbf{e}_{1:d-1})}\left[\mathbb{E}_{y_{1:n}|\mathbf{x}_{1:n}}\left[\|\hat{\mathbf{w}} - \mathbf{w}^*\|^2\right]\right]\right]$$

Here, $S_\alpha(\mathbf{e}_{1:d-1})$ represents the uniform spherical distribution in the subspace orthogonal to $\mathbf{f}^*$. We now exploit the rotational symmetry of this distribution to perform a *data-dependent change of basis*, i.e., we choose basis vectors $\mathbf{u}_1, \ldots, \mathbf{u}_d$ such that $\mathbf{u}_d = \mathbf{f}^*/\|\mathbf{f}^*\|$ and $\mathbf{u}_{1:d-1}$ are functions of $\mathbf{x}_{1:n}$ to be chosen later. As a consequence of the rotational symmetry, $S_\alpha(\mathbf{e}_{1:d-1})$ is equal in distribution to $S_\alpha(\mathbf{u}_{1:d-1})$. Combining this data-dependent change of basis with a careful coupling argument allows us to (approximately) lower bound the Bayes risk by a fixed-design hypothesis testing problem between $\mathbf{w}^*$ and $\mathbf{w}^* + \alpha\mathbf{u}_i$, defined as follows.

$$\mathbb{E}_{\mathbf{w}\sim S_\alpha(\mathbf{e}_{1:d-1})}\left[\mathbb{E}_{y_{1:n}|\mathbf{x}_{1:n}}\left[\|\hat{\mathbf{w}} - \mathbf{w}^*\|^2\right]\right] \geq \frac{\alpha^2}{d}\sum_{i=1}^{d-1}\mathbb{E}_{\mathbf{w}^*\sim S_\alpha(\mathbf{u}_{1:d-1})}\left[1 - \sqrt{\mathsf{KL}\left(P_{\mathbf{w}^*+\alpha\mathbf{u}_i,\mathbf{f}^*} \middle|\middle| P_{\mathbf{w}^*,\mathbf{f}^*}|\mathbf{x}_{1:n}\right)}\right] \tag{3}$$

We choose the vectors $\mathbf{u}_{1:d-1}$ such that $\sum_{i=1}^d \sqrt{\mathsf{KL}\left(P_{\mathbf{w}^*+\alpha\mathbf{u}_i,\mathbf{f}^*} \middle|\middle| P_{\mathbf{w}^*,\mathbf{f}^*}|\mathbf{x}_1,\ldots,\mathbf{x}_n\right)}$ is as small as possible. The key challenge here lies in the fact that $\mathsf{KL}\left(P_{\mathbf{w}^*+\alpha\mathbf{u}_i,\mathbf{f}^*} \middle|\middle| P_{\mathbf{w}^*,\mathbf{f}^*}|\mathbf{x}_1,\ldots,\mathbf{x}_n\right)$ has infinite expectation for fixed $\mathbf{u}_i$. Thus, when $\mathbf{u}_i$ are fixed basis vectors, we obtain a vacuous lower bound since all the KL terms can become moderately large. We resolve this with a careful bucketing argument to choose $\mathbf{u}_i$ such that in the sum $\sum_{i=1}^d \sqrt{\mathsf{KL}\left(P_{\mathbf{w}^*+\alpha\mathbf{u}_i,\mathbf{f}^*} \middle|\middle| P_{\mathbf{w}^*,\mathbf{f}^*}|\mathbf{x}_1,\ldots,\mathbf{x}_n\right)}$, the KL terms corresponding to small i are very large and those corresponding i close to d are small. The



(a) (b) (c) (d)

Figure 2: Estimation of $\mathbf{w}^*$ and $\mathbf{f}^*$ for various d with $n = 10000$ (Figs. 2(a), 2(b)) and for various n with $d = 100$ (Figs. 2(c), 2(d)). Baseline refers to OLS (Algorithm 1) for $\mathbf{w}^*$ estimation and to the spectral method (Algorithm 2) for $\mathbf{f}^*$ estimation. WLS is the weighted least squares (Algorithm 3) using spectral estimator of $\mathbf{f}^*$ (Algorithm 2). SymbLearn (Algorithm 5) significantly outperforms the baselines for both $\mathbf{w}^*$ and $\mathbf{f}^*$ estimation.

properties of the square root ensures that this gives a better upper bound. This is achieved by partitioning the data into buckets B_k defined as,

$$B_k = \{j \in [n] : \langle \mathbf{f}^*, \mathbf{x}_j \rangle^2 \in [2^k\gamma, 2^{k+1}\gamma)\}, k \leq K$$

where $\gamma = c/n^2$ and $K = O(\log(nd))$. The design matrix is accordingly split as $\mathbf{H}_k = \sum_{j \in B_k} \mathbf{x}_j \mathbf{x}_j^\mathsf{T}$. The basis vectors $\mathbf{u}_{1:d}$ are then chosen via Gram-Schmidt orthogonalization on the eigenvectors of $\mathbf{H}_k$ with non-zero eigenvalues starting from $k = 0$ until we obtain $d-1$ such vectors. Under this choice of $\mathbf{u}_{1:d}$, each term of the summation in the hypothesis testing problem (3) can be lower bounded as $1 - \frac{\alpha n}{d}\text{polylog}(nd)$. Plugging this into (3) and choosing α appropriately gives the desired $\tilde{\Omega}\left(\|\mathbf{f}^*\|^2 \frac{d^2}{n^2}\right)$ lower bound. The detailed proof is presented in Appendix G.

7. Experimental Results

We generate data according to the model (1) and estimate $\mathbf{w}^*$ and $\mathbf{f}^*$ with various algorithms and compare it to SymbLearn. Figure 7 shows the cumulative errors $n\|\hat{\mathbf{w}} - \mathbf{w}^*\|_2^2$ and $n\|\hat{\mathbf{f}} - \mathbf{f}^*\|_2^2$, averaged over 5 trials for various algorithms. Since we can only recover $\mathbf{f}^*$ up to a sign, we consider the sign that is closest to the estimate $\hat{\mathbf{f}}$. For SymbLearn (Algorithm 5), we run multiple epochs of the weighted least squares and phase retrieval on the *entire* data to get a final estimate of $\mathbf{w}^*$ and $\mathbf{f}^*$. We observe that SymbLearn not only outperforms the baseline OLS estimator, but is also superior to the WLS estimator (computed using the spectral method) discussed in Section 3.1.

8. Conclusion and Discussion

In this work, we considered heteroscedastic linear regression with Gaussian design and obtained near sample optimal and computationally efficient algorithms. In future work, we plan to explore noise of the form $\epsilon_i \sigma(\langle \mathbf{f}^*, \mathbf{x}_i \rangle)$ for more general functions σ. Developing kernel regression versions of this problem is also an interesting direction. We also hope to study the applicability of our methods in heteroscedastic versions of more challenging problems such as linear contextual bandits and reinforcement learning with linear function approximation.

References

Oren Anava and Shie Mannor. Heteroscedastic sequences: beyond gaussianity. In *International Conference on Machine Learning*, pages 755–763. PMLR, 2016.

Tim Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of econometrics*, 31(3):307–327, 1986.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

Matthew Brennan, Guy Bresler, and Dheeraj Nagaraj. Phase transitions for detecting latent geometry in random graphs. *Probability Theory and Related Fields*, 178(3-4):1215–1289, 2020.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

T Tony Cai, Xiaodong Li, and Zongming Ma. Optimal rates of convergence for noisy sparse phase retrieval via thresholded wirtinger flow. 2016.

Emmanuel J Candes, Yonina C Eldar, Thomas Strohmer, and Vladislav Voroninski. Phase retrieval via matrix completion. *SIAM review*, 57(2):225–251, 2015.

Raymond J Carroll and David Ruppert. *Transformation and weighting in regression*. Chapman and Hall/CRC, 2017.

Kamalika Chaudhuri, Prateek Jain, and Nagarajan Natarajan. Active heteroscedastic regression. In *International Conference on Machine Learning*, pages 694–702. PMLR, 2017.

Yuxin Chen, Yuejie Chi, Jianqing Fan, and Cong Ma. Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval. *Mathematical Programming*, 176:5–37, 2019.

Yuxin Chen, Yuejie Chi, Jianqing Fan, Cong Ma, et al. Spectral methods for data science: A statistical perspective. *Foundations and Trends® in Machine Learning*, 14(5):566–806, 2021.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters, 2018a.

Victor Chernozhukov, Whitney K Newey, and James Robins. Double/de-biased machine learning using regularized riesz representers. Technical report, cemmap working paper, 2018b.

Mark Collier, Rodolphe Jenatton, Basil Mustafa, Neil Houlsby, Jesse Berent, and Effrosyni Kokiopoulou. Massively scaling heteroscedastic classifiers. In *The Eleventh International Conference on Learning Representations*, 2022.

Marie Davidian and Raymond J Carroll. Variance function estimation. *Journal of the American statistical association*, 82(400):1079–1091, 1987.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019.

Robert F Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica: Journal of the econometric society*, pages 987–1007, 1982.

Nitai Fingerhut, Matteo Sesia, and Yaniv Romano. Coordinated double machine learning. In *International Conference on Machine Learning*, pages 6499–6513. PMLR, 2022.

Dylan J Foster and Vasilis Syrgkanis. Orthogonal statistical learning. *Annals of Statistics*, 2023.

Stephen M Goldfeld and Richard E Quandat. Nonlinear methods in econometrics. 1972.

Andrew C Harvey. Estimating regression models with multiplicative heteroscedasticity. *Econometrica: Journal of the Econometric Society*, pages 461–465, 1976.

JD Jobson and WA Fuller. Least squares estimation when the covariance matrix and parameter vector are functionally related. *Journal of the American Statistical Association*, 75(369):176–181, 1980.

Lang Liu, Carlos Cinelli, and Zaid Harchaoui. Orthogonal statistical learning with self-concordant loss. In *Conference on Learning Theory*, pages 5253–5277. PMLR, 2022.

Lester Mackey, Vasilis Syrgkanis, and Ilias Zadik. Orthogonal machine learning: Power and limitations. In *International Conference on Machine Learning*, pages 3375–3383. PMLR, 2018.

Mitchell Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of english: The penn treebank. 1993.

Bhaskar Mukhoty, Govind Gopakumar, Prateek Jain, and Purushottam Kar. Globally-convergent iteratively reweighted least squares for robust regression problems. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 313–322. PMLR, 16–18 Apr 2019. URL https://proceedings.mlr.press/v89/mukhoty19a.html.

Daniel B Nelson. Conditional heteroskedasticity in asset returns: A new approach. *Econometrica: Journal of the econometric society*, pages 347–370, 1991.

Praneeth Netrapalli, Prateek Jain, and Sujay Sanghavi. Phase retrieval using alternating minimization. *Advances in Neural Information Processing Systems*, 26, 2013.

Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1944–1952, 2017.

Yoav Shechtman, Yonina C Eldar, Oren Cohen, Henry Nicholas Chapman, Jianwei Miao, and Mordechai Segev. Phase retrieval with application to optical imaging: a contemporary overview. *IEEE signal processing magazine*, 32(3):87–109, 2015.

Alexandre B. Tsybakov. *Lower bounds on the minimax risk*, pages 77–135. Springer New York, New York, NY, 2009. ISBN 978-0-387-79052-7. doi: 10.1007/978-0-387-79052-7_2. URL https://doi.org/10.1007/978-0-387-79052-7_2.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Peiliang Xu. Improving the weighted least squares estimation of parameters in errors-in-variables models. *Journal of the Franklin Institute*, 356(15):8785–8802, 2019.

Peiliang Xu and Seiichi Shimada. Least squares parameter estimation in multiplicative noise models. *Communications in Statistics-Simulation and Computation*, 29(1):83–96, 2000.

Appendix A. Analysis of MLE, WLS and Self-SymbLearn

In this section, we present our analysis of MLE, WLS and Self-SymbLearn.

A.1. Technical Lemmas

Lemma 9 (Gaussian Linear Combinations of Random Vectors) *Let* $\mathbf{v} = \sum_{i=1}^{n} \epsilon_i \mathbf{y}_i$ *where* $\epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ *and* $\mathbf{y}_i$ *are arbitrary random vectors independent of* ϵ_i*. Then, for any* $\delta \in (0,1)$*, the following holds with probability at least* $1-\delta$*.*

$$\|\mathbf{v}\|^2 \leq \mathrm{Tr}(\mathbf{Y}) \left[1 + \max\left\{8\log(d/\delta), \sqrt{8\log(d/\delta)}\right\}\right]$$

where $\mathbf{Y} = \sum_{i=1}^{n} \mathbf{y}_i \mathbf{y}_i^T$*.*

Proof Let $\mathbf{e}_j, \ j \in [d]$ denote the standard basis of $\mathbb{R}^d$. Since $\epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$, it follows that,

$$\begin{aligned}\mathbb{E}[e^{\mu\langle \mathbf{v}, \mathbf{e}_j\rangle} | \mathbf{y}_{1:n}] &= \prod_{i=1}^{n} \mathbb{E}\left[e^{\mu \epsilon_i \langle \mathbf{y}_i, \mathbf{e}_j\rangle} | \mathbf{y}_{1:n}\right] = \prod_{i=1}^{n} e^{\mu^2 \langle \mathbf{y}_i, \mathbf{e}_j\rangle^2/2} \\ &= \exp\left(\frac{\mu^2 \mathbf{e}_j^T \left(\sum_{i=1}^{n} \mathbf{y}_i \mathbf{y}_i^T\right) \mathbf{e}_j}{2}\right) = e^{\mu^2 \mathbf{e}_j^T \mathbf{Y} \mathbf{e}_j/2}\end{aligned}$$

Thus, we infer that $\langle \mathbf{v}, \mathbf{e}_j\rangle$ is a zero-mean Gaussian random variable conditioned on $\mathbf{y}_{1:n}$. Thus, by χ^2 concentration, the following holds with probability at least $1 - \delta/d$

$$\langle \mathbf{v}, \mathbf{e}_j\rangle^2 \leq (\mathbf{e}_j^T \mathbf{Y} \mathbf{e}_j) \left[1 + \max\left\{8\log(d/\delta), \sqrt{8\log(d/\delta)}\right\}\right]$$

Taking a union bound, we conclude that

$$\begin{aligned}\|\mathbf{v}\|^2 = \sum_{j=1}^{d} \langle \mathbf{v}, \mathbf{e}_j\rangle^2 &\leq (\sum_{j=1}^{d} \mathbf{e}_j^T \mathbf{Y} \mathbf{e}_j) \left[1 + \max\left\{8\log(d/\delta), \sqrt{8\log(d/\delta)}\right\}\right] \\ &\leq \mathrm{Tr}(\mathbf{Y}) \left[1 + \max\left\{8\log(d/\delta), \sqrt{8\log(d/\delta)}\right\}\right]\end{aligned}$$

■

Lemma 10 (Lower Tail Bound for Binomial Random Variables) *Let* X *be a Binomial*(n,p) *random variable. Then,*

$$\mathbb{P}\left\{X \geq np/2\right\} \geq 1 - e^{-np/8}$$

Proof Decomposing $X = \sum_{i=1}^{n} Y_i$ where $Y_i \overset{\text{iid}}{\sim}$ Bernoulli(p) and observing that Y_i are non-negative random variables, the result follows by applying the one-sided Bernstein inequality. ■

Lemma 11 (Tail Bound for Rational Functions of χ^2 RVs - I) *Let $\zeta_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ and $\sigma > 0$. Then, for any $\lambda > 0$,*

$$\mathbb{P}\left\{\sum_{i=1}^{n} \frac{\sigma^2\zeta_i^2}{\lambda+\sigma^2\zeta_i^2} \geq \frac{n\sigma^2 p_0}{2(\sigma^2+\lambda)}\right\} \geq 1 - e^{-np_0/8}$$

where $p_0 = 1 - \sqrt{2/\pi}\int_0^1 e^{-x^2/2}dx = erfc(1/\sqrt{2})$

Proof Let $\lambda = \beta\sigma^2$. Then,

$$\frac{\sigma^2\zeta_i^2}{\lambda+\sigma^2\zeta_i^2} \geq \frac{1}{1+\beta}\mathbb{I}\left\{\frac{\zeta_i^2}{\beta+\zeta_i^2} \geq \frac{1}{1+\beta}\right\} = \frac{1}{1+\beta}\mathbb{I}\left\{\zeta_i^2 \geq 1\right\},$$

where the last equality follows from the fact that $t \to t^2/t^2+\beta$ is monotonic for $t \geq 0$. Hence,

$$\sum_{i=1}^{n} \frac{\zeta_i^2}{\beta+\zeta_i^2} \geq \frac{1}{1+\beta}\sum_{i=1}^{n}\mathbb{I}\left\{\zeta_i^2 \geq 1\right\} = \frac{1}{1+\beta}\text{Bin}(n, p_0),$$

where $p_0 = \mathbb{P}\left\{\zeta_i^2 \geq 1\right\} = \text{erfc}(1/\sqrt{2})$. Hence, by Lemma 10, it follows that,

$$\mathbb{P}\left\{\sum_{i=1}^{n} \frac{\zeta_i^2}{\lambda+\zeta_i^2} \geq \frac{n\sigma^2 p_0}{2(\sigma^2+\lambda)}\right\} \geq 1 - e^{-np_0/8}$$

■

Lemma 12 (Tail Bound for Rational Functions of χ^2 RVs - II) *Let $\gamma_i, \zeta_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ and let $\sigma > 0$. Then, for any $\lambda \leq \sigma^2$*

$$\mathbb{P}\left\{\sum_{i=1}^{n} \frac{\zeta_i^2}{\lambda+\sigma^2\gamma_i^2} \geq \frac{np_0}{16\sigma\sqrt{\lambda}}\right\} \geq 1 - \exp\left(-\frac{np_0\sqrt{\lambda}}{32\sigma}\right)$$

Proof Let $\lambda = \beta\sigma^2$. Using the fact that γ_i and ζ_i are independent.

$$\frac{\zeta_i^2}{\lambda+\sigma^2\gamma_i^2} \geq \frac{1}{2\lambda}\mathbb{I}\left\{\zeta_i^2 \geq 1\right\}\mathbb{I}\left\{\frac{1}{\beta+\gamma_i^2} \geq \frac{1}{2\beta}\right\} = \frac{1}{2\lambda}\mathbb{I}\left\{\zeta_i^2 \geq 1\right\}\mathbb{I}\left\{\gamma_i^2 \leq \beta\right\},$$
$$\sum_{i=1}^{n} \frac{\zeta_i^2}{\lambda+\sigma^2\gamma_i^2} \geq \frac{1}{2\lambda}\sum_{i=1}^{n}\mathbb{I}\left\{\zeta_i^2 \geq 1\right\}\mathbb{I}\left\{\gamma_i^2 \leq \beta\right\} = \frac{1}{2\lambda}\text{Bin}(n, p_0p_1),$$

where,

$$p_0 = \mathbb{P}\left\{\zeta_i^2 \geq 1\right\} = \text{erfc}(1/\sqrt{2}),$$
$$p_1 = \mathbb{P}\left\{\gamma_i^2 \leq \beta\right\} = \sqrt{\frac{2}{\pi}}\int_0^{\sqrt{\beta}} e^{-x^2/2}dx \geq \sqrt{\frac{2\beta}{\pi}}e^{-\sqrt{\beta}/2} \geq \frac{\sqrt{\beta}}{4} \geq \frac{\sqrt{\lambda}}{4\sigma}$$

where the last inequality follows since $\beta = {}^{\lambda}\!/_{\sigma^2} \leq 1$. Thus, by Lemma 10,

$$\mathbb{P}\left\{\sum_{i=1}^{n} \frac{\zeta_i^2}{\lambda + \sigma^2\gamma_i^2} \geq \frac{np_0p_1}{4\lambda}\right\} \geq 1 - e^{-np_0p_1/8}$$

Since $p_1 \geq \frac{\sqrt{\lambda}}{4\sigma}$, we conclude that,

$$\mathbb{P}\left\{\sum_{i=1}^{n} \frac{\zeta_i^2}{\lambda + \sigma^2\gamma_i^2} \geq \frac{np_0}{16\sigma\sqrt{\lambda}}\right\} \geq 1 - \exp\left(-\frac{np_0\sqrt{\lambda}}{32\sigma}\right)$$

■

Lemma 13 (Tail Bound for Rational Functions of Gaussian RVs) *Let $\gamma_i, \zeta_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$. Then, for any $\sigma > 0, \lambda > 0$ and $A \geq 0$,*

$$\mathbb{P}\left\{\sum_{i=1}^{n} \frac{\sigma\gamma_i\zeta_i}{\lambda + \sigma^2\gamma_i^2} \geq -\frac{n\sqrt{2A}}{\lambda^{1/4}}\right\} \geq 1 - e^{-An\sqrt{\lambda}}$$

Proof We show that $\sum_{i=1}^{n} \frac{\gamma_i\zeta_i}{\lambda+\gamma_i^2}$ is a subgaussian random variable. Since $\gamma_i, \zeta_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$, it follows that $\mathbb{E}\left[\sum_{i=1}^{n} \frac{\sigma\gamma_i\zeta_i}{\lambda+\sigma^2\gamma_i^2}\right] = 0$. Moreover, for any $\mu \in \mathbb{R}$,

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\mu\sum_{i=1}^{n} \frac{\sigma\gamma_i\zeta_i}{\lambda + \sigma^2\gamma_i^2}\right)\right] &= \prod_{i=1}^{n} \mathbb{E}\left[\exp\left(\mu\frac{\sigma\gamma_i\zeta_i}{\lambda + \sigma^2\gamma_i^2}\right)\right] \\
&= \prod_{i=1}^{n} \mathbb{E}\left[\mathbb{E}\left[\exp\left(\mu\frac{\sigma\gamma_i\zeta_i}{\lambda + \sigma^2\gamma_i^2}\right) \mid \gamma_i\right]\right] \\
&= \prod_{i=1}^{n} \mathbb{E}\left[\exp\left(\frac{\mu^2\sigma^2\gamma_i^2}{2(\lambda + \sigma^2\gamma_i^2)^2}\right)\right] \\
&\leq \prod_{i=1}^{n} \mathbb{E}\left[\exp\left(\frac{\mu^2}{2\lambda}\frac{\sigma^2\gamma_i^2}{\lambda + \sigma^2\gamma_i^2}\right)\right] \leq \prod_{i=1}^{n} \exp\left(\frac{\mu^2}{2\lambda}\right) = \exp\left(\frac{\mu^2}{2}\frac{n}{\lambda}\right)
\end{aligned}$$

Hence, for any $t \geq 0$,

$$\mathbb{P}\left\{\sum_{i=1}^{n} \frac{\sigma\gamma_i\zeta_i}{\lambda + \sigma^2\gamma_i^2} \geq -t\right\} \geq 1 - e^{-\lambda t^2/2n}$$

Setting $\lambda t^2/2n = An\sqrt{\lambda}$, we get,

$$\mathbb{P}\left\{\sum_{i=1}^{n} \frac{\sigma\gamma_i\zeta_i}{\lambda + \sigma^2\gamma_i^2} \geq -\frac{n\sqrt{2A}}{\lambda^{1/4}}\right\} \geq 1 - e^{-An\sqrt{\lambda}}$$

■

Lemma 14 *Let* $\mathbf{f}$ *be any arbitrary random vector and let* $\mathbf{x}_1, \dots, \mathbf{x}_n$ *be i.i.d samples from* $\mathcal{N}(0, \mathbf{I})$ *that are independent of* $\mathbf{f}$*. Then, for any* $\delta \in (0, 1/2)$*,* $n \geq d\mathsf{polylog}(d/\delta)$ *and any* $\|\mathbf{f}\|^2 \geq \lambda \geq \Omega\left(\|\mathbf{f}\|^2 \frac{d^2}{n^2} \log(nd/\delta)^2\right)$*, the following holds with probability at least* $1 - \delta$*,*

$$\mathrm{Tr}\left(\left[\sum_{i=1}^{n} \frac{\mathbf{x}_i \mathbf{x}_i^T}{\langle \mathbf{f}, \mathbf{x}_i \rangle^2 + \lambda}\right]^{-1}\right) \leq O\left(\|\mathbf{f}\|^2/n + d\|\mathbf{f}\|\sqrt{\lambda}/n\right)$$

Proof Let $\mathbf{X}_\lambda = \sum_{i=1}^{n} \frac{\mathbf{x}_i \mathbf{x}_i^T}{\langle \mathbf{f}, \mathbf{x}_i \rangle^2 + \lambda}$. Furthermore, define $\mathbf{e} = \mathbf{f}/\|\mathbf{f}\|$ and the matrix $\mathbf{B}_\lambda$ as follows,

$$\mathbf{B}_\lambda = \frac{C_1 n}{\|\mathbf{f}\|^2 + \lambda} \mathbf{e}\mathbf{e}^T + \frac{C_2 n}{\|\mathbf{f}\|\sqrt{\lambda}} \left(\mathbf{I} - \mathbf{e}\mathbf{e}^T\right)$$

where $C_1 \leq C_2$ are numerical constants to be chosen later. We note that, since $\mathbf{B}_\lambda$ is symmetric and positive definite and $\lambda \leq \|\mathbf{f}\|^2$,

$$\mathrm{Tr}\left(\mathbf{B}_\lambda^{-1}\right) \leq O\left(\frac{\|\mathbf{f}\|^2 + \lambda}{n} + \frac{(d-1)\|\mathbf{f}\|\sqrt{\lambda}}{n}\right) \leq O\left(\frac{\|\mathbf{f}\|^2}{n} + \frac{d\|\mathbf{f}\|\sqrt{\lambda}}{n}\right)$$

Hence, it suffices to prove that $\mathbf{X}_\lambda \succeq \mathbf{B}_\lambda$ holds with high probability. We establish this by means of a covering argument. To this end, consider any $\epsilon \in (0, 1)$ and let C_ϵ denote an ϵ cover of the unit ball in $\mathbb{R}^d$. Applying a standard discretization argument, we get,

$$\inf_{\|\mathbf{x}\|=1} \mathbf{x}^T \left(\mathbf{X}_\lambda - \mathbf{B}_\lambda\right) \mathbf{x} \geq \inf_{\mathbf{v} \in C_\epsilon} \mathbf{v}^T \left(\mathbf{X}_\lambda - \mathbf{B}_\lambda\right) \mathbf{v} - 2\epsilon\|\mathbf{X}_\lambda\| - 2\epsilon\|\mathbf{B}_\lambda\|$$

Since $\lambda \leq \|\mathbf{f}\|^2$, $\lambda + \|\mathbf{f}\|^2 \geq \|\mathbf{f}\|\sqrt{\lambda}$. Hence, $\|\mathbf{B}_\lambda\| = \frac{C_2 n}{\|\mathbf{f}\|\sqrt{\lambda}}$. Moreover,

$$\|\mathbf{X}_\lambda\| \leq \sum_{i=1}^{n} \frac{\left\|\mathbf{x}_i \mathbf{x}_i^T\right\|}{\lambda + \langle \mathbf{f}, \mathbf{x}_i \rangle^2} \leq \frac{n}{\lambda}\left(\frac{1}{n}\sum_{i=1}^{n} \|\mathbf{x}_i\|^2\right)$$

Since $\mathbf{x}_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \mathbf{I})$, the following holds with probability $1 - \delta/2$ due to χ^2 concentration,

$$\|\mathbf{X}_\lambda\| \leq \frac{n}{\lambda}\left[d + \max\left\{\frac{8\log(2/\delta)}{n}, \sqrt{\frac{8d\log(2/\delta)}{n}}\right\}\right] \leq \frac{n}{\lambda}(d+1)$$

where the last inequality follows from the fact that $n \geq d\mathsf{polylog}(d/\delta)$. Hence, the following holds with probability at least $1 - \delta/2$,

$$\inf_{\|\mathbf{x}\|=1} \mathbf{x}^T \left(\mathbf{X}_\lambda - \mathbf{B}_\lambda\right) \mathbf{x} \geq \inf_{\mathbf{v} \in C_\epsilon} \mathbf{v}^T \left(\mathbf{X}_\lambda - \mathbf{B}_\lambda\right) \mathbf{v} - \frac{2\epsilon n}{\lambda}(d+1) - \frac{2\epsilon C_2 n}{\|\mathbf{f}\|\sqrt{\lambda}} \tag{4}$$

Now consider any $\mathbf{v} \in C_\epsilon$. Since $\|\mathbf{v}\| = 1$, there exists a unit vector $\mathbf{g}$ orthogonal to $\mathbf{f}$ such that $\mathbf{v} = \alpha\mathbf{e} + \beta\mathbf{g}$, with $\alpha^2 + \beta^2 = 1$. We note that,

$$\mathbf{v}^T \mathbf{B}_\lambda \mathbf{v} = \frac{\alpha^2 n C_1}{\|\mathbf{f}\|^2 + \lambda} + \frac{\beta^2 n C_2}{\|\mathbf{f}\|\sqrt{\lambda}}$$

$$\mathbf{v}^T \mathbf{X}_\lambda \mathbf{v} = \frac{\alpha^2}{\|\mathbf{f}\|^2} \sum_{i=1}^{n} \frac{\langle \mathbf{f}, \mathbf{x}_i \rangle^2}{\lambda + \langle \mathbf{f}, \mathbf{x}_i \rangle^2} + \beta^2 \sum_{i=1}^{n} \frac{\langle \mathbf{g}, \mathbf{x}_i \rangle^2}{\lambda + \langle \mathbf{f}, \mathbf{x}_i \rangle^2} + \frac{2\alpha\beta}{\|\mathbf{f}\|} \sum_{i=1}^{n} \frac{\langle \mathbf{f}, \mathbf{x}_i \rangle \langle \mathbf{g}, \mathbf{x}_i \rangle}{\lambda + \langle \mathbf{f}, \mathbf{x}_i \rangle^2}$$

Since $\mathbf{f}$ and $\mathbf{g}$ are orthogonal, $\langle \mathbf{f}, \mathbf{x}_i \rangle$ and $\langle \mathbf{g}, \mathbf{x}_i \rangle$ are independent. From Lemmas 11, 12 and 13, and the fact that $\lambda \leq \|\mathbf{f}\|^2$, we conclude that the following holds with probability at least $1 - e^{-np_0/8} - e^{-\frac{np_0\sqrt{\lambda}}{32\|\mathbf{f}\|}} - e^{-\frac{A_\alpha A_\beta n\sqrt{\lambda}}{\|\mathbf{f}\|}}$,

$$\frac{\alpha^2}{\|\mathbf{f}\|^2}\sum_{i=1}^{n}\frac{\langle \mathbf{f}, \mathbf{x}_i\rangle^2}{\lambda + \langle \mathbf{f}, \mathbf{x}_i\rangle^2} \geq \frac{\alpha^2 n p_0}{2\lambda + 2\|\mathbf{f}\|^2}$$

$$\beta^2\sum_{i=1}^{n}\frac{\langle \mathbf{g}, \mathbf{x}_i\rangle^2}{\lambda + \langle \mathbf{f}, \mathbf{x}_i\rangle^2} \geq \frac{\beta^2 n p_0\sqrt{\lambda}}{16\|\mathbf{f}\|}$$

$$\frac{2\alpha\beta}{\|\mathbf{f}\|}\sum_{i=1}^{n}\frac{\langle \mathbf{f}, \mathbf{x}_i\rangle\langle \mathbf{g}, \mathbf{x}_i\rangle}{\lambda + \langle \mathbf{f}, \mathbf{x}_i\rangle^2} \geq -\frac{2\alpha\beta}{\|\mathbf{f}\|}\frac{n\sqrt{2A_\alpha A_\beta}}{\|\mathbf{f}\|^{1/2}\lambda^{1/4}} \geq -\frac{2\sqrt{2}\alpha^2 n A_\alpha}{\lambda + \|\mathbf{f}\|^2} - \frac{\sqrt{2}\beta^2 n A_\beta}{\|\mathbf{f}\|\sqrt{\lambda}}$$

Setting $C_1 = C_2 = p_0/64$ and $A_\alpha = A_\beta = p_0/64\sqrt{2}$, we conclude that the following holds with probability at least $1 - 3e^{-np_0^2\sqrt{\lambda}/512\|\mathbf{f}\|}$

$$\begin{aligned}\mathbf{v}^T\left(\mathbf{X}_\lambda - \mathbf{B}_\lambda\right)\mathbf{v} &\geq \frac{n\alpha^2}{\lambda + \|\mathbf{f}\|^2}\left(p_0/2 - C_1 - 2\sqrt{2}A_\alpha\right) + \frac{n\beta^2}{\|\mathbf{f}\|\sqrt{\lambda}}\left(p_0/16 - C_2 - \sqrt{2}A_\beta\right)\\ &\geq \frac{np_0/64}{\lambda + \|\mathbf{f}\|^2}\end{aligned}$$

The second inequality follows since $\alpha^2 + \beta^2 = 1$ and $\lambda \leq \|\mathbf{f}\|^2$ implies $\|\mathbf{f}\|\sqrt{\lambda} \leq \|\mathbf{f}\|^2 \leq \lambda + \|\mathbf{f}\|^2$. Taking a union bound over C_ϵ, using the fact that $|C_\epsilon| \leq (3/\epsilon)^d$, we conclude that the following holds with probability at least $1 - \exp\left(\ln 3 + d\ln(3/\epsilon) - np_0^2\sqrt{\lambda}/512\|\mathbf{f}\|\right)$,

$$\inf_{\mathbf{v}\in C_\epsilon}\mathbf{v}^T\left(\mathbf{X}_\lambda - \mathbf{B}_\lambda\right)\mathbf{v} \geq \frac{np_0/64}{\lambda + \|\mathbf{f}\|^2} \tag{5}$$

To ensure that the above event holds with probability at least $1 - \delta/2$, we require λ to be lower bounded as,

$$\frac{\sqrt{\lambda}}{\|\mathbf{f}\|} \geq \frac{512}{np_0^2}\left(\ln(6) + d\ln(3/\epsilon) + \ln(2/\delta)\right) \tag{6}$$

Suppose λ and ϵ appropriately chosen (to be specified later) such that equation (6) is satisfied. Then, from equations (4) and (5), we conclude that the following holds with probability at least $1 - \delta$,

$$\begin{aligned}\inf_{\|\mathbf{x}\|=1}\mathbf{x}^T\left(\mathbf{X}_\lambda - \mathbf{B}_\lambda\right)\mathbf{x} &\geq \frac{np_0/64}{\lambda + \|\mathbf{f}\|^2} - \frac{2\epsilon n(d+1)}{\lambda} - \frac{\epsilon n p_0}{32\|\mathbf{f}\|\sqrt{\lambda}}\\ &\geq \frac{np_0/64}{\lambda + \|\mathbf{f}\|^2} - \frac{\epsilon n}{\lambda}\left(2d + 2 + p_0/32\right)\end{aligned}$$

To ensure that the RHS is non-negative, ϵ must satisfy the following,

$$1/\epsilon \geq \frac{64}{p_0}\left(1 + \|\mathbf{f}\|^2/\lambda\right)\left(2d + 2 + p_0/32\right) \tag{7}$$

Without loss of generality, assume $\epsilon \leq e^{-1}$. Then, we note that equation (6) is satisfied if,

$$\frac{\sqrt{\lambda}}{\|\mathbf{f}\|} \geq \frac{\tau_1 d}{n} \left[\ln(d/\delta) + \ln(2/\epsilon)\right] \tag{8}$$

where $\tau_1 > 10^4$ is a universal constant. It follows that, $\|\mathbf{f}\|^2/\lambda \leq n^2/d^2\tau_1^2$. We observe that, for this choice of λ, equation (7) is satisfied if $1/\epsilon = \tau_2 d^2 n^2$ for some absolute constant $\tau_2 \geq 10^5$. Substituting this choice of $1/\epsilon$ into equation (8), we note that equation (8) is satisfied if,

$$\frac{\sqrt{\lambda}}{\|\mathbf{f}\|} \geq \frac{\tau_1 d}{n} \left[\ln\left(d/\delta\right) + 2\tau_2 \ln(d) + 2\tau_2 \ln(n)\right]$$

Hence, there exists a universal constant τ_3 such that the above is satisfied when,

$$\sqrt{\lambda} \geq \tau_3 \|\mathbf{f}\| d/n \log(nd/\delta)$$

Hence, we conclude that, setting $1/\epsilon = \Theta(n^2 d^2)$ and $\|\mathbf{f}\|^2 \geq \lambda \geq \Omega(\|\mathbf{f}\|^2 \frac{d^2}{n^2} \log(nd/\delta)^2)$ is sufficient to ensure that $\inf_{\|\mathbf{x}\|=1} \mathbf{x}^T \left(\mathbf{X}_\lambda - \mathbf{B}_\lambda\right) \mathbf{x} \geq 0$ with probability at least $1-\delta$, i.e., $\mathbf{X}_\lambda \succeq \mathbf{B}_\lambda$ with probability at least $1-\delta$. Furthermore, the condition $n \geq d\text{polylog}(d/\delta)$ ensures that $\lambda \leq \|\mathbf{f}\|^2$ and $\lambda \geq \Omega(\|\mathbf{f}\|^2 \frac{d^2}{n^2} \log(nd/\delta)^2)$ can be simultaneously satisfied. ■

A.2. Proof of Theorem 5

Proof We recall that, given any relaxation parameter λ and approximate noise model $\hat{\mathbf{f}}$, the WLS estimator $\hat{\mathbf{w}}_{\hat{\mathbf{f}},\lambda}$ is defined as follows.

$$\hat{\mathbf{w}}_{\hat{\mathbf{f}},\lambda} = \left[\sum_{i=1}^{n} \frac{\mathbf{x}_i \mathbf{x}_i^T}{\left\langle \hat{\mathbf{f}}, \mathbf{x}_i \right\rangle^2 + \lambda}\right]^{-1} \left[\sum_{i=1}^{n} \frac{\mathbf{x}_i y_i}{\left\langle \hat{\mathbf{f}}, \mathbf{x}_i \right\rangle^2 + \lambda}\right]$$

We use $\mathbf{X}_{\lambda,\hat{\mathbf{f}}}$ to denote the design matrix of WLS, which is defined as,

$$\mathbf{X}_{\lambda,\hat{\mathbf{f}}} = \left[\sum_{i=1}^{n} \frac{\mathbf{x}_i \mathbf{x}_i^T}{\left\langle \hat{\mathbf{f}}, \mathbf{x}_i \right\rangle^2 + \lambda}\right]$$

Using $\mathbf{y}_i = \langle \mathbf{w}^*, \mathbf{x}_i \rangle + \langle \mathbf{f}^*, \mathbf{x}_i \rangle \epsilon_i$, we observe that,

$$\hat{\mathbf{w}}_{\hat{\mathbf{f}},\lambda} - \mathbf{w}^* = \mathbf{X}_{\lambda,\hat{\mathbf{f}}}^{-1} \left[\sum_{i=1}^{n} \frac{\epsilon_i \langle \mathbf{f}^*, \mathbf{x}_i \rangle \mathbf{x}_i}{\left\langle \hat{\mathbf{f}}, \mathbf{x}_i \right\rangle^2 + \lambda}\right] = \sum_{i=1}^{n} \epsilon_i \mathbf{v}_i,$$

where $\mathbf{v}_i$ is defined as,

$$\mathbf{v}_i = \left[\frac{\langle \mathbf{f}^*, \mathbf{x}_i \rangle}{\left\langle \hat{\mathbf{f}}, \mathbf{x}_i \right\rangle^2 + \lambda}\right] \mathbf{X}_{\lambda,\hat{\mathbf{f}}}^{-1} \mathbf{x}_i$$

We now define the matrix $\mathbf{M}$ as follows

$$\mathbf{M} = \sum_{i=1}^{n} \mathbf{v}_i \mathbf{v}_i^T = \mathbf{X}_{\lambda,\hat{\mathbf{f}}}^{-1} \left[\sum_{i=1}^{n} \frac{\langle \mathbf{f}^*, \mathbf{x}_i \rangle^2 \mathbf{x}_i \mathbf{x}_i^T}{\left[\left\langle \hat{\mathbf{f}}, \mathbf{x}_i \right\rangle^2 + \lambda \right]^2} \right] \mathbf{X}_{\lambda,\hat{\mathbf{f}}}^{-1}$$

We note that, since $\mathbf{X}_{\lambda,\hat{\mathbf{f}}}$ is a symmetric PSD matrix, so is $\mathbf{M}$. Furthermore, since ϵ_i are independent of $\mathbf{v}_i$, we conclude that the following holds with probability at least $1 - \delta/3$,

$$\left\| \hat{\mathbf{w}}_{\hat{\mathbf{f}},\lambda} - \mathbf{w}^* \right\|^2 \leq \mathrm{Tr}(\mathbf{M}) \left[1 + \max \left\{ 8 \log(3d/\delta), \sqrt{8 \log(3d/\delta)} \right\} \right] \tag{9}$$

Motivated by the fact that Lemma 14 allows us to upper bound the trace of $\mathbf{X}_{\lambda,\hat{\mathbf{f}}}^{-1}$, but directly controlling $\mathrm{Tr}(\mathbf{M})$ might be cumbersome, we now aim to establish that $\mathbf{M} \preceq 2\mathbf{X}_{\lambda,\hat{\mathbf{f}}}^{-1}$ holds with high probability, and consequently, so does $\mathrm{Tr}(\mathbf{M}) \leq 2\,\mathrm{Tr}\left(\mathbf{X}_{\lambda,\hat{\mathbf{f}}}^{-1}\right)$. To this end, we recall that $\left\| \hat{\mathbf{f}} - \mathbf{f}^* \right\|^2 \leq \epsilon$ and $\mathbf{x}_1 \ldots, \mathbf{x}_n$ are independent of $\mathbf{x}_i$. Thus, $\left(\hat{\mathbf{f}} - \mathbf{f}^* \right)^T \mathbf{x}_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \epsilon)$. Furthermore, by concentration of suprema of Gaussian random variables, the following holds with probability at least $1 - \delta/3$,

$$\left| \left(\hat{\mathbf{f}} - \mathbf{f}^* \right)^T \mathbf{x}_i \right| \leq \beta = \sqrt{2\epsilon \log\left(6n/\delta\right)} \ \ \forall\, i \in [n]$$

Conditioned on the above event occuring, the following inequalities hold uniformly for every $i \in [n]$ with probability 1.

$$\begin{aligned}
\left| \left(\hat{\mathbf{f}}^T \mathbf{x}_i \right)^2 - \left((\mathbf{f}^*)^T \mathbf{x}_i \right)^2 \right| &= \left| \left(\hat{\mathbf{f}} + \mathbf{f}^* \right)^T \mathbf{x}_i \right| \left| \left(\hat{\mathbf{f}} + \mathbf{f}^* \right)^T \mathbf{x}_i \right| \\
&\leq \beta \left| 2 \left(\mathbf{f}^{*T} \mathbf{x}_i \right) + \beta \right| \\
&\leq 2\beta \left| \left((\mathbf{f}^*)^T \mathbf{x}_i \right) \right| + \beta^2 \\
&\leq \frac{\left((\mathbf{f}^*)^T \mathbf{x}_i \right)^2}{2} + 6\epsilon \log\left(2n/\delta\right) \ \ \forall\, i \in [n]
\end{aligned}$$

Furthermore,

$$\begin{aligned}
\left((\mathbf{f}^*)^T \mathbf{x}_i \right)^2 &\leq \left((\hat{\mathbf{f}})^T \mathbf{x}_i \right)^2 + \left| \left((\hat{\mathbf{f}})^T \mathbf{x}_i \right)^2 - \left((\mathbf{f}^*)^T \mathbf{x}_i \right)^2 \right| \\
&\leq \left((\hat{\mathbf{f}})^T \mathbf{x}_i \right)^2 + \lambda + \frac{\left((\mathbf{f}^*)^T \mathbf{x}_i \right)^2}{2} + 6\epsilon \log\left(6n/\delta\right) \ \ \forall\, i \in [n]
\end{aligned}$$

Setting $\lambda \geq \max\{\epsilon, \|\hat{\mathbf{f}}\|^2 d^2/n^2\}\mathsf{polylog}(nd/\delta) \geq 6\epsilon \log\left(6n/\delta\right)$, we conclude that the following holds with probability at least $1 - \delta/3$

$$\frac{\left((\mathbf{f}^*)^T \mathbf{x}_i \right)^2}{\left((\hat{\mathbf{f}})^T \mathbf{x}_i \right)^2 + \lambda} \leq 2 \ \ \forall\, i \in [n] \tag{10}$$

Since $\mathbf{X}_{\lambda,\hat{\mathbf{f}}}$ and $\mathbf{M}$ are both PSD matrices, we conclude that the following holds with probability at least $1 - \delta/3$,

$$\begin{aligned}
\mathbf{M} &= \mathbf{X}_{\lambda,\hat{\mathbf{f}}}^{-1}\left[\sum_{i=1}^{n}\frac{\langle \mathbf{f}^*, \mathbf{x}_i\rangle^2 \mathbf{x}_i\mathbf{x}_i^T}{\left[\left\langle \hat{\mathbf{f}}, \mathbf{x}_i\right\rangle^2 + \lambda\right]^2}\right]\mathbf{X}_{\lambda,\hat{\mathbf{f}}}^{-1} \\
&\preceq 2\mathbf{X}_{\lambda,\hat{\mathbf{f}}}^{-1}\left[\sum_{i=1}^{n}\frac{\mathbf{x}_i\mathbf{x}_i^T}{\left\langle \hat{\mathbf{f}}, \mathbf{x}_i\right\rangle^2 + \lambda}\right]\mathbf{X}_{\lambda,\hat{\mathbf{f}}}^{-1} \\
&\preceq 2\mathbf{X}_{\lambda,\hat{\mathbf{f}}}^{-1} \qquad (11)
\end{aligned}$$

where the second PSD inequality follows from (10). Furthermore, applying Lemma 14 to $\mathbf{X}_{\lambda,\hat{\mathbf{f}}}$, we ensure that the following holds with probability $1 - \delta/3$

$$\mathrm{Tr}\left(\mathbf{X}_{\lambda,\hat{\mathbf{f}}}^{-1}\right) \leq \left(\|\hat{\mathbf{f}}\|^2/n + d\|\hat{\mathbf{f}}\|\sqrt{\lambda}/n\right)\mathsf{polylog}(nd/\delta) \qquad (12)$$

Finally, from (9), (11) and (12), we conclude that the following guarantee holds with probability at least $1 - \delta$,

$$\begin{aligned}
\left\|\hat{\mathbf{w}}_{\hat{\mathbf{f}},\lambda} - \mathbf{w}^*\right\|^2 &\leq \left(\|\hat{\mathbf{f}}\|^2/n + d\|\hat{\mathbf{f}}\|\sqrt{\lambda}/n\right)\mathsf{polylog}(nd/\delta) \\
&\leq \left(\|\mathbf{f}^*\|^2/n + d\|\mathbf{f}^*\|\sqrt{\lambda}/n + \epsilon/n + d\sqrt{\epsilon\lambda}/n\right)\mathsf{polylog}(nd/\delta)
\end{aligned}$$

■

A.3. Algorithm and results for linear regression with multiplicative noise

In this section, we present a simpler algorithm, Self-SymbLearn, in Algorithm 6 for linear regression with multiplicative noise, that achieves an improved rate compared to OLS.

Algorithm 6 `Self-SymbLearn`

```
Require: $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$.    Steps $K$.    Weights $\lambda_1, \ldots, \lambda_K$.

Divide the data into $K+1$ partitions of size $m = \lfloor n/K \rfloor$
Let $\hat{\mathbf{w}}_0$ be the OLS estimate $\hat{\mathbf{w}}_{\mathsf{OLS}}$ computed on the first data partition
for $k \in \{1, \ldots, K\}$ do
    Compute $\hat{\mathbf{w}}_k$ to be the WLS estimator (Algorithm 3) computed on the $(k+1)^{\text{th}}$ data partition
    using regularization weight $\lambda_k$ and noise model estimate $\hat{\mathbf{w}}_{k-1}$, i.e, $\hat{\mathbf{w}}_k = \hat{\mathbf{w}}_{\hat{\mathbf{w}}_{k-1},\lambda_k}$
end for
Output $\hat{\mathbf{w}}_K$
```

Theorem 15 (Self-SymbLearn) *In the heteroscedastic regression model, we take $\mathbf{w}^* = \pm\mathbf{f}^*$. Consider any $\delta \in (0, 1/2)$ and let $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n) \in \mathbb{R}^d \times \mathbb{R}$ be i.i.d samples from the multiplicative linear regression model. Then, for $n \geq d\mathsf{polylog}(d)$ and $K = \Theta(\log(n))$, there exists an appropriate choice of weights $\lambda_1, \dots, \lambda_K$ such that the output $\hat{\mathbf{w}}_K$ of the Self-SymbLearn algorithm satisfies the following with probability at least $1 - \delta$:*

$$\|\mathbf{w}_K - \mathbf{w}^*\|^2 \leq \tilde{O}\left({\|\mathbf{w}^*\|^2}/{n} + {\|\mathbf{w}^*\|^2 d^2}/{n^2}\right)$$

Proof [Proof of Theorem 15] For ease of exposition, assume $n = mK$ where $K = \lceil \log_2(n) \rceil$ and $m \geq \tilde{\Omega}(d)$. For $k \in \{0, \dots, K\}$, let $e_k = \|\mathbf{w}_k - \mathbf{w}^*\|^2$. Since $\hat{\mathbf{w}}_0$ is the OLS estimate computed on m samples, we know that with probability at least $1 - \delta/(K+1)$, $e_0 = \tilde{O}(\|\mathbf{w}^*\|^2 d/m)$. Set $\lambda_1 = \tilde{O}(\|\hat{\mathbf{w}}_0\|^2 d/m)$. Since $\hat{\mathbf{w}}_1 = \hat{\mathbf{w}}_{\hat{\mathbf{w}}_0, \lambda_1}$, it follows from Theorem 5 and a union bound that the following holds with probability at least $1 - 2\delta/(K+1)$,

$$\begin{aligned}
e_1 = \|\hat{\mathbf{w}}_1 - \mathbf{w}^*\|^2 &\leq \tilde{O}\left(\frac{\|\hat{\mathbf{w}}_0\|^2}{m} + \frac{d\|\hat{\mathbf{w}}_0\|\sqrt{\lambda_1}}{m}\right) \\
&\leq \tilde{O}\left(\|\hat{\mathbf{w}}_0\|^2 \left(1/m + (d/m)^{1.5}\right)\right) \\
&\leq \tilde{O}\left(\|\mathbf{w}^*\|^2 \left(1 + d/m\right) \left(1/m + (d/m)^{1.5}\right)\right) \\
&\leq \tilde{O}\left(\|\mathbf{w}^*\|^2 \left(1/m + (d/m)^{1.5}\right)\right) \\
&= \left(\|\mathbf{w}^*\|^2 \left(1/m + (d/m)^{1.5}\right)\right) \mathsf{polylog}(\tfrac{nd}{\delta})
\end{aligned}$$

where the last inequality follows from the fact that $m \geq \tilde{\Omega}(d)$. We now prove the required convergence guarantee via induction. To this end, we define $S_k = \sum_{j=0}^{k} 1/2^j$ Clearly, $1 \leq S_k \leq 2$ and $S_{k+1} = 1 + S_k/2$. Let $L = \mathsf{polylog}(nd/\delta) > 1$ be a large enough polylog factor independent of k. We now define the event $E_k(L)$ as follows:

1. $\|\hat{\mathbf{w}}_l\| \leq 2\|\mathbf{w}^*\|$ for every $1 \leq l \leq k$
2. $e_l \leq \|\mathbf{w}^*\|^2 \left(\max(\frac{l}{m}, \frac{ld^2}{m^2}) + (\frac{d}{m})^{S_l}\right) L$ for every $1 \leq l \leq k$

We set, $\lambda_{k+1} = \|\hat{\mathbf{w}}_k\|^2 \left(\max(\frac{k}{m}, \frac{kd^2}{m^2}) + (\frac{d}{m})^{S_k}\right) L\mathsf{polylog}(\frac{nd}{\delta})$. Note that $\hat{\mathbf{w}}_{k+1} = \hat{\mathbf{w}}_{\hat{\mathbf{w}}_k, \lambda_{k+1}}$, we conclude from Theorem 5 that whenever polylog is large enough (independent of k, K), when conditioned on the event $E_k(L)$ the following holds with probability at least $1 - \delta/(K+1)$, with polylog factor being independent of k:

$$\begin{aligned}
e_{k+1} &\leq \left(\frac{\|\mathbf{w}^*\|^2}{m} + \frac{d\|\mathbf{w}^*\|\sqrt{\lambda_k}}{m}\right) \mathsf{polylog}(nd/\delta) \\
&\leq \left(\frac{\|\mathbf{w}^*\|^2}{m} + \|\mathbf{w}^*\|^2 \max\left(\frac{d\sqrt{k}}{m^{3/2}}, \frac{d^2\sqrt{k}}{m^2}\right) + \|\mathbf{w}^*\|^2 \left(\frac{d}{m}\right)^{1+S_k/2}\right) \sqrt{L}\mathsf{polylog}(nd/\delta) \\
&\leq \left(\|\mathbf{w}^*\|^2 \max\left(\frac{k+1}{m}, \frac{d^2(k+1)}{m^2}\right) + \|\mathbf{w}^*\|^2 \left(\frac{d}{m}\right)^{1+S_k/2}\right) \sqrt{L}\mathsf{polylog}(nd/\delta)
\end{aligned}$$

In the second step, we have used the fact that $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$. In the third step, we use the fact that $\sqrt{k} \leq k$ and $\frac{d}{m^{3/2}} \leq \max(\frac{1}{m}, \frac{d^2}{m^2})$, since it is the geometric mean of $1/m$ and d^2/m^2. Picking L to be a large enough $\mathsf{polylog}(\frac{nd}{\delta})$ (independent of k), we conclude that conditioned on $E_k(L)$, we must have with probability at-least $1 - \delta/(K+1)$:

$$e_{k+1} \leq \left(\|\mathbf{w}^*\|^2 \max\left(\frac{k+1}{m}, \frac{d^2(k+1)}{m^2}\right) + \|\mathbf{w}^*\|^2 \left(\frac{d}{m}\right)^{1+S_k/2}\right) L$$

If we take $n > d\mathsf{polylog}(d/\delta)$ for large enough $\mathsf{polylog}()$ as in the statement of the theorem, the above equation when combined with the triangle inequality also implies that:

$$\|\hat{\mathbf{w}}_{k+1}\| \leq 2\|\mathbf{w}^*\| \,.$$

Therefore, we conclude: $\mathbb{P}(E_{k+1}|E_k) \geq 1 - \frac{\delta}{K+1}$. Applying a union bound on this and using the fact that $\mathbb{P}(E_1) \geq 1 - \frac{2\delta}{K+1}$, we conclude:

$$\mathbb{P}(E_{K+1}) \geq 1 - \delta$$

When $K \geq \log n$, we have $2 \geq S_{K+1} \geq 2 - 1/n$. Thus, we have under the event E_{K+1}:

$$\begin{aligned} e_{K+1} &\leq \|\mathbf{w}^*\|^2 \left(\max(\frac{1}{m}, \frac{d^2}{m^2}) + \left(\frac{d}{m}\right)^{2-\frac{1}{n}}\right) \mathsf{polylog}(\tfrac{nd}{\delta}) \\ &\leq \|\mathbf{w}^*\|^2 \left(\max(\frac{1}{m}, \frac{d^2}{m^2}) + \left(\frac{d}{m}\right)^{2}\right) \mathsf{polylog}(\tfrac{nd}{\delta}) \end{aligned}$$

Which proves the result. ■

A.4. Proof of Theorem 3

To derive the high probability upper bound, consider any $\delta \in (0, 1/2)$. We note that the OLS estimator for the heteroscedastic regression problem satisfies

$$\hat{\mathbf{w}}_{\mathsf{OLS}} - \mathbf{w}^* = \left(\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T\right)^{-1} \left(\sum_{i=1}^n \langle \mathbf{f}^*, \mathbf{x}_i \rangle \, \epsilon_i \mathbf{x}_i\right)$$

Applying Lemma 9, we note that the following holds with probability at least $1 - \delta/3$

$$\begin{aligned} \|\hat{\mathbf{w}}_{\mathsf{OLS}} - \mathbf{w}^*\| &\leq \mathrm{Tr}(\mathbf{M}) \left(1 + 8\log(3d/\delta)\right) \\ \mathbf{M} &= \left(\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T\right)^{-1} \left(\sum_{i=1}^n \langle \mathbf{f}^*, \mathbf{x}_i \rangle^2 \mathbf{x}_i \mathbf{x}_i^T\right) \left(\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T\right)^{-1} \end{aligned}$$

By suprema of subgaussian random variables, the following holds with probability at least $1 - \delta/3$

$$\langle \mathbf{f}^*, \mathbf{x}_i \rangle^2 \leq \|\mathbf{f}^*\|^2 \log(3n/\delta) \; \forall i \in [n]$$

Furthermore, by concentration of Wishart matrices (Vershynin, 2010), the following holds with probability at least $1 - {}^{\delta}/_{3}$

$$\lambda_{\text{min}} \left(\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T \right) \geq n \left(1 - \sqrt{\frac{d \log({}^{3d}/_{\delta})}{n}} \right) \geq {}^{n}/_{2}$$

where we use the fact that $n \geq \tilde{O}(d)$. Hence, by a union bound, it follows that,

$$\begin{aligned} \text{Tr}\left(\mathbf{M}\right) &\leq \left\|\mathbf{f}^*\right\|^2 \log({}^{3n}/_{\delta}) \,\text{Tr}\left(\left(\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T \right)^{-1} \right) \\ &\leq \left\|\mathbf{f}^*\right\|^2 \log({}^{3n}/_{\delta}) \frac{d}{\lambda_{\text{min}}\left(\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T\right)} \\ &\leq \left\|\mathbf{f}^*\right\|^2 \log({}^{3n}/_{\delta}) \frac{2d}{n} \end{aligned}$$

Then, it follows that,

$$\left\|\hat{\mathbf{w}}_{\mathsf{OLS}} - \mathbf{w}^*\right\|^2 \leq \left\|\mathbf{f}^*\right\|^2 \frac{d}{n} \mathsf{polylog}({}^{nd}/_{\delta})$$

Finally, the guarantee that $\left\|\hat{\mathbf{w}}_{\mathsf{OLS}} - \mathbf{w}^*\right\|^2 = \Theta({}^{d\|\mathbf{f}^*\|^2}/_{n})$ with probability at least $1 - d/n^c$, $c \geq 1$ follows from Theorem 1 of Chaudhuri et al. (2017).

Appendix B. Analysis of Spectral Method – Proof of Theorem 4

We shall use the fact that for any positive semidefinite matrix $\mathbf{A}$, and event E that occurs with probability at least $1 - \delta$ for any $\delta \leq {}^{1}/_{2}$, $\mathbb{E}\left[\mathbf{A}|E\right] \preceq \frac{\mathbb{E}[\mathbf{A}]}{1-\delta} \preceq 2\mathbb{E}[\mathbf{A}]$. $\hat{\mathbf{w}}$ of $\mathbf{w}^*$ and let $\Delta = \mathbf{w}^* - \hat{\mathbf{w}}$. Furthermore, let $\mathbf{e} = {}^{\mathbf{f}^*}/_{\|\mathbf{f}^*\|}$ and define the matrices $\hat{\mathbf{S}}$, $\mathbf{S}$ and Σ as follows,

$$\begin{aligned} \hat{\mathbf{S}} &= \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \langle \hat{\mathbf{w}}, \mathbf{x}_i \rangle\right)^2 \mathbf{x}_i \mathbf{x}_i^T \\ \mathbf{S} &= \frac{1}{n} \sum_{i=1}^{n} \langle \mathbf{f}^*, \mathbf{x}_i \rangle^2 \mathbf{x}_i \mathbf{x}_i^T \\ \Sigma &= 3\|\mathbf{f}^*\|^2 \mathbf{e}\mathbf{e}^T + \|\mathbf{f}^*\|^2 (\mathbf{I} - \mathbf{e}\mathbf{e}^T) \end{aligned}$$

Using $y_i = \langle \mathbf{w}^*, \mathbf{x}_i \rangle + \epsilon_i \langle \mathbf{f}^*, \mathbf{x}_i \rangle$, and writing $\epsilon_i^2 = 1 + z_i$, we expand $\hat{\mathbf{S}}$ as follows,

$$\begin{aligned} \hat{\mathbf{S}} &= \frac{1}{n} \sum_{i=1}^{n} \left(\langle \Delta, \mathbf{x}_i \rangle + \epsilon_i\right)^2 \mathbf{x}_i \mathbf{x}_i^T \\ &= \mathbf{S} + \frac{1}{n} \sum_{i=1}^{n} \langle \Delta, \mathbf{x}_i \rangle^2 \mathbf{x}_i \mathbf{x}_i^T + \frac{2}{n} \sum_{i=1}^{n} \epsilon_i \langle \Delta, \mathbf{x}_i \rangle \langle \mathbf{f}^*, \mathbf{x}_i \rangle \mathbf{x}_i \mathbf{x}_i^T + \frac{1}{n} \sum_{i=1}^{n} z_i \langle \mathbf{f}^*, \mathbf{x}_i \rangle^2 \mathbf{x}_i \mathbf{x}_i^T \end{aligned}$$

We note that $\mathbf{f}^*$ is the top eigenvector of Σ with eigenvalue $3\|\mathbf{f}^*\|^2$. Moreover, by definition $\hat{\mathbf{f}}_{\mathsf{S}}$ is the top eigenvector of $\hat{\Sigma}$ with norm $\sqrt{{}^{\|\hat{\Sigma}\|}/_{3}}$. To this end, our proof is divided into three distinct parts, namely, controlling $\|\mathbf{S} - \Sigma\|$ via Matrix Bernstein, bounding $\left\|\hat{\mathbf{S}} - \mathbf{S}\right\|$ as a function of $\|\Delta\|$, and finally, bounding $\left\|\hat{\mathbf{f}}_{\mathsf{S}} - \mathbf{f}^*\right\|$ as a function of $\left\|\hat{\mathbf{S}} - \Sigma\right\|$ via Davis-Kahan theorem.

B.1. Controlling $\|\mathbf{S} - \Sigma\|$

Since $\mathbf{x}_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \mathbf{I})$, we note that $\mathbb{E}[\hat{\Sigma}] = \Sigma$. Thus we control $\left\|\hat{\Sigma} - \Sigma\right\|$ via Matrix Bernstein's inequality. To this end, define the matrix $\mathbf{A}_i = \langle \mathbf{f}^*, \mathbf{x}_i\rangle^2 \mathbf{x}_i\mathbf{x}_i^T$. Let E_1 and E_2 denote the events $E_1 = \left\{\langle \mathbf{f}^*, \mathbf{x}_i\rangle^2 \leq \|\mathbf{f}^*\|^2 \log(4n/\delta)\ \forall\, i \in [n]\right\}$, $E_2 = \left\{\|\mathbf{x}_i\|^2 \leq d + \log(4n/\delta)\right\}$, and let $E = E_1 \cap E_2$. Gaussian concentration implies $\mathbb{P}(E) \geq 1 - \delta$. Moreover, conditioned on the event E, the following holds,

$$
\begin{aligned}
\|\mathbf{A}_i\| &\leq \langle \mathbf{f}^*, \mathbf{x}_i\rangle^2 \|\mathbf{x}_i\|^2 \leq \|\mathbf{f}^*\|^2 \log(4n/\delta)\, [d + \log(4n/\delta)] \leq \|\mathbf{f}^*\|^2 d \mathsf{polylog}(nd/\delta) \\
\mathbb{E}\left[\mathbf{A}_i\mathbf{A}_i^T | E\right] &= \mathbb{E}\left[\langle \mathbf{f}^*, \mathbf{x}_i\rangle^4 \|\mathbf{x}_i\|^2 \mathbf{x}_i\mathbf{x}_i^T | E\right] \preceq \|\mathbf{f}^*\|^4 \log^2(4n/\delta)\,(d + \log(4n/\delta))\, \mathbb{E}\left[\mathbf{x}_i\mathbf{x}_i^T | E\right] \\
&\preceq \mathbf{I} d \|\mathbf{f}^*\|^4 \mathsf{polylog}(nd/\delta)
\end{aligned}
$$

Notice that conditioned on E, notice that $\mathbf{A}_i$ are still i.i.d random matrices. Hence, by Matrix Bernstein inequality, $P(E_3|E) \geq 1 - \delta$ where E_3 is given by,

$$
E_3 = \left\{\left\|\hat{\Sigma} - \mathbb{E}\left[\mathbf{A}_i|E\right]\right\| \leq \|\mathbf{f}^*\|^2 \sqrt{\frac{d}{n}} \mathsf{polylog}(nd/\delta)\right\}
$$

It follows that

$$
\mathbb{P}\left\{\left\|\hat{\Sigma} - \mathbb{E}\left[\mathbf{A}_i|E\right]\right\| \leq \|\mathbf{f}^*\|^2 \sqrt{\frac{d}{n}} \mathsf{polylog}(nd/\delta)\right\} \geq 1 - 2\delta \tag{13}
$$

We now need to show that $\mathbb{E}\left[\mathbf{A}_i|E\right] \approx \mathbb{E}\left[\mathbf{A}_i\right]$. Consider:

$$
\begin{aligned}
&\|\mathbb{E}\left[\mathbf{A}_i|E\right] - \mathbb{E}\left[\mathbf{A}_i\right]\| = \left\|\frac{\mathbb{E}\left[\mathbf{A}_i \mathbb{1}(E)\right]}{\mathbb{P}(E)} - \mathbb{E}\left[\mathbf{A}_i\right]\right\| = \left\|\frac{\mathbb{E}(\mathbf{A}_i) - \mathbb{E}\left[\mathbf{A}_i \mathbb{1}(E^{\complement})\right]}{\mathbb{P}(E)} - \mathbb{E}\left[\mathbf{A}_i\right]\right\| \\
&\leq \frac{\mathbb{P}(E^{\complement})\|\mathbb{E}[\mathbf{A}_i]\|}{\mathbb{P}(E)} + \frac{\left\|\mathbb{E}[\mathbf{A}_i \mathbb{1}(E^{\complement})]\right\|}{\mathbb{P}(E)} \leq \frac{\mathbb{P}(E^{\complement})\|\mathbb{E}[\mathbf{A}_i]\|}{\mathbb{P}(E)} + \frac{\sqrt{\|\mathbb{E}[\mathbf{A}_i^2]\|\mathbb{P}(E^{\complement})}}{\mathbb{P}(E)} \\
&\leq C\|\mathbf{f}^*\|^2[\delta + \sqrt{d\delta}]
\end{aligned}
$$

The second inequality follows from the Cauchy-Schwarz inequality. The last inequality follows from the fact that $\|\mathbb{E}[\mathbf{A}_i]\| = 3\|\mathbf{f}^*\|^2$ and $\|\mathbb{E}\mathbf{A}_i^2\| \leq C\|\mathbf{f}^*\|^4 d$ for some constant $C > 0$. Replacing δ above with $\delta^2/\mathsf{poly}(nd)$, we conclude from Equation (13) that:

$$
\mathbb{P}\left\{\left\|\hat{\Sigma} - \Sigma\right\| \leq \|\mathbf{f}^*\|^2 \sqrt{\frac{d}{n}} \mathsf{polylog}(nd/\delta)\right\} \geq 1 - \delta \tag{14}
$$

B.2. Controlling $\left\|\hat{\mathbf{S}} - \mathbf{S}\right\|$

We control each term in $\left\|\hat{\mathbf{S}} - \mathbf{S}\right\|$ as follows,

B.2.1. BOUNDING $\left\| 1/n \sum_{i=1}^{n} \langle \Delta, \mathbf{x}_i \rangle^2 \mathbf{x}_i \mathbf{x}_i^T \right\|$

Let E_1 denote the event $E_1 = \left\{ \langle \Delta, \mathbf{x}_i \rangle^2 \leq \|\Delta\|^2 \log(4n/\delta) \right\}$. Then, by suprema of Gaussian random variables, we know that $\mathbb{P}(E_1) \geq 1 - \delta/2$. Furthermore, define the event E_2 as follows,

$$E_2 = \left\{ \left\| \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T - \mathbf{I} \right\| \leq 2\sqrt{\frac{d}{n}} + 2t + \left(\sqrt{\frac{d}{n}} + t \right)^2 \right\}$$

where $t = \sqrt{\frac{2\log(4/\delta)}{n}}$. Taking a union bound over E_1 and E_2, and using the concentration properties of Wishart matrices, we conclude that the following must hold with probability at least $1 - \delta$.

$$\begin{aligned}
\frac{1}{n} \sum_{i=1}^{n} \langle \Delta, \mathbf{x}_i \rangle^2 \mathbf{x}_i \mathbf{x}_i^T &\preceq \frac{\|\Delta\|^2 \log(4n/\delta)}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T \\
\left\| \frac{1}{n} \sum_{i=1}^{n} \langle \Delta, \mathbf{x}_i \rangle^2 \mathbf{x}_i \mathbf{x}_i^T \right\| &\leq \|\Delta\|^2 \log(4n/\delta) \left[1 + 2\sqrt{\frac{d}{n}} + 2\sqrt{\frac{2\log(4/\delta)}{n}} + \left(\sqrt{\frac{d}{n}} + \sqrt{\frac{2\log(4/\delta)}{n}} \right)^2 \right] \\
&\leq \|\Delta\|^2 \mathsf{polylog}(nd/\delta) \qquad (15)
\end{aligned}$$

where the last inequality uses the fact that $n \geq d\mathsf{polylog}(nd/\delta)$

B.2.2. BOUNDING $\left\| \frac{2}{n} \sum_{i=1}^{n} \epsilon_i \langle \Delta, \mathbf{x}_i \rangle \langle \mathbf{f}^*, \mathbf{x}_i \rangle \mathbf{x}_i \mathbf{x}_i^T \right\|$

Define the events E_1, E_2, E_3, E_4 as,

$$\begin{aligned}
E_1 &= \left\{ |\epsilon_i| \leq \sqrt{\log(8n/\delta)} \right\} \\
E_2 &= \left\{ |\langle \Delta, \mathbf{x}_i \rangle| \leq \|\Delta\| \sqrt{\log(8n/\delta)} \right\} \\
E_3 &= \left\{ |\langle \mathbf{f}^*, \mathbf{x}_i \rangle| \leq \|\mathbf{f}^*\| \sqrt{\log(8n/\delta)} \right\} \\
E_4 &= \left\{ \|\mathbf{x}_i\|^2 \leq d + \log(8n/\delta) \right\}
\end{aligned}$$

Let $E = E_1 \cap E_2 \cap E_3 \cap E_4$. It follows from Gaussian concentration that $P(E) \geq 1 - \delta$. We now follow the same steps as the above. In particular, let $\mathbf{B}_i = \epsilon_i \langle \Delta, \mathbf{x}_i \rangle \langle \mathbf{f}^*, \mathbf{x}_i \rangle \mathbf{x}_i \mathbf{x}_i^T$. Then, conditioned on E,

$$\begin{aligned}
\|\mathbf{B}_i\| &\leq |\epsilon_i| |\langle \Delta, \mathbf{x}_i \rangle| \|\mathbf{x}_i\|^2 \\
&\leq \|\Delta\| \|\mathbf{f}^*\| \log^{3/2}(8n/\delta) \left(d + \log(8n/\delta) \right) \leq \|\mathbf{f}^*\| \|\Delta\| d\mathsf{polylog}(nd/\delta) \\
\mathbb{E}\left[\mathbf{B}_i \mathbf{B}_i^T | E \right] &= \mathbb{E}\left[\epsilon_i^2 \langle \Delta, \mathbf{x}_i \rangle^2 \langle \mathbf{f}^*, \mathbf{x}_i \rangle^2 \mathbf{x}_i \mathbf{x}_i^T | E \right] \\
&\preceq \|\Delta\|^2 \|\mathbf{f}^*\|^2 \log^3(8n/\delta) \left(d + \log(8n/\delta) \right) \mathbb{E}\left[\mathbf{x}_i \mathbf{x}_i^T | E \right] \\
&\preceq \|\Delta\|^2 \|\mathbf{f}^*\|^2 \mathsf{polylog}(nd/\delta) \mathbf{I}
\end{aligned}$$

Notice that B_i are still i.i.d. when conditioned on E. Hence, by the matrix Bernstein inequality, $\frac{2}{n}\sum_{i=1}^n \epsilon_i \langle \Delta, \mathbf{x}_i\rangle \langle \mathbf{f}^*, \mathbf{x}_i\rangle \mathbf{x}_i\mathbf{x}_i^T$

$$\mathbb{P}\left\{\left\|\frac{1}{n}\sum_{i=1}^n \epsilon_i \langle \Delta, \mathbf{x}_i\rangle \langle \mathbf{f}^*, \mathbf{x}_i\rangle \mathbf{x}_i\mathbf{x}_i^T - \mathbb{E}\left[\mathbf{B}_i|E\right]\right\| \leq \|\mathbf{f}^*\|\|\Delta\|\sqrt{\frac{d}{n}}\mathsf{polylog}({}^{d}\!/\!{}_{\delta})\right\} \geq 1-2\delta$$

Note that $\mathbb{E}[\mathbf{B}_i|E] = 0$ since $\mathbb{E}[\epsilon_i|E] = 0$. This allows us to conclude:

$$\mathbb{P}\left\{\left\|\frac{1}{n}\sum_{i=1}^n \epsilon_i \langle \Delta, \mathbf{x}_i\rangle \langle \mathbf{f}^*, \mathbf{x}_i\rangle \mathbf{x}_i\mathbf{x}_i^T\right\| \leq \|\mathbf{f}^*\|\|\Delta\|\sqrt{\frac{d}{n}}\mathsf{polylog}({}^{nd}\!/\!{}_{\delta})\right\} \geq 1-2\delta \tag{16}$$

B.2.3. BOUNDING $\left\|{}^1\!/\!{}_n\sum_{i=1}^n z_i\langle \mathbf{f}^*, \mathbf{x}_i\rangle^2\mathbf{x}_i\mathbf{x}_i^T\right\|$

Let $\mathbf{C}_i = z_i\langle \mathbf{f}^*, \mathbf{x}_i\rangle^2\mathbf{x}_i\mathbf{x}_i^T$. Define the events E_1, E_2, E_3 as

$$\begin{aligned}
E_1 &= \{|z_i| \leq \log({}^{6n}\!/\!{}_{\delta})\}\\
E_2 &= \left\{\langle \mathbf{f}^*, \mathbf{x}_i\rangle^2 \leq \|\mathbf{f}^*\|^2 \log({}^{6n}\!/\!{}_{\delta})\right\}\\
E_3 &= \left\{\|\mathbf{x}_i\|^2 \leq d + \log({}^{6n}\!/\!{}_{\delta})\right\}
\end{aligned}$$

Let $E = E_1 \cap E_2 \cap E_3$. Then $P(E) \geq 1-3\delta$. Conditioning on E and following the same steps as before,

$$\begin{aligned}
\|\mathbf{C}_i\| &\leq \|\mathbf{f}^*\|^2 \log^2({}^{6n}\!/\!{}_{\delta})\left[d + \log({}^{6n}\!/\!{}_{\delta})\right]\\
\mathbb{E}\left[\mathbf{C}_i\mathbf{C}_i|E\right] &= \mathbb{E}\left[z_i^2\langle \mathbf{f}^*, \mathbf{x}_i\rangle^4\|\mathbf{x}_i\|^2\mathbf{x}_i\mathbf{x}_i^T|E\right]\\
&\preceq 2\|\mathbf{f}^*\|^4 \log^4({}^{6n}\!/\!{}_{\delta})\left[d + \log({}^{6n}\!/\!{}_{\delta})\right]\mathbf{I}
\end{aligned}$$

Hence, applying the matrix Bernstein inequality in a similar way as above, we obtain:

$$\mathbb{P}\left\{\left\|\frac{1}{n}\sum_{i=1}^n z_i\langle \mathbf{f}^*, \mathbf{x}_i\rangle^2 - \mathbb{E}\left[\mathbf{C}_i|E\right]\right\| \leq \|\mathbf{f}^*\|^2\sqrt{\frac{d}{n}}\mathsf{polylog}({}^{nd}\!/\!{}_{\delta})\right\} \geq 1-2\delta \tag{17}$$

Now, it remains to bound $\|\mathbb{E}[\mathbf{C}_i|E]\|$. Notice that, when conditioned on E, z_i and $\langle \mathbf{f}^*, \mathbf{x}_i\rangle^2\mathbf{x}_i\mathbf{x}_i^T$ are independent. Therefore, we conclude:

$$\begin{aligned}
&\|\mathbb{E}[\mathbf{C}_i|E]\| \lesssim |\mathbb{E}[z_i|E]|\|\mathbf{f}^*\|^2 d\log\left(\tfrac{6n}{\delta}\right)^2 = \frac{\|\mathbf{f}^*\|^2 d\log\left(\frac{6n}{\delta}\right)^2}{\mathbb{P}(E)}|\mathbb{E}[z_i\mathbb{1}(E)]|\\
&= \frac{\|\mathbf{f}^*\|^2 d\log\left(\frac{6n}{\delta}\right)^2}{\mathbb{P}(E)}|\mathbb{E}[z_i\mathbb{1}(E^{\complement})]| \leq \frac{\|\mathbf{f}^*\|^2 d\log\left(\frac{6n}{\delta}\right)^2}{\mathbb{P}(E)}\sqrt{\mathbb{E}z_i^2\mathbb{P}(E^{\complement})}\\
&\lesssim \frac{\|\mathbf{f}^*\|^2 d\log\left(\frac{6n}{\delta}\right)^2}{\mathbb{P}(E)}|\mathbb{E}[z_i\mathbb{1}(E^{\complement})]| \leq \|\mathbf{f}^*\|^2 d\log\left(\tfrac{6n}{\delta}\right)^2\sqrt{\delta}
\end{aligned}$$

In the second line we have used the fact that $\mathbb{E}[z_i] = 0$ and hence $\mathbb{E}[z_i \mathbb{1}(E)] = -\mathbb{E}[z_i \mathbb{1}(E^{\complement})]$. We have also used the Cauchy-Scwharz inequality. Therefore, replacing δ with $\delta^2/\mathsf{poly}(nd)$ in the above discussion and using Equation (17), we conclude:

$$\mathbb{P}\left\{\left\|\frac{1}{n}\sum_{i=1}^{n} z_i \langle \Delta, \mathbf{x}_i\rangle \langle \mathbf{f}^*, \mathbf{x}_i\rangle\right\| \leq \|\mathbf{f}^*\|^2 \sqrt{\frac{d}{n}} \mathsf{polylog}({}^{nd}/_{\delta})\right\} \geq 1-\delta \tag{18}$$

From (14), (15), (16) and (18), we finally conclude that with probability at least $1-\delta$,

$$\left\|\hat{\mathbf{S}} - \Sigma\right\| \leq \left\|\hat{\mathbf{S}} - \mathbf{S}\right\| + \|\mathbf{S} - \Sigma\| \leq \left(\|\Delta\|^2 + \left(\|\mathbf{f}^*\|^2 + \|\mathbf{f}^*\|\|\Delta\|\right)\sqrt{\frac{d}{n}}\right) \mathsf{polylog}({}^{nd}/_{\delta}) \tag{19}$$

B.3. Controlling $\left\|\hat{\mathbf{f}} - \mathbf{f}^*\right\|^2$

Let $\hat{\mathbf{f}}_{\mathsf{S}}$ be the top eigenvector of $\hat{\mathbf{S}}$ with $\left\|\hat{\mathbf{f}}_{\mathsf{S}}\right\| = \sqrt{\|\hat{\mathbf{S}}\|/3}$. Let θ be the angle between $\hat{\mathbf{f}}_{\mathsf{S}}$ and $\mathbf{f}^*$. We assume $\hat{\mathbf{f}}_{\mathsf{S}}$ is aligned with $\mathbf{f}^*$, i.e., $\theta \in [-\pi/2, \pi/2]$. This assumption is without loss of generality because the heteroscedastic regression model is invariant to the sign of $\mathbf{f}^*$. Let $\mathbf{g}$ be a unit vector orthogonal to $\mathbf{f}^*$ such that $\hat{\mathbf{f}}_{\mathsf{S}} = \left\|\hat{\mathbf{f}}_{\mathsf{S}}\right\| \cos\theta \mathbf{e} + \left\|\hat{\mathbf{f}}_{\mathsf{S}}\right\| \sin\theta \mathbf{g}$. Moreover, let $\mathbf{v} = \|\mathbf{f}^*\| \cos\theta \mathbf{e} + \|\mathbf{f}^*\| \sin\theta \mathbf{g}$. Since Σ has a spectral gap of $2\|\mathbf{f}^*\|^2$, it follows from the Davis Kahan theorem that,

$$\|\mathbf{v} - \mathbf{f}^*\| \leq \frac{\sqrt{2}\left\|\hat{\mathbf{S}} - \Sigma\right\|}{\|\mathbf{f}^*\|^2} \tag{20}$$

Furthermore, since $\|\hat{\mathbf{S}}\| = 3\|\hat{\mathbf{f}}_{\mathsf{S}}\|^2$, $\|\hat{\Sigma}\| = 3\|\mathbf{f}^*\|^2$ and $\left\|\hat{\mathbf{S}}\right\| \leq \|\Sigma\| + \left\|\hat{\mathbf{S}} - \Sigma\right\|$, we conclude that,

$$\begin{aligned}
\left\|\hat{\mathbf{f}}_{\mathsf{S}}\right\|^2 &\leq \|\mathbf{f}^*\|^2 \left(1 + \frac{\left\|\hat{\mathbf{S}} - \Sigma\right\|}{3\|\mathbf{f}^*\|^2}\right) \\
\left\|\hat{\mathbf{f}}_{\mathsf{S}}\right\| &\leq \|\mathbf{f}^*\| \left(1 + \frac{\left\|\hat{\mathbf{S}} - \Sigma\right\|}{3\|\mathbf{f}^*\|^2}\right)^{1/2} \\
&\leq \|\mathbf{f}^*\| \left(1 + \frac{\left\|\hat{\mathbf{S}} - \Sigma\right\|}{3\|\mathbf{f}^*\|^2}\right) \\
\left\|\hat{\mathbf{f}}_{\mathsf{S}}\right\| - \|\mathbf{f}^*\| &\leq \frac{\left\|\hat{\mathbf{S}} - \Sigma\right\|}{3\|\mathbf{f}^*\|}
\end{aligned}$$

where we use the fact that $\sqrt{1+x} \leq 1+x$ for any $x \geq 0$. We follow similar steps to lower bound $\left\|\hat{\mathbf{f}}_{\mathsf{S}}\right\| - \|\mathbf{f}^*\|$. In particular, since $\left\|\hat{\mathbf{S}}\right\| \geq \|\Sigma\| - \left\|\hat{\mathbf{S}} - \Sigma\right\|$, we infer that,

$$\left\|\hat{\mathbf{f}}_{\mathsf{S}}\right\|^2 \geq \|\mathbf{f}^*\|^2 \left(1 - \frac{\left\|\hat{\mathbf{S}} - \Sigma\right\|}{3\|\mathbf{f}^*\|^2}\right)$$

Since $\sqrt{1-t} \geq 1 - t \,\forall\, t \in [0,1]$, and $\|\hat{\Sigma}\|, \|\hat{\mathbf{f}}_{\mathsf{S}}\| \geq 0$, we conclude from the above inequality that the following must hold

$$\left\|\hat{\mathbf{f}}_{\mathsf{S}}\right\| \geq \|\mathbf{f}^*\| \left(1 - \frac{\left\|\hat{\mathbf{S}} - \Sigma\right\|}{3\|\mathbf{f}^*\|^2}\right)$$

$$\left\|\hat{\mathbf{f}}_{\mathsf{S}}\right\| - \|\mathbf{f}^*\| \geq -\frac{\left\|\hat{\mathbf{S}} - \Sigma\right\|}{3\|\mathbf{f}^*\|}$$

Hence, $\left\|\hat{\mathbf{f}} - \mathbf{v}\right\| = \left|\left\|\hat{\mathbf{f}}_{\mathsf{S}}\right\| - \|\mathbf{f}^*\|\right| \leq \frac{\|\hat{\mathbf{S}} - \Sigma\|}{3\|\mathbf{f}^*\|}$. From (19) and (20), we obtain the following with probability at least $1 - \delta$

$$\begin{aligned}\left\|\hat{\mathbf{f}} - \mathbf{f}^*\right\|^2 &\leq \frac{5\left\|\hat{\mathbf{S}} - \Sigma\right\|^2}{\|\mathbf{f}^*\|^2} \\ &\leq \left(\frac{\|\Delta\|^4}{\|\mathbf{f}^*\|^2} + \left(\|\mathbf{f}^*\|^2 + \|\Delta\|^2\right)\frac{d}{n}\right)\mathsf{polylog}(nd/\delta) \\ &\leq \left(\frac{\epsilon^2}{\|\mathbf{f}^*\|^2} + \left(\|\mathbf{f}^*\|^2 + \epsilon\right)\frac{d}{n}\right)\mathsf{polylog}(nd/\delta)\end{aligned}$$

Appendix C. Phase Retrieval with Multiplicative Noise

C.1. Derivation of the Pseudo Gradient

First, we will derive the Pseudo Graident $\mathcal{G}$ from the fictitious square loss. Suppose we had access to $\mathbf{w}^*$ and $\mathbf{f}^*$. Then $(y_i - \langle \mathbf{w}^*, \mathbf{x}_i \rangle)^2 = \epsilon_i^2 \langle \mathbf{f}^*, \mathbf{x}_i \rangle^2$. $\mathbb{E}\left[\epsilon_i^2 \langle \mathbf{f}^*, \mathbf{x}_i \rangle^2 | \mathbf{x}_i\right] = \langle \mathbf{f}^*, \mathbf{x}_i \rangle^2$ and $\mathsf{var}(\epsilon_i^2 \langle \mathbf{f}^*, \mathbf{x}_i \rangle^2 | \mathbf{x}_i) = 2\langle \mathbf{f}^*, \mathbf{x}_i \rangle^4$. The (fictitious) square loss function which for recovering f would be:

$$\mathcal{L}^{\mathsf{mul}}(\mathbf{f}) = \frac{1}{m}\sum_{i=1}^{m} \frac{\left[(y_i - \langle \mathbf{w}^*, \mathbf{x}_i \rangle)^2 - \langle \mathbf{f}, \mathbf{x}_i \rangle^2\right]^2}{\langle \mathbf{f}^*, \mathbf{x}_i \rangle^4}$$

Now, the actual gradient of this fictitious loss is:

$$\nabla\mathcal{L}^{\mathsf{mul}}(\mathbf{f}) = \frac{2}{m}\sum_{i=1}^{m} \frac{\left[\langle \mathbf{f}, \mathbf{x}_i \rangle^2 - (y_i - \langle \mathbf{w}^*, \mathbf{x}_i \rangle)^2\right]\langle \mathbf{f}, \mathbf{x}_i \rangle \mathbf{x}_i}{\langle \mathbf{f}^*, \mathbf{x}_i \rangle^4}$$

Note that this loss function cannot be computed. Assuming we have a good enough estimate $\hat{\mathbf{f}} \approx \mathbf{f}^*$ and $\hat{\mathbf{w}} \approx \mathbf{w}^*$, and that $\mathbf{f} \approx \hat{\mathbf{f}}$, we make replace $\mathbf{f}^*$ with $\hat{\mathbf{f}}$ and $\mathbf{w}^*$ with $\hat{\mathbf{w}}$ and for the sake of convenience $\langle f, \mathbf{x}_i \rangle$ with $\langle \hat{\mathbf{f}}, \mathbf{x}_i \rangle$. With these approximation, we obtain:

$$\bar{\mathcal{G}}(\mathbf{f}) := \frac{1}{m}\sum_{i=1}^{m} \frac{\left[\langle \mathbf{f}, \mathbf{x}_i \rangle^2 - (y_i - \langle \hat{\mathbf{w}}, \mathbf{x}_i \rangle)^2\right]\langle \hat{\mathbf{f}}, \mathbf{x}_i \rangle \mathbf{x}_i}{\langle \hat{\mathbf{f}}, \mathbf{x}_i \rangle^4}$$

In order to prevent the denominator from exploding, we add the 'regularization' $\mathbb{1}(|\langle \hat{\mathbf{f}}, \mathbf{x}_i \rangle| \geq \bar{\mu})$ to obtain the defined pseudo-gradient:

$$\mathcal{G}(\mathbf{f}) := \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}(|\langle \hat{\mathbf{f}}, \mathbf{x}_i\rangle| \geq \bar{\mu}) \frac{\left[\langle \mathbf{f}, \mathbf{x}_i\rangle^2 - (y_i - \langle \hat{\mathbf{w}}, \mathbf{x}_i\rangle)^2\right] \langle \hat{\mathbf{f}}, \mathbf{x}_i\rangle \mathbf{x}_i}{\langle \hat{\mathbf{f}}, \mathbf{x}_i\rangle^4}$$

C.2. Noise-Contraction Decomposition of the Pseudo Gradient

We now give the noise-contraction decomposition of the pseudo-gradient $\mathcal{G}_t(\mathbf{f})$, where we write $\mathcal{G}_t(\mathbf{f}) = H_t(\mathbf{f})(\mathbf{f} - \mathbf{f}^*) + N_t$ where $H_t(\mathbf{f})$ is a PSD matrix whenever $\mathbf{f}$ is close to $\mathbf{f}^*$ and N_t is the noise term. For the sake of clarity, only in this derivation, take $\gamma(\mathbf{x}_i) := \frac{\mathbb{1}(|\langle \hat{\mathbf{f}}, \mathbf{x}_i\rangle| \geq \bar{\mu})}{\langle \hat{\mathbf{f}}, \mathbf{x}_i\rangle^4}$.

$$\begin{aligned}\mathcal{G}_t(\mathbf{f}) &= \frac{1}{m}\sum_{i=1}^{m} \gamma(\mathbf{x}_i^{(t)})(\langle \mathbf{f}, \mathbf{x}_i^{(t)}\rangle^2 - (Y_i - \langle \hat{\mathbf{w}}, \mathbf{x}_i\rangle)^2)\mathbf{x}_i \langle \hat{\mathbf{f}}, \mathbf{x}_i^{(t)}\rangle \\ &= \frac{1}{m}\sum_{i=1}^{m} \gamma(\mathbf{x}_i^{(t)})(\langle \mathbf{f}, \mathbf{x}_i^{(t)}\rangle^2 - \langle \mathbf{f}^*, \mathbf{x}_i^{(t)}\rangle^2)\mathbf{x}_i^{(t)} \langle \hat{\mathbf{f}}, \mathbf{x}_i^{(t)}\rangle \\ &\quad + \frac{1}{m}\sum_{i=1}^{m} \gamma(\mathbf{x}_i^{(t)})(\langle \mathbf{f}^*, \mathbf{x}_i^{(t)}\rangle^2 - (Y_i - \langle \hat{\mathbf{w}}, \mathbf{x}_i^{(t)}\rangle)^2)\mathbf{x}_i^{(t)} \langle \hat{\mathbf{f}}, \mathbf{x}_i^{(t)}\rangle \end{aligned} \tag{21}$$

Define $N_t = \frac{1}{m}\sum_{i=1}^{m} \gamma(\mathbf{x}_i^{(t)})(\langle \mathbf{f}^*, \mathbf{x}_i^{(t)}\rangle^2 - (Y_i - \langle \hat{\mathbf{w}}, \mathbf{x}_i^{(t)}\rangle)^2)\mathbf{x}_i^{(t)} \langle \hat{\mathbf{f}}, \mathbf{x}_i^{(t)}\rangle$. Now, consider the first term in Equation (21). We have:

$$\begin{aligned}&\frac{1}{m}\sum_{i=1}^{m} \gamma(\mathbf{x}_i^{(t)})(\langle \mathbf{f}, \mathbf{x}_i^{(t)}\rangle^2 - \langle \mathbf{f}^*, \mathbf{x}_i^{(t)}\rangle^2)\mathbf{x}_i^{(t)} \langle \hat{\mathbf{f}}, \mathbf{x}_i^{(t)}\rangle \\ &= \frac{1}{m}\sum_{i=1}^{m} \gamma(\mathbf{x}_i^{(t)})\langle \mathbf{f} - \mathbf{f}^*, \mathbf{x}_i^{(t)}\rangle \langle f + \mathbf{f}^*, \mathbf{x}_i^{(t)}\rangle \mathbf{x}_i^{(t)} \langle \hat{\mathbf{f}}, \mathbf{x}_i^{(t)}\rangle \\ &=: H_t(\mathbf{f})(\mathbf{f} - \mathbf{f}^*) \end{aligned} \tag{22}$$

Where we define $H_t(\mathbf{f}) := \frac{1}{m}\sum_{i=1}^{m} \gamma(\mathbf{x}_i^{(t)})\langle \mathbf{f} + \mathbf{f}^*, \mathbf{x}_i^{(t)}\rangle \langle \hat{\mathbf{f}}, \mathbf{x}_i^{(t)}\rangle \mathbf{x}_i^{(t)} (\mathbf{x}_i^{(t)})^\intercal$. Plugging these into Equation (21), we conclude that:

$$\mathcal{G}_t(\mathbf{f}) = H_t(\mathbf{f})(\mathbf{f} - \mathbf{f}^*) + N_t \tag{23}$$

C.3. Bounding the Contraction Matrix and the Noise Vector

In this section we will state results regarding the matrix $H_t(\mathbf{f})$ and the vector N_t which will aid us in proving the convergence bounds. In this subsection only, for the sake of clarity, we drop the dependence on t and let $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m)$ to be derived from the model specified in Section 2. We let $\Delta = \mathbf{w}^* - \hat{\mathbf{w}}$ and $\Gamma = \max\Big(\|\mathbf{f}^* - \hat{\mathbf{f}}\|, \|\mathbf{f} - \hat{\mathbf{f}}\|\Big)$.

The following lemma, which we state without proof, follows from Gaussian concentration.

Lemma 16 *For any fixed vector $v \in \mathbb{R}^d$, we have $\mathbb{P}(\sup_i |\langle v, \mathbf{x}_i\rangle| > t\|v\|) \leq 2m \exp\Big(-\frac{t^2}{2}\Big)$.*

First, we turn our attention to the noise term. For the sake of clarity, we will take $\gamma(\mathbf{x}_i) := \frac{\mathbb{1}(|\langle \hat{\mathbf{f}}, \mathbf{x}_i\rangle| \geq \bar{\mu})}{\langle \hat{\mathbf{f}}, \mathbf{x}_i\rangle^4}$. The noise term can be written as:

$$N = \frac{1}{m}\sum_{i=1}^{m} \gamma(\mathbf{x}_i)(\langle \mathbf{f}^*, \mathbf{x}_i\rangle^2(1-\epsilon_i^2)\mathbf{x}_i\langle \hat{\mathbf{f}}, \mathbf{x}_i\rangle + \frac{2}{m}\sum_{i=1}^{m}\gamma(\mathbf{x}_i)\epsilon_i\langle \mathbf{f}^*, \mathbf{x}_i\rangle\langle \Delta, \mathbf{x}_i\rangle \mathbf{x}_i\langle \hat{\mathbf{f}}, \mathbf{x}_i\rangle$$
$$- \frac{1}{m}\sum_{i=1}^{m}\gamma(\mathbf{x}_i)\langle \Delta, \mathbf{x}_i\rangle^2\mathbf{x}_i\langle \hat{\mathbf{f}}, \mathbf{x}_i\rangle \tag{24}$$

Consider

$$\bar{N}_1 := \frac{1}{m}\sum_{i=1}^{m}\gamma(\mathbf{x}_i)(\langle \mathbf{f}^*, \mathbf{x}_i\rangle^2(1-\epsilon_i^2)\mathbf{x}_i\langle \hat{\mathbf{f}}, \mathbf{x}_i\rangle$$
$$\bar{N}_2 := \frac{2}{m}\sum_{i=1}^{m}\gamma(\mathbf{x}_i)\epsilon_i\langle \mathbf{f}^*, \mathbf{x}_i\rangle\langle \Delta, \mathbf{x}_i\rangle \mathbf{x}_i\langle \hat{\mathbf{f}}, \mathbf{x}_i\rangle$$
$$\bar{N}_3 := \frac{1}{m}\sum_{i=1}^{m}\gamma(\mathbf{x}_i)\langle \Delta, \mathbf{x}_i\rangle^2\mathbf{x}_i\langle \hat{\mathbf{f}}, \mathbf{x}_i\rangle \tag{25}$$

Lemmas 17, 18 and 19 bound this quantity. We refer to Section E for their proof.

Lemma 17 *Suppose $u \in \mathbb{R}^d$ and $\|\hat{\mathbf{f}}\| > \bar{\mu}$:*

1. *Let $h_1(X,u) = \frac{1}{m}\sum_{i=1}^{m}\mathbb{1}(|\langle \hat{\mathbf{f}}, \mathbf{x}_i\rangle| > \bar{\mu})\frac{\langle \mathbf{f}^*, \mathbf{x}_i\rangle^4\langle \mathbf{x}_i, u\rangle^2}{\langle \hat{\mathbf{f}}, \mathbf{x}_i\rangle^6}$*

$$\mathbb{P}\left(|\langle \bar{N}_1, u\rangle| > c_0 t\sqrt{\tfrac{h_1(X,u)}{m}}\middle| \mathbf{x}_1, \dots, \mathbf{x}_m\right) \leq \exp(-t)$$

2.

$$\mathbb{P}\left(h_1(X,\hat{\mathbf{f}}) > C(1 + \tfrac{\Gamma^4}{\bar{\mu}^4}\log^2(\tfrac{m}{\delta}))\right) \leq \delta$$

3. *Suppose $u \perp \hat{\mathbf{f}}$ and $\|u\| = 1$. Let $r = \lceil \log\left(\frac{\|\hat{\mathbf{f}}\|_2}{\bar{\mu}}\right)\rceil$. Assume $m \geq C\frac{\|\hat{\mathbf{f}}\|}{\bar{\mu}}\log(\frac{r}{\delta})$ for some large enough constant C. Then:*

$$\mathbb{P}\left(h_1(X,u) > \tfrac{C}{\bar{\mu}\|\hat{\mathbf{f}}\|}(1 + \tfrac{\Gamma^4}{\bar{\mu}^4}\log^2(\tfrac{m}{\delta}))\right) \leq \delta$$

Lemma 18 *Let $u \in \mathbb{R}^d$ be any arbitrary vector. Assume $\|\hat{\mathbf{f}}\| > \bar{\mu}$. Let $r = \lceil \log\left(\frac{\|\hat{\mathbf{f}}\|_2}{\bar{\mu}}\right)\rceil$. Define $h_2(X,u) := \frac{1}{m}\sum_{i=1}^{m}\mathbb{1}(|\langle \hat{\mathbf{f}}, \mathbf{x}_i\rangle| > \bar{\mu})\frac{\langle f^*, \mathbf{x}_i\rangle^2\langle \mathbf{x}_i, u\rangle^2\langle \Delta, \mathbf{x}_i\rangle^2}{\langle \hat{\mathbf{f}}, \mathbf{x}_i\rangle^6}$*

1.

$$\mathbb{P}\left(|\langle \bar{N}_2, u\rangle| > C_1\sqrt{\tfrac{h_2(X,u)\log\frac{1}{\delta}}{m}}\middle| \mathbf{x}_1, \dots, \mathbf{x}_m\right) \leq \delta$$

2. *Assume $m \geq C\frac{\|\hat{\mathbf{f}}\|}{\bar{\mu}}\log\left(\frac{r}{\delta}\right)$ for some large enough constant C.*

$$\mathbb{P}\left(h_2(X,\hat{\mathbf{f}}) > C_1\frac{\|\Delta\|^2}{\bar{\mu}\|\hat{\mathbf{f}}\|}\log\left(\tfrac{m}{\delta}\right)\left(1+\frac{\Gamma^2\log\left(\frac{m}{\delta}\right)}{\bar{\mu}^2}\right)\right) \leq \delta$$

3. *Let $u \perp \hat{\mathbf{f}}$ and $\|u\| = 1$. Assume $m \geq C\frac{\|\hat{\mathbf{f}}\|}{\bar{\mu}}\log\left(\frac{r}{\delta}\right)$ for some large enough constant C.*

$$\mathbb{P}\left(h_2(X,u) > C_1\frac{\|\Delta\|^2}{\bar{\mu}^3\|\hat{\mathbf{f}}\|}\log^2\left(\tfrac{m}{\delta}\right)\left(1+\frac{\Gamma^2\log\left(\frac{m}{\delta}\right)}{\bar{\mu}^2}\right)\right) \leq \delta$$

Lemma 19 *Suppose $\|\hat{\mathbf{f}}\| > \bar{\mu}$. Let $r = \left\lceil \log\left(\frac{\|\hat{\mathbf{f}}\|_2}{\bar{\mu}}\right)\right\rceil$. Assume $m \geq C\frac{\|\hat{\mathbf{f}}\|}{\bar{\mu}}\log\left(\frac{r}{\delta}\right)$, for some large enough constant C.*

1. $$\mathbb{P}\left(|\langle \bar{N}_3, \hat{\mathbf{f}}\rangle| > C\frac{\|\Delta\|^2}{\bar{\mu}\|\hat{\mathbf{f}}\|}\right) \leq \delta$$

2. *Let Q be the projector onto the subspace perpendicular to $\hat{\mathbf{f}}$. Then, we must have:*

$$\|Q\bar{N}_3\| \leq \frac{C\|\Delta\|^2}{\bar{\mu}\|\hat{\mathbf{f}}\|^2} + \sqrt{\frac{C\|\Delta\|^4\log^3\left(\frac{m}{\delta}\right)\left(d+\log\left(\frac{m}{\delta}\right)\right)}{m\bar{\mu}^5\|\hat{\mathbf{f}}\|}}$$

While $H(\mathbf{f})$ need not be a PSD matrix almost surely, we show that whenever Γ is small enough, $H(\mathbf{f})$ is a PSD matrix with high probability as shown below.

Lemma 20 *Suppose with $\bar{\mu} > C\Gamma\sqrt{\log\frac{m}{\delta}}$ for some large enough universal constant C. With probability at-least $1-\delta$, we must have:*

$$\frac{1}{m}\sum_{i=1}^{m}\frac{\mathbb{1}(|\langle\hat{\mathbf{f}},\mathbf{x}_i\rangle| \geq \bar{\mu})}{\langle\hat{\mathbf{f}},\mathbf{x}_i\rangle^4}|\langle\hat{\mathbf{f}},\mathbf{x}_i\rangle|^2\mathbf{x}_i\mathbf{x}_i^{\intercal} \preceq H(f) \preceq \frac{4}{m}\sum_{i=1}^{m}\frac{\mathbb{1}(|\langle\hat{\mathbf{f}},\mathbf{x}_i\rangle| \geq \bar{\mu})}{\langle\hat{\mathbf{f}},\mathbf{x}_i\rangle^4}|\langle\hat{\mathbf{f}},\mathbf{x}_i\rangle|^2\mathbf{x}_i\mathbf{x}_i^{\intercal}$$

Proof Only in this proof, we take $\gamma(\mathbf{x}_i) := \frac{\mathbb{1}(|\langle\hat{\mathbf{f}},\mathbf{x}_i\rangle|\geq\bar{\mu})}{\langle\hat{\mathbf{f}},\mathbf{x}_i\rangle^4}$.

$$H(\mathbf{f}) = \frac{2}{m}\sum_{i=1}^{m}\gamma(\mathbf{x}_i)\langle\hat{\mathbf{f}},\mathbf{x}_i\rangle^2\mathbf{x}_i\mathbf{x}_i^{\intercal} + \frac{1}{m}\sum_{i=1}^{m}\gamma(\mathbf{x}_i)\langle\mathbf{f}+\mathbf{f}^*-2\hat{\mathbf{f}},\mathbf{x}_i\rangle\langle\hat{\mathbf{f}},\mathbf{x}_i\rangle\mathbf{x}_i\mathbf{x}_i^{\intercal} \tag{26}$$

By an application of Lemma 16, we conclude that with probability at-least $1-\delta$, we must have: $\sup_{i\in[m]}|\langle\mathbf{f}+\mathbf{f}^*-\hat{\mathbf{f}},\mathbf{x}_i\rangle| \leq C_0\Gamma\sqrt{\log\frac{m}{\delta}}$ for some universal constant C_0. Therefore, by the definition of $H(\mathbf{f})$, we conclude that under this event:

$$\begin{aligned}
H(\mathbf{f}) &\preceq \frac{2}{m}\sum_{i=1}^{m}\gamma(\mathbf{x}_i)\langle\hat{\mathbf{f}},\mathbf{x}_i\rangle^2\mathbf{x}_i\mathbf{x}_i^{\intercal} + \frac{1}{m}\sum_{i=1}^{m}C_0\gamma(\mathbf{x}_i)\Gamma\sqrt{\log\tfrac{m}{\delta}}|\langle\hat{\mathbf{f}},\mathbf{x}_i\rangle|\mathbf{x}_i\mathbf{x}_i^{\intercal} \\
&= \frac{2}{m}\sum_{i=1}^{m}\gamma(\mathbf{x}_i)|\langle\hat{\mathbf{f}},\mathbf{x}_i\rangle|\left(|\langle\hat{\mathbf{f}},\mathbf{x}_i\rangle| + C_1\Gamma\sqrt{\log\left(\tfrac{m}{\delta}\right)}\right)\mathbf{x}_i\mathbf{x}_i^{\intercal} \\
&\preceq \frac{4}{m}\sum_{i=1}^{m}\gamma(\mathbf{x}_i)|\langle\hat{\mathbf{f}},\mathbf{x}_i\rangle|^2\mathbf{x}_i\mathbf{x}_i^{\intercal}
\end{aligned} \tag{27}$$

In the third step, we have used the fact that $\gamma(\mathbf{x}_i) \neq 0$ iff $|\langle \hat{\mathbf{f}}, \mathbf{x}_i\rangle| \geq \bar{\mu}$ therefore, whenever $\gamma(\mathbf{x}_i) \neq 0$, by the assumption in the statement of the lemma, we must have: $|\langle \hat{\mathbf{f}}, \mathbf{x}_i\rangle| + C_1\Gamma\sqrt{\log\left(\frac{m}{\delta}\right)} \leq 2|\langle \hat{\mathbf{f}}, \mathbf{x}_i\rangle|$. The lower bounds follow in a similar way. The lower bound can be shown in a similar way by considering $|\langle \hat{\mathbf{f}}, \mathbf{x}_i\rangle| - C_1\Gamma\sqrt{\log\left(\frac{m}{\delta}\right)} \geq \frac{1}{2}|\langle \hat{\mathbf{f}}, \mathbf{x}_i\rangle|$. ■

The following result states some useful bounds for $H(\mathbf{f})$. We refer to Section E.4 for the proof of this lemma.

Lemma 21 *Assume $\bar{\mu} < \|\hat{\mathbf{f}}\|$. Let $r := \left\lceil \log\left(\frac{\|\hat{\mathbf{f}}\|_2}{\bar{\mu}}\right)\right\rceil$, $\bar{\mu} > C\Gamma\sqrt{\log\frac{m}{\delta}}$ and $m \geq C\frac{\|\hat{\mathbf{f}}\|}{\bar{\mu}}\log\left(\frac{r}{\delta}\right)$ for some large enough universal constant C. Then, the following bounds hold:*

1.

$$\mathbb{P}\left(\tfrac{1}{2} \leq \langle \hat{\mathbf{f}}, H(\mathbf{f})\hat{\mathbf{f}}\rangle \leq 8\right) \geq 1 - 2\exp(-c_0 m)$$

2. *Let $u \perp \hat{\mathbf{f}}$ be such that $\|u\| = 1$*

$$\mathbb{P}\left(|\langle u, H(\mathbf{f})\hat{\mathbf{f}}\rangle| \leq \frac{C\Gamma}{\bar{\mu}\|\hat{\mathbf{f}}\|} + \sqrt{\frac{C\log\left(\frac{1}{\delta}\right)}{m\bar{\mu}\|\hat{\mathbf{f}}\|}}\right) \geq 1 - \delta$$

3. *Let $\hat{\mathbf{f}}^{\perp} \in \mathsf{span}(\mathbf{f} + \mathbf{f}^* - 2\hat{\mathbf{f}}, \hat{\mathbf{f}})$ and $\hat{\mathbf{f}}^{\perp} \perp \hat{\mathbf{f}}$, $\|\hat{\mathbf{f}}^{\perp}\| = 1$. Suppose $u \perp \hat{\mathbf{f}}^{\perp}$ and $u \perp \hat{\mathbf{f}}$.*

$$\mathbb{P}\left(|\langle u, H(\mathbf{f})\hat{\mathbf{f}}\rangle| \leq \Gamma\sqrt{\frac{C\log\left(\frac{1}{\delta}\right)}{m\bar{\mu}^3\|\hat{\mathbf{f}}\|}} + \sqrt{\frac{C\log(1/\delta)}{m\bar{\mu}\|\hat{\mathbf{f}}\|}}\right) \geq 1 - \delta$$

4.

$$\mathbb{P}\left(\frac{c_0}{\|\hat{\mathbf{f}}\|\bar{\mu}} \leq \langle u, H(\mathbf{f})u\rangle \leq \frac{c_1}{\|\hat{\mathbf{f}}\|\bar{\mu}}\right) \geq 1 - \delta$$

5. *Let Q be the projector to the sup-space perpendicular to $\hat{\mathbf{f}}$. With probability at-least $1 - \delta$, we must have:*

$$\|QH(\mathbf{f})\hat{\mathbf{f}}\|^2 \leq \frac{Cd\log(d/\delta)}{m\bar{\mu}\|\hat{\mathbf{f}}\|} + \frac{Cd\Gamma^2\log^2(\frac{d}{\delta})}{m\bar{\mu}^3\|\hat{\mathbf{f}}\|} + \frac{C\Gamma^2}{\bar{\mu}^2\|\hat{\mathbf{f}}\|^2}$$

6. *Let Q be the projector to the sub-space perpendicular to $\hat{\mathbf{f}}$ and let $u \perp \hat{\mathbf{f}}$ such that $\|u\| = 1$. With probability at-least $1 - \delta$:*

$$\|QH(\mathbf{f})u\| \leq \frac{C}{\|\hat{\mathbf{f}}\|\bar{\mu}} + \sqrt{\frac{Cd\log^2(\frac{4md}{\delta})}{m\bar{\mu}^3\|\hat{\mathbf{f}}\|}} + \sqrt{\frac{Cd\Gamma^2\log\left(\frac{4d}{\delta}\right)\log^3(\frac{8m}{\delta})}{m\bar{\mu}^5\|\hat{\mathbf{f}}\|}}$$

C.4. Proof of Theorem 6

We are now ready to prove Theorem 6. Under the assumption of the theorem, $(\mathbf{x}_i^{(t)}, y_i^{(t)})$ are all i.i.d. Therefore, the iterate $\mathbf{f}_t$ is independent of $(\mathbf{x}_i^{(t)}, y_i^{(t)})$ for $i \in [m]$.

Now, consider the dynamics given in Algorithm 4. Using the noise-contraction decomposition (Equation (23)), we conclude: $\mathbf{f}_{t+1} - \mathbf{f}^* = (I - DH_t(\mathbf{f}_t))(\mathbf{f}_t - \mathbf{f}^*) + DN_t$.

Now, write $\mathbf{f}_t = a_t \frac{\hat{\mathbf{f}}}{\|\hat{\mathbf{f}}\|} + b_t u_t$ where $u_t \perp \hat{\mathbf{f}}$. Similarly, write $N_t = \bar{a}_t \frac{\hat{\mathbf{f}}}{\|\hat{\mathbf{f}}\|} + \bar{b}_t \bar{u}_t$ with $\bar{u}_t \perp \hat{\mathbf{f}}$

Recall that $\hat{\mathbf{f}} \in \mathsf{span}(P)$ and $u_t, \bar{u}_t \in \mathsf{span}(Q)$, where $P = \frac{\hat{\mathbf{f}}\hat{\mathbf{f}}^T}{\|\hat{\mathbf{f}}\|^2}$ and $Q = I - P$. Now, let us track the evolution of a_t, b_t Now,

$$\begin{aligned}
P(\mathbf{f}_{t+1} - \mathbf{f}^*) = a_{t+1}\frac{\hat{\mathbf{f}}}{\|\hat{\mathbf{f}}\|} &= P(I - \mathbf{D}H_t(\mathbf{f}_t))(\mathbf{f}_t - \mathbf{f}^*) + P\mathbf{D}N_t \\
&= P(I - \alpha_0 H_t)(\mathbf{f}_t - \mathbf{f}^*) + \alpha_0 P N_t \\
&= P(I - \alpha_0 H_t(\mathbf{f}_t))P(\mathbf{f}_t - \mathbf{f}^*) + P(I - \alpha_0 H_t(\mathbf{f}_t))Q(\mathbf{f}_t - \mathbf{f}^*) + \alpha_0 \bar{a}_t \frac{\hat{\mathbf{f}}}{\|\hat{\mathbf{f}}\|} \\
&= a_t P(I - \alpha_0 H_t(\mathbf{f}_t))P\frac{\hat{\mathbf{f}}}{\|\hat{\mathbf{f}}\|} + b_t P(I - \alpha_0 H_t(\mathbf{f}_t))Q u_t + \alpha_0 \bar{a}_t \frac{\hat{\mathbf{f}}}{\|\hat{\mathbf{f}}\|} \\
&= a_t P(I - \alpha_0 H_t(\mathbf{f}_t))P\frac{\hat{\mathbf{f}}}{\|\hat{\mathbf{f}}\|} - \alpha_0 b_t P H_t(\mathbf{f}_t) Q u_t + \alpha_0 \bar{a}_t \frac{\hat{\mathbf{f}}}{\|\hat{\mathbf{f}}\|} \qquad (28)
\end{aligned}$$

In the second line we have used the fact that $P\mathbf{D} = \alpha_0 P$ by the definition of $\mathbf{D}$. In the third line, we have used the fact that $P + Q = \mathbf{I}$. In the fourth line, we have used the fact that $PQ = 0$. Similarly, interchanging P and Q, we get

$$Q(\mathbf{f}_{t+1} - \mathbf{f}^*) = b_t Q(I - \alpha_1 H_t(\mathbf{f}_t))Q u_t - \alpha_1 a_t Q H_t(\mathbf{f}_t)\frac{\hat{\mathbf{f}}}{\|\hat{\mathbf{f}}\|} + \alpha_1 \bar{b}_t \bar{u}_t \qquad (29)$$

Now, define $\Gamma_t := \max(\|\mathbf{f}^* - \hat{\mathbf{f}}\|, \|\mathbf{f}_t - \hat{\mathbf{f}}\|)$.

We state the following lemmas which are proved in Sections E.5 and E.6.

Lemma 22 *Assume $\bar{\mu} < \|\hat{\mathbf{f}}\|$. Let $r := \lceil \log\left(\frac{\|\hat{\mathbf{f}}\|_2}{\bar{\mu}}\right) \rceil$, $\bar{\mu} > C^{\mathsf{dist}}\Gamma_t \log\frac{m}{\delta}$, $m \geq C^{\mathsf{par}}\frac{\|\hat{\mathbf{f}}\|}{\bar{\mu}}\log\left(\frac{r}{\delta}\right)$, $C^{\mathsf{par}}\|\Delta\|\log\left(\frac{m}{\delta}\right) \leq \bar{\mu}$ for some large enough constants $C^{\mathsf{par}}, C^{\mathsf{dist}}$, $\alpha_0 \leq c_1\|\hat{\mathbf{f}}\|^2$ for some small enough constant c_1. Then, conditioned on $\mathbf{f}_t$ being such that the above conditions hold, we have with probability at-least $1 - \delta$:*

1. *For some c^{con} which does not depend on $C^{\mathsf{par}}, C^{\mathsf{dist}}$, we must have:*

$$\left\|P(I - \alpha_0 H_t(\mathbf{f}_t))P\frac{\hat{\mathbf{f}}}{\|\hat{\mathbf{f}}\|}\right\|^2 \leq \left(1 - \frac{c^{\mathsf{con}}\alpha_0}{\|\hat{\mathbf{f}}\|^2}\right)$$

2. *For some C which does not depend on $C^{\mathsf{par}}, C^{\mathsf{dist}}$*

$$\|PH_t(\mathbf{f}_t)Qu_t\| \leq \frac{C\Gamma_t}{\bar{\mu}\|\hat{\mathbf{f}}\|^2} + \sqrt{\frac{C\log\left(\frac{1}{\delta}\right)}{m\bar{\mu}\|\hat{\mathbf{f}}\|^3}}$$

3. *For some C which does not depend on $C^{\mathsf{par}}, C^{\mathsf{dist}}$*

$$|\bar{a}_t| \leq C\left[\frac{\log\left(\frac{1}{\delta}\right)}{\|\hat{\mathbf{f}}\|\sqrt{m}} + \frac{\|\Delta\|^2}{\bar{\mu}\|\hat{\mathbf{f}}\|^2}\right] \tag{30}$$

Lemma 23 *Assume $\bar{\mu} < \|\hat{\mathbf{f}}\|$ and let $r := \left\lceil \log\left(\frac{\|\hat{\mathbf{f}}\|_2}{\bar{\mu}}\right)\right\rceil$, $\bar{\mu} > C^{\mathsf{dist}}\Gamma_t \log\frac{m}{\delta}$,*

$$m \geq C^{\mathsf{par}} \max\left(\frac{\|\hat{\mathbf{f}}\|}{\bar{\mu}}\log\left(\tfrac{r}{\delta}\right), \frac{\|\hat{\mathbf{f}}\| d\log^4\left(\frac{m}{\delta}\right)}{\bar{\mu}}, \frac{\|\hat{\mathbf{f}}\|^2\log^4\left(\frac{m}{\delta}\right)}{\bar{\mu}^2}\right),$$

$C^{\mathsf{par}}\|\Delta\|\log\left(\frac{m}{\delta}\right) \leq \bar{\mu}$ for some large enough universal constants $C^{\mathsf{par}}, C^{\mathsf{dist}}$, $\alpha_1 \leq c_1\|\hat{\mathbf{f}}\|\bar{\mu}$ for some small enough constant c_1. Then, conditioned on $\mathbf{f}_t$ being such that the above conditions hold, we have with probability at-least $1-\delta$:

1. *For some c^{con} which does not depend on $C^{\mathsf{par}}, C^{\mathsf{dist}}$, we must have:*

$$\|Q(I - \alpha_1 H_t(\mathbf{f}_t))Qu_t\|^2 \leq \left(1 - \tfrac{c^{\mathsf{con}}\alpha_1}{\|\hat{\mathbf{f}}\|\bar{\mu}}\right)$$

2. *For some C which does not depend on $C^{\mathsf{par}}, C^{\mathsf{dist}}$*

$$\left\|QH_t(\mathbf{f}_t)P\frac{\hat{\mathbf{f}}}{\|\hat{\mathbf{f}}\|}\right\|^2 \leq \frac{Cd\log\left(\frac{d}{\delta}\right)}{m\bar{\mu}\|\hat{\mathbf{f}}\|^3} + \frac{C\Gamma_t^2}{\bar{\mu}^2\|\hat{\mathbf{f}}\|^4}$$

3. *For some C which does not depend on $C^{\mathsf{par}}, C^{\mathsf{dist}}$*

$$|\bar{b}_t| \leq C\log\left(\tfrac{d}{\delta}\right)\sqrt{\frac{d}{m\bar{\mu}\|\hat{\mathbf{f}}\|}} + \frac{C\|\Delta\|^2}{\bar{\mu}\|\hat{\mathbf{f}}\|^2}$$

Under the assumptions of Lemma 22 and Lemma 23, we apply the triangle inequality to Equation (28) and use the bounds in Lemma 22 via the union bound. Given any $c_{ab} > 0$, we can take the constants $C^{\mathsf{par}}, C^{\mathsf{dist}}$ relating Δ, Γ_t, m and $\bar{\mu}$ in Lemmata 22 and 23 to be large enough, such that with probability $1-\delta$:

$$|a_{t+1}| \leq |a_t|\sqrt{\left(1 - \tfrac{c^{\mathsf{con}}\alpha_0}{\|\hat{\mathbf{f}}\|^2}\right)} + \frac{\alpha_0}{\|\hat{\mathbf{f}}\|^2}c_{ab}|b_t| + C\alpha_0\left[\frac{\log\left(\frac{1}{\delta}\right)}{\|\hat{\mathbf{f}}\|\sqrt{m}} + \frac{\|\Delta\|^2}{\bar{\mu}\|\hat{\mathbf{f}}\|^2}\right]$$

Similarly, we have with probability $1-\delta$:

$$|b_{t+1}| \leq |b_t|\sqrt{\left(1 - \tfrac{c^{\mathsf{con}}\alpha_1}{\bar{\mu}\|\hat{\mathbf{f}}\|}\right)} + \alpha_1\frac{c_{ab}}{\|\hat{\mathbf{f}}\|^2}|a_t| + C\alpha_1\left[\log\left(\tfrac{d}{\delta}\right)\sqrt{\frac{d}{m\bar{\mu}\|f\|}} + \frac{\|\Delta\|^2}{\bar{\mu}\|\hat{\mathbf{f}}\|^2}\right]$$

Now define:

$$\kappa_t := \begin{bmatrix} |a_t| \\ |b_t| \end{bmatrix} \in \mathbb{R}^2, \beta := \begin{bmatrix} C\alpha_0 \left[\frac{\log\left(\frac{1}{\delta}\right)}{\|\hat{\mathbf{f}}\|\sqrt{m}} + \frac{\|\Delta\|^2}{\bar{\mu}\|\hat{\mathbf{f}}\|^2}\right] \\ C\alpha_1 \left[\log\left(\frac{d}{\delta}\right)\sqrt{\frac{d}{m\bar{\mu}\|\hat{\mathbf{f}}\|}} + \frac{\|\Delta\|^2}{\bar{\mu}\|\hat{\mathbf{f}}\|^2}\right] \end{bmatrix} \in \mathbb{R}^2, \Theta := \begin{bmatrix} \sqrt{1 - \frac{\alpha_0 c^{\mathsf{con}}}{\|\hat{\mathbf{f}}\|^2}} & \frac{\alpha_0 c_{ab}}{\|\hat{\mathbf{f}}\|^2} \\ \alpha_1 \frac{c_{ab}}{\|\hat{\mathbf{f}}\|^2} & \sqrt{1 - \frac{c^{\mathsf{con}}\alpha_1}{\|\hat{\mathbf{f}}\|\bar{\mu}}} \end{bmatrix} \in \mathbb{R}^{2\times 2}$$

From the recursion for $|b_t|$ and $|a_t|$ above, we conclude that with probability $1-\delta$, we have the following evolution equation for κ when the assumptions above are satisfied.

$$\kappa_{t+1} \leq \Theta\kappa_t + \beta \tag{31}$$

Where the vector inequalities are interpreted to be coordinate-wise. Notice that $\Gamma_t \leq \|\mathbf{f}_t - \mathbf{f}^*\| + \|\mathbf{f}^* - \hat{\mathbf{f}}\|$ and since the initial condition $\mathbf{f}_0 = \hat{\mathbf{f}}$, we must have $\Gamma_t \leq \|\mathbf{f}_t - \mathbf{f}^*\| + \Gamma_0$ and $\|\mathbf{f}_t - \mathbf{f}^*\| = \|\kappa_t\|$. Therefore, we conclude $\Gamma_t \leq \|\kappa_t\| + \Gamma_0$.

Now define the event $\mathcal{C}_{t_0}$ to be the event that the Equation (31) is satisfied for $0 \leq t \leq t_0 - 1$. From the discussions above, we must have:

$$\mathbb{P}\left(\mathcal{C}_{t+1}\middle|\mathcal{C}_t, C^{\mathsf{dist}}\Gamma_{t+1}\log\left(\tfrac{m}{\delta}\right) \leq \bar{\mu}\right) \geq 1 - \delta \tag{32}$$

Below, we will show that $\mathcal{C}_{t+1}$ has a high probability conditioned on $\mathcal{C}_t$ by showing that under the event $\mathcal{C}_t$, Γ_{t+1} is small.

Conditioned on $\mathcal{C}_t$, we unfurl the recursion in Equation (31) the following holds almost surely:

$$\kappa_{t+1} \leq \Theta^{t+1}\kappa_0 + \sum_{s=0}^{t}\Theta^{t-s}\beta \tag{33}$$

Take $\alpha_0 = c_0\|\hat{\mathbf{f}}\|^2$ and $\alpha_1 = c_0\bar{\mu}\|\hat{\mathbf{f}}\|$ for some small enough constant c_0 as in the statement of the theorem (this can be done independently of C^{par} and C^{dist} as per Lemmas 22 and 23). We then pick C^{par} and C^{dist} large enough to make c_{ab} small enough to ensure $0 \preceq \Theta \preceq (1-\gamma)\mathbf{I}$ for some $\gamma \in (0,1)$ for some constant γ which does not depend on $C^{\mathsf{par}}, C^{\mathsf{dist}}$ (by using diagonal dominance for instance). Thus, the following follows from Equation (33):

$$\|\kappa_{t+1}\| \leq \|\Theta\|_{\mathsf{op}}^{t+1}\|\kappa_0\| + \frac{\|\beta\|}{1 - \|\Theta\|_{\mathsf{op}}} = \|\Theta\|_{\mathsf{op}}^{t+1}\Gamma_0 + \frac{\|\beta\|}{1 - \|\Theta\|_{\mathsf{op}}} \tag{34}$$

Using the fact that $\Gamma_t \leq \|\kappa_t\| + \Gamma_0$, we conclude that conditioned on $\mathcal{C}_t$, the following holds almost surely:

$$\Gamma_{t+1} \leq \left(1 + \|\Theta\|_{\mathsf{op}}^{t+1}\right)\Gamma_0 + \frac{\|\beta\|}{1 - \|\Theta\|_{\mathsf{op}}}$$

Therefore, conditioned on $\mathcal{C}_t$, we must have almost surely:

$$
\begin{aligned}
\Gamma_{t+1} &\leq 2\Gamma_0 + \frac{\|\beta\|}{\gamma} \leq 2\Gamma_0 + C\left[\frac{\|\hat{\mathbf{f}}\|\log\left(\frac{1}{\delta}\right)}{\sqrt{m}} + \frac{\|\Delta\|^2}{\bar{\mu}}\right] + C\left[\log\left(\tfrac{d}{\delta}\right)\sqrt{\frac{d\bar{\mu}\|\hat{\mathbf{f}}\|}{m}} + \frac{\|\Delta\|^2}{\|\hat{\mathbf{f}}\|}\right] \\
&\leq 2\Gamma_0 + C\left[\frac{\|\hat{\mathbf{f}}\|\log\left(\frac{1}{\delta}\right)}{\sqrt{m}} + \frac{\|\Delta\|^2}{\bar{\mu}} + \log\left(\tfrac{d}{\delta}\right)\sqrt{\frac{d\bar{\mu}\|\hat{\mathbf{f}}\|}{m}}\right] \\
&\leq 2\Gamma_0 + \frac{C\bar{\mu}}{\sqrt{C^{\mathsf{par}}}\log\left(\frac{m}{\delta}\right)} \leq \frac{C\bar{\mu}}{\sqrt{C^{\mathsf{par}}}\log\left(\frac{m}{\delta}\right)} && (35)
\end{aligned}
$$

In the second line we have used the fact that $\bar{\mu} < \|\hat{\mathbf{f}}\|$. Note that $\Gamma_0 = \|\hat{\mathbf{f}} - \mathbf{f}^*\|$ since $\mathbf{f}_0 = \hat{\mathbf{f}}$.

Taking $C^{\mathsf{dist}} = c\sqrt{C^{\mathsf{par}}}$ for some universal constant c, we check that we can make C^{dist} as large as we wish by making C^{par} large enough. From the above discussion, we conclude that conditioned on $\mathcal{C}_t$, we must have almost surely:

$$
C^{\mathsf{dist}}\Gamma_{t+1}\log\left(\tfrac{m}{\delta}\right) \leq \bar{\mu}
$$

Combining this with Equation (32) and unrolling the recursion, we conclude: $\mathbb{P}(\mathcal{C}_K) \geq 1 - \delta K$. Combining this with Equation (34), with probability at-least $1 - K\delta$, after time K, we must have:

$$
\|\mathbf{f}_K - \mathbf{f}^*\| \leq \exp(-\gamma K)\|\hat{\mathbf{f}} - \mathbf{f}^*\| + C\left[\frac{\|\hat{\mathbf{f}}\|\log\left(\frac{1}{\delta}\right)}{\sqrt{m}} + \frac{\|\Delta\|^2}{\bar{\mu}} + \log\left(\tfrac{d}{\delta}\right)\sqrt{\frac{d\bar{\mu}\|\hat{\mathbf{f}}\|}{m}}\right] \tag{36}
$$

Here, we have used the fact that $\|\kappa_t\| = \|\mathbf{f}_t - \mathbf{f}^*\|$.

Appendix D. Proof of Theorem 7

The proof structure is similar to that of Theorem 15. For ease of exposition, assume $n = mK$ such that $K = \lceil\log_2(n)\rceil\Theta(\log(n))$ and $m \geq d\mathsf{polylog}(d/\delta)$. In the entire proof, we will use these facts freely. Furthermore, denote $e_{\hat{\mathbf{w}}_k} = \|\hat{\mathbf{w}}_k - \mathbf{w}^*\|^2$ and $e_{\hat{\mathbf{f}}_k} = \left\|\hat{\mathbf{f}}_k - \mathbf{f}^*\right\|^2$. Since $\hat{\mathbf{w}}_0$ is the OLS estimate computed on m data points and $\hat{\mathbf{f}}_0$ is the spectral method estimate computed on m data points using $\hat{\mathbf{w}}_0$, it follows that, with probability at least $1 - {}^{2\delta}/_{2(K+1)}$, $e_{\hat{\mathbf{w}}_0} = \tilde{O}(\|\mathbf{f}^*\|^2 {}^{d}/_{m})$ and $e_{\hat{\mathbf{f}}_0} = \tilde{O}(\|\mathbf{f}^*\|^2 {}^{d}/_{m})$. Let $\lambda_1 = \tilde{\Theta}(\|\hat{\mathbf{f}}_0\|^2 {}^{d}/_{m})$. Then, from Theorem 5, it follows that, with probability at least $1 - {}^{3\delta}/_{2(K+1)}$

$$
\begin{aligned}
e_{\hat{\mathbf{w}}_1} &\leq \tilde{O}\left(\left\|\hat{\mathbf{f}}_0\right\|^2 \frac{1}{m} + \left\|\hat{\mathbf{f}}_0\right\| \frac{d\sqrt{\lambda_1}}{m}\right) \\
&\leq \tilde{O}\left(\|\mathbf{f}^*\|^2 \left(1 + {}^{d}/_{m}\right)\left({}^{1}/_{m} + ({}^{d}/_{m})^{1.5}\right)\right)
\end{aligned}
$$

Now, running noisy phase retrieval on m data points with $K_p = \Theta(\log(m))$ steps, using $\hat{\mathbf{f}}_0, \hat{\mathbf{w}}_1$ as estimates and $\bar{\mu}_1 = \tilde{\Theta}(\left\|\hat{\mathbf{f}}_0\right\|\sqrt{{}^{d}/_{m}})$, we conclude from Theorem 6 that the following holds with probability at least $1 - {}^{2\delta}/_{(K+1)}$

$$
\begin{aligned}
e_{\hat{\mathbf{f}}_1} &\leq \tilde{O}\left(\frac{e_{\hat{\mathbf{f}}_0}}{m^2} + \frac{\left\|\hat{\mathbf{f}}_0\right\|^2}{m} + \frac{e^2_{\hat{\mathbf{w}}_1}}{\bar{\mu}_1^2} + \frac{d\left\|\hat{\mathbf{f}}_0\right\|\bar{\mu}_1}{m}\right) \\
&\leq \tilde{O}\left(\|\mathbf{f}^*\|^2\frac{d}{m^3} + \left\|\hat{\mathbf{f}}_0\right\|^2\left({}^1/_m + ({}^d/_m)^{1.5}\right) + \left(\|\mathbf{f}^*\|^2\left({}^1/_m + ({}^d/_m)^{1.5}\right)\right)^2\left(\left\|\hat{\mathbf{f}}_0\right\|^2\frac{d}{m}\right)^{-1}\right) \\
&\leq \tilde{O}\left(\frac{\|\mathbf{f}^*\|^2}{m} + \|\mathbf{f}^*\|^2\left(2 + {}^d/_m\right)\left({}^1/_m + ({}^d/_m)^{1.5}\right)\right) \leq \tilde{O}\left(\|\mathbf{f}^*\|^2\left({}^1/_m + ({}^d/_m)^{1.5}\right)\right)
\end{aligned}
$$

where we use the fact that $m \geq d\mathsf{polylog}(d)$. We shall now complete the proof via an inductive argument. To this end, define $S_k = \sum_{j=0}^{k} {}^1/_{2^j}$ Clearly, $1 \leq S_k \leq 2$ and $S_{k+1} = 1 + {}^{S_k}/_2$. Define the event $E_k(L)$ for some $L > 1$ which does not depend on k and every $1 \leq l \leq k$

1. $\frac{1}{2}\|\mathbf{f}^*\|^2 \leq \|\hat{\mathbf{f}}_l\|^2 \leq 2\|\mathbf{f}^*\|^2$
2. $e_{\hat{\mathbf{f}}_l} \leq L\|\mathbf{f}^*\|^2\left(\max(\frac{l}{m}, \frac{ld^2}{m^2}) + (\frac{d}{m})^{S_l}\right)$

Similarly, for the some $L_1 > 1$ independent of k, we define $F_k(L)$ as the event such that for every $1 \leq l \leq k$

1. $e_{\hat{\mathbf{w}}_l} \leq L_1\|\mathbf{f}^*\|^2\left(\max(\frac{l}{m}, \frac{ld^2}{m^2}) + (\frac{d}{m})^{S_l}\right)$

For some $\mathsf{polylog}()$ independent of k, let

$$
\bar{\mu}_k = \left\|\hat{\mathbf{f}}_k\right\|\sqrt{\max(L_1, L)\max(\tfrac{k}{m}, \tfrac{(k)d^2}{m^2}) + ({}^d/_m)^{S_k}}\,\mathsf{polylog}(\tfrac{nd}{\delta})
$$

$$
\lambda_k = \left\|\hat{\mathbf{f}}_k\right\|^2\left(\max(\tfrac{k}{m}, \tfrac{(k)d^2}{m^2}) + ({}^d/_m)^{S_k}\right)\mathsf{polylog}(\tfrac{nd}{\delta})L
$$

Since $\hat{\mathbf{w}}_{k+1} = \hat{\mathbf{w}}_{\hat{\mathbf{f}}_k,\lambda_k}$, it follows from Theorem 5 that when conditioned on E_k the following holds with probability at least $1 - \frac{\delta}{2(K+1)}$, for $\mathsf{polylog}()$ which does not depend on k:

$$
\begin{aligned}
e_{\hat{\mathbf{w}}_{k+1}} &\leq \left(\frac{\left\|\hat{\mathbf{f}}_k\right\|^2}{m} + \frac{d\left\|\hat{\mathbf{f}}_k\right\|\sqrt{\lambda_k}}{m}\right)\mathsf{polylog}(\tfrac{nd}{\delta}) \\
&\leq \left(\frac{\left\|\hat{\mathbf{f}}_k\right\|^2}{m} + \left\|\hat{\mathbf{f}}_k\right\|^2\max\left(\frac{\sqrt{k}d}{m^{3/2}}, \frac{\sqrt{k}d^2}{m^2}\right) + \left\|\hat{\mathbf{f}}_k\right\|^2\left(\frac{d}{m}\right)^{1+\frac{S_k}{2}}\right)\mathsf{polylog}(\tfrac{nd}{\delta})\sqrt{L} \\
&\leq \left(\|\mathbf{f}^*\|^2\max\left(\frac{(k+1)}{m}, \frac{(k+1)d^2}{m^2}\right) + \|\mathbf{f}^*\|^2\left(\frac{d}{m}\right)^{1+\frac{S_k}{2}}\right)\sqrt{L}\mathsf{polylog}(\tfrac{nd}{\delta}) \qquad (37)
\end{aligned}
$$

In the second step, we have used the definition of λ_{k+1} and the fact that $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$. In the third step, we use the fact that $\frac{d}{m^{3/2}} \leq \max(\frac{1}{m}, \frac{d^2}{m^2})$ since it is their geometric mean. We also use the fact under the event $E_k(L)$, we have $\|\hat{\mathbf{f}}_k\| \leq 2\|\mathbf{f}^*\|$. Therefore for some fixed polylog independent of k:

$$L_1 \geq \sqrt{L}\mathsf{polylog}(nd/\delta) \implies \mathbb{P}(F_{k+1}(L_1)|E_k(L)) \geq 1 - \frac{\delta}{2(K+1)}\,. \tag{38}$$

Conditioned on $E_k \cap F_{k+1}$, running noisy phase retrieval on m data points with $K_p = \Theta(\log(m))$ steps, using $\hat{\mathbf{f}}_k, \hat{\mathbf{w}}_{k+1}$ inputs, we conclude from Theorem 6 that the following holds with probability at least $1 - \delta/2(K+1)$

$$\begin{aligned}
e_{\hat{\mathbf{f}}_{k+1}} &\leq \left(\frac{e_{\hat{\mathbf{f}}_k}}{m^2} + \frac{\left\|\hat{\mathbf{f}}_k\right\|^2}{m} + \frac{e^2_{\hat{\mathbf{w}}_{k+1}}}{\bar{\mu}_k^2} + \frac{d\left\|\hat{\mathbf{f}}_k\right\|\bar{\mu}_k}{m}\right)\mathsf{polylog}(\tfrac{nd}{\delta}) \\
&\leq \left(\frac{\|\mathbf{f}^*\|^2}{m} + e_{\hat{\mathbf{w}}_{k+1}} + \frac{d\|\mathbf{f}^*\|\bar{\mu}_k}{m}\right)\mathsf{polylog}(\tfrac{nd}{\delta}) \\
&\leq \left(\|\mathbf{f}^*\|^2 \max\left(\frac{(k+1)}{m}, \frac{(k+1)d^2}{m^2}\right) + \|\mathbf{f}^*\|^2\left(\frac{d}{m}\right)^{1+\frac{S_k}{2}}\right)(L_1 + \sqrt{\max(L, L_1)})\mathsf{polylog}(\tfrac{nd}{\delta}))
\end{aligned}$$

In the second step, we have used the fact that $\|\hat{\mathbf{f}}_k\| \leq 2\|\mathbf{f}^*\|$ and the fact that $\bar{\mu}_k \geq e_{\hat{\mathbf{w}}_k}$ as per the requirement of Theorem 6 (this can be verified by the choice of parameters and the events E_k, F_{k+1}). The last step follows from a similar calculation as in Equation (37) and substituting the bound on $e_{\hat{\mathbf{w}}_{k+1}}$ implied by the event $F_{k+1}(L_1)$. This allows us to conclude:

$$L \geq (L_1 + \sqrt{\max(L, L_1)})\mathsf{polylog}(\tfrac{nd}{\delta})) \implies \mathbb{P}(E_{k+1}(L)|E_k(L) \cap F_{k+1}(L_1)) \geq 1 - \frac{\delta}{2(K+1)} \tag{39}$$

We note that whenever L is a large enough, fixed poly-log factor $\mathsf{polylog}(\frac{nd}{\delta})$, both the conditions in Equations (38) and (39) are satisfied. Then, from Equations (38) and (39), along with the union bound,

$$\mathbb{P}(E_{K+1}(L) \cap F_{K+1}(L_1)) \geq 1 - \delta$$

Whenever $K = \lceil \log_2 n \rceil$, we conclude that $2 \geq S_l \geq 2 - 1/n$. Thus, from the definition of E_{K+1} and F_{K+1} we conclude that with probability at-least $1 - \delta$, both these inequalities hold:

$$\|\hat{\mathbf{w}}_{K+1} - \mathbf{w}^*\|^2 \leq \|\mathbf{f}^*\|^2 \left(\tfrac{1}{m} + (\tfrac{d}{m})^2\right) \mathsf{polylog}(nd/\delta)$$

$$\|\hat{\mathbf{f}}_{K+1} - \mathbf{f}^*\|^2 \leq \|\mathbf{f}^*\|^2 \left(\tfrac{1}{m} + (\tfrac{d}{m})^2\right) \mathsf{polylog}(nd/\delta)$$

Appendix E. Proofs of Technical Lemmas

The following lemmas will allow us to bound obtain a high-probability bound on certain random variables which we will encounter later. We give their proofs in Section F.

Lemma 24 *Consider i.i.d random vectors $\xi_1, \dots, \xi_n \in \mathbb{R}^k$ a sequence of sets $A_1, \dots, A_r \subseteq \mathbb{R}^k$ such that $A_i \cap A_j = \emptyset$ whenever $i \neq j$. Define $H_i = \sum_{j=1}^{n} \mathbb{1}(\xi_j \in A_i)$. Let $p_i = \mathbb{P}(\xi_1 \in A_i)$ and $\mathcal{H}_i = \{j : \xi_j \in A_i\}$. Then, for some positive constants c_0, c_1:*

$$\mathbb{P}(H_i > 2np_i) \leq \exp(-c_0 np_i)$$

$$\mathbb{P}(H_i < \tfrac{np_i}{2}) \leq \exp(-c_1 np_i)$$

The proof of Lemma 24 follows from an application of Bernstein's inequality for binomial random variables (see Boucheron et al. (2013)).

Lemma 25 *Suppose $z_i = 1 - \epsilon_i^2$, where $\epsilon_i \sim \mathcal{N}(0,1)$. Let $v_1, \dots, v_m \in \mathbb{R}$ be fixed reals. Let $\|v\| := \sqrt{\sum_{i=1}^{m} v_i^2}$. Then, for some positive constants C_0, C_1:*

$$\mathbb{P}(|\sum_{i=1}^{m} z_i v_i| > t\|v\|) \leq C_0 \exp(-C_1 t)\,.$$

Lemma 26 *Assume that $\|\hat{\mathbf{f}}\| > \bar{\mu}$ and define $r = \lceil \log_2(\frac{\|\hat{\mathbf{f}}\|}{\bar{\mu}}) \rceil$. Assume that $m \geq C\frac{\|\hat{\mathbf{f}}\|}{\bar{\mu}} \log\big(\frac{r}{\delta}\big)$. for some large enough constant C. Let $l \geq 1$ be a constant. Let $u \perp \hat{\mathbf{f}}$ and $\|u\| = 1$. Consider the quantities:*

$$L_1(X,u,l) := \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}(|\langle \hat{\mathbf{f}}, \mathbf{x}_i \rangle| > \bar{\mu}) \frac{\langle \mathbf{x}_i, u \rangle^2}{\langle \hat{\mathbf{f}}, \mathbf{x}_i \rangle^{2l}}$$

$$L_2(X,l) := \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}(|\langle \hat{\mathbf{f}}, \mathbf{x}_i \rangle| > \bar{\mu}) \frac{1}{\langle \hat{\mathbf{f}}, \mathbf{x}_i \rangle^{2l}}$$

Then we have the following concentration bounds (making the dependence on l implicit):

1. $\mathbb{P}\left(L_1(X,u) > \frac{C}{\bar{\mu}^{2l-1}\|\hat{\mathbf{f}}\|}\right) \leq \delta$
2. $\mathbb{P}\left(L_1(X,u) < \frac{C_1 2^{-2l}}{\bar{\mu}^{2l-1}\|\hat{\mathbf{f}}\|}\right) \leq \delta$
3. $\mathbb{P}\left(L_2(X) > \frac{C}{\bar{\mu}^{2l-1}\|\hat{\mathbf{f}}\|}\right) \leq \delta$
4. $\mathbb{P}\left(L_2(X) < \frac{C_1 2^{-2l}}{\bar{\mu}^{2l-1}\|\hat{\mathbf{f}}\|}\right) \leq \delta$

We consider symmetrization in the following sense. Draw $\zeta_1, \dots, \zeta_m$ i.i.d rademacher random variables (i.e, uniformly distributed over $\{-1, 1\}$) and independent of $\mathbf{x}_1, \dots, \mathbf{x}_m$. Given any fixed projection matrix R, define $\mathbf{x}_i' = \zeta_i R\mathbf{x}_i + (\mathbf{I} - R)\mathbf{x}_i$.

Lemma 27 *$(X_1', \dots, X_m')$ are jointly distributed as i.i.d $\mathcal{N}(0, \mathbf{I})$.*

E.1. Proof of Lemma 17

Proof [Proof of Lemma 17]

1. Consider $\langle \bar{N}_1, u\rangle = \frac{1}{m}\sum_{i=1}^{m}\gamma(\mathbf{x}_i)\langle \mathbf{f}^*, \mathbf{x}_i\rangle^2\langle \hat{\mathbf{f}}, \mathbf{x}_i\rangle\langle u, \mathbf{x}_i\rangle(1-\epsilon_i^2)$. Invoking Lemma 25, with $v_i = \gamma(\mathbf{x}_i)\langle \mathbf{f}^*, \mathbf{x}_i\rangle^2\langle \hat{\mathbf{f}}, \mathbf{x}_i\rangle\langle u, \mathbf{x}_i\rangle(1-\epsilon_i^2)$ we conclude the result.

2. Let $u \in \mathbb{R}^d$ be arbitrary. By Lemma 16, we have with probability at-least $1-\delta$, for every $i \in [m]$, $|\langle \mathbf{f}^*, \mathbf{x}_i\rangle| \leq \Gamma\sqrt{\log\frac{m}{\delta}} + |\langle \hat{\mathbf{f}}, \mathbf{x}_i\rangle|$. Therefore, we conclude that for some constant $C > 0$,

$$\mathbb{1}(|\langle \hat{\mathbf{f}}, \mathbf{x}_i\rangle| > \bar{\mu})\frac{\langle \mathbf{f}^*, \mathbf{x}_i\rangle^4\langle \mathbf{x}_i, u\rangle^2}{\langle \hat{\mathbf{f}}, \mathbf{x}_i\rangle^6} \leq C\mathbb{1}(|\langle \hat{\mathbf{f}}, \mathbf{x}_i\rangle| > \bar{\mu})\frac{\langle \mathbf{x}_i, u\rangle^2}{\langle \hat{\mathbf{f}}, \mathbf{x}_i\rangle^2}\left(1 + \frac{\Gamma^4\log^2\frac{m}{\delta}}{\bar{\mu}^4}\right)$$

   Therefore, we conclude that with probability at-least $1-\delta$, we must have:

$$h_1(X,u) \leq \frac{C}{m}\sum_{i=1}^{m}\mathbb{1}(|\langle \hat{\mathbf{f}}, \mathbf{x}_i\rangle| > \bar{\mu})\frac{\langle \mathbf{x}_i, u\rangle^2}{\langle \hat{\mathbf{f}}, \mathbf{x}_i\rangle^2}\left(1 + \frac{\Gamma^4\log^2\frac{m}{\delta}}{\bar{\mu}^4}\right) \tag{40}$$

   Taking $u = \hat{\mathbf{f}}$, we conclude the result.

3. We begin with Equation (40), where we replace δ with $\frac{\delta}{2}$. Therefore, we conclude that with probability at-least $1-\frac{\delta}{2}$, we have:

$$h_1(X,u) \leq \frac{C}{m}\sum_{i=1}^{m}\mathbb{1}(|\langle \hat{\mathbf{f}}, \mathbf{x}_i\rangle| > \bar{\mu})\frac{\langle \mathbf{x}_i, u\rangle^2}{\langle \hat{\mathbf{f}}, \mathbf{x}_i\rangle^2}\left(1 + \frac{\Gamma^4\log^2\frac{2m}{\delta}}{\bar{\mu}^4}\right) \tag{41}$$

   Combining Lemma 26 with Equation (41) using the union bound, we conclude that with probability at-least $1-\delta$:

$$h_1(X,u) \leq \tfrac{C}{\bar{\mu}\|\hat{\mathbf{f}}\|}\left(1 + \frac{\Gamma^4\log^2\frac{2m}{\delta}}{\bar{\mu}^4}\right) \tag{42}$$

■

E.2. Proof of Lemma 18

Proof [Proof of Lemma 18]

1. Consider $\langle \bar{N}_2, u\rangle = \frac{2}{m}\sum_{i=1}^{m}\gamma(\mathbf{x}_i)\epsilon_i\langle \mathbf{f}^*, \mathbf{x}_i\rangle\langle \Delta, \mathbf{x}_i\rangle\langle \mathbf{x}_i, u\rangle\langle \hat{\mathbf{f}}, \mathbf{x}_i\rangle$. Conditioned on $\mathbf{x}_1, \dots, \mathbf{x}_m$, this is a one dimensional Gaussian random variable with mean 0 and variance $\frac{4h_2(X,u)}{m}$. The concentration inequality follows from Gaussian concentration.

2. This proof is similar to the proof of Lemma 17. By similar considerations as in Equation (40), we conclude that with probability at-least $1-\delta/4$:

$$h_2(X,u) \leq \frac{C}{m}\sum_{i=1}^{m}\mathbb{1}(|\langle\hat{\mathbf{f}},\mathbf{x}_i\rangle| > \bar{\mu})\frac{\langle\mathbf{x}_i,u\rangle^2\langle\mathbf{x}_i,\Delta\rangle^2}{\langle\hat{\mathbf{f}},\mathbf{x}_i\rangle^4}\left(1+\frac{\Gamma^2\log\frac{4m}{\delta}}{\bar{\mu}^2}\right) \tag{43}$$

Now applying Lemma 16, we have $\sup_{i\in[m]}|\langle\Delta,\mathbf{x}_i\rangle|^2 \leq \|\Delta\|^2\log\left(\frac{4m}{\delta}\right)$ with probability at-least $1-\frac{\delta}{4}$. Combining this with Equation (43) via the union bound, we conclude that with probability at-least $1-\frac{\delta}{2}$, we have:

$$h_2(X,u) \leq \frac{C}{m}\sum_{i=1}^{m}\mathbb{1}(|\langle\hat{\mathbf{f}},\mathbf{x}_i\rangle| > \bar{\mu})\frac{\langle\mathbf{x}_i,u\rangle^2\|\Delta\|^2\log\left(\frac{4m}{\delta}\right)}{\langle\hat{\mathbf{f}},\mathbf{x}_i\rangle^4}\left(1+\frac{\Gamma^2\log\frac{4m}{\delta}}{\bar{\mu}^2}\right) \tag{44}$$

Setting $u=\hat{\mathbf{f}}$, we have with probability at-least $1-\delta/2$:

$$h_2(X,\hat{\mathbf{f}}) \leq \|\Delta\|^2\log\left(\tfrac{4m}{\delta}\right)\left(1+\frac{\Gamma^2\log\frac{4m}{\delta}}{\bar{\mu}^2}\right)\frac{C}{m}\sum_{i=1}^{m}\frac{\mathbb{1}(|\langle\hat{\mathbf{f}},\mathbf{x}_i\rangle| > \bar{\mu})}{\langle\hat{\mathbf{f}},\mathbf{x}_i\rangle^2} \tag{45}$$

Bounding the quantity $\frac{1}{m}\sum_{i=1}^{m}\frac{\mathbb{1}(|\langle\hat{\mathbf{f}},\mathbf{x}_i\rangle|>\bar{\mu})}{\langle\hat{\mathbf{f}},\mathbf{x}_i\rangle^2}$ using Lemma 26. Therefore, with probability at-least $1-\delta/2$:

$$\frac{1}{m}\sum_{i=1}^{m}\frac{\mathbb{1}(|\langle\hat{\mathbf{f}},\mathbf{x}_i\rangle| > \bar{\mu})}{\langle\hat{\mathbf{f}},\mathbf{x}_i\rangle^2} \leq \frac{C}{\bar{\mu}\|\hat{\mathbf{f}}\|} \tag{46}$$

Combining this with Equation (45) via the union bound, we conclude that with probability at-least $1-\delta$:

$$h_2(X,\hat{\mathbf{f}}) \leq C\frac{\|\Delta\|^2\log\left(\frac{4m}{\delta}\right)}{\bar{\mu}\|\hat{\mathbf{f}}\|}\left(1+\frac{\Gamma^2\log\frac{4m}{\delta}}{\bar{\mu}^2}\right) \tag{47}$$

3. Now, assume that $u \perp \hat{\mathbf{f}}$ and $\|u\| = 1$. We begin with Equation (44). Using the union bound, we combine this with the high probability bound for the quantity $\sum_{i=1}^{m}\mathbb{1}(|\langle\hat{\mathbf{f}},\mathbf{x}_i\rangle| > \bar{\mu})\frac{\langle\mathbf{x}_i,u\rangle^2}{\langle\hat{\mathbf{f}},\mathbf{x}_i\rangle^4}$ from Lemma 26. Thus, we conclude that with probability at-least $1-\delta$, we have:

$$h_2(X,u) \leq C\frac{\|\Delta\|^2}{\bar{\mu}^3\|\hat{\mathbf{f}}\|}\log\left(\tfrac{4m}{\delta}\right)\left(1+\frac{\Gamma^2\log\left(\frac{4m}{\delta}\right)}{\bar{\mu}^2}\right)$$

■

E.3. Proof of Lemma 19

Proof [Proof of Lemma 19]

1. By definition, we note that:

$$\langle \bar{N}_3, \hat{\mathbf{f}} \rangle = \frac{1}{m} \sum_{i=1}^{m} \frac{\mathbb{1}(|\langle \hat{\mathbf{f}}, \mathbf{x}_i \rangle| > \bar{\mu})}{\langle \hat{\mathbf{f}}, \mathbf{x}_i \rangle^2} \langle \Delta, \mathbf{x}_i \rangle^2$$

Writing $\Delta = a\hat{\mathbf{f}} + \Delta^{\perp}$ with $\Delta^{\perp} \perp f$, we note that:

$$\langle \bar{N}_3, \hat{\mathbf{f}} \rangle \leq \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}(|\langle \hat{\mathbf{f}}, \mathbf{x}_i \rangle| > \bar{\mu}) \left[a^2 + \frac{\langle \Delta^{\perp}, \mathbf{x}_i \rangle^2}{\langle \hat{\mathbf{f}}, \mathbf{x}_i \rangle^2} \right] \tag{48}$$

Applying Lemma 26, using the fact that $\bar{\mu} \leq \|\hat{\mathbf{f}}\|$ and $\max(a^2\|\hat{\mathbf{f}}\|^2, \|\Delta^{\perp}\|^2) \leq \|\Delta\|^2$, we conclude the result.

2. First, we will write $\Delta = a\hat{\mathbf{f}} + \Delta^{\perp}$ where $\Delta^{\perp} \perp \hat{\mathbf{f}}$. $\langle \Delta, \mathbf{x}_i \rangle^2 = a^2 \langle \hat{\mathbf{f}}, \mathbf{x}_i \rangle^2 + 2a \langle \hat{\mathbf{f}}, \mathbf{x}_i \rangle \langle \Delta^{\perp}, \mathbf{x}_i \rangle + \langle \Delta^{\perp}, \mathbf{x}_i \rangle^2$.

   Therefore, we can write:

$$\begin{aligned} &Q\bar{N}_3 = \\ &\frac{1}{m} \sum_{i=1}^{m} \mathbb{1}(|\langle \hat{\mathbf{f}}, \mathbf{x}_i \rangle| > \bar{\mu}) \left[\frac{a^2 Q\mathbf{x}_i}{\langle \hat{\mathbf{f}}, \mathbf{x}_i \rangle} + \frac{2aQ\mathbf{x}_i \langle \Delta^{\perp}, \mathbf{x}_i \rangle}{\langle \hat{\mathbf{f}}, \mathbf{x}_i \rangle^2} + \frac{Q\mathbf{x}_i \langle \Delta^{\perp}, \mathbf{x}_i \rangle^2}{\langle \hat{\mathbf{f}}, \mathbf{x}_i \rangle^3} \right] \end{aligned} \tag{49}$$

   Now consider the first term in Equation (49). When conditioned on $\langle \hat{\mathbf{f}}, \mathbf{x}_i \rangle$ for $i \in [m]$, $Q\mathbf{x}_i$ are distributed as i.i.d. standard Gaussian in the space orthogonal to $\hat{\mathbf{f}}$. Therefore, by Gaussian concentration, we must have with probability at-least $1 - \delta/6$:

$$\left\| \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}(|\langle \hat{\mathbf{f}}, \mathbf{x}_i \rangle| > \bar{\mu}) \left[\frac{a^2 Q\mathbf{x}_i}{\langle \hat{\mathbf{f}}, \mathbf{x}_i \rangle} \right] \right\|^2 \leq \frac{C}{m^2} \sum_{i=1}^{m} \mathbb{1}(|\langle \hat{\mathbf{f}}, \mathbf{x}_i \rangle| > \bar{\mu}) \frac{a^4 (d + \log(6/\delta))}{\langle \hat{\mathbf{f}}, \mathbf{x}_i \rangle^2}$$

   Using Lemma 26 with the equation above via the union bound, we conclude that with probability at-least $1 - \delta/3$, we have:

$$\left\| \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}(|\langle \hat{\mathbf{f}}, \mathbf{x}_i \rangle| > \bar{\mu}) \left[\frac{a^2 Q\mathbf{x}_i}{\langle \hat{\mathbf{f}}, \mathbf{x}_i \rangle} \right] \right\| \leq \sqrt{\frac{C}{m} \frac{a^4 (d + \log(6/\delta))}{\bar{\mu}\|\hat{\mathbf{f}}\|}} \tag{50}$$

   Only in this proof, we define $\bar{Q} = Q - \frac{\Delta^{\perp}(\Delta^{\perp})^{\intercal}}{\|\Delta^{\perp}\|^2}$. Now consider the second term in Equation (49):

$$\begin{aligned} &\frac{1}{m} \sum_{i=1}^{m} \mathbb{1}(|\langle \hat{\mathbf{f}}, \mathbf{x}_i \rangle| > \bar{\mu}) \left[\frac{2aQ\mathbf{x}_i \langle \Delta^{\perp}, \mathbf{x}_i \rangle}{\langle \hat{\mathbf{f}}, \mathbf{x}_i \rangle^2} \right] \\ &= \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}(|\langle \hat{\mathbf{f}}, \mathbf{x}_i \rangle| > \bar{\mu}) \left[\frac{2a\bar{Q}\mathbf{x}_i \langle \Delta^{\perp}, \mathbf{x}_i \rangle}{\langle \hat{\mathbf{f}}, \mathbf{x}_i \rangle^2} + \frac{2a \langle \Delta^{\perp}, \mathbf{x}_i \rangle^2 \Delta^{\perp}}{\langle \hat{\mathbf{f}}, \mathbf{x}_i \rangle^2 \|\Delta^{\perp}\|^2} \right] \end{aligned} \tag{51}$$

The first term in Equation (51) can be bounded using Gaussian concentration and Lemma 26 just like in Equation (50). With probability at-least $1-\delta/12$, we have:

$$\left\|\frac{1}{m}\sum_{i=1}^{m}\mathbb{1}(|\langle\hat{\mathbf{f}},\mathbf{x}_i\rangle|>\bar{\mu})\left[\frac{2a\bar{Q}\mathbf{x}_i\langle\Delta^{\perp},\mathbf{x}_i\rangle}{\langle\hat{\mathbf{f}},\mathbf{x}_i\rangle^2}\right]\right\|\leq\sqrt{\frac{C}{m}\frac{a^2\|\Delta^{\perp}\|^2\left(d+\log(\frac{12}{\delta})\right)}{\bar{\mu}^3\|\hat{\mathbf{f}}\|}}\tag{52}$$

The second term in Equation (51) can be directly bounded using Lemma 26. Therefore, combining these results using the triangle inequality in Equation (51), we conclude that with probability at-least $1-\delta/6$, we must have:

$$\begin{aligned}&\left\|\frac{1}{m}\sum_{i=1}^{m}\mathbb{1}(|\langle\hat{\mathbf{f}},\mathbf{x}_i\rangle|>\bar{\mu})\left[\frac{2aQ\mathbf{x}_i\langle\Delta^{\perp},\mathbf{x}_i\rangle}{\langle\hat{\mathbf{f}},\mathbf{x}_i\rangle^2}\right]\right\|\\&\leq\sqrt{\frac{C}{m}\frac{a^2\|\Delta^{\perp}\|^2\left(d+\log(\frac{12}{\delta})\right)}{\bar{\mu}^3\|\hat{\mathbf{f}}\|}}+\frac{C|a|\|\Delta^{\perp}\|}{\bar{\mu}\|\hat{\mathbf{f}}\|}\end{aligned}\tag{53}$$

Now consider the third term in Equation (49).

$$\begin{aligned}&\frac{1}{m}\sum_{i=1}^{m}\mathbb{1}(|\langle\hat{\mathbf{f}},\mathbf{x}_i\rangle|>\bar{\mu})\left[\frac{Q\mathbf{x}_i\langle\Delta^{\perp},\mathbf{x}_i\rangle^2}{\langle\hat{\mathbf{f}},\mathbf{x}_i\rangle^3}\right]\\&=\frac{1}{m}\sum_{i=1}^{m}\mathbb{1}(|\langle\hat{\mathbf{f}},\mathbf{x}_i\rangle|>\bar{\mu})\left[\frac{\bar{Q}\mathbf{x}_i\langle\Delta^{\perp},\mathbf{x}_i\rangle^2}{\langle\hat{\mathbf{f}},\mathbf{x}_i\rangle^3}+\frac{\langle\Delta^{\perp},\mathbf{x}_i\rangle^3\Delta^{\perp}}{\langle\hat{\mathbf{f}},\mathbf{x}_i\rangle^3\|\Delta^{\perp}\|^2}\right]\end{aligned}\tag{54}$$

In Equation (54), we can bound the first term using Gaussian concentration again and then conclude that with probability at-least $1-\delta/12$, we must have:

$$\left\|\frac{1}{m}\sum_{i=1}^{m}\mathbb{1}(|\langle\hat{\mathbf{f}},\mathbf{x}_i\rangle|>\bar{\mu})\frac{\bar{Q}\mathbf{x}_i\langle\Delta^{\perp},\mathbf{x}_i\rangle^2}{\langle\hat{\mathbf{f}},\mathbf{x}_i\rangle^3}\right\|\leq\sqrt{\frac{C\|\Delta^{\perp}\|^4\log^3(\frac{12m}{\delta})d}{m\bar{\mu}^5\|\hat{\mathbf{f}}\|}}\tag{55}$$

Now consider the second term in Equation (54). In Lemma 27, we take the projector $R=Q$. Then, we conclude that for i.i.d. rademacher variables $\zeta_1,\dots,\zeta_m$ independent of everything else:

$$\begin{aligned}&\frac{1}{m}\sum_{i=1}^{m}\mathbb{1}(|\langle\hat{\mathbf{f}},\mathbf{x}_i\rangle|>\bar{\mu})\left[\frac{\langle\Delta^{\perp},\mathbf{x}_i\rangle^3\Delta^{\perp}}{\langle\hat{\mathbf{f}},\mathbf{x}_i\rangle^3\|\Delta^{\perp}\|^2}\right]\\&\overset{d}{=}\frac{1}{m}\sum_{i=1}^{m}\mathbb{1}(|\langle\hat{\mathbf{f}},\mathbf{x}_i\rangle|>\bar{\mu})\zeta_i\left[\frac{\langle\Delta^{\perp},\mathbf{x}_i\rangle^3\Delta^{\perp}}{\langle\hat{\mathbf{f}},\mathbf{x}_i\rangle^3\|\Delta^{\perp}\|^2}\right]\end{aligned}\tag{56}$$

Where $\overset{d}{=}$ denotes equality in distribution. By Azuma-Hoeffding inequality for rademacher random variables, we have with probability at-least $1-\delta/6$:

$$\left\|\frac{1}{m}\sum_{i=1}^{m}\mathbb{1}(|\langle\hat{\mathbf{f}},\mathbf{x}_i\rangle|>\bar{\mu})\zeta_i\left[\frac{\langle\Delta^{\perp},\mathbf{x}_i\rangle^3\Delta^{\perp}}{\langle\hat{\mathbf{f}},\mathbf{x}_i\rangle^3\|\Delta^{\perp}\|^2}\right]\right\|$$
$$\leq\sqrt{\frac{C\log\frac{12}{\delta}}{m^2}\sum_{i=1}^{m}\mathbb{1}(|\langle\hat{\mathbf{f}},\mathbf{x}_i\rangle|>\bar{\mu})\left[\frac{\langle\Delta^{\perp},\mathbf{x}_i\rangle^6}{\langle\hat{\mathbf{f}},\mathbf{x}_i\rangle^6\|\Delta^{\perp}\|^2}\right]}\leq\sqrt{\frac{C\|\Delta^{\perp}\|^4\log^4(\frac{12m}{\delta})}{m\bar{\mu}^5\|\hat{\mathbf{f}}\|}} \tag{57}$$

In the last step, we have used the high probability bound on $\sup_i|\langle\mathbf{x}_i,\Delta^{\perp}\rangle|$. Therefore, combining these inequalities, we conclude:

$$\left\|\frac{1}{m}\sum_{i=1}^{m}\mathbb{1}(|\langle\hat{\mathbf{f}},\mathbf{x}_i\rangle|>\bar{\mu})\left[\frac{Q\mathbf{x}_i\langle\Delta^{\perp},\mathbf{x}_i\rangle^2}{\langle\hat{\mathbf{f}},\mathbf{x}_i\rangle^3}\right]\right\|$$
$$=\sqrt{\frac{C\|\Delta^{\perp}\|^4\log^3(\frac{12m}{\delta})(d+\log(\frac{12m}{\delta}))}{m\bar{\mu}^5\|\hat{\mathbf{f}}\|}} \tag{58}$$

Combining Equations (50), (53) and (58) with Equation (49), we conclude:

$$\|QN_3\|\leq\sqrt{\frac{C}{m}\frac{a^4\,(d+\log(6/\delta))}{\bar{\mu}\|\hat{\mathbf{f}}\|}}+\sqrt{\frac{C}{m}\frac{a^2\|\Delta^{\perp}\|^2\,\left(d+\log\left(\frac{12}{\delta}\right)\right)}{\bar{\mu}^3\|\hat{\mathbf{f}}\|}}$$
$$+\frac{C|a|\|\Delta^{\perp}\|}{\bar{\mu}\|\hat{\mathbf{f}}\|}+\sqrt{\frac{C\|\Delta^{\perp}\|^4\log^3(\frac{12m}{\delta})(d+\log(\frac{12m}{\delta}))}{m\bar{\mu}^5\|\hat{\mathbf{f}}\|}}$$
$$\leq\frac{C\|\Delta\|^2}{\bar{\mu}\|\hat{\mathbf{f}}\|^2}+\sqrt{\frac{C\|\Delta\|^4\log^3(\frac{12m}{\delta})(d+\log(\frac{12m}{\delta}))}{m\bar{\mu}^5\|\hat{\mathbf{f}}\|}} \tag{59}$$

The last step follows by the fact that $|a|\leq\frac{\|\Delta\|}{\|\hat{\mathbf{f}}\|}$ and $\|\Delta^{\perp}\|\leq\|\Delta\|$ and the fact that $\bar{\mu}\leq\|\hat{\mathbf{f}}\|$

■

E.4. Proof of Lemma 21

Proof [Proof of Lemma 21]

1. From Lemma 20, we conclude that with probability at-least $1-\frac{\delta}{2}$, we must have:

$$\frac{1}{m}\sum_{i=1}^{m}\mathbb{1}(|\langle\hat{\mathbf{f}},\mathbf{x}_i\rangle|>\bar{\mu})\leq\langle\hat{\mathbf{f}},H(\mathbf{f})\hat{\mathbf{f}}\rangle\leq\frac{4}{m}\sum_{i=1}^{m}\mathbb{1}(|\langle\hat{\mathbf{f}},\mathbf{x}_i\rangle|>\bar{\mu}) \tag{60}$$

Now, note that $\sum_{i=1}^{m}\mathbb{1}(|\langle\hat{\mathbf{f}},\mathbf{x}_i\rangle|>\bar{\mu})$ is distributed as $\mathsf{Bin}(m,\mathbb{P}(|\langle\hat{\mathbf{f}},\mathbf{x}_i\rangle|>\bar{\mu}))$. Applying Lemma 10, we conclude that with probability at-least $1-\exp\Big(-cm\mathbb{P}(|\langle\hat{\mathbf{f}},\mathbf{x}_i\rangle|>\bar{\mu})\Big)$, we must have:

$$\tfrac{\mathbb{P}(|\langle\hat{\mathbf{f}},\mathbf{x}_i\rangle|>\bar{\mu})}{2} \leq \langle \hat{\mathbf{f}}, H(\mathbf{f})\hat{\mathbf{f}}\rangle \leq 8\mathbb{P}(|\langle\hat{\mathbf{f}},\mathbf{x}_i\rangle| > \bar{\mu}) \tag{61}$$

Now, using the fact that $\frac{1}{4} \leq \mathbb{P}(|\langle\hat{\mathbf{f}},\mathbf{x}_i\rangle| > \bar{\mu}) \leq 1$ (since $\bar{\mu} < \|\hat{\mathbf{f}}\|$), we conclude the result.

2. Only in this proof, suppose that $\mathbf{f} + \mathbf{f}^* - 2\hat{\mathbf{f}} = a\frac{\hat{\mathbf{f}}}{\|\hat{\mathbf{f}}\|} + b\hat{\mathbf{f}}^\perp$ such that $\|\hat{\mathbf{f}}^\perp\| = 1$ and $\hat{\mathbf{f}}^\perp \perp \hat{\mathbf{f}}$.

   Note that

$$\begin{aligned}\langle u, H(\mathbf{f})\hat{\mathbf{f}}\rangle &= \frac{1}{m}\sum_{i=1}^m \mathbb{1}(|\langle\hat{\mathbf{f}},\mathbf{x}_i\rangle| \geq \bar{\mu})\frac{\langle \mathbf{f}+\mathbf{f}^*,\mathbf{x}_i\rangle\langle\mathbf{x}_i,u\rangle}{\langle\hat{\mathbf{f}},\mathbf{x}_i\rangle^2} \\ &= \frac{1}{m}\sum_{i=1}^m \mathbb{1}(|\langle\hat{\mathbf{f}},\mathbf{x}_i\rangle| \geq \bar{\mu})\left[\frac{\langle \mathbf{f}+\mathbf{f}^*-2\hat{\mathbf{f}},\mathbf{x}_i\rangle\langle\mathbf{x}_i,u\rangle}{\langle\hat{\mathbf{f}},\mathbf{x}_i\rangle^2} + 2\frac{\langle\mathbf{x}_i,u\rangle}{\langle\hat{\mathbf{f}},\mathbf{x}_i\rangle}\right] \\ &= \frac{1}{m}\sum_{i=1}^m \mathbb{1}(|\langle\hat{\mathbf{f}},\mathbf{x}_i\rangle| \geq \bar{\mu})\left[b\frac{\langle \hat{\mathbf{f}}^\perp,\mathbf{x}_i\rangle\langle\mathbf{x}_i,u\rangle}{\langle\hat{\mathbf{f}},\mathbf{x}_i\rangle^2} + \left(2+\tfrac{a}{\|\hat{\mathbf{f}}\|}\right)\frac{\langle\mathbf{x}_i,u\rangle}{\langle\hat{\mathbf{f}},\mathbf{x}_i\rangle}\right] \end{aligned} \tag{62}$$

   Now, conditioned on $\langle\hat{\mathbf{f}},\mathbf{x}_i\rangle$ for $i \in [m]$, we must have $\langle u,\mathbf{x}_i\rangle$ for $i \in [m]$ to be i.i.d. standard Gaussians and the term $\frac{1}{m}\sum_{i=1}^m \mathbb{1}(|\langle\hat{\mathbf{f}},\mathbf{x}_i\rangle| \geq \bar{\mu})\frac{\langle\mathbf{x}_i,u\rangle}{\langle\hat{\mathbf{f}},\mathbf{x}_i\rangle}$ is a gaussian with variance $\frac{L_2(X)}{m}$ (L_2 as defined in Lemma 26). Therefore, applying Gaussian concentration and Lemma 26, we conclude that with probability at-least $1-\delta/2$, we must have:

$$\left|\frac{1}{m}\sum_{i=1}^m \mathbb{1}(|\langle\hat{\mathbf{f}},\mathbf{x}_i\rangle| \geq \bar{\mu})\frac{\langle\mathbf{x}_i,u\rangle}{\langle\hat{\mathbf{f}},\mathbf{x}_i\rangle}\right| \leq \sqrt{\frac{C\log(4/\delta)}{m\bar{\mu}\|\hat{\mathbf{f}}\|}} \tag{63}$$

   Applying Cauchy Schwarz inequality and Lemma 26, we conclude that with probability $1-\frac{\delta}{2}$ (with L_1 as defined in Lemma 26):

$$\begin{aligned}\left|\frac{1}{m}\sum_{i=1}^m \mathbb{1}(|\langle\hat{\mathbf{f}},\mathbf{x}_i\rangle| \geq \bar{\mu})\frac{\langle \hat{\mathbf{f}}^\perp,\mathbf{x}_i\rangle\langle\mathbf{x}_i,u\rangle}{\langle\hat{\mathbf{f}},\mathbf{x}_i\rangle^2}\right| &\leq \sqrt{L_1(X,u)L_1(X,\hat{\mathbf{f}}^\perp)} \\ &\leq \frac{C}{\bar{\mu}\|\hat{\mathbf{f}}\|}\end{aligned} \tag{64}$$

   The result follows by combining the above equations with the union bound along with the fact that $|a|,|b| \leq 2\Gamma$

3. Now let $u \perp \hat{\mathbf{f}}^\perp$ along with the condition $u \perp \hat{\mathbf{f}}$. We will now find a finer bound than Equation (64). Notice that conditioned on the random variables $\langle u,\mathbf{x}_i\rangle, \langle\mathbf{x}_i,\hat{\mathbf{f}}\rangle$ for $i \in [m]$, the random variables $\langle\hat{\mathbf{f}}^\perp,\mathbf{x}_i\rangle$ are i.i.d. standard Gaussians. Therefore, under this conditioning, we have $\frac{1}{m}\sum_{i=1}^m \mathbb{1}(|\langle\hat{\mathbf{f}},\mathbf{x}_i\rangle| \geq \bar{\mu})\frac{\langle\hat{\mathbf{f}}^\perp,\mathbf{x}_i\rangle\langle\mathbf{x}_i,u\rangle}{\langle\hat{\mathbf{f}},\mathbf{x}_i\rangle^2}$ is a zero mean Gaussian with variance $\frac{L_1(X,u,2)}{m}$ (as defined in Lemma 26)

   We conclude that with probability at-least $1-\delta$, we have:

$$\left|\frac{1}{m}\sum_{i=1}^m \mathbb{1}(|\langle\hat{\mathbf{f}},\mathbf{x}_i\rangle| \geq \bar{\mu})\frac{\langle \hat{\mathbf{f}}^\perp,\mathbf{x}_i\rangle\langle\mathbf{x}_i,u\rangle}{\langle\hat{\mathbf{f}},\mathbf{x}_i\rangle^2}\right| \leq \sqrt{\frac{C\log(\frac{1}{\delta})}{m\bar{\mu}^3\|\hat{\mathbf{f}}\|}}$$

Rest of the proof follows similar to item 2.

4. By similar observations as in item 1, we apply Lemma 20 to conclude that with probability at-least $1 - \frac{\delta}{2}$:

$$\frac{1}{m}\sum_{i=1}^{m}\mathbb{1}(|\langle \hat{\mathbf{f}}, \mathbf{x}_i\rangle| > \bar{\mu})\frac{\langle \mathbf{x}_i, u\rangle^2}{\langle \mathbf{x}_i, \hat{\mathbf{f}}\rangle^2} \leq \langle u, H(\mathbf{f})u\rangle \leq \frac{4}{m}\sum_{i=1}^{m}\mathbb{1}(|\langle \hat{\mathbf{f}}, \mathbf{x}_i\rangle| > \bar{\mu})\frac{\langle \mathbf{x}_i, u\rangle^2}{\langle \mathbf{x}_i, \hat{\mathbf{f}}\rangle^2} \tag{65}$$

The result follows by an application of Lemma 26 along with the union bound.

5. Only in this proof we consider an orthonormal basis $v_1, \dots, v_d$ for $\mathbb{R}^d$ such that $v_1 = \frac{\hat{\mathbf{f}}}{\|\hat{\mathbf{f}}\|}$ and $v_2 = \hat{\mathbf{f}}^{\perp}$

$$\|QH(\mathbf{f})\hat{\mathbf{f}}\|^2 = \sum_{i=2}^{d}\langle v_i, H(\mathbf{f})\hat{\mathbf{f}}\rangle^2 \tag{66}$$

From item 2, we conclude that with probability at-least $1 - \delta/d$:

$$\langle v_2, H(\mathbf{f})\hat{\mathbf{f}}\rangle^2 \leq \frac{C\Gamma^2}{\bar{\mu}^2\|\hat{\mathbf{f}}\|^2} + \left(1 + \tfrac{\Gamma}{\|\hat{\mathbf{f}}\|}\right)^2 \frac{C\log(4d/\delta)}{m\bar{\mu}\|\hat{\mathbf{f}}\|}$$

From item 3, we conclude that for $i \geq 3$, with probability at-least $1 - \delta/d$, we must have:

$$\langle v_i, H(\mathbf{f})\hat{\mathbf{f}}\rangle^2 \leq \frac{C\Gamma^2\log\left(\frac{4d}{\delta}\right)}{m\bar{\mu}^3\|\hat{\mathbf{f}}\|} + \left(1 + \tfrac{\Gamma}{\|\hat{\mathbf{f}}\|}\right)^2 \frac{C\log(4d/\delta)}{m\bar{\mu}\|\hat{\mathbf{f}}\|}$$

Using these in Equation (66) with the union bound, we conclude with probability at least $1 - \delta$, we have:

$$\langle \hat{\mathbf{f}}, H(\mathbf{f})^2\hat{\mathbf{f}}\rangle \leq \frac{C_1}{\|\hat{\mathbf{f}}\|^2} + \left(1 + \tfrac{\Gamma}{\|\hat{\mathbf{f}}\|}\right)^2 \frac{Cd\log(4d/\delta)}{m\bar{\mu}\|\hat{\mathbf{f}}\|} + \frac{Cd\Gamma^2\log^2\left(\frac{4m}{\delta}\right)}{m\bar{\mu}^3\|\hat{\mathbf{f}}\|} + \frac{C\Gamma^2}{\bar{\mu}^2\|\hat{\mathbf{f}}\|^2}$$

6. Consider $QH_t u$. Similar to item 2, take $\mathbf{f} + \mathbf{f}^* - 2\hat{\mathbf{f}} = a\frac{\hat{\mathbf{f}}}{\|\hat{\mathbf{f}}\|} + b\hat{\mathbf{f}}^{\perp}$:

$$\begin{aligned} QH(f)u &= \frac{1}{m}\sum_{i=1}^{m}\frac{\mathbb{1}(|\langle \hat{\mathbf{f}}, \mathbf{x}_i\rangle| \geq \bar{\mu})}{\langle \mathbf{x}_i, \hat{\mathbf{f}}\rangle^3}\langle u, \mathbf{x}_i\rangle Q\mathbf{x}_i\langle \mathbf{f} + \mathbf{f}^*, \mathbf{x}_i\rangle \\ &= \frac{2}{m}\sum_{i=1}^{m}\frac{\mathbb{1}(|\langle \hat{\mathbf{f}}, \mathbf{x}_i\rangle| \geq \bar{\mu})}{\langle \mathbf{x}_i, \hat{\mathbf{f}}\rangle^2}\langle u, \mathbf{x}_i\rangle Q\mathbf{x}_i \\ &\quad + \frac{1}{m}\sum_{i=1}^{m}\frac{\mathbb{1}(|\langle \hat{\mathbf{f}}, \mathbf{x}_i\rangle| \geq \bar{\mu})}{\langle \mathbf{x}_i, \hat{\mathbf{f}}\rangle^3}\langle u, \mathbf{x}_i\rangle\langle f + \mathbf{f}^* - 2\hat{\mathbf{f}}, \mathbf{x}_i\rangle Q\mathbf{x}_i \\ &= \frac{1}{m}\sum_{i=1}^{m}(2 + \tfrac{a}{\|\hat{\mathbf{f}}\|})\frac{\mathbb{1}(|\langle \hat{\mathbf{f}}, \mathbf{x}_i\rangle| \geq \bar{\mu})}{\langle \mathbf{x}_i, \hat{\mathbf{f}}\rangle^2}\langle u, \mathbf{x}_i\rangle Q\mathbf{x}_i \\ &\quad + \frac{b}{m}\sum_{i=1}^{m}\frac{\mathbb{1}(|\langle \hat{\mathbf{f}}, \mathbf{x}_i\rangle| \geq \bar{\mu})}{\langle \mathbf{x}_i, \hat{\mathbf{f}}\rangle^3}\langle u, \mathbf{x}_i\rangle\langle \hat{\mathbf{f}}^{\perp}, \mathbf{x}_i\rangle Q\mathbf{x}_i \end{aligned} \tag{67}$$

In Lemma 27, let R be the projector onto the orthogonal space to $\hat{\mathbf{f}}$. Then for $\zeta_1, \dots, \zeta_m$ i.i.d. rademacher variables independent of everything else, we have:

$$\begin{aligned}
&\frac{1}{m}\sum_{i=1}^{m} \frac{\mathbb{1}(|\langle \hat{\mathbf{f}}, \mathbf{x}_i\rangle| \geq \bar{\mu})}{\langle \mathbf{x}_i, \hat{\mathbf{f}}\rangle^3} \langle u, \mathbf{x}_i\rangle \langle \hat{\mathbf{f}}^{\perp}, \mathbf{x}_i\rangle Q\mathbf{x}_i \\
&\overset{d}{=} \frac{1}{m}\sum_{i=1}^{m} \zeta_i \frac{\mathbb{1}(|\langle \hat{\mathbf{f}}, \mathbf{x}_i\rangle| \geq \bar{\mu})}{\langle \mathbf{x}_i, \hat{\mathbf{f}}\rangle^3} \langle u, \mathbf{x}_i\rangle \langle \hat{\mathbf{f}}^{\perp}, \mathbf{x}_i\rangle Q\mathbf{x}_i
\end{aligned} \tag{68}$$

Where $\overset{d}{=}$ denotes equality in distribution. By rademacher concentration, we have that with probability at-least $1 - \delta/2$, we have:

$$\begin{aligned}
&\left\|\frac{1}{m}\sum_{i=1}^{m} \zeta_i \frac{\mathbb{1}(|\langle \hat{\mathbf{f}}, \mathbf{x}_i\rangle| \geq \bar{\mu})}{\langle \mathbf{x}_i, \hat{\mathbf{f}}\rangle^3} \langle u, \mathbf{x}_i\rangle \langle \hat{\mathbf{f}}^{\perp}, \mathbf{x}_i\rangle Q\mathbf{x}_i\right\|^2 \\
&\leq \frac{C\log\left(\frac{2d}{\delta}\right)}{m^2} \sum_{i=1}^{m} \frac{\mathbb{1}(|\langle \hat{\mathbf{f}}, \mathbf{x}_i\rangle| \geq \bar{\mu})}{\langle \mathbf{x}_i, \hat{\mathbf{f}}\rangle^6} \langle u, \mathbf{x}_i\rangle^2 \langle \hat{\mathbf{f}}^{\perp}, \mathbf{x}_i\rangle^2 \|Q\mathbf{x}_i\|^2
\end{aligned} \tag{69}$$

With probability at-least $1 - \frac{3\delta}{8}$, we have:

$$\sup_{i\in[m]} \langle u, \mathbf{x}_i\rangle^2 \langle \hat{\mathbf{f}}^{\perp}, \mathbf{x}_i\rangle^2 \|Q\mathbf{x}_i\|^2 \leq C\log\left(\tfrac{8m}{\delta}\right)^3 d$$

Using this, along with an application of Lemma 26, we conclude that with probability at-least $1 - \delta/2$:

$$\begin{aligned}
&\left\|\frac{1}{m}\sum_{i=1}^{m} \zeta_i \frac{\mathbb{1}(|\langle \hat{\mathbf{f}}, \mathbf{x}_i\rangle| \geq \bar{\mu})}{\langle \mathbf{x}_i, \hat{\mathbf{f}}\rangle^3} \langle u, \mathbf{x}_i\rangle \langle \hat{\mathbf{f}}^{\perp}, \mathbf{x}_i\rangle Q\mathbf{x}_i\right\|^2 \\
&\leq \frac{Cd\log\left(\frac{4d}{\delta}\right)\log^3\left(\frac{8m}{\delta}\right)}{m\bar{\mu}^5\|\hat{\mathbf{f}}\|}
\end{aligned} \tag{70}$$

Now consider $v_1, \dots, v_{d-1}$ to be any orthonormal basis for the span of Q such that $v_1 = u$.

$$\begin{aligned}
&\|\frac{1}{m}\sum_{i=1}^{m} \frac{\mathbb{1}(|\langle \hat{\mathbf{f}}, \mathbf{x}_i\rangle| \geq \bar{\mu})}{\langle \mathbf{x}_i, \hat{\mathbf{f}}\rangle^2} \langle u, \mathbf{x}_i\rangle Q\mathbf{x}_i\|^2 \\
&= \left\|\frac{1}{m}\sum_{i=1}^{m} \frac{\mathbb{1}(|\langle \hat{\mathbf{f}}, \mathbf{x}_i\rangle| \geq \bar{\mu})}{\langle \mathbf{x}_i, \hat{\mathbf{f}}\rangle^2} \langle u, \mathbf{x}_i\rangle^2\right\|^2 \\
&\quad + \sum_{j=2}^{d-1}\left\|\frac{1}{m}\sum_{i=1}^{m} \frac{\mathbb{1}(|\langle \hat{\mathbf{f}}, \mathbf{x}_i\rangle| \geq \bar{\mu})}{\langle \mathbf{x}_i, \hat{\mathbf{f}}\rangle^2} \langle u, \mathbf{x}_i\rangle \langle v_j, \mathbf{x}_i\rangle\right\|^2 \\
&\leq \frac{C}{\bar{\mu}^2\|\hat{\mathbf{f}}\|^2} + \sum_{j=2}^{d-1}\left\|\frac{1}{m}\sum_{i=1}^{m} \frac{\mathbb{1}(|\langle \hat{\mathbf{f}}, \mathbf{x}_i\rangle| \geq \bar{\mu})}{\langle \mathbf{x}_i, \hat{\mathbf{f}}\rangle^2} \langle u, \mathbf{x}_i\rangle \langle v_j, \mathbf{x}_i\rangle\right\|^2
\end{aligned} \tag{71}$$

We have used Lemma 26 in the second step. The second term is exactly same as that used in item 3 (by replacing $\hat{\mathbf{f}}^{\perp}$ with v_j). Therefore, with probability at-least $1-\delta$, we must have:

$$\left\|\frac{1}{m}\sum_{i=1}^{m}\frac{\mathbb{1}(|\langle\hat{\mathbf{f}},\mathbf{x}_i\rangle|\geq\bar{\mu})}{\langle\mathbf{x}_i,\hat{\mathbf{f}}\rangle^2}\langle u,\mathbf{x}_i\rangle Q\mathbf{x}_i\right\|^2\leq\frac{C}{\bar{\mu}^2\|\hat{\mathbf{f}}\|^2}+\frac{Cd\log^2(\frac{4md}{\delta})}{m\bar{\mu}^3\|\hat{\mathbf{f}}\|}\tag{72}$$

Therefore, using Equations (68) (70), we bound the first term in Equation (67). We bound the second term with Equation (72). Combining these two using the union bound, we conclude:

$$\|QH(f)u\|\leq\left(1+\tfrac{\Gamma}{\|\hat{\mathbf{f}}\|}\right)\frac{C}{\|\hat{\mathbf{f}}\|\bar{\mu}}+\sqrt{\frac{Cd\log^2(\frac{4md}{\delta})}{m\bar{\mu}^3\|\hat{\mathbf{f}}\|}}+\sqrt{\frac{Cd\Gamma^2\log(\frac{4d}{\delta})\log^3(\frac{8m}{\delta})}{m\bar{\mu}^5\|\hat{\mathbf{f}}\|}}$$

■

E.5. Proof of Lemma 22

Proof [Proof of Lemma 22]

1. Consider the following sequence of inequalities:

$$\begin{aligned}\left\|P(I-\alpha_0H_t(\mathbf{f}_t))P\frac{\hat{\mathbf{f}}}{\|\hat{\mathbf{f}}\|}\right\|^2&=\frac{1}{\|\hat{\mathbf{f}}\|^2}\langle P(I-\alpha_0H_t(\mathbf{f}_t))\hat{\mathbf{f}},P(I-\alpha_0H_t(\mathbf{f}_t))\hat{\mathbf{f}}\rangle\\&=1-2\frac{\alpha_0}{\|\hat{\mathbf{f}}\|^2}\langle\hat{\mathbf{f}},H_t(\mathbf{f}_t)\hat{\mathbf{f}}\rangle+\frac{\alpha_0^2}{\|\hat{\mathbf{f}}\|^2}\hat{\mathbf{f}}^{\intercal}H_tP^2H_t(\mathbf{f}_t)\hat{\mathbf{f}}\\&\leq1-2\frac{\alpha_0}{\|\hat{\mathbf{f}}\|^2}\langle\hat{\mathbf{f}},H_t(\mathbf{f}_t)\hat{\mathbf{f}}\rangle+\frac{\alpha_0^2}{\|\hat{\mathbf{f}}\|^4}\langle\hat{\mathbf{f}}^{\intercal},H_t(\mathbf{f}_t)\hat{\mathbf{f}}\rangle^2\end{aligned}\tag{73}$$

In the last step, we use the fact that $P=\frac{\hat{\mathbf{f}}\hat{\mathbf{f}}^{\intercal}}{\|\hat{\mathbf{f}}\|^2}$. By Lemma 21 items 1 and 5, we must have with probability at-least $1-\delta$, the following hold: $C_0\leq\hat{\mathbf{f}}^{\intercal}H_t(\mathbf{f}_t)\hat{\mathbf{f}}\leq C_1$. By the assumption on the step size and Equation (73), we conclude:

$$\left\|P(I-\alpha_0H_t(\mathbf{f}_t))P\frac{\hat{\mathbf{f}}}{\|\hat{\mathbf{f}}\|}\right\|^2\leq\left(1-\tfrac{c\alpha_0}{\|\hat{\mathbf{f}}\|^2}\right)\tag{74}$$

2. Note that $\|PH_t(\mathbf{f}_t)Qu_t\|=\left|\langle\frac{\hat{\mathbf{f}}}{\|\hat{\mathbf{f}}\|},H_t(\mathbf{f}_t)u_t\rangle\right|$. From item 2 of Lemma 21, we conclude that with probability at-least $1-\delta$:

$$\begin{aligned}|\langle\tfrac{\hat{\mathbf{f}}}{\|\hat{\mathbf{f}}\|},H_t(\mathbf{f}_t)u_t\rangle|&\leq\frac{C\Gamma_t}{\bar{\mu}\|\hat{\mathbf{f}}\|^2}+\left(1+\tfrac{\Gamma_t}{\|\hat{\mathbf{f}}\|}\right)\sqrt{\frac{C\log(1/\delta)}{m\bar{\mu}\|\hat{\mathbf{f}}\|^3}}\\&\leq C\left[\frac{\Gamma_t}{\bar{\mu}\|\hat{\mathbf{f}}\|^2}+\sqrt{\frac{\log(1/\delta)}{m\bar{\mu}\|\hat{\mathbf{f}}\|^3}}\right]\end{aligned}\tag{75}$$

In the last step, we have used the assumption to bound $\Gamma_t\leq\bar{\mu}<\|\hat{\mathbf{f}}\|$.

3. We decompose the noise N_t into $N_{t,1}, N_{t,2}$ and $N_{t,3}$ as in Equation (25) to give us $N_t = N_{t,1} + N_{t,2} + N_{t,3}$. Then, by definition of $\bar{a}_t$, and the bounds in Lemmata 17, 18 and 19 we must have:

$$\begin{aligned}
|\bar{a}_t| &\leq \frac{1}{\|\hat{\mathbf{f}}\|}|\langle N_{t,1}, \hat{\mathbf{f}}\rangle| + \frac{1}{\|\hat{\mathbf{f}}\|}|\langle N_{t,2}, \hat{\mathbf{f}}\rangle| + \frac{1}{\|\hat{\mathbf{f}}\|}|\langle N_{t,3}, \hat{\mathbf{f}}\rangle| \\
&\leq \frac{C\log\left(\frac{1}{\delta}\right)}{\|\hat{\mathbf{f}}\|}\sqrt{\frac{1}{m}\left[1 + \frac{\Gamma_t^4}{\bar{\mu}^4}\log^2\frac{m}{\delta}\right]} + C\sqrt{\frac{\|\Delta\|^2\log^2(\frac{2m}{\delta})}{m\bar{\mu}^2\|\hat{\mathbf{f}}\|^2}\left(1 + \frac{\Gamma_t^2\log\frac{2m}{\delta}}{\bar{\mu}^2}\right)} \\
&\quad + C\frac{\|\Delta\|^2}{\bar{\mu}\|\hat{\mathbf{f}}\|^2} \\
&\leq C\left[\frac{\log\left(\frac{1}{\delta}\right)}{\|\hat{\mathbf{f}}\|\sqrt{m}} + \frac{\|\Delta\|^2}{\bar{\mu}\|\hat{\mathbf{f}}\|^2}\right]
\end{aligned} \tag{76}$$

In the last step we have used the bound $C\Gamma_t \leq \bar{\mu}\log\left(\frac{m}{\delta}\right)$ and $C\|\Delta\|\log\left(\frac{m}{\delta}\right) \leq \bar{\mu}$.

■

E.6. Proof of Lemma 23

Proof [Proof of Lemma 23]

1. Notice that $\|Q(I - \alpha_1 H_t(\mathbf{f}_t))Qu_t\|^2 = 1 - 2\alpha_1\langle u_t, H_t(\mathbf{f}_t)u_t\rangle + \alpha_1^2\|QH_t(\mathbf{f}_t)u_t\|^2$

   From item 2 of Lemma 21, we have with probability at-least $1 - \frac{\delta}{2}$: $\frac{C_0}{\|\hat{\mathbf{f}}\|\bar{\mu}} \leq \langle u_t, H_t(\mathbf{f}_t)u_t\rangle \leq \frac{C_1}{\|\hat{\mathbf{f}}\|\bar{\mu}}$. By item 6 of Lemma 21, we have with probability at-least $1 - \frac{\delta}{2}$:

$$\begin{aligned}
\|QH_t(\mathbf{f}_t)u_t\|^2 &\leq \left(1 + \frac{\Gamma_t}{\|\hat{\mathbf{f}}\|}\right)^2\frac{C}{\|\hat{\mathbf{f}}\|^2\bar{\mu}^2} + \frac{Cd\log^2(\frac{4md}{\delta})}{m\bar{\mu}^3\|\hat{\mathbf{f}}\|} + \frac{Cd\Gamma_t^2\log\left(\frac{4d}{\delta}\right)\log^3(\frac{8m}{\delta})}{m\bar{\mu}^5\|\hat{\mathbf{f}}\|} \\
&\leq \frac{C}{\|\hat{\mathbf{f}}\|^2\bar{\mu}^2}
\end{aligned} \tag{77}$$

   In the second step, we have used the bounds on m and Γ_t. Now, picking $\alpha_1 \leq c\bar{\mu}\|\hat{\mathbf{f}}\|$ for small enough c, we conclude the result.

2. Consider $\frac{1}{\|\hat{\mathbf{f}}\|^2}\|QH_t(\mathbf{f}_t)\hat{\mathbf{f}}\|^2$. By item 5 of Lemma 21, we conclude the result and using the bounds on the parameters in the assumptions.

3. $$|\bar{b}_t| \leq \|QN_{t,1}\| + \|QN_{t,2}\| + \|QN_{t,3}\|$$

   Consulting Lemmata 17, 18 and 19, we conclude $\|QN_{t,1}\| \leq C\log\left(\frac{d}{\delta}\right)\sqrt{\frac{d}{m\bar{\mu}\|f\|}}$. $\|QN_{t,2}\| \leq C\sqrt{\frac{d\log\left(\frac{1}{\delta}\right)}{m\bar{\mu}\|f\|}}$. $\|QN_{t,3}\| \leq \frac{C\|\Delta\|^2}{\bar{\mu}\|\hat{\mathbf{f}}\|^2} + \sqrt{\frac{Cd}{m\bar{\mu}\|\hat{\mathbf{f}}\|}}$, from which we conclude the result.

■

Appendix F. Proof of Concentration Lemmas

F.1. Proof of Lemma 25

Proof [Proof of Lemma 25] From (Vershynin, 2010, Section 5.2.4), we conclude that $\mathbb{E}\exp(\lambda z_i) \leq \exp\bigl(c_0\lambda^2\bigr)$ for every $\lambda \in [-c, c]$ for some $c > 0$. Therefore, whenever $|\lambda| \leq \frac{c}{\max_i |v_i|}$ we have:

$$\mathbb{E}\exp\left(\lambda \sum_{i=1}^{n} v_i z_i\right) \leq \exp\left(c_0\lambda^2 \sum_{i=1}^{n} v_i^2\right)$$

The Chernoff bound with $\lambda = \frac{c}{\|v\|}$ implies:

$$\mathbb{P}\left(|\sum_{i=1}^{m} z_i v_i| > t\right) \leq C_0 \exp\left(-C_1 \frac{t}{\|v\|}\right)$$

■

F.2. Proof of Lemma 26

Proof [Proof of Lemma 26]

1. Note that $\langle X_i, \hat{\mathbf{f}}\rangle \sim \mathcal{N}(0, \|\hat{\mathbf{f}}\|^2)$. For integers $r - 1 \geq i \geq 0$, define sets $A_i = \{x \in \mathbb{R} : 2^i\bar{\mu} \leq |x| < 2^{i+1}\bar{\mu}\}$ and $A_r = \{x \in \mathbb{R} : |x| \geq 2^r\bar{\mu}\}$. In Lemma 24, we take $\xi_1, \dots, \xi_n$ to be $\langle X_1, \hat{\mathbf{f}}\rangle, \dots, \langle X_m, \hat{\mathbf{f}}\rangle$. Taking $H_i, \mathcal{H}_i$ to be the frequency as defined in Lemma 24, with $r = \lceil \log_2(\frac{\|\hat{\mathbf{f}}\|}{\bar{\mu}})\rceil$ we conclude that the following holds almost surely:

$$\frac{1}{m}\sum_{i=1}^{m} \mathbb{1}(|\langle \hat{\mathbf{f}}, X_i\rangle| > \bar{\mu})\frac{\langle X_i, u\rangle^2}{\langle \hat{\mathbf{f}}, X_i\rangle^{2l}} \leq \frac{1}{m}\sum_{i=0}^{r} \frac{2^{-2li}}{\bar{\mu}^{2l}} \sum_{j\in\mathcal{H}_i} \langle X_j, u\rangle^2 \tag{78}$$

Note that since $u \perp \hat{\mathbf{f}}$, we must have $\langle X_j, u\rangle$ to be independent of $\mathcal{H}_i$ and is distributed as $\mathcal{N}(0, 1)$. By Lemma 25, we conclude:

$$\mathbb{P}\left(\sum_{j\in\mathcal{H}_i} \langle X_j, u\rangle^2 > H_i + C_0\sqrt{H_i t}\right) \leq \exp(-t) \tag{79}$$

Therefore, with probability at-least $1 - \frac{\delta}{2}$, we must have for every $i = 0, \dots, r$:

$$\sum_{j\in\mathcal{H}_i} \langle X_j, u\rangle^2 \leq H_i + C_0\sqrt{H_i}\log\bigl(\tfrac{2r}{\delta}\bigr) \tag{80}$$

Therefore, combining Equations (78) and (80), we conclude that with probability at-least $1 - \frac{\delta}{2}$, we have:

$$L_1(X) \leq \frac{1}{m}\sum_{i=0}^{r} \frac{2^{-2li}}{\bar{\mu}^{2l}} \left(H_i + C_0\sqrt{H_i}\log\bigl(\tfrac{2r}{\delta}\bigr)\right) \tag{81}$$

Let p_i be the probabilities as defined in Lemma 24. Then, we must have for $i = 0, \dots, r$, $c2^i \frac{\bar{\mu}}{\|\hat{\mathbf{f}}\|} \leq p_i \leq C2^i \frac{\bar{\mu}}{\|\hat{\mathbf{f}}\|}$. Applying Lemma 24, we conclude that with probability at-least $1 - \frac{\delta}{2}$, we must have $i = 0, \dots, r$:

$$H_i \leq C_1 \max \left(\log\left(\tfrac{2r}{\delta}\right), \frac{2^i \bar{\mu} m}{\|\hat{\mathbf{f}}\|} \right) = \frac{C_1 2^i \bar{\mu} m}{\|\hat{\mathbf{f}}\|} \tag{82}$$

Here, we have used the assumption that $m \geq C \frac{\|\hat{\mathbf{f}}\|}{\bar{\mu}} \log^2(\frac{2r}{\delta})$. Under the same assumption and the same event in Equation (82), we conclude that $\sqrt{H_i} \log\left(\frac{2r}{\delta}\right) + H_i \leq \frac{C 2^i \bar{\mu} m}{\|\hat{\mathbf{f}}\|}$ for every $i = 0, \dots, r$. Combining this with Equation (81) using the union bound, we conclude that with probability at-least $1 - \delta$, we must have:

$$L_1(X) \leq \frac{C}{\bar{\mu}^{2l-1} \|\hat{\mathbf{f}}\|}$$

2. The lower bounds proceed in a similar fashion as the upper bounds in item 1. In Equation (78), we use $2^{-2l(i+1)}$ instead to get a lower bound. In Equation (80), we use the high probability lower bound of $H_i - C_0 \sqrt{H}_i \log\left(\frac{2r}{\delta}\right)$ instead of the upper bound. A similar argument as in the proof of item 1 using the high probability lower bounds for H_i bounds from Lemma 24, instead of the upper bounds allows us to conclude the result.

3. This proof proceeds similar to the proof of item 1, but replacing $H_i + C_0 \sqrt{H_i} \log\left(\frac{2r}{\delta}\right)$ with just H_i.

4. This proof proceeds similar to the proof of item 2, but replacing $H_i - C_0 \sqrt{H_i} \log\left(\frac{2r}{\delta}\right)$ with just H_i.

■

Appendix G. Proof of the Lower Bound

G.1. Technical Lemmas

For any $x \in \mathbb{R}$, we write $x^+ = \max\{x, 0\}$

Lemma 28 (Lower Bounding the Expected Risk by a Coupling) *Let $\theta_1, \theta_2 \in \Theta \subset \mathbb{R}$ be two identically distributed continuous random variables. Let $y = g(\theta, \epsilon)$ represent a model parameterized by θ, where g is some measurable function and ϵ is some random variable drawn independently of θ, such that $\mathrm{Law}(y|\theta)$ has a density with respect to the Lebesgue measure for any $\theta \in \Theta$. Furthermore, let $\hat{\theta}(y)$ be an estimator of θ. Then, for any arbitrary coupling $\Pi(\theta_1, \theta_2)$ of θ_1 and θ_2*

$$\mathbb{E}_{\theta,y}\left[(\hat{\theta}(y) - \theta)^2\right] \geq \tfrac{1}{4}\mathbb{E}_{\theta_1,\theta_2 \sim \Pi(\theta_1,\theta_2)}\left[(\theta_1 - \theta_2)^2 \left(1 - \sqrt{2\mathsf{KL}\left(\mathrm{Law}(y \mid \theta_1) \middle|\middle| \mathrm{Law}(y \mid \theta_2)\right)}\right)\right]$$

Proof Since θ_1 and θ_2 are identically distributed,

$$\begin{aligned}\mathbb{E}_{\theta,y}\left[(\hat{\theta}(y)-\theta)^2\right] &= \frac{1}{2}\mathbb{E}_{\theta_1,y}\left[\left(\hat{\theta}(y)-\theta_1\right)^2\right]+\frac{1}{2}\mathbb{E}_{\theta_2,y}\left[\left(\hat{\theta}(y)-\theta_2\right)^2\right] \\ &= \frac{1}{2}\mathbb{E}_{\theta_1,\theta_2\sim\Pi(\theta_1,\theta_2)}\left[\mathbb{E}_{y\,|\theta_1}\left[\left(\hat{\theta}(y)-\theta_1\right)^2\right]+\mathbb{E}_{y\,|\theta_2}\left[\left(\hat{\theta}(y)-\theta_2\right)^2\right]\right]\end{aligned}$$

Let $p(y \mid \theta)$ denote the density of $\mathrm{Law}(y \mid \theta)$ for any $\theta \in \Theta$. It follows that,

$$\begin{aligned}\mathbb{E}_{\theta,y}\left[(\hat{\theta}(y)-\theta)^2\right] &= \frac{1}{2}\mathbb{E}_{\theta_1,\theta_2\sim\Pi(\theta_1,\theta_2)}\left[\int\left(\hat{\theta}(y)-\theta_1\right)^2 p(y\mid\theta_1)\mathrm{d}y+\int\left(\hat{\theta}(y)-\theta_2\right)^2 p(y\mid\theta_2)\mathrm{d}y\right] \\ &\geq \frac{1}{2}\mathbb{E}_{\theta_1,\theta_2\sim\Pi(\theta_1,\theta_2)}\left[\int\left(\left(\hat{\theta}(y)-\theta_1\right)^2+\left(\hat{\theta}(y)-\theta_2\right)^2\right)\min\{p(y\mid\theta_1),p(y\mid\theta_2)\}\mathrm{d}y\right] \\ &\geq \frac{1}{4}\mathbb{E}_{\theta_1,\theta_2\sim\Pi(\theta_1,\theta_2)}\left[(\theta_1-\theta_2)^2\int\min\{p(y\mid\theta_1),p(y\mid\theta_2)\}\mathrm{d}y\right] \\ &\geq \frac{1}{4}\mathbb{E}_{\theta_1,\theta_2\sim\Pi(\theta_1,\theta_2)}\left[(\theta_1-\theta_2)^2\left(1-\mathsf{TV}\left(\mathrm{Law}(y\mid\theta_1),\mathrm{Law}(y\mid\theta_2)\right)\right)\right] \\ &\geq \frac{1}{4}\mathbb{E}_{\theta_1,\theta_2\sim\Pi(\theta_1,\theta_2)}\left[(\theta_1-\theta_2)^2\left(1-\sqrt{2\mathsf{KL}\left(\mathrm{Law}(y\mid\theta_1)\middle|\middle|\mathrm{Law}(y\mid\theta_2)\right)}\right)^+\right]\end{aligned}$$

where the second inequality follows from the identity $a^2+b^2 \geq {(a-b)^2}/{2}$, the third inequality follows from the fact that $\mathsf{TV}(P,Q) = 1-\int\min\left\{\frac{\mathrm{d}P}{\mathrm{d}\lambda},\frac{\mathrm{d}Q}{\mathrm{d}\lambda}\right\}\mathrm{d}\lambda$ for any two probability measures P and Q that are absolutely continuous with respect to some base measure λ, and the last is an application of Pinsker's inequality. ■

G.2. Proof of the Lower Bound

The proof of our lower bound for the heteroscedastic regression problem involves fixing a noise model $\mathbf{f}$ and analyzing the statistical indistinguishability between two instances of the heteroscedastic regression model parameterized by the regressors $\mathbf{w}$ and $\mathbf{w}+\mathbf{v}$ respectively, and sharing a common noise model $\mathbf{f}$. To this end, we use $P_{\mathbf{w},\mathbf{f}}$ to denote the heteroscedastic regression model parameterized by $\mathbf{w}$ and $\mathbf{f}$, i.e., $P_{\mathbf{w},\mathbf{f}}$ is a probability distribution supported on $\mathbb{R}^d \times \mathbb{R}$ such that $(\mathbf{x}, y) \sim P_{\mathbf{w},\mathbf{f}}$ implies $\mathbf{x} \sim \mathcal{N}(0, \mathbf{I})$ and $y \mid \mathbf{x} \sim \mathcal{N}(\langle \mathbf{w}, \mathbf{x}\rangle, \langle \mathbf{f}, \mathbf{x}\rangle^2)$. We first consider the simpler case when $\mathbf{v}$ is parallel to $\mathbf{f}$. This allows us to obtain a loose lower bound of $\Omega\left({\|\mathbf{f}\|^2}/{n}\right)$ via direct application of LeCam's two point method. To this end,

Lemma 29 (Lower Bound for v parallel to f) *Consider a fixed $\mathbf{f} \in \mathbb{R}^d$ and let $\mathbf{v}$ be a vector parallel to $\mathbf{f}$ such that $\|\mathbf{v}\| = \Delta$. Then, for any estimator $\hat{\mathbf{w}}$ which estimates $\mathbf{w}^*$ with inputs $(\mathbf{x}_i, y_i)_{i\in[n]}$:*

$$\inf_{\hat{\mathbf{w}}}\sup_{\mathbf{w}\in\mathbb{R}^d}\mathbb{E}_{(\mathbf{x}_i,y_i)_{i\in[n]}\overset{\mathrm{iid}}{\sim}P_{\mathbf{w},\mathbf{f}}}\left[\left\|\hat{\mathbf{w}}-\mathbf{w}\right\|^2\right] \geq \Omega({\|\mathbf{f}\|^2}/{n})$$

Proof Consider any two instances of the heteroscedastic regression problem with a common noise model $\hat{\mathbf{f}}$ and regressors $\mathbf{w}$ and $\mathbf{w}+\mathbf{v}$ respectively. From a direct computation, it follows that,

$$\begin{aligned}\mathsf{KL}\left(\text{Law}\left(\mathbf{x},\mathbf{y}|\mathbf{w},\mathbf{f}\right)\middle|\middle|\text{Law}\left(\mathbf{x},\mathbf{y}|\mathbf{w}+\mathbf{v},\mathbf{f}\right)\right) &= \mathbb{E}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I})}\left[\mathsf{KL}\left(\mathcal{N}(\langle\mathbf{w},\mathbf{x}\rangle,\langle\mathbf{f},\mathbf{x}\rangle^2)\middle|\middle|\mathcal{N}(\langle\mathbf{w}+\mathbf{v},\mathbf{x}\rangle,\langle\mathbf{f},\mathbf{x}\rangle^2)\right)\right] \\ &= \mathbb{E}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I})}\left[\left(\frac{\langle\mathbf{v},\mathbf{x}\rangle}{\langle\mathbf{f},\mathbf{x}\rangle}\right)^2\right] = \frac{\Delta^2}{\|\mathbf{f}\|^2}\end{aligned}$$

From LeCam's two point method, we obtain, the following for any $\Delta \geq 0$

$$\inf_{\hat{\mathbf{w}}}\sup_{\mathbf{w}\in\mathbb{R}^d}\mathbb{E}_{(\mathbf{x}_i,y_i)_{i\in[n]}\overset{\text{iid}}{\sim}P_{\mathbf{w},\mathbf{f}}}\left[\|\hat{\mathbf{w}}-\mathbf{w}\|^2\right] \geq \Delta^2 e^{-\frac{n\Delta^2}{\|\mathbf{f}\|^2}}$$

Setting $\Delta = \|\hat{\mathbf{f}}\|/\sqrt{n}$, we obtain,

$$\inf_{\hat{\mathbf{w}}}\sup_{\mathbf{w}\in\mathbb{R}^d}\mathbb{E}_{(\mathbf{x}_i,y_i)_{i\in[n]}\overset{\text{iid}}{\sim}P_{\mathbf{w},\mathbf{f}}}\left[\|\hat{\mathbf{w}}-\mathbf{w}\|^2\right] \geq \Omega(\|\mathbf{f}\|^2/n)$$

■

We now present a finer lower bound of $\Omega\left(d^2/n^2\right)$ by considering the case when $\mathbf{v}$ is perpendicular to $\mathbf{f}$. This case encapsulates the key technical challenge of our lower bound analysis.

Lemma 30 (Finer Lower Bound) *Assume $n \geq d$ and $d \geq 2$. For fixed $\mathbf{f}\in\mathbb{R}^d$ and any arbitrary estimator $\hat{\mathbf{w}}$,*

$$\inf_{\hat{\mathbf{w}}}\sup_{\mathbf{w}\in\mathbb{R}^d}\mathbb{E}_{(\mathbf{x}_i,y_i)_{i\in[n]}\overset{\text{iid}}{\sim}P_{\mathbf{w},\mathbf{f}}}\left[\|\hat{\mathbf{w}}-\mathbf{w}\|^2\right] \geq \tilde{\Omega}(\|\mathbf{f}\|^2 d^2/n^2)$$

Here $\tilde{\Omega}$ hides a factor of $\left(\frac{1}{(\log\log n)^{3/2}+(\log\log n)\log d+\frac{(\log\log n)^2(\log d)^2}{\sqrt{d}}}\right)^2$

Proof Let $\mathbf{e}_1,\ldots,\mathbf{e}_d$ denote an orthonormal basis of $\mathbb{R}^d$ such that $\mathbf{e}_d = \mathbf{f}/\|\mathbf{f}\|$. Furthermore, let $S_\alpha(\mathbf{e}_{1:d-1})$ denote uniform distribution over the sphere of radius α on the subspace spanned by $\mathbf{e}_1,\ldots,\mathbf{e}_{d-1}$ centered around the origin. It follows that, for any estimator $\hat{\mathbf{w}}$,

$$\begin{aligned}\sup_{\mathbf{w}\in\mathbb{R}^d}\mathbb{E}_{(\mathbf{x}_i,y_i)_{i\in[n]}\overset{\text{iid}}{\sim}P_{\mathbf{w},\mathbf{f}}}\left[\|\hat{\mathbf{w}}-\mathbf{w}\|^2\right] &\geq \mathbb{E}_{\mathbf{w}\sim S_\alpha(\mathbf{e}_{1:d-1})}\left[\mathbb{E}_{(\mathbf{x}_i,y_i)_{i\in[n]}\overset{\text{iid}}{\sim}P_{\mathbf{w},\mathbf{f}}}\left[\|\hat{\mathbf{w}}-\mathbf{w}\|^2\right]\right] \\ &= \mathbb{E}_{\mathbf{w}\sim S_\alpha(\mathbf{e}_{1:d-1})}\left[\mathbb{E}_{(\mathbf{x}_i)_{i\in[n]}\overset{\text{iid}}{\sim}\mathcal{N}(0,\mathbf{I})}\left[\mathbb{E}_{\mathbf{y}_i\sim\mathcal{N}(\langle\mathbf{w},\mathbf{x}_i\rangle,\langle\mathbf{f},\mathbf{x}_i\rangle^2),i\in[n]}\left[\|\hat{\mathbf{w}}-\mathbf{w}\|^2\right]\right]\right]\end{aligned}$$

For ease of notation, we shall denote $\mathbf{x}_{1:n}$ to be the shorthand for $(\mathbf{x}_i)_{i\in[n]}\overset{\text{iid}}{\sim}\mathcal{N}(0,\mathbf{I})$. Similarly, we shall write $y_i\sim\mathcal{N}(\langle\mathbf{w},\mathbf{x}_i\rangle,\langle\mathbf{f},\mathbf{x}_i\rangle^2), i\in[n]$ as $y_{1:n}|\mathbf{x}_{1:n}$. Thus,

$$\begin{aligned}\sup_{\mathbf{w}\in\mathbb{R}^d}\mathbb{E}_{(\mathbf{x}_i,y_i)_{i\in[n]}\overset{\text{iid}}{\sim}P_{\mathbf{w},\mathbf{f}}}\left[\|\hat{\mathbf{w}}-\mathbf{w}\|^2\right] &\geq \mathbb{E}_{\mathbf{w}\sim S_\alpha(\mathbf{e}_{1:d-1})}\left[\mathbb{E}_{\mathbf{x}_{1:n}}\left[\mathbb{E}_{y_{1:n}|\mathbf{x}_{1:n}}\left[\|\hat{\mathbf{w}}-\mathbf{w}\|^2\right]\right]\right] \\ &= \mathbb{E}_{\mathbf{x}_{1:n}}\left[\mathbb{E}_{\mathbf{w}\sim S_\alpha(\mathbf{e}_{1:d-1})}\left[\mathbb{E}_{y_{1:n}|\mathbf{x}_{1:n}}\left[\|\hat{\mathbf{w}}-\mathbf{w}\|^2\right]\right]\right]\end{aligned}$$

We now proceed by performing a *data-dependent change of co-ordinates*. In particular, we let $\mathbf{u}_1, \dots, \mathbf{u}_d$ be an orthonormal basis of $\mathbb{R}^d$ such that $\mathbf{u}_d = \mathbf{e}_d = {}^{\mathbf{f}}/_{\|\mathbf{f}\|}$ and $\mathbf{u}_1, \dots, \mathbf{u}_{d-1}$ are measurable functions of $\mathbf{x}_1, \dots, \mathbf{x}_n$, to be specified later. We note that, $\text{Span}(\mathbf{e}_1, \dots, \mathbf{e}_{d-1}) = \text{Span}(\mathbf{u}_1, \dots, \mathbf{u}_{d-1})$, and hence, $S_\alpha(\mathbf{e}_{1:d-1})$ is equal to $S_\alpha(\mathbf{u}_{1:d-1})$ as they represent the uniform spherical distribution on the same subspace. Hence,

$$\begin{aligned}\sup_{\mathbf{w}\in\mathbb{R}^d} \mathbb{E}_{(\mathbf{x}_i,y_i)_{i\in[n]} \overset{\text{iid}}{\sim} P_{\mathbf{w},\mathbf{f}}} \left[\|\hat{\mathbf{w}} - \mathbf{w}\|^2\right] &\geq \mathbb{E}_{\mathbf{x}_{1:n}} \left[\mathbb{E}_{\mathbf{w}\sim S_\alpha(\mathbf{u}_{1:d-1})} \left[\mathbb{E}_{y_{1:n}|\mathbf{x}_{1:n}} \left[\|\hat{\mathbf{w}} - \mathbf{w}\|^2\right]\right]\right] \\ &\geq \mathbb{E}_{\mathbf{x}_{1:n}} \left[\mathbb{E}_{\mathbf{w}\sim S_\alpha(\mathbf{u}_{1:d-1})} \left[\mathbb{E}_{y_{1:n}|\mathbf{x}_{1:n}} \left[\sum_{i=1}^{d} \langle \hat{\mathbf{w}} - \mathbf{w}, \mathbf{u}_i\rangle^2\right]\right]\right]\end{aligned} \tag{83}$$

We now construct a family of *co-ordinate flip couplings*, i.e., for every $i \in [d]$, let $(\mathbf{w}, \mathbf{w}^{\setminus i})$ be a coupling such that

$$\left\langle \mathbf{w}^{\setminus i}, \mathbf{u}_j \right\rangle = \begin{cases} \langle \mathbf{w}, \mathbf{u}_j\rangle, & \text{if } j \neq i \\ -\langle \mathbf{w}, \mathbf{u}_j\rangle, & \text{otherwise} \end{cases}$$

By the symmetry of the uniform spherical distribution, both $\mathbf{w}$ and $\mathbf{w}^{\setminus i}$ are distributed identically as $S_\alpha(\mathbf{u}_{1:d-1})$. Applying Lemma 28 to each summand in (83), we conclude that,

$$4R^{\hat{\mathbf{w}}}_{d,n} \geq \mathbb{E}_{\mathbf{x}_{1:n}} \left[\sum_{i=1}^{d-1} \mathbb{E}_{\mathbf{w},\mathbf{w}^{\setminus i}} \left[(\mathbf{w} - \mathbf{w}^{\setminus i})^2 \left(1 - \sqrt{2\mathsf{KL}\left(\text{Law}\left(y_{1:n}\,|\mathbf{x}_{1:n}, \mathbf{w}, \mathbf{f}\right)\middle|\middle|\text{Law}\left(y_{1:n}\,|\mathbf{x}_{1:n}, \mathbf{w}^{\setminus i}, \mathbf{f}\right)\right)}\right)^+\right]\right]$$

where $R^{\hat{\mathbf{w}}}_{d,n}$ denotes the worst-case risk of the estimator $\hat{\mathbf{w}}$ defined as follows:

$$\sup_{\mathbf{w}\in\mathbb{R}^d} \mathbb{E}_{(\mathbf{x}_i,y_i)_{i\in[n]} \overset{\text{iid}}{\sim} P_{\mathbf{w},\mathbf{f}}} \left[\|\hat{\mathbf{w}} - \mathbf{w}\|^2\right]$$

Furthermore, since $\text{Law}\left(y_{1:n} \mid \mathbf{x}_{1:n}, \mathbf{w}, \mathbf{f}\right) = \prod_{j=1}^{n} \mathcal{N}(\mathbf{y}_i \mid \langle \mathbf{w}, \mathbf{x}_i\rangle, \langle \mathbf{f}, \mathbf{x}_i\rangle^2)$, it follows that,

$$\mathsf{KL}\left(\text{Law}\left(y_{1:n}\,|\mathbf{x}_{1:n}, \mathbf{w}, \mathbf{f}\right)\middle|\middle|\text{Law}\left(y_{1:n}\,|\mathbf{x}_{1:n}, \mathbf{w}^{\setminus i}, \mathbf{f}\right)\right) = \frac{\left\langle \mathbf{w} - \mathbf{w}^{\setminus i}, \mathbf{u}_i\right\rangle^2}{\|\mathbf{f}\|^2} \sum_{j=1}^{n} \frac{\langle \mathbf{u}_i, \mathbf{x}_j\rangle^2}{\langle \mathbf{u}_d, \mathbf{x}_j\rangle^2},$$

where we use the fact that $\mathbf{f} = \|\mathbf{f}\|\mathbf{u}_d$. We also note that $\langle \mathbf{w}, \mathbf{u}_i\rangle - \left\langle \mathbf{w}^{\setminus i}, \mathbf{u}_i\right\rangle = 2\langle \mathbf{w}, \mathbf{u}_i\rangle$ almost surely. Furthermore, for ease of notation, denote $\delta_i = \langle \mathbf{w}, \mathbf{u}_i\rangle$. It follows that,

$$R^{\hat{\mathbf{w}}}_{d,n} \geq \mathbb{E}_{\mathbf{x}_{1:n}} \left[\mathbb{E}_{\mathbf{w}\sim S_\alpha(\mathbf{u}_{1:d-1})} \left[\sum_{i=1}^{d-1} \delta_i^2 \left(1 - \sqrt{4\frac{\delta_i^2}{\|\mathbf{f}\|^2} \sum_{j=1}^{n} \left(\frac{\langle \mathbf{u}_i, \mathbf{x}_j\rangle}{\langle \mathbf{u}_d, \mathbf{x}_j\rangle}\right)^2}\right)^+\right]\right] \tag{84}$$

Note that u_d is deterministic and independent of $\mathbf{x}_{1:n}$. We divide the indices $[n]$ in to the following buckets given $1 \geq \gamma > 0$ and $k \geq 0$.

$$B_k = \{j \in [n] : \langle \mathbf{u}_d, \mathbf{x}_j\rangle^2 \in [2^k\gamma, 2^{k+1}\gamma)\}$$

Let $K = \inf\{k : \sum_{l=0}^{k} |B_l| \geq d-1\}$ and $K^{\max}$ be any integer. Define $\mathcal{E}$ to be the event that the following hold simultaneously:

1. $\inf_j \langle \mathbf{u}_d, \mathbf{x}_j \rangle^2 \geq \gamma$
2. $\sup_j \langle \mathbf{u}_d, \mathbf{x}_j \rangle^2 \leq \gamma 2^{K^{\max}}$
3. $K < \min(K^{\max}, \bar{C} + 2\log_2 d)$
4. $|B_k| \leq C_B n \sqrt{\gamma} 2^{k/2} \log(K^{\max}) \quad \forall \quad k \leq K^{\max}$ for some constant C_B.

We pick $\gamma = \frac{c_0}{n^2}$ and $K^{\max} = \max(C_1 \log n, C_2 + \log_2 d)$ for small enough c_0 and large enough constants C_1, C_2. Using the fact that the Gaussian density is bounded above by a constant and that $\langle \mathbf{u}_d, \mathbf{x}_j \rangle$ are i.i.d standard Gaussians we conclude that with probability at-least $9/10$, we must have $\mathbb{P}(\inf_n |\langle \mathbf{u}_d, \mathbf{x}_j \rangle|^2 \geq \gamma) \geq 9/10$. By an application of Lemma 16 we conclude that $\mathbb{P}(\sup_j \langle \mathbf{u}_d, \mathbf{x}_j \rangle^2 \leq \gamma 2^{K^{\max}}) \geq 9/10$. Furthermore, note that $|B_k| \sim \mathsf{Ber}(n, p_k)$ where $p_k \leq C_2 \sqrt{\gamma} 2^{k/2}$. Lemma 24 ensures that $\mathbb{P}(K < \min(K^{\max}, \bar{C} + 2\log_2 d)) \geq 9/10$ since $\mathbb{P}(|B_{\min(K^{\max}, \bar{C}+2\log_2 d)}| \geq d) \geq \frac{9}{10}$. By an application of Bernstein's inequality, we conclude that $\mathbb{P}(|B_k| \leq C_B n \sqrt{\gamma} 2^{k/2} \log(K^{\max})) \geq 1 - \frac{1}{10(K^{\max}+1)}$ by taking C_B to be a large enough universal constant.

Applying union bound on the above events, we conclude that $\mathbb{P}(\mathcal{E}) \geq \frac{6}{10}$. Under the event $\mathcal{E}$, the following holds almost surely

$$\sum_{j=1}^{n} \left(\frac{\langle \mathbf{u}_i, \mathbf{x}_j \rangle}{\langle \mathbf{u}_d, \mathbf{x}_j \rangle} \right)^2 \leq \sum_{k=0}^{K^{\max}} \sum_{j \in B_k} 2^{-k} \frac{\langle \mathbf{u}_i, \mathbf{x}_j \rangle^2}{\gamma} \tag{85}$$

Let Q be the projector onto the space orthogonal to $\mathbf{u}_d$. Then, arguing by the properties of Gaussians, $(Q\mathbf{x}_j)_{j \in [n]}$ is independent of $(B_k)_{k \geq 0}$. Let $\tilde{\mathbf{x}}_j := Q\mathbf{x}_j$. We must have have $\langle \mathbf{u}_i, \mathbf{x}_j \rangle = \langle \mathbf{u}_i, \tilde{\mathbf{x}}_j \rangle$ almost surely whenever $i \neq d$. Therefore, using this in Equation (85), we conclude that under the event $\mathcal{E}$:

$$\begin{aligned} \sum_{j=1}^{n} \left(\frac{\langle \mathbf{u}_i, \mathbf{x}_j \rangle}{\langle \mathbf{u}_d, \mathbf{x}_j \rangle} \right)^2 &\leq \sum_{k=0}^{K^{\max}} \sum_{j \in B_k} 2^{-k} \frac{\langle \mathbf{u}_i, \tilde{\mathbf{x}}_j \rangle^2}{\gamma} \\ &= \sum_{k=0}^{K^{\max}} \frac{2^{-k}}{\gamma} \mathbf{u}_i^{\intercal} H_k \mathbf{u}_i \end{aligned} \tag{86}$$

Where $H_k = \sum_{j \in B_k} \tilde{\mathbf{x}}_j \tilde{\mathbf{x}}_j^{\intercal}$ and we interpret H_k as an operator over $\mathsf{span}(Q)$. Notice that due to the properties of Gaussians, H_k has rank $|B_k|$ almost surely. We are now ready to pick the vectors $\mathbf{u}_1, \dots, \mathbf{u}_{d-1}$. We define, only in this proof, $N_k := \sum_{l=0}^{k} |B_l|$.

Whenever $d-1 \geq |B_0| > 0$, let $\mathbf{u}_1, \dots, \mathbf{u}_{|B_0|}$ be the eigenvectors H_0 with non-zero eigenvalue. If $|B_0| > d-1$, then choose $\mathbf{u}_1, \dots, \mathbf{u}_{d-1}$ to be any arbitrary non-zero eigenvectors of H_0.

Similarly, for $K - 1 \geq k \geq 1$, whenever $|B_k| > 0$, we consider $\mathbf{u}_1^{(k)}, \dots, \mathbf{u}_{|B_k|}^{(k)}$ to be the non-zero eigenvectors of H_k. We define $\mathbf{u}_{1+N_{k-1}}, \dots, \mathbf{u}_{N_k}$ to be $\mathbf{u}_1^{(k)}, \dots, \mathbf{u}_{|B_k|}^{(k)}$, after Gram-Schmidt Orthonormalization with respect to $\mathbf{u}_1, \dots, \mathbf{u}_{N_k}$ (we skip Orthonormalization step in the event $N_{k-1} = 0$). Note that such vectors exist almost surely since the rank of $\sum_{l=0}^{k} H_l$ is N_k almost surely for every $k < K$. If $N_K = d-1$, we do the same procedure as for $1 \leq k \leq K-1$ as described above. If $N_K > d-1$, we take $\mathbf{u}_1^{(K)}, \dots, \mathbf{u}_{d-1-N_{K-1}}^{(K)}$ to be any non-zero eigenvectors of H_K. We ortho-normalize them as above with respect to $\mathbf{u}_1, \dots, \mathbf{u}_{N_{K-1}}$ to obtain $\mathbf{u}_{N_{K-1}+1}, \dots, \mathbf{u}_{d-1}$.

Notice that whenever $0 \leq k < K$, almost surely $\mathbf{u}_i^\intercal H_k \mathbf{u}_i = 0$ whenever $i > N_k$ because we pick such $\mathbf{u}_i$ to be orthogonal to $\mathsf{span}(H_k)$. We also note that whenever $i \leq N_{k-1}$, $\mathbf{u}_i$ is independent of H_k when conditioned on $(B_l)_{l \geq 0}$. We collect these observations in the following claim:

Claim 31

1. *For $k < K$ and $d - 1 \geq i > N_k$,*
$$\mathbf{u}_i^\intercal H_k \mathbf{u}_i = 0\,.$$
2. *For $k \leq K$ and $\min(d-1, N_k) \geq i > N_{k-1}$,*
$$\mathbf{u}_i^\intercal H_k \mathbf{u}_i \leq \|H_k\|\,.$$
3. *When conditioned on $(B_l)_l$ and $\mathcal{E}$, when either*
   (a) $k \leq K$ and $0 < i \leq N_{k-1}$
   (b) $k > K$ and $1 \leq i \leq d-1$

   Then, with probability at-least $1 - \delta$:
$$\mathbf{u}_i^\intercal H_k \mathbf{u}_i \leq |B_k| + C\sqrt{|B_k|}\log\left(\tfrac{1}{\delta}\right)\,.$$
4. *When conditioned on $(B_l)_l$ and $\mathcal{E}$, with probability at-least $1 - \delta$, we must have:*
$$\|H_k\| \leq C(|B_k| + d + \log\left(\tfrac{1}{\delta}\right))$$

Proof Items 1 and 2 are easy to show based on the prior discussion. Note that $(\tilde{\mathbf{x}}_j)_j$ are independent of $(B_l)_l$ and $\mathcal{E}$. Item 3 follows from Lemma 25 after noting that conditioned on $\mathcal{E}$ and $(B_l)_l$, we must have $\mathbf{u}_i^\intercal H_k \mathbf{u}_i$ is a sum of $|B_k|$ i.i.d. χ^2 random variables (since $\mathbf{u}_i$ is independent of H_k). Item 4 follows from (Vershynin, 2010, Theorem 5.29). ■

In the discussion below, we condition on $(B_l)_l$ and $\mathcal{E}$. The probabilities are all conditional probabilities. We pick $K^{\max} > K$ to be a positive integer. By union bound and the claim above, the have with inequalities all hold simultaneously with probability at-least $1 - \delta$:

1. For $0 \leq i \leq \min(N_0, d-1)$,
$$\mathbf{u}_i^\intercal H_0 \mathbf{u}_i \leq C(|B_0| + d + \log\left(\tfrac{K^{\max} d}{\delta}\right))\,.$$
2. For $0 < k \leq K$, $N_{k-1} \leq i \leq \min(N_k, d-1)$, $l < k$ (if such a k exists)
$$\mathbf{u}_i^\intercal H_l \mathbf{u}_i = 0\,.$$
3. For $0 < k \leq K$, $N_{k-1} \leq i \leq \min(N_k, d-1)$ (if such a k exists)
$$\mathbf{u}_i^\intercal H_k \mathbf{u}_i = C(|B_k| + d + \log\left(\tfrac{K^{\max} d}{\delta}\right))\,.$$

4. $0 < k \leq K$ and $0 < i \leq N_{k-1}$ (whenever such k, i exist)
$$\mathbf{u}_i^{\intercal} H_k \mathbf{u}_i \leq |B_k| + C\sqrt{|B_k|}\log\left(\tfrac{K^{\max}d}{\delta}\right).$$

5. $K^{\max} > k > K$ and $1 \leq i \leq d-1$
$$\mathbf{u}_i^{\intercal} H_k \mathbf{u}_i \leq |B_k| + C\sqrt{|B_k|}\log\left(\tfrac{K^{\max}d}{\delta}\right).$$

Therefore, from the above relationships, we conclude that when conditioned on $\mathcal{E}$, with probability at-least $1-\delta$, the following inequalities hold:

1. $1 \leq i \leq \min(N_0, d-1)$ whenever such an i exists:
$$\begin{aligned}\sum_{k=0}^{K^{\max}} \frac{2^{-k}}{\gamma}\mathbf{u}_i^{\intercal} H_k \mathbf{u}_i &\lesssim \frac{1}{\gamma}\left[|B_0| + d + \log\left(\tfrac{K^{\max}d}{\delta}\right) + \sum_{k=1}^{K^{\max}} 2^{-k}\left(|B_k| + \sqrt{|B_k|}\log\left(\tfrac{K^{\max}d}{\delta}\right)\right)\right] \\ &\lesssim \frac{1}{\gamma}\left[\sqrt{\gamma}n\log(K^{\max}) + d + \log\left(\tfrac{K^{\max}d}{\delta}\right) + \sum_{k=1}^{K^{\max}} 2^{-k}\left(\sqrt{\gamma}n2^{k/2}\log(K^{\max}) + \log^2\left(\tfrac{K^{\max}d}{\delta}\right)\right)\right] \\ &\lesssim \frac{n\log(K^{\max})}{\sqrt{\gamma}} + \frac{d}{\gamma} + \frac{\log^2\left(\frac{K^{\max}d}{\delta}\right)}{\gamma} \qquad (87)\end{aligned}$$
In the second step, we have used the bounds on $|B_k|$ in the definition of the event $\mathcal{E}$.

2. $N_{k-1} < i \leq \min(N_k, d-1)$, $K \geq k > 0$ whenever such k, i exist:
$$\begin{aligned}\sum_{l=0}^{K^{\max}} \frac{2^{-l}}{\gamma}\mathbf{u}_i^{\intercal} H_l \mathbf{u}_i &\lesssim \frac{1}{\gamma}\left[2^{-k}\left(|B_k| + d + \log\left(\tfrac{K^{\max}d}{\delta}\right)\right) + \sum_{l=k+1}^{K^{\max}} 2^{-l}\left(|B_l| + \sqrt{|B_l|}\log\left(\tfrac{K^{\max}d}{\delta}\right)\right)\right] \\ &\lesssim 2^{-k/2}\frac{n\log(K^{\max})}{\sqrt{\gamma}} + 2^{-k}\frac{(d + \log^2(\frac{K^{\max}d}{\delta}))}{\gamma} \qquad (88)\end{aligned}$$

Set $\delta = 1/6$. Let us call the event in the which the inequalities given in Equation (86), (87) and (88) hold as $\mathcal{F}$. Now, $\mathbb{P}(\mathcal{F}) \geq \mathbb{P}(\mathcal{F}\cap\mathcal{E}) = \mathbb{P}(\mathcal{F}|\mathcal{E})\mathbb{P}(\mathcal{E}) \geq 5/6 \times 6/10 = 1/2$. Notice that the event $\mathcal{F}$ is measurable with respect to the sigma algebra $\sigma(\mathbf{x}_{1:n})$.

In Equation (84), consider the following term when $\mathbf{x}_{1:n}$ is fixed and satisfies the event $\mathcal{F}$:
$$\begin{aligned}&\mathbb{E}_{\mathbf{w}\sim S_\alpha(\mathbf{u}_{1:d-1})}\left[\sum_{i=1}^{d-1}\delta_i^2\left(1 - \sqrt{4\frac{\delta_i^2}{\|\mathbf{f}\|^2}\sum_{j=1}^{n}\left(\frac{\langle \mathbf{u}_i, \mathbf{x}_j\rangle}{\langle \mathbf{u}_d, \mathbf{x}_j\rangle}\right)^2}\right)^+\right] \\ &\geq \mathbb{E}_{\mathbf{w}\sim S_\alpha(\mathbf{u}_{1:d-1})}\left[\sum_{i=1}^{d-1}\delta_i^2\left(1 - \sqrt{4\frac{\delta_i^2}{\|\mathbf{f}\|^2}\sum_{j=1}^{n}\left(\frac{\langle \mathbf{u}_i, \mathbf{x}_j\rangle}{\langle \mathbf{u}_d, \mathbf{x}_j\rangle}\right)^2}\right)\right] \\ &= \mathbb{E}_{\mathbf{w}\sim S_\alpha(\mathbf{u}_{1:d-1})}\left[\sum_{i=1}^{d-1}\delta_i^2 - \sum_{i=1}^{d-1}|\delta_i|^3\sqrt{\frac{4}{\|\mathbf{f}\|^2}\sum_{j=1}^{n}\left(\frac{\langle \mathbf{u}_i, \mathbf{x}_j\rangle}{\langle \mathbf{u}_d, \mathbf{x}_j\rangle}\right)^2}\right] \\ &\geq \left[\alpha^2 - C_3\frac{\alpha^3}{(d)^{\frac{3}{2}}}\sum_{i=1}^{d-1}\sqrt{\frac{4}{\|\mathbf{f}\|^2}\sum_{j=1}^{n}\left(\frac{\langle \mathbf{u}_i, \mathbf{x}_j\rangle}{\langle \mathbf{u}_d, \mathbf{x}_j\rangle}\right)^2}\right] \qquad (89)\end{aligned}$$

In the last step, we have used the fact that $\sum_{i=1}^{d-1}\delta_i^2 = \alpha^2$ almost surely and that $\mathbb{E}|\delta_i|^3 \leq \frac{C_3\alpha^3}{d^{\frac{3}{2}}}$ for some universal constant C_3 (can be shown easily, see bounds in Brennan et al. (2020)). Since Equation (86), (87) and (88) hold under the event $\mathcal{F}$ by definition, we proceed to bound:

$$\begin{aligned}
\sum_{i=1}^{d-1}\sqrt{\frac{4}{\|\mathbf{f}\|^2}\sum_{j=1}^{n}\left(\frac{\langle \mathbf{u}_i,\mathbf{x}_j\rangle}{\langle \mathbf{u}_d,\mathbf{x}_j\rangle}\right)^2} &\leq \sum_{i=1}^{d-1}\sqrt{\frac{4}{\|\mathbf{f}\|^2}\sum_{k=0}^{K^{\max}}\frac{2^{-k}}{\gamma}\mathbf{u}_i^{\intercal}H_k\mathbf{u}_i} \\
&\lesssim \min(N_0,d-1)\sqrt{\left[\frac{n\log(K^{\max})}{\sqrt{\gamma}}+\frac{d}{\gamma}+\frac{\log^2(K^{\max}d)}{\gamma}\right]} \\
&\quad+\sum_{k=1}^{K}(\min(N_k,d-1)-N_{k-1})\sqrt{\left[2^{-k/2}\frac{n\log(K^{\max})}{\sqrt{\gamma}}+2^{-k}\frac{(d+\log^2(K^{\max}d))}{\gamma}\right]} \\
&\lesssim |B_0|\sqrt{\left[\frac{n\log(K^{\max})}{\sqrt{\gamma}}+\frac{d}{\gamma}+\frac{\log^2(K^{\max}d)}{\gamma}\right]} \\
&\quad+\sum_{k=1}^{K}|B_k|\sqrt{\left[2^{-k/2}\frac{n\log(K^{\max})}{\sqrt{\gamma}}+2^{-k}\frac{(d+\log^2(K^{\max}d))}{\gamma}\right]} \\
&\lesssim n\sqrt{\gamma}\log(K^{\max})\sqrt{\left[\frac{n\log(K^{\max})}{\sqrt{\gamma}}+\frac{d}{\gamma}+\frac{\log^2(K^{\max}d)}{\gamma}\right]} \\
&\quad+\sum_{k=1}^{K}2^{k/2}n\sqrt{\gamma}\log(K^{\max})\sqrt{\left[2^{-k/2}\frac{n\log(K^{\max})}{\sqrt{\gamma}}+2^{-k}\frac{(d+\log^2(K^{\max}d))}{\gamma}\right]} \qquad (90)
\end{aligned}$$

Now, using the fact that $\gamma = \frac{c_0}{n^2}$ in the equation above, we conclude:

$$\begin{aligned}
&\sum_{i=1}^{d-1}\sqrt{\frac{4}{\|\mathbf{f}\|^2}\sum_{j=1}^{n}\left(\frac{\langle \mathbf{u}_i,\mathbf{x}_j\rangle}{\langle \mathbf{u}_d,\mathbf{x}_j\rangle}\right)^2} \\
&\lesssim n(\log(K^{\max}))^{\frac{3}{2}}2^{\frac{K}{4}}+\log(K^{\max})(K+1)n\sqrt{d+\log^2(K^{\max}d)} \\
&\lesssim n(\log(K^{\max}))^{\frac{3}{2}}\sqrt{d}+n\sqrt{d+\log^2(K^{\max}d)}\log(K^{\max})\log(d) \qquad (91)
\end{aligned}$$

In the last step, we have used the fact that $2^{K/4}\lesssim\sqrt{d}$. Plugging the bound above into Equation (89), we conclude that whenever $\mathbf{x}_{1:n}$ satisfies event $\mathcal{F}$, we must have:

$$\begin{aligned}
&\mathbb{E}_{\mathbf{w}\sim S_\alpha(\mathbf{u}_{1:d-1})}\left[\sum_{i=1}^{d-1}\delta_i^2\left(1-\sqrt{4\frac{\delta_i^2}{\|\mathbf{f}\|^2}\sum_{j=1}^{n}\left(\frac{\langle \mathbf{u}_i,\mathbf{x}_j\rangle}{\langle \mathbf{u}_d,\mathbf{x}_j\rangle}\right)^2}\right)^{+}\right] \\
&\geq \alpha^2 - C_4\frac{\alpha^3}{(d)^{\frac{3}{2}}}\left[n\sqrt{d}(\log K^{\max})^{\frac{3}{2}}+n\sqrt{d}\log(K^{\max})\log(d)+n(\log(K^{\max})\log(d))^2\right] \qquad (92)
\end{aligned}$$

Now, we pick $\alpha = \frac{cd}{n}\left[\frac{1}{(\log\log n)^{3/2}+(\log\log n)\log d+\frac{(\log\log n)^2(\log d)^2}{\sqrt{d}}}\right]$ for some small enough constant c. Then, we conclude that under the event $\mathcal{F}$:

$$\mathbb{E}_{\mathbf{w}\sim S_\alpha(\mathbf{u}_{1:d-1})}\left[\sum_{i=1}^{d-1}\delta_i^2\left(1-\sqrt{4\frac{\delta_i^2}{\|\mathbf{f}\|^2}\sum_{j=1}^{n}\left(\frac{\langle \mathbf{u}_i,\mathbf{x}_j\rangle}{\langle \mathbf{u}_d,\mathbf{x}_j\rangle}\right)^2}\right)^+\right] \geq \frac{\alpha^2}{2} \tag{93}$$

Going back to Equation (84), we conclude:

$$\begin{aligned}
R_{d,n}^{\hat{\mathbf{w}}} &\geq \mathbb{E}_{\mathbf{x}_{1:n}}\left[\mathbb{E}_{\mathbf{w}\sim S_\alpha(\mathbf{u}_{1:d-1})}\left[\sum_{i=1}^{d-1}\delta_i^2\left(1-\sqrt{4\frac{\delta_i^2}{\|\mathbf{f}\|^2}\sum_{j=1}^{n}\left(\frac{\langle \mathbf{u}_i,\mathbf{x}_j\rangle}{\langle \mathbf{u}_d,\mathbf{x}_j\rangle}\right)^2}\right)^+\right]\right] \\
&\geq \mathbb{E}_{\mathbf{x}_{1:n}}\mathbb{1}(\mathcal{F})\left[\mathbb{E}_{\mathbf{w}\sim S_\alpha(\mathbf{u}_{1:d-1})}\left[\sum_{i=1}^{d-1}\delta_i^2\left(1-\sqrt{4\frac{\delta_i^2}{\|\mathbf{f}\|^2}\sum_{j=1}^{n}\left(\frac{\langle \mathbf{u}_i,\mathbf{x}_j\rangle}{\langle \mathbf{u}_d,\mathbf{x}_j\rangle}\right)^2}\right)^+\right]\right] \\
&\geq \mathbb{E}_{\mathbf{x}_{1:n}}\mathbb{1}(\mathcal{F})\frac{\alpha^2}{2} = \frac{1}{2}\alpha^2\mathbb{P}(\mathcal{F}) \geq {}^{\alpha^2}\!/_4
\end{aligned} \tag{94}$$

In the second step, we have used Equation (84). Using our choice of α, we conclude the result. ■

Equipped with Lemmas 29 and 30, we finally complete the proof of the lower bound as follows:

G.3. Proof of Theorem 8

Proof [Proof of Theorem 8] Consider a fixed $\mathbf{f} \in \mathbb{R}^d$. By definition, for any estimator $\mathbf{w}$,

$$\inf_{\hat{\mathbf{w}}}\sup_{P_{\mathbf{w}^*,\mathbf{f}^*}\in\mathcal{P}}\mathbb{E}_{(\mathbf{x}_i,y_i)_{i\in[n]}\overset{\text{iid}}{\sim}P_{\mathbf{w}^*,\mathbf{f}^*}}\left[\|\hat{\mathbf{w}}-\mathbf{w}^*\|^2\right] \geq \inf_{\hat{\mathbf{w}}}\sup_{\mathbf{w}^*\in\mathbb{R}^d}\mathbb{E}_{(\mathbf{x}_i,y_i)_{i\in[n]}\overset{\text{iid}}{\sim}P_{\mathbf{w}^*,\mathbf{f}}}\left[\|\hat{\mathbf{w}}-\mathbf{w}^*\|^2\right]$$

From Lemma 29 and 30, it follows that,

$$\begin{aligned}
\inf_{\hat{\mathbf{w}}}\sup_{\mathbf{w}^*\in\mathbb{R}^d}\mathbb{E}_{(\mathbf{x}_i,y_i)_{i\in[n]}\overset{\text{iid}}{\sim}P_{\mathbf{w}^*,\mathbf{f}}}\left[\|\hat{\mathbf{w}}-\mathbf{w}^*\|^2\right] &\geq \Omega\left(\max\left\{{}^{\|\mathbf{f}\|^2}\!/_n, {}^{\|\mathbf{f}\|^2d^2}\!/_{n^2}\right\}\right) \\
&\geq \Omega\left({}^{\|\mathbf{f}\|^2}\!/_n + {}^{\|\mathbf{f}\|^2d^2}\!/_{n^2}\right)
\end{aligned}$$

Hence, we obtain the desired lower bound as follows,

$$\inf_{\hat{\mathbf{w}}}\sup_{P_{\mathbf{w}^*,\mathbf{f}^*}\in\mathcal{P}}\mathbb{E}_{(\mathbf{x}_i,y_i)_{i\in[n]}\overset{\text{iid}}{\sim}P_{\mathbf{w},\mathbf{f}}}\left[\|\hat{\mathbf{w}}-\mathbf{w}^*\|^2\right] \geq \Omega\left({}^{\|\mathbf{f}\|^2}\!/_n + {}^{\|\mathbf{f}\|^2d^2}\!/_{n^2}\right)$$

■