Active Heteroscedastic Regression

Kamalika Chaudhuri [1] **Prateek Jain** [2] **Nagarajan Natarajan** [2]

Abstract

An active learner is given a model class Θ, a large sample of unlabeled data drawn from an underlying distribution and access to a labeling oracle that can provide a label for any of the unlabeled instances. The goal of the learner is to find a model $\theta \in \Theta$ that fits the data to a given accuracy while making as few label queries to the oracle as possible. In this work, we consider a theoretical analysis of the label requirement of active learning for regression under a heteroscedastic noise model.

Previous work has looked at active regression either with no model mismatch (Chaudhuri et al., 2015) or with arbitrary model mismatch (Sabato and Munos, 2014). In the first case, active learning provided no improvement even in the simple case where the unlabeled examples were drawn from Gaussians. In the second case, under arbitrary model mismatch, the algorithm either required a very high running time or a large number of labels. We provide bounds on the convergence rates of active and passive learning for heteroscedastic regression, where the noise depends on the instance. Our results illustrate that just like in binary classification, some partial knowledge of the nature of the noise can lead to significant gains in the label requirement of active learning.

1. Introduction

An active learner is given a model class Θ, a large sample of unlabeled data drawn from an underlying distribution $\mathbb{P}_{\mathbf{x}}$ and access to a labeling oracle $\mathcal{O}$ which can provide a label for any of the unlabeled instances. The goal of the learner is to find a model $\theta \in \Theta$ that fits the data to a given accuracy while making as few label queries to the oracle as possible.

Authors listed in the alphabetical order [1]University of California, San Diego [2]Microsoft Research, India. Correspondence to: Nagarajan Natarajan <t-nanata@microsoft.com>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017.

There has been a lot of theoretical literature on active learning, most of which has been in the context of binary classification in the PAC model (Balcan et al., 2009; Hanneke, 2007; Dasgupta et al., 2007; Beygelzimer et al., 2009; Awasthi et al., 2014; Zhang and Chaudhuri, 2014). For classification, the problem is known to be particularly difficult when there is no perfect classifier in the class that best fits the labeled data induced by the oracle's responses. Prior work in the PAC model has shown that the difficulty of the problem is alleviated when the "noise" is more benign – for example, when there is a ground truth classifier that induces a labeling and the oracle's responses are perturbed versions of these labels (Hanneke, 2007; Awasthi et al., 2014; Zhang and Chaudhuri, 2014; Awasthi et al., 2016) corrupted by certain kinds of noise. In particular, significant improvements in label complexity have been obtained under what is known as the Tsybakov noise conditions, which model the realistic case of noise that decreases as we move further from the decision boundary.

The case of active learning under regression however is significantly less well-understood. In particular, we only have a theoretical understanding of the two extreme cases – no noise (as in, no model mismatch) and arbitrary model mismatch. Chaudhuri et al. (2015) show that allowing the learner to actively select instances for labeling under regression with no model mismatch can only improve the convergence rates by a constant factor; moreover, in many natural cases, such as when the unlabeled data is drawn from a uniform Gaussian, there is no improvement. Sabato and Munos (2014) look at the other extreme case – when arbitrary model mismatch is allowed – and provide an algorithm that attempts to "learn" the locations of the mismatch through increasingly refined partitions of the space, and then learn a model accordingly. However if the model mismatch is allowed to be arbitrary, then this algorithm either requires an extremely refined partition leading to a very high running time, or a large number of labels. More recently, Anava and Mannor (2016) study an online learning approach for estimating heteroscedastic variances and provide general task-dependent regret bounds, but not exact parameter recovery gaurantees.

In this paper we take a step towards closing this gap in understanding by considering active regression under a realistic yet more benign "noise" model – when the variance of

the label noise depends linearly on the example x. Specifically, the oracle's response on an unlabeled example x is distributed as $\mathcal{N}(\langle x, \beta^* \rangle, \sigma_x^2)$ with $\sigma_x = \langle f^*, x \rangle$; here β^* is the unknown vector of regression coefficients and f^* is an unknown parameter vector. In classical statistics, this framework is called heteroscedastic regression, and is known to arise in econometric and medical applications.

While the usual least squares estimator for heteroscedastic regression is still statistically consistent, we find that even in the passive learning case, optimal convergence rates for heteroscedastic regression are not known. We thus begin with a convergence analysis of heteroscedastic regression for passive learning when the distribution $\mathbb{P}_{\mathbf{x}}$ over the unlabeled examples is a spherical Gaussian (in d dimensions). We show that even in this very simple case, the usual least squares estimator is sub-optimal, even when the noise model f^* is known to the learner. Instead, we propose a weighted least squares estimator, and show that its rate of convergence is $\tilde{O}\big(\|f^*\|^2(1/n + d^2/n^2)\big)$ when the noise model is known, and $\tilde{O}\big(\|f^*\|^2(d/n)\big)$ when it needs to be estimated from the data; here, n denotes the number of *labeled* examples used to obtain the estimator. The latter matches the convergence rates of the least squares estimator for the usual homoskedastic linear regression, where $\|f^*\|^2$ plays the role of the variance σ^2.

We next turn to active heteroscedastic regression and propose a two-stage active estimator. We show that when the noise model is known, the convergence rate of active heteroscedastic regression is $\tilde{O}\big(\|f^*\|^2(1/n+d^2/n^4)\big)$, a small improvement over passive. However, in the more realistic case where the noise model is unknown, the rates become $O\big(\|f^*\|^2(1/n+d^2/n^2)\big)$, which improves over the passive estimator by a factor of d. Our results extend to the case when the distribution $\mathbb{P}_{\mathbf{x}}$ over unlabeled examples is an arbitrary Gaussian with covariance matrix Σ and the norm used is the Σ norm. Our work illustrates that just like binary classification, even a partial knowledge of the nature of the model mismatch significantly helps the label complexity of active learning.

Our work is just a first step towards a study of active maximum likelihood estimation under controlled yet realistic forms of noise. There are several avenues for future work. For simplicity, the convergence bounds we present relate to the case when the distribution $\mathbb{P}_{\mathbf{x}}$ is a Gaussian. An open problem is to combine our techniques with the techniques of (Chaudhuri et al., 2015) and establish convergence rates for general unlabeled distributions. Another interesting line of future work is to come up with other, realistic noise models that apply to maximum likelihood estimation problems such as regression and logistic regression, and determine when active learning can help under these noise models. Summary of our main results in this work is given in Table 1. We conclude the paper presenting simulations supporting our theoretical bounds as well as experiments on real-world data.

	NOISE MODEL KNOWN	NOISE MODEL ESTIMATED
PASSIVE	$\tilde{O}\big(\|f^*\|^2(\frac{1}{n} + \frac{d^2}{n^2})\big)$	$\tilde{O}\big(\|f^*\|^2(\frac{d}{n})\big)$
ACTIVE	$\tilde{O}\big(\|f^*\|^2(\frac{1}{n} + \frac{d^2}{n^4})\big)$	$O\big(\|f^*\|^2(\frac{1}{n} + \frac{d^2}{n^2})\big)$

Table 1. Summary of our results: Rates for convergence of estimators, i.e. $\|\widehat{\beta} - \beta^*\|_2^2$, under the heteroscedastic noise model (2). Here, d is the data dimensionality and n is the number of labeled examples used for estimation.

2. Problem Setup and Preliminaries

Let $\mathbf{x}$ denote instances in $\mathbb{R}^d$. Let $\mathbb{P}_{\mathbf{x}}$ denote a fixed unknown distribution over instances $\mathbf{x}$. The response y is generated according to the model: $y = \langle \beta^*, \mathbf{x} \rangle + \eta_{\mathbf{x}}$, where $\eta_{\mathbf{x}}$ denotes *instance-dependent* corruption, and β^* is the ground-truth parameter. In this work, we consider the following heteroscedastic model:

$$\eta_{\mathbf{x}} \sim \mathcal{N}(0, \sigma^2(\mathbf{x})) \,, \tag{1}$$

with a standard parametric model for heteroscedastic noise given by a linear model:

$$\eta_{\mathbf{x}} \sim \mathcal{N}(0, \langle f^*, \mathbf{x} \rangle^2) \,, \tag{2}$$

for some unknown $f^* \neq \beta^*$. Each response is independently corrupted via (2). The goal is to recover β^* using instances drawn from $\mathbb{P}_{\mathbf{x}}$ and their responses sampled from $\mathcal{N}(\langle \beta^*, \mathbf{x} \rangle, \langle f^*, \mathbf{x} \rangle^2)$.

Remark 1. *The noise $\eta_{\mathbf{x}}$ can be sampled from any sub-Gaussian distribution with $\mathbb{E}[\eta_{\mathbf{x}}] = 0$ and bounded second moment $\mathbb{E}[\eta_{\mathbf{x}}^2] \leq \sigma^2$ (for some constant σ). For simplicity, we will consider the Gaussian model (1).*

Our approach is based on maximum likelihood estimator (MLE) for regression. In the homoscedastic setting (i.e. $\sigma(\mathbf{x})^2 = \sigma$ for all $\mathbf{x}$ in (1)), MLE is known to give minimax optimal rates[1]. The standard least squares estimator computed on n iid training instances $(\mathbf{x}_i, y_i)$, $i = 1, 2, \ldots, n$ is given by:

$$\widehat{\beta}_{\text{LS}} = \left(\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T \right)^{-1} \sum_{i=1}^{n} \mathbf{x}_i y_i \,, \tag{3}$$

[1] A notable exception is the Stein's estimator that may do better for high dimensional spaces (Stein et al., 1956)

and is the solution to the minimization problem:

$$\widehat{\beta}_{\text{LS}} = \arg\min_{\beta} \sum_{i=1}^{n} (\langle \beta, \mathbf{x}_i \rangle - y_i)^2.$$

In the heteroscedastic setting, it is easy to show that the standard least squares estimator is consistent.

Remark 2. *Standard least squares estimator is consistent for the heteroscedastic noise model (2): Assuming $\mathbf{x}_i \in \mathbb{R}^d, i = 1, 2, \ldots, n$ are drawn iid from the standard multivariate Gaussian, we have the rate:*

$$\|\widehat{\beta}_{\text{LS}} - \beta^*\|_2^2 = O\left(\|f^*\|^2 \frac{d}{n} \right).$$

While the estimator (3) is consistent, it does not exploit the knowledge of the noise model, and does not give better rates even when the noise model f^* is known exactly. We look at the maximum likelihood estimator for the heteroscedastic noise (1); which is given by the *weighted* least squares estimator (or sometimes referred to as *generalized* least squares estimator):

$$\widehat{\beta}_{\text{GLS}} = \left(\sum_{i=1}^{n} w_i \mathbf{x}_i \mathbf{x}_i^T \right)^{-1} \sum_{i=1}^{n} w_i \mathbf{x}_i y_i, \tag{4}$$

where $w_i = \frac{1}{\sigma(\mathbf{x}_i)^2}$. When the weights are known, it has been shown that the weighted estimator is the "correct" estimator to study; in particular, it is the minimum variance unbiased linear estimator (Theorem 10.7, Greene (2002)). However, we do not know of strong learning rate guarantees for the weighted least squares model in general, or in particular for the model (2), compared to the ordinary least squares estimator. This raises two important questions for which we provide answers in the subsequent sections.

1. What are the rates of convergence of the maximum likelihood estimator for the heteroscedastic model when the noise model, aka, f^* is unknown?
2. Can we achieve a better label requirement via active learning?

The problem is formally stated as follows. Given a set of m instances $\mathcal{U} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\}$ sampled i.i.d. from the underlying $\mathbb{P}_{\mathbf{x}}$, a *label budget* $n \leq m$, and access to label oracle $\mathcal{O}$ that generates responses y_i according to the heteroscedastic noise model (2), we want an estimator $\widehat{\beta}$ of the regression model parameter β^* such that the estimation error is small, i.e. $\|\widehat{\beta} - \beta^*\|_2 \leq O(\epsilon)$.

Remark 3. *Existing active regression methods (Sabato and Munos, 2014; Chaudhuri et al., 2015) do not consider the heteroscedastic noise model. Note that when f^* is known exactly, one can reduce heteroscedastic model to a homoscedastic model, by scaling instances $\mathbf{x}$ and their responses by $1/\langle f^*, \mathbf{x} \rangle$. However, we still may not be able to apply the existing active learning results to the transformed problem, as the modified data distribution may no longer satisfy required nice properties. The resulting estimators do not yield advantages over passive learning, even in simple cases such as when $\mathbb{P}_{\mathbf{x}}$ is spherical Gaussian.*

Notation. I_d denotes the identity matrix of size d. We use bold letters to denote vectors and capital letters to denote matrices.

3. Basic Idea: Noise Model is Known Exactly

To motivate our approach, we begin with the basic heteroscedastic setting: when f^* is known exactly in (2). Even in this arguably simple setting, the rates for passive and active learning are *a priori* not clear, and the exercise turns out to be non-trivial. The results and the analysis here help gain insights into label complexities achievable via passive and active learning strategies.

In the standard (passive) learning setting, we sample n instances uniformly from the set $\mathcal{U}$ and compute the maximum likelihood estimator given in (4) with weights set to $w_i = 1/\langle f^*, \mathbf{x}_i \rangle^2$. The procedure is given in Algorithm 1. The resulting estimator is unbiased, i.e. $\mathbb{E}[\widehat{\beta}_{\text{GLS}}|X] = \beta^*$. Let W denote the diagonal weighting matrix with $W_{ii} = w_i$. The variance of the estimator is given by: $\text{Var}(\widehat{\beta}_{\text{GLS}}|X) = (X^T W X)^{-1}$. The question of interest is if and when the weighted estimator $\widehat{\beta}_{\text{GLS}}$ is qualitatively better than the ordinary least squares estimator $\widehat{\beta}_{\text{LS}}$. The following theorem shows that the variance of the latter, and in turn the estimation error, can be potentially much larger; and in particular, the difference between their estimation errors is at least a factor of dimensionality d.

Theorem 1 (Passive Regression With Noise Oracle)**.** *Let $\widehat{\beta}_{\text{GLS}}$ denote the estimator in (4) (or the output of Algorithm 1) where $\mathbf{x}_i \sim \mathcal{N}(0, I_d)$ i.i.d., with label budget $n \geq 2d \ln d$ and $\widehat{\beta}_{\text{LS}}$ denote the ordinary least squares estimator (3). There exist positive constants C' and $c \geq 1$ such that, with probability at least $1 - \frac{d}{n^c}$, both the statements hold:*

$$\begin{aligned} \|\widehat{\beta}_{\text{GLS}} - \beta^*\|_2^2 &\leq C' \|f^*\|^2 \left(\frac{1}{n} + \frac{d^2 \ln n}{n^2} \right), \\ \|\widehat{\beta}_{\text{LS}} - \beta^*\|_2^2 &= \Theta\left(\|f^*\|^2 \frac{d}{n} \right). \end{aligned}$$

Remark 4. *We present the results for instances sampled from $\mathcal{N}(0, I_d)$ for clarity. The estimation error bounds can be naturally extended to the case of Gaussian distribution with arbitrary covariance matrix Σ. In this case,*

the bounds (in Theorem 1, for example) continue to hold for the estimation error measured w.r.t. Σ, i.e. $(\widehat{\beta} - \beta^*)^T\Sigma(\widehat{\beta} - \beta^*)$. Furthermore, with some calculations, we can obtain analogous bounds for sub-Gaussian distributions, with distribution-specific constants featuring in the resulting bounds.

Remark 5. *In Theorem 1, when $n > d$, d^2/n^2 term is the lower-order term, and thus, up to constants, the error of the weighted least squares estimator is* at most $\|f^*\|^2(1/n)$ *while that of the ordinary least squares estimator is* at least $\|f^*\|^2(d/n)$. *Thus, if the noise model is known, the weighted least squares estimator can give a factor of d improvement in convergence rate.*

Remark 6 (Technical challenges)**.** *The proofs of key results in this paper involve controlling quantities such as sum of ratios of Gaussian random variables, ratios of chi-squared random variables, etc. which do not even have expectation, let alone higher moments; so, standard concentration arguments cannot be made. However, in many of these cases, we can show that our error bounds hold with sufficiently high probability.*

The following lemma is key to showing Theorem 1; the proof sketch illustrates some of the aforementioned technical challenges. Unlike typical results in this domain, which bound $\mathbf{tr}(A^{-1})$ by providing concentration bounds for A, we bound $\mathbf{tr}(A^{-1})$ by providing lower bound on each eigenvalue of A.

Lemma 1. *Let $X \in \mathbb{R}^{n\times d}$ where the rows $\mathbf{x}_i$ are sampled i.i.d. from $\mathcal{N}(\mathbf{0}, I_d)$. Assume $n \geq 2d\ln d$. Let W denote a diagonal matrix, with $W_{ii} = 1/\langle \mathbf{x}_i, f\rangle^2$, for fixed $f \in \mathbb{R}^d$. Let $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_d$ denote the eigenvalues of X^TWX. Then, with probability at least $(1 - \frac{d}{n^c})$:*

1. $\sigma_d(X^TWX) \geq \frac{n}{\|f\|^2}$ *and*
2. $\sigma_i(X^TWX) \geq C'\frac{n^2}{d\|f\|^2\ln n}$ *for $i = 1, 2, \ldots, d-1$,*

where $c \geq 1$ and $C' > 0$ are constants.

Proof. We give a sketch of the proof here (See Appendix B.2 for details). To show a lower bound on the smallest eigenvalue, we first show that the smallest eigenvector is very close to f, with sufficiently large probability. To do so, we exploit the fact that the smallest eigenvalue is at most $n/\|f\|^2$ which can be readily seen. For the second part, we consider the variational characterization of $d-1$st singular value given by:

$$\sigma_{d-1}(X^TWX) = \max_{U:\dim(U)=d-1} \min_{\mathbf{v}\in U, \|\mathbf{v}\|=1} \mathbf{v}^T X^T W X \mathbf{v}.$$

We look at the particular subspace that is orthogonal to f^* to get the desired upper bound. One key challenge here is controlling quantity of the form $\sum_{i=1}^n \frac{g_i^2}{h_i^2}$ where g_i and h_i are iid Gaussian variables. We use a blocking technique based on the magnitude of $\langle f, \mathbf{x}_i\rangle$, and lower bound the quantity with just the first block (as all the quantities involved are positive). This requires proving a bound on the order statistics of iid Gaussian random variables (Lemma 7 in Appendix A). □

Theorem 1 shows that weighting "clean" instances (i.e. $\langle f^*, \mathbf{x}\rangle \approx 0$) much more than "noisy" instances yields a highly accurate estimator of β^*. But can we instead prefer querying labels on instances where we know *a priori* the response will be relatively noise-free? This motivates a simple active learning solution — in principle, if we actually *know* f^*, we could query the labels of instances with low noise, and hope to get an accurate estimator. The active learning procedure is given in Algorithm 2. Besides label budget n, it takes another parameter τ as input, which is a threshold on the noise level.

We state the convergence for Algorithm 2 below:

Theorem 2 (Active Regression with Noise Oracle)**.** *Consider the output estimator $\widehat{\beta}$ of Algorithm 2, with input label budget $n \geq 2d\ln d$, unlabeled set $\mathcal{U}$ with $|\mathcal{U}| = m$ and $\mathbf{x}_i \sim \mathcal{N}(0, I_d)$ i.i.d., and $\tau = 2n/m$. Then, with probability at least $1 - 1/n^c$:*

$$\|\widehat{\beta} - \beta^*\|_2^2 \leq C'\|f^*\|^2\left(\frac{1}{n} + \frac{d^2\ln n}{m^2}\right),$$

for some constants $c > 1$ and $C' > 0$.

Remark 7. *We observe that the estimation error via active learning (Theorem 2) goes to 1/n as the size of the unlabeled examples m becomes larger. Note that O(1/n) is the error for 1-dimensional problem and is much better than d^2/n^2 we get from uniform sampling.*

Remark 8. *If we have $m \geq n^2$ unlabeled samples, then we observe that active learning (Theorem 2) achieves a better convergence rate compared to that of passive learning (Theorem 1) — the lower order term in case of active learning is $O(\frac{d^2}{n^4})$ versus passive learning which is $O(\frac{d^2}{n^2})$. The convergence is superior especially when $n < d^2$ (as we also observe in simulations).*

The proof of Theorem 2 relies on two lemmas stated below.

Lemma 2. *Let $X \in \mathbb{R}^{n\times d}$ denote the design matrix whose rows $\mathbf{x}_i$ are sampled from $\mathcal{U}$ such that they satisfy $|\langle \mathbf{x}_i, f\rangle| \leq \|f\|\tau$ for fixed $f \in \mathbb{R}^d$. Assume $n \geq 2d\ln d$. Let $\sigma_1 \geq \sigma_2 \geq \ldots \sigma_d$ denote the eigenvalues of X^TX. Then, with probability at least $(1 - \frac{1}{n^c})$:*

1. $\sigma_d(X^TX) \geq n\tau^2$,
2. $\sigma_i(X^TX) \geq \frac{1}{2}n$, *for $i = 1, 2, \ldots, d-1$,*

Algorithm 1 Passive Regression With Noise Oracle

Input: Labeling oracle $\mathcal{O}$, instances $\mathcal{U} = \{\mathbf{x}_i, i \in [m]\}$, label budget n, noise model f^*
1. Choose a subset $\mathcal{L}$ of size n from $\mathcal{U}$ uniformly at random from $\mathcal{U}$. Query their labels using $\mathcal{O}$.
2. Estimate $\widehat{\beta}$ using (4) on $\mathcal{L}$, with $w_i = 1/\langle f^*, \mathbf{x}_i\rangle^2$.
Output: $\widehat{\beta}$.

Algorithm 2 Active Regression With Noise Oracle

Input: Labeling oracle $\mathcal{O}$, noise model f^*, instances $\mathcal{U} = \{\mathbf{x}_i, i \in [m]\}$, label budget n, noise tolerance τ.
1. Choose a subset $\mathcal{L}$ of size at most n from $\mathcal{U}$ with expected noise up to the given tolerance τ, i.e. for all $\mathbf{x}_i \in \mathcal{L}$, $|\langle \mathbf{x}_i, f^*\rangle| \leq \tau$. Query their labels using $\mathcal{O}$.
2. Estimate $\widehat{\beta}$ as $\widehat{\beta} = (X^T W X)^{-1} X^T W \mathbf{y}$ where $X \in \mathbb{R}^{n\times d}$ and $\mathbf{y} \in \mathbb{R}^n$, and W is a diagonal matrix with $W_{ii} = \frac{1}{\langle f^*, \mathbf{x}_i\rangle^2}$.
Output: $\widehat{\beta}$.

for some constants $C > 0$ and $c \geq 1$.

Lemma 3. *For each $\mathbf{x}_i \in \mathcal{U}$, where $|\mathcal{U}| = m$, define $g_i = \langle \mathbf{x}_i, \mathbf{z}\rangle$, for any fixed $\mathbf{z} \in \mathbb{R}^d$. Then, with probability at least $\exp(\frac{-m\tau^3}{4\|\mathbf{z}\|^2})$:*

$$\left|\left\{i : |g_i| \leq \|\mathbf{z}\|\tau\right\}\right| \geq \frac{m\tau}{2} .$$

4. Estimating Noise: Algorithms and Guarantees

In this section, we will first show that we can obtain a consistent estimator of f^*, as long as we have a reasonably good estimate of β^*. Let β_0 denote the ordinary least squares estimator of β^*, obtained by using (3), on m_1 labeled instances, chosen i.i.d. from $\mathcal{N}(0, I_d)$. The largest eigenvector of the residual-weighted empirical covariance matrix given by:

$$\widehat{S} = \frac{1}{m_1}\sum_{i=1}^{m_1}(y_i - \langle \mathbf{x}_i, \widehat{\beta}_0\rangle)^2 \mathbf{x}_i \mathbf{x}_i^T . \tag{5}$$

gives a *sufficiently* good estimate of f^*. This is established formally in the following lemma.

Lemma 4. *Let $m_1 = \Omega(d \log^3 d)$. Then, with probability at least $1 - \frac{1}{m_1^5}$, the largest eigenvector $\hat{f}$ of $\widehat{S}$ in (5) converges to f^*:*

$$\|\hat{f} - f^*\|_2^2 \leq C_1 \mathbb{E}[\|\beta_0 - \beta^*\|_2^2] + O\left(\frac{d}{m_1}\right) ,$$

for some positive constant C_1, and expectation is wrt the randomness in the estimator β_0.

We first discuss the implications of using the estimated $\hat{f}$ in order to obtain the generalized least square estimator given in (4) and then present the active learning solution.

4.1. Weighted Least Squares

We now consider a simple (passive) learning algorithm for the setting where the noise model is estimated, based on the weighted least squares solution discussed in Section 3. We first get a good estimator of f^* (as in Lemma 4), and then obtain the weighted least squares estimator: $\widehat{\beta} = (X^T \widehat{W} X)^{-1} X^T \widehat{W} \mathbf{y}$, where $\widehat{W}$ is the diagonal matrix of inverse noise variances obtained using the estimate $\hat{f}$ with a small additive offset γ. The procedure is presented in Algorithm 3.

Remark 9. *Algorithm 3 can be thought of as a special case of the well-known iterative weighted least squares (i.e. with just one iteration), that has been studied in the past (Carroll et al., 1988).*

It is well-known heuristic to first estimate the weights and then obtain the weighted estimator in practice; the approach has been widely in use for decades now in multiple communities including Econometrics and Bioinformatics (Harvey, 1976; Greene, 2002). However, we do not know of strong convergence rates for the solution. To our knowledge, the most comprehensive analysis was done by (Carroll et al., 1988). Their analysis is not directly applicable to us for reasons two-fold: (i) they focus on using a maximum likelihood estimate of the parameters in the heteroscedastic noise model, and does not apply to our noise model (2), and (ii) their analysis relies the noise being smooth (for obtaining tighter Taylor series approximation). More importantly, their analysis conceals a lot of significant factors in both d and n, and the resulting statements about convergence rates are not useful (See Appendix C).

Theorem 3. *Consider the output estimator $\widehat{\beta}$ of Algorithm 3, with label budget $n \geq 2d \ln d$ and offset $\gamma^2 = \sqrt{\frac{d}{n}}$. Then, with probability at least $1 - 1/n^c$:*

$$\|\widehat{\beta} - \beta^*\|_2^2 \leq C\|f^*\|^2 \frac{d \ln^4 n}{n} ,$$

Algorithm 3 Least Squares with Estimated Weights

Input: Labeling oracle $\mathcal{O}$, unlabeled samples $\mathcal{U} = \{\mathbf{x}_i, i \in [m]\}$, label budget n, parameter m_1, offset γ.
1. Draw m_1 examples uniformly at random from $\mathcal{U}$ and query their labels y using $\mathcal{O}$.
2. Estimate $\widehat{\beta}_0$ by solving $\mathbf{y} \approx X\widehat{\beta}_0$ where $X \in \mathbb{R}^{m_1 \times d}$ has $\mathbf{x}_i$ as rows and $\mathbf{y} \in \mathbb{R}^{m_1}$ is the vector of labels.
3. Draw a subset $\mathcal{L}$ of n examples uniformly at random from $\mathcal{U}$. Form $X \in \mathbb{R}^{n \times d}$ and $\mathbf{y} \in \mathbb{R}^n$.
4. Compute $\widehat{f}$ as the largest eigenvector of the residual-weighted empirical covariance given in (5).
5. Set $\widehat{w}_i = \frac{1}{\langle \mathbf{x}_i, \widehat{f}\rangle^2 + \gamma^2}$., for $\mathbf{x}_i \in \mathcal{L}$.
6. Estimate $\widehat{\beta}$ by solving: $\widehat{\beta} = (X^T \widehat{W} X)^{-1} X^T \widehat{W} \mathbf{y}$, where $\widehat{W}$ is diagonal matrix with $\widehat{W}_{ii} = \hat{w}_i$.

Output: $\widehat{\beta}$.

for some constants $C > 0$ and $c \geq 1$.

Remark 10. *Note that the above result holds for the specific choice of γ. When $\gamma = 0$, we get the weighted least squares estimator analogous to the one used in Algorithm 1. However, when estimating weights with $\gamma = 0$, the resulting estimator $\widehat{\beta}$ has poor convergence rate. In particular, we observe empirically that the error $\|\widehat{\beta} - \beta^*\|^2$ scales as $O(\frac{d^3}{n})$.*

4.2. Active Regression

We now show that active learning can help overcome the inadequacy of the passive weighted least squares solution. The proposed active regression algorithm, presented in Algorithm 4, consists of two stages. In the first stage, we obtain an estimate $\widehat{f}$ similar to that in Algorithm 3. Note that if $\widehat{f}$ is indeed very close to f^*, $\widehat{f}$ serves as a good proxy for selecting instances whose labels are nearly noise-free. To this end, we sample instances that have sufficiently small noise: $|\langle \widehat{f}, \mathbf{x}\rangle| \leq \tau$, where τ is an input parameter to the algorithm. If $\widehat{f}$ is exact, then the algorithm reduces to the strategy outlined in Algorithm 2. Our algorithm follows the strategy of using a single-round of interaction (in light of the analysis presented in the passive learning setting) to achieve a good estimate of the underlying β^* akin to the active MLE estimation algorithm studied by Chaudhuri et al. (2015).

Lemma 5. *Let $\widehat{f}$ denote an estimator of f^* satisfying $\|\widehat{f} - f^*\|_2 \leq \Delta$. For a given τ, let $\mathcal{L}$ denote a set of $|\mathcal{L}| \geq 2d \log d$ instances sampled from m unlabeled instances $\mathcal{U}$, such that $|\langle \widehat{f}, \mathbf{x}_i\rangle| \leq \tau$, for all $\mathbf{x}_i \in \mathcal{L}$, and let y_i denote their corresponding labels. Consider the ordinary least squares estimator obtained using $\mathcal{L}$, i.e.:*

$$\widehat{\beta} = \left(\sum_{\mathbf{x}_i \in \mathcal{L}} \mathbf{x}_i \mathbf{x}_i^T \right)^{-1} \sum_{\mathbf{x}_i \in \mathcal{L}} \mathbf{x}_i y_i \ .$$

Then, with probability at least $1 - \frac{1}{n^c}$:

$$\|\widehat{\beta} - \beta^*\|_2^2 \leq C\|f^*\|^2(\tau^2 + \Delta^2)\left(\frac{1}{m\tau^3} + \frac{d-1}{m\tau}\right) .$$

for some constants $C > 0$ and $c \geq 1$.

Remark 11. *The bound in the above theorem recovers the known variance case discussed in Theorem 2, where the estimation error $\Delta^2 = 0$ and the choice of $\tau = \frac{2n}{m}$.*

Compared to the passive learning error bound in Theorem 3, we hope to get leverage — as long we can choose τ sufficiently small, and yet guarantee that the number of samples m_2 in Step 4 of Algorithm 4 is sufficiently large. The following theorem shows that this is indeed the case, and that the proposed active learning solution achieves optimal learning rate.

Theorem 4 (Active Regression with Noise Estimation)**.** *Consider the output estimator $\widehat{\beta}$ of Algorithm 4, with input label budget n, unlabeled examples $m \geq n^2$, $m_1 = \frac{n}{2}$ and $\tau = 2\sqrt{\frac{d}{n}}$. Then, we have, with probability at least $1 - 1/n^c$:*

$$\|\widehat{\beta} - \beta^*\|_2^2 \leq C\|f^*\|^2\left(\frac{1}{n} + \frac{d^2}{n^2}\right) .$$

for some constants $C > 0$ and $c > 1$.

Remark 12. *We observe that active learning (Theorem 4) has the same convergence rate for sufficiently large n, as that of the case when f^* is known exactly (Theorem 2). Note that d^2/n^2 and d^2/m^2 are lower-order terms in the compared bounds.*

Remark 13. *Unlike in the case when noise model was known (Theorem 2), here we can not do better even with infinite unlabeled examples. The source of trouble is the estimation error in $\widehat{f}$, so beyond a point even active learning does not provide improvement. Note that we do not compute weighted least squares estimator in the final step of Algorithm 4 unlike in Algorithm 2, for the same reason.*

5. Simulations

We now present simulations that support the convergence bounds developed in this work. The setup is as follows. We sample unlabeled instances $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$ from $\mathcal{N}(0, I_d)$. Labels are generated according to the heteroscedastic model: $y_i = \langle \beta^*, \mathbf{x}_i\rangle + g_i\langle f^*, \mathbf{x}_i\rangle$, where

Algorithm 4 Active Regression

Input: Labeling oracle $\mathcal{O}$, unlabeled samples $\mathcal{U} = \{\mathbf{x}_i, i \in [m]\}$, label budget n, parameters m_1, τ .
1. Draw m_1 examples uniformly at random from $\mathcal{U}$ and query their labels y using $\mathcal{O}$.
2. Estimate $\widehat{\beta}_0$ by solving $\mathbf{y} \approx X\widehat{\beta}_0$ where $X \in \mathbb{R}^{m_1 \times d}$ and $\mathbf{y} \in \mathbb{R}^{m_1}$.
3. Compute $\widehat{f}$ as the largest eigenvector of $\widehat{S}$ given in (5).
4. Choose a subset $\mathcal{L}$ of $m_2 = n - m_1$ instances from $\mathcal{U}$ with estimated noise variance up to tolerance τ^2, i.e. for all $\mathbf{x}_i \in \mathcal{L}$, $|\langle \mathbf{x}_i, \widehat{f} \rangle|^2 \leq \tau^2$. Query their labels using $\mathcal{O}$.
5. Estimate $\widehat{\beta}$ as $\widehat{\beta} = (X^T X)^{-1} X^T \mathbf{y}$ where $X \in \mathbb{R}^{m_2 \times d}$ and $\mathbf{y} \in \mathbb{R}^{m_2}$.
Output: $\widehat{\beta}$.

g_i are iid standard Gaussian random variables. We fix $\|f^*\|_2 = 1$ and $d = 10$. We look at how the model estimation error (in case of Algorithms 1 and 2) $\|\widehat{\beta} - \beta^*\|$ decays as a function of the label budget n ($m = 2n^2$ for all the simulations). We also check the estimation error of the noise model in case of Algorithms 3 and 4.

The results for convergence of model estimation when the noise model is known are presented in Figure 1 (a)-(d). In passive learning, the bounds in Theorem 1 suggest that when $n \leq d^2$, $\|\beta^* - \widehat{\beta}\| = O(\frac{d}{n})$; but once $n > d^2$, we get a convergence of $O(1/\sqrt{n})$. We observe that the result in Figure 1 (a) closely matches the given bounds[2]. In case of active learning, the bounds in Theorem 2, for the case when $m \geq n^2$, suggest that we get an error rate of $\|\beta^* - \widehat{\beta}\| = O(\frac{d}{n^2})$. We observe a similar phenomenon in the Figure 1 (b). Turning to the noise estimation setting for passive learning, we see in Figure 1 (c) that the estimation error of β^* as well as f^* decay as $\sqrt{d/n}$ (as suggested by Theorem 3); for active learning, we see in Figure 1 (d) that the estimation error of β^* is noticeably better, in particular, better than that of f^*, and approaches $1/\sqrt{n}$ as n becomes larger than d^2.

We also study the performance of the algorithms on two real-world datasets from UCI: (1) WINE QUALITY with $m = 6500$ and $d = 11$, and (2) MSD (a subset of the million song dataset) with $m = 515345$ and $d = 90$. For each dataset, we create a 70-30 train-test split, and learn the best linear regressor using ordinary least squares, which forms our β^*. We then sample labels using β^* and a simulated heteroscedastic noise f^*. We compare active and passive learning algorithms on the root mean square error (RMSE) obtained on the test set. In Figure 1 (e), we see that active learning with noise estimation gives a significant reduction in RMSE early on for WINE QUALITY. We also see that weighted least squares gives slight benefit over ordinary least squares. On MSD dataset [3], again we observe that our active learning algorithm consistently achieves a marginal reduction in RMSE as the number of labeled examples increases.

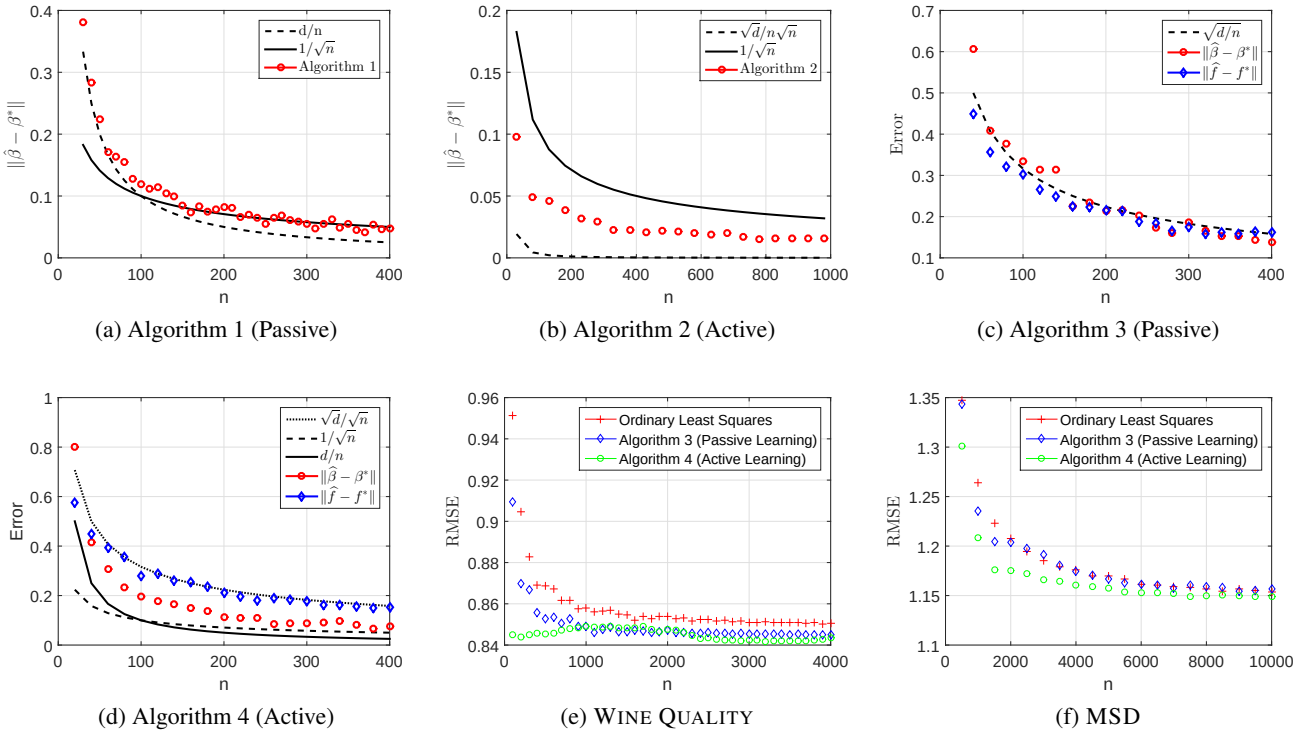
[2]For better resolution, we plot $\|\beta^* - \widehat{\beta}\|$ rather than $\|\beta^* - \widehat{\beta}\|^2$ given in the theorem statements

[3]here, the response variable is the year of the song; we make the response mean zero in our experiments

6. Conclusions and Future Work

In conclusion, we consider active regression under a heteroscedastic noise model. Previous work has looked at active regression either with no model mismatch (Chaudhuri et al., 2015) or arbitrary model mismatch (Sabato and Munos, 2014). In the first case, active learning provided no improvement even in the simple case where the unlabeled examples were drawn from Gaussians. In the second case, under arbitrary model mismatch, the algorithm either required a very high running time or a large number of labels. We provide bounds on the convergence rates of active and passive learning for heteroscedastic regression. Our results illustrate that just like in binary classification, some partial knowledge of the nature of the noise has the potential to lead to significant gains in the label requirement of active learning.

There are several avenues for future work. For simplicity, the convergence bounds we present relate to the case when the distribution $\mathbb{P}_{\mathbf{x}}$ over unlabeled examples is a Gaussian. An open problem is to combine our techniques with the techniques of (Chaudhuri et al., 2015) and establish convergence rates for general unlabeled distributions. Another interesting line of future work is to come up with other, realistic noise models that apply to maximum likelihood estimation problems such as regression and logistic regression, and determine when active learning can help under these noise models.



(a) Algorithm 1 (Passive) (b) Algorithm 2 (Active) (c) Algorithm 3 (Passive)

(d) Algorithm 4 (Active) (e) WINE QUALITY (f) MSD

Figure 1. Plots (a)-(b): convergence of model (β^*) estimation error, when the noise model is known. Plots (c)-(d): convergence of model (β^*) estimation error as well as noise parameter (f^*) estimation error, when the noise model is estimated. Plots (e)-(f): RMSE on test data for two real-world datasets.

References

Oren Anava and Shie Mannor. Heteroscedastic sequences: Beyond gaussianity. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 755–763, 2016.

Pranjal Awasthi, Maria-Florina Balcan, and Philip M. Long. The power of localization for efficiently learning linear separators with noise. In *Symposium on Theory of Computing, STOC 2014, New York, NY, USA, May 31 - June 03, 2014*, pages 449–458, 2014.

Pranjal Awasthi, Maria-Florina Balcan, Nika Haghtalab, and Hongyang Zhang. Learning and 1-bit compressed sensing under asymmetric noise. In *Proceedings of the 29th Conference on Learning Theory, COLT 2016, New York, USA, June 23-26, 2016*, pages 152–192, 2016.

M.-F. Balcan, A. Beygelzimer, and J. Langford. Agnostic active learning. *J. Comput. Syst. Sci.*, 75(1):78–89, 2009.

A. Beygelzimer, S. Dasgupta, and J. Langford. Importance weighted active learning. In *ICML*, 2009.

Raymond J Carroll, CF Jeff Wu, and David Ruppert. The effect of estimating weights in weighted least squares. *Journal of the American Statistical Association*, 83 (404):1045–1054, 1988.

Kamalika Chaudhuri, Sham M Kakade, Praneeth Netrapalli, and Sujay Sanghavi. Convergence rates of active learning for maximum likelihood estimation. In *Advances in Neural Information Processing Systems*, pages 1090–1098, 2015.

S. Dasgupta, D. Hsu, and C. Monteleoni. A general agnostic active learning algorithm. In *NIPS*, 2007.

Yehoram Gordon, Alexander Litvak, Carsten Schütt, and Elisabeth Werner. On the minimum of several random variables. *Proceedings of the American Mathematical Society*, 134(12):3665–3675, 2006.

William H Greene. *Econometric analysis*. Prentice Hall, 2002.

S. Hanneke. A bound on the label complexity of agnostic active learning. In *ICML*, 2007.

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

Prateek Jain and Ambuj Tewari. Alternating minimization for regression problems with vector-valued outputs. In *Advances in Neural Information Processing Systems*, pages 1126–1134, 2015.

Sivan Sabato and Remi Munos. Active regression by stratification. In *Advances in Neural Information Processing Systems*, pages 469–477, 2014.

Charles Stein et al. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley symposium on mathematical statistics and probability*, volume 1, pages 197–206, 1956.

Chicheng Zhang and Kamalika Chaudhuri. Beyond disagreement-based agnostic active learning. In *Neural Information Processing Systems (NIPS)*, 2014.

A. Technical Lemmas

Lemma 6. *Let $\mathbf{z} \in \mathbb{R}^d$ be a fixed vector. Let $U_{\mathbf{z}^\perp}$ denote the subspace orthogonal to $\mathbf{z}$. Assume $n' \geq Cd\log d$. Let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{n'}$ denote random Gaussian vectors from $\mathcal{N}(0, I_d)$ such that $U_{\mathbf{z}^\perp}\mathbf{x}_i$ are mutually independent. Then with probability at least $1 - \exp(-Cn')$, the following holds for all $\mathbf{v} \in U_{\mathbf{z}^\perp}$ such that $\|\mathbf{v}\|_2 = 1$:*

$$\frac{1}{2}n' \leq \mathbf{v}^T \sum_{i=1}^{n'} \left(I_d - \frac{\mathbf{z}\mathbf{z}^T}{\|\mathbf{z}\|^2}\right)\mathbf{x}_i\mathbf{x}_i^T\left(I_d - \frac{\mathbf{z}\mathbf{z}^T}{\|\mathbf{z}\|^2}\right)\mathbf{v} \leq 2n' .$$

Proof. First, note that $\tilde{\mathbf{x}}_i := \left(I - \frac{\mathbf{z}\mathbf{z}^T}{\|\mathbf{z}\|^2}\right)\mathbf{x}_i$ are iid Gaussian random variables drawn from $\mathcal{N}(0, U_{\mathbf{z}^\perp})$. We can apply Lemma 14 of (Jain and Tewari, 2015) to get the statement of the lemma. □

Lemma 7. *Let $R_{(1)} \leq R_{(2)} \leq \cdots \leq R_{(n)}$ be the order statistics of absolute values of a standard Gaussian sample $R_1, R_2, \ldots, R_n$. Then, with probability at least $1 - 1/n^{10}$,*

$$R_{(k)} \leq C_U \frac{k}{n}\ln n,$$

for some positive constant C_U.

Proof. Define the scaled random variable $\tilde{R}_{(k)} = \frac{R_{(k)}}{(k/n)}$. Let $\mu = \mathbb{E}[\tilde{R}_{(k)}]$. For a fixed $p \geq \log n$, and for any $1 \leq k \leq n$, consider the moment:

$$\begin{aligned}\mathbb{E}[|\tilde{R}_{(k)} - \mu|^p] &= \mathbb{E}\left[\left|\sum_{l=1}^{p}(-1)^l\binom{p}{l}\tilde{R}_{(k)}^l\mu^{p-l}\right|\right] \\ &\leq \sum_{l=1}^{p}\binom{p}{l}\mathbb{E}[\tilde{R}_{(k)}^l]\mu^{p-l} \\ &= \sum_{l=1}^{p}\binom{p}{l}\left((\mathbb{E}[\tilde{R}_{(k)}^l])^{1/l}\right)^l\mu^{p-l} \qquad (6)\end{aligned}$$

From Theorem 7 of (Gordon et al., 2006), we have:

$$(\mathbb{E}[\tilde{R}_{(k)}^l])^{1/l} \leq 4\sqrt{\pi}\big(l + \ln(k+1)\big) \leq 4\sqrt{\pi}(p + \ln n).$$

We also know from (Gordon et al., 2006) that:

$$\mu = \mathbb{E}[\tilde{R}_{(k)}] \leq C\ln k \leq C\ln n,$$

for some positive constant C. Substituting these upper bounds in (6), we get:

$$\begin{aligned}\mathbb{E}[|\tilde{R}_{(k)} - \mu|^p] &\leq \sum_{l=1}^{p}\binom{p}{l}\big(4\sqrt{\pi}(p+\ln n)\big)^l(C\ln n)^{p-l} \\ &= \Big(\big(4\sqrt{\pi}(p+\ln n) + C\ln n\Big)^p \\ &\leq \Big(\big(8\sqrt{\pi}+C)p\Big)^p\end{aligned}$$

Finally, by applying Markov inequality, for any $t > 0$:

$$P(|\tilde{R}_{(k)} - \mu|^p \geq t) \leq \frac{\mathbb{E}[|\tilde{R}_{(k)} - \mu|^p]}{t}$$

or

$$P(|\tilde{R}_{(k)} - \mu| \geq \tilde{t}) \leq \frac{\mathbb{E}[|\tilde{R}_{(k)} - \mu|^p]}{\tilde{t}^p}.$$

Choosing $p = 10\ln n$ and $\tilde{t} = e\big(8\sqrt{\pi}+C\big)p$, we get:

$$P(|\tilde{R}_{(k)} - \mu| \geq e\big(80\sqrt{\pi}+C\big)\ln n) \leq e^{-10\ln n} = \frac{1}{n^{10}}$$

Note that $P\big(\tilde{R}_{(k)} \geq 10e\big(8\sqrt{\pi}+C\big)\ln n\big) \leq P\big(|\tilde{R}_{(k)} - \mu| \geq 10e\big(8\sqrt{\pi}+C\big)\ln n\big)$. Finally, observing that $\tilde{R}_{(k)} \geq 10e\big(8\sqrt{\pi}+C\big)\ln n \iff R_{(k)} \geq 10e\big(8\sqrt{\pi}+C\big)\ln n\frac{k}{n}$, we get the statement of the lemma with $C_U = 10e\big(8\sqrt{\pi}+C\big)$. □

B. Proofs

B.1. Proof of Theorem 1

(I) Consider the weighted least squares estimate:

$$\begin{aligned}\widehat{\beta}_{\text{GLS}} &= (X^TWX)^{-1}X^TW\mathbf{y} \\ &= (X^TWX)^{-1}\sum_{i=1}^{n} w_i(\langle\beta^*, \mathbf{x}_i\rangle + g_i\langle\mathbf{x}_i, f^*\rangle)\mathbf{x}_i,\end{aligned}$$

where W is the diagonal matrix with $w_i = 1/\langle f^*, \mathbf{x}_i\rangle^2$ along the diagonal, g_i are i.i.d. $\mathcal{N}(0,1)$ random variables. So we have:

$$\begin{aligned}\widehat{\beta}_{\text{GLS}} - \beta^* &= (X^TWX)^{-1}\sum_{i=1}^{n}\frac{g_i}{\langle f^*, \mathbf{x}_i\rangle}\mathbf{x}_i \\ \|\widehat{\beta}_{\text{GLS}} - \beta^*\|_2^2 &= \mathbf{tr}\Big((X^TWX)^{-2}X^TW^{0.5}\mathbf{g}\mathbf{g}^TW^{0.5}X\Big),\end{aligned}$$

Note that because $\mathbb{E}[\mathbf{g}\mathbf{g}^T] = I_n$ (where the expectation is wrt. to the randomness in the labels given by the oracle $\mathcal{O}$) and $\mathbf{tr}$ is linear operator, we have:

$$\mathbb{E}\|\widehat{\beta}_{\text{GLS}} - \beta^*\|_2^2 = \mathbf{tr}\Big((X^TWX)^{-1}\Big).$$

Consider X^TWX. We can apply Lemma 1 to lower-bound the $(d-1)$ smallest eigenvalues of this matrix by $O(n^2/(\|f^*\|^2 d\ln n))$ and the largest eigenvalue by $O(n/\|f^*\|^2)$, with probability at least $(1 - \frac{d}{n^c})$. This implies an upper-bound for the eigenvalues of $(X^TWX)^{-1}$, and in turn its trace can be bounded by $C'\|f^*\|_2^2\Big(\frac{1}{n} + \frac{(d-1)d\ln n}{n^2}\Big)$, for some constant $C' > 0$. The proof is complete.

B.2. Proof of Lemma 1

1. By definition, the smallest singular value,

$$\begin{aligned}\sigma_d(X^TWX) &= \inf_{\mathbf{v}\in\mathbb{R}^d,\|\mathbf{v}\|=1} \mathbf{v}^T X^T W X \mathbf{v} \\ &\leq \frac{f^T}{\|f\|}\sum_{i=1}^n \frac{1}{(\mathbf{x}_i^T f)^2}\mathbf{x}_i\mathbf{x}_i^T\frac{f}{\|f\|} \\ &= \sum_{i=1}^n \frac{1}{\|f\|^2} = \frac{n}{\|f\|^2} \end{aligned} \tag{7}$$

Let $\mathbf{v}^*$ denote the smallest eigenvector of X^TWX. Write $\mathbf{v}^*$ as:

$$\mathbf{v}^* = \sqrt{1-\alpha_d^2}\mathbf{v}_\perp + \alpha_d f/\|f\|$$

where $\mathbf{v}_\perp$ denotes the component along the subspace orthogonal to f, and $\alpha_d = \mathbf{v}^{*T} f/\|f\|$. Now:

$$\begin{aligned}\mathbf{v}^{*T}X^TWX\mathbf{v}^* &= \alpha_d^2 \,.\, n/\|f\|^2 \\ &+ (1-\alpha_d^2)\sum_{i=1}^n \frac{(\mathbf{v}_\perp^T\mathbf{x}_i)^2}{(f^T\mathbf{x}_i)^2} \\ &+ 2\alpha_d\sqrt{1-\alpha_d^2}\sum_{i=1}^n \frac{\mathbf{v}_\perp^T\mathbf{x}_i}{f^T\mathbf{x}_i} \\ &\leq n/\|f\|^2\end{aligned}$$

where the inequality is due to the upper bound in (7). The second term in the above equation can be lower bounded with probability at least $1-1/n$ by $(1-\alpha_d^2)n^2d/\|f\|^2$. To lower bound the summation in the third term as $\sum_{i=1}^n \frac{\mathbf{v}_\perp^T\mathbf{x}_i}{f^T\mathbf{x}_i} \leq \sqrt{\sum_{i=1}^n \frac{1}{(f^T\mathbf{x}_i)^2}}\sqrt{\sum_{i=1}^n(\mathbf{v}_\perp^T\mathbf{x}_i)^2} \leq \sqrt{2n}\sqrt{\sum_{i=1}^n \frac{1}{(f^T\mathbf{x}_i)^2}}$, with probability at least $1-\exp(-2n)$ (Using Lemma 6). We conjecture that with probability at least $1-1/n$, $\sqrt{\sum_{i=1}^n \frac{1}{(f^T\mathbf{x}_i)^2}} \leq n/\|f\|^2$ (note that it holds in expectation, shown by Gordon et al. (2006)). So we have, with probability at least $1-2/n$:

$$\begin{aligned}\alpha_d^2 n + (1-\alpha_d^2)n^2 d - 2\alpha_d\sqrt{1-\alpha_d^2}n\sqrt{2n} \\ \leq \mathbf{v}^{*T}X^TWX\mathbf{v}^* \leq n\end{aligned}$$

For the above inequality to hold, it must be the case that $\alpha_d^2 \geq 1 - 16\frac{1}{nd^2}$.

2. Consider the variational characterization of the second smallest singular value σ_{d-1} given by:

$$\sigma_{d-1}(X^TWX) = \max_{U:\dim(U)=d-1}\ \min_{\mathbf{v}\in U,\|\mathbf{v}\|=1} \mathbf{v}^TX^TWX\mathbf{v}.$$

Consider the particular $d-1$ dimensional subspace $U_{f^\perp} = \{\mathbf{v}\in\mathbb{R}^d \mid \mathbf{v}^Tf = 0\}$. Note that the projection matrix corresponding to $U_{f^\perp}$ is given by $\left(I_d - \frac{ff^T}{\|f\|^2}\right)$. For any vector $\mathbf{v}\in U_{f^\perp}$, we have:

$$\mathbf{v}^TX^TWX\mathbf{v} = \sum_{i=1}^n \frac{\mathbf{v}^T\mathbf{x}_i\mathbf{x}_i^T\mathbf{v}}{(\mathbf{x}_i^Tf)^2} \tag{8}$$

Note that $g_i = \mathbf{v}^T\mathbf{x}_i$, $i=1,2,\ldots,n$ and $h_i = \mathbf{x}_i^Tf$, $i=1,2,\ldots,n$ are iid Gaussian random variables; in particular, as $\mathbf{v}$ is in the orthogonal subspace of f, g_i and h_i are independent of each other. We will now lower bound $\sum_{i=1}^n \frac{(\mathbf{v}^T\mathbf{x}_i)^2}{(\mathbf{x}_i^Tf)^2}$, by dividing $\mathbf{x}_i, \mathbf{x}_2, \ldots, \mathbf{x}_n$ into $\lceil\frac{n}{2d\ln n}\rceil$ batches of size $s = 2d\ln n$. Let $\mathbf{x}_{(1)}, \mathbf{x}_{(2)}, \ldots, \mathbf{x}_{(n)}$ denote the new ordering of instances, such that

$$|\mathbf{x}_{(1)}^Tf| \leq |\mathbf{x}_{(2)}^Tf| \leq \cdots \leq |\mathbf{x}_{(n)}^Tf|.$$

Let $\mathcal{B}_1$ denote the first s instances according the new ordering. Using Lemma 7, we have, with probability at least $(1-\frac{2d\ln n}{n^{10}})$:

$$\sum_{k=1}^s \frac{\mathbf{v}^T\mathbf{x}_i\mathbf{x}_i^T\mathbf{v}}{(\mathbf{x}_{(k)}^Tf)^2} \geq \frac{1}{C_U^2}\sum_{k=1}^s \frac{n^2}{\|f\|^2k^2\ln^2 n}(\mathbf{v}^T\mathbf{x}_i\mathbf{x}_i^T\mathbf{v})$$

We can replace $\mathbf{v}$ by $\left(I_d - \frac{ff^T}{\|f\|^2}\right)\mathbf{v}$ in the RHS of the above inequality, which is true by definition. Now, we can apply Lemma 6 to control the resulting quantity: with probability at least $1-\frac{1}{n^{4d}}$, over all $\mathbf{v}\in U_{f^{*\perp}}$, we have:

$$\begin{aligned}\sum_{k=1}^s \frac{\mathbf{v}^T\mathbf{x}_i\mathbf{x}_i^T\mathbf{v}}{(\mathbf{x}_{(k)}^Tf)^2} &\geq \frac{1}{C_U^2}\frac{n^2}{4\|f\|^2d^2\ln^2 n}(d\ln n) \\ &= \frac{1}{C_U^2}\frac{n^2}{4\|f\|^2 d\ln n}\end{aligned}$$

Plugging this lower-bound in (8), we get with probability at least $(1-\frac{2d\ln n}{n^{10}} - \frac{1}{n^{4d}})$, $\sigma_{d-1}(X^TX) \geq C'\frac{n^2}{\|f\|^2 d\ln n}$.

B.3. Proof of Lemma 2

1. By definition, the smallest singular value,

$$\begin{aligned}\sigma_d(X^TX) &= \inf_{\mathbf{v}\in\mathbb{R}^d,\|\mathbf{v}\|=1} \mathbf{v}^TX^TX\mathbf{v} \\ &\leq \frac{f}{\|f\|}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^T\frac{f}{\|f\|} = \sum_{i=1}^n \frac{(\mathbf{x}_i^Tf)^2}{\|f\|^2} \\ &\leq n\tau^2\end{aligned} \tag{9}$$

Let $\mathbf{v}^*$ denote the smallest singular vector of X^TX. Write $\mathbf{v}^*$ as:

$$\mathbf{v}^* = \sqrt{1-\alpha_d^2}\mathbf{v}_\perp + \alpha_d f/\|f\|$$

where $\mathbf{v}_\perp$ denotes the component along the subspace orthogonal to f, and $\alpha_d = \mathbf{v}^{*T} f/\|f\|$. Now:

$$\begin{aligned}\mathbf{v}^{*T} X^T X \mathbf{v}^* &= \alpha_d^2 \sum_{i=1}^n \frac{(f^T \mathbf{x}_i)^2}{\|f\|^2} \\ &+ (1-\alpha_d^2) \sum_{i=1}^n (\mathbf{v}_\perp^T \mathbf{x}_i)^2 \\ &+ 2\alpha_d \sqrt{1-\alpha_d^2} \sum_{i=1}^n \frac{(\mathbf{v}_\perp^T \mathbf{x}_i)(f^T \mathbf{x}_i)}{\|f\|} \\ &\leq n\tau^2\end{aligned}$$

The second term in the above equation can be lower bounded with probability at least $1-\exp(-2n)$ by $(1-\alpha_d^2)\frac{n}{2}$. We can upper bound the summation in the third term as $\sum_{i=1}^n \frac{\mathbf{v}_\perp^T \mathbf{x}_i f^T \mathbf{x}_i}{\|f\|} \leq \sqrt{\sum_{i=1}^n \frac{(f^T\mathbf{x}_i)^2}{\|f\|^2}} \sqrt{\sum_{i=1}^n (\mathbf{v}_\perp^T \mathbf{x}_i)^2} \leq \tau\sqrt{n}\sqrt{2n}$, with probability at least $1-\exp(-2n)$. The first term is a positive quantity. So we have, with probability at least $1-2\exp(-2n)$:

$$\begin{aligned}(1-\alpha_d^2)\frac{n}{2} - 2\sqrt{2}\alpha_d\sqrt{1-\alpha_d^2}n\sqrt{n}\tau \\ \leq \mathbf{v}^{*T} X^T X \mathbf{v}^* \leq n\tau^2\end{aligned}$$

This implies,

$$(1-\alpha_d^2) - 4\sqrt{2}\alpha_d\sqrt{1-\alpha_d^2}n\sqrt{n}\tau - 2\tau^2 \leq 0$$

Solving the above, we get $\sqrt{1-\alpha_d^2} \leq 5\sqrt{2}\tau$, and in turn, $\alpha_d^2 \geq 1-50\tau^2$.

2. Consider the variational characterization of the second smallest singular value σ_{d-1} given by:

$$\sigma_{d-1}(X^T X) = \max_{U:\dim(U)=d-1} \min_{\mathbf{v}\in U, \|\mathbf{v}\|=1} \mathbf{v}^T X^T X \mathbf{v}.$$

Consider the particular $d-1$ dimensional subspace $U_{f^\perp} = \{\mathbf{v} \in \mathbb{R}^d \mid \mathbf{v}^T f = 0\}$. Note that the projection matrix corresponding to $U_{f^\perp}$ is given by $\left(I_d - \frac{ff^T}{\|f\|^2}\right)$. For $\mathbf{v} \in U_{f^\perp}$, we have:

$$\begin{aligned}\mathbf{v}^T X^T X \mathbf{v} &= \sum_{i=1}^n \mathbf{v}^T \mathbf{x}_i \mathbf{x}_i^T \mathbf{v} \\ &= \sum_{i=1}^n \mathbf{v}^T \left(I_d - \frac{ff^T}{\|f\|_2}\right)\mathbf{x}_i \mathbf{x}_i^T \left(I_d - \frac{ff^T}{\|f\|_2}\right)\mathbf{v}\end{aligned}$$

where the above equality follows by definition. Even though $\mathbf{x}_i$'s are not independent, observe that $\left(I_d - \frac{ff^T}{\|f\|^2}\right)\mathbf{x}_i$ are iid random variables from the distribution $\mathcal{N}(0, (I_d - \frac{ff^T}{\|f\|^2}))$ and therefore we can invoke Lemma 6 to bound the above quantity uniformly over all $\mathbf{v} \in U_{f^\perp}$, with probability at least $1-\exp(-2n)$, by $n/2$. Thus we have a lower-bound for $\min_{\mathbf{v}\in U_{f^\perp}} \mathbf{v}^T X^T X \mathbf{v}$, which immediately implies a lower bound for σ_{d-1}. We conclude that $\sigma_{d-1}(X^T X) \geq \frac{1}{2}n$ with probability at least $1-\exp(-2n)$.

B.4. Proof of Lemma 3

Note that g_i are iid draws from $\mathcal{N}(0, \|\mathbf{z}\|^2)$. First note that for any i,

$$P(|g_i| \leq \|\mathbf{z}\|\tau) \geq \frac{2\tau}{\sqrt{2\pi}} e^{\frac{-\tau^2}{2\|\mathbf{z}\|^2}} > \frac{1}{2}\tau e^{\frac{-\tau^2}{2\|\mathbf{z}\|^2}}.$$

We divide $\mathcal{U}$ into $\frac{m\tau}{2}$ batches of size $\frac{2}{\tau}$. For each batch $\mathcal{B}_j$, we have:

$$\begin{aligned}P(\min_{i\in\mathcal{B}_j} |g_i| > \tau) &= \left(1 - P(|g_i| \leq \tau)\right)^{2/\tau} \\ &\leq \left(1 - \tau\frac{1}{2}e^{\frac{-\tau^2}{2\|\mathbf{z}\|^2}}\right)^{2/\tau} \\ &\leq 1 - e^{\frac{-\tau^2}{2\|\mathbf{z}\|^2}}\end{aligned}$$

So, $P(\min_{i\in\mathcal{B}_j} |g_i| \leq \tau) > e^{\frac{-\tau^2}{2\|\mathbf{z}\|^2}}$. As the batches are independent, $P\left(\left|\left\{i : |g_i| \leq \tau\right\}\right| \geq \frac{m\tau}{2}\right) \geq P\left(\forall j, \min_{i\in\mathcal{B}_j} |g_i| \leq \tau\right) > e^{-m\tau^3/(4\|\mathbf{z}\|^2)}$.

B.5. Proof of Lemma 4

Recall that $y_i = \langle \beta^*, \mathbf{x}_i\rangle + \eta_i$ where $\eta_i = \langle \mathbf{x}_i, f^*\rangle g_i$ where $g_i \sim N(0,1)$. Consider the following RV:

$$(f^*)^T S f^* = \frac{1}{m_1}\sum_{i=1}^{m_1} (\mathbf{x}_i^T(\beta^* - \beta_0 + g_i f^*))^2 (\mathbf{x}_i^T f^*)^2. \tag{10}$$

As f^*, β^*, β_0, g_i are all fixed w.r.t. $\mathbf{x}_i$. Hence, $\mathbf{x}_i^T(\beta^* - \beta_0 + g_i f^*) \sim N(0, \|\beta^* - \beta_0 + g_i f^*\|^2)$ and $\mathbf{x}_i^T f^* \sim N(0, \|f^*\|^2)$. Hence, for all i, w.p. $\geq 1 - 3\exp(-m_1)$: we have $(\mathbf{x}_i^T(\beta^* - \beta_0 + g_i f^*))^2 (\mathbf{x}_i^T f^*)^2 \leq 2\|f^*\|^2(\|\beta^* - \beta_0\|^2 + \log m_1 \|f^*\|^2)\log^2 m_1$. Using standard Hoeffding bound, we have w.p. $\geq 1 - \frac{1}{m_1^{10}}$:

$$\begin{aligned}&\left|\frac{1}{m_1}\sum_{i=1}^{m_1} (\mathbf{x}_i^T(\beta^* - \beta_0 + g_i f^*))^2 (\mathbf{x}_i^T f^*)^2 - \right. \\ &\quad \left. 3\|f^*\|^4 - \|\beta^* - \beta_0\|^2\|f^*\|^2\right| \\ &\leq \frac{\log^3 m_1}{\sqrt{m_1}} \cdot 2\|f^*\|^2(\|\beta^* - \beta_0\|^2 + \log m_1 \|f^*\|^2).\end{aligned}$$

That is,

$$\frac{1}{m_1}\sum_{i=1}^{m_1}(\mathbf{x}_i^T(\beta^*-\beta_0+g_i f^*))^2(\mathbf{x}_i^T f^*)^2 \geq \left(1-\frac{10\log^3 m_1}{\sqrt{m_1}}\right)(3\|f^*\|^4+\|\beta^*-\beta_0\|^2\|f^*\|^2). \quad (11)$$

Similarly, let $f_\perp$ be a unit vector s.t. $f_\perp^T f^* = 0$. Now, consider the following RV:

$$(f_\perp)^T S f_\perp = \frac{1}{m_1}\sum_{i=1}^{m_1}(\mathbf{x}_i^T(\beta^*-\beta_0+g_i f^*))^2(\mathbf{x}_i^T f_\perp)^2. \quad (12)$$

Using similar argument as above, we have w.p. $\geq 1-3\exp(-m_1)-\delta$:

$$(f_\perp)^T S f_\perp \leq (\|f^*\|^2+\|\beta^*-\beta_0\|^2)\left(1+\frac{10\log^3 m_1\sqrt{\log(1/\delta)}}{\sqrt{m_1}}\right). \quad (13)$$

Hence, using the fact that $m_1 = \Omega(d\log^3 d)$ along with standard ϵ-net argument, we have:

$$\min_{f,\|f\|=1,f\perp f^*} f^T S f \leq 1.1\left(1+\sqrt{\frac{10d\log^3 d}{m_1}}\right)(\|f^*\|^2+\|\beta^*-\beta_0\|^2). \quad (14)$$

Lemma now follows using (11) and (14).

B.6. Proof of Theorem 2

Lemma 3 ensures that, with probability at least $\exp(\frac{-n^3}{4m^2\|f^*\|^2})$, there will be least $n = |\mathcal{L}|$ (by the assumption on n in the statement of the theorem) samples at the end of Step 1 of the Algorithm. Now, consider the weighted least squares estimate computed in Step 2 of Algorithm 2:

$$\begin{aligned}\widehat{\beta} &= (X^TWX)^{-1}X^TW\mathbf{y}\\ &= (X^TWX)^{-1}\sum_{i=1}^{n}\frac{1}{\langle\mathbf{x}_i,f^*\rangle^2}(\langle\beta^*,\mathbf{x}_i\rangle+g_i\langle\mathbf{x}_i,f^*\rangle)\mathbf{x}_i,\end{aligned}$$

where g_i are i.i.d. $\mathcal{N}(0,1)$ random variables. So we have:

$$\widehat{\beta}-\beta^* = (X^TWX)^{-1}\sum_{i=1}^{n}\frac{1}{\langle\mathbf{x}_i,f^*\rangle^2}g_i\langle\mathbf{x}_i,f^*\rangle\mathbf{x}_i\,,$$

and

$$\|\widehat{\beta}-\beta^*\|_2^2 = \mathbf{tr}\big((X^TWX)^{-2}X^TW^{0.5}\mathbf{g}\mathbf{g}^TW^{0.5}X\big),$$

Note that because $\mathbb{E}[\mathbf{g}\mathbf{g}^T] = I_n$ (where the expectation is wrt. to the randomness in the labels given by the oracle $\mathcal{O}$) and $\mathbf{tr}$ is linear operator, we have:

$$\mathbb{E}\|\widehat{\beta}_{\text{GLS}}-\beta^*\|_2^2 = \mathbf{tr}\left((X^TWX)^{-1}\right).$$

We now lower bound each eigenvalue of $(X^TWX)^{-1}$ to obtain the required bound. Note that this claim is similar to Lemma 1.

In particular, $\sigma_d \leq (f^*)^TX^TWX(f^*) = n$. Now, we wish to bound smallest eigenvalue of X^TWX in space orthogonal to f^*. Note that our algorithm selects $\mathbf{x}_i$ s.t. i is amongst n smallest $|\mathbf{x}_i^Tf^*|$. Also, $n \geq 4d$. Let $i_1,\ldots,i_{2d}$ be s.t. $i_k \in \mathcal{L}$ and $|\mathbf{x}_{i_1}^Tf^*| \leq \cdots \leq |\mathbf{x}_{i_{2d\log d}}^Tf^*|$. Note that using Lemma 7, w.h.p. $|\mathbf{x}_{i_{2d\log d}}^Tf^*| = O(\frac{d\log d}{m})$.

Hence, using argument similar to Lemma 1, we have:

$$\sigma_{d-1}(X^TWX) \geq \frac{m^2}{d\log^2 d}.$$

Now, again using same argument as Lemma 1 along with $(f^*)^TX^TWXf^* = n$ and the above bound, we can show that $\sigma_d \geq \frac{n}{2}$.

Theorem now follows by using $\mathbf{tr}\left((X^TWX)^{-1}\right) \leq \frac{1}{\sigma_d(X^TWX)}+\frac{d}{\sigma_{d-1}(X^TWX)}$.

B.7. Proof of Theorem 3

Let $\widehat{W}$ denote the diagonal matrix with estimated weights $\widehat{W}_{ii} := \widehat{w}_i = 1/(\langle\widehat{f},\mathbf{x}_i\rangle^2+\gamma^2)$. Consider the weighted least squares estimate:

$$\begin{aligned}\widehat{\beta}_{\text{GLS}} &= (X^T\widehat{W}X)^{-1}X^T\widehat{W}\mathbf{y}\\ &= (X^T\widehat{W}X)^{-1}\sum_{i=1}^{n}\widehat{w}_i(\langle\beta^*,\mathbf{x}_i\rangle+g_i\langle\mathbf{x}_i,f^*\rangle)\mathbf{x}_i,\end{aligned}$$

where g_i are i.i.d. $\mathcal{N}(0,1)$ random variables. Rearranging, we get:

$$\begin{aligned}\widehat{\beta}_{\text{GLS}}-\beta^* &= (X^T\widehat{W}X)^{-1}\sum_{i=1}^{n}\frac{g_i\langle f^*,\mathbf{x}_i\rangle}{\langle\widehat{f},\mathbf{x}_i\rangle^2+\gamma^2}\mathbf{x}_i\\ \|\widehat{\beta}_{\text{GLS}}-\beta^*\|_2^2 &= \mathbf{tr}\left((X^TWX)^{-2}X^T\widetilde{W}\mathbf{g}\mathbf{g}^T\widetilde{W}X\right),\end{aligned}$$

where $\widetilde{W}$ is the $n\times n$ diagonal matrix with $\widetilde{W}_{ii} = \frac{\langle f^*,\mathbf{x}_i\rangle}{\langle\widehat{f},\mathbf{x}_i\rangle^2+\gamma^2}$. Note that because $\mathbb{E}[\mathbf{g}\mathbf{g}^T] = I_n$ (where the expectation is wrt. to the randomness in the labels given by the oracle $\mathcal{O}$) and $\mathbf{tr}$ is linear operator, we have:

$$\mathbb{E}\|\widehat{\beta}_{\text{GLS}}-\beta^*\|_2^2 = \mathbf{tr}\left((X^T\widehat{W}X)^{-2}X^T\widetilde{W}^2X\right)$$

Write $f^* = \widehat{f} + \delta_f$, where $\|\delta_f\|_2 \leq \Delta$. Let ΔW denote the matrix with $\Delta W_{ii} = \frac{\langle \delta_f, \mathbf{x}_i \rangle}{\langle \widehat{f}, \mathbf{x}_i \rangle^2 + \gamma^2}$. We can bound $\widetilde{W}$ as:

$$\widetilde{W}^2 \leq 2(\widehat{W} + \Delta W^2)$$

So, we have:

$$\mathbb{E}\|\widehat{\beta}_{\text{GLS}} - \beta^*\|_2^2 = 2\mathbf{tr}\Big((X^T\widehat{W}X)^{-1}\Big) + 2\mathbf{tr}\Big((X^T\widehat{W}X)^{-2}X^T\Delta W^2 X\Big) \quad (15)$$

1. Consider the first term $\mathbf{tr}\Big((X^T\widehat{W}X)^{-1}\Big) = \sum_{i=1}^n \frac{1}{\langle \widehat{f}, \mathbf{x}_i \rangle^2 + \gamma^2} \mathbf{x}_i \mathbf{x}_i^T$. It can be bounded readily by $\max_i(\langle \widehat{f}, \mathbf{x}_i \rangle^2 + \gamma^2)\mathbf{tr}\big((X^TX)^{-1}\big)$. We can bound $\mathbf{tr}\big((X^TX)^{-1}\big)$ by $\frac{d}{n}$, using standard arguments. Applying Lemma 7, we can bound $\max_i(\langle \widehat{f}, \mathbf{x}_i \rangle^2 + \gamma^2)$ by $C_U^2 \ln^2 n + \gamma^2$, with probability at least $1 - 1/n^{10}$. Together, we have $\mathbf{tr}\Big((X^T\widehat{W}X)^{-1}\Big) \leq C_U^2 \frac{d}{n} \ln^2 n$.

2. Now to bound the second term $\mathbf{tr}\Big((X^T\widehat{W}X)^{-2}X^T\Delta W^2 X\Big)$, first consider the matrix ΔW^2. The ith entry of this matrix is $\frac{\langle \delta_f, \mathbf{x}_i \rangle^2}{(\langle \widehat{f}, \mathbf{x}_i \rangle^2 + \gamma^2)^2}$. Without loss of generality, assume δ_f is orthogonal to $\widehat{f}$. In expectation, the diagonal entry is at most $\frac{\|\delta_f\|^2}{\gamma^4}$. From Lemma 4, and by the choice of γ in the statement of the theorem, the quantity is at most $\|f^*\|^2$. Thus in expectation ΔW^2 can be bounded by I_n. We can bound $\mathbf{tr}\Big((X^T\widehat{W}X)^{-1}\Big)$ similar to the case above. Together, we have, $\mathbf{tr}\Big((X^T\widehat{W}X)^{-2}X^T\Delta W^2 X\Big) \leq \mathbf{tr}\Big((X^T\widehat{W}X)^{-2})\Big)\|X^TX\|_2 \leq \frac{d \ln^4 n}{n^2}(n) = \frac{\|f^*\|^2 d \ln^4 n}{n}$. Plugging the above two bounds in (15), the proof is complete.

B.8. Proof of Lemma 5

Denote $|\mathcal{L}|$ by n_τ. Let $X \in \mathbb{R}^{n_\tau \times d}$ denote the design matrix with instances in $\mathcal{L}$ as rows. Consider the ordinary least squares estimate:

$$\begin{aligned}\widehat{\beta} &= (X^TX)^{-1}X^T\mathbf{y} \\ &= (X^TX)^{-1}\sum_{i=1}^{n_\tau}(\langle \beta^*, \mathbf{x}_i \rangle + g_i \langle \mathbf{x}_i, f^* \rangle)\mathbf{x}_i,\end{aligned}$$

where g_i are i.i.d. $\mathcal{N}(0,1)$ random variables. So we have:

$$\begin{aligned}\widehat{\beta} - \beta^* &= (X^TX)^{-1}\sum_{i=1}^{n_\tau} g_i \langle \mathbf{x}_i, f^* \rangle \mathbf{x}_i \\ \|\widehat{\beta} - \beta^*\|_2^2 &= \mathbf{tr}\Big((X^TX)^{-2}X^TW^{-0.5}\mathbf{g}\mathbf{g}^{\mathbf{T}}W^{-0.5}X\Big),\end{aligned}$$

where $W^{-0.5}$ is the diagonal matrix with ith diagonal entry $\langle f^*, \mathbf{x}_i \rangle$. Note that because $\mathbb{E}[\mathbf{g}\mathbf{g}^T] = I_{n_\tau}$ and $\mathbf{tr}$ is linear operator, we have:

$$\mathbb{E}\|\widehat{\beta} - \beta^*\|_2^2 = \mathbf{tr}\Big((X^TX)^{-2}X^TW^{-1}X\Big),$$

Now, write $f^* = \widehat{f} + \delta_f$, where $\|\delta_f\| \leq \Delta$ (as given in the statement of the Theorem). So, $\langle f^*, \mathbf{x}_i \rangle = \langle \widehat{f}, \mathbf{x}_i \rangle + \langle \delta_f, \mathbf{x}_i \rangle$, and $\langle f^*, \mathbf{x}_i \rangle^2 \leq 2\|f^*\|^2(\Delta^2 + \tau^2)$.

$$\begin{aligned}\mathbb{E}\|\widehat{\beta} - \beta^*\|_2^2 &= \mathbf{tr}\Big(X(X^TX)^{-2}X^TW^{-1}\Big), \\ &= 2(\tau^2 + \Delta^2)\mathbf{tr}\Big(X(X^TX)^{-2}X^T\Big) \\ &= 2(\tau^2 + \Delta^2)\mathbf{tr}\big((X^TX)^{-1}\big)\end{aligned}$$

We can use identical arguments as in the proof of Lemma 2, we can upper bound the trace quantity in the above RHS by $O\Big(\frac{1}{n_\tau \tau^2} + \frac{d-1}{n_\tau}\Big)$. Using Lemma 3 we can lower bound n_τ by $m\tau$ with probability at least $\exp(-m\tau^3)$. This completes the proof.

B.9. Proof of Theorem 4

From Lemma 3, we know that about $n_\tau = \frac{m\tau}{2}$ instances out of m unlabeled instances satisfy the tolerance condition in Step 4 of the algorithm with high probability. So, we want to choose τ as a function of $\Delta = \|\widehat{f} - f^*\|$, m, and d so that the RHS of the bound in Lemma 5 is minimized. Solving the resulting quadratic problem, we see that $\tau = \Delta$ is optimal choice, up to constant and $\|f^*\|$ factors. From Lemma 4, we have $\Delta = O(\sqrt{d/m_1})$. Choosing $m_1 = n/2$, we then have with probability at least $\exp(-1/n)$, at least n examples satisfying $|\langle \mathbf{x}_i, \widehat{f} \rangle| \leq 2\sqrt{d/n}$ in Step 4 of the algorithm. We can now apply Lemma 5 to recover the statement of the theorem.

C. Iterative Estimation Algorithm of (Carroll et al., 1988)

We now apply the analysis of (Carroll et al., 1988) to bound the estimation error of weighted least squares estimator

with estimated weights (Algorithm 3). In fact, Carroll et al. (1988) develop an iterative algorithm where the estimates $\widehat{f}$ and $\widehat{\beta}$ are iteratively improved. So we will mimic the setup, and derive bounds for the iterative version of Algorithm 3. In the following, $\widehat{\beta}_t$ and $\widehat{f}_t$ denote the estimators at the end of round t. We use the same β_0 as in Algorithm 3. Define the following quantities:

1. $\widehat{r}_i := \widehat{r}_i^{(t)} = y_i - \langle \mathbf{x}_i, \widehat{\beta}_t \rangle$; sometimes we write $\widehat{r}_i$ when t is implicit.

2. $\delta_i = y_i - \tau_i$, where $\tau_i = \langle \mathbf{x}_i, \beta^* \rangle$.

3. $\Psi_i := \Psi(\delta_i, f) = (\delta_i^2 \mathbf{x}_i \mathbf{x}_i^T - \lambda I) f$.

Let:

$$
\begin{aligned}
\mathbb{R}^{d\times d} \ni A_f &= -\frac{1}{n}\sum_{i=1}^n \nabla_f \Psi_i \\
\mathbb{R}^{d\times d} \ni A_\beta &= \frac{1}{n}\sum_{i=1}^n \nabla_\beta \Psi_i \\
\mathbb{R}^{d\times d} \ni A_1 &= \mathbb{E}[A_f] = -\mathbb{E}[\delta_1^2 \mathbf{x}_1 \mathbf{x}_1^T] \\
\mathbb{R}^{d\times d} \ni H_1 &= A_1^{-1}\left(\sqrt{n} A_\beta + \frac{1}{n}\sum_{i=1}^n \nabla_{\tau f}\Psi_i \,.\, g_0 \mathbf{x}_i^T\right) \\
\mathbb{R}^{d^2\times d} \ni W &= \frac{1}{2\sqrt{n}}\sum_{i=1}^n \big(I \otimes \mathbf{x}_i\big) A_1^{-1} \nabla_{\tau\tau}\Psi_i \,.\, \mathbf{x}_i^T \\
\mathbb{R}^{d\times 1} \ni g_0 &= \frac{1}{\sqrt{n}} A_1^{-1} \sum_{i=1}^n \Psi_i
\end{aligned}
$$

Lemma 8 (Bounding $\widehat{f}_t - f$ in terms of $\widehat{\beta}_t - \beta$). *As $n \to \infty$, the error in the estimate $\widehat{f}$ has the expansion:*

$$
\begin{aligned}
\widehat{f}_t - f^* &= \frac{1}{\sqrt{n}} A_f^{-1} A_1 g_0 \\
&+ \left(\frac{1}{\sqrt{n}} H_1 + [I \otimes (\widehat{\beta}_t - \beta^*)^T] W\right)(\widehat{\beta}_t - \beta^*) \\
&+ O_p(n^{-3/2}) \,,
\end{aligned}
$$

where $O_p(n^{-3/2})$ captures lower-order error quantities that converge (in probability) to 0 at or faster than the rate $O\big(\frac{1}{n\sqrt{n}}\big)$.

Define the quantities:

$$
\begin{aligned}
\mathbb{R}^{d\times d} \ni B_0 &= X^T W X \\
\mathbb{R}^{d\times 1} \ni v_0 &= X^T W \delta \\
\mathbb{R} \ni \eta_i &= \delta_i - \mathbf{x}_i^T B_0 v_0 \\
\mathbb{R}^d \ni l_0 &= B_0^{-1} v_0 \\
\mathbb{R}^d \ni l_1 &= B_0^{-1} \sum_{i=1}^n g_0^T \nabla_f w_i \mathbf{x}_i \eta_i \\
\mathbb{R}^d \ni l_2 &= B_0^{-1}\Bigg[\sqrt{n}\sum_{i=1}^n g_0^T (A_f^{-1} A_1 - I)^T \nabla_f w_i \mathbf{x}_i \eta_i \\
&+ 0.5 \sum_{i=1}^n g_0^T \nabla_f^2 w_i g_0 \mathbf{x}_i \eta_i - \\
&\sum_{i,j=1}^n (g_0^T \nabla_f w_i)(g_0^T \nabla_f w_j)(\mathbf{x}_i^T B_0 \mathbf{x}_j)\mathbf{x}_i \eta_j\Bigg] \\
\mathbb{R}^{d\times d} \ni \mathbf{C} &= B_0^{-1}\left[\sum_{i=1}^n \mathbf{x}_i \eta_i \nabla_f w_i^T H_1\right] \\
\mathbb{R}^{d^2\times d} \ni Q &= \sum_{i=1}^n \big(B_0^{-1}\mathbf{x}_i\big) \otimes \big((\nabla_f w_i^T \otimes I) W \eta_i\big)
\end{aligned}
$$

Lemma 9 (Bounding $\widehat{\beta}_{t+1} - \beta$ in terms of $\widehat{\beta}_t - \beta$).

$$
\begin{aligned}
\widehat{\beta}_{t+1} - \beta^* &= l_0 + \frac{1}{\sqrt{n}} l_1 + \frac{1}{n} l_2 \\
&+ \left(\frac{1}{\sqrt{n}}\mathbf{C} + [I \otimes (\widehat{\beta}_t - \beta^*)^T] Q\right)(\widehat{\beta}_t - \beta^*) \\
&+ O_p(n^{-3/2})
\end{aligned}
$$

Corollary 1 (Case f^* is known). *When f is known, we have: $l_1 = l_2 = \mathbf{C} = Q = 0$. So for all $t > 0$, we have:*

$$\widehat{\beta}_t - \beta^* = l_0 = (X^T W X)^{-1} X^T W \delta \,.$$

Note that the initial $\widehat{\beta}_0$ satisfies:

$$(\widehat{\beta}_0 - \beta^*) = (X^T X)^{-1} X^T \delta := \xi_0 \,.$$

Corollary 2 (Case f^* is estimated). *We have:*

$$
\begin{aligned}
1.\ \widehat{\beta}_1 - \beta &= l_0 + \frac{1}{\sqrt{n}} l_1 \\
&+ \frac{1}{n}\big(l_2 + \mathbf{C}\xi_0 + (I \otimes \xi_0^T) Q \xi_0\big) \\
&+ O_p(n^{-3/2}), \qquad (16)
\end{aligned}
$$

and for $t \geq 2$,

$$
\begin{aligned}
2.\ \widehat{\beta}_t - \beta^* &= l_0 + \frac{1}{\sqrt{n}} l_1 \\
&+ \frac{1}{n}\big(l_2 + \mathbf{C} l_0 + (I \otimes l_0^T) Q l_0\big) \\
&+ O_p(n^{-3/2}) \,. \qquad (17)
\end{aligned}
$$

The bounds obtained offer little insight, and importantly, the dependence on factors n and d are not clear. Even for the case when f^* is known, the analysis gives no convergence rates.

D. Active Regression

Algorithm 5 considers a slightly more powerful oracle model, where the same instance can be queried multiple times, and each time the response is generated independent of the other trials. Theorem 5 shows that the learning rate in this setting is $O(1/n)$, as in Theorem 2.

Theorem 5 (Active Regression with Noise Oracle)**.** *Assume $n \geq d$. Consider the output estimator $\widehat{\beta}$ of Algorithm 5. We have, with probability at least $1 - 1/n^c$:*

$$\|\widehat{\beta} - \beta^*\|_2^2 \leq C' \|f^*\|_2^2 \left(\frac{1}{n}\right) ,$$

for some positive constants c, C'.

Proof. First, note that the matrix $N_\perp = I_d - \frac{f^* f^{*T}}{\|f^*\|_2^2}$ corresponds to $(d-1)$ directions orthogal to f^*, and thus we have $N^\perp f^* = 0$. Let $N = \frac{1}{\|f^*\|_2} f^* \mathbf{1}_{n-d}^T$ as in the Step 2 of the algorithm. Clearly, when $n = d+1$, the matrix $X = [N_\perp \; N]^T$ has full rank, with all the d singular values equal to 1. For a general $n > d$, the largest singular value of X is proportional to n, while the other singular values are 1. In this case, notice that the direction of the largest singular vector of X is f^*. Let $\mathbf{x}_i$ denote the rows (instances) of this X.

Now consider the ordinary least squares estimate:

$$\begin{aligned}
\widehat{\beta} &= (X^T X)^{-1} X^T \mathbf{y} \\
&= (X^T X)^{-1} \sum_{i=1}^{n} (\langle \beta^*, \mathbf{x}_i \rangle + g_i \langle \mathbf{x}_i, f^* \rangle) \mathbf{x}_i , \\
&= \beta^* + (X^T X)^{-1} \left(\sum_{i=1}^{d} 0 + \sum_{i=d+1}^{n} g_i f^* \right) ,
\end{aligned}$$

where g_i are i.i.d. $\mathcal{N}(0,1)$ random variables, and the last equality is true by construction of X. So we have:

$$\begin{aligned}
\|\widehat{\beta} - \beta^*\| &= \left\| (X^T X)^{-1} \sum_{i=d+1}^{n} g_i f^* \right\| \\
&\leq \left\| (X^T X)^{-1} f^* \right\| \left\| \sum_{i=d+1}^{n} g_i \right\|
\end{aligned}$$

Notice that f^* is the smallest singular vector of $(X^T X)^{-1}$, and therefore $\left\|(X^T X)^{-1} f^*\right\|$ is proportional to the smallest singular value of $(X^T X)^{-1}$, which is $1/\|(X^T X)\| = 1/n$. So:

$$\|\widehat{\beta} - \beta^*\| \leq \|f^*\| \frac{1}{n} \sum_{i=d+1}^{n} g_i \, .$$

The sum in the above term can be controlled with high probability using Chernoff bounds, which yields, with probability at least $1 - 1/n^c$, $|\sum_{i=d+1}^{n} g_i| \leq C' \sqrt{n-d}$, for $c, C' > 0$. The proof is complete. □

Using essentially identical arguments, we can also prove a lower bound, so that effectively we have:

$$\|\widehat{\beta} - \beta^*\|_2^2 = O\left(\|f^*\|_2^2 \frac{1}{n} \right) .$$

Algorithm 5 Active Regression With Noise Oracle

Input: Labeling oracle $\mathcal{O}$, noise model f^*, label budget $n > d$.

1. Form the matrix $N_\perp = I_d - \frac{f^* f^{*T}}{\|f^*\|_2^2}$, and query $\mathcal{O}$ for (exact) labels of each column of the matrix (call them $y_1, y_2, \ldots, y_d$.

2. Make $n - d$ queries to $\mathcal{O}$ and obtain (noisy) labels along the direction f^*. Call these labels $y_{d+1}, y_{d+2}, \ldots, y_n$. Let $N = \frac{1}{\|f^*\|_2} f^* \mathbf{1}_{n-d}^T$, where $\mathbf{1}_{n-d}^T$ denotes the vector of all ones, in $n - d$ dimensions.

2. Estimate $\widehat{\beta}$ by solving $\mathbf{y} \approx X\widehat{\beta}$ (ordinary least squares) where $X = [N_\perp \;\; N]^T \in \mathbb{R}^{n \times d}$ and $\mathbf{y} \in \mathbb{R}^n$.

Output: $\widehat{\beta}$.