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 $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{n'}$ 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 $R_1, R_2, \dots, R_n$. 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:

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

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

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

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:

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

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

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

Note that $P\big(\tilde{R}_{(k)} \geq 10e\big(8\sqrt{\pi} + C)\ln n\big) \leq P\big(|\tilde{R}_{(k)} - \mu| \geq 10e\big(8\sqrt{\pi} + C)\ln n\big)$. Finally, observing that $\tilde{R}_{(k)} \geq 10e\big(8\sqrt{\pi} + C)\ln n \iff R_{(k)} \geq 10e\big(8\sqrt{\pi} + C)\ln n\frac{k}{n}$, 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:

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

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 $(1-\alpha_d^2)n^2 d/\|f\|^2$. 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:

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

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 any vector $\mathbf{v} \in U_{f^\perp}$, we have:

$$\mathbf{v}^T X^T W X \mathbf{v} = \sum_{i=1}^n \frac{\mathbf{v}^T \mathbf{x}_i \mathbf{x}_i^T \mathbf{v}}{(\mathbf{x}_i^T f)^2} \tag{8}$$

Note that $g_i = \mathbf{v}^T \mathbf{x}_i$, $i = 1, 2, \ldots, n$ and $h_i = \mathbf{x}_i^T f$, $i = 1, 2, \ldots, n$ 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 $\sum_{i=1}^n \frac{(\mathbf{v}^T\mathbf{x}_i)^2}{(\mathbf{x}_i^T f)^2}$, 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

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

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

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

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:

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

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,

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

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:

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

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 $\sum_{i=1}^n \frac{\mathbf{v}_\perp^T \mathbf{x}_i f^T \mathbf{x}_i}{\|f\|} \le \sqrt{\sum_{i=1}^n \frac{(f^T\mathbf{x}_i)^2}{\|f\|^2}} \sqrt{\sum_{i=1}^n (\mathbf{v}_\perp^T \mathbf{x}_i)^2} \le \tau\sqrt{n}\sqrt{2n}$, with probability at least $1-\exp(-2n)$. The first term is a positive quantity. So we have, with probability at least $1-2\exp(-2n)$:

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

This implies,

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

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

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 $\sigma_{d-1}(X^T X) \ge \frac{1}{2} n$ 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,

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

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

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

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

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. $\ge 1 - 3\exp(-m_1)$: we have $(\mathbf{x}_i^T(\beta^* - \beta_0 + g_i f^*))^2 (\mathbf{x}_i^T f^*)^2 \le 2\|f^*\|^2(\|\beta^* - \beta_0\|^2 + \log m_1 \|f^*\|^2)\log^2 m_1$. Using standard Hoeffding bound, we have w.p. $\ge 1 - \frac{1}{m_1^{10}}$:

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

That is,

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

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

$$(f_\perp)^TSf_\perp=\frac{1}{m_1}\sum_{i=1}^{m_1}(\mathbf{x}_i^T(\beta^*-\beta_0+g_if^*))^2(\mathbf{x}_i^Tf_\perp)^2. \tag{12}$$

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

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

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

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

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:

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

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 $i_1,\dots,i_{2d}$ 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 $\mathbf{tr}\bigg((X^TWX)^{-1}\bigg)\leq\frac{1}{\sigma_d(X^TWX)}+\frac{d}{\sigma_{d-1}(X^TWX)}$.

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:

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

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:

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

Write $f^* = \widehat{f} + \delta_f$, where $\|\delta_f\|_2 \le \Delta$. 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:

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

So, we have:

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

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 $\mathbf{tr}\Big((X^T\widehat{W}X)^{-1}\Big) \le C_U^2 \frac{d}{n} \ln^2 n$.

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, $\mathbf{tr}\Big((X^T\widehat{W}X)^{-2}X^T\Delta W^2 X\Big) \le \mathbf{tr}\Big((X^T\widehat{W}X)^{-2})\Big)\|X^TX\|_2 \le \frac{d \ln^4 n}{n^2}(n) = \frac{\|f^*\|^2 d \ln^4 n}{n}$. 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 $\|\delta_f\| \le \Delta$ (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 $\langle f^*, \mathbf{x}_i \rangle^2 \le 2\|f^*\|^2(\Delta^2 + \tau^2)$.

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

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 $|\langle \mathbf{x}_i, \widehat{f} \rangle| \le 2\sqrt{d/n}$ 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: $\widehat{\beta}$.
```