Supplementary Material

A Omitted Proofs

A.1 Proof of Theorem 4.1

In order to prove Theorem 4.1, let us first state the Danskin theorem.

Lemma A.1 (Danskin). *Let $\mathcal{B}$ be nonempty compact topological space and $h : \mathbb{R}^d \times \mathcal{B} \to \mathbb{R}$ be such that $h(\cdot, \delta)$ is differentiable for every $\delta \in \mathcal{B}$ and $\nabla_\theta h(\theta, \delta)$ is continuous on $\mathbb{R}^d \times \mathcal{B}$. Also, let $\delta^*(\theta) = \{\delta \in \arg\max_{\delta \in \mathcal{B}} h(\theta, \delta)\}$.*
Then, the corresponding max-function

$$\varsigma(\theta) = \max_{\delta \in \mathcal{B}} h(\theta, \delta)$$

is locally Lipschitz continuous, directionally differentiable, and its directional derivatives satisfy

$$\varsigma'(\theta, r) = \sup_{\delta \in \delta^*(\theta)} r^T \nabla_\theta h(\theta, \delta).$$

In particular, if for some $\theta \in \mathbb{R}^d$ the set $\delta^*(\theta) = \{\delta^*_\theta\}$ is a singleton, the max-function is differentiable at θ and

$$\nabla \varsigma(\theta) = \nabla_\theta h(\theta, \delta^*_\theta).$$

By this lemma, we can easily obtain the following lemma:

Lemma A.2. *For any $\tilde{\theta}$ that minimize $\varsigma(\theta)$ and lying in the interior, we can obtain*

$$\nabla_\theta h(\tilde{\theta}, \delta) = 0.$$

Proof. Since $\tilde{\theta}$ minimizes $\varsigma(\theta)$ and lies in the interior of Θ , we can obtain

$$\varsigma'(\tilde{\theta}, r) = 0$$

for any direction vector r.

If there is a $\delta \in \delta^*(\tilde{\theta})$, such that $\nabla_\theta h(\tilde{\theta}, \delta) \neq 0$, then we take $r = \nabla_\theta h(\tilde{\theta}, \delta)/\|\nabla_\theta h(\tilde{\theta}, \delta)\|$, we have

$$\varsigma'(\tilde{\theta}, r) = \sup_{\delta \in \delta^*(\tilde{\theta})} r^T \nabla_\theta h(\tilde{\theta}, \delta) \geq \|\nabla_\theta h(\tilde{\theta}, \delta)\|_2 > 0,$$

which is contradictory to the fact $\varsigma'(\tilde{\theta}, r) = 0$. $\square$

[**Proof of Theorem 1**] Now we are ready to give the formal proof. In order for simplicity, we here use $L(\theta^{\mathcal{M}}, x, y)$ instead of $L(\theta^{\mathcal{M}}, x, y, \mathcal{M})$. With lemma A.2, we can obtain that

$$\frac{1}{n} \sum_{i=1}^{n} \nabla_\theta L(\theta^{\mathcal{M}}, x_i + \delta_i, y_i)|_{\theta^{\mathcal{M}} = \hat{\theta}^{\mathcal{M}}_{\varepsilon,\min}} = 0.$$

With Taylor expansion and under the assumption of Lemma A.2, we can obtain

$$0 = \frac{1}{n} \sum_{i=1}^{n} [\nabla_\theta L(\hat{\theta}^{\mathcal{M}}_{\varepsilon,\min}, x_i, y_i) + \nabla_{x,\theta} L(\hat{\theta}^{\mathcal{M}}_{\varepsilon,\min}, x_i, y_i)\delta_i + O(\|\delta\|_2^2)].$$

Here the assumption of compactness and continuity can help us to write the remainder $\frac{1}{2}\delta_i^T H_{\tilde{\theta}}\delta_i$ into $O(\|\delta_i\|_2^2)$ since we can bound every entry of $H_{\tilde{\theta}}$. We use the same property repeatedly and will not reiterate it.

Now, let us perform taylor expansion on $\nabla_\theta L(\hat{\theta}^{\mathcal{M}}_{\varepsilon,\min}, x_i, y_i)$ and $\nabla_{x,\theta} L(\hat{\theta}^{\mathcal{M}}_{\varepsilon,\min}, x_i, y_i)$.

$$\nabla_\theta L(\hat{\theta}^{\mathcal{M}}_{\varepsilon,\min}, x_i, y_i) = \nabla_\theta L(\hat{\theta}^{\mathcal{M}}_{\min}, x_i, y_i) + \nabla^2_\theta L(\hat{\theta}^{\mathcal{M}}_{\min}, x_i, y_i)(\hat{\theta}^{\mathcal{M}}_{\varepsilon,\min} - \hat{\theta}^{\mathcal{M}}_{\min}) + +O(\|\hat{\theta}^{\mathcal{M}}_{\varepsilon,\min} - \hat{\theta}^{\mathcal{M}}_{\min}\|_2^2)$$

and

$$\nabla_{x,\theta} L(\hat{\theta}^{\mathcal{M}}_{\varepsilon,\min}, x_i, y_i) = \nabla_{x,\theta} L(\hat{\theta}^{\mathcal{M}}_{\min}, x_i, y_i) + O(\|\hat{\theta}^{\mathcal{M}}_{\varepsilon,\min} - \hat{\theta}^{\mathcal{M}}_{\min}\|_2).$$

By simple algebra,

$$\begin{aligned}\hat{\theta}^{\mathcal{M}}_{\varepsilon,\min} - \hat{\theta}^{\mathcal{M}}_{\min} + O(\|\hat{\theta}^{\mathcal{M}}_{\varepsilon,\min} - \hat{\theta}^{\mathcal{M}}_{\min}\|_2^2) =& \Big(-\frac{1}{n}\sum_{i=1}^n \nabla^2_\theta L(\hat{\theta}^{\mathcal{M}}_{\min}, x_i, y_i)\Big)^{-1}\Big(\frac{1}{n}\sum_{i=1}^n \nabla_{x,\theta} L(\hat{\theta}^{\mathcal{M}}_{\min}, x_i, y_i)\delta_i \\ &+ \|\hat{\theta}^{\mathcal{M}}_{\varepsilon,\min} - \hat{\theta}^{\mathcal{M}}_{\min}\|_2\|\delta_i\|_2\Big).\end{aligned}$$

We know if we divided ε on both sides, we know when ε goes to 0, the limit of the right handside exists if we assume the limit of $\lim_{\varepsilon\to 0}\delta_i/\varepsilon$ exist (notice δ_i is a implicit function of ε). Thus, $\|\hat{\theta}^{\mathcal{M}}_{\varepsilon,\min} - \hat{\theta}^{\mathcal{M}}_{\min}\|/\varepsilon$ cannot goes to infinity as ε goes to 0. In orther words, AIF must exist.

Now the only thing left is to prove $\lim_{\varepsilon\to 0}\delta_i/\varepsilon$ exist. We prove that

$$\lim_{\varepsilon\to 0}\frac{\delta_i}{\varepsilon} = \gamma_i,$$

where

$$\gamma_{i,k} = \frac{b_k^{q-1}}{(\sum_{k=1}^m b_k^q)^{\frac{1}{p}}}\operatorname{sgn}\Big(\frac{\partial}{\partial x_{\cdot,k}} L(\hat{\theta}^{\mathcal{M}}_{\min}, x_i, y_i)\Big),$$

with $b_k = |\frac{\partial}{\partial x_{\cdot,k}} L(\hat{\theta}^{\mathcal{M}}_{\min}, x_i, y_i)|$. By Hölder inequality, we know

$$|\nabla_x L(\hat{\theta}^{\mathcal{M}}_{\min}, x_i, y_i)\cdot\delta_i| \le \varepsilon\|\nabla_x L(\hat{\theta}^{\mathcal{M}}_{\min}, x_i, y_i\|_q$$

the equality holds if and only if $\delta_i = \varepsilon\gamma_i$. Since

$$L(\hat{\theta}^{\mathcal{M}}_{\varepsilon,\min}, x_i + \delta_i, y_i) = L(\hat{\theta}^{\mathcal{M}}_{\varepsilon,\min}, x_i, y_i) + \nabla_x L(\hat{\theta}^{\mathcal{M}}_{\varepsilon,\min}, x_i, y_i)\delta_i + O(\|\delta_i\|_2^2),$$

we know the reminder is ignorable

$$\frac{O(\|\delta_i\|_2^2)}{|L(\hat{\theta}^{\mathcal{M}}_{\varepsilon,\min}, x_i, y_i)\delta_i|} \to 0$$

as ε goes to 0. So, we must have

$$\lim_{\varepsilon\to 0}\frac{\delta_i}{\varepsilon} = \gamma_i.$$

As a result,

$$\hat{\mathcal{I}}(\mathcal{M}) = -H^{-1}_{\hat{\theta}^{\mathcal{M}}_{\min}}\Phi$$

as described in the theorem.

A.2 Proof of Theorem 5.1

Let us first compute the AIF for linear models.

Specifically, let us consider the regression setting $(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$ are *i.i.d.* draws from a joint distribution $P_{x,y}$, for $i = 1, 2, ..., n$. Note that we don't assume linear relationship, but the linear regression model tries to find the best linear approximation by solving

$$\hat{\theta} = \arg\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} l(\theta, x_i, y_i) := \arg\min_{\theta} \frac{1}{2n} \sum_{i=1}^{n} (y_i - \theta^T x_i)^2,$$

where we use $l(\theta, x_i, y_i) = \frac{1}{2}(y_i - \theta^T x_i)^2$ as the loss function.

Further, let us define

$$\theta^* = \arg\min_{\theta} \mathbb{E}_{P_{x,y}}[\frac{1}{2}(Y - \theta^T X)^2],$$

denoting the best population linear approximation to Y.

When the true model is $Y = X^\top \beta_1^* + (X^\top \beta_2^*)^2 + \xi$, and $X \sim \mathcal{N}(0, \sigma_x^2 I)$, we have

$$\theta^* = (\mathbb{E}[XX^\top])^{-1}\mathbb{E}[XY] = (\mathbb{E}[XX^\top])^{-1}(\mathbb{E}[XX^\top \beta_1^*] + \mathbb{E}[X(X^\top \beta_2^*)^2]) = \beta_1^*.$$

Further, denote $\epsilon_i = y_i - \theta^{*\top} x_i$, and

$$\hat{\theta} = \arg\min_{\theta} \frac{1}{2n} \sum_{i=1} (y_i - \theta^T x_i)^2,$$

and we have $\|\hat{\theta} - \theta^*\|_2 = O_p(\sqrt{\frac{m}{n}})$.

By definition, for $k \in [m]$,

$$b_k = |\frac{\partial}{\partial x_{\cdot,k}} l(\hat{\theta}^{\mathcal{M}}_{\min}, x_i, y_i, \mathcal{M})| = |y_i - \hat{\theta}^\top x_i| \cdot |\hat{\theta}_k|,$$

and therefore, by letting $p = q = 2$, in Eqn (5) of Theorem 4.1,

$$\begin{aligned}
\psi_k^i &= \frac{b_k^{q-1}}{(\sum_{k=1}^d b_k^q)^{\frac{1}{p}}} \mathrm{sgn}(\frac{\partial}{\partial x_{\cdot,k}} l(\hat{\theta}^{\mathcal{M}}_{\min}, x_i, y_i, \mathcal{M})) = \frac{b_k}{(\sum_{k=1}^d b_k^2)^{1/2}} \mathrm{sgn}((y_i - \hat{\theta}^\top x_i) \cdot \hat{\theta}_k) \\
&= \frac{(y_i - \hat{\theta}^\top x_i) \cdot \hat{\theta}_k}{|y_i - \theta^\top x_i| \cdot \|\hat{\theta}\|_2} = \frac{\hat{\theta}_k}{\|\hat{\theta}\|} \cdot \mathrm{sgn}(y_i - \hat{\theta}^\top x_i).
\end{aligned}$$

As a result

$$\phi_i = (\psi_1^i, \psi_2^i, \cdots, \psi_m^i)^T = \mathrm{sgn}(y_i - \hat{\theta}^\top x_i) \cdot \frac{1}{\|\hat{\theta}\|} \cdot \hat{\theta},$$

and

$$
\begin{aligned}
\frac{\Phi}{\mathbb{E}_{x\sim\hat{P}_x}\|x\|_2} =&\frac{1}{n}\sum_{i=1}^{n}\nabla_{x,\theta}l(\hat{\theta}^{\mathcal{M}}_{\min},x_i,y_i,\mathcal{M})\phi_i = \frac{1}{n}\sum_{i=1}^{n}[(\hat{\theta}^\top x_i - y_i)\cdot I_d + \hat{\theta}x_i^\top]\cdot \mathrm{sgn}(y_i - \hat{\theta}^\top x_i)\cdot\frac{1}{\|\hat{\theta}\|}\cdot\hat{\theta}\\
=&\frac{1}{n\|\hat{\theta}\|}\sum_{i=1}^{n}[(\hat{\theta}^\top x_i - y_i)\cdot\hat{\theta} + \hat{\theta}x_i^\top\hat{\theta}]\cdot\mathrm{sgn}(y_i - \hat{\theta}^\top x_i)\\
=&-\frac{1}{n\|\hat{\theta}\|}\sum_{i=1}^{n}(y_i\cdot\hat{\theta})\cdot\mathrm{sgn}(y_i - \hat{\theta}^\top x_i)\\
=&-\frac{\hat{\theta}}{n\|\hat{\theta}\|}\sum_{i=1}^{n}y_i\cdot\mathrm{sgn}(y_i - \hat{\theta}^\top x_i)\\
=&-\frac{\hat{\theta}}{n\|\hat{\theta}\|}\sum_{i=1}^{n}(\epsilon_i + \theta^{*\top}x_i)\cdot\mathrm{sgn}(y_i - \hat{\theta}^\top x_i)\\
=&-\frac{\hat{\theta}}{n\|\hat{\theta}\|}\sum_{i=1}^{n}\theta^{*\top}x_i\cdot\mathrm{sgn}(y_i - \hat{\theta}^\top x_i) - \frac{\hat{\theta}}{n\|\hat{\theta}\|}\sum_{i=1}^{n}\epsilon_i\cdot\mathrm{sgn}(y_i - \hat{\theta}^\top x_i)\\
=&-\frac{\hat{\theta}}{n\|\hat{\theta}\|}\sum_{i=1}^{n}\theta^{*\top}x_i\cdot\mathrm{sgn}(\epsilon_i) - \frac{\hat{\theta}}{n\|\hat{\theta}\|}\sum_{i=1}^{n}\epsilon_i\cdot\mathrm{sgn}(\epsilon_i)\\
&+\frac{\hat{\theta}}{n\|\hat{\theta}\|}\sum_{i=1}^{n}\theta^{*\top}x_i\cdot(\mathrm{sgn}(\epsilon_i) - \mathrm{sgn}(\epsilon_i - (\hat{\theta}-\theta^*)^\top x_i))\\
&+\frac{\hat{\theta}}{n\|\hat{\theta}\|}\sum_{i=1}^{n}\epsilon_i\cdot(\mathrm{sgn}(\epsilon_i) - \mathrm{sgn}(\epsilon_i - (\hat{\theta}-\theta^*)^\top x_i)).
\end{aligned}
$$

$$
\mathbb{P}(\mathrm{sgn}(\epsilon_i)\neq\mathrm{sgn}(\epsilon_i - (\hat{\theta}-\theta^*)^\top x_i)\leq\mathbb{P}(|\epsilon|\leq|(\hat{\theta}-\theta^*)^\top x_i|) = O(\sqrt{\frac{1}{n}}) = o(1).
$$

Recall that $\epsilon_i = y_i - \theta^{*\top}x_i = (x_i^\top\beta_2^*)^2 + \xi_i$, then we obtain $x_i\mathrm{sgn}((x_i^\top\beta_2^*)^2 + \xi_i) \overset{d}{=} -x_i\mathrm{sgn}((x_i^\top\beta_2^*)^2 + \xi_i)$. As a result, we have $\mathbb{E}[x_i\mathrm{sgn}(\epsilon_i)] = \mathbb{E}[x_i\mathrm{sgn}((x_i^\top\beta_2^*)^2 + \xi_i)] = 0$, yielding

$$
\frac{1}{n}\sum_{i=1}^{n}\theta^{*\top}x_i\cdot\mathrm{sgn}(\epsilon_i) = O_p(\frac{1}{\sqrt{n}}).
$$

Then, we have

$$
\frac{\Phi}{\mathbb{E}_{x\sim\hat{P}_x}\|x\|_2} = -\frac{\hat{\theta}}{\|\hat{\theta}\|}(\frac{1}{n}\sum_{i=1}^{n}|\epsilon_i| + O_p(\frac{1}{\sqrt{n}})) = -\frac{\hat{\theta}}{\|\hat{\theta}\|}(\mathbb{E}|\epsilon_i| + O_p(\frac{1}{\sqrt{n}}))
$$

Moreover, the Hessian matrix

$$
H_\theta(X^e, Y^e) = 1/n'\sum_{i=1}^{n'}\nabla_\theta^2 l(\hat{\theta}^{\mathcal{M}}_{\min}, x_i^e, y_i^e;\mathcal{A}) = \frac{1}{n}X^{e\top}X^e = \sigma_x^2 I + O_p(\sqrt{\frac{m}{n}}).
$$

Then

$$\hat{S}_\epsilon(\mathcal{L}) = \Phi^\top H_\theta^{-1}(X^e, Y^e)\Phi = (\mathbb{E}_{x\sim\hat{P}_x}\|x\|_2)^2(\mathbb{E}|\epsilon_i| + O_p(\frac{1}{\sqrt{n}}))^2\frac{\hat{\theta}^\top(\sigma_x^2 I + O_p(\sqrt{\frac{m}{n}}))^{-1}\hat{\theta}}{\|\hat{\theta}\|^2}$$
$$= (\mathbb{E}_{x\sim\hat{P}_x}\|x\|_2)^2 \cdot [(\mathbb{E}|\epsilon_i|)^2 + O_p(\frac{1}{\sqrt{n}})] \cdot (\sigma_x^{-2} + O_p(\sqrt{\frac{m}{n}})).$$

Then, let us consider the quadratic basis of the regression setting $(x_i, y_i) \in \mathbb{R}^m \times \mathbb{R}$ are *i.i.d.* draws from a joint distribution $P_{x,y}$, for $i = 1, 2, ..., n$. Suppose we use the basis $v(x) = (v_1(x), ..., v_d(x)) = (x_1, ..., x_m, x_1^2/2, ..., x_m^2/2, \{x_j x_k\}_{j<k})$, to approximate y, and try to solve

$$\hat{\theta} = \arg\min_{\theta\in\mathbb{R}^m} \frac{1}{n}\sum_{i=1}^n l(\theta, x_i, y_i) := \arg\min_\theta \frac{1}{2n}\sum_{i=1}^n (y_i - \theta^T v(x_i))^2.$$

Further, let us define

$$\theta^* = \arg\min_\theta \mathbb{E}_{P_{x,y}}[\frac{1}{2}(Y - \theta^T v(X))^2],$$

denoting the best population linear approximation to Y.

Denote $\epsilon_i = y_i - \theta^{*\top} v(x_i)$. Since the true model is $Y = X^\top\beta_1^* + (X^\top\beta_2^*)^2 + \xi$, and $X \sim N(0, \sigma_x^2 I)$, then $\epsilon_i = \xi_i$ and we have $\mathbb{E}[\mathrm{sgn}(\epsilon_i)x_i] = 0$.

Further, denote
$$\hat{\theta} = \arg\min_\theta \frac{1}{2n}\sum_{i=1} (y_i - \theta^T v(x_i))^2.$$

We have $\|\hat{\theta} - \theta^*\| = O_p(\sqrt{\frac{m}{n}})$.

By definition, for $k \in [m]$,

$$b_k = |\frac{\partial}{\partial x_{\cdot,k}} l(\hat{\theta}^{\mathcal{M}}_{\min}, x_i, y_i)| = |y_i - \hat{\theta}^\top v(x_i)| \cdot |\hat{\theta}^\top \frac{\partial}{\partial x_{\cdot,k}} v(x_i)| = |y_i - \hat{\theta}^\top v(x_i)| \cdot |\hat{\theta}^\top \frac{\partial}{\partial x} v(x_i) e_k|.$$

Therefore, by letting $p = q = 2$ in Eqn (5) of Theorem 4.1,

$$\psi_k^i = \frac{b_k^{q-1}}{(\sum_{k=1}^d b_k^q)^{\frac{1}{p}}}\mathrm{sgn}(\frac{\partial}{\partial x_{\cdot,k}} l(\hat{\theta}^{\mathcal{M}}_{\min}, x_i, y_i)) = \frac{b_k}{(\sum_{k=1}^d b_k^2)^{1/2}}\mathrm{sgn}(\frac{\partial}{\partial x_{\cdot,k}} l(\hat{\theta}^{\mathcal{M}}_{\min}, x_i, y_i))$$
$$= \frac{(\hat{\theta}^\top v(x_i) - y_i)\cdot \hat{\theta}^\top \frac{\partial}{\partial x} v(x_i) e_k}{|y_i - \hat{\theta}^\top v(x_i)| \cdot \|\hat{\theta}^\top \frac{\partial}{\partial x} v(x_i)\|_2} = \frac{\hat{\theta}^\top \frac{\partial}{\partial x} v(x_i) e_k}{\|\hat{\theta}^\top \frac{\partial}{\partial x} v(x_i)\|} \cdot \mathrm{sgn}(\hat{\theta}^\top v(x_i) - y_i).$$

As a result

$$\phi_i^\top = (\psi_1^i, \psi_2^i, \cdots, \psi_m^i) = \mathrm{sgn}(\hat{\theta}^\top v(x_i) - y_i) \cdot \frac{1}{\|\hat{\theta}^\top \frac{\partial}{\partial x} v(x_i)\|} \cdot \hat{\theta}^\top \frac{\partial}{\partial x} v(x_i),$$

and
$$\nabla_x l(\hat{\theta}, x_i, y_i) = (\hat{\theta}^\top v(x_i) - y_i) \cdot (\frac{\partial}{\partial x} v(x_i))^\top \hat{\theta}$$

$$\nabla_{x,\theta} l(\hat{\theta}, x_i, y_i) = v(x_i)\hat{\theta}^\top \frac{\partial}{\partial x} v(x_i) + (\hat{\theta}^\top v(x_i) - y_i) \cdot \frac{\partial}{\partial x} v(x_i)$$

Then

$$
\begin{aligned}
\frac{\Phi}{\mathbb{E}_{x\sim\hat{P}_x}\|x\|_2} =&\frac{1}{n}\sum_{i=1}^{n}\nabla_{x,\theta}l(\hat{\theta}^{\mathcal{M}}_{\min},x_i,y_i)\phi_i\\
=&\frac{1}{n}\sum_{i=1}^{n}[(\hat{\theta}^\top v(x_i)-y_i)\cdot\frac{\partial}{\partial x}v(x_i)+v(x_i)\hat{\theta}^\top\frac{\partial}{\partial x}v(x_i)]\cdot\operatorname{sgn}(\hat{\theta}^\top v(x_i)-y_i)\\
&\cdot\frac{1}{\|\hat{\theta}^\top\frac{\partial}{\partial x}v(x_i)\|}\cdot(\frac{\partial}{\partial x}v(x_i))^\top\hat{\theta}\\
=&\frac{1}{n\|\hat{\theta}^\top\frac{\partial}{\partial x}v(x_i)\|}\sum_{i=1}^{n}[(\hat{\theta}^\top v(x_i)-y_i)\cdot\frac{\partial}{\partial x}v(x_i)(\frac{\partial}{\partial x}v(x_i))^\top\hat{\theta}+v(x_i)\\
&\cdot\|\hat{\theta}^\top\frac{\partial}{\partial x}v(x_i)\|^2]\cdot\operatorname{sgn}(\hat{\theta}^\top v(x_i)-y_i)\\
=&\frac{1}{n\|\hat{\theta}^\top\frac{\partial}{\partial x}v(x_i)\|}\sum_{i=1}^{n}|\hat{\theta}^\top v(x_i)-y_i|\cdot\frac{\partial}{\partial x}v(x_i)(\frac{\partial}{\partial x}v(x_i))^\top\hat{\theta}\\
&+\frac{1}{n}\sum_{i=1}^{n}v(x_i)\cdot\|\hat{\theta}^\top\frac{\partial}{\partial x}v(x_i)\|\cdot\operatorname{sgn}(\hat{\theta}^\top v(x_i)-y_i)\\
=&\mathbb{E}[\frac{|\hat{\theta}^\top v(x_i)-y_i|}{\|\hat{\theta}^\top\frac{\partial}{\partial x}\boldsymbol{v}(x_i)\|}\frac{\partial}{\partial x}v(x_i)(\frac{\partial}{\partial x}v(x_i))^\top\hat{\theta}]+O_p(\sqrt{\frac{m^2}{n}})
\end{aligned}
$$

Then we have

$$
\begin{aligned}
&\|\mathbb{E}[\frac{|\hat{\theta}^\top v(x_i)-y_i|}{\|\hat{\theta}^\top\frac{\partial}{\partial x}\boldsymbol{v}(x_i)\|}\frac{\partial}{\partial x}\boldsymbol{v}(x_i)(\frac{\partial}{\partial x}\boldsymbol{v}(x_i))^\top\hat{\theta}]\|^2\\
\le&\mathbb{E}[\|\frac{|\hat{\theta}^\top v(x_i)-y_i|}{\|\hat{\theta}^\top\frac{\partial}{\partial x}v(x_i)\|}\frac{\partial}{\partial x}v(x_i)(\frac{\partial}{\partial x}v(x_i))^\top\hat{\theta}\|^2]\\
\le&\mathbb{E}[\|(\frac{\partial}{\partial x}v(x_i))^\top\frac{\partial}{\partial x}v(x_i)\|_2\cdot|\hat{\theta}^\top v(x_i)-y_i|^2]\\
\le&\mathbb{E}[\|(\frac{\partial}{\partial x}v(x_i))^\top\frac{\partial}{\partial x}v(x_i)\|_2\cdot|\theta^{*\top}v(x_i)-y_i|^2]+O_p(\sqrt{\frac{m^2}{n}})\\
\le&\mathbb{E}[\lambda_{\max}(\frac{\partial}{\partial x}v(x_i))^\top\frac{\partial}{\partial x}v(x_i))]\cdot((\mathbb{E}|\xi_i|)^2+O_p(\sqrt{\frac{1}{n}}))+O_p(\sqrt{\frac{m^2}{n}})
\end{aligned}
$$

By similar argument, we have

$$
\begin{aligned}
\|\mathbb{E}[\frac{1}{\|\hat{\theta}^\top\frac{\partial}{\partial x}v(x_i)\|}\frac{\partial}{\partial x}v(x_i)(\frac{\partial}{\partial x}v(x_i))^\top\hat{\theta}]\|^2\ge&\mathbb{E}[\lambda_{\min}(\frac{\partial}{\partial x}v(x_i))^\top\frac{\partial}{\partial x}v(x_i))]\cdot((\mathbb{E}|\epsilon_i|)^2\\
&+O_p(\sqrt{\frac{1}{n}}))+O_p(\sqrt{\frac{m^2}{n}}).
\end{aligned}
$$

Recall that $v(x)=(x_1,...,x_m,x_1^2/2,...,x_m^2/2,\{x_jx_k\}_{j<k})$, and the quadratic term in the true model is $(\beta_2^{*\top}x)^2$ (so then $\mathbb{E}|\epsilon|$ is easy to compute), then

$$
\frac{\partial}{\partial x}v(x)=(I_m,\operatorname{diag}(x_1,...,x_m),Perm(x_ix_j))^\top=\begin{bmatrix} I_m\\ D_x\\ Perm(x_ix_j)\end{bmatrix}\in\mathbb{R}^{(m^2+m)\times m},
$$

where $D_x = \text{diag}(x_1, ..., x_d)$, and $Perm(x_i x_j) \in \mathbb{R}^{m\times(m^2-m)}$ with each column being $x_j e_k + x_k e_j$ for $1 \le j < k \le m$.

Then we have

$$(\frac{\partial}{\partial x} v(x_i))^\top \frac{\partial}{\partial x} v(x_i) = I_m + D_Q,$$

where $(D_Q)_{jj} = (x_1^2 + ... + x_m^2)$, $(D_Q)_{jk} = x_j x_k$. As a result, $D_Q = xx^\top + (x_1^2 + ... + x_m^2)I_m - D_x^2$

$$\inf_{v:\|v\|=1} v^\top (xx^\top + (x_1^2 + ... + x_d^2)I_d)v = (x_1^2 + ... + x_d^2) + \inf_v (x^\top v)^2$$

Therefore,

$$1 + (x_1^2 + ... + x_d^2) - \max x_j^2 \le \lambda_{\min}(I_d + D_Q) \le \lambda_{\max}(I_d + D_Q) \le 1 + 2(x_1^2 + ... + x_d^2).$$

Moreover,

$$\begin{aligned} H_\theta(X^e, Y^e) &= \frac{1}{n}\sum_{i=1}^n v(x_i)v(x_i)^\top = E[v(x_i)v(x_i)^\top] + O_p(\sqrt{\frac{m^2}{n}}) \\ &= diag(\sigma_x^2 I_m, \frac{3}{4}\sigma_x^4 I_m, \sigma_x^4 I_{m(m-1)}) + O_p(\sqrt{\frac{m^2}{n}}) \end{aligned}$$

Then

$$\begin{aligned} \hat{S}_\epsilon(\mathcal{Q}) \le& \frac{1}{\max\{\sigma_x^2, \sigma_x^4\}} \mathbb{E}[\lambda_{\max}(\frac{\partial}{\partial x}\boldsymbol{v}(x_i))^\top \frac{\partial}{\partial x}\boldsymbol{v}(x_i))] \cdot ((\mathbb{E}|\xi_i|)^2 + O_p(\sqrt{\frac{1}{n}})) + O_p(\sqrt{\frac{m^2}{n}}) \\ \le& \frac{1}{\max\{\sigma_x^2, \sigma_x^4\}} \mathbb{E}[1 + 2(x_1^2 + ... + x_m^2)] \cdot ((\mathbb{E}|\xi_i|)^2 + O_p(\sqrt{\frac{1}{n}})) + O_p(\sqrt{\frac{m^2}{n}}) \\ \le& \frac{1}{\max\{\sigma_x^2, \sigma_x^4\}} (1 + 2m\sigma_x^2) \cdot ((\mathbb{E}|\xi_i|)^2 + O_p(\sqrt{\frac{1}{n}})) + O_p(\sqrt{\frac{m^2}{n}}) \end{aligned}$$

Similarly,

$$\begin{aligned} \hat{S}_\epsilon(\mathcal{Q}) \ge& \frac{1}{\min\{\sigma_x^2, \frac{3}{4}\sigma_x^4\}} \mathbb{E}[\lambda_{\min}(\frac{\partial}{\partial x}v(x_i))^\top \frac{\partial}{\partial x}v(x_i))] \cdot ((\mathbb{E}|\xi_i|)^2 + O_p(\sqrt{\frac{1}{n}})) + O_p(\sqrt{\frac{m^2}{n}}) \\ \ge& \frac{1}{\min\{\sigma_x^2, \frac{3}{4}\sigma_x^4\}} \mathbb{E}[1 + (x_1^2 + ... + x_m^2) - \max x_j^2] \cdot ((\mathbb{E}|\xi_i|)^2 + O_p(\sqrt{\frac{1}{n}})) + O_p(\sqrt{\frac{m^2}{n}}) \\ \ge& \frac{1}{\min\{\sigma_x^2, \frac{3}{4}\sigma_x^4\}} (1 + (m - 2\log m)\sigma_x^2) \cdot ((\mathbb{E}|\xi_i|)^2 + O_p(\sqrt{\frac{1}{n}})) + O_p(\sqrt{\frac{m^2}{n}}) \end{aligned}$$

Recall that for linear model,

$$\begin{aligned} \hat{S}_\epsilon(\mathcal{L}) &= \Phi^\top H_\theta^{-1}(X^e, Y^e)\Phi = (\mathbb{E}_{x\sim\hat{P}_x}\|x\|_2)^2 (\mathbb{E}|\epsilon_i| + O_p(\frac{1}{\sqrt{n}}))^2 \frac{\hat{\theta}^\top (\sigma_x^2 I + O_p(\sqrt{\frac{m}{n}}))^{-1}\hat{\theta}}{\|\hat{\theta}\|^2} \\ &= (\mathbb{E}_{x\sim\hat{P}_x}\|x\|_2)^2 \cdot [(\mathbb{E}|\epsilon_i|)^2 + O_p(\frac{1}{\sqrt{n}})] \cdot (\sigma_x^{-2} + O_p(\sqrt{\frac{m}{n}})). \end{aligned}$$

Since the true model is $y = \beta_1^{*\top}x + (\beta_2^{*\top}x)^2 + \xi$ with $\xi \sim \mathcal{N}(0, \sigma_\xi^2)$, and $x \sim \mathcal{N}(0, \sigma_x^2 I_m)$, we have

$$(\mathbb{E}|\epsilon_i|)^2 = (\mathbb{E}|(\beta_2^{*\top}x)^2 + \xi|)^2 \in [(\|\beta_2^*\|_2^2\sigma_x^2 - \sqrt{\frac{2}{\pi}}\sigma_\xi)^2, (\|\beta_2^*\|_2^2\sigma_x^2 + \sqrt{\frac{2}{\pi}}\sigma_\xi)^2].$$

Therefore, when $\frac{1}{\sigma_x^2}(\|\beta_2^*\|_2^2\sigma_x^2 - \sqrt{\frac{2}{\pi}}\sigma_\xi)^2 \geq \frac{1}{\max\{\sigma_x^2, \sigma_x^4\}}(1+2m\sigma_x^2)\cdot\frac{2}{\pi}\sigma_\xi^2$, that is, $(\|\beta_2^*\|_2^2\sigma_x^2 - \sqrt{\frac{2}{\pi}}\sigma_\xi)^2 \geq \frac{1+2m\sigma_x^2}{\max\{\sigma_x^2, 1\}} \cdot \frac{2}{\pi}\sigma_\xi^2$,

$$\hat{\Delta}(\mathcal{L}) \geq \hat{\Delta}(\mathcal{Q}) + O_p(\sqrt{\frac{m^2}{n}}).$$

On the other hand, $\frac{1}{\sigma_x^2}(\|\beta_2^*\|_2^2\sigma_x^2 + \sqrt{\frac{2}{\pi}}\sigma_\xi)^2 \leq \frac{1}{\min\{\sigma_x^2, \frac{3}{4}\sigma_x^4\}}(1 + m\sigma_x^2 - 2\sigma_x^2 \cdot \log m) \cdot \frac{3}{2\pi}\sigma_\epsilon^2$, that is, $(\|\beta_2^*\|_2^2\sigma_x^2 + \sqrt{\frac{2}{\pi}}\sigma_\xi)^2 \leq \frac{1}{\min\{1, \frac{3}{4}\sigma_x^2\}}(1 + m\sigma_x^2 - 2\sigma_x^2 \cdot \log m) \cdot \frac{3}{2\pi}\sigma_\epsilon^2$,

$$\hat{\Delta}(\mathcal{L}) \leq \hat{\Delta}(\mathcal{Q}) + O_p(\sqrt{\frac{m^2}{n}}).$$

Then let us consider the case where $p = \infty, q = 1$ in Eqn (5) of Theorem 4.1,

$$\begin{aligned}
\nu_k^i &= \frac{b_k^{q-1}}{(\sum_{k=1}^d b_k^q)^{\frac{1}{p}}}\operatorname{sgn}(\frac{\partial}{\partial x_{\cdot,k}} l(\hat{\theta}_{\min}^{\mathcal{M}}, x_i, y_i)) = \operatorname{sgn}(\frac{\partial}{\partial x_{\cdot,k}} l(\hat{\theta}_{\min}^{\mathcal{M}}, x_i, y_i)) \\
&= \operatorname{sgn}(\theta^\top \frac{\partial}{\partial x}\boldsymbol{v}(x_i)e_k) \cdot \operatorname{sgn}(\theta^\top \boldsymbol{v}(x_i) - y_i).
\end{aligned}$$

Then

$$\begin{aligned}
\Phi =& \frac{1}{n}\sum_{i=1}^n \nabla_{x,\theta} l(\hat{\theta}_{\min}^{\mathcal{M}}, x_i, y_i)\phi_i \\
=& \frac{1}{n}\sum_{i=1}^n [(\theta^\top v(x_i) - y_i)\cdot \frac{\partial}{\partial x}v(x_i) + v(x_i)\theta^\top \frac{\partial}{\partial x}v(x_i)] \cdot \operatorname{sgn}(\theta^\top v(x_i) - y_i) \cdot \operatorname{sgn}((\frac{\partial}{\partial x}\boldsymbol{v}(x_i))^\top\theta) \\
=& \frac{1}{n}\sum_{i=1}^n |\theta^\top v(x_i) - y_i| \cdot \frac{\partial}{\partial x}v(x_i)\operatorname{sgn}((\frac{\partial}{\partial x}v(x_i))^\top\theta) + \frac{1}{n}\sum_{i=1}^n v(x_i)\cdot\|\theta^\top\frac{\partial}{\partial x}v(x_i)\|_1 \cdot \operatorname{sgn}(\theta^\top v(x_i) - y_i) \\
=& \mathbb{E}[|\theta^\top v(x_i) - y_i| \cdot \frac{\partial}{\partial x}v(x_i)\operatorname{sgn}((\frac{\partial}{\partial x}v(x_i))^\top\theta)] + O_p(\frac{m}{\sqrt{n}})
\end{aligned}$$

By similar argument, since $\|\operatorname{sgn}((\frac{\partial}{\partial x}v(x_i))^\top\theta\| = \sqrt{d}$ we have

$$\begin{aligned}
&\|\mathbb{E}[|\theta^\top v(x_i) - y_i| \cdot \frac{\partial}{\partial x}v(x_i)\operatorname{sgn}((\frac{\partial}{\partial x}v(x_i))^\top\theta\|^2/d \\
\geq& \mathbb{E}[\lambda_{\min}(\frac{\partial}{\partial x}v(x_i))^\top\frac{\partial}{\partial x}v(x_i))] \cdot ((\mathbb{E}|\epsilon_i|)^2 + O_p(\sqrt{\frac{1}{n}})) + O_p(\sqrt{\frac{m^2}{n}}).
\end{aligned}$$

$$\begin{aligned}
&\|\mathbb{E}[|\theta^\top v(x_i) - y_i| \cdot \frac{\partial}{\partial x}\boldsymbol{v}(x_i)\operatorname{sgn}((\frac{\partial}{\partial x}v(x_i))^\top\theta\|^2/d \\
\leq& \mathbb{E}[\lambda_{\max}(\frac{\partial}{\partial x}v(x_i))^\top\frac{\partial}{\partial x}v(x_i))] \cdot ((\mathbb{E}|\epsilon_i|)^2 + O_p(\sqrt{\frac{1}{n}})) + O_p(\sqrt{\frac{m^2}{n}}).
\end{aligned}$$

In addition, for the class of linear models, we have

$$\hat{\Delta}(\mathcal{A}_{lin}) = d \cdot (\mathbb{E}|(\beta_2^\top x)^2 + \epsilon|)^2 \in [(\|\beta_2\|_2^2\sigma_x^2 - \sqrt{\frac{2}{\pi}}\sigma_\epsilon)^2, (\|\beta_2\|_2^2\sigma_x^2 + \sqrt{\frac{2}{\pi}}\sigma_\epsilon)^2].$$

Therefore, using the exact same statement as the previous case where $p = q = 2$, we get the desired result.

A.3 Proof of Theorem 5.2 and Corollary 5.1

Now let us first compute the AIF for linear models.

Specifically, let us consider the regression setting $(x_i, y_i) \in \mathbb{R}^m \times \mathbb{R}$ are *i.i.d.* draws from a joint distribution $P_{x,y}$, for $i = 1, 2, ..., n$. Note that we don't assume linear relationship, but the linear regression model tries to find the best linear approximation by solving

$$\hat{\theta} = \arg\min_\theta \frac{1}{n}\sum_{i=1}^n l(\theta, x_i, y_i) := \arg\min_\theta \frac{1}{2n}\sum_{i=1}^n (y_i - \theta^T x_i)^2,$$

where we use $l(\theta, x_i, y_i) = \frac{1}{2}(y_i - \theta^T x_i)^2$ as the loss function.

Further, let us define

$$\beta_{\min}^{\mathcal{L}} = \arg\min_\theta \mathbb{E}_{P_{x,y}}[\frac{1}{2}(Y - \theta^T X)^2],$$

denoting the best population linear approximation to Y, which makes $Cov(x_i, y_i - \theta^{*\top} x_i) = 0$. Denote $\eta_i^{\mathcal{L}} = y_i - \beta_{\min}^{\mathcal{L}\top} x_i$, we then have $\mathbb{E}[\eta_i^{\mathcal{L}} x_i] = 0$.

Further, denote $\eta_i^{\mathcal{L}} = y_i - \beta_{\min}^{\mathcal{L}\top} x_i$, and

$$\hat{\beta} = \arg\min_\theta \frac{1}{2n}\sum_{i=1}(y_i - \theta^T x_i)^2,$$

and we have $\|\hat{\beta} - \beta_{\min}^{\mathcal{L}}\|_2 = O_p(\sqrt{\frac{m}{n}})$.

By definition, for $k \in [m]$,

$$b_k = |\frac{\partial}{\partial x_{\cdot,k}} l(\hat{\beta}, x_i, y_i, \mathcal{L})| = |y_i - \hat{\beta}^\top x_i| \cdot |\hat{\beta}_k|,$$

and therefore, by letting $p = q = 2$ in Eqn (5) of Theorem 4.1,

$$\begin{aligned}\psi_k^i &= \frac{b_k^{q-1}}{(\sum_{k=1}^d b_k^q)^{\frac{1}{p}}}\mathrm{sgn}(\frac{\partial}{\partial x_{\cdot,k}} l(\hat{\beta}, x_i, y_i, \mathcal{M})) = \frac{b_k}{(\sum_{k=1}^d b_k^2)^{1/2}}\mathrm{sgn}((y_i - \hat{\beta}^\top x_i)\cdot\hat{\beta}_k) \\ &= \frac{(y_i - \hat{\beta}^\top x_i)\cdot\hat{\beta}_k}{|y_i - \theta^\top x_i|\cdot\|\hat{\beta}\|_2} = \frac{\hat{\beta}_k}{\|\hat{\beta}\|}\cdot\mathrm{sgn}(y_i - \hat{\beta}^\top x_i).\end{aligned}$$

As a result

$$\phi_i = (\psi_1^i, \psi_2^i, \cdots, \psi_m^i)^T = \mathrm{sgn}(y_i - \hat{\beta}^\top x_i)\cdot\frac{1}{\|\hat{\beta}\|}\cdot\hat{\beta},$$

and

$$\begin{aligned}
\frac{\Phi}{\mathbb{E}_{x\sim\hat{P}_x}\|x\|_2} =&\frac{1}{n}\sum_{i=1}^{n}\nabla_{x,\theta}l(\hat{\beta},x_i,y_i,\mathcal{M})\phi_i = \frac{1}{n}\sum_{i=1}^{n}[(\hat{\beta}^\top x_i - y_i)\cdot I_d + \hat{\beta}x_i^\top]\cdot \mathrm{sgn}(y_i-\hat{\beta}^\top x_i)\cdot\frac{1}{\|\hat{\beta}\|}\cdot\hat{\beta}\\
=&\frac{1}{n\|\hat{\beta}\|}\sum_{i=1}^{n}[(\hat{\beta}^\top x_i - y_i)\cdot\hat{\beta} + \hat{\beta}x_i^\top\hat{\beta}]\cdot\mathrm{sgn}(y_i-\hat{\beta}^\top x_i)\\
=&-\frac{1}{n\|\hat{\beta}\|}\sum_{i=1}^{n}(y_i\cdot\hat{\beta})\cdot\mathrm{sgn}(y_i-\hat{\beta}^\top x_i)\\
=&-\frac{\hat{\beta}}{n\|\hat{\beta}\|}\sum_{i=1}^{n}y_i\cdot\mathrm{sgn}(y_i-\hat{\beta}^\top x_i)\\
=&-\frac{\hat{\beta}}{n\|\hat{\beta}\|}\sum_{i=1}^{n}(\eta_i^{\mathcal{L}}+\beta_{\min}^{\mathcal{L}\top}x_i)\cdot\mathrm{sgn}(y_i-\hat{\beta}^\top x_i)\\
=&-\frac{\hat{\beta}}{n\|\hat{\beta}\|}\sum_{i=1}^{n}\beta_{\min}^{\mathcal{L}\top}x_i\cdot\mathrm{sgn}(y_i-\hat{\beta}^\top x_i)-\frac{\hat{\beta}}{n\|\hat{\beta}\|}\sum_{i=1}^{n}\eta_i^{\mathcal{L}}\cdot\mathrm{sgn}(y_i-\hat{\beta}^\top x_i)\\
=&-\frac{\hat{\beta}}{n\|\hat{\beta}\|}\sum_{i=1}^{n}\beta_{\min}^{\mathcal{L}\top}x_i\cdot\mathrm{sgn}(\eta_i^{\mathcal{L}})-\frac{\hat{\beta}}{n\|\hat{\beta}\|}\sum_{i=1}^{n}\eta_i^{\mathcal{L}}\cdot\mathrm{sgn}(\eta_i^{\mathcal{L}})\\
&+\frac{\hat{\beta}}{n\|\hat{\beta}\|}\sum_{i=1}^{n}\beta_{\min}^{\mathcal{L}\top}x_i\cdot(\mathrm{sgn}(\eta_i^{\mathcal{L}})-\mathrm{sgn}(\eta_i^{\mathcal{L}}-(\hat{\beta}-\beta_{\min}^{\mathcal{L}})^\top x_i))\\
&+\frac{\hat{\beta}}{n\|\hat{\beta}\|}\sum_{i=1}^{n}\eta_i^{\mathcal{L}}\cdot(\mathrm{sgn}(\eta_i^{\mathcal{L}})-\mathrm{sgn}(\eta_i^{\mathcal{L}}-(\hat{\beta}-\beta_{\min}^{\mathcal{L}})^\top x_i)).
\end{aligned}$$

Then we have

$$\mathbb{P}(\mathrm{sgn}(\eta_i^{\mathcal{L}})\neq\mathrm{sgn}(\eta_i^{\mathcal{L}}-(\hat{\beta}-\beta_{\min}^{\mathcal{L}})^\top x_i)\leq\mathbb{P}(|\epsilon|\leq|(\hat{\beta}-\beta_{\min}^{\mathcal{L}})^\top x_i|)=O(\sqrt{\frac{1}{n}})=o(1).$$

Recall that $\mathbb{E}[x_i\mathrm{sgn}(\eta_i^{\mathcal{L}})]=\mathbb{E}[x_i\mathrm{sgn}((x_i^\top\beta_2^*)^2+\xi_i)]=0$, we have

$$\frac{1}{n}\sum_{i=1}^{n}\beta_{\min}^{\mathcal{L}\top}x_i\cdot\mathrm{sgn}(\eta_i^{\mathcal{L}})=O_p(\frac{1}{\sqrt{n}}).$$

Then, we have

$$\frac{\Phi}{\mathbb{E}_{x\sim\hat{P}_x}\|x\|_2}=-\frac{\hat{\beta}}{\|\hat{\beta}\|}(\frac{1}{n}\sum_{i=1}^{n}|\eta_i^{\mathcal{L}}|+O_p(\frac{1}{\sqrt{n}}))=-\frac{\hat{\beta}}{\|\hat{\beta}\|}(\mathbb{E}|\eta_i^{\mathcal{L}}|+O_p(\frac{1}{\sqrt{n}}))$$

Moreover, the Hessian matrix

$$H_\theta(X^e,Y^e)=1/n'\sum_{i=1}^{n'}\nabla_\theta^2 l(\hat{\beta},x_i^e,y_i^e;\mathcal{A})=\frac{1}{n}X^{e\top}X^e=\sigma_x^2 I+O_p(\sqrt{\frac{m}{n}}).$$

Then, we have

$$\hat{S}_\epsilon(\mathcal{L}) = \Phi^\top H_\theta^{-1}(X^e, Y^e)\Phi = (\mathbb{E}_{x\sim\hat{P}_x}\|x\|_2)^2(\mathbb{E}|\eta_i^{\mathcal{L}}| + O_p(\frac{1}{\sqrt{n}}))^2\frac{\hat{\beta}^\top(\sigma_x^2 I + O_p(\sqrt{\frac{m}{n}}))^{-1}\hat{\beta}}{\|\hat{\beta}\|^2}$$
$$= \epsilon^2(\mathbb{E}_{x\sim\hat{P}_x}\|x\|_2)^2\cdot[(\mathbb{E}|\eta_i^{\mathcal{L}}|)^2 + O_p(\frac{1}{\sqrt{n}})]\cdot(\sigma_x^{-2} + O_p(\sqrt{\frac{m}{n}}))$$
$$= \epsilon^2(\mathbb{E}_{x\sim\hat{P}_x}\|x\|_2)^2\cdot(\mathbb{E}|\eta_i^{\mathcal{L}}|)^2\cdot\sigma_x^{-2} + O_p(\sqrt{\frac{m}{n}}). \tag{10}$$

Now, let us consider the random effect model in Corollary 5.1, when the true model is $y = \beta^\top x + \xi$, where $x \in \mathbb{R}^M$, $\beta_1, ..., \beta_M \overset{i.i.d.}{\sim} N(0,1)$, $\xi \sim \mathcal{N}(0, \sigma_\xi^2)$, and $x_1, ..., x_n \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_x^2 I_M)$. Then when we only include m features in the linear predictive model, the residual

$$\eta_i^{\mathcal{L}} = \xi + x_{i,m+1}\beta_{m+1} + ... + x_{i,M}\beta_M.$$

Then conditional on β, we have

$$\eta_i^{\mathcal{L}} \sim N(0, \sigma_\xi^2 + (\beta_{m+1}^2 + ... + \beta_M^2)\sigma_x^2).$$

We then have $\mathbb{E}[|\eta_i^{\mathcal{L}}|]^2 = \frac{2}{\pi}(\sigma_\xi^2 + (\beta_{m+1}^2 + ... + \beta_M^2)\sigma_x^2)$ Take expectation w.r.t β, we have $\mathbb{E}[|\eta_i^{\mathcal{L}}|]^2 = \frac{2}{\pi}(\sigma_\xi^2 + (M-m)\sigma_x^2)$.

For $\mathbb{E}_{x\sim\hat{P}_x}\|x\|_2$, we have

$$\mathbb{E}_{x\sim\hat{P}_x}\|x\|_2 = \mathbb{E}[\sqrt{\beta_1^2 + ... + \beta_m^2}] = \sqrt{\frac{\frac{m+1}{2}}{\frac{m}{2}}}.$$

Plug into (10), we get

$$\mathbb{E}[\hat{S}_\varepsilon(\mathcal{L})] = \frac{4\epsilon^2}{\pi\sigma_x^2}\frac{\Gamma^2(\frac{m+1}{2})}{\Gamma^2(\frac{m}{2})}\cdot((M-m)\sigma_x^2 + \sigma_\xi^2) + O_p(\epsilon^2\sqrt{\frac{m^2}{n}}).$$

A.4 Proof of Theorem 5.3

Now let us consider the general basis of the regression setting $(x_i, y_i) \in \mathbb{R}^m \times \mathbb{R}$ are *i.i.d.* draws from a joint distribution $P_{x,y}$, for $i = 1, 2, ..., n$. Suppose we use the basis $v(x) = (v_1(x), ..., v_d(x)) = (x_1, ..., x_m, x_1^2/2, ..., x_m^2/2, \{x_j x_k\}_{j<k})$, to approximate y, and try to solve

$$\hat{\theta} = \arg\min_{\theta\in\mathbb{R}^m}\frac{1}{n}\sum_{i=1}^n l(\theta, x_i, y_i) := \arg\min_\theta\frac{1}{2n}\sum_{i=1}^n(y_i - \theta^T v(x_i))^2.$$

Further, let us define

$$\theta^* = \arg\min_\theta \mathbb{E}_{P_{x,y}}[\frac{1}{2}(Y - \theta^T v(X))^2],$$

denoting the best population linear approximation to Y.

Denote $\xi_i = y_i - \theta^{*\top}v(x_i)$, and let

$$\hat{\theta} = \arg\min_\theta\frac{1}{2n}\sum_{i=1}(y_i - \theta^T v(x_i))^2.$$

We have $\|\hat{\theta}-\theta^*\| = O_p(\sqrt{\frac{m}{n}})$.

By definition, for $k \in [m]$,

$$b_k = |\frac{\partial}{\partial x_{\cdot,k}} l(\hat{\theta}^{\mathcal{M}}_{\min}, x_i, y_i)| = |y_i - \hat{\theta}^\top v(x_i)| \cdot |\hat{\theta}^\top \frac{\partial}{\partial x_{\cdot,k}} v(x_i)| = |y_i - \hat{\theta}^\top v(x_i)| \cdot |\hat{\theta}^\top \frac{\partial}{\partial x} v(x_i) e_k|.$$

Therefore, by letting $p = q = 2$ in Eqn (5) of Theorem 4.1,

$$\begin{aligned}
\psi_k^i &= \frac{b_k^{q-1}}{(\sum_{k=1}^d b_k^q)^{\frac{1}{p}}} \mathrm{sgn}(\frac{\partial}{\partial x_{\cdot,k}} l(\hat{\theta}^{\mathcal{M}}_{\min}, x_i, y_i)) = \frac{b_k}{(\sum_{k=1}^d b_k^2)^{1/2}} \mathrm{sgn}(\frac{\partial}{\partial x_{\cdot,k}} l(\hat{\theta}^{\mathcal{M}}_{\min}, x_i, y_i)) \\
&= \frac{(\hat{\theta}^\top v(x_i) - y_i) \cdot \hat{\theta}^\top \frac{\partial}{\partial x} v(x_i) e_k}{|y_i - \hat{\theta}^\top v(x_i)| \cdot \|\hat{\theta}^\top \frac{\partial}{\partial x} v(x_i)\|_2} = \frac{\hat{\theta}^\top \frac{\partial}{\partial x} v(x_i) e_k}{\|\hat{\theta}^\top \frac{\partial}{\partial x} v(x_i)\|} \cdot \mathrm{sgn}(\hat{\theta}^\top v(x_i) - y_i).
\end{aligned}$$

As a result

$$\phi_i^\top = (\psi_1^i, \psi_2^i, \cdots, \psi_m^i) = \mathrm{sgn}(\hat{\theta}^\top v(x_i) - y_i) \cdot \frac{1}{\|\hat{\theta}^\top \frac{\partial}{\partial x} v(x_i)\|} \cdot \hat{\theta}^\top \frac{\partial}{\partial x} v(x_i),$$

and

$$\nabla_x l(\hat{\theta}, x_i, y_i) = (\hat{\theta}^\top v(x_i) - y_i) \cdot (\frac{\partial}{\partial x} v(x_i))^\top \hat{\theta}$$

$$\nabla_{x,\theta} l(\hat{\theta}, x_i, y_i) = v(x_i) \hat{\theta}^\top \frac{\partial}{\partial x} v(x_i) + (\hat{\theta}^\top v(x_i) - y_i) \cdot \frac{\partial}{\partial x} v(x_i)$$

Then

$$\begin{aligned}
\frac{\Phi}{\mathbb{E}_{x\sim \hat{P}_x} \|x\|_2} =& \frac{1}{n} \sum_{i=1}^n \nabla_{x,\theta} l(\hat{\theta}^{\mathcal{M}}_{\min}, x_i, y_i) \phi_i \\
=& \frac{1}{n} \sum_{i=1}^n [(\hat{\theta}^\top v(x_i) - y_i) \cdot \frac{\partial}{\partial x} v(x_i) + v(x_i) \hat{\theta}^\top \frac{\partial}{\partial x} v(x_i)] \cdot \mathrm{sgn}(\hat{\theta}^\top v(x_i) - y_i) \cdot \\
& \frac{1}{\|\hat{\theta}^\top \frac{\partial}{\partial x} v(x_i)\|} \cdot (\frac{\partial}{\partial x} v(x_i))^\top \hat{\theta} \\
=& \frac{1}{n \|\hat{\theta}^\top \frac{\partial}{\partial x} v(x_i)\|} \sum_{i=1}^n [(\hat{\theta}^\top v(x_i) - y_i) \cdot \\
& \frac{\partial}{\partial x} v(x_i) (\frac{\partial}{\partial x} v(x_i))^\top \hat{\theta} + v(x_i) \cdot \|\hat{\theta}^\top \frac{\partial}{\partial x} v(x_i)\|^2] \cdot \mathrm{sgn}(\hat{\theta}^\top v(x_i) - y_i) \\
=& \frac{1}{n \|\hat{\theta}^\top \frac{\partial}{\partial x} v(x_i)\|} \sum_{i=1}^n |\hat{\theta}^\top v(x_i) - y_i| \cdot \frac{\partial}{\partial x} v(x_i) (\frac{\partial}{\partial x} v(x_i))^\top \hat{\theta} \\
& + \frac{1}{n} \sum_{i=1}^n v(x_i) \cdot \|\hat{\theta}^\top \frac{\partial}{\partial x} v(x_i)\| \cdot \mathrm{sgn}(\hat{\theta}^\top v(x_i) - y_i).
\end{aligned}$$

Recall that we assume $\mathbb{E}[\mathrm{sgn}(\epsilon_i) x_i] = 0$, then we have

$$\frac{\Phi}{\mathbb{E}_{x\sim \hat{P}_x} \|x\|_2} = \mathbb{E}[\frac{|\hat{\theta}^\top v(x_i) - y_i|}{\|\hat{\theta}^\top \frac{\partial}{\partial x} \boldsymbol{v}(x_i)\|} \frac{\partial}{\partial x} \boldsymbol{v}(x_i) (\frac{\partial}{\partial x} \boldsymbol{v}(x_i))^\top \hat{\theta}] + O_p(\sqrt{\frac{d}{n}}).$$

Then, since

$$
\begin{aligned}
&\|\mathbb{E}[\frac{|\hat{\theta}^\top v(x_i)-y_i|}{\|\hat{\theta}^\top\frac{\partial}{\partial x}v(x_i)\|}\frac{\partial}{\partial x}v(x_i)(\frac{\partial}{\partial x}v(x_i))^\top\hat{\theta}]\|^2\\
\leq&\mathbb{E}[\|\frac{|\hat{\theta}^\top v(x_i)-y_i|}{\|\hat{\theta}^\top\frac{\partial}{\partial x}v(x_i)\|}\frac{\partial}{\partial x}v(x_i)(\frac{\partial}{\partial x}v(x_i))^\top\hat{\theta}\|^2]\\
\leq&\mathbb{E}[\|(\frac{\partial}{\partial x}v(x_i))^\top\frac{\partial}{\partial x}v(x_i)\|_2\cdot|\hat{\theta}^\top v(x_i)-y_i|^2]\\
\leq&\mathbb{E}[\|(\frac{\partial}{\partial x}v(x_i))^\top\frac{\partial}{\partial x}v(x_i)\|_2\cdot|\theta^{*\top} v(x_i)-y_i|^2]+O_p(\sqrt{\frac{d}{n}})\\
\leq&\mathbb{E}[\lambda_{\max}(\frac{\partial}{\partial x}\boldsymbol{v}(x_i))^\top\frac{\partial}{\partial x}v(x_i))]\cdot((\mathbb{E}|\xi_i|)^2+O_p(\sqrt{\frac{1}{n}}))+O_p(\sqrt{\frac{d}{n}})
\end{aligned}
$$

Moreover,

$$
H_\theta(X^e,Y^e)=\frac{1}{n}\sum_{i=1}^n v(x_i)v(x_i)^\top=E[v(x_i)v(x_i)^\top]+O_p(\sqrt{\frac{d}{n}})
$$

Then

$$
\hat{S}_\epsilon(\mathcal{GL})\leq\frac{1}{\lambda_{\min}(E[v(x_i)v(x_i)^\top])}\mathbb{E}[\lambda_{\max}(\frac{\partial}{\partial x}v(x_i))^\top\frac{\partial}{\partial x}v(x_i))]\cdot((\mathbb{E}|\xi_i|)^2+O_p(\sqrt{\frac{1}{n}}))+O_p(\sqrt{\frac{m^2}{n}}).
$$

A.5 Proof of Theorem 5.4

Now let us first recall the AIF for linear models.

Specifically, let us consider the regression setting $(x_i,y_i)\in\mathbb{R}^d\times\mathbb{R}$ are *i.i.d.* draws from a joint distribution $P_{x,y}$, for $i=1,2,...,n$. Note that we don't assume linear relationship, but the linear regression model tries to find the best linear approximation by solving

$$
\hat{\theta}=\arg\min_\theta\frac{1}{n}\sum_{i=1}^n l(\theta,\tilde{x}_i,y_i):=\arg\min_\theta\frac{1}{2n}\sum_{i=1}^n(y_i-\theta^T\tilde{x}_i)^2,
$$

where $\tilde{x}_i=x_i+\vartheta_i$ we use $l(\theta,\tilde{x}_i,y_i)=\frac{1}{2}(y_i-\theta^T\tilde{x}_i)^2$ as the loss function.

Further, let us define

$$
\beta^*=\arg\min_\theta\mathbb{E}_{P_{x,y}}[\frac{1}{2}(y-\theta^Tx)^2]=(\mathbb{E}[xx^\top])^{-1}\mathbb{E}[xy]=(\sigma_x^2)^{-1}\mathbb{E}[xy],
$$

$$
\beta_{\min}^{\mathcal{L}}=\arg\min_\theta\mathbb{E}_{P_{x,y}}[\frac{1}{2}(y-\theta^T\tilde{x})^2]=(\mathbb{E}[\tilde{x}\tilde{x}^\top])^{-1}\mathbb{E}[\tilde{x}y]=(\sigma_x^2+\sigma_r^2)^{-1}\mathbb{E}[xy]=\frac{\sigma_x^2}{\sigma_x^2+\sigma_r^2}\beta^*,
$$

denoting the best population linear approximation to Y, which makes $Cov(x_i,y_i-\theta^{*\top}x_i)=0$. Denote $\eta_i^{\mathcal{L}}=y_i-\beta_{\min}^{\mathcal{L}\top}x_i$, we then have $\mathbb{E}[\eta_i^{\mathcal{L}}x_i]=0$.

Suppose $y_i=\beta^*x_i+\xi_i$, then $y_i-\beta_{\min}^{\mathcal{L}\top}\tilde{x}_i=y_i-\beta_{\min}^{\mathcal{L}\top}x_i-\beta_{\min}^{\mathcal{L}\top}\vartheta_i=(\beta^*-\beta_{\min}^{\mathcal{L}})^\top x_i+\xi_i-\beta_{\min}^{\mathcal{L}\top}\vartheta_i$

$\mathbb{E}([y_i-\beta_{\min}^{\mathcal{L}\top}\tilde{x}_i)\tilde{x}_i]=0$

$$
\begin{aligned}
Var(y_i - \beta_{\min}^{\mathcal{L}\top}\tilde{x}_i) =& Var(y_i - \frac{\sigma_x^2}{\sigma_x^2+\sigma_r^2}\beta^{*\top}\tilde{x}_i) \\
=& Var(y_i - \frac{\sigma_x^2}{\sigma_x^2+\sigma_r^2}\beta^{*\top}x_i) + Var(\frac{\sigma_x^2}{\sigma_x^2+\sigma_r^2}\beta^{*\top}\vartheta_i) \\
=& Var(\beta^{*\top}x_i + \xi_i - \frac{\sigma_x^2}{\sigma_x^2+\sigma_r^2}\beta^{*\top}x_i) + Var(\frac{\sigma_x^2}{\sigma_x^2+\sigma_r^2}\beta^{*\top}\vartheta_i) \\
=& Var(\frac{\sigma_r^2}{\sigma_x^2+\sigma_r^2}\beta^{*\top}x_i) + Var(\xi_i) + Var(\frac{\sigma_x^2}{\sigma_x^2+\sigma_r^2}\beta^{*\top}\vartheta_i) \\
=& (\frac{\sigma_r^2\sigma_x^2}{\sigma_x^2+\sigma_r^2})\|\beta^*\|_2^2 + \sigma_\xi^2 + (\frac{\sigma_r^2\sigma_x^2}{\sigma_x^2+\sigma_r^2})\|\beta^*\|_2^2 \\
=& (\frac{2\sigma_r^2\sigma_x^2}{\sigma_x^2+\sigma_r^2})\|\beta^*\|_2^2 + \sigma_\xi^2
\end{aligned}
$$

Further, denote $\eta_i^{\mathcal{L}} = y_i - \beta_{\min}^{\mathcal{L}\top}x_i$, and

$$
\hat{\beta} = \arg\min_\theta \frac{1}{2n}\sum_{i=1}(y_i - \theta^T x_i)^2,
$$

and we have $\|\hat{\beta} - \beta_{\min}^{\mathcal{L}}\|_2 = O_p(\sqrt{\frac{m}{n}})$. Then, we have

$$
\frac{\Phi}{\mathbb{E}_{x\sim\hat{P}_x}\|x\|_2} = -\frac{\hat{\beta}}{\|\hat{\beta}\|}(\frac{1}{n}\sum_{i=1}^n |\eta_i^{\mathcal{L}}| + O_p(\frac{1}{\sqrt{n}})) = -\frac{\hat{\beta}}{\|\hat{\beta}\|}(\mathbb{E}|\eta_i^{\mathcal{L}}| + O_p(\frac{1}{\sqrt{n}}))
$$

Moreover, the Hessian matrix on the test data

$$
H_\theta(X^e, Y^e) = 1/n' \sum_{i=1}^{n'} \nabla_\theta^2 l(\hat{\beta}, x_i^e, y_i^e; \mathcal{A}) = \frac{1}{n}X^{e\top}X^e = \sigma_x^2 I + O_p(\sqrt{\frac{m}{n}}).
$$

Then we have

$$
\begin{aligned}
\hat{S}_\epsilon(\mathcal{L}) &= \Phi^\top H_\theta^{-1}(X^e, Y^e)\Phi = (\mathbb{E}_{x\sim\hat{P}_x}\|x\|_2)^2(\mathbb{E}|\eta_i^{\mathcal{L}}| + O_p(\frac{1}{\sqrt{n}}))^2\frac{\hat{\beta}^\top(\sigma_x^2 I + O_p(\sqrt{\frac{m}{n}}))^{-1}\hat{\beta}}{\|\hat{\beta}\|^2} \\
&= (\mathbb{E}_{x\sim\hat{P}_x}\|x\|_2)^2 \cdot \epsilon^2 \cdot [(\mathbb{E}|\eta_i^{\mathcal{L}}|)^2 + O_p(\frac{1}{\sqrt{n}})] \cdot (\sigma_x^{-2} + O_p(\sqrt{\frac{m}{n}})).
\end{aligned}
$$

and

$$
\begin{aligned}
\hat{S}_\epsilon(\mathcal{L}_{noise}) &= \Phi^\top H_\theta^{-1}(X^e, Y^e)\Phi = (\mathbb{E}_{x\sim\hat{P}_x}\|x\|_2)^2(\mathbb{E}|\eta_i^{\mathcal{L}_{noise}}| + O_p(\frac{1}{\sqrt{n}}))^2\frac{\hat{\beta}^\top((\sigma_x^2+\sigma_r^2)I + O_p(\sqrt{\frac{m}{n}}))^{-1}\hat{\beta}}{\|\hat{\beta}\|^2} \\
&= (\mathbb{E}_{x\sim\hat{P}_x}\|x\|_2)^2 \cdot \epsilon^2 \cdot [(\mathbb{E}|\eta_i^{\mathcal{L}_{noise}}|)^2 + O_p(\frac{1}{\sqrt{n}})] \cdot ((\sigma_x^2+\sigma_r^2)^{-1} + O_p(\sqrt{\frac{m}{n}})).
\end{aligned}
$$

Then

$$
\begin{aligned}
\frac{\hat{S}_\epsilon(\mathcal{L}_{noise})}{\hat{S}_\epsilon(\mathcal{L})} =& \frac{\sigma_x^2}{\sigma_x^2+\sigma_r^2} \cdot \frac{(\frac{2\sigma_r^2\sigma_x^2}{\sigma_x^2+\sigma_r^2})\|\beta_{\min}^{\mathcal{L}}\|_2^2 + \sigma_\xi^2}{\sigma_\xi^2} + O(\sqrt{\frac{m}{n}}) \\
=& \frac{\sigma_x^2/\sigma_\xi^2}{\sigma_x^2+\sigma_r^2} \cdot \left((\frac{2\sigma_r^2\sigma_x^2}{\sigma_x^2+\sigma_r^2})\|\beta_{\min}^{\mathcal{L}}\|_2^2 + \sigma_\xi^2\right) + O(\sqrt{\frac{m}{n}})
\end{aligned}
$$

A.6 Proof of Corollary 6.1

We consider kernel regression in the following form:

$$\hat{\mathcal{L}}_n(\theta, X, Y) = \frac{1}{n}\sum_{i=1}^{n}\big(y_i - \sum_{j=1}^{n} K(x_i, x_j)\theta_j\big)^2 + \lambda\|\theta\|_2^2.$$

Let us denote $K(x_i) := \big(K(x_i, x_1), K(x_i, x_2), \cdots, K(x_i, x_n)\big)^T$. The proof of Corollary 6.1 is almost the same as Theorem 4.1, with slightly modification. Actually, the loss can be in a general form as $\hat{\mathcal{L}}_n(\theta, X, Y)$, our proof for Theorem 4.1can still be applied. Since

$$\nabla_\theta \hat{\mathcal{L}}_n(\theta, X, Y) = \frac{1}{n}\sum_{i=1}^{n} 2\Big(K(x_i)^T\theta - y_i\Big)K(x_i) + 2\lambda\theta,$$

we have

$$\begin{aligned}
\nabla_{x_k,\theta}\hat{\mathcal{L}}_n(\theta, X, Y) &= \frac{2}{n}\sum_{i=1}^{n}\nabla_{x_k}\Big(K(x_i)^T\theta - y_i\Big)K(x_i)\\
&= \frac{2}{n}\sum_{i=1}^{n}(K(x_i)^T\theta\mathcal{K}_{x_i,x_k} + K(x_i)\theta^T\mathcal{K}_{x_i,x_k} - y_i\mathcal{K}_{x_i,x_k}),
\end{aligned}$$

where $\mathcal{K}_{x_i,x_k}$ is a $n\times m$ matrix in the following form:

$$\mathcal{K}_{x_i,x_k} = \begin{pmatrix} \left(\frac{\partial K(x_i,x_1)}{\partial x_k}\right)^T \\ \vdots \\ \left(\frac{\partial K(x_i,x_n)}{\partial x_k}\right)^T \end{pmatrix}.$$

Meanwhile,

$$\nabla^2_{\theta\theta}\hat{\mathcal{L}}_n(\theta, X, Y) = \frac{2}{n}\sum_{i=1}^{n}K(x_i)K(x_i)^T + 2\lambda I.$$

Thus, we have

$$\begin{aligned}
\hat{\theta}_{\varepsilon,\min} - \hat{\theta}_{\min} + O(\|\hat{\theta}_{\varepsilon,\min} - \hat{\theta}_{\min}\|_2^2) =&\big(-\nabla^2_{\theta\theta}\hat{\mathcal{L}}_n(\hat{\theta}_{\min}, X, Y)\big)^{-1}\Big(\sum_{i=1}^{n}\nabla_{x_i,\theta}\hat{\mathcal{L}}_n(\hat{\theta}_{\min}, X, Y)\delta_i\\
&+\|\hat{\theta}_{\varepsilon,\min} - \hat{\theta}_{\min}\|_2\|\delta_i\|_2\Big).
\end{aligned}$$

Besides,

$$\nabla_{x_k}\hat{\mathcal{L}}_n(\theta, X, Y) = \frac{2}{n}\sum_{i=1}^{n}\Big(K(x_i)^T\theta - y_i\Big)\mathcal{K}_{x_i,x_k}^T\theta,$$

By the argument in Theorem 4.1, we know

$$\lim_{\varepsilon\to 0}\frac{\delta_i}{\varepsilon} = \beta_i$$

where

$$\beta_{i,k} = \frac{c_k^{q-1}}{(\sum_{k=1}^{m}c_k^q)^{\frac{1}{p}}}\mathrm{sgn}\Big(\nabla_{x_i}\hat{\mathcal{L}}_n(\hat{\theta}_{\min}, k)\Big),$$

with $c_k = |\nabla_{x_i}\hat{\mathcal{L}}_n(\hat{\theta}_{\min}, k)|$ and $\nabla_{x_i}\hat{\mathcal{L}}_n(\hat{\theta}_{\min}, k)$ is short for the k-th coordinate of $\nabla_{x_i}\hat{\mathcal{L}}_n(\theta, X, Y)$.

A.7 Proof of Theorem 6.1

In this subsection, we state more rigorously about our theorem. Firstly, we notice that if we let n and ε in Lemma 6.1 to be independent with each other, actually, the definition of $\hat{\mathcal{I}}^{DRO}(\mathcal{M})$ should be defined as

$$\frac{d\hat{\theta}^{\mathcal{M},DRO}_{\varepsilon,\min}}{d\varepsilon}\Big|_{\varepsilon=0}$$

and will obtain a limit that is dependent of sample size n and as n goes to infinity, $\hat{\mathcal{I}}^{DRO}(\mathcal{M})$ will goes to a trivial solution 0. In order to obtain a limit independent of sample size, we need n and ε to be dependent. Besides, if we do let n and ε to be dependent, the limit of $\hat{\mathcal{I}}^{DRO}(\mathcal{M})$ depends on the dependent relationship between ε and n. So unlike AIF, $\hat{\mathcal{I}}^{DRO}(\mathcal{M})$ is algorithm dependent.

We assume n is a function of ε and $n\varepsilon^u = 1$. Then, we still have

$$\hat{\theta}^{\mathcal{M},DRO}_{\varepsilon,\min} - \hat{\theta}^{\mathcal{M}}_{\min} \approx \big(-\frac{1}{n}\sum_{i=1}^{n}\nabla^2_\theta L(\hat{\theta}^{\mathcal{M}}_{\min}, x_i, y_i)\big)^{-1}\Big(\frac{1}{n}\sum_{i=1}^{n}\nabla_{x,\theta}L(\hat{\theta}^{\mathcal{M}}_{\min}, x_i, y_i)\delta_i\Big).$$

Notice, we can put all the mass in 6.1 on one of δ_i, so, we can put all on the δ_i with largest $\|\nabla_x L(\hat{\theta}^{\mathcal{M}}_{\min}, x_i, y_i)\|_q$ in order to achieve the maximum, which yields the final result.

References

[1] Naman Agarwal, Alon Gonen, and Elad Hazan. Learning in non-convex games with an optimization oracle. *arXiv preprint arXiv:1810.07362*, 2018.

[2] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *arXiv preprint arXiv:1802.00420*, 2018.

[3] Erhan Bayaktar and Lifeng Lai. On the adversarial robustness of robust estimators. *arXiv preprint arXiv:1806.03801*, 2018.

[4] Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust optimization*, volume 28. Princeton University Press, 2009.

[5] Hans-Georg Beyer and Bernhard Sendhoff. Robust optimization–a comprehensive survey. *Computer methods in applied mechanics and engineering*, 196(33-34):3190–3218, 2007.

[6] Andreas Christmann and Ingo Steinwart. On robustness properties of convex risk minimization methods for pattern recognition. *Journal of Machine Learning Research*, 5(Aug):1007–1034, 2004.

[7] Jeremy M Cohen, Elan Rosenfeld, and J Zico Kolter. Certified adversarial robustness via randomized smoothing. *arXiv preprint arXiv:1902.02918*, 2019.

[8] Christophe Croux and Gentiane Haesbroeck. Influence function and efficiency of the minimum covariance determinant scatter matrix estimator. *Journal of Multivariate Analysis*, 71(2):161–190, 1999.

[9] Michiel Debruyne, Mia Hubert, and Johan AK Suykens. Model selection in kernel based regression using the influence function. *Journal of Machine Learning Research*, 9(Oct):2377–2400, 2008.

[10] Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. *arXiv preprint arXiv:1810.02054*, 2018.

[11] Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Analysis of classifiers robustness to adversarial perturbations. *Machine Learning*, 107(3):481–508, 2018.

[12] Rui Gao and Anton J Kleywegt. Distributionally robust stochastic optimization with wasserstein distance. *arXiv preprint arXiv:1604.02199*, 2016.

[13] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[15] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pages 8571–8580, 2018.

[16] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1885–1894. JMLR. org, 2017.

[17] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial machine learning at scale. 2017.

[18] José Lezama, Qiang Qiu, Pablo Musé, and Guillermo Sapiro. Ole: Orthogonal low-rank embedding-a plug and play geometric loss for deep learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8109–8118, 2018.

[19] Xuanqing Liu and Cho-Jui Hsieh. Rob-gan: Generator, discriminator, and adversarial attacker. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11234–11243, 2019.

[20] Yong Liu, Shali Jiang, and Shizhong Liao. Efficient approximation of cross-validation for kernel methods using bouligand influence function. In *International Conference on Machine Learning*, pages 324–332, 2014.

[21] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[22] Michel Minoux. Robust network optimization under polyhedral demand uncertainty is np-hard. *Discrete Applied Mathematics*, 158(5):597–603, 2010.

[23] Gyung-Jin Park, Tae-Hee Lee, Kwon Hee Lee, and Kwang-Hyeon Hwang. Robust design: an overview. *AIAA journal*, 44(1):181–191, 2006.

[24] Amartya Sanyal, Varun Kanade, and Philip HS Torr. Learning low-rank representations. *arXiv preprint arXiv:1804.07090*, 2018.

[25] Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! In *Advances in Neural Information Processing Systems*, pages 3353–3364, 2019.

[26] Matthew Staib and Stefanie Jegelka. Distributionally robust deep learning as a generalization of adversarial training. In *NIPS workshop on Machine Learning and Computer Security*, 2017.

[27] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[28] Genichi Taguchi and Madhav S Phadke. Quality engineering through design optimization. In *Quality Control, Robust Design, and the Taguchi Method*, pages 77–96. Springer, 1989.

[29] Dong Yin, Kannan Ramchandran, and Peter Bartlett. Rademacher complexity for adversarially robust generalization. *arXiv preprint arXiv:1810.11914*, 2018.