Appendix

A. Proofs of the technical results in Section 2

A.1. Proof of Proposition 1

Proposition 1 *Given a fixed training set $\mathcal{S}$, let $\boldsymbol{\mu} = [\mu_q]_{q\in[Q]}$ be the Lagrangian multipliers for the constraints $\{\frac{1}{|V_q|}\sum_{j\in V_q}(y_j - h_{\boldsymbol{w}}(\boldsymbol{x}_j))^2 \le \delta + \xi_q\}_{q\in[Q]}$ in the optimization problem (2) and $F(\boldsymbol{w},\boldsymbol{\mu},\mathcal{S})$ be defined as follows:*

$$F(\boldsymbol{w},\boldsymbol{\mu},\mathcal{S}) = \sum_{i\in\mathcal{S}}[\lambda\left\|\boldsymbol{w}\right\|^2 + (y_i - h_{\boldsymbol{w}}(\boldsymbol{x}_i))^2] + \sum_{q\in[Q]}\mu_q\left[\frac{\sum_{j\in V_q}(y_j - h_{\boldsymbol{w}}(\boldsymbol{x}_j))^2}{|V_q|} - \delta\right] \tag{16}$$

Then, for the fixed set $\mathcal{S}$, the dual of the optimization problem (2) for estimating $\boldsymbol{w}$ and $\{\xi_q\}$ is given by,

$$\underset{\boldsymbol{0}\le\boldsymbol{\mu}\le C\boldsymbol{1}}{\text{maximize}}\ \underset{\boldsymbol{w}}{\text{minimize}}\ F(\boldsymbol{w},\boldsymbol{\mu},\mathcal{S}) \tag{17}$$

Proof The dual problem of our data selection problem (2) is given as:

$$\underset{\boldsymbol{\mu}\ge 0,\boldsymbol{\nu}}{\text{maximize}}\ \underset{\boldsymbol{w},\{\xi_q\}_{q\in[Q]}}{\text{minimize}} \sum_{i\in\mathcal{S}}[\lambda\left\|\boldsymbol{w}\right\|^2 + (y_i - h_{\boldsymbol{w}}(\boldsymbol{x}_i))^2] + C\sum_{q\in V_q}\xi_q + \sum_{q\in[Q]}\mu_q\left[\frac{\sum_{j\in V_q}(y_j - h_{\boldsymbol{w}}(\boldsymbol{x}_j))^2}{|V_q|} - \delta - \xi_q\right] - \nu_q\xi_q$$

Differentiating with respect to $\boldsymbol{\xi}$, we get $\boldsymbol{\mu}+\boldsymbol{\nu} = C\boldsymbol{1}$, which proves the Proposition (giving us the constraint $\boldsymbol{0} \le \boldsymbol{\mu} \le C\boldsymbol{1}$). ■

A.2. Proof of Proposition 3

Proposition 3 *Both the variants of the data selection problems (4) and (8) are NP-Hard.*

Proof Consider our data selection problem as follows:

$$\underset{\mathcal{S}\subset\mathcal{D},\boldsymbol{w},\{\xi_q\}_{q\in[Q]}}{\text{minimize}} \sum_{i\in\mathcal{S}}[\lambda\left\|\boldsymbol{w}\right\|^2 + (y_i - h_{\boldsymbol{w}}(\boldsymbol{x}_i))^2] + C\sum_{q\in V_q}\xi_q,$$

$$\text{such that, } \frac{\sum_{j\in V_q}(y_j - h_{\boldsymbol{w}}(\boldsymbol{x}_j))^2}{|V_q|} \le \delta + \xi_q \quad \forall q\in[Q],$$

$$\xi_q \ge 0 \quad \forall\, q\in[Q] \text{ and, } |\mathcal{S}| = k \tag{18}$$

We make $C = 0$ and $h_{\boldsymbol{w}}(\boldsymbol{x}) = \boldsymbol{w}^\top\boldsymbol{x}$. Then the problem becomes equivalent to the robust regression problem (Bhatia et al., 2017), i.e.,

$$\underset{\mathcal{S}\subset\mathcal{D},\boldsymbol{w}}{\text{minimize}} \sum_{i\in\mathcal{S}}[\lambda\left\|\boldsymbol{w}\right\|^2 + (y_i - \boldsymbol{w}^\top\boldsymbol{x})^2], \qquad \text{such that, } |\mathcal{S}| = k, \tag{19}$$

which is known to be NP-hard. ■

B. Techninal results on Section 3 and their proofs

B.1. Proof of Proposition 5

Proposition 5 *For any model $h_{\boldsymbol{w}}$, $f(\mathcal{S})$ is monotone,* i.e., $f(\mathcal{S}\cup a) - f(\mathcal{S}) \geq 0$ *for all $\mathcal{S}\subset\mathcal{D}$ and $a\in\mathcal{D}\backslash\mathcal{S}$.*

Proof We note that

$$
\begin{aligned}
f(\mathcal{S}\cup a) - f(\mathcal{S}) &= F\left(\boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{S}\cup a),\mathcal{S}\cup a),\boldsymbol{\mu}^*(\mathcal{S}\cup a),\mathcal{S}\cup a\right) - F\left(\boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{S}),\mathcal{S}),\boldsymbol{\mu}^*(\mathcal{S}),\mathcal{S}\right) \quad (20)\\
&= \underbrace{F\left(\boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{S}\cup a),\mathcal{S}\cup a),\boldsymbol{\mu}^*(\mathcal{S}\cup a),\mathcal{S}\cup a\right) - F\left(\boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{S}),\mathcal{S}\cup a),\boldsymbol{\mu}^*(\mathcal{S}),\mathcal{S}\cup a\right)}_{\geq 0}\\
&\quad + F\left(\boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{S}),\mathcal{S}\cup a),\boldsymbol{\mu}^*(\mathcal{S}),\mathcal{S}\cup a\right) - F\left(\boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{S}),\mathcal{S}),\boldsymbol{\mu}^*(\mathcal{S}),\mathcal{S}\right) \quad (21)\\
&\overset{(i)}{\geq} F\left(\boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{S}),\mathcal{S}\cup a),\boldsymbol{\mu}^*(\mathcal{S}),\mathcal{S}\cup a\right) - F\left(\boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{S}),\mathcal{S}),\boldsymbol{\mu}^*(\mathcal{S}),\mathcal{S}\right) \quad (22)\\
&= F\left(\boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{S}),\mathcal{S}\cup a),\boldsymbol{\mu}^*(\mathcal{S}),\mathcal{S}\cup a\right) - F\left(\boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{S}),\mathcal{S}\cup a),\boldsymbol{\mu}^*(\mathcal{S}),\mathcal{S}\right)\\
&\quad + \underbrace{F\left(\boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{S}),\mathcal{S}\cup a),\boldsymbol{\mu}^*(\mathcal{S}),\mathcal{S}\right) - F\left(\boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{S}),\mathcal{S}),\boldsymbol{\mu}^*(\mathcal{S}),\mathcal{S}\right)}_{\geq 0} \quad (23)\\
&\overset{(ii)}{\geq} F\left(\boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{S}),\mathcal{S}\cup a),\boldsymbol{\mu}^*(\mathcal{S}),\mathcal{S}\cup a\right) - F\left(\boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{S}),\mathcal{S}\cup a),\boldsymbol{\mu}^*(\mathcal{S}),\mathcal{S}\right)\\
&= \sum_{i\in\mathcal{S}\cup a}\left[\lambda\left\|\boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{S}),\mathcal{S}\cup a)\right\|^2 + (y_i - h_{\boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{S}),\mathcal{S}\cup a)}(\boldsymbol{x}_i))^2\right]\\
&\quad + \sum_{q\in[Q]}\mu_q^*(\mathcal{S})\left[\frac{\sum_{j\in V_q}(y_j - h_{\boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{S}),\mathcal{S}\cup a)}(\boldsymbol{x}_j))^2}{|V_q|} - \delta\right] \quad (24)\\
&\quad - \sum_{i\in\mathcal{S}}\left[\lambda\left\|\boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{S}),\mathcal{S}\cup a)\right\|^2 + (y_i - h_{\boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{S}),\mathcal{S}\cup a)}(\boldsymbol{x}_i))^2\right]\\
&\quad - \sum_{q\in[Q]}\mu_q^*(\mathcal{S})\left[\frac{\sum_{j\in V_q}(y_j - h_{\boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{S}),\mathcal{S}\cup a)}(\boldsymbol{x}_j))^2}{|V_q|} - \delta\right] \quad (25)\\
&= \lambda\left\|\boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{S}),\mathcal{S}\cup a)\right\|^2 + (y_a - h_{\boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{S}),\mathcal{S}\cup a)}(\boldsymbol{x}_a))^2 \quad (26)
\end{aligned}
$$

Here, inequality (i) is due to the fact that: $\boldsymbol{\mu}^*(\mathcal{S}\cup a) = \text{argmax}_{0\leq\boldsymbol{\mu}\leq C}\, F\left(\boldsymbol{w}^*(\boldsymbol{\mu},\mathcal{S}\cup a),\boldsymbol{\mu},\mathcal{S}\cup a\right)$; and, inequality (ii) is due to the fact that: $\boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{S}),\mathcal{S}) = \text{argmin}_{\boldsymbol{w}}\; F\left(\boldsymbol{w},\boldsymbol{\mu}^*(\mathcal{S}),\mathcal{S}\right)$. ■

B.2. Proof of Theorem 6

Theorem 6 *Assume that $|y|\leq y_{\max}$; $h_{\boldsymbol{w}}(\boldsymbol{x}) = 0$ if $\boldsymbol{w} = \mathbf{0}$,* i.e., *$h_{\boldsymbol{w}}(\boldsymbol{x})$ has no bias term; $h_{\boldsymbol{w}}$ is H-Lipschitz, i.e., $|h_{\boldsymbol{w}}(\boldsymbol{x})| \leq H\left\|\boldsymbol{w}\right\|$; the eigenvalues of the Hessian matrix of $(y - h_{\boldsymbol{w}}(\boldsymbol{x}))^2)$ have a finite upper bound,* i.e., $\text{Eigenvalue}(\nabla^2_{\boldsymbol{w}}(y - h_{\boldsymbol{w}}(\boldsymbol{x}))^2) \leq 2\chi^2_{\max}$; *and, define $\ell^* = \min_{a\in\mathcal{D}}\min_{\boldsymbol{w}}\;\chi^2_{\max}\cdot\left\|\boldsymbol{w}\right\|^2 + (y_a - h_{\boldsymbol{w}}(\boldsymbol{x}_a))^2 > 0$. Then, for $\lambda \geq \max\left\{\chi^2_{\max}, 32(1+CQ)^2 y^2_{\max}H^2/\ell^*\right\}$, $f(\mathcal{S})$ is a α-submodular set function, where*

$$
\alpha \geq \widehat{\alpha}_f = 1 - \frac{32(1+CQ)^2 y^2_{\max}H^2}{\lambda\ell^*}, \quad (27)
$$

Proof We assume that: $\mathcal{S}\subset\mathcal{T}$. Hence, $|\mathcal{T}|>0$. Let us define: $\ell_a(\boldsymbol{w}) = \lambda\left\|\boldsymbol{w}\right\|^2 + (y_a - h_{\boldsymbol{w}}(\boldsymbol{x}_a))^2$, $\overline{\boldsymbol{w}} = \text{argmin}_{\boldsymbol{w}}\ell_a(\boldsymbol{w})$. Finally, we denote $\ell^* = \min_{a\in\mathcal{D}}\min_{\boldsymbol{w}}\chi^2_{\max}\left\|\boldsymbol{w}\right\|^2 + (y_a - h_{\boldsymbol{w}}(\boldsymbol{x}_a))^2$. Next, we have that:

$$
\begin{aligned}
\frac{f(\mathcal{S}\cup a) - f(\mathcal{S})}{f(\mathcal{T}\cup a) - f(\mathcal{T})} &\geq \frac{\ell_a\left(\boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{S}),\mathcal{S}\cup a)\right)}{\ell_a\left(\boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{T}\cup a),\mathcal{T})\right)} \quad \text{(Due to Lemma 12)}\\
&\geq \frac{\ell_a\left(\overline{\boldsymbol{w}}\right)}{\ell_a\left(\boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{T}\cup a),\mathcal{T})\right)} \quad \text{(Since } \ell_a\left(\boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{S}),\mathcal{S}\cup a)\right) \geq \ell_a\left(\overline{\boldsymbol{w}}\right)\text{)}
\end{aligned}
$$

$$
\begin{aligned}
&\overset{(i)}{\geq} \frac{\ell_a\left(\boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{T}), \mathcal{T}\cup a)\right) - (\lambda + \chi^2_{\max}) \left\|\overline{\boldsymbol{w}} - \boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{T}\cup a), \mathcal{T})\right\|^2}{\ell_a\left(\boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{T}\cup a), \mathcal{T})\right)} \\
&\geq 1 - \frac{(\lambda + \chi^2_{\max}) \left\|\overline{\boldsymbol{w}} - \boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{T}\cup a), \mathcal{T})\right\|^2}{\ell_a\left(\boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{T}\cup a), \mathcal{T})\right)} \\
&\geq 1 - \frac{(\lambda + \chi^2_{\max})}{\ell_a(\overline{\boldsymbol{w}})} \left\|\overline{\boldsymbol{w}} - \boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{T}\cup a), \mathcal{T})\right\|^2 \quad (\text{Since } \ell_a\left(\boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{S}), \mathcal{S}\cup a)\right) \geq \ell_a\left(\overline{\boldsymbol{w}}\right)) \\
&\overset{(ii)}{\geq} 1 - \frac{32\lambda}{\ell^*} \frac{(1+CQ)^2 y^2_{\max} H^2}{\lambda^2} \\
&= 1 - \frac{32(1+CQ)^2 y^2_{\max} H^2}{\lambda \ell^*}, \qquad (28)
\end{aligned}
$$

Inequality (i) is due to the following:

$$
\begin{aligned}
&\ell_a(\boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{T}\cup a), \mathcal{T})) && (29) \\
&= \ell_a(\overline{\boldsymbol{w}}) + \nabla\ell_a(\overline{\boldsymbol{w}})^\top (\boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{T}\cup a), \mathcal{T}) - \overline{\boldsymbol{w}}) + (\boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{T}\cup a), \mathcal{T}) - \overline{\boldsymbol{w}})^\top \nabla^2\ell_a(\boldsymbol{w}')(\boldsymbol{w} - \overline{\boldsymbol{w}})^\top && (30) \\
&\leq \ell_a(\overline{\boldsymbol{w}}) + \frac{\max_{\{\text{eig}(\nabla^2\ell_a)\}}}{2} \left\|\boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{T}\cup a), \mathcal{T}) - \overline{\boldsymbol{w}}\right\|^2 \quad (\nabla\ell_a(\overline{\boldsymbol{w}}) = 0) && (31) \\
&\leq \ell_a(\overline{\boldsymbol{w}}) + (\lambda + \chi^2_{\max}) \left\|\boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{T}\cup a), \mathcal{T}) - \overline{\boldsymbol{w}}\right\|^2; && (32)
\end{aligned}
$$

and inequality (ii) follows from

$$
\begin{aligned}
&(1) \quad \ell_a(\overline{\boldsymbol{w}}) = \lambda \left\|\overline{\boldsymbol{w}}\right\|^2 + (y_a - h_{\overline{\boldsymbol{w}}}(\boldsymbol{x}_a))^2 \geq \chi^2_{\max} \left\|\overline{\boldsymbol{w}}\right\|^2 + (y_a - h_{\overline{\boldsymbol{w}}}(\boldsymbol{x}_a))^2 \geq \min_{\boldsymbol{w}} \chi^2_{\max} \left\|\boldsymbol{w}\right\|^2 + (y_a - h_{\boldsymbol{w}}(\boldsymbol{x}_a))^2 = \ell^*, \\
&(2) \quad \left\|\boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{T}\cup a), \mathcal{T}) - \overline{\boldsymbol{w}}\right\| \leq 2w_{\max} = \frac{4(1+CQ)y_{\max}H}{\lambda} \qquad (\text{Due to Lemma } 13), \\
&(3) \quad \lambda \geq \chi^2_{\max}. \qquad (33)
\end{aligned}
$$

■

B.3. Proof of Proposition 7

Proposition 7 *Given* $0 < y_{\min} \leq |y| \leq y_{\max}$, $h_{\boldsymbol{w}}(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x}$, $\|\boldsymbol{x}\| \leq x_{\max}$, *we set the regularizing coefficient as* $\lambda \geq \max\left\{x^2_{\max}, 16(1+CQ)^2 y^2_{\max} x^2_{\max}/y^2_{\min} \cdot\right\}$. *Then* $f(\mathcal{S})$ *is a* α*-submodular set function, where*

$$
\alpha \geq \widehat{\alpha}_f = 1 - \frac{16(1+CQ)^2 y^2_{\max} x^2_{\max}}{\lambda y^2_{\min}}. \qquad (34)
$$

Proof The proof exactly follows the previous proof, except in the highlighted part. We assume that: $\mathcal{S} \subset \mathcal{T}$. Hence, $|\mathcal{T}| > 0$ and define $\ell_a(\boldsymbol{w}) = \lambda \left\|\boldsymbol{w}\right\|^2 + (y_a - h_{\boldsymbol{w}}(\boldsymbol{x}_a))^2$, $\overline{\boldsymbol{w}} = \text{argmin}_{\boldsymbol{w}} \ell_a(\boldsymbol{w})$; $\ell^* = \min_{a\in\mathcal{D}} \min_{\boldsymbol{w}} \chi^2_{\max} \left\|\boldsymbol{w}\right\|^2 + (y_a - h_{\boldsymbol{w}}(\boldsymbol{x}_a))^2$. Then, we have that:

$$
\begin{aligned}
\frac{f(\mathcal{S}\cup a) - f(\mathcal{S})}{f(\mathcal{T}\cup a) - f(\mathcal{T})} &\geq \frac{\ell_a\left(\boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{S}), \mathcal{S}\cup a)\right)}{\ell_a\left(\boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{T}\cup a), \mathcal{T})\right)} \\
&\geq \frac{\ell_a\left(\overline{\boldsymbol{w}}\right)}{\ell_a\left(\boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{T}\cup a), \mathcal{T})\right)} \\
&\geq \frac{\ell_a\left(\boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{T}), \mathcal{T}\cup a)\right) - (\lambda + \chi^2_{\max}) \left\|\overline{\boldsymbol{w}} - \boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{T}\cup a), \mathcal{T})\right\|^2}{\ell_a\left(\boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{T}\cup a), \mathcal{T})\right)} \\
&\geq 1 - \frac{(\lambda + \chi^2_{\max}) \left\|\overline{\boldsymbol{w}} - \boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{T}\cup a), \mathcal{T})\right\|^2}{\ell_a\left(\boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{T}\cup a), \mathcal{T})\right)} \\
&\geq 1 - \frac{(\lambda + \chi^2_{\max})}{\ell_a(\overline{\boldsymbol{w}})} \left\|\overline{\boldsymbol{w}} - \boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{T}\cup a), \mathcal{T})\right\|^2 \\
&\geq 1 - \frac{8\lambda}{\ell^*} \frac{(1+CQ)^2 y^2_{\max} x^2_{\max}}{\lambda^2}
\end{aligned}
$$

$$
\begin{aligned}
&= 1 - \frac{8\,(1+CQ)^2 y_{\max}^2 x_{\max}^2}{\lambda \ell^*} \\
&\geq 1 - \frac{16\,(1+CQ)^2 y_{\max}^2 x_{\max}^2}{\lambda\, y_{\min}^2} \qquad (35)
\end{aligned}
$$

where the highlighted part is due to second part of Lemma 13 which gives:

$$
\|\boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{T}\cup a),\mathcal{T}) - \overline{\boldsymbol{w}}\| \leq 2w_{\max} = \frac{2(1+CQ)y_{\max}x_{\max}}{\lambda}; \qquad (36)
$$

and, Claim 1 which shows that $\ell^* = \frac{\lambda y_{\min}^2}{\lambda + x_{\max}^2} \geq y_{\min}^2/2$. ■

B.4. Proof of Proposition 8

Proposition 8 *Given the assumptions stated in Theorem 6. the generalized curvature $k_f(\mathcal{S})$ for any set $\mathcal{S}$ satisfies* $\kappa_f(\mathcal{S}) \leq \widehat{\kappa}_f = 1 - \dfrac{\ell^*}{(CQ+1)y_{\max}^2}$.

Proof Let us define: $\ell_a(\boldsymbol{w}) = \lambda\|\boldsymbol{w}\|^2 + (y_a - h_{\boldsymbol{w}}(\boldsymbol{x}_a))^2$, $\overline{\boldsymbol{w}} = \text{argmin}_{\boldsymbol{w}}\ell_a(\boldsymbol{w})$. Finally, we denote $\ell^* = \min_{a\in\mathcal{D}}\min_{\boldsymbol{w}}\chi_{\max}^2\|\boldsymbol{w}\|^2 + (y_a - h_{\boldsymbol{w}}(\boldsymbol{x}_a))^2$. By definition, we have $1-\kappa_f(\mathcal{S}) = \min_{a\in\mathcal{D}}\frac{f(a|\mathcal{S}\setminus a)}{f(a|\emptyset)}$. We show that, from Lemma 12, we have that:

$$
f(a|\mathcal{S}\setminus a) \geq \lambda\|\boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{S}\setminus a),\mathcal{S})\|^2 + (y_a - h_{\boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{S}\setminus a),\mathcal{S})}(\boldsymbol{x}_a))^2 \geq \ell_a(\overline{\boldsymbol{w}}) \geq \ell^* \qquad (37)
$$

Next, we note that:

$$
\begin{aligned}
f(a|\emptyset) &= f(a) - f(\emptyset) \qquad (38)\\
&= \lambda\|\boldsymbol{w}^*(\boldsymbol{\mu}^*(a),a)\|^2 + (y_a - h_{\boldsymbol{w}^*(\boldsymbol{\mu}^*(a),a)}(\boldsymbol{x}_a))^2 \\
&\quad + \sum_{q\in[Q]}\mu_q^*(a)\sum_{j\in V_q}\left[\frac{(y_j - h_{\boldsymbol{w}^*(\boldsymbol{\mu}^*(a),a)}(\boldsymbol{x}_j))^2}{|V_q|} - \delta\right] - \sum_{q\in[Q]}\mu_q^*(\emptyset)\sum_{j\in V_q}\left[\frac{(y_j - h_{\boldsymbol{w}^*(\boldsymbol{\mu}^*(\emptyset),\emptyset)}(\boldsymbol{x}_j))^2}{|V_q|} - \delta\right] \\
&\overset{(i)}{\leq} \lambda\|\boldsymbol{w}^*(\boldsymbol{\mu}^*(a),a)\|^2 + (y_a - h_{\boldsymbol{w}^*(\boldsymbol{\mu}^*(a),a)}(\boldsymbol{x}_a))^2 \\
&\quad + \sum_{q\in[Q]}\mu_q^*(a)\sum_{j\in V_q}\left[\frac{(y_j - h_{\boldsymbol{w}^*(\boldsymbol{\mu}^*(a),a)}(\boldsymbol{x}_j))^2}{|V_q|} - \delta\right] - \sum_{q\in[Q]}\mu_q^*(a)\sum_{j\in V_q}\left[\frac{(y_j - h_{\boldsymbol{w}^*(\boldsymbol{\mu}^*(\emptyset),\emptyset)}(\boldsymbol{x}_j))^2}{|V_q|} - \delta\right] \\
&= \lambda\|\boldsymbol{w}^*(\boldsymbol{\mu}^*(a),a)\|^2 + (y_a - h_{\boldsymbol{w}^*(\boldsymbol{\mu}^*(a),a)}(\boldsymbol{x}_a))^2 \\
&\quad + \sum_{q\in[Q]}\mu_q^*(a)\sum_{j\in V_q}\left[\frac{(y_j - h_{\boldsymbol{w}^*(\boldsymbol{\mu}^*(a),a)}(\boldsymbol{x}_j))^2}{|V_q|}\right] - \sum_{q\in[Q]}\mu_q^*(a)\sum_{j\in V_q}\left[\frac{(y_j - h_{\boldsymbol{w}^*(\boldsymbol{\mu}^*(\emptyset),\emptyset)}(\boldsymbol{x}_j))^2}{|V_q|}\right] \\
&\leq \lambda\|\boldsymbol{w}^*(\boldsymbol{\mu}^*(a),a)\|^2 + (y_a - h_{\boldsymbol{w}^*(\boldsymbol{\mu}^*(a),a)}(\boldsymbol{x}_a))^2 \\
&\quad + \sum_{q\in[Q]}\mu_q^*(a)\sum_{j\in V_q}\left[\frac{(y_j - h_{\boldsymbol{w}^*(\boldsymbol{\mu}^*(a),a)}(\boldsymbol{x}_j))^2}{|V_q|}\right] \qquad (39)\\
&\overset{(ii)}{\leq} (CQ+1)\,y_{\max}^2 \qquad (40)
\end{aligned}
$$

Here, (i) is because $\boldsymbol{\mu}^*(\emptyset) = \text{argmax}_{\boldsymbol{\mu}}\sum_{q\in[Q]}\mu_q\sum_{j\in V_q}\left[\dfrac{(y_j - h_{\boldsymbol{w}^*(\boldsymbol{\mu},\emptyset)}(\boldsymbol{x}_j))^2}{|V_q|} - \delta\right]$, (ii) is obtained by putting $\boldsymbol{w} = \mathbf{0}$ in Eq. (39) which is now at the minimum, i.e.,

$$
\boldsymbol{w}^*(\boldsymbol{\mu}^*(a),a) = \underset{\boldsymbol{w}}{\text{argmin}}\,\lambda\|\boldsymbol{w}\|^2 + (y_a - h_{\boldsymbol{w}}(\boldsymbol{x}_a))^2 + \sum_{q\in[Q]}\mu_q^*(a)\sum_{j\in V_q}\left[\frac{(y_j - h_{\boldsymbol{w}}(\boldsymbol{x}_j))^2}{|V_q|}\right] \qquad (41)
$$

Hence, Eqs. (37) and (40) show that, $\kappa_f(\mathcal{S}) \leq 1 - \dfrac{\ell^*}{(CQ+1)\,y_{\max}^2}$. ■

B.5. Auxiliary Lemmas

Lemma 12 *If $f(\cdot)$ defined in Eq. (7), we have that*

$$f(\mathcal{S} \cup a) - f(\mathcal{S}) \geq \lambda \left\| \boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{S}), \mathcal{S} \cup a) \right\|^2 + (y_a - h_{\boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{S}), \mathcal{S} \cup a)}(\boldsymbol{x}_a))^2. \tag{42}$$

and,

$$f(\mathcal{S} \cup a) - f(\mathcal{S}) \leq \lambda \left\| \boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{S} \cup a), \mathcal{S}) \right\|^2 + (y_a - h_{\boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{S} \cup a), \mathcal{S})}(\boldsymbol{x}_a))^2. \tag{43}$$

Proof The proof of the lower bound of the marginal gain

$$f(\mathcal{S} \cup a) - f(\mathcal{S}) \geq \lambda \left\| \boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{S}), \mathcal{S} \cup a) \right\|^2 + (y_a - h_{\boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{S}), \mathcal{S} \cup a)}(\boldsymbol{x}_a))^2. \tag{44}$$

follows from the proof of Proposition 5.

Next we prove that

$$f(\mathcal{S} \cup a) - f(\mathcal{S}) \leq \lambda \left\| \boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{S} \cup a), \mathcal{S}) \right\|^2 + (y_a - h_{\boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{S} \cup a), \mathcal{S})}(\boldsymbol{x}_a))^2. \tag{45}$$

To show this, we prove that:

$$\begin{aligned}
& f(\mathcal{S} \cup a) - f(\mathcal{S}) \\
& = F\left(\boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{S} \cup a), \mathcal{S} \cup a), \boldsymbol{\mu}^*(\mathcal{S} \cup a), \mathcal{S} \cup a\right) - F\left(\boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{S}), \mathcal{S}), \boldsymbol{\mu}^*(\mathcal{S}), \mathcal{S}\right) && (46) \\
& = F\left(\boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{S} \cup a), \mathcal{S} \cup a), \boldsymbol{\mu}^*(\mathcal{S} \cup a), \mathcal{S} \cup a\right) - F\left(\boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{S} \cup a), \mathcal{S}), \boldsymbol{\mu}^*(\mathcal{S} \cup a), \mathcal{S}\right) \\
& \quad + \underbrace{F\left(\boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{S} \cup a), \mathcal{S}), \boldsymbol{\mu}^*(\mathcal{S} \cup a), \mathcal{S}\right) - F\left(\boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{S}), \mathcal{S}), \boldsymbol{\mu}^*(\mathcal{S}), \mathcal{S}\right)}_{\leq 0} && (47) \\
& \overset{(i)}{\leq} F\left(\boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{S} \cup a), \mathcal{S} \cup a), \boldsymbol{\mu}^*(\mathcal{S} \cup a), \mathcal{S} \cup a\right) - F\left(\boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{S} \cup a), \mathcal{S}), \boldsymbol{\mu}^*(\mathcal{S} \cup a), \mathcal{S}\right) && (48) \\
& = \underbrace{F\left(\boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{S} \cup a), \mathcal{S} \cup a), \boldsymbol{\mu}^*(\mathcal{S} \cup a), \mathcal{S} \cup a\right) - F\left(\boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{S} \cup a), \mathcal{S}), \boldsymbol{\mu}^*(\mathcal{S} \cup a), \mathcal{S} \cup a\right)}_{\leq 0} \\
& \quad + F\left(\boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{S} \cup a), \mathcal{S}), \boldsymbol{\mu}^*(\mathcal{S} \cup a), \mathcal{S} \cup a\right) - F\left(\boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{S} \cup a), \mathcal{S}), \boldsymbol{\mu}^*(\mathcal{S} \cup a), \mathcal{S}\right) \\
& \overset{(ii)}{\leq} F\left(\boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{S} \cup a), \mathcal{S}), \boldsymbol{\mu}^*(\mathcal{S} \cup a), \mathcal{S} \cup a\right) - F\left(\boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{S} \cup a), \mathcal{S}), \boldsymbol{\mu}^*(\mathcal{S} \cup a), \mathcal{S}\right) \\
& = \sum_{i \in \mathcal{S} \cup a} \left[\lambda \left\| \boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{S} \cup a), \mathcal{S}) \right\|^2 + (y_i - h_{\boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{S} \cup a), \mathcal{S})}(\boldsymbol{x}_i))^2\right] \\
& \quad + \sum_{q \in [Q]} \mu_q^*(\mathcal{S} \cup a) \left[\frac{\sum_{j \in V_q} (y_j - h_{\boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{S} \cup a), \mathcal{S})}(\boldsymbol{x}_j))^2}{|V_q|} - \delta\right] && (49) \\
& \quad - \sum_{i \in \mathcal{S}} \left[\lambda \left\| \boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{S} \cup a), \mathcal{S}) \right\|^2 + (y_i - h_{\boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{S} \cup a), \mathcal{S})}(\boldsymbol{x}_i))^2\right] \\
& \qquad - \sum_{q \in [Q]} \mu_q^*(\mathcal{S} \cup a) \left[\frac{\sum_{j \in V_q} (y_j - h_{\boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{S} \cup a), \mathcal{S})}(\boldsymbol{x}_j))^2}{|V_q|} - \delta\right] && (50) \\
& = \lambda \left\| \boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{S} \cup a), \mathcal{S}) \right\|^2 + (y_a - h_{\boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{S} \cup a), \mathcal{S})}(\boldsymbol{x}_a))^2. && (51)
\end{aligned}$$

Here (i) is due to the fact that,

$$\boldsymbol{\mu}^*(\mathcal{S}) = \operatorname*{argmax}_{\boldsymbol{\mu}} F(\boldsymbol{w}^*(\boldsymbol{\mu}, \mathcal{S}), \boldsymbol{\mu}, \mathcal{S}) \tag{52}$$

and (ii) is due to the fact that:

$$\boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{S} \cup a), \mathcal{S} \cup a) = \operatorname*{argmin}_{\boldsymbol{w}} F(\boldsymbol{w}, \boldsymbol{\mu}^*(\mathcal{S} \cup a), \mathcal{S} \cup a) \tag{53}$$

■

Lemma 13 *Given that $\mathcal{S} \neq \emptyset$; $h_{\boldsymbol{w}}(\boldsymbol{x}) = 0$ for $\boldsymbol{w} = \mathbf{0}$; and, $h_{\boldsymbol{w}}(\boldsymbol{x})$ is H-Lipschitz, i.e., $|h_{\boldsymbol{w}}(\boldsymbol{x})| \leq H\,\|w\|$. Then, we have $\|\boldsymbol{w}^*(\boldsymbol{\mu}, \mathcal{S})\| \leq \dfrac{2(1+CQ)y_{\max}H}{\lambda}$. Moreover, if $h_{\boldsymbol{w}}(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x}$, we have that $\|\boldsymbol{w}^*(\boldsymbol{\mu}, \mathcal{S})\| \leq \dfrac{(1+CQ)y_{\max}x_{\max}}{\lambda}$. Note that, for the linear model, we are able to exploit the structure of the model much better and therefore the bound is tighter.*

Proof First we define $\nabla h_0(\boldsymbol{x}) = \nabla_{\boldsymbol{w}} h_{\boldsymbol{w}}(\boldsymbol{x})|_{\boldsymbol{w}=\mathbf{0}}$.

$$\begin{aligned}
&F(\boldsymbol{w}^*(\boldsymbol{\mu},\mathcal{S}),\boldsymbol{\mu},\mathcal{S}) \\
&= \lambda\,\|\boldsymbol{w}^*(\boldsymbol{\mu},\mathcal{S})\|^2\,|\mathcal{S}| + \sum_{i\in\mathcal{S}}(y_i - h_{\boldsymbol{w}^*(\boldsymbol{\mu},\mathcal{S})}(\boldsymbol{x}_i))^2 + \sum_{q\in[Q]}\frac{\mu_q}{|V_q|}\sum_{j\in V_q}\left[(y_j - h_{\boldsymbol{w}^*(\boldsymbol{\mu},\mathcal{S})}(\boldsymbol{x}_j))^2 - \delta\right] \\
&= \lambda\,\|\boldsymbol{w}^*(\boldsymbol{\mu},\mathcal{S})\|^2\,|\mathcal{S}| + \sum_{i\in\mathcal{S}} y_i^2 + \sum_{q\in[Q]}\frac{\mu_q}{V_q}\sum_{j\in V_q} y_j^2 - \sum_{q\in[Q]}\frac{\mu_q}{|V_q|}\sum_{j\in V_q}\delta \\
&\quad - 2\sum_{i\in\mathcal{S}} y_i h_{\boldsymbol{w}^*(\boldsymbol{\mu},\mathcal{S})}(\boldsymbol{x}_i) - 2\sum_{q\in[Q]}\frac{\mu_q}{|V_q|}\sum_{j\in V_q} y_j h_{\boldsymbol{w}^*(\boldsymbol{\mu},\mathcal{S})}(\boldsymbol{x}_j) + \underbrace{\sum_{i\in\mathcal{S}} h^2_{\boldsymbol{w}^*(\boldsymbol{\mu},\mathcal{S})}(\boldsymbol{x}_i) + \sum_{q\in[Q]}\frac{\mu_q}{|V_q|}\sum_{j\in V_q} h^2_{\boldsymbol{w}^*(\boldsymbol{\mu},\mathcal{S})}(\boldsymbol{x}_j)}_{\geq 0} \\
&\overset{(i)}{\geq} \lambda\,\|\boldsymbol{w}^*(\boldsymbol{\mu},\mathcal{S})\|^2\,|\mathcal{S}| + \overbrace{\sum_{i\in\mathcal{S}} y_i^2 + \sum_{q\in[Q]}\frac{\mu_q}{V_q}\sum_{j\in V_q} y_j^2 - \sum_{q\in[Q]}\frac{\mu_q}{|V_q|}\sum_{j\in V_q}\delta}^{F(\mathbf{0},\mu,\mathcal{S})} \\
&\quad - 2\sum_{i\in\mathcal{S}} y_i h_{\boldsymbol{w}^*(\boldsymbol{\mu},\mathcal{S})}(\boldsymbol{x}_i) - 2\sum_{q\in[Q]}\frac{\mu_q}{|V_q|}\sum_{j\in V_q} y_j h_{\boldsymbol{w}^*(\boldsymbol{\mu},\mathcal{S})}(\boldsymbol{x}_j).
\end{aligned} \tag{54}$$

Here (i) is due to $\sum_{i\in\mathcal{S}} h^2_{\boldsymbol{w}^*(\boldsymbol{\mu},\mathcal{S})}(\boldsymbol{x}_i) + \sum_{q\in[Q]}\frac{\mu_q}{|V_q|}\sum_{j\in V_q} h^2_{\boldsymbol{w}^*(\boldsymbol{\mu},\mathcal{S})}(\boldsymbol{x}_j) \geq 0$. Now since $F(\mathbf{0},\mu,\mathcal{S}) = \sum_{i\in\mathcal{S}} y_i^2 + \sum_{q\in[Q]}\frac{\mu_q}{V_q}\sum_{j\in V_q} y_j^2 - \sum_{q\in[Q]}\frac{\mu_q}{|V_q|}\sum_{j\in V_q}\delta$, Eq. (54) gives us:

$$\begin{aligned}
F(\boldsymbol{w}^*(\boldsymbol{\mu},\mathcal{S}),\boldsymbol{\mu},\mathcal{S}) - F(\mathbf{0},\mu,\mathcal{S}) &\geq \lambda\,\|\boldsymbol{w}^*(\boldsymbol{\mu},\mathcal{S})\|^2\,|\mathcal{S}| \\
&\quad - 2\sum_{i\in\mathcal{S}} y_i h_{\boldsymbol{w}^*(\boldsymbol{\mu},\mathcal{S})}(\boldsymbol{x}_i) - 2\sum_{q\in[Q]}\frac{\mu_q}{|V_q|}\sum_{j\in V_q} y_j h_{\boldsymbol{w}^*(\boldsymbol{\mu},\mathcal{S})}(\boldsymbol{x}_j)
\end{aligned} \tag{55}$$

Now since $\boldsymbol{w}^*(\boldsymbol{\mu},\mathcal{S}) = \operatorname{argmin}_{\boldsymbol{w}} F(\boldsymbol{w},\boldsymbol{\mu},\mathcal{S})$, we have that $F(\boldsymbol{w}^*(\boldsymbol{\mu},\mathcal{S}),\boldsymbol{\mu},\mathcal{S}) \leq F(\mathbf{0},\mu,\mathcal{S})$. Then, Eq. (55) implies that

$$\begin{aligned}
&\lambda\,\|\boldsymbol{w}^*(\boldsymbol{\mu},\mathcal{S})\|^2\,|\mathcal{S}| - 2\sum_{i\in\mathcal{S}} y_i h_{\boldsymbol{w}^*(\boldsymbol{\mu},\mathcal{S})}(\boldsymbol{x}_i) - 2\sum_{q\in[Q]}\frac{\mu_q}{|V_q|}\sum_{j\in V_q} y_j h_{\boldsymbol{w}^*(\boldsymbol{\mu},\mathcal{S})}(\boldsymbol{x}_j) \leq 0 \\
&\overset{(i)}{\Longrightarrow} \lambda\,\|\boldsymbol{w}^*(\boldsymbol{\mu},\mathcal{S})\|^2\,|\mathcal{S}| \leq 2|\mathcal{S}|\,y_{\max}H\,\|\boldsymbol{w}^*(\boldsymbol{\mu},\mathcal{S})\| + 2\sum_{q\in[Q]}\frac{\mu_q}{|V_q|}\sum_{j\in V_q} y_{\max}H\,\|\boldsymbol{w}^*(\boldsymbol{\mu},\mathcal{S})\| \\
&\qquad\qquad \leq 2(|\mathcal{S}| + CQ)y_{\max}H\,\|\boldsymbol{w}^*(\boldsymbol{\mu},\mathcal{S})\| \\
&\Longrightarrow \|\boldsymbol{w}^*(\boldsymbol{\mu},\mathcal{S})\| \leq \frac{2(|\mathcal{S}| + CQ)y_{\max}H}{\lambda|\mathcal{S}|} \leq \frac{2(1+CQ)y_{\max}H}{\lambda}
\end{aligned} \tag{56}$$

Here (i) is due to H-Lipschitzness of $\boldsymbol{h}_{\boldsymbol{w}^*(\boldsymbol{\mu},\mathcal{S})}(\boldsymbol{x})$. For linear model, we have $H = x_{\max}$. However, we use the structure of the model to obtain a better bound. More specifically, for linear model, we have:

$$\begin{aligned}
\boldsymbol{w}^*(\boldsymbol{\mu},\mathcal{S}) &= \left(\lambda|\mathcal{S}|\mathbb{I} + \sum_{i\in\mathcal{S}}\boldsymbol{x}_i\boldsymbol{x}_i^\top + \sum_{q\in[Q]}\frac{\mu_q}{|V_q|}\sum_{j\in V_q}\boldsymbol{x}_j\boldsymbol{x}_j^\top\right)^{-1}\left(\sum_{i\in\mathcal{S}} y_i\boldsymbol{x}_i + \sum_{q\in[Q]}\frac{\mu_q}{|V_q|}\sum_{j\in V_q} y_j\boldsymbol{x}_j\right) \\
\Longrightarrow \|\boldsymbol{w}^*(\boldsymbol{\mu},\mathcal{S})\| &\leq \frac{(|\mathcal{S}| + CQ)x_{\max}y_{\max}}{\mathrm{Eig}_{\min}\left(\lambda|\mathcal{S}|\mathbb{I} + \sum_{i\in\mathcal{S}}\boldsymbol{x}_i\boldsymbol{x}_i^\top + \sum_{q\in[Q]}\frac{\mu_q}{|V_q|}\sum_{j\in V_q}\boldsymbol{x}_j\boldsymbol{x}_j^\top\right)} \\
&\leq \frac{(|\mathcal{S}| + CQ)x_{\max}y_{\max}}{\lambda|\mathcal{S}|} \\
&\leq \frac{(1+CQ)x_{\max}y_{\max}}{\lambda}.
\end{aligned} \tag{57}$$

■

Claim 1 $\min_{\boldsymbol{w}}[\lambda \left\|\boldsymbol{w}\right\|^2 + (y_a - \boldsymbol{w}^\top \boldsymbol{x}_a)^2] = \frac{\lambda y_a^2}{\lambda + \|\boldsymbol{x}_a\|^2}$

Proof We note that:

$$\overline{\boldsymbol{w}} = y_a(\lambda + \boldsymbol{x}_a \boldsymbol{x}_a^\top)^{-1}\boldsymbol{x}_a \tag{58}$$

Hence, we have that:

$$\begin{aligned}
\lambda \left\|\overline{\boldsymbol{w}}\right\|^2 + (y_a - \overline{\boldsymbol{w}}^\top \boldsymbol{x}_a)^2 &= y_a^2 - 2y_a \overline{\boldsymbol{w}}^\top \boldsymbol{x}_a + \overline{\boldsymbol{w}}^\top (\lambda \mathbb{I} + \boldsymbol{x}_a \boldsymbol{x}_a^\top)\overline{\boldsymbol{w}} && (59)\\
&= y_a^2 - y_a \overline{\boldsymbol{w}}^\top \boldsymbol{x}_a && (60)\\
&= y_a^2 - y_a^2 \boldsymbol{x}_a^\top (\lambda + \boldsymbol{x}_a \boldsymbol{x}_a^\top)^{-1} \boldsymbol{x}_a && (61)\\
&= y_a^2 - y_a^2 \boldsymbol{x}_a^\top \left[\frac{1}{\lambda} - \frac{\boldsymbol{x}_a \boldsymbol{x}_a^\top / \lambda^2}{1 + \boldsymbol{x}_a^\top \boldsymbol{x}_a / \lambda}\right] \boldsymbol{x}_a \ \ \text{(Due to Sherman Morrison formula)} && (62)\\
&= \frac{\lambda y_a^2}{\lambda + \left\|\boldsymbol{x}_a\right\|^2}\\
&\geq \frac{\lambda y_{\min}^2}{\lambda + x_{\max}^2} && (63)
\end{aligned}$$

■

C. Proofs of the technical results in Section 4

C.1. Proof of Lemma 9

Lemma 9 *Given a fixed set $\widehat{\mathcal{S}}$ and an α-submodular function $f(\mathcal{S})$, let the modular function $m^f_{\widehat{\mathcal{S}}}[\mathcal{S}]$ be defined as follows:*

$$\begin{aligned} m^f_{\widehat{\mathcal{S}}}[\mathcal{S}] =& f(\widehat{\mathcal{S}}) - \sum_{i\in\widehat{\mathcal{S}}} \alpha f(i|\widehat{\mathcal{S}}\backslash\{i\}) \\ &+ \sum_{i\in\widehat{\mathcal{S}}\cap\mathcal{S}} \alpha f(i|\widehat{\mathcal{S}}\backslash\{i\}) + \sum_{i\in\mathcal{S}\backslash\widehat{\mathcal{S}}} \frac{f(i|\emptyset)}{\alpha}. \end{aligned} \tag{64}$$

Then, $f(\mathcal{S}) \leq m^f_{\widehat{\mathcal{S}}}[\mathcal{S}]$ for all $\mathcal{S} \subseteq \mathcal{D}$.

Proof Recall that f is α-submodular with coefficient $\widehat{\alpha}_f$ if $f(a|\mathcal{S}) \geq \widehat{\alpha}_f f(a|\mathcal{T}), a \notin \mathcal{T}, \mathcal{S} \subseteq \mathcal{T}$. Given this, the following inequalities follow directly from:

$$\widehat{\alpha}_f[f(\mathcal{S}) - f(\widehat{\mathcal{S}}\cap\mathcal{S})] \leq \sum_{i\in\mathcal{S}\backslash\widehat{\mathcal{S}}} f(i|\emptyset) \tag{65}$$

and similarly,

$$[f(\widehat{\mathcal{S}}) - f(\widehat{\mathcal{S}}\cap\mathcal{S})] \geq \widehat{\alpha}_f \sum_{i\in\widehat{\mathcal{S}}\backslash\mathcal{S}} f(i|\widehat{\mathcal{S}}\backslash i) \tag{66}$$

The inequalities above hold by considering a chain of sets from $\widehat{\mathcal{S}}\cap\mathcal{S}$ to either $\widehat{\mathcal{S}}$ or $\mathcal{S}$ and applying the weak-submodularity definition by considering sets $\mathcal{S}$ and $\mathcal{T}$ appropriately. We then multiply -1 to inequality (66), multiply $1/\widehat{\alpha}_f$ to equation (65) and add both of them together. We then achieve:

$$f(\mathcal{S}) \leq f(\widehat{\mathcal{S}}) - \widehat{\alpha}_f \sum_{i\in\widehat{\mathcal{S}}\backslash\mathcal{S}} f(i|\widehat{\mathcal{S}}\backslash i) + \frac{1}{\widehat{\alpha}_f} \sum_{i\in\mathcal{S}\backslash\widehat{\mathcal{S}}} f(i|\emptyset) \tag{67}$$

Rearranging this, we get the expression for the Lemma. ■

C.2. Proof of Theorem 10

Theorem 10 *If the training algorithm in Algorithm 1 (lines 3, 6, 8) provides perfect estimates of the model parameters, it obtains a set $\widehat{\mathcal{S}}$ which satisfies:*

$$f(\widehat{\mathcal{S}}) \leq \frac{k}{\widehat{\alpha}_f(1+(k-1)(1-\widehat{\kappa}_f)\widehat{\alpha}_f)} f(\mathcal{S}^*) \tag{68}$$

where $\widehat{\alpha}_f$ and $\widehat{\kappa}_f$ are as stated in Theorem 6 and Proposition 8 respectively.

Proof From the definition of α-submodularity, note that $\widehat{\alpha}_f f(\mathcal{S}) \leq \sum_{i\in\mathcal{S}} f(i)$. Next, we can obtain the following inequality for any $k \in \mathcal{S}$ using weak submodularity:

$$f(\mathcal{S}) - f(k) \geq \widehat{\alpha}_f \sum_{j\in\mathcal{S}\backslash k} (f(j|\mathcal{S}\backslash j) \tag{69}$$

We can add this up for all $k \in \mathcal{S}$ and obtain:

$$\begin{aligned} |\mathcal{S}|f(\mathcal{S}) - \sum_{k\in\mathcal{S}} f(k) &\geq \widehat{\alpha}_f \sum_{k\in\mathcal{S}} \sum_{j\in\mathcal{S}\backslash k} (f(j|\mathcal{S}\backslash j) \\ &\geq \widehat{\alpha}_f(|\mathcal{S}|-1) \sum_{k\in\mathcal{S}} f(k|\mathcal{S}\backslash k) \end{aligned} \tag{70}$$

Finally, from the definition of curvature, note that $f(k|\mathcal{S}\backslash k) \leq (1-\widehat{\kappa}_f)f(k)$. Combining all this together, we obtain:

$$|\mathcal{S}|f(\mathcal{S}) \geq (1+\widehat{\alpha}_f(1-\widehat{\kappa}_f)(|\mathcal{S}|-1)) \sum_{j\in\mathcal{S}} f(j) \tag{71}$$

which implies:

$$\sum_{j\in\mathcal{S}} f(j) \leq \frac{|\mathcal{S}|}{1+\widehat{\alpha}_f(1-\widehat{\kappa}_f)(|\mathcal{S}|-1)} f(\mathcal{S}) \tag{72}$$

Combining this with the fact that $\widehat{\alpha}_f f(\mathcal{S}) \leq \sum_{i\in\mathcal{S}} f(i)$, we obtain that:

$$f(\mathcal{S}) \leq \frac{1}{\widehat{\alpha}_f}\sum_{i\in\mathcal{S}} f(i) \leq \frac{|\mathcal{S}|}{\widehat{\alpha}_f(1+\widehat{\alpha}_f(1-\widehat{\kappa}_f)(|\mathcal{S}|-1))} f(\mathcal{S}) \tag{73}$$

The approximation guarantee then follows from some simple observations. In particular, given an approximation

$$m^f(\mathcal{S}) = \frac{1}{\widehat{\alpha}_f}\sum_{i\in\mathcal{S}} f(i) \tag{74}$$

which satisfies $f(\mathcal{S}) \leq m^f(\mathcal{S}) \leq \beta_f f(\mathcal{S})$, we claim that optimizing m^f essentially gives a β_f approximation factor. To prove this, let $\mathcal{S}^*$ be the optimal subset, and $\hat{\mathcal{S}}$ be the subset obtained after optimizing m^f. The following chain of inequalities holds:

$$f(\hat{\mathcal{S}}) \leq m^f(\hat{\mathcal{S}}) \leq m^f(\mathcal{S}^*) \leq \beta_f f(\mathcal{S}^*) \tag{75}$$

This shows that $\hat{\mathcal{S}}$ is a β_f approximation of $\mathcal{S}^*$. Finally, note that this is just the first iteration of SELCON, and with subsequent iterations, SELCON is guaranteed to reduce the objective value (see Appendix C.4). ■

C.3. Proof of Theorem 11

Theorem 11 *If the training algorithm (lines 3, 6, 8) in Algorithm 1 provides imperfect estimates, so that* $\|F(\widehat{\boldsymbol{w}},\widehat{\boldsymbol{\mu}},\mathcal{S}) - F(\boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{S}),\mathcal{S}),\boldsymbol{\mu}^*(\mathcal{S}),\mathcal{S})\| \leq \epsilon$ *for any* $\mathcal{S}$*, then Algorithm 1 obtains a set* $\widehat{\mathcal{S}}$ *that satisfies:*

$$f(\widehat{\mathcal{S}}) \leq \left(\frac{k}{\widehat{\alpha}_f(1+(k-1)(1-\widehat{\kappa}_f)\widehat{\alpha}_f)} + \frac{2k\epsilon}{\ell}\right) f(S^*),$$

where $\ell = \min_{a\in\mathcal{D}}\min_{\boldsymbol{w}} \lambda||\boldsymbol{w}||^2 + (y_i - h_{\boldsymbol{w}}(\boldsymbol{x}_i))^2$, $\widehat{\alpha}_f$ *and* $\widehat{\kappa}_f$ *are obtained in Theorem 6 and Proposition 8, respectively.*

Proof Define:

$$\beta_f = \frac{k}{(1+(k-1)(1-\widehat{\kappa}_f)\widehat{\alpha}_f)} \tag{76}$$

and also define, $\hat{f}(\mathcal{S}) = F(\widehat{\boldsymbol{w}},\widehat{\boldsymbol{\mu}},\mathcal{S})$ and $f(\mathcal{S}) = F(\boldsymbol{w}^*(\boldsymbol{\mu}^*(\mathcal{S}),\mathcal{S}),\boldsymbol{\mu}^*(\mathcal{S}),\mathcal{S})$. Note that instead of having access to f, the algorithm has access to $\hat{f}$ which satisfies:

$$|f(\mathcal{S}) - \hat{f}(\mathcal{S})| \leq \epsilon, \forall\mathcal{S} \tag{77}$$

Let us assume that $\hat{f}$ is always smaller compared to f, i.e. in other words,

$$f(\mathcal{S}) \leq \hat{f}(\mathcal{S}) \leq f(\mathcal{S}) + \epsilon \tag{78}$$

Combining this with the fact that:

$$f(\mathcal{S}) \leq \frac{1}{\widehat{\alpha}_f}\sum_{j\in\mathcal{S}} f(j) \leq \frac{\beta_f}{\widehat{\alpha}_f} f(\mathcal{S}) \tag{79}$$

we obtain the following chain of inequalities:

$$f(\mathcal{S}) \leq \frac{1}{\widehat{\alpha}_f}\sum_{j\in\mathcal{S}} f(j) \leq \frac{1}{\widehat{\alpha}_f}\sum_{j\in\mathcal{S}}[\hat{f}(j)] \leq \frac{1}{\widehat{\alpha}_f}\sum_{j\in\mathcal{S}}[f(j)+\epsilon] \leq \frac{\beta_f}{\widehat{\alpha}_f} f(\mathcal{S}) + \frac{k\epsilon}{\widehat{\alpha}_f} \tag{80}$$

where $|\mathcal{S}| = k$. Finally, we get the approximation factor by dividing by a lower bound of $l = \min_{\mathcal{S}:|\mathcal{S}|=k} f(\mathcal{S})$ which can be obtained via a very similar proof technique to the weak submodularity and curvature results. Hence we get the final approximation factor as $\frac{\beta_f}{\widehat{\alpha}_f} + \frac{k\epsilon}{l\widehat{\alpha}_f}$.

We end by pointing out that we can get a similar result even if we do not assume that $\hat{f}$ is always smaller compared to f and in fact, assume the more general condition:

$$f(\mathcal{S}) - \epsilon \leq \hat{f}(\mathcal{S}) \leq f(\mathcal{S}) + \epsilon \tag{81}$$

The only difference is we have an additional factor of 2 in the additive bound. In particular, we get the following chain of inequalities:

$$f(\mathcal{S}) \leq \frac{1}{\widehat{\alpha}_f} \sum_{j\in\mathcal{S}} f(j) \leq \frac{1}{\widehat{\alpha}_f} \sum_{j\in\mathcal{S}} [\hat{f}(j) + \epsilon] \leq \frac{1}{\widehat{\alpha}_f} \sum_{j\in\mathcal{S}} [f(j) + 2\epsilon] \leq \frac{\beta_f}{\widehat{\alpha}_f} f(\mathcal{S}) + \frac{2k\epsilon}{\widehat{\alpha}_f} \tag{82}$$

The chain of inequalities holds because $f(j) \leq \hat{f}(j) + \epsilon$ and $\hat{f}(j) \leq f(j) + \epsilon$. ■

C.4. Convergence property

We begin this section by showing that SELCON is guaranteed to reduce the objective value at every iteration as long as we obtain perfect solutions from the training algorithm (lines 3, 6, 8 in Algorithm 1).

Lemma 14 SELCON *(Algorithm 1) is guaranteed to reduce the objective value of f at every iteration as long as we obtain perfect solutions from the training sub-routine.*

Proof SELCON essentially uses modular upper bounds m^f of f at every iteration. Denote $\mathcal{S}_l$ as the set obtained in the lth iteration and let $\mathcal{S}_{l+1}$ be the one from the $l+1$th iteration. Then the following chain of inequalities hold:

$$f(\mathcal{S}_{l+1}) \leq m^f(\mathcal{S}_{l+1}) \leq m^f(\mathcal{S}_l) = f(\mathcal{S}_l) \tag{83}$$

The first inequality holds because m^f is a modular upper bound, the second inequality holds because $\mathcal{S}_{l+1}$ is the solution of minimizing m^f (and hence $m^f(\mathcal{S}_{l=1})$ is lower in value compared to $m^f(\mathcal{S}_l)$). The last equality holds because m^f is a modular upper bound which is tight at $\mathcal{S}_l$ and hence $m^f(\mathcal{S}_l) = f(\mathcal{S}_l)$. This shows that $f(\mathcal{S}_{l+1}) \leq f(\mathcal{S}_l)$. ■

We end this section by pointing out that this chain of inequalities does not hold if we get inexact or approximation solutions to the training sub-routine. In practice, we observe that the objective value of f still reduces even though we obtain only inexact solutions since the inexact solutions are often close to the true solutions of the training step.

D. Additional details about experimental setup

D.1. Dataset details

- **Cadata:** California housing dataset is obtained from the LIBSVM package [6]. This spatial dataset contains 20,640 observations on housing prices with 9 economic covariates. As described in (Pace & Barry, 1997), here $\boldsymbol{x}$ are information about households in a block, say median age, median income, total rooms/population, bedrooms/population, population/households,households etc. and y is median price in median housing prices by all California census blocks. It has $\text{dimension}(\boldsymbol{x}) = 8$.
- **Law:** This refers to the dataset on Law School Admissions Council's National Longitudinal Bar Passage Study (Wightman, 1998). Here $\boldsymbol{x}$ is information about a law student, including information on gender, race, family income, age, *etc.* and y indicates GPA normalised to $[0, 1]$. We use race as a protected attribute for the fairness experiments. It has $\text{dimension}(\boldsymbol{x}) = 10$.
- **NYSE-High:** This dataset is obtained from the New York stock exchange (NYSE) [7] dataset as follows. Given the set $\{s_i\}$ with s_i corresponding to the the highest stock price of the i^{th} day, we define $s_{k+1} = \sum_{i\in[100]} w_i s_{k+1-i}$. Here $y_k = s_{k+1}$ and $\boldsymbol{x}_k = [s_k, s_{k-1}, ..., s_{k-99}]$. This dataset has $\text{dimension}(\boldsymbol{x}) = 100$.
- **NYSE-close:** This dataset is obtained from the New York stock exchange (NYSE) [8] dataset as follows. Given the set $\{s_i\}$ with s_i corresponding to the the closing stock price of the i^{th} day, we define $s_{k+1} = \sum_{i\in[100]} w_i s_{k+1-i}$. Here $y_k = s_{k+1}$ and $\boldsymbol{x}_k = [s_k, s_{k-1}, ..., s_{k-99}]$. This dataset has $\text{dimension}(\boldsymbol{x}) = 100$.

D.2. Implementation details

Our models. We use two models— a simple linear regression model and a two layer neural network that consists of a linear layer of 5 hidden nodes and a ReLU activation unit. In all our experiments, we use a learning rate of 0.01. We choose the value of δ as the 30% of the mean validation error obtained using Full-selection.

Implementation of CRAIG. CRAIG (Mirzasoleiman et al., 2020) requires computing a $\mathcal{D} \times \mathcal{D}$ matrix with similarity measure for each pair of points in the training set. For the larger datasets, i.e., NYSE-close and NYSE-high, such a computation requires a large amount of memory. Hence, we use a stochastic version where we randomly select R points and build $R \times R$ matrix and select $\frac{kR}{\mathcal{D}}$ each time and repeat the process $\frac{\mathcal{D}}{R}$ times. We use $R = 50000$. Note that, for other datasets, since $|\mathcal{D}| < 50000$ the stochastic version is same as the original version.

CRAIG requires us to select the subset only once, since features will not change even as the training proceeds. However, since CRAIG is an adaptive method, for the non-linear setting, we need to run CRAIG every epoch. Despite using the stochastic version, we found CRAIG to be very slow in the non-linear setting and therefore we don't report it.

Implementation of GLISTER. GLISTER (Killamsetty et al., 2021b), an another adaptive subset selection method where we select a new subset every 35th epoch to help make a fair comparison against SELCON. We update the model parameters after every selection step.

Machine configuration. We performed our experiments on a computer system with Ubuntu 16.04.6 LTS, an i-7 with 6 cores CPU and a total RAM of 125 GBs. The system had a single GeForce GTX 1080 GPU which was employed in our experiments.

E. Additional experiments

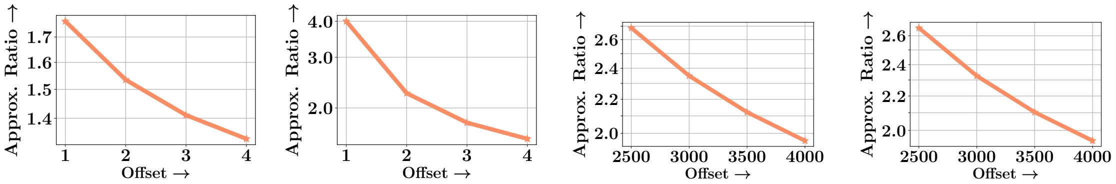
E.1. Discussion on adding offsets to the response variable y

The approximation ratio of SELCON is $f(\widehat{\mathcal{S}})/f(\mathcal{S}^*) \leq \dfrac{k}{\widehat{\alpha}_f(1+(k-1)(1-\widehat{\kappa}_f)\widehat{\alpha}_f)}$ when the training method is accurate. A trite calculation shows that this quantity is $O(y_{\text{max}}^4/y_{\text{min}}^4)$. If $y_{\text{max}}/y_{\text{min}}$ is very high, the approximation ratio is affected. Such a problem can be easily overcome by adding an offset to y and then augmenting the feature $\boldsymbol{x}$ with an additional term 1— which incorporates the effect of the added offset. We summarize the effect of this offset on the approximation ratio (for different datasets) in Figure 5 which shows that adding an offset improves the approximation factor. Note that in the case of

[6] https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/regression.html

[7] https://www.kaggle.com/dgawlik/nyse

[8] https://www.kaggle.com/dgawlik/nyse



(a) Cadata (b) Law (c) NYSE-High (d) NYSE-Close

Figure 5. Variation in the approximation ratio with respect to the offset added to the response variables y.

Cadata, y indicates the house price; whereas in the case of Law, y indicates student GPA. Therefore, the approximation factor of these datasets is reasonable even without adding an offset. Whereas, for NYSE-High and NYSE-Clone, the approximation factor is somewhat poorer at lower values of the offset (not shown in the plot).

E.2. Significance Tests

	Random-selection	Random-with-constraints	CRAIG	GLISTER	SelCon-without-constraints	SelCon
Random-selection						
Random-with-constraints	0.000089					
CRAIG	0.00012	0.50159				
GLISTER	0.040043	0.00014	0.00078			
SelCon-without-constraints	0.001713	0.601212	0.88129	0.00803		
SelCon	0.0001	0.00014	0.00059	0.0001	0.0001	

Table 6. Pairwise significance p-values using Wilcoxon signed rank test.

In Table 6, we show the p-values of two-tailed Wilcoxon signed-rank test (Wilcoxon, 1992) performed on every possible pair of data selection strategies to determine whether there is a significant statistical difference between the strategies in each pair, across all datasets. Our null hypothesis is that there is no difference between each pair of data selection strategies. From the results, it is evident that SelCon significantly outperforms other baselines at $p < 0.01$.