Supplementary material for "Submodular Observation Selection and Information Gathering for Quadratic Models"

Abolfazl Hashemi [1] **Mahsa Ghasemi** [1] **Haris Vikalo** [1] **Ufuk Topcu** [1]

Proof of Proposition 1

First note that we can define c_f equivalently as $c_f = \max_{l=1}^{n-1} \mathcal{C}_l$ where

$$\mathcal{C}_l = \max_{(\mathcal{S},\mathcal{T},i)\in\mathcal{X}_l} f_i(\mathcal{T})/f_i(\mathcal{S}), \tag{1}$$

and $\mathcal{X}_l = \{(\mathcal{S},\mathcal{T},i) | \mathcal{S} \subseteq \mathcal{T} \subset \mathcal{X}, i \in \mathcal{X}\backslash\mathcal{T}, |\mathcal{T}\backslash\mathcal{S}| = l\}$. Now, let $\mathcal{S} \subset \mathcal{T}$ and $\mathcal{T}\backslash\mathcal{S} = \{j_1,\dots,j_r\}$. Then,

$$\begin{aligned} f(\mathcal{T}) - f(\mathcal{S}) &= f(\mathcal{S} \cup \{j_1,\dots,j_r\}) - f(\mathcal{S}) \\ &= f_{j_1}(\mathcal{S}) + f_{j_2}(\mathcal{S}\cup\{j_1\}) + \dots \\ &\quad + f_{j_r}(\mathcal{S}\cup\{j_1,\dots,j_{r-1}\}). \end{aligned} \tag{2}$$

Applying (1) yields

$$\begin{aligned} f(\mathcal{T}) - f(\mathcal{S}) &\le f_{j_1}(\mathcal{S}) + \mathcal{C}_1 f_{j_2}(\mathcal{S}) + \dots + \mathcal{C}_{r-1} f_{j_r}(\mathcal{S}) \\ &= f_{j_1}(\mathcal{S}) + \sum_{l=1}^{r-1} \mathcal{C}_l f_{j_t}(\mathcal{S}). \end{aligned} \tag{3}$$

Note that (3) is invariant to the ordering of elements in $\mathcal{T}\backslash\mathcal{S}$. In fact, it is straightforward to see that given ordering $\{j_1,\dots,j_r\}$, one can choose a set $\mathcal{Q} = \{\mathcal{P}_1,\dots,\mathcal{P}_r\}$ with r permutations – e.g., by defining the right circular-shift operator $\mathcal{P}_t(\{j_1,\dots,j_r\}) = \{j_{r-t+1},\dots,j_1,\dots\}$ for $1 \le t \le r$ – such that $\mathcal{P}_p(j) \ne \mathcal{P}_q(j)$ for $p \ne q$ and $\forall j \in \mathcal{T}\backslash\mathcal{S}$. Hence, (3) holds for r such permutations. Summing all of these r inequalities we obtain

$$\begin{aligned} f(\mathcal{T}) - f(\mathcal{S}) &\le \frac{1}{r}\left(1 + \sum_{l=1}^{r-1}\mathcal{C}_l\right)\sum_{j\in\mathcal{T}\backslash\mathcal{S}} f_j(\mathcal{S}) \\ &\le \frac{1}{r}\left(1 + (r-1)c_f\right)\sum_{j\in\mathcal{T}\backslash\mathcal{S}} f_j(\mathcal{S}). \end{aligned} \tag{4}$$

Next, we prove the second inequality. Note that we can define $\epsilon_f = \max_{l=1}^{n-1}\epsilon_l$ where $\epsilon_l = \max_{(\mathcal{S},\mathcal{T},i)\in\mathcal{X}_l} f_i(\mathcal{T}) - f_i(\mathcal{S})$. Using a similar argument as the one that we used for c_f, for any $\mathcal{S} \subset \mathcal{T}$ and $\mathcal{T}\backslash\mathcal{S} = \{j_1,\dots,j_r\}$, it holds that

$$\begin{aligned} f(\mathcal{T}) - f(\mathcal{S}) &\le \sum_{l=1}^{r-1}\epsilon_l + \sum_{j\in\mathcal{T}\backslash\mathcal{S}} f_j(\mathcal{S}) \\ &\le (r-1)\epsilon_f + \sum_{j\in\mathcal{T}\backslash\mathcal{S}} f_j(\mathcal{S}), \end{aligned} \tag{5}$$

which completes the proof.

Proof of Proposition 2

The proof follows the classical proof of greedy maximization of submodular functions given in (Nemhauser et al., 1978). We first prove the performance bound stated in terms of c_f. Consider $\mathcal{S}_i$, the set generated at the end of the i^{th} iteration of the greedy algorithm and assume $|\mathcal{S}^\star\backslash S_t^{(i)}| = r \le k$. Employing Proposition 1 with $\mathcal{S} = \mathcal{S}_i$ and $\mathcal{T} = \mathcal{S}^\star \cup \mathcal{S}_i$, and using monotonicity of f yields

$$\begin{aligned} \frac{f(\mathcal{S}^\star) - f(\mathcal{S}_i)}{\frac{1}{r}\left(1+(r-1)c_f\right)} &\le \frac{f(\mathcal{S}^\star\cup\mathcal{S}_i) - f(\mathcal{S}_i)}{\frac{1}{r}\left(1+(r-1)c_f\right)} \\ &\le \sum_{j\in\mathcal{S}^\star\backslash\mathcal{S}_i} f_j(\mathcal{S}_i) \\ &\le r(f(\mathcal{S}_{i+1}) - f(\mathcal{S}_i)), \end{aligned} \tag{6}$$

where we use the fact that the greedy algorithm selects the element with the maximum marginal gain in each iteration. It is easy to verify, e.g., by taking the derivative, that $\frac{1}{r}\left(1+(r-1)c_f\right)$ is decreasing (increasing) with respect to r if $c_f < 1$ ($c_f > 1$). Let $c = \max\{c_f, 1\}$. Then $\frac{1}{r}(1+(r-1)\mathcal{C}_{\max}) \le c$. Therefore, using the fact that $r \le k$ we get

$$f(\mathcal{S}^\star) - f(\mathcal{S}_i) \le ck(f(\mathcal{S}_{i+1}) - f(\mathcal{S}_i)). \tag{7}$$

By induction and due to the fact that $f(\emptyset) = 0$ we obtain

$$f(\mathcal{S}_g) \ge \left(1 - \left(1 - \frac{1}{kc}\right)^k\right) f(\mathcal{S}^\star) \ge \left(1 - e^{-\frac{1}{c}}\right) f(\mathcal{S}^\star), \tag{8}$$

---

[1]University of Texas at Austin. Correspondence to: Abolfazl Hashemi <abolfazl@utexas.edu>, Mahsa Ghasemi <mahsa.ghasemi@utexas.edu>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

where we use the fact that $(1+x)^y \leq e^{xy}$ for $y > 0$. The proof of second inequality is almost identical except we employ the second result of Proposition 1 to begin the proof.

Centering θ in Quadratic Models

In (19), defining $\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta} - \mathbb{E}[\boldsymbol{\theta}]$ yields

$$\begin{aligned} y_i &= \frac{1}{2}(\tilde{\boldsymbol{\theta}} + \mathbb{E}[\boldsymbol{\theta}])^\top \mathbf{X}_i (\tilde{\boldsymbol{\theta}} + \mathbb{E}[\boldsymbol{\theta}]) + \mathbf{z}_i^\top (\tilde{\boldsymbol{\theta}} + \mathbb{E}[\boldsymbol{\theta}]) + \nu_i \\ &= \frac{1}{2}\tilde{\boldsymbol{\theta}}^\top \mathbf{X}_i \tilde{\boldsymbol{\theta}} + \frac{1}{2}(\mathbf{X}_i\,\mathbb{E}[\boldsymbol{\theta}] + \mathbf{X}_i^\top\,\mathbb{E}[\boldsymbol{\theta}] + 2\mathbf{z}_i)^\top \tilde{\boldsymbol{\theta}} \\ &\quad + \frac{1}{2}\,\mathbb{E}[\boldsymbol{\theta}]^\top \mathbf{X}_i\,\mathbb{E}[\boldsymbol{\theta}] + \nu_i. \end{aligned} \tag{9}$$

Thus, we obtain a new quadratic model $\tilde{y}_i = \frac{1}{2}\tilde{\boldsymbol{\theta}}^\top \mathbf{X}_i \tilde{\boldsymbol{\theta}} + \tilde{\mathbf{z}_i}^\top \tilde{\boldsymbol{\theta}} + \nu_i$ with zero-mean unknown parameters $\tilde{\boldsymbol{\theta}}$, where $\tilde{y}_i = y_i - \frac{1}{2}\,\mathbb{E}[\boldsymbol{\theta}]^\top \mathbf{X}_i\,\mathbb{E}[\boldsymbol{\theta}]$, and $\tilde{\mathbf{z}_i} = \frac{1}{2}(\mathbf{X}_i\,\mathbb{E}[\boldsymbol{\theta}] + \mathbf{X}_i^\top\,\mathbb{E}[\boldsymbol{\theta}] + 2\mathbf{z}_i)$.

Proof of Theorem 2

Let $q_{\boldsymbol{\theta}}(\boldsymbol{\Theta}) = p_{\boldsymbol{\theta},\mathbf{y}_\mathcal{S}}(\boldsymbol{\Theta};\mathbf{y})$ denote the posterior distribution of $\boldsymbol{\theta}$ given $\mathbf{y}_\mathcal{S}$, $\boldsymbol{\Gamma}_\mathcal{S} = \mathrm{diag}(\{\sigma_i^2\}_{i\in\mathcal{S}})$ denote the noise covariance matrix $\mathrm{Cov}(\boldsymbol{\nu}_\mathcal{S})$, and define

$$\boldsymbol{\mu}_\mathcal{S} = \mathrm{vec}(\{\frac{1}{2}\boldsymbol{\theta}^\top \mathbf{X}_i \boldsymbol{\theta} + \mathbf{z}_i^\top \boldsymbol{\theta}\}_{i\in\mathcal{S}}).$$

Then, the Van Trees' bound is found as

$$\begin{aligned} \mathbf{B}_\mathcal{S}^{-1} &= \mathbb{E}_{\mathbf{y}_\mathcal{S},\boldsymbol{\theta}}[(\nabla_{\boldsymbol{\Theta}} \log q_{\boldsymbol{\theta}}(\boldsymbol{\Theta}))(\nabla_{\boldsymbol{\Theta}} \log q_{\boldsymbol{\theta}}(\boldsymbol{\Theta}))^\top] \\ &= \mathbb{E}_{\mathbf{y}_\mathcal{S},\boldsymbol{\theta}}[(\nabla_{\boldsymbol{\Theta}} \log p_{\mathbf{y}_\mathcal{S}|\boldsymbol{\theta}}(\mathbf{y};\boldsymbol{\Theta}) p_{\boldsymbol{\theta}}(\boldsymbol{\Theta})) \\ &\qquad (\nabla_{\boldsymbol{\Theta}} \log p_{\mathbf{y}_\mathcal{S}|\boldsymbol{\theta}}(\mathbf{y};\boldsymbol{\Theta}) p_{\boldsymbol{\theta}}(\boldsymbol{\Theta}))^\top] \\ &= \mathbb{E}_{\mathbf{y}_\mathcal{S},\boldsymbol{\theta}}[(\nabla_{\boldsymbol{\Theta}} \log p_{\mathbf{y}_\mathcal{S}|\boldsymbol{\theta}}(\mathbf{y};\boldsymbol{\Theta})) \\ &\qquad (\nabla_{\boldsymbol{\Theta}} \log p_{\mathbf{y}_\mathcal{S}|\boldsymbol{\theta}}(\mathbf{y};\boldsymbol{\Theta}))^\top] + \mathbf{I}_x, \end{aligned} \tag{10}$$

where

$$\mathbf{I}_x = \mathbb{E}_{\mathbf{y}_\mathcal{S},\boldsymbol{\theta}}\left[(\nabla_{\boldsymbol{\Theta}} \log p_{\boldsymbol{\theta}}(\boldsymbol{\Theta}))(\nabla_{\boldsymbol{\Theta}} \log p_{\boldsymbol{\theta}}(\boldsymbol{\Theta}))^\top\right]$$

is the prior Fisher information on $\boldsymbol{\theta}$ (e.g., if $p_{\boldsymbol{\theta}}(\boldsymbol{\Theta}) = \mathcal{N}(\mathbf{0},\mathbf{P})$ then $\mathbf{I}_x = \mathbf{P}^{-1}$). Note that the conditional distribution $p_{\mathbf{y}_\mathcal{S}|\boldsymbol{\theta}}(\mathbf{y};\boldsymbol{\Theta})$ is normal $\mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\theta}},\boldsymbol{\Gamma})$. Therefore,

$$\nabla_{\boldsymbol{\Theta}} \log p_{\mathbf{y}_\mathcal{S}|\boldsymbol{\theta}}(\mathbf{y};\boldsymbol{\Theta}) = -(\nabla_{\boldsymbol{\Theta}}\boldsymbol{\mu}_\mathcal{S})\boldsymbol{\Gamma}_\mathcal{S}^{-1}(\mathbf{y}_\mathcal{S} - \boldsymbol{\mu}_\mathcal{S}), \tag{11}$$

where $[\nabla_{\boldsymbol{\Theta}}\boldsymbol{\mu}_\mathcal{S}]_i = \mathbf{X}_i\boldsymbol{\theta} + \mathbf{z}_i$. Using this result we obtain

$$\begin{aligned} \mathbf{B}_\mathcal{S}^{-1} &= \mathbb{E}_{\mathbf{y}_\mathcal{S},\boldsymbol{\theta}}[(\nabla_{\boldsymbol{\Theta}}\boldsymbol{\mu}_\mathcal{S})\boldsymbol{\Gamma}_\mathcal{S}^{-1}(\mathbf{y}_\mathcal{S} - \boldsymbol{\mu}_\mathcal{S}) \\ &\qquad (\mathbf{y}_\mathcal{S} - \boldsymbol{\mu}_\mathcal{S})^\top \boldsymbol{\Gamma}_\mathcal{S}^{-1}(\nabla_{\boldsymbol{\Theta}}\boldsymbol{\mu}_\mathcal{S})^\top] + \mathbf{I}_x \\ &\overset{(a)}{=} \mathbb{E}_{\mathbf{y}_\mathcal{S},\boldsymbol{\theta}}[\mathbb{E}_{\mathbf{y}_\mathcal{S}|\boldsymbol{\theta}}[(\nabla_{\boldsymbol{\Theta}}\boldsymbol{\mu}_\mathcal{S})\boldsymbol{\Gamma}_\mathcal{S}^{-1}(\mathbf{y}_\mathcal{S} - \boldsymbol{\mu}_\mathcal{S}) \\ &\qquad (\mathbf{y}_\mathcal{S} - \boldsymbol{\mu}_\mathcal{S})^\top \boldsymbol{\Gamma}_\mathcal{S}^{-1}(\nabla_{\boldsymbol{\Theta}}\boldsymbol{\mu}_\mathcal{S})^\top]] + \mathbf{I}_x \\ &= \mathbb{E}_{\mathbf{y}_\mathcal{S},\boldsymbol{\theta}}[(\nabla_{\boldsymbol{\Theta}}\boldsymbol{\mu}_\mathcal{S})\boldsymbol{\Gamma}_\mathcal{S}^{-1}\,\mathbb{E}_{\mathbf{y}_\mathcal{S}|\boldsymbol{\theta}}\left[(\mathbf{y}_\mathcal{S} - \boldsymbol{\mu}_\mathcal{S})(\mathbf{y}_\mathcal{S} - \boldsymbol{\mu}_\mathcal{S})^\top\right] \\ &\qquad \boldsymbol{\Gamma}_\mathcal{S}^{-1}(\nabla_{\boldsymbol{\Theta}}\boldsymbol{\mu}_\mathcal{S})^\top] + \mathbf{I}_x \\ &\overset{(b)}{=} \mathbb{E}_{\mathbf{y}_\mathcal{S},\boldsymbol{\theta}}\left[(\nabla_{\boldsymbol{\Theta}}\boldsymbol{\mu}_\mathcal{S})\boldsymbol{\Gamma}_\mathcal{S}^{-1}\boldsymbol{\Gamma}_\mathcal{S}\boldsymbol{\Gamma}_\mathcal{S}^{-1}(\nabla_{\boldsymbol{\Theta}}\boldsymbol{\mu}_\mathcal{S})^\top\right] + \mathbf{I}_x \\ &= \mathbb{E}_{\mathbf{y}_\mathcal{S},\boldsymbol{\theta}}\left[(\nabla_{\boldsymbol{\Theta}}\boldsymbol{\mu}_\mathcal{S})\boldsymbol{\Gamma}_\mathcal{S}^{-1}(\nabla_{\boldsymbol{\Theta}}\boldsymbol{\mu}_\mathcal{S})^\top\right] + \mathbf{I}_x \\ &= \mathbb{E}_{\mathbf{y}_\mathcal{S},\boldsymbol{\theta}}\left[\sum_{i\in\mathcal{S}} \frac{1}{\sigma_i^2}(\mathbf{X}_i\boldsymbol{\theta} + \mathbf{z}_i)(\mathbf{X}_i\boldsymbol{\theta} + \mathbf{z}_i)^\top\right] + \mathbf{I}_x \\ &= \sum_{i\in\mathcal{S}} \frac{1}{\sigma_i^2}(\mathbb{E}_{\mathbf{y}_\mathcal{S},\boldsymbol{\theta}}\left[\mathbf{X}_i\boldsymbol{\theta}\boldsymbol{\theta}^\top\mathbf{X}_i^\top\right] + \mathbb{E}_{\mathbf{y}_\mathcal{S},\boldsymbol{\theta}}\left[2\mathbf{X}_i\boldsymbol{\theta}\mathbf{z}_i^\top\right] \\ &\qquad + \mathbb{E}_{\mathbf{y}_\mathcal{S},\boldsymbol{\theta}}\left[\mathbf{z}_i\mathbf{z}_i^\top\right]) + \mathbf{I}_x \\ &= \sum_{i\in\mathcal{S}} \frac{1}{\sigma_i^2}\left(\mathbf{X}_i\mathbf{P}\mathbf{X}_i^\top + \mathbf{z}_i\mathbf{z}_i^\top\right) + \mathbf{I}_x \end{aligned} \tag{12}$$

where to obtain (a) we use the law of total expectation, (b) follows by the definition of covariance matrices (see (2) in the paper), and the last equality follows since we assumed $\mathbb{E}[\boldsymbol{\theta}] = \mathbf{0}$ and $\mathrm{Cov}(\boldsymbol{\theta}) = \mathbb{E}[\boldsymbol{\theta}\boldsymbol{\theta}^\top] = \mathbf{P}$. Inverting the last line that consists of an invertible positive definite matrix establishes the stated results which in turn completes the proof.

Proof of Theorem 3

The marginal gain of adding a new observation to a subset $\mathcal{S}$ is

$$\begin{aligned} f_j^T(\mathcal{S}) &= \mathrm{Tr}\left(\mathbf{B}_{\mathcal{S}\cup\{j\}}^{-1}\right) - \mathrm{Tr}\left(\mathbf{I}_x\right) - \mathrm{Tr}\left(\mathbf{B}_\mathcal{S}^{-1}\right) + \mathrm{Tr}\left(\mathbf{I}_x\right) \\ &= \mathrm{Tr}\left(\mathbf{B}_{\mathcal{S}\cup\{j\}}^{-1} - \mathbf{B}_\mathcal{S}^{-1}\right) \\ &= \mathrm{Tr}\left(\frac{1}{\sigma_j^2}\left(\mathbf{X}_j\mathbf{P}\mathbf{X}_j^\top + \mathbf{z}_j\mathbf{z}_j^\top\right)\right). \end{aligned} \tag{13}$$

Therefore, the marginal gain is trace of a positive semi-definite matrix and hence $f_j^T(\mathcal{S}) \geq 0$ and the function is monotone. Furthermore, since the marginal gain does not depend on set $\mathcal{S}$ it is a modular function.

Proof of Theorem 4

Let $\mathbf{I}_j = \frac{1}{\sigma_j^2}\left(\mathbf{X}_j\mathbf{P}\mathbf{X}_j^\top + \mathbf{z}_j\mathbf{z}_j^\top\right)$. The marginal gain of adding a new observation to a subset $\mathcal{S}$ is

$$\begin{aligned} f_j^D(\mathcal{S}) &= \log\det\left(\mathbf{B}_{\mathcal{S}\cup\{j\}}^{-1}\right) - \log\det\left(\mathbf{I}_x\right) - \log\det\left(\mathbf{B}_{\mathcal{S}}^{-1}\right) \\ &\qquad + \log\det\left(\mathbf{I}_x\right) \\ &= \log\det\left(\mathbf{B}_{\mathcal{S}}^{-1} + \mathbf{I}_j\right) - \log\det\left(\mathbf{B}_{\mathcal{S}}^{-1}\right) \\ &\stackrel{(a)}{=} \log\frac{\det\mathbf{B}_{\mathcal{S}}^{-1}\det\left(\mathbf{I} + \mathbf{B}_{\mathcal{S}}^{1/2}\mathbf{I}_j\mathbf{B}_{\mathcal{S}}^{1/2}\right)}{\det\mathbf{B}_{\mathcal{S}}^{-1}} \\ &= \log\det\left(\mathbf{I} + \mathbf{B}_{\mathcal{S}}^{1/2}\mathbf{I}_j\mathbf{B}_{\mathcal{S}}^{1/2}\right) \\ &\stackrel{(b)}{\geq} 0, \end{aligned} \tag{14}$$

where (a) follows from the fact that $\det(\mathbf{A}+\mathbf{B}) = \det(\mathbf{A})\det\left(1 + \mathbf{A}^{-1/2}\mathbf{B}\mathbf{A}^{-1/2}\right)$, according to Sylvester's determinant identity, for any positive definite matrix $\mathbf{A}$ and Hermitian matrix $\mathbf{B}$ (Bellman, 1997), and (b) holds due to $\det(\mathbf{I}+\mathbf{A}) \geq (1 + \det\mathbf{A})$ for any positive semi-definite matrix $\mathbf{A}$. Therefore f^D is monotonically increasing.

Now consider $\mathcal{S} \subseteq \mathcal{T} \subset \mathcal{X}$ and $j \in \mathcal{X}\backslash\mathcal{T}$. Using the Sylvester's determinant identity we obtain

$$f_j^D(\mathcal{T})/f_j^D(\mathcal{S}) = \frac{\log\det\left(\mathbf{I} + \mathbf{B}_{\mathcal{T}}^{1/2}\mathbf{I}_j\mathbf{B}_{\mathcal{T}}^{1/2}\right)}{\log\det\left(\mathbf{I} + \mathbf{B}_{\mathcal{S}}^{1/2}\mathbf{I}_j\mathbf{B}_{\mathcal{S}}^{1/2}\right)} \leq 1. \tag{15}$$

Hence, $c_{f^D} = \max_{(\mathcal{S},\mathcal{T},j)\in\tilde{\mathcal{X}}} f_j^D(\mathcal{T})/f_j^D(\mathcal{S}) \leq 1$ which in turn proves submodularity of D-optimality.

Proof of Theorem 5

The marginal gain of adding a new observation to the subset $\mathcal{S}$ is

$$\begin{aligned} f_j^E(\mathcal{S}) &= \lambda_{\min}\left(\mathbf{B}_{\mathcal{S}\cup\{j\}}^{-1}\right) - \lambda_{\min}\left(\mathbf{I}_x\right) - \lambda_{\min}\left(\mathbf{B}_{\mathcal{S}}^{-1}\right) \\ &\qquad + \lambda_{\min}\left(\mathbf{I}_x\right) \\ &= \lambda_{\min}\left(\mathbf{B}_{\mathcal{S}}^{-1} + \mathbf{I}_j\right) - \lambda_{\min}\left(\mathbf{B}_{\mathcal{S}}^{-1}\right) \\ &\stackrel{(a)}{\geq} \lambda_{\min}\left(\mathbf{I}_j\right) \end{aligned} \tag{16}$$

where (a) follows from $\lambda_{\min}(\mathbf{A}+\mathbf{B}) \geq \lambda_{\min}(\mathbf{A}) + \lambda_{\min}(\mathbf{B})$ according to Weyl's inequality for two Hermitian matrices (Bellman, 1997). The positive semi-definiteness of $\mathbf{I}_j$ implies $f_j^E(\mathcal{S}) \geq 0$ and hence, monotonicity of f^E is established.

We now provide bounds on additive and multiplicative weak-submodularity constants of $f^E(\mathcal{S})$ (Note that it can be shown using simple examples that f^E is not in general weak submodular). Let $\mathcal{S} \subseteq \mathcal{T} \subset \mathcal{X}$ and $j \in \mathcal{X}\backslash\mathcal{T}$.

$$\begin{aligned} f_j^E(\mathcal{T})/f_j^E(\mathcal{S}) &= \frac{\lambda_{\min}\left(\mathbf{B}_{\mathcal{T}}^{-1} + \mathbf{I}_j\right) - \lambda_{\min}\left(\mathbf{B}_{\mathcal{T}}^{-1}\right)}{\lambda_{\min}\left(\mathbf{B}_{\mathcal{S}}^{-1} + \mathbf{I}_j\right) - \lambda_{\min}\left(\mathbf{B}_{\mathcal{S}}^{-1}\right)} \\ &\stackrel{(b)}{\leq} \frac{\lambda_{\min}\left(\mathbf{B}_{\mathcal{T}}^{-1}\right) + \lambda_{\max}\left(\mathbf{I}_j\right) - \lambda_{\min}\left(\mathbf{B}_{\mathcal{T}}^{-1}\right)}{\lambda_{\min}\left(\mathbf{B}_{\mathcal{S}}^{-1}\right) + \lambda_{\min}\left(\mathbf{I}_j\right) - \lambda_{\min}\left(\mathbf{B}_{\mathcal{S}}^{-1}\right)} \\ &\leq \frac{\lambda_{\max}\left(\mathbf{I}_j\right)}{\lambda_{\min}\left(\mathbf{I}_j\right)}, \end{aligned} \tag{17}$$

where (b) follows from Weyl's inequality (Bellman, 1997). Therefore,

$$c_{f^E} = \max_{(\mathcal{S},\mathcal{T},j)\in\tilde{\mathcal{X}}} f_j^D(\mathcal{T})/f_j^D(\mathcal{S}) \leq \max_{j\in\mathcal{X}} \frac{\lambda_{\max}\left(\mathbf{I}_j\right)}{\lambda_{\min}\left(\mathbf{I}_j\right)}. \tag{18}$$

For the additive weak-submodularity constant, we have

$$\begin{aligned} f_j^E(\mathcal{T}) - f_j^E(\mathcal{S}) &= \lambda_{\min}\left(\mathbf{B}_{\mathcal{T}}^{-1} + \mathbf{I}_j\right) - \lambda_{\min}\left(\mathbf{B}_{\mathcal{T}}^{-1}\right) \\ &\qquad - \lambda_{\min}\left(\mathbf{B}_{\mathcal{S}}^{-1} + \mathbf{I}_j\right) + \lambda_{\min}\left(\mathbf{B}_{\mathcal{S}}^{-1}\right) \\ &\stackrel{(c)}{\leq} \lambda_{\max}\left(\mathbf{I}_j\right) - \lambda_{\min}\left(\mathbf{I}_j\right) \end{aligned} \tag{19}$$

where (c) follows from Weyl's inequality. Hence,

$$\begin{aligned} \epsilon_{f^E} &= \max_{(\mathcal{S},\mathcal{T},j)\in\tilde{\mathcal{X}}} f_j^D(\mathcal{T}) - f_j^D(\mathcal{S}) \\ &\leq \max_{j\in\mathcal{X}} \left(\lambda_{\max}\left(\mathbf{I}_j\right) - \lambda_{\min}\left(\mathbf{I}_j\right)\right). \end{aligned} \tag{20}$$

Proof of Theorem 6

We first prove the monotonicity. Let $\mathbf{I}_j = \frac{1}{\sigma_j^2}\left(\mathbf{X}_j\mathbf{P}\mathbf{X}_j^\top + \mathbf{z}_j\mathbf{z}_j^\top\right)$. For any set $\mathcal{S}$ and $j \in \mathcal{X}\backslash\mathcal{S}$, define

$$\tilde{\mathbf{F}}_{\mathcal{S},j} = \mathbf{I}_x + \sum_{i\in\mathcal{S}} \mathbf{I}_i + \sigma_j^2\mathbf{I}_j = \mathbf{F}_{\mathcal{S}} + \sigma_j^2\mathbf{I}_j, \tag{21}$$

where both $\tilde{\mathbf{F}}_{\mathcal{S},j}$ and $\mathbf{F}_{\mathcal{S}}$ are invertible and positive definite (PSD) matrices. Using the matrix inversion lemma (Bellman, 1997) as well as some algebraic simplifications, we obtain an expression for the marginal gain according to

$$\begin{aligned} f_j^A(\mathcal{S}) &= \frac{\mathbf{z}_j^\top\tilde{\mathbf{F}}_{\mathcal{S},j}^{-2}\mathbf{z}_j}{\sigma_j^2 + \mathbf{z}_j^\top\tilde{\mathbf{F}}_{\mathcal{S},j}^{-1}\mathbf{z}_j} \\ &\qquad + \operatorname{Tr}\left(\mathbf{F}_{\mathcal{S}}^{-1}\mathbf{X}_j(\sigma_j^2\mathbf{P}^{-1} + \mathbf{X}_j^\top\mathbf{F}_{\mathcal{S}}^{-1}\mathbf{X}_j)^{-1}\mathbf{X}_j^\top\mathbf{F}_{\mathcal{S}}^{-1}\right). \end{aligned} \tag{22}$$

Notice the first term on the right-hand side is positive since $\tilde{\mathbf{F}}_{\mathcal{S},j}$ is PSD and hence the quadratic form $\mathbf{z}_j^\top\tilde{\mathbf{F}}_{\mathcal{S},j}^{-2}\mathbf{z}_j$ is also positive. Further, The second term on the right-hand side is also positive as it is trace of the quadratic form $\mathbf{F}_{\mathcal{S}}^{-1}\mathbf{X}_j(\sigma_j^2\mathbf{P}^{-1} + \mathbf{X}_j^\top\mathbf{F}_{\mathcal{S}}^{-1}\mathbf{X}_j)^{-1}\mathbf{X}_j^\top\mathbf{F}_{\mathcal{S}}^{-1}$ which is also PSD because the matrix $(\sigma_j^2\mathbf{P}^{-1} + \mathbf{X}_j^\top\mathbf{F}_{\mathcal{S}}\mathbf{X}_j)^{-1}$ is itself PSD. Thus, the marginal gain is positive and the function is monotonically increasing.

We now provide bounds on additive and multiplicative weak-submodularity constants of $f^E(\mathcal{S})$ (Note that it can be shown f^E is not in general submodular). Finding these bounds in the general form form of model requires intense algebraic techniques and the resulting bounds will not be interpretable. In stead, we here provide bounds in scenarios where $\mathbf{z}_i = \mathbf{0}$ and $\mathbf{X}_i = \mathbf{x}_i\mathbf{x}_i^\top$ (rank 1) which is motivated by the phase retrieval applications. In this setting, it can be shown the marginal gain simplifies to

$$f_j^A(\mathcal{S}) = \frac{\mathbf{x}_j^\top \tilde{\mathbf{F}}_{\mathcal{S}}^{-2}\mathbf{x}_j}{\mathbf{x}_j^\top(\sigma_j^2\mathbf{P} + \tilde{\mathbf{F}}_{\mathcal{S}}^{-1})\mathbf{x}_j}, \tag{23}$$

where $\tilde{\mathbf{F}}_{\mathcal{S}} = \mathbf{I}_x + \sum_{i\in\mathcal{S}} \frac{1}{\sigma_j^2}\mathbf{x}_i\mathbf{x}_i^\top\mathbf{P}\mathbf{x}_i\mathbf{x}_i^\top$. Hence, the definition of multiplicative weak-submodularity constant yields,

$$\begin{aligned} c_{f^A} &= \max_{(\mathcal{S},\mathcal{T},j)\in\tilde{\mathcal{X}}} f_j^A(\mathcal{T})/f_j^A(\mathcal{S}) \\ &= \max_{(\mathcal{S},\mathcal{T},j)\in\tilde{\mathcal{X}}} \frac{(\mathbf{x}_j^\top\tilde{\mathbf{F}}_{\mathcal{T}}^{-2}\mathbf{x}_j)(\mathbf{x}_j^\top(\sigma_j^2\mathbf{P}+\tilde{\mathbf{F}}_{\mathcal{S}}^{-1})\mathbf{x}_j)}{(\mathbf{x}_j^\top\tilde{\mathbf{F}}_{\mathcal{S}}^{-2}\mathbf{x}_j)(\mathbf{x}_j^\top(\sigma_j^2\mathbf{P}+\tilde{\mathbf{F}}_{\mathcal{T}}^{-1})\mathbf{x}_j)} \\ &\le \max_{(\mathcal{S},\mathcal{T},j)\in\tilde{\mathcal{X}}} \frac{\lambda_{\max}(\tilde{\mathbf{F}}_{\mathcal{T}}^{-2})\lambda_{\max}(\sigma_j^2\mathbf{P}+\tilde{\mathbf{F}}_{\mathcal{S}}^{-1})}{\lambda_{\min}(\tilde{\mathbf{F}}_{\mathcal{S}}^{-2})\lambda_{\min}(\sigma_j^2\mathbf{P}+\tilde{\mathbf{F}}_{\mathcal{T}}^{-1})}, \end{aligned} \tag{24}$$

where the last inequality follows from the Courant–Fischer min-max theorem (Bellman, 1997). Notice that by Weyl's inequality $\lambda_{\max}(\tilde{\mathbf{F}}_{\mathcal{S}}^{-1}) = \lambda_{\min}(\tilde{\mathbf{F}}_{\mathcal{S}})^{-1}$ and $\lambda_{\min}(\tilde{\mathbf{F}}_{\mathcal{T}}) \ge \lambda_{\min}(\tilde{\mathbf{F}}_{\mathcal{S}}) \ge \lambda_{\min}(\tilde{\mathbf{F}}_{\emptyset}) = \lambda_{\min}(\mathbf{I}_x)$. Therefore,

$$\begin{aligned} c_{f^A} &\le \max_j \frac{\lambda_{\max}(\mathbf{I}_x^{-1})^2\lambda_{\max}(\sigma_j^2\mathbf{P}+\mathbf{I}_x^{-1})}{\lambda_{\min}(\tilde{\mathbf{F}}_{[n]}^{-1})^2\lambda_{\min}(\sigma_j^2\mathbf{P}+\tilde{\mathbf{F}}_{[n]}^{-1})} \\ &\le \max_j \frac{\lambda_{\max}(\mathbf{I}_x^{-1})^3(\frac{\lambda_{\max}(\sigma_j^2\mathbf{P})}{\lambda_{\max}(\mathbf{I}_x^{-1})}+1)}{\lambda_{\min}(\tilde{\mathbf{F}}_{[n]}^{-1})^3(\frac{\lambda_{\min}(\sigma_j^2\mathbf{P})}{\lambda_{\min}(\tilde{\mathbf{F}}_{[n]}^{-1})}+1)}. \end{aligned} \tag{25}$$

Noting $\tilde{\mathbf{F}}_{[n]}^{-1} = \mathbf{B}_{[n]}$ completes the proof of bounded c_{f^A}. We can also use more applications of Weyl's inequality to achieve looser yet more intuitive and compact bounds. Using similar techniques such as applications of Courant–Fischer min-max theorem and Weyl's inequality we obtain the stated results for ϵ_{f^A}.

References

Bellman, R. (1997). *Introduction to matrix analysis*. SIAM.

Nemhauser, G. L., Wolsey, L. A., and Fisher, M. L. (Dec. 1978). An analysis of approximations for maximizing submodular set functions I. *Mathematical Programming*, 14(1):265–294.