We now prove Lemma 3.

We make use of the following result that provides an approximation bound for greedy selections for weakly submodular functions.

**Lemma 10.** [Das and Kempe, 2011] Let $S^*$ be the optimal k-sized set that maximizes $f(\cdot)$ under the k-sparsity constraint (see (2)). Let $S_G$ be the set returned by greedy forward selection (Algorithm 1), then

\[ f(S_G) \geq (1 - \exp(-\gamma_{S_G,k}^*))f(S^*). \]

### A.1 Distributed Greedy

**Lemma 2.** $f(G_j) \geq (1 - \exp(-\gamma_{G_j,k}^*))f(T_j)$.

**Proof.** From Lemma 1, we know that running greedy on $A_j \cup T_j$ instead of $A_j$ will still return the set $G_j$,

\[
\begin{align*}
  f(G_j) &\overset{\text{Lemma 10}}{\geq} (1 - \exp(-\gamma_{G_j,k}^*)) \max_{S \subseteq A_j \cup T_j} f(S) \\
  &\geq (1 - \exp(-\gamma_{G_j,k}^*))f(T_j).
\end{align*}
\]

For proving Lemma 3, we require another auxiliary result.

**Lemma 11.** For any $x \in A_i, \mathbb{P}(x \in \cup_j G_j) = \frac{1}{t} \sum_j \mathbb{P}(x \in S_j)$.

**Proof.** We have

\[
\begin{align*}
  \mathbb{P}(x \in \cup_j G_j) &= \sum_j \mathbb{P}(x \in A_i \cap x \in \text{GREEDY}(A_i, k)) \\
  &= \sum_j \mathbb{P}(x \in A_i)\mathbb{P}(x \in \text{GREEDY}(A_i, k) | x \in A_i) \\
  &= \sum_j \mathbb{P}(x \in A_i)\mathbb{P}(x \in S_j) \\
  &= \frac{1}{t} \mathbb{P}(x \in S_j).
\end{align*}
\]

We now prove Lemma 3.

**Lemma 3.** $\exists j \in [t], \mathbb{E}[f(G)] \geq (1 - \frac{1}{e^{-\gamma}}} f(S_j)$.

**Proof.** For $i \in [k]$, let $B_i : \text{GREEDY}(\cup_j G_j, i)$, so that $B_k = G$ in step 3 of Algorithm 2. Then,

\[
\mathbb{E}[f(B_{i+1}) - f(B_i)] \geq \frac{1}{k} \sum_{x \in \mathbb{A}^*} \mathbb{P}(x \in \cup_j G_j)\mathbb{E}[f(B_i \cup x) - f(B_i)]
\]

\[
\overset{\text{Lemma 11}}{=} \frac{1}{k} \sum_{x \in \mathbb{A}^*} \left( \sum_{j=1}^k \mathbb{P}(x \in S_j) \right) \mathbb{E}(f(B_i \cup x) - f(B_i))
\]

\[
= \frac{1}{k} \sum_{j=1}^k \sum_{x \in S_j} \mathbb{E}(f(B_i \cup x) - f(B_i))
\]

\[
= \frac{1}{k} \sum_{j=1}^k \gamma_{B_i,s_j \setminus B_i} \mathbb{E}(f(B_i \cup S_j - f(B_i)))
\]

\[
\overset{10}{\geq} \frac{1}{k} \sum_{j=1}^k \gamma_{B_i,s_j \setminus B_i} \mathbb{E}(f(S_j) - f(B_i))
\]

\[
= \frac{1}{k} \sum_{j=1}^k \gamma_{B_i,s_j \setminus B_i} \mathbb{E}(f(S_j) - f(B_i))
\]

Using $\gamma_{B_i,s_j \setminus B_i} \geq \gamma_{G_j,k}^*$ and proceeding now as in the proof of Theorem 2, we get the desired result.

\[\square\]

### A.2 Stochastic Greedy

**Lemma 4.** Let $A, B \subseteq \{1, \ldots, n\}$ with $|B| \leq k$. Consider another set $C$ drawn randomly from $\{1, \ldots, n\} \setminus A$ with $|C| = \lceil \frac{n\log k}{k} \rceil$. Then

\[
\mathbb{E}\left[\max_{v \in C} f(v \cup A) - f(A)\right] \geq \frac{(1 - \delta)\gamma_{A,B \setminus A}^*}{k} f(B) - f(A).
\]

**Proof.** To relate the best possible marginal gain from $C$ to the total gain of including the set $B \setminus A$ into $A$, we must upper bound the probability of overlap between $C$ and $B \setminus A$ as follows:

\[
\mathbb{P}(C \cap (B \setminus A) \neq \emptyset) = 1 - \left(1 - \frac{|B \setminus A|}{|n \setminus A|}\right)^{|C|}
\]

\[
= 1 - \left(1 - \frac{|B \setminus A|}{|n \setminus A|}\right)^{\lceil \log k \rceil / k}
\]

\[
\geq 1 - \exp\left(-\frac{n \log l / k}{|n \setminus A|}\right)
\]

\[
\geq 1 - \exp\left(-\frac{|B \setminus A| \log l / k}{k}\right)
\]

\[
= (1 - \delta)\frac{|B \setminus A|}{k}
\]

where (14) is because $\frac{|B \setminus A|}{k} \leq 1$. Let $S = C \cap (B \setminus A)$. Since
A similar analysis also gives

\[ f(v \cup A) - f(A) \text{ is nonnegative,} \]

\[ \mathbb{E}[\max_{v \in C} f(v \cup A) - f(A)] \]
\[ \geq \mathbb{P}(S \neq \emptyset) \mathbb{E}[\max_{v \in C} f(v \cup A) - f(A)|S \neq \emptyset] \]
\[ \geq (1 - \delta) \frac{|B\setminus A|}{k} \mathbb{E}[\max_{v \in C \setminus (B \setminus A)} f(v \cup B) - f(B)|S \neq \emptyset] \]
\[ \geq (1 - \delta) \frac{|B\setminus A|}{k} \max_{v \in C \setminus (B \setminus A)} \max(f(v \cup A) - f(A)|S \neq \emptyset) \]
\[ \geq (1 - \delta) \frac{|B\setminus A|}{k} \sum_{v \in B \setminus A} \max_{v \in C \setminus (B \setminus A)} \|f(v \cup A) - f(A)|S \neq \emptyset\| \]
\[ \geq (1 - \delta) \frac{|B\setminus A|}{k} \max_{v \in C \setminus (B \setminus A)} (f(B) - f(A)). \]

\[ \Box \]

A.3 Linear regression

**Lemma’ 6.** For the maximization of the \( R^2 \) \( \nu_S \geq \frac{\lambda_{\max}(C_S)}{\lambda_{\max}(C)} \), where \( C_S \) is the submatrix of \( C \) with rows and columns indexed by \( S \).

**Proof.** Say \( A, B \) is an arbitrary partition of \( S \). Consider,

\[ \|P_{AY}\|^2 = y^T P_{AY} \]
\[ = y^T X_A (X_A X_A)^{-1} X_A^T y \]
\[ \geq \|X_A y\|^2 \frac{1}{\lambda_{\min}((X_A X_A)^{-1})} \]
\[ = \|X_A y\|^2 \frac{1}{\lambda_{\max}(X_A X_A)} \]
\[ \geq \|X_A y\|^2 \frac{1}{\lambda_{\max}(X_S X_S)}. \]

where (15) results from the fact that all orthogonal projection matrices are symmetric and idempotent. Repeating a similar analysis for \( B \) instead of \( A \), we get

\[ \|P_{BY}\|^2 + \|P_{B^T Y}\|^2 \geq \frac{\|X_A y\|^2 + \|X_B y\|^2}{\lambda_{\max}(X_S X_S)} \]
\[ = \frac{\|X_S y\|^2}{\lambda_{\min}(X_S X_S)} \]
\[ \frac{\|X_S y\|^2}{\lambda_{\max}(X_S X_S)} \]

A similar analysis also gives \( \|P_{S^T Y}\|^2 \leq \frac{\|X_S y\|^2}{\lambda_{\min}(X_S X_S)} \), which gives the desired result.

A.4 RSC implies weak subadditivity

Let \( \beta^{(S)} := \max_{\supp(x) \subseteq S} g(x) \).

**Lemma’ 8.** For a support set \( S \subset [d] \), \( f(S) \geq \frac{1}{2\pi} \|\nabla g(0)\|_2^2 \).

**Proof.** For any \( v \) with support in \( S \),

\[ g(\beta^{(S)}) - g(0) \geq g(v) - g(0) \]
\[ \geq \langle \nabla g(0), v \rangle - \frac{L}{2} \|v\|_2^2. \]

Using \( v = \frac{1}{L} \nabla g(0)_S \), we get the desired result.

**\Box**

**Lemma’ 9.** For any support set \( S \), \( f(S) \leq \frac{1}{2m} \|\nabla g(0)_S\|_2^2 \).

**Proof.** By strong concavity,

\[ g(\beta^{(S)}) - g(0) \leq \langle \nabla g(0), \beta^{(S)} \rangle - \frac{m}{2} \|\beta^{(S)}\|_2^2 \]
\[ \leq \max_v \langle \nabla g(0), v \rangle - \frac{m}{2} \|v\|_2^2, \]

where \( v \) is an arbitrary vector that has support only on \( S \). Optimizing the RHS over \( v \) gives the desired result.

**\Box**

B Additional experiments

Figures 4, 5 shows the performance of all algorithms on the following metrics: log likelihood (normalized with respect to a null model), generalization to new test measurements from the same true support parameter, area under ROC, and percentage of the true support recovered for \( l = 2 \). Recall that Figure 1 presents the results from the same experiment with \( l = 10 \). Clearly, the greedy algorithms benefit more from increased number of partitions.

Figure 4: Distributed linear regression, \( l = 2 \) partitions, \( n = 800 \) training and test samples, \( \alpha = 0.5 \). Training/testing performance
Figure 5: Distributed linear regression, \( l = 2 \) partitions, \( n = 800 \) training and test samples, \( \alpha = 0.5 \). Support Recovery Performance