9 Supplementary Material

9.1 Proof of Lemma 1

Use the definition $d_t := (1 - \rho_t) d_{t-1} + \rho_t \nabla \tilde{F}(x_t, z_t)$ to write $\|\nabla F(x_t) - d_t\|^2$ as

$$\|\nabla F(x_t) - d_t\|^2 = \|\nabla F(x_t) - (1 - \rho_t) d_{t-1} - \rho_t \nabla \tilde{F}(x_t, z_t)\|^2. \quad (36)$$

Add and subtract the term $(1 - \rho_t) \nabla F(x_{t-1})$ to the right hand side of (36), regroup the terms to obtain

$$\|\nabla F(x_t) - d_t\|^2 = \|\rho_t(\nabla F(x_t) - \nabla \tilde{F}(x_t, z_t)) + (1 - \rho_t)(\nabla F(x_t) - \nabla F(x_t)) + (1 - \rho_t)(\nabla F(x_t - 1) - d_{t-1})\|^2. \quad (37)$$

Define $\mathcal{F}_t$ as a sigma algebra that measures the history of the system up until time $t$. Expanding the square and computing the conditional expectation $E[\cdot | \mathcal{F}_t]$ of the resulted expression yield

$$E \left[ \|\nabla F(x_t) - d_t\|^2 \mid \mathcal{F}_t \right] = \rho_t^2 E \left[ \|\nabla F(x_t) - \nabla \tilde{F}(x_t, z_t)\|^2 \mid \mathcal{F}_t \right] + (1 - \rho_t)^2 \|\nabla F(x_{t-1}) - d_{t-1}\|^2
+ (1 - \rho_t)^2 \|\nabla F(x_t) - \nabla F(x_{t-1})\|^2 + 2(1 - \rho_t)^2 \langle \nabla F(x_t) - \nabla F(x_{t-1}), \nabla F(x_{t-1}) - d_{t-1} \rangle. \quad (38)$$

The term $E \left[ \|\nabla F(x_t) - \nabla \tilde{F}(x_t, z_t)\|^2 \mid \mathcal{F}_t \right]$ can be bounded above by $\sigma^2$ according to Assumption 3. Based on Assumptions 1 and 2, we can also show that the squared norm $\|\nabla F(x_t) - \nabla F(x_{t-1})\|^2$ is upper bounded by $L^2 D^2 / T^2$. Moreover, the inner product $2 \nabla F(x_t) - \nabla F(x_{t-1}), \nabla F(x_{t-1}) - d_{t-1}$ can be upper bounded by $\beta_t \|\nabla F(x_{t-1}) - d_{t-1}\|^2 + (1/\beta_t) L^2 D^2 / T^2$ using Young’s inequality (i.e., $2 \langle a, b \rangle \leq \beta \|a\|^2 + \|b\|^2 / \beta$ for any $a, b \in \mathbb{R}^n$ and $\beta > 0$) and the conditions in Assumptions 1 and 2, where $\beta_t > 0$ is a free scalar. Applying these substitutions into (38) leads to

$$E \left[ \|\nabla F(x_t) - d_t\|^2 \mid \mathcal{F}_t \right] \leq \rho_t^2 \sigma^2 + (1 - \rho_t)^2 (1 + \frac{1}{\beta_t}) \frac{L^2 D^2}{T^2} + (1 - \rho_t)^2 (1 + \beta_t) \|\nabla F(x_{t-1}) - d_{t-1}\|^2. \quad (39)$$

Replace $(1 - \rho_t)^2$ by $(1 - \rho_t)$, set $\beta := \rho_t / 2$, and compute the expectation with respect to $\mathcal{F}_0$ to obtain

$$E \left[ \|\nabla F(x_t) - d_t\|^2 \right] \leq \rho_t^2 \sigma^2 + \frac{L^2 D^2}{T^2} \frac{2L^2 D^2}{\rho_t T^2} + \left(1 - \frac{\rho_t}{2}\right) \frac{E \left[ \|\nabla F(x_{t-1}) - d_{t-1}\|^2 \right]}{2}, \quad (40)$$

and the claim in (14) follows.

9.2 Proof of Lemma 2

Define $a_t := E \left[ \|\nabla F(x_t) - d_t\|^2 \right]$. Also, assume $\rho_t = \frac{4}{(t+s)^{2/3}}$, where $s$ is a fixed scalar and satisfies the condition $8 \leq s \leq T$ (so the proof is slightly more general). Apply these substitutions into (14) to obtain

$$a_t \leq \left(1 - \frac{2}{(t+s)^{2/3}}\right) a_{t-1} + \frac{16\sigma^2}{(t+s)^{4/3}} + \frac{L^2 D^2}{T^2} + \frac{L^2 D^2 (t+s)^{2/3}}{2T^2}. \quad (41)$$

Now use the conditions $s \leq T$ and $t \leq T$ to replace $1/T$ in (41) by its upper bound $2/(t+s)$. Applying this substitution leads to

$$a_t \leq \left(1 - \frac{2}{(t+s)^{2/3}}\right) a_{t-1} + \frac{16\sigma^2}{(t+s)^{4/3}} + \frac{4L^2 D^2}{(t+s)^2} + \frac{2L^2 D^2}{(t+s)^{4/3}}. \quad (42)$$

Since $t+s \geq 8$ we can write $(t+s)^2 = (t+s)^{4/3}(t+s)^{2/3} \geq (t+s)^{4/3}(t+s)^{2/3} \geq 4(t+s)^{4/3}$. Replacing the term $(t+s)^2$ in (42) by $4(t+s)^{4/3}$ and regrouping the terms lead to

$$a_t \leq \left(1 - \frac{2}{(t+s)^{2/3}}\right) a_{t-1} + \frac{16\sigma^2 + 3L^2 D^2}{(t+s)^{4/3}} \quad (43)$$

Now we prove by induction that for $t = 0, \ldots, T$ we can write

$$a_t \leq \frac{Q}{(t+s + 1)^{2/3}}, \quad (44)$$
where $Q := \max\{a_0(s+1)^{2/3}, 16\sigma^2 + 3L^2D^2\}$. First, note that $Q \geq a_0(s+1)^{2/3}$ and therefore $a_0 \leq Q/(s+1)^{2/3}$ and the base step of the induction holds true. Now assume that the condition in (44) holds for $t = k - 1$, i.e.,

$$a_{k-1} \leq \frac{Q}{(k+s)^{2/3}}. \quad (45)$$

The goal is to show that (44) also holds for $t = k$. To do so, first set $t = k$ in the expression in (43) to obtain

$$a_k \leq \left(1 - \frac{2}{(k+s)^{2/3}}\right) a_{k-1} + \frac{16\sigma^2 + 3L^2D^2}{(k+s)^{4/3}}. \quad (46)$$

According to the definition of $Q$, we know that $Q \geq 16\sigma^2 + 3L^2D^2$. Moreover, based on the induction hypothesis it holds that $a_{k-1} \leq \frac{Q}{(k+s)^{2/3}}$. Using these inequalities and the expression in (46) we can write

$$a_k \leq \left(1 - \frac{2}{(k+s)^{2/3}}\right) \frac{Q}{(k+s)^{2/3}} + \frac{Q}{(k+s)^{4/3}}. \quad (47)$$

Pulling out $\frac{Q}{(k+s)^{2/3}}$ as a common factor and simplifying and reordering terms it follows that (47) is equivalent to

$$a_k \leq Q \left(\frac{(k+s)^{2/3} - 1}{(k+s)^{4/3}}\right). \quad (48)$$

Based on the inequality

$$((k+s)^{2/3} - 1)((k+s)^{2/3} + 1) < (k+s)^{4/3}, \quad (49)$$

the result in (48) implies that

$$a_k \leq \left(\frac{Q}{(k+s)^{2/3} + 1}\right). \quad (50)$$

Since $(k+s)^{2/3} + 1 \geq (k+s+1)^{2/3}$, the result in (50) implies that

$$a_k \leq \left(\frac{Q}{(k+s+1)^{2/3}}\right), \quad (51)$$

and the induction step is complete. Therefore, the result in (44) holds for all $t = 0, \ldots, T$. Indeed, by setting $s = 8$, the claim in (15) follows.

### 9.3 How to Construct an Unbiased Estimator of the Gradient in Multilinear Extensions

Recall that $f(S) = \mathbb{E}_{\mathbf{z} \sim P}[\hat{f}(S, \mathbf{z})]$. In terms of the multilinear extensions, we obtain $F(\mathbf{x}) = \mathbb{E}_{\mathbf{z} \sim P}[\hat{F}(\mathbf{x}, \mathbf{z})]$, where $F$ and $\hat{F}$ denote the multilinear extension for $f$ and $\hat{f}$, respectively. So $\nabla \hat{F}(\mathbf{x}, \mathbf{z})$ is an unbiased estimator of $\nabla F(\mathbf{x})$ when $\mathbf{z} \sim P$. Note that $\hat{F}(\mathbf{x}, \mathbf{z})$ is a multilinear extension.

It remains to provide an unbiased estimator for the gradient of a multilinear extension. We thus consider an arbitrary submodular set function $g$ with multilinear $G$. Our goal is to provide an unbiased estimator for $\nabla G(\mathbf{x})$.

We have $G(\mathbf{x}) = \sum_{S \subseteq V} \prod_{i \in S} x_i \prod_{j \notin S} (1 - x_j) g(S)$. Now, it can easily be shown that

$$\frac{\partial G}{\partial x_i} = G(\mathbf{x}; x_i \leftarrow 1) - G(\mathbf{x}; x_i \leftarrow 0). \quad (52)$$

where for example by $(\mathbf{x}; x_i \leftarrow 1)$ we mean a vector which has value 1 on its $i$-th coordinate and is equal to $\mathbf{x}$ elsewhere. To create an unbiased estimator for $\frac{\partial G}{\partial x_i}$ at a point $\mathbf{x}$ we can simply sample a set $S$ by including each element in it independently with probability $x_i$ and use $g(S \cup \{i\}) - g(S \setminus \{i\})$ as an unbiased estimator for the $i$-th partial derivative. We can sample one single set $S$ and use the above trick for all the coordinates. This involves $n$ function computations for $g$. Having a mini-batch size $B$ we can repeat this procedure $B$ times and then average.
9.4 Proof of Lemma 3

Based on the mean value theorem, we can write
\[
\nabla F(x_t + \frac{1}{T} v_t) - \nabla F(x_t) = \frac{1}{T} H(\tilde{x}_t) v_t,
\]
where \( \tilde{x}_t \) is a convex combination of \( x_t \) and \( x_t + \frac{1}{T} v_t \) and \( H(\tilde{x}_t) := \nabla^2 F(\tilde{x}_t) \). This expression shows that the difference between the coordinates of the vectors \( \nabla F(x_t + \frac{1}{T} v_t) \) and \( \nabla F(x_t) \) can be written as
\[
\nabla_j F(x_t + \frac{1}{T} v_t) - \nabla_j F(x_t) = \frac{1}{T} \sum_{i=1}^{n} H_{j,i}(\tilde{x}_t) v_{i,t},
\]
where \( v_{i,t} \) is the \( i \)-th element of the vector \( v_t \) and \( H_{j,i} \) denotes the component in the \( j \)-th row and \( i \)-th column of the matrix \( H \). Hence, the norm of the difference \( |\nabla_j F(x_t + \frac{1}{T} v_t) - \nabla_j F(x_t)| \) is bounded above by
\[
|\nabla_j F(x_t + \frac{1}{T} v_t) - \nabla_j F(x_t)| \leq \frac{1}{T} \left| \sum_{i=1}^{n} H_{j,i}(\tilde{x}_t) v_{i,t} \right|.
\]
(55)

Note here that the elements of the matrix \( H(\tilde{x}_t) \) are less than the maximum marginal value (i.e. \( \max_{i,j} |H_{i,j}(\tilde{x}_t)| \leq \max_{i \in \{1, \ldots, n\}} f(i) \equiv m_f \)). We thus get
\[
|\nabla_j F(x_t + \frac{1}{T} v_t) - \nabla_j F(x_t)| \leq \frac{m_f}{T} \sum_{i=1}^{n} |v_{i,t}|.
\]
(56)

Note that at each round \( t \) of the algorithm, we have to pick a vector \( v_t \in \mathcal{C} \) s.t. the inner product \( \langle v_t, d_t \rangle \) is maximized. Hence, without loss of generality we can assume that the vector \( v_t \) is one of the extreme points of \( \mathcal{C} \), i.e. it is of the form \( 1_I \) for some \( I \in \mathcal{I} \) (note that we can easily force integer vectors). Therefore by noticing that \( v_t \) is an integer vector with at most \( r \) ones, we have
\[
|\nabla_j F(x_t + \frac{1}{T} v_t) - \nabla_j F(x_t)| \leq \frac{m_f \sqrt{T}}{T} \sqrt{\sum_{i=1}^{n} |v_{i,t}|^2},
\]
(57)

which yields the claim in (28).

9.5 Proof of Theorem 2

According to the Taylor’s expansion of the function \( F \) near the point \( x_t \) we can write
\[
F(x_{t+1}) = F(x_t) + \langle \nabla F(x_t), x_{t+1} - x_t \rangle + \frac{1}{2} \langle x_{t+1} - x_t, H(\tilde{x}_t)(x_{t+1} - x_t) \rangle
\]
\[
= F(x_t) + \frac{1}{T} \langle \nabla F(x_t), v_t \rangle + \frac{1}{2T^2} \langle v_t, H(\tilde{x}_t) v_t \rangle,
\]
(58)

where \( \tilde{x}_t \) is a convex combination of \( x_t \) and \( x_t + \frac{1}{T} v_t \) and \( H(\tilde{x}_t) := \nabla^2 F(\tilde{x}_t) \). Note that based on the inequality \( \max_{i,j} |H_{i,j}(\tilde{x}_t)| \leq \max_{i \in \{1, \ldots, n\}} f(i) \equiv m_f \), we can lower bound \( H_{i,j} \) by \(-m_f\). Therefore,
\[
\langle v_t, H(\tilde{x}_t) v_t \rangle = \sum_{j=1}^{n} \sum_{i=1}^{n} v_{i,j} v_{i,j} H_{i,j}(\tilde{x}_t) \geq -m_f \sum_{j=1}^{n} \sum_{i=1}^{n} v_{i,t} v_{i,t} = -m_f \left( \sum_{i=1}^{n} v_{i,t} \right)^2 = -m_f r \|v_t\|^2,
\]
(59)

where the last inequality is because \( v_t \) is a vector with \( r \) ones and \( n - r \) zeros (see the explanation in the proof of Lemma 3). Replace the expression \( \langle v_t, H(\tilde{x}_t) v_t \rangle \) in (58) by its lower bound in (59) to obtain
\[
F(x_{t+1}) \geq F(x_t) + \frac{1}{T} \langle \nabla F(x_t), v_t \rangle - \frac{m_f r}{2T^2} \|v_t\|^2.
\]
(60)

In the following lemma we derive a variant of the result in Lemma 2 for the multilinear extension setting.
Lemma 4. Consider Stochastic Continuous Greedy (SCG) outlined in Algorithm 1, and recall the definitions of the function $F$ in (27), the rank $r$, and $m_f \triangleq \max_{i \in \{1, \ldots, n\}} f(i)$. If we set $\rho_t = \frac{4}{(t+8)^{2/3}}$, then for $t = 0, \ldots, T$ and for $j = 1, \ldots, n$ it holds

$$E \left[ \|\nabla F(x_t) - d_t\|^2 \right] \leq \frac{Q}{(t+9)^{2/3}},$$

where $Q := \max\{9^{2/3}\|\nabla F(x_0) - d_0\|^2, 16\sigma^2 + 3m_f^2 rD^2\}$.

Proof. The proof is similar to the proof of Lemma 1. The main difference is to write the analysis for the $j$-th coordinate and replace $L$ by $m_f \sqrt{r}$ as shown in Lemma 3. Then using the proof techniques in Lemma 2 the claim in Lemma 4 follows.

The rest of the proof is identical to the proof of Theorem 1, by following the steps from (17) to (25) and considering the bound in (61) we obtain

$$E [F(x_T)] \geq (1 - 1/e)F(x^*) - \frac{2DQ^{1/2}}{T^{1/3}} - \frac{m_f r D^2}{2T},$$

where $Q := \max\{\|\nabla F(x_0) - d_0\|^2 9^{2/3}, 16\sigma^2 + 3m_f^2 rD^2\}$. Therefore, the claim in Theorem 2 follows.