9 Supplementary Material

9.1 Proof of Lemma 1

Use the definition $\mathbf{d}_t := (1-\rho_t)\mathbf{d}_{t-1} + \rho_t \nabla \tilde{F}(\mathbf{x}_t, \mathbf{z}_t)$ to write $\|\nabla F(\mathbf{x}_t) - \mathbf{d}_t\|^2$ as

$$\|\nabla F(\mathbf{x}_t) - \mathbf{d}_t\|^2 = \|\nabla F(\mathbf{x}_t) - (1-\rho_t)\mathbf{d}_{t-1} - \rho_t \nabla \tilde{F}(\mathbf{x}_t, \mathbf{z}_t)\|^2. \tag{36}$$

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

$$\|\nabla F(\mathbf{x}_t) - \mathbf{d}_t\|^2 = \|\rho_t(\nabla F(\mathbf{x}_t) - \nabla \tilde{F}(\mathbf{x}_t, \mathbf{z}_t)) + (1-\rho_t)(\nabla F(\mathbf{x}_t) - \nabla F(\mathbf{x}_{t-1})) + (1-\rho_t)(\nabla F(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1})\|^2. \tag{37}$$

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 $\mathbb{E}[\cdot \mid \mathcal{F}_t]$ of the resulted expression yield

$$\begin{aligned}\mathbb{E}\left[\|\nabla F(\mathbf{x}_t) - \mathbf{d}_t\|^2 \mid \mathcal{F}_t\right] &= \rho_t^2 \mathbb{E}\left[\|\nabla F(\mathbf{x}_t) - \nabla \tilde{F}(\mathbf{x}_t, \mathbf{z}_t)\|^2 \mid \mathcal{F}_t\right] + (1-\rho_t)^2 \|\nabla F(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}\|^2 \\ &+ (1-\rho_t)^2 \|\nabla F(\mathbf{x}_t) - \nabla F(\mathbf{x}_{t-1})\|^2 + 2(1-\rho_t)^2 \langle \nabla F(\mathbf{x}_t) - \nabla F(\mathbf{x}_{t-1}), \nabla F(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}\rangle. \end{aligned} \tag{38}$$

The term $\mathbb{E}\left[\|\nabla F(\mathbf{x}_t) - \nabla \tilde{F}(\mathbf{x}_t, \mathbf{z}_t)\|^2 \mid \mathcal{F}_t\right]$ can be bounded above by σ^2 according to Assumption 3. Based on Assumptions 1 and 2, we can also show that the squared norm $\|\nabla F(\mathbf{x}_t) - \nabla F(\mathbf{x}_{t-1})\|^2$ is upper bounded by L^2D^2/T^2. Moreover, the inner product $2\langle \nabla F(\mathbf{x}_t) - \nabla F(\mathbf{x}_{t-1}), \nabla F(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}\rangle$ can be upper bounded by $\beta_t \|\nabla F(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}\|^2 + (1/\beta_t) L^2D^2/T^2$ using Young's inequality (i.e., $2\langle \mathbf{a}, \mathbf{b}\rangle \leq \beta \|\mathbf{a}\|^2 + \|\mathbf{b}\|^2/\beta$ for any $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$ 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

$$\mathbb{E}\left[\|\nabla F(\mathbf{x}_t) - \mathbf{d}_t\|^2 \mid \mathcal{F}_t\right] \leq \rho_t^2 \sigma^2 + (1-\rho_t)^2 (1 + \frac{1}{\beta_t}) \frac{L^2D^2}{T^2} + (1-\rho_t)^2 (1+\beta_t) \|\nabla F(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}\|^2. \tag{39}$$

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

$$\mathbb{E}\left[\|\nabla F(\mathbf{x}_t) - \mathbf{d}_t\|^2\right] \leq \rho_t^2 \sigma^2 + \frac{L^2D^2}{T^2} + \frac{2L^2D^2}{\rho_t T^2} + \left(1 - \frac{\rho_t}{2}\right) \mathbb{E}\left[\|\nabla F(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}\|^2\right], \tag{40}$$

and the claim in (14) follows.

9.2 Proof of Lemma 2

Define $a_t := \mathbb{E}\left[\|\nabla F(\mathbf{x}_t) - \mathbf{d}_t\|^2\right]$. 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

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

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

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

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

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

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

$$a_t \leq \frac{Q}{(t+s+1)^{2/3}}, \tag{44}$$

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.,

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

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

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

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

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

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

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

Based on the inequality

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

the result in (48) implies that

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

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

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

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

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

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

$$\nabla F(\mathbf{x}_t + \frac{1}{T}\mathbf{v}_t) - \nabla F(\mathbf{x}_T) = \frac{1}{T}\mathbf{H}(\tilde{\mathbf{x}}_t)\mathbf{v}_t, \tag{53}$$

where $\tilde{\mathbf{x}}_t$ is a convex combination of $\mathbf{x}_t$ and $\mathbf{x}_t + \frac{1}{T}\mathbf{v}_t$ and $\mathbf{H}(\tilde{\mathbf{x}}_t) := \nabla^2 F(\tilde{\mathbf{x}}_t)$. This expression shows that the difference between the coordinates of the vectors $\nabla F(\mathbf{x}_t + \frac{1}{T}\mathbf{v}_t)$ and $\nabla F(\mathbf{x}_t)$ can be written as

$$\nabla_j F(\mathbf{x}_t + \frac{1}{T}\mathbf{v}_t) - \nabla_j F(\mathbf{x}_t) = \frac{1}{T}\sum_{i=1}^{n} H_{j,i}(\tilde{\mathbf{x}}_t)v_{i,t}, \tag{54}$$

where $v_{i,t}$ is the i-th element of the vector $\mathbf{v}_t$ and $H_{j,i}$ denotes the component in the j-th row and i-th column of the matrix $\mathbf{H}$. Hence, the norm of the difference $|\nabla_j F(\mathbf{x}_t + \frac{1}{T}\mathbf{v}_t) - \nabla_j F(\mathbf{x}_t)|$ is bounded above by

$$|\nabla_j F(\mathbf{x}_t + \frac{1}{T}\mathbf{v}_t) - \nabla_j F(\mathbf{x}_t)| \leq \frac{1}{T}\left|\sum_{i=1}^{n} H_{j,i}(\tilde{\mathbf{x}}_t)v_{i,t}\right|. \tag{55}$$

Note here that the elements of the matrix $\mathbf{H}(\tilde{\mathbf{x}}_t)$ are less than the maximum marginal value (i.e. $\max_{i,j}|H_{i,j}(\tilde{\mathbf{x}}_t)| \leq \max_{i\in\{1,\cdots,n\}} f(i) \triangleq m_f$). We thus get

$$|\nabla_j F(\mathbf{x}_t + \frac{1}{T}\mathbf{v}_t) - \nabla_j F(\mathbf{x}_t)| \leq \frac{m_f}{T}\sum_{i=1}^{n}|v_{i,t}|. \tag{56}$$

Note that at each round t of the algorithm, we have to pick a vector $\mathbf{v}_t \in \mathcal{C}$ s.t. the inner product $\langle \mathbf{v}_t, \mathbf{d}_t\rangle$ is maximized. Hence, without loss of generality we can assume that the vector $\mathbf{v}_t$ 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 $\mathbf{v}_t$ is an integer vector with at most r ones, we have

$$|\nabla_j F(\mathbf{x}_t + \frac{1}{T}\mathbf{v}_t) - \nabla_j F(\mathbf{x}_t)| \leq \frac{m_f\sqrt{r}}{T}\sqrt{\sum_{i=1}^{n}|v_{i,t}|^2}, \tag{57}$$

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 $\mathbf{x}_t$ we can write

$$\begin{aligned} F(\mathbf{x}_{t+1}) &= F(\mathbf{x}_t) + \langle \nabla F(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}_t\rangle + \frac{1}{2}\langle \mathbf{x}_{t+1} - \mathbf{x}_t, \mathbf{H}(\tilde{\mathbf{x}}_t)(\mathbf{x}_{t+1} - \mathbf{x}_t)\rangle \\ &= F(\mathbf{x}_t) + \frac{1}{T}\langle \nabla F(\mathbf{x}_t), \mathbf{v}_t\rangle + \frac{1}{2T^2}\langle \mathbf{v}_t, \mathbf{H}(\tilde{\mathbf{x}}_t)\mathbf{v}_t\rangle, \end{aligned} \tag{58}$$

where $\tilde{\mathbf{x}}_t$ is a convex combination of $\mathbf{x}_t$ and $\mathbf{x}_t + \frac{1}{T}\mathbf{v}_t$ and $\mathbf{H}(\tilde{\mathbf{x}}_t) := \nabla^2 F(\tilde{\mathbf{x}}_t)$. Note that based on the inequality $\max_{i,j}|H_{i,j}(\tilde{\mathbf{x}}_t)| \leq \max_{i\in\{1,\cdots,n\}} f(i) \triangleq m_f$, we can lower bound H_{ij} by $-m_f$. Therefore,

$$\langle \mathbf{v}_t, \mathbf{H}(\tilde{\mathbf{x}}_t)\mathbf{v}_t\rangle = \sum_{j=1}^{n}\sum_{i=1}^{n} v_{i,t}v_{j,t}H_{ij}(\tilde{\mathbf{x}}_t) \geq -m_f\sum_{j=1}^{n}\sum_{i=1}^{n} v_{i,t}v_{j,t} = -m_f\left(\sum_{i=1}^{n} v_{i,t}\right)^2 = -m_f r\|\mathbf{v}_t\|^2, \tag{59}$$

where the last inequality is because $\mathbf{v}_t$ is a vector with r ones and $n - r$ zeros (see the explanation in the proof of Lemma 3). Replace the expression $\langle \mathbf{v}_t, \mathbf{H}(\tilde{\mathbf{x}}_t)\mathbf{v}_t\rangle$ in (58) by its lower bound in (59) to obtain

$$F(\mathbf{x}_{t+1}) \geq F(\mathbf{x}_t) + \frac{1}{T}\langle \nabla F(\mathbf{x}_t), \mathbf{v}_t\rangle - \frac{m_f r}{2T^2}\|\mathbf{v}_t\|^2. \tag{60}$$

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 $m_f \triangleq \max_{i\in\{1,\cdots,n\}} f(i)$. If we set $\rho_t = \frac{4}{(t+8)^{2/3}}$, then for $t = 0,\dots,T$ and for $j = 1,\dots,n$ it holds*

$$\mathbb{E}\left[\|\nabla F(\mathbf{x}_t) - \mathbf{d}_t\|^2\right] \leq \frac{Q}{(t+9)^{2/3}}, \tag{61}$$

where $Q := \max\{9^{2/3}\|\nabla F(\mathbf{x}_0) - \mathbf{d}_0\|^2, 16\sigma^2 + 3m_f^2 r D^2\}$.

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 and 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

$$\mathbb{E}\left[F(\mathbf{x}_T)\right] \geq (1-1/e)F(\mathbf{x}^*) - \frac{2DQ^{1/2}}{T^{1/3}} - \frac{m_f r D^2}{2T}, \tag{62}$$

where $Q := \max\{\|\nabla F(\mathbf{x}_0) - \mathbf{d}_0\|^2 9^{2/3}, 16\sigma^2 + 3rm_f^2 D^2\}$. Therefore, the claim in Theorem 2 follows.