A Experiments

In this section, we empirically validate the efficiency of the proposed 1-SFW algorithm by comparing it with the baseline methods: Stochastic Frank-Wolfe (SFW) [Hazan and Luo [2016]] and Stochastic Conditional Gradient (SCG) [Mokhtari et al. [2018b]]. Note that SCG is the only existing provably convergent Frank-Wolfe variant that accepts a constant per-iteration mini-batch size (possibly 1). Denote the constant mini-batch size of 1-SFW and SCG by $m$. The growing mini-batch size of SFW is set to $m \cdot t^2$, where $t$ is the iteration count.

We study three types of problems, i.e., $\ell_1$-constrained logistic-regression (convex), robust low rank matrix recovery (nonconvex), and maximization of multilinear extensions of monotone discrete submodular functions (DR-submodular).

A.1 Logistic Regression

In this task, we consider $\ell_1$-constrained logistic regression problem. Concretely, denote each data point $i$ by $(a_i, y_i) \in \mathbb{R}^d \times \{-1, +1\}$, where $a_i$ is a feature vector and $y_i \in \{1, \ldots, C\}$ is the corresponding label. Our goal is to minimize the following loss

$$ F(W) = \frac{1}{n} \sum_{i=1}^{n} \log(1 + \exp(-y_i W^T c a_i)),$$

over the constraint $C = \{W \in \mathbb{R}^{d \times C} : \|W\|_1 \leq r\}$ for some constant $r \in \mathbb{R}_+$, where $\|W\|_1$ is the matrix $\ell_1$ norm, i.e., $\|W\|_1 = \max_{1 \leq j \leq C} \sum_{i=1}^{d} |[W]_{ij}|$. We note that the loss function $F$ is convex.

Two datasets are used in our experiments: MNIST (digit 2 and 4 as positive and negative class respectively) and CIFAR10 (cat and dog as positive and negative class respectively). In terms of the parameter setting, we grid search the step size $\eta_t$ for all three methods over the set $\{\min\{1, c/(t + 1)^a\} | c \in \{0.1, 0.25, 0.5, 1.0, 2.0\}, a \in \{1, 2/3, 1/2\}\}$, set the mixing weights $\rho_t$ of SCG and 1-SFW to $1/(t + 1)^{2/3}$, and set the constant mini-batch parameter $m = 16$. We report the results in Figure 1. We can see the advantage of 1-SFW over its competitors.

A.2 Robust Low-Rank Matrix Recovery

LRMR plays a key role in solving many important learning tasks, such as collaborative filtering [Koren et al., 2009], dimensionality reduction [Weinberger and Saul, 2006], and multi-class learning [Xu et al., 2013]. The loss of LRMR is defined as

$$ \min_{X \in \mathbb{R}^{M \times N}} \sum_{(i,j) \in \Omega} \psi(X_{ij} - Y_{ij}) $$

$$ s.t. \quad \|X\|_* \leq B, \quad (13) $$
where $\psi : \mathbb{R} \rightarrow \mathbb{R}$ is the potentially nonconvex empirical loss function, $X_{ij}$ is the $i, j$th element of matrix $X$, and $\Omega$ is the set of observed indices in target matrix $Y \in \mathbb{R}^{M \times N}$. Here we focus on a robust version of LRMR with the loss $\psi$ being:

$$\psi(z; \sigma) = 1 - \exp(-z^2/2\sigma),$$

where $\sigma$ is a tunable parameter. Loss (14) is less sensitive to the discrepancy $X_{ij} - Y_{ij}$, compared to the common least square loss $\psi(z) = z^2/2$, and hence is robust to adversarial outliers [Qu et al., 2017].

In each trial, we first generate an underlying matrix $M$ of size $200 \times 200$ and rank $\gamma = 15$. The singular values of $M$ are set as $2^{\gamma}/2^{\gamma} \times 50$ and hence $\|M\|_\infty \leq C = 100$, where $[\gamma] = \{1, \ldots, \gamma\}$. We then inject adversarial noise into $M$ by (1) uniformly sampling $5\%$ of the entries in $M$ and (2) adding random noise uniformly sampled from $[-\rho, \rho]$ to each selected entry, where the noise level $\rho$ equals 10. Denote $\hat{M}$ as the matrix after noise injection. We uniformly sample $10\%$ of the entries in $\hat{M}$ to obtain the observations, i.e., $Y_{ij}$. Hence $|\Omega|$, the number of observation is $M \times N \times 10\% = 4,000$.

In terms of algorithmic parameter setting, we set the mini-batch size $m$ to $|\Omega|/20$. The number of epoch $T$ is set to 50 for all cases, and the step size parameter $\eta_t$ is set to $1/(T \times |\Omega|/m) = 1/1000$ in all cases for all methods.

We present the comparison of listed methods in Figure 2 where we observe that 1-SFW has the best performance in terms of the Frank-Wolfe gap (a), gradient estimation accuracy (b), and the Root Mean Square Error (RMSE) between the prediction matrix and the underlying true matrix.

### A.3 Discrete Monotone Submodular Maximization with Matroid Constraint

In this section, we consider the discrete monotone submodular maximization subject to a matroid constraint via the maximizing the corresponding multilinear extension. Let $V$ be a finite set of $d$ elements and $\mathcal{I}$ be a collection of its subsets. It is proved that to maximize a discrete monotone submodular function $f : 2^V \rightarrow \mathbb{R}_+$ subject to the matroid constraint $\mathcal{M} \triangleq \{V, \mathcal{I}\}$ is equivalent to maximize its multilinear extension, defined as

$$F(x) = \sum_{S \subseteq [d]} f(S) \prod_{j \in S} |x_j| \prod_{\ell \notin S} (1 - |x_\ell|),$$

subject to the constraint $x \in \mathcal{C}$, where $\mathcal{C}$ is the base polytope of $\mathcal{M}$. Further, it is known that $F$ is monotone DR-submodular.

We now focus on a concrete recommendation problem which can be formulated as discrete monotone submodular maximization. We use $r(u, j)$ to denote user $u$’s rating for item $j \in [d]$ and set $r(u, j) = 0$ if item $j$ is not rated by user $u$. Our goal is to recommend a set of $k = 10$ items to all users such that they have the highest total rating. Two types of utility functions can be defined for such task: facility location

$$f(S) = \sum_u \max_{j \in S} r(u, j),$$
\[ f(S) = \sum_u \left( \sum_j r(u, j) \right)^{1/2}. \] (17)

Here the matroid is \( \{ V, I \} \coloneqq \{ S \subseteq V | |S| = k \} \). Two datasets are used in this experiment, Jester \( 1^1 \) and movielens \( 1M^2 \) with the results presented in Figure 3 and Figure 4 respectively. We observe that 1-SFW always achieves the highest utility after sufficient function evaluations.

B Proof of Lemma \(^2\)

**Proof.** Let \( A_t = \| \nabla F(x_t) - d_t \|^2 \). By definition, we have

\[ A_t = \| \nabla F(x_{t-1}) - d_{t-1} + \nabla F(x_t) - \nabla F(x_{t-1}) - (d_t - d_{t-1}) \|^2. \]

Note that

\[ d_t - d_{t-1} = -\rho_t d_{t-1} + \rho_t \nabla \hat{F}(x_t, z_t) + (1 - \rho_t) \tilde{\Delta}_t, \]

and define \( \Delta_t = \nabla F(x_t) - \nabla F(x_{t-1}) \), we have

\[
A_t = \| \nabla F(x_{t-1}) - d_{t-1} + \Delta_t - \rho_t \nabla \hat{F}(x_t, z_t) + \rho_t d_{t-1} \|^2 \\
= \| \nabla F(x_{t-1}) - d_{t-1} + (1 - \rho_t) (\Delta_t - \tilde{\Delta}_t) + \rho_t (\nabla F(x_t) - \nabla \hat{F}(x_t, z_t) + \rho_t (d_{t-1} - \nabla F(x_{t-1}))) \|^2 \\
= \| (1 - \rho_t) (\nabla F(x_{t-1}) - d_{t-1}) + (1 - \rho_t) (\Delta_t - \tilde{\Delta}_t) + \rho_t (\nabla F(x_t) - \nabla \hat{F}(x_t, z_t)) \|^2.
\]

Since \( \tilde{\Delta}_t \) is an unbiased estimator of \( \Delta_t \), \( \mathbb{E}[A_t] \) can be decomposed as

\[
\mathbb{E}[A_t] = \mathbb{E}\{(1 - \rho_t)^2 \| \nabla F(x_{t-1}) - d_{t-1} \|^2 \} + (1 - \rho_t)^2 \| \Delta_t - \tilde{\Delta}_t \|^2 + \rho_t^2 \| \nabla F(x_t) - \nabla \hat{F}(x_t, z_t) \|^2 \\
+ 2\rho_t(1 - \rho_t) \langle \nabla F(x_{t-1}) - d_{t-1}, \nabla F(x_t) - \nabla \hat{F}(x_t, z_t) \rangle + 2\rho_t(1 - \rho_t) \langle \Delta_t - \tilde{\Delta}_t, \nabla F(x_t) - \nabla \hat{F}(x_t, z_t) \rangle.
\] (18)

---

\(^1\)http://eigentaste.berkeley.edu/dataset/
\(^2\)https://grouplens.org/datasets/movielens/
Then we turn to upper bound the items above. First, by Lemma 1, we have

\[ \mathbb{E}[\|\tilde{\Delta}_t - \Delta_t\|^2] = \mathbb{E}[\|\nabla^2_t(x_t - x_{t-1}) - (\nabla F(x_t) - \nabla F(x_{t-1}))\|^2] \leq \mathbb{E}[\|\nabla^2_t(x_t - x_{t-1})\|^2] = \mathbb{E}[\|\nabla^2_t(\eta_{t-1}(v_{t-1} - x_{t-1}))\|^2] \leq \eta^2_{t-1}D^2 \mathbb{E}[\|\nabla^2_t\|^2] \leq \eta^2_{t-1}D^2 \bar{L}^2. \]  

(19)

By Jensen’s inequality, we have

\[ \mathbb{E}[\|\tilde{\Delta}_t - \Delta_t\|] \leq \sqrt{\mathbb{E}[\|\tilde{\Delta}_t - \Delta_t\|^2]} \leq \eta_{t-1}D\bar{L}, \]  

(20)

and

\[ \mathbb{E}[\|\nabla F(x_t) - d_t\|] = \sqrt{\mathbb{E}[\|\nabla F(x_t) - d_t\|^2]} = \sqrt{\mathbb{E}[A_t]]. \]  

(21)

Note that \( z_t \) is sampled according to \( p(z|x_t(a)) \), where \( x_t(a) = ax_t + (1-a)x_{t-1} \). Thus \( \bar{F}(x_t, z_t) \) is NOT an unbiased estimator of \( \nabla F(x_t) \) when \( a \neq 1 \), which occurs with probability 1. However, we will show that \( \bar{F}(x_t, z_t) \) is still a good estimator. Let \( \mathcal{F}_{t-1} \) be the \( \sigma \)-field generated by all the randomness before round \( t \), then by Law of Total Expectation, we have

\[ \mathbb{E}[2\rho_t(1 - \rho_t)\langle \nabla F(x_{t-1}) - d_{t-1}, \nabla F(x_t) - \nabla \bar{F}(x_t, z_t) \rangle] = \mathbb{E}[2\rho_t(1 - \rho_t)\langle \nabla F(x_{t-1}) - d_{t-1}, \nabla F(x_t) - \nabla \bar{F}(x_t, z_t) \rangle|\mathcal{F}_{t-1}, x_t(a)] \]  

(22)

where

\[ \mathbb{E}[\nabla F(x_t) - \nabla \bar{F}(x_t, z_t)|\mathcal{F}_{t-1}] = \nabla F(x_t) - \nabla F(x_t(a)) + \nabla F(x_t(a)) - \mathbb{E}[\nabla \bar{F}(x_t, z_t)|\mathcal{F}_{t-1}, x_t(a)]. \]

By Lemma 1 \( F \) is \( \bar{L} \)-smooth, thus

\[ \|\nabla F(x_t) - \nabla F(x_t(a))\| \leq \bar{L}\|x_t - x_t(a)\| = \bar{L}(1 - a)\|\eta_{t-1}(v_{t-1} - x_{t-1})\| \leq \eta_{t-1}D\bar{L}. \]
We also have

\[
\|\nabla F(x_t(a)) - E[\nabla \tilde{F}(x_t, z_t)|\mathcal{F}_{t-1}, x_t(a)]\| = \| \int [\nabla \tilde{F}(x_t(a); z) - \nabla \tilde{F}(x_t; z)]p(z; x_t(a))dz \|
\]

\[
\leq \int \|\nabla \tilde{F}(x_t(a); z) - \nabla \tilde{F}(x_t; z)\|p(z; x_t(a))dz
\]

\[
\leq \int L\|x_t(a) - x_t\|p(z; x_t(a))dz
\]

\[
\leq \eta_{t-1}DL,
\]

where the second inequality holds because of Assumption 4. Combine the analysis above with Eq. (22), we have

\[
E[2\rho_t(1 - \rho_t)[\nabla F(x_{t-1}) - d_{t-1}, \nabla F(x_t) - \nabla \tilde{F}(x_t, z_t)]]
\]

\[
\leq E[2\rho_t(1 - \rho_t)\|\nabla F(x_{t-1}) - d_{t-1}\|.E[\|\nabla F(x_t) - \nabla \tilde{F}(x_t, z_t)|\mathcal{F}_{t-1}]]
\]

\[
\leq 2\rho_t(1 - \rho_t)E[\|\nabla F(x_{t-1}) - d_{t-1}\|. (\eta_{t-1}DL + \eta_{t-1}DL)
\]

\[
\leq 2\eta_{t-1}\rho_t(1 - \rho_t)\sqrt{E[A_{t-1}]D(\bar{L} + L)}.
\]

Finally, by Assumption 3 we have \(\|\nabla F(x_t) - \nabla \tilde{F}(x_t, z_t)\| \leq 2G\). Thus

\[
\rho_t^2\|\nabla F(x_t) - \nabla \tilde{F}(x_t, z_t)\|^2 \leq 4\rho_t^2G^2,
\]

and

\[
E[2\rho_t(1 - \rho_t)[\Delta_t - \tilde{\Delta}_t, \nabla F(x_t) - \nabla \tilde{F}(x_t, z_t)]] \leq E[2\rho_t(1 - \rho_t)\|\Delta_t - \tilde{\Delta}_t\|.\|\nabla F(x_t) - \nabla \tilde{F}(x_t, z_t)\|]
\]

\[
\leq 4\eta_{t-1}\rho_t(1 - \rho_t)GD\bar{L}.
\]

Combine Eqs. (18), (19) and (23) to (25), we have

\[
E[A_t] \leq (1 - \rho_t)^2E[A_{t-1}] + (1 - \rho_t)^2\eta_{t-1}^2D^2\bar{L}^2 + \rho_t^24G^2 + 2\eta_{t-1}\rho_t(1 - \rho_t)\sqrt{E[A_{t-1}]D(\bar{L} + L)} + 4\eta_{t-1}\rho_t(1 - \rho_t)GD\bar{L}
\]

For the simplicity of analysis, we replace \(t\) by \(t + 1\), and have

\[
E[A_{t+1}]
\]

\[
\leq (1 - \rho_{t+1})^2E[A_t] + (1 - \rho_{t+1})^2\eta_{t+1}^2D^2\bar{L}^2 + \rho_{t+1}^24G^2 + 2\eta_{t+1}\rho_{t+1}(1 - \rho_{t+1})\sqrt{E[A_{t}]D(\bar{L} + L)} + 4\eta_{t+1}\rho_{t+1}(1 - \rho_{t+1})GD\bar{L}
\]

\[
\leq (1 - \frac{1}{t^\alpha})^2E[A_t] + \frac{D^2\bar{L}^2 + 4G^2 + 4GD\bar{L}}{t^{2\alpha}} + \frac{2D(\bar{L} + L)}{t^{2\alpha}}\sqrt{E[A_t]}.
\]

(26)

We claim that \(E[A_t]\leq Ct^{-\alpha}\), and prove it by induction. Before the proof, we first analyze one item in the definition of \(C\). Define \(h(\alpha) = 2 - 2^{-\alpha} - \alpha\). Since \(h'(\alpha) = 2^{-\alpha} ln(2) - 1 \leq 0\) for \(\alpha \in (0, 1]\), so \(1 = h(0) \geq h(\alpha) \geq h(1) = 1/2 > 0\), \(\forall \alpha \in (0, 1]\). As a result, \(2 \leq \frac{2}{2^{-\alpha} - \alpha} \leq 4\).

When \(t = 1\), we have

\[
E[A_1] = E[\|\nabla F(x_1) - \nabla \tilde{F}(x_1; z_1)\|^2] \leq (2G)^2 \leq \frac{2(2G + DL)^2}{2 - 2^{-\alpha} - \alpha}/1 \leq C \cdot 1^{\alpha}
\]

When \(t = 2\), since \(\rho_2 = 1\), we have

\[
E[A_2] = E[\|\nabla \tilde{F}(x_2, z_2) - \nabla F(x_2)\|^2] \leq (2G)^2 \leq \frac{2(2G + DL)^2}{2 - 2^{-\alpha} - \alpha}/2 \leq C \cdot 2^{-\alpha}.
\]
Now assume for $t \geq 2$, we have $\mathbb{E}[A_t] \leq Ct^{-\alpha}$, by Eq. (26) and the definition of $C$, we have

$$
\mathbb{E}[A_{t+1}] \leq (1 - \frac{1}{t^\alpha})^2 \cdot Ct^{-\alpha} + \frac{(2G + 2D\bar{L})^2}{t^{2\alpha}} + \frac{2D(\bar{L} + L)}{t^{(5/2)\alpha}} \sqrt{C}
$$

$$
\leq Ct^{-\alpha} - 2Ct^{-2\alpha} + Ct^{-3\alpha} + \frac{(2 - 2^{-\alpha} - \alpha)C}{2t^{2\alpha}} + \frac{C^{3/4}}{t^{(5/2)\alpha}}
$$

$$
\leq \frac{C}{t^\alpha} - \frac{2C + Ct^{-\alpha} + (2 - 2^{-\alpha} - \alpha)C/2 + (2 - 2^{-\alpha} - \alpha)/2}{t^{2\alpha}}
$$

Define $g(t) = t^{-\alpha}$, then $g(t)$ is a convex function for $\alpha \in (0, 1]$. Thus we have $g(t + 1) - g(t) \geq g'(t)$, i.e., $(t + 1)^{-\alpha} - t^{-\alpha} \geq -\alpha t^{-(\alpha + 1)}$. So we have

$$
\frac{C}{t^\alpha} - \frac{\alpha C}{t^{2\alpha}} \leq C(t^{-\alpha} - \alpha t^{-(\alpha + 1)}) \leq C(t + 1)^{-\alpha}.
$$

Combine with Eq. (27), we have $\mathbb{E}[A_{t+1}] \leq C(t + 1)^{-\alpha}$. Thus by induction, we have $\mathbb{E}[A_t] \leq Ct^{-\alpha}, \forall t \geq 1$.

## C Proof of Lemma 3

The only difference with the proof of Lemma 2 is the bound for $\mathbb{E}[\|\tilde{\Delta}_t - \Delta_t\|]$. Specifically, we have

$$
\mathbb{E}[\|\tilde{\Delta}_t - \Delta_t\|^2] = \mathbb{E}[\|\tilde{\Delta}_t - \tilde{\nabla}_t^2(x_t - x_{t-1}) + \tilde{\nabla}_t^2(x_t - x_{t-1}) - (\nabla F(x_t) - \nabla F(x_{t-1}))\|^2]
$$

$$
= \mathbb{E}[\|\tilde{\Delta}_t - \tilde{\nabla}_t^2(x_t - x_{t-1})\|^2] + \mathbb{E}[\|\tilde{\nabla}_t^2(x_t - x_{t-1}) - (\nabla F(x_t) - \nabla F(x_{t-1}))\|^2]
$$

$$
\leq [D^2L_2\delta_t(1 + \bar{F}(x_t(a), z_t))]^2 + \eta_{t-1}^2D^2L^2
$$

$$
\leq (1 + B)^2L_2^2D^4\delta_t^2 + \eta_{t-1}^2D^2L^2
$$

Then by the analysis same to the proof of Lemma 2, we have

$$
\mathbb{E}[A_{t+1}] \leq (1 - \frac{1}{t^\alpha})^2\mathbb{E}[A_t] + \frac{4(D^2\bar{L}^2 + G^2 + GD\bar{L})}{t^{2\alpha}} + \frac{4D(\bar{L} + L)}{t^{(5/2)\alpha}} \sqrt{\mathbb{E}[A_t]},
$$

and thus $\mathbb{E}[A_{t+1}] \leq C(t + 1)^{-\alpha}$, where $C = \max\left\{\frac{8(D^2\bar{L}^2 + G^2 + GD\bar{L})}{2\cdot 2^{-\alpha} - \alpha}, \left[\frac{2}{2\cdot 2^{-\alpha} - \alpha}\right]^4, 4D(\bar{L} + L)\right\}^4$.

## D Proof of Theorem 1

First, since $x_{t+1} = (1 - \eta_t)x_t + \eta_t v_t$ is a convex combination of $x_t, v_t$, and $x_t \in K, v_t \in K, \forall t$, we can prove $x_t \in K, \forall t$ by induction. So $x_{t+1} \in K$.

Then we present an auxiliary lemma.

**Lemma 4.** Under the condition of Theorem 7 in Algorithm 3, we have

$$
F(x_{t+1}) - F(x^*) \leq (1 - \eta_t)(F(x_t) - F(x^*)) + \eta_tD\|\nabla F(x_t) - d_t\| + \frac{LD^2\eta_t^2}{2}.
$$

By Jensen’s inequality and Lemma 2 with $\alpha = 1$, we have

$$
\mathbb{E}[\|\nabla F(x_t) - d_t\|] \leq \mathbb{E}[\|\nabla F(x_t) - d_t\|^2] \leq \frac{\sqrt{C}}{\sqrt{t}}.
$$
where $C = \max\{4(2G + 2L)^2, 256, [2D(L + 2L)]^2\}$. Then by Lemma 4 we have
\[
\begin{align*}
&E[F(x_{t+1}) - F(x^*)] \\
\leq & (1 - \eta_t)E[F(x_t) - F(x^*)] + \eta_t D^2 \|\nabla F(x_t) - d_t\| + \frac{\bar{L}^2 \eta_t^2}{2} \\
= & \prod_{i=1}^{T} (1 - \eta_i)E[F(x_i) - F(x^*)] + D \sum_{k=1}^{T} \eta_k \|\nabla F(x_k) - d_k\| \prod_{i=k+1}^{T} (1 - \eta_i) + \frac{\bar{L}^2 D^2}{2} \sum_{k=1}^{T} \eta_k^2 \prod_{i=k+1}^{T} (1 - \eta_i) \\
\leq & 0 + D \sum_{k=1}^{T} \frac{1}{k!} + \frac{\bar{L}^2 D^2}{2} \sum_{k=1}^{T} k^{-1}.
\end{align*}
\] (29)

Since
\[
\sum_{k=1}^{T} \frac{1}{k!} \leq \int_{0}^{T} x^{-1/2} dx = 2\sqrt{T},
\]
and
\[
\sum_{k=1}^{T} k^{-1} \leq 1 + \int_{1}^{T} x^{-1} dx = 1 + \ln T,
\]
by Eq. (29), we have
\[
E[F(x_{t+1}) - F(x^*)] \leq \frac{2\sqrt{C}D}{\sqrt{T}} + \frac{\bar{L}^2 D^2}{2T}(1 + \ln T).
\]

### E  Proof of Theorem 2

First, since $x_{t+1} = (1 - \eta_t)x_t + \eta_t v_t$ is a convex combination of $x_t, v_t,$ and $x_1 \in K, v_t \in K, \forall t,$ we can prove $x_t \in K, \forall t$ by induction. So $x_0 \in K.$

Note that if we define $v'_t = \arg\min_{v \in K} \langle v, \nabla F(x_t) \rangle,$ then $G(x_t) = \langle v'_t - x_t, -\nabla F(x_t) \rangle = -\langle v'_t - x_t, \nabla F(x_t) \rangle.$ So we have
\[
\begin{align*}
F(x_{t+1}) & \overset{(a)}{\leq} F(x_t) + \langle \nabla F(x_t), x_{t+1} - x_t \rangle + \frac{\bar{L}}{2} \|x_{t+1} - x_t\|^2 \\
& = F(x_t) + \langle \nabla F(x_t), \eta_t(v_t - x_t) \rangle + \frac{\bar{L}}{2} \|\eta_t(v_t - x_t)\|^2 \\
& \overset{(b)}{\leq} F(x_t) + \eta_t \langle \nabla F(x_t), v_t - x_t \rangle + \frac{\bar{L} \eta_t^2 D^2}{2} \\
& = F(x_t) + \eta_t \langle d_t, v_t - x_t \rangle + \eta_t \langle \nabla F(x_t) - d_t, v_t - x_t \rangle + \frac{\bar{L} \eta_t^2 D^2}{2} \\
& \overset{(c)}{\leq} F(x_t) + \eta_t \langle d_t, v'_t - x_t \rangle + \eta_t \langle \nabla F(x_t) - d_t, v_t - x_t \rangle + \frac{\bar{L} \eta_t^2 D^2}{2} \\
& = F(x_t) + \eta_t \langle \nabla F(x_t), v'_t - x_t \rangle + \eta_t \langle d_t - \nabla F(x_t), v'_t - x_t \rangle \\
& \quad + \eta_t \langle \nabla F(x_t) - d_t, v_t - x_t \rangle + \frac{\bar{L} \eta_t^2 D^2}{2} \\
& \overset{(d)}{\leq} F(x_t) - \eta_t G(x_t) + \eta_t \|\nabla F(x_t) - d_t\| \|v_t - v'_t\| + \frac{\bar{L} \eta_t^2 D^2}{2} \\
& \overset{(e)}{\leq} F(x_t) - \eta_t G(x_t) + \eta_t D \|\nabla F(x_t) - d_t\| + \frac{\bar{L} \eta_t^2 D^2}{2},
\end{align*}
\]
where we used the fact that $F$ is $\hat{L}$-smooth in inequality (a). Inequalities (b), (e) hold because of Assumption 1. Inequality (c) is due to the optimality of $v_t$, and in (d), we applied the Cauchy-Schwarz inequality.

Rearrange the inequality above, we have

$$\eta_t G(x_t) \leq F(x_t) - F(x_{t+1}) + \eta_t D \|\nabla F(x_t) - d_t\| + \frac{\hat{L} \eta_t^2 D^2}{2}. \quad (30)$$

Apply Eq. (30) recursively for $t = 1, 2, \cdots, T$, and take expectations, we attain the following inequality:

$$\sum_{t=1}^{T} \eta_t E[G(x_t)] \leq F(x_1) - F(x_{T+1}) + D \sum_{t=1}^{T} \eta_t E[\|\nabla F(x_t) - d_t\|] + \frac{\hat{L} D^2}{2} \sum_{t=1}^{T} \eta_t^2.$$

By Jensen’s inequality Lemma 2 with $\alpha = 2/3$, we have

$$E[\|\nabla F(x_t) - d_t\|] \leq \sqrt{E[\|\nabla F(x_t) - d_t\|^2]} \leq \sqrt{C_{\frac{1}{T}^{1/3}}},$$

where $C = \max\{\frac{2(2G + D \hat{L})^2}{4/3 - 2/3 - \alpha}, \left(\frac{2}{4/3 - 2/3 - \alpha}\right)^4, [2D(\hat{L} + L)]^4\}$. Since $\eta_t = T^{-2/3}$, we have

$$E[G(x_0)] = \frac{\sum_{t=1}^{T} E[G(x_t)]}{T} \leq \frac{1}{T \cdot T^{-2/3}} [F(x_1) - F(x_{T+1}) + D \sum_{t=1}^{T} T^{-2/3} \sqrt{C_{\frac{1}{T}^{1/3}}} + \frac{\hat{L} D^2}{2} \sum_{t=1}^{T} T^{-4/3}]$$

$$\leq \frac{1}{T^{1/3}} [2B + D\sqrt{C} T^{-2/3} \frac{3}{2} T^{2/3} + \frac{\hat{L} D^2}{2T^{1/3}}]$$

$$= \frac{2B + 3\sqrt{C} D/2}{T^{1/3}} + \frac{\hat{L} D^2}{2T^{2/3}},$$

where the second inequality holds because $\sum_{t=1}^{T} t^{-1/3} \leq \int_{0}^{T} x^{-1/3} dx = \frac{3}{2} T^{2/3}$.

**F  Proof of Theorem 3**

First, since $x_{t+1} = x_t + \eta_t v_t = x_t + T^{-1} v_t$, we have $x_{T+1} = \frac{\sum_{t=1}^{T} v_t}{T} \in K$. Also, because now $\|x_{t+1} - x_t\| = \|\eta_t v_t\| \leq \eta_t R$, (rather than $\eta_t D$), Lemma 2 holds with new constant $C = \max\{\frac{2(2G + R \hat{L})^2}{4/3 - 2/3 - \alpha}, \left(\frac{2}{4/3 - 2/3 - \alpha}\right)^4, [2R(\hat{L} + L)]^4\}$. Since $\alpha = 1$, we have $C = \max\{4(2G + R \hat{L})^2, 256, [2R(\hat{L} + L)]^4\}$. Then by Jensen’s inequality, we have

$$E[\|\nabla F(x_t) - d_t\|] \leq \sqrt{E[\|\nabla F(x_t) - d_t\|^2]} \leq \sqrt{C_{\frac{1}{\sqrt{t}}}}.$$
We observe that

\[
F(x_{t+1}) \overset{(a)}{=} F(x_t) + \langle \nabla F(x_t), x_{t+1} - x_t \rangle - \frac{L}{2} \| x_{t+1} - x_t \|
\]

\[
= F(x_t) + \frac{1}{T} \langle \nabla F(x_t), v_t \rangle - \frac{L}{2T^2} \| v_t \|
\]

\[
\overset{(b)}{\geq} F(x_t) + \frac{1}{T} \langle d_t, v_t \rangle + \frac{1}{T} \langle \nabla F(x_t) - d_t, v_t \rangle - \frac{LR^2}{2T^2}
\]

\[
\overset{(c)}{\geq} F(x_t) + \frac{1}{T} \langle d_t, x^* \rangle + \frac{1}{T} \langle \nabla F(x_t) - d_t, v_t \rangle - \frac{LR^2}{2T^2}
\]

\[
\overset{(d)}{=} F(x_t) + \frac{1}{T} \langle \nabla F(x_t), x^* \rangle + \frac{1}{T} \langle \nabla F(x_t) - d_t, v_t - x^* \rangle - \frac{LR^2}{2T^2}
\]

\[
\overset{(e)}{\geq} F(x_t) + \frac{F(x^*) - F(x_t)}{T} - \frac{1}{T} \langle \nabla F(x_t) - d_t, v_t - x^* \rangle - \frac{LR^2}{2T^2}
\]

\[
\overset{(f)}{\geq} F(x_t) + \frac{F(x^*) - F(x_t)}{T} - \frac{1}{T} \langle \nabla F(x_t) - d_t, v_t - x^* \rangle - \frac{LR^2}{2T^2}
\]

where inequality (a) holds because of the $\bar{L}$-smoothness of $F$, inequalities (b), (c) comes from Assumption \[.\] We used the optimality of $v_t$ in inequality (c), and applied the Cauchy-Schwarz inequality in (e). Inequality (d) is a little involved, since $F$ is monotone and concave in positive directions, we have

\[
F(x^*) - F(x_t) \leq F(x^* \lor x_t) - F(x_t) \leq \langle \nabla F(x_t), x^* \lor x_t - x_t \rangle = \langle \nabla F(x_t), (x^* - x_t) \lor 0 \rangle \leq \langle \nabla F(x_t), x^* \rangle.
\]

Taking expectations on both sides of Eq. (31),

\[
\mathbb{E}[F(x_{t+1})] \geq \mathbb{E}[F(x_t)] + \frac{F(x^*) - \mathbb{E}[F(x_t)]}{T} - \frac{2R \sqrt{C}}{T} \sqrt{t} - \frac{LR^2}{2T^2}.
\]

Or

\[
F(x^*) - \mathbb{E}[F(x_{t+1})] \leq (1 - \frac{1}{T})[F(x^*) - \mathbb{E}[F(x_t)]] + \frac{2R \sqrt{C}}{T} \sqrt{t} + \frac{LR^2}{2T^2}
\]

Apply the inequality above recursively for $t = 1, 2, \cdots, T$, we have

\[
F(x^*) - \mathbb{E}[F(x_{T+1})] \leq (1 - \frac{1}{T})^T [F(x^*) - F(x_1)] + \frac{2R \sqrt{C}}{T} \sum_{t=1}^{T} t^{-1/2} + \frac{LR^2}{2T}
\]

\[
\leq e^{-1} F(x^*) + \frac{4R \sqrt{C}}{T^{1/2}} + \frac{LR^2}{2T},
\]

where the second inequality holds since $\sum_{t=1}^{T} t^{-1/2} \leq \int_0^{T} x^{-1/2}dx = 2T^{1/2}$. Thus we have

\[
\mathbb{E}[F(x_{T+1})] \geq (1 - e^{-1}) F(x^*) - \frac{4R \sqrt{C}}{T^{1/2}} - \frac{LR^2}{2T}.
\]