# 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

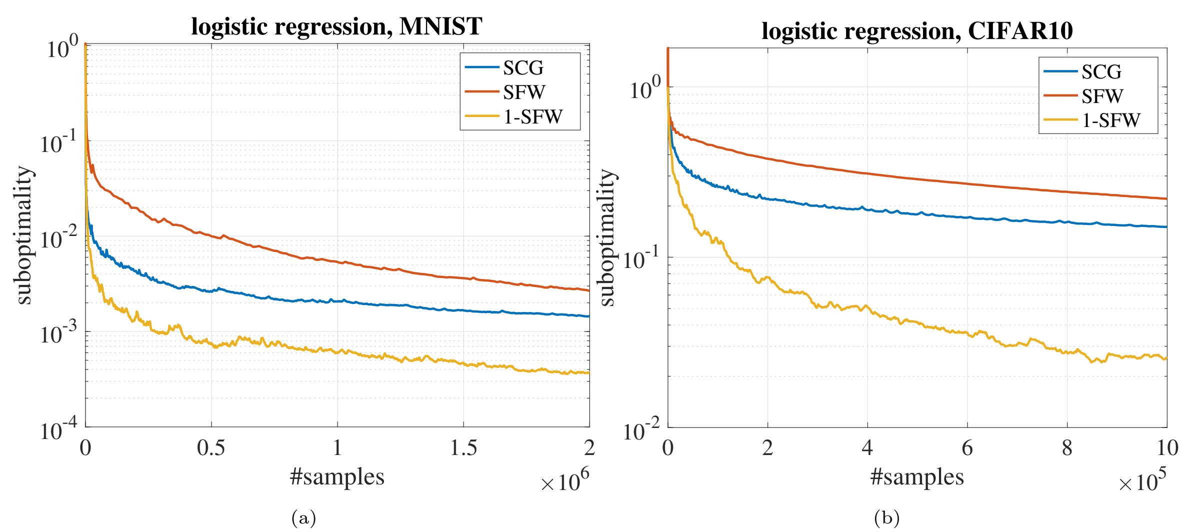



Figure 1: Logistic Regression. (a) uses digit 2 and 4 in MNIST, (b) uses cat and dog in CIFAR10.

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

$$F(\mathbf{W}) = \frac{1}{n}\sum_{i=1}^{n} \log(1 + \exp(-y_i \mathbf{W}_c^T \mathbf{a}_i)),$$

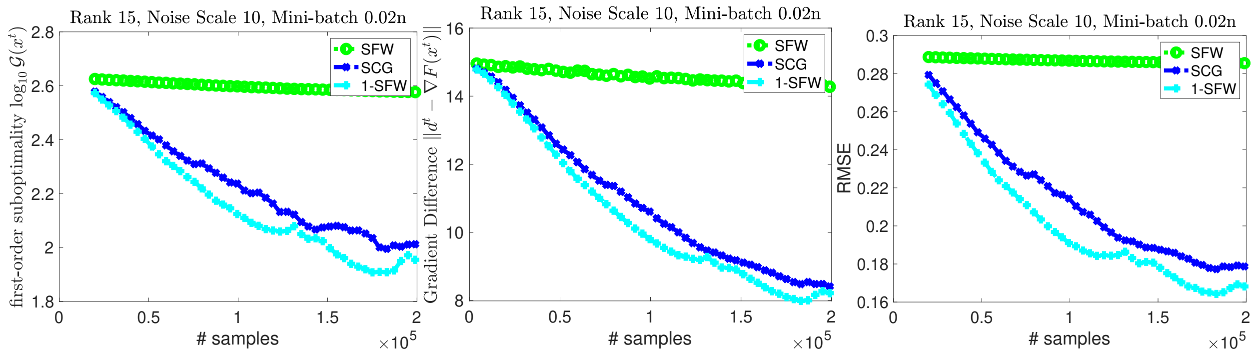
over the constraint $\mathcal{C} = \{\mathbf{W} \in \mathbb{R}^{d\times C} : \|\mathbf{W}\|_1 \le r\}$ for some constant $r \in \mathbb{R}_+$, where $\|\mathbf{W}\|_1$ is the matrix $\ell_1$ norm, *i.e.*, $\|\mathbf{W}\|_1 = \max_{1\le j\le C} \sum_{i=1}^{d} |[\mathbf{W}]_{ij}|$. We note that the loss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

Two datasets are used in our experiments: MNIST (digit 2 and 4 as positive and negative class respectively) and CIFART10 (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

$$\begin{aligned} \min_{\mathbf{X}\in\mathbb{R}^{M\times N}} \quad & \sum_{(i,j)\in\Omega} \psi(\mathbf{X}_{ij} - \mathbf{Y}_{ij}) \\ s.t. \quad & \|\mathbf{X}\|_* \le B, \end{aligned} \tag{13}$$




(a) (b) (c)

Figure 2: Matrix Recovery. (a) compares the Frank-Wolfe gap, (b) compares the accuracy of gradient estimation, (c) compares the Root Mean Square Error (RMSE) between the prediction matrix and the underlying true matrix.

where $\psi : \mathbb{R} \to \mathbb{R}$ is the potentially nonconvex empirical loss function, $\mathbf{X}_{ij}$ is the $i, j$ th element of matrix $\mathbf{X}$, and $\Omega$ is the set of observed indices in target matrix $\mathbf{Y} \in \mathbb{R}^{M \times N}$. Here we focus on a robust version of LRMR with the loss $\psi$ being:

$$\psi(z; \sigma) = 1 - \exp(-z^2/2\sigma), \tag{14}$$

where $\sigma$ is a tunable parameter. Loss (14) is less sensitive to the discrepancy $\mathbf{X}_{ij} - \mathbf{Y}_{ij}$ 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 $\mathbf{M}$ of size $200 \times 200$ and rank $\gamma = 15$. The singular values of $\mathbf{M}$ are set as $2^{[\gamma]}/2^\gamma \times 50$ and hence $\|\mathbf{M}\|_* \leq C = 100$, where $[\gamma] = \{1, \ldots, \gamma\}$. We then inject adversarial noise into $\mathbf{M}$ by (1) uniformly sampling 5% of the entries in $\mathbf{M}$ and (2) adding random noise uniformly sampled from $[-\rho, \rho]$ to each selected entry, where the noise level $\rho$ equals 10. Denote $\hat{\mathbf{M}}$ as the matrix after noise injection. We uniformly sample 10% of the entries in $\hat{\mathbf{M}}$ to obtain the observations, *i.e.*, $\mathbf{Y}_{ij}$. 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 $1/(T * |\Omega|/m) = 1/1000$ 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

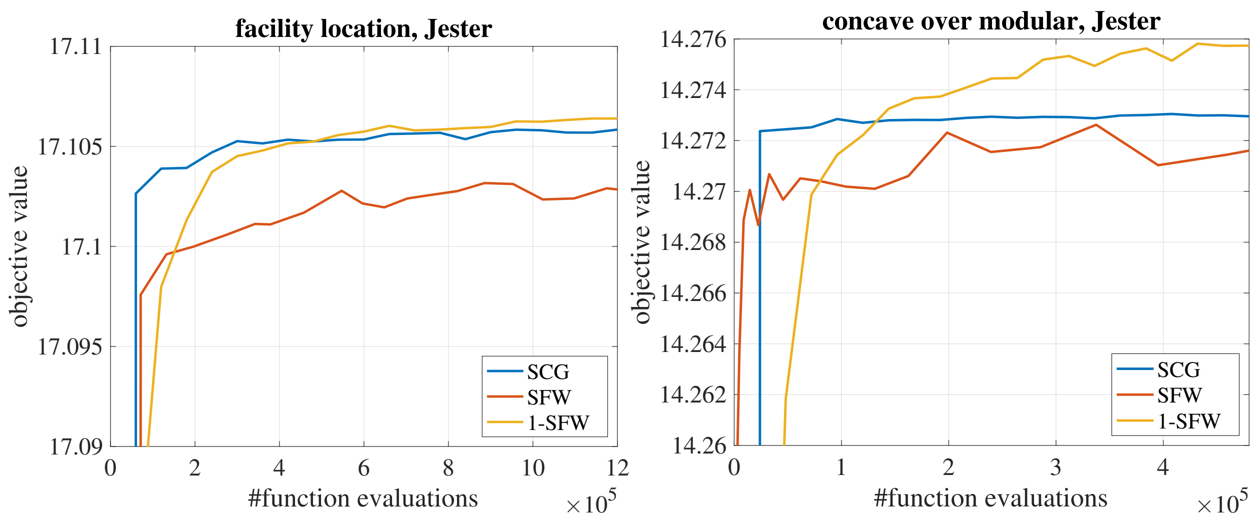
In this section, we consider the discrete monotone submodular maximization subject to a matroid constraint via the maximizing the corresponding multilinear extension. Let $V$ be a finite set of $d$ elements and $\mathcal{I}$ be a collection of its subsets. It is proved that to maximize a discrete monotone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

$$F(\mathbf{x}) = \sum_{S \subset [d]} f(S) \prod_{j \in S} [\mathbf{x}]_j \prod_{\ell \notin S} (1 - [\mathbf{x}]_\ell), \tag{15}$$

subject to the constraint $\mathbf{x} \in \mathcal{C}$, 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

$$f(S) = \sum_u \max_{j \in S} \ r(u, j), \tag{16}$$




(a) (b)

Figure 3: Submodular Maximization on Jester dataset. (a) uses the facility location utility and (b) uses the concave over modular utility.

or concave over modular

$$f(S) = \sum_u \Big(\sum_{j \in S} r(u, j)\Big)^{1/2}. \tag{17}$$

Here the matroid is $\{V, \mathcal{I} \stackrel{\text{def}}{=} \{S \subseteq V \,|\, |S| = k\}\}$. 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 $A_t = \|\nabla F(\mathbf{x}_t) - \mathbf{d}_t\|^2$. By definition, we have

$$A_t = \|\nabla F(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1} + \nabla F(\mathbf{x}_t) - \nabla F(\mathbf{x}_{t-1}) - (\mathbf{d}_t - \mathbf{d}_{t-1})\|^2.$$

Note that

$$\mathbf{d}_t - \mathbf{d}_{t-1} = -\rho_t \mathbf{d}_{t-1} + \rho_t \nabla \tilde{F}(\mathbf{x}_t, \mathbf{z}_t) + (1 - \rho_t)\tilde{\Delta}_t,$$

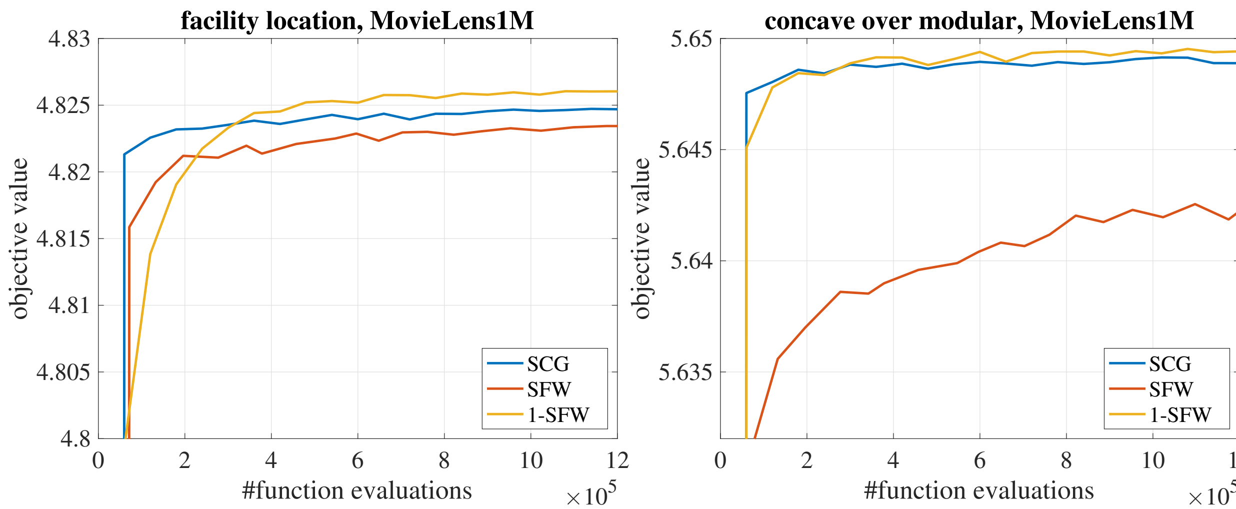
and define $\Delta_t = \nabla F(\mathbf{x}_t) - \nabla F(\mathbf{x}_{t-1})$, we have

$$\begin{aligned} A_t &= \|\nabla F(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1} + \Delta_t - (1 - \rho_t)\tilde{\Delta}_t - \rho_t \nabla \tilde{F}(\mathbf{x}_t, \mathbf{z}_t) + \rho_t \mathbf{d}_{t-1}\|^2 \\ &= \|\nabla F(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1} + (1 - \rho_t)(\Delta_t - \tilde{\Delta}_t) + \rho_t(\nabla F(\mathbf{x}_t) - \nabla \tilde{F}(\mathbf{x}_t, \mathbf{z}_t) + \rho_t(\mathbf{d}_{t-1} - \nabla F(\mathbf{x}_{t-1})))\|^2 \\ &= \|(1 - \rho_t)(\nabla F(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}) + (1 - \rho_t)(\Delta_t - \tilde{\Delta}_t) + \rho_t(\nabla F(\mathbf{x}_t) - \nabla \tilde{F}(\mathbf{x}_t, \mathbf{z}_t))\|^2 \end{aligned}$$

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

$$\begin{aligned} \mathbb{E}[A_t] = \mathbb{E}\{&(1 - \rho_t)^2\|\nabla F(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}\|^2 + (1 - \rho_t)^2\|\Delta_t - \tilde{\Delta}_t\|^2 + \rho_t^2\|\nabla F(\mathbf{x}_t) - \nabla \tilde{F}(\mathbf{x}_t, \mathbf{z}_t)\|^2 \\ &+ 2\rho_t(1 - \rho_t)\langle \nabla F(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}, \nabla F(\mathbf{x}_t) - \nabla \tilde{F}(\mathbf{x}_t, \mathbf{z}_t)\rangle + 2\rho_t(1 - \rho_t)\langle \Delta_t - \tilde{\Delta}_t, \nabla F(\mathbf{x}_t) - \nabla \tilde{F}(\mathbf{x}_t, \mathbf{z}_t)\rangle\}. \end{aligned} \tag{18}$$

[1] http://eigentaste.berkeley.edu/dataset/
[2] https://grouplens.org/datasets/movielens/




(a)

(b)

Figure 4: Submodular Maximization on Movielens dataset. (a) uses the facility location utility and (b) uses the concave over modular utility.

Then we turn to upper bound the items above. First, by Lemma 1, we have

$$\begin{aligned}\mathbb{E}[\|\tilde{\Delta}_t - \Delta_t\|^2] &= \mathbb{E}[\|\tilde{\nabla}_t^2(\mathbf{x}_t - \mathbf{x}_{t-1}) - (\nabla F(\mathbf{x}_t) - \nabla F(\mathbf{x}_{t-1}))]\|^2] \\ &\le \mathbb{E}[\|\tilde{\nabla}_t^2(\mathbf{x}_t - \mathbf{x}_{t-1})\|^2] \\ &= \mathbb{E}[\|\tilde{\nabla}_t^2(\eta_{t-1}(\mathbf{v}_{t-1} - \mathbf{x}_{t-1}))\|^2] \\ &\le \eta_{t-1}^2 D^2 \mathbb{E}[\|\tilde{\nabla}_t^2\|^2] \\ &\le \eta_{t-1}^2 D^2 \bar{L}^2.\end{aligned} \tag{19}$$

By Jensen's inequality, we have

$$\mathbb{E}[\|\tilde{\Delta}_t - \Delta_t\|] \le \sqrt{\mathbb{E}[\|\tilde{\Delta}_t - \Delta_t\|^2]} \le \eta_{t-1} D\bar{L}, \tag{20}$$

and

$$\mathbb{E}[\|\nabla F(\mathbf{x}_t) - \mathbf{d}_t\|] = \sqrt{\mathbb{E}[\|\nabla F(\mathbf{x}_t) - \mathbf{d}_t\|^2]} = \sqrt{\mathbb{E}[A_t]}. \tag{21}$$

Note that $\mathbf{z}_t$ is sampled according to $p(\mathbf{z}; \mathbf{x}_t(a))$, where $\mathbf{x}_t(a) = a\mathbf{x}_t + (1-a)\mathbf{x}_{t-1}$. Thus $\nabla\tilde{F}(\mathbf{x}_t, \mathbf{z}_t)$ is NOT an unbiased estimator of $\nabla F(\mathbf{x}_t)$ when $a \neq 1$, which occurs with probability 1. However, we will show that $\nabla\tilde{F}(\mathbf{x}_t, \mathbf{z}_t)$ 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

$$\begin{aligned}&\mathbb{E}[2\rho_t(1-\rho_t)\langle\nabla F(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}, \nabla F(\mathbf{x}_t) - \nabla\tilde{F}(\mathbf{x}_t, \mathbf{z}_t)\rangle] \\ =&\mathbb{E}[\mathbb{E}[2\rho_t(1-\rho_t)\langle\nabla F(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}, \nabla F(\mathbf{x}_t) - \nabla\tilde{F}(\mathbf{x}_t, \mathbf{z}_t)\rangle|\mathcal{F}_{t-1}, \mathbf{x}_t(a)]] \\ =&\mathbb{E}[2\rho_t(1-\rho_t)\langle\nabla F(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}, \mathbb{E}[\nabla F(\mathbf{x}_t) - \nabla\tilde{F}(\mathbf{x}_t, \mathbf{z}_t)|\mathcal{F}_{t-1}, \mathbf{x}_t(a)]\rangle],\end{aligned} \tag{22}$$

where

$$\mathbb{E}[\nabla F(\mathbf{x}_t) - \nabla\tilde{F}(\mathbf{x}_t, \mathbf{z}_t)|\mathcal{F}_{t-1}]\rangle] = \nabla F(\mathbf{x}_t) - \nabla F(\mathbf{x}_t(a)) + \nabla F(\mathbf{x}_t(a)) - \mathbb{E}[\nabla\tilde{F}(\mathbf{x}_t, \mathbf{z}_t)|\mathcal{F}_{t-1}, \mathbf{x}_t(a)].$$

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

$$\|\nabla F(\mathbf{x}_t) - \nabla F(\mathbf{x}_t(a))\| \le \bar{L}\|\mathbf{x}_t - \mathbf{x}_t(a)\| = \bar{L}(1-a)\|\eta_{t-1}(\mathbf{v}_{t-1} - \mathbf{x}_{t-1})\| \le \eta_{t-1}D\bar{L}.$$

We also have

$$\begin{aligned}\|\nabla F(\mathbf{x}_t(a)) - \mathbb{E}[\nabla\tilde{F}(\mathbf{x}_t,\mathbf{z}_t)|\mathcal{F}_{t-1},\mathbf{x}_t(a)]\| &= \|\int[\nabla\tilde{F}(\mathbf{x}_t(a);\mathbf{z}) - \nabla\tilde{F}(\mathbf{x}_t;\mathbf{z})]p(\mathbf{z};\mathbf{x}_t(a))\mathrm{d}\mathbf{z}\| \\ &\le \int\|\nabla\tilde{F}(\mathbf{x}_t(a);\mathbf{z}) - \nabla\tilde{F}(\mathbf{x}_t;\mathbf{z})\|p(\mathbf{z};\mathbf{x}_t(a))\mathrm{d}\mathbf{z} \\ &\le \int L\|\mathbf{x}_t(a) - \mathbf{x}_t\|p(\mathbf{z};\mathbf{x}_t(a))\mathrm{d}\mathbf{z} \\ &\le \eta_{t-1}DL,\end{aligned}$$

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

$$\begin{aligned}&\mathbb{E}[2\rho_t(1-\rho_t)\langle\nabla F(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}, \nabla F(\mathbf{x}_t) - \nabla\tilde{F}(\mathbf{x}_t,\mathbf{z}_t)\rangle] \\ \le&\mathbb{E}[2\rho_t(1-\rho_t)\|\nabla F(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}\|\cdot\|\mathbb{E}[\nabla F(\mathbf{x}_t) - \nabla\tilde{F}(\mathbf{x}_t,\mathbf{z}_t)|\mathcal{F}_{t-1}]\|] \\ \le&2\rho_t(1-\rho_t)\mathbb{E}[\|\nabla F(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}\|]\cdot(\eta_{t-1}D\bar{L} + \eta_{t-1}DL) \\ \le&2\eta_{t-1}\rho_t(1-\rho_t)\sqrt{\mathbb{E}[A_{t-1}]}D(\bar{L}+L).\end{aligned} \tag{23}$$

Finally, by Assumption 3, we have $\|\nabla F(\mathbf{x}_t) - \nabla\tilde{F}(\mathbf{x}_t,\mathbf{z}_t)\|\le 2G$. Thus

$$\rho_t^2\|\nabla F(\mathbf{x}_t) - \nabla\tilde{F}(\mathbf{x}_t,\mathbf{z}_t)\|^2\le 4\rho_t^2G^2, \tag{24}$$

and

$$\begin{aligned}\mathbb{E}[2\rho_t(1-\rho_t)\langle\Delta_t - \tilde{\Delta}_t, \nabla F(\mathbf{x}_t) - \nabla\tilde{F}(\mathbf{x}_t,\mathbf{z}_t)\rangle] &\le \mathbb{E}[2\rho_t(1-\rho_t)\|\Delta_t - \tilde{\Delta}_t\|\cdot\|\nabla F(\mathbf{x}_t) - \nabla\tilde{F}(\mathbf{x}_t,\mathbf{z}_t)]\| \\ &\le 4\eta_{t-1}\rho_t(1-\rho_t)GD\bar{L}.\end{aligned} \tag{25}$$

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

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

For the simplicity of analysis, we replace $t$ by $t+1$, and have

$$\begin{aligned}&\mathbb{E}[A_{t+1}] \\ \le&(1-\rho_{t+1})^2\mathbb{E}[A_t] + (1-\rho_{t+1})^2\eta_t^2D^2\bar{L}^2 + \rho_{t+1}^2 4G^2 + 2\eta_t\rho_{t+1}(1-\rho_{t+1})\sqrt{\mathbb{E}[A_t]}D(\bar{L}+L) + 4\eta_t\rho_{t+1}(1-\rho_{t+1})GD\bar{L} \\ \le&(1-\frac{1}{t^\alpha})^2\mathbb{E}[A_t] + \frac{D^2\bar{L}^2+4G^2+4GD\bar{L}}{t^{2\alpha}} + \frac{2D(\bar{L}+L)}{t^{2\alpha}}\sqrt{\mathbb{E}[A_t]}.\end{aligned} \tag{26}$$

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

When $t = 1$, we have

$$\mathbb{E}[A_1] = \mathbb{E}[\|\nabla F(\mathbf{x}_1) - \nabla\tilde{F}(\mathbf{x}_1;\mathbf{z}_1)\|^2] \le (2G)^2 \le \frac{2(2G+D\bar{L})^2}{2-2^{-\alpha}-\alpha}/1 \le C\cdot 1^{-\alpha}$$

When $t = 2$, since $\rho_2 = 1$, we have

$$\mathbb{E}[A_2] = \mathbb{E}[\|\nabla\tilde{F}(\mathbf{x}_2,\mathbf{z}_2) - \nabla F(\mathbf{x}_2)\|^2] \le (2G)^2 \le \frac{2(2G+D\bar{L})^2}{2-2^{-\alpha}-\alpha}/2 \le C\cdot 2^{-\alpha}.$$

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

$$\begin{aligned}\mathbb{E}[A_{t+1}] &\leq (1-\frac{1}{t^\alpha})^2 \cdot Ct^{-\alpha} + \frac{(2G+D\bar{L})^2}{t^{2\alpha}} + \frac{2D(\bar{L}+L)}{t^{(5/2)\alpha}}\sqrt{C} \\ &\leq Ct^{-\alpha} - 2Ct^{-2\alpha} + Ct^{-3\alpha} + \frac{(2-2^{-\alpha}-\alpha)C}{2t^{2\alpha}} + \frac{C^{3/4}}{t^{(5/2)\alpha}} \\ &\leq \frac{C}{t^\alpha} + \frac{-2C + Ct^{-\alpha} + (2-2^{-\alpha}-\alpha)C/2 + t^{-\alpha/2}C/C^{1/4}}{t^{2\alpha}} \\ &\leq \frac{C}{t^\alpha} + \frac{C[-2+2^{-\alpha}+(2-2^{-\alpha}-\alpha)/2+(2-2^{-\alpha}-\alpha)/2]}{t^{2\alpha}} \\ &\leq \frac{C}{t^\alpha} - \frac{\alpha C}{t^{2\alpha}}.\end{aligned} \tag{27}$$

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

$$\frac{C}{t^\alpha} - \frac{\alpha C}{t^{2\alpha}} \leq C(t^{-\alpha} - \alpha t^{-(1+\alpha)}) \leq C(t+1)^{-\alpha}. \tag{28}$$

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

$$\begin{aligned}\mathbb{E}[\|\tilde{\Delta}_t - \Delta_t\|^2] &= \mathbb{E}[\|\tilde{\Delta}_t - \tilde{\nabla}_t^2(\mathbf{x}_t - \mathbf{x}_{t-1}) + \tilde{\nabla}_t^2(\mathbf{x}_t - \mathbf{x}_{t-1}) - (\nabla F(\mathbf{x}_t) - \nabla F(\mathbf{x}_{t-1}))]\|^2] \\ &= \mathbb{E}[\|\tilde{\Delta}_t - \tilde{\nabla}_t^2(\mathbf{x}_t - \mathbf{x}_{t-1})\|^2] + \mathbb{E}[\|\tilde{\nabla}_t^2(\mathbf{x}_t - \mathbf{x}_{t-1}) - (\nabla F(\mathbf{x}_t) - \nabla F(\mathbf{x}_{t-1}))\|^2] \\ &\leq [D^2 L_2 \delta_t (1 + \tilde{F}(\mathbf{x}_t(a), \mathbf{z}_t))]^2 + \eta_{t-1}^2 D^2 \bar{L}^2 \\ &\leq (1+B)^2 L_2^2 D^4 \delta_t^2 + \eta_{t-1}^2 D^2 \bar{L}^2 \\ &\leq 4\eta_{t-1}^2 D^2 \bar{L}^2.\end{aligned}$$

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

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

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

# D Proof of Theorem 1

First, since $\mathbf{x}_{t+1} = (1-\eta_t)\mathbf{x}_t + \eta_t \mathbf{v}_t$ is a convex combination of $\mathbf{x}_t, \mathbf{v}_t$, and $\mathbf{x}_1 \in \mathcal{K}, \mathbf{v}_t \in \mathcal{K}, \forall\, t$, we can prove $\mathbf{x}_t \in \mathcal{K}, \forall\, t$ by induction. So $\mathbf{x}_{T+1} \in \mathcal{K}$.

Then we present an auxiliary lemma.

**Lemma 4.** *Under the condition of Theorem 1, in Algorithm 1, we have*

$$F(\mathbf{x}_{t+1}) - F(\mathbf{x}^*) \leq (1-\eta_t)(F(\mathbf{x}_t) - F(\mathbf{x}^*)) + \eta_t D\|\nabla F(\mathbf{x}_t) - \mathbf{d}_t\| + \frac{\bar{L}D^2\eta_t^2}{2}.$$

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

$$\mathbb{E}[\|\nabla F(\mathbf{x}_t) - \mathbf{d}_t\|] \leq \sqrt{\mathbb{E}[\|\nabla F(\mathbf{x}_t) - \mathbf{d}_t\|^2]} \leq \frac{\sqrt{C}}{\sqrt{t}},$$

where $C = \max\{4(2G + D\bar{L})^2, 256, [2D(\bar{L} + L)]^4\}$. Then by Lemma 4, we have

$$
\begin{aligned}
&\mathbb{E}[F(\mathbf{x}_{T+1}) - F(\mathbf{x}^*)] \\
\leq &(1-\eta_T)\mathbb{E}[F(\mathbf{x}_T) - F(\mathbf{x}^*)] + \eta_T D\mathbb{E}[\|\nabla F(\mathbf{x}_T) - \mathbf{d}_T\|] + \frac{\bar{L}D^2\eta_T^2}{2} \\
= &\prod_{i=1}^{T}(1-\eta_i)\mathbb{E}[F(\mathbf{x}_1) - F(\mathbf{x}^*)] + D\sum_{k=1}^{T}\eta_k\mathbb{E}[\|\nabla F(\mathbf{x}_k) - \mathbf{d}_k\|]\prod_{i=k+1}^{T}(1-\eta_i) + \frac{\bar{L}D^2}{2}\sum_{k=1}^{T}\eta_k^2\prod_{i=k+1}^{T}(1-\eta_i) \\
\leq &\,0 + D\sum_{k=1}^{T}k^{-1}\frac{\sqrt{C}}{\sqrt{k}}\prod_{i=k+1}^{T}\frac{i-1}{i} + \frac{\bar{L}D^2}{2}\sum_{k=1}^{T}k^{-2}\prod_{i=k+1}^{T}\frac{i-1}{i} \\
= &\frac{\sqrt{C}D}{T}\sum_{k=1}^{T}\frac{1}{\sqrt{k}} + \frac{\bar{L}D^2}{2T}\sum_{k=1}^{T}k^{-1}.
\end{aligned}
\tag{29}
$$

Since

$$\sum_{k=1}^{T}\frac{1}{\sqrt{k}} \leq \int_0^T x^{-1/2}\mathrm{d}x = 2\sqrt{T},$$

and

$$\sum_{k=1}^{T}k^{-1} \leq 1 + \int_1^T x^{-1}\mathrm{d}x = 1 + \ln T,$$

by Eq. (29), we have

$$\mathbb{E}[F(\mathbf{x}_{T+1}) - F(\mathbf{x}^*)] \leq \frac{2\sqrt{C}D}{\sqrt{T}} + \frac{\bar{L}D^2}{2T}(1 + \ln T).$$

## E Proof of Theorem 2

First, since $\mathbf{x}_{t+1} = (1-\eta_t)\mathbf{x}_t + \eta_t\mathbf{v}_t$ is a convex combination of $\mathbf{x}_t, \mathbf{v}_t$, and $\mathbf{x}_1 \in \mathcal{K}, v_t \in \mathcal{K}, \forall\, t$, we can prove $\mathbf{x}_t \in \mathcal{K}, \forall\, t$ by induction. So $\mathbf{x}_o \in \mathcal{K}$.

Note that if we define $\mathbf{v}_t' = \arg\min_{\mathbf{v}\in\mathcal{K}}\langle \mathbf{v}, \nabla F(\mathbf{x}_t)\rangle$, then $\mathcal{G}(\mathbf{x}_t) = \langle \mathbf{v}_t' - \mathbf{x}_t, -\nabla F(\mathbf{x}_t)\rangle = -\langle \mathbf{v}_t' - \mathbf{x}_t, \nabla F(\mathbf{x}_t)\rangle$. So we have

$$
\begin{aligned}
F(\mathbf{x}_{t+1}) &\overset{(a)}{\leq} F(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}_t\rangle + \frac{\bar{L}}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \\
&= F(\mathbf{x}_t) + \langle \nabla F(\mathbf{x}_t), \eta_t(\mathbf{v}_t - \mathbf{x}_t)\rangle + \frac{\bar{L}}{2}\|\eta_t(\mathbf{v}_t - \mathbf{x}_t)\|^2 \\
&\overset{(b)}{\leq} F(\mathbf{x}_t) + \eta_t\langle \nabla F(\mathbf{x}_t), \mathbf{v}_t - \mathbf{x}_t\rangle + \frac{\bar{L}\eta_t^2 D^2}{2} \\
&= F(\mathbf{x}_t) + \eta_t\langle \mathbf{d}_t, \mathbf{v}_t - \mathbf{x}_t\rangle + \eta_t\langle \nabla F(\mathbf{x}_t) - \mathbf{d}_t, \mathbf{v}_t - \mathbf{x}_t\rangle + \frac{\bar{L}\eta_t^2 D^2}{2} \\
&\overset{(c)}{\leq} F(\mathbf{x}_t) + \eta_t\langle \mathbf{d}_t, \mathbf{v}_t' - \mathbf{x}_t\rangle + \eta_t\langle \nabla F(\mathbf{x}_t) - \mathbf{d}_t, \mathbf{v}_t - \mathbf{x}_t\rangle + \frac{\bar{L}\eta_t^2 D^2}{2} \\
&= F(\mathbf{x}_t) + \eta_t\langle \nabla F(\mathbf{x}_t), \mathbf{v}_t' - \mathbf{x}_t\rangle + \eta_t\langle \mathbf{d}_t - \nabla F(\mathbf{x}_t), \mathbf{v}_t' - \mathbf{x}_t\rangle \\
&\qquad + \eta_t\langle \nabla F(\mathbf{x}_t) - \mathbf{d}_t, \mathbf{v}_t - \mathbf{x}_t\rangle + \frac{\bar{L}\eta_t^2 D^2}{2} \\
&= F(\mathbf{x}_t) - \eta_t\mathcal{G}(\mathbf{x}_t) + \eta_t\langle \nabla F(\mathbf{x}_t) - \mathbf{d}_t, \mathbf{v}_t - \mathbf{v}_t'\rangle + \frac{\bar{L}\eta_t^2 D^2}{2} \\
&\overset{(d)}{\leq} F(\mathbf{x}_t) - \eta_t\mathcal{G}(\mathbf{x}_t) + \eta_t\|\nabla F(\mathbf{x}_t) - \mathbf{d}_t\|\|\mathbf{v}_t - \mathbf{v}_t'\| + \frac{\bar{L}\eta_t^2 D^2}{2} \\
&\overset{(e)}{\leq} F(\mathbf{x}_t) - \eta_t\mathcal{G}(\mathbf{x}_t) + \eta_t D\|\nabla F(\mathbf{x}_t) - \mathbf{d}_t\| + \frac{\bar{L}\eta_t^2 D^2}{2},
\end{aligned}
$$

where we used the fact that $F$ is $\bar{L}$-smooth in inequality (a). Inequalities (b), (e) hold because of Assumption 1. Inequality (c) is due to the optimality of $\mathbf{v}_t$, and in (d), we applied the Cauchy-Schwarz inequality.

Rearrange the inequality above, we have

$$\eta_t \mathcal{G}(\mathbf{x}_t) \leq F(\mathbf{x}_t) - F(\mathbf{x}_{t+1}) + \eta_t D\|\nabla F(\mathbf{x}_t) - \mathbf{d}_t\| + \frac{\bar{L}\eta_t^2 D^2}{2}. \tag{30}$$

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

$$\sum_{t=1}^{T} \eta_t \mathbb{E}[\mathcal{G}(\mathbf{x}_t)] \leq F(\mathbf{x}_1) - F(\mathbf{x}_{T+1}) + D\sum_{t=1}^{T} \eta_t \mathbb{E}[\|\nabla F(\mathbf{x}_t) - \mathbf{d}_t\|] + \frac{\bar{L}D^2}{2}\sum_{t=1}^{T}\eta_t^2.$$

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

$$\mathbb{E}[\|\nabla F(\mathbf{x}_t) - \mathbf{d}_t\|] \leq \sqrt{\mathbb{E}[\|\nabla F(\mathbf{x}_t) - \mathbf{d}_t\|^2]} \leq \frac{\sqrt{C}}{t^{1/3}},$$

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

$$\begin{aligned}
\mathbb{E}[\mathcal{G}(\mathbf{x}_o)] &= \frac{\sum_{t=1}^{T}\mathbb{E}[\mathcal{G}(\mathbf{x}_t)]}{T} \\
&\leq \frac{1}{T\cdot T^{-2/3}}[F(\mathbf{x}_1) - F(\mathbf{x}_{T+1}) + D\sum_{t=1}^{T} T^{-2/3}\frac{\sqrt{C}}{t^{1/3}} + \frac{\bar{L}D^2}{2}\sum_{t=1}^{T} T^{-4/3}] \\
&\leq \frac{1}{T^{1/3}}[2B + D\sqrt{C}T^{-2/3}\frac{3}{2}T^{2/3} + \frac{\bar{L}D^2}{2T^{1/3}}] \\
&= \frac{2B + 3\sqrt{C}D/2}{T^{1/3}} + \frac{\bar{L}D^2}{2T^{2/3}},
\end{aligned}$$

where the second inequality holds because $\sum_{t=1}^{T} t^{-1/3} \leq \int_0^T x^{-1/3}\mathrm{d}x = \frac{3}{2}T^{2/3}$.

# F Proof of Theorem 3

First, since $\mathbf{x}_{t+1} = \mathbf{x}_t + \eta_t \mathbf{v}_t = \mathbf{x}_t + T^{-1}\mathbf{v}_t$, we have $\mathbf{x}_{T+1} = \frac{\sum_{t=1}^{T}\mathbf{v}_t}{T} \in \mathcal{K}$. Also, because now $\|\mathbf{x}_{t+1} - \mathbf{x}_t\| = \|\eta_t \mathbf{v}_t\| \leq \eta_t R$, (rather than $\eta_t D$), Lemma 2 holds with new constant $C = \max\{\frac{2(2G+R\bar{L})^2}{2-2^{-\alpha}-\alpha}, \left(\frac{2}{2-2^{-\alpha}-\alpha}\right)^4, [2R(\bar{L}+L)]^4\}$. Since $\alpha = 1$, we have $C = \max\{4(2G+R\bar{L})^2, 256, [2R(\bar{L}+L)]^4\}$. Then by Jensen's inequality, we have

$$\mathbb{E}[\|\nabla F(\mathbf{x}_t) - \mathbf{d}_t\|] \leq \sqrt{\mathbb{E}[\|\nabla F(\mathbf{x}_t) - \mathbf{d}_t\|^2]} \leq \frac{\sqrt{C}}{\sqrt{t}}.$$

We observe that

$$
\begin{aligned}
F(\mathbf{x}_{t+1}) &\overset{(a)}{\geq} F(\mathbf{x}_t) + \langle \nabla F(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}_t \rangle - \frac{\bar{L}}{2} \|\mathbf{x}_{t+1} - \mathbf{x}_t\| \\
&= F(\mathbf{x}_t) + \frac{1}{T} \langle \nabla F(\mathbf{x}_t), \mathbf{v}_t \rangle - \frac{\bar{L}}{2T^2} \|\mathbf{v}_t\| \\
&\overset{(b)}{\geq} F(\mathbf{x}_t) + \frac{1}{T} \langle \mathbf{d}_t, \mathbf{v}_t \rangle + \frac{1}{T} \langle \nabla F(\mathbf{x}_t) - \mathbf{d}_t, \mathbf{v}_t \rangle - \frac{\bar{L}R^2}{2T^2} \\
&\overset{(c)}{\geq} F(\mathbf{x}_t) + \frac{1}{T} \langle \mathbf{d}_t, \mathbf{x}^* \rangle + \frac{1}{T} \langle \nabla F(\mathbf{x}_t) - \mathbf{d}_t, \mathbf{v}_t \rangle - \frac{\bar{L}R^2}{2T^2} \\
&= F(\mathbf{x}_t) + \frac{1}{T} \langle \nabla F(\mathbf{x}_t), \mathbf{x}^* \rangle + \frac{1}{T} \langle \nabla F(\mathbf{x}_t) - \mathbf{d}_t, \mathbf{v}_t - \mathbf{x}^* \rangle - \frac{\bar{L}R^2}{2T^2} \\
&\overset{(d)}{\geq} F(\mathbf{x}_t) + \frac{F(\mathbf{x}^*) - F(\mathbf{x}_t)}{T} - \frac{1}{T} \langle \nabla F(\mathbf{x}_t) - \mathbf{d}_t, -\mathbf{v}_t + \mathbf{x}^* \rangle - \frac{\bar{L}R^2}{2T^2} \\
&\overset{(e)}{\geq} F(\mathbf{x}_t) + \frac{F(\mathbf{x}^*) - F(\mathbf{x}_t)}{T} - \frac{1}{T} \|\nabla F(\mathbf{x}_t) - \mathbf{d}_t\| \cdot \| -\mathbf{v}_t + \mathbf{x}^* \| - \frac{\bar{L}R^2}{2T^2} \\
&\overset{(f)}{\geq} F(\mathbf{x}_t) + \frac{F(\mathbf{x}^*) - F(\mathbf{x}_t)}{T} - \frac{1}{T} 2R \|\nabla F(\mathbf{x}_t) - \mathbf{d}_t\| - \frac{\bar{L}R^2}{2T^2},
\end{aligned}
\tag{31}
$$

where inequality $(a)$ holds because of the $\bar{L}$-smoothness of $F$, inequalities $(b), (e)$ comes from Assumption 1. We used the optimality of $\mathbf{v}_t$ 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

$$
F(\mathbf{x}^*) - F(\mathbf{x}_t) \leq F(\mathbf{x}^* \vee \mathbf{x}_t) - F(\mathbf{x}_t) \leq \langle \nabla F(\mathbf{x}_t), \mathbf{x}^* \vee \mathbf{x}_t - \mathbf{x}_t \rangle = \langle \nabla F(\mathbf{x}_t), (\mathbf{x}^* - \mathbf{x}_t) \vee 0 \rangle \leq \langle \nabla F(\mathbf{x}_t), \mathbf{x}^* \rangle.
$$

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

$$
\mathbb{E}[F(\mathbf{x}_{t+1})] \geq \mathbb{E}[F(\mathbf{x}_t)] + \frac{F(\mathbf{x}^*) - \mathbb{E}[F(\mathbf{x}_t)]}{T} - \frac{2R}{T} \frac{\sqrt{C}}{\sqrt{t}} - \frac{\bar{L}R^2}{2T^2}.
$$

Or

$$
F(\mathbf{x}^*) - \mathbb{E}[F(\mathbf{x}_{t+1})] \leq (1 - \frac{1}{T})[F(\mathbf{x}^*) - \mathbb{E}[F(\mathbf{x}_t)]] + \frac{2R}{T} \frac{\sqrt{C}}{\sqrt{t}} + \frac{\bar{L}R^2}{2T^2}
$$

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

$$
\begin{aligned}
F(\mathbf{x}^*) - \mathbb{E}[F(\mathbf{x}_{T+1})] &\leq (1 - \frac{1}{T})^T [F(\mathbf{x}^*) - F(\mathbf{x}_1)] + \frac{2R\sqrt{C}}{T} \sum_{t=1}^{T} t^{-1/2} + \frac{\bar{L}R^2}{2T} \\
&\leq e^{-1} F(\mathbf{x}^*) + \frac{4R\sqrt{C}}{T^{1/2}} + \frac{\bar{L}R^2}{2T},
\end{aligned}
$$

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

$$
\mathbb{E}[F(\mathbf{x}_{T+1})] \geq (1 - e^{-1}) F(\mathbf{x}^*) - \frac{4R\sqrt{C}}{T^{1/2}} - \frac{\bar{L}R^2}{2T}.
$$