Additional Experiments

In this section, we further consider online matrix completion with strongly convex loss functions, and verify the efficiency and effectiveness of our Multi-OCG+. All algorithms are implemented wtih Matlab R2016b and tested on a linux machine with 2.4GHz CPU and 768GB RAM.

The settings are mainly following our main paper, and we only make two slight changes as follows.

- The original loss function is replaced with
$$f_t(X) = \sum_{(i,j)\in \mathrm{OB}_t} |X_{ij} - M_{ij}| + \lambda \|X\|_F^2$$
which is 2λ-strongly convex, where we set $\lambda = 1e-4$.
- Instead of $T = 3000$, we equally divided the dataset used in previous experiments into $T = 300$ partitions according to its original sequence.

The first baseline is the strongly convex variant of RFTL (SC-RFTL), which updates as

$$\begin{aligned}\mathbf{x}_{t+1} =& \operatorname*{argmin}_{\mathbf{x}\in\mathcal{K}} \sum_{i=1}^{t} \left(\nabla f_i(\mathbf{x})^\top \mathbf{x} + \frac{\lambda}{2}\|\mathbf{x}-\mathbf{x}_i\|_2^2\right) \\ &+ \frac{\lambda}{2}\|\mathbf{x}-\mathbf{x}_1\|_2^2.\end{aligned}$$

The second baseline is Multi-SC-RFTL that is a projection-based variant of our Multi-OCG+ by only replacing the line 12 of Algorithm 2 with

$$\mathbf{x}_{t+1}^{\gamma} = \operatorname*{argmin}_{\mathbf{x}\in\mathcal{K}} F_{t+1}^{\gamma}(\mathbf{x}).$$

For strongly convex functions, it is not hard to verify that SC-RFTL achieves the $O(\log T)$ static regret bound, and Multi-SC-RFTL attains the same dynamic regret bound as our Multi-OCG+.

In this experiment, we set $K_\gamma = 8$ for Multi-OCG+. Moreover, for both Multi-SC-RFTL and Multi-OCG+, the parameter τ is set to be $1e-3$. Figure 2 shows the cumulative loss and runtime of each algorithm for online matrix completion with strongly convex loss functions. We find that the performance of SC-RFTL becomes worse after the environment changes, which shows that SC-RFTL cannot deal with dynamic environments. By contrast, Multi-SC-RFTL and our Multi-OCG+ can catch up with changing environments. Moreover, our Multi-OCG+ matches the performance of Multi-SC-RFTL, and is faster than it, which verifies the advantage of our algorithm in time cost.

Detailed Proofs

Proof of Lemma 1

We will utilize the property of strongly convex function, and the convergence of conditional gradient. If $f(\mathbf{x}) : \mathcal{K} \to \mathbb{R}$ is α-strongly convex and $\mathbf{x}^* = \operatorname{argmin}_{\mathbf{x}\in\mathcal{K}} f(\mathbf{x})$, combining Definition 3 with the first order optimally condition (Boyd and Vandenberghe 2004), Hazan and Kale (2012) have proved that

$$\frac{\alpha}{2}\|\mathbf{x}-\mathbf{x}^*\|_2^2 \le f(\mathbf{x}) - f(\mathbf{x}^*) \tag{10}$$

for any $\mathbf{x} \in \mathcal{K}$. The following lemma gives the convergence rate of conditional gradient.

Lemma 3 *(Derived from Theorem 1 of Jaggi (2013)) If $F(\mathbf{x}) : \mathcal{K} \to \mathbb{R}$ is a convex and α-smooth function and Assumption 1 holds, Algorithm 1 ensures*

$$F(\mathbf{x}_{\mathrm{out}}) - F(\mathbf{x}_*) \le \frac{2\alpha D^2}{K+2}.$$

where $\mathbf{x}_* \in \operatorname{argmin}_{\mathbf{x}\in\mathcal{K}} F(\mathbf{x})$.

Let $F_t^\gamma(\mathbf{x}) = \eta_\gamma \sum_{i=q_j}^{t-1} \nabla f_i(\mathbf{x}_i^\gamma)^\top \mathbf{x} + \|\mathbf{x} - \mathbf{x}_{q_j}^\gamma\|_2^2$ and $\hat{\mathbf{x}}_t^* = \operatorname{argmin}_{\mathbf{x}\in\mathcal{K}} F_t^\gamma(\mathbf{x})$ for any $t \in [q_j, q_{j+1}]$. According to the convexity of f_t, we have

$$\begin{aligned}&\sum_{t=q_j}^{q_{j+1}-1} f_t(\mathbf{x}_t^\gamma) - \sum_{t=q_j}^{q_{j+1}-1} f_t(\mathbf{x}^*) \\ \le& \sum_{t=q_j}^{q_{j+1}-1} \nabla f_t(\mathbf{x}_t^\gamma)^\top (\mathbf{x}_t^\gamma - \mathbf{x}^*) \\ =& \underbrace{\sum_{t=q_j}^{q_{j+1}-1} \nabla f_t(\mathbf{x}_t^\gamma)^\top (\mathbf{x}_t^\gamma - \hat{\mathbf{x}}_t^*)}_{:=A} \\ &+ \underbrace{\sum_{t=q_j}^{q_{j+1}-1} \nabla f_t(\mathbf{x}_t^\gamma)^\top (\hat{\mathbf{x}}_t^* - \mathbf{x}^*)}_{:=B}.\end{aligned} \tag{11}$$

Therefore, we can establish the regret bound by bounding A and B, respectively.

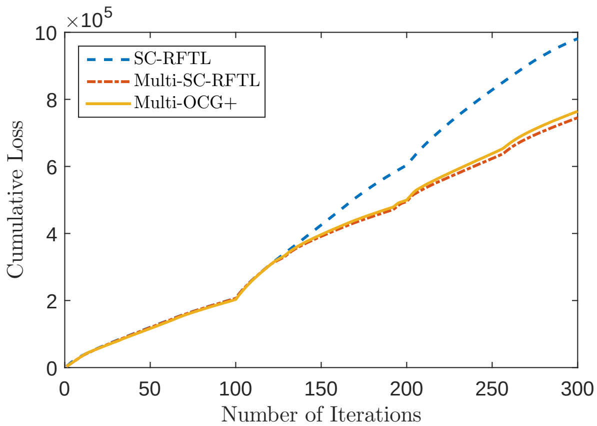
Note that for any $t \in [q_j, q_{j+1}]$, $F_t^\gamma(\mathbf{x})$ is 2-strongly convex and 2-smooth. We can bound A as

$$\begin{aligned}&\sum_{t=q_j}^{q_{j+1}-1} \nabla f_t(\mathbf{x}_t^\gamma)^\top (\mathbf{x}_t^\gamma - \hat{\mathbf{x}}_t^*) \\ \le& \sum_{t=q_j}^{q_{j+1}-1} \|\nabla f_t(\mathbf{x}_t^\gamma)\|_2 \|\mathbf{x}_t^\gamma - \hat{\mathbf{x}}_t^*\|_2 \\ \le& G \sum_{t=q_j}^{q_{j+1}-1} \|\mathbf{x}_t^\gamma - \hat{\mathbf{x}}_t^*\|_2 \\ \le& G \sum_{t=q_j}^{q_{j+1}-1} \sqrt{F_t^\gamma(\mathbf{x}_t^\gamma) - F_t^\gamma(\hat{\mathbf{x}}_t^*)} \\ \le& G\gamma \sqrt{\frac{4D^2}{\gamma+2}} \le 2GD\sqrt{\gamma}\end{aligned} \tag{12}$$

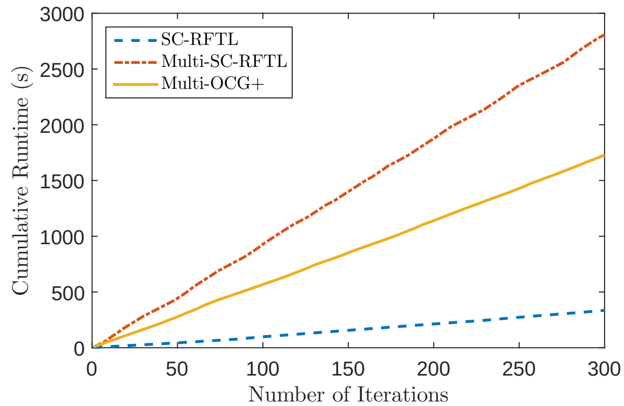
where the third inequality is due to (10) and the fourth inequality is due to Lemma 3.

To bound B, we introduce the following lemma.

Lemma 4 *(Lemma 2.3 of Shalev-Shwartz (2011)) Let $\hat{\mathbf{x}}_t^* = \operatorname*{argmin}_{\mathbf{x}\in\mathcal{K}} \left\{\sum_{i=1}^{t-1} f_i(\mathbf{x}) + \mathcal{R}(\mathbf{x})\right\}, \forall t \in [T]$. Then, $\forall \mathbf{x} \in \mathcal{K}$,*



(a) Comparison of cumulative loss



(b) Comparison of cumulative runtime

Figure 2: Experimental results for online matrix completion with strongly convex losses in dynamic environments

it holds that

$$\begin{aligned}&\sum_{t=1}^{T}\left(f_t(\hat{\mathbf{x}}_t^*)-f_t(\mathbf{x})\right)\\ \leq&\mathcal{R}(\mathbf{x})-\mathcal{R}(\hat{\mathbf{x}}_1^*)+\sum_{t=1}^{T}\left(f_t(\hat{\mathbf{x}}_t^*)-f_t(\hat{\mathbf{x}}_{t+1}^*)\right).\end{aligned}$$

Applying Lemma 4 with the linear loss functions $\{\nabla f_t(\mathbf{x}_t^\gamma)^\top\mathbf{x}\}_{t=q_j}^{q_{j+1}-1}$ and the regularizer $\mathcal{R}(\mathbf{x}) = \frac{\|\mathbf{x}-\mathbf{x}_{q_j}^\gamma\|_2^2}{\eta_\gamma}$, we can bound B as

$$\begin{aligned}&\sum_{t=q_j}^{q_{j+1}-1}\nabla f_t(\mathbf{x}_t^\gamma)^\top(\hat{\mathbf{x}}_t^*-\mathbf{x}^*)\\ \leq&\frac{\|\mathbf{x}^*-\mathbf{x}_{q_j}^\gamma\|_2^2}{\eta_\gamma}-0+\sum_{t=q_j}^{q_{j+1}-1}\nabla f_t(\mathbf{x}_t^\gamma)^\top(\hat{\mathbf{x}}_t^*-\hat{\mathbf{x}}_{t+1}^*)\\ \leq&\frac{D^2}{\eta_\gamma}+\sum_{t=q_j}^{q_{j+1}-1}\|\nabla f_t(\mathbf{x}_t^\gamma)\|_2\|\hat{\mathbf{x}}_t^*-\hat{\mathbf{x}}_{t+1}^*\|_2\\ \leq&\frac{D^2}{\eta_\gamma}+G\sum_{t=q_j}^{q_{j+1}-1}\|\hat{\mathbf{x}}_t^*-\hat{\mathbf{x}}_{t+1}^*\|_2.\end{aligned} \tag{13}$$

Moreover, because for any $t \in [q_j, q_{j+1}]$, $F_t^\gamma(\mathbf{x})$ is 2-strongly convex, we have

$$\begin{aligned}&\|\hat{\mathbf{x}}_t^*-\hat{\mathbf{x}}_{t+1}^*\|_2^2\\ \leq&F_{t+1}^\gamma(\hat{\mathbf{x}}_t^*)-F_{t+1}^\gamma(\hat{\mathbf{x}}_{t+1}^*)\\ =&F_t^\gamma(\hat{\mathbf{x}}_t^*)+\eta_\gamma\nabla f_t(\mathbf{x}_t^\gamma)^\top\hat{\mathbf{x}}_t^*-F_t^\gamma(\hat{\mathbf{x}}_{t+1}^*)\\ &-\eta_\gamma\nabla f_t(\mathbf{x}_t^\gamma)^\top\hat{\mathbf{x}}_{t+1}^*\\ =&F_t^\gamma(\hat{\mathbf{x}}_t^*)-F_t^\gamma(\hat{\mathbf{x}}_{t+1}^*)+\eta_\gamma\nabla f_t(\mathbf{x}_t^\gamma)^\top\left(\hat{\mathbf{x}}_t^*-\hat{\mathbf{x}}_{t+1}^*\right)\\ \leq&\eta_\gamma\|\nabla f_t(\mathbf{x}_t^\gamma)\|_2\|\hat{\mathbf{x}}_t^*-\hat{\mathbf{x}}_{t+1}^*\|_2\end{aligned}$$

which implies that

$$\|\hat{\mathbf{x}}_t^*-\hat{\mathbf{x}}_{t+1}^*\|_2\leq\eta_\gamma\|\nabla f_t(\mathbf{x}_t^\gamma)\|_2. \tag{14}$$

Substituting (14) in to (13), we further have

$$\begin{aligned}&\sum_{t=q_j}^{q_{j+1}-1}\nabla f_t(\mathbf{x}_t^\gamma)^\top(\hat{\mathbf{x}}_t^*-\mathbf{x}^*)\\ \leq&\frac{D^2}{\eta_\gamma}+\eta_\gamma G\sum_{t=q_j}^{q_{j+1}-1}\|\nabla f_t(\mathbf{x}_t^\gamma)\|_2\\ \leq&\frac{D^2}{\eta_\gamma}+\eta_\gamma\gamma G^2\leq 2GD\sqrt{\gamma}.\end{aligned} \tag{15}$$

Substituting (12) and (15) into (11), we complete this proof.

Proof of Lemma 2

Since OCG+ essentially performs the same steps on time intervals

$$[q_1, q_2-1], [q_2, q_3-1], \cdots, [q_r, q_{r+1}-1]$$

successively, we only need to prove this lemma for $j = 1$, i.e.,

$$\begin{aligned}&\sum_{t=1}^{q_2-1}f_t(\mathbf{x}_t^\gamma)-\sum_{t=1}^{q_2-1}f_t(\mathbf{x}^*)\\ \leq&\frac{\lambda D^2}{2}+2(G+\lambda D)D+\frac{2(G+\lambda D)^2\ln(\gamma+1)}{\lambda}.\end{aligned}$$

For any $j = 2, \cdots, r$, we can adopt the same proof steps.

Let $\tilde{f}_t(\mathbf{x}) = \nabla f_t(\mathbf{x}_t^\gamma)^\top\mathbf{x} + \frac{\lambda}{2}\|\mathbf{x}-\mathbf{x}_t^\gamma\|_2^2$ for any $t \in [1, q_2-1]$ and $\tilde{f}_0(\mathbf{x}) = \frac{\lambda}{2}\|\mathbf{x}-\mathbf{x}_1^\gamma\|_2^2$. Moreover, let $F_t^\gamma(\mathbf{x}) = \sum_{i=0}^{t-1}\tilde{f}_i(\mathbf{x})$ and $\hat{\mathbf{x}}_t^* = \operatorname{argmin}_{\mathbf{x}\in\mathcal{K}} F_t^\gamma(\mathbf{x})$ for any $t \in [1, q_2]$.

Since each $f_t(\mathbf{x})$ is λ-strongly convex, we have

$$
\begin{aligned}
&\sum_{t=1}^{q_2-1} f_t(\mathbf{x}_t^\gamma) - \sum_{t=1}^{q_2-1} f_t(\mathbf{x}^*) \\
\leq &\sum_{t=1}^{q_2-1} \left(\nabla f_t(\mathbf{x}_t^\gamma)^\top (\mathbf{x}_t^\gamma - \mathbf{x}^*) - \frac{\lambda}{2}\|\mathbf{x}_t^\gamma - \mathbf{x}^*\|_2^2\right) \\
= &\sum_{t=1}^{q_2-1} (\tilde{f}_t(\mathbf{x}_t^\gamma) - \tilde{f}_t(\mathbf{x}^*)) \\
= &\underbrace{\sum_{t=1}^{q_2-1} (\tilde{f}_t(\mathbf{x}_t^\gamma) - \tilde{f}_t(\hat{\mathbf{x}}_{t+1}^*))}_{:=A} + \underbrace{\sum_{t=1}^{q_2-1} (\tilde{f}_t(\hat{\mathbf{x}}_{t+1}^*) - \tilde{f}_t(\mathbf{x}^*))}_{:=B}.
\end{aligned} \tag{16}
$$

Therefore, we can establish the regret bound by bounding A and B, respectively.

For any $\mathbf{x}, \mathbf{y} \in \mathcal{K}$ and $t \in [1, q_2 - 1]$, we have

$$
\begin{aligned}
&\tilde{f}_t(\mathbf{x}) - \tilde{f}_t(\mathbf{y}) \\
\leq &\nabla \tilde{f}_t(\mathbf{x})^\top (\mathbf{x} - \mathbf{y}) \\
= &(\nabla f_t(\mathbf{x}_t^\gamma) + \lambda(\mathbf{x} - \mathbf{x}_t^\gamma))^\top (\mathbf{x} - \mathbf{y}) \\
\leq &\|\nabla f_t(\mathbf{x}_t^\gamma) + \lambda(\mathbf{x} - \mathbf{x}_t^\gamma)\|_2 \|\mathbf{x} - \mathbf{y}\|_2 \\
\leq &(G + \lambda D)\|\mathbf{x} - \mathbf{y}\|_2.
\end{aligned}
$$

Furthermore, for any $t \in [1, q_2 - 1]$, we have

$$
\begin{aligned}
&F_{t+1}^\gamma(\hat{\mathbf{x}}_t^*) - F_{t+1}^\gamma(\hat{\mathbf{x}}_{t+1}^*) \\
= &F_t^\gamma(\hat{\mathbf{x}}_t^*) - F_t^\gamma(\hat{\mathbf{x}}_{t+1}^*) + \tilde{f}_t(\hat{\mathbf{x}}_t^*) - \tilde{f}_t(\hat{\mathbf{x}}_{t+1}^*) \\
\leq &(G + \lambda D)\|\hat{\mathbf{x}}_t^* - \hat{\mathbf{x}}_{t+1}^*\|_2.
\end{aligned} \tag{17}
$$

Moreover, since each $F_t(\mathbf{x})$ is $t\lambda$-strongly convex, for any $t \in [1, q_2 - 1]$, we have

$$
\|\hat{\mathbf{x}}_t^* - \hat{\mathbf{x}}_{t+1}^*\|_2^2 \leq \frac{2(F_{t+1}^\gamma(\hat{\mathbf{x}}_t^*) - F_{t+1}^\gamma(\hat{\mathbf{x}}_{t+1}^*))}{(t+1)\lambda}. \tag{18}
$$

Combining (17) and (18), for any $t \in [1, q_2 - 1]$, we have

$$
\|\hat{\mathbf{x}}_t^* - \hat{\mathbf{x}}_{t+1}^*\|_2 \leq \frac{2(G + \lambda D)}{(t+1)\lambda}. \tag{19}
$$

Note that for any $t \in [1, q_2 - 1]$, $F_t^\gamma(\mathbf{x})$ is also $t\lambda$-smooth. Then, we can bound A as

$$
\begin{aligned}
&\sum_{t=1}^{q_2-1} (\tilde{f}_t(\mathbf{x}_t^\gamma) - \tilde{f}_t(\mathbf{x}_{t+1}^*)) \\
\leq &\sum_{t=1}^{q_2-1} (G + \lambda D)\|\mathbf{x}_t^\gamma - \hat{\mathbf{x}}_{t+1}^*\|_2 \\
\leq &(G + \lambda D) \sum_{t=1}^{q_2-1} \|\mathbf{x}_t^\gamma - \hat{\mathbf{x}}_t^*\|_2 \\
&+ (G + \lambda D) \sum_{t=1}^{q_2-1} \|\hat{\mathbf{x}}_t^* - \hat{\mathbf{x}}_{t+1}^*\|_2 \\
\leq &(G + \lambda D) \sum_{t=1}^{q_2-1} \sqrt{\frac{2(F_t^\gamma(\mathbf{x}_t^\gamma) - F_t^\gamma(\hat{\mathbf{x}}_t^*))}{t\lambda}} \\
&+ (G + \lambda D) \sum_{t=1}^{q_2-1} \frac{2(G + \lambda D)}{(t+1)\lambda} \\
\leq &(G + \lambda D) \sum_{t=1}^{q_2-1} \sqrt{\frac{4D^2}{\gamma^2 + 2}} \\
&+ (G + \lambda D) \sum_{t=1}^{q_2-1} \frac{2(G + \lambda D)}{(t+1)\lambda} \\
\leq &2(G + \lambda D)D + \frac{2(G + \lambda D)^2 \ln(\gamma + 1)}{\lambda}
\end{aligned} \tag{20}
$$

where the fourth inequality is due to Lemma 3.

To bound B, we introduce the following lemma.

Lemma 5 *(Lemma 6.6 of Garber and Hazan (2016)) Let $\{f_t(\mathbf{x})\}_{t=1}^T$ be a sequence of loss functions and let $\mathbf{x}_t^* \in \operatorname{argmin}_{\mathbf{x} \in \mathcal{K}} \sum_{\tau=1}^t f_t(\mathbf{x})$ for any $t \in [T]$. Then, it holds that*

$$
\sum_{t=1}^T f_t(\mathbf{x}_t^*) - \min_{\mathbf{x} \in \mathcal{K}} \sum_{t=1}^T f_t(\mathbf{x}) \leq 0.
$$

Applying Lemma 5 with the loss functions $\{\tilde{f}(\mathbf{x})\}_{t=0}^{q_2-1}$, we have

$$
\sum_{t=0}^{q_2-1} (\tilde{f}_t(\hat{\mathbf{x}}_{t+1}^*) - \tilde{f}_t(\mathbf{x}^*)) \leq 0
$$

which further implies that

$$
\begin{aligned}
B = &\sum_{t=1}^{q_2-1} (\tilde{f}_t(\hat{\mathbf{x}}_{t+1}^*) - \tilde{f}_t(\mathbf{x}^*)) \leq \tilde{f}_0(\mathbf{x}^*) - \tilde{f}_0(\hat{\mathbf{x}}_1^*) \\
= &\frac{\lambda}{2}\|\mathbf{x}^* - \mathbf{x}_1^\gamma\|_2^2 - \frac{\lambda}{2}\|\hat{\mathbf{x}}_1^* - \mathbf{x}_1^\gamma\|_2^2 \leq \frac{\lambda D^2}{2}.
\end{aligned} \tag{21}
$$

Substituting (20) and (21) into (16), we complete this proof.

Proof of Corollary 1

If $V_T \geq \sqrt{\frac{1}{T}}$, we can set $1 \leq \gamma = \left\lfloor \left(\frac{T}{V_T}\right)^{2/3} \right\rfloor \leq T$, and have

$$\begin{aligned}&\sum_{t=1}^{T} f_t(\mathbf{x}_t^{\gamma}) - \sum_{t=1}^{T} \min_{\mathbf{x}\in\mathcal{K}} f_t(\mathbf{x})\\ \leq& 8\sqrt{2}GDT^{2/3}V_T^{1/3} + 2T^{2/3}V_T^{1/3}.\end{aligned}$$

Conversely, we can simply set $\gamma = T$ and achieve

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t^{\gamma}) - \sum_{t=1}^{T} \min_{\mathbf{x}\in\mathcal{K}} f_t(\mathbf{x}) \leq 8GD\sqrt{T} + 2\sqrt{T}.$$

Proof of Corollary 2

If $V_T \geq \frac{\ln(T+1)}{T}$, we can set $1 \leq \gamma = \left\lfloor \sqrt{\frac{T\ln(T+1)}{V_T}} \right\rfloor \leq T$, and have

$$\begin{aligned}&\sum_{t=1}^{T} f_t(\mathbf{x}_t^{\gamma}) - \sum_{t=1}^{T} \min_{\mathbf{x}\in\mathcal{K}} f_t(\mathbf{x})\\ \leq& \frac{4T\sqrt{V_T}(c_1 + c_2\ln(\gamma+1))}{\sqrt{T\ln(T+1)}} + 2\sqrt{TV_T\ln(T+1)}\\ \leq&(4c_1 + 4c_2 + 2)\sqrt{TV_T\ln(T+1)}.\end{aligned}$$

Conversely, we can simply set $\gamma = T$ and achieve

$$\begin{aligned}&\sum_{t=1}^{T} f_t(\mathbf{x}_t^{\gamma}) - \sum_{t=1}^{T} \min_{\mathbf{x}\in\mathcal{K}} f_t(\mathbf{x})\\ \leq& 2(c_1 + c_2\ln(\gamma+1)) + 2\ln(T+1)\\ \leq& 2c_1 + (2c_2+2)\ln(T+1).\end{aligned}$$

Proof of Theorem 3

First, for any $\gamma \in \mathcal{H}$, we have

$$\begin{aligned}&\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} \min_{\mathbf{x}\in\mathcal{K}} f_t(\mathbf{x})\\ =&\underbrace{\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{x}_t^{\gamma})}_{:=A_\gamma}\\ &+\underbrace{\sum_{t=1}^{T} f_t(\mathbf{x}_t^{\gamma}) - \sum_{t=1}^{T} \min_{\mathbf{x}\in\mathcal{K}} f_t(\mathbf{x})}_{:=B_\gamma}.\end{aligned} \tag{22}$$

To bound A_γ, we first introduce the following lemma.

Lemma 6 *(Lemma 1 in Zhang, Lu, and Zhou (2018)) Under Assumptions 1 and 2, Algorithm 3 has*

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \min_{\gamma\in\mathcal{H}}\left(\sum_{t=1}^{T} f_t(\mathbf{x}_t^{\gamma}) + \frac{1}{\tau}\ln\frac{1}{w_1^{\gamma}}\right) \leq \frac{\tau TG^2D^2}{8}.$$

Using Lemma 6 with $\tau = \sqrt{\frac{8}{TG^2D^2}}$, for any $\gamma \in \mathcal{H}$, we have

$$\begin{aligned}&\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{x}_t^{\gamma})\\ \leq&\frac{1}{\tau}\ln\frac{1}{w_1^{\gamma}} + \frac{\tau TG^2D^2}{8}\\ \leq&\sqrt{\frac{TG^2D^2}{8}}\left(1 + \ln\frac{1}{w_1^{\gamma}}\right)\\ \leq&\sqrt{\frac{TG^2D^2}{8}}\left(1 + 2\ln N\right).\end{aligned} \tag{23}$$

Then, we need to bound B_γ. If $V_T \geq \sqrt{\frac{1}{T}}$, we define $1 \leq \gamma^* = \left\lfloor \left(\frac{T}{V_T}\right)^{2/3} \right\rfloor \leq T$. Because of

$$\mathcal{H} = \left\{\gamma_i = 2^i | i = 0, \cdots, N\right\}$$

where $N = \lfloor \log_2(T) \rfloor$, there must exist a $\gamma_i \in \mathcal{H}$ such that

$$\gamma_i \leq \gamma^* < 2\gamma_i.$$

Therefore, for γ_i, we have

$$\begin{aligned}&\sum_{t=1}^{T} f_t(\mathbf{x}_t^{\gamma_i}) - \sum_{t=1}^{T} \min_{\mathbf{x}\in\mathcal{K}} f_t(\mathbf{x})\\ \leq&\frac{8TGD}{\sqrt{\gamma_i}} + 2\gamma_i V_T\\ \leq&\frac{8\sqrt{2}TGD}{\sqrt{\gamma^*}} + 2\gamma^* V_T\\ \leq&16GDT^{2/3}V_T^{1/3} + 2T^{2/3}V_T^{1/3}\end{aligned} \tag{24}$$

where the first inequality is due to Theorem 1.

Conversely, we can simply set $\gamma^* = T$. Similarly, there must exist a $\gamma_i \in \mathcal{H}$ such that

$$\gamma_i \leq \gamma^* < 2\gamma_i.$$

Therefore, we have

$$\begin{aligned}&\sum_{t=1}^{T} f_t(\mathbf{x}_t^{\gamma_i}) - \sum_{t=1}^{T} \min_{\mathbf{x}\in\mathcal{K}} f_t(\mathbf{x})\\ \leq&\frac{8TGD}{\sqrt{\gamma_i}} + 2\gamma_i V_T\\ \leq&\frac{8\sqrt{2}GDT}{\sqrt{\gamma^*}} + 2\gamma^* V_T\\ \leq&8\sqrt{2}GD\sqrt{T} + 2\sqrt{T}\end{aligned} \tag{25}$$

Combining (22), (23), (24) and (25), we achieve

$$\begin{aligned}&\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} \min_{\mathbf{x}\in\mathcal{K}} f_t(\mathbf{x})\\ \leq&\max\left\{c_3\sqrt{T}, c_4T^{2/3}V_T^{1/3}\right\}\\ &+\sqrt{\frac{TG^2D^2}{8}}\left(1 + 2\ln N\right).\end{aligned}$$

where $c_3 = 8\sqrt{2}GD + 2$ and $c_4 = 16GD + 2$.

Proof of Theorem 4

This proof is similar as that of Theorem 3. First, it is not hard to verify that (22) still holds. So we only need to bound A_γ and B_γ in (22).

To bound A_γ, we introduce the following lemma.

Lemma 7 *If $f_t(\mathbf{x})$ is λ-strongly convex for any $t \in [T]$ and Assumptions 1 and 2 hold, Algorithm 3 with $\tau = \frac{\lambda}{G^2}$ has*

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \min_{\gamma\in\mathcal{H}} \sum_{t=1}^{T} f_t(\mathbf{x}_t^\gamma) \leq \frac{2G^2}{\lambda} \ln N$$

where $N = \lfloor \log_2(T) \rfloor$.

Using Lemma 7, for any $\gamma \in \mathcal{H}$, we have

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{x}_t^\gamma) \leq \frac{2G^2}{\lambda} \ln N. \tag{26}$$

Then, we bound B_γ by utilizing Theorem 2. If $V_T \geq \frac{\ln(T+1)}{T}$, we define $1 \leq \gamma^* = \left\lfloor \sqrt{\frac{T\ln(T+1)}{V_T}} \right\rfloor \leq T$. Because

$$\mathcal{H} = \left\{\gamma_i = 2^i | i = 0, \cdots, N\right\}$$

where $N = \lfloor \log_2(T) \rfloor$, there must exist a $\gamma_i \in \mathcal{H}$ such that

$$\gamma_i \leq \gamma^* < 2\gamma_i.$$

Therefore, for γ_i, we have

$$\begin{aligned}
&\sum_{t=1}^{T} f_t(\mathbf{x}_t^{\gamma_i}) - \sum_{t=1}^{T} \min_{\mathbf{x}\in\mathcal{K}} f_t(\mathbf{x}) \\
\leq & \frac{2T(c_1 + c_2 \ln(\gamma_i + 1))}{\gamma_i} + 2\gamma_i V_T \\
\leq & \frac{4T(c_1 + c_2 \ln(\gamma_i + 1))}{\gamma^*} + 2\gamma^* V_T \\
\leq & \frac{8T\sqrt{V_T}(c_1 + c_2 \ln(\gamma_i + 1))}{\sqrt{T\ln(T+1)}} + 2\sqrt{TV_T \ln(T+1)} \\
\leq & (8c_1 + 8c_2 + 2)\sqrt{TV_T \ln(T+1)}
\end{aligned} \tag{27}$$

where the first inequality is due to Theorem 2.

Conversely, we can simply set $\gamma^* = T$. Similarly, there must exist a $\gamma_i \in \mathcal{H}$ such that

$$\gamma_i \leq \gamma^* < 2\gamma_i.$$

Therefore, we have

$$\begin{aligned}
&\sum_{t=1}^{T} f_t(\mathbf{x}_t^{\gamma_i}) - \sum_{t=1}^{T} \min_{\mathbf{x}\in\mathcal{K}} f_t(\mathbf{x}) \\
\leq & \frac{2T(c_1 + c_2 \ln(\gamma_i + 1))}{\gamma_i} + 2\gamma_i V_T \\
\leq & \frac{4T(c_1 + c_2 \ln(\gamma_i + 1))}{\gamma^*} + 2\gamma^* V_T \\
\leq & 4(c_1 + c_2 \ln(\gamma_i + 1)) + 2TV_T \\
\leq & 4c_1 + (4c_2 + 2)\ln(T+1).
\end{aligned} \tag{28}$$

Combining (22), (26), (27) and (28), we achieve

$$\begin{aligned}
&\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} \min_{\mathbf{x}\in\mathcal{K}} f_t(\mathbf{x}) \\
\leq & \max\Big\{4c_1 + (4c_2 + 2)\ln(T+1), \\
& (8c_1 + 8c_2 + 2)\sqrt{TV_T \ln(T+1)}\Big\} + \frac{2G^2}{\lambda} \ln N.
\end{aligned}$$

Proof of Lemma 7

We will utilize the theoretical guarantee of the exponentially weighted average forecaster for exponentially concave (abbr. exp-concave) functions (Cesa-Bianchi and Lugosi 2006). So, we first introduce the standard definition of exp-concave functions (Cesa-Bianchi and Lugosi 2006).

Definition 3 *Let $f(\mathbf{x}) : \mathcal{K} \to \mathbb{R}$ be a function over $\mathcal{K}$. It is called α-exp-concave if $\exp(-\alpha f(\mathbf{x}))$ is concave over $\mathcal{K}$.*

Furthermore, we introduce the following lemma, which reveals the relationship between strongly convex and exp-concave functions.

Lemma 8 *(Lemma 2 of Zhang et al. (2018)) Suppose $f(\mathbf{x}) : \mathcal{K} \to \mathbb{R}$ is λ-strongly convex and $\|\nabla f(\mathbf{x})\|_2 \leq G$ for all $\mathbf{x} \in \mathcal{K}$. Then, $f(\mathbf{x})$ is $\frac{\lambda}{G^2}$-exp-concave.*

Note that we have set $\tau = \frac{\lambda}{G^2}$ in Algorithm 3 and each $f_t(\mathbf{x})$ is λ-strongly convex loss function for $t \in [T]$. Applying Lemma 8, for any $t \in [T]$, $f_t(\mathbf{x})$ is τ-exp-concave, which further implies that

$$e^{-\tau f_t(\mathbf{x}_t)} = e^{-\tau f_t\left(\sum_{\gamma\in\mathcal{H}} w_t^\gamma \mathbf{x}_t^\gamma\right)} \geq \sum_{\gamma\in\mathcal{H}} w_t^\gamma e^{-\tau f_t(\mathbf{x}_t^\gamma)}$$

where the last inequality is due to the concavity of $e^{-\tau f_t(\mathbf{x})}$ and Jensen's inequality.

Taking logarithm, we have

$$f_t(\mathbf{x}_t) \leq \frac{-\ln \sum_{\gamma\in\mathcal{H}} w_t^\gamma e^{-\tau f_t(\mathbf{x}_t^\gamma)}}{\tau}.$$

Then, for any $\gamma \in \mathcal{H}$, we have

$$\begin{aligned}
&f_t(\mathbf{x}_t) - f_t(\mathbf{x}_t^\gamma) \\
\leq & \frac{-\ln \sum_{\gamma\in\mathcal{H}} w_t^\gamma e^{-\tau f_t(\mathbf{x}_t^\gamma)}}{\tau} - \frac{-\ln e^{-\tau f_t(\mathbf{x}_t^\gamma)}}{\tau} \\
= & \frac{1}{\tau} \ln \frac{e^{-\tau f_t(\mathbf{x}_t^\gamma)}}{\sum_{\gamma\in\mathcal{H}} w_t^\gamma e^{-\tau f_t(\mathbf{x}_t^\gamma)}} \\
= & \frac{1}{\tau} \ln \left(\frac{1}{w_t^\gamma} \cdot \frac{w_t^\gamma e^{-\tau f_t(\mathbf{x}_t^\gamma)}}{\sum_{\gamma\in\mathcal{H}} w_t^\gamma e^{-\tau f_t(\mathbf{x}_t^\gamma)}}\right) \\
= & \frac{1}{\tau} \ln \frac{w_{t+1}^\gamma}{w_t^\gamma}.
\end{aligned}$$

Therefore, for any $\gamma \in \mathcal{H}$, we have

$$
\begin{aligned}
& \sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{x}_t^{\gamma}) \\
\leq & \frac{1}{\tau} \sum_{t=1}^{T} \left(\ln w_{t+1}^{\gamma} - \ln w_t^{\gamma} \right) \\
\leq & \frac{1}{\tau} \left(\ln w_{T+1}^{\gamma} - \ln w_1^{\gamma} \right) \\
\leq & -\frac{G^2}{\lambda} \ln w_1^{\gamma} \leq \frac{G^2}{\lambda} \ln \frac{N(N+1)}{C} \\
= & \frac{2G^2}{\lambda} \ln N
\end{aligned}
$$

where the last equality is due to $C = 1 + \frac{1}{N}$.