Supplementary Material: Projection-Free Bandit Convex Optimization over Strongly Convex Sets

Chenxu Zhang[1], Yibo Wang[2], Peng Tian[3], Xiao Cheng[3], Yuanyu Wan[1✉], and Mingli Song[4]

[1] School of Software Technology, Zhejiang University, Ningbo 315048, China
{zhangchenxu, wanyy}@zju.edu.cn
[2] National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China
wangyb@lamda.nju.edu.cn
[3] State Grid Chongqing Electric Power Research Institute, China
{tianpeng_sg, chenxiao_sg}@outlook.com
[4] College of Computer Science and Technology, Zhejiang University, Hangzhou 310058, China
brooksong@zju.edu.cn

A Proof of Theorem 1

We start this proof by introducing standard definitions for smooth functions and strongly convex functions[5].

Definition 1. *Let $f(\mathbf{x}) : \mathcal{K} \to \mathbb{R}$ be a function over $\mathcal{K}$. It is called β-smooth over $\mathcal{K}$ if for all $\mathbf{x}, \mathbf{y} \in \mathcal{K}$*

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\beta}{2}\|\mathbf{y} - \mathbf{x}\|_2^2.$$

Definition 2. *Let $f(\mathbf{x}) : \mathcal{K} \to \mathbb{R}$ be a function over $\mathcal{K}$. It is called α-strongly convex over $\mathcal{K}$ if for all $\mathbf{x}, \mathbf{y} \in \mathcal{K}$*

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\alpha}{2}\|\mathbf{y} - \mathbf{x}\|_2^2.$$

Then, let $\mathbf{x}^* \in \operatorname{argmin}_{\mathbf{x} \in \mathcal{K}} \sum_{t=1}^T f_t(\mathbf{x})$ and $\mathbf{y}^* = (1 - \delta/r)\mathbf{x}^*$. It is easy to verify that

$$\begin{aligned}
\text{Regret}(T) &= \sum_{t=1}^T f_t(\mathbf{y}_t + \delta \mathbf{u}_t) - \sum_{t=1}^T f_t(\mathbf{x}^*) \\
&\leq \sum_{t=1}^T \left(f_t(\mathbf{y}_t) + \delta G \|\mathbf{u}_t\|_2 - f_t(\mathbf{y}^*) + G \left\| \frac{\delta}{r} \mathbf{x}^* \right\|_2 \right) \qquad (1) \\
&\leq \sum_{t=1}^T (f_t(\mathbf{y}_t) - f_t(\mathbf{y}^*)) + \delta G T + \frac{\delta G R T}{r}
\end{aligned}$$

where the first inequality is due to Assumption 2, and the last inequality is due to $\|\mathbf{u}_t\|_2 = 1$ and Assumption 1.

To further bound the right side of (1), we define the δ-smooth version of $f_t(\mathbf{x})$ as

$$\widehat{f}_{t,\delta}(\mathbf{x}) = \mathbb{E}_{\mathbf{u}\sim\mathcal{B}^d}[f_t(\mathbf{x}+\delta\mathbf{u})]$$

and introduce the following lemma.

Lemma 1 (Lemma 2.6 in Hazan[20]). *Let $f(\mathbf{x}) : \mathbb{R}^d \to \mathbb{R}$ be α-strongly convex and G-Lipschitz over a convex and compact set $\mathcal{K} \subseteq \mathbb{R}^d$. Then, its δ-smooth version defined in (5) has the following properties:*

- *$\widehat{f}_\delta(\mathbf{x})$ is α-strongly convex over $\mathcal{K}_\delta$;*
- *$|\widehat{f}_\delta(\mathbf{x}) - f(\mathbf{x})| \leq \delta G$ for any $\mathbf{x} \in \mathcal{K}_\delta$;*
- *$\widehat{f}_\delta(\mathbf{x})$ is G-Lipschitz over $\mathcal{K}_\delta$.*

By applying Lemma 1, we have

$$\begin{aligned}
&\sum_{t=1}^{T} f_t\left(\mathbf{y}_t\right) - \sum_{t=1}^{T} f_t\left(\mathbf{y}^*\right) \\
\leq &\sum_{t=1}^{T}\left(\widehat{f}_{t,\delta}\left(\mathbf{y}_t\right) + \delta G\right) - \sum_{t=1}^{T}\left(\widehat{f}_{t,\delta}\left(\mathbf{y}^*\right) - \delta G\right) \\
\leq &\sum_{t=1}^{T} \widehat{f}_{t,\delta}\left(\mathbf{y}_t\right) - \sum_{t=1}^{T} \widehat{f}_{t,\delta}\left(\mathbf{y}^*\right) + 2\delta G T.
\end{aligned} \tag{2}$$

Then, by utilizing the convexity of $\widehat{f}_{t,\delta}\left(\cdot\right)$, we have

$$\begin{aligned}
&\sum_{t=1}^{T} \widehat{f}_{t,\delta}\left(\mathbf{y}_t\right) - \sum_{t=1}^{T} \widehat{f}_{t,\delta}\left(\mathbf{y}^*\right) \\
\leq &\sum_{t=1}^{T}\left\langle \nabla \widehat{f}_{t,\delta}\left(\mathbf{y}_t\right), \mathbf{y}_t - \mathbf{y}^*\right\rangle \\
\leq &\sum_{t=1}^{T}\left\langle \nabla \widehat{f}_{t,\delta}\left(\mathbf{y}_t\right), \mathbf{y}_t^* - \mathbf{y}^*\right\rangle + \sum_{t=1}^{T}\left\langle \nabla \widehat{f}_{t,\delta}\left(\mathbf{y}_t\right), \mathbf{y}_t - \mathbf{y}_t^*\right\rangle \\
\leq &\sum_{t=1}^{T}\left\langle \nabla \widehat{f}_{t,\delta}\left(\mathbf{y}_t\right), \mathbf{y}_t^* - \mathbf{y}^*\right\rangle + \sum_{t=1}^{T} G\left\|\mathbf{y}_t - \mathbf{y}_t^*\right\|_2.
\end{aligned} \tag{3}$$

Combining (1), (2) and (3), we have

$$\begin{aligned}
&\mathbb{E}[\mathrm{Regret}(T)] \\
\leq &\mathbb{E}\left[\sum_{t=1}^{T}\left\langle\nabla \widehat{f}_{t,\delta}\left(\mathbf{y}_t\right), \mathbf{y}_t^* - \mathbf{y}^*\right\rangle\right] + \mathbb{E}\left[\sum_{t=1}^{T} G\left\|\mathbf{y}_t - \mathbf{y}_t^*\right\|_2\right] \\
&+ 3\delta G T + \frac{\delta G R T}{r} \\
= &\mathbb{E}\left[\sum_{t=1}^{T}\left\langle \mathbf{g}_t, \mathbf{y}_t^* - \mathbf{y}^*\right\rangle\right] + \mathbb{E}\left[\sum_{t=1}^{T} G\left\|\mathbf{y}_t - \mathbf{y}_t^*\right\|_2\right] \\
&+ 3\delta G T + \frac{\delta G R T}{r}
\end{aligned} \tag{4}$$

where the last equality is due to Lemma 1.

To bound the term $\|\mathbf{y}_t - \mathbf{y}_t^*\|_2$ in (4), we notice that as proved by [16], if a function $f(\mathbf{x}) : \mathcal{K} \to \mathbb{R}$ is α-strongly convex, it holds that

$$\frac{\alpha}{2}\|\mathbf{x} - \mathbf{x}^*\|_2^2 \leq f(\mathbf{x}) - f(\mathbf{x}^*) \tag{5}$$

for $\mathbf{x}^* = \operatorname{argmin}_{\mathbf{x}\in\mathcal{K}} f(\mathbf{x})$ and any $\mathbf{x} \in \mathcal{K}$.It is easy to verify that $F_{t-1}(\mathbf{y}) = \eta\sum_{\tau=1}^{t-1}\langle \mathbf{g}_\tau, \mathbf{y}\rangle + \|\mathbf{y} - \mathbf{y}_1\|_2^2$ is 2-strongly convex. Therefore, according to (5), we have

$$\|\mathbf{y}_t - \mathbf{y}_t^*\|_2 \leq \sqrt{F_{t-1}\left(\mathbf{y}_t\right) - F_{t-1}\left(\mathbf{y}_t^*\right)} \leq \frac{\sqrt{C}}{(t+2)^{1/4}} \tag{6}$$

where the last inequality is due to Lemma 3.

Moreover, to bound the term $\sum_{t=1}^{T}\langle \mathbf{g}_t, \mathbf{y}_t^* - \mathbf{y}^*\rangle$, we introduce the following lemma.

Lemma 2 (Lemma 2.3 in Shalev-Shwartz [34]). *Let $\{\ell_t(\mathbf{x})\}_{t=1}^{T}$ be a sequence of loss functions over a convex set $\mathcal{K}'$ and let $\mathbf{x}_t^* = \operatorname{argmin}_{\mathbf{x}\in\mathcal{K}'}\sum_{\tau=1}^{t-1}\ell_\tau(\mathbf{x}) + \mathcal{R}(\mathbf{x})$ for any $t \in [T+1]$, where $\mathcal{R}(\mathbf{x})$ is a strongly convex function over $\mathcal{K}'$. Then, for any $\mathbf{x} \in \mathcal{K}'$, it holds that*

$$\sum_{t=1}^{T}\ell_t\left(\mathbf{x}_t^*\right) - \sum_{t=1}^{T}\ell_t(\mathbf{x}) \leq \mathcal{R}(\mathbf{x}) - \mathcal{R}\left(\mathbf{x}_1^*\right) + \sum_{t=1}^{T}\left(\ell_t\left(\mathbf{x}_t^*\right) - \ell_t\left(\mathbf{x}_{t+1}^*\right)\right).$$

By applying Lemma 2 with $\ell_t(\mathbf{x}) = \langle \mathbf{g}_t, \mathbf{x}\rangle$, $\mathcal{R}(\mathbf{x}) = \|\mathbf{x} - \mathbf{y}_1\|_2^2/\eta$, and $\mathcal{K}' = \mathcal{K}_\delta$, we have

$$\begin{aligned}
\sum_{t=1}^{T}\left\langle \mathbf{g}_t, \mathbf{y}_t^* - \mathbf{y}^*\right\rangle \leq &\frac{\left\|\mathbf{y}^* - \mathbf{y}_1\right\|_2^2}{\eta} - \frac{\left\|\mathbf{y}_1^* - \mathbf{y}_1\right\|_2^2}{\eta} + \sum_{t=1}^{T}\left\langle \mathbf{g}_t, \mathbf{y}_t^* - \mathbf{y}_{t+1}^*\right\rangle \\
\leq &\frac{4R^2}{\eta} + \sum_{t=1}^{T}\left\|\mathbf{g}_t\right\|_2\left\|\mathbf{y}_t^* - \mathbf{y}_{t+1}^*\right\|_2.
\end{aligned} \tag{7}$$

Since $F_t(\mathbf{x})$ is 2-strongly convex and 2-smooth over $\mathcal{K}_\delta$ for $t = 0, \cdots, T$. Then, from (5), for any $t \in [T]$, we have

$$\begin{aligned}\left\|\mathbf{y}_t^* - \mathbf{y}_{t+1}^*\right\|_2^2 \leq & F_t\left(\mathbf{y}_t^*\right) - F_t\left(\mathbf{y}_{t+1}^*\right) \\ = & F_{t-1}\left(\mathbf{y}_t^*\right) - F_{t-1}\left(\mathbf{y}_{t+1}^*\right) + \left\langle \eta \mathbf{g}_t, \mathbf{y}_t^* - \mathbf{y}_{t+1}^*\right\rangle \\ \leq & \eta \left\|\mathbf{g}_t\right\|_2 \left\|\mathbf{y}_t^* - \mathbf{y}_{t+1}^*\right\|_2\end{aligned}$$

which implies that

$$\left\|\mathbf{y}_t^* - \mathbf{y}_{t+1}^*\right\|_2 \leq \eta \left\|\mathbf{g}_t\right\|_2 \leq \frac{\eta d M}{\delta}. \tag{8}$$

Therefore, substituting (8) into (7), we have

$$\sum_{t=1}^{T} \left\langle \mathbf{g}_t, \mathbf{y}_t^* - \mathbf{y}^* \right\rangle \leq \frac{4R^2}{\eta} + \sum_{t=1}^{T} \eta \left\|\mathbf{g}_t\right\|_2^2 \leq \frac{4R^2}{\eta} + \frac{\eta d^2 M^2 T}{\delta^2}. \tag{9}$$

By combining (4) with (6) and (9), we finally have

$$\begin{aligned}\mathbb{E}[\text{Regret}(T)] \leq & \frac{4R^2}{\eta} + \frac{\eta d^2 M^2 T}{\delta^2} + \sum_{t=1}^{T} \frac{\sqrt{C} G}{(t+2)^{1/4}} + 3\delta G T + \frac{\delta G R T}{r} \\ \leq & \frac{4RdM(T+2)^{3/4}}{c} + \frac{RdMT^{3/4}}{c} + \frac{4\sqrt{C}G(T+2)^{3/4}}{3} \\ & + 3cGT^{3/4} + \frac{cGRT^{3/4}}{r}\end{aligned} \tag{10}$$

where the second inequality is derived by substituting the value of η and δ.

B Proof of Theorem 2

By combining (1), (2) and (3), we first have

$$\text{Regret}(T) \leq \sum_{t=1}^{T} \langle \nabla \widehat{f}_{t,\delta}(\mathbf{y}_t), \mathbf{y}_t^* - \mathbf{y}^* \rangle + \sum_{t=1}^{T} G \|\mathbf{y}_t - \mathbf{y}_t^*\|_2 + 3\delta G T + \frac{\delta G R T}{r}. \tag{11}$$

To bound $\sum_{t=1}^{T} \langle \nabla \widehat{f}_{t,\delta}(\mathbf{y}_t), \mathbf{y}_t^* - \mathbf{y}^* \rangle$, we introduce the following lemma.

Lemma 3. *Let* $\mathbf{y}_t^* = \operatorname{argmin}_{\mathbf{y} \in \mathcal{K}_\delta} F_{t-1}(\mathbf{y})$, $\mathbf{y}^* = (1 - \delta/r)\mathbf{x}^*$, *where* $\mathbf{x}^* \in \operatorname{argmin}_{\mathbf{x} \in \mathcal{K}} \sum_{t=1}^{T} f_t(\mathbf{x})$. *Under Assumption 1, 2 and 3, with probability at least* $1 - \gamma$, *Algorithm 1 ensures*

$$\sum_{t=1}^{T} \langle \nabla \widehat{f}_{t,\delta}(\mathbf{y}_t), \mathbf{y}_t^* - \mathbf{y}^* \rangle \leq 2R\left(G + \frac{dM}{\delta}\right)\sqrt{2T \ln \frac{1}{\gamma}} + \frac{4R^2}{\eta} + \frac{\eta d^2 M^2 T}{\delta^2}.$$

By combining (6), (11), and Lemma 3, with probability at least $1-\gamma$, we have

$$\begin{aligned}\mathrm{Regret}(T) \leq & 2R\left(G+\frac{dM}{\delta}\right)\sqrt{2T\ln\frac{1}{\gamma}}+\frac{4R^2}{\eta}+\frac{\eta d^2M^2T}{\delta^2}\\ &+\sum_{t=1}^{T}\frac{G\sqrt{C}}{(t+2)^{1/4}}+3\delta GT+\frac{\delta GRT}{r}.\end{aligned}$$

By substituting $\eta=\frac{cR}{dM(T+2)^{3/4}}$ and $\delta=cT^{1/4}$ into the above inequality, we have

$$\begin{aligned}\mathrm{Regret}(T) \leq & 2RG\sqrt{2\ln\frac{1}{\gamma}}T^{1/2}+\frac{2RdM}{c}\sqrt{2\ln\frac{1}{\gamma}}T^{3/4}\\ &+\frac{4RdM(T+2)^{3/4}}{c}+\frac{4\sqrt{C}G(T+2)^{3/4}}{3}\\ &+\frac{RdMT^{3/4}}{c}+3cGT^{3/4}+\frac{cGRT^{3/4}}{r}.\end{aligned}\tag{12}$$

C Proof of Lemma 2

Let $\mathbf{x},\mathbf{y}$ be any vectors in $\mathcal{K}_\delta$. According to the definition of $\mathcal{K}_\delta$, we must have

$$\frac{\mathbf{x}}{1-(\delta/r)}\in\mathcal{K},\quad \frac{\mathbf{y}}{1-(\delta/r)}\in\mathcal{K}.$$

Then, since $\mathcal{K}$ is α_K-strongly convex with respect to a norm $\|\cdot\|$, according to Definition 1, for any $\gamma\in[0,1]$ and $\mathbf{z}\in\mathbb{R}^d$ such that $\|\mathbf{z}\|=1$, it holds that

$$\gamma\frac{\mathbf{x}}{1-(\delta/r)}+(1-\gamma)\frac{\mathbf{y}}{1-(\delta/r)}+\gamma(1-\gamma)\frac{\alpha_K}{2}\left\|\frac{\mathbf{x}-\mathbf{y}}{1-(\delta/r)}\right\|^2\mathbf{z}\in\mathcal{K}.$$

By using the definition of $\mathcal{K}_\delta$ again, for any $\gamma\in[0,1]$ and $\mathbf{z}\in\mathbb{R}^d$ such that $\|\mathbf{z}\|=1$, it holds that

$$\gamma\mathbf{x}+(1-\gamma)\mathbf{y}+\gamma(1-\gamma)\frac{\alpha_K}{2(1-(\delta/r))}\|\mathbf{x}-\mathbf{y}\|^2\mathbf{z}\in\mathcal{K}_\delta.$$

Finally, combining the above result with Definition 1, $\mathcal{K}_\delta$ is $\frac{\alpha_K}{1-(\delta/r)}$-strongly convex with respect to the norm $\|\cdot\|$.

D Proof of Lemma 3

For brevity, we define $h_t=F_{t-1}(\mathbf{y}_t)-F_{t-1}(\mathbf{y}_t^*)$ for $t=1,\cdots,T$ and $h_t(\mathbf{y}_{t-1})=F_{t-1}(\mathbf{y}_{t-1})-F_{t-1}(\mathbf{y}_t^*)$ for $t=2,\cdots,T$.

For $t=1$, since $\mathbf{y}_1=\mathrm{argmin}_{\mathbf{x}\in\mathcal{K}_\delta}\|\mathbf{x}-\mathbf{y}_1\|_2^2$, we have

$$h_1=F_0\left(\mathbf{y}_1\right)-F_0\left(\mathbf{y}_1^*\right)=0\leq\frac{C}{\sqrt{1+2}}=\epsilon_1.\tag{13}$$

For any $T+1 \geq t \geq 2$, assuming $h_{t-1} \leq \epsilon_{t-1}$, we first note that

$$\begin{aligned}
h_t(\mathbf{y}_{t-1}) =& F_{t-1}(\mathbf{y}_{t-1}) - F_{t-1}(\mathbf{y}_t^*) \\
=& F_{t-2}(\mathbf{y}_{t-1}) - F_{t-2}(\mathbf{y}_t^*) + \langle \eta \mathbf{g}_{t-1}, \mathbf{y}_{t-1} - \mathbf{y}_t^* \rangle \\
\leq& F_{t-2}(\mathbf{y}_{t-1}) - F_{t-2}\left(\mathbf{y}_{t-1}^*\right) + \langle \eta \mathbf{g}_{t-1}, \mathbf{y}_{t-1} - \mathbf{y}_t^* \rangle \\
\leq& \epsilon_{t-1} + \eta \|\mathbf{g}_{t-1}\|_2 \|\mathbf{y}_{t-1} - \mathbf{y}_t^*\|_2 \\
\leq& \epsilon_{t-1} + \eta \|\mathbf{g}_{t-1}\|_2 \left\|\mathbf{y}_{t-1} - \mathbf{y}_{t-1}^*\right\|_2 + \eta \|\mathbf{g}_{t-1}\|_2 \left\|\mathbf{y}_{t-1}^* - \mathbf{y}_t^*\right\|_2 \\
\leq& \epsilon_{t-1} + \frac{\eta\sqrt{\epsilon_{t-1}}dM}{\delta} + \frac{\eta^2 d^2 M^2}{\delta^2}
\end{aligned} \tag{14}$$

where the first inequality is due to $\mathbf{y}_{t-1}^* = \operatorname{argmin}_{\mathbf{x}\in\mathcal{K}_\delta} F_{t-2}(\mathbf{x})$ and the last inequality is due to $\|\mathbf{y}_{t-1} - \mathbf{y}_{t-1}^*\|_2 \leq \sqrt{F_{t-2}(\mathbf{y}_{t-1}) - F_{t-2}(\mathbf{y}_{t-1}^*)}$ and (8).

Then, by substituting $\eta = \frac{cR}{dM(T+2)^{3/4}}$ and $\delta = cT^{-1/4}$ into (14), we have

$$\begin{aligned}
h_t(\mathbf{y}_{t-1}) &\leq \epsilon_{t-1} + \frac{R\sqrt{C}}{\sqrt{T+2}(t+1)^{1/4}} + \frac{R^2}{(T+2)} \\
&\leq \epsilon_{t-1} + \frac{\epsilon_{t-1}}{4(t+1)^{1/4}} + \frac{\epsilon_{t-1}}{16(t+1)^{1/4}} \\
&\leq \left(1 + \frac{1}{2(t+1)^{1/4}}\right)\epsilon_{t-1}
\end{aligned} \tag{15}$$

where the second inequality is due to $T \geq t-1$ and $R \leq \frac{\sqrt{C}}{4}$.

Then, to bound $h_t = F_{t-1}(\mathbf{y}_t) - F_{t-1}(\mathbf{y}_t^*)$ by ϵ_t, we further introduce the following lemma.

Lemma 4 (Derived from Lemma 1 of Garber and Hazan[15]). *Let $f(\mathbf{x}) : \mathcal{K}' \to \mathbb{R}$ be a convex and β_f-smooth function, where $\mathcal{K}'$ is $\alpha_{\mathcal{K}'}$-strongly convex with respect to the ℓ_2 norm. Moreover, let $\mathbf{x}_{\text{in}} \in \mathcal{K}$ and $\mathbf{x}_{\text{out}} = \mathbf{x}_{\text{in}} + \sigma'(\mathbf{v} - \mathbf{x}_{\text{in}})$, where $\mathbf{v} \in \operatorname{argmin}_{\mathbf{x}\in\mathcal{K}'} \langle \nabla f(\mathbf{x}_{\text{in}}), \mathbf{x}\rangle$ and $\sigma' = \operatorname{argmin}_{\sigma\in[0,1]} \langle \sigma(\mathbf{v} - \mathbf{x}_{\text{in}}), \nabla f(\mathbf{x}_{\text{in}})\rangle + \frac{\sigma^2\beta_f}{2}\|\mathbf{v} - \mathbf{x}_{\text{in}}\|_2^2$. For any $\mathbf{x}^* \in \operatorname{argmin}_{\mathbf{x}\in\mathcal{K}} f(\mathbf{x})$, we have*

$$f(\mathbf{x}_{\text{out}}) - f(\mathbf{x}^*) \leq (f(\mathbf{x}_{\text{in}}) - f(\mathbf{x}^*)) \max\left(\frac{1}{2}, 1 - \frac{\alpha_{K'}\|\nabla f(\mathbf{x}_{\text{in}})\|_2}{8\beta_f}\right).$$

We note that $F_{t-1}(\mathbf{x})$ is 2-smooth over $\mathcal{K}_\delta$ for any $t \in [T+1]$ and $\mathcal{K}_\delta$ is $\frac{\alpha_K}{1-(\delta/r)}$-strongly convex with respect to the ℓ_2 norm. By applying Lemma 4 with $f(\mathbf{x}) = F_{t-1}(\mathbf{x}), \mathcal{K}' = \mathcal{K}_\delta, \alpha_{K'} = \frac{\alpha_K}{1-(\delta/r)}$, and $\mathbf{x}_{\text{in}} = \mathbf{y}_{t-1}$, for any $t \in [T+1]$, we have $\mathbf{x}_{\text{out}} = \mathbf{y}_t$ and

$$h_t \leq h_t(\mathbf{y}_{t-1}) \max\left(\frac{1}{2}, 1 - \frac{\alpha_{K'}\|\nabla F_{t-1}(\mathbf{y}_{t-1})\|_2}{16}\right). \tag{16}$$

Because of (15), (16), and $1 + \frac{1}{2(t+1)^{1/4}} \leq \frac{3}{2}$, if $\frac{1}{2} \leq \frac{\alpha_{K'}\|\nabla F_{t-1}(\mathbf{y}_{t-1})\|_2}{16}$, it is easy to verify that

$$h_t \leq \frac{3}{4}\epsilon_{t-1} = \frac{3}{4}\frac{C}{\sqrt{t+1}} = \frac{C}{\sqrt{t+2}}\frac{3\sqrt{t+2}}{4\sqrt{t+1}} \leq \frac{C}{\sqrt{t+2}} = \epsilon_t \tag{17}$$

where the last inequality is due to $\frac{3\sqrt{t+2}}{4\sqrt{t+1}} \leq 1$ for any $t \geq 2$.

Then, if $\frac{1}{2} > \frac{\alpha_{K'}\|\nabla F_{t-1}(\mathbf{y}_{t-1})\|_2}{16}$, there exist two cases. First , if $h_t(\mathbf{y}_{t-1}) \leq \frac{3C}{4\sqrt{t+1}}$, it is easy to verify that

$$h_t \leq h_t(\mathbf{y}_{t-1}) \leq \frac{3C}{4\sqrt{t+1}} \leq \epsilon_t \tag{18}$$

where the last inequality has been proved in (17).

Second, if $h_t(\mathbf{y}_{t-1}) > \frac{3C}{4\sqrt{t+1}}$, we have

$$\begin{aligned} h_t &\leq h_t(\mathbf{y}_{t-1})\left(1 - \frac{\alpha_{K'}\|\nabla F_{t-1}(\mathbf{y}_{t-1})\|_2}{16}\right) \\ &\leq \epsilon_{t-1}\left(1 + \frac{1}{2(t+1)^{1/4}}\right)\left(1 - \frac{\alpha_{K'}\|\nabla F_{t-1}(\mathbf{y}_{t-1})\|_2}{16}\right) \end{aligned}$$

where the last inequality is due to (15). Then, we notice that as proved by Garber and Hazan[16], if a function $f(\mathbf{x}) : \mathcal{K} \to \mathbb{R}$ is α-strongly convex, it holds that

$$\|\nabla f(\mathbf{x})\|_2 \geq \sqrt{\frac{\alpha}{2}}\sqrt{f(\mathbf{x}) - f(\mathbf{x}^*)} \tag{19}$$

for $\mathbf{x}^* = \operatorname{argmin}_{\mathbf{x}\in\mathcal{K}} f(\mathbf{x})$ and any $\mathbf{x} \in \mathcal{K}$. Because of (19) and the fact that $F_{t-1}(\mathbf{y})$ is 2-strongly convex, we further have

$$\begin{aligned} h_t &\leq \epsilon_{t-1}\left(1 + \frac{1}{2(t+1)^{1/4}}\right)\left(1 - \frac{\alpha_{K'}\sqrt{h_t(\mathbf{y}_{t-1})}}{16}\right) \\ &\leq \epsilon_t\frac{\sqrt{t+2}}{\sqrt{t+1}}\left(1 + \frac{1}{2(t+1)^{1/4}}\right)\left(1 - \frac{\alpha_K\sqrt{3C}}{32(t+1)^{1/4}}\right) \end{aligned}$$

where the last inequality is due to $\alpha_{K'} = \frac{\alpha_K}{1-(\delta/r)} \geq \alpha_K$.

Since $\frac{\sqrt{t+2}}{\sqrt{t+1}} \leq 1 + \frac{1}{\sqrt{t+1}}$ for any $t \geq 0$, it is easy to verify that

$$\frac{\sqrt{t+2}}{\sqrt{t+1}}\left(1 + \frac{1}{2(t+1)^{1/4}}\right) \leq 1 + \frac{2}{(t+1)^{1/4}}$$

which further implies that

$$\begin{aligned} h_t &\leq \epsilon_t\left(1 + \frac{2}{(t+1)^{1/4}}\right)\left(1 - \frac{\alpha_K\sqrt{3C}}{32(t+1)^{1/4}}\right) \\ &\leq \epsilon_t\left(1 + \frac{2}{(t+1)^{1/4}}\right)\left(1 - \frac{2}{(t+1)^{1/4}}\right) \\ &\leq \epsilon_t \end{aligned} \tag{20}$$

where the second inequality is due to $\alpha_K\sqrt{3C} \geq 64$.

By combining (13), (17), (18) and (20), we complete the proof.

E Proof of Lemma 3

We introduce the classical Azuma's inequality [4] for martingales in the following lemma.

Lemma 5. *Suppose $D_1, \cdots, D_r$ is a martingale difference sequence and*

$$|D_j| \leq c_j$$

almost surely. Then, we have

$$\Pr\left(\sum_{j=1}^{r} D_j \geq \Delta\right) \leq \exp\left(\frac{-\Delta^2}{2\sum_{j=1}^{r} c_j^2}\right).$$

To apply Lemma 5, we define

$$D_t = \left(\nabla \widehat{f}_{t,\delta}(\mathbf{y}_t) - \mathbf{g}_t(\mathbf{x}_t)\right)(\mathbf{y}_t^* - \mathbf{y}^*) \tag{21}$$

where $\mathbf{y}_t^* = \operatorname{argmin}_{\mathbf{x}\in\mathcal{K}_\delta} f_t(\mathbf{x})$, $\mathbf{y}^* = (1-\delta/r)\mathbf{x}^*$ and $\mathbf{x}^* \in \operatorname{argmin}_{\mathbf{x}\in\mathcal{K}} \sum_{t=1}^{T} f_t(\mathbf{x})$. According to Algorithm 1 and Lemma 1, we have

$$\mathbb{E}\left[D_t \mid \mathbf{y}_t, \mathbf{y}_t^*\right] = 0$$

which further implies that $D_1, \cdots, D_T$ is a martingale difference sequence with

$$\begin{aligned}
|D_t| &= \left|\left(\nabla \widehat{f}_{t,\delta}(\mathbf{y}_t) - \mathbf{g}_t(\mathbf{x}_t)\right)(\mathbf{y}_t^* - \mathbf{y}^*)\right| \\
&\leq \left\|\nabla \widehat{f}_{t,\delta}(\mathbf{y}_t) - \mathbf{g}_t(\mathbf{x}_t)\right\|_2 \|\mathbf{y}_t^* - \mathbf{y}^*\|_2 \\
&\leq 2R\left(\left\|\nabla \widehat{f}_{t,\delta}(\mathbf{y}_t)\right\|_2 + \|\mathbf{g}_t(\mathbf{x}_t)\|_2\right) \\
&\leq 2R\left(G + \frac{dM}{\delta}\right)
\end{aligned}$$

where the last inequality is due to Lemma 1 and Assumption 3.
Then, applying Lemma 5 with $\Delta = 2R(G + \frac{dM}{\delta})\sqrt{2T\ln\frac{1}{\gamma}}$, with probability at least $1-\gamma$, we have

$$\sum_{t=1}^{T} D_t \leq \Delta = 2R(G + \frac{dM}{\delta})\sqrt{2T\ln\frac{1}{\gamma}}. \tag{22}$$

Additionally, combining (21), we further have

$$\sum_{t=1}^{T}\langle\nabla \widehat{f}_{t,\delta}(\mathbf{y}_t), \mathbf{y}_t^* - \mathbf{y}^*\rangle = \sum_{t=1}^{T} D_t + \sum_{t=1}^{T} \mathbf{g}_t(\mathbf{x}_t)^\top(\mathbf{y}_t^* - \mathbf{y}^*). \tag{23}$$

By combining (9), (22) and (23), we complete the proof.