Supplementary Material

No Author Given

No Institute Given

Abstract. In this supplementary material, we give the detailed proof for main theorems and lemmas, and the omitted algorithms.

1 Property of $\mathrm{AO_2MD}$ with the Square Regularizer

Let $\psi_{t,i}(f) = \frac{1}{2\lambda_{t,i}}\|f\|^2_{\mathcal{H}_i}$. The Bregman divergence between any $f, g \in \mathcal{H}_i$ is

$$\mathcal{D}_{\psi_{t,i}}(f,g) = \frac{1}{2\lambda_{t,i}}\|f\|^2_{\mathcal{H}_i} - \frac{1}{2\lambda_{t,i}}\|g\|^2_{\mathcal{H}_i} - \frac{1}{\lambda_{t,i}}\langle g, f-g\rangle_{\mathcal{H}_i} = \frac{1}{2\lambda_{t,i}}\|f-g\|^2_{\mathcal{H}_i}.$$

$\mathrm{AO_2MD}$ is as follows,

$$\begin{aligned} f_{t,i} &= \mathop{\arg\min}_{f\in\mathbb{H}_i}\left\{\langle f, \bar{\nabla}_{t,i}\rangle + \mathcal{D}_{\psi_{t,i}}(f, f'_{t-1,i})\right\},\\ f'_{t,i} &= \mathop{\arg\min}_{f\in\mathbb{H}_i}\left\{\langle f, \nabla_{t,i}\rangle + \mathcal{D}_{\psi_{t,i}}(f, f'_{t-1,i})\right\}, \end{aligned}$$

which is identical to

$$\begin{aligned} \overline{f}_{t,i} &= \mathop{\arg\min}_{f\in\mathcal{H}_i}\left\{\langle f, \bar{\nabla}_{t,i}\rangle + \mathcal{D}_{\psi_{t,i}}(f, f'_{t-1,i})\right\}, \quad f_{t,i} = \mathop{\arg\min}_{f\in\mathbb{H}_i}\|f - \overline{f}_{t,i}\|^2_{\mathcal{H}_i}, \quad (1)\\ \overline{f'}_{t,i} &= \mathop{\arg\min}_{f\in\mathcal{H}_i}\left\{\langle f, \nabla_{t,i}\rangle + \mathcal{D}_{\psi_{t,i}}(f, f'_{t-1,i})\right\}, \quad f'_{t,i} = \mathop{\arg\min}_{f\in\mathbb{H}_i}\|f - \overline{f'}_{t,i}\|^2_{\mathcal{H}_i}. \end{aligned}$$

Then we further analyze the hypothesis updating procedure.

$$\overline{f}_{t,i} = \mathop{\arg\min}_{f\in\mathcal{H}_i}\left\{\langle f, \bar{\nabla}_{t,i}\rangle + \frac{\|f\|_{\mathcal{H}_i}}{2\lambda_{t,i}} - \frac{1}{2\lambda_{t,i}}\|f'_{t-1,i}\|_{\mathcal{H}_i} - \frac{1}{\lambda_{t,i}}\langle f'_{t-1,i}, f - f'_{t-1,i}\rangle\right\}.$$

Solving the optimization problem yields

$$\overline{f}_{t,i} = f'_{t-1,i} - \lambda_{t,i}\bar{\nabla}_{t,i}, \quad \text{and} \quad \overline{f'}_{t,i} = f'_{t-1,i} - \lambda_{t,i}\nabla_{t,i}, \tag{2}$$

which coincides with online gradient descent.

According to the proof of Proposition 7 in [2], we can derive the following inequality, which is critical for subsequent analyses.

$$\|f_{t,i} - f'_{t,i}\|_{\mathcal{H}_i} \leq \|\overline{f}_{t,i} - \overline{f'}_{t,i}\|_{\mathcal{H}_i}. \tag{3}$$

It can be proved that the projection in (1) is defined by $f_{t,i} = \min\{1, \frac{U}{\|\overline{f}_{t,i}\|_{\mathcal{H}_i}}\}\overline{f}_{t,i}$.

2 Bernstein's inequality

Lemma 1. *Let $X_1, \ldots, X_n$ be independent centered, real random variables, and assume that each one is uniformly bounded:*

$$\mathbb{E}[X_k] = 0 \quad \text{and} \quad |X_k| \le a \text{ for each } k = 1, \ldots, n.$$

Introduce the sum $Z = \sum_{k=1}^{n} X_k$, and let Σ_n^2 be the sum of the variances

$$\Sigma_n^2 = \mathbb{E}[Z^2] = \sum_{k=1}^{n} \mathbb{E}X_k^2.$$

Then, with probability at least $1-\delta$,

$$Z < \frac{2}{3} a \ln \frac{1}{\delta} + \sqrt{2\Sigma_n^2 \ln \frac{1}{\delta}}.$$

Lemma 1 is derived from Theorem 1.6.1 in [5], which is only suitable for independent random variables. Next we show a counterpart for martingale.

Lemma 2 (Bernstein's inequality for martingale). *Let $X_1, \ldots, X_n$ be a bounded martingale difference sequence w.r.t. the filtration $\mathcal{F} = (\mathcal{F}_k)_{1 \le k \le n}$ and with $|X_k| \le a$. Let $Z_t = \sum_{k=1}^{t} X_k$ be the associated martingale. Denote the sum of the conditional variances by*

$$\Sigma_n^2 = \sum_{k=1}^{n} \mathbb{E}\left[X_k^2 | \mathcal{F}_{k-1}\right] \le v.$$

Then for all constants $a, v > 0$, with probability at least $1-\delta$,

$$\max_{t=1,\ldots,n} Z_t < \frac{2}{3} a \ln \frac{1}{\delta} + \sqrt{2v \ln \frac{1}{\delta}}.$$

Note that v must be a constant. Lemma 2 is derived from Lemma 1.8 in [1].

3 Proof of Theorem 1

To start with, we give the pseudo-code of the algorithm proposed in Theorem 1.

Let $\mathcal{E}(K)$ be the weighted majority algorithm that enjoys an expected small-loss regret bound [1]. At any round t, $\mathcal{E}(K)$ outputs a probability distribution $\mathbf{p}_t \in \Delta_{K-1}$ as follows

$$p_{t,i} = \frac{\exp(-\eta \sum_{\tau=1}^{t-1} c_{\tau,i})}{\sum_{j=1}^{K} \exp(-\eta \sum_{\tau=1}^{t-1} c_{\tau,j})},$$

where η is the learning rate. We select a kernel $\kappa_{I_t} \sim \mathbf{p}_t$, and execute the first mirror updating for obtaining f_{t,I_t}. Then we make the prediction $\mathrm{sign}(f_t(\mathbf{x}_t))$, where $f_t = f_{t,I_t}$. After obtaining the true gradient ∇_{t,I_t}, we execute the second mirror updating. For the other kernel $\kappa_j \neq \kappa_{I_t}$, we execute a similar procedure.

Next we prove Theorem 1.

Algorithm 1 the Algorithm in Theorem 1

Input: $\psi_{t,i}(f)=\frac{1}{\lambda_{t,i}}\|f\|^2_{\mathcal{H}_i}, i=1,\ldots,K,\ D,\ U.$

Initialization: $\forall \kappa_i\in\mathcal{K},\ f'_{0,i}=0,\ V=\emptyset.$

for $t=1,2,\ldots,T$ **do**
  Select a kernel $\kappa_{I_t}\sim \mathbf{p}_t$ ($\mathbf{p}_t$ is output by $\mathcal{E}(K)$),
  Compute $f_{t,I_t}=\arg\min_{f\in\mathbb{H}_{I_t}}\left\{\langle f,\bar{\nabla}_{t,I_t}\rangle+\mathcal{D}_{\psi_{t,I_t}}(f,f'_{t-1,I_t})\right\}$,
  Output prediction $\hat{y}_t=\mathrm{sign}(f_{t,I_t}(\mathbf{x}_t))$,
  for $\kappa_i\in\mathcal{K}$ **do**
    if $i\neq I_t$, **then** compute $f_{t,i}$ and $f_{t,i}(\mathbf{x}_t)$,
    if $y_tf_{t,i}(\mathbf{x}_t)<1$ **then**
      Compute $\nabla_{t,i}=-y_t\kappa_i(\mathbf{x}_t,\cdot)$,
      Compute $f'_{t,i}=\arg\min_{f\in\mathbb{H}_i}\left\{\langle f,\nabla_{t,i}\rangle+\mathcal{D}_{\psi_{t,i}}(f,f'_{t-1,i})\right\}$,
    end if
    Compute $c_{t,i}=\frac{1}{1+U\sqrt{D}}\max\{0,1-y_tf_{t,i}(\mathbf{x}_t)\}$,
  end for
  Send $\mathbf{c}_t=(c_{t,1},\ldots,c_{t,K})$ to $\mathcal{E}(K)$,
end for

Proof. We analyze the regret w.r.t. any $f\in\mathbb{H}_i$, which can be decomposed by

$$\begin{aligned}\mathrm{Reg}_T(\mathbb{H}_i)&=\sum_{t=1}^T\ell(f_t(\mathbf{x}_t),y_t)-\sum_{t=1}^T\ell(f(\mathbf{x}_t),y_t)\\&=\underbrace{\sum_{t=1}^T\left[\ell(f_t(\mathbf{x}_t),y_t)-\ell(f_{t,i}(\mathbf{x}_t),y_t)\right]}_{\mathcal{T}_1}+\underbrace{\sum_{t=1}^T\left[\ell(f_{t,i}(\mathbf{x}_t),y_t)-\ell(f(\mathbf{x}_t),y_t)\right]}_{\mathcal{T}_2},\end{aligned}\tag{4}$$

where $f_t=f_{t,I_t}\in\mathbb{H}_{I_t}$. Next we separately prove an upper bound of $\mathcal{T}_1$ and $\mathcal{T}_2$.

3.1 Analyzing $\mathcal{T}_1$

Recalling that we reduce online kernel selection to a problem of prediction with expert advice, and define $c_{t,i}=\frac{\ell(f_{t,i}(\mathbf{x}_t),y_t)}{1+U\sqrt{D}}$. Thus $\mathcal{T}_1$ can be rewritten as follows,

$$\mathcal{T}_1:=\sum_{t=1}^T\ell(f_t(\mathbf{x}_t),y_t)-\sum_{t=1}^T\ell(f_{t,i}(\mathbf{x}_t),y_t)=(1+U\sqrt{D})\cdot\sum_{t=1}^T\left[c_{t,I_t}-c_{t,i}\right].$$

$\mathcal{T}_1$ can be upper bounded by the high-probability regret of $\mathcal{E}(K)$. We first state the following expected regret bound of $\mathcal{E}(K)$.

Lemma 3 (Corollary 2.4 in [1]). *For any loss vector $c_t\in[0,1]^K$, let $\eta=\ln\left(1+\sqrt{2\ln(K)/\sum_{t=1}^T c_{t,i^*}}\right)$, the expected regret of $\mathcal{E}(K)$ satisfies,*

$$\sum_{t=1}^T\langle\mathbf{p}_t,c_t\rangle-\sum_{t=1}^T c_{t,i^*}\leq\sqrt{2\sum_{t=1}^T c_{t,i^*}\ln K}+\ln K,$$

where $i^* = \operatorname{argmin}_{i\in[K]} \sum_{t=1}^T c_{t,i}$.

Next we prove a lemma that relates the expected cumulative losses with the actual cumulative losses.

Lemma 4. *For any loss vector* $c_t \in [0,1]^K$*, with probability at least* $1-\delta$*,*

$$\sum_{t=1}^T [c_{t,I_t} - \langle \mathbf{p}_t, c_t\rangle] \le \ln \frac{\lceil \ln(2T)\rceil + 1}{\delta} + \sqrt{4\mathrm{e}\sum_{t=1}^T c_{t,i^*} + 2\mathrm{e}\ln K} \cdot \sqrt{\ln \frac{\lceil \ln(2T)\rceil + 1}{\delta}}.$$

Proof. Define a random variable $X_t = c_{t,I_t} - \langle \mathbf{p}_t, c_t\rangle$ satisfying $|X_t| \le 1$. Conditioned on $c_1, \ldots, c_{t-1}$, we have $\mathbb{E}_t[X_t] = 0$. Thus, $X_1, \ldots, X_{t-1}$ forms bounded martingale difference sequence. The sum of conditional variances satisfies

$$\sum_{t=1}^T \mathbb{E}_t[(X_t)^2] \le \sum_{t=1}^T \mathbb{E}_t[c_{t,I_t}^2] \le \sum_{t=1}^T \langle \mathbf{p}_t, c_t\rangle \le \underbrace{\sum_{t=1}^T c_{t,i^*} + \sqrt{2\sum_{t=1}^T c_{t,i^*}\ln K} + \ln K}_{C^*},$$

where the last inequality comes from Lemma 3. If the examples are generated by a non-oblivious adversary, or the hypothesis sequence that generates $\sum_{t=1}^T c_{t.i^*}$ is random, then C^* is a random variable. Thus we can not use Lemma 2 directly. To solve this issue, we use the exponential grid technique. It is easy to find that $C^* \in [0, 2T]$. We divide the interval $[0, 2T]$ as follows

$$[0, 2T] = \left[0, \mathrm{e}^{-(\lceil \ln T\rceil + 3)}\right] \bigcup_{\mathrm{e}=-\lceil \ln T\rceil - 2}^{\lceil \ln(2T)\rceil} (\mathrm{e}^{i-1}, \mathrm{e}^i].$$

First, we consider the case $C^* > \mathrm{e}^{-(\lceil \ln T\rceil+3)}$. Let $\epsilon = \frac{2}{3}\ln\frac{1}{\delta} + \sqrt{2\mathrm{e}C^*\ln\frac{1}{\delta}}$. We decompose the random event as follows,

$$\begin{aligned}
&\mathbb{P}\left[\max_{t=1,\ldots,T}\sum_{\tau=1}^t X_\tau > \epsilon, \Sigma_T^2 \le C^*, C^* > \mathrm{e}^{-(\lceil \ln T\rceil+3)}\right] \\
=&\mathbb{P}\left[\max_{t=1,\ldots,T}\sum_{\tau=1}^t X_\tau > \epsilon, \Sigma_T^2 \le C^*, \cup_{i=-\lceil \ln T\rceil-2}^{\lceil \ln(2T)\rceil}\mathrm{e}^{i-1} < C^* \le \mathrm{e}^i\right] \\
\le&\sum_{i=-\lceil \ln T\rceil-2}^{\lceil \ln(2T)\rceil}\mathbb{P}\left[\max_{t=1,\ldots,T}\sum_{\tau=1}^t X_\tau > \epsilon_i, \Sigma_T^2 \le \mathrm{e}^i\right],
\end{aligned}$$

where $\epsilon_i = \frac{2}{3}\ln\frac{1}{\delta} + \sqrt{2\mathrm{e}\cdot\mathrm{e}^{i-1}\ln\frac{1}{\delta}}$. For each sub-event, we can use Lemma 2, and obtain

$$\mathbb{P}\left[\max_{t=1,\ldots,T}\sum_{\tau=1}^t X_\tau > \epsilon, \Sigma_T^2 \le C^*, C^* > \mathrm{e}^{-(\lceil \ln T\rceil+3)}\right] < (\lceil \ln(2T)\rceil + \lceil \ln T\rceil + 3)\delta.$$

Next we consider the case $C^* \leq \mathrm{e}^{-(\lceil \ln T \rceil + 3)} \leq \frac{1}{\mathrm{e}^3 T}$. Using Lemma 2 yields, with probability at least $1-\delta$,

$$\max_{t=1,\ldots,T} \sum_{\tau=1}^{t} X_\tau \leq \frac{2}{3}\ln\frac{1}{\delta} + \sqrt{2\mathrm{e}^{-(\lceil \ln T\rceil+3)}\ln\frac{1}{\delta}} \leq \frac{2}{3}\ln\frac{1}{\delta} + \sqrt{\frac{2}{\mathrm{e}^3 T}\ln\frac{1}{\delta}} \leq \ln\frac{1}{\delta},$$

where we use the fact $\sqrt{\frac{2}{\mathrm{e}^3 T}} < \frac{1}{3}$ for $T \geq 1$.
Summing over the two cases, with probability at least $1-(2\lceil \ln(2T)\rceil + 4)\delta$,

$$\max_{t=1,\ldots,T} \sum_{\tau=1}^{t} X_\tau \leq \ln\frac{1}{\delta} + \sqrt{2\mathrm{e}C^* \ln\frac{1}{\delta}},$$

which completes the proof.

Based on Lemma 3 and Lemma 4, we have, with probability at least $1-\delta$,

$$\sum_{t=1}^{T} c_{t,I_t} - \sum_{t=1}^{T} c_{t,i} \leq \ln\frac{K(2\lceil\ln(2T)\rceil+4)}{\delta} + 6\sqrt{\sum_{t=1}^{T} c_{t,i^*}\ln\frac{2\lceil\ln(2T)\rceil+4}{\delta}\ln K},$$

where we merge the lower order terms. Furthermore, we can obtain, with probability at least $1-\delta$,

$$\begin{aligned}\mathcal{T}_1 \leq & 6\sqrt{(1+U\sqrt{D})\min_{i\in[K]}\left[\sum_{t=1}^{T}\ell(f_{t,i}(\mathbf{x}_t),y_t)\right]\ln(K)\ln\frac{2\lceil\ln(2T)\rceil+4}{\delta}} \\ & + (1+U\sqrt{D})\ln\frac{K(2\lceil\ln(2T)\rceil+4)}{\delta}.\end{aligned} \tag{5}$$

3.2 Analyzing $\mathcal{T}_2$

We consider a fixed $\mathbb{H}_i$. The analyses follows some techniques in [2]. Thus we just show the main steps. Using the convexity of the hinge loss, we have

$$\begin{aligned}\sum_{t=1}^{T}\ell(f_{t,i}(\mathbf{x}_t),y_t) - \sum_{t=1}^{T}\ell(f(\mathbf{x}_t),y_t) &\leq \sum_{t\in E_i}\ell(f_{t,i}(\mathbf{x}_t),y_t) - \sum_{t\in E_i}\ell(f(\mathbf{x}_t),y_t) \\ &\leq \sum_{t\in E_i}\langle f_{t,i}-f, \nabla_{t,i}\rangle \\ &= \sum_{t\in E_i}[\underbrace{\langle f_{t,i}-f'_{t,i}, \bar{\nabla}_{t,i}\rangle}_{\mathcal{T}_{2,1,t}} + \underbrace{\langle f'_{t,i}-f, \nabla_{t,i}\rangle}_{\mathcal{T}_{2,2,t}} + \underbrace{\langle f_{t,i}-f'_{t,i}, \nabla_{t,i}-\bar{\nabla}_{t,i}\rangle}_{\mathcal{T}_{2,3,t}}],\end{aligned}$$

where $E_i = \{t\in[T] : y_t f_{t,i}(\mathbf{x}_t) < 1\}$. For $t \notin E_i$, we have $\ell(f_{t,i}(\mathbf{x}_t),y_t) = 0$. Thus the first inequality holds on. For convenience, let $E_i = \{s_1,\ldots,s_j,\ldots,s_{|E_i|}\}$, where $s_j \in [T]$. For $\mathcal{T}_{2,1,t}$ and $\mathcal{T}_{2,2,t}$, we have

$$\mathcal{T}_{2,1,t} \leq \mathcal{D}_{\psi_{t,i}}(f'_{t,i}, f'_{t-1,i}) - \mathcal{D}_{\psi_{t,i}}(f'_{t,i}, f_{t,i}) - \mathcal{D}_{\psi_{t,i}}(f_{t,i}, f'_{t-1,i}), \tag{6}$$

$$\mathcal{T}_{2,2,t} \leq \mathcal{D}_{\psi_{t,i}}(f, f'_{t-1,i}) - \mathcal{D}_{\psi_{t,i}}(f, f'_{t,i}) - \mathcal{D}_{\psi_{t,i}}(f'_{t,i}, f'_{t-1,i}). \tag{7}$$

Combining (6) and (7), and taking a summation over $t \in E_i$ yields

$$
\begin{aligned}
&\sum_{t\in E_i}\left[\mathcal{T}_{2,1,t}+\mathcal{T}_{2,2,t}\right]\\
\leq&\sum_{t\in E_i}\left[\mathcal{D}_{\psi_{t,i}}(f,f'_{t-1,i})-\mathcal{D}_{\psi_{t,i}}(f,f'_{t,i})-\mathcal{D}_{\psi_{t,i}}(f'_{t,i},f_{t,i})-\mathcal{D}_{\psi_{t,i}}(f_{t,i},f'_{t-1,i})\right]\\
=&\mathcal{D}_{\psi_{s_1,i}}(f,f'_{s_1-1,i})+\sum_{j=1}^{|E_i|-1}\left[\mathcal{D}_{\psi_{s_{j+1},i}}(f,f'_{s_{j+1}-1,i})-\mathcal{D}_{\psi_{s_j,i}}(f,f'_{s_j,i})\right]\\
&\qquad-\mathcal{D}_{\psi_{s_{|E_i|},i}}(f,f'_{s_{|E_i|},i})-\sum_{t\in E_i}\left[\mathcal{D}_{\psi_{t,i}}(f'_{t,i},f_{t,i})+\mathcal{D}_{\psi_{t,i}}(f_{t,i},f'_{t-1,i})\right]\\
\leq&\mathcal{D}_{\psi_{s_1,i}}(f,f'_{s_1-1,i})-\sum_{t\in E_i}\left[\mathcal{D}_{\psi_{t,i}}(f'_{t,i},f_{t,i})+\mathcal{D}_{\psi_{t,i}}(f_{t,i},f'_{t-1,i})\right].
\end{aligned}
$$

If $s_{j+1}-1\in E_i$, it must be $f'_{s_{j+1}-1}=f'_{s_j}$. Now we need to prove that the equality holds on when $s_{j+1}-1\notin E_i$. Note that for $t\notin E_i$, our algorithm will not execute the second mirror updating, thus $f'_{t,i}=f'_{\tau,i}$ where $\tau=\max_{s<t}s\in E_i$. Let $t=s_{j+1}-1$. Then we have $f'_{s_{j+1}-1}=f'_{s_j}$. Using a constant learning rate λ_i for $\psi_{t,i}$, the last inequality holds on. Recalling that $f'_{0,i}=0$ and $\bar{\nabla}_{1,i}=0$, we have $f_{1,i}=f'_{0,i}=0$, which implies $\ell(f_{1,i}(\mathbf{x}_1),y_1)=1$. Thus it must be $s_1=1$ and $\mathcal{B}_{\psi_{s_1,i}}(f,f'_{s_1-1,i})=\frac{1}{2\lambda_i}\|f\|^2_{\mathcal{H}_i}$.

Next we analyze $\mathcal{T}_{2,3,t}$. Using (3), we have

$$
\begin{aligned}
\langle f_{t,i}-f'_{t,i},\nabla_{t,i}-\bar{\nabla}_{t,i}\rangle\leq&\|f_{t,i}-f'_{t,i}\|_{\mathcal{H}_i}\cdot\|\nabla_{t,i}-\bar{\nabla}_{t,i}\|_{\mathcal{H}_i}\\
\leq&\|\overline{f}_{t,i}-\overline{f'}_{t,i}\|_{\mathcal{H}_i}\cdot\|\nabla_{t,i}-\bar{\nabla}_{t,i}\|_{\mathcal{H}_i}\\
\leq&\lambda_i\|\nabla_{t,i}-\bar{\nabla}_{t,i}\|^2_{\mathcal{H}_i}.
\end{aligned}
$$

Furthermore, we can obtain

$$
\begin{aligned}
\sum_{t\in E_i}\langle f_{t,i}-f,\nabla_{t,i}\rangle\leq&\frac{1}{\lambda_i}\|f\|^2_{\mathcal{H}_i}+\lambda_i\sum_{t\in E_i}\|\nabla_{t,i}-\bar{\nabla}_{t,i}\|^2_{\mathcal{H}_i}\\
\leq&\frac{1}{2\lambda_i}\|f\|^2_{\mathcal{H}_i}+2\lambda_i\sum_{t\in E_i}\left[\|\nabla_{t,i}-\mu_{T,i}\|^2_{\mathcal{H}_i}+\left\|\mu_{T,i}-\nabla_{r_i(t),i}\right\|^2_{\mathcal{H}_i}\right]\\
\leq&\frac{1}{2\lambda_i}\|f\|^2_{\mathcal{H}_i}+2\lambda_i\left(2\mathcal{A}(\mathcal{I}_T,\kappa_i)+D_i\right).
\end{aligned}
$$

Let the learning rate $\lambda_i=(8\mathcal{A}(\mathcal{I}_T,\kappa_i))^{-\frac{1}{2}}$. Then we have

$$
\mathcal{T}_2:=\sum_{t=1}^{T}\left[\ell(f_{t,i}(\mathbf{x}_t),y_t)-\ell(f(\mathbf{x}_t),y_t)\right]=O\left((\|f\|^2_{\mathcal{H}_i}+1)\sqrt{\mathcal{A}(\mathcal{I}_T,\kappa_i)}\right),
$$

where we omit the lower order terms.

3.3 The final regret

Summing over $\mathcal{T}_1$ and $\mathcal{T}_2$ yields that, with probability at least $1-\delta$, we have

$$\begin{aligned}\mathrm{Reg}_T(\mathbb{H}_i) =& O\left(\sqrt{\min_{i\in[K]}\left[\sum_{t=1}^{T}\ell(f_{t,i}(\mathbf{x}_t),y_t)\right]\ln(K)\ln\frac{\ln(2T)}{\delta}}+\ln\frac{K\ln(2T)}{\delta}\right)\\ &+O\left((\|f\|^2_{\mathcal{H}_i}+1)\sqrt{\mathcal{A}(\mathcal{I}_T,\kappa_i)}\right)\\ =& O\left(\sqrt{L_T(f)\ln(K)\ln\frac{\ln(2T)}{\delta}}+(\|f\|^2_{\mathcal{H}_i}+1)\sqrt{\mathcal{A}(\mathcal{I}_T,\kappa_i)}\right).\end{aligned}$$

Let $f^\ast_i=\mathrm{argmin}_{f\in\mathbb{H}_i}L_T(f)$ and $\tilde{f}=\frac{1}{T}\sum_{\tau=1}^{T}y_\tau\kappa_i(\mathbf{x}_\tau,\cdot)$. It can be verified that

$$L_T(f^\ast_i)\leq L_T(\tilde{f})\leq T-\frac{1}{T}\mathbf{Y}^\top_T\mathbf{K}_{\kappa_i}\mathbf{Y}_T\leq\mathcal{A}(\mathcal{I}_T,\kappa_i),$$

where we use the fact $\kappa_i(\mathbf{x},\mathbf{x})\geq 1$. Besides, we have $\|\tilde{f}\|_{\mathcal{H}_i}\leq D_i\leq D$. Thus $\tilde{f}\in\mathbb{H}_i$. In this way, we complete the proof.

4 Proof of Lemma 1

Proof. Define a random variable X_t as follows,

$$X_t=\|\tilde{\mu}_{t,i}-\mu_{t,i}\|^2_{\mathcal{H}_i}-\mathbb{E}_t\left[\|\tilde{\mu}_{t,i}-\mu_{t,i}\|^2_{\mathcal{H}_i}\right].$$

Under the condition of $\tilde{\mu}_{1,i},\ldots,\tilde{\mu}_{t-1,i}$, it can be verified that $\mathbb{E}_t[X_t]=0$, and $|X_t|\leq 4D_i$. Since $\mathbb{E}_t[\tilde{\mu}_{t,i}]=\mu_{t,i}$ for $t>M$ and $\tilde{\mu}_{t,i}=\mu_{t,i}$ for $t\leq M$ [3], we have

$$\mathbb{E}_t\left[\|\tilde{\mu}_{t,i}-\mu_{t,i}\|^2_{\mathcal{H}_i}\right]=\mathrm{Var}(\tilde{\mu}_{t,i})\leq\frac{1}{Mt}\sum_{\tau=1}^{t}\left\|-y_\tau\kappa_i(\mathbf{x}_\tau,\cdot)+\frac{1}{t}\sum_{s=1}^{t}y_s\kappa_i(\mathbf{x}_s,\cdot)\right\|^2_{\mathcal{H}_i}.$$

The sum of conditional variances can be upper bounded as follows,

$$\begin{aligned}\sum_{t=M+1}^{T}\mathbb{E}_t[(X_t)^2]=&\sum_{t=M+1}^{T}\mathbb{E}_t\left[\left(\|\tilde{\mu}_{t,i}-\mu_{t,i}\|^2_{\mathcal{H}_i}-\mathbb{E}_t\left[\|\tilde{\mu}_{t,i}-\mu_{t,i}\|^2_{\mathcal{H}_i}\right]\right)^2\right]\\ =&\sum_{t=M+1}^{T}\mathbb{E}_t\left[\|\tilde{\mu}_{t,i}-\mu_{t,i}\|^4_{\mathcal{H}_i}\right]-\left(\mathbb{E}_t\left[\|\tilde{\mu}_{t,i}-\mu_{t,i}\|^2_{\mathcal{H}_i}\right]\right)^2\\ \leq&4D_i\sum_{t=M+1}^{T}\mathbb{E}_t\left[\|\tilde{\mu}_{t,i}-\mu_{t,i}\|^2_{\mathcal{H}_i}\right]\\ \leq&4D_i\sum_{t=M+1}^{T}\frac{1}{Mt}\sum_{\tau=1}^{t}\left\|-y_\tau\kappa_i(\mathbf{x}_\tau,\cdot)+\frac{1}{t}\sum_{s=1}^{t}y_s\kappa_i(\mathbf{x}_s,\cdot)\right\|^2_{\mathcal{H}_i}\\ \leq&4D_i\sum_{t=M+1}^{T}\frac{1}{Mt}\mathcal{A}(\mathcal{I}_T,\kappa_i)\leq\frac{4D_i\mathcal{A}(\mathcal{I}_T,\kappa_i)}{M}\ln\frac{T}{M}.\end{aligned}$$

The above analysis also implies $\sum_{t=M+1}^{T}\mathbb{E}_t\left[\|\tilde{\mu}_{t,i}-\mu_{t,i}\|^2_{\mathcal{H}_i}\right]\leq\frac{\mathcal{A}(\mathcal{I}_T,\kappa_i)}{M}\ln\frac{T}{M}$. Using Lemma 2 yields the desired result.

5 Proof of Theorem 2

Proof. The proof is similar with the analysis of $\mathcal{T}_2$ in the proof of Theorem 1. We decompose the regret as follows,

$$\sum_{t=1}^{T}\left[\ell(f_{t,i}(\mathbf{x}_t),y_t)-\ell(f(\mathbf{x}_t),y_t)\right]\leq\underbrace{\sum_{t\in E_i}\langle f_{t,i}-f,\tilde{\nabla}_{t,i}\rangle}_{\mathcal{T}_{2,1}}+\underbrace{\sum_{t\in E_i}\langle f_{t,i}-f,\nabla_{t,i}-\tilde{\nabla}_{t,i}\rangle}_{\mathcal{T}_{2,2}}.$$

We need to analyze $\mathcal{T}_{2,1}$ and $\mathcal{T}_{2,2}$.

5.1 Analyzing $\mathcal{T}_{2,1}$

$$\mathcal{T}_{2,1}=\sum_{t\in E_i}[\underbrace{\langle f_{t,i}-f'_{t,i},\bar{\nabla}_{t,i}\rangle}_{\mathcal{T}_{2,1,t}}+\underbrace{\langle f'_{t,i}-f,\tilde{\nabla}_{t,i}\rangle}_{\mathcal{T}_{2,2,t}}+\underbrace{\langle f_{t,i}-f'_{t,i},\tilde{\nabla}_{t,i}-\bar{\nabla}_{t,i}\rangle}_{\mathcal{T}_{2,3,t}}]. \tag{8}$$

Combining (6) and (7), and taking a summation over $t\in E_i$ yields

$$\sum_{t\in E_i}[\mathcal{T}_{2,1,t}+\mathcal{T}_{2,2,t}]\leq\frac{1}{2\lambda_i}\|f\|^2_{\mathcal{H}_i}-\sum_{t\in E_i}[\mathcal{D}_{\psi_{t,i}}(f'_{t,i},f_{t,i})+\mathcal{D}_{\psi_{t,i}}(f_{t,i},f'_{t-1,i})].$$

Substituting into (8) yields

$$\mathcal{T}_{2,1}\leq\frac{\|f\|^2_{\mathcal{H}_i}}{2\lambda_i}+\sum_{t\in E_i}\langle f_{t,i}-f'_{t,i},\tilde{\nabla}_{t,i}-\bar{\nabla}_{t,i}\rangle\leq\frac{\|f\|^2_{\mathcal{H}_i}}{2\lambda_i}+\sum_{t\in E_i}\lambda_i\|\tilde{\nabla}_{t,i}-\bar{\nabla}_{t,i}\|^2_{\mathcal{H}_i}, \tag{9}$$

where the last equality comes from (2) and (3). Recalling that

$$\tilde{\nabla}_{t,i}=\frac{\nabla_{t,i}-\bar{\nabla}_{t,i}}{\mathbb{P}[i=J_t]}\mathbb{I}_{i=J_t}+\bar{\nabla}_{t,i},\ \forall i\in[K].$$

Substituting into (9) and using Lemma 5 yields, with probability at least $1-2\delta$,

$$\begin{aligned}\mathcal{T}_{2,1}\leq&\frac{1}{2\lambda_i}\|f\|^2_{\mathcal{H}_i}+\lambda_i\sum_{t\in E_i}\frac{\|\nabla_{t,i}-\bar{\nabla}_{t,i}\|^2_{\mathcal{H}_i}}{(\mathbb{P}[i=J_t])^2}\mathbb{I}_{i=J_t}\\ \leq&\frac{1}{2\lambda_i}\|f\|^2_{\mathcal{H}_i}+\lambda_i\left[11\sqrt{D_i}K\frac{M+D_i\ln\frac{T}{M}}{M}\mathcal{A}(\mathcal{I}_T,\kappa_i)\ln^{\frac{3}{4}}\frac{1}{\delta}+20K^2D_i\ln\frac{1}{\delta}\right].\end{aligned}$$

5.2 Analyzing $\mathcal{T}_{2,2}$

Using Lemma 6, with probability at least $1-2\delta$,

$$\begin{aligned}\mathcal{T}_{2,2} :=& \sum_{t\in E_i} \langle f_{t,i} - f, \nabla_{t,i} - \tilde{\nabla}_{t,i}\rangle\\ \leq& 13K\sqrt{D_i}U\ln\frac{1}{\delta} + 7U\sqrt{K\frac{M+D_i\ln\frac{T}{M}}{M}\mathcal{A}(\mathcal{I}_T,\kappa_i)}\ln^{\frac{3}{4}}\frac{1}{\delta}.\end{aligned}$$

5.3 The final regret

Summing over $\mathcal{T}_{2,1}$ and $\mathcal{T}_{2,2}$ yields, with probability at least $1-4\delta$,

$$\begin{aligned}\sum^{T}_{t=1}\left[\ell(f_{t,i}(\mathbf{x}_t),y_t) - \ell(f(\mathbf{x}_t),y_t)\right] \leq& \frac{\|f\|^2_{\mathcal{H}_i}}{2\lambda_i} + 11\sqrt{D_i}K\lambda_i g_i(T,M)\mathcal{A}(\mathcal{I}_T,\kappa_i)\ln^{\frac{3}{4}}\frac{1}{\delta}\\ &+ 20\lambda_i K^2 U D_i \ln\frac{1}{\delta} + 13K\sqrt{D_i}U\ln\frac{1}{\delta} + 7U\sqrt{Kg_i(T,M)\mathcal{A}(\mathcal{I}_T,\kappa_i)}\ln^{\frac{3}{4}}\frac{1}{\delta}.\end{aligned}$$

where $g_i(T,M) = \frac{M+D_i\ln\frac{T}{M}}{M}$. Using a union-of-events bound to $i = 1,\ldots,K$ completes the proof.

Lemma 5. *For a fixed $i\in[K]$, with probability at least $1-2\delta$,*

$$\sum_{t\in E_i}\frac{\|\nabla_{t,i}-\bar{\nabla}_{t,i}\|^2_{\mathcal{H}_i}\mathbb{I}_{i=J_t}}{(\mathbb{P}[i=J_t])^2} \leq 11\sqrt{D_i}K\frac{M+D_i\ln\frac{T}{M}}{M}\mathcal{A}(\mathcal{I}_T,\kappa_i)\ln^{\frac{3}{4}}\frac{1}{\delta} + 20K^2D_i\ln\frac{1}{\delta}.$$

Proof. Define a random variable

$$X_t = \frac{\|\nabla_{t,i}-\bar{\nabla}_{t,i}\|^2_{\mathcal{H}_i}\mathbb{I}_{i=J_t}}{(\mathbb{P}[i=J_t])^2} - \frac{\|\nabla_{t,i}-\bar{\nabla}_{t,i}\|^2_{\mathcal{H}_i}}{\mathbb{P}[i=J_t]}.$$

Conditioned on $J_1,\ldots,J_{t-1}$, $\mathbb{E}_t[X_{t,i}] = 0$ and $|X_{t,i}| \leq 2K^2D_i$. The sum of conditional variances satisfies

$$\begin{aligned}\Sigma^2_T =& \sum_{t\in E_i}\mathbb{E}_t[(X_t-\mathbb{E}_t[X_t])^2] \leq \sum_{t\in E_i}\frac{\|\nabla_{t,i}-\bar{\nabla}_{t,i}\|^4_{\mathcal{H}_i}}{(\mathbb{P}[i=J_t])^3}\\ \leq& \sum_{t\in E_i} 2D_iK^3\|\nabla_{t,i}-\bar{\nabla}_{t,i}\|^2_{\mathcal{H}_i}\\ \leq& 2D_iK^3\sum_{t\in E_i} 3\left[\|\nabla_{t,i}-\mu_{t,i}\|^2_{\mathcal{H}_i} + \|\mu_{t,i}-\mu_{t-1,i}\|^2_{\mathcal{H}_i} + \|\mu_{t-1,i}-\tilde{\mu}_{t-1,i}\|^2_{\mathcal{H}_i}\right],\end{aligned}$$

where $\mu_{t,i} = \frac{-1}{t}\sum^t_{\tau=1} y_\tau\kappa_i(\mathbf{x}_\tau,\cdot)$ and $\bar{\nabla}_{t,i} = \tilde{\mu}_{t-1,i}$. Next we analyze the three terms in the square brackets, respectively.
For the first term, we have

$$\sum_{t\in E_i}\|\nabla_{t,i}-\mu_{t,i}\|^2_{\mathcal{H}_i} \leq \sum^T_{t=1}\|\nabla_{t,i}-\mu_{t,i}\|^2_{\mathcal{H}_i} \leq \sum^T_{t=1}\|\nabla_{t,i}-\mu_{T,i}\|^2_{\mathcal{H}_i} \leq \mathcal{A}(\mathcal{I}_T,\kappa_i).$$

For the second term, we have

$$\sum_{t\in E_i}\|\mu_{t,i}-\mu_{t-1,i}\|^2_{\mathcal{H}_i}\le\sum_{t=2}^{T}\left\|\frac{1}{t-1}\sum_{\tau=1}^{t-1}y_\tau\kappa_i(\mathbf{x}_\tau,\cdot)-\frac{1}{t}\sum_{\tau=1}^{t}y_\tau\kappa_i(\mathbf{x}_\tau,\cdot)\right\|^2_{\mathcal{H}_i}+D_i$$
$$\le\sum_{t=2}^{T}\left\|\frac{1}{t(t-1)}\sum_{\tau=1}^{t-1}y_\tau\kappa_i(\mathbf{x}_\tau,\cdot)-\frac{1}{t}y_t\kappa_i(\mathbf{x}_t,\cdot)\right\|^2_{\mathcal{H}_i}+D_i$$
$$\le\sum_{t=2}^{T}\frac{2}{t^2(t-1)^2}\left\|\sum_{\tau=1}^{t-1}y_\tau\kappa_i(\mathbf{x}_\tau,\cdot)\right\|^2_{\mathcal{H}_i}+D_i\sum_{t=2}^{T}\frac{2}{t^2}+D_i\le 5D_i,$$

where we use the inequality $\sum_{t=2}^{T}\frac{1}{t^2}\le 1$.
For the third term, using Lemma 1, with probability at least $1-\delta$, we have

$$\sum_{t\in E_i}\|\mu_{t-1,i}-\tilde{\mu}_{t-1,i}\|^2_{\mathcal{H}_i}\le\sum_{t=M+2}^{T}\|\mu_{t-1,i}-\tilde{\mu}_{t-1,i}\|^2_{\mathcal{H}_i}$$
$$\le\frac{D_i\mathcal{A}(\mathcal{I}_T,\kappa_i)}{M}\ln\frac{T}{M}+\frac{8D_i}{3}\ln\frac{1}{\delta}+\sqrt{\frac{8D_i\mathcal{A}(\mathcal{I}_T,\kappa_i)}{M}\ln\frac{T}{M}\ln\frac{1}{\delta}},$$

where the first inequality comes from $\mu_{t,i}=\tilde{\mu}_{t,i}$ for $t\le M$.
Summing over the three above terms, with probability at least $1-\delta$,

$$\Sigma_T^2\le 6D_i^2K^3\left[\frac{M+D_i\ln\frac{T}{M}}{D_iM}\mathcal{A}(\mathcal{I}_T,\kappa_i)+5+\frac{8}{3}\ln\frac{1}{\delta}+\sqrt{\frac{8\mathcal{A}(\mathcal{I}_T,\kappa_i)}{D_iM}\ln\frac{T}{M}\ln\frac{1}{\delta}}\right]$$
$$<24D_iK^3\left[\frac{M+D_i\ln\frac{T}{M}}{M}\mathcal{A}(\mathcal{I}_T,\kappa_i)\sqrt{\ln\frac{1}{\delta}}+2D_i\ln\frac{1}{\delta}\right].$$

Using Lemma 2, with probability at least $1-2\delta$, we have

$$\sum_{t\in E_i}\frac{\|\nabla_{t,i}-\bar{\nabla}_{t,i}\|^2_{\mathcal{H}_i}\mathbb{I}_{i=J_t}}{(\mathbb{P}[i=J_t])^2}\le\sum_{t\in E_i}\frac{\|\nabla_{t,i}-\bar{\nabla}_{t,i}\|^2_{\mathcal{H}_i}}{\mathbb{P}[i=J_t]}+\frac{4K^2D_i}{3}\ln\frac{1}{\delta}$$
$$+\sqrt{48D_iK^3\left[\frac{M+D_i\ln\frac{T}{M}}{M}\mathcal{A}(\mathcal{I}_T,\kappa_i)\sqrt{\ln\frac{1}{\delta}}+2D_i\ln\frac{1}{\delta}\right]\ln\frac{1}{\delta}}.$$

Similarly, we can obtain

$$\sum_{t\in E_i}\frac{\|\nabla_{t,i}-\bar{\nabla}_{t,i}\|^2_{\mathcal{H}_i}}{\mathbb{P}[i=J_t]}\le 4K\left[\frac{M+D_i\ln\frac{T}{M}}{M}\mathcal{A}(\mathcal{I}_T,\kappa_i)\sqrt{\ln\frac{1}{\delta}}+2D_i\ln\frac{1}{\delta}\right].$$

Assuming $K<\mathcal{A}(\mathcal{I}_T,\kappa_i)$ and merging the lower order terms gives, with probability at least $1-2\delta$,

$$\sum_{t\in E_i}\frac{\|\nabla_{t,i}-\bar{\nabla}_{t,i}\|^2_{\mathcal{H}_i}\mathbb{I}_{i=J_t}}{(\mathbb{P}[i=J_t])^2}\le 11\sqrt{D_i}Kg_i(M,T)\mathcal{A}(\mathcal{I}_T,\kappa_i)\ln^{\frac{3}{4}}\frac{1}{\delta}+20K^2D_i\ln\frac{1}{\delta},$$

where $g_i(M,T) = \frac{M+D_i \ln \frac{T}{M}}{M}$. Thus we complete the proof.

Lemma 6. *For a fixed $i \in [K]$, with probability at least $1-2\delta$,*

$$\sum_{t\in E_i} \langle f_{t,i} - f, \nabla_{t,i} - \tilde{\nabla}_{t,i}\rangle \leq 13K\sqrt{D_i}U \ln\frac{1}{\delta} + 7U\sqrt{Kg_i(M,T)\mathcal{A}(\mathcal{I}_T,\kappa_i)}\ln^{\frac{3}{4}}\frac{1}{\delta}.$$

Proof. To start with, we state an intermediate result derived from the technique adopted in the proof of Lemma 5. With probability at least $1-\delta$,

$$\sum_{t=1}^{T} \|\nabla_{t,i} - \bar{\nabla}_{t,i}\|^2_{\mathcal{H}_i} \leq 12\left[\frac{M + D_i \ln\frac{T}{M}}{M}\mathcal{A}(\mathcal{I}_T,\kappa_i)\sqrt{\ln\frac{1}{\delta}} + 2D_i\ln\frac{1}{\delta}\right]. \tag{10}$$

For any $t \in E_i$, we define a random variable $X_{t,i}$ as follows,

$$\begin{aligned}
X_{t,i} =& \langle f_{t,i} - f, \nabla_{t,i} - \tilde{\nabla}_{t,i}\rangle \\
=& \left\langle f_{t,i} - f, \nabla_{t,i} - \tilde{\mu}_{t-1,i} - \frac{\nabla_{t,i} - \tilde{\mu}_{t-1,i}}{\mathbb{P}[i = J_t]}\mathbb{I}_{i=J_t}\right\rangle \\
=& \Xi_t(f_{t,i}) - \Xi_t(f) - \frac{\Xi_t(f_{t,i}) - \Xi_t(f)}{\mathbb{P}[i = J_t]}\mathbb{I}_{i=J_t},
\end{aligned}$$

where $\Xi_t(f) = \sum_{(\mathbf{x},y)\in V} \frac{yf(\mathbf{x})}{|V|} - y_t f(\mathbf{x}_t)$. Under the condition of $J_1,\ldots,J_{t-1}$, we have $\mathbb{E}_t[X_{t,i}] = 0$. Besides

$$|\langle f_{t,i} - f, \nabla_{t,i} - \tilde{\nabla}_{t,i}\rangle| \leq \|f_{t,i} - f\|_{\mathcal{H}_i} \cdot \|\nabla_{t,i} - \tilde{\nabla}_{t,i}\|_{\mathcal{H}_i} \leq 4UK\sqrt{D_i},$$

where we use the fact $\|f\|_{\mathcal{H}_i} \leq U$ for any $f \in \mathbb{H}_i$ and $\mathbb{P}[i = J_t] = \frac{1}{K}$. Thus $X_{t,i}, t \in E_i$ forms bounded martingale difference sequence. Using (10), with probability at least $1-\delta$, the sum of conditional variances satisfies

$$\begin{aligned}
\Sigma_T^2 = \sum_{t\in E_i} \mathbb{E}_t\left[(X_{t,i})^2\right] \leq& K\sum_{t\in E_i}\left(\Xi_t(f_{t,i}) - \Xi_t(f)\right)^2 \\
=& K\sum_{t\in E_i}\left\langle f_{t,i} - f, \nabla_{t,i} - \bar{\nabla}_{t,i}\right\rangle^2_{\mathcal{H}_i} \\
\leq& 2KU^2\sum_{t\in E_i}\|\nabla_{t,i} - \tilde{\mu}_{t-1,i}\|^2_{\mathcal{H}_i} \\
\leq& 24KU^2\left[\frac{M + D_i\ln\frac{T}{M}}{M}\mathcal{A}(\mathcal{I}_T,\kappa_i)\sqrt{\ln\frac{1}{\delta}} + 2D_i\ln\frac{1}{\delta}\right].
\end{aligned}$$

Using Lemma 2 gives the desired result.

6 Proof of Theorem 3

Proof. We analyze the regret w.r.t. any $f \in \mathbb{H}_i$. According to the regret decomposition (4), we also have

$$\mathrm{Reg}_T(\mathbb{H}_i) = \underbrace{\sum_{t=1}^{T} \left[\ell(f_t(\mathbf{x}_t), y_t) - \ell(f_{t,i}(\mathbf{x}_t), y_t)\right]}_{\mathcal{T}_1} + \underbrace{\sum_{t=1}^{T} \left[\ell(f_{t,i}(\mathbf{x}_t), y_t) - \ell(f(\mathbf{x}_t), y_t)\right]}_{\mathcal{T}_2},$$

where $f_t = f_{t,I_t} \in \mathbb{H}_{I_t}$. Next we separately provide upper bounds on $\mathcal{T}_1$ and $\mathcal{T}_2$.

Analyzing $\mathcal{T}_1$ We reduce online kernel selection to a K-armed adversarial bandit problem, in which $c_{t,i} = \frac{\ell(f_{t,i}(\mathbf{x}_t), y_t)}{1+U\sqrt{D}}$. Thus $\mathcal{T}_1$ can be rewritten as follows,

$$\mathcal{T}_1 = (1 + U\sqrt{D}) \cdot \sum_{t=1}^{T} \left[c_{t,I_t} - c_{t,i}\right].$$

$\mathcal{T}_1$ can be upper bounded by the regret of $\mathcal{M}(K)$. The following theorem shows there exists an algorithm that satisfies Assumption 1.

Theorem 1 ([4]). *For any loss vector $c_t \in [0,1]^K$, with probability at least $1-\delta$, the regret of GREEN-IX satisfies,*

$$\sum_{t=1}^{T} c_{t,I_t} - \sum_{t=1}^{T} c_{t,i} = \tilde{O}\left(\sqrt{\min_{i\in[K]} \sum_{t=1}^{T} c_{t,i} K \ln \frac{K}{\delta}} + K \ln \frac{K}{\delta}\right).$$

Based on Theorem 1, with probability at least $1-\delta$, we have

$$\mathcal{T}_1 = \tilde{O}\left(\sqrt{(1+U\sqrt{D}) \min_{i\in[K]} \left[\sum_{t=1}^{T} \ell(f_{t,i}(\mathbf{x}_t), y_t)\right] K \ln \frac{K}{\delta}} + K \ln \frac{K}{\delta}\right).$$

Analyzing $\mathcal{T}_2$ $\mathcal{T}_2$ has been shown in Theorem 2.

The final regret Combining $\mathcal{T}_1$ and $\mathcal{T}_2$ yields, with probability at least $1-5\delta$,

$$\begin{aligned}
\mathrm{Reg}_T(\mathbb{H}_i) \leq& \tilde{O}\left(\sqrt{(1+U\sqrt{D}) \min_{i\in[K]} \left[\sum_{t=1}^{T} \ell(f_{t,i}(\mathbf{x}_t), y_t)\right] K \ln \frac{K}{\delta}} + K \ln \frac{K}{\delta}\right) \\
&+ O\left((\|f\|^2_{\mathcal{H}_i} + U)\sqrt{K\mathcal{A}(\mathcal{I}_T, \kappa_i)} \ln^{\frac{3}{4}} \frac{K}{\delta}\right) \\
\leq& \tilde{O}\left(\sqrt{L_T(f) K \ln \frac{K}{\delta}} + (\|f\|^2_{\mathcal{H}_i} + U)\sqrt{K\mathcal{A}(\mathcal{I}_T, \kappa_i)} \ln^{\frac{3}{4}} \frac{K}{\delta}\right).
\end{aligned}$$

Replacing f with $f_i^* = \mathrm{argmin}_{f\in\mathbb{H}_i} L_T(f)$ recoveries the desired result.

7 Proof of Theorem 4

Proof. At the end of round t, the support vectors in each $S_i, i = 1, \dots, K$ contains three sources: (i) the first M examples in V $(t \le M)$, (ii) the ones inserted by reservoir updating from $t \ge M+1$, and (iii) the ones inserted by the Bernoulli sampling procedure. Recalling that $(\mathbf{x}_t, y_t)$ will be added into V with probability $\min\{1, \frac{M}{t}\}$. Define a Bernoulli random variable δ_t. If $\delta_t = 1$, then $(\mathbf{x}_t, y_t)$ is added into V. Besides, if $y_t f_{t,i}(\mathbf{x}_t) < 1$ and $b_{t,i} = 1$, then $(\mathbf{x}_t, y_t)$ will still be added into S_i. Define a random variable $\nu_{t,i} = \mathbb{I}_{y_t f_{t,i}(\mathbf{x}_t)<1} \cdot \mathbb{I}_{b_{t,i}=1}$. At the end of the $(T-1)$-th round, the size of S_i satisfies

$$|S_i| \le M + \sum_{t=M+1}^{T-1} (\delta_t + \nu_{t,i} - \delta_t \cdot \nu_{t,i}) \le M + \sum_{t=M+1}^{T-1} \delta_t + \sum_{t\in F_i} \mathbb{I}_{b_{t,i}=1},$$

where $F_i = \{t \ge M+1 : y_t f_{t,i}(\mathbf{x}_t) < 1\}$.
Analyzing $\sum_{t=M+1}^{T-1} \delta_t$. Using Lemma 7, with probability at least $1-\delta$,

$$\sum_{t=M+1}^{T-1} \delta_t \le M \ln \frac{T}{M} + \frac{2}{3} \ln \frac{1}{\delta} + \sqrt{2M \ln \frac{T}{M} \ln \frac{1}{\delta}}.$$

Analyzing $\sum_{t\in F_i} \mathbb{I}_{b_{t,i}=1}$. For each $t \in F_i$, we define a random variable $X_{t,i} = \mathbb{I}_{b_{t,i}=1} - \mathbb{P}[b_{t,i} = 1]$. Under the condition of $\{b_{\tau,i}\}_{\tau\in F_i, \tau<t}$, we have $\mathbb{E}_t[X_{t,i}] = 0$ and $|X_{t,i}| \le 1$. Thus $X_{t,i}, t \in F_i$ forms bounded martingale difference sequence. Using (10), with probability at least $1-\delta$, the sum of conditional variances can be upper bounded as follows

$$\begin{aligned}
\Sigma_T^2 = \sum_{t\in F_i} \mathbb{E}_t[(X_{t,i} - \mathbb{E}[X_{t,i}])^2] \le& \sqrt{\frac{T}{\beta_i^2} \sum_{t=M+1}^{T-1} \|\nabla_{t,i} - \bar{\nabla}_{t,i}\|_{\mathcal{H}_i}^2} \\
\le& \sqrt{\frac{12T}{\beta_i^2}\left[g_i(M,T)\mathcal{A}(\mathcal{I}_T, \kappa_i)\sqrt{\ln \frac{1}{\delta}} + 2D_i \ln \frac{1}{\delta}\right]} \\
\le& 6\sqrt{\frac{T g_i(M,T)}{\beta_i^2} \mathcal{A}(\mathcal{I}_T, \kappa_i) \ln \frac{1}{\delta}},
\end{aligned}$$

where $g_i(M,T) = \frac{M + D_i \ln \frac{T}{M}}{M}$ and $Z_{t,i} \ge \beta_i$.
Using Lemma 2, with probability at least $1-\delta$, we have

$$\begin{aligned}
\sum_{t\in F_i} \mathbb{I}_{b_{t,i}=1} \le& 6\sqrt{\frac{T}{\beta_i^2} g_i(M,T)\mathcal{A}(\mathcal{I}_T, \kappa_i) \ln \frac{1}{\delta}} + \frac{2}{3} \ln \frac{1}{\delta} \\
&+ \sqrt{12\sqrt{\frac{T}{\beta_i^2} g_i(M,T)\mathcal{A}(\mathcal{I}_T, \kappa_i) \ln \frac{1}{\delta}} \ln \frac{1}{\delta}} \\
\le& \frac{10}{\beta_i} \sqrt{T g_i(M,T)\mathcal{A}(\mathcal{I}_T, \kappa_i)} \ln^{\frac{3}{4}} \left(\frac{1}{\delta}\right) + \frac{2}{3} \ln \frac{1}{\delta},
\end{aligned}$$

in which we merge the lower order terms.

Summing over all of the above results, with probability at least $1-2\delta$,

$$\begin{aligned}|S_i| \leq & M + \sum_{t=M+1}^{T-1} \delta_t + \sum_{t \in F_i} \mathbb{I}_{b_{t,i}=1} \\ \leq & 4M \ln \frac{T}{M} \sqrt{\ln \frac{1}{\delta}} + \frac{4}{3} \ln \frac{1}{\delta} + \frac{10}{\beta_i} \sqrt{T g_i(M,T) \mathcal{A}(\mathcal{I}_T, \kappa_i)} \ln^{\frac{3}{4}} \left(\frac{1}{\delta}\right),\end{aligned}$$

which completes the proof.

Lemma 7. *At any round t, with probability at least $1-\delta$, the updating time of V is $M + M\ln\frac{T}{M} + \frac{2}{3}\ln\frac{1}{\delta} + \sqrt{2M\ln\frac{T}{M}\ln\frac{1}{\delta}}$ at most.*

Proof. The proof is based on Lemma 1. Let δ_t be a Bernoulli random variable satisfying $\mathbb{P}[\delta_t = 1] = \min\{1, \frac{M}{t}\}$. We consider the case $t > M$. At the end of the $(T-1)$-th round, the updating times of V equals $M + \sum_{t=M+1}^{T-1} \delta_t$. It is obvious that $\delta_1, \ldots, \delta_{T-1}$ are independent. Define a random variable

$$X_t = \delta_t - \mathbb{P}[\delta_t = 1].$$

It can be verified that $\mathbb{E}[X_t] = 0$ and $|X_t| \leq 1$. The sum of the variances is

$$\Sigma_{T-1}^2 = \sum_{t=M+1}^{T-1} \mathbb{E}[X_t^2] = \sum_{t=M+1}^{T-1} \mathbb{P}[\delta_t = 1] \leq M \sum_{t=M+1}^{T-1} \frac{1}{t} \leq M \ln \frac{T}{M+1}.$$

Using Lemma 1, with probability at least $1-\delta$,

$$\sum_{t=M+1}^{T-1} \delta_t \leq \sum_{t=M+1}^{T-1} \mathbb{P}[\delta_t = 1] + \frac{2}{3}\ln\frac{1}{\delta} + \sqrt{2M\ln\frac{T}{M+1}\ln\frac{1}{\delta}}.$$

Thus the updating time is upper bounded as follows

$$M + \sum_{t=M+1}^{T-1} \delta_t \leq M + M\ln\frac{T}{M} + \frac{2}{3}\ln\frac{1}{\delta} + \sqrt{2M\ln\frac{T}{M}\ln\frac{1}{\delta}},$$

which completes the proof.

8 Proof of Theorem 5

Proof. The proof is similar with that of Theorem 2. the regret can still be decomposed into $\mathcal{T}_{2,1}$ and $\mathcal{T}_{2,2}$. We will separately analyze the two terms.

8.1 Analyzing $\mathcal{T}_{2,1}$

We start from (9),

$$\mathcal{T}_{2,1} \leq \frac{1}{2\lambda_i}\|f\|^2_{\mathcal{H}_i} + \sum_{t\in E_i} \lambda_i \left[\|\tilde{\nabla}_{t,i} - \bar{\nabla}_{t,i}\|^2_{\mathcal{H}_i} - \|\bar{\nabla}_{t,i}\|^2_{\mathcal{H}_i}\right].$$

Recalling that

$$\tilde{\nabla}_{t,i} = \frac{\nabla_{t,i} - \bar{\nabla}_{t,i}}{\mathbb{P}[b_{t,i}=1]}\mathbb{I}_{b_{t,i}=1} + \bar{\nabla}_{t,i}, \quad \forall t \in E_i,$$
$$\mathbb{P}[b_{t,i}=1] = \frac{\|\nabla_{t,i} - \bar{\nabla}_{t,i}\|_{\mathcal{H}_i}}{Z_{t,i}},\ Z_{t,i} = \beta_i\left(\|\nabla_{t,i} - \bar{\nabla}_{t,i}\|_{\mathcal{H}_i} + \|\bar{\nabla}_{t,i}\|_{\mathcal{H}_i}\right).$$

Thus we can obtain, with probability at least $1-2\delta$,

$$\begin{aligned}\mathcal{T}_{2,1} - \frac{1}{2\lambda_i}\|f\|^2_{\mathcal{H}_i} \leq & \sum_{t\in E_i} \lambda_i \left[\frac{\|\nabla_{t,i} - \bar{\nabla}_{t,i}\|^2_{\mathcal{H}_i}\mathbb{I}_{b_{t,i}=1}}{(\mathbb{P}[b_{t,i}=1])^2} - \|\bar{\nabla}_{t,i}\|^2_{\mathcal{H}_i}\right]\\ \leq & \lambda_i \left[18\beta_i\sqrt{D_i T g_i(M,T)\mathcal{A}(\mathcal{I}_T,\kappa_i)\ln\frac{1}{\delta}}\right.\\ & \left. +6\beta_i^2 D_i^2 \ln\frac{1}{\delta} + 18 D_i^{\frac{3}{4}}\beta_i^{\frac{3}{2}} g_i(M,T)^{\frac{1}{4}} T^{\frac{1}{4}}\mathcal{A}(\mathcal{I}_T,\kappa_i)^{\frac{1}{4}}\ln^{\frac{3}{4}}\frac{1}{\delta}\right],\end{aligned}$$

where the last inequality comes from Lemma 8.

8.2 Analyzing $\mathcal{T}_{2,2}$

Using Lemma 9, we have, with probability at least $1-2\delta$,

$$\sum_{t\in E_i}\langle f_{t,i} - f, \nabla_{t,i} - \tilde{\nabla}_{t,i}\rangle \leq \frac{20}{3}U\beta_i\sqrt{D_i}\ln\frac{1}{\delta} + a_i\sqrt{\beta_i}UT^{\frac{1}{4}}\mathcal{A}(\mathcal{I}_T,\kappa_i)^{\frac{1}{4}}\ln^{\frac{3}{4}}\frac{1}{\delta},$$

where $a_i = 9D_i^{\frac{1}{4}} g_i^{\frac{1}{4}}(M,T)$.

8.3 The final regret

Combining $\mathcal{T}_{2,1}$ with $\mathcal{T}_{2,2}$ yields, with probability at least $1-4\delta$,

$$\begin{aligned}&\sum_{t=1}^{T}[\ell(f_{t,i}(\mathbf{x}_t),y_t) - \ell(f(\mathbf{x}_t),y_t)] \leq \frac{\|f\|^2_{\mathcal{H}_i}}{2\lambda_i} + 18\lambda_i\beta_i\sqrt{D_i T g_i(M,T)\mathcal{A}(\mathcal{I}_T,\kappa_i)\ln\frac{1}{\delta}}\\ &+ (6\lambda_i\beta_i^2 + 7U\beta_i)D_i^{\theta}\ln\frac{1}{\delta} + 9(2\lambda_i\beta_i^{\frac{3}{2}} + U\beta_i^{\frac{1}{2}})D_i^{\vartheta} g_i(M,T)^{\frac{1}{4}}T^{\frac{1}{4}}\mathcal{A}(\mathcal{I}_T,\kappa_i)^{\frac{1}{4}}\ln^{\frac{3}{4}}\frac{1}{\delta},\end{aligned}$$

where $\theta = \{2, \frac{1}{2}\}$, $\vartheta = \{\frac{3}{4}, \frac{1}{4}\}$. Using a union-of-events bound to $i = 1, \ldots, K$ completes the proof.

Lemma 8. *For a fixed $i \in [K]$, with probability at least $1-2\delta$,*

$$\sum_{t\in E_i} \frac{\|\nabla_{t,i}-\bar{\nabla}_{t,i}\|^2_{\mathcal{H}_i}\mathbb{I}_{b_{t,i}=1}}{(\mathbb{P}[b_{t,i}=1])^2} \leq 18\beta_i\sqrt{D_i T g(M,T)\mathcal{A}(\mathcal{I}_T,\kappa_i)\ln\frac{1}{\delta}} + 6\beta_i^2 D_i^2\ln\frac{1}{\delta} + 18D_i^{\frac{3}{4}}\beta_i^{\frac{3}{2}}g(M,T)^{\frac{1}{4}}T^{\frac{1}{4}}\mathcal{A}(\mathcal{I}_T,\kappa_i)^{\frac{1}{4}}\ln^{\frac{3}{4}}\frac{1}{\delta}.$$

Proof. For each $t \in E_i$, we define a random variable

$$X_{t,i} = \frac{\|\nabla_{t,i}-\bar{\nabla}_{t,i}\|^2_{\mathcal{H}_i}\mathbb{I}_{b_{t,i}=1}}{(\mathbb{P}[b_{t,i}=1])^2} - \frac{\|\nabla_{t,i}-\bar{\nabla}_{t,i}\|^2_{\mathcal{H}_i}}{\mathbb{P}[b_{t,i}=1]}.$$

Conditioned on $\{b_{\tau,i}\}_{\tau\in E_i}$, $\mathbb{E}_t[X_{t,i}]=0$ and $|X_{t,i}|\leq Z^2_{t,i}\leq 9\beta_i^2 D_i$. Using (10), with probability at least $1-\delta$, the sum of conditional variances satisfies

$$\begin{aligned}
\varSigma_T^2 = \sum_{t\in E_i}\mathbb{E}_t[X^2_{t,i}] &\leq \sum_{t\in E_i}\frac{\|\nabla_{t,i}-\bar{\nabla}_{t,i}\|^4_{\mathcal{H}_i}}{(\mathbb{P}[b_{t,i}=1])^3}\\
&\leq \sum_{t\in E_i} Z^3_{t,i}\|\nabla_{t,i}-\bar{\nabla}_{t,i}\|_{\mathcal{H}_i}\\
&\leq \beta_i^3(3\sqrt{D_i})^3\sqrt{12T\left[\frac{M+D_i\ln\frac{T}{M}}{M}\mathcal{A}(\mathcal{I}_T,\kappa_i)\sqrt{\ln\frac{1}{\delta}}+2D_i\ln\frac{1}{\delta}\right]}\\
&\leq 162\beta_i^3 D_i^{\frac{3}{2}}\sqrt{Tg_i(M,T)\mathcal{A}(\mathcal{I}_T,\kappa_i)\ln\frac{1}{\delta}}.
\end{aligned}$$

Similarly, we can obtain

$$\frac{\|\nabla_{t,i}-\bar{\nabla}_{t,i}\|^2_{\mathcal{H}_i}}{\mathbb{P}[b_{t,i}=1]} \leq 18\beta_i\sqrt{D_i T g_i(M,T)\mathcal{A}(\mathcal{I}_T,\kappa_i)\ln\frac{1}{\delta}}.$$

Using Lemma 2 yields the desired result.

Lemma 9. *For a fixed $i \in [K]$, with probability at least $1-2\delta$,*

$$\sum_{t\in E_i}\langle f_{t,i}-f,\nabla_{t,i}-\tilde{\nabla}_{t,i}\rangle \leq \frac{20}{3}U\beta_i\sqrt{D_i}\ln\frac{1}{\delta} + a_i\sqrt{\beta_i}UT^{\frac{1}{4}}\mathcal{A}(\mathcal{I}_T,\kappa_i)^{\frac{1}{4}}\ln^{\frac{3}{4}}\frac{1}{\delta},$$

where $a_i = 9D_i^{\frac{1}{4}}g_i^{\frac{1}{4}}(M,T)$.

Proof. The proof is similar with that of Lemma 6. For each $t \in E_i$, we define a random variable $X_{t,i}$,

$$X_{t,i} = \langle f_{t,i}-f,\nabla_{t,i}-\tilde{\nabla}_{t,i}\rangle = (\varXi_t(f_{t,i})-\varXi_t(f))\left[1-\frac{\mathbb{I}_{b_{t,i}=1}}{\mathbb{P}[b_{t,i}=1]}\right],$$

where $\Xi_t(f) = \sum_{(\mathbf{x},y)\in V} \frac{yf(\mathbf{x})}{|V|} - y_t f(\mathbf{x}_t)$. Under the condition of $\{b_{\tau,i}\}_{\tau\in E_i,\tau<t}$, we have $\mathbb{E}_t[X_{t,i}] = 0$. Besides,

$$|X_{t,i}| = |\langle f_{t,i} - f, \nabla_{t,i} - \tilde{\nabla}_{t,i}\rangle| \leq 2U\left(2\sqrt{D_i} + 3\beta_i\sqrt{D_i}\right) \leq 10U\beta_i\sqrt{D_i},$$

where we use the fact $\beta_i \geq 1$. Therefore, $X_{t,i}, t \in E_i$ forms bounded martingale difference sequence. Using inequality (10), we obtain, with probability at least $1-\delta$, the sum of conditional variances satisfies

$$\begin{aligned}
\Sigma_T^2 =& \sum_{t\in E_i} \mathbb{E}_t\left[(\Xi_t(f_{t,i}) - \Xi_t(f))^2\left(1 - \frac{\mathbb{I}_{b_{t,i}=1}}{\mathbb{P}[b_{t,i}=1]}\right)^2\right] \\
\leq& \sum_{t\in E_i} \frac{\langle f_{t,i} - f, \nabla_{t,i} - \bar{\nabla}_{t,i}\rangle^2_{\mathcal{H}_i}}{\mathbb{P}[b_{t,i}=1]} \\
\leq& 6\beta_i\sqrt{D_i}U^2 \sum_{t\in E_i} \left\|\nabla_{t,i} - \bar{\nabla}_{t,i}\right\|_{\mathcal{H}_i} \\
\leq& 36\beta_i U^2\sqrt{TD_i g_i(M,T)\mathcal{A}(\mathcal{I}_T,\kappa_i)\ln\frac{1}{\delta}}.
\end{aligned}$$

Finally, using Lemma 2 yields the desired result.

9 Proof of Theorem 6

Proof. We restate the regret decomposition (4) as follows,

$$\mathrm{Reg}_T(\mathbb{H}_i) = \underbrace{\sum_{t=1}^T [\ell(f_t(\mathbf{x}_t), y_t) - \ell(f_{t,i}(\mathbf{x}_t), y_t)]}_{\mathcal{T}_1} + \underbrace{\sum_{t=1}^T [\ell(f_{t,i}(\mathbf{x}_t), y_t) - \ell(f(\mathbf{x}_t), y_t)]}_{\mathcal{T}_2},$$

in which $\mathcal{T}_1$ can still be bounded by (5), and $\mathcal{T}_2$ has been shown in Theorem 5. Summing over $\mathcal{T}_1$ and $\mathcal{T}_2$ yields that, with probability at least $1-5\delta$, we have

$$\begin{aligned}
\mathrm{Reg}_T(\mathbb{H}_i) =& \tilde{O}\left(\sqrt{(1+U\sqrt{D})\min_{i\in[K]}\left[\sum_{t=1}^T \ell(f_{t,i}(\mathbf{x}_t), y_t)\right]\ln(K)\ln\frac{\ln(2T)}{\delta}}\right) \\
&+ O\left((\|f\|^2_{\mathcal{H}_i} + U)\beta_i^{\frac{1}{2}} T^{\frac{1}{4}}\mathcal{A}(\mathcal{I}_T,\kappa_i)^{\frac{1}{4}}\ln^{\frac{3}{4}}\frac{K}{\delta}\right) \\
=& \tilde{O}\left(\sqrt{L_T(f)\ln(K)\ln\frac{\ln(2T)}{\delta}} + (\|f\|^2_{\mathcal{H}_i} + U)\sqrt{\beta_i K}T^{\frac{1}{4}}\mathcal{A}(\mathcal{I}_T,\kappa_i)^{\frac{1}{4}}\ln^{\frac{3}{4}}\frac{K}{\delta}\right).
\end{aligned}$$

Replacing f with $f_i^* = \mathrm{argmin}_{f\in\mathbb{H}_i} L_T(f)$ completes the proof.

Algorithm 2 BEA$_2$OKS

Input: $\lambda_i, i = 1, \ldots, K$, D, U, β_i, B, M.
Initialization: $\forall \kappa_i \in \mathcal{K}$, $f'_{0,i} = 0, S_i = \emptyset$, $V = \emptyset$.
for $t = 1, 2, \ldots, T$ **do**
  Select a kernel $\kappa_{I_t} \sim \mathbf{p}_t$ ($\mathbf{p}_t$ is output by $\mathcal{E}(K)$),
  Compute $\bar{\nabla}_{t,I_t} = \frac{-1}{|V|} \sum_{(\mathbf{x},y)\in V} y\kappa_{I_t}(\mathbf{x}, \cdot)$,
  Update f_{t,I_t}, and output prediction $\hat{y}_t = \mathrm{sign}(f_{t,I_t}(\mathbf{x}_t))$,
  for $\kappa_i \in \mathcal{K}$ **do**
    if $|S_i| = B$ **then**
      $S_i = \emptyset$,
      $f'_{t-1,i} = 0$,
    end if
    if $\kappa_i \neq \kappa_{I_t}$, **then** update $f_{t,i}$ according to the first mirror updating,
    if $y_t f_{t,i}(\mathbf{x}_t) < 1$ **then**
      Compute $\nabla_{t,i} = -y_t\kappa_i(\mathbf{x}_t, \cdot)$,
      Compute $\mathbb{P}[b_{t,i} = 1] = \|\nabla_{t,i} - \bar{\nabla}_{t,i}\|_{\mathcal{H}_i}/Z_{t,i}$,
      Sample $b_{t,i} \sim \mathrm{Ber}(\mathbb{P}[b_{t,i} = 1], 1)$,
      if $b_{t,i} = 1$, **then** $S_i = S_i \cup (\mathbf{x}_t, y_t)$,
      Compute $\tilde{\nabla}_{t,i} = \frac{\nabla_{t,i}\mathbb{I}_{b_{t,i}=1}}{\mathbb{P}[b_{t,i}=1]} + \left(1 - \frac{\mathbb{I}_{b_{t,i}=1}}{\mathbb{P}[b_{t,i}=1]}\right)\bar{\nabla}_{t,i}$,
      Update $f'_{t,i}$ according to the second mirror updating,
      Compute $c_{t,i} = \frac{1}{1+U\sqrt{D}} \max\{0, 1 - y_t f_{t,i}(\mathbf{x}_t)\}$,
    end if
  end for
  Send $\mathbf{c}_t = (c_{t,1}, \ldots, c_{t,K})$ to $\mathcal{E}(K)$,
  Update Reservoir V (line 16-18 in Algorithm 1),
end for

10 Pseudo-code of BEA$_2$OKS

We give the pseudo-code of BEA$_2$OKS in Algorithm 2

11 Incremental Computing

In this section, we show some incremental computing procedures.

11.1 Computing $\|\bar{\nabla}_{t+1,i}\|^2_{\mathcal{H}_i}$

For the case $t \leq M$.

$$\|\bar{\nabla}_{t+1,i}\|^2_{\mathcal{H}_i} = \frac{(t-1)^2}{t^2}\|\bar{\nabla}_{t,i}\|^2_{\mathcal{H}} + \frac{\kappa_i(\mathbf{x}_t, \mathbf{x}_t)}{t^2} + \frac{2}{t^2}\sum_{\tau=1}^{t-1} y_t y_\tau \kappa_i(\mathbf{x}_\tau, \mathbf{x}_t).$$

For $t > M$, assuming that the removed example is $(\mathbf{x}_{j_t}, y_{j_t})$.

$$\begin{aligned}&\|\bar{\nabla}_{t+1,i}\|^2_{\mathcal{H}_i}\\ =&\|\bar{\nabla}_{t,i}\|^2_{\mathcal{H}} + \frac{2(1 - y_{j_t} y_t \kappa(\mathbf{x}_{j_t}, \mathbf{x}_t))}{M^2} - \frac{2}{M^2}\sum_{(\mathbf{x}_\tau, y_\tau)} [y_\tau y_{j_t}\kappa(\mathbf{x}_\tau, \mathbf{x}_{j_t}) - y_\tau y_t \kappa(\mathbf{x}_\tau, \mathbf{x}_t)].\end{aligned}$$

Next we compute $\|\nabla_{t,i} - \bar{\nabla}_{t,i}\|^2_{\mathcal{H}_i}$.

$$\|\nabla_{t,i} - \bar{\nabla}_{t,i}\|^2_{\mathcal{H}_i} = \kappa_i(\mathbf{x}_t, \mathbf{x}_t) + \|\bar{\nabla}_{t,i}\|^2_{\mathcal{H}_i} - 2 \sum_{(\mathbf{x}_\tau, y_\tau)\in V} \frac{y_t y_\tau \kappa_i(\mathbf{x}_\tau, \mathbf{x}_t)}{M}.$$

The projection operation needs to compute $\|\overline{f}_{t,i}\|_{\mathcal{H}_i}$ and $\|\overline{f'}_{t,i}\|_{\mathcal{H}_i}$. In the next subsections, we will show how to execute incremental computing.

11.2 Computing $\|\overline{f}_{t,i}\|_{\mathcal{H}_i}$

We show the computing procedure for B(AO)$_2$KS and EA$_2$OKS, respectively.

B(AO)$_2$KS We only need to compute $\|\overline{f}_{t,i}\|$ for $\kappa_i \in \{\kappa_{I_t}, \kappa_{J_t}\}$. Recalling that $\overline{f}_{t,i} = f'_{t-1,i} - \lambda_i \bar{\nabla}_{t,i}$. Thus, we have

$$\|\overline{f}_{t,i}\|^2 = \|f'_{t-1,i}\|^2 + \lambda_i^2 \|\bar{\nabla}_{t,i}\|^2 - 2\lambda_i \langle f'_{t-1,i}, \bar{\nabla}_{t,i}\rangle. \tag{11}$$

EA$_2$OKS For all $\kappa_i \in \mathcal{K}$, we need to compute $\|\overline{f}_{t,i}\|$. which is same with (11).

11.3 Computing $\|f'_{t,i}\|$

B(AO)$_2$KS Recalling that

$$\overline{f'}_{t,i} = f'_{t-1,i} - \lambda_i \tilde{\nabla}_{t,i}, \quad \text{and} \quad \tilde{\nabla}_{t,i} = \frac{\nabla_{t,i} - \bar{\nabla}_{t,i}}{\mathbb{P}[i = J_t]} \mathbb{I}_{i=J_t} + \bar{\nabla}_{t,i},$$

where $\mathbb{P}[i = J_t] = \frac{1}{K}$. For $\kappa_i \neq \kappa_{J_t}$, we have

$$\|\overline{f'}_{t,i}\|^2_{\mathcal{H}_i} = \|f'_{t-1,i} - \lambda_i \bar{\nabla}_{t,i}\|^2_{\mathcal{H}_i} = \|f'_{t-1,i}\|^2_{\mathcal{H}_i} + \lambda_i^2 \|\bar{\nabla}_{t,i}\|^2 - 2\lambda_i \langle f'_{t-1,i}, \bar{\nabla}_{t,i}\rangle.$$

For $\kappa_i = \kappa_{J_t}$, we consider two cases.

(1) $y_t f_{t,i}(\mathbf{x}_t) \geq 1$, i.e, $\nabla_{t,i} = 0$.
In this case, $\overline{f'}_{t,i} = f'_{t-1,i} - (1-K)\lambda_i \bar{\nabla}_{t,i}$, thus

$$\|\overline{f'}_{t,i}\|^2_{\mathcal{H}_i} = \|f'_{t-1,i}\|^2_{\mathcal{H}_i} + (1-K)^2 \lambda_i^2 \|\bar{\nabla}_{t,i}\|^2_{\mathcal{H}_i} - 2\lambda_i (1-K) \langle f'_{t-1,i}, \bar{\nabla}_{t,i}\rangle.$$

(2) $y_t f_{t,i}(\mathbf{x}_t) < 1$, i.e, $\nabla_{t,i} \neq 0$.
In this case, $\overline{f'}_{t,i} = f'_{t-1,i} - \lambda_i (1-K) \bar{\nabla}_{t,i} - K\lambda_i \nabla_{t,i}$. Thus

$$\begin{aligned}
\|\overline{f'}_{t,i}\|^2_{\mathcal{H}_i} =& \left\|f'_{t-1,i} - \lambda_i(1-K)\bar{\nabla}_{t,i}\right\|^2_{\mathcal{H}_i} + K^2\lambda_i^2 \|\nabla_{t,i}\|^2_{\mathcal{H}_i} \\
&- 2K\lambda_i \langle f'_{t-1,i}, \nabla_{t,i}\rangle + 2\lambda_i^2 K(1-K)\langle \bar{\nabla}_{t,i}, \nabla_{t,i}\rangle \\
=& \|f'_{t-1,i}\|^2_{\mathcal{H}_i} - 2\lambda_i(1-K)\langle f'_{t-1,i}, \bar{\nabla}_{t,i}\rangle + \lambda_i^2 (1-K)^2 \|\bar{\nabla}_{t,i}\|^2_{\mathcal{H}_i} \\
&+ K^2\lambda_i^2 \|\nabla_{t,i}\|^2_{\mathcal{H}_i} - 2K\lambda_i \langle f'_{t-1,i}, \nabla_{t,i}\rangle + 2\lambda_i^2 K(1-K)\langle \bar{\nabla}_{t,i}, \nabla_{t,i}\rangle.
\end{aligned}$$

$\mathbf{EA_2OKS}$ Recalling that $\overline{f'}_{t,i} = f'_{t-1,i} - \lambda_i \tilde{\nabla}_{t,i}$ where

$$\tilde{\nabla}_{t,i} = \nabla_{t,i}\mathbb{I}_{y_t f_{t,i}(\mathbf{x}_t)\geq 1} + \left[\frac{\nabla_{t,i} - \bar{\nabla}_{t,i}}{\mathbb{P}[b_{t,i}=1]}\mathbb{I}_{b_{t,i}=1} + \bar{\nabla}_{t,i}\right]\mathbb{I}_{y_t f_{t,i}(\mathbf{x}_t)<1}, \forall i = 1,\ldots,K.$$

For all $\kappa_i \in \mathcal{K}$, we consider three cases.

(1) $y_t f_{t,i}(\mathbf{x}_t) \geq 1$, i.e, $\nabla_{t,i} = 0$. In this case, we have $f'_{t,i} = f'_{t-1,i}$.
(2) $y_t f_{t,i}(\mathbf{x}_t) < 1$, i.e, $\nabla_{t,i} \neq 0$.
We further consider two cases.
Case 1: $b_{t,i} = 0$. We have $\overline{f'}_{t,i} = f'_{t-1,i} - \lambda_i \bar{\nabla}_{t,i}$. Thus

$$\|\overline{f'}_{t,i}\|^2_{\mathcal{H}_i} = \|f'_{t-1,i}\|^2_{\mathcal{H}_i} + \lambda_i^2\|\bar{\nabla}_{t,i}\|^2_{\mathcal{H}_i} - 2\lambda_i\langle f'_{t-1,i}, \bar{\nabla}_{t,i}\rangle.$$

Case 2: $b_{t,i} = 1$. We have

$$\overline{f'}_{t,i} = f'_{t-1,i} - \lambda_i\left(1 - \frac{1}{\mathbb{P}[b_{t,i}=1]}\right)\bar{\nabla}_{t,i} - \frac{\lambda_i\nabla_{t,i}}{\mathbb{P}[b_{t,i}=1]}.$$

Thus

$$\begin{aligned}\|\overline{f'}_{t,i}\|^2_{\mathcal{H}_i} =& \left\|f'_{t-1,i} - \lambda_i\left(1 - \frac{1}{\mathbb{P}[b_{t,i}=1]}\right)\bar{\nabla}_{t,i}\right\|^2_{\mathcal{H}_i} + \frac{\lambda_i^2\|\nabla_{t,i}\|^2_{\mathcal{H}_i}}{\mathbb{P}[b_{t,i}=1])^2}\\ &- 2\lambda_i\frac{\langle f'_{t-1,i}, \nabla_{t,i}\rangle}{\mathbb{P}[b_{t,i}=1]} + 2\lambda_i^2\left(1 - \frac{1}{\mathbb{P}[b_{t,i}=1]}\right)\frac{\langle\bar{\nabla}_{t,i}, \nabla_{t,i}\rangle}{\mathbb{P}[b_{t,i}=1]},\end{aligned}$$

in which

$$\begin{aligned}&\left\|f'_{t-1,i} - \lambda_i\left(1 - \frac{1}{\mathbb{P}[b_{t,i}=1]}\right)\bar{\nabla}_{t,i}\right\|^2_{\mathcal{H}_i} = \|f'_{t-1,i}\|^2_{\mathcal{H}_i}\\ &- 2\lambda_i\left[1 - \frac{1}{\mathbb{P}[b_{t,i}=1]}\right]\langle f'_{t-1,i}, \bar{\nabla}_{t,i}\rangle + \lambda_i^2\left[1 - \frac{1}{\mathbb{P}[b_{t,i}=1]}\right]^2\|\bar{\nabla}_{t,i}\|^2_{\mathcal{H}_i}\end{aligned}$$

It can be find that the key is to compute $\langle f'_{t-1,i}, \bar{\nabla}_{t,i}\rangle$ incrementally.

11.4 Computing $\langle \boldsymbol{f'_{t-1,i}}, \bar{\boldsymbol{\nabla}}_{\boldsymbol{t,i}}\rangle$

$\mathbf{B(AO)_2KS}$ If $2 \leq t \leq M+1$, we just compute $\langle f'_{t-1,i}, \bar{\nabla}_{t,i}\rangle$ directly. The time complexity is of order $O(M^2)$. Since $M = O(\ln T)$ or M is a small constant, the time complexity can be omitted. Next we consider the case $t \geq M+2$. Let $\delta_{t-1} \in \{0,1\}$. If the reservoir is updated after round $t-1$, then $\delta_{t-1} = 1$. Otherwise, $\delta_{t-1} = 0$. If $\delta_{t-1} = 1$, we still compute $\langle f'_{t-1,i}, \bar{\nabla}_{t,i}\rangle$ directly. The reason is that the times of $\delta_t = 1$ for $t > M+2$ is upper bounded by $O(M\ln(T))$, which is a lower order term. Next we compute $\langle f'_{t-1,i}, \bar{\nabla}_{t,i}\rangle$ incrementally when $\delta_{t-1} = 0$. Note that, in this case, we have $\bar{\nabla}_{t,i} = \bar{\nabla}_{t-1,i}$. We consider three different cases. Recalling that $f'_{t-1,i} = \min\left\{1, \frac{1}{\|\overline{f'}_{t-1,i}\|_{\mathcal{H}_i}}U\right\}\overline{f'}_{t-1,i}$. For all of the following three cases, we assume that $\|\overline{f'}_{t-1,i}\|_{\mathcal{H}_i} \leq U$.

(1) $\kappa_i \neq \kappa_{J_{t-1}}$. We have $f'_{t-1,i} = f'_{t-2,i} - \lambda_i \bar{\nabla}_{t-1,i}$, and

$$\langle f'_{t-1,i}, \bar{\nabla}_{t,i} \rangle = \langle f'_{t-2,i} - \lambda_i \bar{\nabla}_{t-1,i}, \bar{\nabla}_{t-1,i} \rangle = \langle f'_{t-2,i}, \bar{\nabla}_{t-1,i} \rangle - \lambda_i \|\bar{\nabla}_{t-1,i}\|^2_{\mathcal{H}_i}.$$

(2) $\kappa_i = \kappa_{J_{t-1}}$ and $y_{t-1} f_{t-1,i}(\mathbf{x}_{t-1}) < 1$.

$$\begin{aligned}\langle f'_{t-1,i}, \bar{\nabla}_{t,i} \rangle &= \left\langle f'_{t-2,i} - \lambda_i K \nabla_{t-1,i} - \lambda_i (1-K) \bar{\nabla}_{t-1,i}, \bar{\nabla}_{t-1,i} \right\rangle \\ &= \left\langle f'_{t-2,i}, \bar{\nabla}_{t-1,i} \right\rangle - \lambda_i K \langle \nabla_{t-1,i}, \bar{\nabla}_{t-1,i} \rangle - \lambda_i (1-K) \|\bar{\nabla}_{t-1,i}\|^2_{\mathcal{H}_i}.\end{aligned}$$

(3) $\kappa_i = \kappa_{J_{t-1}}$ and $y_{t-1} f_{t-1,i}(\mathbf{x}_{t-1}) \geq 1$.

$$\begin{aligned}\langle f'_{t-1,i}, \bar{\nabla}_{t,i} \rangle &= \langle f'_{t-2,i} - \lambda_i (1-K) \bar{\nabla}_{t-1,i}, \bar{\nabla}_{t-1,i} \rangle \\ &= \langle f'_{t-2,i}, \bar{\nabla}_{t-1,i} \rangle - \lambda_i (1-K) \|\bar{\nabla}_{t-1,i}\|^2_{\mathcal{H}_i}.\end{aligned}$$

If $\|\overline{f'}_{t-1,i}\|_{\mathcal{H}_i} > U$, then we just multiply a factor $\frac{1}{\|\overline{f'}_{t-1,i}\|_{\mathcal{H}_i}} U$.

EA$_\mathbf{2}$OKS We also consider the case $\delta_{t-1} = 0$. Besides, if $t-1 \notin E_i$, then $f'_{t-1,i} = f'_{t-2,i}$, which implies $\langle f'_{t-1,i}, \bar{\nabla}_{t,i} \rangle = \langle f'_{t-2,i}, \bar{\nabla}_{t-1,i} \rangle$. Next we consider $t-1 \in E_i$. For the following two cases, we still assume $\|\overline{f'}_{t-1,i}\|_{\mathcal{H}_i} \leq U$.

(1) $b_{t-1,i} = 0$. In this case, $\tilde{\nabla}_{t-1,i} = \bar{\nabla}_{t-1,i}$. Thus we have

$$\langle f'_{t-1,i}, \bar{\nabla}_{t,i} \rangle = \langle f'_{t-2,i} - \lambda_i \tilde{\nabla}_{t-1,i}, \bar{\nabla}_{t-1,i} \rangle = \langle f'_{t-2,i}, \bar{\nabla}_{t-1,i} \rangle - \lambda_i \|\bar{\nabla}_{t-1,i}\|^2_{\mathcal{H}_i}.$$

(2) $b_{t-1,i} = 1$, $y_{t-1} f_{t-1,i}(\mathbf{x}_{t-1}) < 1$.

$$\begin{aligned}&\langle f'_{t-1,i}, \bar{\nabla}_{t,i} \rangle \\ =& \left\langle f'_{t-2,i} - \lambda_i \frac{\nabla_{t-1,i}}{\mathbb{P}[b_{t-1,i} = 1]} - \lambda_i \left(1 - \frac{1}{\mathbb{P}[b_{t-1,i} = 1]}\right) \bar{\nabla}_{t-1,i}, \bar{\nabla}_{t-1,i} \right\rangle \\ =& \left\langle f'_{t-2,i}, \bar{\nabla}_{t-1,i} \right\rangle - \lambda_i \frac{\langle \nabla_{t-1,i}, \bar{\nabla}_{t-1,i} \rangle}{\mathbb{P}[b_{t-1,i} = 1]} - \lambda_i \left(1 - \frac{1}{\mathbb{P}[b_{t-1,i} = 1]}\right) \|\bar{\nabla}_{t-1,i}\|^2_{\mathcal{H}_i}.\end{aligned}$$

If $\|\overline{f'}_{t-1,i}\|_{\mathcal{H}_i} > U$, then we just multiply a factor $\frac{1}{\|\overline{f'}_{t-1,i}\|_{\mathcal{H}_i}} U$.

References

1. Cesa-Bianchi, N., Lugosi, G.: Prediction, Learning, and Games. Cambridge University Press (2006)
2. Chiang, C., Yang, T., Lee, C., Mahdavi, M., Lu, C., Jin, R., Zhu, S.: Online optimization with gradual variations. In: Proceedings of the 25th Annual Conference on Learning Theory. pp. 6.1–6.20 (2012)
3. Hazan, E., Kale, S.: Better algorithms for benign bandits. In: Proceedings of the Twentieth Annual ACM-SIAM Symposium on Discrete Algorithms. pp. 38–47 (2009)
4. Lykouris, T., Sridharan, K., Tardos, É.: Small-loss bounds for online learning with partial information. In: Proceedings of the 31st Conference on Learning Theory. pp. 979–986 (2018)
5. Tropp, J.A.: An introduction to matrix concentration inequalities. Foundations and Trends in Machine Learning **8**(1-2), 1–230 (2015)