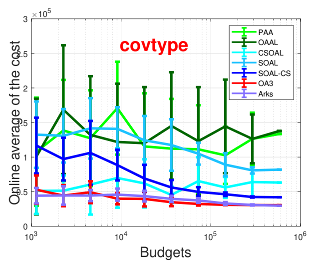
Supplementary Material: Cost-aware Online Kernel Learning on Imbalanced Data

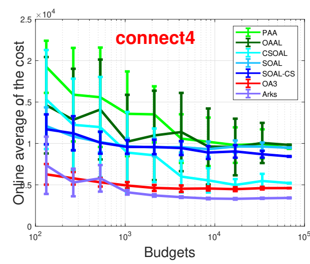
ACM Reference Format:
. 2021. Supplementary Material: Cost-aware Online Kernel Learning on Imbalanced Data. In *SIGKDD '21: Cost-aware Online Kernel Learning on Imbalanced Data, August 14–18, 2021, Singapore, Singapore.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/1122445.1122456

Experimental Results

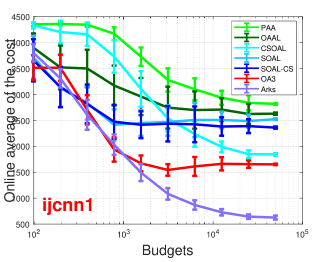
The experimental results on evaluation of cost with varying budgets are shown in Figure 1. The proposed algorithm Arks achieve a promising result on most of varied setting, and demonstrate the effectiveness and robustness of this proposed algorithm.



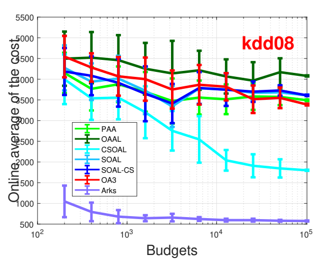
(a) Covtype



(b) Connect4



(c) Ijcnn1



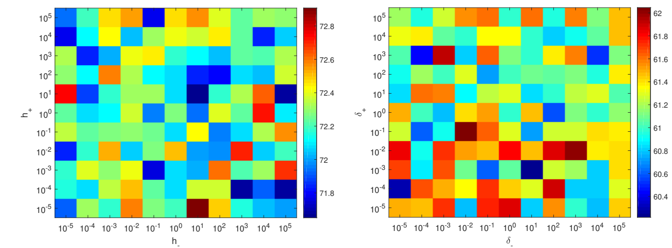
(d) KDDCUP08

Figure 1: Evaluation of cost with varying budgets.

The performance with Query biases is shown on Figure 2. For Connect4, better performances are obtained when $h_+ \in \{10^4, 10^5\}$ and $h_- \in \{10^{-2}, 10^{-1}, 1\}$ (i.e., upper middle part in Figure 2 (a)). With an imbalanced ratio in Connect4 (i.e., #Pos:#Neg =1:2), the model can boost the perform via biasing to the minority class. For the balanced data Covtype (i.e., #Pos:#Neg =1:1.1), the better performances are usually achieved under symmetric small biases (i.e., bottom left part in Figure 2 (b)). This observation also indicates that

SIGKDD '21, August 14–18, 2021, Singapore, Singapore

ACM ISBN 978-1-4503-XXXX-X/21/06...$15.00
https://doi.org/10.1145/1122445.1122456



(a) Connect4 (b) Covtype

Figure 2: Performance of Sum with varying Query Biases.

this algorithm could perform decently via querying a small amount of labels.

Proof of Theorem

PROPOSITION 1. *For any $t > 1$, $\overline{\Phi}_t = [\overline{\phi}_1, \cdots, \overline{\phi}_t] \in \mathbb{R}^{D\times t}$, $\overline{\mathbf{K}}_t = \overline{\Phi}_t^\top \overline{\Phi} \in \mathbb{R}^{t\times t}$, $\overline{\mathbf{y}}_t = [\overline{y}_1, \cdots, \overline{y}_t]^\top$ and $b > 0$. Then the problem (5) could be solved with an optimal solution,*

$$\mathbf{w}_t = \overline{\Phi}_{t-1}\left(\overline{\mathbf{K}}_{t-1} + b\mathbf{I}\right)^{-1}\overline{\mathbf{y}}_{t-1}, \tag{1}$$

where the predicted label $\widehat{y}_t = sgn(\mathbf{w}_t^\top \phi_t)$ can be computed without explicit representation ϕ_t,

$$\widehat{y}_t = sgn\left(\overline{\mathbf{y}}_{t-1}^\top\left(\overline{\mathbf{K}}_{t-1} + b\mathbf{I}\right)^{-1}\overline{\mathbf{k}}_{[t-1],t}\right).$$

PROOF. Given that $\overline{\Phi}_T = [\overline{\phi}_1, \cdots, \overline{\phi}_T] \in \mathbb{R}^{D\times T}$ with $\overline{\phi}_t = \sqrt{a_t}\phi_t$ and $\overline{\mathbf{y}}_T = [\overline{y}_1, \cdots, \overline{y}_T] \in \mathbb{R}^T$ with $\overline{y}_t = \sqrt{a_t}y_t$, we have

$$\begin{aligned}
&G_T(\mathbf{w}) = b\|\mathbf{w}\|^2 + \sum_{t=1}^{T} a_t(y_t - \mathbf{w}^\top\phi_t)^2 \\
&= b\|\mathbf{w}\|^2 + \sum_{t=1}^{T} a_t\left(y_t^2 - 2\left(y_t\mathbf{w}^\top\phi_t\right) + \mathbf{w}^\top\phi_t\phi_t^\top\mathbf{w}\right) \\
&= \mathbf{w}^\top\left(b\mathbf{I} + \sum_{t=1}^{T} a_t\phi_t\phi_t^\top\right)\mathbf{w} - 2\mathbf{w}^\top\left(\sum_{t=1}^{T} a_t\phi_t y_t\right) + \sum_{t=1}^{T} a_t y_t^2 \\
&= \mathbf{w}^\top(b\mathbf{I} + \overline{\Phi}_T\overline{\Phi}_T^\top)\mathbf{w} - 2\mathbf{w}^\top(\overline{\Phi}_T\overline{\mathbf{y}}_T) + \overline{\mathbf{y}}_T^\top\overline{\mathbf{y}}_T.
\end{aligned}$$

Then follows that $\nabla G_T(\mathbf{w}) = 2(b\mathbf{I} + \overline{\Phi}_T\overline{\Phi}_T^\top)\mathbf{w} - 2\overline{\Phi}_T\overline{\mathbf{y}}_T$, $\nabla^2 G_T(\mathbf{w}) = 2(b\mathbf{I} + \overline{\Phi}_T\overline{\Phi}_T^\top) \succ \mathbf{0}$. Thus $G_T(\mathbf{w})$ is convex and it is minimal if $\nabla G_T(\mathbf{w}) = (b\mathbf{I} + \overline{\Phi}_T\overline{\Phi}_T^\top)\mathbf{w} - \overline{\Phi}_T\overline{\mathbf{y}}_T = 0$ with

$$\mathbf{w} = (b\mathbf{I} + \overline{\Phi}_T\overline{\Phi}_T^\top)^{-1}\overline{\Phi}_T\overline{\mathbf{y}}_T.$$

For any $\overline{\Phi}_T = [\overline{\phi}_1, \ldots, \overline{\phi}_T] \in \mathbb{R}^{D\times T}$ matrix and $b > 1$,

$$\overline{\Phi}_T\overline{\Phi}_T^\top(\overline{\Phi}_T\overline{\Phi}_T^\top + b\mathbf{I}_D)^{-1} = \overline{\Phi}_T(\overline{\Phi}_T^\top\overline{\Phi}_T + b\mathbf{I}_T)^{-1}\overline{\Phi}_T^\top = \overline{\Phi}_T(\overline{\mathbf{K}}_T + b\mathbf{I}_T)^{-1}\overline{\Phi}_T^\top$$

With above equation, we have

$$(\overline{\Phi}_T\overline{\Phi}_T^\top + b\mathbf{I}_D)^{-1} = \frac{1}{b}b\mathbf{I}_D(\overline{\Phi}_T\overline{\Phi}_T^\top + b\mathbf{I}_D)^{-1}$$
$$=\frac{1}{b}\left(\overline{\Phi}_T\overline{\Phi}_T^\top + b\mathbf{I}_D - \overline{\Phi}_T\overline{\Phi}_T^\top\right)\left(\overline{\Phi}_T\overline{\Phi}_T^\top + b\mathbf{I}_D\right)^{-1}$$
$$=\frac{1}{b}\left(\mathbf{I}_D - \overline{\Phi}_T\overline{\Phi}_T^\top\left(\overline{\Phi}_T\overline{\Phi}_T^\top + b\mathbf{I}_D\right)^{-1}\right) = \frac{1}{b}\left(\mathbf{I}_D - \overline{\Phi}_T\left(\overline{\mathbf{K}}_T + b\mathbf{I}_T\right)^{-1}\overline{\Phi}_T^\top\right)$$

Substituting the two conclusions, we obtain

$$\begin{aligned}
\mathbf{w}_{T+1} &= (b\mathbf{I}_D + \overline{\Phi}_T\overline{\Phi}_T^\top)^{-1}\overline{\Phi}_T\overline{\mathbf{y}}_T \\
&= \frac{1}{b}\left(\mathbf{I}_D - \overline{\Phi}_T\left(\overline{\mathbf{K}}_T + b\mathbf{I}_T\right)^{-1}\overline{\Phi}_T^\top\right)\overline{\Phi}_T\overline{\mathbf{y}}_T \\
&= \frac{1}{b}\overline{\Phi}_T\left(\overline{\mathbf{y}}_T - \left(\overline{\mathbf{K}}_T + b\mathbf{I}_T\right)^{-1}\overline{\mathbf{K}}_T\overline{\mathbf{y}}_T\right) \\
&= \frac{1}{b}\overline{\Phi}_T\left(\mathbf{I}_T - \left(\overline{\mathbf{K}}_T + b\mathbf{I}_T\right)^{-1}\overline{\mathbf{K}}_T\right)\overline{\mathbf{y}}_T \\
&= \frac{1}{b}\overline{\Phi}_T\left(\left(\overline{\mathbf{K}}_T + b\mathbf{I}_T\right)^{-1}\left(\overline{\mathbf{K}}_T + b\mathbf{I}_T - \overline{\mathbf{K}}_T\right)\right)\overline{\mathbf{y}}_T \\
&= \overline{\Phi}_T\left(\overline{\mathbf{K}}_T + b\mathbf{I}_T\right)^{-1}\overline{\mathbf{y}}_T.
\end{aligned}$$

Substituting the $\mathbf{w}_T$ back to $\widehat{y}_T = \mathrm{sgn}(\mathbf{w}_T^\top\phi_T)$, we finish the proof. □

Proof on Lemma 1

LEMMA 1. *Let $(\phi_1, y_1), \cdots, (\phi_T, y_T)$ be a sequence of input samples, where $\phi_t \in \mathcal{H}$ and $y_t \in \{\pm 1\}$ for all t. The model parameter $\mathbf{w}_t$ is learned by Eq. (1) with $b > 0$. For any $\mathbf{w} \in \mathcal{H}$, it satisfies:*

$$\begin{aligned}
&\sum_{t=1}^T a_t(y_t - \mathbf{w}_t^\top\phi_t)^2 - \sum_{t=1}^T a_t(y_t - \mathbf{w}^\top\phi_t)^2 \\
&\le b\|\mathbf{w}\|^2 + 2\alpha(1 + C^2)\log\det(\frac{1}{b}\overline{\mathbf{K}}_T + \mathbf{I}),
\end{aligned} \tag{2}$$

where $\alpha = \frac{\mu_p T_n}{\mu_n T_p}$ or $\frac{c_p}{c_n}$ and $|\mathbf{w}_t^\top\phi_t| \le C$ for $t \in [T]$.

PROOF. Defined that $\mathbf{A}_T = \left(b\mathbf{I} + \overline{\Phi}_T\overline{\Phi}_T^\top\right)$ with $\mathbf{A}_T = \mathbf{A}_{T-1} + \overline{\phi}_T\overline{\phi}_T^\top$, and $\mathbf{b}_T = \overline{\Phi}_T\overline{\mathbf{y}}_T$ with $\mathbf{b}_T = \mathbf{b}_{T-1} + \overline{y}_T\overline{\phi}_T$, so that $\mathbf{w}_T = \mathbf{A}_{T-1}^{-1}\mathbf{b}_{T-1}$. According to the Woodbury formulation, $\mathbf{A}_T^{-1} = \mathbf{A}_{T-1}^{-1} - \frac{\mathbf{A}_{T-1}^{-1}\overline{\phi}_T\overline{\phi}_T^\top\mathbf{A}_{T-1}^{-1}}{1+\overline{\phi}_T^\top\mathbf{A}_{T-1}^{-1}\overline{\phi}_T}$, and with annotation $\tau_t = \overline{\phi}_t^\top\mathbf{A}_{t-1}^{-1}\overline{\phi}_t$, we have

$$\overline{\phi}_t^\top\mathbf{A}_t^{-1}\overline{\phi}_t = \overline{\phi}_t^\top\left(\mathbf{A}_{t-1}^{-1} - \frac{\mathbf{A}_{t-1}^{-1}\overline{\phi}_t\overline{\phi}_t^\top\mathbf{A}_{t-1}^{-1}}{1+\overline{\phi}_t^\top\mathbf{A}_{t-1}^{-1}\overline{\phi}_t}\right)\overline{\phi}_t = \frac{\tau_t}{1+\tau_t}.$$

Substituting $\mathbf{w}_T = \mathbf{A}_T^{-1}\mathbf{b}_T$ back to F_T, $F_T = -\mathbf{b}_T^\top\mathbf{A}_T^{-1}\mathbf{b}_T + \overline{\mathbf{y}}_T^\top\overline{\mathbf{y}}_T$. It derives that

$$\begin{aligned}
&F_{T-1} - F_T \\
=&\left(-\mathbf{b}_{T-1}^\top\mathbf{A}_{T-1}^{-1}\mathbf{b}_{T-1} + \overline{\mathbf{y}}_{T-1}^\top\overline{\mathbf{y}}_{T-1}\right) - \left(-\mathbf{b}_T^\top\mathbf{A}_T^{-1}\mathbf{b}_T + \overline{\mathbf{y}}_T^\top\overline{\mathbf{y}}_T\right) \\
=&(\mathbf{b}_{T-1} + \overline{y}_T\overline{\phi}_T)^\top\mathbf{A}_T^{-1}(\mathbf{b}_{T-1} + \overline{y}_T\overline{\phi}_T) - \mathbf{b}_{T-1}^\top\mathbf{A}_{T-1}^{-1}\mathbf{b}_{T-1} - \overline{y}_T^2 \\
=&\overline{y}_T^2\left(\overline{\phi}_T^\top\mathbf{A}_T^{-1}\overline{\phi}_T - 1\right) + 2\overline{y}_T\mathbf{b}_{T-1}^\top\mathbf{A}_T^{-1}\overline{\phi}_T + \mathbf{b}_{T-1}^\top\left(\mathbf{A}_T^{-1} - \mathbf{A}_{T-1}^{-1}\right)\mathbf{b}_{T-1} \\
=&\frac{-a_T y_T^2}{1+\tau_T} + \frac{2a_T y_T f_T}{1+\tau_T} + \frac{-a_T f_T^2}{1+\tau_T} \\
=&-a_T(y_T - f_T)^2(1 - \phi_t^\top\mathbf{A}_t^{-1}\phi_t)
\end{aligned}$$

where the fourth equation holds since $f_t = \mathbf{w}_t^\top\phi_t = \mathbf{b}_t^\top\mathbf{A}_t^{-1}\phi_t$, and $\overline{\phi}_t^\top\mathbf{A}_t^{-1}\overline{\phi}_t = \frac{\tau_t}{1+\tau_t}$, and the fifth equation holds since $\frac{1}{1+\tau_t} = 1 - \frac{\tau_t}{1+\tau_t} = 1 - \overline{\phi}_t^\top\mathbf{A}_t^{-1}\overline{\phi}_t$. Summing over t from 1 to T, it yields $\sum_{t=1}^T(F_t - F_{t-1}) = F_T = \min_{\mathbf{w}\in\mathcal{H}}\left(b\|\mathbf{w}\|^2 + \sum_{t=1}^T\alpha_t(y_t - \mathbf{w}^\top\phi_t)^2\right)$. For any $\mathbf{w} \in \mathcal{H}$, we have

$$\sum_{t=1}^T a_t(y_t - f_t)^2(1 - \overline{\phi}_t^\top\mathbf{A}_t^{-1}\overline{\phi}_t) \le b\|\mathbf{w}\|^2 + \sum_{t=1}^T a_t(y_t - \mathbf{w}^\top\phi_t)^2$$

Rearrange the terms, and bound $|\mathbf{w}_t^\top\phi_t| \le C$ and $a_t \le \alpha$ for $t \in [T]$,

$$\begin{aligned}
&\sum_{t=1}^T a_t(y_t - \mathbf{w}_t^\top\phi_t)^2 - a_t(y_t - \mathbf{w}^\top\phi_t)^2 \\
&\le b\|\mathbf{w}\|^2 + 2\alpha(1 + C^2)\sum_{t=1}^T\overline{\phi}_t^\top\mathbf{A}_t^{-1}\overline{\phi}_t
\end{aligned} \tag{3}$$

Since $\mathbf{A}_t = \mathbf{A}_{t-1} + \overline{\phi}_t\overline{\phi}_t^\top \Rightarrow \mathbf{A}_{t-1}\mathbf{A}_t^{-1} = \mathbf{A}_t\mathbf{A}_t^{-1} - \overline{\phi}_t\overline{\phi}_t^\top\mathbf{A}_t^{-1}$, according to Sylvester's determinant theorem,

$$\det\left(\mathbf{A}_t\mathbf{A}_t^{-1} - \overline{\phi}_t\overline{\phi}_t^\top\mathbf{A}_t^{-1}\right) = \det\left(I_D - \overline{\phi}_t\overline{\phi}_t^\top\mathbf{A}_t^{-1}\right) = 1 - \overline{\phi}_t^\top\mathbf{A}_t^{-1}\overline{\phi}_t,$$

while $\det\left(\mathbf{A}_{t-1}\mathbf{A}_t^{-1}\right) = \det(\mathbf{A}_{t-1})\det(\mathbf{A}_t)^{-1}$. With these equations, we obtain $\overline{\phi}_t^\top\mathbf{A}_t^{-1}\overline{\phi}_t = 1 - \frac{\det(\mathbf{A}_{t-1})}{\det(\mathbf{A}_t)} \le \log\left(\frac{\det(\mathbf{A}_t)}{\det(\mathbf{A}_{t-1})}\right)$, where the inequality holds since $1 - \frac{1}{x} \le \log x$ for $x > 0$. Summing over $t = 1, \cdots, T$,

$$\begin{aligned}
&\sum_{t=1}^T\overline{\phi}_t^\top\mathbf{A}_t^{-1}\overline{\phi}_t \le \sum_{t=1}^T\log\left(\frac{\det(\mathbf{A}_t)}{\det(\mathbf{A}_{t-1})}\right) = \log\left(\frac{\det(\mathbf{A}_T)}{\det(\mathbf{A}_0)}\right) \\
&= \log\left(\det\left(\frac{1}{b}\overline{\Phi}_T\overline{\Phi}_T^\top + \mathbf{I}_D\right)\right) = \log\det\left(\frac{1}{b}\overline{\Phi}_T^\top\overline{\Phi}_T + \mathbf{I}_T\right).
\end{aligned}$$

where the last equality holds according to Sylvester's determinant theorem, and we obtain $\sum_{t=1}^T\overline{\phi}_t^\top\mathbf{A}_t^{-1}\overline{\phi}_t \le \log\left(\det\left(\frac{1}{b}\overline{\mathbf{K}}_T + \mathbf{I}\right)\right)$. Substituting this bound back to Eq. (3), it concludes the proof. □

Proof of Theorem 1

THEOREM 1. *Given an arbitrary sequence $\{(\phi_t, y_t)\}_{t=1}^T$, the algorithm learns on only queried trails $\{Z_t\phi_t\}_{t=1}^T$ where $Z_t \sim \mathcal{B}(1, p_t)$ with $p_t = \frac{h_+}{h_+ + \max(0,\Theta_t)}$ if $\widehat{f}_t \ge 0$ and $p_t = \frac{h_-}{h_- + \max(0,\Theta_t)}$ if $\widehat{f}_t < 0$. Let $\ell_h(\cdot)$ be the hinge loss, set $h_+ = \sqrt{\frac{\alpha\log\det(\frac{1}{b}\overline{\mathbf{K}}_{T_p}+\mathbf{I})}{(b+T_p)C^2}}$ and $h_- = \sqrt{\frac{\alpha\log\det(\frac{1}{b}\overline{\mathbf{K}}_{T_n}+\mathbf{I})}{(b+T_n)C^2}}$. For any $\mathbf{w} \in \mathcal{H}$, the expected weighted mistake number is bounded by:*

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^T a_t M_t\right] &= \mathbb{E}\left[\sum_{t=1, y_t=+1}^T a_t M_t + \sum_{t=1, y_t=-1}^T a_t M_t\right] \\
&\le \sum_{t=1}^T a_t\ell_h(\mathbf{w}) + C\sqrt{\alpha(b+T)\log\det\left(\frac{1}{b}\overline{\mathbf{K}}_T + \mathbf{I}\right)}
\end{aligned}$$

PROOF. According to the Lemma 1, for any $\mathbf{w} \in \mathcal{H}$,

$$\sum_{t=1}^T\frac{a_t(y_t - f_t)^2}{1+\tau_t} \le \mathbf{w}^\top\left(b\mathbf{I} + \overline{\Phi}_T\overline{\Phi}_T^\top\right)\mathbf{w} + \sum_{t=1}^T a_t(y_t^2 - 2y_t\mathbf{w}^\top\phi_t)$$

Since $\mathbf{w}$ is a random variable, we use $h\mathbf{w}$ to replace $\mathbf{w}$, and add $\sum_{t=1}^{T} 2a_t h$ on both hands of the inequality,

$$\sum_{t=1}^{T} 2a_t \left(h + \frac{-y_t f_t - \tau_t/2}{1+\tau_t}\right)$$
$$\leq h^2 \mathbf{w}^\top \mathbf{A}_T \mathbf{w} + \sum_{t=1}^{T} 2ha_t(1 - y_t \mathbf{w}^\top \phi_t) - \sum_{t=1}^{T} \frac{a_t f_t^2}{1+\tau_t}$$

Since $1 - y_t \mathbf{w}^\top \phi_t \leq \max(0, 1 - y_t \mathbf{w}^\top \phi_t) = \ell_t(\mathbf{w})$, where ℓ_t is the hinge loss, we obtain

$$\sum_{t=1}^{T} a_t \left(h + \frac{-y_t f_t - \tau_t/2}{1+\tau_t}\right) \leq \frac{h^2}{2} \mathbf{w}^\top \mathbf{A}_T \mathbf{w} + h\sum_{t=1}^{T} a_t \ell_t(\mathbf{w}) - \sum_{t=1}^{T} \frac{a_t f_t^2}{2(1+\tau_t)}$$

Omitting the term $-\sum_{t=1}^{T} \frac{a_t f_t^2}{2(1+\tau_t)}$ that does not affect inequality, and considering the inequality with the scenarios $M_t Z_t$,

$$\sum_{t=1}^{T} M_t Z_t a_t \left(h + \frac{-y_t f_t - \tau_t/2}{1+\tau_t}\right) \leq \frac{h^2}{2} \mathbf{w}^\top \mathbf{A}_T \mathbf{w} + h\sum_{t=1}^{T} a_t \ell_t(\mathbf{w}),$$

One can easily prove that this inequality still holds for $M_t Z_t = 0$. If an error occurs at trial t, i.e., $M_t Z_t = 1$, we have $-y_t f_t = |f_t|$. Then the left hand of the inequality becomes

$$\sum_{t=1}^{T} M_t Z_t a_t \left(\frac{|f_t| - \tau_t/2}{1+\tau_t}\right) = \sum_{t=1}^{T} M_t Z_t a_t \Theta_t$$

Considering that the algorithm queries an input and suffers a mistake at round t, i.e., $Z_t = 1$ and $M_t = 1$. There are two mistake cases: (1) False negative: true label $y_t = +1$ while the prediction $f_t < 0$; (2) False positive: true label $y_t = -1$ while the prediction $f_t \geq 0$. For the first case, we have:

$$\sum_{t, y_t=+1} M_t Z_t a_t (h_+ + \Theta_t)$$
$$\leq \sum_{t, y_t=+1} h_+ a_t \ell_h(\mathbf{w}) + \frac{h_+^2}{2} \mathbf{w}^\top \left(b\mathbf{I} + \sum_{t, y_t=+1} \overline{\phi}_t \overline{\phi}_t^\top\right) \mathbf{w}$$

Now we would like to remove the random variable Z_t. First, when the confidence score $\Theta_t > 0$, taking the expectation over variables $\mathbb{E}(Z_t) = \frac{h_+}{h_+ + \Theta_t}$, we have:

$$\mathbb{E}\left[\sum_{t, y_t=+1} h_+ M_t a_t\right] \leq \frac{h_+^2}{2} \mathbf{w}^\top \left(b\mathbf{I} + \sum_{t, y_t=+1} \overline{\phi}_t \overline{\phi}_t^\top\right) \mathbf{w} + \sum_{t=1, y_t=+1}^{T} h_+ a_t \ell_h(\mathbf{w})$$

In addition, one can easily prove this inequality holds for $M_t = 0$. Second, when $\Theta_t \leq 0$, the random variable is assigned to $\mathbb{E}(Z_t) = 1$ and $M_t \in \{0, 1\}$,

$$\mathbb{E}[M_t Z_t a_t (h_+ + \Theta_t)] = \mathbb{E}[Z_t] \mathbb{E}\left[M_t a_t \left(h_+ + \frac{|f_t| - \tau_t/2}{1+\tau_t}\right)\right]$$
$$\geq \mathbb{E}\left[M_t a_t (h_+ - \frac{\tau_t/2}{1+\tau_t})\right] \geq \mathbb{E}[h_+ M_t a_t] - \mathbb{E}[\frac{a_t \tau_t}{2(1+\tau_t)}]$$
$$= \mathbb{E}[h_+ M_t a_t] - \frac{a_t}{2} \overline{\phi}_t^\top \mathbf{A}_t^{-1} \overline{\phi}_t.$$

Combining above two scenarios for Θ_t, we obtain

$$\mathbb{E}\left[\sum_{t, y_t=+1} a_t M_t\right] \leq \frac{h_+}{2} \mathbf{w}^\top \left(b\mathbf{I} + \sum_{t, y_t=+1} \overline{\phi}_t \overline{\phi}_t^\top\right) \mathbf{w} + \sum_{t, y_t=+1} a_t \ell_h(\mathbf{w})$$
$$+ \sum_{t, y_t=+1, \Theta_t \leq 0} \frac{a_t}{2h_+} \overline{\phi}_t^\top \mathbf{A}_t^{-1} \overline{\phi}_t$$
$$\leq \frac{h_+}{2} (b + T_p) C^2 + \sum_{t, y_t=+1} a_t \ell_h(\mathbf{w}) + \frac{\alpha}{2h_+} \log\left(\det\left(\frac{1}{b}\overline{\mathbf{K}}_{T_p} + \mathbf{I}\right)\right)$$

where $\|\mathbf{w}\|_2 \leq C$ and T_p is number of positive labels in the last inequality. If we set $h_+ = \sqrt{\frac{\alpha \log\left(\det(\frac{1}{b}\overline{\mathbf{K}}_{T_p} + \mathbf{I})\right)}{(b+T_p)C^2}}$, then

$$\mathbb{E}\left[\sum_{t, y_t=+1} a_t M_t\right] \leq \sum_{t, y_t=+1} a_t \ell_h(\mathbf{w}) + C\sqrt{\alpha(b + T_p) \log\det\left(\frac{1}{b}\overline{\mathbf{K}}_{T_p} + \mathbf{I}\right)} \tag{4}$$

When $y_t = -1$, setting $h_- = \sqrt{\frac{\alpha \log\left(\det(\frac{1}{b}\overline{\mathbf{K}}_{T_n} + \mathbf{I})\right)}{(b+T_n)C^2}}$, we have

$$\mathbb{E}\left[\sum_{t, y_t=-1} a_t M_t\right] \leq \sum_{t, y_t=-1} a_t \ell_h(\mathbf{w}) + C\sqrt{\alpha(b + T_n) \log\left(\det\left(\frac{1}{b}\overline{\mathbf{K}}_{T_n} + \mathbf{I}\right)\right)} \tag{5}$$

Summing Eq. (4) and (5) will give:

$$\mathbb{E}\left[\sum_{t=1}^{T} a_t M_t\right] = \mathbb{E}\left[\sum_{t=1, y_t=+1}^{T} a_t M_t + \sum_{t=1, y_t=-1}^{T} a_t M_t\right]$$
$$\leq \sum_{t=1}^{T} a_t \ell_h(\mathbf{w}) + 2C\sqrt{\alpha(b + T) \log\left(\det\left(\frac{1}{b}\overline{\mathbf{K}}_T + \mathbf{I}\right)\right)}$$

Then, we conclude the proofs of Theorem.

□

Proof of Theorem 2

THEOREM 2. *Under the same condition in Theorem 1, by setting* $\alpha = \frac{\mu_p T_n}{\mu_n T_p}$, *the proposed algorithm satisfies for any* $\mathbf{w} \in \mathcal{H}$*:*

$$\mathbb{E}[sum] \geq 1 - \frac{\mu_n}{T_n}\left[\sum_{t=1}^{T} a_t \ell_h(\mathbf{w}) + 2C\sqrt{\frac{\mu_p T_n}{\mu_n T_p}(b + T) \log\det\left(\frac{1}{b}\overline{\mathbf{K}}_T + \mathbf{I}\right)}\right]$$

By setting $\mu_p = \mu_n = 0.5$, *we can easily obtain the bound of the balanced accuracy.*

PROOF. Associating the cost-aware sum with the cost-aware loss function, we have

$$sum = \mu_p \frac{T_p - M_p}{T_p} + \mu_n \frac{T_n - M_n}{T_n}$$
$$= 1 - \frac{\mu_n}{T_n}\left[\sum_{y_i=+1} \frac{\mu_p T_n}{\mu_n T_p} \mathbb{I}_{(\widehat{y}_i < 0)} + \sum_{y_i=-1} \mathbb{I}_{(\widehat{y}_i \geq 0)}\right].$$

By setting $a_t = \frac{\mu_p T_n}{\mu_n T_p}\mathbb{I}_{(y_t=+1)} + \mathbb{I}_{(y_t=-1)}$, we have

$$\begin{aligned}\mathbb{E}[sum] &= 1 - \frac{\mu_n}{T_n}\mathbb{E}\left[\sum_{t=1}^{T} a_t M_t\right] \\ &\geq 1 - \frac{\mu_n}{T_n}\left[\sum_{t=1}^{T} a_t \ell_h(\mathbf{w}) + 2C\sqrt{\frac{\mu_p T_n}{\mu_n T_p}(b+T)\log\left(\det\left(\frac{1}{b}\overline{\mathbf{K}}_T + \mathbf{I}\right)\right)}\right]\end{aligned}$$

□

Proof of Theorem 3

THEOREM 3. *Under the same condition in Theorem 1, by setting $\alpha = \frac{c_p}{c_n}$, the proposed algorithm satisfies for any $\mathbf{w} \in \mathcal{H}$:*

$$\mathbb{E}[cost] \leq c_n\left[\sum_{t=1}^{T} a_t \ell_h(\mathbf{w}) + 2C\sqrt{\frac{c_p}{c_n}(b+T)\log\left(\det\left(\frac{1}{b}\overline{\mathbf{K}}_T + \mathbf{I}\right)\right)}\right]$$

By setting $c_p = c_n$, we can easily obtain the bound of the balanced penalty.

PROOF. Associating the cost-aware cost with the cost-aware loss function, we have

$$\cdot cost = c_p \times M_p + c_n \times M_n = c_n\left[\sum_{y_i=+1}\frac{c_p}{c_n}\mathbb{I}_{(\widehat{y}_i<0)} + \sum_{y_i=-1}\mathbb{I}_{(\widehat{y}_i\geq 0)}\right]$$

By setting $a_t = \frac{c_p}{c_n}\mathbb{I}_{(y_t=+1)} + \mathbb{I}_{(y_t=-1)}$, we have

$$\begin{aligned}\mathbb{E}[cost] &= c_n\mathbb{E}\left[\sum_{t=1}^{T} a_t M_t\right] \\ &\leq c_n\left[\sum_{t=1}^{T} a_t \ell_h(\mathbf{w}) + 2C\sqrt{\frac{c_p}{c_n}(b+T)\log\left(\det\left(\frac{1}{b}\overline{\mathbf{K}}_T + \mathbf{I}\right)\right)}\right]\end{aligned}$$

□