# Appendix: Kernelized Online Imbalanced Learning with Fixed Budgets

**Junjie Hu**[1,2], **Haiqin Yang**[1,2], **Irwin King**[1,2], **Michael R. Lyu**[1,2], **and Anthony Man-Cho So**[3]
[1]Shenzhen Key Laboratory of Rich Media Big Data Analytics and Applications
Shenzhen Research Institute, The Chinese University of Hong Kong
[2]Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong
[3]Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Hong Kong
{jjhu, hqyang, king, lyu}@cse.cuhk.edu.hk, manchoso@se.cuhk.edu.hk

In the following, we provide self-contained proofs and more experimental descriptions for our proposed Kernelized Online Imbalanced Learning (KOIL) with fixed budgets. The Appendix is organized as follows: In the section **Main Theoretical Results**, we present the main theoretical results of our KOIL by exploiting the pair-wise hinge loss function as a surrogate of AUC metric. In the section **More Experiments**, we present more detailed experimental descriptions and results. Finally, in the section **Pseudocodes**, we show all pesudocodes of KOIL.

# Main Theoretical Results

In the following, we first present the assumption and define the notation and the objective of KOIL. Next, we prove the bound of the norm of the decision function, the bound of the pair-wise hinge loss function. We then present the sophisticated compensation updating policy and the bound of the norm of the compensated updating decision function, $f_t^{++}$. Finally, we prove the regret bounds of our KOIL with pair-wise hinge loss function and infinite budgets and our proposed KOIL with corresponding compensation updating policy.

## Assumptions and Notation

In this work, we focus on learning a nonlinear decision function $f : \mathbb{R}^d \to \mathbb{R}$ from a sequence of imbalanced feature-labeled pair instances in binary classification, $\{\mathbf{z}_t = (\mathbf{x}_t, y_t) \in \mathcal{Z}, t \in [T]\}$, where $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, $\mathbf{x}_t \in \mathcal{X} \subseteq \mathbb{R}^d$, $y_t \in \mathcal{Y} = \{-1, +1\}$ and $[T] = \{1, \ldots, T\}$. Without loss of generality, we assume the positive class is the minority class while the negative class is the majority class. We denote $N_{t,k}^{\tilde{y}}(\mathbf{z})$ as the set of feature-labeled pair instances which are the $k$-nearest neighbors of $\mathbf{z}$ and have the label of $\tilde{y}$. Here, the neighborhood is defined by the distance or the similarity between two instances, i.e., the smaller the distance (or the more similarity) of instances, the close the neighbors. Besides, we define the index sets, $I_t^+$ and $I_t^-$, to record the indexes of positive support vectors and negative support vectors at the $t$-th trial. Moreover, for simplicity, we define two buffers, $\mathcal{K}_t^+$ and $\mathcal{K}_t^-$, to store the learned information for two classes at the $t$-th trial, respectively:

$$\mathcal{K}_t^+.\mathcal{A} = \{\alpha_{i,t}^+ \,|\, \alpha_{i,t}^+ \neq 0, i \in I_t^+\}, \;\; \mathcal{K}_t^+.\mathcal{B} = \{\mathbf{z}_i \,|\, y_i = +1, i \in I_t^+\}, \tag{1}$$

$$\mathcal{K}_t^-.\mathcal{A} = \{\alpha_{i,t}^- \,|\, \alpha_{i,t}^- \neq 0, i \in I_t^-\}, \;\; \mathcal{K}_t^-.\mathcal{B} = \{\mathbf{z}_i \,|\, y_i = -1, i \in I_t^-\}, \tag{2}$$

where $\alpha_{i,t}$ denotes the weight of the support vector firstly occurred at the $i$-th trial and updated at the $t$-th trial. Here, we fix the budgets, i.e., the buffer sizes, to $N$. That is, $|I^+| = |I^-| = N$.

The objective of KOIL is to seek a decision function at the $t$-th trial expressed as follows:

$$f_t(\mathbf{x}) = \sum\nolimits_{i \in I_t^+} \alpha_{i,t}^+ k(\mathbf{x}_i, \mathbf{x}) + \sum\nolimits_{j \in I_t^-} \alpha_{j,t}^- k(\mathbf{x}_j, \mathbf{x}), \tag{3}$$

where the information of support vectors is stored at $\mathcal{K}_t^+$ and $\mathcal{K}_t^-$, respectively. More generally, $f_t(\mathbf{x})$ is an element of a Reproducing Kernel Hilbert Space (RKHS) (Schölkopf and Smola 2002) and can be expressed as $f_t(\mathbf{x}) = \langle f_t(\cdot), k(\mathbf{x}, \cdot)\rangle_{\mathcal{H}}$. The prediction of a new sample $\mathbf{x}$ is made by $\text{sgn}(f_t(\mathbf{x}))$.

Here, we adopt the pair-wise hinge loss function as the loss function for a convex surrogate of the AUC maximization (Gao et al. 2013; Zhao et al. 2011), which is defined by,

$$\ell_h(f, \mathbf{z}_t, \mathbf{z}_i) = \left[1 - \frac{1}{2}(y_t - y_i)(f(\mathbf{x}_t) - f(\mathbf{x}_i))\right]_+, \quad \text{where } [v]_+ = \max\{0, v\}. \tag{4}$$

Our proposed Kernelized Online Imbalanced Learning (KOIL) aims at minimizing the *localized instantaneous regularized risk of AUC* on a single instance $\mathbf{z}_t$, which is defined as:

$$\hat{\mathcal{L}}_t(f) := \hat{\mathcal{L}}_t(f) = \frac{1}{2}\|f\|_{\mathcal{H}}^2 + C \sum\nolimits_{\mathbf{z}_i \in N_{t,k}^{-y_t}(\mathbf{z}_t)} \ell_h(f, \mathbf{z}_t, \mathbf{z}_i), \tag{5}$$

where $\|f\|_{\mathcal{H}}^2$ is the same as $\langle f, f\rangle_{\mathcal{H}}$.

Using the standard stochastic gradient descent method to update the decision function, we have $f_1 = 0$ and the updating rule as follows:

$$f_{t+1} := f_t - \eta \partial_f \hat{\mathcal{L}}_t(f)|_{f=f_t}, \tag{6}$$

where $\eta > 0$ is the learning rate, and $\partial_f$ is shorthand for $\partial/\partial f$ (the gradient with respect to $f$). In the following, without further explanation, $\partial$ implies $\partial_f$.

The partial derivative in Eq. (6) is calculated by

$$\partial \hat{\mathcal{L}}_t(f_t) = f_t - C \sum_{\mathbf{z}_i \in N_{t,k}^{-y_t}(\mathbf{z}_t)} \mathbb{I}[\ell_h(f, \mathbf{z}_t, \mathbf{z}_i) > 0] \cdot \varphi(\mathbf{z}_t, \mathbf{z}_i), \tag{7}$$

where $\varphi(\mathbf{z}_t, \mathbf{z}_i) = y_t(k(\mathbf{x}_t, \cdot) - k(\mathbf{x}_i, \cdot))$.

For simplicity, we define the valid set $V_t$ and its complementary set $\overline{V_t}$ at the $t$-th trial as follows:

$$V_t := \{i \in I_t^{-y_t} | \mathbf{z}_i \in N_{t,k}^{-y_t}(\mathbf{z}_t) \wedge \ell_h(f, \mathbf{z}_t, \mathbf{z}_i) > 0\}, \overline{V_t} := I_t^{-y_t} \setminus V_t. \tag{8}$$

Hence, the corresponding updating rule for the kernel weights at the $t$-th trial is derived as follows:

$$\alpha_{i,t} = \begin{cases} \eta C y_t |V_t|, & i = t \\ (1-\eta)\alpha_{i,t-1} - \eta C y_t, & \forall i \in V_t \\ (1-\eta)\alpha_{i,t-1}, & \forall i \in I_t^{y_t} \cup \overline{V_t} \end{cases} \tag{9}$$

Details about the algorithms are in Algorithm 1 and **UpdateKernel** of Algorithm 2 in Section .

## Proof of Lemma 1

We first present the bound of the norm of the decision function.

**Lemma 1.** *Suppose for all $\mathbf{x} \in \mathbb{R}^d$, $k(\mathbf{x}, \mathbf{x}) \leq X^2$, where $X > 0$. Let $\xi_1$ be in $[0, X]$, such that $k(\mathbf{x}_t, \mathbf{x}_i) \geq \xi_1^2$, $\forall\, \mathbf{z}_i = (\mathbf{x}_i, y_i) \in N_t^{-y_t}(\mathbf{z}_t)$. With $f_1 = 0$, we have*

$$\|f_{t+1}\|_{\mathcal{H}} \leq Ck\sqrt{2X^2 - 2\xi_1^2}. \tag{10}$$

*Proof.* First, since $\mathbf{z}_i$ is one of the $k$-nearest opposite support vectors to $\mathbf{z}_t$, i.e., $\mathbf{z}_i \in N_{t,k}^{-y_t}(\mathbf{z}_t)$, the assumption $k(\mathbf{z}_t, \mathbf{z}_i) \geq \xi_1^2$ with $\xi_1^2 > 0$ makes sense. We then have

$$\|\varphi(\mathbf{z}_t, \mathbf{z}_i)\|_{\mathcal{H}} = \sqrt{k(\mathbf{x}_t, \mathbf{x}_t) - 2k(\mathbf{x}_t, \mathbf{x}_i) + k(\mathbf{x}_i, \mathbf{x}_i)} \leq \sqrt{2X^2 - 2\xi_1^2}. \tag{11}$$

Now we bound the norm of the decision function $f_{t+1}$

$$\begin{aligned} \|f_{t+1}\|_{\mathcal{H}} &= \left\| (1-\eta) f_t + \eta C \sum_{\mathbf{z}_i \in N_{t,k}^{-y_t}(\mathbf{z}_t)} \mathbb{I}[\ell_h(f, \mathbf{z}_t, \mathbf{z}_i) > 0] \cdot \varphi(\mathbf{z}_t, \mathbf{z}_i) \right\|_{\mathcal{H}} \\ &\leq (1-\eta)\|f_t\|_{\mathcal{H}} + \eta C \sum_{\mathbf{z}_i \in N_{t,k}^{-y_t}(\mathbf{z}_t)} \mathbb{I}[\ell_h(f, \mathbf{z}_t, \mathbf{z}_i) > 0] \cdot \|\varphi(\mathbf{z}_t, \mathbf{z}_i)\|_{\mathcal{H}} \\ &\leq (1-\eta)\|f_t\|_{\mathcal{H}} + \eta C \sum_{\mathbf{z}_i \in N_{t,k}^{-y_t}(\mathbf{z}_t)} \mathbb{I}[\ell_h(f, \mathbf{z}_t, \mathbf{z}_i) > 0] \cdot \sqrt{2X^2 - 2\xi_1^2} \\ &\leq (1-\eta)\|f_t\|_{\mathcal{H}} + \eta C k \sqrt{2X^2 - 2\xi_1^2}. \end{aligned} \tag{12}$$

In the above, the first equality is by substituting Eq. (7) into Eq. (6) to calculate $f_{t+1}$. The first inequality is attained by the triangle inequality. The second inequality is attained by Eq. (11). The third inequality holds since the number of elements in $N_{t,k}^{-y_t}(\mathbf{z}_t)$ is at most $k$.

By expanding $\|f_t\|_{\mathcal{H}}$ iteratively, we have

$$\|f_{t+1}\|_{\mathcal{H}} \leq (1-\eta)^t \|f_1\|_{\mathcal{H}} + \left( \frac{1 - (1-\eta)^t}{\eta} \right) \eta C k \sqrt{2X^2 - 2\xi_1^2} \leq Ck\sqrt{2X^2 - 2\xi_1^2}.$$

The second inequality holds when $\eta < 1$, $1 - (1-\eta)^t \leq 1$ for $t \in [T]$ and $f_1 = 0$. □

## Proof of Lemma 2

In the following, we prove that the pair-wise hinge loss function is bounded.

**Lemma 2.** *With the same assumption in Lemma 1 and the pair-wise hinge loss function $\ell : \mathcal{H} \times \mathcal{Z} \times \mathcal{Z} \to [0, U]$ defined by Eq. (4), we can determine the bound by*

$$U = 1 + 2Ck(X^2 - \xi_1^2). \tag{13}$$

*Proof.*

$$\begin{aligned}
\ell_h(f_t, \mathbf{z}_t, \mathbf{z}_i) &= \left[1 - \frac{1}{2}(y_t - y_i)(f_t(\mathbf{x}_t) - f_t(\mathbf{x}_i))\right]_+ \\
&\leq 1 + |f_t(\mathbf{x}_t) - f_t(\mathbf{x}_i)| \\
&= 1 + |\langle f_t, k(\mathbf{x}_t, \cdot) - k(\mathbf{x}_i, \cdot)\rangle_{\mathcal{H}}| \\
&\leq 1 + \|f_t\|_{\mathcal{H}} \cdot \|k(\mathbf{x}_t, \cdot) - k(\mathbf{x}_i, \cdot)\|_{\mathcal{H}} \\
&\leq 1 + Ck\sqrt{2X^2 - 2\xi_1^2} \cdot \|k(\mathbf{x}_t, \cdot) - k(\mathbf{x}_i, \cdot)\|_{\mathcal{H}} \\
&\leq 1 + Ck\sqrt{2X^2 - 2\xi_1^2} \cdot \sqrt{2X^2 - 2\xi_1^2} \\
&= 1 + 2Ck(X^2 - \xi_1^2) \quad (:= U).
\end{aligned}$$

In the above, the first inequality is valid due to the triangle inequality and $y_t - y_i = -2, 0,$ or $2$. The second inequality is due to the Cauchy-Schwarz inequality. The third inequality is due to the bound of the decision function in Lemma 1. The forth inequality is due to the assumption that $k(\mathbf{x}_t, \mathbf{x}_i) \geq \xi_1^2, \forall\, \mathbf{z}_i = (\mathbf{x}_i, y_i) \in N_t^{-y_t}(\mathbf{z}_t)$ and Eq. (11). □

**Note:** In Lemma 2, we only prove the bound for pair-wise hinge loss function. For other convex loss functions, the proof is similar.

## Proof of Lemma 3

In the following, we first present the sophisticated compensation updating policy. We then prove the bound of the norm of the compensated decision function.

**Update Buffers** To avoid information loss, we need to design a compensation scheme. Let the removed support vector be $\mathbf{z}_r = (\mathbf{x}_r, y_r)$, we find the most similar support vector $\mathbf{z}_c = (\mathbf{x}_c, y_c)$ with $y_c = y_r$ in $\mathcal{K}_t^{y_r}$ and update its corresponding weight. Note that at the $t$-th trial, the updating rule is Eq. (6). When the support vector $\mathbf{z}_r$ is removed, the decision function becomes

$$\hat{f}_{t+1}(\mathbf{x}) = f_{t+1}(\mathbf{x}) - \alpha_{r,t} k(\mathbf{x}_r, \mathbf{x}).$$

Now we find the compensated support vector $\mathbf{z}_c$ and determine its updated weight $\Delta\alpha_{c,t}$. We want to keep track of all information and do not change the decision function after the compensation:

$$f_{t+1}(\mathbf{x}) = \hat{f}_{t+1}(\mathbf{x}) + \Delta\alpha_{c,t} \cdot k(\mathbf{x}_c, \mathbf{x}) = f_{t+1}(\mathbf{x}) - \alpha_{r,t} k(\mathbf{x}_r, \mathbf{x}) + \Delta\alpha_{c,t} \cdot k(\mathbf{x}_c, \mathbf{x}). \tag{14}$$

By Eq. (14), we set

$$\Delta\alpha_{c,t} = \alpha_{r,t} \frac{k(\mathbf{x}_r, \mathbf{x})}{k(\mathbf{x}_c, \mathbf{x})} \approx \alpha_{r,t}. \tag{15}$$

The above approximation is due to the similarity of the removed support vector, $\mathbf{x}_r$ and the compensated support vector, $\mathbf{x}_c$.

We then express the updating rule of $f_{t+1}$ with compensation by $f_{t+1}^{++}$ as:

$$f_{t+1}^{++} = (1 - \eta) f_t^{++} + \eta \partial_f \hat{\mathcal{L}}_t(f)|_{f=f_t^{++}} + \alpha_{r,t} \left(k(\mathbf{x}_c, \cdot) - k(\mathbf{x}_r, \cdot)\right), \tag{16}$$

where $f_t^{++}$ is the previous decision function. When either buffer is not full, $f_t^{++}$ corresponds to the original decision function without compensation, i.e., $f_t$ updated by Eq. (6). Ideally, if $k(\mathbf{x}_c, \mathbf{x})$ equals $k(\mathbf{x}_r, \mathbf{x})$, $f_t^{++}$ is equivalent to the one learned with infinite budgets, which reserves all the revealed instances as support vectors. Hence, we call the replacement with the compensation scheme as the extended updating policy. For the Reservoir Sampling policy, it is named **RS++**, while for the FIFO policy, it is named **FIFO++**. Details about the compensation is shown in Algorithm 3.

**Proof** Now we show the bound of the norm of the decision function using the compensation technique. We assume that $\forall i \in I_t^+ \cup I_t^-, |\alpha_{i,t}| \in [0, \gamma\eta]$. Otherwise, $f_t^{++} = 0$ is a better solution as it may approach to infinity by the objective function in Eq. (5). Similarly, we can use the projection operation in (Hoi et al. 2012) to project $|\alpha_{i,t}|$ to $[0, \gamma\eta]$.

**Lemma 3.** *Suppose for all $\mathbf{x} \in \mathbb{R}^d$, $k(\mathbf{x}, \mathbf{x}) \leq X^2$, where $X > 0$. $\forall i \in I_t^+ \cup I_t^-$, $\alpha_{i,t} \in [0, \gamma\eta]$. Let $\xi_1$ be in $[0, X]$, such that $k(\mathbf{x}_t, \mathbf{x}_i) \geq \xi_1^2, \forall\, \mathbf{z}_i = (\mathbf{x}_i, y_i) \in N_t^{-y_t}(\mathbf{z}_t)$. Let $\mathbf{x}_r$ and $\mathbf{x}_c$ be the removed and compensated support vectors at the $t$-th trial respectively such that $k(\mathbf{x}_r, \mathbf{x}_c) \geq \xi_2^2$, where $0 < \xi_2 \leq X$. With $f_1^{++} = 0$ and updating rule in Eq. (16), we have*

$$\|f_t^{++}\|_{\mathcal{H}} \leq Ck\sqrt{2X^2 - 2\xi_1^2} + \gamma\sqrt{2X^2 - 2\xi_2^2}. \tag{17}$$

*Proof.* For every pair of removed and compensated support vectors at each trial, we have,

$$\|k(\mathbf{x}_c,\cdot)-k(\mathbf{x}_r,\cdot)\|_{\mathcal{H}}=\sqrt{k(\mathbf{x}_r,\mathbf{x}_r)-2k(\mathbf{x}_c,\mathbf{x}_c)+k(\mathbf{x}_r,\mathbf{x}_c)}\leq\sqrt{2X^2-2\xi_2^2}. \tag{18}$$

Now we bound the norm of the decision function $f_{t+1}^{++}$. We consider the following two cases:
Case I: When the buffer is full at the $t$-th trial, we have

$$\begin{aligned}
&\|f_{t+1}^{++}\|_{\mathcal{H}}\\
&=\left\|(1-\eta)f_t^{++}+\eta C\left[\sum_{\mathbf{z}_i\in N_{t,k}^{-y_t}(\mathbf{z}_t)}\mathbb{I}[\ell_h(f,\mathbf{z}_t,\mathbf{z}_i)>0]\cdot\varphi(\mathbf{z}_t,\mathbf{z}_i)\right]+\alpha_{r,t}(k(\mathbf{x}_c,\cdot)-k(\mathbf{x}_r,\cdot))\right\|_{\mathcal{H}}\\
&\leq(1-\eta)\|f_t^{++}\|_{\mathcal{H}}+\eta C\left[\sum_{\mathbf{z}_i\in N_{t,k}^{-y_t}(\mathbf{z}_t)}\mathbb{I}[\ell_h(f,\mathbf{z}_t,\mathbf{z}_i)>0]\cdot\|\varphi(\mathbf{z}_t,\mathbf{z}_i)\|_{\mathcal{H}}\right]+|\alpha_{r,t}|\|k(\mathbf{x}_c,\cdot)-k(\mathbf{x}_r,\cdot)\|_{\mathcal{H}}\\
&\leq(1-\eta)\|f_t^{++}\|_{\mathcal{H}}+\eta C\left[\sum_{\mathbf{z}_i\in N_{t,k}^{-y_t}(\mathbf{z}_t)}\mathbb{I}[\ell_h(f,\mathbf{z}_t,\mathbf{z}_i)>0]\cdot\sqrt{2X^2-2\xi_1^2}\right]+|\alpha_{r,t}|\sqrt{2X^2-2\xi_2^2}\\
&\leq(1-\eta)\|f_t^{++}\|_{\mathcal{H}}+\eta Ck\sqrt{2X^2-2\xi_1^2}+|\alpha_{r,t}|\sqrt{2X^2-2\xi_2^2}\\
&\leq(1-\eta)\|f_t^{++}\|_{\mathcal{H}}+\eta Ck\sqrt{2X^2-2\xi_1^2}+\gamma\eta\sqrt{2X^2-2\xi_2^2}.
\end{aligned}\tag{19}$$

In the above, the first equality is by substituting Eq. (16) for $f_{t+1}$. The first inequality is attained by the triangle inequality. The second inequality is attained by Eq. (11) and Eq. (18). The third inequality holds since the number of elements in $N_{t,k}^{-y_t}(\mathbf{z}_t)$ is at most $k$. The forth inequality holds since $|\alpha_{r,t}|$ is bounded by $\gamma\eta$.

Case II: When the buffer is not full at the $t$-th trail, $f_{t+1}^{++}=f_{t+1}$. Similarly in Eq. (12), we have

$$\begin{aligned}
\|f_{t+1}^{++}\|_{\mathcal{H}}&\leq(1-\eta)\|f_t^{++}\|_{\mathcal{H}}+\eta Ck\sqrt{2X^2-2\xi_1^2}\\
&\leq(1-\eta)\|f_t^{++}\|_{\mathcal{H}}+\eta Ck\sqrt{2X^2-2\xi_1^2}+\gamma\eta\sqrt{2X^2-2\xi_2^2}.
\end{aligned}\tag{20}$$

The above second inequality is due to $\gamma\eta\sqrt{2X^2-2\xi_2^2}\geq 0$.

In sum, by Eq. (19) and Eq. (20), we have the following inequality:

$$\|f_{t+1}^{++}\|_{\mathcal{H}}\leq(1-\eta)\|f_t^{++}\|_{\mathcal{H}}+\eta Ck\sqrt{2X^2-2\xi_1^2}+\gamma\eta\sqrt{2X^2-2\xi_2^2}. \tag{21}$$

By expanding $\|f_t^{++}\|_{\mathcal{H}}$ iteratively, we have

$$\begin{aligned}
\|f_{t+1}^{++}\|_{\mathcal{H}}&\leq(1-\eta)^t\|f_1^{++}\|_{\mathcal{H}}+\sum_{i=0}^{t-1}(1-\eta)^i\left(\eta Ck\sqrt{2X^2-2\xi_1^2}+\gamma\eta\sqrt{2X^2-2\xi_2^2}\right)\\
&\leq(1-\eta)^t\|f_1^{++}\|_{\mathcal{H}}+\left(\frac{1-(1-\eta)^t}{\eta}\right)\left(\eta Ck\sqrt{2X^2-2\xi_1^2}+\gamma\eta\sqrt{2X^2-2\xi_2^2}\right)\\
&\leq Ck\sqrt{2X^2-2\xi_1^2}+\gamma\sqrt{2X^2-2\xi_2^2}.
\end{aligned}$$

The third inequality holds when $\eta<1$, $1-(1-\eta)^t\leq 1$ for $t\in[1,T]$ and $f_1^{++}=0$. □

## Proof of Theorem 1

We define the regret bound as the difference between the objective value up to the $T$-th step and the smallest objective value from hindsight.

$$R_T=\sum\nolimits_{t=1}^{T}\hat{\mathcal{L}}_t(f_t)-\hat{\mathcal{L}}(f^*,\mathbf{z}_t),\quad R_T^{++}=\sum\nolimits_{t=1}^{T}\hat{\mathcal{L}}(f_t^{++},\mathbf{z}_t)-\hat{\mathcal{L}}(f^*,\mathbf{z}_t), \tag{22}$$

where $f^*$ is the optimal decision function obtained from hindsight, and $f_t$ and $f_t^{++}$ corresponds to the updating in Eq. (6) and Eq. (16), respectively.

**Theorem 1.** *Suppose for all $\mathbf{x} \in \mathbb{R}^d$, $k(\mathbf{x},\mathbf{x}) \leq X^2$, where $X > 0$. Let $\xi_1$ be in $[0, X]$, such that $k(\mathbf{x}_t, \mathbf{x}_i) \geq \xi_1^2$, $\forall\, \mathbf{z}_i = (\mathbf{x}_i, y_i) \in N_t^{-y_t}(\mathbf{z}_t)$. Given $k > 0, C > 0, \eta > 0$ and a bounded convex loss function $\ell : \mathcal{H} \times \mathcal{Z} \times \mathcal{Z} \to [0, U]$ for $f_t$ updated by Eq. (6), with $f_1 = 0$, we have*

$$R_T \leq \frac{\|f^*\|_{\mathcal{H}}^2}{2\eta} + \eta C k \sum_{t=1}^{T} \left((U-1) + (k+1)C(X^2 - \xi_1^2)\right). \tag{23}$$

*Moreover, assume that $\forall i \in I_t^+ \cup I_t^-$, $\alpha_{i,t} \in [0, \gamma\eta]$ and $k(\mathbf{x}_r, \mathbf{x}_c) \geq \xi_2^2$ with $0 < \xi_2 \leq 0$ for any removed support vector $\mathbf{x}_r$ and compensated support vector $\mathbf{x}_c$ at any trial. With $f_1 = 0$ and $f_t^{++}$ updated by Eq. (16), we have*

$$R_T^{++} \leq R_T + T\left(4\gamma C k \sqrt{(X^2 - \xi_2^2)(X^2 - \xi_1^2)} + 2\gamma^2(X^2 - \xi_2^2)\right). \tag{24}$$

*Proof.* Let $f^*$ be the optimal solution from hindsight. We define the distance between $f_t$ and $f^*$ at the $t$-th trial as $\|f_t - f^*\|_{\mathcal{H}}$. Then we have

$$\begin{aligned}
&\|f_{t+1} - f^*\|_{\mathcal{H}}^2 - \|f_t - f^*\|_{\mathcal{H}}^2 \\
=&\|f_t - \eta \partial \hat{\mathcal{L}}_t(f_t) - f^*\|_{\mathcal{H}}^2 - \|f_t - f^*\|_{\mathcal{H}}^2 \\
=&\eta^2 \|\partial \hat{\mathcal{L}}_t(f_t)\|_{\mathcal{H}}^2 - 2\eta \langle \partial \hat{\mathcal{L}}_t(f_t), f_t - f^* \rangle_{\mathcal{H}}.
\end{aligned}$$

By summing over $t = 1, \ldots, T$, we have

$$\begin{aligned}
&\|f_{T+1} - f^*\|_{\mathcal{H}}^2 - \|f_1 - f^*\|_{\mathcal{H}}^2 \\
=& -2\eta \sum_{t=1}^{T} \langle \partial \hat{\mathcal{L}}_t(f_t), f_t - f^* \rangle_{\mathcal{H}} + \eta^2 \sum_{t=1}^{T} \|\partial \hat{\mathcal{L}}_t(f_t)\|_{\mathcal{H}}^2.
\end{aligned}$$

Due to the convexity of $\hat{\mathcal{L}}_t(f_t)$, we have

$$R_T \leq \sum_{t=1}^{T} \langle \partial \hat{\mathcal{L}}_t(f_t), f_t - f^* \rangle_{\mathcal{H}} \leq \frac{\|f_1 - f^*\|_{\mathcal{H}}^2}{2\eta} - \frac{\|f_{T+1} - f^*\|_{\mathcal{H}}^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \|\partial \hat{\mathcal{L}}_t(f_t)\|_{\mathcal{H}}^2.$$

Since $f_1 = 0$ and $\|f_{T+1} - f^*\|_{\mathcal{H}}^2 \geq 0$, we have

$$R_T \leq \frac{\|f^*\|_{\mathcal{H}}^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \|\partial \hat{\mathcal{L}}_t(f_t)\|_{\mathcal{H}}^2.$$

We now bound $\|\partial \hat{\mathcal{L}}_t(f_t)\|_{\mathcal{H}}^2$. That is

$$\begin{aligned}
&\|\partial \hat{\mathcal{L}}_t(f_t)\|_{\mathcal{H}}^2 \\
=&\left\| f_t - C \sum_{\mathbf{z}_i \in N_{t,k}^{-y_t}(\mathbf{z}_t)} \mathbb{I}[\ell_h(f, \mathbf{z}_t, \mathbf{z}_i) > 0] \cdot \varphi(\mathbf{z}_t, \mathbf{z}_i) \right\|_{\mathcal{H}}^2 \\
=&\|f_t\|_{\mathcal{H}}^2 + \left\| C \sum_{\mathbf{z}_i \in N_{t,k}^{-y_t}(\mathbf{z}_t)} \mathbb{I}[\ell_h(f, \mathbf{z}_t, \mathbf{z}_i) > 0] \cdot \varphi(\mathbf{z}_t, \mathbf{z}_i) \right\|_{\mathcal{H}}^2 \\
&- 2C \sum_{\mathbf{z}_i \in N_{t,k}^{-y_t}(\mathbf{z}_t)} \mathbb{I}[\ell_h(f, \mathbf{z}_t, \mathbf{z}_i) > 0] \langle f_t, \varphi(\mathbf{z}_t, \mathbf{z}_i) \rangle_{\mathcal{H}}.
\end{aligned} \tag{25}$$

From Lemma 1, we know that the first term of Eq. (25) is bounded by

$$\|f_t\|_{\mathcal{H}}^2 \leq C^2 k^2 (2X^2 - 2\xi_1^2). \tag{26}$$

Now, we bound the second term of Eq.(25). For any $\mathbf{z}_i \in N_{t,k}^{-y_t}(\mathbf{z}_t)$, we have

$$\begin{aligned}
\|\varphi(\mathbf{z}_t, \mathbf{z}_i)\|_{\mathcal{H}}^2 &= k(\mathbf{x}_t, \mathbf{x}_t) - 2k(\mathbf{x}_t, \mathbf{x}_i) + k(\mathbf{x}_i, \mathbf{x}_i) \\
&\leq 2X^2 - 2\xi_1^2.
\end{aligned} \tag{27}$$

Therefore, we have

$$\begin{aligned}
&\left\| C \sum_{\mathbf{z}_i \in N_{t,k}^{-y_t}(\mathbf{z}_t)} \mathbb{I}[\ell_h(f, \mathbf{z}_t, \mathbf{z}_i) > 0] \cdot \varphi(\mathbf{z}_t, \mathbf{z}_i) \right\|_{\mathcal{H}}^2 \\
&\leq C^2 \sum_{\mathbf{z}_i \in N_{t,k}^{-y_t}(\mathbf{z}_t)} \mathbb{I}[\ell_h(f, \mathbf{z}_t, \mathbf{z}_i) > 0] \cdot \|\varphi(\mathbf{z}_t, \mathbf{z}_i)\|_{\mathcal{H}}^2 \\
&\leq C^2 k(2X^2 - 2\xi_1^2). \qquad (28)
\end{aligned}$$

The first inequality holds due to the triangle inequality. The second inequality holds since the number of elements in $N_{t,k}^{-y_t}(\mathbf{z}_t)$ is at most $k$ and the bound derived in Eq. (27).

Next, we bound the third term of Eq.(25). First, using the facts that the decision function is an element of a Reproducing Kernel Hilbert Space (RKHS) and the pairwise loss function $\ell : \mathcal{H} \times \mathcal{Z} \times \mathcal{Z} \to [0, U]$ is bounded, we have

$$\begin{aligned}
\langle f_t, \varphi(\mathbf{z}_t, \mathbf{z}_i) \rangle_{\mathcal{H}} &= \frac{1}{2}(y_t - y_i)\left(f_t(\mathbf{x}_t) - f_t(\mathbf{x}_i)\right) \\
&\geq (1 - U). \qquad (29)
\end{aligned}$$

Hence, we have

$$\begin{aligned}
&- 2C \sum_{\mathbf{z}_i \in N_{t,k}^{-y_t}(\mathbf{z}_t)} \mathbb{I}[\ell_h(f, \mathbf{z}_t, \mathbf{z}_i) > 0] \langle f_t, \varphi(\mathbf{z}_t, \mathbf{z}_i) \rangle_{\mathcal{H}} \\
&\leq 2C \sum_{\mathbf{z}_i \in N_{t,k}^{-y_t}(\mathbf{z}_t)} \mathbb{I}[\ell_h(f, \mathbf{z}_t, \mathbf{z}_i) > 0](U - 1) \\
&\leq 2Ck(U - 1). \qquad (30)
\end{aligned}$$

In the above, the first inequality holds due to the fact of Eq. (29). The second inequality holds since the number of elements in $N_{t,k}^{-y_t}(\mathbf{z}_t)$ is at most $k$.

By combining Eq. (26), Eq. (28), and Eq. (30), we have a bound for Eq. (25). That is,

$$\|\partial \hat{\mathcal{L}}_t(f_t)\|_{\mathcal{H}}^2 \leq C^2 k(k+1)(2X^2 - 2\xi_1^2) + 2Ck(U - 1).$$

We then obtain the bound of $R_T$ in Eq. (23) by summing $\partial \hat{\mathcal{L}}_t(f_t)$ for all $t \in [T]$.

Next, we proof the regret bound $R_T^{++}$ for decision function $f_t^{++}$. We first define $G_t(f)$ as follow:

$$\begin{aligned}
G_t(f) &= \hat{\mathcal{L}}(f, \mathbf{z}_t) - \hat{\mathcal{L}}(f^*, \mathbf{z}_t) \\
&= \left( \frac{1}{2}\|f\|_{\mathcal{H}}^2 + C \sum_{\mathbf{z}_i \in N_{t,k}^{-y_t}(\mathbf{z}_t)} \ell_h(f, \mathbf{z}_t, \mathbf{z}_i) \right) - \left( \frac{1}{2}\|f^*\|_{\mathcal{H}}^2 + C \sum_{\mathbf{z}_i \in N_{t,k}^{-y_t}(\mathbf{z}_t)} \ell_h(f^*, \mathbf{z}_t, \mathbf{z}_i) \right). \qquad (31)
\end{aligned}$$

Hence, we have

$$R_T = \sum_{i=1}^{T} G_t(f_t).$$

Since $G_t(f)$ is convex, we get

$$G_t(f_t) \geq G_t(f_t^{++}) + \langle f_t - f_t^{++}, \partial G_t(f_t^{++}) \rangle_{\mathcal{H}}. \qquad (32)$$

Hence by reordering Eq. (32), we have,

$$\begin{aligned}
G_t(f_t^{++}) &\leq G_t(f_t) + \langle f_t^{++} - f_t, \partial G_t(f_t^{++}) \rangle_{\mathcal{H}} \\
&\leq G_t(f_t) + \|f_t^{++} - f_t\|_{\mathcal{H}} \cdot \|\partial G_t(f_t^{++})\|_{\mathcal{H}}. \qquad (33)
\end{aligned}$$

From Eq. (6) and Eq. (14), we have,

$$\begin{aligned}
\|f_t^{++} - f_t\|_{\mathcal{H}} &= \left\| \sum_{i=s}^{t} (1-\eta)^{t-i} \alpha_{r,i} (k(\mathbf{x}_{c,i}, \cdot) - k(\mathbf{x}_{r,i}, \cdot)) \right\|_{\mathcal{H}} \\
&\leq \left\| \sum_{i=s}^{t} (1-\eta)^{t-i} \alpha_{r,i} \sqrt{2X^2 - 2\xi_2^2} \right\|_{\mathcal{H}} \\
&\leq \sum_{i=s}^{t} (1-\eta)^{t-i} |\alpha_{r,i}| \sqrt{2X^2 - 2\xi_2^2} \\
&\leq \sum_{i=s}^{t} (1-\eta)^{t-i} \gamma\eta \sqrt{2X^2 - 2\xi_2^2} \\
&\leq \left( \frac{1 - (1-\eta)^{t-s+1}}{\eta} \right) \gamma\eta \sqrt{2X^2 - 2\xi_2^2} \\
&\leq \gamma \sqrt{2X^2 - 2\xi_2^2}. \qquad (34)
\end{aligned}$$

where $s$ denotes the first trial that the replacement and the compensation are conducted, and $\mathbf{x}_{c,i}$ and $\mathbf{x}_{r,i}$ are the compensated and the removed support vectors at the $i$-th trial.

Next, we have

$$\begin{aligned}
\|\partial G_t(f_t^{++})\|_{\mathcal{H}} &= \left\| f_t^{++} - C \sum_{\mathbf{z}_i \in N_{t,k}^{-y_t}(\mathbf{z}_t)} \mathbb{I}[\ell_h(f_t^{++}, \mathbf{z}_t, \mathbf{z}_i) > 0] \cdot \varphi(\mathbf{z}_t, \mathbf{z}_i) \right\|_{\mathcal{H}} \\
&\leq \|f_t^{++}\|_{\mathcal{H}} + C \sum_{\mathbf{z}_i \in N_{t,k}^{-y_t}(\mathbf{z}_t)} \mathbb{I}[\ell_h(f_t^{++}, \mathbf{z}_t, \mathbf{z}_i) > 0] \cdot \|\varphi(\mathbf{z}_t, \mathbf{z}_i)\|_{\mathcal{H}} \\
&\leq \|f_t^{++}\|_{\mathcal{H}} + Ck\sqrt{2X^2 - 2\xi_1^2} \\
&\leq 2Ck\sqrt{2X^2 - 2\xi_1^2} + \gamma\sqrt{2X^2 - 2\xi_2^2}. \qquad (35)
\end{aligned}$$

In the above, the first inequality is due to the triangle inequality. The second inequality is due to the property of $k$-nearest opposite support vectors. The third inequality is due to the bound derived in Lemma 3.

Hence by substitute Eq. (34) and Eq. (35) for Eq. (33), we have,

$$\begin{aligned}
G_t(f_t^{++}) &\leq G_t(f_t) + \|f_t^{++} - f_t\|_{\mathcal{H}} \cdot \|\partial G_t(f_t^{++})\|_{\mathcal{H}} \\
&\leq G_t(f_t) + \gamma\sqrt{2X^2 - 2\xi_2^2} \left( 2Ck\sqrt{2X^2 - 2\xi_1^2} + \gamma\sqrt{2X^2 - 2\xi_2^2} \right) \\
&\leq G_t(f_t) + 4\gamma Ck\sqrt{(X^2 - \xi_2^2)(X^2 - \xi_1^2)} + 2\gamma^2(X^2 - \xi_2^2). \qquad (36)
\end{aligned}$$

Hence by suming over $t = 1, \ldots, T$, we have

$$\begin{aligned}
R_T^{++} &\leq \sum_{t=1}^{T} \left( G_t(f_t) + 4\gamma Ck\sqrt{(X^2 - \xi_2^2)(X^2 - \xi_1^2)} + 2\gamma^2(X^2 - \xi_2^2) \right) \\
&\leq \sum_{t=1}^{T} G_t(f_t) + T \left( 4\gamma Ck\sqrt{(X^2 - \xi_2^2)(X^2 - \xi_1^2)} + 2\gamma^2(X^2 - \xi_2^2) \right) \\
&= R_T + T \left( 4\gamma Ck\sqrt{(X^2 - \xi_2^2)(X^2 - \xi_1^2)} + 2\gamma^2(X^2 - \xi_2^2) \right).
\end{aligned}$$

□

**Remarks.** It is noted that although $R_T^{++}$ is a little larger than $R_T$, if the compensated support vector is close enough to the removed support vector, we have $\xi_2 = X$, which yield the same regret bound. This result implies that the decision function learned by the replacement with compensation updating policy can seek the same decision function learned with infinite budgets.

# More Experiments

In this section, we conduct extensive experiments on benchmark real-world datasets to evaluate the performance of our proposed KOIL [1] algorithm with fixed budgets.

## Compared Algorithms

We compare our proposed KOIL with the state-of-the-art online learning algorithms. Since we only focus on online imbalanced learning, for fair comparison, we do not compare with existing batch-trained imbalanced learning algorithms. Specifically, we compare online linear algorithms and kernel-based online learning algorithms with a finite or infinite buffer size.

- "**Perceptron**": the classical perceptron algorithm (Rosenblatt 1958);
- "**OAM$_{\text{seq}}$**": an online linear AUC maximization algorithm (Zhao et al. 2011);
- "**OPAUC**": One-pass AUC maximization (Gao et al. 2013);
- "**NORMA**": online learning with kernels (Kivinen, Smola, and Williamson 2004);
- "**RBP**": Randomized budget perceptron (Cavallanti, Cesa-Bianchi, and Gentile 2007);
- "**Forgetron**": a kernel-based perceptron on a fixed budget (Dekel, Shalev-Shwartz, and Singer 2008);
- "**Projectron/Projectron++**": a bounded kernel-based perceptron (Orabona, Keshet, and Caputo 2009);
- "**KOIL$_{\text{RS++}}$/KOIL$_{\text{FIFO++}}$**": our proposed kernelized online imbalanced learning algorithm with fixed budgets updated by RS++ and FIFO++, respectively.

## Experimental Setup

To make fair comparisons, all algorithms adopt the same setup. We set the learning rate to a small constant $\eta = 0.01$ and apply a 5-fold cross validation to find the penalty cost $C \in 2^{[-10:10]}$. For kernel-based methods, we use the Gaussian kernel and tune its parameter $\sigma \in 2^{[-10:10]}$ by a 5-fold cross validation. For NORMA, we apply a 5-fold cross validation to select $\lambda$ and $\nu \in 2^{[-10:10]}$. For Projectron, we apply a similar 5-fold cross validation to select the parameter of projection difference $\eta \in 2^{[-10:10]}$.

Table 1: Summary of the benchmark datasets.

| Dataset | Samples | Dimensions | $T^-/T^+$ |
|---|---|---|---|
| sonar | 208 | 60 | 1.144 |
| australian | 690 | 14 | 1.248 |
| heart | 270 | 13 | 1.250 |
| ionosphere | 351 | 34 | 1.786 |
| diabetes | 768 | 8 | 1.866 |
| glass | 214 | 9 | 2.057 |
| german | 1,000 | 24 | 2.333 |
| svmguide2 | 391 | 20 | 2.342 |
| segment | 2,310 | 19 | 6.000 |
| satimage | 4,435 | 36 | 9.687 |
| vowel | 528 | 10 | 10.000 |
| letter | 15,000 | 16 | 26.881 |
| poker | 25,010 | 10 | 47.752 |
| shuttle | 43,500 | 9 | 328.546 |

## Experiments on Benchmark Real-world Datasets

We conduct experiments on 14 benchmark datasets obtained from the UCI and LIBSVM websites. The imbalanced ratio ranges from 1.144 to 328.546. The detailed statistics of the datasets is summarized in Table 1.

For each dataset, we conduct 5-fold cross validation on all the algorithms, where four folds of the data are used for training while the rest for test. The 5-fold cross validation is independently repeated four times. We set the buffer size to 100 for each class for all related algorithms, including OAM$_{\text{seq}}$, RBP, and Forgetron. We then average the AUC performance of 20 runs and report the results in Table 2.

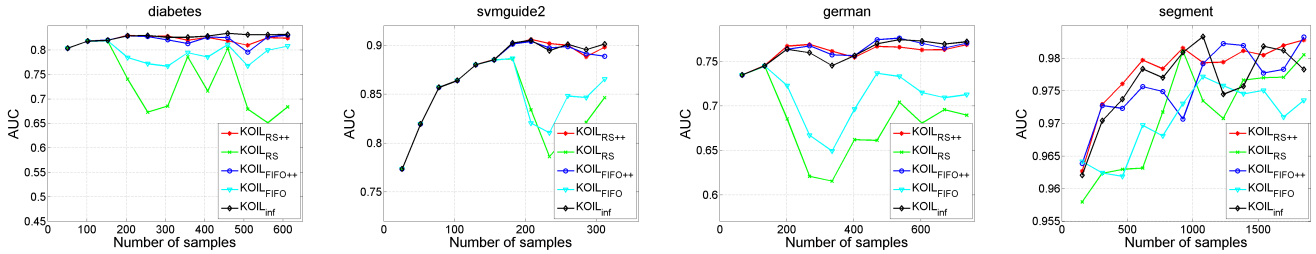
Several observations can be drawn as described in the following:

- Our KOIL with RS++ and FIFO++ updating policies perform better than online linear AUC maximization algorithms in most datasets. By examining the results of OAM$_{\text{seq}}$ on the datasets of australian, heart, diabetes, german, and shuttle and those of OPAUC on australian and german, we speculate that for these datasets, a linear classifier is enough to achieve good performance, while a nonlinear classifier can be affected by outliers.

[1] Demo codes in both C++ and Matlab can be downloaded in `https://www.dropbox.com/sh/nuepinmqzepx54r/AAAKuL4NSZe0IRpGuNIsuxQxa?dl=0`.

Table 2: Average AUC performance (mean±std) on the benchmark datasets, •/◦ (-) indicates that both/one of $\text{KOIL}_{\text{RS++}}$ and $\text{KOIL}_{\text{FIFO++}}$ are/is significantly better (worse) than the corresponding method (pairwise $t$-tests at 95% significance level).

| Data | $\text{KOIL}_{\text{RS++}}$ | $\text{KOIL}_{\text{FIFO++}}$ | Perceptron | $\text{OAM}_{\text{seq}}$ | OPAUC | NORMA | RBP | Forgetron | Projectron | Projectron++ |
|---|---|---|---|---|---|---|---|---|---|---|
| sonar | **.955**±.028 | **.955**±.028 | .803±.083• | .843±.056• | .844±.077• | .925±.044• | .913±.032• | .896±.054• | .896±.049• | .896±.049• |
| australian | .923±.023 | .922±.026 | .869±.035• | **.925**±.024 | .923±.025 | .919±.023 | .911±.017• | .912±.026• | .923±.024 | .923±.024 |
| heart | .908±.040 | .910±.040 | .876±.066• | **.912**±.040 | .901±.043◦ | .890±.051• | .865±.043• | .900±.053 | .902±.038 | .905±.042 |
| ionosphere | **.985**±.015 | **.985**±.015 | .851±.056• | .905±.041• | .888±.046• | .961±.016• | .960±.030• | .945±.031• | .964±.025• | .963±.027• |
| diabetes | .826±.036 | .830±.030 | .726±.059• | .827±.033 | .805±.035• | .792±.032• | .828±.034 | .820±.027◦ | .832±.033 | **.833**±.033 |
| glass | **.887**±.053 | .884±.054 | .810±.065• | .827±.064• | .800±.074• | .811±.077• | .811±.071• | .813±.075• | .811±.070• | .781±.076• |
| german | .769±.032 | .778±.031 | .748±.033• | .777±.027 | **.787**±.026 - | .766±.032◦ | .699±.038• | .712±.054• | .769±.028◦ | .770±.024 |
| svmguide2 | **.897**±.040 | .885±.043 | .860±.037• | .886±.045◦ | .859±.050• | .865±.046• | .890±.038 | .864±.045• | .886±.044◦ | .886±.045◦ |
| segment | .983±.008 | **.985**±.012 | .875±.020• | .919±.020• | .882±.019• | .910±.042• | .969±.017• | .943±.038• | .979±.013• | .978±.016• |
| satimage | **.924**±.012 | .923±.015 | .700±.015• | .755±.018• | .724±.016• | .914±.025• | .899±.018• | .892±.032• | .910±.015• | .904±.011• |
| vowel | **1.000**±.000 | **1.000**±.001 | .848±.070• | .905±.024• | .885±.034• | .996±.005• | .968±.017• | .987±.027• | .982±.013• | .994±.019• |
| letter | .933±.021 | **.942**±.017 | .767±.029• | .827±.021• | .823±.018• | .910±.027• | .928±.011◦ | .815±.102• | .926±.016• | .926±.015• |
| poker | .681±.031 | **.693**±.032 | .514±.030• | .503±.024• | .509±.031• | .577±.040• | .501±.031• | .572±.029• | .675±.027• | .675±.027• |
| shuttle | .950±.040 | .956±.021 | .520±.134• | **.999**±.000 - | .754±.043• | .725±.053• | .844±.041• | .839±.060• | .873±.063• | .795±.063• |
| win/tie/loss | | | 14/0/0 | 9/4/1 | 12/1/1 | 13/1/0 | 12/2/0 | 13/1/0 | 11/3/0 | 10/4/0 |




(a) diabetes (b) svmguide2 (c) german (d) segment

Figure 1: Average AUC performance of four datasets obtained by different updating policies of KOIL.

- In most datasets, kernel-based algorithms show better AUC performance than the linear algorithms in most of datasets. This again demonstrates the power of kernel methods in classifying real-world datasets.
- Our proposed KOIL significantly outperforms all competing kernel-based algorithms in nearly all datasets. The results demonstrate the effectiveness of our KOIL in imbalanced learning.
- We observe that the performance of $\text{OAM}_{\text{seq}}$ on satimage dataset is not as good as that in (Zhao et al. 2011) and (Yang et al. 2013). We check that this is mainly due to the different partition of the training and test data.

## Evaluation on Updating Policies

We test the improvement of our updating policies, RS++ and FIFO++, with the original updating policies, RS and FIFO. We show in Figure 1 for the average AUC performance of 20 runs on four typical datasets. The results of $\text{KOIL}_{\text{inf}}$, i.e., learning with infinite budgets, are provided for reference. We have the following observations:

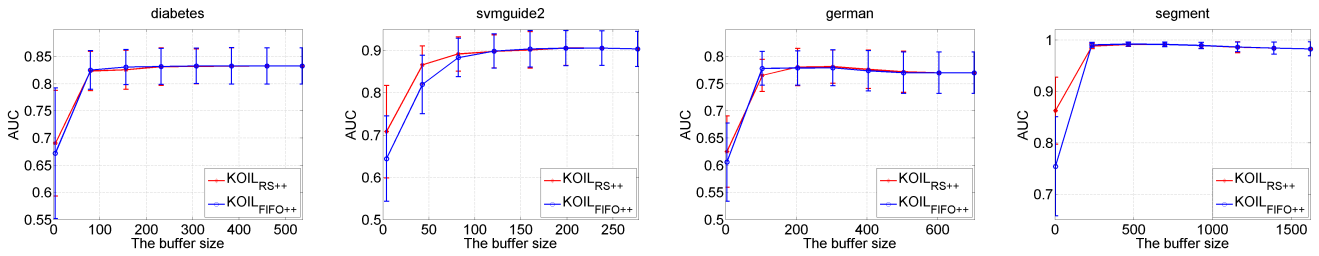
- $\text{KOIL}_{\text{RS++}}$ and $\text{KOIL}_{\text{FIFO++}}$ attain nearly the same performance as $\text{KOIL}_{\text{inf}}$. The results confirm that the extended policies maintain all available classification information during the training.
- Our KOIL with extended updating policies significant outperform the corresponding with original stream oblivious policy when either buffer is full. Without compensation, the performance fluctuates and is easily affected by noisy samples. Differently, with compensation, KOIL can maintain the performance smoothly.

## Sensitivity Evaluation of KOIL

We first test the performance of KOIL with different buffer sizes. From Figure 2, we observe that the performance increases gradually with the increase of the buffer size and it is saturated when the size is relatively large. This is similar to the observations in (Yang et al. 2013; Zhao et al. 2011).

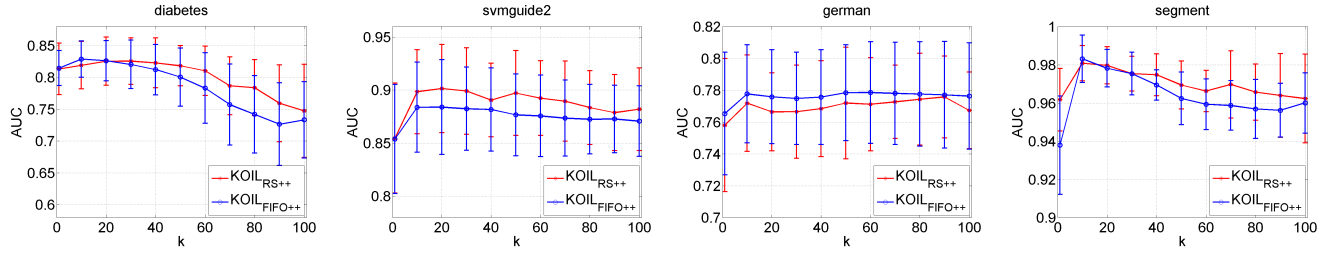
Next, we test the performance of KOIL with different $k$, which determines the number of localized support vectors. From Figure 3, we have the following observations:

- When $k$ is extremely small, say $k = 1$, KOIL only considers the pairwise loss yielded by the nearest opposite support vector of the new instance and can not fully utilize the localized information. The updating weight is similar to NORMA, which adds a constant weight, $|\eta C y_t|$, to the misclassified new instance.




(a) (b) (c) (d)

Figure 2: Average AUC of KOIL with different buffer sizes.




(a) (b) (c) (d)

Figure 3: Average AUC of KOIL with different $k$. Here $k = [1, 10{:}10{:}100]$ and the budget is 100 for each buffer.

- KOIL usually attains the best performance when $k$ equals 10% of the buffer size. The performance decreases when $k$ increases. The results consistently show that by only utilizing the local information of the new instance indeed prevents the effect of outliers.
- For some datasets, e.g., svmguide2 and german, the performance is not so sensitive to $k$. The reason may be that the learned support vectors in these datasets are well-separated when the buffers are full. Hence, new instances play little influence on seeking the decision function.

## Pseudocodes

Here we present the pseudocode for our proposed KOIL algorithm and its key components **UpdateKernel** and **UpdateBuffer**.

**Algorithm 1** Kernelized Online Imbalanced Learning (KOIL) with Fixed Budgets

1: **Input:**
- the penalty parameter $C$ and the learning rate $\eta$
- the maximum budget size $N^+$ and $N^-$
- the number of nearest neighbors $k$

2: **Initialize** $\mathcal{K}^+.\mathcal{A} = \mathcal{K}^-.\mathcal{A} = \emptyset, \mathcal{K}^+.\mathcal{B} = \mathcal{K}^-.\mathcal{B} = \emptyset, N_p = N_n = 0$
3: **for** $t = 1$ **to** $T$ **do**
4: Receive a training sample $\mathbf{z}_t = (\mathbf{x}_t, y_t)$
5: **if** $y_t = +1$ **then**
6: $N_p = N_p + 1$
7: $[\mathcal{K}^-, \mathcal{K}^+, \alpha_t^+] = \text{UpdateKernel}(\mathbf{z}_t, \mathcal{K}^-, \mathcal{K}^+, C, \eta, k)$
8: $\mathcal{K}^+ = \text{UpdateBuffer}(\alpha_t^+, \mathbf{z}_t, \mathcal{K}^+, k, N^+, N_p)$
9: **else**
10: $N_n = N_n + 1$
11: $[\mathcal{K}^+, \mathcal{K}^-, \alpha_t^-] = \text{UpdateKernel}(\mathbf{z}_t, \mathcal{K}^+, \mathcal{K}^-, C, \eta, k)$
12: $\mathcal{K}^- = \text{UpdateBuffer}(\alpha_t^-, \mathbf{z}_t, \mathcal{K}^-, k, N^-, N_n)$
13: **end if**
14: **end for**

Algorithm 3 shows the procedure of the extended Reservoir Sampling (RS++).

- In line 3 to line 4, if the buffer is not full, i.e., $|\mathcal{K}.\mathcal{B}| < N$, the new instance becomes a new support vector and is directly added into the buffer $\mathcal{K}$.
- In line 6 to line 10, if the buffer is full, reservoir sampling is performed. That is, with probability $\frac{N}{N_t}$, we update the buffer by randomly replacing one support vector $\mathbf{z}_r$ in $\mathcal{K}.\mathcal{B}$ with $\mathbf{z}_t$.
- In line 12, if replacement is not conducted, the new instance $\mathbf{z}_t$ is set as the removed support vector $\mathbf{z}_r$.
- In line 14 to line 15, this is the extension of RS. We find the most similar support vector $\mathbf{z}_c$ to the removed support vector $\mathbf{z}_r$, update its weight and put its weight back to the buffer $\mathcal{K}.\mathcal{A}$.

Similarly, we can define the extended FIFO policy, namely **FIFO++**. For FIFO++, the line 6 to line 13 in Algorithm 3 is replaced by removing the first support vector in the buffer and adding the new instance as a new support vector to the end of

**Algorithm 2** UpdateKernel

**Input:**
- the newly received sample with label $\mathbf{z}_t$,
- $\mathcal{K}$ and $\mathcal{K}'$ for support vectors with the opposite and the same label to $\mathbf{z}_t$ respectively,
- the penalty parameter $C$, the learning rate $\eta$, and the number of the nearest neighbors $k$.

**Output:** the updated $\mathcal{K}$, $\mathcal{K}'$ and the weight $\alpha_t$ for $\mathbf{z}_t$
**Initialize:** $V_t = \emptyset$, compute $f_t$ by Eq. (3)
**for** $i \in I_t^{-y_t}$ **do**
  **if** $1 > y_t(f_t(\mathbf{x}_t) - f_t(\mathbf{x}_i))$ **then**
    $V_t = V_t \cup \{i\}$
  **end if**
**end for**
**if** $|V_t| > k$ **then**
  $Sim(i) = k(\mathbf{x}_t, \mathbf{x}_i),\ \forall\, i \in V_t$
  $[Sim', idx]$ =Sort($Sim$,'descend')
  $idx_k = idx(1:k)$
  $V_t = V_t(idx_k)$
**end if**
Update $\alpha_{i,t}$ by Eq. (9)
**return** $\mathcal{K}, \mathcal{K}', \alpha_{t,t}$

**Algorithm 3** UpdateBuffer–RS++

**Input:**
- the received sample $\mathbf{z}_t$ and its weight $\alpha_t$
- the buffer $\mathcal{K}$ to be updated
- the buffer size $N$
- the number of instances received until trial $t$, $N_t$

**Output:** the updated buffer $\mathcal{K}$
**if** $|\mathcal{K}.\mathcal{B}| < N$ **then**
  $\mathcal{K}.\mathcal{A} = \mathcal{K}.\mathcal{A} \cup \{\alpha_t\}, \mathcal{K}.\mathcal{B} = \mathcal{K}.\mathcal{B} \cup \{\mathbf{z}_t\}$
**else**
  Sample $Z$ from a Bernoulli distribution with $\Pr(Z = 1) = N/N_t$
  **if** $Z = 1$ **then**
    Uniformly select an instance $\mathbf{z}_r$
    Update $\mathcal{K}.\mathcal{A}$: $\mathcal{K}.\mathcal{A} = \mathcal{K}.\mathcal{A} \setminus \{\alpha_{r,t}\} \cup \{\alpha_{t,t}\}$
    Update $\mathcal{K}.\mathcal{B}$: $\mathcal{K}.\mathcal{B} = \mathcal{K}.\mathcal{B} \setminus \{\mathbf{z}_r\} \cup \mathbf{z}_t$
  **else**
    $\mathbf{z}_r = \mathbf{z}_t, \alpha_{r,t} = \alpha_{t,t}$
  **end if**
  Find $\mathbf{z}_c = \arg\max_{\mathbf{z}_i \in \mathcal{K}.\mathcal{B}} \{k(\mathbf{x}_r, \mathbf{x}_i)\}$
  Set $\alpha_{c,t} = \alpha_{c,t} + \alpha_{r,t}$ and update $\alpha_{c,t}$ in $\mathcal{K}.\mathcal{A}$
**end if**
**return** $\mathcal{K}$

the buffer.

## References


Cavallanti, G.; Cesa-Bianchi, N.; and Gentile, C. 2007. Tracking the best hyperplane with a simple budget perceptron. *Machine Learning* 69(2-3):143–167.

Dekel, O.; Shalev-Shwartz, S.; and Singer, Y. 2008. The Forgetron: A kernel-based perceptron on a budget. *SIAM Journal on Computing* 37(5):1342–1372.

Gao, W.; Jin, R.; Zhu, S.; and Zhou, Z.-H. 2013. One-pass AUC optimization. In *ICML*, 906–914.

Hoi, S. C. H.; Wang, J.; Zhao, P.; Jin, R.; and Wu, P. 2012. Fast bounded online gradient descent algorithms for scalable kernel-based online learning. In *ICML*.

Kivinen, J.; Smola, A. J.; and Williamson, R. C. 2004. Online learning with kernels. *IEEE Transactions on Signal Processing* 52(8):2165–2176.

Orabona, F.; Keshet, J.; and Caputo, B. 2009. Bounded kernel-based online learning. *Journal of Machine Learning Research* 10:2643–2666.

Rosenblatt, F. 1958. The Perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review* 65(6):386.

Schölkopf, B., and Smola, A. 2002. *Learning with Kernels*. Cambridge, MA: MIT Press.

Yang, H.; Hu, J.; Lyu, M. R.; and King, I. 2013. Online imbalanced learning with kernels. In *NIPS Workshop on Big Learning*.

Zhao, P.; Hoi, S. C. H.; Jin, R.; and Yang, T. 2011. Online AUC maximization. In *ICML*, 233–240.