Supplementary Materials

Online Adaptive Asymmetric Active Learning for Budgeted Imbalanced Data

Anonymous Submission

ABSTRACT

This supplemental file provides the proofs of theorems and additional experiments in our paper of "Online Adaptive Asymmetric Active Learning for Budgeted Imbalanced Data".

A PROOFS OF THEOREMS

This section presents the proofs for all the theorems. For convenience, we introduce the following notations:

$$M_t = \mathbb{I}_{(\hat{y}_t \neq y_t)},\ \rho = \frac{\alpha_p T_n}{\alpha_n T_p} \text{ or } \frac{c_p}{c_n},$$

$\rho_t{=}\rho\mathbb{I}_{(y_t=+1)}{+}\mathbb{I}_{(y_t=-1)},\ \rho_{max}{=}\max\{1,\rho\},\ \rho_{min}{=}\min\{1,\rho\}.$

A.1 Proof of Lemma 1

Lemma 1. *Let $(x_1, y_1), ..., (x_T, y_T)$ be a sequence of input samples, where $x_t \in \mathbb{R}^d$ and $y_t \in \{-1, +1\}$ for all t. Let T_B be the round that runs out of the budgets, i.e., $B_{T_B+1} = B$. For any $\mu \in \mathbb{R}^d$ and any $\delta > 0$, OA3 algorithm satisfies:*

$$\sum_{t=1}^{T_B} M_t Z_t(\delta{+}q_t){\leq}\frac{\delta}{\rho_{min}}\sum_{t=1}^{T_B}\ell_t(\mu){+}\frac{1}{\eta\rho_{min}}\mathrm{Tr}(|\Sigma_{T_B+1}^{-1}|)\times \\ \big[M(\mu) + (1-\delta)^2||\mu||^2\big],$$

where $M(\mu) = \max_t ||\mu_t{-}\mu||^2$.

Proof. Consider that OA3 queries a label but makes a mistake at the round t, so that $Z_t = 1$ and $M_t = 1$. Then, based on the adaptive asymmetric update strategy, we have:

$$\mu_{t+1} = \arg\min_\mu f_t(\mu, \Sigma) = \arg\min_\mu h_t(\mu),$$

where $h_t(\mu) = \frac{1}{2}||\mu_t - \mu||^2_{\Sigma_{t+1}^{-1}} + \eta g_t^\top \mu$.

Since h_t is convex and continuous, one can easily obtain the following inequality:

$$\begin{aligned}&\partial h_t(\mu_{t+1})^\top(\mu - \mu_{t+1}) \\ &{=}\big[(\mu_{t+1} - \mu_t)^\top\Sigma_{t+1}^{-1} + \eta g_t^\top\big](\mu - \mu_{t+1}) \geq 0, \forall\mu.\end{aligned}$$

Rearranging the inequality will give:

$$\begin{aligned}(\eta g_t)^\top(\mu_{t+1} - \mu) \leq&(\mu_{t+1} - \mu_t)^\top\Sigma_{T+1}^{-1}(\mu - \mu_{t+1}) \\ {=}&\frac{1}{2}\Big[||\mu_t - \mu||^2_{\Sigma_{t+1}^{-1}} - ||\mu_{t+1} - \mu||^2_{\Sigma_{t+1}^{-1}} \\ &- ||\mu_t - \mu_{t+1}||^2_{\Sigma_{t+1}^{-1}}\Big]. \qquad (1)\end{aligned}$$

Now, we provide a lower bound for $g_t^\top(\mu_{t+1} - \mu)$. Since $\ell_t(\mu) = \rho_t \max(0, 1{-}y_t x_t^\top \mu_t)$ is a convex function, and based on $g_t^\top = M_t(-\rho_t y_t x_t^\top)$ and

$$\partial h_t(\mu_{t+1}) = 0 \Longleftrightarrow (\mu_{t+1} - \mu_t)^\top\Sigma_{t+1}^{-1} + \eta g_t^\top = 0, \qquad (2)$$

we have:

$$\begin{aligned}&g_t^\top(\mu_{t+1} - \mu) = g_t^\top(\mu_{t+1} - \mu + \mu_t - \mu_t) \\ &{=}M_t(-\rho_t y_t x_t^\top \mu_t){+}M_t(\rho_t y_t x_t^\top \mu){-}\frac{1}{\eta}||\mu_t{-}\mu_{t+1}||^2_{\Sigma_{t+1}^{-1}}. \qquad (3)\end{aligned}$$

Combining the above Equation (3) with the facts:

$$\rho_t M_t(-y_t x_t^\top \mu_t) = \rho_t M_t|y_t x_t^\top \mu_t| = \rho_t M_t|p_t|,$$

and

$$\delta\ell_t(\frac{\mu}{\delta}) \geq \delta\rho_t(1 - y_t x_t^\top\frac{\mu}{\delta}) \ \Leftrightarrow\ y_t x_t^\top \mu \geq \delta - \frac{\delta}{\rho_t}\ell_t(\frac{\mu}{\delta}),$$

we get the following bound for $g_t^\top(\mu_{t+1} - \mu_t)$, *i.e.*,

$$\begin{aligned}&g_t^\top(\mu_{t+1}{-}\mu_t) \\ &\quad\geq \rho_t M_t|p_t|{+}\rho_t M_t\Big[\delta{-}\frac{\delta}{\rho_t}\ell_t(\frac{\mu}{\delta})\Big]{-}\frac{1}{\eta}||\mu_t{-}\mu_{t+1}||^2_{\Sigma_{t+1}^{-1}} \\ &\quad= \rho_t M_t(\delta{+}|p_t|){-}M_t\delta\ell_t(\frac{\mu}{\delta}){-}\frac{1}{\eta}||\mu_t{-}\mu_{t+1}||^2_{\Sigma_{t+1}^{-1}}. \qquad (4)\end{aligned}$$

Combining Equations (1) and (4) will give the following important inequality:

$$\begin{aligned}M_t Z_t(\delta{+}|p_t|) \leq&\frac{Z_t}{2\eta\rho_t}\Big[||\mu_t - \delta\mu||^2_{\Sigma_{t+1}^{-1}} - ||\mu_{t+1} - \delta\mu||^2_{\Sigma_{t+1}^{-1}} \\ &{+}||\mu_t - \mu_{t+1}||^2_{\Sigma_{t+1}^{-1}}\Big]{+}M_t Z_t\Big[\frac{\delta}{\rho_t}\ell_t(\mu)\Big], \quad (5)\end{aligned}$$

where we replace $\delta\mu$ with μ.

Then, according to Equation (2), we have:

$$\begin{aligned}||\mu_t - \mu_{t+1}||^2_{\Sigma_{t+1}^{-1}} &= \eta^2 g_t^\top \Sigma_{t+1} g_t \\ &= M_t \eta^2 \rho_t^2 x_t^\top \Sigma_{t+1} x_t \\ &= M_t \eta^2 \rho_t^2 \Big(x_t^\top \Sigma_t x_t - \frac{x_t^\top \Sigma_t x_t x_t^\top \Sigma_t x_t}{\gamma + x_t^\top \Sigma_t x_t}\Big) \\ &= M_t \eta^2 \rho_t^2 \frac{\gamma v_t}{\gamma + v_t} \\ &= M_t \frac{\eta^2 \rho_t^2}{\frac{1}{\gamma} + \frac{1}{v_t}},\end{aligned}$$

where we use the updating rule of Σ.

KDD'18, August 2018, London, UK
2018. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM. . . $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Then, according to $M_t \leq 1$ and $Z_t \leq 1$, we rearrange Equation (5):

$$\begin{aligned} M_t Z_t(\delta + q_t) &= M_t Z_t\big(\delta{+}|p_t| + c_t\big) \\ &= M_t Z_t\Big(\delta{+}|p_t| - \frac{1}{2}\frac{\eta\rho_{max}}{\frac{1}{\gamma}+\frac{1}{v_t}}\Big) \\ &\leq M_t Z_t\Big(\delta{+}|p_t| - \frac{1}{2}\frac{\eta\rho_t}{\frac{1}{\gamma}+\frac{1}{v_t}}\Big) \\ &\leq \frac{Z_t}{2\eta\rho_t}\Big[||\mu_t - \delta\mu||^2_{\Sigma^{-1}_{t+1}} - ||\mu_{t+1} - \delta\mu||^2_{\Sigma^{-1}_{t+1}}\Big] \\ &\quad + M_t Z_t\Big[\frac{\delta}{\rho_t}\ell_t(\mu)\Big] \\ &\leq \frac{Z_t}{2\eta\rho_t}\Big[||\mu_t - \delta\mu||^2_{\Sigma^{-1}_{t+1}} - ||\mu_{t+1} - \delta\mu||^2_{\Sigma^{-1}_{t+1}}\Big] \\ &\quad + \frac{\delta}{\rho_{min}}\ell_t(\mu). \qquad (6) \end{aligned}$$

We highlight the analysis here provides the theoretical guarantees for the definition of query confidence c_t, which fascinates the theoretical studies of the proposed algorithm. Next, summing the first right term of above inequality over $t = 1, ..., T_B$, we have:

$$\begin{aligned} &\sum_{t=1}^{T_B}\frac{Z_t}{\rho_t}\Big[||\mu_t - \delta\mu||^2_{\Sigma^{-1}_{t+1}} - ||\mu_{t+1} - \delta\mu||^2_{\Sigma^{-1}_{t+1}}\Big] \\ &\leq \frac{1}{\rho_{min}}\Big\{||\mu_1{-}\delta\mu||^2_{\Sigma_2^{-1}}{+}\sum_{t=2}^{T_B}\big[||\mu_t{-}\delta\mu||^2_{\Sigma^{-1}_{t+1}}{-}||\mu_t{-}\delta\mu||^2_{\Sigma_t^{-1}}\big]\Big\} \\ &= \frac{1}{\rho_{min}}\Big[||\mu_1 - \delta\mu||^2_{\Sigma_2^{-1}} + \sum_{t=2}^{T_B}||\mu_t - \delta\mu||^2_{(\Sigma^{-1}_{t+1}-\Sigma^{-1}_t)}\Big] \\ &\leq \frac{1}{\rho_{min}}\Big[||\mu_1{-}\delta\mu||^2\lambda_{max}(\Sigma_2^{-1}){+}\sum_{t=2}^{T_B}||\mu_t{-}\delta\mu||^2\lambda_{max}(\Sigma^{-1}_{t+1}{-}\Sigma^{-1}_t)\Big] \\ &\leq \frac{1}{\rho_{min}}\Big[||\mu_1{-}\delta\mu||^2\mathrm{Tr}(\Sigma_2^{-1}){+}\sum_{t=2}^{T_B}||\mu_t{-}\delta\mu||^2\mathrm{Tr}(\Sigma^{-1}_{t+1}{-}\Sigma^{-1}_t)\Big] \\ &\leq \frac{1}{\rho_{min}}\max_{t\leq T_B}||\mu_t{-}\delta\mu||^2\mathrm{Tr}(\Sigma^{-1}_{T_B+1}) \\ &\leq \frac{2}{\rho_{min}}\big[M(\mu) + (1-\delta)^2||\mu||^2\big]\mathrm{Tr}(|\Sigma^{-1}_{T_B+1}|), \qquad (7) \end{aligned}$$

where $M(\mu) = \max_t ||\mu_t - \mu||^2$ and $\lambda_{max}(\Sigma)$ is the largest eigenvalue of Σ.

Now, combining Inequalities (6) and (7), we can easily obtain:

$$\sum_{t=1}^{T_B} M_t Z_t(\delta + q_t) \leq \frac{\delta}{\rho_{min}}\sum_{t=1}^{T_B}\ell_t(\mu) + \frac{1}{\eta\rho_{min}}\mathrm{Tr}(|\Sigma^{-1}_{T_B+1}|)\times \big[M(\mu) + (1-\delta)^2||\mu||^2\big].$$

Consider another situation that $M_t Z_t = 0$, and we can find above inequality still holds. As results, we conclude the proofs of Lemma 1. □

A.2 Proof of Theorem 1

Theorem 1. *Let $(x_1, y_1), ..., (x_T, y_T)$ be a sequence of input samples, where $x_t \in \mathbb{R}^d$ and $y_t \in \{-1, +1\}$ for all t. Let T_B be the round that runs out of the budgets, i.e., $B_{T_B+1} = B$. For any $\mu \in \mathbb{R}^d$, the expected mistake number of OA3 within budgets is bounded by:*

$$\begin{aligned} \mathbb{E}\Big[\sum_{t=1}^{T_B} M_t\Big] &= \mathbb{E}\Big[\sum_{\substack{t=1\\ y_t=+1}}^{T_B} M_t + \sum_{\substack{t=1\\ y_t=-1}}^{T_B} M_t\Big] \\ &\leq \frac{1}{\rho_{min}}\Big[\sum_{t=1}^{T_B}\ell_t(\mu) + \frac{1}{\eta}D(\mu)\mathrm{Tr}(|\Sigma^{-1}_{T_B+1}|)\Big], \end{aligned}$$

where $D(\mu) = \max\Big\{\frac{M(\mu)+(1-\delta_+)^2||\mu||^2}{\delta_+}, \frac{M(\mu)+(1-\delta_-)^2||\mu||^2}{\delta_-}\Big\}$.

Proof. Considering that OA3 queries a label but makes a mistake at the round t, so that $Z_t = 1$ and $M_t = 1$, there are two scenarios. That is, $p_t \geq 0$ with $M_t Z_t = 1$ represents our estimated class of sample x_t is positive, but true label is negative; while $p_t < 0$ with $M_t Z_t = 1$ represents our estimated class of sample x_t is negative, but true label is positive.

First, if $p_t \geq 0$, based on Lemma 1, for any $\mu \in \mathbb{R}^d$ and any $\delta_+ > 0$, we have :

$$\sum_{\substack{t=1\\ y_t=-1}}^{T_B} M_t Z_t(\delta_+ + q_t) \leq \frac{\delta_+}{\rho_{min}}\sum_{\substack{t=1\\ y_t=-1}}^{T_B}\ell_t(\mu) + \frac{\mathbb{I}_{(y_t=-1)}}{\eta\rho_{min}}\times \big[M(\mu) + (1-\delta_+)^2||\mu||^2\big]\mathrm{Tr}(|\Sigma^{-1}_{T_B+1}|).$$

One can easily prove that this inequality still holds for $M_t Z_t = 0$.

Now, we would like to remove the random variable Z_t. First, when the query parameter $q_t > 0$, taking the expectation over random variables $\mathbb{E}(Z_t) = \frac{\delta_+}{\delta_+ + q_t}$ for $p_t \geq 0$, we have:

$$\mathbb{E}\Big[\sum_{\substack{t=1\\ y_t=-1}}^{T_B} \delta_+ M_t\Big] \leq \frac{\delta_+}{\rho_{min}}\sum_{\substack{t=1\\ y_t=-1}}^{T_B}\ell_t(\mu) + \frac{\mathbb{I}_{(y_t=-1)}}{\eta\rho_{min}}\times \big[M(\mu) + (1-\delta_+)^2||\mu||^2\big]\mathrm{Tr}(|\Sigma^{-1}_{T_B+1}|).$$

On the other hand, when the query parameter $q_t \leq 0$, we set $q_t = 0$, and the random variables satisfy $\mathbb{E}(Z_t) = 1$. Then, we find the above inequality still holds. In addition, one can easily prove this inequality holds for $M_t = 0$.

Now, we obtain:

$$\mathbb{E}\Big[\sum_{\substack{t=1\\ y_t=-1}}^{T_B} M_t\Big] \leq \sum_{\substack{t=1\\ y_t=-1}}^{T_B}\frac{\ell_t(\mu)}{\rho_{min}} + \frac{\mathbb{I}_{(y_t=-1)}}{\eta\rho_{min}\delta_+}\times \big[M(\mu) + (1-\delta_+)^2||\mu||^2\big]\mathrm{Tr}(|\Sigma^{-1}_{T_B+1}|). \quad (8)$$

Similarly, when $p_t < 0$, for any $\mu \in \mathbb{R}^d$ and any $\delta_- > 0$, we obtain:

$$\mathbb{E}\Big[\sum_{\substack{t=1\\ y_t=+1}}^{T_B} M_t\Big] \leq \sum_{\substack{t=1\\ y_t=+1}}^{T_B}\frac{\ell_t(\mu)}{\rho_{min}} + \frac{\mathbb{I}_{(y_t=+1)}}{\eta\rho_{min}\delta_-}\times \big[M(\mu) + (1-\delta_-)^2||\mu||^2\big]\mathrm{Tr}(|\Sigma^{-1}_{T_B+1}|). \quad (9)$$

Summing Equations (8) and (9) will give:

$$
\begin{aligned}
\mathbb{E}\Big[\sum_{t=1}^{T_B} M_t\Big] &= \mathbb{E}\Big[\sum_{\substack{t=1\\y_t=+1}}^{T_B} M_t + \sum_{\substack{t=1\\y_t=-1}}^{T_B} M_t\Big]\\
&\leq \sum_{t=1}^{T_B} \frac{\ell_t(\mu)}{\rho_{min}} + \frac{1}{\eta\rho_{min}} D(\mu)\mathrm{Tr}(|\Sigma_{T_B+1}^{-1}|)\\
&= \frac{1}{\rho_{min}}\Big[\sum_{t=1}^{T_B} \ell_t(\mu) + \frac{1}{\eta} D(\mu)\mathrm{Tr}(|\Sigma_{T_B+1}^{-1}|)\Big],
\end{aligned}
$$

where $D(\mu){=}\max\big\{\frac{M(\mu)+(1-\delta_+)^2||\mu||^2}{\delta_+}, \frac{M(\mu)+(1-\delta_-)^2||\mu||^2}{\delta_-}\big\}$.

Then, we conclude the proofs of Theorem 1. □

A.3 Proof of Theorem 2

Theorem 2. *Under the same condition in Theorem 1, by setting $\rho = \frac{\alpha_p T_n}{\alpha_n T_p}$, the proposed OA3 within budgets satisfies for any $\mu \in \mathbb{R}^d$:*

$$
\mathbb{E}\Big[sum\Big] \geq 1 - \frac{\alpha_n \rho_{max}}{T_n \rho_{min}}\Big[\sum_{t=1}^{T_B} \ell_t(\mu){+}\frac{1}{\eta} D(\mu)\mathrm{Tr}(|\Sigma_{T_B+1}^{-1}|)\Big].
$$

Proof. According to Equation (9), we have:

$$
\begin{aligned}
\mathbb{E}\Big[\sum_{\substack{t=1\\y_t=+1}}^{T_B} \rho M_t\Big] \leq \sum_{\substack{t=1\\y_t=+1}}^{T_B} \frac{\rho\ell_t(\mu)}{\rho_{min}} + \frac{\rho\mathbb{I}_{(y_t=+1)}}{\eta\rho_{min}\delta_-}\times&\\
\big[M(\mu) + (1-\delta_-)^2||\mu||^2\big]\mathrm{Tr}(|\Sigma_{T_B+1}^{-1}|)&.
\end{aligned}
$$

Now, combining with Equation (8) will give:

$$
\begin{aligned}
\mathbb{E}\Big[\sum_{t=1}^{T_B} \rho_t M_t\Big] &{=} \mathbb{E}\Big[\sum_{\substack{t=1\\y_t=+1}}^{T_B} \rho M_t + \sum_{\substack{t=1\\y_t=-1}}^{T_B} M_t\Big]\\
&\leq \frac{\max\{1,\rho\}}{\rho_{min}}\Big[\sum_{t=1}^{T_B} \ell_t(\mu) + \frac{1}{\eta} D(\mu)\mathrm{Tr}(|\Sigma_{T_B+1}^{-1}|)\Big]\\
&= \frac{\rho_{max}}{\rho_{min}}\Big[\sum_{t=1}^{T_B} \ell_t(\mu){+}\frac{1}{\eta} D(\mu)\mathrm{Tr}(|\Sigma_{T_B+1}^{-1}|)\Big]. \quad (10)
\end{aligned}
$$

Now, from the definition of the weighted *sum*, we have:

$$
\begin{aligned}
&\mathbb{E}\Big[\sum_{t=1}^{T_B} \rho_t M_t\Big] {=} \rho\mathbb{E}\big[M_p\big] + \mathbb{E}\big[M_n\big] {=} \Big(\frac{\alpha_p T_n}{\alpha_n T_p}\Big)\mathbb{E}\big[M_p\big] {+} \mathbb{E}\big[M_n\big]\\
&{=} \frac{T_n}{\alpha_n}\Big[\alpha_p \frac{\mathbb{E}\big[M_p\big]}{T_p} {+} \alpha_n \frac{\mathbb{E}\big[M_n\big]}{T_n}\Big] {=} \frac{T_n}{\alpha_n}\Big(1{-}\mathbb{E}\big[sum\big]\Big), \quad (11)
\end{aligned}
$$

where we used $\alpha_p + \alpha_n = 1$.

Combining Equations (10) and (11), we have:

$$
\mathbb{E}\Big[sum\Big] \geq 1 - \frac{\alpha_n \rho_{max}}{T_n \rho_{min}}\Big[\sum_{t=1}^{T_B} \ell_t(\mu){+}\frac{1}{\eta} D(\mu)\mathrm{Tr}(|\Sigma_{T_B+1}^{-1}|)\Big].
$$

Then, we conclude Theorem 2. □

A.4 Proof of Theorem 3

Theorem 3. *Under the same condition in Theorem 1, by setting $\rho{=}\frac{c_p}{c_n}$, the proposed OA3 within budgets satisfies for any $\mu \in \mathbb{R}^d$:*

$$
\mathbb{E}\Big[cost\Big] \leq \frac{c_n \rho_{max}}{\rho_{min}}\Big[\sum_{t=1}^{T_B} \ell_t(\mu){+}\frac{1}{\eta} D(\mu)\mathrm{Tr}(|\Sigma_{T_B+1}^{-1}|)\Big].
$$

Proof. From the definition of *cost* metric, we have:

$$
\begin{aligned}
&\mathbb{E}\Big[\sum_{t=1}^{T_B} \rho_t M_t\Big] {=} \rho\mathbb{E}\big[M_p\big] {+} \mathbb{E}\big[M_n\big] {=} \frac{c_p}{c_n}\mathbb{E}\big[M_p\big] + \mathbb{E}\big[M_n\big]\\
&= \frac{1}{c_n}\Big(c_p\mathbb{E}\big[M_p\big] {+} c_n\mathbb{E}\big[M_n\big]\Big) {=} \frac{1}{c_n}\mathbb{E}\Big[cost\Big]. \quad (12)
\end{aligned}
$$

Combining both Equations (10) and (12) concludes Theorem 3. □

A.5 Proof of Theorem 4

Theorem 4. *Let $(x_1, y_1), ..., (x_T, y_T)$ be a sample stream, where $x_t \in \mathbb{R}^d$ and $y_t \in \{-1, +1\}$. Let T_B be the round that uses up the budgets, i.e., $B_{T_B+1}{=}B$. For any $\mu{\in}\mathbb{R}^d$, the expected mistakes of OA3 over budgets is bounded by:*

$$
\mathbb{E}\Big[\sum_{T_B+1}^{T} M_t\Big] \leq \sum_{T_B+1}^{T}\Big[\frac{\ell_t(\mu)}{\rho_{min}} + y_t x_t^\top \mu_{T_B+1}\Big],
$$

where μ_{T_B+1} is the predictive vector of model, trained by all the previous queried samples.

Proof. When running out of budget, the sample sequence is from $(x_{T_B+1}, y_{T_B+1}), ..., (x_T, y_T)$. Now, for any t after T_B, the predictive vector $\mu_{t+1} = \mu_t = \mu_{T_B+1}$. Combining this with the fact:

$$
\ell_t(\mu) \geq \rho_t(1 - y_t x_t^\top \mu) \;\Leftrightarrow\; y_t x_t^\top \mu \geq 1 - \frac{1}{\rho_t}\ell_t(\mu),
$$

we have:

$$
\begin{aligned}
M_t &\leq M_t\Big[y_t x_t^\top \mu_{T_B+1} + \frac{\ell_t(\mu)}{\rho_t}\Big]\\
&\leq y_t x_t^\top \mu_{T_B+1} + \frac{\ell_t(\mu)}{\rho_{min}},
\end{aligned}
$$

where we use $M_t \leq 1$. Then, we obtain:

$$
\mathbb{E}\Big[\sum_{T_B+1}^{T} M_t\Big] \leq \sum_{T_B+1}^{T}\Big[\frac{\ell_t(\mu)}{\rho_{min}} + y_t x_t^\top \mu_{T_B+1}\Big],
$$

which concludes Theorem 4. □

A.6 Proof of Theorem 5

Theorem 5. *Under the same condition in Theorem 4, by setting $\rho = \frac{\alpha_p T_n}{\alpha_n T_p}$, the sum performance of OA3 over budgets satisfies for any $\mu \in \mathbb{R}^d$:*

$$
\mathbb{E}\Big[sum\Big] \geq 1 - \frac{\alpha_n \rho_{max}}{T_n}\sum_{T_B+1}^{T}\Big[\frac{\ell_t(\mu)}{\rho_{min}} + y_t x_t^\top \mu_{T_B+1}\Big].
$$

Proof. From Theorem 4, we have:

$$\mathbb{E}\Big[\sum_{\substack{T_B+1\\ y_t=+1}}^{T} \rho M_t\Big] \le \sum_{\substack{T_B+1\\ y_t=+1}}^{T} \rho\Big[\ell_t(\mu)+y_t x_t^\top \mu_{T_B+1}\Big], \tag{13}$$

$$\mathbb{E}\Big[\sum_{\substack{T_B+1\\ y_t=-1}}^{T} M_t\Big] \le \sum_{\substack{T_B+1\\ y_t=-1}}^{T} \Big[\ell_t(\mu)+y_t x_t^\top \mu_{T_B+1}\Big]. \tag{14}$$

According to Equations (11), (13) and (14), we have:

$$\mathbb{E}\Big[sum\Big] \ge 1 - \frac{\alpha_n}{T_n} \times \max\{1,\rho\} \sum_{T_B+1}^{T} \Big[\frac{\ell_t(\mu)}{\rho_{min}} + y_t x_t^\top \mu_{T_B+1}\Big]$$

$$\ge 1 - \frac{\alpha_n \rho_{max}}{T_n} \sum_{T_B+1}^{T} \Big[\frac{\ell_t(\mu)}{\rho_{min}} + y_t x_t^\top \mu_{T_B+1}\Big],$$

which concludes Theorem 5. □

A.7 Proof of Theorem 6

Theorem 6. *Under the same condition in Theorem 4, by setting* $\rho = \frac{c_p}{c_n}$*, the misclassification cost of OA3 over budgets satisfies for any* $\mu \in \mathbb{R}^d$*:*

$$\mathbb{E}\Big[cost\Big] \le c_n \rho_{max} \sum_{T_B+1}^{T} \Big[\frac{\ell_t(\mu)}{\rho_{min}} + y_t x_t^\top \mu_{T_B+1}\Big].$$

Proof. Based on Equations (12), (13) and (14), we have:

$$\mathbb{E}\Big[cost\Big] \le c_n \max\{1,\rho\} \sum_{T_B+1}^{T} \Big[\frac{\ell_t(\mu)}{\rho_{min}} + y_t x_t^\top \mu_{T_B+1}\Big]$$

$$\le c_n \rho_{max} \sum_{T_B+1}^{T} \Big[\frac{\ell_t(\mu)}{\rho_{min}} + y_t x_t^\top \mu_{T_B+1}\Big],$$

which concludes Theorem 6. □

B ADDITIONAL EXPERIMENTS

This section provides the additional experimental results.

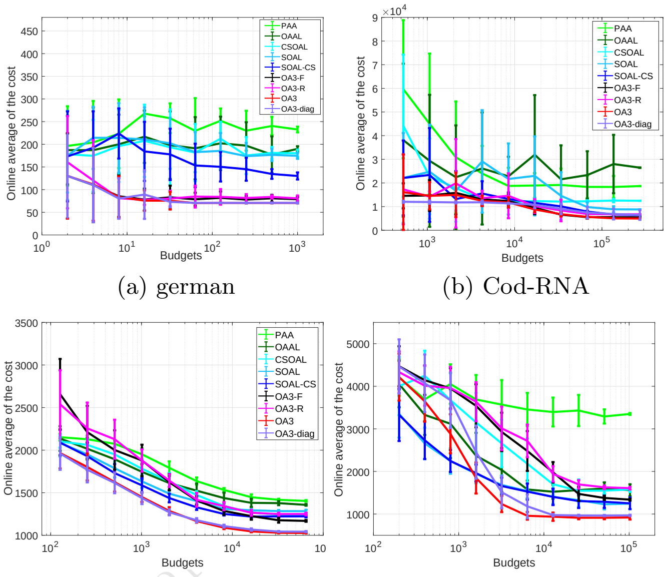
B.1 Cost Evaluation on Varying Budgets

In this subsection, we examine all the algorithms based on the *cost* metric with varying budgets. From Figure 1, our proposed algorithms consistently outperform all other algorithms over a wide range of budgets, which is consistent with the *sum* results, and confirms the effectiveness and robustness of our algorithms again.

B.2 Cost Evaluation of Query Biases

This subsection evaluates the influence of the query biases on *cost* results under fixed budgets, where both query biases (δ_+ and δ_-) are selected from $[10^{-5}, ..., 10^{5}]$, and other parameters are fixed. From Figure 2, we draw several observations.

First, the best results (*i.e.,* deep blue) are usually achieved when $\delta_+ \in \{10, 10^2, 10^3, 10^4\}$ and $\delta_- \in \{1, 10\}$. This observation suggests the potential settings of query biases.



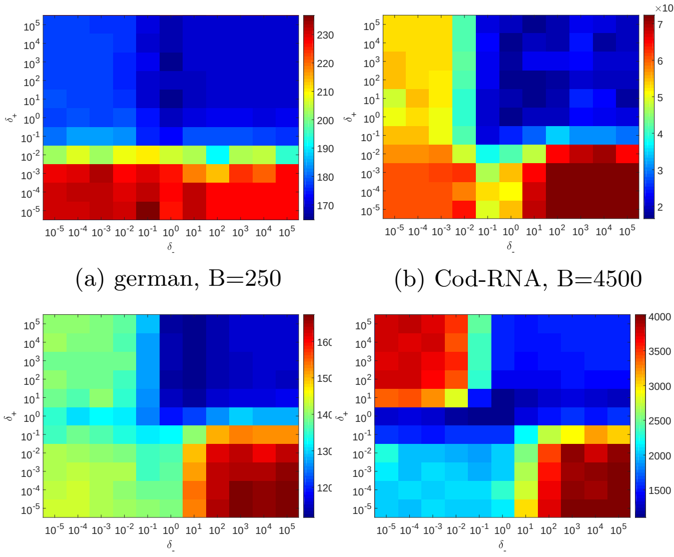
(a) german (b) Cod-RNA

(c) w8a (d) KDDCUP08

Figure 1: Evaluation of cost with varying budgets.

Note that the best result on **KDDCUP08** are obtained when $\delta_+ \in \{1, 10\}$ and $\delta_- \in \{10^{-2}, 10^{-1}, 1\}$. This implies there are no absolutely unified parameter settings for all applications. Nevertheless, the performance of recommended settings is still good. Thus, it would be better to adjust parameters using the recommended settings until desirable performance.

Second, when both δ_+ and δ_- are large (*i.e.,* the upper right corner), OA3 obtains relatively good performance; while when both δ_+ and δ_- are small (*i.e.,* the bottom left corner), OA3 performs relatively bad. This observation further validates the findings in *sum* results.



(a) german, B=250 (b) Cod-RNA, B=4500

(c) w8a, B=16000 (d) KDDCUP08, B=12000

Figure 2: Evaluation of query biases.

Finally, OA3 with large δ_+ and small δ_- (the upper left corner) outperforms the performance with large δ_- and small δ_+ (the bottom right corner). This means that OA3 performs better when querying more samples with the positive prediction and training itself by more positive samples. The main reason is that the positive samples are often more important in real-world anomaly detection. Thus, our algorithms would be more effective in practical tasks due to the good algorithm characteristics, compared with the algorithms that treat all data equally, or tend to query more negative data.

B.3 Evaluation of Robustness

We have demonstrated the robustness of our algorithms from the perspectives of standard deviations and varying budgets in previous experiments. In this subsection, we further explore the stability based on the performance distribution. In detail, since we conduct experiments over 20 random permutations on each dataset, we can exhibit the performance distributions on each algorithm in violin diagrams, *i.e.*, one updated version of the box plot. As results, Figure 3 and Figure 4 record the distributions of all algorithm performance on both metrics under fixed budgets.

By evaluating both *sum* and *cost*, our algorithms outperform all other algorithms with fewer performance fluctuations and lower standard deviations. This result confirms the effectiveness and robustness of our algorithms again.

By comparing the algorithms against imbalance problems, including OAAL, CSOAL, SOAL-CS and our algorithms, we draw several observations based on performance volatility and standard deviations.

First, OAAL shows relatively poor performance with higher performance volatility and higher standard deviations, which means OAAL is more volatile than other algorithms.

Second, CSOAL shows better performance with smaller volatility and lower standard deviations than OAAL. Since CSOAL is "asymmetric update" and OAAL is "asymmetric query", we infer that asymmetric update rules are more stable than asymmetric query rules in imbalance problems.

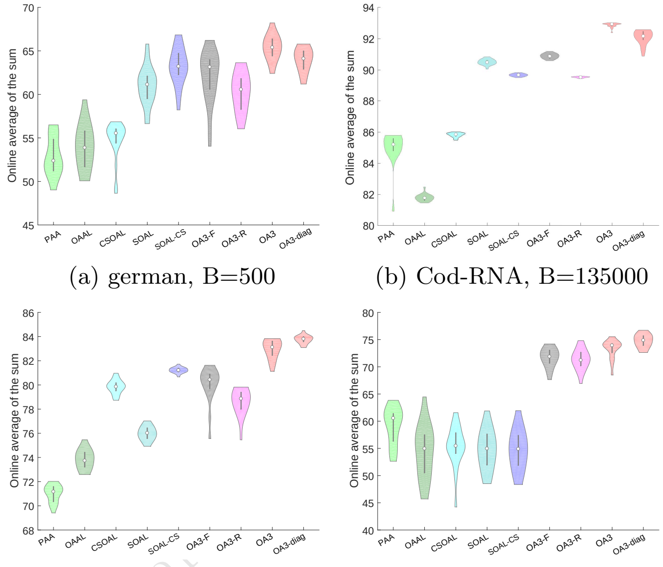
Next, SOAL-CS displays better performance than CSOAL in most cases, which verifies the effectiveness of samples' second-order information in practical applications.

Finally, our proposed algorithms outperform all other algorithms. This discovery confirms the superiority of our proposed asymmetric strategy in solving imbalance problems, and also validates the effectiveness of second-order information in budgeted online active learning.

B.4 Evaluation of Learning Rate

This subsection evaluates the influence of the learning rate on both *sum* and *cost* metrics, where the learning rate is selected from $[10^{-5}, 10^{-4}, ..., 10^{4}, 10^{5}]$.

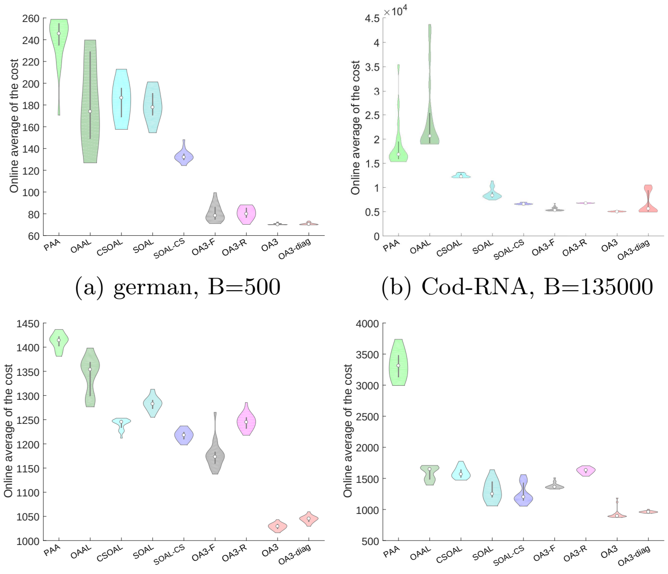
From Figure 5 and Figure 6, we find OA3 algorithms achieve the best result on most datasets when selecting the learning rate from $[10^{-1}, 1, 10^{1}]$. This observation provides a potential choice of the learning rate for algorithm engineers.



(a) german, B=500 (b) Cod-RNA, B=135000

(c) w8a, B=32000 (d) KDDCUP08, B=50000

Figure 3: Evaluation of sum for robustness.

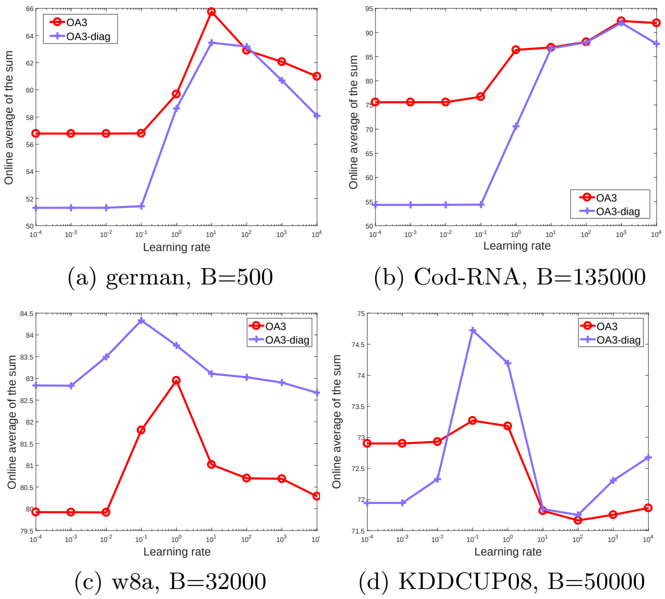


(a) german, B=500 (b) Cod-RNA, B=135000

(c) w8a, B=32000 (d) KDDCUP08, B=50000

Figure 4: Evaluation of cost for robustness.

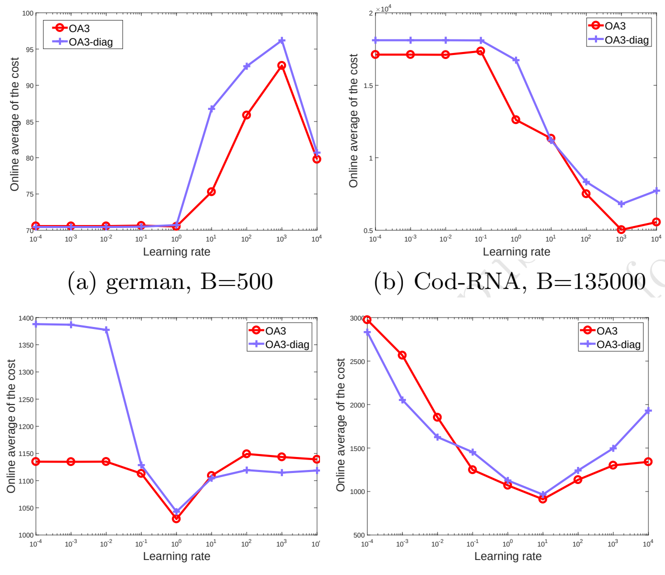
Moreover, given a suitable learning rate, OA3$_{diag}$ achieves a relatively good performance on most datasets, and sometimes even better, compared with OA3. This confirms the competitive power of OA3$_{diag}$. Since we have demonstrated that OA3$_{diag}$ is more efficient than OA3 in previous experiments, we conclude the OA3$_{diag}$ is a favorable choice to balance the performance and efficiency in online anomaly detection tasks.



(a) german, B=500 (b) Cod-RNA, B=135000

(c) w8a, B=32000 (d) KDDCUP08, B=50000

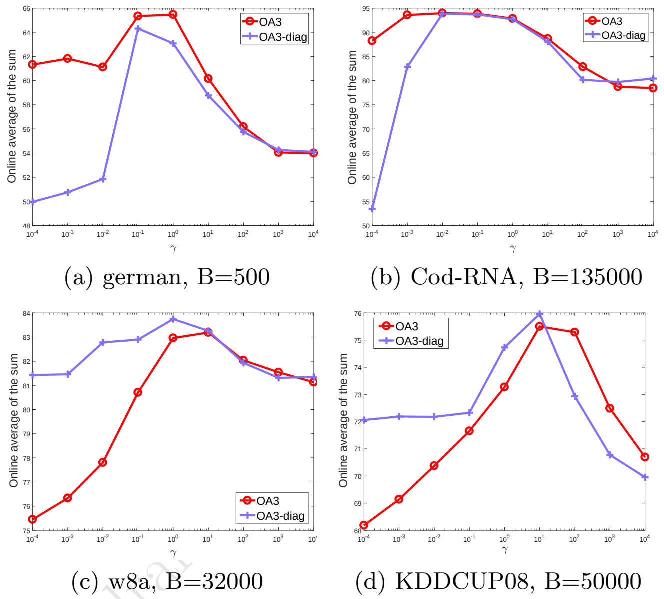
Figure 5: Sum under varying learning rates.



(a) german, B=500 (b) Cod-RNA, B=135000

(c) w8a, B=32000 (d) KDDCUP08, B=50000

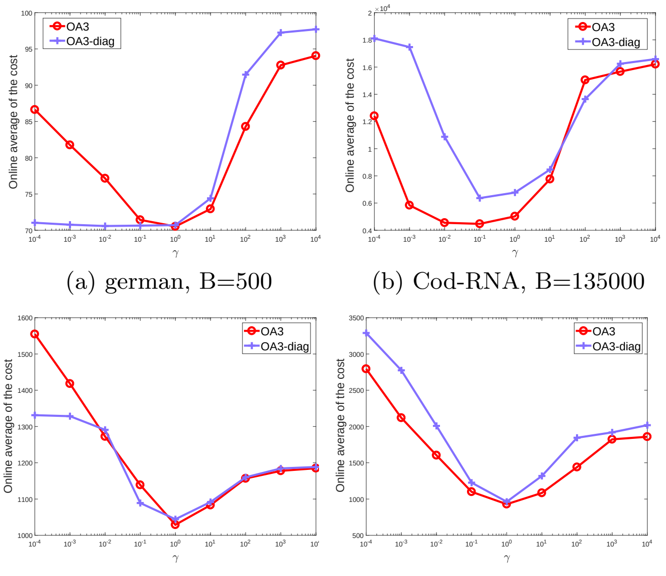
Figure 6: Cost under varying learning rates.



(a) german, B=500 (b) Cod-RNA, B=135000

(c) w8a, B=32000 (d) KDDCUP08, B=50000

Figure 7: Sum with varying regularized parameters.



(a) german, B=500 (b) Cod-RNA, B=135000

(c) w8a, B=32000 (d) KDDCUP08, B=50000

Figure 8: Cost with varying regularized parameters.

B.5 Evaluation of Regularized Parameter

This subsection evaluates the influence of the regularized parameter γ. Let us recall the training process of our algorithms. When receiving a new sample, the model computes the query parameter $q_t = |p_t| + c_t$, where $c_t = -\frac{1}{2}\frac{\eta\rho_{max}}{\frac{1}{v_t}+\frac{1}{\gamma}}$ with default setting γ=1. Then, if deciding to query but making a incorrect prediction, the model trains itself by $\Sigma_{t+1} = \Sigma_t - \frac{\Sigma_t x_t x_t^\top \Sigma_t}{\gamma + x_t^\top \Sigma_t x_t}$ with default setting γ=1. Nevertheless, the rationality of this default setting has not been verified.

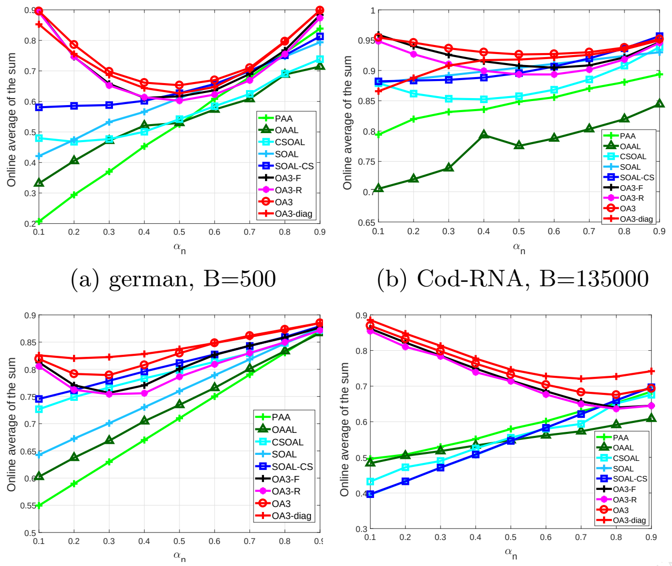
Thus, in this subsection, we examine the performance of our algorithms with different regularized parameters γ from $[10^{-5}, 10^{-4}, ..., 10^4, 10^5]$.

From results in Figure 7 and Figure 8, we find the optimal selection of γ diverse according to datasets. Nevertheless, in most cases, the setting γ=1 achieves the best or fairly good performance. This validates the practical value of our algorithms with the default setting in real-world online anomaly detection tasks.

B.6 Evaluation of Cost Weights

In this subsection, we evaluate the influence of different cost weights, *i.e.*, α_n and c_n, where $\alpha_p = 1 - \alpha_n$ and $c_p = 1 - c_n$. Figure 9 and Figure 10 summarize the results of both metrics under fixed budgets.

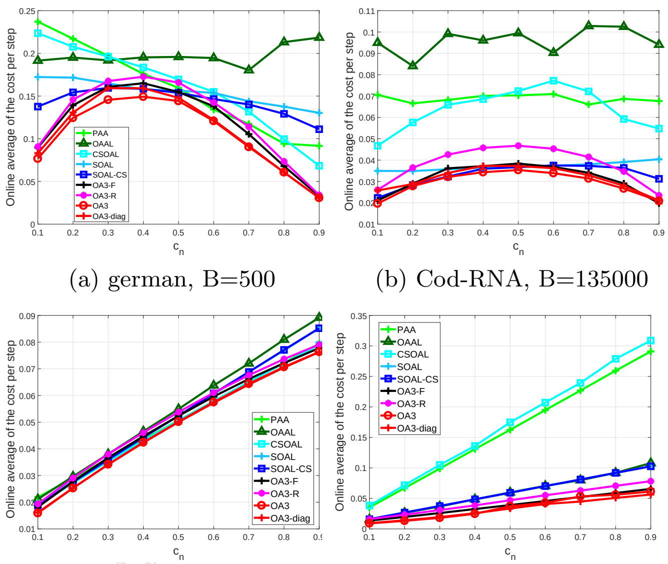
From the results, we find our proposed algorithms consistently outperform all other algorithms with different weights. This observation shows that OA3 and OA3$_{diag}$ algorithms have a wide selection range of cost weights, which further validates the effectiveness of our proposed methods in online anomaly detection tasks.



(a) german, B=500 (b) Cod-RNA, B=135000

(c) w8a, B=32000 (d) KDDCUP08, B=50000

Figure 9: Sum with varying cost weights.



(a) german, B=500 (b) Cod-RNA, B=135000

(c) w8a, B=32000 (d) KDDCUP08, B=50000

Figure 10: Cost with varying cost weights.