Online Passive-Aggressive Multilabel Classification Algorithms

CONTENTS

I. DERIVATION OF PAML-I UPDATE

We mainly focus on the case when $Y_t \neq \varnothing$ and $\bar{Y}_t \neq \varnothing$ since the derivation for the case $Y_t = \varnothing$ or $\bar{Y}_t = \varnothing$ is very simple and we just omit it.

First define the *Lagrangian* associated with problem (4) as

$$\mathcal{L}(\boldsymbol{w}^{(1)}, \cdots, \boldsymbol{w}^{(L+1)}, \xi_1, \xi_2, \alpha, \beta, \lambda, \mu) = \frac{1}{2}\sum_{i=1}^{L+1} ||\boldsymbol{w}^{(i)} - \boldsymbol{w}_t^{(i)}||^2 + C(\xi_1 + \xi_2) - \lambda\xi_1 - \mu\xi_2$$
$$+ \alpha[1 - \xi_1 - (\boldsymbol{x}_t^\top \boldsymbol{w}^{(r_t)} - \boldsymbol{x}_t^\top \boldsymbol{w}^{(L+1)})] + \beta[1 - \xi_2 - (\boldsymbol{x}_t^\top \boldsymbol{w}^{(L+1)} - \boldsymbol{x}_t^\top \boldsymbol{w}^{(s_t)})]$$

where α, β, λ and μ are the Lagrangian multipliers.

Remember that $\boldsymbol{w}_{t+1}^{(1)}, \cdots, \boldsymbol{w}_{t+1}^{(L+1)}, \xi_1^*, \xi_2^*$ are the optimal solutions of the primal problem (4). Let α_t, β_t, λ_t and μ_t denote the dual optimal solutions of (4). Then KKT conditions for problem

(4) include

$$\begin{cases} \nabla_{\boldsymbol{w}^{(r_t)}}\mathcal{L}(\boldsymbol{w}_{t+1}^{(1)},\cdots,\boldsymbol{w}_{t+1}^{(L+1)},\xi_1^*,\xi_2^*,\alpha_t,\beta_t,\lambda_t,\mu_t) = \boldsymbol{w}_{t+1}^{(r_t)} - \boldsymbol{w}_t^{(r_t)} - \alpha_t\boldsymbol{x}_t = 0 \\ \nabla_{\boldsymbol{w}^{(s_t)}}\mathcal{L}(\boldsymbol{w}_{t+1}^{(1)},\cdots,\boldsymbol{w}_{t+1}^{(L+1)},\xi_1^*,\xi_2^*,\alpha_t,\beta_t,\lambda_t,\mu_t) = \boldsymbol{w}_{t+1}^{(s_t)} - \boldsymbol{w}_t^{(s_t)} + \beta_t\boldsymbol{x}_t = 0 \\ \nabla_{\boldsymbol{w}^{(L+1)}}\mathcal{L}(\boldsymbol{w}_{t+1}^{(1)},\cdots,\boldsymbol{w}_{t+1}^{(L+1)},\xi_1^*,\xi_2^*,\alpha_t,\beta_t,\lambda_t,\mu_t) = \boldsymbol{w}_{t+1}^{(L+1)} - \boldsymbol{w}_t^{(L+1)} + (\alpha_t-\beta_t)\boldsymbol{x}_t = 0 \\ \nabla_{\boldsymbol{w}^{(i)}}\mathcal{L}(\boldsymbol{w}_{t+1}^{(1)},\cdots,\boldsymbol{w}_{t+1}^{(L+1)},\xi_1^*,\xi_2^*,\alpha_t,\beta_t,\lambda_t,\mu_t) = \boldsymbol{w}_{t+1}^{(i)} - \boldsymbol{w}_t^{(i)} = 0,\ \forall i \notin \{r_t,s_t,L+1\} \\ \frac{\partial\mathcal{L}(\boldsymbol{w}_{t+1}^{(1)},\cdots,\boldsymbol{w}_{t+1}^{(L+1)},\xi_1^*,\xi_2^*,\alpha_t,\beta_t,\lambda_t,\mu_t)}{\partial\xi_1} = C-\alpha_t-\lambda_t = 0 \\ \frac{\partial\mathcal{L}(\boldsymbol{w}_{t+1}^{(1)},\cdots,\boldsymbol{w}_{t+1}^{(L+1)},\xi_1^*,\xi_2^*,\alpha_t,\beta_t,\lambda_t,\mu_t)}{\partial\xi_2} = C-\beta_t-\mu_t = 0 \\ \alpha_t\ge 0,\beta_t\ge 0,\lambda_t\ge 0,\mu_t\ge 0,\xi_1^*\ge 0,\xi_2^*\ge 0 \\ \lambda_t\xi_1^*=0,\mu_t\xi_2^*=0 \\ 1-\xi_1^*-(\boldsymbol{x}_t^\top\boldsymbol{w}_{t+1}^{(r_t)} - \boldsymbol{x}_t^\top\boldsymbol{w}_{t+1}^{(L+1)}) \le 0 \\ 1-\xi_2^*-(\boldsymbol{x}_t^\top\boldsymbol{w}_{t+1}^{(L+1)} - \boldsymbol{x}_t^\top\boldsymbol{w}_{t+1}^{(s_t)}) \le 0 \\ \alpha_t[1-\xi_1^*-(\boldsymbol{x}_t^\top\boldsymbol{w}_{t+1}^{(r_t)} - \boldsymbol{x}_t^\top\boldsymbol{w}_{t+1}^{(L+1)})] = 0 \\ \beta_t[1-\xi_2^*-(\boldsymbol{x}_t^\top\boldsymbol{w}_{t+1}^{(L+1)} - \boldsymbol{x}_t^\top\boldsymbol{w}_{t+1}^{(s_t)})] = 0 \end{cases}$$

The first four equality constraints give us the same update rule as Eq.(3). So, the key is to solve $\alpha_t,\beta_t,\lambda_t,\mu_t,\xi_1^*$ and ξ_2^*. Now by plugging the first three equality constraints into the last four constraints, and using the definition of $f_{t,1}$ and $f_{t,2}$, we get all conditions that these variables should satisfy,

$$\begin{cases} \alpha_t\ge 0,\beta_t\ge 0,\lambda_t\ge 0,\mu_t\ge 0,\xi_1^*\ge 0,\xi_2^*\ge 0 \\ \lambda_t\xi_1^*=0,\mu_t\xi_2^*=0 \\ \alpha_t+\lambda_t=C,\beta_t+\mu_t=C \\ f_{t,1}+(\beta_t-2\alpha_t)||\boldsymbol{x}_t||^2 \le \xi_1^* \\ f_{t,2}+(\alpha_t-2\beta_t)||\boldsymbol{x}_t||^2 \le \xi_2^* \\ \alpha_t(f_{t,1}+(\beta_t-2\alpha_t)||\boldsymbol{x}_t||^2-\xi_1^*)=0 \\ \beta_t(f_{t,2}+(\alpha_t-2\beta_t)||\boldsymbol{x}_t||^2-\xi_2^*)=0 \end{cases}$$

Next we will take account of different cases of $f_{t,1}$ and $f_{t,2}$, as displayed in Fig. 1.

1) $f_{t,1}\le 0$ and $f_{t,2}\le 0$ (Area ① in Fig. 1)

In this case, $\boldsymbol{W}_t$ is clearly the optimal solution of (4), so $\alpha_t=\beta_t=0$.

2) $f_{t,1} \leq -\frac{1}{2}f_{t,2}$ and $0 < f_{t,2} \leq 2C||\boldsymbol{x}_t||^2$ (Area ② in Fig. 1)

First we prove by contradiction that $\alpha_t < 2\beta_t$. Indeed, if $\alpha_t \geq 2\beta_t$, we will get contradiction with the constraint $\alpha_t \leq C$:

$$\left.\begin{aligned} \alpha_t \geq 2\beta_t, \quad f_{t,2} > 0 \\ f_{t,2} + (\alpha_t - 2\beta_t)||\boldsymbol{x}_t||^2 \leq \xi_2^* \end{aligned}\right\} \Rightarrow \xi_2^* > 0 \Rightarrow \mu_t = 0 \Rightarrow \beta_t = C \Rightarrow \alpha_t \geq 2C$$

Therefore, it holds that $\alpha_t < 2\beta_t$. Since $\alpha_t \geq 0$, we get $\beta_t > 0$, which leads to that $f_{t,2} + (\alpha_t - 2\beta_t)||\boldsymbol{x}_t||^2 = \xi_2^*$. Now using the condition $f_{t,1} \leq -\frac{1}{2}f_{t,2}$, we can get

$$f_{t,1} + (\beta_t - 2\alpha_t)||\boldsymbol{x}_t||^2 - \xi_1^* \leq -\frac{3}{2}\alpha_t||\boldsymbol{x}_t||^2 - \xi_1^* - \frac{1}{2}\xi_2^* \leq 0$$

which further implies that $\alpha_t = 0$. So we can get that $f_{t,2} - 2\beta_t||\boldsymbol{x}_t||^2 = \xi_2^*$. Given that $f_{t,2} \leq 2C||\boldsymbol{x}_t||^2$, we can derive that $\xi_2^* \leq 2(C - \beta_t)||\boldsymbol{x}_t||^2$. Combining the constraint $\xi_2^* \geq 0$, we get $\beta_t \leq C$. If $0 < \beta_t < C$, we have

$$0 < \beta_t < C \Rightarrow \mu_t > 0 \Rightarrow \xi_2^* = 0 \Rightarrow f_{t,2} - 2\beta_t||\boldsymbol{x}_t||^2 = 0 \Rightarrow \beta_t = \frac{f_{t,2}}{2||\boldsymbol{x}_t||^2}$$

If $\beta_t = C$, we have

$$\beta_t = C \Rightarrow \xi_2^* = 0 \Rightarrow \beta_t = \frac{f_{t,2}}{2||\boldsymbol{x}_t||^2}$$

Therefore, if $(f_{t,1}, f_{t,2}) \in$ Area ② in Fig. 1, we can get $\alpha_t = 0$ and $\beta_t = \frac{f_{t,2}}{2||\boldsymbol{x}_t||^2}$.

3) $f_{t,1} \leq -C||\boldsymbol{x}_t||^2$ and $f_{t,2} > 2C||\boldsymbol{x}_t||^2$ (Area ③ in Fig. 1)

First we can prove that

$$\left.\begin{aligned} f_{t,2} > 2C||\boldsymbol{x}_t||^2 \\ 0 \leq \alpha_t \leq C, 0 \leq \beta_t \leq C \\ f_{t,2} + (\alpha_t - 2\beta_t)||\boldsymbol{x}_t||^2 \leq \xi_2^* \end{aligned}\right\} \Rightarrow \xi_2^* > 0 \Rightarrow \mu_t = 0 \Rightarrow \beta_t = C$$

Further, combining the condition $f_{t,1} \leq -C||\boldsymbol{x}_t||^2$, we can get

$$f_{t,1} + (\beta_t - 2\alpha_t)||\boldsymbol{x}_t||^2 - \xi_1^* \leq -2\alpha_t||\boldsymbol{x}_t||^2 - \xi_1^* \leq 0$$

which implies that $\alpha_t = 0$.

4) $-C||\boldsymbol{x}_t||^2 < f_{t,1} \leq C||\boldsymbol{x}_t||^2$ and $f_{t,2} \geq -\frac{1}{2}f_{t,1} + \frac{3C}{2}||\boldsymbol{x}_t||^2$ (Area ④ in Fig. 1)

First we can prove that

$$\left.\begin{aligned} f_{t,1} + (\beta_t - 2\alpha_t)||\boldsymbol{x}_t||^2 \leq \xi_1^* \\ f_{t,2} + (\alpha_t - 2\beta_t)||\boldsymbol{x}_t||^2 \leq \xi_2^* \\ f_{t,2} + \frac{1}{2}f_{t,1} \geq \frac{3C}{2}||\boldsymbol{x}_t||^2 \end{aligned}\right\} \Rightarrow \frac{1}{2}\xi_1^* + \xi_2^* \geq \frac{3(C - \beta_t)}{2}||\boldsymbol{x}_t||^2 \Rightarrow \beta_t = C$$

Otherwise, if $\beta_t < C$, we will get a paradox,

$$\beta_t < C \Rightarrow \left.\begin{array}{r} \mu_t > 0 \Rightarrow \xi_2^* = 0 \\ \frac{1}{2}\xi_1^* + \xi_2^* > 0 \end{array}\right\} \Rightarrow \xi_1^* > 0 \Rightarrow \lambda_t = 0 \Rightarrow \alpha_t = C$$

$$\left.\begin{array}{r} \Rightarrow f_{t,1} + (\beta_t - 2C)||\boldsymbol{x}_t||^2 = \xi_1^* \\ \xi_1^* > 0,\ f_{t,1} \leq C||\boldsymbol{x}_t||^2 \end{array}\right\} \Rightarrow \beta_t > C$$

Next we prove by contradiction that $\alpha_t > 0$:

$$\alpha_t = 0 \Rightarrow \lambda_t = C \Rightarrow \xi_1^* = 0 \Rightarrow f_{t,1} \leq (2\alpha_t - \beta_t)||\boldsymbol{x}_t||^2 = -C||\boldsymbol{x}_t||^2$$

which contradicts with the condition that $f_{t,1} > -C||\boldsymbol{x}_t||^2$. Thus $\alpha_t > 0$ holds. Further,

$$\left.\begin{array}{r} f_{t,1} + (C - 2\alpha_t)||\boldsymbol{x}_t||^2 = \xi_1^* \\ f_{t,1} \leq C||\boldsymbol{x}_t||^2 \end{array}\right\} \left.\begin{array}{r} \Rightarrow \xi_1^* \leq 2(C - \alpha_t)||\boldsymbol{x}_t||^2 \\ \xi_1^* \geq 0 \end{array}\right\} \Rightarrow \alpha_t \leq C$$

Two different cases of α_t are given into account:

$$\left.\begin{array}{r} \alpha_t < C \Rightarrow \lambda_t > 0 \Rightarrow \xi_1^* = 0 \\ \alpha_t = C \Rightarrow \xi_1^* = 0 \end{array}\right\} \Rightarrow f_{t,1} + (C - 2\alpha_t)||\boldsymbol{x}_t||^2 = 0 \Rightarrow \alpha_t = \frac{1}{2}(C + \frac{f_{t,1}}{||\boldsymbol{x}_t||^2})$$

In summary, if $(f_{t,1}, f_{t,2})$ resides in Area ④ in Fig. 1, then $\alpha_t = \frac{1}{2}(C + \frac{f_{t,1}}{||\boldsymbol{x}_t||^2})$ and $\beta_t = C$.

5) $f_{t,1} > C||\boldsymbol{x}_t||^2$ and $f_{t,2} > C||\boldsymbol{x}_t||^2$ (Area ⑤ in Fig. 1)

First we can prove that

$$\left.\begin{array}{r} f_{t,1} + (\beta_t - 2\alpha_t)||\boldsymbol{x}_t||^2 \leq \xi_1^* \\ f_{t,2} + (\alpha_t - 2\beta_t)||\boldsymbol{x}_t||^2 \leq \xi_2^* \\ f_{t,1} > C||\boldsymbol{x}_t||^2, f_{t,2} > C||\boldsymbol{x}_t||^2 \end{array}\right\} \left.\begin{array}{r} \Rightarrow \xi_1^* + \xi_2^* > (2C - \alpha_t - \beta_t)||\boldsymbol{x}_t||^2 \\ \alpha_t \leq C, \beta_t \leq C \end{array}\right\} \Rightarrow \xi_1^* + \xi_2^* > 0$$

If $\xi_1^* = 0$, we will get a paradox,

$$\xi_1^* = 0 \Rightarrow \left.\begin{array}{r} \xi_2^* > 0 \Rightarrow \mu_t = 0 \Rightarrow \beta_t = C \\ f_{t,1} + (\beta_t - 2\alpha_t)||\boldsymbol{x}_t||^2 \leq 0 \end{array}\right\} \left.\begin{array}{r} \Rightarrow f_{t,1} \leq (2\alpha_t - C)||\boldsymbol{x}_t||^2 \\ f_{t,1} > C||\boldsymbol{x}_t||^2 \end{array}\right\} \Rightarrow \alpha_t > C$$

which contradicts with $\alpha_t \leq C$. Therefore, it holds that $\xi_1^* > 0$. Similarly, if $\xi_2^* = 0$, we will get contradiction with $\beta_t \leq C$. Thus, $\xi_2^* > 0$. Combining these results, we get that

$$\left.\begin{array}{r} \xi_1^* > 0 \\ \xi_2^* > 0 \end{array}\right\} \Rightarrow \left.\begin{array}{r} \lambda_t = 0 \\ \mu_t = 0 \end{array}\right\} \Rightarrow \left.\begin{array}{r} \alpha_t = C \\ \beta_t = C \end{array}\right\}$$

6) $f_{t,1} \geq -\frac{1}{2}f_{t,2} + \frac{3C}{2}||\boldsymbol{x}_t||^2$ and $-C||\boldsymbol{x}_t||^2 < f_{t,2} \leq C||\boldsymbol{x}_t||^2$ (Area ⑥ in Fig. 1)

   The derivation progress is similar to that for Area ④ in Fig. 1.

7) $f_{t,1} > 2C||\boldsymbol{x}_t||^2$ and $f_{t,2} \leq -C||\boldsymbol{x}_t||^2$ (Area ⑦ in Fig. 1)

   The derivation progress is similar to that for Area ③ in Fig. 1.

8) $f_{t,2} \leq -\frac{1}{2}f_{t,1}$ and $0 < f_{t,1} \leq 2C||\boldsymbol{x}_t||^2$ (Area ⑧ in Fig. 1)

   The derivation progress is similar to that for Area ② in Fig. 1.

9) $-\frac{1}{2}f_{t,2} < f_{t,1} < -\frac{1}{2}f_{t,2} + \frac{3C}{2}||\boldsymbol{x}_t||^2$ and $-\frac{1}{2}f_{t,1} < f_{t,2} < -\frac{1}{2}f_{t,1} + \frac{3C}{2}||\boldsymbol{x}_t||^2$ (Area ⑨ in Fig. 1)

   First we prove by contradiction that $\alpha_t > 0$:

$$\alpha_t = 0 \Rightarrow \lambda_t = C \Rightarrow \xi_1^* = 0 \Rightarrow f_{t,1} \leq -\beta_t||\boldsymbol{x}_t||^2$$

   Given that $-\frac{1}{2}f_{t,2} < f_{t,1}$, we get that $f_{t,2} > 2\beta_t||\boldsymbol{x}_t||^2$. Further,

$$\left.\begin{aligned} f_{t,2} > 2\beta_t||\boldsymbol{x}_t||^2 \\ f_{t,2} - 2\beta_t||\boldsymbol{x}_t||^2 \leq \xi_2^* \end{aligned}\right\} \Rightarrow \xi_2^* > 0 \Rightarrow \mu_t = 0 \Rightarrow \beta_t = C \Rightarrow f_{t,2} > 2C||\boldsymbol{x}_t||^2$$

   which contradicts with the fact $f_{t,2} < 2C||\boldsymbol{x}_t||^2$ in Area ⑨. So it holds that $\alpha_t > 0$. Similarly, we can also prove that $\beta_t > 0$. Using $\alpha_t > 0$ and $\beta_t > 0$, we can derive that

$$\left.\begin{aligned} f_{t,1} + (\beta_t - 2\alpha_t)||\boldsymbol{x}_t||^2 = \xi_1^* \\ f_{t,2} + (\alpha_t - 2\beta_t)||\boldsymbol{x}_t||^2 = \xi_2^* \end{aligned}\right\} \Rightarrow \left.\begin{aligned} f_{t,1} + \frac{1}{2}f_{t,2} = \xi_1^* + \frac{1}{2}\xi_2^* + \frac{3\alpha_t}{2}||\boldsymbol{x}_t||^2 \\ f_{t,2} + \frac{1}{2}f_{t,1} = \xi_2^* + \frac{1}{2}\xi_1^* + \frac{3\beta_t}{2}||\boldsymbol{x}_t||^2 \\ f_{t,1} + \frac{1}{2}f_{t,2} < \frac{3C}{2}||\boldsymbol{x}_t||^2,\ \xi_1^* \geq 0 \\ f_{t,2} + \frac{1}{2}f_{t,1} < \frac{3C}{2}||\boldsymbol{x}_t||^2,\ \xi_2^* \geq 0 \end{aligned}\right\} \Rightarrow \left.\begin{aligned} \alpha_t < C \\ \beta_t < C \end{aligned}\right\}$$

$$\Rightarrow \left.\begin{aligned} \lambda_t > 0 \\ \mu_t > 0 \end{aligned}\right\} \Rightarrow \left.\begin{aligned} \xi_1^* = 0 \\ \xi_2^* = 0 \end{aligned}\right\} \Rightarrow \left.\begin{aligned} f_{t,1} + (\beta_t - 2\alpha_t)||\boldsymbol{x}_t||^2 = 0 \\ f_{t,2} + (\alpha_t - 2\beta_t)||\boldsymbol{x}_t||^2 = 0 \end{aligned}\right\} \Rightarrow \left.\begin{aligned} \alpha_t = \frac{2f_{t,1} + f_{t,2}}{3||\boldsymbol{x}_t||^2} \\ \beta_t = \frac{f_{t,1} + 2f_{t,2}}{3||\boldsymbol{x}_t||^2} \end{aligned}\right\}$$

By merging similar cases into one case, we get the desired solution for problem (4).

II. Derivation of PAML-II update

Similarly, our derivation focuses on the case where $Y_t \neq \varnothing$ and $\bar{Y}_t \neq \varnothing$. The *Lagrangian* associated with problem (5) is defined as

$$\mathcal{L}(\boldsymbol{w}^{(1)}, \cdots, \boldsymbol{w}^{(L+1)}, \xi_1, \xi_2, \alpha, \beta) = \frac{1}{2}\sum_{i=1}^{L+1} ||\boldsymbol{w}^{(i)} - \boldsymbol{w}_t^{(i)}||^2 + C(\xi_1^2 + \xi_2^2)$$
$$+ \alpha[1 - \xi_1 - (\boldsymbol{x}_t^\top \boldsymbol{w}^{(r_t)} - \boldsymbol{x}_t^\top \boldsymbol{w}^{(L+1)})] + \beta[1 - \xi_2 - (\boldsymbol{x}_t^\top \boldsymbol{w}^{(L+1)} - \boldsymbol{x}_t^\top \boldsymbol{w}^{(s_t)})]$$

Let α_t and β_t denote the dual optimal solutions of problem (5). Remember that $\boldsymbol{w}_{t+1}^{(1)}, \cdots, \boldsymbol{w}_{t+1}^{(L+1)}, \xi_1^*$ and ξ_2^* are the primal optimal solutions of (5). Then KKT conditions for problem (5) include

$$\begin{cases}
\nabla_{\boldsymbol{w}^{(r_t)}} \mathcal{L}(\boldsymbol{w}_{t+1}^{(1)}, \cdots, \boldsymbol{w}_{t+1}^{(L+1)}, \xi_1^*, \xi_2^*, \alpha_t, \beta_t) = \boldsymbol{w}_{t+1}^{(r_t)} - \boldsymbol{w}_t^{(r_t)} - \alpha_t \boldsymbol{x}_t = 0 \\
\nabla_{\boldsymbol{w}^{(s_t)}} \mathcal{L}(\boldsymbol{w}_{t+1}^{(1)}, \cdots, \boldsymbol{w}_{t+1}^{(L+1)}, \xi_1^*, \xi_2^*, \alpha_t, \beta_t) = \boldsymbol{w}_{t+1}^{(s_t)} - \boldsymbol{w}_t^{(s_t)} + \beta_t \boldsymbol{x}_t = 0 \\
\nabla_{\boldsymbol{w}^{(L+1)}} \mathcal{L}(\boldsymbol{w}_{t+1}^{(1)}, \cdots, \boldsymbol{w}_{t+1}^{(L+1)}, \xi_1^*, \xi_2^*, \alpha_t, \beta_t) = \boldsymbol{w}_{t+1}^{(L+1)} - \boldsymbol{w}_t^{(L+1)} + (\alpha_t - \beta_t) \boldsymbol{x}_t = 0 \\
\nabla_{\boldsymbol{w}^{(i)}} \mathcal{L}(\boldsymbol{w}_{t+1}^{(1)}, \cdots, \boldsymbol{w}_{t+1}^{(L+1)}, \xi_1^*, \xi_2^*, \alpha_t, \beta_t) = \boldsymbol{w}_{t+1}^{(i)} - \boldsymbol{w}_t^{(i)} = 0, \ \forall i \notin \{r_t, s_t, L+1\} \\
\frac{\partial \mathcal{L}(\boldsymbol{w}_{t+1}^{(1)}, \cdots, \boldsymbol{w}_{t+1}^{(L+1)}, \xi_1^*, \xi_2^*, \alpha_t, \beta_t)}{\partial \xi_1} = 2C\xi_1^* - \alpha_t = 0 \\
\frac{\partial \mathcal{L}(\boldsymbol{w}_{t+1}^{(1)}, \cdots, \boldsymbol{w}_{t+1}^{(L+1)}, \xi_1^*, \xi_2^*, \alpha_t, \beta_t)}{\partial \xi_2} = 2C\xi_2^* - \beta_t = 0 \\
\alpha_t \geq 0, \beta_t \geq 0 \\
1 - \xi_1^* - (\boldsymbol{x}_t^\top \boldsymbol{w}_{t+1}^{(r_t)} - \boldsymbol{x}_t^\top \boldsymbol{w}_{t+1}^{(L+1)}) \leq 0 \\
1 - \xi_2^* - (\boldsymbol{x}_t^\top \boldsymbol{w}_{t+1}^{(L+1)} - \boldsymbol{x}_t^\top \boldsymbol{w}_{t+1}^{(s_t)}) \leq 0 \\
\alpha_t[1 - \xi_1^* - (\boldsymbol{x}_t^\top \boldsymbol{w}_{t+1}^{(r_t)} - \boldsymbol{x}_t^\top \boldsymbol{w}_{t+1}^{(L+1)})] = 0 \\
\beta_t[1 - \xi_2^* - (\boldsymbol{x}_t^\top \boldsymbol{w}_{t+1}^{(L+1)} - \boldsymbol{x}_t^\top \boldsymbol{w}_{t+1}^{(s_t)})] = 0
\end{cases}$$

Using the first four equality constraints, we get the same update rule as Eq.(3) for PAML-II. By plugging the first six constraints of the above KKT conditions into the last four constraints and using the definition of $f_{t,1}$ and $f_{t,2}$, we can get all conditions that α_t and β_t should satisfy,

$$\begin{cases}
\alpha_t \geq 0, \ \beta_t \geq 0 \\
f_{t,1} + (\beta_t - 2\alpha_t)||\boldsymbol{x}_t||^2 \leq \frac{\alpha_t}{2C} \\
f_{t,2} + (\alpha_t - 2\beta_t)||\boldsymbol{x}_t||^2 \leq \frac{\beta_t}{2C} \\
\alpha_t(f_{t,1} + (\beta_t - 2\alpha_t)||\boldsymbol{x}_t||^2 - \frac{\alpha_t}{2C}) = 0 \\
\beta_t(f_{t,2} + (\alpha_t - 2\beta_t)||\boldsymbol{x}_t||^2 - \frac{\beta_t}{2C}) = 0
\end{cases}$$

Next consider different cases of $f_{t,1}$ and $f_{t,2}$.

1) If $f_{t,1} \leq 0$ and $f_{t,2} \leq 0$, then $\boldsymbol{W}_t$ is the optimal solution of problem (5). So $\alpha_t = 0$ and $\beta_t = 0$.

2) If $f_{t,1} > 0$ and $f_{t,2} \leq -\kappa f_{t,1}$, then we can derive that

$$\left.\begin{aligned} f_{t,1} > 0 \\ f_{t,1} + (\beta_t - 2\alpha_t)||\boldsymbol{x}_t||^2 \leq \frac{\alpha_t}{2C} \end{aligned}\right\} \Rightarrow \left.\begin{aligned} \alpha_t > \kappa\beta_t \\ \kappa > 0, \beta_t \geq 0 \end{aligned}\right\} \Rightarrow \alpha_t > 0 \Rightarrow f_{t,1} + (\beta_t - 2\alpha_t)||\boldsymbol{x}_t||^2 = \frac{\alpha_t}{2C}$$

Since $f_{t,2} \leq -\kappa f_{t,1}$, we can get that $f_{t,2} \leq (\kappa\beta_t - \alpha_t)||\boldsymbol{x}_t||^2$. Further, we can derive that

$$\left.\begin{aligned} f_{t,2} + (\alpha_t - 2\beta_t)||\boldsymbol{x}_t||^2 \leq (\kappa - 2)\beta_t||\boldsymbol{x}_t||^2 \leq 0 \\ \beta_t(f_{t,2} + (\alpha_t - 2\beta_t)||\boldsymbol{x}_t||^2 - \frac{\beta_t}{2C}) = 0 \end{aligned}\right\} \Rightarrow \beta_t = 0 \Rightarrow \alpha_t = \frac{\kappa f_{t,1}}{||\boldsymbol{x}_t||^2}$$

3) If $f_{t,1} \leq -\kappa f_{t,2}$ and $f_{t,2} > 0$, the derivation progress is similar to that in the previous case.

4) If $f_{t,1} > -\kappa f_{t,2}$ and $f_{t,2} > -\kappa f_{t,1}$, we can get

$$\left.\begin{aligned} f_{t,1} > -\kappa f_{t,2} \\ f_{t,1} + (\beta_t - 2\alpha_t)||\boldsymbol{x}_t||^2 \leq \frac{\alpha_t}{2C} \end{aligned}\right\} \Rightarrow \left.\begin{aligned} f_{t,2} > \frac{1}{\kappa}(\beta_t - \frac{\alpha_t}{\kappa})||\boldsymbol{x}_t||^2 \\ f_{t,2} + (\alpha_t - 2\beta_t)||\boldsymbol{x}_t||^2 \leq \frac{\beta_t}{2C} \end{aligned}\right\} \Rightarrow \alpha_t(1 - \frac{1}{\kappa^2})||\boldsymbol{x}_t||^2 < 0$$

Given that $0 < \kappa < \frac{1}{2}$, we can derive that $\alpha_t > 0$. On the other hand, we also have that

$$\left.\begin{aligned} f_{t,2} > -\kappa f_{t,1} \\ f_{t,2} + (\alpha_t - 2\beta_t)||\boldsymbol{x}_t||^2 \leq \frac{\beta_t}{2C} \end{aligned}\right\} \Rightarrow \left.\begin{aligned} f_{t,1} > \frac{1}{\kappa}(\alpha_t - \frac{\beta_t}{\kappa})||\boldsymbol{x}_t||^2 \\ f_{t,1} + (\beta_t - 2\alpha_t)||\boldsymbol{x}_t||^2 \leq \frac{\alpha_t}{2C} \end{aligned}\right\} \Rightarrow \beta_t(1 - \frac{1}{\kappa^2})||\boldsymbol{x}_t||^2 < 0$$

So, we can derive that $\beta_t > 0$. Using $\alpha_t > 0$ and $\beta_t > 0$, we can derive that

$$\left.\begin{aligned} f_{t,1} + (\beta_t - 2\alpha_t)||\boldsymbol{x}_t||^2 - \frac{\alpha_t}{2C} = 0 \\ f_{t,2} + (\alpha_t - 2\beta_t)||\boldsymbol{x}_t||^2 - \frac{\beta_t}{2C} = 0 \end{aligned}\right\} \Rightarrow \left.\begin{aligned} \alpha_t = \frac{f_{t,1} + \kappa f_{t,2}}{(\frac{1}{\kappa} - \kappa)||\boldsymbol{x}_t||^2} \\ \beta_t = \frac{\kappa f_{t,1} + f_{t,2}}{(\frac{1}{\kappa} - \kappa)||\boldsymbol{x}_t||^2} \end{aligned}\right\}$$

In conclusion, we get the desired solution for problem (5).

III. Detailed proof for the theorems in Section ANALYSIS

Theorem 1. *Let $(\boldsymbol{x}_1, Y_1), \cdots, (\boldsymbol{x}_T, Y_T)$ be an arbitrary sequence of input examples, where $\boldsymbol{x}_t \in \mathbb{R}^d$, $Y_t \subseteq \mathcal{Y}$ and $||\boldsymbol{x}_t|| \leq R$ for all t. Assume that there exists some $\boldsymbol{U} \in \mathbb{R}^{d\times(L+1)}$ such that $\ell^*_{t,1} = 0$ and $\ell^*_{t,2} = 0$ for all t. Then the cumulative squared loss suffered by PAML on this sequence of examples is bounded by*

$$\sum_{t=1}^{T}(\ell_{t,1} + \ell_{t,2})^2 \leq 2R^2||\boldsymbol{U}||_F^2$$

Proof. Since $\ell^*_{t,1} = \ell^*_{t,2} = 0$ for all t, Lemma 1 implies that

$$\sum_{t=1}^{T} \left[\alpha_t f_{t,1} + \beta_t f_{t,2} - (\alpha_t^2 + \beta_t^2 - \alpha_t\beta_t)||\boldsymbol{x}_t||^2\right] \leq \frac{1}{2}||\boldsymbol{U}||_F^2 \tag{1}$$

For simplicity, let

$$Z_t = \alpha_t f_{t,1} + \beta_t f_{t,2} - (\alpha_t^2 + \beta_t^2 - \alpha_t\beta_t)||\boldsymbol{x}_t||^2.$$

We will derive the lower bound for Z_t. First focus on the online rounds $\{t : Y_t \neq \varnothing \text{ and } \bar{Y}_t \neq \varnothing\}$.

1) If $f_{t,1} > 0$ and $f_{t,2} \leq -\frac{1}{2}f_{t,1}$, then $\alpha_t = \frac{f_{t,1}}{2||\boldsymbol{x}_t||^2}$, $\beta_t = 0$, $\ell_{t,1} = f_{t,1}$ and $\ell_{t,2} = 0$. Using these knowledge, we can derive that

$$Z_t = \frac{\ell_{t,1}^2}{4||\boldsymbol{x}_t||^2} = \frac{(\ell_{t,1} + \ell_{t,2})^2}{4||\boldsymbol{x}_t||^2}$$

2) If $f_{t,1} \leq -\frac{1}{2}f_{t,2}$ and $f_{t,2} > 0$, then $\alpha_t = 0$, $\beta_t = \frac{f_{t,2}}{2||\boldsymbol{x}_t||^2}$, $\ell_{t,2} = f_{t,2}$ and $\ell_{t,1} = 0$. Further, we can derive that

$$Z_t = \frac{\ell_{t,2}^2}{4||\boldsymbol{x}_t||^2} = \frac{(\ell_{t,1} + \ell_{t,2})^2}{4||\boldsymbol{x}_t||^2}$$

3) If $f_{t,1} > -\frac{1}{2}f_{t,2}$ and $f_{t,2} > -\frac{1}{2}f_{t,1}$, then $\alpha_t = \frac{2f_{t,1}+f_{t,2}}{3||\boldsymbol{x}_t||^2}$ and $\beta_t = \frac{f_{t,1}+2f_{t,2}}{3||\boldsymbol{x}_t||^2}$. Further, we get

$$Z_t = \frac{f_{t,1}^2 + f_{t,2}^2 + f_{t,1}f_{t,2}}{3||\boldsymbol{x}_t||^2} \geq \frac{(\ell_{t,1} + \ell_{t,2})^2}{4||\boldsymbol{x}_t||^2}$$

We now prove why the last inequality holds. The area defined by $f_{t,1} > -\frac{1}{2}f_{t,2}$ and $f_{t,2} > -\frac{1}{2}f_{t,1}$ can be further divided into three small areas:

a) If $f_{t,1} > -\frac{1}{2}f_{t,2}$ and $f_{t,1} \leq 0$, then $\ell_{t,1} = 0$ and $\ell_{t,2} = f_{t,2}$. Therefore,

$$Z_t = \frac{(f_{t,1} + \frac{1}{2}f_{t,2})^2 + \frac{3}{4}f_{t,2}^2}{3||\boldsymbol{x}_t||^2} > \frac{\ell_{t,2}^2}{4||\boldsymbol{x}_t||^2} = \frac{(\ell_{t,1} + \ell_{t,2})^2}{4||\boldsymbol{x}_t||^2}$$

b) If $f_{t,2} > -\frac{1}{2}f_{t,1}$ and $f_{t,2} \leq 0$, then $\ell_{t,1} = f_{t,1}$ and $\ell_{t,2} = 0$. Therefore,

$$Z_t = \frac{(f_{t,2} + \frac{1}{2}f_{t,1})^2 + \frac{3}{4}f_{t,1}^2}{3||\boldsymbol{x}_t||^2} > \frac{\ell_{t,1}^2}{4||\boldsymbol{x}_t||^2} = \frac{(\ell_{t,1} + \ell_{t,2})^2}{4||\boldsymbol{x}_t||^2}$$

c) If $f_{t,1} > 0$ and $f_{t,2} > 0$, then $\ell_{t,1} = f_{t,1}$ and $\ell_{t,2} = f_{t,2}$. Therefore,

$$Z_t = \frac{\frac{3}{4}(f_{t,1}+f_{t,2})^2 + \frac{1}{4}(f_{t,1}-f_{t,2})^2}{3||\boldsymbol{x}_t||^2} \geq \frac{(\ell_{t,1} + \ell_{t,2})^2}{4||\boldsymbol{x}_t||^2}$$

On those online rounds where $Y_t = \varnothing$ or $\bar{Y}_t = \varnothing$, it is easy to check that $Z_t = \frac{(\ell_{t,1}+\ell_{t,2})^2}{4||\boldsymbol{x}_t||^2}$. Therefore, the following inequality holds for all online rounds,

$$Z_t \geq \frac{(\ell_{t,1} + \ell_{t,2})^2}{4||\boldsymbol{x}_t||^2} \geq \frac{(\ell_{t,1} + \ell_{t,2})^2}{4R^2}$$

Summing the inequality over $t=1$ to T and combining with Eq.(1) gives the desired bound. $\square$

Theorem 2. *Let $(\boldsymbol{x}_1, Y_1), \cdots, (\boldsymbol{x}_T, Y_T)$ be an arbitrary sequence of input examples, where $\boldsymbol{x}_t \in \mathbb{R}^d$, $Y_t \subseteq \mathcal{Y}$ and $||\boldsymbol{x}_t|| = 1$ for all t. Then for any $\boldsymbol{U} \in \mathbb{R}^{d\times(L+1)}$, the cumulative squared loss of PAML on this sequence of examples is bounded by*

$$\sum_{t=1}^{T}(\ell_{t,1}+\ell_{t,2})^2 \le \left(\frac{8}{3}\sqrt{\sum_{t=1}^{T}(\ell^*_{t,1}+\ell^*_{t,2})^2}+\sqrt{2}||\boldsymbol{U}||_F\right)^2$$

Proof. Since $||\boldsymbol{x}_t|| = 1$ for all t, Lemma 1 implies that

$$\sum_{t=1}^{T}\left[\alpha_t(f_{t,1}-\ell^*_{t,1})+\beta_t(f_{t,2}-\ell^*_{t,2})-(\alpha_t^2+\beta_t^2-\alpha_t\beta_t)\right] \le \frac{1}{2}||\boldsymbol{U}||_F^2 \tag{2}$$

For simplicity, let

$$Z_t = \alpha_t(f_{t,1}-\ell^*_{t,1})+\beta_t(f_{t,2}-\ell^*_{t,2})-(\alpha_t^2+\beta_t^2-\alpha_t\beta_t).$$

We derive the lower bound for Z_t. First focus on the online rounds $\{t : Y_t \neq \varnothing \text{ and } \bar{Y}_t \neq \varnothing\}$.

1) If $f_{t,1} > 0$ and $f_{t,2} \le -\frac{1}{2}f_{t,1}$, then $\alpha_t = \frac{f_{t,1}}{2}$, $\beta_t = 0$, $\ell_{t,1} = f_{t,1}$ and $\ell_{t,2} = 0$. Therefore, we can derive that

$$Z_t = \frac{\ell_{t,1}^2}{4} - \frac{\ell_{t,1}\ell^*_{t,1}}{2} \ge \frac{(\ell_{t,1}+\ell_{t,2})^2}{4} - \frac{2}{3}(\ell_{t,1}+\ell_{t,2})(\ell^*_{t,1}+\ell^*_{t,2})$$

2) If $f_{t,1} \le -\frac{1}{2}f_{t,2}$ and $f_{t,2} > 0$, then $\alpha_t = 0$, $\beta_t = \frac{f_{t,2}}{2}$, $\ell_{t,2} = f_{t,2}$ and $\ell_{t,1} = 0$. Further, we can derive that

$$Z_t = \frac{\ell_{t,2}^2}{4} - \frac{\ell_{t,2}\ell^*_{t,2}}{2} \ge \frac{(\ell_{t,1}+\ell_{t,2})^2}{4} - \frac{2}{3}(\ell_{t,1}+\ell_{t,2})(\ell^*_{t,1}+\ell^*_{t,2})$$

3) If $f_{t,1} > -\frac{1}{2}f_{t,2}$ and $f_{t,2} > -\frac{1}{2}f_{t,1}$, then $\alpha_t = \frac{2f_{t,1}+f_{t,2}}{3}$ and $\beta_t = \frac{f_{t,1}+2f_{t,2}}{3}$. Further, we get

$$\begin{aligned}Z_t &= \frac{f_{t,1}^2+f_{t,2}^2+f_{t,1}f_{t,2}}{3} - \frac{2f_{t,1}+f_{t,2}}{3}\ell^*_{t,1} - \frac{f_{t,1}+2f_{t,2}}{3}\ell^*_{t,2} \\ &\ge \frac{(\ell_{t,1}+\ell_{t,2})^2}{4} - \frac{2}{3}(\ell_{t,1}+\ell_{t,2})(\ell^*_{t,1}+\ell^*_{t,2})\end{aligned}$$

In summary, for all online rounds including the ones $\{t : Y_t = \varnothing \text{ or } \bar{Y}_t = \varnothing\}$, the following inequality holds,

$$Z_t \ge \frac{(\ell_{t,1}+\ell_{t,2})^2}{4} - \frac{2}{3}(\ell_{t,1}+\ell_{t,2})(\ell^*_{t,1}+\ell^*_{t,2})$$

Summing the above inequality over $t = 1$ to T and combining with Eq.(2) gives us that,

$$\sum_{t=1}^{T}(\ell_{t,1}+\ell_{t,2})^2 \leq \frac{8}{3}\sum_{t=1}^{T}(\ell_{t,1}+\ell_{t,2})(\ell_{t,1}^{*}+\ell_{t,2}^{*}) + 2||\boldsymbol{U}||_F^2$$

Using Cauchy-Schwartz inequality, we get that

$$\sum_{t=1}^{T}(\ell_{t,1}+\ell_{t,2})^2 \leq \frac{8}{3}\sqrt{\sum_{t=1}^{T}(\ell_{t,1}+\ell_{t,2})^2}\sqrt{\sum_{t=1}^{T}(\ell_{t,1}^{*}+\ell_{t,2}^{*})^2} + 2||\boldsymbol{U}||_F^2$$

Let $P_T = \sqrt{\sum_{t=1}^{T}(\ell_{t,1}+\ell_{t,2})^2}$, $Q_T = \sqrt{\sum_{t=1}^{T}(\ell_{t,1}^{*}+\ell_{t,2}^{*})^2}$. Then we can get

$$P_T^2 \leq \frac{8}{3}P_T Q_T + 2||\boldsymbol{U}||_F^2.$$

Solving this inequality, we get

$$P_T \leq \frac{4}{3}Q_T + \sqrt{2||\boldsymbol{U}||_F^2 + (\frac{4}{3}Q_T)^2} \leq \frac{8}{3}Q_T + \sqrt{2}||\boldsymbol{U}||_F$$

where the last inequality is owing to the fact that $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$.

Finally, taking the square on both sides of the above inequality and plugging into the definition of P_T and Q_T gives the desired bound. □

Theorem 3. *Let $(\boldsymbol{x}_1, Y_1), \cdots, (\boldsymbol{x}_T, Y_T)$ be an arbitrary sequence of input examples, where $\boldsymbol{x}_t \in \mathbb{R}^d$, $Y_t \subseteq \mathcal{Y}$, and $||\boldsymbol{x}_t|| \leq R$ for all t. Then for any $\boldsymbol{U} \in \mathbb{R}^{d\times(L+1)}$, the number of wrong predictions made by PAML-I on this sequence of examples is bounded from above by*

$$\max\{2R^2, 1/C\}\left(||\boldsymbol{U}||_F^2 + 2C\sum_{t=1}^{T}(\ell_{t,1}^{*}+\ell_{t,2}^{*})\right)$$

where C is the aggressiveness parameter provided to PAML-I.

Proof. First we analyze different types of mistakes that PAML-I may made. If PAML-I makes a wrong prediction at round t, namely, $Y_t \neq \hat{Y}_t$, three types of mistakes may occur:

- "Type-I" mistakes: there exists some irrelevant labels that are wrongly predicted as relevant. For making such mistakes, two cases may occur. In the first case that $Y_t \neq \varnothing$ and $\bar{Y}_t \neq \varnothing$, it follows that $\boldsymbol{x}_t^\top \boldsymbol{w}_t^{(r_t)} > \boldsymbol{x}_t^\top \boldsymbol{w}_t^{(L+1)}$ and $\boldsymbol{x}_t^\top \boldsymbol{w}_t^{(L+1)} < \boldsymbol{x}_t^\top \boldsymbol{w}_t^{(s_t)}$, which implies that $0 \leq \ell_{t,1} < 1$ and $\ell_{t,2} > 1$. In the second case that $Y_t = \varnothing$, it follows that $\boldsymbol{x}_t^\top \boldsymbol{w}_t^{(L+1)} < \boldsymbol{x}_t^\top \boldsymbol{w}_t^{(s_t)}$, which implies that $\ell_{t,2} > 1$.
- "Type-II" mistakes: there exists some relevant labels that are wrongly predicted as irrelevant. Similarly, if $Y_t \neq \varnothing$ and $\bar{Y}_t \neq \varnothing$, it follows that $\boldsymbol{x}_t^\top \boldsymbol{w}_t^{(r_t)} \leq \boldsymbol{x}_t^\top \boldsymbol{w}_t^{(L+1)}$ and $\boldsymbol{x}_t^\top \boldsymbol{w}_t^{(L+1)} \geq$

$\boldsymbol{x}_t^\top \boldsymbol{w}_t^{(s_t)}$, which leads to the results that $\ell_{t,1} \geq 1$ and $0 \leq \ell_{t,2} \leq 1$. If $\bar{Y}_t = \varnothing$, then $\boldsymbol{x}_t^\top \boldsymbol{w}_t^{(r_t)} \leq \boldsymbol{x}_t^\top \boldsymbol{w}_t^{(L+1)}$, which implies that $\ell_{t,1} \geq 1$.

- "Type-III" mistakes: there exists some relevant labels that are wrongly predicted as irrelevant, and also exists irrelevant labels that are wrongly predicted as relevant. In this case, it must have that $Y_t \neq \varnothing$ and $\bar{Y}_t \neq \varnothing$. So it follows that $\boldsymbol{x}_t^\top \boldsymbol{w}_t^{(r_t)} \leq \boldsymbol{x}_t^\top \boldsymbol{w}_t^{(L+1)}$ and $\boldsymbol{x}_t^\top \boldsymbol{w}_t^{(L+1)} < \boldsymbol{x}_t^\top \boldsymbol{w}_t^{(s_t)}$, which further implies that $\ell_{t,1} \geq 1$ and $\ell_{t,2} > 1$.

Next we start to bound the number of mistakes PAML-I made on the entire sequence. According to the definition of α_t and β_t for PAML-I, we have $\alpha_t \leq C$ and $\beta_t \leq C$ for all t. Thus, Lemma 1 implies that

$$\sum_{t=1}^{T} \left[\alpha_t f_{t,1} + \beta_t f_{t,2} - (\alpha_t^2 + \beta_t^2 - \alpha_t \beta_t)||\boldsymbol{x}_t||^2\right] \leq \frac{1}{2}||\boldsymbol{U}||_F^2 + C\sum_{t=1}^{T}(\ell_{t,1}^* + \ell_{t,2}^*) \tag{3}$$

Similarly, for simplicity, let

$$Z_t = \alpha_t f_{t,1} + \beta_t f_{t,2} - (\alpha_t^2 + \beta_t^2 - \alpha_t \beta_t)||\boldsymbol{x}_t||^2.$$

The lower bound for Z_t will be derived. First focus on the online rounds where $Y_t \neq \varnothing$ and $\bar{Y}_t \neq \varnothing$.

1) If $(f_{t,1}, f_{t,2}) \in$ Area ② ∪ ③ in Fig. 1, then $\alpha_t = 0$, $\beta_t = \min\{\frac{f_{t,2}}{2||\boldsymbol{x}_t||^2}, C\}$, $\ell_{t,1} = 0$ and $\ell_{t,2} = f_{t,2}$. Using these facts, we can get

$$Z_t = \beta_t(f_{t,2} - \beta_t||\boldsymbol{x}_t||^2) \geq \beta_t(f_{t,2} - \frac{f_{t,2}}{2}) = \frac{1}{2}\beta_t \ell_{t,2}$$

Further, if PAML-I made wrong predictions in this case, only "Type-I" mistakes can be made, which implies that $\ell_{t,2} > 1$. Given that $||\boldsymbol{x}_t|| \leq R$ for all t, we can get

$$\frac{1}{2}\beta_t \ell_{t,2} \geq \frac{1}{2}\min\{\frac{1}{2R^2}, C\}$$

2) If $(f_{t,1}, f_{t,2}) \in$ Area ⑦ ∪ ⑧ in Fig. 1, then $\beta_t = 0$, $\alpha_t = \min\{\frac{f_{t,1}}{2||\boldsymbol{x}_t||^2}, C\}$, $\ell_{t,2} = 0$ and $\ell_{t,1} = f_{t,1}$. Further, we can get

$$Z_t = \alpha_t(f_{t,1} - \alpha_t||\boldsymbol{x}_t||^2) \geq \alpha_t(f_{t,1} - \frac{f_{t,1}}{2}) = \frac{1}{2}\alpha_t \ell_{t,1}$$

If PAML-I made mistakes in this case, only "Type-II" mistakes can be made, which implies that $\ell_{t,1} \geq 1$. So we can get

$$\frac{1}{2}\alpha_t \ell_{t,1} \geq \frac{1}{2}\min\{\frac{1}{2R^2}, C\}$$

3) If $(f_{t,1}, f_{t,2}) \in$ Area ④ in Fig. 1, then $\alpha_t = \frac{1}{2}(C + \frac{f_{t,1}}{||\boldsymbol{x}_t||^2})$, $\beta_t = C$ and $f_{t,2} = \ell_{t,2}$. By plugging into the definition of α_t and β_t, we can get

$$Z_t = \frac{C}{2} f_{t,1} + C f_{t,2} + \frac{f_{t,1}^2}{4||\boldsymbol{x}_t||^2} - \frac{3C^2}{4}||\boldsymbol{x}_t||^2 \tag{4}$$

Area ④ can be divided into two small areas.

a) When $0 < f_{t,1} \leq C||\boldsymbol{x}_t||^2$ and $\frac{1}{2} f_{t,1} + f_{t,2} \geq \frac{3C}{2}||\boldsymbol{x}_t||^2$, we have that $\ell_{t,1} = f_{t,1}$ and $\ell_{t,2} = f_{t,2}$. Using these facts, we can get

$$\begin{aligned} \text{Eq. (4)} &\geq \frac{C}{4} f_{t,1} + \frac{C}{2} f_{t,2} + \frac{f_{t,1}^2}{4||\boldsymbol{x}_t||^2} \\ &= \frac{\alpha_t}{2} f_{t,1} + \frac{\beta_t}{2} f_{t,2} > \frac{C}{2} \ell_{t,2} \end{aligned}$$

Further, both "Type-I" and "Type-III" mistakes may be made in this case. For either type of mistakes, it holds that $\ell_{t,2} > 1$, which implies that

$$\frac{C}{2}\ell_{t,2} > \frac{C}{2} \geq \frac{1}{2}\min\{\frac{1}{2R^2}, C\}$$

b) When $-C||\boldsymbol{x}_t||^2 < f_{t,1} \leq 0$ and $\frac{1}{2} f_{t,1} + f_{t,2} \geq \frac{3C}{2}||\boldsymbol{x}_t||^2$, only "Type-I" mistakes may be made in this case. We will consider the minimum of Eq. (4) in three different situations.

i) If $\frac{3C}{2}||\boldsymbol{x}_t||^2 \geq 1$, then $(f_{t,1}, f_{t,2})$ may locate in any position in Area ④ when PAML-I made mistakes. Since $f_{t,2}$ has no upper bound in Area ④, Eq. (4) achieves the minimum when $\frac{1}{2} f_{t,1} + f_{t,2} = \frac{3C}{2}||\boldsymbol{x}_t||^2$. It is easy to derive that the minimum point is at $(f_{t,1}, f_{t,2}) = (0, \frac{3C}{2}||\boldsymbol{x}_t||^2)$. Plugging into the point gives us that

$$\text{Eq. (4)} \geq \frac{3C^2}{4}||\boldsymbol{x}_t||^2 \geq \frac{C}{2}$$

ii) If $\frac{3C}{2}||\boldsymbol{x}_t||^2 < 1 < 2C||\boldsymbol{x}_t||^2$, we can derive that when $f_{t,2} > 1$, the minimum value point of Eq. (4) is at $(f_{t,1}, f_{t,2}) = (3C||\boldsymbol{x}_t||^2 - 2, 1)$. Plugging into the point, we get

$$\begin{aligned} \text{Eq. (4)} &> 3C^2||\boldsymbol{x}_t||^2 + \frac{1}{||\boldsymbol{x}_t||^2} - 3C \\ &= (\sqrt{3}C||\boldsymbol{x}_t|| - \frac{\sqrt{3}}{2||\boldsymbol{x}_t||})^2 + \frac{1}{4||\boldsymbol{x}_t||^2} \geq \frac{1}{4R^2} \end{aligned}$$

iii) If $1 \geq 2C||\boldsymbol{x}_t||^2$, we can derive that when $f_{t,2} > 1$, the minimum value point of Eq. (4) is at $(f_{t,1}, f_{t,2}) = (-C||\boldsymbol{x}_t||^2, 1)$. Plugging into the point gives us that

$$\text{Eq. (4)} > C - C^2||\boldsymbol{x}_t||^2 \geq C - \frac{C}{2} = \frac{C}{2}$$

In summary, if $(f_{t,1}, f_{t,2}) \in$ Area ④ and PAML-I makes prediction mistakes in such case, then we can get that

$$Z_t \geq \frac{1}{2}\min\{\frac{1}{2R^2}, C\}$$

4) If $(f_{t,1}, f_{t,2}) \in$ Area ⑥ in Fig. 1, then $\beta_t = \frac{1}{2}(C + \frac{f_{t,2}}{||\boldsymbol{x}_t||^2})$ and $\alpha_t = C$. The derivation process is similar to that in the previous case.
5) If $(f_{t,1}, f_{t,2}) \in$ Area ⑤ in Fig. 1, then $\alpha_t = \beta_t = C$, $\ell_{t,1} = f_{t,1}$ and $\ell_{t,2} = f_{t,2}$. So we have

$$\begin{aligned} Z_t &= C(f_{t,1} + f_{t,2} - C||\boldsymbol{x}_t||^2) \\ &> \frac{C}{2}(f_{t,1} + f_{t,2}) = \frac{C}{2}(\ell_{t,1} + \ell_{t,2}) \end{aligned}$$

where the inequality is due to that $f_{t,1} + f_{t,2} > 2C||\boldsymbol{x}_t||^2$ in Area ⑤. Whichever type of mistakes PAML-I may made in this case, it always holds that

$$\frac{C}{2}(\ell_{t,1} + \ell_{t,2}) > \frac{C}{2} \geq \frac{1}{2}\min\{\frac{1}{2R^2}, C\}$$

6) If $(f_{t,1}, f_{t,2}) \in$ Area ⑨ in Fig. 1, then $\alpha_t = \frac{2f_{t,1}+f_{t,2}}{3||\boldsymbol{x}_t||^2}$ and $\beta_t = \frac{f_{t,1}+2f_{t,2}}{3||\boldsymbol{x}_t||^2}$. We can get

$$Z_t \geq \frac{(\ell_{t,1} + \ell_{t,2})^2}{4||\boldsymbol{x}_t||^2}$$

where the proof for the inequality is similar to that in Theorem 1. Similarly, whichever type of mistakes PAML-I may made in this case, it holds that $(\ell_{t,1} + \ell_{t,2}) > 1$. Given that $||\boldsymbol{x}_t|| \leq R$ for all t, we can get

$$\frac{(\ell_{t,1} + \ell_{t,2})^2}{4||\boldsymbol{x}_t||^2} > \frac{1}{4R^2} \geq \frac{1}{2}\min\{\frac{1}{2R^2}, C\}$$

On the online rounds where $Y_t = \varnothing$ or $\bar{Y}_t = \varnothing$, if PAML-I makes wrong predictions, we can also get that $Z_t \geq \frac{1}{2}\min\{\frac{1}{2R^2}, C\}$. Thus, the following inequality holds for all online rounds where prediction mistakes are made,

$$Z_t \geq \frac{1}{2}\min\{\frac{1}{2R^2}, C\}$$

Let M denote the number of wrong predictions PAML-I made on the entire sequence. Given that Z_t is always non-negative, it holds that

$$\sum_{t=1}^{T} Z_t \geq \frac{1}{2}\min\{\frac{1}{2R^2}, C\}M$$

Combining the above inequality with Eq. (3), we get

$$\min\{\frac{1}{2R^2}, C\}M \leq ||\boldsymbol{U}||_F^2 + 2C\sum_{t=1}^{T}(\ell_{t,1}^* + \ell_{t,2}^*)$$

Our theorem follows from multiplying both sides of the above inequality by $\max\{2R^2, 1/C\}$.

□

Theorem 4. *Let $(\boldsymbol{x}_1, Y_1), \cdots, (\boldsymbol{x}_T, Y_T)$ be an arbitrary sequence of input examples, where $\boldsymbol{x}_t \in \mathbb{R}^d$, $Y_t \subseteq \mathcal{Y}$ and $||\boldsymbol{x}_t|| \leq R$ for all t. Then for any $\boldsymbol{U} \in \mathbb{R}^{d\times(L+1)}$, the cumulative squared loss of PAML-II on this sequence of examples is bounded from above by*

$$\sum_{t=1}^{T}(\ell_{t,1}^2 + \ell_{t,2}^2) \leq \left(4R^2 + \frac{1}{C}\right)\left(\frac{1}{2}||\boldsymbol{U}||_F^2 + C\sum_{t=1}^{T}\big[(\ell_{t,1}^*)^2 + (\ell_{t,2}^*)^2\big]\right)$$

Proof. Again, for simplicity, let

$$Z_t = \alpha_t(f_{t,1} - \ell_{t,1}^*) + \beta_t(f_{t,2} - \ell_{t,2}^*) - (\alpha_t^2 + \beta_t^2 - \alpha_t\beta_t)||\boldsymbol{x}_t||^2.$$

Our analyses will mainly focus on the online rounds where $Y_t \neq \varnothing$ and $\bar{Y}_t \neq \varnothing$. Next we start to bound from below the left hand side of the inequality in Lemma 1. Different cases of $f_{t,1}$ and $f_{t,2}$ will be taken into account.

1) If $f_{t,1} > 0$ and $f_{t,2} \leq -\kappa f_{t,1}$, then $\alpha_t = \frac{\kappa f_{t,1}}{||\boldsymbol{x}_t||^2}$, $\beta_t = 0$, $\ell_{t,1} = f_{t,1}$ and $\ell_{t,2} = 0$. Using these facts, we get that

$$\begin{aligned}
Z_t &= \alpha_t(f_{t,1} - \ell_{t,1}^*) - \alpha_t^2||\boldsymbol{x}_t||^2 \\
&\geq \alpha_t(f_{t,1} - \ell_{t,1}^*) - \alpha_t^2||\boldsymbol{x}_t||^2 - \frac{1}{2}(\frac{1}{\sqrt{2C}}\alpha_t - \sqrt{2C}\ell_{t,1}^*)^2 \\
&= \alpha_t f_{t,1} - (||\boldsymbol{x}_t||^2 + \frac{1}{4C})\alpha_t^2 - C(\ell_{t,1}^*)^2 \\
&=_1 \frac{1}{2}\alpha_t f_{t,1} - C(\ell_{t,1}^*)^2 = \frac{\kappa}{2||\boldsymbol{x}_t||^2}\ell_{t,1}^2 - C(\ell_{t,1}^*)^2 \\
&\geq \frac{\kappa}{2||\boldsymbol{x}_t||^2}(\ell_{t,1}^2 + \ell_{t,2}^2) - C\big[(\ell_{t,1}^*)^2 + (\ell_{t,2}^*)^2\big]
\end{aligned}$$

where $=_1$ is owing to the fact that $||\boldsymbol{x}_t||^2 + \frac{1}{4C} = \frac{||\boldsymbol{x}_t||^2}{2\kappa}$ and $\alpha_t||\boldsymbol{x}_t||^2 = \kappa f_{t,1}$.

2) If $f_{t,1} \leq -\kappa f_{t,2}$ and $f_{t,2} > 0$, then $\alpha_t = 0$, $\beta_t = \frac{\kappa f_{t,2}}{||\boldsymbol{x}_t||^2}$, $\ell_{t,1} = 0$ and $\ell_{t,2} = f_{t,2}$. Following the similar derivation to that of the first case, we can get that

$$Z_t \geq \frac{\kappa}{2||\boldsymbol{x}_t||^2}(\ell_{t,1}^2 + \ell_{t,2}^2) - C\big[(\ell_{t,1}^*)^2 + (\ell_{t,2}^*)^2\big]$$

3) If $f_{t,1} > -\kappa f_{t,2}$ and $f_{t,2} > -\kappa f_{t,1}$, then $\alpha_t = \frac{f_{t,1}+\kappa f_{t,2}}{(\frac{1}{\kappa}-\kappa)||\boldsymbol{x}_t||^2}$ and $\beta_t = \frac{\kappa f_{t,1}+f_{t,2}}{(\frac{1}{\kappa}-\kappa)||\boldsymbol{x}_t||^2}$. We get that

$$\begin{aligned}
Z_t &\geq \alpha_t(f_{t,1}-\ell^*_{t,1}) - \frac{1}{2}(\frac{1}{\sqrt{2C}}\alpha_t - \sqrt{2C}\ell^*_{t,1})^2 + \beta_t(f_{t,2}-\ell^*_{t,2}) - \frac{1}{2}(\frac{1}{\sqrt{2C}}\beta_t - \sqrt{2C}\ell^*_{t,2})^2 \\
&\quad - \alpha_t^2||\boldsymbol{x}_t||^2 - \beta_t^2||\boldsymbol{x}_t||^2 + \alpha_t\beta_t||\boldsymbol{x}_t||^2 \\
&= \alpha_t f_{t,1} + \beta_t f_{t,2} - (||\boldsymbol{x}_t||^2 + \frac{1}{4C})(\alpha_t^2+\beta_t^2) - C\big[(\ell^*_{t,1})^2 + (\ell^*_{t,2})^2\big] + \alpha_t\beta_t||\boldsymbol{x}_t||^2 \\
&= \frac{\kappa}{2(1-\kappa^2)||\boldsymbol{x}_t||^2}(f_{t,1}^2 + f_{t,2}^2 + 2\kappa f_{t,1}f_{t,2}) - C\big[(\ell^*_{t,1})^2 + (\ell^*_{t,2})^2\big] \\
&> \frac{\kappa}{2||\boldsymbol{x}_t||^2}(\ell_{t,1}^2 + \ell_{t,2}^2) - C\big[(\ell^*_{t,1})^2 + (\ell^*_{t,2})^2\big]
\end{aligned}$$

We now prove why the last inequality holds. The area defined by $f_{t,1} > -\kappa f_{t,2}$ and $f_{t,2} > -\kappa f_{t,1}$ consists of three smaller areas:

a) when $f_{t,1} > -\kappa f_{t,2}$ and $f_{t,1} \leq 0$, we have $\ell_{t,1} = 0$ and $\ell_{t,2} = f_{t,2}$. Then,

$$f_{t,1}^2 + f_{t,2}^2 + 2\kappa f_{t,1}f_{t,2} = (f_{t,1}+\kappa f_{t,2})^2 + (1-\kappa^2)f_{t,2}^2 > (1-\kappa^2)(\ell_{t,1}^2 + \ell_{t,2}^2)$$

b) when $f_{t,2} > -\kappa f_{t,1}$ and $f_{t,2} \leq 0$, we have $\ell_{t,2} = 0$ and $\ell_{t,1} = f_{t,1}$. Then,

$$f_{t,1}^2 + f_{t,2}^2 + 2\kappa f_{t,1}f_{t,2} = (\kappa f_{t,1}+f_{t,2})^2 + (1-\kappa^2)f_{t,1}^2 > (1-\kappa^2)(\ell_{t,1}^2 + \ell_{t,2}^2)$$

c) when $f_{t,1} > 0$ and $f_{t,2} > 0$, we have $\ell_{t,1} = f_{t,1}$ and $\ell_{t,2} = f_{t,2}$. Then,

$$f_{t,1}^2 + f_{t,2}^2 + 2\kappa f_{t,1}f_{t,2} > (1-\kappa^2)(\ell_{t,1}^2 + \ell_{t,2}^2)$$

Given that $||\boldsymbol{x}_t|| \leq R$ for all t, the following inequality holds for all online rounds,

$$Z_t \geq \frac{\kappa}{2||\boldsymbol{x}_t||^2}(\ell_{t,1}^2 + \ell_{t,2}^2) - C\big[(\ell^*_{t,1})^2 + (\ell^*_{t,2})^2\big] \geq \frac{1}{4R^2 + \frac{1}{C}}(\ell_{t,1}^2 + \ell_{t,2}^2) - C\big[(\ell^*_{t,1})^2 + (\ell^*_{t,2})^2\big]$$

Summing the above inequality over $t = 1$ to T and combining with Lemma 1, we get the desired bound. □

IV. Parameter settings in our comparative experiments

By performing 10×10 cross validation on each training dataset, we find good parameter values for each algorithm on each dataset, as shown in Table I.

TABLE I

PARAMETER SETTINGS OF EACH ALGORITHM ON EACH DATASET

Dataset	OSML-ELM	ELM-OMLL		PA-I-BR		PA-II-BR		PAML	PAML-I		PAML-II	
	HNs	HNs	ρ	C	δ^2	C	δ^2	δ^2	C	δ^2	C	δ^2
Rcv1v2(industries)	1000	1000	$2^{-6.5}$	2^2	–	2^9	–	–	2^{-7}	–	2^{-4}	–
Rcv1v2(regions)	1000	1000	$2^{-6.5}$	2^2	–	2^9	–	–	2^{-6}	–	2^{-4}	–
Rcv1v2(topics)	1000	1000	$2^{-6.5}$	2	–	2	–	–	$2^{-7.5}$	–	$2^{-7.5}$	–
Bibtex	1000	1000	2^{-6}	$2^{-2.5}$	–	$2^{-1.625}$	–	–	$2^{-10.5}$	–	2^{-14}	–
Birds	900	700	2^{-5}	$2^{-2.5}$	$2^{-3.75}$	2^{-6}	$2^{-3.75}$	$2^{-4.25}$	2^{-8}	$2^{-3.5}$	2^{-8}	$2^{-3.5}$
Scene	1000	400	2^{-3}	2	$2^{1.5}$	$2^{5.5}$	$2^{1.5}$	$2^{1.25}$	$2^{-2.5}$	$2^{1.25}$	2^{-3}	$2^{1.25}$
Emotions	400	500	$2^{-1.5}$	$2^{0.5}$	2^{-2}	2^4	$2^{-2.25}$	$2^{-2.25}$	2^{-3}	2^{-2}	2^{-6}	$2^{-2.25}$
Yeast	100	1000	2^{-5}	2	$2^{-2.25}$	1	2^{-2}	$2^{-2.25}$	2^{-4}	$2^{-2.375}$	2^{-7}	$2^{-2.5}$
Mediamill	1000	1000	2^{-7}	2	2^{-7}	2^7	2^{-7}	2^{-7}	2^{-6}	$2^{-7.25}$	$2^{-8.5}$	2^{-7}

[1] "HNs" : the number of hidden layer neurons; "ρ": the regularization factor for ELM-OMLL.