Online Boosting for Anytime Transfer and Multitask Learning Supplementary Materials

Boyu Wang and **Joelle Pineau**
School of Computer Science
McGill University, Montreal, Canada
boyu.wang@mail.mcgill.ca, jpineau@cs.mcgill.ca

Proof of Theorem 1

Let D^b_m be the weight distribution of batch TrAdaBoost algorithm, D^o_m be the weight distribution of OTB, which can be viewed as the normalized version of Poisson parameter λ in Algorithm 2. Lemma 1 shows the convergence property of D^o_m.

Lemma 1. *As $N_S \to \infty$ and $N_T \to \infty$, $D^o_1 \xrightarrow{P} D^b_1$.*

Define h^b_m as the mth base learner of batch TrAdaBoost, and h^o_m as the analogous base learner of OTB. Lemma 2 states that if the weight vector D^o_m converges to D^b_m, the base learner h^o_m also converges to h^b_m.

Lemma 2. *If $D^o_m \xrightarrow{P} D^b_m$, and the base learners are naive Bayes classifiers, then $h^o_m \xrightarrow{P} h^b_m$.*

Let $\epsilon^b_{S,m} = \sum_{x_n \in S_S} D^b_m(n) I(h^b_m(x_n) \neq y_n)$, $\epsilon^b_{T,m} = \sum_{x_n \in S_T} D^b_m(n) I(h^b_m(x_n) \neq y_n)$, $D^b_{T,m} = \sum_{x_n \in S_T} D^b_m(n)$; and $\epsilon^o_{S,m}$, $\epsilon^o_{T,m}$, $D^o_{T,m}$ be their online approximation defined in line 22-24 of Algorithm 2. Lemma 3 states that $\epsilon^o_{S,m}$, $\epsilon^o_{T,m}$, $D^o_{T,m}$ also converge to their batch counterparts given h^o_m converging to h^b_m.

Lemma 3. *If $D^o_m \xrightarrow{P} D^b_m$, $h^o_m \xrightarrow{P} h^b_m$, and the base learners are naive Bayes classifiers, then $\epsilon^o_{S,m} \xrightarrow{P} \epsilon^b_{S,m}$, $\epsilon^o_{T,m} \xrightarrow{P} \epsilon^b_{T,m}$, and $D^o_{T,m} \xrightarrow{P} D^b_{T,m}$.*

To prove the convergence of the ensemble of classifiers, we also need Lemma 4.

Lemma 4. *If X_1, X_2,... and X are discrete random variables and $X_n \xrightarrow{P} X$, then $I(X_n = x) \xrightarrow{P} I(X = x)$ for all possible values x.*

We omit the proofs of these these lemmas since they follows quite readily from Theorem in (Oza and Russell 2001), Lemma 2, Lemma 8, Lemma 9, and Lemma 4 in (Oza 2001). We only give the proof of the main theorem.

Theorem 1. *As $N_S \to \infty$ and $N_T \to \infty$, if the base learners are naive Bayes classifiers, OTB converges to batch TrAdaBoost algorithm.*

Sketch of the Proof. The convergence of OTB can be proved by induction. For the first base learner, we have $D^o_1 \xrightarrow{P} D^b_1$ by Lemma 1. Then by Lemma 2 and Lemma 3, we have $h^o_1 \xrightarrow{P} h^b_1$, $\epsilon^o_{S,1} \xrightarrow{P} \epsilon^b_{S,1}$, $\epsilon^o_{T,1} \xrightarrow{P} \epsilon^b_{T,1}$, and $D^o_{T,1} \xrightarrow{P} D^b_{T,1}$, which completes the proof of the base case.

Now suppose we have $D^o_m \xrightarrow{P} D^b_m$, we need to prove $D^o_{m+1} \xrightarrow{P} D^b_{m+1}$, which can be shown as follow.

Note that $D^o_m(n)$ is normalized version of the Poisson parameter λ of the nth sample of online data stream. Therefore, by (2) and (3) in *Algorithm Outline* section, we have

$$D^o_{m+1}(n) = \begin{cases} \frac{D^o_m(n)}{1+D^o_{T,m}-(1-\beta)\epsilon^o_{S,m}-2\epsilon^o_{T,m}}, & h^o_m(x_n) = y_n \\ \frac{\beta D^o_m(n)}{1+D^o_{T,m}-(1-\beta)\epsilon^o_{S,m}-2\epsilon^o_{T,m}}, & h^o_m(x_n) \neq y_n \end{cases}$$

for a sample from source domain, and

$$D^o_{m+1}(n) = \begin{cases} \frac{D^o_m(n)}{1+D^o_{T,m}-(1-\beta)\epsilon^o_{S,m}-2\epsilon^o_{T,m}}, & h^o_m(x_n) = y_n \\ \frac{D^o_m(n)(D^o_{T,m}-\epsilon_{T,m})}{\epsilon_{T,m}(1+D^o_{T,m}-(1-\beta)\epsilon^o_{S,m}-2\epsilon^o_{T,m})}, & h^o_m(x_n) \neq y_n \end{cases}$$

for a sample from target domain. It can be verified that this weight update mechanism is identical to the distribution update step of batch TrAdaBoost (line 7 and line 9 of Algorithm 1). By the assumption $D^o_m \xrightarrow{P} D^b_m$, we have $h^o_m \xrightarrow{P} h^b_m$ (Lemma 2), $\epsilon^o_{S,m} \xrightarrow{P} \epsilon^b_{S,m}$, $\epsilon^o_{T,m} \xrightarrow{P} \epsilon^b_{T,m}$, and $D^o_{T,m} \xrightarrow{P} D^b_{T,m}$ (Lemma 3). Also, note that both $D^b_{m+1}(n)$ and $D^o_{m+1}(n)$ are continuous functions of these convergent quantities, we have $D^b_{m+1}(n) \xrightarrow{P} D^o_{m+1}(n)$. Again, by Lemma 2 and Lemma 3, we have $h^o_{m+1,N} \xrightarrow{P} h^b_{m+1,N}$, $\epsilon^o_{S,m+1} \xrightarrow{P} \epsilon^b_{S,m+1}$, $\epsilon^o_{T,m+1} \xrightarrow{P} \epsilon^b_{T,m+1}$, and $D^o_{T,m+1} \xrightarrow{P} D^b_{T,m+1}$, which implies that all of the base learners returned by OTB converges to that returned by batch TrAdaBoost. By Lemma 4, we have $\sum_{m=\lceil \frac{1}{M/2} \rceil}^{M} log(\frac{1-\epsilon^o_{T,m}}{\epsilon^o_{T,m}}) I(h^o_m(x) = y) \xrightarrow{P} \sum_{m=\lceil \frac{1}{M/2} \rceil}^{M} log(\frac{1-\epsilon^b_{T,m}}{\epsilon^b_{T,m}}) I(h^b_m(x) = y)$, which implies $H^o \xrightarrow{P} H^b$. □

References

Oza, N. C., and Russell, S. 2001. Online bagging and boosting. In *AISTATS*, 105–112.

Oza, N. C. 2001. *Online Ensemble Learning*. Ph.D. Dissertation, University of California, Berkeley.