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 $D_m^b$ be the weight distribution of batch TrAdaBoost algorithm, $D_m^o$ be the weight distribution of OTB, which can be viewed as the normalized version of Poisson parameter $\lambda$ in Algorithm 2. Lemma 1 shows the convergence property of $D_m^o$.

**Lemma 1.** As $N_S \to \infty$ and $N_T \to \infty$, $D_1^o \overset{P}{\to} D_1^b$.

Define $h_m^b$ as the $m$th base learner of batch TrAdaBoost, and $h_m^o$ as the analogous base learner of OTB. Lemma 2 states that if the weight vector $D_m^o$ converges to $D_m^b$, the base learner $h_m^o$ also converges to $h_m^b$.

**Lemma 2.** If $D_m^o \overset{P}{\to} D_m^b$, and the base learners are naive Bayes classifiers, then $h_m^o \overset{P}{\to} h_m^b$.

Let $\epsilon_{S,m}^b = \sum_{x_n \in S_T} D_m^o(n) I(h_m^b(x_n) \neq y_n)$, $\epsilon_{T,m}^b = \sum_{x_n \in S_T} D_m^o(n) I(h_m^b(x_n) \neq y_n)$, $D_m^b = \sum_{x_n \in S_T} D_m^o(n)$; and $\epsilon_{S,m}^o$, $\epsilon_{T,m}^o$, $D_m^o$ be their online approximation defined in line 22-24 of Algorithm 2. Lemma 3 states that $\epsilon_{S,m}^o$, $\epsilon_{T,m}^o$, $D_m^o$ converge to their batch counterparts given $h_m^o$ for a sample from source domain, and $h_m^b$ for a sample from target domain.

**Lemma 3.** If $D_m^o \overset{P}{\to} D_m^b$, $h_m^o \overset{P}{\to} h_m^b$, and the base learners are naive Bayes classifiers, then $\epsilon_{S,m}^o \overset{P}{\to} \epsilon_{S,m}^b$, $\epsilon_{T,m}^o \overset{P}{\to} \epsilon_{T,m}^b$, and $D_m^o \overset{P}{\to} D_m^b$.

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

**Lemma 4.** If $X_1, X_2, \ldots$ and $X_n \overset{P}{\to} X$, then $I(X_n = x) \overset{P}{\to} I(X = x)$ for all possible values $x$.

We omit the proofs of these last four lemmas since they follow 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 $D_1^o \overset{P}{\to} D_1^b$ by Lemma 1. Then by Lemma 2 and Lemma 3, we have $h_1^o \overset{P}{\to} h_1^b$, $\epsilon_{S,1}^o \overset{P}{\to} \epsilon_{S,1}^b$, $\epsilon_{T,1}^o \overset{P}{\to} \epsilon_{T,1}^b$, and $D_{1}^o \overset{P}{\to} D_{1}^b$, which completes the proof of the base case.

Now suppose we have $D_m^o \overset{P}{\to} D_m^b$, we need to prove $D_{m+1}^o \overset{P}{\to} D_{m+1}^b$, which can be shown as follow.

Note that $D_m^o(n)$ is normalized version of Poisson parameter $\lambda$ of the $n$th sample of online data stream. Therefore, by (2) and (3) in Algorithm Outline section, we have

$$D_{m+1}^o(n) = \begin{cases} \frac{D_m^o(n) - \epsilon_{S,m}^o}{\epsilon_{T,m}^o + \epsilon_{S,m}^o - \epsilon_{T,m}^o} & h_m^o(x_n) = y_n \\ \frac{D_m^o(n) - \epsilon_{S,m}^o}{\epsilon_{T,m}^o + \epsilon_{S,m}^o - \epsilon_{T,m}^o} & h_m^o(x_n) \neq y_n \end{cases}$$

for a sample from source domain, and

$$D_{m+1}^o(n) = \begin{cases} \frac{D_m^o(n) - \epsilon_{S,m}^o}{\epsilon_{T,m}^o + \epsilon_{S,m}^o - \epsilon_{T,m}^o} & h_m^o(x_n) = y_n \\ \frac{D_m^o(n) - \epsilon_{S,m}^o}{\epsilon_{T,m}^o + \epsilon_{S,m}^o - \epsilon_{T,m}^o} & h_m^o(x_n) \neq y_n \end{cases}$$

for a sample from target domain.

References
