Abstract

We study the problem of learning accurate ensemble predictors, in particular boosting, in the presence of multiple source domains. We show that the standard convex combination ensembles in general cannot succeed in this scenario and adopt instead a domain-weighted combination. We introduce and analyze a new boosting algorithm, MULTIBOOST, for this scenario and show that it benefits from favorable theoretical guarantees. We also report the results of several experiments with our algorithm demonstrating that it outperforms natural baselines on multi-source text-based, image-based and tabular data. We further present an extension of our algorithm to the federated learning scenario and report favorable experimental results for that setting as well. Additionally, we describe in detail an extension of our algorithm to the multi-class setting, MCMULTIBOOST, for which we also report experimental results.

1 Motivation

Ensemble methods such as Bagging, AdaBoost, Stacking, error-correction techniques, Bayesian averaging, AdaNet or other adaptive methods for learning neural networks are general machine learning techniques used to combine several predictors to devise a more accurate one (Breiman, 1996; Freund and Schapire, 1997; Smyth and Wolpert, 1999; MacKay, 1991; Freund et al., 2004; Cortes et al., 2014; Kuznetsov et al., 2014; Cortes et al., 2017). These techniques are often very effective in practice and benefit from favorable margin-based learning guarantees (Schapire et al., 1997). These algorithms assume access to a training sample drawn from the target distribution. But, in many applications, the learner receives labeled data from multiple source domains that it seeks to use to find a good predictor for target domains, which are typically assumed to be mixtures of source distributions (Mansour et al., 2008, 2009a; Hoffman et al., 2018, 2021; Cortes et al., 2021; Muandet et al., 2013; Xu et al., 2014; Hoffman et al., 2012; Saito et al., 2019; Wang et al., 2019a). How can we generalize ensemble methods such as boosting to this scenario?

Several related problems have been tackled in previous work. In the special case of a single target domain, boosting solutions have been derived (Dai et al., 2007; Yao and Doretto, 2010). These algorithms have been further improved by boosting base predictors jointly with source features (or feature views) that are predictive in the target (Yuan et al., 2017; Cheng et al., 2013; Zhang et al., 2014; Xu and Sun, 2012, 2011). Such methods have been widely adopted in various domain adaptation scenarios for different types of data. For example, Huang et al. (2010, 2012) showed that by selecting a base learner jointly with a feature that is predictive across multiple domains at every boosting step, one can achieve higher accuracy than standard transfer learning methods. Moreover, the margin
provided by boosting-style algorithms can aid in transfer learning where target domain is unlabelled. Habrard et al. (2013) have developed an algorithm that jointly minimizes the source domain error and margin violation proportion on the target domain. Wang et al. (2019a) have demonstrated that boosting classifiers from different domains can be carried out online. Taherkhani et al. (2020) and Becker et al. (2013) have shown that multi-source boosting can be combined with Deep Neural Networks for multi-task learning on large scale datasets.

We give a more detailed discussion of this related prior work in Appendix A. However, as we demonstrate in this paper, several critical questions related to the presence of multiple domains have not been fully addressed by existing boosting methods from both the algorithmic and the theoretical point of view. First, what should be the form of the ensemble solutions? We show that, in general, the standard convex combinations used in much of previous work may not succeed in this problem (Section 2). Instead, we put forward Q-ensembles, which are convex combinations weighted by a domain classifier Q, that is, $Q(k|x)$ is the conditional probability of domain $k$ given input point $x$. This is inspired by similar Q-ensembles (Cortes, Mohri, Suresh, and Zhang, 2021) or distribution-weighted combinations (Mansour, Mohri, and Rostamizadeh, 2008, 2009a; Hoffman, Mohri, and Zhang, 2018, 2021) used in the context of multiple-source adaptation and crucially differentiates our work from that of past studies of ensemble methods. Our learning scenario strictly generalizes previous algorithms. In the special case of a single domain, the form of our ensemble solutions coincides with that of the familiar ensembles used in AdaBoost. In Appendix J, we further compare our results and guarantees with the multiple-source adaptation algorithms of (Hoffman, Mohri, and Zhang, 2018, 2021; Cortes, Mohri, Suresh, and Zhang, 2021), when using AdaBoost to derive a predictor for each domain.

Second, what should be the form of the objective? Unlike existing boosting methods which optimize for a specific target distribution or domain, we seek a solution that is accurate for any mixture of the source distributions, where the mixture weights may be constrained to be in a subset of the simplex. This conforms with the learning goal in this scenario where the target distribution is typically assumed to be a mixture of the source ones and is further inspired by the formulation of the agnostic loss in the related context of federated learning (Mohri, Sivek, and Suresh, 2019). Additionally, this formulation can be viewed as more risk-averse and robust, and it also ensures more favorable fairness properties, since it does not favor any distribution possibly biased towards some subset of the source domains. In Section 3, we introduce and analyze an algorithm, MULTIBOOST, that is based on such an objective. This further distinguishes our boosting algorithm from related prior work. In Appendix L, we further give a detailed description of a multi-class extension of our algorithm, MCMULTIBOOST, including its pseudocode, the discussion of some of its properties, and the results of several experiments.

Third, in Section 4, we present a theoretical analysis of our MULTIBOOST algorithm, including margin-based generalization bounds that hold for any target mixture of the source distributions. In Appendix F, we derive finer and more flexible learning bounds that can provide more favorable guarantees, in particular when the family of target mixtures is more constrained.

Fourth, in Section 5, we describe FEDMULTIBOOST, an extension of our MULTIBOOST algorithm adapted to the distributed learning scenario of federated learning, which is increasingly important in a wide array of applications (Kairouz et al., 2021), including healthcare (Brisimi et al., 2018). FEDMULTIBOOST allows the server to train ensemble models that benefit from clients’ data, while the data remains with clients. We provide a detailed description of our algorithm, as well as a brief analysis of the communication costs, which are critical in this scenario. Appendix G contains experimental results comparing FEDMULTIBOOST with several baselines.

Finally, in Section 6, we report the results of extensive experiments with our MULTIBOOST algorithm. We start with a detailed description of the learning scenario we consider.

2 Learning scenario

Let $X$ denote the input space and $Y = \{-1, +1\}$ the output space associated to binary classification. We consider a scenario where the learner receives labeled samples from $p$ source domains, each defined by a distribution $D_k$ over $X \times Y$, $k \in [p]$. We denote by $S_k = ((x_{k,1}, y_{k,1}), \ldots, (x_{k,m_k}, y_{k,m_k})) \in (X \times Y)^{m_k}$ the labeled sample of size $m_k$ received from source $k$, which is drawn i.i.d. from $D_k$. For any function $f: X \to \mathbb{R}$ and distribution $D$ over $X \times Y$, let
$\mathcal{L}(\mathcal{D}, f)$ be the expected loss of $f$, that is $\mathcal{L}(\mathcal{D}, f) = \mathbb{E}_{(x, y) \sim \mathcal{D}}[\ell(f(x), y)]$, where $\ell$ is the binary loss $\ell(f(x), y) = 1$ if $y f(x) \leq 0$.

For any $k \in [p]$, let $\mathcal{H}_k$ denote a hypothesis set of functions mapping from $\mathcal{X}$ to $[-1, +1]$, $|\mathcal{H}_k| = N_k$.

The objective of the learner is to find a predictor $f$ that is accurate for any target distribution $\mathcal{D}_\lambda$ that is a mixture of the source distributions, where $\lambda$ may be in a subset $\Lambda$ of the simplex. Thus, $\mathcal{D}_\lambda$ can be written as $\mathcal{D}_\lambda = \sum_{k=1}^p \lambda_k \mathcal{D}_k$, with $\lambda = (\lambda_1, \ldots, \lambda_p) \geq 0$ and $\sum_{k=1}^p \lambda_k = 1$. This leads to the following agnostic loss of a predictor $f$ (Mohri et al., 2019):

$$\mathcal{L}(\mathcal{D}_\lambda, f) = \max_{\lambda \in \Lambda} \mathcal{L}(\mathcal{D}_\lambda, f).$$

To come up with a predictor $f$, the learner seeks an ensemble of functions from the base classes $\mathcal{H}_k$. What should be the form of such ensembles? A natural solution is a convex combinations of the form

$$f = \sum_{l=1}^p \sum_{j=1}^{N_l} \alpha_{l,j} h_{l,j},$$

where $h_{l,j} \in \mathcal{H}_l$ and $\alpha_{l,j} \geq 0$, $\sum_{l=1}^p \sum_{j=1}^{N_l} \alpha_{l,j} = 1$, as in prominent algorithms such as BAGGING (Breiman, 1999) or AdaBoost (Freund and Schapire, 1997). It is not hard to show, however, that for some distributions, any such convex combination of base predictors would lead to a poor solution.

**Proposition 1.** There exist distributions $\mathcal{D}_1$ and $\mathcal{D}_2$ and hypotheses $h_1$ and $h_2$ with $\mathcal{L}(\mathcal{D}_1, h_1) = 0$ and $\mathcal{L}(\mathcal{D}_2, h_2) = 0$ such that $\mathcal{L} \left( \frac{1}{2} (\mathcal{D}_1 + \mathcal{D}_2), \alpha h_1 + (1 - \alpha) h_2 \right) \geq \frac{1}{2}$ for any $\alpha \in [0, 1]$.

This result is similar to (Mansour et al., 2008, Theorem 1) or similar statements in (Mansour et al., 2008, 2009a; Hoffman et al., 2018, 2021; Cortes et al., 2021), however, there are some technical differences. In particular, we are considering here a binary classification loss and not the absolute loss. The proof is given in Appendix B.

Note that in the example of the proposition, the base predictors $h_1, h_2$ are both perfect for their respective domains. Nevertheless, unlike the common case of a single source domain, standard convex combinations in general may perform poorly. Thus, we will consider instead the following form for the ensembles of base predictors:

$$\forall x \in \mathcal{X}, \quad f(x) = \sum_{l=1}^p \sum_{j=1}^{N_l} \alpha_{l,j} Q(l|x) h_{l,j}(x),$$

where $Q(l|x)$ denotes the conditional probability of domain $l$ given $x$. This helps us account for the fact that the learner combines base predictors from different domains. For a given point $x$, experts from the domains where $x$ is more likely to appear are allocated more weight in the voted combination. The following result further substantiates the hope for this method to succeed.

**Proposition 2.** For the same distributions $\mathcal{D}_1$ and $\mathcal{D}_2$ and hypotheses $h_1$ and $h_2$ as in Proposition 1, the equality $\mathcal{L}(\mathcal{D}_\lambda, (\alpha Q(1|\cdot) h_1 + (1 - \alpha) Q(2|\cdot) h_2)) = 0$ holds for any $\lambda \in \Delta$ and any $\alpha \in (0, 1)$.

The proof is also given in Appendix B. Thus, our $Q$-ensembles can be substantially more effective in the multiple-source adaptation problem considered. Note that, in the special case of a single domain, the $Q$-combinations coincide with standard convex combinations, since $Q(k|x) = 1$ for all $x$.

As in the standard boosting scenario, we will adopt a weak-learning assumption. However, our assumption here must hold for each source domain: for each domain $k \in [p]$ and any distribution $\mathcal{D}$ over $\mathcal{X}_k$, there exists a base classifier $h \in \mathcal{H}_k$ such that the weighted loss of $h$ is $\gamma$-better than random: $\frac{1}{2} \left[ 1 - \mathbb{E}_{i \sim \mathcal{D}}[y_{k,i} h(x_{k,i})] \right] \leq \frac{1}{2} - \gamma$, for some edge value $\gamma > 0$. This is equivalent to the existence of a weak-learning algorithm for each domain, which is a mild assumption. As in the standard boosting scenario, this corresponds to the existence of a good rule of thumb for each domain. Unlike the standard scenario, however, here, we seek a $Q$-ensemble and further require that ensemble to be accurate for any target mixture $\mathcal{D}_\lambda$, $\lambda \in \Lambda$. Note that the definition of the hypothesis set $\mathcal{H}_k$ and the weak-learning assumption for domain $k$ are intimately related: the elements of $\mathcal{H}_k$ are typically simple predictors sufficient to guarantee the weak-learning assumption for that domain.

In the next section, we present an algorithm, MULTIBoost, for deriving an accurate $Q$-ensemble for any target mixture domain that belongs to the convex combination of the source domains. In Appendix L, we further present a multi-class extension of our algorithm, MCMultiBoost.
3 Algorithm

Let \( \Phi \) be a convex, increasing and differentiable function such that \( u \mapsto \Phi(-u) \) upper-bounds the binary loss \( u \mapsto 1_{u \leq 0} \). \( \Phi \) could be the exponential function as in AdaBoost or the logistic function, as in logistic regression. Using \( \Phi \) to upper-bound the agnostic loss leads to the following objective function for an ensemble \( f \) defined by (3) for any \( \alpha = (\alpha_{t,j})_{(t,j) \in [p] \times [N_t]} \geq 0 \):

\[
F(\alpha) = \max_{\lambda \in \Lambda} \sum_{k=1}^{p} \lambda_k \sum_{i=1}^{m_k} \Phi \left( -y_{k,i} \sum_{l=1}^{p} \sum_{j=1}^{N_t} \alpha_{l,j} Q(l|x_{k,i}) h_{l,j}(x_{k,i}) \right).
\]

(4)

Since a convex function composed with an affine function is convex and a sum of convex functions is convex, \( F(\alpha) \) is then convex as the maximum of a set of convex functions. In this section, we will consider the case where the set \( \Lambda \) coincides with the full simplex, that is \( \Lambda = \Delta \). \( F(\alpha) \) can then be expressed more simply as \( F = \max_{k=1}^{p} F_k \), with \( F_k(\alpha) = \frac{1}{m_k} \sum_{i=1}^{m_k} \Phi \left( -y_{k,i} \sum_{l=1}^{p} \sum_{j=1}^{N_t} \alpha_{l,j} Q(l|x_{k,i}) h_{l,j}(x_{k,i}) \right) \).

It is known that AdaBoost coincides with coordinate descent applied to an exponential loss objective function (Duffy and Helmbold, 1999; Schapire and Freund, 2012; Mohri et al., 2018). Our algorithm similarly applies coordinate descent to the objective just described, as with other boosting-type algorithms (Mason et al., 1999; Cortes et al., 2014; Kuznetsov et al., 2014). While convex, \( F(\alpha) \) is not differentiable and, in general, coordinate descent may not succeed in such cases (Tseng, 2001; Luo and Tseng, 1992). This is because the algorithm may be stuck at a point where no progress is possible along any of the active directions, while there exists a favorable descent along some other direction. However, we will show that, under the weak-learning assumption we adopted, at any point \( \alpha \) and for each active function \( f_k \), that is \( F_k(\alpha) = F(\alpha) \), there exists a coordinate direction along which a descent progress is possible. We will assume that these directions are also descent directions for \( F(\alpha) \). More generally, it suffices in fact that one such direction admits this guarantee. Alternatively, one can replace the maximum with a soft-max, that is the \((x_1, \ldots, x_k) \mapsto \frac{1}{\mu} \log \left( \sum_{k=1}^{p} e^{\mu x_k} \right) \) for \( \mu > 0 \), which leads to a continuously differentiable function with a close approximation for \( \mu \) sufficiently large.

Description. Let \( \alpha_{t-1} \) denote the value of the parameter vector \( \alpha = (\alpha_{t,j}) \) at the end of the \( (t-1) \)th iteration and \( f_{t-1} \) the corresponding function: \( f_{t-1} = \sum_{l=1}^{p} \sum_{j=1}^{N_t} \alpha_{l,j} Q(l|x) h_{l,j}(x) \). Coordinate descent at iteration \( t \) consists of choosing a direction \( e_{q,r} \) corresponding to base classifier \( h_{q,r} \) and a step value \( \eta > 0 \) to minimize \( F(\alpha_{t-1} + \eta e_{q,r}) \).

To select a direction, we consider the subdifferential of \( F(\alpha_{t-1}) \) along any \( e_{q,r} \). Since functions \( F_k \) are differentiable, by the subdifferential calculus for the maximum of functions, the subdifferential of \( F(\alpha_{t-1}) \) along the direction \( e_{q,r} \) is given by:

\[
\partial F(\alpha_{t-1}, e_{q,r}) = \text{conv}\{ F'_k(\alpha_{t-1}, e_{q,r}) : k \in K_t \},
\]

where \( F'_k(\alpha_{t-1}, e_{q,r}) \) is the directional derivative of \( F_k(\alpha_{t-1}) \) along the direction \( e_{q,r} \). We will therefore consider the direction \( e_{q,r} \) with the largest absolute directional derivative value \( F'_k(\alpha_{t-1}, e_{q,r}) \), \( k \in K_t \), but will restrict ourselves to \( q = k \) since, as we shall see, that will suffice to guarantee a non-zero directional gradient. To do so, we will express \( F'_k(\alpha_{t-1}, e_{q,r}) \) in terms of the distribution \( D_t(k, \cdot) \) over \( S_k \) defined by \( D_t(k,i) = \Phi'(-y_{k,i} f_{t-1}(x_{k,i})) \), with \( Z_{t,k} = \sum_{i=1}^{m_k} \Phi'(-y_{k,i} f_{t-1}(x_{k,i})) \), for all \( i \in [m_k] \):

\[
F'_k(\alpha_{t-1}, h_{k,r}) = \frac{1}{m_k} \sum_{i=1}^{m_k} -y_{k,i} Q(k|x_{k,i}) h_{k,r}(x_{k,i}) \Phi'(-y_{k,i} f_{t-1}(x_{k,i})) = \frac{Z_{t,k}}{m_k} [2 \epsilon_{t,k,r} - 1],
\]

where \( \epsilon_{t,k,r} = \frac{1}{2} \left[ 1 - \mathbb{E}_{i \sim D_t(k, \cdot)} [y_{k,i} Q(k|x_{k,i}) h_{k,r}(x_{k,i})] \right] \) denotes the \textit{weighted error} of \( Q(k|\cdot) h_{k,r} \).

For any \( s \in [m_k] \), since \( x_{k,s} \) is a sample drawn from \( D_t(k, \cdot) \), we have \( Q(k|x_{k,s}) > 0 \) and therefore we have:

\[
\mathbb{Q}_{t,k} = \sum_{i=1}^{m_k} D_t(k,i) Q(k|x_{k,i}) > 0.
\]

Thus, we can write \( \mathbb{E}_{i \sim D_t(k, \cdot)} [y_{k,i} Q(k|x_{k,i}) h_{k,r}(x_{k,i})] \) as

\[
\sum_{i=1}^{m_k} D_t(k,i) Q(k|x_{k,i}) y_{k,i} h_{k,r}(x_{k,i}) = \frac{1}{\mathbb{Q}_{t,k}} \mathbb{Q}_{t,k} \mathbb{Q}_{t,k} \geq \gamma > 0.
\]

where \( \mathbb{Q}_{t,k} = \mathbb{Q}_{t,k} Q(k|x_{k,i}) \). By our weak-learning assumption, there exists \( r \in N_k \) such that \( \mathbb{E}_{i \sim D_t(k, \cdot)} [y_{k,i} h_{k,r}(x_{k,i})] \geq \gamma > 0 \). For that choice of \( r \), we have \( \epsilon_{t,k,r} < \frac{1}{2} - \gamma \), with \( \gamma = \gamma \mathbb{Q}_{t,k} > 0 \).
In view of that, it suffices for us to search along the directions $h_{k,r}$ and we do not need to consider the directional derivative of $F_k$ along directions $h_q$ with $q \neq k$. The direction chosen by our coordinate descent algorithm is thus defined by: $\arg\max_{k \in K, r \in [N_k]} \frac{Z_l}{m_l} [1 - 2\epsilon_t(k,r)]$. Given the direction $e_{k,r}$, the optimal step value $\eta$ is $\arg\min_{\eta > 0} F(\alpha_t - 1 + \eta \epsilon_{k,r})$. The pseudocode of our algorithm, MULTIBOOST, is provided in Figure 1. In the most general case, $\eta$ can be found via a line search or other numerical methods.

**Step size.** In some special cases, the line search can be executed using a simpler expression by using an upper bound on $F(\alpha_t - 1 + \eta \epsilon_{k,r})$. The convexity of $\Phi$, since $y_{t,i} Q(k|x_i) h_{k,r}(x_i) = \frac{1+y_{t,i} Q(k|x_i) h_{k,r}(x_i)}{2} \cdot (+1) + \frac{1-y_{t,i} Q(k|x_i) h_{k,r}(x_i)}{2} \cdot (-1)$, the following holds for all $\eta \in \mathbb{R}$:

$$
\Phi(-y_{t,i} f_{t-1}(x_i) - \eta y_{t,i} Q(k|x_i) h_{k,r}(x_i)) \leq \frac{1+y_{t,i} Q(k|x_i) h_{k,r}(x_i)}{2} \Phi(-y_{t,i} f_{t-1}(x_i) - \eta) + \frac{1-y_{t,i} Q(k|x_i) h_{k,r}(x_i)}{2} \Phi(-y_{t,i} f_{t-1}(x_i) + \eta).
$$

In the case of exponential and logistic functions, the following upper bounds can then be derived.

**Lemma 1.** For any $k \in [p]$, the following upper bound holds when $\Phi$ is the exponential or the logistic function:

$$
F(\alpha_t - 1 + \eta \epsilon_{k,r}) \leq \max_{l \in [p]} \frac{Z_{l,t}}{m_l} \left[ (1 - \epsilon_{t,l,k,r}) e^{-\eta} + \epsilon_{t,l,k,r} e^\eta \right],
$$

where $\epsilon_{t,l,k,r} = \frac{1}{2} \left[ 1 - \frac{1}{2} \log \frac{1 - \epsilon_{t,l,k,r}}{\epsilon_{t,l,k,r}} \right]$ used in ADABoost.

**Q-function.** The conditional probability functions $Q(k|\cdot)$ are crucial to the definition of our algorithm. As pointed out earlier, Qensembles can help achieve accurate solutions in some cases that cannot be realized using convex combinations. Furthermore, for any $k \in [p]$, since $D_t(k, \cdot)$ is a distribution over the sample $S_k$, it is natural to assume that for any $j \neq k$ we have $\mathbb{E}_{s \sim D_t(k, \cdot)} Q(k|x_{k,s}) \geq \mathbb{E}_{s \sim D_t(k, \cdot)} Q(j|x_{k,s})$. This implies the following lower bound: $\mathbb{E}_{s \sim D_t(k, \cdot)} Q(k|x_{k,s}) \geq \frac{1}{p}$, which in turn implies $\tau \geq \frac{1}{p}$, since for any $x \in X$, $\sum_{j=1}^{p} Q(j|x) = 1$. In the special case where all domains coincide, we have $Q(k|x_{k,s}) = \frac{1}{m}$ for all $s$ and this lower bound is reached. At another extreme, when all domains admit distinct supports, we have $Q(k|x_{k,s}) = 1$ for all $s \in [m_k]$ and thus $\tau = \gamma$.

In practice, we do not have access to the true conditional probabilities $Q(k|\cdot)$. Instead, we can derive accurate estimates $\hat{Q}(k|\cdot)$ using large unlabeled samples from the source domains, the label $l$ used for training being simply the domain index. This can be done using algorithms such as conditional maximum entropy models (Berger et al., 1996) (or multinomial logistic regression), which benefit from strong theoretical guarantees (Mohri et al., 2018, Chapter 13), or other rich models based on neural networks.

**Other variants of MULTIBOOST.** In Appendix D, we present and discuss other variants of our algorithm, including their convergence guarantees.
4 Theoretical analysis

In this section, we present a theoretical analysis of our algorithm, including margin-based learning bounds for multiple-source Q-ensembles.

For any \(k \in [p]\), define the family \(S_k\) as follows: \(S_k = \{Q(k_i) \mid h \in \mathcal{H}_k\}\). Then, the family of ensemble functions \(\mathcal{F}\) that we consider can be defined as \(\mathcal{F} = \text{conv}(\bigcup_{k=1}^{p} S_k)\). For any \(\lambda \in \Delta\), let \(\mathcal{D}_\lambda\) be the distribution defined by \(\mathcal{D}_\lambda = \sum_{k=1}^{p} \lambda_k \mathcal{D}_k\), where \(\mathcal{D}_k\) is the empirical distribution associated to an i.i.d. sample \(S_k\) drawn from \(\mathcal{D}_k\). \(\mathcal{D}_\lambda\) is distinct from the distribution associated to \(\mathcal{D}_\lambda\).

We seek to derive margin-based bounds for ensemble functions \(f \in \mathcal{F}\) with respect to target mixture distributions \(\mathcal{D}_\lambda\), while the empirical data is from multiple source distributions \(\mathcal{D}_k\), or from \(\mathcal{D}_\lambda\) itself.

Thus, these guarantees differ from standard learning bounds where the source and target distribution coincide.

For any distribution \(\mathcal{D}\) over \(X \times Y\) and \(\rho > 0\), let \(L_\rho(\mathcal{D}, f)\) denote the expected \(\rho\)-margin loss of \(f\) with respect to \(\mathcal{D}\): \(L_\rho(\mathcal{D}, f) = \mathbb{E}_{(x,y) \sim \mathcal{D}} [\mathbb{I}(y f(x) \leq \rho)]\). We denote by \(\mathcal{R}_m(\mathcal{F})\) the Rademacher complexity of \(\mathcal{F}\) for samples \(S = (x_1, \ldots, x_m)\) of size \(m\), defined by \(\mathcal{R}_m(\mathcal{F}) = \mathbb{E}_{s \sim \text{Bernoulli}_m} \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^{m} \mathbb{I}(s_i f(x_i)) \right|\), with \(s\) independent random variables taking values in \([-1, +1]\).

The following gives a margin-bound for our multiple-source learning with Q-ensembles.

**Theorem 1.** Fix \(\rho > 0\). Then, for any \(\delta > 0\), with probability at least \(1 - \delta\) over the draw of a sample \(S = (S_1, \ldots, S_p) \sim \mathcal{D}_1^{m_1} \otimes \cdots \otimes \mathcal{D}_p^{m_p}\), the following inequality holds for all ensemble functions \(f = \sum_{t=1}^{T} \alpha_t Q(k_t) h_t \in \mathcal{F}\) and all \(\lambda \in \Delta\):

\[
L(\mathcal{D}_\lambda, f) \leq L(\mathcal{D}_\lambda, f) + \sum_{k=1}^{p} \lambda_k \left[ \frac{2}{\rho} \max_{t \in [p]} \mathcal{R}_m(\mathcal{H}_t) + \sqrt{\frac{\log^2 \frac{p}{\delta}}{2m_k}} \right].
\]

Note that the complexity term depends only on the Rademacher complexities of the families of base predictors \(\mathcal{H}_t\) and does not involve the domain classification function \(Q\). Theorem 1 improves upon previous known bounds of Mohri et al. (2019). In particular, it provides a dimension-independent margin guarantee for the zero-one loss, while the bound of Mohri et al. (2019), when applied to the zero-one loss, is dimension-dependent (Mohri et al., 2019, Lemma 3). Additionally, our bound holds for Q-ensembles. Our margin bound can be straightforwardly generalized to hold uniformly for all \(\rho \in (0, 1)\), at the cost of an additional mild term depending on \(\log \log_2(2/\rho)\) (see Theorem 4 in Appendix E). For the ensemble functions returned by our algorithm, the bound applies to \(x \mapsto \frac{f(x)}{\|\alpha\|_1}\).

These learning guarantees suggest choosing \(\alpha \geq 0\) and \(\rho\) to minimize the following:

\[
\max_{\lambda \in \Lambda} \sum_{k=1}^{p} \lambda_k \left\{ \mathbb{E}_{(x,y) \sim S_k} \left[ \mathbb{I}(y \sum_{t=1}^{T} \frac{\alpha_t}{\|\alpha\|_1} Q(k_t) h_t(x) \leq \rho) \right] \right\} + r_k \frac{\rho}{\rho},
\]

where \(r_k = 2 \max_{\alpha \in [p]} \mathcal{R}_m(\mathcal{H}_t)\). Choosing \(\frac{\rho}{\|\alpha\|_1}\) and upper bounding the binary loss using a convex function \(\Phi\), the problem can then be equivalently cast as the following unconstrained minimization, for example in the case of the exponential function:

\[
\min_{\alpha \geq 0} \max_{\lambda \in \Lambda} \sum_{k=1}^{p} \lambda_k \left\{ \mathbb{E}_{(x,y) \sim S_k} \left[ \exp \left( -y \sum_{t=1}^{T} \alpha_t Q(k_t) h_t(x) \right) \right] + r_k \|\alpha\|_1 \right\}.
\]

Thus, the learning guarantees justify a posteriori an \(\ell_1\)-regularized version of our algorithm. It is straightforward to derive that extension of our algorithm and its pseudocode. This is similar to the extension to the \(\ell_1\)-AdaBoost (Rätsch et al., 2001) of the original unregularized AdaBoost (Freund and Schapire, 1997). Let us add that finer learning guarantees such as those for Deep Boosting (Cortes et al., 2014) ones can be derived similarly here, which would lead to a weighted \(\ell_1\)-regularization, where the weights are the Rademacher complexities of the families \(\mathcal{H}_k\). For the sake of simplicity, we do not detail that extension. Let us point out, however, that such an extension would also lead to an algorithm generalizing AdaNet (Cortes et al., 2017) to multiple sources.

In some applications, prior knowledge can be used to constrain \(\lambda\) to a strict subset of the simplex \(\Delta\). Finer margin-based learning guarantees can be derived to cover that scenario. Let \(\bar{\Lambda}\) be the set of
vertices of a subsimplicial cover of $\Lambda$, that is a decomposition of a cover of $\Lambda$ into subsimplices and let $\bar{X}_c$ be the set of $\lambda c$-close to $X$ in $\ell_1$-distance. Then, the following guarantees hold.

Theorem 2. Fix $\rho > 0$ and $c > 0$. Then, for any $\delta > 0$, with probability at least $1 - \delta$ over the draw of a sample $S = (S_1, \ldots, S_p) \sim D_{\rho}^{m_1} \otimes \cdots \otimes D_{\rho}^{m_p}$, each of the following inequalities holds:

1. for all ensemble functions $f = \sum_{t=1}^T \alpha_t Q(k_t^\dagger) h_t \in \mathcal{F}$ and all $\lambda \in \bar{X}_c$:

$$\mathcal{L}(D, \lambda) \leq \mathcal{L}(D, \lambda) + \frac{2}{\rho} \max_{r \in [p]} \mathcal{R}_{m_2} (\mathcal{H}_r, \lambda) + \frac{3\epsilon}{2} + \frac{p}{\rho} \sum_{k=1}^p \frac{\lambda_k^2}{2m_k} \log \frac{|\Lambda|}{\delta}. \quad (6)$$

2. for all ensemble functions $f = \sum_{t=1}^T \alpha_t Q(k_t^\dagger) h_t \in \mathcal{F}$ and all $\lambda = \sum_{k=1}^p \mu_k \beta_k \in \text{conv}(\bar{X})$:

$$\mathcal{L}(D, \lambda) \leq \mathcal{L}(D, \lambda) + \frac{2}{\rho} \sum_{k=1}^p \mu_k \max_{r \in [p]} \mathcal{R}_{m_2} (\mathcal{H}_r, \beta_k) + \sum_{k=1}^p \mu_k \frac{\beta_k^2}{2m_k} \log \frac{|\Lambda|}{\delta} \quad (7).$$

The proof of the theorem and further discussion are presented in Appendix F. Here, $\mathcal{R}_{m_2} (\mathcal{H}_r, \lambda)$ is the $\lambda$-weighted Rademacher complexity of $\mathcal{H}_r$ (see Appendix F). For $\lambda$ a strict subset of the simplex, these learning bounds suggest an algorithm with a regularization term of the form $\sum_{k=1}^p \lambda_k^2 / m_k$. Our analysis and learning guarantees can be similarly extended to cover the multi-class setting.

### 5 Federated MultiBoost algorithm

In this section, we present an extension of our algorithm to the federated learning scenario, called 

**FedMultiBoost.** The pseudocode is given in Figure 2. Federated learning is a distributed learning scenario where a global model is trained at the server level, based on data remaining at clients (Konečný et al., 2016; McMahan et al., 2017; Yang et al., 2019). Clients are typically mobile phones, network sensors, or other IoT devices. While the data remains at the clients, the trained model significantly benefits from user data, as demonstrated, for example, in next word prediction (Hardt et al., 2018; Yang et al., 2018) and healthcare applications (Brisimi et al., 2018). We refer the reader to (Kairouz et al., 2021) for a more detailed survey of federated learning.

A widely used algorithm in this scenario is federated averaging, where the server trains the global model via stochastic updates from the clients (McMahan et al., 2017). It has been argued by Mohri et al. (2019), however, that minimizing the expected loss with respect to a particular training distribution, as done by standard federated learning algorithms, may be risk-prone and benefit only a specific subset of clients. This compromises fairness, which is one of the critical questions in federated learning, where clients are often heterogeneous (Bickel et al., 1975; Hardt et al., 2016; Abay et al., 2020; Li et al., 2020). This issue can be further
We also compare our results with the a multiple-source adaptation algorithm, as well as experimental results can be found in Appendix. The errors and their standard deviations are reported based on 10-fold cross validation. Each source dataset.

Before presenting our algorithm, we first describe the learning scenario and the notation we adopt. Let $n$ denote the number of clients, which in the cross-device setting can be in the order of millions. Let $m_c$ denote the number of samples in client $c$ and $(x_{c,i}, y_{c,i})$ denote the $i$th sample of client $c$. The server has access to a set of base-predictor classes $H_k$ for $k \in [p]$ and a domain classifier $Q$ that assigns for each sample $x$, the likelihood of it belonging to each of the $p$ domains. The base classifiers can be obtained by training on public data or training on client data with federated learning.

The domain classifier $Q$ can be trained using unlabeled client data. Additional discussion of the algorithm as well as experimental results can be found in Appendix G. At each round, the algorithm randomly selects $r$ clients and $s$ base classifiers from each domain. Next, it chooses the best classifier out of the previously selected subset, based on the data from the subsampled clients. Since FedMULTIBoost operates only on a small set of base classifiers at each round, the algorithm is communication-efficient. At round $t$, the server needs to send $Q$, $f_{t-1}$, and $s$ base classifiers to the selected clients. Hence, the server-to-client communication cost at round $t$ is $O((t+s)K + K')$, where $K$ is the number of parameters in a single base classifier and $K'$ is the number of parameters required to specify $Q$. Furthermore, the clients communicate only the aggregate statistics $\hat{F}_c, Z_{t,c}, \beta_{t,c,r}$ to the server and hence the client-to-server communication is $O(p \cdot s)$ real numbers, which is independent of the number of parameters of the base classifiers. We note that FedMULTIBoost can be extended to the multi-class setting using techniques presented in Appendix L.

6 Experiments

In this section, we present experimental results for the MULTIBoost algorithm on several multiple-source datasets. Our study is restricted to learning an ensemble of decision stumps $H_{stumps}$ using the exponential surrogate loss $\Phi(u) = e^{-u}$. To estimate the probabilities $Q(k|\cdot)$ for $k \in [p]$, we assigned the label $k$ to each sample from domain $k$ and used multinomial logistic regression. This step can be facilitated with unlabeled examples, as only their domain membership information is required.

We compare MULTIBoost with a set of boosting-style algorithms that operate on the same hypotheses class $H_{stumps}$. Those include ADABOOST-$k$ for $k \in [p]$ and ADABOOST-all. The former is a standard ADABoost algorithm trained only on a single source $S_k$ and the latter is ADABoost trained on the union of all sources $\cup_{k=1}^p S_k$. It is natural to compare MULTIBoost against ADABoost-all, since both of them have access to all sources during training. The difference is that while ADABoost-all treats $\cup_{k=1}^p S_k$ as a single dataset, MULTIBoost trains base learners separately for each source and weights examples by domain probabilities $Q(k|\cdot)$. We used $T = 100$ boosting steps for all benchmarks. $T$ up to 1,000 were explored, but did not change any results significantly.

We also compare our results with the a multiple-source adaptation algorithm, DMSA, designed for a scenario where no labeled data is available, and where only an accurate predictor per domain and unlabeled data are supplied. DMSA was shown to outperform other algorithms designed for that scenario. To apply DMSA, we use the domain predictors ADABoost-$k, k \in [p]$.

We report classification errors on the test data for various mixtures of the source distributions $\lambda_1, \ldots, \lambda_p$, including: errors for $\lambda$ on the simplex $\Delta$ edges (errors on each domain separately); errors on the uniform mixture $\forall k: \lambda_k = \frac{1}{p}$; agnostic error, which is the maximum error across all sources. The errors and their standard deviations are reported based on 10-fold cross validation. Each source
\(S_k\) is independently split into 10 folds \(S_k^1, \ldots, S_k^{10}\). For the \(i\)-th cross-validation step, the test set is \(\{S_k^1, \ldots, S_k^i\}\), while the rest is used for training.

Datasets and preprocessing steps used are described below with additional dataset details provided in Appendix H. Note, that all datasets are public and do not contain any personal identifiers or offensive information. The experiments were performed on Linux and Mac workstations with Quad-Core Intel Core i7 2.9 GHz and Intel Xeon 2.20 GHz respectively. The time taken to perform 10-fold cross-validation varies per dataset: from roughly 1 hour on CPU for all benchmarks on tabular data to 8 hours on CPU for all benchmarks on digits data. The run-times of a single MULTIBOOST iteration is almost identical to that of ADABOOST-all when a closed form solution is used for step size \(\eta_t\) (line 12 in Figure 1) for the exponential surrogate loss function. Alternatively, for some experiments we used line search with 100 steps. If training time is critical, one can use \(\eta_t = C/\sqrt{t}\) instead. The cost of slightly worse convergence guarantee. Convergence plots are illustrated in Appendix I.

**Sentiment analysis.** The Sentiment dataset (Blitzer et al., 2007) consists of text reviews and rating labels for products sold on Amazon in various categories. We selected \(p = 3\) sources from this dataset, where \(k = 1\) corresponds to books, \(k = 2\) means dvd, and \(k = 3\) is electronics.

We converted the 5-star rating to binary labels, assigning \(y = +1\) for a positive review and \(y = -1\) for a negative one. Neutral reviews with 3-star labels are removed from the dataset. The benchmarks are trained using bigram features generated from the raw text reviews. Since the Sentiment dataset contains only 2,000 instances for each domain, we used random train/test splits with 10 different seeds instead of the 10-fold cross-validation.

To illustrate the importance of the conditional domain probabilities \(Q(K|x)\), for each domain, we illustrate its values along the projections of \(x\) onto the first principal component of the joint Sentiment dataset in Figure 3. The figure shows that our domain classifiers are able to provide coherent separation between the three domains and narrow the applicability of the base classifiers.

In Appendix K, we further discuss the importance of obtaining a high-quality \(Q\) function and illustrate the performance degradation resulting from a lower-quality \(Q\) function.

**Digits recognition.** For this problem, we aggregated 32x32 pixel handwritten digits images from \(p = 3\) sources: MNIST \((k = 1)\), SVHN \((k = 2)\), MNIST-M \((k = 3)\). We compared algorithms on two binary classification problems: digits 4 vs. 9 and digits 1 vs. 7.

**Object recognition.** We divided images of clothes items from Fashion-MNIST (Xiao et al., 2017) dataset into two classes: tops and bottoms from \(p = 3\) sources. \(k = 1\) consists of t-shirts, pullovers, trousers; \(k = 2\) consists of dresses, coats, sandals; \(k = 3\) consists of shirts, bags, sneakers, ankle-boots.

**Tabular data.** In addition to text and images, we tested MULTIBOOST on tabular data. We used the Adult dataset (Kohavi, 1996), which consists of numerical and categorical features describing an individual’s socioeconomic characteristics, given the task to predict if an person’s income exceeds USD 50,000. We divided this dataset into \(p = 3\) sources based on individual’s educational background. Source \(k = 1\) consisted of individuals with a university degree (BSc, MS or PhD), source \(k = 2\) those with only a High School diploma, and source \(k = 2\), none of the above.

**Discussion** As can be seen from Table 4, MULTIBOOST provides agnostic and uniform errors that are significantly better than the baselines on all datasets. While ADAABOOST-\(k\) predictors on digits recognition and object recognition tasks each show the lowest error on their specific domains \(D_1, D_2, D_3\), they fail to generalize to other target distribution in \(\text{conv}\{D_1, D_2, D_3\}\). Moreover, as suggested by the discussion in Section 2, the standard convex combination of domain-specific predictors (ADAABOOST-all) also performs poorly on several target distributions in \(\text{conv}\{D_1, D_2, D_3\}\), such as the uniform and agnostic targets. As the experimental results confirm, MULTIBOOST, by leveraging domain predictors \(Q(k|x)\) and the agnostic loss, is able to produce an ensemble of domain-specific predictors that generalizes to different targets. Appendix I provides additional illustrations of \(Q(k|x)\) by projecting each domain on the first principal component of the joint data.

Moreover, for tabular data and the digits (4 vs. 9) classification problem, MULTIBOOST benefits from positive knowledge transfer and obtains an error on individual domains that is even smaller than that...
We presented a new boosting algorithm, learning scenario, which is a critical distributed learning setting in modern applications. We showed that our algorithm benefits from strong theoretical guarantees and exhibits setting in domain adaptation problems – where the target distribution can be any mixture of the source distributions. We also highlighted the extension of our work to the federated learning scenario, which is a critical distributed learning setting in modern applications.

7 Conclusion

We presented a new boosting algorithm, MultiBoost, for a multiple-source scenario – a common setting in domain adaptation problems – where the target distribution can be any mixture of the source distributions. We showed that our algorithm benefits from strong theoretical guarantees and exhibits favorable empirical performance.
Acknowledgments

This work was partly supported by NSF CCF-1535987, NSF IIS-1618662, and a Google Research Award.

References


