Axiomatic Characterization of AdaBoost and the Multiplicative Weight Update Procedure

Ibrahim Alabdulmohsin(✉)

Business Intelligence Division, Saudi Aramco, Dhahran 31311, Saudi Arabia
ibrahim.alabdulmohsin@kaust.edu.sa

Abstract. AdaBoost was introduced for binary classification tasks by Freund and Schapire in 1995. Ever since its publication, numerous results have been produced, which revealed surprising links between AdaBoost and related fields, such as information geometry, game theory, and convex optimization. This remarkably comprehensive set of connections suggests that adaBoost is a *unique* approach that may, in fact, arise out of *axiomatic* principles. In this paper, we prove that this is indeed the case. We show that three natural axioms on adaptive re-weighting and combining algorithms, also called arcing, suffice to construct adaBoost and, more generally, the multiplicative weight update procedure as the unique family of algorithms that meet those axioms. Informally speaking, our three axioms only require that the arcing algorithm satisfies some elementary notions of additivity, objectivity, and utility. We prove that any method that satisfies these axioms must be minimizing the composition of an exponential loss with an additive function, and that the weights must be updated according to the multiplicative weight update procedure. This conclusion holds in the general setting of learning, which encompasses regression, classification, ranking, and clustering.

Keywords: Ensemble methods · Boosting · AdaBoost · Axioms

1 Introduction

In an axiomatic treatment, the goal is to formalize broad intuitive notions into precise mathematical terms, called *axioms*. Often, a collection of axioms would pinpoint unequivocally to a *unique* solution but, in some striking cases, it could result in an *impossibility theorem*, where it is concluded that no solution could possibly satisfy all of the postulated axioms. In both instances, an in-depth insight is provided by axiomatization. Unfortunately, despite the fact that axiomatic methods are quite abundant in related fields, developing an axiomatic basis for artificial intelligence, in general, and machine learning, in particular, is not a common practice today.

Consider, for the sake of comparison, the closely-related field of information theory. In information theory, the Shannon entropy alone has been characterized

M. Berlingerio et al. (Eds.): ECML PKDD 2018, LNAI 11051, pp. 591–604, 2019.
https://doi.org/10.1007/978-3-030-10925-7_36

axiomatically via, at least, seven different approaches [2,11,22,33]. These different axiomatizations are built on intuitive notions, such as continuity, additivity, monotonicity, recursivity and symmetry. Similarly, the maximum-entropy method has been axiomatized in various ways [10,34,35]. Moreover, axiomatic characterizations have been developed for the Kullback-Leibler divergence by Kannappan and Ng, for the Rényi entropies by Daróczy, and for the broad class of f-divergences by Csiszár [11,24].

In machine learning, on the other hand, only a few areas within such a broad discipline have witnessed some axiomatic treatments. These include clustering, rank aggregation, Bayesian inference, and collaborative filtering. Even more, in some of those few axiomatic approaches, the conclusion turned out to be implicitly imposed in the axioms themselves, thus significantly weakening the implications of their results. For instance, [3] proposed an axiomatic approach for defining *relevance* in feature subset selection. Because the two axioms of [3] required that the *mutual information* between the instance and the target be preserved and that the description length be minimized, it is not, perhaps, quite surprising that the proposed definition of relevance was expressed in terms of the mutual information itself. Similarly, the axiomatic basis of [17] for clustering algorithms identified *single-linkage* as the only clustering method that met those axioms. However, as pointed out by [1], this is arguably a consequence of the fact that one of the axioms used in [17] was, implicitly, the optimization objective of that particular clustering algorithm.

Nevertheless, there are celebrated stories of successful axiomatic characterizations in machine learning. One of the most prominent examples is the axiomatization of Bayesian inference using Cox's theorem. In 1946, Cox established that the laws of probability were the *only* method of manipulating one's degrees of belief in a manner that was both consistent and agreed with common sense [9]. Jaynes, in his posthumous book "*Probability Theory: The Logic of Science*", viewed this theorem as the cornerstone of the Bayesian interpretation of probability theory [18].

A second area in machine learning that has received a fair amount of axiomatic treatment is clustering. In [21], Kleinberg developed an impossibility theorem, showing that no clustering function could achieve scale-invariance and richness, while also simultaneously satisfying a third condition, which he called "consistency". This negative result was interpreted as a formal proof of the ill-defined nature of the clustering task. However, while the first two of Kleinberg's axioms were quite natural, the third axiom was, in fact, quite strong. One, arguably, more natural approach would be to state the third axiom in terms of the "refinement" of the partition under "transformations", rather than requiring the partitions to be identical, *per se*, or, at least, to fix the number of clusters in advance. Indeed, using similar arguments, a different axiomatization was derived in [1] that captured the same principles of Kleinberg while sidestepping his impossibility result.

One particular axiomatic tool that has found many applications in machine learning are the axioms of social choice theory. In [26], an analog to the

celebrated Arrow's impossibility theorem was derived for combining the predictions of weak learners in ensemble methods in the multiclass setting, and axiomatic characterizations of weighted averaging and majority voting were provided as well. Similarly, [27] derived analogs of the axioms of social choice theory for rank aggregation, while [25] derived analogs for collaborative filtering.

In this paper, we develop an axiomatic characterization of boosting algorithms that operate via adaptive re-weighting and combining, which are also called arcing methods [5]. By requiring that these procedures satisfy three elementary notions of additivity, objectivity and utility, we prove that adaBoost and its variants are the *unique* family of boosting methods that meet those axioms. More precisely, we prove that any boosting procedure that satisfies these axioms must also be minimizing the composition of an exponential loss with an additive function and that the weights must be updated according to the multiplicative weight update procedure. This conclusion holds even in the general setting of learning [32,36], which encompasses regression, classification, ranking, and clustering, despite the fact that *weak learnability* is hard, if not impossible, to define in such a broad setting. To the best of our knowledge, this is the first axiomatic treatment of boosting in the literature.

Definition 1 (General Setting of Learning). *In the general setting of learning, we have a hypothesis space $\mathcal{H}$ and a stochastic loss $l : \mathcal{H} \to \mathbb{R}$. The learner is provided with m realizations of the stochastic loss $S = \{l_1, \ldots, l_m\}$ drawn i.i.d. from some unknown probability measure $\mathcal{D}$. The goal is to select a hypothesis $\boldsymbol{h} \in \mathcal{H}$ according to the sample S such that the expected risk $\mathbb{E}_{\boldsymbol{l} \sim \mathcal{D}}[\boldsymbol{l}(\boldsymbol{h})]$ is small.*

There are several families of boosting algorithms that have been proposed in the literature. Two prominent families include gradient boosting [16,23], of which *anyBoost* is a prominent example, and the adaptive re-weighting and combining (arcing) procedure [5], which includes algorithms such as *adaBoost* [14], arc-x4 [5], and the averaging algorithm in [19]. In this paper, we focus on the latter class of algorithms. We will show that while, in principle, many adaptive re-weighting and combining schemes can be devised, as argued by Breiman in [5], there exists a *unique* method that satisfies some natural axioms. This unique approach coincides with adaBoost in the binary classification setting.

AdaBoost was introduced for binary classification by Freund and Schapire in [14]. Ever since its publication, numerous results have been produced, which revealed surprising links between AdaBoost and related fields, such as information geometry, game theory, and convex optimization [29]. This remarkably comprehensive set of connections suggests that adaBoost is a fundamental approach that may, in fact, arise out of *axiomatic* principles. We establish in this paper that this is indeed the case. In the literature, several variants of adaBoost have been proposed that minimize different loss functionals, such as the logistic loss, the hinge loss, the square loss, and so on [23,29,37,39]. Our axiomatic characterization establishes in what sense do those algorithms essentially differ from adaBoost.

AdaBoost is, in turn, a particular instance of a more general learning strategy, called the Multiplicative Weight Update procedure [29]. The Multiplicative Weight Update procedure arises in game theory as a learning algorithm in mixed games with repeated play. Similarly, it arises in the online prediction setting as a generalization of the Weighted Majority Algorithm. As will be clear later, the axioms we present for AdaBoost can also serve as axioms for the Multiplicative Weight Update procedure.

2 The Boosting Framework

Before we present an axiomatic characterization of adaBoost and related algorithms, we need to define what a "boosting (arcing) procedure" is. We begin with an informal description, first.

2.1 Informal Description

Boosting is an instance of a broad category of machine learning algorithms, called *ensemble methods*, which combine weak learners into strong aggregated rules. Ensemble methods differ in how they instantiate weak learning algorithms and how they combine them afterward. For instance, in a *bagging* approach, multiple weak learners are trained on bootstrap subsamples of the training set, which, in turn, are aggregated by averaging or majority voting. The most well-known example of this approach is the *random forests* algorithm introduced by [6]. In a *stacking* approach, on the other hand, the predictions of weak learners form a new representation of the data, and aggregation is carried out by training a weak learner on the newly learned representation [12].

In a boosting approach, by contrast, a weak learner is supplied with *both* a training sample and a set of weights on those training examples. At each stage of the algorithm, the task of a weak learner is to do well with respect to a weighted training set. The predictions of those weak learners are, then, combined using an appropriate aggregation rule. As mentioned by [14,29] and studied empirically in [5], this setting can be relaxed by subsampling from the training set according to the weights. However, since subsampling can be considered as a part of the weak learner's algorithm, we adopt the convention of having weighted training examples here for generality.

In our axiomatic approach, which holds in the general setting of learning (see Definition 1), we will not formally define the notion of "weak learnability"[1]. In fact, we will not even require it! Hence, we will refer to weak learners from now on as *base* learners.

[1] Weak learnability, informally speaking, only ensures that the learner performs better than random guessing rather than mandating the learner to achieve an arbitrarily optimal performance.

2.2 Formal Definitions

In this section, we introduce our notation and some preliminary definitions. Let $\mathcal{H}$ be a hypothesis space and let $\mathbb{H}$ be a space that is formed from $\mathcal{H}$ using an appropriate aggregation rule. For example, $\mathbb{H}$ might be the space of all linear combinations of hypotheses in $\mathcal{H}$. We assume throughout this paper that $\mathcal{H} \subseteq \mathbb{H}$ and that $\mathbb{H}$ is a *vector space* on $\mathbb{R}$ (or $\mathbb{C}$). That is, $\mathbb{H}$ is closed under addition and scalar multiplication on $\mathbb{R}$ (or $\mathbb{C}$). In particular, notions such as addition, scalar multiplication and linear maps of hypotheses in $\mathcal{H}$ are meaningful. Examples of hypothesis spaces that satisfy this assumption include the Euclidean plane $\mathbb{R}^d$, such as in linear classification or regression problems, and function spaces, such as Hilbert spaces in kernel methods. Moreover, the probability simplex in $\mathbb{R}^m$ will be denoted $\mathcal{P}^m$.

Definition 2 (Span). *Let $\mathcal{H}$ be a set of hypotheses that reside in some vector space on $\mathbb{R}$ (or $\mathbb{C}$). Then, the span of $\mathcal{H}$ is the set of all possible combinations of finite elements in $\mathcal{H}$. That is:* $\mathbf{Span}(\mathcal{H}) = \left\{h : \exists h_1, \ldots, h_K \in \mathcal{H} \;:\; h = \sum_{k=1}^{K} h_k\right\}$.

The goal of learning via boosting is to select an aggregated rule $\mathbf{h} \in \mathbb{H}$ that minimalizes some weighted loss $\sum_{i=1}^{m} w_1(i)\, l_i(\cdot)$, for some fixed set of loss functions $l_i : \mathbb{H} \to \mathbb{R}$ and some initial distribution $w_1 \in \mathcal{P}^m$. In the general setting of learning, a loss function $l_i(\cdot)$ is often of the form $l(\cdot, z_i)$, where $z_i \in \mathcal{Z}$ is the i-th training example. In the latter case, $S = \{l_1, \ldots, l_m\}$ is fixed in all rounds of the boosting algorithm because the training sample $(z_1, \ldots, z_m)$ is fixed. This is often referred to as the *boosting-by-sampling* setting [14,31]. Since, for a fixed sample S, only the *weights* on the loss functions determine the outcome of a learning algorithm, one may view a base learner as a mapping from the probability simplex $\mathcal{P}^m$ to the hypothesis space $\mathcal{H}$. This brings us to the following definition:

Definition 3 (Base Learner). *For a fixed set of loss functions $S = \{l_1, \ldots, l_m\}$, a base learner f_S is a (possibly randomized) mapping $f_S : \mathcal{P}^m \to \mathcal{H}$. The set of all base learners will be denoted $\mathcal{F}_S$.*

Informally, we interpret f_S as follows. A base learner f_S is supplied with a distribution $w_t \in \mathcal{P}^m$ on the loss functions in S. Then, the task of f_S is to select a hypothesis $h_t \in \mathcal{H}$ whose weighted loss $\sum_{i=1}^{m} w_t(i)\, l_i(h_t)$ is small. Note that this is merely an informal interpretation since f_S can be any *arbitrary* map. Examples of base learners include the support vector machine (SVM), decision stumps, and the classification and regression tree (CART) algorithm [7,8,29].

Definition 4 (Boosting Procedure). *For a fixed sample $S = \{l_1, \ldots, l_m\}$, a boosting (arcing) procedure is a mapping $g : \mathcal{P}^m \times \mathcal{F}_S \to \mathbb{H}$, which operates as follows. Initially, g is provided with a distribution $w_1 \in \mathcal{P}^m$ and a base learner f_S. Then, g operates sequentially in T rounds. At round t, it assigns $h_t = f_S(w_t)$ and updates the weight $w_t \to w_{t+1}$ according to $(h_1, \ldots, h_t)$. The final output is $\boldsymbol{h} \in \mathbb{H}$, which is an aggregation of all the base hypotheses $(h_1, \ldots, h_T)$.*

According to Definition 4, a boosting procedure generates several base hypotheses in sequence via adaptive re-weighting. Finally, it combines them into an aggregated rule. Needless to mention, several possible aggregation methods exist, such as by using majority voting, averaging, random sampling, or by simply selecting the single hypothesis obtained in the final round[2].

3 Axiomatic Characterization

As shown in Definition 4, boosting procedures, which operate via adaptive re-weighting and combining, can vary according to how they update the weights w_t and how they combine hypotheses afterwards. In fact, they also vary depending on the choice of the loss functions $S = \{l_1, \ldots, l_m\}$ that they minimalize. Indeed, many boosting algorithms that have been proposed in the past can be categorized along these lines.

For example, Breiman introduced a boosting algorithm, called arc-x4, in which the weight of a training example at a given round is proportional to $1+\epsilon(i)$, where $\epsilon(i)$ is the number of times the i-th training example had been misclassified by hypotheses in previous rounds [5]. This is similar in spirit to adaBoost; it forces the classifier to focus on the training examples that are harder to predict correctly. Similarly, Ji and Ma proposed a different scheme, in which training examples are partitioned into "cares" and "don't-cares", according to whether or not they are classified correctly by the combined classifier [19]. Needless to mention, other possibilities also exist. For instance, [20] studied adaptive re-weighting schemes, where $w_t(i) \propto (1 + \epsilon(i)^n)$ for different choices of $n \in \mathbb{N}$, and so on.

The fact that many adaptive re-weighting schemes can be (and have been) proposed raises the following fundamental question: What natural axioms should a boosting procedure satisfy? And, do such axioms lead to a unique approach? We will answer these questions in the remainder of the paper. We will show that once three natural axioms are imposed, the range of possibilities is greatly reduced. Essentially, the only boosting algorithm that operates via adaptive re-weighting and combining and also satisfies the postulated axioms will turn out to be a slight generalization of adaBoost. Surprisingly, this holds even though (1) we operate in the general setting of learning and (2) we do not impose any constraints on the base learner f_S and the base hypotheses $(h_1, \ldots, h_T)$.

3.1 The Axioms

Our axioms are three: ADDITIVITY, OBJECTIVITY and UTILITY. We describe each axiom in details next.

[2] The main results of this paper can be extended to the setting where the base learner f_S is different at each round t. However, we assume in this paper, with no loss of generality, that f_S is fixed to simplify the discussion and notation.

Axiom 1 (Additivity). *Let $w_T = g_T(h_1, \ldots, h_{T-1})$ be the adaptive weights selected by the boosting procedure g at round T when $h_1, \ldots, h_{T-1}$ are the base hypotheses provided by the base learner $f_S : \mathcal{P}^m \to \mathcal{H}$ at all preceding rounds. Let $h_t = f_S(w_t)$ for all rounds $t \geq 1$. Then, the aggregated rule is $\sum_{t=1}^{T} h_t$.*

Our first axiom states that the boosting procedure is a stagewise *additive* model; at each round, it improves its prediction rule by reweighting the sample and *adding* a new hypothesis to the aggregated rule. These additive models have a long history in statistics and signal processing (see for instance [15] and the references therein). Note that in the case of binary classification problems, where the target set is $\mathcal{Y} = \{-1, +1\}$ and $\mathbf{sign}(\mathbf{h}(\cdot)) : \mathcal{X} \to \mathcal{Y}$ is the prediction rule, the aggregation rule in Axiom 1 reduces to *majority voting*. For regression problems, the aggregation rule in Axiom 1 reduces to *averaging*. Hence, it has a wide applicability.

Axiom 2 (Objectivity). *For any $T \geq 1$ and any $i \in \{1, \ldots, m\}$, we have $w_T(i) \to 0$ as $w_1(i) \to 0$. More precisely: $\forall\big(\epsilon \geq 0, i \in \{1, \ldots, m\}, T \geq 1\big) : \exists \delta_T \geq 0 :\quad w_1(i) \leq \delta_T \;\Rightarrow\; w_T(i) \leq \epsilon$.*

Informally, our second axiom requires that the boosting procedure aims at minimalizing its original objective function *only*. More precisely, if a loss function $l_i(\cdot)$ has an initial weight of zero, such as when it is not in the training sample to begin with, then its weight $w_t(i)$ at all rounds $t \geq 1$ will remain zero. Formally, Axiom 2 requires that for any fixed $t \geq 1$, we have $w_t(i) \to 0$ as $w_1(i) \to 0$. In particular, we note that the formal specification of Axiom 2 implies that $w_t(i) = 0$ for all $t \geq 1$ if $w_1(i) = 0$ as described earlier[3].

Axiom 3 (Utility). *At any round $T \geq 1$ and for any $h_T \in \mathcal{H}$, we have $\sum_{i=1}^{m} w_T(i)\, l_i(h_T) \propto \sum_{i=1}^{m} w_1(i)\, l_i(\sum_{t=1}^{T} h_t)$, with a proportionality constant that is independent of h_T.*

The last axiom can be intuitively understood in light of the aggregation rule $\mathbf{h} = \sum_{t=1}^{T} h_t$ stipulated by Axiom 1. Prior to round T, the aggregated hypothesis is $\sum_{t=1}^{T-1} h_t$. At round T, the task of the base learner f_S is to select a hypothesis $h_T \in \mathcal{H}$ that performs well according to the objective function $\sum_{i=1}^{m} w_T(i) l_i(\cdot)$. This hypothesis h_T will, then, be added to the aggregated rule, which, ideally, should result in a better aggregated hypothesis with respect to the *original* objective function $\sum_{i=1}^{m} w_1(i) l_i(\mathbf{h})$. Axiom 3 states that the better h_T is for the learning problem at round T, the better it is for the original optimization problem.

To recall, the base hypotheses $(h_1, \ldots, h_T)$ selected by the base learner $f_S : \mathcal{P}^m \to \mathcal{H}$ can be entirely arbitrary. As mentioned earlier, we have not imposed any notion of goodness on f_S, such as weak learnability. However, in

[3] Axiom 2 can also be interpreted as a *stability* constraint on the boosting procedure. It can be argued that the main advantage of ensemble methods is their ability to improve stability, which results in a reduced over-fitting risk and an improved generalization [4,5,32].

order for the boosting algorithm to be of any utility, the performance of the base hypothesis h_t at round t should factor into the performance of the overall boosting procedure. In particular, we should impose a condition, which qualitatively states that having a "better" hypothesis at round t would be more "helpful" to the overall boosting algorithm. This is achieved by Axiom 3.

3.2 Proof of Independence

Next, we show that the three axioms are mutually independent.

Proposition 1. *Any two of the three axioms* ADDITIVITY, OBJECTIVITY, *and* UTILITY *can be satisfied without satisfying the third axiom. In other words, the three axioms are independent of each other.*

Proof. First, we show that ADDITIVITY and OBJECTIVITY can be satisfied without satisfying the UTILITY axiom. Let $l_i(h) = \langle z_i, h\rangle$ be a linear cost function. The objective is to minimalize $\sum_{i=1}^{m} w_1(i)\,\langle z_i, h\rangle$ in some hypothesis space $\mathcal{H}$. Let g be the boosting procedure, which always sets $w_t(i) = w_1(i)$ and combines hypotheses according to Axiom 1. Then, g satisfies the ADDITIVITY and OBJECTIVITY axioms trivially. However:

$$\sum_{i=1}^{m} w_T(i) l_i(h_T) = \sum_{i=1}^{m} w_1(i)\langle z_i, h_T\rangle = \sum_{i=1}^{m} w_1(i)\langle z_i, \sum_{t=1}^{T} h_t\rangle \; + \beta,$$

where $\beta = -\sum_{i=1}^{m} w_1(i)\langle z_i, \sum_{t=1}^{T-1} h_t\rangle$. Hence, Axiom 3 is not satisfied unless $\beta = 0$, but the value of β is determined by the base learner f_S, not the boosting procedure g.

Second, consider the boosting procedure g that always selects $w_t(i) = w_1(i)$ and uses h_T as the final aggregated hypothesis. In other words, $\mathbb{H} = \mathcal{H}$ and the boosting procedure aggregates the hypotheses $(h_1, \ldots, h_T)$ by selecting h_T only. This is a boosting algorithm that trivially satisfies the OBJECTIVITY and UTILITY axioms, but not ADDITIVITY.

Finally, we show that ADDITIVITY and UTILITY can be satisfied without satisfying the OBJECTIVITY axiom. As will be proved later, AdaBoost satisfies the three axioms. Let g be the boosting procedure that coincides with adaBoost except that if $l_j(h) = l_k(h)$ for some $j, k \in \{1, \ldots, m\}$ and all $h \in \mathcal{H}$, then the boosting procedure always sets $w_t(j) = w_t(k)$ for all $t \geq 2$. In other words, it distributes the weight equally between the two loss functions if they are identical. More precisely, let $\hat{w}_t(j)$ and $\hat{w}_t(k)$ be the weights assigned by the adaBoost procedure at round t and let the corresponding weights assigned by the new boosting procedure g be $w_t(j) = w_t(k) = (\hat{w}_t(j) + \hat{w}_t(k))/2$. Then, g satisfies the ADDITIVITY and UTILITY axioms but without satisfying the OBJECTIVITY axiom because $w_t(k) \to 0$ as $w_1(k) \to 0$ only if $w_1(j) = 0$. Therefore, the three axioms are independent. □

4 Implications of the Axioms

Before we present our main uniqueness theorem, we elaborate on an important definition first. Suppose that in the fixed set of loss functions $S = \{l_1, \ldots, l_m\}$, two of those functions were, in fact, identical. That is, suppose that there exists $i, j \in \{1, \ldots, m\}$ with $i \neq j$ such that $l_i(\cdot) = l_j(\cdot)$. Then, it is clear that no algorithm can make a *meaningful* distinction between the weights $w_t(i)$ and $w_t(j)$ for all $t \geq 1$. Similar conclusions hold when a loss function $l_i(\cdot)$ can be written as a *linear combination* of the others. Therefore, linearly dependent loss functions pose an *inherent* source of ambiguity. This brings us to the following definition [28].

Definition 5 (Linear Independence). *A set of loss functions $\{l_1, \ldots, l_m\}$, with $l_i : \mathcal{H} \to \mathbb{R}$, are called linearly independent on $\mathcal{H}$ if and only if there exists $h_1, \ldots, h_m \in \mathcal{H}$ such that the column vectors $\boldsymbol{v}_j = (l_1(h_j), l_2(h_j), \ldots, l_m(h_j))^T$ for all $j = 1, \ldots, m$ are linearly independent.*

Informally, the set of loss functions are linearly independent if the training examples are sufficiently different. We analyze the implications of our axioms on linearly independent loss functions, next.

Lemma 1. *Let $S = \{l_1, \ldots, l_m\}$ comprises of m linearly independent functions on $\mathcal{H}$. Let $w_T = g_T(h_1, \ldots, h_{T-1})$ be the adaptive weights selected by the boosting procedure g at round T when $h_1, \ldots, h_{T-1}$ are the base hypotheses provided by the base learner $f_S : \mathcal{P}^m \to \mathcal{H}$ at all preceding rounds. Let $h_t = f_S(w_t)$ for all rounds $t \geq 1$. Then, the* ADDITIVITY, OBJECTIVITY, *and* UTILITY *axioms are satisfiable only if $w_T(i) = \zeta_i(\sum_{t=1}^{T-1} h_t)$ for some function $\zeta_i : \mathbf{Span}(\mathcal{H}) \to [0, 1]$.*

Proof. First, the statement trivially holds when $T = 1$. Suppose that $T > 1$. Then, by the UTILITY and ADDITIVITY axioms, we have:

$$\sum_{i=1}^{m} w_T(i)\, l_i(h_T) = c \cdot \sum_{i=1}^{m} w_1(i)\, l_i(\sum_{t=1}^{T} h_t), \tag{1}$$

with a proportionality constant c that is independent of h_T. Since $w_T(i)$ is independent of h_T and the loss functions in S are linearly independent, there exists m hypotheses $\tilde{h}_1, \tilde{h}_2, \ldots, \tilde{h}_m$ such that the matrix:

$$L = \begin{bmatrix} l_1(\tilde{h}_1) & l_2(\tilde{h}_1) & \cdots & l_m(\tilde{h}_1) \\ l_1(\tilde{h}_2) & l_2(\tilde{h}_2) & \cdots & l_m(\tilde{h}_2) \\ \cdots & \cdots & \cdots & \cdots \\ l_1(\tilde{h}_m) & l_2(\tilde{h}_m) & \cdots & l_m(\tilde{h}_m) \end{bmatrix}$$

is non-singular. Consequently, Eq. (1) implies that:

$$\begin{bmatrix} w_T(1) \\ w_T(2) \\ \cdots \\ w_T(m) \end{bmatrix} = c\, L^{-1} \cdot \begin{bmatrix} \sum_{i=1}^{m} w_1(i)\, l_i(\sum_{t=1}^{T-1} h_t \; + \hat{h}_1) \\ \sum_{i=1}^{m} w_1(i)\, l_i(\sum_{t=1}^{T-1} h_t \; + \hat{h}_2) \\ \cdots \\ \sum_{i=1}^{m} w_1(i)\, l_i(\sum_{t=1}^{T-1} h_t \; + \hat{h}_m) \end{bmatrix}$$

Therefore, the weights $w_T(i)$ are determined by the sequence of base hypotheses $(h_1, \ldots, h_{T-1})$ only via the aggregated rule $\sum_{t=1}^{T-1} h_t$. □

Lemma 1 shows that a boosting procedure that satisfies the three axioms is a Markov chain; the future of the boosting procedure is conditionally independent of the sequence of base hypotheses given the aggregated rule. Now, we are ready to state the main uniqueness theorem. To recall, a mapping M is called *additive* if it satisfies $M(x+y) = M(x) + M(y)$. For example, linear mappings are additive.

Theorem 1. *If $S = \{l_1, \ldots, l_m\}$ comprises of m linearly independent functions on $\mathcal{H}$ and $0 \in \mathcal{H}$, then the* ADDITIVITY, OBJECTIVITY, *and* UTILITY *axioms are satisfiable simultaneously for all initial distributions $w_1 \in \mathcal{P}^m$ if and only if the following two conditions hold:*

1. *We have $l_i(h) \propto \exp\{\Lambda_i(h) + \lambda_i\}$ for some additive mapping $\Lambda_i : \mathbf{Span}(\mathcal{H}) \rightarrow \mathbb{R}$ and some constant $\lambda_i \in \mathbb{R}$.*
2. *The weights are updated according to the multiplicative weight update procedure:*

$$w_T(i) \propto w_{T-1}(i) \cdot \exp\{\Lambda_i(h_{T-1})\}$$

Proof. [PROOF OF NECESSITY]: Lemma 1 implies that there exists some functions $\zeta_i : \mathbf{Span}(\mathcal{H}) \rightarrow [0, 1]$ such that:

$$w_T(i) = \zeta_i\big(\sum_{t=1}^{T-1} h_t\big), \tag{2}$$

for all $T \geq 1$. In other words, the weights at round T depend on the sequence of hypotheses $(h_1, \ldots, h_{T-1})$ only via their aggregated rule.

In addition, Axiom 3 states that for any $h \in \mathcal{H}$ and any round $T \geq 1$:

$$\sum_{i=1}^{m} w_T(i)\, l_i(h) \propto \sum_{i=1}^{m} w_1(i)\, l_i\big(h + \sum_{t=1}^{T-1} h_t\big) \tag{3}$$

However, $\sum_{t=1}^{T-1} h_t$ is arbitrary, so we denote it by $u \in \mathbf{Span}(\mathcal{H})$. Therefore, by (2), we conclude that for any $u \in \mathbf{Span}(\mathcal{H})$ and any $h \in \mathcal{H}$, the following equality must hold for some constant $c > 0$, which is independent of h:

$$\sum_{i=1}^{m} \zeta_i(u)\, l_i(h) = c \cdot \sum_{i=1}^{m} w_1(i) \cdot l_i\big(h + u\big) \tag{4}$$

Because c is independent of h and $0 \in \mathcal{H}$, we set $h = 0$ to conclude that:

$$c = \frac{\sum_{i=1}^{m} \zeta_i(u)\, l_i(0)}{\sum_{i=1}^{m} w_1(i) \cdot l_i(u))}$$

Hence, (4) may be rewritten as:

$$\frac{\sum_{i=1}^{m} \zeta_i(u)\, l_i(h)}{\sum_{i=1}^{m} \zeta_i(u)\, l_i(0)} = \frac{\sum_{i=1}^{m} w_1(i) \cdot l_i\big(h + u\big)}{\sum_{i=1}^{m} w_1(i) \cdot l_i(u)} \tag{5}$$

Since we require that the axioms hold simultaneously for all initial probability distributions $w_1 \in \mathcal{P}^m$, consider the following initial distribution:

$$w_1^{(i)}(k) = \begin{cases} 1-\epsilon, & \text{if } k = i \\ \frac{\epsilon}{m-1}, & \text{otherwise} \end{cases}$$

By Axiom 2, we know that $\zeta_i(u) \to 1$ as $\epsilon \to 0^+$. In the latter case, (5) reduces to the functional equation:

$$l_i(h+u) = \frac{l_i(h) \cdot l_i(u)}{l_i(0)} \tag{6}$$

Since we require the axioms to hold simultaneously for all initial weights, the above functional equation must hold as well for all loss functions $l_i(\cdot)$. Now, consider the function $\Lambda_i : \mathbf{Span}(\mathcal{H}) \to \mathbb{R}$ defined by $\Lambda_i(u) = \log l_i(u) - \log l_i(0)$. Then, $\Lambda_i(\cdot)$ satisfies $\Lambda_i(u+h) = \Lambda_i(u) + \Lambda_i(h)$. Hence, $\Lambda_i(\cdot)$ is an additive function on $\mathbf{Span}(\mathcal{H})$, which implies that $l_i(\cdot)$ must be equal to $\exp\{\Lambda_i(w) + \lambda_i\}$ for some additive mapping $\Lambda_i : \mathbf{Span}(\mathcal{H}) \to \mathbb{R}$ and some constant $\lambda_i \in \mathbb{R}$.

Now, (4) can be rewritten as:

$$\forall h \in \mathcal{H} : \sum_{i=1}^{m} \zeta_i(u)\, l_i(h) = c \cdot \sum_{i=1}^{m} \frac{w_1(i) \cdot l_i(u)}{l_i(0)}\, l_i(h), \tag{7}$$

where c does not depend on h. This defines a system of linear equations on $\zeta_i(u)$ for different choices of $h \in \mathcal{H}$. Because the set of loss functions are linearly independent on $\mathcal{H}$, the above condition is satisfiable if and only if $\forall i \in \{1, \ldots, m\} : \zeta_i(u) = c\, w_1(i) \cdot l_i(u)/l_i(0)$. However, $\zeta_i(u)$ is a probability distribution so c is absorbed in the normalization constant, which we can ignore. We have:

$$\zeta_i(u) \propto w_1(i) \cdot \frac{l_i(u)}{l_i(0)} \tag{8}$$

$$= w_1(i)\, e^{\Lambda_i(h_1)} \prod_{t=2}^{T-1} \exp\{\Lambda_i(h_t)\} \;\propto\; w_{T-1} \cdot \exp\{\Lambda_i(h_t)\}, \tag{9}$$

where the last line holds by induction and the fact that $\Lambda_i(\cdot)$ is an additive function. This proves that the conditions are necessary.

[Proof of Sufficiency]: Next, we prove that the conditions are sufficient. First, from the multiplicative weight update mechanism, it is clear that Axiom 2 is satisfied. Moreover, (9) shows that the weights can be determined at any round T using only the aggregated rule $\sum_{t=1}^{T-1} h_t$. In particular, we have:

$$w_T(i) \;\propto\; w_1(i) \cdot e^{\Lambda_i(h_1)} \prod_{t=2}^{T-1} \exp\{\Lambda_i(h_t)\} \;=\; w_1(i) \cdot \exp\left\{\Lambda_i\Big(\sum_{t=1}^{T-1} h_t\Big)\right\}$$

Hence, Axiom 1 is satisfied. Finally, by plugging the functional equation in (6) and the expression in (8) into (5), we deduce that Axiom 3 is satisfied as well. Therefore, the conditions are also sufficient for the three axioms to hold. □

Theorem 1 reveals an axiomatic characterization of adaBoost and related algorithms, such as the extension of adaBoost to confidence-rated predictions [30], the RankBoost algorithm [13], the Real-AdaBoost for probabilistic classifiers [15], and the Multi-class AdaBoost method [38]. In particular, the ADDITIVITY, OBJECTIVITY and UTILITY axioms are satisfied if and only if the loss functions were of the exponential type and the weights were updated according to the multiplicative weight update mechanism. Therefore, even though many possible adaptive re-weighting methods could be (and have been) proposed, such as the methods studied in [5,19,20], the adaptive re-weighting method employed by adaBoost and its variants can be *uniquely* constructed axiomatically. This sheds some insight on the rich set of connections that have been established between adaBoost and related fields, such as information geometry, game theory, and convex optimization [29][4].

As mentioned earlier, the ADDITIVITY, OBJECTIVITY, and UTILITY axioms can also serve as axioms for the more general Multiplicative Weight Update procedure. This follows from the fact that the axioms are satisfied if and only if the multiplicative weight update procedure is used without imposing any additional conditions on the base learners. That is, the base learners are entirely arbitrary.

5 Concluding Remarks

Boosting procedures, which operate via adaptive re-weighting and combining, can vary according to the choice of the loss function they minimalize and how they adaptively update the weights. Not surprisingly, different algorithms for adaptive re-weighting have been proposed in the literature, such as the arc-x4 algorithm [5], its generalization to higher order polynomials [20], and the partitioning scheme in [19]. This raises the fundamental questions: What natural axioms should an adaptive re-weighting and combining procedure satisfy? And, do such axioms lead to a *unique* solution?

In this work, we address these questions. We establish that three natural axioms on boosting algorithms are satisfied *if and only if* the boosting algorithm minimalizes the sum of exponential-additive loss functions and the weights are updated according to the multiplicative weight update procedure. Surprisingly, this conclusion holds even though (1) we operate in the general setting of learning, which encompasses regression, classification, ranking, and clustering, and (2) we do not impose any constraints on the base learner and the base hypotheses.

The fact that the loss functions have to be of the form specified in Theorem 1 might appear to be overly restrictive at first sight. For instance, it is not immediately obvious how one might define a loss function for regression tasks, while also being a composition of the exponential with additive functions. However,

[4] The assumption of linear independence in Theorem 1 is required to eliminate an *inherent* source of ambiguity. Without this assumption, no uniqueness theorem can be established. However, this result does not imply that a boosting algorithm must guarantee linear independence. Rather, it states that *up to this inherent source of ambiguity*, adaBoost and its variants arise uniquely out of three natural axioms.

this function class is by no means restrictive. For instance, if the hypothesis space is a subset of $\mathbb{R}^d$, then the Fourier theorem states that *any* desired function in a compact domain can be approximated arbitrarily well using a sum of exponential-additive functions. Hence, the general class of exponential-additive functions is quite rich. Indeed, many extensions of adaBoost have been proposed that also satisfy the EFFICIENCY, STABILITY and LINEAR UTILITY axioms, including the extension of adaBoost to confidence-rated predictions [30], the RankBoost algorithm [13], the Real-AdaBoost for probabilistic classifiers [15], and the Multi-class AdaBoost method [38].

References

1. Ackerman, M., Ben-David, S.: Measures of clustering quality: a working set of axioms for clustering. In: NIPS, pp. 121–128 (2009)
2. Aczél, J., Forte, B., Ng, C.T.: Why the shannon and hartley entropies are natural. Adv. Appl. Probab. **6**(01), 131–146 (1974)
3. Bell, D.A., Wang, H.: A formalism for relevance and its application in feature subset selection. Mach. Learn. **41**(2), 175–195 (2000)
4. Bousquet, O., Elisseeff, A.: Stability and generalization. J. Mach. Learn. Res. **2**, 499–526 (2002)
5. Breiman, L.: Prediction games and arcing algorithms. Neural Comput. **11**(7), 1493–1517 (1999)
6. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
7. Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A.: Classification and Regression Trees. CRC Press, Boca Raton (1984)
8. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**(3), 273–297 (1995)
9. Cox, R.T.: Probability, frequency and reasonable expectation. Am. J. Phys. **14**(1), 1–13 (1946)
10. Csiszar, I.: Why least squares and maximum entropy? an axiomatic approach to inference for linear inverse problems. Ann. Statist. **19**, 2032–2066 (1991)
11. Csiszár, I.: Axiomatic characterizations of information measures. Entropy **10**(3), 261–273 (2008)
12. Džeroski, S., Ženko, B.: Is combining classifiers with stacking better than selecting the best one? Mach. Learn. **54**(3), 255–273 (2004)
13. Freund, Y., Iyer, R., Schapire, R.E., Singer, Y.: An efficient boosting algorithm for combining preferences. JMLR **4**, 933–969 (2003)
14. Freund, Y., Schapire, R.E.: A desicion-theoretic generalization of on-line learning and an application to boosting. In: Vitányi, P. (ed.) EuroCOLT 1995. LNCS, vol. 904, pp. 23–37. Springer, Heidelberg (1995). https://doi.org/10.1007/3-540-59119-2_166
15. Friedman, J., Hastie, T., Tibshirani, R., et al.: Additive logistic regression: a statistical view of boosting. Ann. Stat. **28**(2), 337–407 (2000)
16. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Ann. Stat. **29**, 1189–1232 (2001)
17. Jardine, N., Sibson, R.: The construction of hierarchic and non-hierarchic classifications. Comput. J. **11**(2), 177–184 (1968)
18. Jaynes, E.T.: Probability theory: The Logic of science. Cambridge University Press, Cambridge (2003)

19. Ji, C., Ma, S.: Combined weak classifiers. NIPS **9**, 494–500 (1997)
20. Khanchel, R., Limam, M.: Empirical comparison of arcing algorithms (2005)
21. Kleinberg, J.: An impossibility theorem for clustering. In: NIPS, vol. 15, pp. 463–470 (2002)
22. Lee, P.: On the axioms of information theory. Ann. Math. Stat. **35**(1), 415–418 (1964)
23. Mason, L., Baxter, J., Bartlett, P.L., Frean, M.R.: Boosting algorithms as gradient descent. In: NIPS, pp. 512–518 (1999)
24. Österreicher, F.: Csiszár's f-divergences-basic properties. Technical report (2002)
25. Pennock, D.M., Horvitz, E.: Analysis of the axiomatic foundations of collaborative filtering. Ann Arbor **1001**, 48109–2110 (1999)
26. Pennock, D.M., Maynard-Reid II, P., Giles, C.L., Horvitz, E.: A normative examination of ensemble learning algorithms. In: ICML, pp. 735–742 (2000)
27. Prasad, A., Pareek, H.H., Ravikumar, P.: Distributional rank aggregation, and anaxiomatic analysis. In: ICML, pp. 2104–2112 (2015)
28. Sansone, G.: Orthogonal Functions. Dover Publications, New York (1991)
29. Schapire, R.E., Freund, Y.: Boosting: Foundations and Algorithms. MIT Press, Cambridge (2012)
30. Schapire, R.E., Singer, Y.: Improved boosting algorithms using confidence-rated predictions. Mach. Learn. **37**(3), 297–336 (1999)
31. Servedio, R.A.: Smooth boosting and learning with malicious noise. J. Mach. Learn. Res. (JMLR) **4**, 633–648 (2003)
32. Shalev-Shwartz, S., Shamir, O., Srebro, N., Sridharan, K.: Learnability, stability and uniform convergence. J. Mach. Learn. Res. (JMLR) **11**, 2635–2670 (2010)
33. Shannon, C.: A mathematical theory of communication. Bell Syst. Tech. J. **27**, 379–423 (1948)
34. Shore, J., Johnson, R.: Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. IEEE Trans. Inf. Theory **26**(1), 26–37 (1980)
35. Skilling, J.: The axioms of maximum entropy. In: Erickson, G.J., Smith, C.R. (eds.) Maximum-Entropy and Bayesian Methods in Science and Engineering, pp. 173–187. Springer, Dordrecht (1988). https://doi.org/10.1007/978-94-009-3049-0_8
36. Vapnik, V.N.: An overview of statistical learning theory. IEEE Trans. Neural Netw. **10**(5), 988–999 (1999)
37. Wang, Z., et al.: Multi-class hingeboost. Methods Inf. Med. **51**(2), 162–167 (2012)
38. Zhu, J., Zou, H., Rosset, S., Hastie, T.: Multi-class adaboost. Stat. Interface **2**(3), 349–360 (2009)
39. Zou, H., Zhu, J., Hastie, T.: New multicategory boosting algorithms based on multicategory fisher-consistent losses. Ann. Appl. Stat. **2**(4), 1290 (2008)