ECAI 2025
I. Lynce et al. (Eds.)

doi:10.3233/FAIA251422

An Evidence-Theoretic Framework for Online Learning from Expert Advice

Andrea Campagner [a,*], **Francesca Arredondo**[b], **Davide Ciucci** [b] **and Federico Cabitza** [a,b]

[a]IRCCS Ospedale Galeazzi Sant'Ambrogio, Milan, Italy
[b]University of Milano-Bicocca, Milan, Italy
ORCID (Andrea Campagner): https://orcid.org/0000-0002-0027-5157, ORCID (Davide Ciucci): https://orcid.org/0000-0002-8083-7809, ORCID (Federico Cabitza): https://orcid.org/0000-0002-4065-3415

Abstract. The use of belief function theory (BFT) in machine learning has gained attention as researchers seek more principled foundations for decision-making in uncertain environments. However, research has mostly focused on the setting of batch learning. In this article, in contrast and to our knowledge for the first time in the literature, we study the application of BFT to the setting of online (machine) learning. Within this context, online learning from expert advice (LEA) offers a framework where learners iteratively update their predictions based on experts' input and (adversarially labeled) observed outcomes. Despite extensive study and strong theoretical results, the epistemological underpinnings of LEA remain largely heuristic. This work addresses this gap by proposing belief function theory (BFT) as a formal foundation for LEA. Here we report a theoretical and algorithmic integration of BFT into LEA, showing that classical LEA algorithms such as Halving and Weighted Majority can be derived as special cases of evidential reasoning. We further introduce two novel LEA algorithms—Evidential Halving and Evidential Weighted Majority—which fully exploit BFT and support cautious prediction through abstention. These new algorithms demonstrate improved regret bounds over traditional methods, under mild assumptions. These findings open a new direction in online learning by leveraging the full expressive power of BFT to design theoretically grounded algorithms.

1 Introduction

Online learning encompasses a class of machine learning problems in which data are not provided in a single batch but are instead presented as a sequential data stream [32]. In this setting, the learner maintains a *state*—for example, a probability distribution over different models—and receives data instances one at a time. At each time step, the learner makes a prediction, after which the true target value is revealed and used to assess the prediction. This feedback is then employed to update the learner's state [41].

As a relevant framework in online learning, in the *learning from expert advice* (LEA) framework [20, 38, 52], the learner generates predictions based on the classifications provided by a set of *experts*. In this setting, the learner's state is typically represented by a set of weights—often interpreted as probabilities—assigned to the experts. Both the input data and corresponding target values are viewed as *evidence* about which expert performs best. The learner's objective is to incorporate this evidence into its state in order to minimize *regret* relative to the (unknown) best expert, with no assumption about the input stream of data (so-called adversarial setting).

From the point of view of computational learning theory, the LEA problem is well understood [8, 41]: indeed, several algorithms have been proposed to solve this problem [30, 41], and tight computational bounds are known [10, 54], as well as connections with other fields such as bandit theory [3], convex optimization [30], game theory [8] and finance [49]. However, despite recent advances in the field, particularly those stemming from research on game-theoretic probability [45], the epistemological foundations of the LEA setting remain insufficiently defined. In particular, the concept of evidence—and the principles by which it should inform updates to the learner's state—has not been fully formalized. Practical algorithms generally rely on intuitive interpretations of these notions, with their design guided largely by heuristic principles often inspired by convex optimization [47] that, despite being proven highly effective in practice, do not provide a unified conceptual foundation of the field.

Conversely, the notion of evidence has been more explicitly studied in *belief function theory* (BFT) (also called *evidence theory* and *Dempster-Shafer theory*), originally proposed by Dempster [17] in the context of statistical inference, and then formalized by Shafer in [44] as well as by other authors in subsequent work [16, 21, 40, 48]. BFT can be understood as a generalization of probability theory (as well as possibility theory [16], and as a special case of imprecise probability [35]), in which evidence is represented as a *mass function*: a probability distribution defined on the power set of a given universe of interest (representing possible answers to a question), where the mass assigned to a set quantifies the amount of available evidence that the true answer lies in that set (and nothing more). BFT then provides mathematical tools for reasoning about evidence (so-called evidential reasoning), including methods to combine different sources of evidence [18, 22, 53], as well as to express evidence on a given universe of discourse at different granularity levels [44].

Despite the notion of evidence being central in both LEA (implicitly) and BFT (explicitly), up to our knowledge, no previous work has explored the connections between these two fields of artificial intelligence [4]. In this article[1], we address this gap by proposing BFT as a theoretical foundation for the LEA problem. Focusing on the setting of LEA with a finite set of experts and binary learning

* Email:andrea.campagner@unimib.it

[1] For complete proofs, as well as additional material, we refer the reader to the technical appendix [7].

problems, we provide two main contributions. First, we show how classical algorithms proposed in the LEA literature can be understood as special cases of evidential reasoning and, hence, naturally emerge from the application of BFT as a way to model the problem of online learning. Second, we introduce the application of BFT as a reasoning framework for devising novel LEA algorithms. To this end, we first illustrate how BFT provides a natural setting to model *cautious learners* that can abstain from providing predictions [13], thus drawing a connection with selective prediction [25, 6], and then use the introduced setup to propose two novel such algorithms, showing that, under certain conditions, they can provide better performance than any classical (i.e., non-cautious) LEA algorithm.

2 Background

Let X be a feature space and Y a target space. We focus on binary classification problems, hence we assume that $Y = \{0, 1\}$. For each $t \in \mathbb{N}$, let $Z^t = (X \times Y)^t$ and $Z^* = \cup_{t\in\mathbb{N}} Z^t$: we denote with $z_t = ((x_1, y_1), \ldots, (x_t, y_t))$ a sequence in Z^t. We assume the existence of a set of experts $\mathcal{H}$: an expert is a (possibly probabilistic) algorithm taking input from $Z^* \times X$ and returning an output in Y. We assume that $\mathcal{H}$ is finite. A learning algorithm A is a (possibly probabilistic) algorithm taking as input a set of experts $\mathcal{H}$, a sequence in $Z^* \times X$ and returning an output in Y. A loss function is a function $l : Y \times Y \to \mathbb{R}$. In the LEA problem, we assume that a learning algorithm A receives data according to a sequential protocol, as illustrated in Algorithm 1. The aim of the LEA problem is to find an algorithm that minimizes the regret (as defined in Algorithm 1):

$$Regret = \sum_{t=1}^{T} l(y_t, A(\mathcal{H}, z^t \cup (x_t))) - \min_{h\in\mathcal{H}} \sum_{t=1}^{T} l(y_t, h(x_t)),$$

as a function of the time horizon T (i.e., the maximum sequence length observed by the learning algorithm, see Algorithm 1), without any assumption (adversarial setting) on the sequence of incoming data. If we assume that the incoming data stream is such that the condition $\min_{h\in\mathcal{H}} Err(h) = 0$, then we say the LEA problem is *realizable*. It is a well-known result from LEA theory [9, 26] that, under general assumption, one typically aims for learning algorithms A that achieve a regret on the order (up to logarithmic terms[2]) of $C\sqrt{T\log(|\mathcal{H}|)}$, where C is a universal constant: in fact there exist simple algorithms (e.g., Weighted Majority [38], Hedge [27] and other methods based on multiplicative weight update) that achieve the mentioned bound. For additional details on online learning, we refer the interested reader to [8, 30, 41].

Example 2.1. *As a simple example of a LEA problem, we describe an instantiation of the Halving algorithm (see Section 3). Let $X = \{0, 1, \ldots, 9\}$ and $Y = \{0, 1\}$. Assume that $\mathcal{H} = Y^X$, that is, the collection of all (deterministic) functions from X to Y. Additionally, assume $l(y, y') = \mathbb{1}_{y\neq y'}$ is the 0-1 loss. Given a sequence z_t, let $\mathcal{H}_t = \{h \in \mathcal{H} : \forall(x_i, y_i) \in z_t, h(x_i) = y_i\}$. Further, given an instance x, define $\mathcal{H}_0 = \{h \in \mathcal{H}_t : h(x) = 0\}$ and $\mathcal{H}_1 = \{h \in \mathcal{H}_t : h(x) = 1\}$. Let A be the learning algorithm defined by:*

$$A(\mathcal{H}, z_t, x) = \begin{cases} 1 & |\mathcal{H}_1| \geq |\mathcal{H}_0| \\ 0 & otherwise. \end{cases}$$

Intuitively, A always selects the label that has been predicted by a majority of experts, considering only those experts that have never made an incorrect prediction on the observed sequence. Assuming realizability (i.e., $\exists h \in \mathcal{H}$ s.t. $\sum_{t=1}^{T} l(y_t, h(x_t)) = 0$), the worst-case regret of A can be easily seen to be at most $\log(|\mathcal{H}|)$, since at each mistake the set $\mathcal{H}_t$ is halved.

In the same setting as above, we define a *cautious predictor* as a (possibly probabilistic) algorithm taking input from $Z^* \times X$ and returning an output in $Y \cup \{\perp\}$, where the symbol $\perp$ is interpreted as *abstaining from prediction*: that is, a cautious predictor can sometimes refrain from providing a prediction. Cautious predictors have been first proposed in [13] and studied mostly in the context of offline learning [2, 6, 25], while only more recently their application has been extended to the setting of online learning [2, 12, 15][3]. Cortes et al. [15], in particular, focus on the setting where each expert in $\mathcal{H}$ is a cautious predictor and study LEA algorithms in this setting, proving the existence of algorithms that can achieve the above-mentioned regret bound of $C\sqrt{T\log(|\mathcal{H}|)}$: nonetheless, the authors mostly focus on the stochastic (rather than adversarial) setting. By contrast, Cheung et al [12] show that, in the same setting as [15] but assuming realizability, there exist problems for which allowing cautious predictors results in a finite regret bound, but the regret of any *disambiguation*[4] of $\mathcal{H}$ cannot be bounded from above, showing that online learning with cautious predictors is intrinsically more powerful than its classical counterpart.

Example 2.2. *We present a simplified example of LEA with cautious predictors, in which we assume that the experts are as in Example 2.1, but the algorithm can select its predictions from the set $Y \cup \{\perp\}$. Specifically, assume that A is defined as:*

$$A(\mathcal{H}, z_t, x) = \begin{cases} y' & \forall h \in \mathcal{H}_t, h(x) = y' \\ \perp & otherwise. \end{cases}$$

Thus, A abstains from providing a prediction whenever there is a disagreement among the valid experts (i.e., the experts that did not make any mistake in labeling the sequence z_t). Assume that A is penalized according to the 0-1 loss with abstention, defined as:

$$l(y, y') = \begin{cases} \alpha & y' = \perp \\ \mathbb{1}_{y\neq y'} & otherwise. \end{cases}$$

Under the same assumptions as in Example 2.1, the worst-case regret of A is $\alpha(|\mathcal{H}| - 1)$, as the algorithm abstains from making a prediction whenever there is a possibility for error (which, in turn, implies that at least one expert's advice was incorrect).

We now turn our attention to evidence theory. Evidence theory is an uncertainty representation formalism that generalizes probability theory, providing a general framework to represent epistemic uncertainty and evidence, as well as tools to combine evidence sources. Formally, let Ω be a reference set. A *mass function* is a function $m : 2^\Omega \to [0, 1]$ such that $\sum_{A\subseteq\Omega} m(A) = 1$b. The *focal sets* of m are the sets in the support of m, that is $\mathcal{F}(m) = \{A \subseteq \Omega : m(A) \neq 0\}$. If $\emptyset \notin \mathcal{F}(m)$ we say that m is *normalized*. A normalized mass function represents the evidence about the possible answers to some question of interest: in particular, the value $m(A)$ represents the evidence that the correct answer lies in $A \subseteq \Omega$ and nothing more.

[2] The restriction of "up to logarithmic terms" is necessary since one can, in certain cases, obtain bounds of the form $C\sqrt{TL(\mathcal{H})}$, where $L(\mathcal{H}) \leq \log(|\mathcal{H}|)$ is the Littlestone dimension of $\mathcal{H}$ [37].

[3] Related models have been previously proposed in the literature, e.g. the sleeping experts model [28], the selective sampling model [11], and the Knows-What-It-Knows model [36], however these models differ from the notion of cautious prediction we consider in this article.

[4] Intuitively, a disambiguation of a cautious predictor is a classical algorithm that replaces $\perp$ with either 0 or 1. See [2, 12] for a formal definition.

Algorithm 1 General scheme of a learning from expert advice algorithm .

$z = []$
$Regret = 0$
For each $h \in \mathcal{H}$, $Err(h) = 0$
for t in $\{1, \dots, T\}$ **do**
  Receive $x \in X$
  Predict $y' := A(\mathcal{H}, z.append(x))$
  Receive $y \in Y$
  For each $h \in \mathcal{H}$, $Err(h)$ += $l(y, h(z.append((x, y))$
  $Regret$ += $l(y, y')$
  $z = z.append((x, y))$
end for
$Regret$ −= $\min_{h \in \mathcal{H}} Err(h)$

Given m, one can define a *belief function* $Bel(A) = \sum_{B \subseteq A} m(B)$, a *plausibility function* $Pl(A) = \sum_{B \cap A \neq \emptyset} m(B)$ and a *commonality function* $Q(A) = \sum_{A \subseteq B} m(B)$. The belief $Bel(A)$ can be interpreted as measuring the evidence supporting the claim that the correct answer is in A, or, equivalently, the probability that the correct answer necessarily lies in A [24]. Similarly, the plausibility $Pl(A)$ can be interpreted as a measure of the lack of support for the complement of A, or, equivalently, as the probability that the correct answer may possibly lie in A.

As mentioned in the introduction, evidence theory can be seen as a generalization of other uncertainty representation formalisms, chiefly among them probability and possibility theory. In the first case, a mass function m is *Bayesian* if all its focal sets are singletons, i.e., $\forall A \in \mathcal{F}(m), |A| = 1$: in this case, for each $A \subseteq \Omega$, $Bel(A) = Pl(A)$ and m is equivalent to a *probability distribution*. As for the case of possibility theory, the *contour function* $pl_m : \Omega \to [0, 1]$ of a mass function m is defined as $pl_m(x) = Pl(\{x\}) = Q(\{x\})$ for each $x \in \Omega$: irrespective of m, the corresponding countour function pl_m always defines a *possibility distribution* over Ω [21, 44].

This connection between evidence theory, on the one hand, and probability and possibility theory, on the other hand, is particularly relevant in the context of decision-making, as one of the most commonly adopted approaches to derive a decision theory based on evidence theory is based on defining a way to transform a mass function into a decisional probability distribution [19]. In this sense, we define the *plausibility transform* [14, 51] of m as the Bayesian mass function t_m defined by $t_m(x) = \frac{pl_m(x)}{\sum_{x \in X} pl_m(x)}$, for each $x \in X$. The *pignistic transform* [48] of m is the Bayesian mass function $Bet(m)$ defined by $Bet(m)(x) = \sum_{A \subseteq \Omega : x \in A} \frac{m(A)}{|A|}$. A mass function m is *simple* if it is in the form, for $s \in [0, 1]$:

$$m_C(A) = \begin{cases} 1 - s & A = C \neq \emptyset \\ s & A = \Omega \\ 0 & \text{otherwise.} \end{cases}$$

Example 2.3. *We illustate the above abstract definition of evidence theory through an example originally discussed by Shafer in [44]. Assume we need to assess whether a vase, presented as a product of the Ming dynasty, is genuine (G) or counterfeit (C). In this case, $\Omega = \{G, C\}$. To decide upon the genuinity of the vase, we involve a rater that provides their judgment: assume it is G (so, the rater believes the vase is genuine). Given that the rater could be wrong in their assessment, we can represent our uncertainty about the genuinity of the vase as a (simple) mass function $m(\{G\}) = p$, $m(\Omega) = 1 - p$, where p represents our subjective belief that the rater is correct in its judgment (in which case the vase is genuine): for example, this could be measured as the previous accuracy of the rater in assessing other similar items. In this case, we cannot exclude that the judgment of the rater may be not sufficiently convincing, in which case we have no information about the genuinity of the vase: this is represented by the evidence mass $1 - p$ assigned to the vacuous event $\Omega = \{G, C\}$.*

In this case, the belief and plausibility of all subsets of Ω can be easily calculated as $Bel(\{G\}) = p, Pl(\{G\}) = 1$, $Bel(\{C\}) = 0, Pl(\{C\}) = 1 - p$ and $Bel(\Omega) = Pl(\Omega) = 1$. If asked to take a decision about the genuinity of the vase, we could convert our evidence assessment into a probability distribution by using either the pignistic or the plausibility transform. In the first case, we obtain $Bet(m)(G) = p + \frac{1-p}{2}$ and $Bet(m)(C) = \frac{1-p}{2}$. In the second case, $t_m(G) = \frac{1}{2-p}$ and $t_m(C) = \frac{1-p}{2-p}$. It is easy to observe that, in both cases, the assignment of probability would be different from what we could have obtained had we directly adopted a probabilistic model of our uncertainty in which $Pr(G) = p$ and $Pr(C) = 1 - p$.

Given two mass functions m_1, m_2, we define their *unnormalized combination* as the mass function $m_1 \cap m_2$ defined by $m_1 \cap m_2(A) = \sum_{B \cap C = A} m_1(B) m_2(C)$: it can be shown that it holds that $Q_{m_1 \cap m_2}(A) = Q_{m_1}(A) Q_{m_2}(A)$. The *degree of conflict* between m_1, m_2 is defined as $\mathcal{K}_U = m_1 \cap m_2(\emptyset)$. Given two mass functions such that $\mathcal{K}_U \neq 1$, we define their Dempster combination (or, orthogonal sum) as the mass function $m_\oplus$ s.t. $m_\oplus(\emptyset) = 0$ and $m_\oplus(A) = \frac{m_1 \cap m_2(A)}{1 - \mathcal{K}_U}$. Let Ω_1, Ω_2 two sets s.t. $|\Omega_1| < |\Omega_2|$ and let $\rho : \Omega_2 \to \Omega_1$ be a surjective function: we say that Ω_2 is a *refinement* of Ω_1 and, symmetrically, Ω_1 is a *coarsening* of Ω_2. Given a mass function m on Ω_1, we call the mass function $m^{\Omega_1 \uparrow \Omega_2}$ on Ω_2 defined by $m^{\Omega_1 \uparrow \Omega_2}(\rho^{-1}(A)) = m(A)$, the *vacuous extensions* of m induced by ρ. Symmetrically, given a mass function m on Ω_2, we call the mass function $m^{\Omega_2 \downarrow \Omega_1}$ on Ω_1 defined by $m^{\Omega_2 \downarrow \Omega_1}(A) = \sum_{B \subseteq \Omega_2 : \{\rho(b) : b \in B\} = A} m(B)$, the *restriction* of m. For additional details on BFT, we refer the reader to [16, 44].

Example 2.4. *To continue on Example 2.3, assume that we also involved a second rater. We are absolutely certain of the competence of this rater. However, the rater is unsure about the genuinity of the vase, and provides a judgment in the form of a Bayesian mass function (hence, a probability distribution) $m(\{G\}) = a, m(\{C\}) = 1 - a$. We can measure our evidence concerning the genuinity of the vase by combining the two mass function. Adopting the Dempster combination rule, we obtain the mass function m_D defined by $m_D(\{G\}) = \frac{pa + (1-p)a}{1 - p(1-a)}, m_D(\{C\}) = \frac{(1-p)(1-a)}{1 - p(1-a)}$: it is easy to see that m_D is a Bayesian mass function.*

3 Methods

In this section we will study how the LEA problem can be seen as an instance of the evidential reasoning expressed in BFT. In particular, we first show how two optimal (up to logarithmic terms) LEA algorithms can be derived as restricted forms of evidential reasoning; then, we propose two new LEA algorithms directly inspired by evidential reasoning, while also studying their properties. In the following, we will reformulate the LEA problem as an evidential reasoning protocol following the general scheme defined by Algorithm 2, where $\mathcal{H}$ is the set of experts, and we interpret the prediction $\{0, 1\}$ as equivalent to an abstention (i.e., the symbol $\perp$, see Section 2).

As hinted at in the Introduction, the protocol defined in Algorithm 2 provides a natural setting to describe cautious prediction methods in the LEA setting. Nonetheless, we note that this approach is

Algorithm 2 General evidential reasoning protocol for describing online learning algorithms .

The learner defines a mass function m over $\mathcal{H}$
for $t \in \{1, \dots, T\}$ **do**
  The learner receives $x \in X$
  The learner predicts $y' \in Y \cup \{\{0,1\}\}$ based on m
  The learner receives a target y and a mass function u over Y
  The learner is penalized based on a loss function $l(y, y')$
  The learner updates m based on u
end for

not entirely equivalent to that assumed in [12, 15]. While in the latter case each expert is assumed to be a cautious predictor, in Algorithm 2, experts in $\mathcal{H}$ are *classical* binary predictors. In contrast, abstentions in Algorithm 2 derive from identifying *sets of experts* as cautious predictors: that is, given $A \subseteq \mathcal{H}$ and $x \in X$, we define $A(x) = \{y \in Y : \exists h \in A, h(x) = y\}$. In this sense, Algorithm 2 can be understood as a special case of the setting described in [12, 15], where each cautious predictor corresponds to a set of classical experts, similarly to the *version space*-based selective prediction framework proposed in [25]. Differently from this latter approach, however, our evidential framework is based on evidence theory, which is a generalization of version space theory [23, 34], and is applied to the online learning, rather than batch learning, setting.

In the following, with the notation $m^{\mathcal{H}\downarrow_x\{0,1\}}$ we denote the restriction of a mass function m, originally defined on the set of experts $\mathcal{H}$, to the set Y obtained by means of the surjective function $\rho_x : \mathcal{H} \to Y$, defined by $\rho_x(h) = h(x)$. Thus, intuitively, $m^{\mathcal{H}\downarrow_x\{0,1\}}$ represents the projection of the expert set $\mathcal{H}$ to the set of possible predictions for instance x.

3.1 Learning from Expert Advice Algorithms as Restricted Evidential Reasoning

In this first section, we will show that two classical algorithms, namely *Halving* [46] and *Weighted Majority* [38], can be derived as special cases of the evidential reasoning protocol described in Algorithm 2. We focus on these two algorithms since they provide optimal regret bounds (up to logarithmic terms) in the realizable and agnostic (i.e., non-realizable) settings [1], respectively. As we show in this section, these two algorithms represent approximations of the fully evidential approach described in Algorithm 2, where evidence is restricted to, respectively, probabilistic and possibilistic terms.

Halving The Halving algorithm can be derived as a special case of Algorithm 2 where the evidence about the optimal expert is represented as a (uniform) *probability distribution* and each new instance (x, y) provides a way to update the evidence obtained so far by *conditioning*, that is, by combining a Bayesian mass function with a binary mass function: a binary mass function is employed since, under the assumption of realizability, the optimal expert cannot make an incorrect prediction, and hence any mistake committed by an expert constitutes evidence that this latter cannot be the optimal one. This idea is formalized in Algorithm 3, which we show to be equivalent to the Halving algorithm in Theorem 3.1.

Theorem 3.1. *Let $l_{0-1}(y, y') = \mathbb{1}_{y \neq y'}$ be the $0-1$ loss function. Then, Algorithm 3 is equivalent to the Halving algorithm: in particular, the two algorithms always make the same predictions. Consequently, under realizability, Algorithm 3 enjoys $O(\log_2(|\mathcal{H}|))$ regret.*

Proof. See Appendix [7]. □

Algorithm 3 Evidential reasoning protocol for the Halving algorithm

For each $h \in \mathcal{H}$, $m^1(\{h\}) = \frac{1}{|\mathcal{H}|}$
for $t \in \{1, \dots, T\}$ **do**
  Receive x
  Compute $m^{\mathcal{H}\downarrow_x Y}$, restriction of m^t
  Predict $y' = \arg\max_{y \in \{0,1\}} m^{\mathcal{H}\downarrow_x\{0,1\}}(y)$
  Receive y and construct $m^y := \begin{cases} 1 & \{h \in \mathcal{H} : h(x) = y\} \\ 0 & \text{otherwise} \end{cases}$
  Penalize learner based on $l_{0-1}(y, y')$
  $w = m^{\mathcal{H}\downarrow_x Y} \oplus m^y$
  Compute $w^{Y\uparrow\mathcal{H}}$
  $m^{t+1} = Bet(w)$
end for

Weighted Majority Similarly to the Halving algorithm, also the Weighted Majority algorithm can be derived as a special case of Algorithm 2. In this case, however and as shown in Algorithm 4, evidence about the optimal expert is represented in terms of a (initially vacuous) mass function (rather than a probability distribution), and each new instance provides a way to update the available evidence based on a *simple mass function* using Dempster rule of combination. In this case, a simple mass function is employed since, in the agnostic setting, even the best expert can issue mistaken predictions: therefore, the evidence of any expert being the optimal one can never be fully discounted, as represented by the mass assigned to the full expert set $\mathcal{H}$. Interestingly, while Algorithm 4 maintains a mass function m^t defined over the set of experts $\mathcal{H}$, each prediction is made solely based on the corresponding contour function pl_{m^t}, which is a *possibility distribution* over $\mathcal{H}$. As we prove the equivalence of Weighted Majority and Algorithm 4 in Theorem 3.2, this implies that similarly as to how Halving can be seen as probabilistic approximation of the evidential reasoning protocol in Algorithm 2, Weighted Majority can instead be seen as a *possibilistic* approximation.

Algorithm 4 Evidential reasoning protocol for the Weighted Majority algorithm

Set $m^1(\mathcal{H}) = 1$
Set $\eta \leq \sqrt{\frac{2\log(|\mathcal{H}|)}{T}}$
for $t \in \{1, \dots, T\}$ **do**
  Receive x
  Compute t^m, plausibility transform of m^t
  Select $h \in \mathcal{H}$ randomly according to t^m
  Predict $y' := h(x)$
  Receive y
  Penalize learner based on $l(y, y')$
  Construct $m^y(A) := \begin{cases} s = e^{-\eta} & A = \mathcal{H} \\ 1 - s & A = \{h \in \mathcal{H} : h(x) = y\} \\ 0 & \text{otherwise} \end{cases}$
  $m^{t+1} = m^t \cap m^y$
end for

Theorem 3.2. *Assume that the loss function l satisfies $l(y, y') \in [0, 1]$, for any $y, y' \in Y$. Then, Algorithm 4 is equivalent to the Weighted Majority algorithm, in the sense that, at each time step $t = \{1, \dots, T\}$, the two algorithms will select a prediction $y' \in \{0, 1\}$ according to the same probability distribution. Consequently, Algorithm 4 enjoys a regret bound of $\sqrt{2\log(|\mathcal{H}|)T}$.*

Proof. See Appendix [7]. □

3.2 Novel Learning from Expert Advice Algorithms Inspired by Evidential Reasoning

In the previous section, we showed how two optimal LEA algorithms can be derived as restricted cases of the evidential reasoning protocol defined in Algorithm 2, where we require evidence to be represented in an approximated form. A natural question, then, would be to understand the effects of adopting a fully evidential approach, rather than an approximate one. In this section, we will derive two new cautious LEA algorithms, all of which are based on adopting a fully evidential approach. We will discuss, in particular, one algorithm for the realizable setting and one for the agnostic one: in both cases, we will show that, under certain conditions, these algorithms provide better regret bounds than any non-cautious LEA algorithm.

Evidential Halving Evidential Halving is a generalization of the Halving algorithm, described in Algorithm 5. Evidential Halving initially assumes no knowledge about the optimal expert, as represented by the vacuous mass function $m^1(\mathcal{H}) = 1$. At each step t, the Algorithm construct a new mass function w which is obtained by first transforming the current evidence m^t into a probability distribution (using the pignistic transform) $Bet(m^t)$ and then computing its restriction to Y. The mass w is then transformed into a binary mass function by moving all mass to, either, the prediction $y' \in Y$ associated with highest probability, if this latter is larger than a pre-defined threshold β, or to the abstention decision $\{0, 1\}$. After receiving the correct label y, the evidence is updated by combining the *original* mass function m^t with a binary mass function (assigning mass 1 to all experts that predicted correctly). It is easy to observe that, when $\beta = \frac{1}{2}$, Algorithm 5 is equivalent to the Halving algorithm. More in general, however, the following result holds.

Algorithm 5 Evidential Halving

$m^1(\mathcal{H}) = 1$
for $t \in \{1, \dots, T\}$ **do**
  Receive x
  $w = Bet(m^t) \downarrow_x Y$
  $q = \arg\max_{y \in \{0,1\}} w(\{y\})$
  $s = \max_{y \in \{0,1\}} w(\{y\})$
  $m_p(A) = \begin{cases} \mathbb{1}_{s \geq \beta} & A = \{q\} \\ \mathbb{1}_{s < \beta} & A = \{0, 1\} \\ 0 & \text{otherwise} \end{cases}$
  Predict $y' := \arg\max_{A \in \mathcal{F}(m_p)} m_p(A)$
  Receive y and construct $m^y := \begin{cases} 1 & \{h \in \mathcal{H} : h(x) = y\} \\ 0 & \text{otherwise} \end{cases}$
  Penalize learner based on $l_\alpha(y, y')$
  $m^{t+1} = m^t \oplus m^y$
end for

Theorem 3.3. *Assume realizability, and that*

$$l_\alpha(y, A) = \begin{cases} \alpha & A = \{0, 1\} \\ 0 & A = \{y\} \\ 1 & \textit{otherwise}, \end{cases}$$

with $\alpha < 1$. Then, Algorithm 5 enjoys a regret bound of $Regret \leq \frac{\log(|\mathcal{H}|)}{\log(\frac{1}{1-\beta})} + \frac{\alpha \log(|\mathcal{H}|)}{\log(\frac{1}{\beta})}$. For any $\alpha \leq 0.08$ there exists $\beta \in (0.6, 1)$ such that the above bound is smaller (up to logarithmic terms) than the minimum possible regret obtained by any (non-cautious) LEA algorithm. In particular, this holds for the Halving algorithm.

Proof. See Appendix [7]. □

Theorem 3.3 shows that when the cost associated with an abstention is small, Evidential Halving can obtain better performance (up to logarithmic terms) than any non-cautious LEA algorithm. In particular, if $\alpha \to 0$, setting $\beta = 1$, would provide an algorithm that never makes an error (as Algorithm 5 would make a prediction only when all experts agree which, under realizability, ensures that the prediction must be correct) but can abstain up to $|\mathcal{H}| - 1$ times: hence, this setting makes Evidential Halving equivalent to a cautious version of the classical Consistent algorithm [46].

Evidential Weighted Majority Similarly to the case of Evidential Halving, Evidential Weighted Majority, described in Algorithm 6, can be understood as a fully evidential generalization of Weighted Majority. Similarly to Weighted Majority, Algorithm 6 initially assumes no knowledge about the optimal expert, as represented by the vacuous mass function m^1. However, Algorithm 6 differs from Weighted Majority in two main points: 1) at each step t, the current evidence is not approximated by means of a possibility distribution, but is represented as a full (unnormalized) mass function w^t; 2) predictions are not made by sampling from a probability distribution over experts, but rather by directly sampling from the mass function representing the current evidence m^t. Hence, Algorithm 6 directly defines a cautious predictor by allowing to sample sets of experts that disagree on the predictions associated with the given labels. We provide an analysis of the behavior of Algorithm 6 in Theorem 3.4.

Algorithm 6 Evidential reasoning protocol for the Evidential Weighted Majority algorithm

For each $h \in \mathcal{H}$, set $w^1(\mathcal{H}) = 1$
Set $\eta \in (0.02, \ln(2)]$
for $t \in \{1, \dots, T\}$ **do**
  Receive x
  Select $\emptyset \neq H \subseteq \mathcal{H}$ randomly according to $m^t(H) = \frac{w^t(H)}{1 - w^t(\emptyset)}$
  Predict $y' := H(x)$
  Receive y
  Penalize learner based on $l(y, y')$
  Construct $m^y(A) := \begin{cases} s = e^{-\eta} & A = \mathcal{H} \\ 1 - s & A = \{h \in \mathcal{H} : h(x) = y\} \\ 0 & \text{otherwise} \end{cases}$
  $w^{t+1} = w^t \cap m^y$
end for

Theorem 3.4. *Assume that $l(y, y') \in [0, 1]$ for any $y \in Y, y' \subseteq Y$ and, furthermore $l(y, \{0, 1\}) \leq l(y, y')$ whenever $y' \neq \{y\}$. Then, Algorithm 6, with $\eta \in (0.2, \ln(2)]$, satisfies the regret bound:*

$$\sum_{t=1}^{T} \langle m^t, l(y_t, m^t) \rangle - \min_{h \in \mathcal{H}} \sum_{t=1}^{T+1} v_h^t < \frac{T}{2} \left(\frac{1 - e^{-\eta}}{2\eta} + \eta \right), \quad (1)$$

where $l(y_t, \mathcal{H}) = (l(y_t, A(x_t)))_{A \subseteq \mathcal{H}}$. For $\mathcal{H}$ sufficiently large, the above bound is smaller than the regret obtained by any (non-cautious) LEA algorithm achieving a bound of the form $C\sqrt{T \log |\mathcal{H}|}$. In particular, this holds for Weighted Majority.

Proof. See Appendix [7]. □

Theorem 3.4 shows two remarkable properties of Algorithm 6. First, the proven regret bound has no dependence on the size of the

expert set $\mathcal{H}$: this is in stark contrast with the standard LEA framework, where the best known regret bounds show a logarithmic dependence of the form $\sqrt{\log|\mathcal{H}|}$. In particular, this allows Algorithm 6 to enjoy small regret bounds even for very large (actually, even infinite) experts sets, while the same is not known to be achievable by any non-cautious LEA algorithm. Thus, Theorem 3.4 provides a stronger result than those previously known in the literature [15], showing that cautious prediction algorithms can potentially obtain better performance than non-cautious ones. In general, however, neither Weighted Majority nor Evidential Weighted Majority is uniformly better than the other: for example, when $\mathcal{H}$ is small compared to e^T, Weighted Majority can provide better performance than Evidential Weighted Majority. However, Evidential Weighted Majority offers another advantage in comparison with other non-cautious LEA algorithms. Indeed, while these latter typically require a learning rate η which vanishes with the time horizon T (i.e., $\lim_{T\to\infty}\eta = 0$), this is not the case for Algorithm 6 (as Theorem 3.4 requires $\eta > 0.02$). To understand why this can be advantageous, it is easy to see from Algorithm 4, that smaller learning rates η will have the effect of *slowing down* the accumulation of evidence, as $m^y(\mathcal{H}) \sim 1$, thus making learning in general harder. Indeed, it is important to note that the regret bounds in Theorems 3.2 and 3.4 are *worst-case* bounds: while they can, in principle, be enforced by an adversary, in practice one could expect to see an algorithm obtaining better performance, especially if it is able to quickly pinpoint the optimal experts in $\mathcal{H}$: the rate of this convergence process depends on the learning rate, with larger values of η leading to faster learning. In this sense, even though the worst-case optimal setup of η in Algorithm 6 would be $\eta \sim 0.02$ (resulting in a regret bound of approximately $\frac{T}{4}$, meaning that Evidential Weighted Majority does not make more than $\frac{T}{4}$ incorrect predictions than the best expert), the increased flexibility resulting from the possibility to select larger learning rates allows Evidential Weighted Majority to, at the same time, enjoy a worst-case regret bound (indeed, the bound in Theorem 3.4 is non-vacuous for any $\eta \leq 0.5$) while also speeding-up practical learning. In contrast, such faster learning can be achieved using Weighted Majority only at the price of losing any informative worst-case regret bound. Finally, we note that the bound in Theorem 3.4 is not as sharp as possible. Indeed, under the assumptions of Theorem 3.4, a strictly sharper upper bound can easily be derived as:

$$\frac{T}{2}\left(\frac{1-e^{-\eta}}{2\eta}+\eta\right) - \sum_{t=1}^{T}\langle m^t, g^t\rangle + \underbrace{\min_{A\subset\mathcal{H}}\sum_{t=1}^{T+1} v_A^t - \min_{h\in\mathcal{H}}\sum_{t=1}^{T+1} v_h^t}_{<0}, \tag{2}$$

where $g^t = v_A^t - l(y_t, m^t)$. We note that this bound provides strictly better performance guarantees than the one provided above and, in general, can also provide a negative regret bound (meaning that the Algorithm 6 is better than any of the experts in $\mathcal{H}$) when the gap in performance between the experts $h \in \mathcal{H}$ and the cautious predictors $A \subseteq \mathcal{H}$ is large enough. Nonetheless, the bound given in Theorem 3.4 is mathematically simpler, as it does not depend on any non-constant quantity and, hence, can be used to evaluate the worst-case performance of Evidential Weighted Majority a priori, before observing any sequence of instances.

Computational Complexity We conclude our analysis with a discussion of the computational complexity of the analyzed algorithms. To this aim, we assume that the expert set $\mathcal{H}$ is finite: while, in principle, certain algorithm can be implemented efficiently even for infinite expert sets, the restriction to a finite a number of experts is common in the LEA literature and simplifies the analysis. Under the finiteness assumption it is easy to observe that Algorithms 3 and 5 can be trivially implemented in time polynomial (actually, linear) in $|\mathcal{H}|$, with a time complexity of $O(T|\mathcal{H}|)$. For the case of Algorithm 4, a similar result can be obtained by maintaining the contour function of m^t rather than the mass function itself. These results are summarized in the following proposition.

Proposition 3.1. *Algorithms 3, 4 and 5 can be implemented in time complexity $O(T|\mathcal{H}|)$.*

The preceding positive result, however, cannot be applied for Algorithm 6. Indeed, at each step t, Algorithm 6 requires computing the Dempster combination of two non-binary mass functions: in particular, m^y is simple, while m^t is *separable*, i.e. obtained as the Dempster combination of simple mass functions [44]. Unfortunately, in general and without any further assumption on $\mathcal{H}$, this problem is known to be #P-complete [42], even for simple mass functions [43]. This shows that while Algorithm 6 can provide better guarantees than any non-cautious LEA algorithm, in practice, its execution can be computationally unfeasible. We provide, however, a case in which Algorithm 6 can be executed in time polynomial in $|\mathcal{H}|$. More general assumptions ensuring that Algorithm 6 is (fixed-parameter) tractable can be derived from [43], Theorem 6.6.

Proposition 3.2. *Let m^y be defined as in Algorithm 6, and assume that for each $t \in \{1,\ldots,T\}$, it holds that $|\{h \in \mathcal{H} : h(x) = y\}| \leq c$, where c is a constant. Then, Algorithm 6 can be executed in time polynomial[5] in $|\mathcal{H}|$.*

Proof. Under the assumptions of the result, the result directly follows from [43], Prop. 3.2. □

4 Discussion and Interpretation

In the Introduction, we motivated our study of evidence theory as a conceptual foundation for LEA; in this section, we examine this aspect in greater detail. First of all, we describe more in detail how the evidence-theoretical interpretation of uncertainty connects with the algorithms studied. In particular the notion of uncertainty in BFT is reflected in the following aspects: 1) updating in LEA algorithms can be naturally interpreted as a combination of evidence, revealing how a purely algorithmic operation can be re-interpreted through an evidence-theoretic lens; 2) states in LEA algorithms can be represented as mass functions, wherein different structural constraints yield different algorithms; 3) predictions can be interpreted as the result of an evidence-theoretic decision-making process, involving either the approximation of the available evidence into a probability distribution, aligning with the decision theories proposed in [48] and [14], or the selection of a set of experts, which relates more closely to the random set semantics of BFT [40].

We then focus on the studied algorithms. We begin our analysis with the Halving algorithm: as shown by Theorem 3.1, the evidence-theoretic perspective clarifies how the Halving algorithm can be understood as a Bayesian approach. Specifically, the algorithm assumes a *uniform prior* over the set of experts $\mathcal{H}$, which is then updated via *conditioning*. Notably, the Halving algorithm corresponds to the Bayes-optimal prediction rule under the assumptions of realizability and a uniform prior. This Bayesian interpretation also

[5] Note, however, that the time complexity can have exponential dependency on c, so in general Algorithm 6 is only fixed-parameter tractable.

suggests that variants of the algorithm could be defined by modifying the initial prior over the experts. In this sense, the evidence-theoretic perspective provides a useful link between the Halving algorithm and developments in structural risk minimization [50] and minimum description length (MDL)-inspired machine learning [29]. Although such algorithms are necessarily sub-optimal in adversarial settings—since their worst-case regret bounds cannot improve upon $O(\log|\mathcal{H}|)$—they may yield better performance in non-adversarial scenarios. A detailed analysis of these variants is left for future work.

A Bayesian interpretation has also been proposed for the Weighted Majority algorithm [27]. By analogy with the AdaBoost algorithm, it has been shown that, under certain assumptions, Weighted Majority can be interpreted as a Bayes-optimal learner. However, this interpretation has two limitations in the context of adversarial LEA. First, it assumes that the predictions made by the learner (and the experts) at different time steps are independent. In practice, this assumption often fails, as both the learner and the adversary may adapt to each other's behavior over time. Second, the Bayesian interpretation only establishes an equivalence with the Bayes-optimal predictor at the *final* time step. It does not provide a justification for the specific update rule employed by the algorithm. Theorem 3.2 offers a clearer interpretation, suggesting that Weighted Majority is better understood as a possibilistic, rather than probabilistic, algorithm. Indeed, our results show that experts' weights can be interpreted as a possibility distribution, updated iteratively via Dempster's rule of combination with a simple mass function that discounts evidence attributed to unreliable experts (i.e., those who made incorrect predictions). Theorem 3.2 further shows that, under the independence assumption of [27]—which holds, for example, in batch or stochastic, non-adaptive settings—a possibilistic approximation of evidential reasoning results in a Bayes-optimal decision rule. More generally, this evidence-theoretic analysis also sheds light on the optimization-based view of the same algorithm. In this sense, while the use of exponential weights in the algorithm is known to correspond to mirror descent with an exponential loss function [31], this perspective does not clearly justify the choice of this loss. Our analysis shows that these mirror updates emerge naturally as evidence combination in BFT. In this light, our contribution parallels the Bayesian interpretation of batch machine learning: in the latter case, natural loss functions arise from specific priors on model parameters, with regularized risk minimization corresponding to MAP inference (e.g., ridge regression from a normal prior). Analogously, mirror descent (with exponential discounting) in online learning can be seen as resulting from a vacuous prior updated through a possibilistic revision of evidence using unnormalized Dempster's rule. This connection also suggests a principled way to design algorithm variants by modifying either the initial prior (e.g., using a non-vacuous mass function) or the form of the update mechanism (e.g., likelihood- based discounting, as recently proposed in the context of uncertainty quantification [34]), which, in turn, may uncover deeper links with mirror descent. We plan to explore these connections in future work.

The evidence-theoretic perspective also clarifies how classical algorithms can be naturally interpreted as special or approximated cases of a more general, fully evidential approach to online learning, as exemplified by Algorithms 5 and 6. In both cases, while the update mechanisms mirror those of the corresponding classical algorithms, the evidential variants adopt an evidential representation of uncertainty through vacuous mass functions, rather than probabilistic (uniform distributions) or possibilistic (vacuous possibility distributions) approximations. In the case of Evidential Halving, the fully evidential version modifies the classical Halving algorithm by not discarding the entire mass associated with the minority prediction. Instead, this mass is reallocated to an abstention decision. This modification helps avoid scenarios in which a prediction y is made over $1-y$ even when the evidence masses for both are similar. In such situations, Evidential Halving abstains from predicting, thereby reducing regret. In the most extreme case, as previously discussed, this approach leads to an algorithm that never makes errors. Notably, this result demonstrates that Evidential Halving (with $\beta = 1$) qualifies as a KWIK (Knows-What-It-Knows) algorithm [36][6].

Similarly, Evidential Weighted Majority extends the classical algorithm by maintaining a full mass function over experts, rather than resorting to a possibilistic approximation. This fully evidential reasoning framework naturally enables the learner to abstain from making predictions by allowing it to sample *sets* of experts larger than singletons. In this sense, Evidential Weighted Majority can be interpreted as a strict generalization of Weighted Majority, in the sense that it assigns evidence mass in a more general way: this result in increased flexibility (as shown by the fact that in some cases Evidential Weighted Majority can also achieve negative regret with respect to the best expert) but, at the same time, in increased computational costs, as the algorithm generally requires both time and space on the order of $O(2^{\mathcal{H}})$. In both cases—Evidential Halving and Evidential Weighted Majority—further algorithmic variants may be developed by modifying the initial assignment of evidence (e.g., to use a non-vacuous mass function) or the update mechanism (e.g., to use a non-simple mass function). We leave this development for future work.

5 Conclusion

In this article, we examined the application of evidence theory to the domain of online (machine) learning, and in particular, demonstrated how it can be used to formalize the *learning from expert advice* (LEA) problem. To this end, we made two main contributions. First, we showed that two classical LEA algorithms can be derived as special cases of evidence theory, where evidence is approximated using, respectively, probability and possibility distributions. Second, we introduced two novel, fully evidential LEA algorithms and analyzed their theoretical properties. We believe these results lay the groundwork for further investigations into the relationship between evidence theory and online learning. In particular, we consider the following research directions to be especially promising: 1) While this article focused on specific algorithms, future work should explore the general strengths and limitations of the evidence-theoretic approach, including learning-theoretic lower bounds. 2) Our study highlighted significant differences between approximate inference and fully evidential methods. Further research is warranted to deepen the understanding of how various uncertainty representation frameworks interact with online learning. 3) This work was limited to finite expert sets; extending the approach to other settings—such as the widely studied *convex online learning* framework [30]—remains an important direction for future research. 4) Finally, in this work we focused on settings where all available information, including the advice of the experts, are *precise*: that is, they are *complete* and not affected by noise. Evidence theory, and related formalisms, have also been used to model weakly supervised learning tasks in the batch setting [5, 33, 39]: the extension of weakly supervised learning to online learning, using tools from evidence theory, represents an interesting avenue for future work.

[6] In this case, Evidential Halving achieves the same regret bound as the Enumeration algorithm introduced in the KWIK literature.

References

[1] S. Ahmadi, A. Blum, and K. Yang. Fundamental bounds on online strategic classification. In *Proceedings of the 24th ACM Conference on Economics and Computation*, pages 22–58, 2023.

[2] N. Alon, S. Hanneke, R. Holzman, and S. Moran. A theory of pac learnability of partial concept classes. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 658–671. IEEE, 2022.

[3] S. Bubeck, N. Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

[4] T. Burger. Bridging belief function theory to modern machine learning. *arXiv preprint arXiv:1504.03874*, 2015.

[5] A. Campagner. Credal learning: Weakly supervised learning from credal sets. In *ECAI 2023*, pages 327–334. IOS Press, 2023.

[6] A. Campagner and D. Ciucci. Three-way learnability: a learning theoretic perspective on three-way decision. In *2022 17th Conference on Computer Science and Intelligence Systems (FedCSIS)*, pages 243–246. IEEE, 2022.

[7] A. Campagner, F. Arredondo, D. Ciucci, and F. Cabitza. *An Evidence-Theoretic Framework for Online Learning from Expert Advice - Supplementary Material*, Aug. 2025. URL https://doi.org/10.5281/zenodo.16686356. Available on Zenodo. DOI: 10.5281/zenodo.16686356.

[8] N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

[9] N. Cesa-Bianchi, Y. Freund, D. P. Helmbold, and M. K. Warmuth. Online prediction and conversion strategies. *Machine Learning*, 25:71–110, 1996.

[10] N. Cesa-Bianchi, Y. Freund, D. Haussler, D. P. Helmbold, R. E. Schapire, and M. K. Warmuth. How to use expert advice. *Journal of the ACM (JACM)*, 44(3):427–485, 1997.

[11] N. Cesa-Bianchi, C. Gentile, L. Zaniboni, and M. Warmuth. Worst-case analysis of selective sampling for linear classification. *Journal of Machine Learning Research*, 7(7), 2006.

[12] T.-M. Cheung, H. Hatami, P. Hatami, and K. Hosseini. Online learning and disambiguations of partial concept classes. *arXiv preprint arXiv:2303.17578*, 2023.

[13] C. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on information theory*, 16(1):41–46, 1970.

[14] B. R. Cobb and P. P. Shenoy. On the plausibility transformation method for translating belief function models to probability models. *International journal of approximate reasoning*, 41(3):314–330, 2006.

[15] C. Cortes, G. DeSalvo, C. Gentile, M. Mohri, and S. Yang. Online learning with abstention. In *international conference on machine learning*, pages 1059–1067. PMLR, 2018.

[16] F. Cuzzolin. *The geometry of uncertainty: the geometry of imprecise probabilities*. Springer Nature, 2020.

[17] A. P. Dempster. Upper and lower probability inferences based on a sample from a finite univariate population. *Biometrika*, 54(3-4):515–528, 1967.

[18] T. Denœux. Conjunctive and disjunctive combination of belief functions induced by nondistinct bodies of evidence. *Artificial Intelligence*, 172 (2-3):234–264, 2008.

[19] T. Denoeux. Decision-making with belief functions: A review. *International Journal of Approximate Reasoning*, 109:87–110, 2019.

[20] A. DeSantis, G. Markowsky, and M. N. Wegman. Learning probabilistic prediction functions. In *[Proceedings 1988] 29th Annual Symposium on Foundations of Computer Science*, pages 110–119. IEEE Computer Society, 1988.

[21] D. Dubois and H. Prade. A set-theoretic view of belief functions logical operations and approximations by fuzzy sets. *International Journal Of General System*, 12(3):193–226, 1986.

[22] D. Dubois and H. Prade. Representation and combination of uncertainty with belief functions and possibility measures. *Computational intelligence*, 4(3):244–264, 1988.

[23] D. Dubois and H. Prade. Reasoning and learning in the setting of possibility theory-overview and perspectives. *International Journal of Approximate Reasoning*, 171:109028, 2024.

[24] D. Dubois, L. Godo, and H. Prade. An elementary belief function logic. *Journal of Applied Non-Classical Logics*, 33(3-4):582–605, 2023.

[25] R. El-Yaniv et al. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(5), 2010.

[26] Y. Filmus, S. Hanneke, I. Mehalel, and S. Moran. Optimal prediction using expert advice and randomized littlestone dimension. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 773–836. PMLR, 2023.

[27] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

[28] Y. Freund, R. E. Schapire, Y. Singer, and M. K. Warmuth. Using and combining predictors that specialize. In *Proceedings of the twenty-ninth annual ACM symposium on Theory of computing*, pages 334–343, 1997.

[29] P. D. Grünwald. *The minimum description length principle*. MIT press, 2007.

[30] E. Hazan et al. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

[31] D. Hoeven, T. Erven, and W. Kotłowski. The many faces of exponential weights in online learning. In *Conference On Learning Theory*, pages 2067–2092. PMLR, 2018.

[32] S. C. Hoi, D. Sahoo, J. Lu, and P. Zhao. Online learning: A comprehensive survey. *Neurocomputing*, 459:249–289, 2021.

[33] E. Hüllermeier. Learning from imprecise and fuzzy observations: Data disambiguation through generalized loss minimization. *International Journal of Approximate Reasoning*, 55(7):1519–1534, 2014.

[34] E. Hüllermeier and W. Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine learning*, 110(3):457–506, 2021.

[35] H. E. Kyburg Jr. Bayesian and non-bayesian evidential updating. *Artificial intelligence*, 31(3):271–293, 1987.

[36] L. Li, M. L. Littman, and T. J. Walsh. Knows what it knows: a framework for self-aware learning. In *Proceedings of the 25th international conference on Machine learning*, pages 568–575, 2008.

[37] N. Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine learning*, 2:285–318, 1988.

[38] N. Littlestone and M. K. Warmuth. The weighted majority algorithm. *Information and computation*, 108(2):212–261, 1994.

[39] Z.-g. Liu, Q. Pan, and J. Dezert. Classification of uncertain and imprecise data based on evidence theory. *Neurocomputing*, 133:459–470, 2014.

[40] H. T. Nguyen. On random sets and belief functions. *Studies in Fuzziness and Soft Computing, Volume 219*, page 105, 1978.

[41] F. Orabona. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2019.

[42] P. Orponen. Dempster's rule of combination is# p-complete. *Artificial Intelligence*, 44(1-2):245–253, 1990.

[43] D. P. Prieto and R. de Haan. Using hierarchies to efficiently combine evidence with dempster's rule of combination. In *Uncertainty in Artificial Intelligence*, pages 1634–1643. PMLR, 2022.

[44] G. Shafer. *A mathematical theory of evidence*, volume 42. Princeton university press, 1976.

[45] G. Shafer and V. Vovk. *Game-theoretic foundations for probability and finance*. John Wiley & Sons, 2019.

[46] S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[47] S. Shalev-Shwartz et al. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2012.

[48] P. Smets and R. Kennes. The transferable belief model. *Artificial intelligence*, 66(2):191–234, 1994.

[49] C. Tommaso, R. Colomboni, et al. An online learning theory of trading-volume maximization. In *The Thirteenth International Conference on Learning Representations*, pages 1–18, 2025.

[50] V. Vapnik. Principles of risk minimization for learning theory. *Advances in neural information processing systems*, 4, 1991.

[51] F. Voorbraak. A computationally efficient approximation of dempster-shafer theory. *International Journal of Man-Machine Studies*, 30(5): 525–536, 1989.

[52] V. Vovk. Aggregating strategies. In *Proceedings of the Third Annual Workshop on Computational Learning*, pages 371–383, 1990.

[53] R. R. Yager. On the dempster-shafer framework and new combination rules. *Information sciences*, 41(2):93–137, 1987.

[54] R. Yaroshinsky, R. El-Yaniv, and S. S. Seiden. How to better use expert advice. *Machine Learning*, 55(3):271–309, 2004.