A Potential-based Framework for Online Learning with Mistakes and Abstentions

Chicheng Zhang
University of California, San Diego
9500 Gilman Drive, La Jolla, CA 92093
chz038@eng.ucsd.edu

Kamalika Chaudhuri
University of California, San Diego
9500 Gilman Drive, La Jolla, CA 92093
kamalika@cs.ucsd.edu

Abstract

This paper studies the problem of online selective classification, where for each new example, the algorithm has the option to predict Don't Know (abstain). The goal is to make as few abstention as possible, subject to that the number of mistakes made is bounded over time.

Previous work has left a major open challenge, that is, to design tractable algorithms that works in nonrealizable case. In this paper, we provide such an algorithm. We develop an algorithmic framework for designing online learning algorithms with mistakes and abstentions, utilizing a notion called *admissible potential functions*. This framework immediately yields natural generalizations of existing algorithms (e.g. Binomial Weight [CFHW96] or Weighted Majority [LW94, Vov95]) onto online learning with abstentions.

1 Introduction

In many applications of machine learning, misclassification may be costly, but the learning algorithm has the option to occassionally abstain from prediction. For example, in an online credit card fraud detection system, classifying an arriving transaction as fraudulent can result in asset losses of the customers; however the system has the option to predict "Don't know"($\perp$) and pass the transaction on to a human expert. Another example is a medical diagnosis system. When the system is in doubt about a patient's symptom, it has to option to say "Don't Know" to ask for more examinations on the patient [TS13], or ask a physician for assistance.

To ensure reliable learning in these applications, it is therefore essential to develop good algorithms that can trade off classification mistakes for abstention. The performance of the learning algorithm is measured by two quantities: mistakes, the total number of times when the algorithm outputs a wrong label, and abstentions, the total number of Don't Know's ($\perp$) output. The problem has been formulated in the context of online learning recently [LLWS11, SZB10].

Previous work has proposed efficient algorithms that work for finite hypothesis class and realizable setting [SZB10, DZ13], but it is unclear how to extend it to nonrealizable case. Recently, [ZC16] provides an algorithm that works in nonrealizable case. However the algorithm requires computing the *Extended Littlestone's Dimension*, which, similar to computing the Littlestone's Dimension [Lit87], is believed to be intractable. Thus, a major open question is to design tractable algorithms that works in nonrealizable case.

In this paper, we provide such an algorithm. It is based on our two key contributions, which we outline as follows. We first develop an algorithmic framework for designing online learning algorithms with mistakes and abstentions, utilizing a notion called potential function. A potential function quantifies the complexity of the learning problem. We show that if the potential function satisfies an *admissibility* condition, then the algorithm has the desired performance guarantees.

29th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

Secondly, we provide examples of admissible potential functions, e.g. binomial weight potential, exponential potential, etc. These potentials, when combined with our algorithmic framework, yields generalizations of existing efficient online binary prediction algorithms (e.g. Binomial Weight [CFHW96], Weighted Majority [LW94, Vov95]) to online prediction with abstentions.

Related Work. The problem of online prediction with abstention has not received attention until recently. [LLWS11] proposes the KWIK model, where the goal is to make online prediction with $\perp$ option while no mistakes are allowed. [SZB10] proposes an extension of this model, where the goal is to make as few abstentions as possible, subject to the number of mistakes is at most k. It also gives an algorithm in this model, which only works for finite hypothesis class and realizable setting. [DZ13] studies efficient algorithms for learning disjunctions in the above setup. Recently, [ZC16] provides a minimax analysis that exploits structures in hypothesis classes, giving optimal algorithms for the realizable case and mistake-abstention traedeoff upper bounds in non-realizable case, but the algorithm is computationally inefficient.

In the batch setting, the problem is commonly referred to as selective classification or confidence-rated prediction. The pioneering work of [Cho70] studies the setting when the conditional probability of label y given the instance x is known. [BW08, YW10] considers surrogate risk minimization and provide threshold-based abstention rules consistent with the loss functions proposed. [EYW10] studies *perfect* selective classification, where the goal is to find a selective classifier that minimizes the abstention rate, subject to the error rate being zero. [ZC14] proposes an algorithm for *imperfect* selective classification, and shows its tight connection to active learning. [Bal16] gives a trasductive selective classification algorithm by incoroprating constraints on the labels associated with the unlabeled examples.

2 Algorithm

2.1 Setting

We study binary classification in online setting. At each round $t = 1, 2, \ldots$, the algorithm is presented with an example x_t chosen from instance domain $\mathcal{X}$. Then, it is asked to make a prediction $\hat{y}_t$, which can be -1, $+1$, or $\perp$. Subsequently, the true label of the example $y_t \in \{-1, +1\}$, is revealed.

The performance of the algorithm is measured by two quantities: the number of mistakes $\sum_t I(\hat{y}_t = -y_t)$, and the number of abstentions $\sum_t I(\hat{y}_t = \perp)$. We say that an algorithm achieves a (k, d)-SZB bound, if throughout the learning process, it makes at most k mistakes, and at most d abstentions. A round t is called nontrivial if the algorithm incurs a mistake or abstention on that round.

Some constraints need to be imposed on adversary for the proposed algorithm to have nontrivial guarantees. Throughout we make the l-mistake assumption, studied by [CFHW96, ALW06]. When $l = 0$, this degrades to realizability.

Assumption 1 (l-Mistake)**.** *There is a hypothesis h in $\mathcal{H}$ that makes at most l mistakes throughout, that is, $\sum_t I(h(x_t) \neq y_t) \leq l$.*

2.2 Algorithmic Framework

We present Algorithm 1 below, called the *Generalized Weighted Majority Algorithm*. It resembles the Halving Algorithm [Ang87, Lit87], but with a threshold Φ_t set adaptively, inspired by [SZB10]. First, is *conservative*, that is, it only makes state updates in nontrivial rounds. The algorithm keeps a counter c, the number of nontrivial rounds incurred so far.

Second, given a set of examples S_c of size c, $\Phi_{c,T_0+1}(S_c)$ represents the total potential remaining given the examples S_c seen. When a new example x_t arrives, the potential is split into two parts, $\Phi_{c+1,T_0+1}(S_c \cup \{(x_t, -1)\})$ and $\Phi_{c+1,T_0+1}(S_c \cup \{(x_t, +1)\})$, representing the weight voting for -1 (resp. $+1$). The algorithm takes a majority vote over the weights. If the majority only beats the minority by a small margin, then it predicts $\perp$. [1] This guarantees that when a mistake or an abstention happens, the potential drops by a large fraction.

[1] If the algorithm always output a weighted majority label (with no abstentions), then it degrades to an algorithm in the classic Mistake Bound model [Lit87, Ang87].

Algorithm 1 Generalized Weighted Majority Algorithm

Input: admissible potential function $\left\{\Phi_{c,T}(\cdot), 0 \le c \le T\right\}$, mistake budget k.
Precompute horizon $T_0 := \min\left\{T : \binom{T+1}{\le k+1} > \Phi_{0,T+1}(\emptyset)\right\}$.
Initialization: set of examples $S_0 \leftarrow \emptyset$, nontrivial round counter $c \leftarrow 0$, mistake budget $m \leftarrow k$.
for $t = 1, 2, \ldots,$ **do**
  Set threshold $\Phi_t = \binom{T_0-c}{\le m-1}$. # Make Prediction (lines 5 – 11)
  if $\Phi_{c+1,T_0+1}(S_c \cup \{(x_t, -1)\}) < \Phi_t$ **then**
    predict $\hat{y}_t = +1$.
  else if $\Phi_{c+1,T_0+1}(S_c \cup \{(x_t, +1)\}) < \Phi_t$ **then**
    predict $\hat{y}_t = -1$.
  else
    predict $\hat{y}_t = \perp$.
  Receive feedback y_t. # State Update (lines 13 – 17)
  if $\hat{y}_t = -y_t$ **then**
    Mistake budget $m \leftarrow m - 1$.
  if $\hat{y}_t = -y_t$ or $\perp$ **then**
    Examples seen $S_{c+1} \leftarrow S_c \cup \{(x_t, y_t)\}$.
    Nontrivial round counter $c \leftarrow c + 1$.

2.3 Admissible Potential Functions

Algorithm 1 works if the potential function $\{\Phi_{c,T}(\cdot)\}$ has desirable properties, formalized in the definition below.

Definition 1. *A family of potential functions $\left\{\Phi_{c,T}(S), 0 \le c \le T\right\}$ is called* admissible*, if the following holds:*

1. *Uniform Lower Bound. For any S of size c, $\Phi_{c,T}(S) \ge 1$.*[2]
2. *Divisibility. For any T_0, set S of size $c \le T_0 - 1$, and example $x \in \mathcal{X}$,*
$$\Phi_{c,T}(S) \ge \Phi_{c+1,T}(S \cup \{(x, -1)\}) + \Phi_{c+1,T}(S \cup \{(x, +1)\}).$$

We give canonical examples of admissible potential functions below.

Example 1: Binomial Potential. Given a finite hypothesis class $\mathcal{H}$, define
$$\Phi^{\text{bin}}_{c,T}(S) := \sum_{h \in \mathcal{H}} \binom{T-c}{\le l - e(h,S)},$$
where $e(h, S) = \sum_{(x,y)\in S} I(h(x) \ne y)$ is the number of mistakes made by h on S and $\binom{n}{\le k} := \sum_{i=0}^{k} \binom{n}{i}$. It can be verified that $\left\{\Phi^{\text{bin}}_{c,T}(S)\right\}$ is admissible under l-Mistake Assuption.

Example 2: Exponential Potential. Given a finite hypothesis class $\mathcal{H}$, define
$$\Phi^{\text{exp}}_{c,T}(S) := \sum_{h \in \mathcal{H}} (1+\beta)^{T-c} \beta^{e(h,S)-l}.$$
It can be verified that $\left\{\Phi^{\text{exp}}_{c,T}(S)\right\}$ is admissible under l-Mistake Assuption.

Example 3: Potential Functions for Infinite Hypothesis Classes. Given a possibly infinite hypothesis class $\mathcal{H}$ with Littlestone's dimension $\text{Ldim}(\mathcal{H})$, define
$$\Phi^{\text{bin}}_{c,T}(S) := \sum_{(\tilde{y}_1,\ldots,\tilde{y}_c)\in\{-1,+1\}^c} \binom{T-c}{\le \text{Ldim}(\mathcal{H}[(x_1,\tilde{y}_1),\ldots,(x_c,\tilde{y}_c)])} \binom{T-c}{\le l - e(\tilde{y}_1,\ldots,\tilde{y}_c,S)},$$

[2] We lose no generality in setting the potential lower bound as 1, as one can scale the potential by a constant.

where for a labeled dataset S, $\mathcal{H}[S]$ is defined as the set of hypotheses in $\mathcal{H}$ that agrees with the labeled examples in S, i.e. $\mathcal{H}[S] := \big\{h \in \mathcal{H} : h(x) = y \text{ for all } (x, y) \in S\big\}$. Meanwhile, $e(\tilde{y}_1, \ldots, \tilde{y}_c, S) = \sum_{i=1}^{c} I(\tilde{y}_i \neq y_i)$ is the number of mistakes made by labeling $\tilde{y}_1, \ldots, \tilde{y}_c$. It can be verified that $\Big\{\Phi^{\text{bin}}_{c,T}(S)\Big\}$ is admissible under l-Mistake Assuption.

Alternatively, define

$$\Phi^{\text{exp}}_{c,T}(S) := \sum_{(\tilde{y}_1,\ldots,\tilde{y}_c)\in\{-1,+1\}^c} (1+\beta)^{T-c}\beta^{-\operatorname{Ldim}(\mathcal{H}[(x_1,\tilde{y}_1),\ldots,(x_c,\tilde{y}_c)])}(1+\gamma)^{T-c}\gamma^{e(\tilde{y}_1,\ldots,\tilde{y}_c,S)-l},$$

which is also admissible under l-Mistake Assuption.

2.4 Performance Guarantees

We formally provide performance guarantees of Algorithm 1.

Theorem 1. *Suppose Algorithm 1 is run over admissible potential function family $\big\{\Phi_{c,T}(\cdot)\big\}$ with mistake budget k. Then it has a (k, T_0)-SZB bound, where $T_0 := \min\Big\{T \in \mathbb{N} : \binom{T+1}{\leq k+1} > \Phi_{0,T+1}(\emptyset)\Big\}$.*

Plugging into specific potential functions, we get the following corollaries.

Finite Classes. Define T_1^* as the real-valued solution of the equation $\binom{T}{\leq k+1} = |\mathcal{H}|(\frac{eT}{l})^l$. [3] It can be checked by algebra that $T_1^* \leq e(k+1)|\mathcal{H}|^{\frac{1}{k-l+1}}$.

Corollary 1. *Given a finite hypothesis class $\mathcal{H}$, suppose the l-Mistake Assuption holds.*

1. *Algorithm 1, over $\Big\{\Phi^{\text{bin}}_{c,T}(\cdot)\Big\}$ with mistake budget k, has a $(k, (k+1)|\mathcal{H}|^{\frac{1}{k-l+1}})$-SZB bound.*
2. *Algorithm 1, over $\Big\{\Phi^{\text{exp}}_{c,T}(\cdot)\Big\}$ with mistake budget k and $\beta = \frac{l}{T_1^*-l}$. has a $(k, (k+1)|\mathcal{H}|^{\frac{1}{k-l+1}})$-SZB bound.*

Infinite Classes. Define T_2^* as the real-valued solution of the equation $\binom{T}{\leq k+1} = (\frac{eT}{d})^d(\frac{eT}{l})^l$. It can be checked by algebra that $T_2^* \leq (k+1)e^{\frac{k+1}{k+1-l-d}}$.

Corollary 2. *Given a hypothesis class $\mathcal{H}$ of Littlestone's dimension d, suppose the l-Mistake Assuption holds.*

1. *Algorithm 1, over $\Big\{\Phi^{\text{bin}}_{c,T}(\cdot)\Big\}$ with mistake budget k, has a $(k, (k+1)e^{\frac{k+1}{k+1-l-d}})$-SZB bound.*
2. *Algorithm 1, over $\Big\{\Phi^{\text{exp}}_{c,T}(\cdot)\Big\}$ with mistake budget k and $\beta = \frac{l}{T_2^*-l}$, $\gamma = \frac{d}{T_2^*-d}$, has a $(k, (k+1)e^{\frac{k+1}{k+1-l-d}})$-SZB bound.*

3 Conclusions and Future Work

In this paper, we have developed a general potential-based framework for designing online learning algorithms with an abstenion option. This yields tractable prediction algorithms which naturally generalizes existing ones. Several directions are well worth exploration:

1. Can this framework be generalized to analyze multiclass online learning, with perhaps bandit feedback [DH13]? More generally, can this be used to analyze online KWIK regression [SS11]?
2. Can one design other natural potential functions that yield parameter-free algorithms?
3. Our proposed algorithm is deterministic. Can this framework be used to analyze randomized prediction?

[3] The combination number $\binom{a}{b} := \frac{a(a-1)\ldots(a-b+1)}{b(b-1)\ldots 1}$ is still well-defined.

References

[ALW06] Jacob Abernethy, John Langford, and Manfred K. Warmuth. Continuous experts and the binning algorithm. In *19th Annual Conference on Learning Theory, COLT 2006*, pages 544–558, 2006.

[Ang87] Dana Angluin. Queries and concept learning. *Machine Learning*, 2(4):319–342, 1987.

[Bal16] Akshay Balsubramani. Learning to abstain from binary prediction. *arXiv preprint arXiv:1602.08151*, 2016.

[BW08] P. L. Bartlett and M. H. Wegkamp. Classification with a reject option using a hinge loss. *JMLR*, 9, 2008.

[CFHW96] Nicolò Cesa-Bianchi, Yoav Freund, David P. Helmbold, and Manfred K. Warmuth. On-line prediction and conversion strategies. *Machine Learning*, 25(1):71–110, 1996.

[Cho70] C.K. Chow. On optimum error and reject trade-off. *IEEE Trans. on Information Theory*, 1970.

[DH13] Amit Daniely and Tom Helbertal. The price of bandit information in multiclass online classification. In *COLT*, pages 93–104, 2013.

[DZ13] Erik D. Demaine and Morteza Zadimoghaddam. Learning disjunctions: Near-optimal trade-off between mistakes and "i don't know's". In *Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2013*, pages 1369–1379, 2013.

[EYW10] R. El-Yaniv and Y. Wiener. On the foundations of noise-free selective classification. *JMLR*, 11, 2010.

[Lit87] Nick Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, 2(4):285–318, 1987.

[LLWS11] Lihong Li, Michael L. Littman, Thomas J. Walsh, and Alexander L. Strehl. Knows what it knows: a framework for self-aware learning. *Machine Learning*, 82(3):399–443, 2011.

[LW94] Nick Littlestone and Manfred K. Warmuth. The weighted majority algorithm. *Inf. Comput.*, 108(2):212–261, 1994.

[SS11] István Szita and Csaba Szepesvári. Agnostic kwik learning and efficient approximate reinforcement learning. In *COLT*, pages 739–772, 2011.

[SZB10] Amin Sayedi, Morteza Zadimoghaddam, and Avrim Blum. Trading off mistakes and don't-know predictions. In *Advances in Neural Information Processing Systems 23*, pages 2092–2100, 2010.

[TS13] Kirill Trapeznikov and Venkatesh Saligrama. Supervised sequential classification under budget constraints. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, pages 581–589, 2013.

[Vov95] Vladimir G Vovk. A game of prediction with expert advice. In *Proceedings of the eighth annual conference on Computational learning theory*, pages 51–60. ACM, 1995.

[YW10] M. Yuan and M. H. Wegkamp. Classification methods with reject option based on convex risk minimization. *JMLR*, 11, 2010.

[ZC14] C. Zhang and K. Chaudhuri. Beyond disagreement-based agnostic active learning. In *NIPS*, 2014.

[ZC16] Chicheng Zhang and Kamalika Chaudhuri. The extended littlestone's dimension for learning with mistakes and abstentions. page 1584–1616, 2016.