Supplement: A Theory of Multiclass Boosting

Indraneel Mukherjee **Robert E. Schapire**

Princeton University
Department of Computer Science
Princeton NJ 08540
{imukherj,schapire}@cs.princeton.edu

We give formal proofs for various claims made in the paper, roughly in their order of appearance. Recall that we have assumed that the true label y_i of example i in our training set is always 1. Nevertheless, we may occasionally continue to refer to the true labels as y_i.

S.1 Minimax Theorem

We will make use of the following minimax result, that appears as Corollary 37.3.2 of [2].

Theorem. *(Minimax Theorem) Let C, D be non-empty closed convex subsets of $\mathbb{R}^m, \mathbb{R}^n$ respectively, and let K be a continuous finite concave-convex function on $C \times D$. If either C or D is bounded, one has*

$$\min_{v \in D} \max_{u \in C} K(u, v) = \max_{u \in C} \min_{v \in D} K(u, v).$$

S.2 Proof of Theorem 1

Applying the minimax theorem yields

$$0 \geq \max_{\mathbf{C} \in \mathcal{C}^{\text{eor}}} \min_{h \in \mathcal{H}} \mathbf{C} \bullet (\mathbf{1}_h - \mathbf{B}) = \min_{\boldsymbol{\lambda} \in \Delta(\mathcal{H})} \max_{\mathbf{C} \in \mathcal{C}^{\text{eor}}} \mathbf{C} \bullet (\mathbf{H}_{\boldsymbol{\lambda}} - \mathbf{B}),$$

where

$$\mathbf{H}_{\boldsymbol{\lambda}} \triangleq \sum_{h \in \mathcal{H}} \lambda(h) \mathbf{1}_h,$$

and where the first inequality follows from the definition (2) of the weak-learning condition. Let $\boldsymbol{\lambda}^*$ be a minimizer of the min-max expression. Unless the first entry of each-row of $(\mathbf{H}_{\boldsymbol{\lambda}^*} - \mathbf{B})$ is the largest, the right hand side of the min-max expression can be made arbitrarily large by choosing $\mathbf{C} \in \mathcal{C}^{\text{eor}}$ appropriately. For example, if in some row i, the j_0th element is strictly larger than the first element, by choosing

$$C(i,j) = \begin{cases} -1 & \text{if } j = 1 \\ 1 & \text{if } j = j_0 \\ 0 & \text{otherwise,} \end{cases}$$

we get a matrix in $\mathcal{C}^{\text{eor}}$ which causes $\mathbf{C} \bullet (\mathbf{H}_{\boldsymbol{\lambda}^*} - \mathbf{B})$ to be equal to $C(i, j_0) - C(i, 1) > 0$, an impossibility by the first inequality.

Therefore, the convex combination of the weak classifiers, obtained by choosing each weak classifier with weight given by $\boldsymbol{\lambda}^*$, perfectly classifies the training data, in fact with a margin γ.

□

S.3 Proof of Theorem 2

We will reuse notation from the proof of Theorem 1 above. $\mathcal{H}$ is boostable implies there exists some distribution $\boldsymbol{\lambda}^* \in \Delta(\mathcal{H})$ such that

$$\forall j \neq 1, i : \mathbf{H}_{\boldsymbol{\lambda}^*}(i, 1) - \mathbf{H}_{\boldsymbol{\lambda}^*}(i, j) > 0.$$

Let $\gamma > 0$ be the minimum of the above expression over all possible (i, j), and let $\mathbf{B} = \mathbf{H}_{\boldsymbol{\lambda}^*}$. Then $\mathbf{B} \in \mathcal{B}^{\text{eor}}_\gamma$, and

$$\max_{\mathbf{C}\in\mathcal{C}^{\text{eor}}} \min_{h\in\mathcal{H}} \mathbf{C} \bullet (\mathbf{1}_h - \mathbf{B}) \le \min_{\boldsymbol{\lambda}\in\Delta(\mathcal{H})} \max_{\mathbf{C}\in\mathcal{C}^{\text{eor}}} \mathbf{C} \bullet (\mathbf{H}_{\boldsymbol{\lambda}} - \mathbf{B}) \le \max_{\mathbf{C}\in\mathcal{C}^{\text{eor}}} \mathbf{C} \bullet (\mathbf{H}_{\boldsymbol{\lambda}^*} - \mathbf{B}) = 0,$$

where the equality follows since by definition $\mathbf{H}_{\boldsymbol{\lambda}^*} - \mathbf{B} = \mathbf{0}$. The max-min expression is at most zero is another way of saying that $\mathcal{H}$ satisfies the weak-learning condition $(\mathcal{C}^{\text{eor}}, \mathbf{B} \in \mathcal{B}^{\text{eor}}_\gamma)$ as in (2). $\square$

S.4 Each edge-over-random condition is too strong

In Section 3 we mention that any single edge-over-random condition is too strong. Here we provide, for any $\gamma > 0$ and edge-over-random baseline $\mathbf{B} \in \mathcal{B}^{\text{eor}}_\gamma$, a dataset and weak classifier space that is boostable but fails to satisfy the condition $(\mathcal{C}^{\text{eor}}, \mathbf{B})$.

Pick $m > 1/\gamma$ so that $\lfloor m(1/2 + \gamma)\rfloor > m/2$. Our data-set will have m labeled examples $\{(0, y_0), \ldots, (m-1, y_{m-1})\}$, and m weak classifiers. We want the following symmetries in our weak classifiers:

- Each weak classifier correctly classifies $\lfloor m(1/2+\gamma)\rfloor$ examples and misclassifies the rest.
- On each example, $\lfloor m(1/2 + \gamma)\rfloor$ weak classifiers predict correctly.

Note the second property implies boostability, since the uniform convex combination of all the weak classifiers is a perfect predictor.

The two properties can be satisfied by the following design. A window is a contiguous sequence of examples that may wrap around; for example $\{i, (i+1) \mod m, \ldots, (i+k) \mod m\}$ is a window containing k elements, which may wrap around if $i + k \ge m$. For each window of length $\lfloor m(1/2 + \gamma)\rfloor$ create a hypothesis that correctly classifies within the window, and misclassifies outside. This weak-hypothesis space has size m, and has the required properties.

We still have flexibility as to how the misclassifications occur, and which cost-matrix to use, which brings us to the next two choices:

- Whenever a hypothesis misclassifies on example i, it predicts label $\hat{y}_i \triangleq \operatorname{argmin}\{B(i,l) : l \ne y_i\}$.
- A cost-matrix is chosen so that the cost of predicting $\hat{y}_i$ on example i is 1, but for any other prediction the cost is zero. Observe this cost-matrix belongs to $\mathcal{C}^{\text{eor}}$.

Therefore, every time a weak classifier predicts incorrectly, it also suffers cost 1. Since each weak classifier predicts correctly only within a window of length $\lfloor m(1/2+\gamma)\rfloor$, it suffers cost $\lceil m(1/2 - \gamma)\rceil$. On the other hand, by definition, $B(i, \hat{y}_i) \le 1/k - \gamma$. So the cost of $\mathbf{B}$ on the chosen cost-matrix is $m(1/k - \gamma)$, which is less than the cost $\lceil m(1/2-\gamma)\rceil$ of any weak classifier whenever the number of labels k is more than two. Hence our boostable space of weak classifiers fails to satisfy $(\mathcal{C}^{\text{eor}}, \mathbf{B})$. $\square$

S.5 Conditions for AdaBoost.MH and AdaBoost.MR in our framework

In Section 3, we have stated the conditions in our framework corresponding to AdaBoost.MH and AdaBoost.MR [4]. Here we provide proofs showing that our conditions match the ones in the original paper.

Theorem. *The weak-learning condition used by AdaBoost.MH [4] is equivalent to $(\mathcal{C}^{MH}, \mathbf{B}^{MH}_\gamma)$, and that used by AdaBoost.MR [4] is equivalent to $(\mathcal{C}^{MR}, \mathbf{B}^{MR}_\gamma)$.*

Proof. AdaBoost.MH [4] was originally designed to use weak-hypotheses that return a prediction for every example and every label. They require that for any matrix with non-negative entries $d(i, l)$, the weak-hypothesis should achieve $1/2+\gamma$ accuracy

$$\begin{aligned}\sum_{i=1}^m \left(\mathbb{1}\left[h(x_i) \neq y_i\right] d(i,y_i) + \sum_{l \neq y_i} \mathbb{1}\left[h(x_i) = l\right] d(i,l) \right) \\ \leq \quad (1/2-\gamma) \sum_{i=1}^m \sum_{l=1}^k d(i,l). \end{aligned} \tag{S.1}$$

This can be rewritten as

$$\begin{aligned}\sum_{i=1}^m \left(-\mathbb{1}\left[h(x_i) = y_i\right] d(i,y_i) + \sum_{l \neq y_i} \mathbb{1}\left[h(x_i) = l\right] d(i,l) \right) \\ \leq \quad \sum_{i=1}^m \left((1/2-\gamma) \sum_{l \neq y_i} d(i,l) - (1/2+\gamma)\, d(i,y_i) \right). \end{aligned}$$

Using the mapping

$$C(i,l) = \begin{cases} d(i,l) & \text{if } l \neq y_i \\ -d(i,l) & \text{if } l = y_i, \end{cases}$$

their weak-learning condition may be rewritten as follows

$$\begin{aligned} &\forall \mathbf{C} \in \mathbb{R}^{m \times k} \text{ satisfying } \left\{ C(i,y_i) \leq 0, C(i,l) \geq 0 \text{ for } l \neq y_i \right\}, \exists h \in \mathcal{H} : \\ &\sum_{i=1}^m C(i,h(x_i)) \leq \sum_{i=1}^m \left((1/2+\gamma) C(i,y_i) + (1/2-\gamma) \sum_{l \neq y_i} C(i,l) \right). \end{aligned}$$

Finally using the fact that we have assumed (without loss of generality) that $\forall i : y_i = 1$, the above condition is the same as

$$\forall \mathbf{C} \in \mathcal{C}^{\text{MH}}, \exists h \in \mathcal{H} : \mathbf{C} \bullet \left(\mathbf{1}_h - \mathbf{B}_\gamma^{\text{MH}} \right) \leq 0,$$

i.e. the $(\mathcal{C}^{\text{MH}}, \mathbf{B}_\gamma^{\text{MH}})$ weak-learning condition.

AdaBoost.MR [4] is a variant of AdaBoost.MH. For any non-negative cost-vectors $\{d(i,l)\}_{l \neq y_i}$, the weak-hypothesis returned should satisfy the following

$$\begin{aligned} \sum_{i=1}^m \sum_{l \neq y_i} \left(\mathbb{1}\left[h(x_i) = l\right] - \mathbb{1}\left[h(x_i) = y_i\right] \right) d(i,l) \quad &\leq \quad -2\gamma \sum_{i=1}^m \sum_{l \neq y_i} d(i,l) \\ \text{i.e. } \sum_{i=1}^m \left(-1\left[h(x_i) = y_i\right] \sum_{l \neq y_i} d(i,l) + \sum_{l \neq y_i} \mathbb{1}\left[h(x_i) = l\right] d(i,l) \right) \quad &\leq \quad -2\gamma \sum_{i=1}^m \sum_{l \neq y_i} d(i,l) \end{aligned}$$

Substituting

$$C(i,l) = \begin{cases} d(i,l) & l \neq y_i \\ -\sum_{l \neq y_i} d(i,l) & l = y_i, \end{cases}$$

we may rewrite AdaBoost.MR's weak-learning condition as

$$\begin{aligned} &\forall \mathbf{C} \in \mathbb{R}^{m \times k} \text{ satisfying } \left\{ C(i,l) \geq 0 \text{ for } l \neq y_i, C(i,y_i) = -\sum_{l \neq y_i} C(i,l) \right\}, \exists h \in \mathcal{H} : \\ &\sum_{i=1}^m C(i,h(x_i)) \leq -\gamma \sum_{i=1}^m \left(-C(i,y_i) + \sum_{l \neq y_i} C(i,l) \right). \end{aligned}$$

Again using the fact that we have assumed $\forall i : y_i = 1$, the above condition is the same as

$$\forall \mathbf{C} \in \mathcal{C}^{\text{MR}}, \exists h \in \mathcal{H} : \mathbf{C} \bullet \left(\mathbf{1}_h - \mathbf{B}_\gamma^{\text{MR}} \right) \leq 0,$$

i.e. the $(\mathcal{C}^{\text{MR}}, \mathbf{B}_\gamma^{\text{MR}})$ weak-learning condition.

$\square$

S.6 Weak-learning conditions of AdaBoost.MH and AdaBoost.M1 are same in our framework

Here we prove the claim, made in Section 1, that the weak-learning conditions of AdaBoost.MH and AdaBoost.M1 [1] are identical in our framework.

We first rewrite the conditons used by AdaBoost.M1 in the language of our framework. Adaboost.M1 [1] requires $1/2 + \gamma$ accuracy with respect to any non-negative weights $d(1), \ldots, d(m)$ on the training set,

$$\begin{aligned} \sum_{i=1}^{m} d(i)\mathbb{1}\left[h(x_i) \neq y_i\right] &\leq (1/2 - \gamma)\sum_{i=1}^{m} d(i), \qquad \text{(S.2)} \\ \text{i.e. } \sum_{i=1}^{m} d(i)[\![h(x_i) \neq y_i]\!] &\leq -2\gamma\sum_{i=1}^{m} d(i). \end{aligned}$$

where $[\![\cdot]\!]$ is the $\pm$ indicator function, taking value $+1$ when its argument is true, and -1 when false. Using the transformation

$$C(i,l) = [\![l \neq y_i]\!] d(i)$$

we may rewrite the above condition as

$$\forall C \in \mathbb{R}^{m \times k} \text{ satisfying } \left\{0 \leq -C(i, y_i) = C(i,l) \text{ for } l \neq y_i\right\}, \qquad \text{(S.3)}$$

$$\exists h \in \mathcal{H} : \sum_{i=1}^{m} C(i, h(x_i)) \leq 2\gamma \sum_{i=1}^{m} C(i, y_i)$$

$$\text{i.e.} \quad \forall \mathbf{C} \in \mathcal{C}^{\mathrm{M1}}, \exists h \in \mathcal{H} : \mathbf{C} \bullet \left(\mathbf{1}_h - \mathbf{B}^{\mathrm{M1}}_\gamma\right) \leq 0, \qquad \text{(S.4)}$$

where $\mathbf{B}^{\mathrm{M1}}_\gamma(i,l) = 2\gamma\mathbb{1}\left[l = y_i\right]$, and $\mathcal{C}^{\mathrm{M1}} \subset \mathbb{R}^{m \times k}$ consists of matrices satisfying the constraints in (S.3).

We now show the equivalence of the weak-learning conditions of AdaBoost.M1 and AdaBoost.MH.

Lemma. *A weak classifier space $\mathcal{H}$ satisfies $(\mathcal{C}^{M1}, \mathbf{B}^{M1}_\gamma)$ if and only if it satisfies $(\mathcal{C}^{MH}, \mathbf{B}^{MH}_\gamma)$.*

Proof. We will refer to $(\mathcal{C}^{\mathrm{M1}}, \mathbf{B}^{\mathrm{M1}}_\gamma)$ by M1 and $(\mathcal{C}^{\mathrm{MH}}, \mathbf{B}^{\mathrm{MH}}_\gamma)$ by MH for brevity. The proof is in three steps.

Step (i): $\mathcal{H}$ satisfies M1 implies $\mathcal{H}$ satisfies MH. This follows since any constraint (S.2) imposed by M1 on $\mathcal{H}$ can be reproduced by MH by plugging the following values of $d(i,l)$ in (S.1)

$$d(i,l) = \begin{cases} d(i) & \text{if } l = y_i \\ 0 & \text{if } l \neq y_i. \end{cases}$$

Step (ii): $\mathcal{H}$ satisfies M1 implies there is a convex combination $\mathbf{H}_{\boldsymbol{\lambda}}$ of the matrices $\mathbf{1}_h \in \mathcal{H}$ such that

$$\forall i : \left(\mathbf{H}_{\boldsymbol{\lambda}} - \mathbf{B}^{\mathrm{MH}}_\gamma\right)(i,l) \begin{cases} \geq 0 & \text{if } l = y_i \\ \leq 0 & \text{if } l \neq y_i. \end{cases}$$

Indeed, the minmax theorem yields

$$\min_{\boldsymbol{\lambda} \in \Delta(\mathcal{H})} \max_{\mathbf{C} \in \mathcal{C}^{\mathrm{M1}}} \mathbf{C} \bullet \left(\mathbf{H}_{\boldsymbol{\lambda}} - \mathbf{B}^{\mathrm{M1}}_\gamma\right) = \max_{\mathbf{C} \in \mathcal{C}^{\mathrm{M1}}} \min_{h \in \mathcal{H}} \mathbf{C} \bullet \left(\mathbf{1}_h - \mathbf{B}^{\mathrm{M1}}_\gamma\right) \leq 0,$$

where the inequality is a restatement of our assumption that $\mathcal{H}$ satisfies M1. If $\boldsymbol{\lambda}$ is a minimizer of the minmax expression, then $\mathbf{H}_{\boldsymbol{\lambda}}$ must satisfy

$$\forall i : \mathbf{H}_{\boldsymbol{\lambda}}(i,l) \begin{cases} \geq 1/2 + \gamma & \text{if } l = y_i \\ \leq 1/2 - \gamma & \text{if } l \neq y_i, \end{cases} \qquad \text{(S.5)}$$

or else some choice of $\mathbf{C} \in \mathcal{C}^{\mathrm{M1}}$ can cause $\mathbf{C} \bullet \left(\mathbf{H}_{\boldsymbol{\lambda}} - \mathbf{B}^{\mathrm{M1}}\right)$ to exceed 0. In particular, if $\mathbf{H}_{\boldsymbol{\lambda}}(i_0, l) < 1/2 + \gamma$, then

$$\left(\mathbf{H}_{\boldsymbol{\lambda}} - \mathbf{B}^{\mathrm{M1}}_\gamma\right)(i_0, y_{i_0}) < \sum_{l \neq y_{i_0}} \left(\mathbf{H}_{\boldsymbol{\lambda}} - \mathbf{B}^{\mathrm{M1}}_\gamma\right)(i_0, l).$$

Now, if we choose $\mathbf{C} \in \mathcal{C}^{\mathrm{M1}}$ as

$$C(i,l) = \begin{cases} 0 & \text{if } i \neq i_0 \\ 1 & \text{if } i = i_0, l \neq y_{i_0} \\ -1 & \text{if } i = i_0, l = y_{i_0}, \end{cases}$$

then,

$$\mathbf{C} \bullet \left(\mathbf{H}_{\boldsymbol{\lambda}} - \mathbf{B}_\gamma^{\mathrm{M1}}\right) = -\left(\mathbf{H}_{\boldsymbol{\lambda}} - \mathbf{B}_\gamma^{\mathrm{M1}}\right)(i_0, y_{i_0}) + \sum_{l \neq y_{i_0}} \left(\mathbf{H}_{\boldsymbol{\lambda}} - \mathbf{B}_\gamma^{\mathrm{M1}}\right)(i_0, l) > 0,$$

contradicting the minmax inequality. Therefore some $\mathbf{H}_{\boldsymbol{\lambda}}$ satisfying (S.5) exists. Step (ii) now follows by observing that $\mathbf{B}_\gamma^{\mathrm{MH}}$ satisfies

$$\forall i : \mathbf{B}_\gamma^{\mathrm{MH}}(i,l) = \begin{cases} 1/2 + \gamma & \text{if } l = y_i \\ 1/2 - \gamma & \text{if } l \neq y_i. \end{cases}$$

Step (iii) If $\mathcal{H}$ satisfies M1's conditions, then Step (ii) implies

$$0 \geq \min_{\boldsymbol{\lambda} \in \Delta(\mathcal{H})} \max_{\mathbf{C} \in \mathcal{C}^{\mathrm{MH}}} \mathbf{C} \bullet \left(\mathbf{H}_{\boldsymbol{\lambda}} - \mathbf{B}_\gamma^{\mathrm{MH}}\right) = \max_{\mathbf{C} \in \mathcal{C}^{\mathrm{MH}}} \min_{h \in \mathcal{H}} \mathbf{C} \bullet \left(\mathbf{1}_h - \mathbf{B}_\gamma^{\mathrm{MH}}\right),$$

where the equality follows from the minimax theorem. The $\max\min$ expression at most zero encodes $\mathbf{B}^{\mathrm{MH}}$'s weak-learning condition. Hence $\mathcal{H}$ satisfies M1 implies $\mathcal{H}$ satisfies MH. Together with Step (i), this completes the proof. □

S.7 Proof of Theorem 3

We will show the following three conditions are equivalent:

(A) $\mathcal{H}$ is boostable

(B) $\exists \gamma > 0$ such that $\forall \mathbf{C} \in \mathcal{C}^{\mathrm{eor}}, \exists h \in \mathcal{H} : \mathbf{C} \bullet \mathbf{1}_h \leq \max_{\mathbf{B} \in \mathcal{B}_\gamma^{\mathrm{eor}}} \mathbf{C} \bullet \mathbf{B}$

(C) $\exists \gamma > 0 : \mathcal{H}$ satisfies $(\mathcal{C}^{\mathrm{MR}}, \mathbf{B}_\gamma^{\mathrm{MR}})$.

We will show (A) implies (B), (B) implies (C), and (C) implies (A) to achieve the above.

(A) implies (B): Immediate from Theorem 2.

(B) implies (C): Suppose (B) is satisfied with 2γ. We will show that this implies $\mathcal{H}$ satisfies $(\mathcal{C}^{\mathrm{MR}}, \mathbf{B}_\gamma^{\mathrm{MR}})$. Notice $\mathcal{C}^{\mathrm{MR}} \subset \mathcal{C}^{\mathrm{eor}}$. Therefore it suffices to show that

$$\forall \mathbf{C} \in \mathcal{C}^{\mathrm{MR}}, \mathbf{B} \in \mathcal{B}_{2\gamma}^{\mathrm{eor}} : \mathbf{C} \bullet \left(\mathbf{B} - \mathbf{B}_\gamma^{\mathrm{MR}}\right) \leq 0.$$

Notice that $\mathbf{B} \in \mathcal{Q}^{2\gamma}$ implies $\mathbf{B}' = \mathbf{B} - \mathbf{B}_\gamma^{\mathrm{MR}}$ belongs to $\mathcal{B}_0^{\mathrm{eor}}$. Then, for any $\mathbf{C} \in \mathcal{C}^{\mathrm{MR}}$, $\mathbf{C} \bullet \mathbf{B}'$ can be written as

$$\mathbf{C} \bullet \mathbf{B}' = \sum_{i=1}^{m} \sum_{j=2}^{k} C(i,j) \left(B'(i,j) - B'(i,1)\right).$$

Since $C(i,j) \geq 0$ for $j > 1$, and $B'(i,j) - B'(i,1) \leq 0$, we have our result.

(C) implies (A): Applying the minimax theorem,

$$0 \geq \max_{\mathbf{C} \in \mathcal{C}^{\mathrm{MR}}} \min_{h \in \mathcal{H}} \mathbf{C} \bullet \left(\mathbf{1}_h - \mathbf{B}_\gamma^{\mathrm{MR}}\right) = \min_{\boldsymbol{\lambda} \in \Delta(\mathcal{H})} \max_{\mathbf{C} \in \mathcal{C}^{\mathrm{MR}}} \mathbf{C} \bullet \left(\mathbf{H}_{\boldsymbol{\lambda}} - \mathbf{B}_\gamma^{\mathrm{MR}}\right).$$

For any i_0 and $l_0 \neq 1$, the following cost-matrix $\mathbf{C}$ satisfies $\mathbf{C} \in \mathcal{C}^{\mathrm{MR}}$,

$$\mathbf{C}(i,l) = \begin{cases} 0 & \text{if } i \neq i_0 \text{ or } l \notin \{1, l_0\} \\ 1 & \text{if } i = i_0, l = l_0 \\ -1 & \text{if } i = i_0, l = 1. \end{cases}$$

Let $\boldsymbol{\lambda}$ belong to the argmin of the $\min\max$ expression. Then $\mathbf{C} \bullet \left(\mathbf{H}_{\boldsymbol{\lambda}} - \mathbf{B}_\gamma^{\mathrm{MR}}\right) \leq 0$ implies $\mathbf{H}_{\boldsymbol{\lambda}}(i_0, 1) - \mathbf{H}_{\boldsymbol{\lambda}}(i_0, l_0) \geq 2\gamma$. Since this is true for all i_0 and $l_0 \neq 1$, we conclude that the $(\mathcal{C}^{\mathrm{MR}}, \mathbf{B}_\gamma^{\mathrm{MR}})$ condition implies boostability.

This concludes the proof of equivalence. □

S.8 Proof of Theorem 5

Let $\mathcal{C}_0^{\text{eor}} \subseteq \mathbb{R}^k$ denote all vectors $\mathbf{c}$ satisfying $\forall l : c(1) \leq c(l)$. Then, we have

$$\begin{aligned}
\phi_t^{\mathbf{b}}(\mathbf{s}) &= \min_{\mathbf{c}\in\mathcal{C}_0^{\text{eor}}} \max_{\mathbf{p}\in\Delta\{1,\ldots,k\}} \quad \mathbb{E}_{l\sim\mathbf{p}}\left[\phi_{t-1}\left(\mathbf{s}+\mathbf{e}_l\right)\right] \quad \text{s.t.} \quad \mathbb{E}_{l\sim\mathbf{p}}[c(l)] \leq \mathbb{E}_{l\sim\mathbf{b}}\left[c(l)\right], \quad \text{(by (4))} \\
&= \min_{\mathbf{c}\in\mathcal{C}_0^{\text{eor}}} \max_{\mathbf{p}\in\Delta} \min_{\lambda\geq 0} \left\{\mathbb{E}_{l\sim\mathbf{p}}\left[\phi_{t-1}\left(\mathbf{s}+\mathbf{e}_l\right)\right] + \lambda\left(\mathbb{E}_{l\sim\mathbf{b}}\left[c(l)\right] - \mathbb{E}_{l\sim\mathbf{p}}[c(l)]\right)\right\} \text{ (Lagrangean)} \\
&= \min_{\mathbf{c}\in\mathcal{C}_0^{\text{eor}}} \min_{\lambda\geq 0} \max_{\mathbf{p}\in\Delta} \mathbb{E}_{l\sim\mathbf{p}}\left[\phi_{t-1}\left(\mathbf{s}+\mathbf{e}_l\right)\right] + \lambda\left\langle \mathbf{b}-\mathbf{p},\mathbf{c}\right\rangle \text{ (min-max theorem)} \\
&= \min_{\mathbf{c}\in\mathcal{C}_0^{\text{eor}}} \max_{\mathbf{p}\in\Delta} \mathbb{E}_{l\sim\mathbf{p}}\left[\phi_{t-1}\left(\mathbf{s}+\mathbf{e}_l\right)\right] + \left\langle \mathbf{b}-\mathbf{p},\mathbf{c}\right\rangle \text{ (absorb } \lambda \text{ into } \mathbf{c}\text{)} \\
&= \max_{\mathbf{p}\in\Delta} \min_{\mathbf{c}\in\mathcal{C}_0^{\text{eor}}} \mathbb{E}_{l\sim\mathbf{p}}\left[\phi_{t-1}\left(\mathbf{s}+\mathbf{e}_l\right)\right] + \left\langle \mathbf{b}-\mathbf{p},\mathbf{c}\right\rangle \text{ (min-max theorem) .}
\end{aligned}$$

Unless $q(1)-p(1) \leq 0$ and $q(l)-p(l) \geq 0$ for each $l > 1$, the quantity $\langle \mathbf{b}-\mathbf{p},\mathbf{c}\rangle$ can be made arbitrarily small for appropriate choices of $\mathbf{c} \in \mathcal{C}_0^{\text{eor}}$. The max-player is therefore forced to constrain its choices of $\mathbf{p}$, and the above expression becomes

$$\max_{\substack{\mathbf{p}\in\Delta \\ p(1)\geq q(1), \forall l>1: p(l)\leq q(l)}} \quad \mathbb{E}_{l\sim\mathbf{p}}\left[\phi_{t-1}\left(\mathbf{s}+\mathbf{e}_l\right)\right]$$

Lemma 6 of [3] states that if L is *proper* (as defined in *our* paper), so is ϕ_t; the same result can be extended to our drifting games. This implies the optimal choice of $\mathbf{p}$ in the above expression is in fact the distribution that puts as small weight as possible in the first coordinate, namely $\mathbf{b}$.

Therefore the optimum choice of $\mathbf{p}$ is $\mathbf{b}$, and the potential is the same as

$$\phi_t(\mathbf{s}) = \mathbb{E}_{l\sim\mathbf{b}}\left[\phi_{t-1}\left(\mathbf{s}+\mathbf{e}_l\right)\right].$$

Inductively assuming $\phi_{t-1}(\mathbf{x}) = \mathbb{E}\left[L(\mathcal{R}_{\mathbf{b}}^{t-1}(\mathbf{x}))\right]$,

$$\begin{aligned}
\phi_t(\mathbf{s}) &= \mathbb{E}_{l\sim\mathbf{b}}\left[L(\mathcal{R}_{\mathbf{b}}^{t-1}(\mathbf{s})+\mathbf{e}_l)\right] \\
&= \mathbb{E}\left[L(\mathcal{R}_{\mathbf{b}}^{t}(\mathbf{s}))\right].
\end{aligned}$$

The last equality follows by observing that the random position $\mathcal{R}_{\mathbf{b}}^{t-1}(\mathbf{s})+\mathbf{e}_l$ is distributed as $\mathcal{R}_{\mathbf{b}}^{t}(\mathbf{s})$ when l is sampled from $\mathbf{b}$. □

S.9 Calculations for the Adaptive case

While discussing the adaptive algorithm we mention how to choose the weights α_t in each round. Here are formal proofs to back up some of the claims made in that section.

Lemma. *Suppose cost matrix $\mathbf{C}_t$ is chosen as in (7), and the returned weak classifier h_t beats $\mathbf{U}_{\delta_t}$ on $\mathbf{C}_t$ i.e. $\mathbf{C}_t \bullet \mathbf{1}_{h_t} \leq \mathbf{C}_t \bullet \mathbf{U}_{\delta_t}$. Then choosing any weight $\alpha_t > 0$ for h_t makes the loss at time t, $\sum_{i=1}^m \sum_{l=2}^k e^{\{f_t(i,l)-f_t(i,1)\}}$, at most a factor*

$$1 - \frac{1}{2}(e^{\alpha_t} - e^{-\alpha_t})\delta_t + \frac{1}{2}(e^{\alpha_t} + e^{-\alpha_t} - 2)$$

of the loss before choosing, $\sum_{i=1}^m \sum_{l=2}^k e^{\{f_{t-1}(i,l)-f_{t-1}(i,1)\}}$.

Proof. Let S_+, S_- denote the set of examples where h_t classified correctly, incorrectly resp. Also let $L_t(i)$ denote the sum $\sum_{l=2}^k e^{f_t(i,l)-f_t(i,1)}$. Then the loss after t rounds is $\sum_{i\in S_+\cup S_-} L_t(i)$. Further $C_t(i,1) = -L_{t-1}(i)$. By the edge-condition

$$\begin{aligned}
-\sum_{i\in S_+} L_{t-1}(i) + \sum_{i\in S_-} e^{\{f_{t-1}(i,h_t(x_i))-f_{t-1}(i,1)\}} = \mathbf{C}_t \bullet \mathbf{1}_{h_t} &\leq \mathbf{C}_t \bullet \mathbf{U}_{\delta_t} = -\delta_t \sum_{i\in S_+\cup S_-} L_{t-1}(i), \\
\text{i.e, } \sum_{i\in S_+} L_{t-1}(i) - \sum_{i\in S_-} e^{\{f_{t-1}(i,h_t(x_i))-f_{t-1}(i,1)\}} &\geq \delta_t \sum_{i\in S_+\cup S_-} L_{t-1}(i).
\end{aligned}$$

On the other hand, the drop in loss after choosing h_t with weight α_t is

$$\begin{aligned}
&\sum_{i\in S_+}\left(1-e^{-\alpha_t}\right)L_{t-1}(i)-\sum_{i\in S_-}\left(e^{\alpha_t}-1\right)e^{\{f_{t-1}(i,h_t(x_i))-f_{t-1}(i,1)\}}\\
&= \left(\frac{e^{\alpha_t}-e^{-\alpha_t}}{2}\right)\left\{\sum_{i\in S_+}L_{t-1}(i)-\sum_{i\in S_-}e^{\{f_{t-1}(i,h_t(x_i))-f_{t-1}(i,1)\}}\right\}\\
&- \left(\frac{e^{\alpha_t}+e^{-\alpha_t}-2}{2}\right)\left\{\sum_{i\in S_+}L_{t-1}(i)+\sum_{i\in S_-}e^{\{f_{t-1}(i,h_t(x_i))-f_{t-1}(i,1)\}}\right\}.
\end{aligned}$$

Now $e^{\{f_{t-1}(i,h_t(x_i))-f_{t-1}(i,1)\}}$ is upper bounded by $L_{t-1}(i)$, so that the second term in curly-brackets is upper bounded by the loss after $t-1$ rounds. We have already shown the first term in curly brackets is at least δ_t times the loss after $t-1$ rounds. Hence the loss in round t is at most a factor $1-\frac{1}{2}(e^{\alpha_t}-e^{-\alpha_t})\delta_t+\frac{1}{2}(e^{\alpha_t}+e^{-\alpha_t}-2)$ of the loss in round $t-1$. □

Corollary. *Suppose $\mathbf{C}_t$ is chosen as in (7). Then if h_t beats $\mathbf{U}_{\delta_t}$, for some $\delta_t\in[0,1]$, on $\mathbf{C}_t$, then for any $\alpha_t>0$, there is a $\gamma_t\in[1-k,1]$ such that*

- *h_t beats $\mathbf{U}_{\gamma_t}$ on $\mathbf{C}_{\alpha_t}$, where $\mathbf{C}_{\alpha_t}$ is defined as in (7), and*
- *$\kappa(\gamma_t,\alpha_t)\le g(\alpha_t,\delta_t)\triangleq 1-\frac{1}{2}(e^{\alpha_t}-e^{-\alpha_t})\delta_t+\frac{1}{2}(e^{\alpha_t}+e^{-\alpha_t}-2)$.*

Proof. Recall $\kappa(\gamma_t,\alpha_t)=1+\frac{1-\gamma_t}{k}\left(e^{\alpha_t}-e^{-\alpha_t}\right)-\gamma_t(1-e^{-\alpha_t})$. If $g(\alpha_t,\delta_t)>\sup_{\gamma_t\in[1-k,1]}\kappa(\gamma_t,\alpha_t)$, then the choice of $\gamma_t=1-k$ satisfies the requirements in the statement of the corollary. Otherwise observe

$$\kappa(0,\alpha_t)=e^{-\alpha_t}\le g(\alpha_t,\delta_t),$$

so that, by continuity of κ, we may pick a value of γ_t such that $\kappa(\gamma_t,\alpha_t)=g(\alpha_t,\delta_t)$. As before, define $L_t(i)=\sum_{l=2}^{k}e^{\{f_t(i,l)-f_t(i,1)\}}$. By expanding out one may see

$$\sum_{i=1}^{m}L_{t-1}(i)+\alpha_t\mathbf{C}_{\alpha_t}\bullet\mathbf{U}_{\gamma_t}=\kappa(\gamma_t,\alpha_t)\sum_{i=1}^{m}L_{t-1}(i).$$

Similarly one may verify,

$$\sum_{i=1}^{m}L_{t-1}(i)+\alpha_t\mathbf{C}_{\alpha_t}\bullet\mathbf{1}_{h_t}=\sum_{i=1}^{m}L_t(i).$$

The previous lemma yields $\sum_{i=1}^{m}L_t(i)\le g(\alpha_t,\delta_t)\sum_{i=1}^{m}L_{t-1}(i)=\kappa(\gamma_t,\alpha_t)\sum_{i=1}^{m}L_{t-1}(i)$. This shows h_t beats $\mathbf{U}_{\gamma_t}$ on $\mathbf{C}_{\alpha_t}$. □

References

[1] Yoav Freund and Robert 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, August 1997.

[2] R. Tyrell Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

[3] Robert E. Schapire. Drifting games. *Machine Learning*, 43(3):265–291, June 2001.

[4] Robert E. Schapire and Yoram Singer. Improved boosting algorithms using confidence-rated predictions. *Machine Learning*, 37(3):297–336, December 1999.