THE WEAK-LEARNING GAME OF FACTORIZED ADABOOST.MH: THE EXACT VALUE OF AN INSTANCE, AND THE EFFECTIVE CLASS NUMBER

KOREA SUPERINTELLIGENCE LABS

Abstract. Let W range over the nonnegative $n \times K$ matrices of total mass one, let Y^ℓ be the signed one-hot label matrix of a label map $\ell : [n] \to [K]$, and let v range over $\{\pm 1\}^K$. The quantity $\max_v \|(W \odot Y^\ell)v\|_1$ is the binary weight mass that a factorized multi-class base classifier can induce, and Kégl asked in 2014 whether it admits a positive lower bound independent of n. Zou and Xu have recently determined its minimum over *both* W and ℓ exactly, $\mathfrak{W}_{n,K} = C_{\min\{n+1,K\}}$ with $C_q = q/(3q-4)$ for even $q \geq 2$ and $C_q = (q+1)/(3q-1)$ for odd $q \geq 3$, and they conjecture that some *effective class number* governs the hardness contributed by the labels. We determine the value of a *fixed* instance. For every n, every K and every label map using exactly m distinct labels with $2 \leq m \leq 9$,

$$\min_W \max_v \|(W \odot Y^\ell)v\|_1 \;=\; C_{\min\{m+1,\,K\}}.$$

The value therefore depends on the instance only through the *number* of labels it uses: not on the number of examples, and not on how the examples are distributed among those labels. This identifies the effective class number of an instance as $\min\{m+1, K\}$ and contains $\mathfrak{W}_{n,K} = C_{\min\{n+1,K\}}$ as the case $m = \min\{n, K\}$. The lower half of the identity is already contained, unstated, in the refined lower-bound argument of the source; what is new is the matching *upper* bound at every m, obtained by re-instantiating the source's own two weight-matrix constructions at m, and the exact value and its two independence properties that follow. The lower bounds are certified by optimal maximiser strategies of the underlying ± 1 matrix game, in closed form, to which the source's analytic argument has no counterpart. All the finite content is machine-checked in Lean 4; Section 8 says exactly what is and what is not formalised.

1. Introduction

The objects. Fix $n \geq 1$ examples and $K \geq 1$ classes, and write $[n] = \{1, \ldots, n\}$. AdaBoost.MH [5] reduces multi-class classification to K binary problems by carrying a weight $W_{ij} \geq 0$ on each example–class pair (i, j) together with the *signed one-hot* label matrix Y: row i of Y is $+1$ in the true-class column of example i and -1 in every other column. Following the source, we write

$$\mathcal{W}_{n,K} = \Big\{ W \in \mathbb{R}^{n \times K}_{\geq 0} \;:\; \textstyle\sum_{i,j} W_{ij} = 1 \Big\}, \qquad \mathcal{V}_K = \{\pm 1\}^K,$$

and $\mathcal{Y}_{n,K}$ for the set of signed one-hot label matrices. *Factorized* AdaBoost.MH [4] uses base classifiers of the form $h(x) = \alpha\, v\, \varphi(x)$: a single binary rule φ shared across the classes, with all class dependence carried by a *vote vector* $v \in \mathcal{V}_K$. Fixing v collapses the weighted example–class problem back to a binary problem on the examples, but only with the induced total mass

$$w_\Sigma(W, Y, v) \;=\; \big\|(W \odot Y)\, v\big\|_1 \;=\; \sum_{i=1}^{n} \Big| \sum_{j=1}^{K} W_{ij}\, Y_{ij}\, v_j \Big|, \tag{1}$$

where $\odot$ is the entrywise product. A weak-learning guarantee for the factorized rule is therefore useful only if *some* vote vector keeps (1) away from zero, for every weight matrix and every labelling; the worst case over both is

$$\mathfrak{W}_{n,K} \;:=\; \min_{W \in \mathcal{W}_{n,K},\, Y \in \mathcal{Y}_{n,K}} \; \max_{v \in \mathcal{V}_K} \; w_\Sigma(W, Y, v). \tag{2}$$

A label matrix $Y \in \mathcal{Y}_{n,K}$ is the same datum as a *label map* $\ell : [n] \to [K]$, namely $Y^{\ell}_{ij} = +1$ if $j = \ell(i)$ and -1 otherwise, and we write $w_\Sigma(W, \ell, v)$ for (1) with $Y = Y^{\ell}$. The object this paper is about is the value of a *single* instance,

$$V(\ell) \;:=\; \min_{W \in \mathcal{W}_{n,K}} \; \max_{v \in \mathcal{V}_K} \; w_\Sigma(W, \ell, v), \qquad \text{so} \qquad \mathfrak{W}_{n,K} = \min_{\ell : [n] \to [K]} V(\ell), \tag{3}$$

and about the single statistic of ℓ on which it turns out to depend, the number

$$m(\ell) \;=\; \left|\operatorname{im} \ell\right| \;=\; \#\{\text{classes that some example actually has}\}.$$

The source, and what it leaves open. Kégl [3] posed the question for the factorized rule. In his own words, verbatim ([3, Remark 3]; his w'_Σ is our w_Σ):

> "So the question is whether there exists a setup (X, W, Y, and function class) in which all of the 2^K different vote vectors $v = \{\pm 1\}^K$ lead to arbitrarily small (or zero) w'_Σ, or we can find a constant (independent of n) lower bound ω such that with at least one vote vector v and classifier ϕ, $w'_\Sigma \geq \omega$ holds."

The body of the same paper states the target as "*to find a lower bound on w'_Σ which may depend on the number of classes K but is independent of the number of training points n*"; the source [1, §2] quotes the question in a lightly reworded form. Zou, Zhou, Xu and Liu [2] answered the first half in the negative, with $\mathfrak{W}_{n,K} \geq \max\{1/n,\, 1/\sqrt{2K}\}$ (their Theorems 3.3 and 3.4), which still degrades in both parameters. Zou and Xu [1] then computed (2) exactly. Their Theorem 3.5, verbatim: *"For integers $n \geq 1$ and $K \geq 2$, we have: $\mathfrak{W}_{n,K} = C_{\min\{n+1,K\}}$"*, where

$$C_1 = 1, \qquad C_q = \frac{q}{3q-4} \;\; (q \geq 2 \text{ even}), \qquad C_q = \frac{q+1}{3q-1} \;\; (q \geq 2 \text{ odd}). \tag{4}$$

So the answer to Kégl's question is a constant independent of *both* n and K: the sequence (4) decreases to $1/3$ and never reaches it.

What (2) does not say is how hard a *particular* problem is. The minimum over Y in (2) is attained at a label map that spreads the examples over as many classes as possible, and the source notices exactly this gap, writing ([1, §3], verbatim):

> "However, when $n < K$, the n examples can not use up all K labels. So we conjecture that there might be **effective class number** that really captures the intrinsic hardness caused by the labels."

Their theorem pins that number to $\min\{n+1, K\}$, a function of the two size parameters alone. An instance with a thousand examples in three classes and an instance with three examples in three classes get the same answer only if n is the right statistic — and it is not.

What this paper settles.

- *The exact value of an instance* (Theorem 4.2). For every n, every K and every label map ℓ with $m = m(\ell)$ and $2 \leq m \leq 9$,
$$V(\ell) \;=\; C_{\min\{m+1,\,K\}}.$$
Both halves are proved: no weight matrix does better, and an explicit one attains the value.
- *Two independence properties, and the effective class number* (Corollary 4.3). In that range $V(\ell)$ does not depend on n and does not depend on the block sizes $|\ell^{-1}(c)|$: two instances with the same number of used labels and the same K have the same value, whatever their sizes and however unbalanced they are. The effective class number of an instance is $\min\{m(\ell) + 1,\, K\}$, and $\mathfrak{W}_{n,K} = C_{\min\{n+1,K\}}$ is the case $m = \min\{n, K\}$.
- *Optimal maximiser strategies in closed form* (Section 3). The lower half rests on a dual mixed strategy for the underlying ± 1 matrix game which is exactly tight — all its correlation constraints hold with equality — and which is described in closed form at every m (Table 1). Eighteen instances of it, at $m \leq 9$, are verified as integer certificates. The bound they certify is the source's, as the next paragraph explains; the strategies are not in it.

- *The source's own theorem, machine-checked* (Theorem 5.1): $\mathfrak{W}_{n,K} = C_{\min\{n+1,K\}}$ for every $n \geq 1$ and $K \geq 1$ with $\min\{n, K\} \leq 9$. This is published mathematics; only the certification is ours.
- *Negative controls* (Section 6), including that the value of an instance is unique, so that a wrong value is refuted rather than merely unproved.

Attribution. Three things must be said plainly, and they are said again where they are used.

First, the *lower* half of Theorem 4.2 is the source's Appendix B.2 in all but name. That appendix argues for an arbitrary *fixed* pair (W, Y): it sets $I_+ = \{i : \sum_j W_{ij} > 0\}$, $S = \{c_i : i \in I_+\}$, $r = |S|$, $q = r + 1$, and concludes, verbatim, "*By the proof of the lower bounds of Theorem 3.1, we have* $\max_{u\in\mathcal{V}_q} w_\Sigma(\bar{W}, \bar{Y}, u) \geq C_q$. *So we can conclude that when* $r < K$, $\max_{v\in\mathcal{V}_K} w_\Sigma(W, Y, v) \geq C_q \geq C_{\min\{n+1,K\}}$" — where only the *last* inequality minimises over Y, the case $r = K$ having been settled a few lines earlier by $\mathfrak{W}_{n,K} \geq C_K$. Since $r \leq m(\ell)$ and C is non-increasing, the per-instance bound $V(\ell) \geq C_{\min\{m+1,K\}}$ follows from their text by bookkeeping alone. Proposition 3.1 is that bound; it is stated as a re-derivation, with an explicit maximiser strategy the source does not exhibit, and not as a new result.

Second, what is new is the matching per-instance *upper* bound at every m, and hence the exact value and its independence from n and from the block sizes. The source's Appendix C.1 constructs a weight matrix only for the particular label map with $\min\{n, K\}$ distinct labels, and its Appendix C.2 only for n even with $n + 1 \leq K$; neither pins the value of an arbitrary instance. The ingredients of Proposition 4.1 are nevertheless theirs: it is their two tables, re-instantiated at m in place of $\min\{n, K\}$, on a set of representatives of the used labels.

Third, *effective class number* is the source's phrase, and the conjecture that such a quantity exists is theirs; what we add is its identification with $\min\{m + 1, K\}$ at the level of an instance.

Relation to game-theoretic boosting. That boosting has a minimax reading is old: Freund and Schapire [6] present boosting as a repeated two-player game, Rätsch and Warmuth [7] and Shalev-Shwartz and Singer [8] use von Neumann duality to characterise weak learnability against a convex hull of hypotheses, and Mukherjee and Schapire [9] build the multiclass theory on *edge-over-random* conditions that are themselves minimax constraints. None of these computes a constant of the shape (4), and none computes the value or the optimal strategies of the game (3); a search of the literature accessible to us found the constants (4) only in [1]. One point of contact should be disclosed: [2] already randomises the vote vector. Its proof of Theorem 3.4 bounds $\max_v w_\Sigma$ below by $\mathbb{E}_{v\sim D}\, w_\Sigma$ "*for any distribution D on $\mathcal{V}$*", takes the coordinates of v to be independent Rademacher variables, and finishes with Khintchine's inequality. That is a distribution over v alone: no sign pattern on the examples is randomised, and the resulting bound $1/\sqrt{2K}$ is not tight. The strategies of Section 3 are joint distributions over the pair (sign pattern, vote vector) and are exactly tight, which is what makes them optimal rather than merely feasible.

2. Preliminaries

Notation. Throughout, $n \geq 1$ and $K \geq 1$ are integers, $\ell : [n] \to [K]$ is a label map, $Y^\ell_{ij} = +1$ if $j = \ell(i)$ and -1 otherwise, and $w_\Sigma(W, \ell, v)$ is (1). We write $m = m(\ell) = |\operatorname{im} \ell|$; note $1 \leq m \leq \min\{n, K\}$. Since $|\sum_j W_{ij} Y_{ij} v_j| \leq \sum_j W_{ij}$ for $v \in \mathcal{V}_K$, every instance satisfies $w_\Sigma(W, \ell, v) \leq 1$, so $V(\ell) \leq 1$ always.

All the statements below are proved in the two-sided form

$$(5) \qquad \big(\forall W \in \mathcal{W}_{n,K}\ \exists v \in \mathcal{V}_K : w_\Sigma(W, \ell, v) \geq c\big) \quad \text{and} \quad \big(\exists W \in \mathcal{W}_{n,K}\ \forall v \in \mathcal{V}_K : w_\Sigma(W, \ell, v) \leq c\big),$$

which is what "$V(\ell) = c$" abbreviates; no compactness or attainment argument is needed, and (5) determines c uniquely (Proposition 6.1).

The constants.

Proposition 2.1. *Let C_q be given by (4), extended by $C_0 = 1$. Then for every $q \geq 1$,*

$$C_q \;=\; \frac{t}{3t-2}, \qquad t = \lceil q/2 \rceil,$$

and: C is non-increasing; $C_{q+1} = C_q$ for odd q; and $1/3 < C_q \leq 1$ for every q. The first values are

q	1	2	3	4	5	6	7	8	9	10
C_q	1	1	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{3}{7}$	$\frac{3}{7}$	$\frac{2}{5}$	$\frac{2}{5}$	$\frac{5}{13}$	$\frac{5}{13}$

The three properties are the source's own Proposition 3.2 and the remark following its Theorem 3.5 [1]; they are restated here because every statement below is phrased in terms of C, and because the one-case form $t/(3t-2)$, which merges the source's three cases, is what makes the parity collapse and the monotonicity immediate. That form is an algebraic rewriting of (4), not a result.

Proof. Write $t = \lceil q/2 \rceil$. For even $q \geq 2$, $t = q/2$ and $q/(3q-4) = 2t/(6t-4) = t/(3t-2)$; for odd $q \geq 3$, $t = (q+1)/2$ and $(q+1)/(3q-1) = 2t/(6t-4) = t/(3t-2)$; and $t = 1$ gives 1 at $q \in \{1, 2\}$. The map $t \mapsto t/(3t-2)$ is decreasing on $t \geq 1$ with value 1 at $t = 1$ and limit $1/3$, and $\lceil q/2 \rceil$ is non-decreasing in q and constant on $\{q, q+1\}$ for odd q. $\square$

The instance as a ± 1 matrix game. Because $\|x\|_1 = \max_{s \in \{\pm 1\}^n} \langle s, x \rangle$, the inner maximisation in (3) can be enlarged to a maximisation over pairs:

$$V(\ell) \;=\; \min_{W \in \mathcal{W}_{n,K}} \; \max_{(s,v) \in \{\pm 1\}^n \times \mathcal{V}_K} \; \sum_{i,j} s_i \, v_j \, Y^{\ell}_{ij} \, W_{ij}. \tag{6}$$

This is the value of a finite matrix game. The minimiser mixes over the nK cells (i,j) — the set of mixed strategies *is* the simplex $\mathcal{W}_{n,K}$ — the maximiser plays a pair (s, v), and the payoff is $s_i v_j Y^{\ell}_{ij} \in \{\pm 1\}$. In particular $V(\ell)$ is a rational number, and the maximiser side of the game reads: a mixed strategy p over pairs certifies $V(\ell) \geq c$ as soon as its induced correlations satisfy $Y^{\ell}_{ij} \, \mathbb{E}_p[s_i v_j] \geq c$ at every cell (i,j). That implication is the only form of duality we use, and it is elementary.

Proposition 2.2 (Weak duality)**.** *Let $W \in \mathbb{R}^{n \times K}$ be nonnegative with $\sum_{i,j} W_{ij} = 1$, let $A \in \mathbb{R}^{n \times K}$, and let p be a probability vector on N pairs $(s^k, v^k) \in \{\pm 1\}^n \times \mathbb{R}^K$. If*

$$A_{ij} \sum_k p_k \, s^k_i v^k_j \;\geq\; c \qquad \text{for every } (i,j),$$

then $\sum_i \bigl| \sum_j W_{ij} A_{ij} v^k_j \bigr| \geq c$ for at least one k.

Proof. Put $F_k = \sum_{i,j} W_{ij} A_{ij} s^k_i v^k_j$. Exchanging the order of summation and using $W \geq 0$ together with the hypothesis cellwise,

$$\sum_k p_k F_k \;=\; \sum_{i,j} W_{ij} \Bigl(A_{ij} \sum_k p_k s^k_i v^k_j \Bigr) \;\geq\; c \sum_{i,j} W_{ij} \;=\; c,$$

so $F_k \geq c$ for some k. Finally, since $s^k_i \in \{\pm 1\}$,

$$F_k \;=\; \sum_i s^k_i \Bigl(\sum_j W_{ij} A_{ij} v^k_j \Bigr) \;\leq\; \sum_i \Bigl| \sum_j W_{ij} A_{ij} v^k_j \Bigr|. \qquad \square$$

Applied with $A = Y^{\ell}$, Proposition 2.2 turns a maximiser strategy into a lower bound on w_Σ valid for *every* weight matrix at once. What makes such a bound usable at all n and K simultaneously is a compression of the strategy space, and it is worth isolating because everything below rests on it.

Remark 2.3 (Two compressions). In the game (6) the payoff at cell (i,j) depends on the row i only through the label $\ell(i)$, and every column that is nobody's label sees the same entry -1 from every row. So it costs nothing to restrict the maximiser to strategies in which (i) s_i depends on i only through $\ell(i)$ — a sign on the m *used labels* rather than on the n examples — and (ii) v_j depends on j only through which used label j is, all unused columns carrying one common sign. Such a strategy lives on m *label slots* and $M \in \{m, m+1\}$ *column slots*, and mentions neither n nor K.

Definition 2.4. A *dual certificate* D consists of integers $m, M \geq 1$, a finite list of *atoms* (p_a, σ_a, τ_a) with $p_a \in \mathbb{Z}_{\geq 0}$ a weight and $\sigma_a \in \{\pm 1\}^m$, $\tau_a \in \{\pm 1\}^M$ sign patterns on the label and column slots, a denominator $d = \sum_a p_a > 0$, and a claimed value $c = \nu/\delta$ with $\delta > 0$. Writing $\operatorname{corr}(b,j) = \sum_a p_a\, \sigma_a(b)\, \tau_a(j)$, the certificate *passes* when

$$\nu\, d \;\leq\; \delta \cdot \operatorname{corr}(b,b) \quad \text{and} \quad \nu\, d \;\leq\; -\,\delta \cdot \operatorname{corr}(b,j) \qquad (b \in [m],\; j \in [M],\; j \neq b),$$

that is, when the maximiser's correlation with the true-class slot is at least c and with every other column slot at most $-c$.

Lemma 2.5 (Certificate soundness)**.** *Let D be a certificate that passes, with value c.*

(i) *Let $\ell : [n] \to [K]$ and let $g : [K] \to \{1, \dots, M\}$ send each used column to its own label slot, distinct used columns to distinct slots, and every column to a slot $\leq M$. Then for every $W \in \mathcal{W}_{n,K}$ there is $v \in \mathcal{V}_K$ with $w_\Sigma(W, \ell, v) \geq c$. In particular this holds for every n and every K: if $M = m+1$ it suffices that ℓ use at most m labels, and if $M = m$ that in addition ℓ be onto $[K]$.*

(ii) *Dually, let $r_1, \dots, r_m$ be examples carrying m distinct labels $c_1, \dots, c_m$, let $c_1, \dots, c_M$ be distinct columns extending them, and let $\omega \in \mathbb{Z}_{\geq 0}^{m \times M}$ with $\sum_{a,b} \omega_{ab} = d' > 0$. If*

$$\delta' \sum_{a=1}^{m} \Big| \sum_{b=1}^{M} \omega_{ab}\, y_{ab}\, e_b \Big| \;\leq\; \nu'\, d' \qquad \text{for all } e \in \{\pm 1\}^M, \qquad y_{ab} = \begin{cases} +1, & b = a \\ -1, & b \neq a, \end{cases}$$

then the matrix W with $W_{r_a c_b} = \omega_{ab}/d'$ and zeros elsewhere lies in $\mathcal{W}_{n,K}$ and satisfies $w_\Sigma(W, \ell, v) \leq \nu'/\delta'$ for every $v \in \mathcal{V}_K$.

Proof. (i) Transport the atoms along g: put $s_i^a = \sigma_a(g(\ell(i)))$ and $v_j^a = \tau_a(g(j))$, and p_a/d for the probabilities. The hypotheses on g say exactly that at a cell (i,j) with $j = \ell(i)$ the pair of slots is (b,b), and at a cell with $j \neq \ell(i)$ it is (b, j') with $j' \neq b$; so the two families of certificate inequalities are precisely $Y^\ell_{ij} \mathbb{E}[s_i v_j] \geq c$ cellwise, and Proposition 2.2 applies with $A = Y^\ell$. (ii) W is nonnegative with total mass 1 by construction, the rows outside $\{r_1, \dots, r_m\}$ contribute nothing to (1), and on the row r_a only the columns $c_1, \dots, c_M$ carry weight, so $w_\Sigma(W, \ell, v)$ equals the left-hand side of the displayed inequality divided by d' with $e_b = v_{c_b}$. The inequality is assumed for all 2^M patterns e. □

3. The value of an instance from below

This section re-derives, with an explicit optimal strategy, a bound that is already contained in [1, Appendix B.2]; see the Attribution paragraph of Section 1. It is stated because Theorem 4.2 needs it in per-instance form and because the maximiser strategies are the objects the machine verification actually carries.

Proposition 3.1. *Let $1 \leq m \leq 9$. For every n, every K, every label map $\ell : [n] \to [K]$ using at most m distinct labels and every $W \in \mathcal{W}_{n,K}$ there is a vote vector $v \in \mathcal{V}_K$ with*

$$w_\Sigma(W, \ell, v) \;\geq\; C_{m+1};$$

and if moreover every class is somebody's label, that is ℓ is onto $[K]$, then there is $v \in \mathcal{V}_K$ with $w_\Sigma(W, \ell, v) \geq C_m$.

Applied at $m = m(\ell)$ this gives $\max_v w_\Sigma(W, \ell, v) \geq C_{\min\{m(\ell)+1,\, K\}}$ for every W: when $K > m(\ell)$ the first bound applies, and when $K = m(\ell)$ the second does. For $m > m(\ell)$ the statement is weaker, C being non-increasing.

The maximiser strategies. The certificates are instances of one closed-form strategy, which is what the proof of Proposition 3.1 really uses. Let $\sigma \in \{\pm 1\}^m$ be the sign carried by the used labels (so $s_i = \sigma_{\ell(i)}$), let the vote vector be $v = \sigma$ on the used columns and $v = \varepsilon$ on the unused ones, and let $a \in [0, 1)$. The strategy is:

- with probability a, play the *anti* atom $\sigma \equiv +1$, $v \equiv -1$;
- otherwise draw σ uniformly from a prescribed level set of $\sum_c \sigma_c$ and put $\varepsilon = -\operatorname{sign}(\sum_c \sigma_c)$ on the unused columns, splitting the two signs evenly when $\sum_c \sigma_c = 0$.

TABLE 1. The optimal maximiser strategies. In each case all three correlations $\mathbb{E}[\sigma_c v_c]$, $\mathbb{E}[\sigma_c v_j]$ ($j \neq c$ a used column) and $\mathbb{E}[\sigma_c \varepsilon]$ equal $+C$, $-C$, $-C$ exactly.

case	a	law of σ off the anti atom	value C
m odd	$\dfrac{m-1}{3m-1}$	uniform on $\lvert\sum_c \sigma_c\rvert = 1$	$C_m = C_{m+1}$
m even, $K = m$	$\dfrac{m-2}{3m-4}$	uniform on $\sum_c \sigma_c = 0$	C_m
m even, $K > m$	$\dfrac{m}{3m+2}$	$\sum_c \sigma_c = 0$ w.p. $\dfrac{m+2}{2m+2}$, $\lvert\sum_c \sigma_c\rvert = 2$ w.p. $\dfrac{m}{2m+2}$	C_{m+1}

Table 1 records the three cases. That all three correlations come out as *equalities* is what makes each strategy optimal and not merely feasible: a feasible primal of the same value appears in Section 4, and the two match. The source's lower-bound argument is analytic — it tracks the true-class mass $\rho = \sum_i W_{i\ell(i)}$ and combines the all-ones vote vector with a balanced family of vote vectors — and exhibits no maximiser strategy, so Table 1 has no counterpart there.

Proof of Proposition 3.1, and what was computed.

Proof sketch. For each $m \leq 9$ the strategy of Table 1 is written out as an integer certificate in the sense of Definition 2.4: E_m with $M = m + 1$ column slots and claimed value C_{m+1}, and F_m with $M = m$ and claimed value C_m. Each is checked to pass, which is the evaluation of $m \cdot M$ integer inequalities, each a sum over the atom list, plus the nonnegativity and total-weight conditions. Lemma 2.5(i), applied with the slot map that sends a used column to its rank among the used columns and every unused column to the extra slot, then gives the two conclusions for all n and K at once. □

The computation is exhaustive over a finite, explicitly bounded index set — there is no search inside the verification. The eighteen certificates are:

m	1	2	3	4	5	6	7	8	9
atoms of E_m	1	7	7	21	21	71	71	253	253
value C_{m+1}	1	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{3}{7}$	$\frac{3}{7}$	$\frac{2}{5}$	$\frac{2}{5}$	$\frac{5}{13}$	$\frac{5}{13}$
atoms of F_m	1	2	7	7	21	21	71	71	253
value C_m	1	1	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{3}{7}$	$\frac{3}{7}$	$\frac{2}{5}$	$\frac{2}{5}$	$\frac{5}{13}$

The atom counts are those of Table 1 after collecting equal atoms: for odd m the $\binom{m}{(m\pm1)/2}$ patterns with $\lvert\sum\sigma\rvert = 1$ and the anti atom; for even m with $K > m$ the $\binom{m}{m/2}$ balanced patterns, each split in two by the sign ε, the patterns with $\lvert\sum\sigma\rvert = 2$, and the anti atom. So, for instance, E_4 has $2\binom{4}{2} + \binom{4}{3} + \binom{4}{1} + 1 = 21$ atoms of total weight 28, of which the anti atom carries weight $8 = 28 \cdot \frac{m}{3m+2}$. The certificates were produced outside the verification, by solving the game (6) as a linear program at small m and averaging the optimum over the symmetry group $S_m \times S_{K-m}$; only the passing of the check is used inside, so no property of that generator is relied upon.

4. The value of an instance from above, and the exact value

Proposition 4.1. *Let $2 \le m \le 9$ and let $\ell : [n] \to [K]$ use exactly m distinct labels $c_1, \dots, c_m$, with representatives $r_1, \dots, r_m$, $\ell(r_a) = c_a$.*

(i) *Let $a_0 = \dfrac{m-2}{3m-4}$ for even m and $a_0 = \dfrac{m-1}{3m-1}$ for odd m, and let W^{d} put*

$$W^{\mathrm{d}}_{r_a c_a} = \frac{a_0}{m}, \qquad W^{\mathrm{d}}_{r_a c_b} = \frac{1-a_0}{m(m-1)} \quad (b \ne a),$$

and zero in every other cell. Then $W^{\mathrm{d}} \in \mathcal{W}_{n,K}$ and $w_\Sigma(W^{\mathrm{d}}, \ell, v) \le C_m$ for every $v \in \mathcal{V}_K$.

(ii) *Suppose $K > m$ and m is even, let c_0 be a column that no example uses, put $D = 3m+2$, and let W^+ put*

$$W^{+}_{r_a c_a} = \frac{1}{D}, \qquad W^{+}_{r_a c_b} = \frac{2}{mD} \quad (b \ne a), \qquad W^{+}_{r_a c_0} = \frac{4}{mD},$$

and zero in every other cell. Then $W^+ \in \mathcal{W}_{n,K}$ and $w_\Sigma(W^+, \ell, v) \le C_{m+1}$ for every $v \in \mathcal{V}_K$.

Consequently $V(\ell) \le C_{\min\{m+1,K\}}$: for $K = m$ use (i); for $K > m$ and m even use (ii); for $K > m$ and m odd use (i) again, since $C_{m+1} = C_m$.

The two tables are the source's Appendices C.1 and C.2 with m in place of $\min\{n, K\}$, put on a set of representatives of the used labels rather than on the first $\min\{n, K\}$ examples. What this buys is the arbitrary instance: the source's constructions choose the label map, and the choice of representatives, together with the spare column c_0 in (ii), is what removes that choice.

Proof sketch. Both tables are integer tables after scaling — by $m(m-1)q$ in (i), where $a_0 = p/q$, and by $m(3m+2)$ in (ii) — with row sums $(m-1)q$, resp. $3m+2$, so the total mass is 1. Lemma 2.5(ii) then reduces the claim to the finite check over the 2^M sign patterns of the vote vector on the columns $c_1, \dots, c_M$, with $M = m$ in (i) and $M = m+1$ in (ii); the vote vector's values on the remaining columns do not enter, since those cells carry no weight. Each check is exhaustive over $2^M \le 2^{10} = 1024$ patterns, each a sum of m absolute values of M-term sums, and is carried out for each $m \le 9$ separately. The representatives and, in (ii), the unused column exist because $|\operatorname{im} \ell| = m$ and, in (ii), $m < K$. □

Theorem 4.2 (The exact value of an instance)**.** *Let $n \ge 1$ and $K \ge 1$, and let $\ell : [n] \to [K]$ be a label map using exactly m distinct labels, $2 \le m \le 9$. Then*

$$V(\ell) \;=\; \min_{W \in \mathcal{W}_{n,K}} \; \max_{v \in \mathcal{V}_K} \left\| (W \odot Y^\ell) v \right\|_1 \;=\; C_{\min\{m+1,\,K\}}.$$

Explicitly: for every $W \in \mathcal{W}_{n,K}$ some vote vector achieves at least $C_{\min\{m+1,K\}}$, and the weight matrix of Proposition 4.1 keeps every vote vector at or below it.

Proof. Since $\operatorname{im} \ell \subseteq [K]$ we have $m \le K$. If $K = m$ then ℓ is onto and Proposition 3.1 gives the bound $C_m = C_{\min\{m+1,K\}}$ from below, and Proposition 4.1(i) gives C_m from above. If $K > m$ then Proposition 3.1 gives $C_{m+1} = C_{\min\{m+1,K\}}$ from below, and Proposition 4.1(ii) — or (i) together with $C_{m+1} = C_m$ when m is odd — gives it from above. □

Corollary 4.3 (Independence, and the effective class number)**.** *Fix K and let $2 \le m \le 9$. Then all label maps into $[K]$ that use exactly m labels have the same value $C_{\min\{m+1,K\}}$, whatever their number of examples n and whatever the block sizes $|\ell^{-1}(c)|$. In particular the effective class number of an instance — the index at which the sequence (4) must be read to give its value — is*

$$q(\ell) \;=\; \min\big\{ \, |\operatorname{im} \ell| + 1, \; K \, \big\},$$

and the source's $\mathfrak{W}_{n,K} = C_{\min\{n+1,K\}}$ *is the case* $m = \min\{n, K\}$*, in which* $\min\{m+1, K\} = \min\{n+1, K\}$.

Proof. The first sentence is Theorem 4.2, whose right-hand side mentions only m and K. For the last: if $n \geq K$ then $\min\{n, K\} = K$ and both expressions equal K; if $n < K$ then $\min\{n, K\} = n$ and both equal $n + 1$. □

Neither independence was predictable from (2). The value moves with the number of parts of the label partition but not with their sizes: at $n = 6$, $K = 5$ the partitions $(2, 2, 2)$ and $(2, 2, 1, 1)$ — the same six examples, the same five classes — have values $C_4 = 1/2$ and $C_5 = 3/7$, differing exactly because one uses three labels and the other four, while $(3, 1)$ and $(2, 2)$ at $n = 4$ agree. Section 7 records the exhaustive rational computation in which this pattern was found.

Remark 4.4 ($m = 1$). The excluded case is degenerate rather than open: if every example has the same label then $C_{\min\{2,K\}} = 1$ and $V(\ell) = 1$, since $w_\Sigma \leq 1$ always and Proposition 3.1 at $m = 1$ gives $w_\Sigma \geq C_2 = 1$ for some vote vector.

5. The minimum over instances

For completeness, and as a check on the certificate route, the per-instance statements assemble into the source's own theorem. This is published mathematics [1, Theorem 3.5], proved there for all n and K; what is added here is the certification, and only on the range of the finite certificates.

Theorem 5.1. *Let* $n \geq 1$ *and* $K \geq 1$ *with* $\min\{n, K\} \leq 9$*. Then* $\mathfrak{W}_{n,K} = C_{\min\{n+1,K\}}$*; that is, for every label map* $\ell : [n] \to [K]$ *and every* $W \in \mathcal{W}_{n,K}$ *some vote vector achieves* $w_\Sigma(W, \ell, v) \geq C_{\min\{n+1,K\}}$*, and there are a label map and a weight matrix for which no vote vector exceeds it.*

Proof. Lower half: a label map uses $m(\ell) \leq \min\{n, K\} \leq 9$ labels, so Proposition 3.1 gives $\max_v w_\Sigma \geq C_{\min\{m(\ell)+1,K\}} \geq C_{\min\{n+1,K\}}$, the last step by monotonicity of C and $m(\ell) \leq n$. Upper half: the label map $i \mapsto \min\{i, \min\{n, K\}\}$ uses exactly $\min\{n, K\}$ labels, so Theorem 4.2 (or, at $\min\{n, K\} = 1$, the preceding remark) applies to it and Corollary 4.3 identifies its value as $C_{\min\{n+1,K\}}$. □

Two small differences from the source's statement are worth recording. It assumes $K \geq 2$, while the statement above also covers $K = 1$, where both sides are 1. And the range $\min\{n, K\} \leq 9$ is a restriction the source does not have: within it, n and K are otherwise unrestricted, so for example every n with $K \leq 9$ classes is covered, but $n = K = 10$ is not.

6. Negative controls

A value statement of the form (5) could in principle be satisfied vacuously, or by a certificate check that passes for the wrong reason. The following are the controls; the first is what makes the others genuine refutations rather than mere absences of proof.

Proposition 6.1. *For a fixed instance* ℓ*, at most one real number* c *satisfies* (5)*.*

Proof. Let c, c' both satisfy it, with W' the matrix witnessing the upper half for c'. The lower half for c applied to W' gives v with $c \leq w_\Sigma(W', \ell, v) \leq c'$; symmetrically $c' \leq c$. □

Proposition 6.2. *Let* $\ell_0 : [4] \to [5]$ *be injective — four examples in four of five classes, so* $m = 4$ *and* $V(\ell_0) = C_5 = 3/7$ *by Theorem 4.2. Then:*

(i) (a value one step too large is refuted) *there is* $W \in \mathcal{W}_{4,5}$ *with* $w_\Sigma(W, \ell_0, v) < 1/2$ *for every* $v \in \mathcal{V}_5$*; in particular* $1/2$ *is not the value of* ℓ_0*;*

(ii) (a value one step too small is refuted) *there is no* $W \in \mathcal{W}_{4,5}$ *with* $w_\Sigma(W, \ell_0, v) \leq 2/5$ *for every* $v \in \mathcal{V}_5$*; in particular* $2/5$ *is not the value of* ℓ_0*;*

(iii) (the number of classes matters) *the same four examples with only four classes available,* $\ell_0' : [4] \to [4]$ *bijective, have value* $C_4 = 1/2 \neq 3/7$*;*

(iv) (the certificate check has teeth) *the* 21 *atoms of the certificate* E_4 *with their claimed value raised from* $3/7$ *to* $1/2$ *fail the check of Definition 2.4.*

Moreover the hypotheses of the statements above are not vacuous: a label map with exactly four labels, a vote vector and a weight matrix are all exhibited.

Note that $1/2$ and $2/5$ are the neighbours of $3/7$ in the sequence (4), so (i) and (ii) exclude the two ways of being off by one step; (iii) shows the answer really does move with K at fixed examples and labels, which is the content of the $\min\{m+1, K\}$ in Theorem 4.2; and (iv) shows that passing the check of Definition 2.4 is not automatic.

7. Computations outside the formal development

Everything in this section was computed in exact rational arithmetic and is *not* machine-checked; each item says what was run.

The exhaustive value table. Before any of the statements above were formulated, $V(\ell)$ was computed exactly for every label partition of every $n \le 6$ at every $K \le 7$ — one instance per pair (partition, K) for which K is at least the number of parts. The method: solve $\min t$ subject to $AW \le t\mathbf{1}$, $\sum W = 1$, $W \ge 0$, where the rows of A are the pairs (s, v) of (6) modulo the global flip $(s, v) \mapsto (-s, -v)$, in floating point to locate the primal support and the active rows; then solve the reduced systems exactly over $\mathbb{Q}$ and verify exactly, both a primal W (by recomputing $\max_v w_\Sigma$ by brute force over all 2^{K-1} vote vectors) and a dual mixed strategy (by checking its correlation at every one of the nK cells). Both halves exact means the value is proved to be t for that instance; the floating-point solver is used only to guess which constraints are tight. Cost: 11.3 CPU-seconds for the whole sweep.

In all 36 pairs (n, K) with $n \le 6$, $2 \le K \le 7$ the minimum over instances reproduced $\mathfrak{W}_{n,K} = C_{\min\{n+1,K\}}$, values $1, 1/2, 3/7, 2/5$. Per instance, the table is much flatter than its minimum; at $n = 6$ it reads, in full,

λ	6	51	42	411	33	321	3111	222	2211	21111	1^6
$K = 5$	1	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{3}{7}$	$\frac{1}{2}$	$\frac{3}{7}$	$\frac{3}{7}$	—
$K = 7$	1	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{3}{7}$	$\frac{1}{2}$	$\frac{3}{7}$	$\frac{3}{7}$	$\frac{2}{5}$

(λ the multiset of block sizes, written without separators; the entry "—" is the partition with more parts than K). In every instance of the sweep the value equals $C_{\min\{m+1,K\}}$ with m the number of parts, and this is the computation in which the shape of Theorem 4.2 was found. The independence from the block sizes is the agreement of $(5, 1)$, $(4, 2)$, $(3, 3)$ with each other and of $(4, 1, 1)$, $(3, 2, 1)$, $(2, 2, 2)$ with each other; the two sharpest separations, where partitions of the same n at the same K differ by a single part, are $(2, 2, 2)$ against $(2, 2, 1, 1)$ at $K = 5$ and $(2, 1, 1, 1, 1)$ against (1^6) at $K = 7$.

The closed-form strategies beyond $m = 9$. The strategies of Table 1 were verified in exact rational arithmetic for $m = 1, \ldots, 14$ in both variants ($K = m$ and $K > m$), 28 cases in all. In the 27 of them in which all three correlations are present — at $m = 1$ with $K = m$ there is neither a second used column nor an unused one — they equal $+C$, $-C$, $-C$ exactly. Cost: 36.8 CPU-seconds. This is evidence for, not a proof of, the general-m lower bound; the formalised range is $m \le 9$ (Proposition 3.1), and Section 9 says what a proof for all m needs.

A consequence for the round count.

Remark 7.1 (not part of the formal development). The reason the constant matters downstream is the source's round count. Its equation (3.7) reads $T_{\text{fac}} := \lceil 2 \log(nK)/(\delta^2 w_\Sigma^2) \rceil + 1$, so w_Σ — and hence $\mathfrak{W}$, which is what bounds it uniformly — enters only through the factor w_Σ^{-2}; substituting $\mathfrak{W} > 1/3$ gives its equation (3.9), $T_{\text{fac},2} := \lceil 18 \log(nK)/\delta^2 \rceil + 1$, and indeed $18 = 2/(1/3)^2$. Replacing the uniform $1/3$ by the value of the instance at hand replaces 18 by $2/C^2_{\min\{m+1,K\}}$, which is 8 for an instance using two labels when $K \geq 3$. This is a substitution in the source's derivation, not a theorem proved here, and neither the derivation nor the substitution is machine-checked.

8. Formal verification

Propositions 2.1, 2.2, 3.1, 4.1, 6.1 and 6.2, Lemma 2.5, Theorems 4.2 and 5.1 are formalised and machine-checked in Lean 4 [10] against *Mathlib* [11], in eight modules totalling about 1,950 lines and 105 lemmas and theorems, a large part of the latter being certificate data. The game is set up over $\mathbb{R}$ exactly as in (1)–(3), with $\mathcal{W}_{n,K}$, $\mathcal{V}_K$ and the two-sided form (5) as definitions, so no numerical or floating-point statement occurs anywhere; the certificates and weight tables are integer data, and the constants C_q are rationals. Quantification over n and K is genuine: Proposition 3.1 and Theorem 4.2 are single statements about all n and all K, the finiteness being confined to the number of labels used, through the slot compression of Remark 2.3. What is *not* formalised: everything in Section 7, including the exhaustive value table, the verification of Table 1 beyond $m = 9$ and Remark 7.1; the derivation of Table 1 itself, of which only the resulting eighteen certificates are checked; and the statements attributed to [1], [2] and [3], which are quoted, not re-proved — with the exception of $\mathfrak{W}_{n,K} = C_{\min\{n+1,K\}}$ on $\min\{n, K\} \leq 9$, which is Theorem 5.1. Corollary 4.3 is a reading of Theorem 4.2 and carries no verification of its own. The negative controls of Section 6 are part of the development, not a separate script. Axiom sets were printed for every headline statement: each is exactly propositional extensionality, quotient soundness and choice, and the certificate checks use propositional extensionality alone. Every finite check runs inside ordinary kernel evaluation — no compiled evaluation and no extra axiom of any kind — the largest single check taking about one CPU-minute, and the whole development about four CPU-minutes. There is no incomplete proof anywhere in it. The repository is private: repository reference to be added at submission.

9. The frontier

Question 9.1. Does $V(\ell) = C_{\min\{m+1,K\}}$ hold for every m?

We expect so, and both halves have a route. For the lower half, the strategies of Table 1 are defined at every m and are exactly tight there — verified in exact arithmetic to $m = 14$ — so what is missing is a proof of tightness in general, which amounts to two hypergeometric identities for $\mathbb{E}[\sigma_c\sigma_j]$ over the level sets of $\sum_c \sigma_c$, namely $\mathbb{E}[\sigma_c\sigma_j] = (\mathbb{E}[(\sum_c \sigma_c)^2] - m)/(m(m-1))$ evaluated on the two laws of Table 1. For the upper half, the source's own analysis of its constructions, re-instantiated at m, should give Proposition 4.1 for all m at once. Widening the verified window from $m \leq 9$ to $m \leq 12$ is by contrast mechanical rather than mathematical: the larger of the two certificates at $m = 12$ has 3,433 atoms against 253 at $m = 9$, and nothing in the argument changes.

Question 9.2. Is the value of an instance governed by the number of labels used, or by the number of labels used *by the rows of positive weight*?

The source's refined lower bound is stated in terms of r, the number of distinct labels among the examples that the weight matrix charges, which is at most m and depends on W; ours is stated in terms of m, which depends only on the instance. Since the minimiser is free to abandon examples, the two can only be reconciled by an argument about which supports are optimal, and the exhaustive table of Section 7 does not separate them. A per-support refinement of Theorem 4.2 would.

Finally, the boosting-theoretic point of the refinement is Remark 7.1: the round count of factorized AdaBoost.MH is governed not by the number of classes in the problem statement but by the number of classes the sample actually exhibits, at any sample size. Making that a theorem about the algorithm, rather than about the constant it uses, is the natural next step.

References

[1] X. Zou and J. Xu, *Factorized AdaBoost.MH Achieves the Same Convergence Rate as AdaBoost.MH*, arXiv: 2608.01091v2, primary and only classification cs.LG. The arXiv record, read on 3 September 2026, shows "[Submitted on 2 Aug 2026 (v1), last revised 8 Aug 2026 (this version, v2)]", with no comments field and no journal

reference. All quotations above are from the v2 e-print source, and all numbering used above is v2's: equation (2.10) is (2), equation (3.2) is (4), Proposition 3.2 together with the remark following Theorem 3.5 is Proposition 2.1, Theorem 3.5 is the exact value of $\mathfrak{W}_{n,K}$, Theorem 3.1 carries the bounds that its Appendix B.2 invokes, equations (3.7) and (3.9) are the round counts of Remark 7.1, Appendix B.2 is the refined lower bound, and Appendices C.1 and C.2 are the two constructions of Proposition 4.1. The version matters: the v1 abstract claims only $\max\{1/n, C_K\} \le \mathfrak{W}_{n,K} \le C_{\min\{n,K\}}$ and does not use the phrase "effective class number"; both the exact value and that language were added at v2, whose Conclusion nevertheless still states the v1 bounds — a leftover rather than an error in the mathematics.

[2] X. Zou, Z. Zhou, J. Xu and W. Liu, *A Boosting-Type Convergence Result for AdaBoost.MH with Factorized Multi-Class Classifiers*, Advances in Neural Information Processing Systems **37** (NeurIPS 2024), 79764–79782; DOI `10.52202/079017-2533`. The prior art on (2): its Theorem 3.3 gives $\mathfrak{W}_{n,K} \ge 1/n$ and its Theorem 3.4 gives $\mathfrak{W}_{n,K} \ge 1/\sqrt{2K}$. Read in the proceedings version.

[3] B. Kégl, *Open Problem: A (Missing) Boosting-type Convergence Result for AdaBoost.MH with Factorized Multi-class Classifiers*, Proceedings of the 27th Conference on Learning Theory (COLT 2014), Proceedings of Machine Learning Research vol. 35, 1268–1275. The paper that poses the question quoted in Section 1, verbatim from its Remark 3; [1] renders that question in a lightly reworded form. It has no arXiv identifier: an arXiv API search for `all:"AdaBoost.MH"` on 3 September 2026 returned exactly five papers, none of them this one.

[4] B. Kégl, *The return of AdaBoost.MH: multi-class Hamming trees*, ICLR 2014; arXiv:1312.6086, whose record, read on 3 September 2026, shows v1 as the only version, 20 December 2013. The source of the factorized base classifier $h(x) = \alpha\, v\, \varphi(x)$.

[5] R. E. Schapire and Y. Singer, *Improved boosting algorithms using confidence-rated predictions*, Machine Learning **37** (1999), no. 3, 297–336; DOI `10.1023/A:1007614523901`. AdaBoost.MH itself.

[6] Y. Freund and R. E. Schapire, *Game theory, on-line prediction and boosting*, in: Proceedings of the Ninth Annual Conference on Computational Learning Theory (COLT '96), ACM, 1996, 325–332; DOI `10.1145/238061.238163`. Boosting as a repeated two-player game; no constant of the shape (4).

[7] G. Rätsch and M. K. Warmuth, *Efficient margin maximizing with boosting*, Journal of Machine Learning Research **6** (2005), 2131–2152. Minimax and duality for the binary weak-learning condition.

[8] S. Shalev-Shwartz and Y. Singer, *On the equivalence of weak learnability and linear separability: new relaxations and efficient boosting algorithms*, Machine Learning **80** (2010), no. 2–3, 141–163; DOI `10.1007/s10994-010-5173-z`. Von Neumann duality applied to the weak-learning condition.

[9] I. Mukherjee and R. E. Schapire, *A theory of multiclass boosting*, Journal of Machine Learning Research **14** (2013), 437–497; arXiv:1108.2989 (v1 only, record read 3 September 2026). The multiclass edge-over-random conditions. Its full text carries no constant of the shape $q/(3q-4)$.

[10] L. de Moura and S. Ullrich, *The Lean 4 theorem prover and programming language*, in: Automated Deduction — CADE 28, Lecture Notes in Computer Science, Springer, 2021, 625–635; DOI `10.1007/978-3-030-79876-5_37`.

[11] The mathlib Community, *The Lean mathematical library*, in: Proceedings of the 9th ACM SIGPLAN International Conference on Certified Programs and Proofs (CPP 2020), ACM, 2020, pp. 367–381; DOI `10.1145/3372885.3373824`.