Open Problem: Lower bounds for Boosting with Hadamard Matrices

Jiazhong Nie \texttt{jnie@ucsc.edu} and Manfred K. Warmuth \texttt{manfred@ucsc.edu}
Department of Computer Science UC Santa Cruz

S.V.N. Vishwanathan \texttt{vishy@stat.purdue.edu}
Departments of Statistics and Computer Science Purdue University

Xinhua Zhang \texttt{xinhua.zhang.cs@gmail.com}
NICTA, Canberra, Australia

Abstract

Boosting algorithms can be viewed as a zero-sum game. At each iteration a new column / hypothesis is chosen from a game matrix representing the entire hypotheses class. There are algorithms for which the gap between the value of the sub-matrix (the $t$ columns chosen so far) and the value of the entire game matrix is $O\left(\sqrt{\frac{\log n}{t}}\right)$. A matching lower bound has been shown for random game matrices for $t$ up to $n^\alpha$ where $\alpha \in (0, \frac{1}{2})$. We conjecture that with Hadamard matrices we can build a certain game matrix for which the game value grows at the slowest possible rate for $t$ up to a fraction of $n$.

1. Boosting as a zero-sum game

Boosting algorithms follow the following protocol in each iteration (e.g. Freund and Schapire, 1997; Freund, 1995): The algorithm provides a distribution $d$ on a given set of $n$ examples. Then an oracle provides “weak hypothesis” from some hypotheses class and the distribution is updated. At the end, the algorithm outputs a convex combination $w$ of the hypotheses it received from the oracle.

One can view Boosting as a zero-sum game between a row and a column player (Freund and Schapire, 1997). Each possible hypothesis provided by the oracle is a column chosen from an underlying game matrix $U$ that represents the entire hypotheses class available to the oracle. The examples correspond to the rows of this matrix. At the end of iteration $t$, the algorithm has received $t$ columns/hypotheses so far, and we use $U_t$ to denote this sub-matrix of $U$. The minimax value of $U_t$ is defined as follows:

$$\text{val}(U_t) = \min_{d \in S^n} \max_{w \in S^t} d^\top U_t w = \max_{w \in S^t} \min_{r=1,\ldots,n} [U_t w]_r.$$  \hfill (1)

Here $d$ is the distribution on the rows/examples and $w$ represents a convex combination of the $t$ columns of $U_t$. Finally $[U_t w]_r$ is the margin of row/example $r$ wrt the convex combination $w$ of the current hypotheses set. So in Boosting the value of $U_t$ is the maximum minimum margin of all examples achievable with the current $t$ columns of $U_t$.

The value of $U_t$ increases as columns are added and in this view of Boosting, the goal is to raise the value of $U_t$ as quickly as possible to the value of the entire underlying game matrix $U$. There are boosting algorithms that guarantee that after $O\left(\frac{\log n}{\epsilon^2}\right)$ iterations, the

gap \text{val}(U) - \text{val}(U_t) is at most $\epsilon$ (Freund and Schapire, 1997; Rätsch and Warmuth, 2005; Warmuth et al., 2008). In other words, the gap at iteration $t$ is at most $O(\sqrt{\log n}/t)$. Here we are interested in finding game matrices with a matching lower bound for the value gap. The lower bound should hold for any boosting algorithm, and therefore the gap in this case is defined as the maximum over \text{all submatrices} $U_t$ of $t$ columns of $U$: 

\[
gap_t(U) := \text{val}(U) - \max_{U_t} \text{val}(U_t).
\]

First notice that the gap is non-zero only when $t \leq n$, since for any $n \times m$ ($m > n$) game matrix, its value is always attained by one of its sub-matrices of size $n \times (n + 1)$. This follows from Carathéodory theorem which implies that for any column player $w \in S^m$, there is $\hat{w}$ with support of size at most $n + 1$ satisfying $U\hat{w} = U\hat{w}$. So wlog $m \leq n$.

Klein and Young (1999) showed that for a limited range of $t$ ($\log n \leq t \leq n^{\alpha}$ with $\alpha \in (0, \frac{1}{2})$), the gap is $\Omega(\sqrt{\log n}/t)$ with high probability for random bit matrices $U$.

We claim that with certain game matrices the range of $t$ in this lower bound can be increased.

2. Lower bounds with Hadamard matrices

Hadamard matrices have been used before for proving hardness results in Machine Learning (eg Kivinen et al., 1997; Warmuth and Vishwanathan, 2005) and for iteratively constructing game matrices with large gaps (Nemirovski and Yudin, 1983; Ben-Tal et al., 2001). We begin by giving a simple but weak lower bound using these matrices (an adaptation of Proposition 4.2 of Ben-Tal et al. (2001)).

Let $n = 2^k$ and $\hat{H}$ be the $n \times n$ Hadamard matrix. Define $\hat{H}$ to be $H$ with first row removed. We use game matrix $U = \begin{bmatrix} \hat{H} \\ -\hat{H} \end{bmatrix}$ and let $\text{val}_D(U)$ denote $\text{val}(\begin{bmatrix} U \\ -U \end{bmatrix})$. Notice that by definition 1, $\text{val}_D(U) = -\min_{w \in S^n} \|Uw\|_\infty \leq 0$.

**Theorem** For $1 \leq t \leq \frac{n}{2}$, $\text{val}_D(\hat{H}) - \max_{\hat{H}_t} \text{val}_D(\hat{H}_t) \geq \frac{1}{2t}$, where the maximum is over all sub-matrices $\hat{H}_t$ of $t$ columns of $\hat{H}$.

**Proof** First we show $\text{val}_D(\hat{H}) = 0$. Notice that $\hat{H}$ has row sum zero and

$$\text{val}_D(\hat{H}) = -\min_{w \in S^n} \|\hat{H}w\|_\infty \geq -\|\hat{H}\frac{1}{n}w\|_\infty = 0.$$ 

Since $H$ has orthogonal columns, we have that for any $\hat{H}_t$, $\hat{H}_t^\top \hat{H}_t = nI_t - 1_t1_t^\top$ and

$$\min_{w \in S^t} \|\hat{H}_t w\|_\infty \geq \min_{w \in S^t} \|\hat{H}_t w\|_2 \geq \min_{w \in S^t} \sqrt{n-1} = \min_{w \in S^t} \sqrt{n} \frac{n}{n} - \frac{1}{n-1} \geq \sqrt{(n-t)/(n-1)t}. $$

1. Freund (1995) originally gave an adversarial oracle that iteratively produces a hypothesis of error $\epsilon$ w.r.t. the current distribution, and for any particular algorithm, the oracle can make this go on for $\Omega(\frac{\log n}{\epsilon^2})$ iterations. A lower bound of $\Omega(\sqrt{\log n}/t)$ on the value gap is a much stronger type of lower bound.
2. The same lower bound translates to random $\pm 1$ matrices via shifting and scaling.

---

Nie Warmuth Vishwanathan Zhang
Finally we have for \( t \leq \frac{n}{2} \), \( \text{val}_D(\mathbf{H}) - \max_\mathbf{H}_t \text{val}_D(\mathbf{H}_t) \geq \sqrt{\frac{n-t}{(n-t)t}} \geq \sqrt{\frac{1}{2t}}.\)

Note that this weaker lower bound holds for a larger range of \( t \) \((1 \leq t \leq \frac{n}{2})\) than the stronger lower bound of \( \sqrt{\frac{\log n}{t}} \) proven by Klein and Young (1999) for a restricted range. We first conjecture that the stronger lower bound holds for the larger range for our matrices:

**Conjecture 1** There are fixed fractions \( c, c' \in (0, 1) \) and \( n_0 \) such that the gap of \( \hat{\mathbf{H}} \) is lower bounded as follows: \( \forall n \geq n_0 \) and \( \log n \leq t \leq cn \): \( \text{val}_D(\hat{\mathbf{H}}) - \max_\mathbf{H}_t \text{val}_D(\hat{\mathbf{H}}_t) \geq c' \sqrt{\frac{\log n}{t}}.\)

We further conjecture that our modified Hadamard matrices give the largest gaps among all \( \pm 1 \) matrices with game value 0. We have verified this conjecture by tedious combinatorial arguments for \( n = 2, 4, 8 \) and \( t \leq n \) as well as for \( n = 2^k \) and \( n - 2 \leq t \leq n \).

**Conjecture 2** For any \((n-1) \times n\) dimensional \( \pm 1 \) valued matrix \( \mathbf{U} \) satisfying \( \text{val}_D(\mathbf{U}) = 0 \), the following inequality holds for \( 1 \leq t \leq n \): \( \max_\mathbf{H}_t \text{val}_D(\hat{\mathbf{H}}_t) \leq \max_\mathbf{U}_t \text{val}_D(\mathbf{U}_t) \), where \( \hat{\mathbf{H}}_t \) is any \( t \) column sub-matrix of \( \hat{\mathbf{H}} \) and \( \mathbf{U}_t \) is any \( t \) column sub-matrix of \( \mathbf{U} \).

**References**


