

Lower bounds for Boosting with Hadamard Matrices

Jiazhong Nie JNIE@UCSC.EDU and **Manfred K. Warmuth** MANFRED@UCSC.EDU

Department of Computer Science UC Santa Cruz

S.V.N. Vishwanathan

VISHY@STAT.PURDUE.EDU

Departments of Statistics and Computer Science Purdue University

Xinhua Zhang

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 $O(\sqrt{\frac{\log n}{t}})$. A matching lower bound has been shown for random game matrices for t up to n^α 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 \mathbf{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 \mathbf{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 \mathbf{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 \mathbf{U}_t to denote this sub-matrix of \mathbf{U} . The minimax value of \mathbf{U}_t is defined as follows:

$$\text{val}(\mathbf{U}_t) = \min_{\mathbf{d} \in \mathcal{S}^n} \max_{\mathbf{w} \in \mathcal{S}^t} \mathbf{d}^\top \mathbf{U}_t \mathbf{w} = \max_{\mathbf{w} \in \mathcal{S}^t} \min_{r=1, \dots, n} [\mathbf{U}_t \mathbf{w}]_r. \quad (1)$$

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

The value of \mathbf{U}_t increases as columns are added and in this view of Boosting, the goal is to raise the value of \mathbf{U}_t as quickly as possible to the value of the entire underlying game matrix \mathbf{U} . There are boosting algorithms that guarantee that after $O(\frac{\log n}{\epsilon^2})$ iterations, the *gap* $\text{val}(\mathbf{U}) - \text{val}(\mathbf{U}_t)$ is at most ϵ ([Freund and Schapire, 1997](#); [Rätsch and Warmuth, 2005](#);

Warmuth et al., 2008). In other words, the gap at iteration t is at most $O(\sqrt{\frac{\log n}{t}})$. 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 *all submatrices \mathbf{U}_t of t columns of \mathbf{U}* :¹

$$\text{gap}_t(\mathbf{U}) := \text{val}(\mathbf{U}) - \max_{\mathbf{U}_t} \text{val}(\mathbf{U}_t).$$

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 Carathodory theorem which implies that for any column player $\mathbf{w} \in \mathcal{S}^m$, there is $\hat{\mathbf{w}}$ with support of size at most $n + 1$ satisfying $\mathbf{U}\mathbf{w} = \mathbf{U}\hat{\mathbf{w}}$. 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 $\Omega(\sqrt{\frac{\log n}{t}})$ with high probability for random bit matrices \mathbf{U} .² 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 \mathbf{H} be the $n \times n$ Hadamard matrix. Define $\hat{\mathbf{H}}$ to be \mathbf{H} with first row removed. We use game matrix $\mathbf{U} = \begin{bmatrix} \hat{\mathbf{H}} \\ -\hat{\mathbf{H}} \end{bmatrix}$ and let $\text{val}_D(\mathbf{U})$ denote $\text{val}\left(\begin{bmatrix} \mathbf{U} \\ -\mathbf{U} \end{bmatrix}\right)$. Notice that by definition 1, $\text{val}_D(\mathbf{U}) = -\min_{\mathbf{w} \in \mathcal{S}^n} \|\mathbf{U}\mathbf{w}\|_\infty \leq 0$.

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

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

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

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

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

Finally we have for $t \leq \frac{n}{2}$, $\text{val}_D(\mathbf{H}) - \max_{\hat{\mathbf{H}}_t} \text{val}_D(\hat{\mathbf{H}}_t) \geq \sqrt{\frac{n-t}{(n-1)t}} \geq \sqrt{\frac{1}{2t}}$. ■

1. Freund (1995) originally gave an adversarial oracle that iteratively produces a hypothesis of error ϵ 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 ± 1 matrices via shifting and scaling.

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 $\log n \leq t \leq cn$: $\text{val}_D(\hat{\mathbf{H}}) - \max_{\hat{\mathbf{H}}_t} \text{val}_D(\hat{\mathbf{H}}_t) \geq c' \sqrt{\frac{\log n}{t}}$.*

We further conjecture that our modified Hadamard matrices give the largest gaps among all ± 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 ± 1 valued matrix \mathbf{U} satisfying $\text{val}_D(\mathbf{U}) = 0$, the following inequality holds for $1 \leq t \leq n$: $\max_{\hat{\mathbf{H}}_t} \text{val}_D(\hat{\mathbf{H}}_t) \leq \max_{\mathbf{U}_t} \text{val}_D(\mathbf{U}_t)$, 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

- Ahron Ben-Tal, Tamar Margalit, and Arkadi Nemirovski. The ordered subsets mirror descent optimization method with applications to tomography. *SIAM Journal on Optimization*, 12(1):79–108, July 2001.
- Yoav Freund. Boosting a weak learning algorithm by majority. *Inform. Comput.*, 121(2):256–285, September 1995. Also appeared in COLT90.
- 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.
- J. Kivinen, M. K. Warmuth, and P. Auer. The perceptron algorithm vs. winnow: linear vs. logarithmic mistake bounds when few input variables are relevant. *Artificial Intelligence*, 97:325–343, December 1997.
- Philip Klein and Neal Young. On the number of iterations for dantzig-wolfe optimization and packing-covering approximation algorithms. In *In Proceedings of the 7th International IPCO Conference*, pages 320–327. Springer, 1999.
- Arkadi Nemirovski and D Yudin. *Problem Complexity and Method Efficiency in Optimization*. John Wiley and Sons, 1983.
- Gunnar Rätsch and Manfred K. Warmuth. Efficient margin maximization with boosting. *J. Mach. Learn. Res.*, 6:2131–2152, December 2005.
- M. K. Warmuth and S. V. N. Vishwanathan. Leaving the span. In P. Auer and R. Meir, editors, *Proc. Annual Conf. Computational Learning Theory*, number 3559 in Springer Lecture Notes in Artificial Intelligence, pages 365–380, Bertinoro, Italy, June 2005.
- Manfred K. Warmuth, Karen A. Glocer, and S. V. N. Vishwanathan. Entropy regularized LPBoost. In Yoav Freund, Yoav László Györfi, and György Turán, editors, *Proc. Intl. Conf. Algorithmic Learning Theory*, number 5254 in Lecture Notes in Artificial Intelligence, pages 256 – 271, Budapest, October 2008. Springer-Verlag.