LEARNING LINEAR TRANSFORMATIONS

Alan Frieze* Mark Jerrum† Ravi Kannan‡

April 5, 1996

Abstract We present a polynomial time algorithm to learn (in Valiant's PAC model) cubes in $n-$ space (with general sides - not necessarily axes parallel) given **uniformly** distributed samples from the cube. In fact, we solve a more general problem of learning in polynomial time linear transformations in $n-$space. I.e., suppose x is an $n-$vector whose coordinates are mutually independent random variables with unknown (possibly different) probability distributions and A is an unknown nonsingular $n \times n$ matrix. Then given polynomially many samples of $y = Ax$, we are able to learn the columns of A approximately. Geometrically, this is equivalent to learning a parallelepiped given uniformly distributed samples from it. Actually, we will only need a weak 4-way independence which we will describe later; also we will handle the case when $y = Ax + b$ where b is an unknown vector.

We first show that using some standard Linear Algebra, we can learn parallelepipeds upto rotations. This only involves analyzing the matrix of second moments of the "observed" variables y.

The central problem is determining the rotation. We first prove that certain fourth moments of y determine the rotation; we actually show that the maxima (and minima) of the fourth moment function give us the columns of A. Then we show a constructive (polynomial time) version of this result; i.e., we show that the maxima and minima can be found approximately by a nonlinear (fourth degree) optimization algorithm.

While our primary motivation comes from Learning Theory, the problem has some similarities to problems in Factor Analysis, a branch of Statistics. There, no assumption is made about the independence of the x_i, so the problem is more general; but also the results are weaker in that one only finds A upto rotations. Then one uses heuristics to find a "pleasing" rotation of A. (See for example [6].)

The paper closes with some generalizations of the result and open problems.

*Department of Mathematics, Carnegie Mellon University, Pittsburgh PA15213, U.S.A.,Supported in part by NSF grant CCR-9225008. **E-mail: af1p@andrew.cmu.edu**

†Department of Computer Science, University of Edinburgh, The King's Buildings, Edinburgh EH9 3JZ, United Kingdom. Supported in part by grant GR/F 90363 of the UK Science and Engineering Research Council, and Esprit Working Group 7097 "RAND". **E-mail: mrj@dcs.ed.ac.uk**

‡Department of Computer Science, Carnegie Mellon University, Pittsburgh PA15213, U.S.A., Supported in part by NSF grant CCR-9208597 and CCR-9528973. **E-mail: kannan@convex.sp.cs.cmu.edu**

1. Introduction The class of intersections of halfspaces is a widely studied concept class in machine learning theory (e.g., [5, 2, 1, 4]). Not only are they quite natural geometrically, but they also correspond to functions computed by simple neural networks. The case of one half space is well-solved by Linear Programming.

Unfortunately, the case of two half spaces already presents a problem : In the Valiant distribution-free PAC model, even an intersection of 2 halfspaces cannot be learned in a representation dependent sense (the learners hypothesis must also be an intersection of 2 halfspaces) unless **RP**=**NP** [4, 10]. Some intuitive arguments for the difficulty of learning this class in a representation-independent manner (the learners hypothesis may be any polynomial-time prediction algorithm) are given by Baum [2]. Also, Long and Warmuth [8] have shown that the class of convex polytopes given by their vertices is prediction complete for **P**.

In restricted distribution models, however, there have been some positive results. In particular, Baum [1] showed that an intersection of two homogeneous halfspaces (the hyperplanes that define them must pass through the origin) is learnable in polynomial time over any distribution Δ such that $\Delta(x) = \Delta(-x)$ for all x. Also, Blum and Kannan [3] have shown that if the underlying distribution is uniform over the unit ball, we can PAC learn in polynomial time the intersection of a constant number of half spaces.

Here we give the first result that tackles the intersection of a non constant number of half spaces. Our result (already described in the Abstract) raises the question of whether other convex sets can also be PAC learnt when we restrict attention to the uniform distribution. These remain interesting open questions.

We note that there has been some prior work on learning cubes, but generally, attention has been restricted to axes-parallel cubes and some generalizations - see for example Maass and Warmuth [9].

A special case of our result says that we can learn product distributions when the axes of the independent variables are unknown. We also note that Kearns, Mansour, Ron, Rubinfeld, Schapire and Sellie [7] have considered the problem of learning a probability distribution from given samples. Their focus is mainly on discrete distributions on $\{0,1\}^n$ and their methods and results are of a different flavor.

We also note that while our algorithm is polynomial time bounded, both its time complexity and the number of samples it needs need to be substantially improved before it becomes practical.

In the next section, we present a preliminary result that finds A upto "rotations" using the first and second moments of y. This follows from standard Linear Algebra. We have tried to state this result here in a "clean" form without invoking quantities like the condition number of A as is sometimes done because the result is applicable in other contexts. This preliminary result makes no assumptions about the independence of the x_i.

Then, in section 3 , we present the main result about finding the rotation. This assumes a 4-way independence of the x_i. [We do not need any more independence than 4-way.]

2. Using Second Moment Information The "variance-covariance" matrix $M(\mathcal{P})$ of a probability distribution $\mathcal{P}$ on $\mathbf{R}^n$ is an $n \times n$ matrix whose (i,j) th entry is $E_{\mathcal{P}}(x_i x_j)$. If x denotes a column vector (as we will use throughout), we can write it in matrix notation as

$$M = E_{\mathcal{P}}(xx^T).$$

We say that $\mathcal{P}$ is "uncorrelated" if $E_{\mathcal{P}}(x_i) = 0 \forall i$ and $M(\mathcal{P})$ is the identity matrix.

If we are given sufficiently many samples each drawn independently according to $\mathcal{P}$, then we can first estimate $\overline{x}_i = E_{\mathcal{P}}(x_i)$ for each i and after moving the origin to $\overline{x}$, we have that $E_{\mathcal{P}}(x_i) = 0$. We could also estimate $M(\mathcal{P})$. This is a positive semi definite matrix and we may decompose it as $M = S^2$ where S is symmetric and if M is nonsingular, then so is S. In this case, if we transform space by S^{-1}, we see that in the transformed space, $\mathcal{P}$ will be uncorrelated. This process is useful in many contexts including the present one. We state below a clean version of this. Since there are errors, we cannot expect that in the transformed space, the variance-covariance matrix is the identity. Instead, we will be satisfied with making its eigenvalues all close to 1 in absolute value. [Recall that a square symmetric matrix has all eigenvalues equal to 1 iff it is the identity.] We get the following result. (Proof deferred to the final paper.)

Lemma 1 *Suppose $\mathcal{P}$ is any probability distribution in $\mathbf{R}^n$ with a nonsingular variance-covariance matrix. Without loss of generality, assume that $E_{\mathcal{P}}(x_i) = 0 \forall i$ and $E_{\mathcal{P}}(x_i^2) = 1 \forall i$. Let $\mu_4 = \max_i E_{\mathcal{P}}(x_i^4)$. Also, let $\varepsilon \in (0,\ 1/10)$. Then given $10n^2\mu_4\varepsilon^{-2}$ samples each drawn independently according to $\mathcal{P}$, we can find in polynomial time a linear transformation τ such that*

$$E_{\mathcal{P}}((\tau x)(\tau x)^T)$$

has all its eigenvalues between $1-\varepsilon$ and $1+\varepsilon$.

We apply this lemma in the present context as follows : we have $y = Ax + b$, where x is a vector random variable. A is an unknown nonsingular matrix and b is an unknown vector. We also assume that the variance-covariance matrix of x is nonsingular. We may then assume without loss of generality that (after changing A, b if necessary) that $E(x_i) = 0 \forall i$ and that the variance-covariance matrix of x is the identity. The above result will imply that we can find a matrix B such that the eigenvalues of $B^{-1}A$ are all close to 1 in absolute value. (The proof is simple and is deferred to the final paper.) This says that $B^{-1}A$ is close to a an orthonormal matrix. [Recall that a square (not necessarily symmetric) matrix has all eigenvalues equal to 1 iff it is orthonormal.] With some abuse of terminology, we will refer to this in the paper as "finding A approximately upto rotations". Here is the result.

Lemma 2 *Suppose A is a nonsingular $n \times n$ matrix and b any vector. Suppose x is a random variable with values in $\mathbf{R}^n$ with $E(x_i) = 0 \forall i$ and variance-covariance matrix equal to the identity. Let $\mu_4 = \max_i E(x_i^4)$. Let $\varepsilon \in (0\ 1/10)$. Then given $10n^2\mu_4\varepsilon^{-2}$ independent samples of $y = Ax + b$, we can find a matrix B such that the eigenvalues of $B^{-1}A$ are all between $1-\varepsilon$ and $1+\varepsilon$ in absolute value.*

3. Main Results The general problem we consider is the following.

There are n real valued random variables $x = (x_1, x_2, \ldots x_n)$. We are given observations of

$$y = Ax + b,$$

where A is an unknown nonsingular matrix and b is an unknown vector. Our aim is to find A, b approximately from polynomially many samples of y. First, we observe that after changing b suitably and scaling A, we may assume without loss of generality that

Assumption 1 $E(x_i) = 0 \forall i$ and $E(x_i^2) = 1 \forall i$.

Now b_i may be estimated as $E(y_i)$ and replacing y by $y - E(y)$, we will assume henceforth that $b = 0$. [We defer a careful error analysis of this to the final paper.]

Assumption 2 We will assume that in fact the variance-covariance matrix of x is the identity.

Unlike the situation in the last section, this assumption does entail a loss of generality here, since we will also need Assumption 3 below.

Under Assumptions 1 and 2, we may find a B as in Lemma 2 of the last section. In what follows, we let

$$z = B^{-1}y.$$

From the observations of y, we may obviously now obtain observations of z. Remembering that $b = 0$, we see that

$$z = Rx \quad \text{where} \quad R = B^{-1}A \text{ is a nearly orthonormal matrix.}$$

We will use the observations of z to find R. For this, we need one more assumption, namely that of 4-way independence as mentioned earlier. We make the precise assumption now.

Assumption 3 We will assume a weak form of 4-way independence. I.e., we will assume that the expectation of each monomial of degree 4 in the x_i 's is the product of the expectations of each variable to the suitable power. More precisely, we assume that

$E(x_i x_j x_k x_l) = 0$ whenever for any s, x_s occurs an odd number of times in the product $x_i x_j x_k x_l$; we also assume that $E(x_i^2 x_j^2) = 1$.

Note that independence of all the coordinates of x implies the above. In general, our assumption is of course much weaker than total independence.

The central idea is contained in the following lemma which is a theoretical result that ignores errors. The proof of this is relatively straightforward. The constructive version which actually finds the maxima and minima in the presence of errors is more complicated and will be discussed later.

Lemma 3 *Suppose we have random variables $x = (x_1, x_2, \ldots x_n)$ satisfying assumptions 1,2 and 3 above. Suppose R is an orthonormal matrix. Consider the function $F(u)$ (where u is a column vector in $\mathbf{R}^n$) defined by*

$$F(u) = E((u^T Rx)^4).$$

The local maxima (respectively, the local minima) of $F()$ over the unit sphere ($\{u : |u| = 1\}$) are precisely the rows of A^{-1} corresponding to i such that $E(x_i^4) > 3$ (respectively, $E(x_i^4) < 3$.)

Proof Sketch : Since R is orthonormal, the vector $v^T = u^T R$ varies over the unit sphere as u does. Let $F(u) = G(v)$ with this change of variables. Then $G(v) = E((v^T x)^4)$. Expanding the fourth power and using the assumptions, we can see after some manipulation that

$$G(v) = 3 + \sum_{i=1}^{n} v_i^4 [E(x_i^4) - 3].$$

From this, it is not difficult to derive the lemma.

Exceptional variables

This still leaves open the i for which we have $E(x_i^4) = 3$. We call these the "exceptional" i. It is easy to see that we cannot in general avoid the exceptional i. In the case that the x_i is a standard normal variable, i will be exceptional as one may see by direct calculation. In fact, if all the x_i are independent standard normals, then the resulting distribution is rotation invariant and so in fact, we cannot find the actual rotation R (since all R look alike).

We perturb the distribution of the exceptional x_i to make it non-exceptional and then find the rotation. The technical details of the perturbation are complicated, and so as not to obscure the main ideas in this extended abstract, we defer this to the final paper.

Error Analysis and Computing the local minima and maxima

We will only outline this here. As remarked earlier, we have observations of $z = Rx$ where now R is approximately an orthonormal matrix. First we will indicate the number N of samples of z we use.

Let $\mu_{i,t} = \mathbf{E}(|\mathrm{x}_i^t|)$ and $\mu_t = \max_{1 \le i \le n} \mu_{i,t}$ for $i, t \ge 1$. Thus by assumption $\mu_1 = 0$ and $\mu_2 = 1$.

Let

$$\Delta_i = \mathbf{E}(\mathrm{x}_i^4) - 3, \qquad 1 \le i \le n,$$

and

$$\Delta_{\max} = \max_{1 \le i \le n} |\Delta_i|, \ \Delta_{\min} = \min_{1 \le i \le n} |\Delta_i|.$$

For this Extended Abstract, we assume that $\Delta_{\min} > 0$.

$$\text{Let } N = \frac{2 \times 10^5 \times n^{10} \mu_8}{\Delta_{\min}^2 \varepsilon^4}$$

and let $z^{(j)}, j = 1, 2, \ldots N$ be the N independent observations of z.

Define

$$\phi_\infty(u) \quad = \quad \mathbf{E}((u^T \mathbf{z})^4), \qquad u \in S_{n-1},$$

$$\begin{aligned}\phi_N(u) &= \frac{1}{N}\sum_{j=1}^{N}(u^T\mathbf{z}^{(j)})^4, && u \in S_{n-1},\\ \psi(v) &= \sum_{i=1}^{n}\Delta_i v_i^4 + 3, && v \in S_{n-1}.\end{aligned}$$

The relevance of ψ will be apparent when we see what ϕ_∞ looks like in the $\mathbf{x}$ coordinate system:

$$\begin{aligned}\phi_\infty(u)/|R^Tu|^4 &= \mathbf{E}((u^TR\mathbf{x})^4/|R^Tu|^4)\\ &= \mathbf{E}((v^T\mathbf{x})^4)\\ &= \sum_{i=1}^{n}\mu_{i,4}v_i^4 + 3\sum_{i\neq j}v_i^2v_j^2\\ &= \sum_{i=1}^{n}\Delta_i v_i^4 + 3. \qquad (1)\end{aligned}$$

Let

$$\zeta_N(u) = \bigtriangledown\phi_N(u) - (u^T \bigtriangledown \phi_N(u))u \qquad (2)$$

denote the projection of $\bigtriangledown\phi_N(u)$ orthogonal to u.

For a twice differentiable function $F(u)$, let Hess(F) denote the matrix $\left[\frac{\partial^2 F}{\partial u_i \partial u_j}(u)\right]$. Let $H_N(u) = \text{Hess}(\phi_N(u))$.

We now describe our ascent algorithm. The algorithm as described finds a local maximum of the function $\phi_N(u)$. Intuitively, the algorithm would make either first order moves (along the component of the gradient tangential to the sphere) or if this component is negligible, it makes second order moves dictated by an eigenvector of the Hessian. Actually, we combine the two into one local optimization problem at each step. The crucial part will be to prove that at the termination of the algorithm, when no first or second order moves are possible, we are in fact done.

ASCEND

Step 0 Choose $u \in S_{n-1}$ e.g. $(1, 0, 0, \ldots, 0)$.
Assume that $\phi_N(u) \geq 3$. (See Remark 2, below).

Step 1 Solve Problem $Q_N(u)$:

Maximize

$$f_N(u,\xi) = \zeta_N(u)^T\xi + \frac{1}{2}\xi^T(H_N(u) - 4\phi_N(u)I)\xi$$

subject to

$$\begin{aligned}|\xi| &\leq \delta\\ u^T\xi &= 0.\end{aligned}$$

Let ξ^* be a solution to $Q_N(u)$ and

$$\tilde{u} = \frac{u + \xi^*}{|u + \xi^*|}.$$

If

$$\phi_N(\tilde{u}) \geq \phi_N(u) + \frac{\Delta_{\min}\delta^2}{30n}$$

then repeat Step 1, with u replaced by $\tilde{u}$, otherwise

Step 2 Terminate: output u.

Throughout this section u will always denote a vector in S_{n-1}. Also, the relations $w = R^T u$ and $v = w/|w|$ will always hold. Similar relations will be valid for $u' \in S_{n-1}, v', w'$ etc.

Remark 1: Q_N is easy to solve: After a bit of simple linear algebra, it reduces to a maximum eigenvalue calculation.
Remark 2: $\psi(v) \geq 3$ implies that there exists at least one positive Δ_i. Now $\phi_N(u)$ is close to $\psi(v)$ – Lemma 4 below and so $\phi_N(u) \geq 3$ should yield $\psi(v) \geq 3$. It is conceivable, that through sampling error, we have $\phi_N(u) \geq 3$ and yet $\Delta_i < 0$ for all i. In which case we maximizing ϕ_N would be a mistake. We can recognize this as follows: if there is a positive Δ_i then after $O(n\delta^{-2})$ iterations the value of ϕ_N will have increased by at least $\Delta_{\min}/2$. But if there are no positive Δ_i, ϕ_N can only increase by at most $\Delta_{\min}\delta^2/(100n)$. So we proceed under the assumption that there is at least one positive Δ_i.

Our analysis will track the changes in v as u is changed by ASCEND. We will show that when the algorithm terminates we should be close to a local maximum of ψ which means that $\pm v$ is close to a standard basis vector $(0, 0, \ldots, 1, 0, \ldots)$ and then $\pm u$ must be close to a column of R. A further calculation is needed to get the sign correct.

Now $\psi(v) \leq \Delta_{\max} + 3$ and (Lemma 4) $\phi_N(u)$ and $\psi(v)$ are close. So, ASCEND terminates after at most $O(n\delta^{-2})$ iterations.

4. Proof of Correctness of the Algorithm The proof of correctness is very technical, mainly owing to the fact that we only have approximations to the real $\mathbf{E}((u^T z)^4)$ and its derivatives. We defer most of the long proof to the final paper

Our aim now is to show that when ASCEND terminates, it is likely to have produced a column of R. Our first lemma explains that **whp** $\phi_N(u)$ and $\psi(v)$ are close.

Lemma 4 *With contrary probability at most*

$$p_3 = \frac{25000n^{10}\mu_8(1+\alpha)^2}{N\Delta_{\min}^2\delta^2}$$

we find that for all $u, u' \in S_{n-1}$,

(a)

$$|\phi_N(u) - \psi(v)| \leq \frac{\Delta_{\min}\delta^2}{100n},$$

(b)

$$|u' - u| \leq 3|v' - v|.$$

Proof (deferred to full paper).

Note that (a) implies that $\psi(v)$ also increases with each iteration of ASCEND.

We now continue under the asumption that (a), (b) of the above lemma hold. Fix $u \in S_{n-1}$ and let $u' = (u+h)/|u+h|$ where $h^T u = 0$ and $\min\{|u'-u|, |h|\} \leq .1$.

Lemma 5

$$\frac{3}{4}|h| \leq |u' - u| \leq |h|. \tag{3}$$

Proof (deferred to full paper).

We now compare $\phi_N(u)$ and $\phi_N(u')$ when $|h|$ is small.

$$\begin{aligned}
\phi_N(u') &= \frac{1}{|u+h|^4}\phi_N(u+h) \\
&= \frac{1}{(1+|h|^2)^2}(\phi_N(u) + \zeta_N^T h + \frac{1}{2}h^T H_N h + \mathrm{err}_1(u,h)) &(4)\\
&= \phi_N(u) + f_N(u,h) + \mathrm{err}(u,h), &(5)
\end{aligned}$$

We will show in the full paper that with contrary probability at most p_4,

$$|\mathrm{err}(u,h)| \leq \frac{31(1+\alpha)^2 n^2 \mu_4 |h|^3}{N p_4}, \tag{6}$$

for all $u \in S_{n-1}$ and sufficiently small $|h|$.

We can now claim that on termination u is (almost) a local maximum.

Lemma 6 *On termination of* ASCEND

$$u' \in S_{n-1},\ |u'-u| \leq 3\delta/4 \textit{ implies } \phi_N(u') < \phi_N(u) + \frac{\Delta_{\min}\delta^2}{20n}.$$

Proof See Appendix

We now translate Lemma 6 to the v domain.

Lemma 7 *On termination of* ASCEND

$$v' \in S_{n-1},\ |v'-v| \leq \delta/4 \textit{ implies } \psi(v') < \psi(v) + \frac{7\Delta_{\min}\delta^2}{100n}.$$

Proof See Appendix

We can now show that on termination v is close to some standard basis vector e_i. Let $\zeta = \zeta(v)$ (see (2) denote the projection of $\nabla\psi(v)$ orthogonal to v. We consider two possibilities: $|\zeta| \geq \delta$ or $|\zeta| < \delta$.

Lemma 8 *If $|\zeta| \geq \delta$ then there exists v' such that $|v' - v| \leq \delta/4$ and*

$$\psi(v') \geq \psi(v) + \frac{\delta^2}{64\Delta_{\max}(\Delta_{\max} + 3)}.$$

Proof See Appendix

So we can deduce from Lemmas 7 and 8 that ASCEND terminates with $|\zeta| \leq \delta$. This puts a significant restriction on the shape of v.

Lemma 9 *$|\zeta(v)| \leq \delta$ implies that for $1 \leq j \leq n$,*

$$|v_j| \leq \delta^{1/2},$$

or

$$\sqrt{\frac{D}{4\Delta_j}}\left(1 - \frac{n\delta^{1/2}}{3\Delta_{\min}}\right) \leq |v_j| \leq \sqrt{\frac{D}{4\Delta_j}}\left(1 + \frac{n\delta^{1/2}}{3\Delta_{\min}}\right),$$

where $D = 4\sum_{j=1}^{n} \Delta_j v_j^4$.

Proof See Appendix.

We are left with the case where

$$|\zeta(v)| \leq \delta \text{ and } \max\{|v_i| : 1 \leq i \leq n\} \leq 1 - \delta^{1/2}. \tag{7}$$

Lemma 10 *If (7) holds then there exists v' such that $|v' - v| \leq \delta/4$ and*

$$\psi(v') \geq \psi(v) + \frac{3\Delta_{\min}\delta^2}{4n}.$$

Proof See Appendix.

We now give a lemma which summarizes the above discussion.

Lemma 11 *When* ASCEND *terminates, there exists i and $\kappa \in \pm 1$ such that $|u - \kappa R_i| \leq (1 - \alpha)^{-1/2}(\delta^{1/2} + \alpha^{1/2})$.*

Proof See Appendix

Having computed plus or minus one column of R (approximately) we find the remaining columns by working in the subspace orthogonal to it. We then use third moment information to correct signs. We left-multiply our estimate for R by B to obtain our estimate for A.

5. Open Problems, Conclusion It would be very interesting to extend this to other convex sets than parallelepipeds. The immediate target should be simplices. In this connection, it is worthwhile noting a result of Fiedler that the Graph Isomorphism problem is reducible to the problem of the problem of determining whether there is a rotation (an orthonormal transformation) that maps a simplex into another.

It would also be interesting to dispense with the assumption of 4-way independence and consider situations closer to what is done in Factor Analysis - namely where one assumes that the joint distributions of the x_i are in a known class like the normal. This also is related to the problem of learning cubes under other distributions than the uniform.

References

[1] E. B. Baum. Polynomial time algorithms for learning neural nets. In *Proceedings of the Third Annual Workshop on Computational Learning Theory*, pages 258–272. Morgan Kaufmann, 1990.

[2] E. B. Baum. On learning a union of half spaces. *Journal of Complexity*, 6(1):67–101, March 1990.

[3] A. Blum and R. Kannan Learning the intersection of k halfspaces over a uniform distribution *Proceedings of the IEEE Symposium on the Foundations of Computer Science* 1993.

[4] A. Blum and R. Rivest. Training a 3-node neural network is NP-Complete. *Neural Networks*, 5:117–127, 1992.

[5] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM*, 36(4):929–965, 1989.

[6] R. Christensen Linear Models for Multivariate Time Series, and Spatial Data. Springer Texts in Statistics, Springer-Verlag 1991.

[7] M. Kearns, Y. Mansour, D. Ron,R. Rubinfeld,R. Schapire, and L. Sellie. On the Learnability of Discrete Distributions. *Proceedings of the 26 th ACM Symposium on Theory of Computing.*, 1994.

[8] P. M. Long and M. K. Warmuth. Composite geometric concepts and polynomial predictability. In *Proceedings of the Third Annual Workshop on Computational Learning Theory*, pages 273–287. Morgan Kaufmann, 1990.

[9] W. Maass and M.K.Warmuth. Efficient learning with virtual threshold gates. to appear.

[10] N. Megiddo. On the complexity of polyhedral separability. Technical Report RJ 5252, IBM Almaden Research Center, August 1986.

[11] L. G. Valiant. A theory of the learnable. *CACM* 27(11):1134-1142, 1984.

Appendix

Proof of Lemma 6 Write $u' = (u+h)/|u+h|$ where $h = \frac{u'}{u^T u'} - u$ satisfies $h^T u = 0$. Lemma 5 implies that $|h| \leq \delta$. Then

$$\begin{aligned}
\phi_N(u') &= \phi_N(u) + f_N(h) + \mathrm{err}(u,h) \\
&\leq \phi_N(u) + f_N(\xi^*) + \mathrm{err}(u,h) \\
&= \phi_N(\tilde{u}) + \mathrm{err}(u,h) + \mathrm{err}(\tilde{u},\xi^*) \\
&\leq \phi_N(u) + \frac{\Delta_{\min}\delta^2}{30n} + \mathrm{err}(u,h) + \mathrm{err}(\tilde{u},\xi^*) \\
&< \phi_N(u) + \frac{\Delta_{\min}\delta^2}{20n}.
\end{aligned}$$

Proof of Lemma 7 $|v'-v| \leq \delta/4$ implies that $|u'-u| \leq 3\delta/4$ (Lemma 4(b)). So

$$\begin{aligned}
\psi(v') &\leq \phi_N(u') + \frac{\Delta_{\min}\delta^2}{100n} && \text{Lemma 4(a)} \\
&\leq \phi_N(u) + \frac{3\Delta_{\min}\delta^3}{50n} && \text{Lemma 6} \\
&\leq \psi(v) + \frac{7\Delta_{\min}\delta^2}{100n} && \text{Lemma 4(a).}
\end{aligned}$$

Proof of Lemma 8 Let

$$\begin{aligned}
\lambda &= \frac{1}{8(4\Delta_{\max}+3)}\frac{\delta}{|\zeta|} \\
&\geq \frac{\delta}{32\Delta_{\max}(4\Delta_{\max}+3)} && (|\zeta| \leq 4\Delta_{\max})
\end{aligned}$$

and

$$v' = \frac{v+\lambda\zeta}{|v+\lambda\zeta|}.$$

It follows from Lemma 5 that $|v'-v| \leq |\lambda\zeta| < \delta/4$. Also,

$$\psi(v') - 3 = \frac{1}{|v+\lambda\zeta|^4}\left(\psi(v) - 3 + \lambda|\zeta|^2 + \frac{\lambda^2 t^2}{2}\zeta^T H\zeta\right),$$

where $0 \leq t \leq 1$ and $H = \mathrm{Hess}(\psi(v))$. Now

$$|v+\lambda\zeta|^{-4} = (1+\lambda^2|\zeta|^2)^{-2} \geq 1 - 2\lambda^2|\zeta|^2$$

and $|\zeta^T H\zeta| \leq 12\Delta_{\max}|\zeta|^2$ (since $H = \mathrm{diag}(12\Delta_i v_i^2)$ and so

$$\begin{aligned}
\psi(v') - \psi(v) &\geq \lambda|\zeta|^2(1 - (2\lambda + \lambda^3|\zeta|^2)(\psi(v)-3) - 6\Delta_{\max}\lambda) \\
&\geq \frac{\lambda|\zeta|^2}{2}, \qquad (8)
\end{aligned}$$

since $\psi(v) - 3 \leq \Delta_{\max}$ and $|\zeta| \leq 4$.

Proof of Lemma 9

$$\zeta_j(v) = 4\Delta_j v_j^3 - (4\sum_{j=1}^{n} \Delta_j v_j^4) v_j.$$

Suppose $|\zeta(v)| \leq \delta$. Then,

$$|v_j||4\Delta_j v_j^2 - D| \leq \delta.$$

So either

$$\begin{aligned} |v_j| &\leq \delta^{1/2}, \quad \text{or} \\ |4\Delta_j v_j^2 - D| &\leq \delta^{1/2} \quad \text{and so} \\ \left| v_j^2 - \frac{D}{4\Delta_j} \right| &\leq \frac{\delta^{1/2}}{4|\Delta_j|} \quad \text{or} \end{aligned}$$

$$\frac{D}{4\Delta_j}\left(1 - \frac{\delta^{1/2}}{D}\right) \leq v_j^2 \leq \frac{D}{4\Delta_j}\left(1 + \frac{\delta^{1/2}}{D}\right).$$

Let $J = \{j : |v_j| > \delta^{1/2}\}$. Then $\sum_{j \in J} v_j^2 \geq 1 - n\delta$ and so

$$\left(\frac{D + \delta^{1/2}}{4}\right) \sum_{j \in J} \frac{1}{\Delta_j} \geq 1 - n\delta.$$

This implies that

$$\begin{aligned} D &\geq \frac{4(1 - n\delta)\Delta_{\min}}{n} - \delta^{1/2} \\ &\geq \frac{3\Delta_{\min}}{n}. \end{aligned}$$

We are using the fact that $D + \delta^{1/2} > 0$, which follows from the fact that D increases and is initially at least $-\Delta_{\min}\delta^2/n$. □

Proof of Lemma 10 We apply Lemma 9. Putting $K_j = \sqrt{D/(4\Delta_j)}$, let

$$\begin{aligned} J_1 &= \{j : |v_j| \leq \delta^{1/2}\} \\ J_2 &= \{j : K_j(1 - n\delta^{1/2}/(3\Delta_{\min})) \leq |v_j| \leq K_j(1 + n\delta^{1/2}/(3\Delta_{\min}))\}. \end{aligned}$$

Now $|J_2| \geq 2$ else,

$$\begin{aligned} \sum_{i=1}^{n} v_i^2 &\leq (1 - \delta^{1/2})^2 + (n-1)\delta \\ &< 1. \end{aligned}$$

Choose $k, \ell \in J_2$. Define h by

$$\begin{aligned} h_j &= 0, & j \neq k, \ell, \\ h_k &= \tau v_\ell, \\ h_\ell &= -\tau v_k, \end{aligned}$$

where $\tau = \delta/4(v_k^2 + v_\ell^2)^{1/2}$. We can assume that $h^T\zeta \geq 0$, otherwise we replace it by $-h$. Observe that $h^T v = 0$ and $|h| = \delta/4$. Let $v' = (v+h)/|v+h|$. Arguing as in (5) we see that

$$\begin{aligned} \psi(v') - 3 &= \psi(v) - 3 + \zeta^T h + \frac{1}{2} h^T (H - DI) h + 4n\Delta_{\max}|h|^3 \\ &\geq \psi(v) - 3 + 6\tau^2 v_k^2 v_\ell^2 (\Delta_k + \Delta_\ell) - D\delta^2/32 + 4n\Delta_{\max}|h|^3 \\ &\geq \psi(v) - 3 + \frac{3D\delta^2}{8}\left(1 - \frac{n\delta^{1/2}}{3\Delta_{\min}}\right)^{1/2} - D\delta^2/32 + 4n\Delta_{\max}|h|^3 \\ &\geq \psi(v) - 3 + \frac{D\delta^2}{4} \\ &\geq \psi(v) - 3 + \frac{3\Delta_{\min}\delta^2}{4n}. \end{aligned}$$

Proof of Lemma 11 From Lemmas 7, 8 and 10 we see that there exists i such that $|v_i| \geq 1 - \delta^{1/2}$. But then for some $\kappa \in \pm 1$,

$$\begin{aligned} |u - \kappa R_i| &\leq (1-\alpha)^{-1/2} |R^T(u - \kappa R_i)| \\ &\leq (1-\alpha)^{-1/2} (|(v - \kappa e_i| + |\kappa e_i - R^T R e_i|) \\ &\leq (1-\alpha)^{-1/2} (\delta^{1/2} + \alpha). \end{aligned}$$