Learning convex polytopes with margin

Lee-Ad Gottlieb
Ariel University
leead@ariel.ac.il

Eran Kaufman
Ariel University
erankfmn@gmail.com

Aryeh Kontorovich
Ben-Gurion University
karyeh@bgu.sc.il

Gabriel Nivasch
Ariel University
gabrieln@ariel.ac.il

Abstract

We present a near-optimal algorithm for properly learning convex polytopes in the realizable PAC setting from data with a margin. Our first contribution is to identify distinct generalizations of the notion of *margin* from hyperplanes to polytopes and to understand how they relate geometrically; this result may be of interest beyond the learning setting. Our novel learning algorithm constructs a consistent polytope as an intersection of about $t \log t$ halfspaces in time polynomial in t (where t is the number of halfspaces forming an optimal polytope). This is an exponential improvement over the state of the art [Arriaga and Vempala, 2006]. We also improve over the super-polynomial-in-t algorithm of Klivans and Servedio [2008], while achieving a better sample complexity. Finally, we provide the first nearly matching hardness-of-approximation lower bound, whence our claim of near optimality.

1 Introduction

In the theoretical PAC learning setting [Valiant, 1984], one considers an abstract *instance space* $\mathcal{X}$ — which, most commonly, is either the Boolean cube $\{0,1\}^d$ or the Euclidean space $\mathbb{R}^d$. For the former setting, an extensive literature has explored the statistical and computational aspects of learning Boolean functions [Angluin, 1992, Hellerstein and Servedio, 2007]. Yet for the Euclidean setting, a corresponding theory of learning geometric concepts is still being actively developed [Kwek and Pitt, 1998, Jain and Kinber, 2003, Anderson et al., 2013, Kane et al., 2013]. The focus of this paper is the latter setting.

The simplest nontrivial geometric concept is perhaps the halfspace. These concepts are well-known to be hard to agnostically learn [Höffgen et al., 1995] or even approximate [Amaldi and Kann, 1995, 1998, Ben-David et al., 2003]. Even the realizable case, while commonly described as "solved" via the Perceptron algorithm or linear programming (LP), is not straightforward: The Perceptron's runtime is quadratic in the inverse-margin, while solving the consistent hyperplane problem in strongly polynomial time would provide a solution to the general LP problem ([Matoušek and Gärtner, 2006, p. 84] and personal communication from [Anonymous]), a question which has been open for decades [Bárász and Vempala, 2010]. Thus, an unconditional (i.e., infinite-precision and independent of data configuration in space) polynomial-time solution for the consistent hyperplane problem hinges on the strongly polynomial LP conjecture.

If we consider not a single halfspace, but polytopes defined by the intersection of multiple halfspaces, the algorithmic and computational bounds rapidly become more pessimistic. Megiddo [1988] showed that the problem of deciding whether two sets of points in general space can be separated by the intersection of two hyperplanes is NP-complete, and Khot and Saket [2011] showed that "unless

Preprint. Work in progress.

$\mathrm{NP} = \mathrm{RP}$, it is hard to (even) weakly PAC-learn intersection of two halfspaces", even when allowed the richer class of $O(1)$ intersecting halfspaces. Under cryptographic assumptions, Klivans and Sherstov [2009] showed that learning an intersection of n^{ε} halfspaces is intractable regardless of hypothesis representation.

Our results. Since the margin assumption is what allows one to find a consistent hyperplane in provably strongly polynomial time, it is natural to seek to generalize this scheme to intersections of halfspaces. To this end, we identify two distinct notions of polytope margin: These are the γ-*envelope* of a convex polytope, defined as all points within distance γ of the polytope boundary, and the γ-*margin* of the polytope, defined as the intersection of the γ-margins of the hyperplanes forming the polytope. (See Figure 2 for an illustration, and Section 2 for precise definitions.) Note that these two objects may exhibit vastly different behaviors, particularly at a sharp intersection of two or more hyperplanes.

It seems to us that the envelope of a polytope is a more natural structure than its margin, yet we find the margin more amenable to the derivation of both VC-bounds (Lemma 1) and algorithms (Theorem 8). Our first contribution is in demonstrating that results derived for margins can be adapted to apply to envelopes as well. This result may be of independent geometric interest. We prove that when confined to the unit ball, the γ-envelope fully contains within it the $(\gamma^2/2)$-margin (Theorem 5), and this implies that statistical and algorithmic results for the latter hold for the former as well.

We then present the central contribution of the paper, improving algorithmic runtimes for computing separating polytopes. The current state of the art is the algorithm of Arriaga and Vempala [2006], whose runtime is exponential in t/γ^2 (where t is the number of halfspaces forming the polytope, and γ is their margin). In contrast, we give an algorithm whose runtime has only polynomial dependence on t — an *exponential* improvement in speed (Theorem 8). Complementing our algorithm, we provide the first nearly matching hardness-of-approximation bounds, which demonstrate that an exponential dependence on γ^{-2} (but not $t!$) is unavoidable under standard complexity-theoretic assumptions (Theorem 7).

Related work. When general convex bodies are considered under the uniform distribution[1], exponential (in dimension and accuracy) sample-complexity bounds were obtained by Rademacher and Goyal [2009]. This may motivate the consideration of convex polytopes, and indeed a number of works have studied the problem of learning convex polytopes, including Hegedüs [1994], Kwek and Pitt [1998], Anderson et al. [2013], Kane et al. [2013], Kantchelian et al. [2014]. Hegedüs [1994] examines query-based exact identification of convex polytopes with integer vertices, with runtime polynomial in the number of vertices (which could be exponential in the number of faces [Matoušek, 2002]). Kwek and Pitt [1998] also rely on membership queries (see also references therein regarding prior results, as well as strong positive results in dimension 2. Anderson et al. [2013] efficiently approximately recover an unknown simplex from uniform samples inside it. Kane et al. [2013] learn halfspaces under the log-concave distributional assumption.

The recent work of Kantchelian et al. [2014] bears a superficial resemblance to ours, but the two are actually not directly comparable. What they term *worst case margin* will indeed correspond to our *margin*. However, their optimization problem is non-convex, and the solution relies on heuristics without rigorous run-time guarantees. Their generalization bounds exhibit a better dependence on the number t of halfspaces than our Lemma 3 ($O(\sqrt{t})$ vs. our $O(t \log t)$). However, the hinge loss appearing in their Rademacher-based bound could be significantly worse than the 0-1 error appearing in our VC-based bound. We stress, however, that the main contribution of our paper is algorithmic rather than statistical.

The works of Arriaga and Vempala [2006] and Klivans and Servedio [2008] are most comparable to ours. They also consider learning polytopes defined by t hyperplanes with margins, and give learning algorithms for this problem. (The former paper actually constructs a candidate polytope [proper learning], while the latter constructs a function that approximates the polytope's behavior, without constructing the polytope itself [improper learning].) Arriaga and Vempala [2006] present a runtime

[1] Since the concept class of convex sets has infinite VC-dimension, without distribution assumptions, an adversarial distribution can require an arbitrarily large sample size, even in 2 dimensions [Kearns and Vazirani, 1997].

of

$$\left(\frac{\log t}{\gamma}\right)^{O(t \log t \log(1/\gamma))}$$

and sample complexity

$$m = O\left((t/(\varepsilon\gamma^2)) \log(t/\varepsilon) \log(t/(\varepsilon\gamma\delta))\right).$$

Klivans and Servedio [2008] achieve runtime

$$\min\left\{ d(t/\gamma)^{O(t \log t \log(1/\gamma))}, d\left(\frac{\log t}{\gamma}\right)^{O(\sqrt{1/\gamma} \log t)} \right\}$$

(in dimension d) and sample complexity

$$m = O\left((1/\varepsilon)[(1/\gamma)\log(1/(\delta\varepsilon))]^{t \log t \log(1/\gamma)}\right).$$

these algorithms are respectively exponential and super-polynomial in t, while our algorithm has runtime only polynomial in t with near-optimal dependence on the margin.

2 Preliminaries

Notation. For $\mathbf{x} \in \mathbb{R}^d$, we denote its Euclidean norm $\|\mathbf{x}\|_2 := \sqrt{\sum_{i=1}^d \mathbf{x}(i)^2}$ by $\|\mathbf{x}\|$ and for $n \in \mathbb{N}$, we write $[n] := \{1, \ldots, n\}$. Our **instance space** $\mathcal{X}$ is the unit ball in $\mathbb{R}^d$: $\mathcal{X} = \left\{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\| \leq 1\right\}$. We assume familiarity with the notion of VC-dimension as well as with basic PAC definitions such as *generalization error* (see, e.g., Kearns and Vazirani [1997]).

Polytopes. A (convex) polytope $P \subset \mathbb{R}^d$ is the convex hull of finitely many points: $P = \mathrm{conv}(\{\mathbf{x}_1, \ldots, \mathbf{x}_n\})$. Alternatively, it can be defined by t hyperplanes $(\mathbf{w}_i, b_i) \in \mathbb{R}^d \times \mathbb{R}$ where $\|\mathbf{w}_i\| = 1$ for each i:

$$P = \left\{\mathbf{x} \in \mathbb{R}^d : \min_{i \in [t]} \mathbf{w}_i \cdot \mathbf{x} + b_i \geq 0\right\}. \tag{1}$$

A hyperplane $(\mathbf{w}, b)$ is said to classify a point $\mathbf{x}$ as positive (resp., negative) with margin γ if $\mathbf{w} \cdot \mathbf{x} + b \geq \gamma$ (resp., $\leq -\gamma$). Since $\|\mathbf{w}\| = 1$, this means that $\mathbf{x}$ is γ-far from the hyperplane $\left\{\mathbf{x}' \in \mathbb{R}^d : \mathbf{w} \cdot \mathbf{x}' + b = 0\right\}$, in ℓ_2 distance.

Margins and envelopes. We consider two natural ways of extending this notion to polytopes: the γ-margin and the γ-envelope. For a polytope defined by t hyperplanes as in (1), we say that $\mathbf{x}$ is in the *inner γ-margin* of P if

$$0 \leq \min_{i \in [t]} \mathbf{w}_i \cdot \mathbf{x} + b_i \leq \gamma$$

and that $\mathbf{x}$ is in the *outer γ-margin* of P if

$$0 \geq \min_{i \in [t]} \mathbf{w}_i \cdot \mathbf{x} + b_i \geq -\gamma.$$

Similarly, we say that $\mathbf{x}$ is in the *outer γ-envelope* of P if $\mathbf{x} \notin P$ and $\inf_{\mathbf{p} \in P} \|\mathbf{x} - \mathbf{p}\| \leq \gamma$ and that $\mathbf{x}$ is in the *inner γ-envelope* of P if $\mathbf{x} \in P$ and $\inf_{\mathbf{p} \notin P} \|\mathbf{x} - \mathbf{p}\| \leq \gamma$.

We call the union of the inner and the outer γ-margins the *γ-margin*, and we denote it by $\partial P^{[\gamma]}$. Similarly, we call the union of the inner and the outer γ-envelopes the *γ-envelope*, and we denote it by $\partial P^{(\gamma)}$.

The two notions are illustrated in Figure 2. As we show in Section 3 below, the inner envelope coincides with the inner margin, but this is not the case for the outer objects: The outer margin always contains the outer envelope, and could be of arbitrarily larger volume.

Fat hyperplanes and polytopes. Binary classification requires a collection of *concepts* mapping the instance space (in our case, the unit ball in $\mathbb{R}^d$) to $\{-1, 1\}$. However, given a hyperplane $(\mathbf{w}, b)$ and a margin γ, the function $f_{\mathbf{w},b} : \mathbb{R}^d \to \mathbb{R}$ given by $f_{\mathbf{w},b}(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b$ partitions $\mathbb{R}^d$ into three regions: *positive* $\{\mathbf{x} \in \mathbb{R}^d : f_{\mathbf{w},b}(\mathbf{x}) \geq \gamma\}$, *negative* $\{\mathbf{x} \in \mathbb{R}^d : f_{\mathbf{w},b}(\mathbf{x}) \leq -\gamma\}$, and *ambiguous* $\{\mathbf{x} \in \mathbb{R}^d : |f_{\mathbf{w},b}(\mathbf{x})| < \gamma\}$. We use a standard device (see, e.g., Hanneke and Kontorovich [2017, Section 4]) of defining an auxiliary instance space $\mathcal{X}' = \mathcal{X} \times \{-1, 1\}$ together with the concept class $\mathcal{H}_\gamma = \{h_{\mathbf{w},b} : \mathbf{w} \in \mathbb{R}^d, b \in \mathbb{R}, \|\mathbf{w}\| = 1/\gamma\}$, where, for all $(\mathbf{x}, y) \in \mathcal{X}'$,

$$h_{\mathbf{w},b}(\mathbf{x}, y) = \begin{cases} \operatorname{sign}(y(\mathbf{w} \cdot x + b)), & |\mathbf{w} \cdot \mathbf{x} + b| \geq \gamma \\ -1, & \text{else.} \end{cases}$$

It is shown in [Hanneke and Kontorovich, 2017, Lemma 6] that [2]

Lemma 1. *The VC-dimension of $\mathcal{H}_\gamma$ is at most $(2/\gamma + 1)^2$.*

Analogously, we define the concept class $\mathcal{P}_{t,\gamma}$ of γ-fat t-polytopes as follows. Each $h_P \in \mathcal{P}_{t,\gamma}$ is induced by some t-polytope intersection P as in (1). The label of a pair $(\mathbf{x}, y) \in \mathcal{X}'$ is determined as follows: If $\mathbf{x}$ is in the γ-margin of P, then the pair is labeled -1 irrespective of y. Otherwise, if $\mathbf{x} \in P$ and $y = 1$, or else $\mathbf{x} \notin P$ and $y = -1$, then the pair is labeled 1. Otherwise, the pair is labeled -1.

Lemma 2. *The VC-dimension of $\mathcal{P}_{t,\gamma}$ in d dimensions is at most*

$$\max\left\{2(d+1)t\log(3t), 2(v+1)t\log(3t)\right\},$$

where $v = (2/\gamma + 1)^2$.

Proof. The family of intersections of t concept classes of VC-dimension at most v is bounded by $2vt\log(3t)$ Blumer et al. [1989, Lemma 3.2.3]. Since the class of d-dimensional hyperplances has VC-dimension $d + 1$ [Long and Warmuth, 1994], the family of polytopes has VC-dimension at most $2(d + 1)t\log(3t)$. The second part of the bound is obtained by applying Blumer et al. [1989, Lemma 3.2.3] to the VC bound in Lemma 1.

□

Generalization bounds. The following VC-based generalization bounds are well-known; the first one may be found in, e.g., Cristianini and Shawe-Taylor [2000], while the second one in Anthony and Bartlett [1999].

Lemma 3. *Let H be a class of learners with VC-dimension d_{VC}. If a learner $h \in H$ is consistent on a random sample S of size m, then with probability at least $1 - \delta$ its generalization error is*

$$\operatorname{err}(h) \leq \frac{2}{m}\big(d_{\text{VC}}\log(2em/d_{\text{VC}}) + \log(2/\delta)\big).$$

If the learner has empirical error $\widehat{\operatorname{err}}(h)$ on the sample, then with probability at least $1 - \delta$ its generalization error is

$$\operatorname{err}(h) \leq \widehat{\operatorname{err}}(h) + c\sqrt{\frac{d_{\text{VC}} + \log(1/\delta)}{m}}$$

for come universal constant c.

Dimension reduction. The Johnson-Lindenstrauss (JL) transform takes a set S of m vectors in $\mathbb{R}^d$ and projects them into $k = (c\log m)/\varepsilon^2$ dimensions, while preserving all inter-point distances and vectors norms up to $1 + \varepsilon$ distortion. That is, if $f : \mathbb{R}^d \to \mathbb{R}^k$ is a linear embedding realizing the guarantees of the JL transform on S, then for every $\mathbf{x} \in S$ we have

$$(1 - \varepsilon)\|\mathbf{x}\| \leq \|f(\mathbf{x})\| \leq (1 + \varepsilon)\|\mathbf{x}\|,$$

[2] Such estimates may be found in the literature for homogeneous (i.e., $b = 0$) hyperplanes (see, e.g., Bartlett and Shawe-Taylor [1999, Theorem 4.6]), but dealing with polytopes, it is important for us to allow offsets. As discussed in Hanneke and Kontorovich [2017], the standard non-homogeneous to homogeneous conversion can degrade the margin by an arbitrarily large amount, and hence the non-homogeneus case warrants an independent analysis.

and for every $\mathbf{x}, \mathbf{y} \in S$ we have

$$(1-\varepsilon)\|\mathbf{x}-\mathbf{y}\| \leq \|f(\mathbf{x}-\mathbf{y})\| \leq (1+\varepsilon)\|\mathbf{x}-\mathbf{y}\|.$$

The JL transform can be realized with probability $1 - n^{-c}$ for any constant $c \geq 1$ by a randomized linear embedding, for example a projection matrix with entries drawn from a normal distribution [Achlioptas, 2003]. This embedding is *oblivious*, in the sense that the matrix can be chosen without knowledge of the set S.

It is an easy matter to show that the JL transform can also be used to approximately preserve distances to hyperplanes, as in the following lemma.

Lemma 4. *Let S be set of d-dimensional vectors in the unit ball, T be a set of normalized vectors with $|T| \leq |S|$, and $f : \mathbb{R}^d \to \mathbb{R}^k$ a linear embedding realizing the guarantees of the JL transform. Then for any $0 < \varepsilon < 1$ and $k = O((\log |S|)/\varepsilon^2)$ and with probability $1 - |S|^{-c}$ (for any constant $c > 1$), we have for all $\mathbf{x} \in S$ and $\mathbf{t} \in T$ that*

$$f(\mathbf{t}) \cdot f(\mathbf{x}) \in \mathbf{t} \cdot \mathbf{x} \pm \varepsilon.$$

Proof. Let the constant in k be chosen so that the JL transform preserves distances and norms within a factor $1 + \varepsilon'$ for $\varepsilon' = \varepsilon/5$. By the guarantees of the JL transform for the chosen value of k, we have that

$$\begin{aligned}
f(\mathbf{w}) \cdot f(\mathbf{x}) &= \frac{1}{2}\left[\|f(\mathbf{w})\|^2 + \|f(\mathbf{x})\|^2 - \|f(\mathbf{w}) - f(\mathbf{x})\|^2\right] \\
&\leq \frac{1}{2}\left[(1+\varepsilon')^2(\|\mathbf{w}\|^2 + \|\mathbf{x}\|^2) - (1-\varepsilon')^2\|\mathbf{w}-\mathbf{x}\|^2\right] \\
&< \frac{1}{2}\left[(1+3\varepsilon')(\|\mathbf{w}\|^2 + \|\mathbf{x}\|^2) - (1-2\varepsilon')\|\mathbf{w}-\mathbf{x}\|^2\right] \\
&< \frac{1}{2}\left[5\varepsilon'(\|\mathbf{w}\|^2 + \|\mathbf{x}\|^2) + \mathbf{w}\cdot\mathbf{x}\right] \\
&= 5\varepsilon' + \mathbf{w}\cdot\mathbf{x}. \\
&= \varepsilon + \mathbf{w}\cdot\mathbf{x}.
\end{aligned}$$

A similar argument gives that $f(\mathbf{w}) \cdot f(\mathbf{x}) > -\varepsilon + \mathbf{w} \cdot \mathbf{x}$. □

3 Polytope margin and envelope

In this section, we show that the notions of margin and envelope defined in Section 2 are, in general, quite distinct. Fortunately, when confined to the unit ball $\mathcal{X}$, one can be used to approximate the other.

Given two sets $S_1, S_2 \subseteq \mathbb{R}^d$, their *Minkowski sum* is given by $S_1 + S_2 = \{\mathbf{p} + \mathbf{q} : \mathbf{p} \in S_1, \mathbf{q} \in S_2\}$, and their *Minkowski difference* is given by $S_1 - S_2 = \{\mathbf{p} \in \mathbb{R}^d : \{\mathbf{p}\} + S_2 \subseteq S_1\}$. Let $B_\gamma = \{\mathbf{p} \in \mathbb{R}^d : \|\mathbf{p}\| \leq \gamma\}$ be a ball of radius γ centered at the origin.

Given a polytope $P \in \mathbb{R}^d$ an a real number $\gamma > 0$, let

$$P^{(+\gamma)} = P + B_\gamma,$$
$$P^{(-\gamma)} = P - B_\gamma.$$

Hence, $P^{(+\gamma)}$ and $P^{(-\gamma)}$ are the results of expanding or contracting, in a certain sense, the polytope P.

Also, let $P^{[+\gamma]}$ be the result of moving each halfspace defining a facet of P outwards by distance γ, and similarly, let $P^{[-\gamma]}$ be the result of moving each such halfspace inwards by distance γ. Put differently, we can think of the halfspaces defining the facets of P as moving outwards at unit speed, so P expands with time. Then $P^{[\pm\gamma]}$ is P at time $\pm\gamma$. See Figure 1.

Observation 1. *We have $P^{(-\gamma)} = P^{[-\gamma]}$.*

Proof. Each point in $P^{[-\gamma]}$ is at distance at least γ from each hyperplane containing a facet of P, hence, it is at distance at least γ from the boundary of P, so it is in $P^{(-\gamma)}$. Now, suppose for a

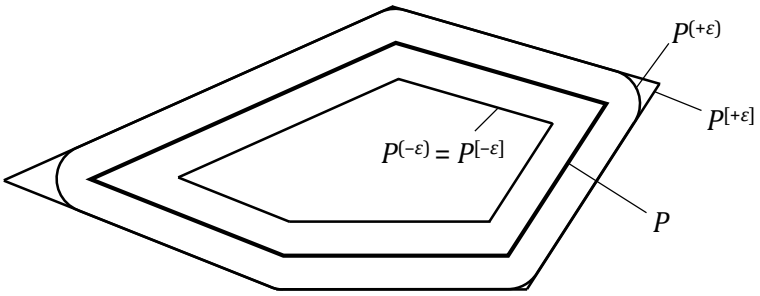

Figure 1: Expansion and contraction of a polytope by γ.

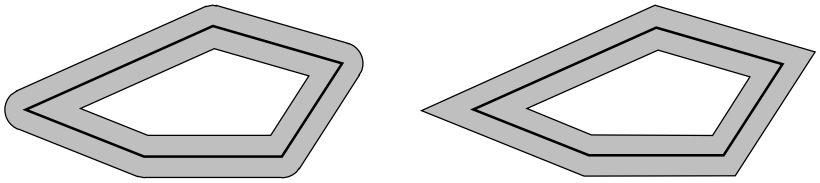

Figure 2: The γ-envelope $\partial P^{(\gamma)}$ (left) and γ-margin $\partial P^{[\gamma]}$ (right) of a polytope P.

contradiction that there exists a point $\mathbf{p} \in P^{(-\gamma)} \setminus P^{[-\gamma]}$. Then $\mathbf{p}$ is at distance less than γ from a point $\mathbf{q} \in \partial h \setminus f$, where f is some facet of P and ∂h is the hyperplane containing f. But then the segment $\mathbf{pq}$ must intersect another facet of P. □

However, in the other direction we have $P^{(+\gamma)} \subsetneq P^{[+\gamma]}$. Furthermore, the Hausdorff distance between them could be arbitrarily large (see again Figure 1).

Then the γ-envelope of P is given by $\partial P^{(\gamma)} = P^{(+\gamma)} \setminus P^{(-\gamma)}$, and the γ-margin of P is given by $\partial P^{[\gamma]} = P^{[+\gamma]} \setminus P^{[-\gamma]}$. See Figure 2.

Since the γ-margin of P is not contained in the γ-envelope of P, we would like to find some sufficient condition under which, for some $\gamma' < \gamma$, the γ'-margin of P is contained in the γ-envelope of P. Our solution to this problem is given in the following theorem. Recall that $\mathcal{X}$ is the unit ball in $\mathbb{R}^d$.

Theorem 5. *Let $P \subset \mathbb{R}^d$ be a polytope, and let $0 < \gamma < 1$. Suppose that $P^{[-\gamma]} \cap \mathcal{X} \neq \emptyset$. Then, within $\mathcal{X}$, the $(\gamma^2/2)$-margin of P is contained in the γ-envelope of P; meaning, $\partial P^{[\gamma^2/2]} \cap \mathcal{X} \subseteq \partial P^{(\gamma)}$.*

The proof uses the following general observation:

Observation 2. *Let $Q = Q(t)$ be an expanding polytope whose defining halfspaces move outwards with time, each one at its own constant speed. Let $\mathbf{p} = \mathbf{p}(t)$ be a point that moves in a straight line at constant speed. Suppose $t_1 < t_2 < t_3$ are such that $\mathbf{p}(t_1) \in Q(t_1)$ and $\mathbf{p}(t_3) \in Q(t_3)$. Then $\mathbf{p}(t_2) \in Q(t_2)$ as well.*

Proof. Otherwise, $\mathbf{p}$ exits one of the halfspaces and enters it again, which is impossible. □

Proof of Theorem 5. By Observation 1, it suffices to show that $P^{[+\gamma^2/2]} \cap \mathcal{X} \subseteq P^{(+\gamma)}$. Hence, let $\mathbf{p} \in P^{[+\gamma^2/2]} \cap \mathcal{X}$ and $\mathbf{q} \in P^{[-\gamma]} \cap \mathcal{X}$. Let s be the segment $\mathbf{pq}$. Let $\mathbf{r}$ be the point in s that is at distance γ from $\mathbf{p}$. Suppose for a contradiction that $\mathbf{p} \notin P^{(+\gamma)}$. Then $\mathbf{r} \notin P$. Consider $P = P(t)$ as a polytope that expands with time, as above. Let $\mathbf{z} = \mathbf{z}(t)$ be a point that moves along s at constant speed, such that $\mathbf{z}(-\gamma) = \mathbf{q}$ and $\mathbf{z}(\gamma^2/2) = \mathbf{p}$. Since $\|\mathbf{r} - \mathbf{q}\| \leq 2$, the speed of s is at most $2/\gamma$. Hence, between $t = 0$ and $t = \gamma^2/2$, $\mathbf{z}$ moves distance at most γ, so $\mathbf{z}(0)$ is already between $\mathbf{r}$ and $\mathbf{p}$. In other words, $\mathbf{z}$ exits P and reenters it, contradicting Observation 2. □

It follows immediately from Theorem 5 and Lemma 2 that the VC-dimension of the class of t-polytopes with envelope γ is at most

$$\max\left\{2(d+1)t\log(3t), 2(v+1)t\log(3t)\right\},$$

where $v = (4/\gamma^2 + 1)^2$. Likewise, we can approximate the optimal t-polytope with envelope γ by the algorithms of Theorem 8 (with parameter $\gamma' = \gamma^2/2$).

4 Computing and learning separating polytopes

In this section, we present algorithms to compute and learn γ-fat t-polytopes. We begin with hardness results for this problem, and show that these hardness results justify algorithms with run time exponential in the dimension or the square of the reciprocal of the margin, but not exponential in t. We then present our algorithms.

4.1 Hardness

We show that computing separating polytopes is NP-hard, and even hard to approximate. We begin with the case of a single hyperplane. The following preliminary lemma builds upon [Amaldi and Kann, 1995, Theorem 10].

Lemma 6. *Given a labelled point set S, let h^* be a hyperplane that places all positive points of S on its positive side, and maximizes the number of negative points on its negative size — call this quantity* opt. *Then it is* NP*-hard to find a hyperplane $\tilde{h}$ consistent with all positive points, and which places at least* $\mathrm{opt}/|S|^{1-o(1)}$ *negative points on on the negative side of $\tilde{h}$. This holds even when the optimal hyperplane has margin exactly $1/(4\sqrt{\mathrm{opt}})$.*

Proof. We reduce from maximum independent set, which is hard to approximate to within $n^{1+o(1)}$ [Zuckerman, 2007]. Given a graph $G = (V, E)$, for each vetex $v_i \in V$ place a negative point on the basis vector $\mathbf{e}_i$. Now place a positive point at the origin, and for each edge $(v_i, v_j) \in E$, place a positive point at $(\mathbf{e}_i + \mathbf{e}_j)/2$.

Consider a hyperplane consistent with the positive points and placing m negative points on the negative side: These negative points must represent an independent set in G, for if $(v_i, v_j) \in E$, then by construction the midpoint of $\mathbf{e}_i, \mathbf{e}_j$ is positive, and so both $\mathbf{e}_i, \mathbf{e}_j$ cannot lie on the negative side of the plane.

Likewise, if G contained an independent set $V' \subset V$ of size m, then we consider the hyperplane defined by the equation $\mathbf{w} \cdot \mathbf{x} + 3/(4\sqrt{m}) = 0$, where coordinate $\mathbf{w}(j) = -m^{-1/2}$ if $\mathbf{v}_j \in V'$ and $\mathbf{w}(j) = 0$ otherwise. It is easily verified that the distance from the hyperplane to a negative point is $-1/(4\sqrt{m})$, to the origin is $3/(4\sqrt{m})$, and to other positive points are at least $1/(4\sqrt{m})$. □

Theorem 7. *Given a labelled point set S, let H^* be a collection of t hyperplanes whose intersection partitions S into positive and negative sets. Then it is* NP*-hard to find a collection $\tilde{H}$ of size less than $t|S|^{1-o(1)}$ whose intersection partitions S into positive and negative sets. This holds even when all hyperplanes have margin $\sqrt{t/n}$ or greater.*

Proof. The reduction is from minimum coloring, which is hard to approximate within a factor of $n^{1-o(1)}$ [Zuckerman, 2007]. The construction is identical to that of the proof of Lemma 6. The only difference is that if one color covers more than n/t vertices, we break it up into a set of colors, each covering at most n/t vertices. This increases the total number of colors to at most $2t$. The claim follows. □

4.2 Algorithms

Here we present algorithms for computing polytopes, and use them to give an efficient algorithm for learning polytopes. As a consequence of Lemma 6 and Theorem 7, we cannot hope to find in polynomial time even a single hyperplane consistent with the positive point which correctly classifies many negative points, let alone a t-polytope consistent with all the data. In fact, by the Exponential Time Hypothesis, we cannot achieve a runtime better than exponential in either the dimension or $(1/\gamma)^{2-o(1)}$ — but not necessarily exponential in t.

In what follows, we give two algorithms inspired by the polytope algorithm presented by Arriaga and Vempala [2006]. Both have runtime faster than the algorithm of Arriaga and Vempala [2006], and the second is only polynomial in t.

Theorem 8. *Given a labelled point set S ($n = |S|$) for which some γ-fat t-polytope ($t \leq n$) correctly separates the positive and negative points (i.e., the polytope is* consistent*), we can compute the following with high probability:*

1. *A consistent $(\gamma/2)$-fat t-polytope in time $n^{O((t/\gamma^2)\log(1/\gamma))}$.*
2. *A consistent $(\gamma/2)$-fat $O(t \log n)$-polytope in time $n^{O(\gamma^{-2}\log(1/\gamma))}$.*

Before proving Theorem 8, we will need a preliminary lemma:

Lemma 9. *Given any $0 < \delta < 1$, there exists a set V of unit vectors of size $|V| \leq (1 + 2/\delta)^d$ with the following property: For any unit vector $\mathbf{w}$, there exists a $\mathbf{v} \in V$ that satisfies $\mathbf{v} \cdot \mathbf{x} \in \mathbf{w} \cdot \mathbf{x} \pm \delta$ for all vectors $\mathbf{x}$ with $\|\mathbf{x}\| \leq 1$. The set V can be constructed in time $\delta^{-O(d)}$ with high probability.*

This implies that if a set S admits a hyperplane with margin γ, the set V contains a hyperplane with margin at least $\gamma - \delta$.

Proof. We take V to be a δ-net of the unit ball, a set satisfying that every point on the ball is within distance δ of some point in V. Then $|V| \leq (1 + 2/\delta)^d$ [Vershynin, 2010, Lemma 5.2]. For any unit vector $\mathbf{v}$ we have for some $\mathbf{w} \in V$ that $\|\mathbf{w} - \mathbf{v}\| \leq \delta$. Now let $\mathbf{w}, \mathbf{v}$ be vectors normal to their respective hyperplanes, and so for any vector $\mathbf{x}$ satisfying $\|\mathbf{x}\| \leq 1$ we have that its distance from the respective hyperplanes is $\mathbf{w} \cdot \mathbf{x}$ and $\mathbf{v} \cdot \mathbf{x}$. It follows that

$$|\mathbf{w} \cdot \mathbf{x} - \mathbf{v} \cdot \mathbf{x}| = |(\mathbf{w} - \mathbf{v}) \cdot \mathbf{x}| \leq \|\mathbf{w} - \mathbf{v}\| \leq \delta.$$

The net can be constructed by a randomized greedy algorithm. By coupon-collector analysis, it suffices to construct $O(|V| \log |V|)$ random vectors. For example, each can be chosen by sampling each coordinate from a (0,1)-Normal distribution, and then normalizing the vector. Then the resulting set contains within in a δ-net, which can be extracted in the stated time. □

Proof of Theorem 8. We first apply the Johnson-Lindenstrauss transform to reduce dimension of the points in S to $k = O((\log n)/\gamma^2)$, achieving the guarantees of Lemma 4 with parameter $\varepsilon = \gamma/8$. We then restrict our attention to the k-dimensional hyperplanes implied by the δ-net V of Lemma 9 with parameter $\delta = \gamma/8$. Then for each d-dimensional hyperplane $\mathbf{w}$ forming the original γ-fat t-polytope, there is a k-dimensional hyperplane $\mathbf{w}'$ of the net satisfying $\mathbf{w}' \cdot f(\mathbf{x}) \in \mathbf{w} \cdot \mathbf{x} \pm \gamma/4$ for every $\mathbf{x} \in S$. Given this $\mathbf{w}'$, we can recover an approximation to $\mathbf{w}$ thus: Let $S' \subset S$ include only those points $\mathbf{x} \in S$ for which $|\mathbf{w}' \cdot f(\mathbf{x})| \geq (3/4)\gamma$, from which it follows that $|\mathbf{w} \cdot \mathbf{x}| \geq \gamma/2$. We then run the Perceptron algorithm on S' in time $O(dn/\gamma^2)$, and find a hyperplane $\mathbf{w}''$ consistent with $\mathbf{w}$ on all points at distance $\gamma/2$ or more from $\mathbf{w}$. We will refer to $\mathbf{w}''$ as the d-dimensional mirror of $\mathbf{w}'$.

Having enumerated all vectors in V and computed their d-dimensional mirrors, we enumerate all possible t-polytopes by taking all combinations of t mirror hyperplanes, in total time

$$(1/\gamma)^{O(kt)} = n^{O((t/\gamma^2)\log(1/\gamma))},$$

and choose the best one consistent with S. The first part of the theorem follows.

Alternatively, we may give a greedy algorithm with better runtime: First note that as the intersection of t hyperplanes correctly classifies all points, the best hyperplane among them correctly classifies at least a $(1/t)$-fraction of the negative points with margin γ. Hence it suffices to compute the mirror hyperplane which is consistent with all positive points and maximizes the number of correct negative points, all with margin $\gamma/2$. We choose this hyperplane, remove from S the correctly placed negative points, and iteratively search for the next best hyperplane. After $ct \log n$ iterations (for an appropriate constant c), the number of remaining points is

$$n(1 - \Omega(1/t))^{ct \log n} = ne^{-\ln n/2} < 1,$$

and the algorithm terminates. □

Having given an algorithm to *compute* γ-fat t-polytopes, we can now give an efficient algorithm to *learn* γ-fat t-polytopes. We sample m points, and use the second item of Theorem 8 to find a $(\gamma/2)$-fat $O(t \log m)$-polytope consistent with the sample. By Lemma 2, the class of polytopes has VC-dimension $(1/\gamma^2)t \log m$. The size of m is chosen according to Lemma 3, and we conclude:

Theorem 10. *There exists an algorithm that learns γ-fat t-polytopes with sample complexity*

$$m = \min\left\{O\Big(\frac{t}{\varepsilon\gamma^2}\log^2\frac{t}{\varepsilon\gamma^2} + \log\frac{1}{\delta}\Big), O\Big(\frac{1}{\varepsilon^2}\Big(\frac{t^2}{\gamma^2}\log\frac{t}{(\varepsilon\gamma)^2} + \log\frac{1}{\delta}\Big)\Big)\right\}$$

in time $m^{O((1/\gamma^2)\log(1/\gamma))}$, where δ is the desired confidence level.

The stated runtime is only polynomial in t, improving over that of Arriaga and Vempala [2006], whose algorithm was exponential in t, and over that of Klivans and Servedio [2008], whose algorithm was super-polynomial in t (and also had worse sample complexity).

References

Dimitris Achlioptas. Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *J. Comput. Syst. Sci.*, 66(4):671–687, 2003. doi: 10.1016/S0022-0000(03)00025-4. URL https://doi.org/10.1016/S0022-0000(03)00025-4.

Edoardo Amaldi and Viggo Kann. The complexity and approximability of finding maximum feasible subsystems of linear relations. *Theoretical Computer Science*, 147(1):181 – 210, 1995. ISSN 0304-3975. doi: https://doi.org/10.1016/0304-3975(94)00254-G. URL http://www.sciencedirect.com/science/article/pii/030439759400254G.

Edoardo Amaldi and Viggo Kann. On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems. *Theoretical Computer Science*, 209(1):237 – 260, 1998. ISSN 0304-3975. doi: https://doi.org/10.1016/S0304-3975(97)00115-1. URL http://www.sciencedirect.com/science/article/pii/S0304397597001151.

Joseph Anderson, Navin Goyal, and Luis Rademacher. Efficient learning of simplices. In *COLT 2013 - The 26th Annual Conference on Learning Theory, June 12-14, 2013, Princeton University, NJ, USA*, pages 1020–1045, 2013. URL http://jmlr.org/proceedings/papers/v30/Anderson13.html.

Dana Angluin. Computational learning theory: Survey and selected bibliography. In *Proceedings of the 24th Annual ACM Symposium on Theory of Computing, May 4-6, 1992, Victoria, British Columbia, Canada*, pages 351–369, 1992. doi: 10.1145/129712.129746. URL http://doi.acm.org/10.1145/129712.129746.

Martin Anthony and Peter L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, Cambridge, 1999. ISBN 0-521-57353-X. doi: 10.1017/CBO9780511624216. URL http://dx.doi.org/10.1017/CBO9780511624216.

Rosa I. Arriaga and Santosh Vempala. An algorithmic theory of learning: Robust concepts and random projection. *Machine Learning*, 63(2):161–182, 2006. doi: 10.1007/s10994-006-6265-7. URL https://doi.org/10.1007/s10994-006-6265-7.

Mihály Bárász and Santosh Vempala. A new approach to strongly polynomial linear programming. In *Innovations in Computer Science - ICS 2010, Tsinghua University, Beijing, China, January 5-7, 2010. Proceedings*, pages 42–48, 2010. URL http://conference.itcs.tsinghua.edu.cn/ICS2010/content/papers/4.html.

Peter Bartlett and John Shawe-Taylor. *Generalization performance of support vector machines and other pattern classifiers*, pages 43–54. MIT Press, Cambridge, MA, USA, 1999. ISBN 0-262-19416-3.

Shai Ben-David, Nadav Eiron, and Philip M. Long. On the difficulty of approximately maximizing agreements. *J. Comput. Syst. Sci.*, 66(3):496–514, 2003. doi: 10.1016/S0022-0000(03)00038-2. URL https://doi.org/10.1016/S0022-0000(03)00038-2.

Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *J. Assoc. Comput. Mach.*, 36(4):929–965, 1989. ISSN 0004-5411.

Nello Cristianini and John Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, 2000. ISBN 0521780195. URL https://www.amazon.com/Introduction-Support-Machines-Kernel-based-Learning/dp/0521780195?SubscriptionId=0JYN1NVW651KCA56C102&tag=techkie-20&linkCode=xm2&camp=2025&creative=165953&creativeASIN=0521780195.

Steve Hanneke and Aryeh Kontorovich. Optimality of SVM: Novel proofs and tighter bounds. 2017. URL https://www.cs.bgu.ac.il/~karyeh/opt-svm.pdf.

Tibor Hegedüs. Geometrical concept learning and convex polytopes. In *Proceedings of the Seventh Annual ACM Conference on Computational Learning Theory, COLT 1994, New Brunswick, NJ, USA, July 12-15, 1994.*, pages 228–236, 1994. doi: 10.1145/180139.181124. URL http://doi.acm.org/10.1145/180139.181124.

Lisa Hellerstein and Rocco A. Servedio. On PAC learning algorithms for rich boolean function classes. *Theor. Comput. Sci.*, 384(1):66–76, 2007. doi: 10.1016/j.tcs.2007.05.018. URL https://doi.org/10.1016/j.tcs.2007.05.018.

Klaus-Uwe Höffgen, Hans Ulrich Simon, and Kevin S. Van Horn. Robust trainability of single neurons. *J. Comput. Syst. Sci.*, 50(1):114–125, 1995. doi: 10.1006/jcss.1995.1011. URL https://doi.org/10.1006/jcss.1995.1011.

Sanjay Jain and Efim B. Kinber. Intrinsic complexity of learning geometrical concepts from positive data. *J. Comput. Syst. Sci.*, 67(3):546–607, 2003. doi: 10.1016/S0022-0000(03)00067-9. URL https://doi.org/10.1016/S0022-0000(03)00067-9.

Daniel M. Kane, Adam R. Klivans, and Raghu Meka. Learning halfspaces under log-concave densities: Polynomial approximations and moment matching. In *COLT 2013 - The 26th Annual Conference on Learning Theory, June 12-14, 2013, Princeton University, NJ, USA*, pages 522–545, 2013. URL http://jmlr.org/proceedings/papers/v30/Kane13.html.

Alex Kantchelian, Michael Carl Tschantz, Ling Huang, Peter L. Bartlett, Anthony D. Joseph, and J. Doug Tygar. Large-margin convex polytope machine. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3248–3256, 2014. URL http://papers.nips.cc/paper/5511-large-margin-convex-polytope-machine.

Micheal Kearns and Umesh Vazirani. *An Introduction to Computational Learning Theory*. The MIT Press, 1997.

Subhash Khot and Rishi Saket. On the hardness of learning intersections of two halfspaces. *J. Comput. Syst. Sci.*, 77(1):129–141, 2011. doi: 10.1016/j.jcss.2010.06.010. URL https://doi.org/10.1016/j.jcss.2010.06.010.

Adam R. Klivans and Rocco A. Servedio. Learning intersections of halfspaces with a margin. *J. Comput. Syst. Sci.*, 74(1):35–48, 2008. doi: 10.1016/j.jcss.2007.04.012. URL https://doi.org/10.1016/j.jcss.2007.04.012.

Adam R. Klivans and Alexander A. Sherstov. Cryptographic hardness for learning intersections of halfspaces. *J. Comput. Syst. Sci.*, 75(1):2–12, 2009. doi: 10.1016/j.jcss.2008.07.008. URL https://doi.org/10.1016/j.jcss.2008.07.008.

Stephen Kwek and Leonard Pitt. PAC learning intersections of halfspaces with membership queries. *Algorithmica*, 22(1/2):53–75, 1998. doi: 10.1007/PL00013834. URL https://doi.org/10.1007/PL00013834.

Philip M. Long and Manfred K. Warmuth. Composite geometric concepts and polynomial predictability. *Inf. Comput.*, 113(2):230–252, 1994. doi: 10.1006/inco.1994.1071. URL https://doi.org/10.1006/inco.1994.1071.

Jiří Matoušek. *Lectures on discrete geometry*, volume 212 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 2002. ISBN 0-387-95373-6. doi: 10.1007/978-1-4613-0039-7. URL https://doi.org/10.1007/978-1-4613-0039-7.

Jiří Matoušek and Bernd Gärtner. *Understanding and Using Linear Programming (Universitext)*. Springer, 2006. ISBN 3540306978. URL https://www.amazon.com/Understanding-Using-Linear-Programming-Universitext/dp/3540306978?SubscriptionId=0JYN1NVW651KCA56C102&tag=techkie-20&linkCode=xm2&camp=2025&creative=165953&creativeASIN=3540306978.

Nimrod Megiddo. On the complexity of polyhedral separability. *Discrete & Computational Geometry*, 3(4):325–337, Dec 1988. ISSN 1432-0444. doi: 10.1007/BF02187916. URL https://doi.org/10.1007/BF02187916.

Luis Rademacher and Navin Goyal. Learning convex bodies is hard. In *COLT 2009 - The 22nd Conference on Learning Theory, Montreal, Quebec, Canada, June 18-21, 2009*, 2009. URL http://www.cs.mcgill.ca/~colt2009/papers/030.pdf#page=1.

Leslie G. Valiant. A theory of the learnable. *Commun. ACM*, 27(11):1134–1142, 1984.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *CoRR*, abs/1011.3027, 2010. URL http://arxiv.org/abs/1011.3027.

David Zuckerman. Linear degree extractors and the inapproximability of max clique and chromatic number. *Theory of Computing*, 3(6):103–128, 2007. doi: 10.4086/toc.2007.v003a006. URL http://www.theoryofcomputing.org/articles/v003a006.