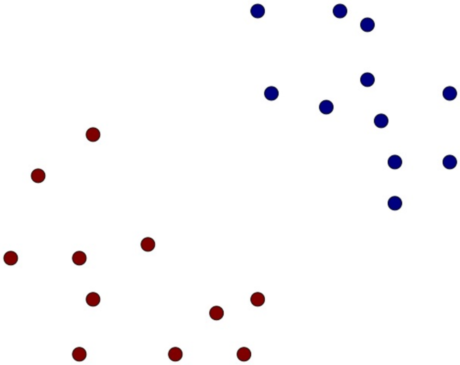
On the Furthest Hyperplane Problem and Maximal Margin Clustering

Zohar S. Karnin [1] Edo Liberty[2] Shachar Lovett [3] Roy Schwartz [4] Omri Weinstein[5]

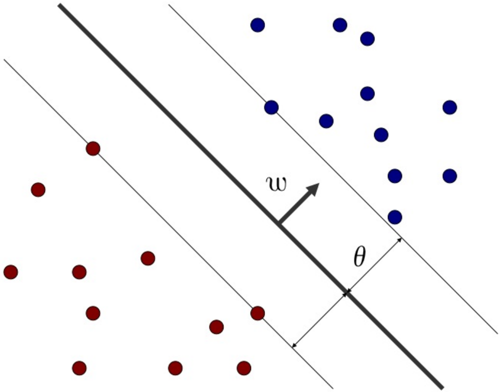
[1]Yahoo! Research zkarnin@yahoo-inc.com.
[2]Yahoo! Research, edo@yahoo-inc.com.
[3]IAS, slovett@math.ias.edu. Supported by DMS-0835373.
[4]Technion Institute of Technology and Yahoo! Research roys@yahoo-inc.com.
[5]Princeton University and Yahoo! Research oweinste@cs.princeton.edu.

Supervised SVMs



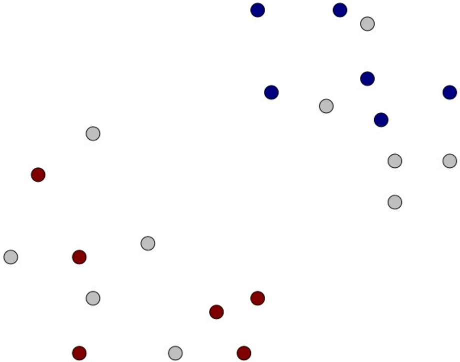
Solving fully separable SVMs is a textbook classic.

Supervised SVMs



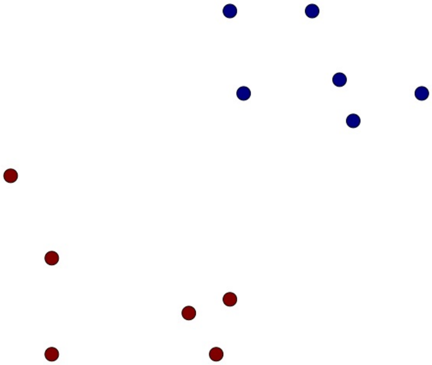
The solution w maximizes the margin $(\langle w, x^{(i)} \rangle + b)y_i \geq \theta$.

Semi-supervised SVMs



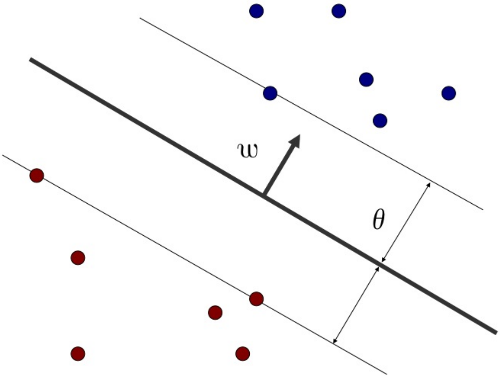
In reality most example labels are not known (that's why we learn).

Semi-supervised SVMs



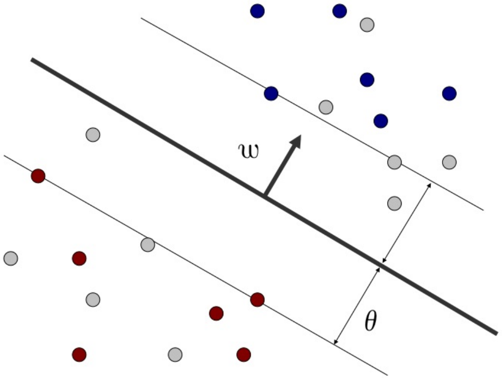
One option is to ignore the unlabeled points....

Semi-supervised SVMs



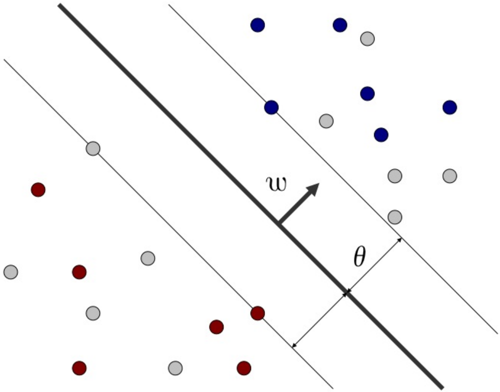
... and solve the SVM problem on the labeled ones.

Semi-supervised SVMs



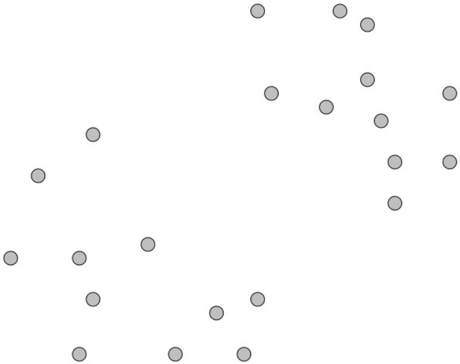
This might lead to suboptimal results.

Semi-supervised SVMs



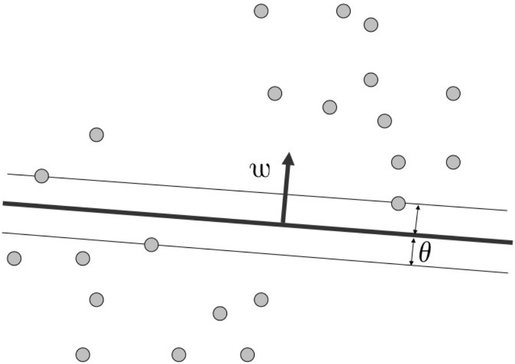
Semi-supervised SVMs were shown to be practicaly useful [1][2][3][4].

Unsupervised SVMs



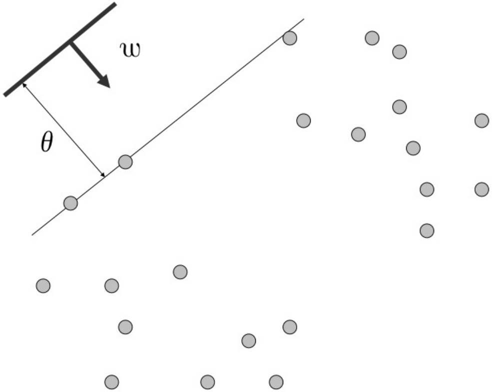
How about completely unsupervised SVMs?

Unsupervised SVMs



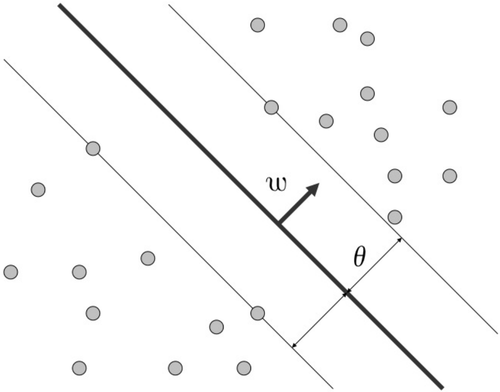
These are always fully separable.

Unsupervised SVMs



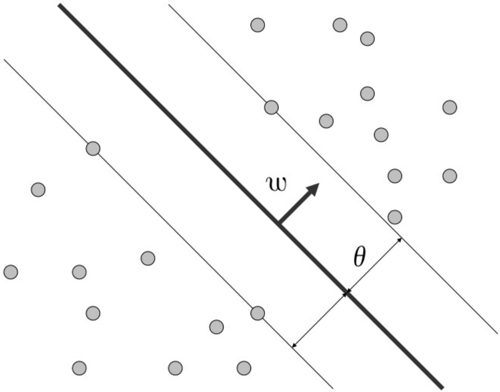
There are also trivial unbounded solutions.

Unsupervised SVMs



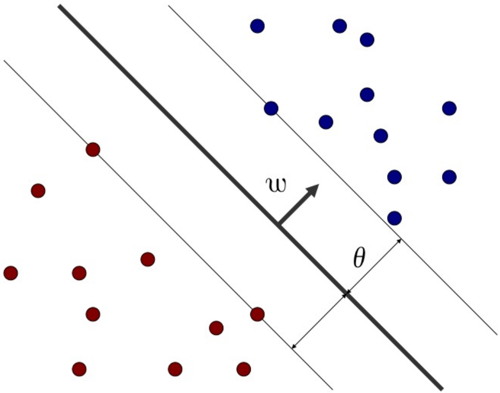
But there is one separator which maximizes the margin $|\langle w, x^{(i)} \rangle + b| \geq \theta$

Unsupervised SVMs



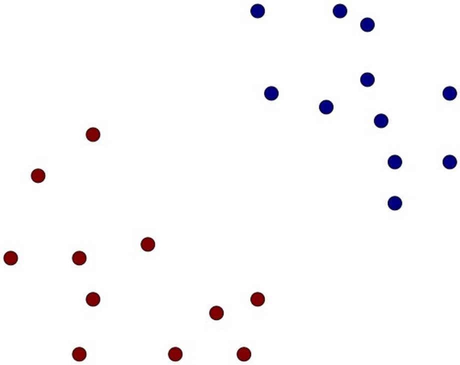
Consider the labels obtained by the separator $\text{sign}(\langle w, x^{(i)} \rangle + b)$

Unsupervised SVMs



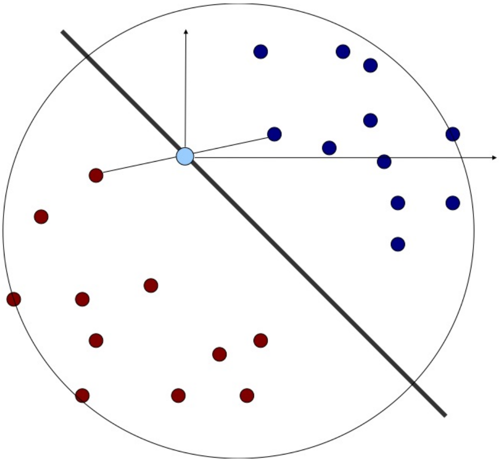
They should be correct under the right assumptions.

Unsupervised SVMs



They should be correct under the right assumptions.

Furthest hyperplane problem



W.l.o.g., hyperplane passes through origin ($b = 0$), and $\|x_i\| \leq 1$.

Furthest hyperplane problem

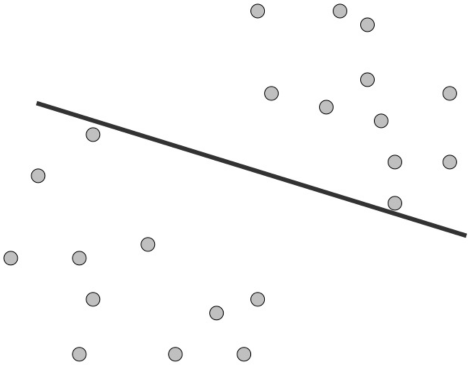
FHP

$$\begin{aligned} \text{Maximize} \quad & \theta' \\ \text{s.t} \quad \|w\|^2 &= 1 \\ \forall\, 1 \le i \le n \quad |\langle w \cdot x_i \rangle| &\ge \theta' \end{aligned}$$

(alternatively)

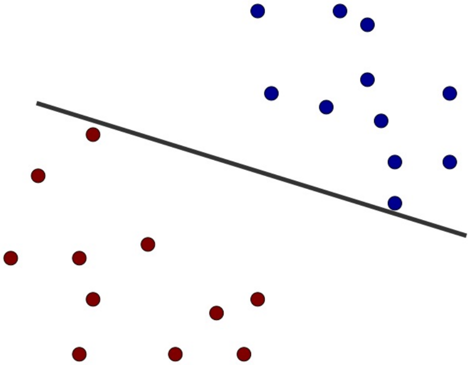
$$\begin{aligned} \text{Minimize} \quad & \|w\|^2 \\ \forall\, 1 \le i \le n \quad |\langle w \cdot x_i \rangle| &\ge 1 \end{aligned}$$

Exact solution



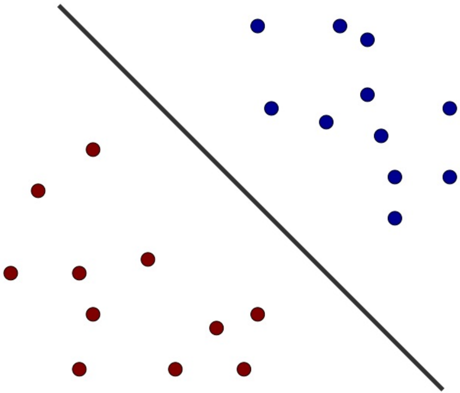
Observation: many separators are “optimal” in a sense.

Exact solution



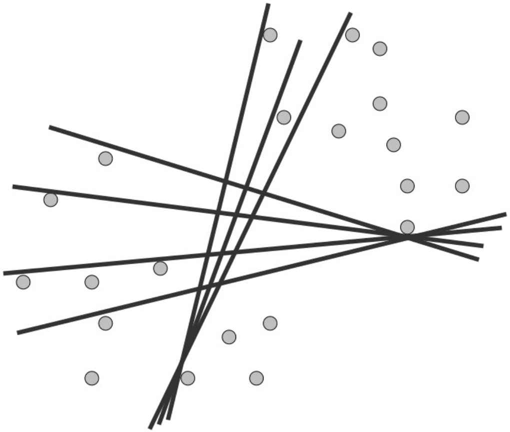
Those that generate the correct labeling.

Exact solution



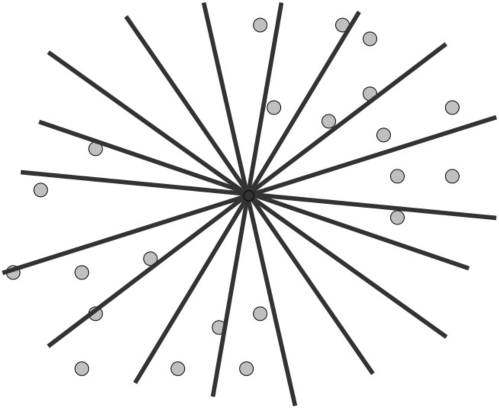
From the correct labeling it is possible to solve exactly.

Exact solution



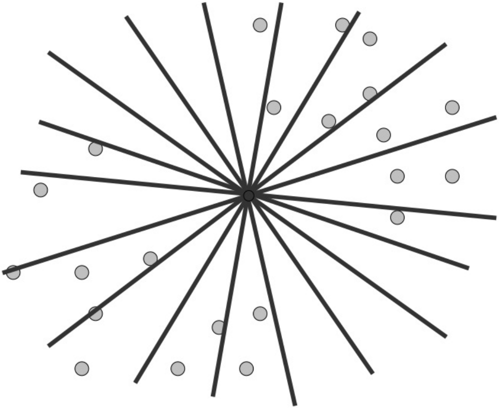
Solution 1: Consider $O(n^d)$ linear partitions (Sauer's Lemma + VC dim)

Exact solution



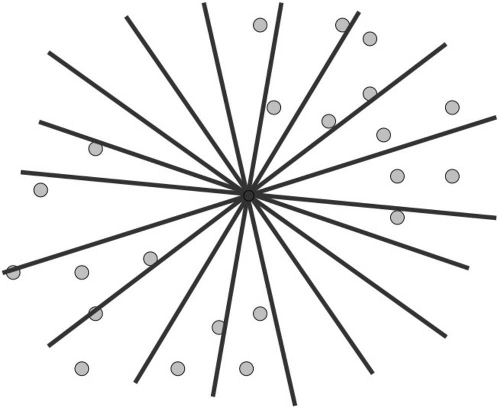
Solution 2: Consider $(1/\theta)^{O(d)}$ separators from and ε-net.

Exact solution



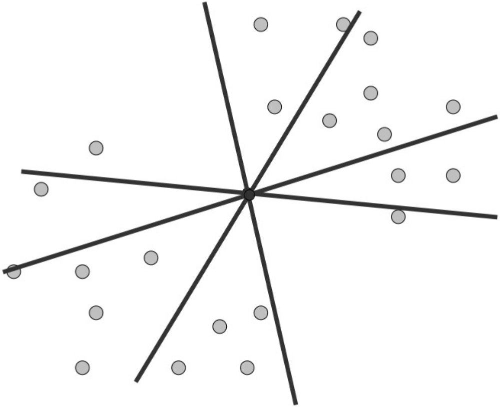
Solution 3: Randomly project to $k = O(\log(n)/\theta^2)$ (margin preserved [5]).

Exact solution



Solution 3: ε-net yields $(1/\theta)^{O(k)} = n^{O(\log(1/\theta)/\theta^2)}$ candidates.

Exact solution



Solution 4: Choose $n^{O(1/\theta^2)}$ random hyperplanes.

Hardness of approximation

There is no PTAS for FHP unless P=NP.

1. MAX-3SAT(13) is hard to approximate [6].
2. MAX-3SAT(13) reduces to SYM(30) (Symmetric CNF).
3. SYM(30) reduces of FHP.

Theorem

It is NP-hard to distinguish whether FHP *admits margin* $\frac{1}{\sqrt{d}}$ *or at most* $(1-\varepsilon)\frac{1}{\sqrt{d}}$ *for some constant* ε

The consequence of this is that:

Lemma

The random hyperplane solution is optimal.
Otherwise 3-SAT is solvable in $2^{o(n)}$.

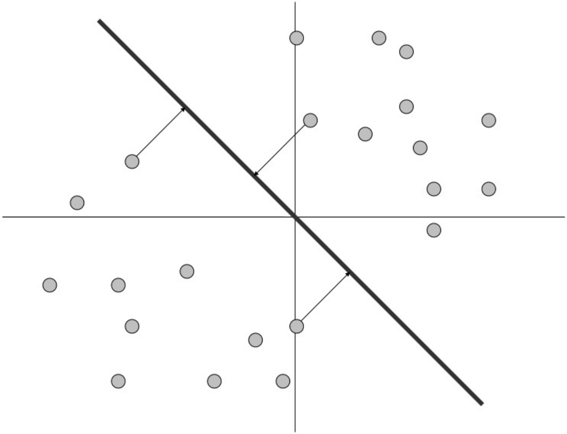
FHP approximation algorithm

Input: Set of points $x_1, \ldots, x_n \in \mathbb{R}^d$
Output: $w \in \mathbb{S}^{d-1}$
$\forall i \in [n]\ \tau_1(i) \leftarrow 1 \ ; \quad j \leftarrow 1$
while $\sum_{i=1}^{n} \tau_j(i) \geq 1/n$ **do**
  $A_j \leftarrow n \times d$ matrix whose i'th row is $\sqrt{\tau_j(i)} \cdot x_i$
  $w^{(j)} \leftarrow$ top right singular vector of A_j
  $\sigma_j(i) \leftarrow \left|\langle x_i, w^{(j)} \rangle\right|$
  $\tau_{j+1}(i) \leftarrow \tau_j(i) c^{-\sigma_j^2(i)}$
  $j \leftarrow j + 1$
end while
$w' \leftarrow \sum_{j=1}^{t} g_j \cdot w^{(j)}$ for $g_j \sim \mathcal{N}(0,1)$
return: $w \leftarrow w'/\|w'\|$

Theorem

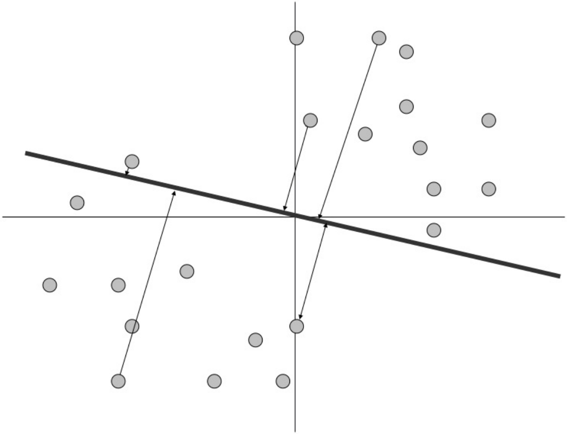
The algorithm returns a hyperplane whose margin is $\alpha\theta$ for at least $n(1-3\alpha)$ of the points (for any $\alpha \in [0,1]$) w.p. at least $1/147$.

FHP approximation algorithm



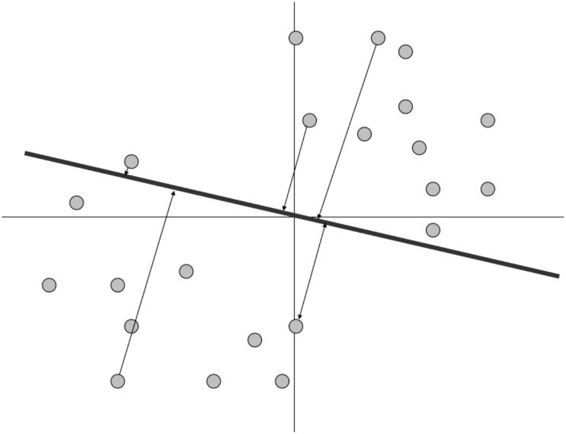
Maximize: $\max_{\|w\|^2=1} \min_i \langle w, x_i \rangle^2$

FHP approximation algorithm



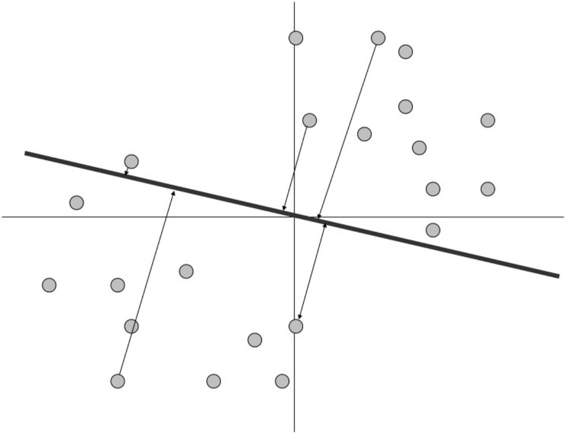
Maximize: $\max_{\|w\|^2=1} \mathbb{E}_i \langle w, x_i \rangle^2$

FHP approximation algorithm



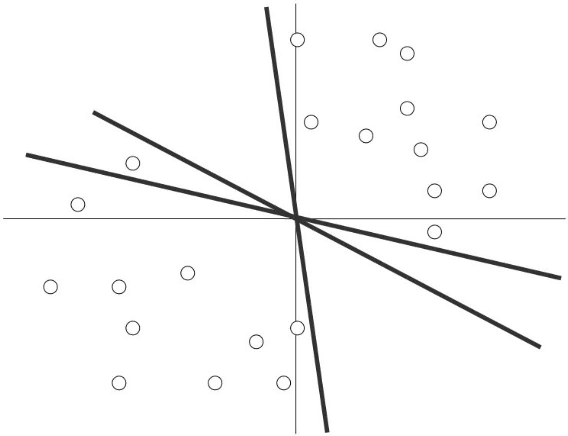
$$w_1 \leftarrow SVD([x_1, \ldots, x_n]) \quad \text{yields} \quad \mathbb{E}_i \langle w, x_i \rangle^2 \geq \theta^2$$

FHP approximation algorithm



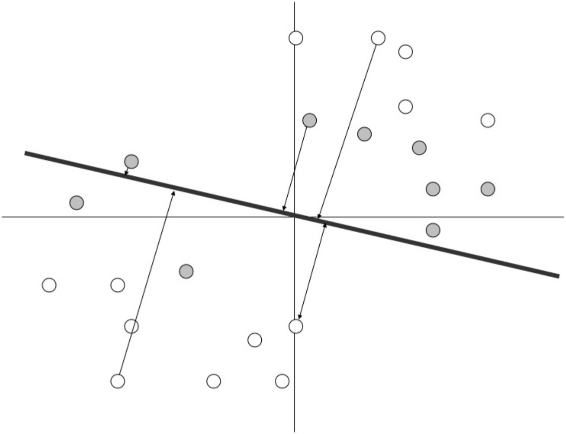
$$[\langle w_1, x_1\rangle^2, \ldots, \langle w_1, x_n\rangle^2] = [1, 1, \ldots 1, 1_{\theta^2 n}, 0, 0, 0, 0, \ldots, 0, 0, 0, 0]$$

FHP approximation algorithm



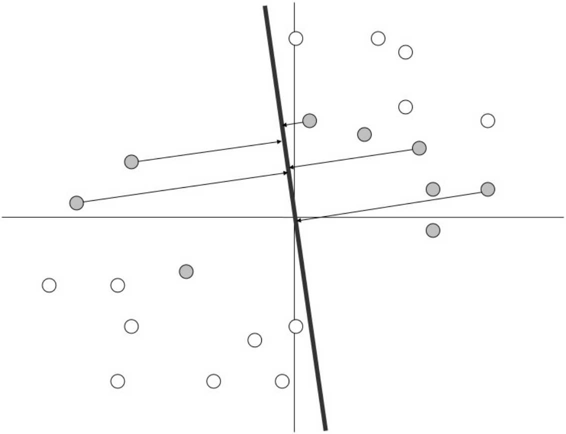
We need a set $\{w_1, \ldots, w_t\}$ such that $\quad \forall_i \; \mathbb{E}_j \; \langle w_j, x_i \rangle^2 \in \Omega(\theta^2)$

FHP approximation algorithm



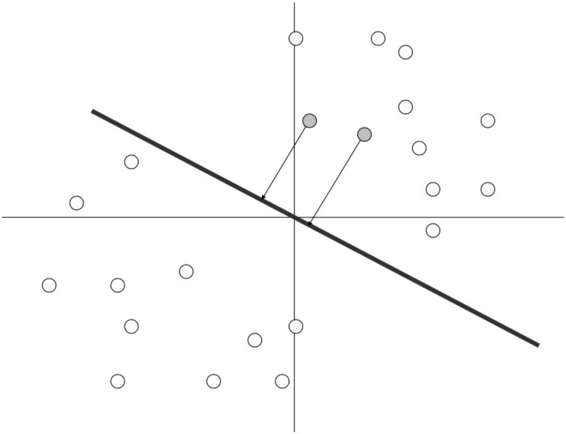
$$\tau_1(i) = 1 \qquad \tau_2(i) = \tau_1(i)c^{-\langle w_1, x_i\rangle^2}$$

FHP approximation algorithm



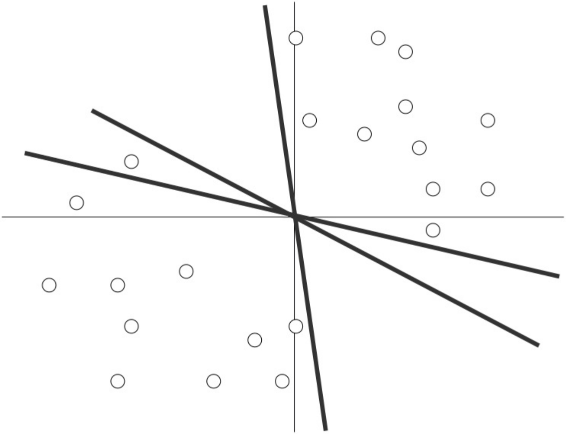
$$w_2 \leftarrow SVD([\sqrt{\tau_2(1)}x_1, \ldots, \sqrt{\tau_2(n)}x_n])$$

FHP approximation algorithm



$$w_t \leftarrow SVD([\sqrt{\tau_t(1)}x_1, \ldots, \sqrt{\tau_t(n)}x_n])$$

FHP approximation algorithm



The algorithm produce t hyperplanes $\{w_1, \ldots, w_t\}$ (one per itereation).

FHP approximation algorithm

Claim

The algorithm terminates after t iterations

$$t \leq 2\ln(n)/\left(\theta^2(1-1/c)\right).$$

Claim

When the algorithm terminates, for each i it holds

$$\sum_{j=1}^{t} \sigma_j^2(i) \geq \ln(n)/\ln(c).$$

Claim

Let $\{w_1, \ldots, w_t\}$ be the output of the above algorithm then:

$$\forall_i \ \mathbb{E}_j \left\langle w_j, x_i \right\rangle^2 \geq \theta^2/2.$$

FHP approximation algorithm

Claim

Let $\{w_1, \ldots, w_t\}$ *be the output of the above algorithm then:*

$$\forall_i \; \mathbb{E}_j \; \langle w_j, x_i \rangle^2 \geq \theta^2/2.$$

Claim

Let $w' = \sum_j g_j w_j$ *(*$g_j \sim \mathcal{N}(0,1)$ *independently) and* $w = w'/\|w'\|$ *then:*

$$|\langle w, x_i \rangle| \geq \alpha\theta$$

for at lease $n(1 - 3\alpha)$ *points with probability at least* $1/147$ *for any* $\alpha \in [0, 1]$.

this concludes the algorithm description.

Recap

- FHP is an important building block (not only in machine learning).
- There is an exact poly-time algorithm when the margin is constant.
- There is no PTAS in general.
- The random hyperplane algorithm is optimal unless 3SAT is solvable in $2^{o(n)}$ time.
- There is an efficient approximation algorithm (for most points...)

Future work and open questions

1. A connection to the multiplicative updates framework [7] (noticed by Elad Hazan) is being explored further.
2. Improve the naïve Gaussian combination of $w_1, \ldots, w_t$ (unclear if possible)
3. It seems that a more careful tweaking of the parameters will yield slightly better constants.
4. More general case, minimizing Hinge loss (in progress with Elad Hazan and Zohar Karnin).
5. Are there more efficient algorithm when the margin is large? (the random algorithm optimality only holds for small margins...)

Thanks for listening

Linli Xu, James Neufeld, Bryce Larson, and Dale Schuurmans.
Maximum margin clustering.
In *Advances in Neural Information Processing Systems 17*, pages 1537–1544. MIT Press, 2005.

Kristin P. Bennett and Ayhan Demiriz.
Semi-supervised support vector machines.
In *Advances in Neural Information Processing Systems*, pages 368–374. MIT Press, 1998.

Tijl De Bie and Nello Cristianini.
Convex methods for transduction.
In *Advances in Neural Information Processing Systems 16*, pages 73–80. MIT Press, 2003.

Thorsten Joachims.
Transductive inference for text classification using support vector machines.
In *Proceedings of the Sixteenth International Conference on Machine Learning*, ICML '99, pages 200–209, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc.

Rosa I. Arriaga and Santosh Vempala.
An algorithmic theory of learning: Robust concepts and random projection.
In *IEEE Symposium on Foundations of Computer Science*, pages 616–623, 1999.

Sanjeev Arora.
Probabilistic checking of proofs and hardness of approximation problems.
Revised version of a dissertation submitted at CS Division, U C Berkeley, CS-TR-476-94, August 1994.

Sanjeev Arora, Elad Hazan, and Satyen Kale.
The multiplicative weights update method: a meta algorithm and applications.
Technical report, 2005.