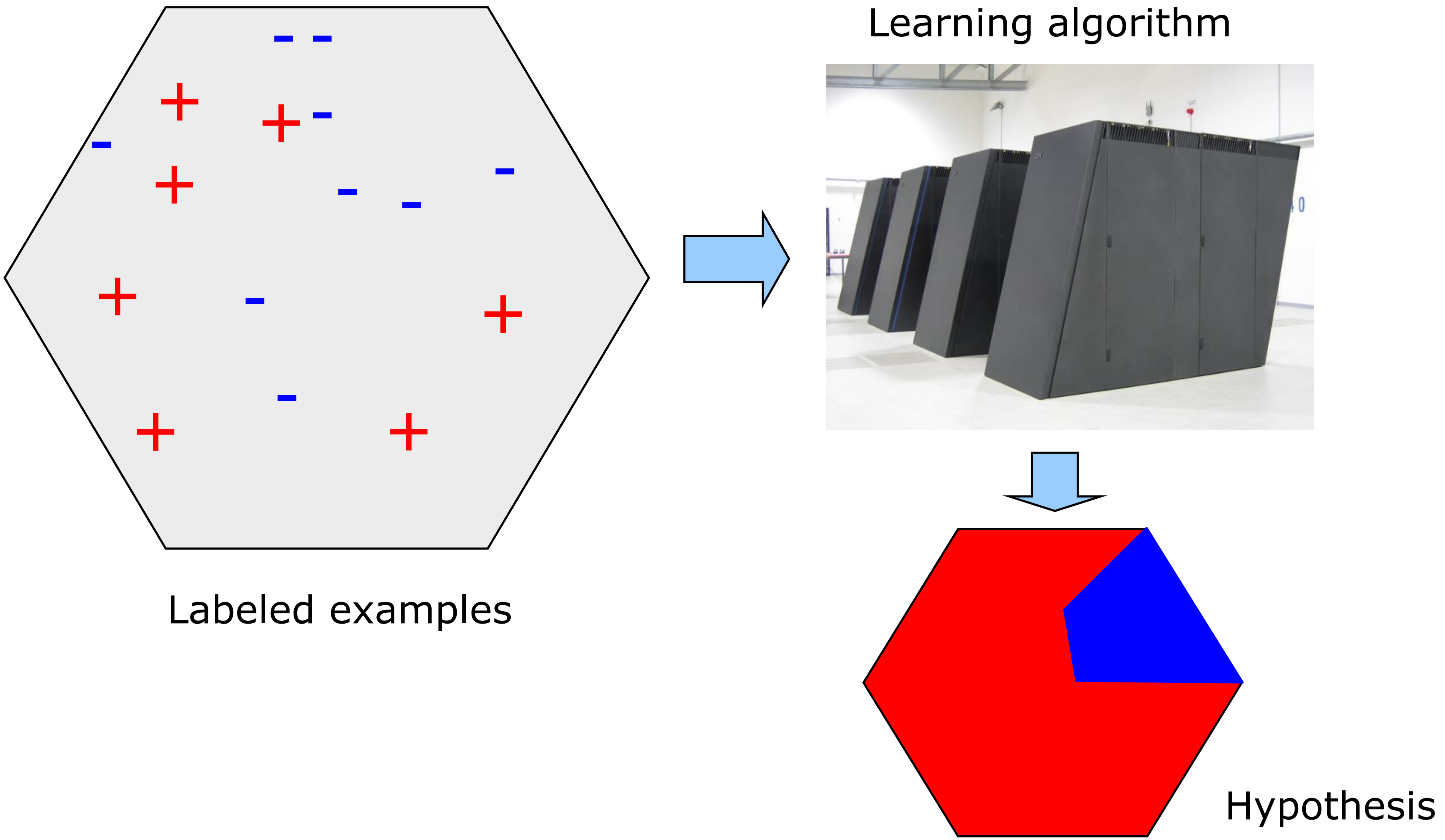
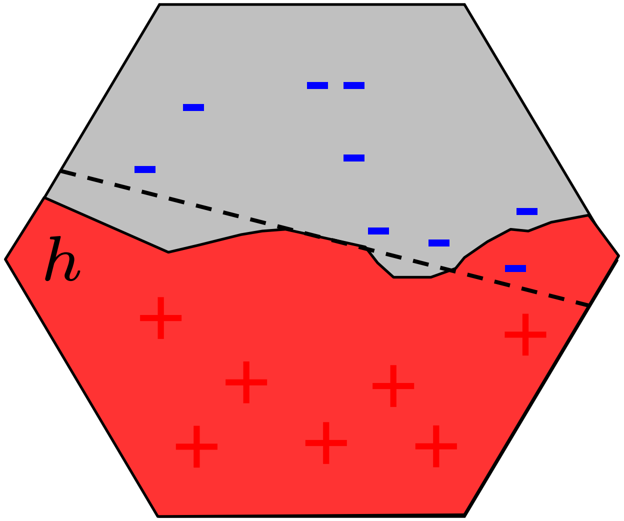
Distribution Specific Agnostic Boosting

Vitaly Feldman
CS Theory Group
IBM Almaden Research Center

Learning from examples



PAC learning [Valiant 84]



X domain

f: $X \to$ {-1,+1} unknown function

Random example: (x, $f(x)$)

$x \sim D$: unknown distribution over X

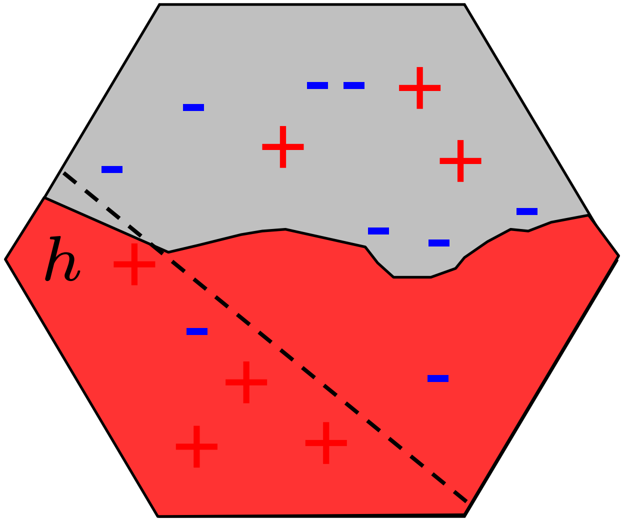
- PAC learning of a class of functions C:

$\forall D$, $f \in C$, and $\varepsilon > 0$, w.h.p. produce hypothesis h s.t. $\Pr_D[f(x) \neq h(x)] \leq \varepsilon$

Efficient: polynomial time in n (problem size) and $1/\varepsilon$

- *Distribution-specific* learning over D. D is fixed
- Some known learnable classes:
  - Boolean dis-/conjunctions over $\{0,1\}^n$ [Valiant 84]
  - Linear threshold functions (halfspaces) over $\mathbf{R}^n$ [BEHW 87]
  - Parity functions over $\{0,1\}^n$ [HSW 92]

Agnostic learning [Haussler; Kearns,Schapire,Sellie 92]



Example: (x, b)

$(x, b) \sim A$: unknown distribution over $X \times \{-1,1\}$

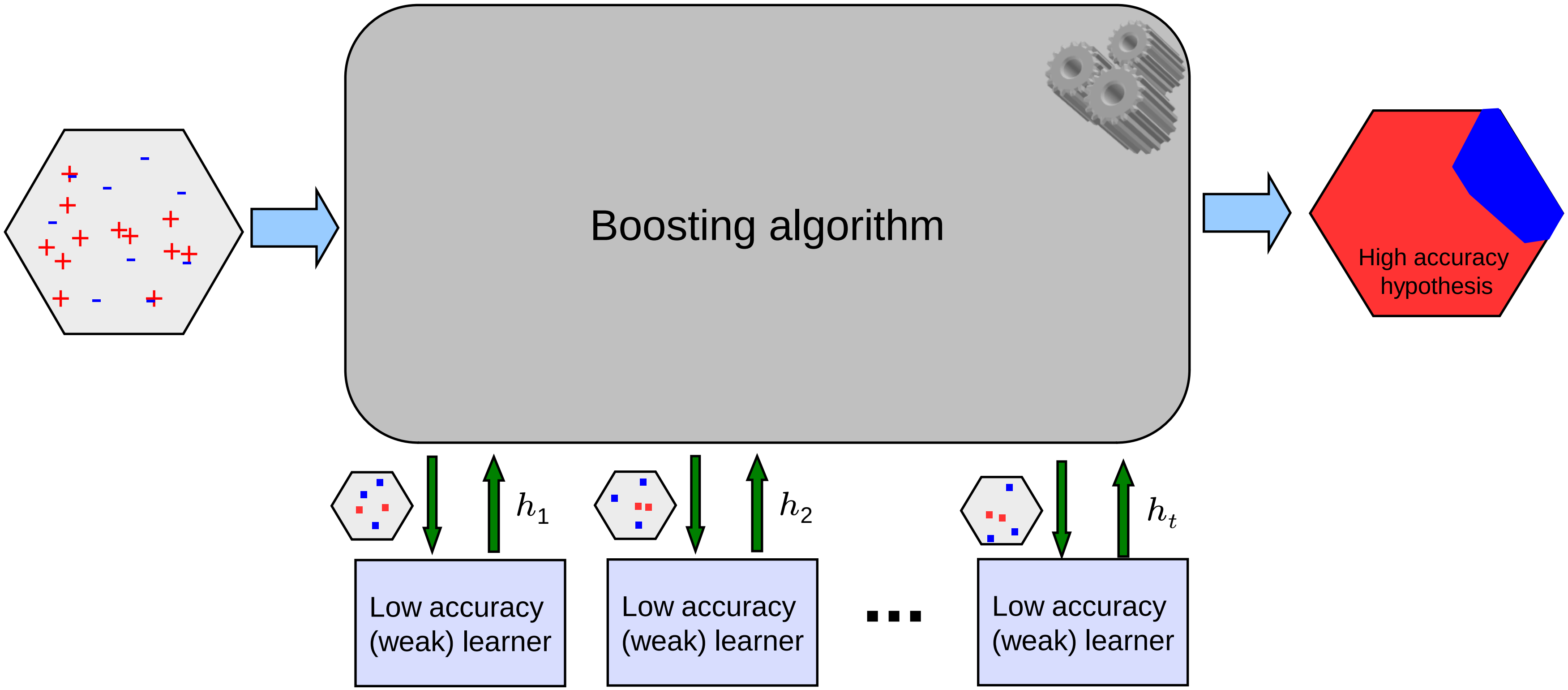
$\text{Opt}_A(C) = \min_{g \in C}\{\Pr_A[b \neq g(x)]\}$

- Agnostic learning of a class of functions C

$\forall A$, and $\varepsilon>0$, produce w.h.p. h such that $\Pr_A[b \neq h(x)] \leq \text{Opt}_A(C) + \varepsilon$

- *Distribution-specific* learning over D. Marginal of A on X equals to a fixed D
- Some known agnostically learnable function classes:
  - Uniform distribution over $\{0,1\}^n$:
    - Parities using queries [Goldreich,Levin 89]
    - Halfspaces [Kalai,Klivans,Mansour,Servedio 05]
    - Decision trees using queries [Gopalan,Kalai,Klivans 08]

Accuracy boosting



- *Weak* PAC learning [Kearns,Valiant 87]: $\Pr_D[f \neq h] \leq$ ½ - 1/poly(n)
- Weak PAC learning implies (strong) PAC learning [Schapire 90]
  - Only for distribution-independent learning!

Agnostic boosting [Ben-David,Long,Mansour 00]

- α-*weak* agnostic learning: output h s.t. $\Pr_A[b \neq h(x)] \leq$ ½-1/poly(n) whenever $\mathrm{Opt}_A(C) \leq$ ½ - α
- α-*weak* agnostic learning implies α-*optimal* agnostic learning

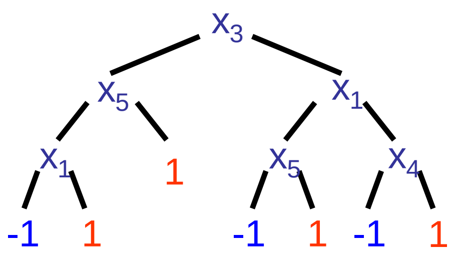
  [Kalai,Mansour,Verbin 08]
  - Outputs a hypothesis h s.t. $\Pr_A[b \neq h(x)] \leq \mathrm{Opt}_A(C) + \alpha + \varepsilon$
  - Distribution-independent
  - Based on boosting by branching programs [Mansour,McAllester 99]
  - Obtained the first non-trivial algorithm for agnostic learning of parities

Our results

- α-*weak* agnostic learning over D implies α-*optimal* agnostic learning over D
  - Simple and more efficient boosting algorithm

- Agnostic boosting algorithms from hardcore set constructions with the optimal set size parameter
  - Given a function f hard to δ-approximate construct a subset of the domain of weight 2δ where f is hard to weakly approximate
  - Hardcore set constructions [Impagliazzo 95] are closely related to boosting algorithms [Klivans,Servedio 99]
  - Known constructions: [Holenstein 05; Barak,Hardt,Kale 09]
  - Obtained new simple hardcore set construction

Results: applications



- Decision trees are agnostically learnable over U using queries
  - Known [Gopalan,Kalai,Klivans 08]
- Proof
  - For every distribution A uniform over X and a DT c of size s if $\Pr_A[b \neq c(x)] \leq ½ - \gamma$ then $\Pr_A[b \neq p(x)] \leq ½ - \gamma/s$ for some parity function $p(x)$ [Kushilevitz,Mansour 91]
  - Agnostic parity learning algorithm [Goldreich,Levin 89] gives weak agnostic learning
  - Boost

Applications to PAC learning

- MAJ(C,t) : majorities of at most t functions from C
- Agnostic learning of C implies PAC learning of MAJ(C,poly(n)) [KSS 92]
  - For every $f \in$ MAJ(C,t) and D, exists $c \in C$ s.t. $\Pr_D[f(x) \neq c(x)] \leq$ ½ - 1/(2t)
- Our result: agnostic learning of C over D implies PAC learning of MAJ(C,poly(n)) over D
- Corollary: DNF formulas are learnable over U using queries

  $$(x_2 \wedge x_3 \wedge \bar{x}_{10}) \bigvee (\bar{x}_2 \wedge x_5 \wedge x_6)$$

  - Known [Jackson 95]
- Proof
  - DNF $\subseteq$ MAJ(PARITY,poly(n)) [Jackson 95]

Some intuition

- Classical boosting: example $(x,b) \rightarrow (x,b)$ of weight $\gamma \in [0,1]$

- Here: example $(x,b) \rightarrow \begin{cases} (x,b) \text{ of weight } (1+\gamma)/2 \\ (x,-b) \text{ of weight } (1-\gamma)/2 \end{cases}$

- Total weight = 1. Error contribution γ

- General technique: gradient descent
  - Projection step
  - Balancing step

Conclusions and further work

- Agnostic boosting does not require modifying the marginal distribution over X
  - Useful in theoretical problems
- Agnostic boosting is natural
  - Several new algorithms
  - Avoids overfitting of some PAC boosters (e.g. Adaboost [Freund 95])
- Distribution-specific agnostic boosting and application to learning of decision trees also given by Kalai and Kanade
- Further directions:
  - More general understanding of agnostic boosting
  - More efficient agnostic boosting
  - Can PAC boosters be converted to agnostic ones automatically?
  - Behavior in practice