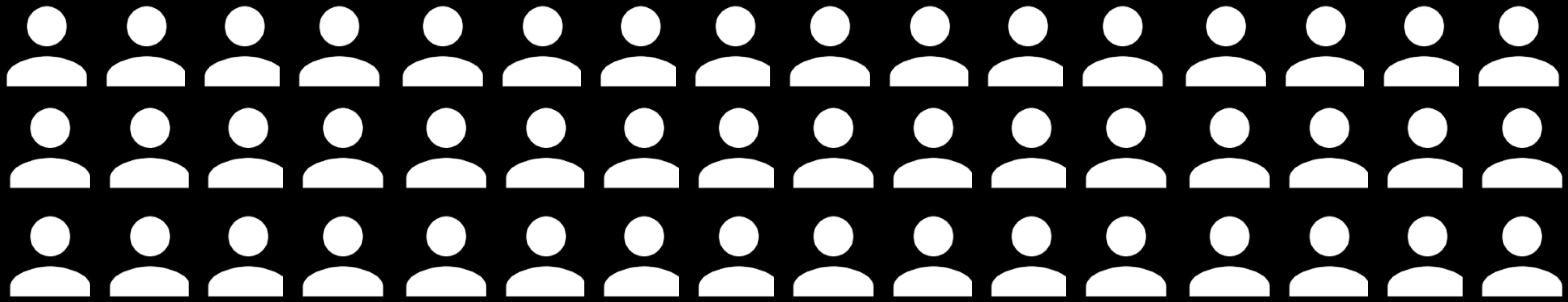
On the Complexity of Learning from Label Proportions

Lev Reyzin
UIC, Math Dept.
MMLS 2017

Reference

Benjamin Fish, Lev Reyzin. *On the Complexity of Learning from Label Proportions.* IJCAI 2017

A Motivating Example: Elections



A Motivating Example: Elections

$\hat{p}$

$1-\hat{p}$

Elections

- We know who voted.
- We know the result.
- But we don't know who voted for whom!
- We want to train a model that predicts future elections.

Supervised Learning (PAC)

A class of functions H is **PAC learnable** if there is an efficient algorithm A such that for every target function c in H, any distribution D over $\{0, 1\}^n$, and for any ε, δ > 0, given

$$m \geq \text{poly}(1/\varepsilon, 1/\delta, n, \text{size}(c))$$

labeled examples drawn i.i.d. from D, returns a hypothesis h in H such that

$$P[1_{c(x)\neq h(x)} \leq \varepsilon] \geq 1 - \delta.$$

Learning from Proportions (LLP)

A class of functions H is **PAC learnable from label proportions** if there is an efficient algorithm A such that for every target function c in H, any distribution D over $\{0, 1\}^n$, and for any $\varepsilon, \delta > 0$, given

$$m \geq \text{poly}(1/\varepsilon, 1/\delta, n, \text{size}(c))$$

examples drawn i.i.d. from D and $\hat{p}$, returns a hypothesis h in H such that

$$P[|p(c) - p(h)| \leq \varepsilon] \geq 1 - \delta.$$

Label Proportions vs PAC

A class of functions H is **PAC learnable** if there is an efficient algorithm A such that for every target function c in H, any distribution D over $\{0, 1\}^n$, and for any ε, δ > 0, given m ≥ poly(1/ ε , 1/δ, n, size(c)) examples drawn i.i.d. from D and their labels, returns a hypothesis h in H such that $P[1_{c(x)\neq h(x)} \leq \varepsilon] \geq 1 - \delta$.

A class of functions H is **PAC learnable from label proportions** if there is an efficient algorithm A such that for every target function c in H, any distribution D over $\{0, 1\}^n$, and for any ε, δ > 0, given m ≥ poly(1/ ε , 1/δ, n, size(c)) examples drawn i.i.d. from D and $\hat{p}$, returns a hypothesis h in H such that $P[|p(c) - p(h)| \leq \varepsilon] \geq 1 - \delta$.

Main Question

What is the complexity of Learning from Label Proportions? And how does LLP learning relate to normal PAC learning?

An Occam's Razor Bound for LLP

For target function c, with probability at least $1 - \delta$, for all $h \in H$,

$|p_c - p_h| \leq |\hat{p}_c - \hat{p}_h| + \tilde{O}(\ 1/\delta\ (VC(H)/m)^{1/2})$

Results

- LLP $\subsetneq$ PAC [1]
- VC-dimension hardness for LLP[1] (not a complete characterization)
- A nontrivial problem in LLP

[1] under standard complexity-theoretic assumptions

Results

- LLP $\subsetneq$ PAC [1]

- VC-dimension hardness for LLP[1]
  (not a complete characterization)

- A nontrivial problem in LLP

[1] under standard complexity-theoretic assumptions

LLP vs PAC

Theorem: Suppose NP $\neq$ RP. Then if a hypothesis class H is efficiently learnable from label proportions, it is also efficiently (properly) PAC learnable.

LLP vs PAC

Theorem: Suppose NP ≠ RP. Then if a hypothesis class H is efficiently learnable from label proportions, it is also efficiently (properly) PAC learnable.

proof. involves a reduction from PAC to LLP. Idea is to make LLP distribution that forces all (and only the) + examples to be labeled + by LLP learner if its threshold is to be met.

LLP vs PAC

Theorem: Suppose NP ≠ RP. Then if a hypothesis class H is efficiently learnable from label proportions, it is also efficiently (properly) PAC learnable.

$$D'(x) = \left\{ \begin{array}{cc} \frac{m}{km+m-k} & \text{if } x \in S \text{ and } c(x) = 1 \\ \frac{1}{km+m-k} & \text{if } x \in S \text{ and } c(x) = 0 \\ 0 & \text{otherwise} \end{array} \right\}$$

where k = number of positive examples and $\varepsilon = 1/(2m^2)$

Results

- LLP $\subsetneq$ PAC [1]
- VC-dimension hardness for LLP[1]
  (not a complete characterization)
- A nontrivial problem in LLP

[1] under standard complexity-theoretic assumptions

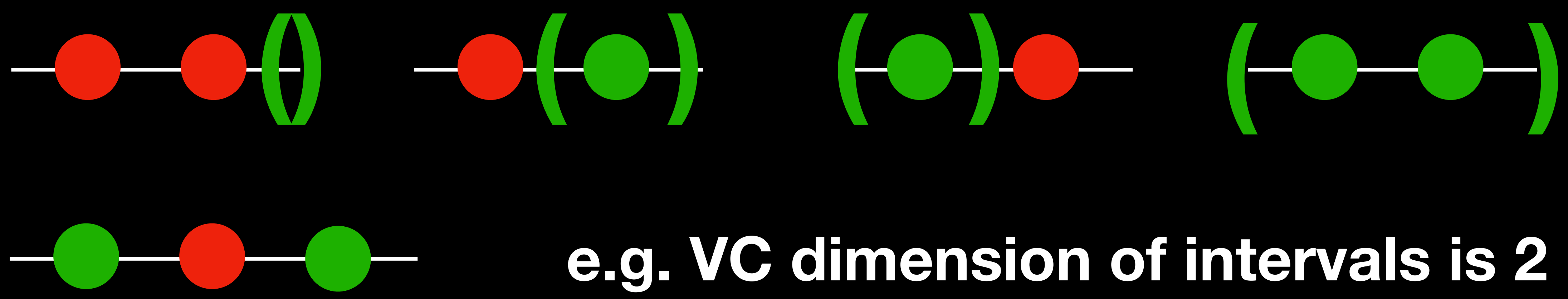
VC-dim for LLP

Theorem: Let C be a hypothesis class such that $VC(C) \geq n^{\gamma}$ for some constant $\gamma > 0$. There is no efficient algorithm for PAC learning C from label proportions unless NP = RP.

VC-dim for LLP

Theorem: Let C be a hypothesis class s.t. $VC(C) \geq n^{\gamma}$ for some constant $\gamma > 0$. There is no efficient algorithm for PAC learning C from label proportions unless NP = RP.

reminder: the **VC-dim** is the maximum number of points a class of functions can **shatter**.



e.g. VC dimension of intervals is 2

VC-dim for LLP

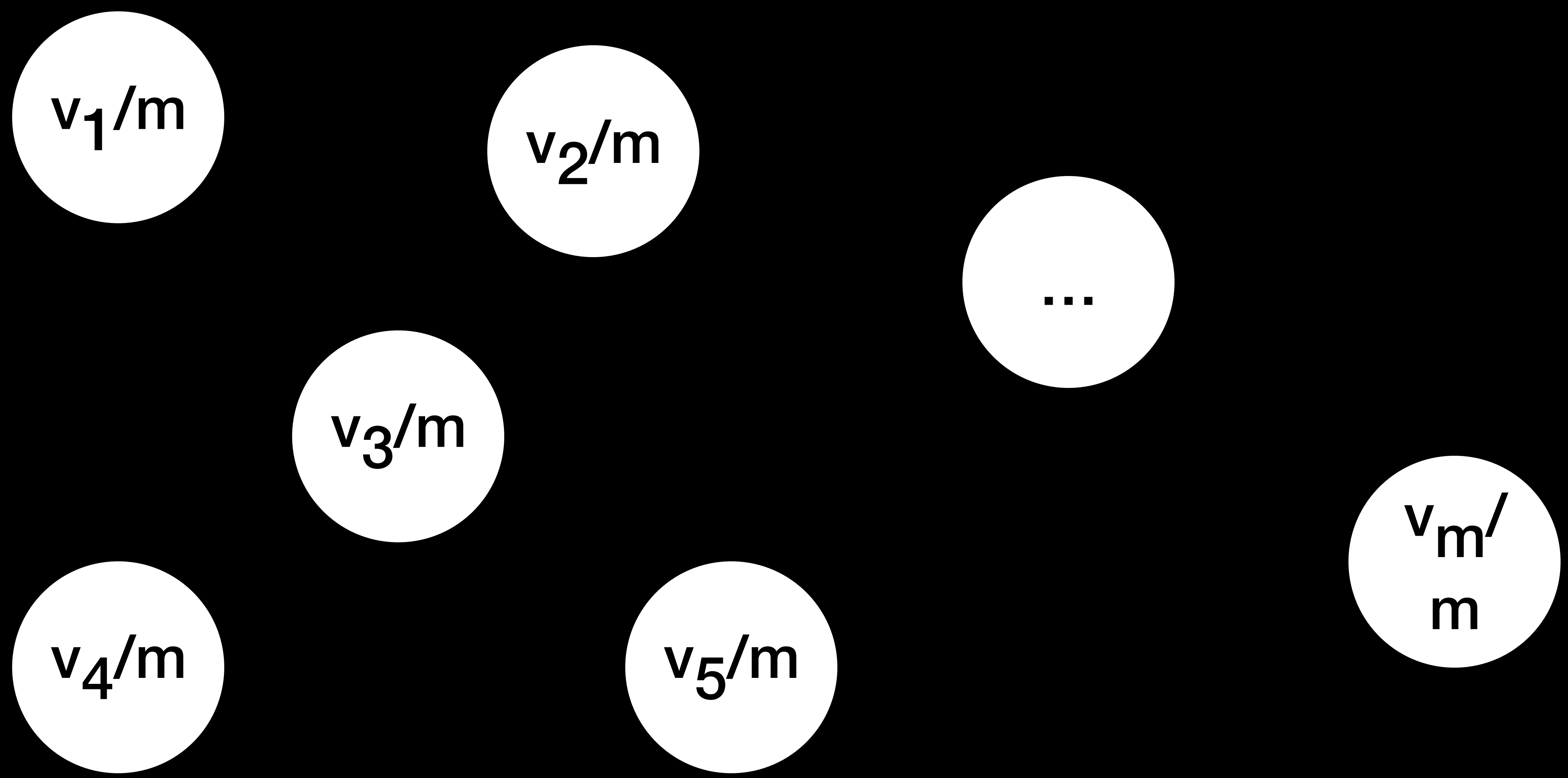
Theorem: Let C be a hypothesis class s.t. $VC(C) \geq n^{\gamma}$ for some constant $\gamma > 0$. There is no efficient algorithm for PAC learning C from label proportions unless NP = RP.

proof idea: can reduce from subset sum to LLP learning any class with large VC dimension.

How? choose a set of shattered points, make LLP learner solve subset sum to get the "right" threshold.

VC-dim for LLP

Theorem: Let C be a hypothesis class s.t. $VC(C) \geq n^{\gamma}$ for some constant $\gamma > 0$. There is no efficient algorithm for PAC learning C from label proportions unless NP = RP.



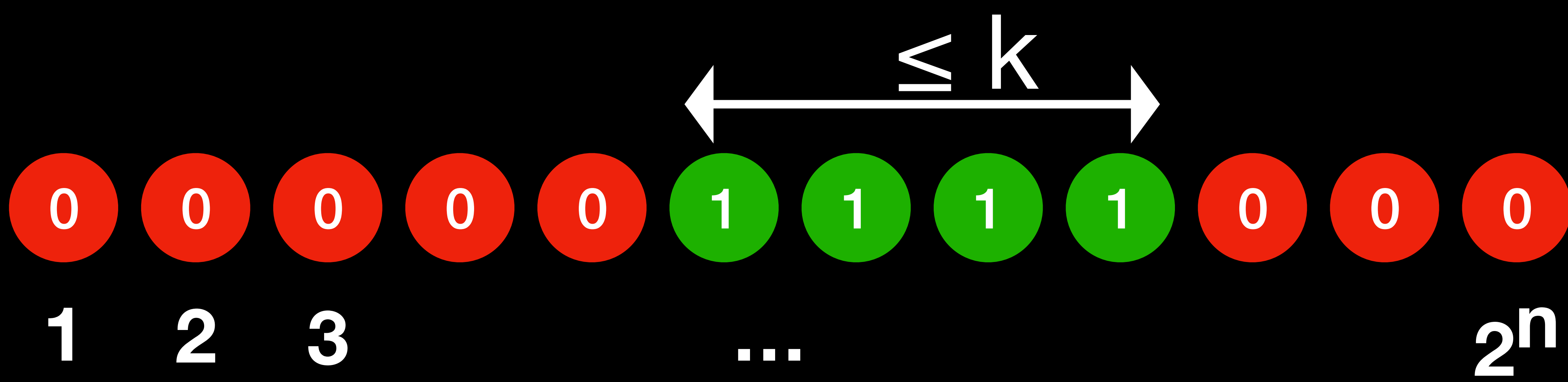
Results

- LLP $\subsetneq$ PAC [1]
- VC-dimension hardness for LLP[1] (not a complete characterization)
- A nontrivial problem in LLP

[1] under standard complexity-theoretic assumptions

An LLP-Learnable Class

$$\{h : \{1, . . . , 2^n\} \to \{0, 1\} : \max_{h(i)=h(j)=1} |i - j| \leq k\}$$

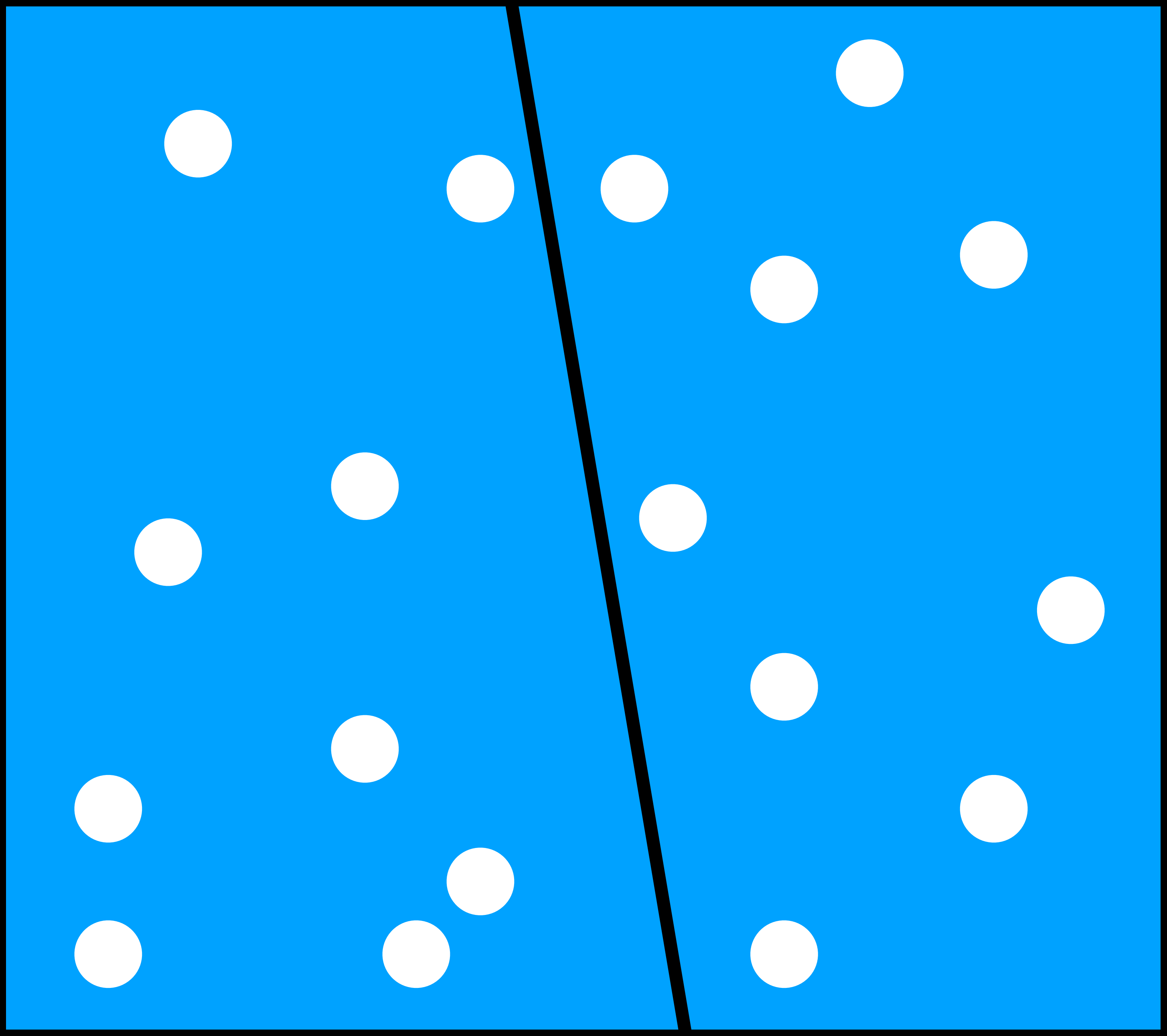


An LLP-Learnable Class

$\{h : \{1, \ldots, 2^n\} \to \{0, 1\} : \max_{h(i)=h(j)=1} |i - j| \leq k\}$

For k=log(n), VC-dimension is super-constant, yet there is a poly-time algorithm.

LLP is Also Easy for Nice Distributions



Conclusions

- I presented the beginnings of a learning theory for LLP.
- LLP is the simplest case of multi-bag learning.
- Extensions include to multiclass learning and regression.
- Would also be interesting to develop practical LLP algorithms.