A Computational Teaching Theory for Bayesian Learners

Xiaojin Zhu

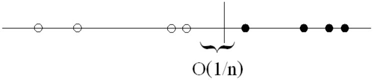
Department of Computer Sciences
University of Wisconsin-Madison

NIPS 2012 Workshop
Personalizing Education With Machine Learning

Teaching needs a different theory

Learning a threshold classifier in 1D

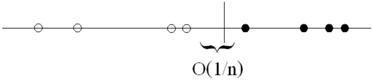
- passive learning $(x_i, y_i) \overset{iid}{\sim} p$, risk $\approx O(\frac{1}{n})$



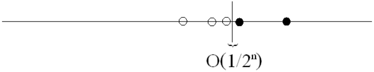
Teaching needs a different theory

Learning a threshold classifier in 1D

- passive learning $(x_i, y_i) \overset{iid}{\sim} p$, risk $\approx O(\frac{1}{n})$



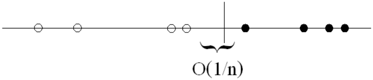
- active learning risk $\approx \frac{1}{2^n}$



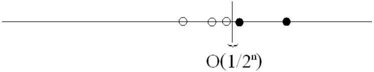
Teaching needs a different theory

Learning a threshold classifier in 1D

- passive learning $(x_i, y_i) \overset{iid}{\sim} p$, risk $\approx O(\frac{1}{n})$

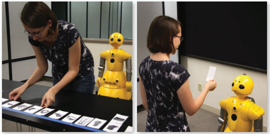


- active learning risk $\approx \frac{1}{2^n}$

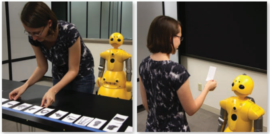


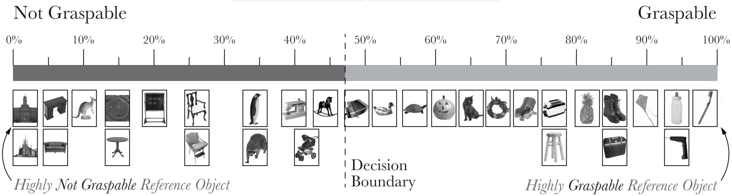
- taught: $n = 2$. Teaching dimension [Goldman and Kearns 1995]

Teaching dimension $\neq$ curriculum learning [Bengio et al. 2009]?

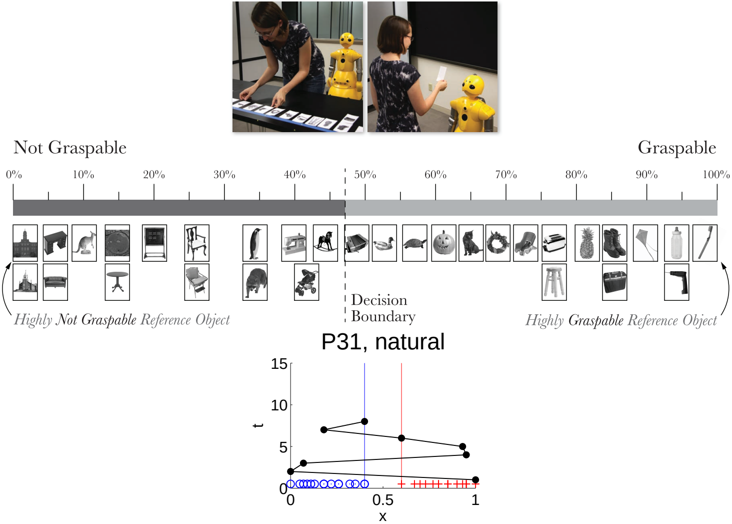


Teaching dimension $\neq$ curriculum learning [Bengio et al. 2009]?



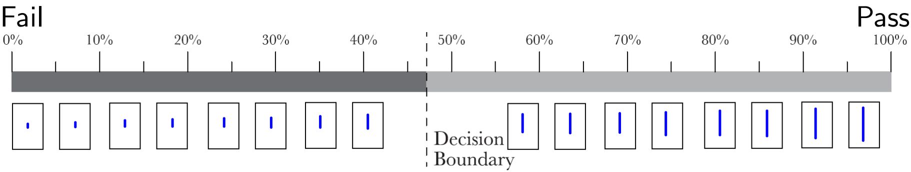


Teaching dimension ≠ curriculum learning [Bengio et al. 2009]?



No human teachers started at the boundary [Khan et al. NIPS11]

More to the story



The master card

56% human teachers started at the boundary.

A refined framework for teaching

Three actors:

- World: $p(x, y \mid \theta^*)$, loss function $\ell(f(x), y)$

A refined framework for teaching

Three actors:

- World: $p(x, y \mid \theta^*)$, loss function $\ell(f(x), y)$
- Learner: Bayesian.

A refined framework for teaching

Three actors:

- World: $p(x, y \mid \theta^*)$, loss function $\ell(f(x), y)$
- Learner: Bayesian.
  - prior over Θ ($\theta^* \in \Theta$), likelihood $p(x, y \mid \theta)$

A refined framework for teaching

Three actors:

- World: $p(x, y \mid \theta^*)$, loss function $\ell(f(x), y)$
- Learner: Bayesian.
  - prior over Θ ($\theta^* \in \Theta$), likelihood $p(x, y \mid \theta)$
  - maintains posterior $p(\theta \mid \mathrm{data})$ by Bayesian update

A refined framework for teaching

Three actors:

- World: $p(x, y \mid \theta^*)$, loss function $\ell(f(x), y)$
- Learner: Bayesian.
  - prior over Θ ($\theta^* \in \Theta$), likelihood $p(x, y \mid \theta)$
  - maintains posterior $p(\theta \mid \mathrm{data})$ by Bayesian update
  - makes prediction $f(x \mid \mathrm{data})$ using the posterior

A refined framework for teaching

Three actors:

- World: $p(x, y \mid \theta^*)$, loss function $\ell(f(x), y)$
- Learner: Bayesian.
  - prior over Θ ($\theta^* \in \Theta$), likelihood $p(x, y \mid \theta)$
  - maintains posterior $p(\theta \mid \mathrm{data})$ by Bayesian update
  - makes prediction $f(x \mid \mathrm{data})$ using the posterior
- Teacher:

A refined framework for teaching

Three actors:

- World: $p(x, y \mid \theta^*)$, loss function $\ell(f(x), y)$
- Learner: Bayesian.
  - prior over Θ ($\theta^* \in \Theta$), likelihood $p(x, y \mid \theta)$
  - maintains posterior $p(\theta \mid \text{data})$ by Bayesian update
  - makes prediction $f(x \mid \text{data})$ using the posterior
- Teacher:
  - clairvoyant, knows everything above

A refined framework for teaching

Three actors:

- World: $p(x, y \mid \theta^*)$, loss function $\ell(f(x), y)$
- Learner: Bayesian.
  - prior over Θ ($\theta^* \in \Theta$), likelihood $p(x, y \mid \theta)$
  - maintains posterior $p(\theta \mid \text{data})$ by Bayesian update
  - makes prediction $f(x \mid \text{data})$ using the posterior
- Teacher:
  - clairvoyant, knows everything above
  - can teach only by giving (x, y) to the learner

A refined framework for teaching

Three actors:

- World: $p(x, y \mid \theta^*)$, loss function $\ell(f(x), y)$
- Learner: Bayesian.
  - prior over Θ ($\theta^* \in \Theta$), likelihood $p(x, y \mid \theta)$
  - maintains posterior $p(\theta \mid \mathrm{data})$ by Bayesian update
  - makes prediction $f(x \mid \mathrm{data})$ using the posterior
- Teacher:
  - clairvoyant, knows everything above
  - can teach only by giving (x, y) to the learner
  - goal: choose the smallest teaching set $D = (x, y)_{1:n}$ to minimize the learner's future loss

$$\mathbb{E}_{\theta^*}[\ell(f(x \mid D), y)]$$

A refined framework for teaching

Three actors:

- World: $p(x, y \mid \theta^*)$, loss function $\ell(f(x), y)$
- Learner: Bayesian.
  - prior over Θ ($\theta^* \in \Theta$), likelihood $p(x, y \mid \theta)$
  - maintains posterior $p(\theta \mid \mathrm{data})$ by Bayesian update
  - makes prediction $f(x \mid \mathrm{data})$ using the posterior
- Teacher:
  - clairvoyant, knows everything above
  - can teach only by giving (x, y) to the learner
  - goal: choose the smallest teaching set $D = (x, y)_{1:n}$ to minimize the learner's future loss

$$\mathbb{E}_{\theta^*}[\ell(f(x \mid D), y)]$$

  - if the future loss approaches Bayes risk, D is a teaching set and n is the (generalized) teaching dimension

A refined framework for teaching

Three actors:

- World: $p(x, y \mid \theta^*)$, loss function $\ell(f(x), y)$
- Learner: Bayesian.
  - prior over Θ ($\theta^* \in \Theta$), likelihood $p(x, y \mid \theta)$
  - maintains posterior $p(\theta \mid \text{data})$ by Bayesian update
  - makes prediction $f(x \mid \text{data})$ using the posterior
- Teacher:
  - clairvoyant, knows everything above
  - can teach only by giving (x, y) to the learner
  - goal: choose the smallest teaching set $D = (x, y)_{1:n}$ to minimize the learner's future loss
  $$\mathbb{E}_{\theta^*}[\ell(f(x \mid D), y)]$$
  - if the future loss approaches Bayes risk, D is a teaching set and n is the (generalized) teaching dimension
  - may have computational limitations

Case study on graspability and lines

- Unify "curriculum learning" and "teaching at the boundary" both as greedy (learner) risk minimization

Case study on graspability and lines

- Unify "curriculum learning" and "teaching at the boundary" both as greedy (learner) risk minimization
- Key difference: dimension of X

Case study on graspability and lines

- Unify "curriculum learning" and "teaching at the boundary" both as greedy (learner) risk minimization
- Key difference: dimension of X
  - graspability: d large.

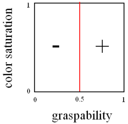
    e.g., squirrel = Boolean vector (graspable, shy, store supplies for the winter, is not poisonous, has four paws, has teeth, has two ears, has two eyes, is beautiful, is brown, lives in trees, rodent, doesn't herd, doesn't sting, drinks water, eats nuts, feels soft, fluffy, gnaws on everything, has a beautiful tail, has a large tail, has a mouth, has a small head, has gnawing teeth, has pointy ears, has short paws, is afraid of people, is cute, is difficult to catch, is found in Belgium, is light, is not a pet, is not very big, is short haired, is sweet , jumps, lives in Europe, lives in the wild, short front legs, small ears, smaller than a horse, soft fur, timid animal, can't fly, climbs in trees, collects nuts, crawls up trees, eats acorns, eats plants, does not lay eggs ...)

Case study on graspability and lines

- Unify "curriculum learning" and "teaching at the boundary" both as greedy (learner) risk minimization
- Key difference: dimension of X
  - graspability: d large.

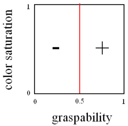
    e.g., squirrel = Boolean vector (graspable, shy, store supplies for the winter, is not poisonous, has four paws, has teeth, has two ears, has two eyes, is beautiful, is brown, lives in trees, rodent, doesn't herd, doesn't sting, drinks water, eats nuts, feels soft, fluffy, gnaws on everything, has a beautiful tail, has a large tail, has a mouth, has a small head, has gnawing teeth, has pointy ears, has short paws, is afraid of people, is cute, is difficult to catch, is found in Belgium, is light, is not a pet, is not very big, is short haired, is sweet , jumps, lives in Europe, lives in the wild, short front legs, small ears, smaller than a horse, soft fur, timid animal, can't fly, climbs in trees, collects nuts, crawls up trees, eats acorns, eats plants, does not lay eggs ...)
  - lines: $d = 1$.

Problem setting



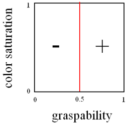
- World: $p(x) = \mathrm{Unif}[0,1]^d, p_{y=1|x} = 1_{(x_1 \geq \frac{1}{2})}$, 0-1 loss

Problem setting

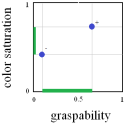


- World: $p(x) = \mathrm{Unif}[0,1]^d, p_{y=1|x} = 1_{\left(x_1 \geq \frac{1}{2}\right)}$, 0-1 loss
- Learner:

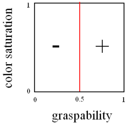
Problem setting



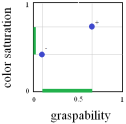
- World: $p(x) = \mathrm{Unif}[0,1]^d, p_{y=1|x} = 1_{\left(x_1 \geq \frac{1}{2}\right)}$, 0-1 loss
- Learner:
  - axis-parallel version space V



Problem setting

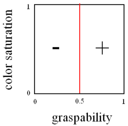


- World: $p(x) = \mathrm{Unif}[0,1]^d, p_{y=1|x} = 1_{(x_1 \geq \frac{1}{2})}$, 0-1 loss
- Learner:

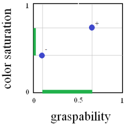


  - axis-parallel version space V
  - Gibbs classifier $f(x) \equiv \hat{y} \sim p(y \mid x, D)$

Problem setting

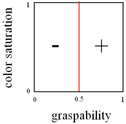


- World: $p(x) = \mathrm{Unif}[0,1]^d, p_{y=1|x} = 1_{\left(x_1 \geq \frac{1}{2}\right)}$, 0-1 loss
- Learner:

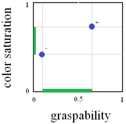


  - axis-parallel version space V
  - Gibbs classifier $f(x) \equiv \hat{y} \sim p(y \mid x, D)$
- Teacher:

Problem setting

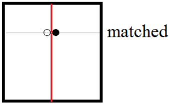


- World: $p(x) = \mathrm{Unif}[0,1]^d, p_{y=1|x} = 1_{(x_1 \geq \frac{1}{2})}$, 0-1 loss
- Learner:

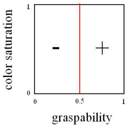


  - axis-parallel version space V
  - Gibbs classifier $f(x) \equiv \hat{y} \sim p(y \mid x, D)$
- Teacher:
  - ideally match irrelevant dimensions $\Rightarrow n = 2$, doesn't match human

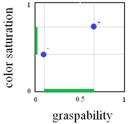
teacher behaviors



Problem setting

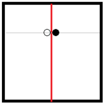


- World: $p(x) = \mathrm{Unif}[0,1]^d, p_{y=1|x} = 1_{(x_1 \geq \frac{1}{2})}$, 0-1 loss
- Learner:



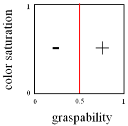
  - axis-parallel version space V
  - Gibbs classifier $f(x) \equiv \hat{y} \sim p(y \mid x, D)$
- Teacher:
  - ideally match irrelevant dimensions $\Rightarrow n = 2$, doesn't match human

teacher behaviors

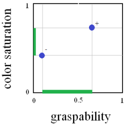


  - let's limit the teacher's power:

Problem setting

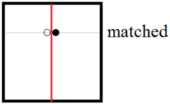


- World: $p(x) = \text{Unif}[0,1]^d, p_{y=1|x} = 1_{(x_1 \geq \frac{1}{2})}$, 0-1 loss
- Learner:



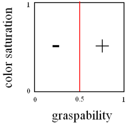
  - axis-parallel version space V
  - Gibbs classifier $f(x) \equiv \hat{y} \sim p(y \mid x, D)$
- Teacher:
  - ideally match irrelevant dimensions $\Rightarrow n = 2$, doesn't match human

    teacher behaviors

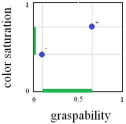


  - let's limit the teacher's power:
    - ★ pool-based teaching $\mathbf{x}_1, \ldots, \mathbf{x}_n \sim \text{unif}[0,1]^d$

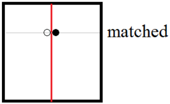
Problem setting



- World: $p(x) = \mathrm{Unif}[0,1]^d, p_{y=1|x} = 1_{(x_1 \geq \frac{1}{2})}$, 0-1 loss
- Learner:



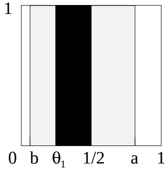
  - axis-parallel version space V
  - Gibbs classifier $f(x) \equiv \hat{y} \sim p(y \mid x, D)$
- Teacher:
  - ideally match irrelevant dimensions $\Rightarrow n = 2$, doesn't match human teacher behaviors



  - let's limit the teacher's power:
    - ★ pool-based teaching $\mathbf{x}_1, \ldots, \mathbf{x}_n \sim \mathrm{unif}[0,1]^d$
    - ★ only pays attention to the target dimension

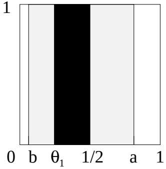
After the first two teaching items

- if hypothesis $f = x_{\cdot 1} \geq \theta_1$ selected from dim 1, error$=|\theta_1 - \frac{1}{2}|$

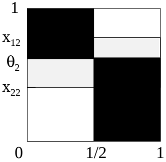


After the first two teaching items

- if hypothesis $f = x_{\cdot 1} \geq \theta_1$ selected from dim 1, error$=|\theta_1 - \frac{1}{2}|$

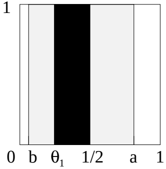


- if from dim 2, error$=\frac{1}{2}$

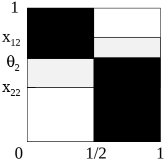


After the first two teaching items

- if hypothesis $f = x_{\cdot 1} \geq \theta_1$ selected from dim 1, error$=|\theta_1 - \frac{1}{2}|$



- if from dim 2, error$=\frac{1}{2}$

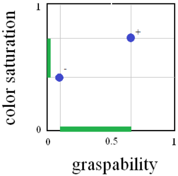


- The learner's risk

$$R = \frac{1}{|V|}\left(\int_b^a |\theta_1 - \frac{1}{2}|d\theta_1 + \sum_{k=2}^{d} \int_{\min(x_{1k},x_{2k})}^{\max(x_{1k},x_{2k})} \frac{1}{2} d\theta_k\right)$$

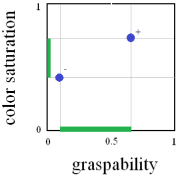
Risk minimization

- The teacher chooses two items with dim1$= a, b$ to minimize R. (The computational limitation assumption)



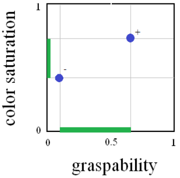
Risk minimization

- The teacher chooses two items with dim1$= a, b$ to minimize R. (The computational limitation assumption)
- Trade off:



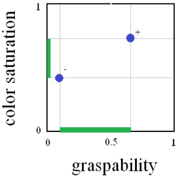
Risk minimization

- The teacher chooses two items with dim1$= a, b$ to minimize R. (The computational limitation assumption)
- Trade off:
  - $b - a$ too small: learner frequently picks f in irrelevant dimensions $\Rightarrow$ large error



Risk minimization

- The teacher chooses two items with dim1$= a, b$ to minimize R. (The computational limitation assumption)
- Trade off:
  - $b - a$ too small: learner frequently picks f in irrelevant dimensions $\Rightarrow$ large error
  - $b - a$ too large: learner picks very wrong f in the relevant dimension $\Rightarrow$ large error



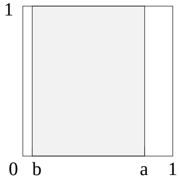
Risk minimization

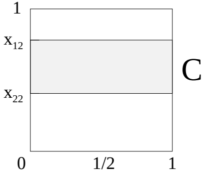
Theorem

The risk R is minimized by

$$\begin{aligned} a^* &= \frac{\sqrt{c^2+2c}-c+1}{2} \\ b^* &= 1-a^* \end{aligned}$$

where $c \equiv \sum_{k=2}^{d} |x_{1k} - x_{2k}|$ is the version subspace size in irrelevant dimensions.





d decides where to start teaching

- $|x_{1k} - x_{2k}| \sim \mathrm{Beta}(1, 2)$ for $k = 2, \dots, d$ (order statistics)

d decides where to start teaching

- $|x_{1k} - x_{2k}| \sim \mathrm{Beta}(1, 2)$ for $k = 2, \ldots, d$ (order statistics)
- $c \equiv \sum_{k=2}^{d} |x_{1k} - x_{2k}|$ is the sum of $d - 1$ $\mathrm{Beta}(1, 2)$ random variables.

d decides where to start teaching

- $|x_{1k} - x_{2k}| \sim \mathrm{Beta}(1,2)$ for $k = 2, \ldots, d$ (order statistics)
- $c \equiv \sum_{k=2}^{d} |x_{1k} - x_{2k}|$ is the sum of $d-1$ $\mathrm{Beta}(1,2)$ random variables.

Corollary

When $d \to \infty$, the minimizer of R is $a^* = 1, b^* = 0$. (curriculum)
When $d = 1$, the minimizer of R is $a^* \to \frac{1}{2}_-, b^* \to \frac{1}{2}_+$. (boundary)

d decides where to start teaching

- $|x_{1k} - x_{2k}| \sim \mathrm{Beta}(1,2)$ for $k = 2, \ldots, d$ (order statistics)
- $c \equiv \sum_{k=2}^{d} |x_{1k} - x_{2k}|$ is the sum of $d-1$ $\mathrm{Beta}(1,2)$ random variables.

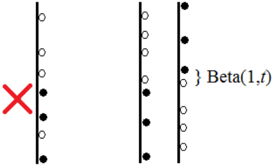
Corollary

When $d \to \infty$, the minimizer of R is $a^* = 1, b^* = 0$. (curriculum)
When $d = 1$, the minimizer of R is $a^* \to \frac{1}{2}_-, b^* \to \frac{1}{2}_+$. (boundary)

- Matches graspability and lines

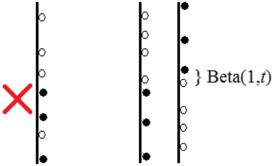
d also decides convergence toward boundary

- Version subspace V_k survives t teaching items if the items are linearly separable in dimension k



d also decides convergence toward boundary

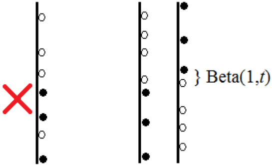
- Version subspace V_k survives t teaching items if the items are linearly separable in dimension k



- This happens with probability $\frac{2}{\binom{t}{t_0}}$ where t_0 is the number of positive items

d also decides convergence toward boundary

- Version subspace V_k survives t teaching items if the items are linearly separable in dimension k



- This happens with probability $\frac{2}{\binom{t}{t_0}}$ where t_0 is the number of positive items
- If V_k does survive, its size $\sim \mathrm{Beta}(1, t)$ (order statistics)

Teaching items should approach decision boundary

Theorem

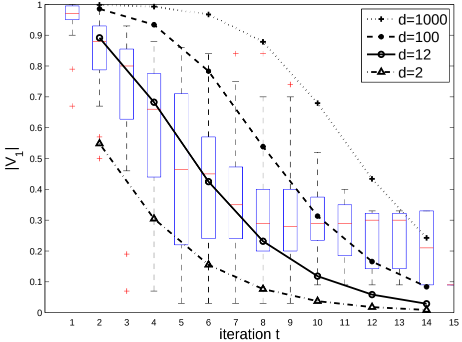
Let the teaching sequence contain t_0 negative labels and $t - t_0$ positive ones. Then the version space in dim k has size $|V_k| = \alpha_k \beta_k$, where

$$\begin{aligned} \alpha_k &\sim \mathrm{Bernoulli}\left(2/\binom{t}{t_0}, 1 - 2/\binom{t}{t_0}\right) \\ \beta_k &\sim \mathrm{Beta}(1, t) \end{aligned}$$

independently for $k = 2 \ldots d$. Consequently, $\mathbb{E}(c) = \frac{2(d-1)}{\binom{t}{t_0}(1+t)}$.

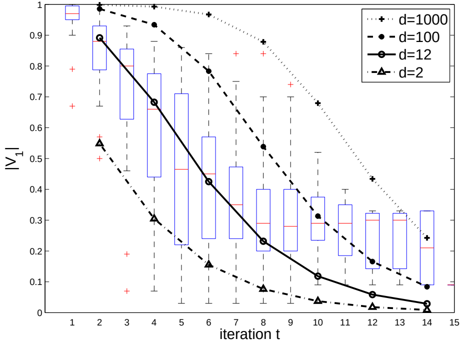
Comparing theory to behaviors

- On the "graspability" task with assumed d's:



Comparing theory to behaviors

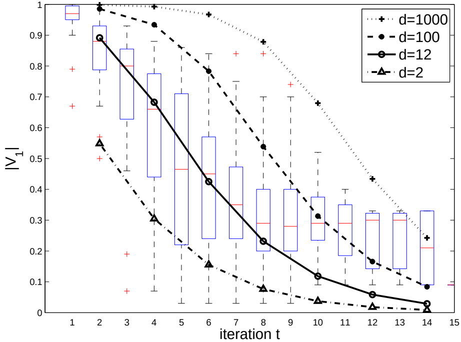
- On the "graspability" task with assumed d's:



- On the "lines" task, theory predicts $|V_1|$ at minimum in iteration 2

Comparing theory to behaviors

- On the "graspability" task with assumed d's:



- On the "lines" task, theory predicts $|V_1|$ at minimum in iteration 2
- Curriculum learning and teaching dimension both correct: different cases of the same theory

Conclusion

- A general teaching framework

Conclusion

- A general teaching framework
- Case studies match human teacher behaviors on graspability and lines

Conclusion

- A general teaching framework
- Case studies match human teacher behaviors on graspability and lines
  - sequential risk minimization

Conclusion

- A general teaching framework
- Case studies match human teacher behaviors on graspability and lines
  - sequential risk minimization
  - small d: boundary; large d: curriculum

Conclusion

- A general teaching framework
- Case studies match human teacher behaviors on graspability and lines
  - sequential risk minimization
  - small d: boundary; large d: curriculum
- Open questions:

Conclusion

- A general teaching framework
- Case studies match human teacher behaviors on graspability and lines
  - sequential risk minimization
  - small d: boundary; large d: curriculum
- Open questions:
  - optimal teaching strategy beyond the special cases?

Conclusion

- A general teaching framework
- Case studies match human teacher behaviors on graspability and lines
  - sequential risk minimization
  - small d: boundary; large d: curriculum
- Open questions:
  - optimal teaching strategy beyond the special cases?
  - can we use this theory to improve human learners?

Conclusion

- A general teaching framework
- Case studies match human teacher behaviors on graspability and lines
  - sequential risk minimization
  - small d: boundary; large d: curriculum
- Open questions:
  - optimal teaching strategy beyond the special cases?
  - can we use this theory to improve human learners?
- Acknowledgments

Conclusion

- A general teaching framework
- Case studies match human teacher behaviors on graspability and lines
  - sequential risk minimization
  - small d: boundary; large d: curriculum
- Open questions:
  - optimal teaching strategy beyond the special cases?
  - can we use this theory to improve human learners?
- Acknowledgments
  - Collaborators: Kwangsung Jun, Faisal Khan, Bilge Mutlu, Burr Settles

Conclusion

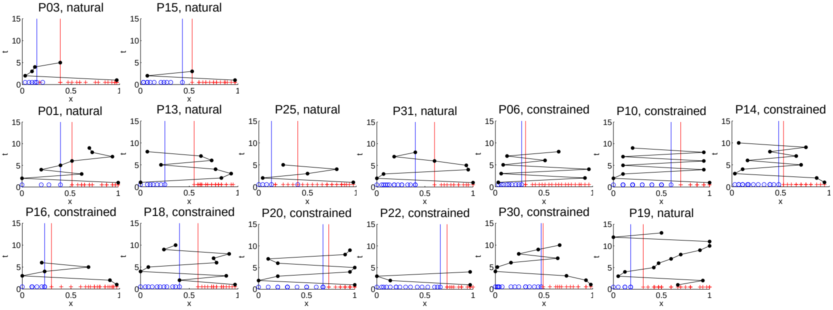
- A general teaching framework
- Case studies match human teacher behaviors on graspability and lines
  - sequential risk minimization
  - small d: boundary; large d: curriculum
- Open questions:
  - optimal teaching strategy beyond the special cases?
  - can we use this theory to improve human learners?
- Acknowledgments
  - Collaborators: Kwangsung Jun, Faisal Khan, Bilge Mutlu, Burr Settles
  - NSF CAREER IIS-0953219, AFOSR FA9550-09-1-0313, The Wisconsin Alumni Research Foundation

Backup slides

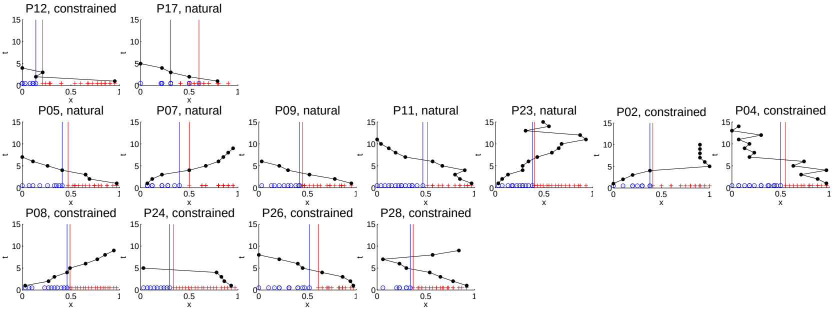
Graspability Strategy 1: “decision boundary” (0% subjects)

None

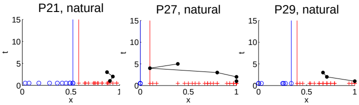
Strategy 2: "curriculum learning" (48% subjects)



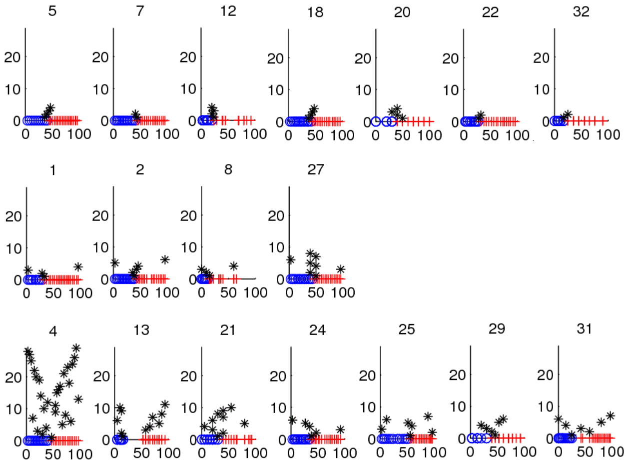
Strategy 3: "linear" (42% subjects)



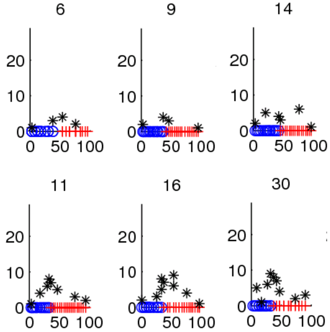
Strategy 4: "positive only" (10% subjects)



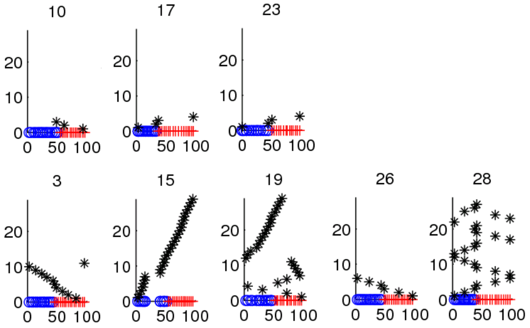
Line strategy 1: “decision boundary” (56% subjects)



Strategy 2: "curriculum learning" (19% subjects)



Strategy 3: “linear” (25% subjects)



Strategy 4: "positive only" (0% subjects)

None

Comparing the two experiments

strategy	boundary	curriculum	linear	positive
"graspability" $(n = 31)$	0%	48%	42%	10%
"lines" $(n = 32)$	56%	19%	25%	0%

The hidden dimensionality

- Humans represent objects by $\mathcal{X} \subseteq \mathbb{R}^d, d \gg 1$.

The hidden dimensionality

- Humans represent objects by $\mathcal{X} \subseteq \mathbb{R}^d, d \gg 1$.
- e.g., squirrel = Boolean vector (graspable, shy, store supplies for the winter, is not poisonous, has four paws, has teeth, has two ears, has two eyes, is beautiful, is brown, lives in trees, rodent, doesn't herd, doesn't sting, drinks water, eats nuts, feels soft, fluffy, gnaws on everything, has a beautiful tail, has a large tail, has a mouth, has a small head, has gnawing teeth, has pointy ears, has short paws, is afraid of people, is cute, is difficult to catch, is found in Belgium, is light, is not a pet, is not very big, is short haired, is sweet , jumps, lives in Europe, lives in the wild, short front legs, small ears, smaller than a horse, soft fur, timid animal, can't fly, climbs in trees, collects nuts, crawls up trees, eats acorns, eats plants, does not lay eggs ...)

The hidden dimensionality

- Humans represent objects by $\mathcal{X} \subseteq \mathbb{R}^d, d \gg 1$.
- e.g., squirrel = Boolean vector (graspable, shy, store supplies for the winter, is not poisonous, has four paws, has teeth, has two ears, has two eyes, is beautiful, is brown, lives in trees, rodent, doesn't herd, doesn't sting, drinks water, eats nuts, feels soft, fluffy, gnaws on everything, has a beautiful tail, has a large tail, has a mouth, has a small head, has gnawing teeth, has pointy ears, has short paws, is afraid of people, is cute, is difficult to catch, is found in Belgium, is light, is not a pet, is not very big, is short haired, is sweet , jumps, lives in Europe, lives in the wild, short front legs, small ears, smaller than a horse, soft fur, timid animal, can't fly, climbs in trees, collects nuts, crawls up trees, eats acorns, eats plants, does not lay eggs ...)
- "Graspability" is probably a 1D subspace in $\mathcal{X}$