Part 3: Beyond Disagreement-Based Active Learning – Current Directions

- Subregion-Based Active Learning
- Margin-Based Active Learning: Linear Separators
- Shattering-Based Active Learning
- Distribution-Free Analysis, Optimality
- TicToc: Adapting to Heterogeneous Noise
- Tsybakov Noise

Tutorial on Active Learning: Theory to Practice

Steve Hanneke
Toyota Technological Institute at Chicago
steve.hanneke@gmail.com

Robert Nowak
University of Wisconsin - Madison
rdnowak@wisc.edu

ICML | 2019
Thirty-sixth International Conference on Machine Learning

Subregion-Based Active Learning

Zhang & Chaudhuri, 2014

$$\mathrm{DIS}(\mathcal{H}) := \{x \in \mathcal{X} : \exists f, f' \in \mathcal{H}, f(x) \neq f'(x)\}$$

A^2 (Agnostic Active)

for $t = 1, 2, \ldots$ (til *stopping-criterion*)

1. **sample** 2^t unlabeled points S
2. **label** points in $Q = \mathrm{DIS}(\mathcal{H}) \cap S$
3. **optimize** $\hat{f} \leftarrow \operatorname*{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$
4. **reduce** $\mathcal{H}$: remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$.

output final $\hat{f}$

Subregion-Based Active Learning

Zhang & Chaudhuri, 2014

$$\mathrm{DIS}(\mathcal{H}) := \{x \in \mathcal{X} : \exists f, f' \in \mathcal{H}, f(x) \neq f'(x)\}$$

Subregion-based Active Learning

for $t = 1, 2, \ldots$ (til *stopping-criterion*)

1. **sample** 2^t unlabeled points S
2. **label** points in $Q = \mathcal{R}_{\epsilon'_t}(\mathcal{H}) \cap S$
3. **optimize** $\hat{f} \leftarrow \operatorname*{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$
4. **reduce** $\mathcal{H}$: remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$.

output final $\hat{f}$

Instead, pick **region $\mathcal{R}_{\epsilon'}(\mathcal{H})$** s.t.
$\forall f, f' \in \mathcal{H},\ P_X(x \notin \mathcal{R}_{\epsilon'}(\mathcal{H}) : f(x) \neq f'(x)) \leq \epsilon'$.

Pick ϵ' carefully each round,
$R(\hat{f}) - R(f^*) \leq \epsilon$ at end

e.g., Bounded noise: $\epsilon' \propto d2^{-t}$

Subregion-Based Active Learning

Zhang & Chaudhuri, 2014

$\mathrm{DIS}(\mathcal{H}) := \{x \in \mathcal{X} : \exists f, f' \in \mathcal{H}, f(x) \neq f'(x)\}$

Subregion-based Active Learning

for $t = 1, 2, \ldots$ (til *stopping-criterion*)

1. **sample** 2^t unlabeled points S
2. **label** points in $Q = \mathcal{R}_{\epsilon'_t}(\mathcal{H}) \cap S$
3. **optimize** $\hat{f} \leftarrow \operatorname*{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$
4. **reduce** $\mathcal{H}$: remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$.

output final $\hat{f}$

Pick region $\mathcal{R}_{\epsilon'}(\mathcal{H})$ s.t.

$\forall f, f' \in \mathcal{H},\ P_X(x \notin \mathcal{R}_{\epsilon'}(\mathcal{H}) : f(x) \neq f'(x)) \leq \epsilon'$.

$$\varphi_c := \sup_{r > \epsilon} \frac{P_X(\mathcal{R}_{r/c}(\mathrm{B}(f^*, r)))}{r}$$

Theorem: with **Bounded noise**,

$R(\hat{f}) \leq R(f^*) + \epsilon$ using # labels

$$\approx \varphi_c d \log\left(\tfrac{1}{\epsilon}\right)$$

Subregion-Based Active Learning

Zhang & Chaudhuri, 2014

$\mathrm{DIS}(\mathcal{H}) := \{x \in \mathcal{X} : \exists f, f' \in \mathcal{H}, f(x) \neq f'(x)\}$

Subregion-based Active Learning

for $t = 1, 2, \ldots$ (til *stopping-criterion*)

1. **sample** 2^t unlabeled points S
2. **label** points in $Q = \mathcal{R}_{\epsilon'_t}(\mathcal{H}) \cap S$
3. **optimize** $\hat{f} \leftarrow \underset{f \in \mathcal{H}}{\operatorname{argmin}} \hat{R}_Q(f)$
4. **reduce** $\mathcal{H}$: remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$

output final $\hat{f}$

Pick region $\mathcal{R}_{\epsilon'}(\mathcal{H})$ s.t.
$\forall f, f' \in \mathcal{H},\ P_X(x \notin \mathcal{R}_{\epsilon'}(\mathcal{H}) : f(x) \neq f'(x)) \leq \epsilon'.$

$$\varphi_c := \sup_{r > \epsilon} \frac{P_X(\mathcal{R}_{r/c}(\mathrm{B}(f^*, r)))}{r}$$

Theorem: with **Bounded noise**,
$R(\hat{f}) \leq R(f^*) + \epsilon$ using # labels

$$\approx \varphi_c d \log\left(\tfrac{1}{\epsilon}\right)$$

Agnostic case: $\varphi'_c := \sup_{r > \epsilon} \frac{P_X(\mathcal{R}_{r/c}(\mathrm{B}(f^*, 2\beta + r)))}{2\beta + r}$

Theorem:
$R(\hat{f}) \leq R(f^*) + \epsilon$ using # labels

$$\approx \varphi'_c d \frac{\beta^2}{\epsilon^2}$$

Subregion-Based Active Learning

Zhang & Chaudhuri, 2014

How to find such an $\mathcal{R}_{\epsilon'}(\mathcal{H})$?

- $\mathcal{R}_{\epsilon'}(\mathcal{H}) = \mathrm{DIS}(\mathcal{H})$ works
- Empirically (Zhang & Chaudhuri, 2014)
- Nice structure: e.g., **Linear separators**

Pick region $\mathcal{R}_{\epsilon'}(\mathcal{H})$ s.t.
$\forall f, f' \in \mathcal{H},\ P_X(x \notin \mathcal{R}_{\epsilon'}(\mathcal{H}) : f(x) \neq f'(x)) \leq \epsilon'$.

$$\varphi_c := \sup_{r > \epsilon} \frac{P_X(\mathcal{R}_{r/c}(\mathrm{B}(f^*, r)))}{r}$$

Theorem: with **Bounded noise**,
$R(\hat{f}) \leq R(f^*) + \epsilon$ using # labels

$$\approx \varphi_c d \log\left(\frac{1}{\epsilon}\right)$$

Subregion-Based Active Learning

Zhang & Chaudhuri, 2014

How to find such an $\mathcal{R}_{\epsilon'}(\mathcal{H})$?

- $\mathcal{R}_{\epsilon'}(\mathcal{H}) = \mathrm{DIS}(\mathcal{H})$ works
- Empirically (Zhang & Chaudhuri, 2014)
- Nice structure: e.g., **Linear separators**

  Margin-based Active Learning
  (Dasgupta, Kalai, Monteleoni, 2005;
  Balcan, Broder, Zhang, 2007; ...)

Pick region $\mathcal{R}_{\epsilon'}(\mathcal{H})$ s.t.
$\forall f, f' \in \mathcal{H}, P_X(x \notin \mathcal{R}_{\epsilon'}(\mathcal{H}) : f(x) \neq f'(x)) \leq \epsilon'$.

$$\varphi_c := \sup_{r>\epsilon} \frac{P_X(\mathcal{R}_{r/c}(\mathrm{B}(f^*, r)))}{r}$$

Theorem: with **Bounded noise**,
$R(\hat{f}) \leq R(f^*) + \epsilon$ using # labels

$$\approx \varphi_c d \log\left(\frac{1}{\epsilon}\right)$$

Subregion-Based Active Learning

Zhang & Chaudhuri, 2014

How to find such an $\mathcal{R}_{\epsilon'}(\mathcal{H})$?

- Nice structure: e.g., **Linear separators**

  Margin-based Active Learning
  (Dasgupta, Kalai, Monteleoni, 2005;
  Balcan, Broder, Zhang, 2007; ...)

Pick region $\mathcal{R}_{\epsilon'}(\mathcal{H})$ s.t.
$\forall f, f' \in \mathcal{H},\ P_X(x \notin \mathcal{R}_{\epsilon'}(\mathcal{H}) : f(x) \neq f'(x)) \leq \epsilon'.$

$$\varphi_c := \sup_{r>\epsilon} \frac{P_X(\mathcal{R}_{r/c}(\mathrm{B}(f^*,r)))}{r}$$

Theorem: with **Bounded noise**,
$R(\hat{f}) \leq R(f^*) + \epsilon$ using # labels

$$\approx \varphi_c d \log\left(\frac{1}{\epsilon}\right)$$

Subregion-Based Active Learning

How to find such an $\mathcal{R}_{\epsilon'}(\mathcal{H})$?

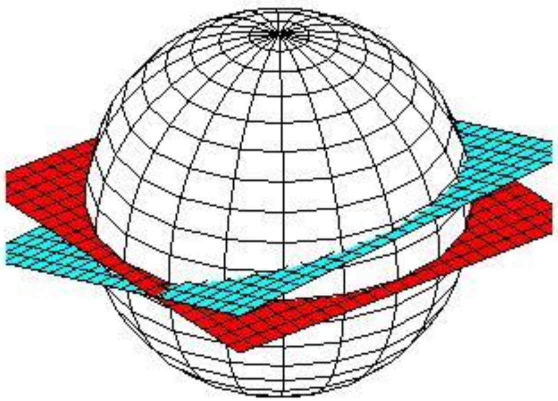
- Nice structure: e.g., **Linear separators**

  Margin-based Active Learning
  (Dasgupta, Kalai, Monteleoni, 2005;
  Balcan, Broder, Zhang, 2007; ...)

Pick **region $\mathcal{R}_{\epsilon'}(\mathcal{H})$** s.t.
$\forall f, f' \in \mathcal{H}, P_X(x \notin \mathcal{R}_{\epsilon'}(\mathcal{H}) : f(x) \neq f'(x)) \leq \epsilon'$.

Uniform P_X on d-dim sphere

For $w \in \mathrm{B}(w^*, r)$, **project** to $\mathrm{Span}(w, w^*)$



Subregion-Based Active Learning

How to find such an $\mathcal{R}_{\epsilon'}(\mathcal{H})$?

- Nice structure: e.g., **Linear separators**

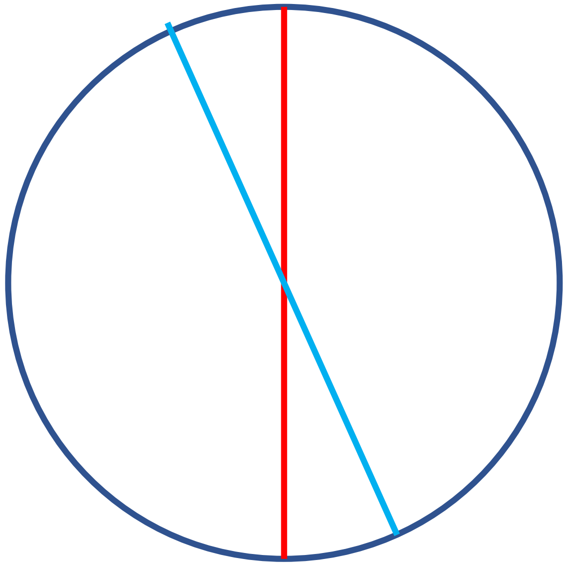
  Margin-based Active Learning
  (Dasgupta, Kalai, Monteleoni, 2005;
  Balcan, Broder, Zhang, 2007; ...)

Pick region $\mathcal{R}_{\epsilon'}(\mathcal{H})$ s.t.
$\forall f, f' \in \mathcal{H},\ P_X(x \notin \mathcal{R}_{\epsilon'}(\mathcal{H}) : f(x) \neq f'(x)) \leq \epsilon'$.

Uniform P_X on d-dim sphere

For $w \in \mathrm{B}(w^*, r)$, **project** to Span(w,w^*)

Most projected prob mass toward middle



Subregion-Based Active Learning

How to find such an $\mathcal{R}_{\epsilon'}(\mathcal{H})$?

- Nice structure: e.g., **Linear separators**

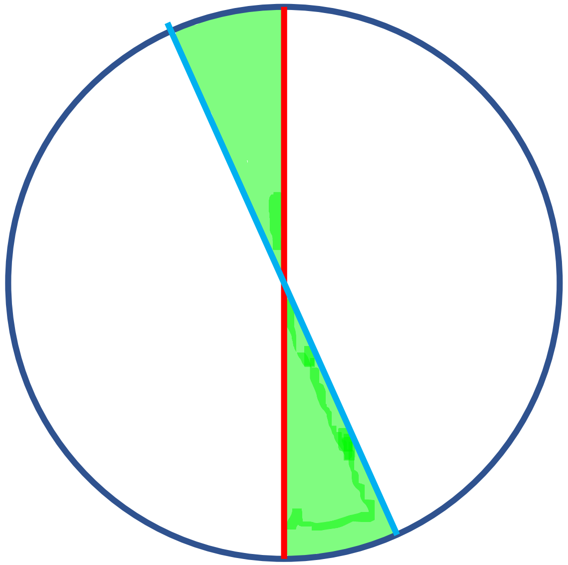
  Margin-based Active Learning
  (Dasgupta, Kalai, Monteleoni, 2005;
  Balcan, Broder, Zhang, 2007; ...)

Pick **region $\mathcal{R}_{\epsilon'}(\mathcal{H})$** s.t.
$\forall f, f' \in \mathcal{H},\ P_X(x \notin \mathcal{R}_{\epsilon'}(\mathcal{H}) : f(x) \neq f'(x)) \leq \epsilon'$.

Uniform P_X on d-dim sphere

For $w \in \mathrm{B}(w^*, r)$, **project** to Span(w,w^*)

Most projected prob mass toward middle



Subregion-Based Active Learning

How to find such an $\mathcal{R}_{\epsilon'}(\mathcal{H})$?

- Nice structure: e.g., **Linear separators**

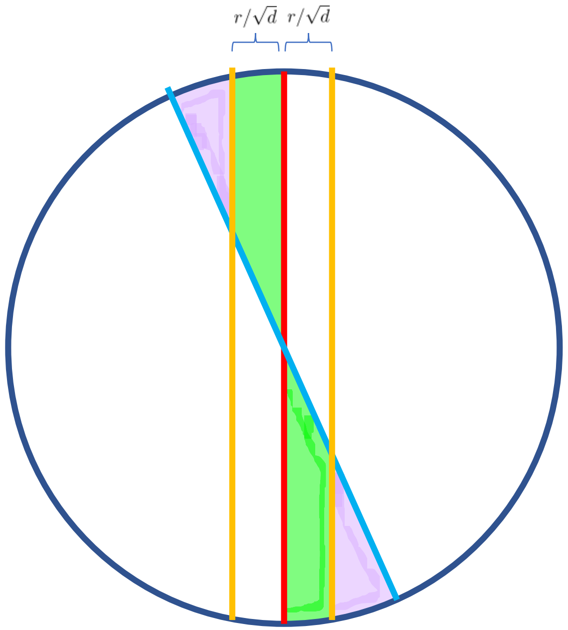
  Margin-based Active Learning
  (Dasgupta, Kalai, Monteleoni, 2005;
  Balcan, Broder, Zhang, 2007; ...)

Pick region $\mathcal{R}_{\epsilon'}(\mathcal{H})$ s.t.
$\forall f, f' \in \mathcal{H},\ P_X(x \notin \mathcal{R}_{\epsilon'}(\mathcal{H}) : f(x) \neq f'(x)) \leq \epsilon'$.

Uniform P_X on d-dim sphere

For $w \in \mathrm{B}(w^*, r)$, **project** to Span(w,w^*)

Most projected prob mass toward middle



DIS($\{w, w^*\}$) in
slab of width $\approx r$

Most of its prob in
slab of width $\approx r/\sqrt{d}$

Subregion-Based Active Learning

How to find such an $\mathcal{R}_{\epsilon'}(\mathcal{H})$?

- Nice structure: e.g., **Linear separators**

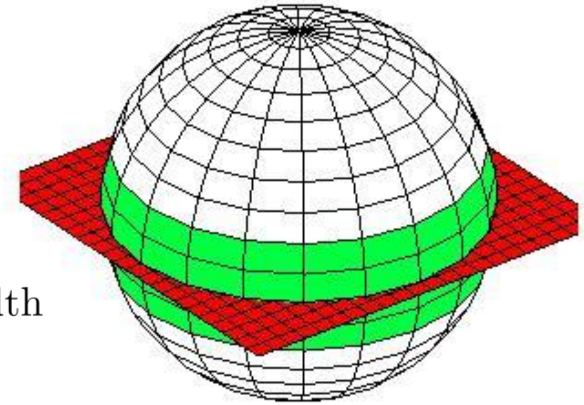
  Margin-based Active Learning
  (Dasgupta, Kalai, Monteleoni, 2005;
  Balcan, Broder, Zhang, 2007; ...)

DIS(B(f^*, r)) =
slab of width $\approx r$

Take $\mathcal{R}_{r/c}(\mathrm{B}(f^*, r))$ =
slab of width $\approx r/\sqrt{d}$

Prob in slab $\approx \sqrt{d} \times$ width

$\Rightarrow \varphi_c \leq$ constant

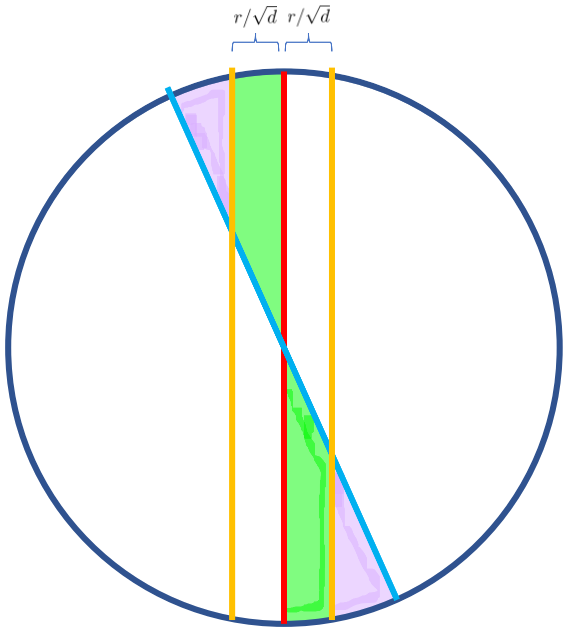


Pick region $\mathcal{R}_{\epsilon'}(\mathcal{H})$ s.t.
$\forall f, f' \in \mathcal{H}$, $P_X(x \notin \mathcal{R}_{\epsilon'}(\mathcal{H}) : f(x) \neq f'(x)) \leq \epsilon'$.

Uniform P_X on d-dim sphere

For $w \in \mathrm{B}(w^*, r)$, **project** to Span(w,w^*)

Most projected prob mass toward middle



DIS($\{w, w^*\}$) in
slab of width $\approx r$

Most of its prob in
slab of width $\approx r/\sqrt{d}$

Subregion-Based Active Learning

How to find such an $\mathcal{R}_{\epsilon'}(\mathcal{H})$?

- Nice structure: e.g., **Linear separators**

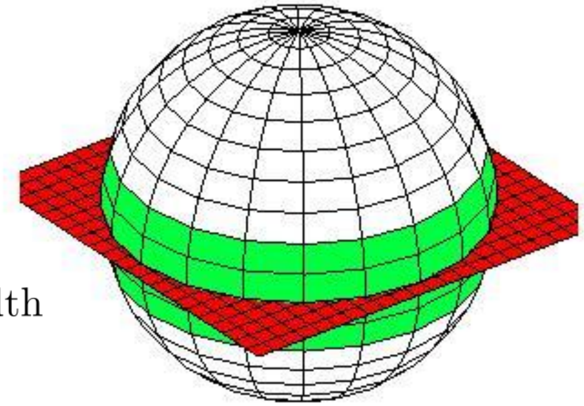
  Margin-based Active Learning
  (Dasgupta, Kalai, Monteleoni, 2005;
  Balcan

DIS(B(f^*, r)) =
slab of width $\approx r$

Take $\mathcal{R}_{r/c}(\mathrm{B}(f^*, r))$ =
slab of width $\approx r/\sqrt{d}$

Prob in slab $\approx \sqrt{d} \times$ width

$\Rightarrow \varphi_c \leq$ constant



Pick region $\mathcal{R}_{\epsilon'}(\mathcal{H})$ s.t.
$\forall f, f' \in \mathcal{H},\ P_X(x \notin \mathcal{R}_{\epsilon'}(\mathcal{H}) : f(x) \neq f'(x)) \leq \epsilon'$.

$$\varphi_c := \sup_{r>\epsilon} \frac{P_X(\mathcal{R}_{r/c}(\mathrm{B}(f^*,r)))}{r}$$

Theorem: with **Bounded noise**,
$R(\hat{f}) \leq R(f^*) + \epsilon$ using # labels
$\approx \varphi_c d \log\left(\frac{1}{\epsilon}\right)$

$\Rightarrow$ # labels $\approx d \log(\frac{1}{\epsilon})$ suffice

Recall:
Passive $\approx \frac{d}{\epsilon}$

Comparison:
Recall $\theta \approx \sqrt{d}$
$\Rightarrow A^2$ # labels $\approx d^{3/2} \log(\frac{1}{\epsilon})$

Margin-Based Active Learning

(Balcan, Broder, Zhang, 2007; ...)

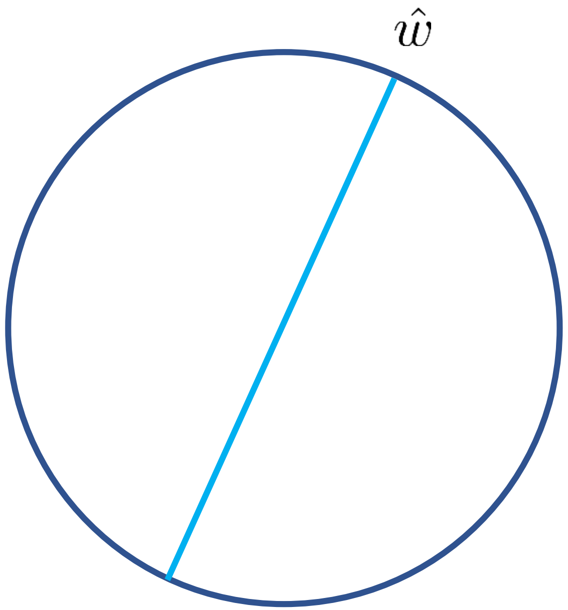
Margin-based Active Learning

Initialize $\hat{w}$

for $t = 1, 2, \ldots$ (til *stopping-criterion*)

1. **sample** $d2^t$ unlabeled points S
2. **label** points in $Q =$ all $x \in S$ s.t. $<\hat{w}, x> \;\leq\; c2^{-t}/\sqrt{d}$
3. **optimize** $\hat{w} \leftarrow \underset{w:\|w-\hat{w}\|\leq c'2^{-t}}{\operatorname{argmin}} \hat{R}_Q(w)$

output final $\hat{w}$



Uniform P_X on d-dim sphere

Margin-Based Active Learning

(Balcan, Broder, Zhang, 2007; ...)

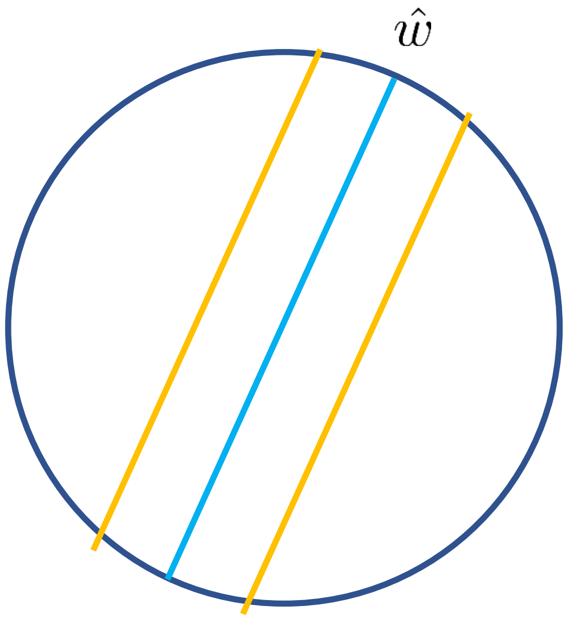
Margin-based Active Learning

Initialize $\hat{w}$

for $t = 1, 2, \ldots$ (til *stopping-criterion*)

1. **sample** $d2^t$ unlabeled points S
2. **label** points in $Q =$ all $x \in S$ s.t. $<\hat{w}, x> \;\leq\; c2^{-t}/\sqrt{d}$
3. **optimize** $\hat{w} \leftarrow \underset{w:\|w-\hat{w}\|\leq c'2^{-t}}{\operatorname{argmin}} \hat{R}_Q(w)$

output final $\hat{w}$



Uniform P_X on d-dim sphere

Margin-Based Active Learning

(Balcan, Broder, Zhang, 2007; ...)

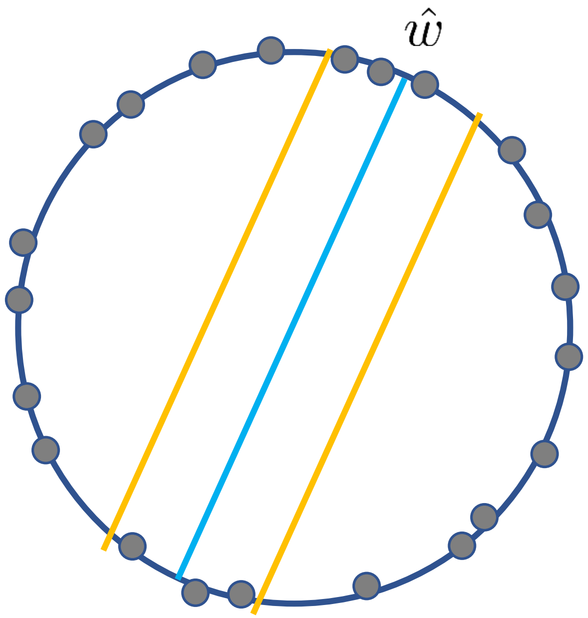
Margin-based Active Learning

Initialize $\hat{w}$

for $t = 1, 2, \ldots$ (til *stopping-criterion*)

1. **sample** $d2^t$ unlabeled points S
2. **label** points in $Q = $ all $x \in S$ s.t. $<\hat{w}, x> \;\leq\; c2^{-t}/\sqrt{d}$
3. **optimize** $\hat{w} \leftarrow \underset{w:\|w-\hat{w}\| \leq c' 2^{-t}}{\operatorname{argmin}} \hat{R}_Q(w)$

output final $\hat{w}$



Uniform P_X on d-dim sphere

Margin-Based Active Learning

(Balcan, Broder, Zhang, 2007; ...)

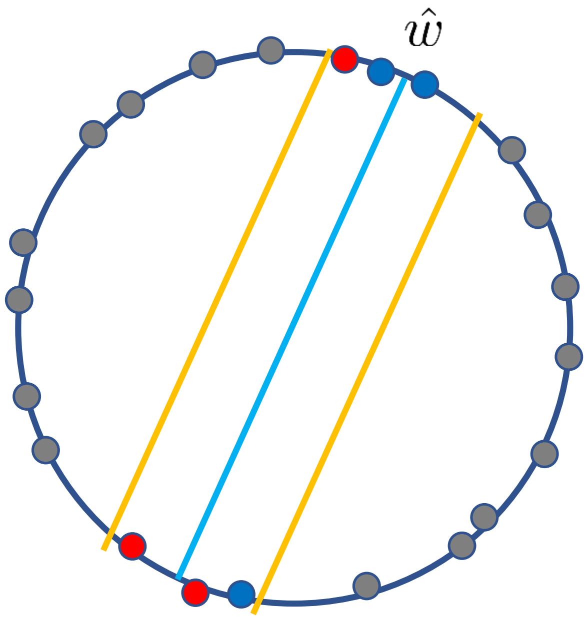
Margin-based Active Learning

Initialize $\hat{w}$

for $t = 1, 2, \ldots$ (til *stopping-criterion*)

1. **sample** $d2^t$ unlabeled points S
2. **label** points in $Q =$ all $x \in S$ s.t. $<\hat{w}, x> \;\leq\; c2^{-t}/\sqrt{d}$
3. **optimize** $\hat{w} \leftarrow \underset{w:\|w-\hat{w}\|\leq c'2^{-t}}{\operatorname{argmin}} \hat{R}_Q(w)$

output final $\hat{w}$



Uniform P_X on d-dim sphere

Margin-Based Active Learning

(Balcan, Broder, Zhang, 2007; ...)

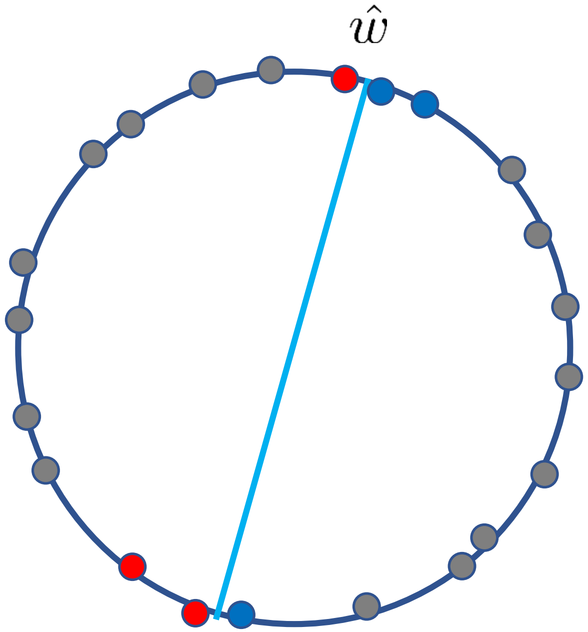
Margin-based Active Learning

Initialize $\hat{w}$

for $t = 1, 2, \ldots$ (til *stopping-criterion*)

1. **sample** $d2^t$ unlabeled points S
2. **label** points in $Q =$ all $x \in S$ s.t. $< \hat{w}, x > \;\; \leq \;\; c2^{-t}/\sqrt{d}$
3. **optimize** $\hat{w} \leftarrow \underset{w : \|w - \hat{w}\| \leq c' 2^{-t}}{\operatorname{argmin}} \hat{R}_Q(w)$

output final $\hat{w}$



Uniform P_X on d-dim sphere

Margin-Based Active Learning

(Balcan, Broder, Zhang, 2007; ...)

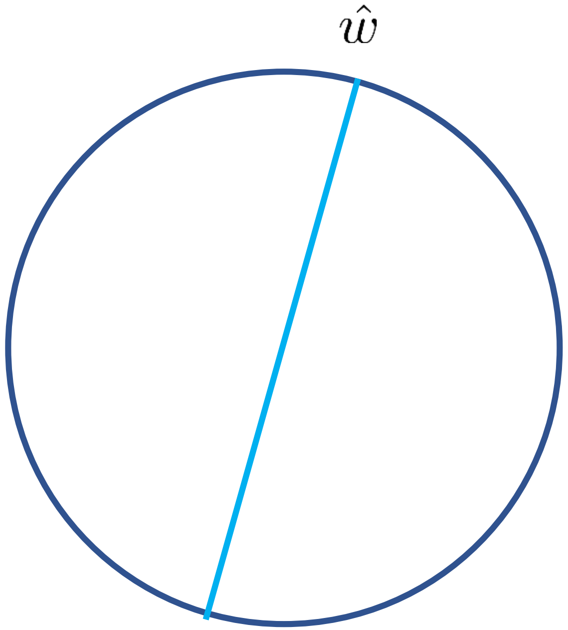
Margin-based Active Learning

Initialize $\hat{w}$

for $t = 1, 2, \ldots$ (til *stopping-criterion*)

1. **sample** $d2^t$ unlabeled points S
2. **label** points in $Q =$ all $x \in S$ s.t. $<\hat{w}, x> \;\leq\; c2^{-t}/\sqrt{d}$
3. **optimize** $\hat{w} \leftarrow \underset{w:\|w-\hat{w}\|\leq c'2^{-t}}{\operatorname{argmin}} \hat{R}_Q(w)$

output final $\hat{w}$



Uniform P_X on d-dim sphere

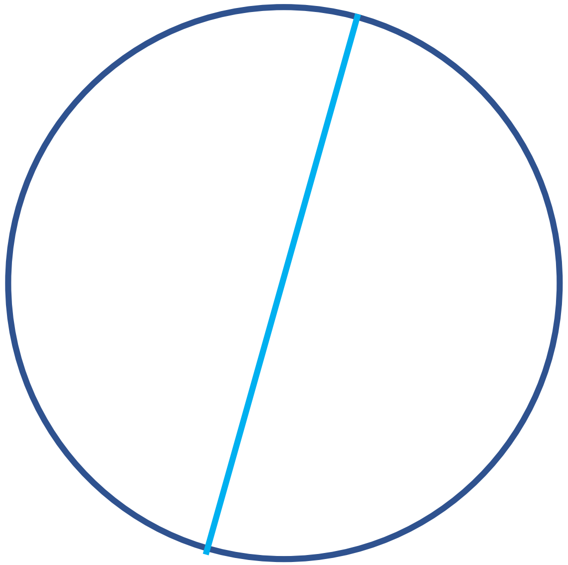
Theorem: with **Bounded noise**,
$R(\hat{f}) \leq R(f^*) + \epsilon$ using # labels
$\approx d \log\left(\frac{1}{\epsilon}\right)$

(also works for isotropic log-concave distributions)

Computational Efficiency

(Awasthi, Balcan, Long, 2014,...)

Uniform P_X on d-dim sphere



Efficient Alg

Initialize $\hat{w}$

for $t = 1, 2, \ldots$ (til *stopping-criterion*)

1. **sample** $d2^t$ unlabeled points S
2. **label** points in $Q =$ all $x \in S$ s.t. $<\hat{w}, x> \;\; \leq \;\; c2^{-t}/\sqrt{d}$
3. **optimize** $\hat{w} \leftarrow \underset{w:\|w-\hat{w}\| \leq c' 2^{-t}}{\operatorname{argmin}} \hat{R}_Q^{\ell_t}(w)$

output final $\hat{w}$

Surrogate loss

$$\ell_t(<w, x>, y) \approx \max\{1 - 2^t\sqrt{d}(y<w, x>), 0\}$$

Hinge loss slope **changes** each round

Computational Efficiency

(Awasthi, Balcan, Long, 2014,...)

Efficient Alg

Initialize $\hat{w}$

for $t = 1, 2, \ldots$ (til *stopping-criterion*)

1. **sample** $d2^t$ unlabeled points S
2. **label** points in $Q =$ all $x \in S$ s.t. $<\hat{w}, x> \;\; \leq \;\; c2^{-t}/\sqrt{d}$
3. **optimize** $\hat{w} \leftarrow \underset{w: \|w - \hat{w}\| \leq c' 2^{-t}}{\text{argmin}} \hat{R}_Q^{\ell_t}(w)$

output final $\hat{w}$

Surrogate loss

$$\ell_t(<w, x>, y) \approx \max\{1 - 2^t \sqrt{d}(y <w, x>), 0\}$$

Hinge loss slope **changes** each round

Uniform P_X on d-dim sphere

Theorem: with **Bounded noise**,
$R(\hat{f}) \leq R(f^*) + \epsilon$ using # labels
$$\approx d \log\left(\tfrac{1}{\epsilon}\right)$$
and running in polynomial time

Computational Efficiency

(Awasthi, Balcan, Long, 2014,...)

Efficient Alg

Initialize $\hat{w}$

for $t = 1, 2, \ldots$ (til *stopping-criterion*)

1. **sample** $d2^t$ unlabeled points S
2. **label** points in $Q =$ all $x \in S$ s.t. $<\hat{w}, x> \ \leq \ c2^{-t}/\sqrt{d}$
3. **optimize** $\hat{w} \leftarrow \underset{w:\|w-\hat{w}\| \leq c' 2^{-t}}{\operatorname{argmin}} \hat{R}_Q^{\ell_t}(w)$

output final $\hat{w}$

Surrogate loss

$$\ell_t(<w, x>, y) \approx \max\{1 - 2^t\sqrt{d}(y<w, x>), 0\}$$

Hinge loss slope **changes** each round

Uniform P_X on d-dim sphere

Theorem: with **Bounded noise**, $R(\hat{f}) \leq R(f^*) + \epsilon$ using # labels $\approx d \log\left(\frac{1}{\epsilon}\right)$ **and running in polynomial time**

Theorem: with **Agnostic** case, $R(\hat{f}) \leq C R(f^*)$ **in polynomial time**

(was first alg. known to achieve these; even passively)

(also works for isotropic log-concave distributions)

Up Next:
Shattering-Based Active Learning

Shattering-Based Active Learning

(Hanneke, 2009, 2012)

Recall: $\mathcal{H}$ **shatters** $x_1, \ldots, x_k$ if all 2^k classifications realized by $\mathcal{H}$

$\mathrm{DIS}(\mathcal{H})$ checks for shattering 1 point.

Idea: Generalize to shattering ≥ 1 points.

Shattering-Based Active Learning

(Hanneke, 2009, 2012)

Recall: $\mathcal{H}$ **shatters** $x_1, \ldots, x_k$ if all 2^k classifications realized by $\mathcal{H}$

A^2 (Agnostic Active)

for $t = 1, 2, \ldots$ (til *stopping-criterion*)

1. **sample** 2^t unlabeled points S
2. **label** points in $Q = \mathrm{DIS}(\mathcal{H}) \cap S$
3. **optimize** $\hat{f} \leftarrow \underset{f \in \mathcal{H}}{\operatorname{argmin}} \hat{R}_Q(f)$
4. **reduce** $\mathcal{H}$: remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$.

output final $\hat{f}$

$\mathrm{DIS}(\mathcal{H})$ checks for shattering 1 point.

Idea: Generalize to shattering ≥ 1 points.

Shattering-Based Active Learning

(Hanneke, 2009, 2012)

Recall: $\mathcal{H}$ **shatters** $x_1, \ldots, x_k$ if all 2^k classifications realized by $\mathcal{H}$

Shattering-based Active Learning

for $t = 1, 2, \ldots$ (til *stopping-criterion*)

1. **sample** 2^t unlabeled points S
2. **label** points in $Q =$ all $x \in S$ s.t.
   $P_X^k(A \in \mathcal{X}^k : \mathcal{H} \text{ shatters } A \cup \{x\} | \mathcal{H} \text{ shatters } A) \geq \frac{1}{2}$
3. **optimize** $\hat{f} \leftarrow \underset{f \in \mathcal{H}}{\operatorname{argmin}} \hat{R}_Q(f)$
4. **reduce** $\mathcal{H}$: remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$.

output final $\hat{f}$

DIS($\mathcal{H}$) checks for shattering 1 point.

Idea: Generalize to shattering ≥ 1 points.

Shattering-Based Active Learning

(Hanneke, 2009, 2012)

Recall: $\mathcal{H}$ **shatters** $x_1, \ldots, x_k$ if all 2^k classifications realized by $\mathcal{H}$

Shattering-based Active Learning

for $t = 1, 2, \ldots$ (til *stopping-criterion*)

1. **sample** 2^t unlabeled points S
2. **label** points in $Q =$ all $x \in S$ s.t.
   $P_X^k(A \in \mathcal{X}^k : \mathcal{H} \text{ shatters } A \cup \{x\} | \mathcal{H} \text{ shatters } A) \geq \frac{1}{2}$
3. **add** the remaining points $x \in S$ to Q with label
   $\hat{y}_x := \operatorname*{argmax}_y P_X^k(A \in \mathcal{X}^k : \mathcal{H}_{x,y} \text{ shatters } A | \mathcal{H} \text{ shatters } A)$
4. **optimize** $\hat{f} \leftarrow \operatorname*{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$
5. **reduce** $\mathcal{H}$: remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$

output final $\hat{f}$

DIS($\mathcal{H}$) checks for shattering 1 point.

Idea: Generalize to shattering ≥ 1 points.

Denote $\mathcal{H}_{x,y} := \{h \in \mathcal{H} : h(x) = y\}$

Shattering-Based Active Learning

Recall: $\mathcal{H}$ **shatters** $x_1, \ldots, x_k$ if all 2^k classifications realized by $\mathcal{H}$

Shattering-based Active Learning

for $t = 1, 2, \ldots$ (til *stopping-criterion*)

1. **sample** 2^t unlabeled points S
2. **label** points in $Q =$ all $x \in S$ s.t.
   $P_X^k(A \in \mathcal{X}^k : \mathcal{H} \text{ shatters } A \cup \{x\} | \mathcal{H} \text{ shatters } A) \geq \frac{1}{2}$
3. **add** the remaining points $x \in S$ to Q with label
   $\hat{y}_x := \operatorname*{argmax}_y P_X^k(A \in \mathcal{X}^k : \mathcal{H}_{x,y} \text{ shatters } A | \mathcal{H} \text{ shatters } A)$
4. **optimize** $\hat{f} \leftarrow \operatorname*{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$
5. **reduce** $\mathcal{H}$: remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$

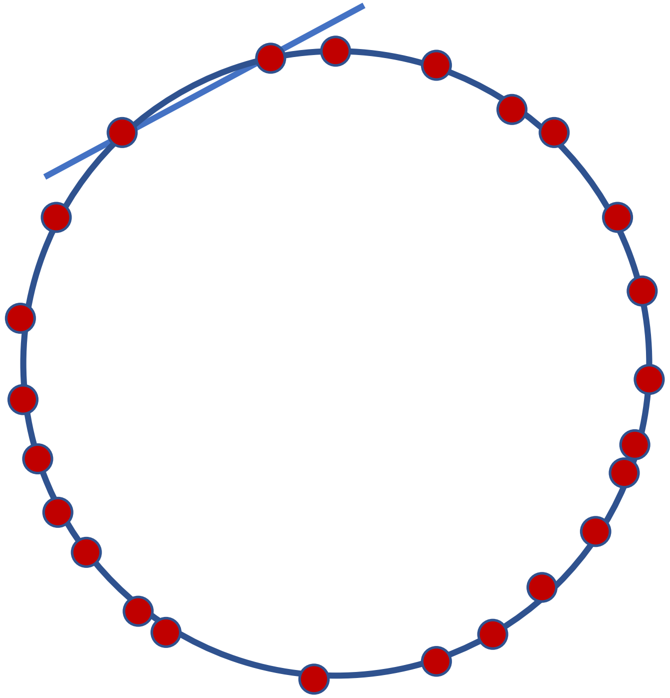
output final $\hat{f}$

Denoting $\mathcal{H}_{x,y} := \{h \in \mathcal{H} : h(x) = y\}$

Example: Linear separators, Uniform P_X on circle

Suppose true labels are **all** -1

$\text{DIS}(\mathcal{H}) =$ **entire circle**



Shattering-Based Active Learning

Recall: $\mathcal{H}$ **shatters** $x_1, \ldots, x_k$ if all 2^k classifications realized by $\mathcal{H}$

Shattering-based Active Learning

for $t = 1, 2, \ldots$ (til *stopping-criterion*)

1. **sample** 2^t unlabeled points S
2. **label** points in $Q =$ all $x \in S$ s.t.
   $P_X^k(A \in \mathcal{X}^k : \mathcal{H} \text{ shatters } A \cup \{x\} | \mathcal{H} \text{ shatters } A) \geq \frac{1}{2}$
3. **add** the remaining points $x \in S$ to Q with label
   $\hat{y}_x := \underset{y}{\operatorname{argmax}}\, P_X^k(A \in \mathcal{X}^k : \mathcal{H}_{x,y} \text{ shatters } A | \mathcal{H} \text{ shatters } A)$
4. **optimize** $\hat{f} \leftarrow \underset{f \in \mathcal{H}}{\operatorname{argmin}}\, \hat{R}_Q(f)$
5. **reduce** $\mathcal{H}$: remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f})\frac{d}{|Q|}}$

output final $\hat{f}$

Denoting $\mathcal{H}_{x,y} := \{h \in \mathcal{H} : h(x) = y\}$

Example: Linear separators, Uniform P_X on circle

Suppose true labels are **all** $\mathbf{-1}$

$\text{DIS}(\mathcal{H}) =$ **entire circle**

Try $k = 1$

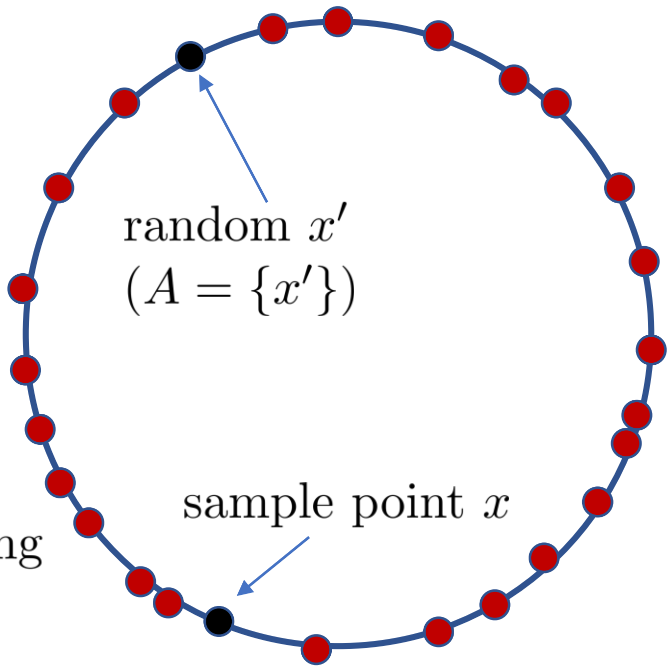
Given sample x

Rand x' probably not close

Can't shatter $\{x, x'\}$

without a lot of points wrong

So won't query x



Shattering-Based Active Learning

Recall: $\mathcal{H}$ **shatters** $x_1, \ldots, x_k$ if all 2^k classifications realized by $\mathcal{H}$

Shattering-based Active Learning

for $t = 1, 2, \ldots$ (til *stopping-criterion*)

1. **sample** 2^t unlabeled points S
2. **label** points in $Q =$ all $x \in S$ s.t. $P_X^k(A \in \mathcal{X}^k : \mathcal{H} \text{ shatters } A \cup \{x\} | \mathcal{H} \text{ shatters } A) \geq \frac{1}{2}$
3. **add** the remaining points $x \in S$ to Q with label $\hat{y}_x := \operatorname*{argmax}_y P_X^k(A \in \mathcal{X}^k : \mathcal{H}_{x,y} \text{ shatters } A | \mathcal{H} \text{ shatters } A)$
4. **optimize** $\hat{f} \leftarrow \operatorname*{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$
5. **reduce** $\mathcal{H}$: remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$

output final $\hat{f}$

Denoting $\mathcal{H}_{x,y} := \{h \in \mathcal{H} : h(x) = y\}$

Example: Linear separators, Uniform P_X on circle
Suppose true labels are **all** -1

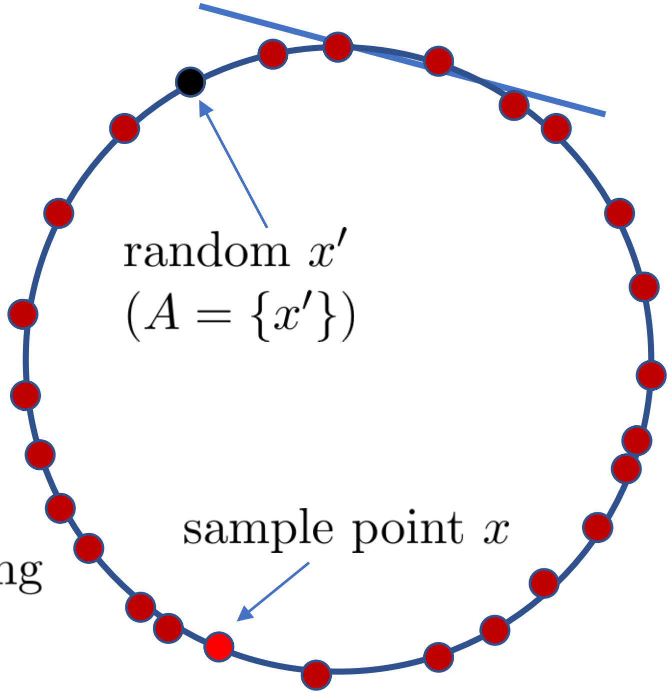
$\text{DIS}(\mathcal{H}) =$ **entire circle**

Try $k = 1$

Given sample x
Rand x' probably not close

Can't shatter $\{x, x'\}$
without a lot of points wrong

So won't query x



$\text{DIS}(\mathcal{H}_{x,-1})$ still entire circle (minus x)

$\text{DIS}(\mathcal{H}_{x,+1})$ **small** region

$\Rightarrow \hat{y}_x = -1$

Shattering-Based Active Learning

Recall: $\mathcal{H}$ **shatters** $x_1, \ldots, x_k$ if all 2^k classifications realized by $\mathcal{H}$

Shattering-based Active Learning

for $t = 1, 2, \ldots$ (til *stopping-criterion*)

1. **sample** 2^t unlabeled points S
2. **label** points in $Q =$ all $x \in S$ s.t.
   $P_X^k(A \in \mathcal{X}^k : \mathcal{H} \text{ shatters } A \cup \{x\} | \mathcal{H} \text{ shatters } A) \geq \frac{1}{2}$
3. **add** the remaining points $x \in S$ to Q with label
   $\hat{y}_x := \underset{y}{\operatorname{argmax}} P_X^k(A \in \mathcal{X}^k : \mathcal{H}_{x,y} \text{ shatters } A | \mathcal{H} \text{ shatters } A)$
4. **optimize** $\hat{f} \leftarrow \underset{f \in \mathcal{H}}{\operatorname{argmin}} \hat{R}_Q(f)$
5. **reduce** $\mathcal{H}$: remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$

output final $\hat{f}$

Denoting $\mathcal{H}_{x,y} := \{h \in \mathcal{H} : h(x) = y\}$

Example: Linear separators, Uniform P_X on circle

Suppose true labels are **all** $\mathbf{-1}$

$\mathrm{DIS}(\mathcal{H}) =$ **entire circle**

Try $k = 1$

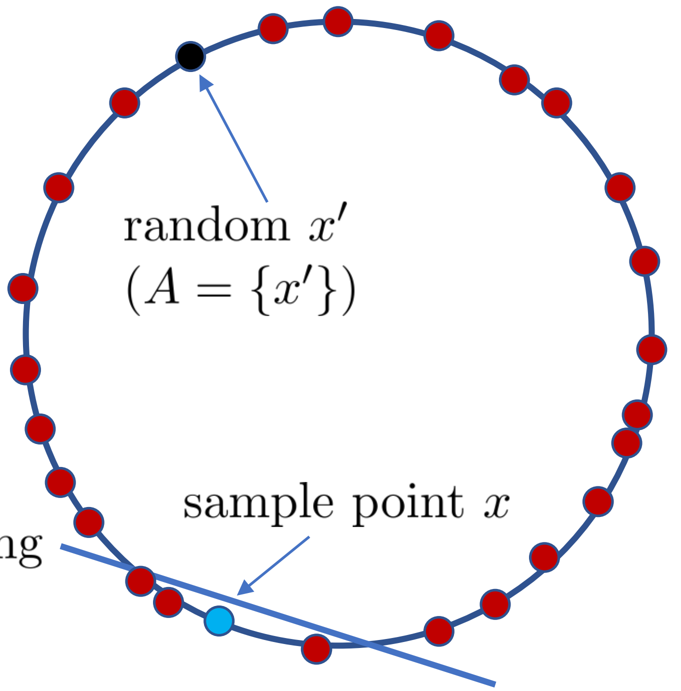
Given sample x

Rand x' probably not close

Can't shatter $\{x, x'\}$

without a lot of points wrong

So won't query x



$\mathrm{DIS}(\mathcal{H}_{x,-1})$ still entire circle (minus x)

$\mathrm{DIS}(\mathcal{H}_{x,+1})$ **small** region

$\Rightarrow \hat{y}_x = -1$

Shattering-Based Active Learning

Recall: $\mathcal{H}$ **shatters** $x_1, \ldots, x_k$ if all 2^k classifications realized by $\mathcal{H}$

Shattering-based Active Learning

for $t = 1, 2, \ldots$ (til *stopping-criterion*)

1. **sample** 2^t unlabeled points S
2. **label** points in $Q =$ all $x \in S$ s.t.
   $P_X^k(A \in \mathcal{X}^k : \mathcal{H} \text{ shatters } A \cup \{x\} | \mathcal{H} \text{ shatters } A) \geq \frac{1}{2}$
3. **add** the remaining points $x \in S$ to Q with label
   $\hat{y}_x := \operatorname*{argmax}_y P_X^k(A \in \mathcal{X}^k : \mathcal{H}_{x,y} \text{ shatters } A | \mathcal{H} \text{ shatters } A)$
4. **optimize** $\hat{f} \leftarrow \operatorname*{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$
5. **reduce** $\mathcal{H}$: remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$

output final $\hat{f}$

Denoting $\mathcal{H}_{x,y} := \{h \in \mathcal{H} : h(x) = y\}$

Example: Linear separators, Uniform P_X on circle
Suppose true labels are **all** $\mathbf{-1}$

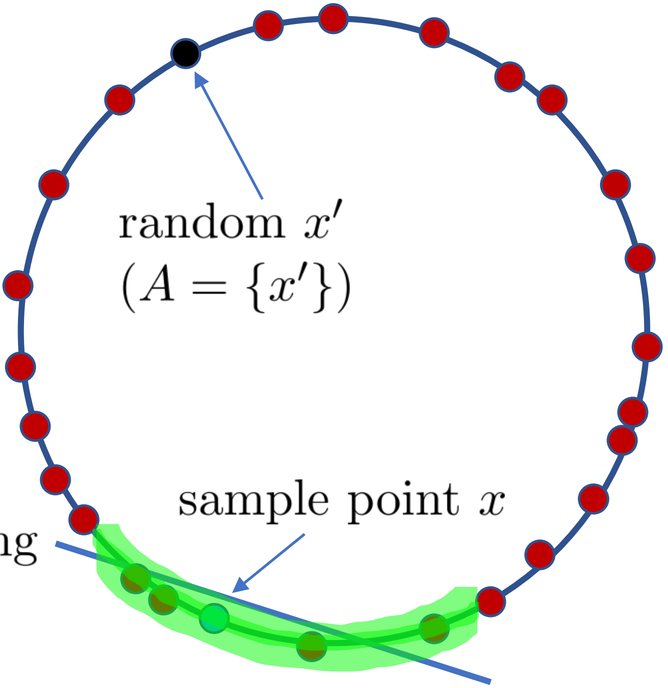
$\text{DIS}(\mathcal{H}) =$ **entire circle**

Try $k = 1$

Given sample x
Rand x' probably not close

Can't shatter $\{x, x'\}$
without a lot of points wrong

So won't query x



$\text{DIS}(\mathcal{H}_{x,-1})$ still entire circle (minus x)

$\text{DIS}(\mathcal{H}_{x,+1})$ **small** region

$\Rightarrow \hat{y}_x = -1$

Shattering-Based Active Learning

Recall: $\mathcal{H}$ **shatters** $x_1, \ldots, x_k$ if
all 2^k classifications realized by $\mathcal{H}$

Shattering-based Active Learning

for $t = 1, 2, \ldots$ (til *stopping-criterion*)

1. **sample** 2^t unlabeled points S
2. **label** points in $Q =$ all $x \in S$ s.t.
   $P_X^k(A \in \mathcal{X}^k : \mathcal{H} \text{ shatters } A \cup \{x\} | \mathcal{H} \text{ shatters } A) \geq \frac{1}{2}$
3. **add** the remaining points $x \in S$ to Q with label
   $\hat{y}_x := \operatorname*{argmax}_y P_X^k(A \in \mathcal{X}^k : \mathcal{H}_{x,y} \text{ shatters } A | \mathcal{H} \text{ shatters } A)$
4. **optimize** $\hat{f} \leftarrow \operatorname*{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$
5. **reduce** $\mathcal{H}$: remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$

output final $\hat{f}$

Generally, need to try various k and pick one
(See the papers)

Denoting $\mathcal{H}_{x,y} := \{h \in \mathcal{H} : h(x) = y\}$

Shattering-Based Active Learning

Recall: $\mathcal{H}$ **shatters** $x_1, \ldots, x_k$ if all 2^k classifications realized by $\mathcal{H}$

Shattering-based Active Learning

for $t = 1, 2, \ldots$ (til *stopping-criterion*)

1. **sample** 2^t unlabeled points S
2. **label** points in $Q =$ all $x \in S$ s.t.
   $P_X^k(A \in \mathcal{X}^k : \mathcal{H} \text{ shatters } A \cup \{x\} | \mathcal{H} \text{ shatters } A) \geq \frac{1}{2}$
3. **add** the remaining points $x \in S$ to Q with label
   $\hat{y}_x := \operatorname*{argmax}_y P_X^k(A \in \mathcal{X}^k : \mathcal{H}_{x,y} \text{ shatters } A | \mathcal{H} \text{ shatters } A)$
4. **optimize** $\hat{f} \leftarrow \operatorname*{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$
5. **reduce** $\mathcal{H}$: remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$

output final $\hat{f}$

Denoting $\mathcal{H}_{x,y} := \{h \in \mathcal{H} : h(x) = y\}$

Generally, need to try various k and pick one
(See the papers)

$$\theta^{(k)} := \sup_{r > \epsilon} \frac{P_X^k(A \in \mathcal{X}^k : \mathrm{B}(f^*, r) \text{ shatters } A)}{r}$$

$$\tilde{d} := \min\left\{k : P_X^k(A \in \mathcal{X}^k : \mathrm{B}(f^*, r) \text{ shatters } A) \xrightarrow[r \to 0]{} 0\right\}$$

$$\tilde{\theta} := \theta^{(\tilde{d})}$$

Theorem: For Bounded noise, $R(\hat{f}) \leq R(f^*) + \epsilon$
with # labels

$$\approx C\tilde{\theta} d \log\left(\frac{1}{\epsilon}\right)$$

Note: $\tilde{\theta} \ll \frac{1}{\epsilon}$

Shattering-Based Active Learning

Recall: $\mathcal{H}$ **shatters** $x_1, \ldots, x_k$ if
all 2^k classifications realized by $\mathcal{H}$

Shattering-based Active Learning

for $t = 1, 2, \ldots$ (til *stopping-criterion*)

1. **sample** 2^t unlabeled points S
2. **label** points in $Q =$ all $x \in S$ s.t.
   $P_X^k(A \in \mathcal{X}^k : \mathcal{H} \text{ shatters } A \cup \{x\} | \mathcal{H} \text{ shatters } A) \geq \frac{1}{2}$
3. **add** the remaining points $x \in S$ to Q with label
   $\hat{y}_x := \underset{y}{\operatorname{argmax}}\, P_X^k(A \in \mathcal{X}^k : \mathcal{H}_{x,y} \text{ shatters } A | \mathcal{H} \text{ shatters } A)$
4. **optimize** $\hat{f} \leftarrow \underset{f \in \mathcal{H}}{\operatorname{argmin}}\, \hat{R}_Q(f)$
5. **reduce** $\mathcal{H}$: remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$

output final $\hat{f}$

Denoting $\mathcal{H}_{x,y} := \{h \in \mathcal{H} : h(x) = y\}$

Generally, need to try various k and pick one
(See the papers)

$$\theta^{(k)} := \sup_{r > \epsilon} \frac{P_X^k(A \in \mathcal{X}^k : \mathrm{B}(f^*, r) \text{ shatters } A)}{r}$$

$$\tilde{d} := \min\left\{k : P_X^k(A \in \mathcal{X}^k : \mathrm{B}(f^*, r) \text{ shatters } A) \xrightarrow[r \to 0]{} 0\right\}$$

$$\tilde{\theta} := \theta^{(\tilde{d})}$$

Theorem: For Bounded noise, $R(\hat{f}) \leq R(f^*) + \epsilon$
with # labels

$$\approx C \tilde{\theta} \tilde{d} \log\left(\frac{1}{\epsilon}\right)$$

Note: $\tilde{\theta} \ll \frac{1}{\epsilon}$

In the example: $\tilde{\theta} = 2$, $\theta = \frac{1}{\epsilon}$

Shattering-Based Active Learning

Recall: $\mathcal{H}$ **shatters** $x_1, \ldots, x_k$ if
all 2^k classifications realized by $\mathcal{H}$

Shattering-based Active Learning

for $t = 1, 2, \ldots$ (til *stopping-criterion*)

1. **sample** 2^t unlabeled points S
2. **label** points in Q = all $x \in S$ s.t.
   $P_X^k(A \in \mathcal{X}^k : \mathcal{H} \text{ shatters } A \cup \{x\} | \mathcal{H} \text{ shatters } A) \geq \frac{1}{2}$
3. **add** the remaining points $x \in S$ to Q with label
   $\hat{y}_x := \operatorname{argmax}_y P_X^k(A \in \mathcal{X}^k : \mathcal{H}_{x,y} \text{ shatters } A | \mathcal{H} \text{ shatters } A)$
4. **optimize** $\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$
5. **reduce** $\mathcal{H}$: remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$

output final $\hat{f}$

Denoting $\mathcal{H}_{x,y} := \{h \in \mathcal{H} : h(x) = y\}$

Generally, need to try various k and pick one
(See the papers)

$$\theta^{(k)} := \sup_{r > \epsilon} \frac{P_X^k(A \in \mathcal{X}^k : \mathrm{B}(f^*, r) \text{ shatters } A)}{r}$$

$$\tilde{d} := \min\left\{k : P_X^k(A \in \mathcal{X}^k : \mathrm{B}(f^*, r) \text{ shatters } A) \xrightarrow[r \to 0]{} 0\right\}$$

$$\tilde{\theta} := \theta^{(\tilde{d})}$$

Theorem: For Bounded noise, $R(\hat{f}) \leq R(f^*) + \epsilon$
with # labels

$$\approx C \tilde{\theta} \tilde{d} \log\left(\frac{1}{\epsilon}\right)$$

Note: $\tilde{\theta} \ll \frac{1}{\epsilon}$ (may depend on f^*, P_X)

In the example: $\tilde{\theta} = 2$, $\theta = \frac{1}{\epsilon}$

Up Next:
Distribution-free Analysis

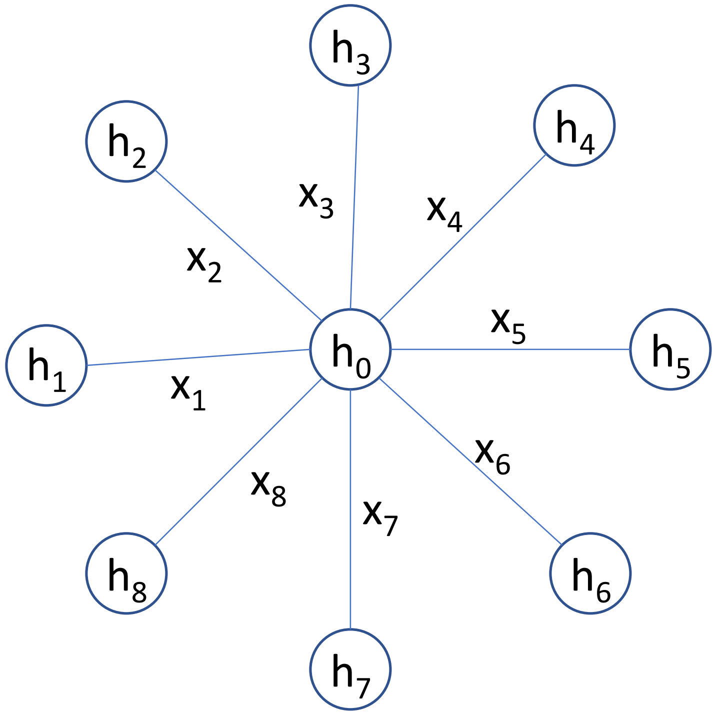
Distribution-Free Analysis

(Hanneke & Yang, 2015)

θ, φ, $\tilde{\theta}$ depend on f^*, P_X.

Can we do sample complexity analysis **without** distribution-dependence?

Definition: The **star number** $\mathfrak{s}$ is the largest k s.t. $\exists h_0, h_1, \ldots, h_k \in \mathcal{H}$, $\exists x_1, \ldots, x_k \in \mathcal{X}$ s.t. $\forall i \in \{1, \ldots, k\}$, $\{x_j : h_i(x_j) \neq h_0(x_j)\} = \{x_i\}$.



Distribution-Free Analysis

(Hanneke & Yang, 2015)

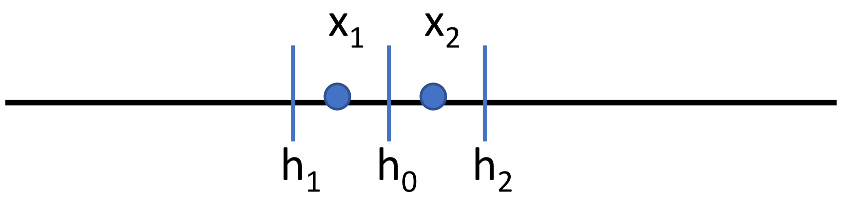
θ, φ, $\tilde{\theta}$ depend on f^*, P_X.

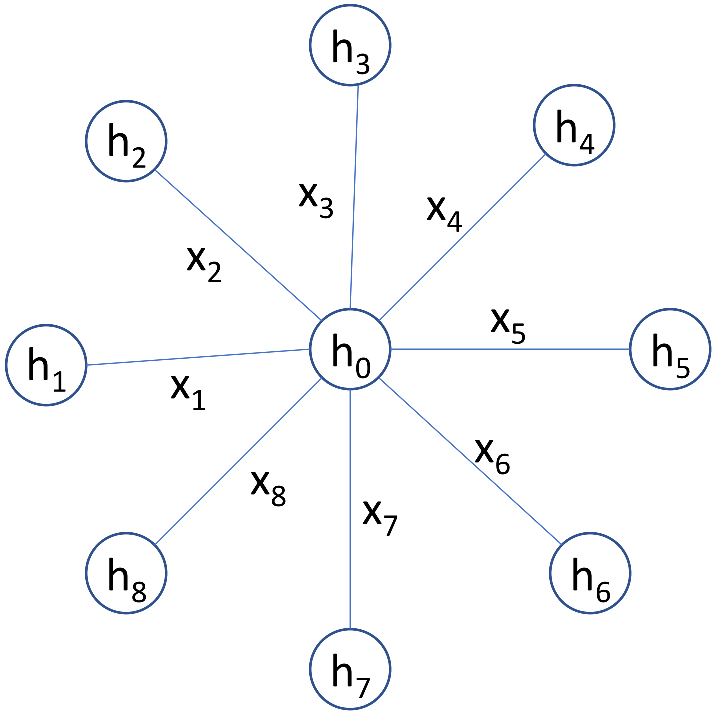
Can we do sample complexity analysis **without** distribution-dependence?

Definition: The **star number** $\mathfrak{s}$ is the largest k s.t. $\exists h_0, h_1, \ldots, h_k \in \mathcal{H}$, $\exists x_1, \ldots, x_k \in \mathcal{X}$ s.t. $\forall i \in \{1, \ldots, k\}$, $\{x_j : h_i(x_j) \neq h_0(x_j)\} = \{x_i\}$.

Example: Thresholds: $f(x) = \mathbb{I}[x \geq t]$.

$\mathfrak{s} = 2$.





Distribution-Free Analysis

(Hanneke & Yang, 2015)

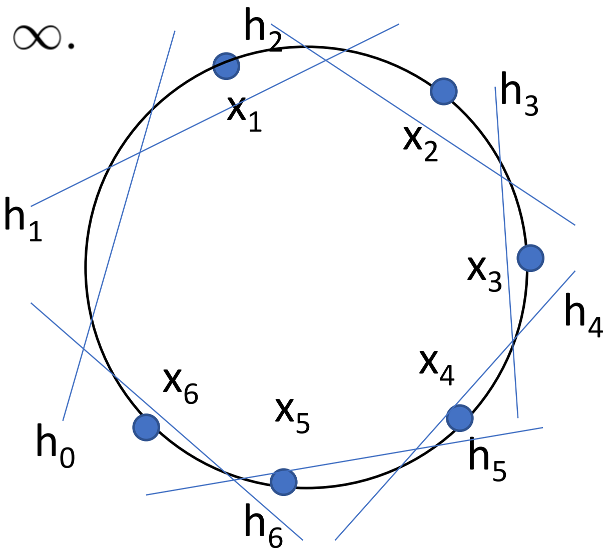
θ, φ, $\tilde{\theta}$ depend on f^*, P_X.

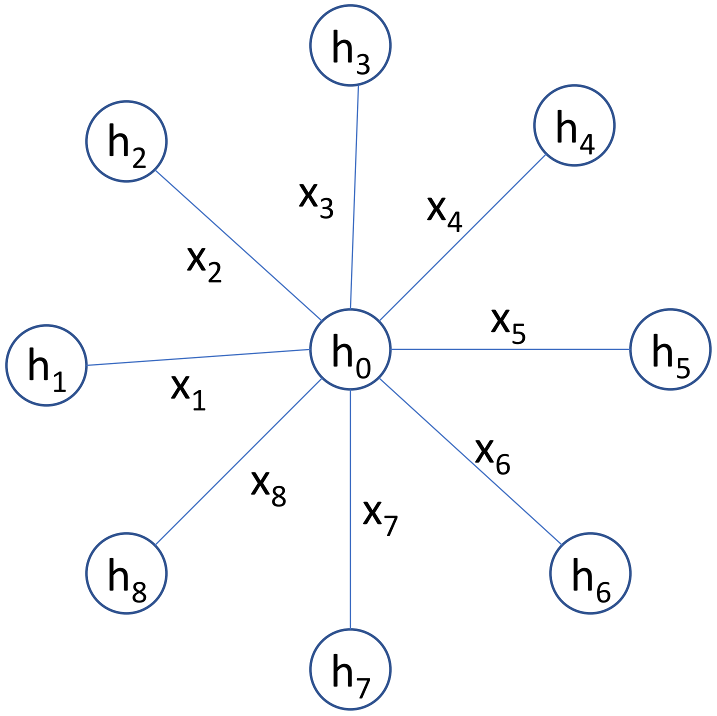
Can we do sample complexity analysis **without** distribution-dependence?

Definition: The **star number** $\mathfrak{s}$ is the largest k s.t. $\exists h_0, h_1, \ldots, h_k \in \mathcal{H}$, $\exists x_1, \ldots, x_k \in \mathcal{X}$ s.t. $\forall i \in \{1, \ldots, k\}$, $\{x_j : h_i(x_j) \neq h_0(x_j)\} = \{x_i\}$.

Example: Linear Separators in $\mathbb{R}^n$, $n \geq 2$:

$\mathfrak{s} = \infty$.





Distribution-Free Analysis

(Hanneke & Yang, 2015)

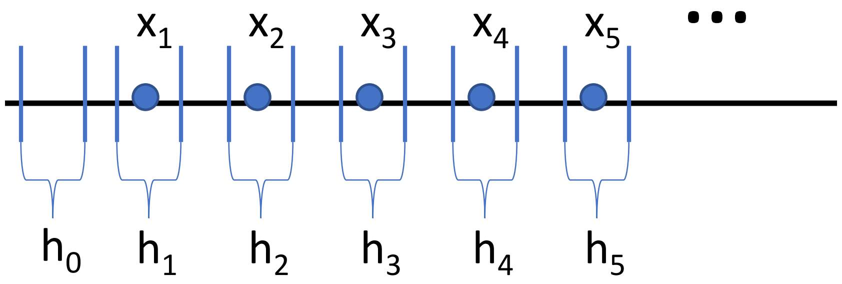
θ, φ, $\tilde{\theta}$ depend on f^*, P_X.

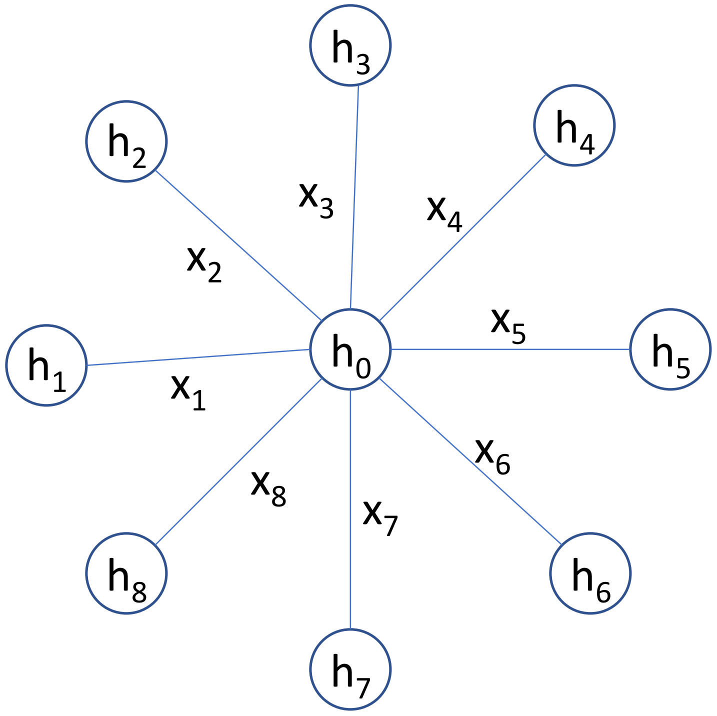
Can we do sample complexity analysis **without** distribution-dependence?

Definition: The **star number** $\mathfrak{s}$ is the largest k s.t. $\exists h_0, h_1, \ldots, h_k \in \mathcal{H}$, $\exists x_1, \ldots, x_k \in \mathcal{X}$ s.t. $\forall i \in \{1, \ldots, k\}$, $\{x_j : h_i(x_j) \neq h_0(x_j)\} = \{x_i\}$.

Example: Intervals: $x \mapsto \mathbb{I}[a \leq x \leq b]$

$\mathfrak{s} = \infty$.





Intervals of width w ($b - a = w > 0$) on $\mathcal{X} = [0,1]$: $\mathfrak{s} \approx \left\lfloor \frac{1}{w} \right\rfloor$.

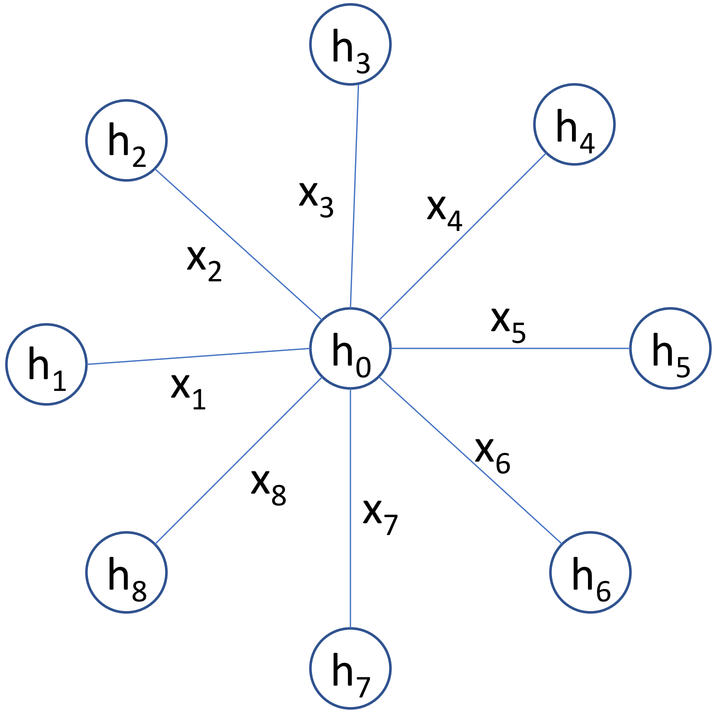
Distribution-Free Analysis

(Hanneke & Yang, 2015; Hanneke, 2016)

θ, φ, $\tilde{\theta}$ depend on f^*, P_X.

Can we do sample complexity analysis **without** distribution-dependence?

Definition: The **star number** $\mathfrak{s}$ is the largest k s.t. $\exists h_0, h_1, \ldots, h_k \in \mathcal{H}$, $\exists x_1, \ldots, x_k \in \mathcal{X}$ s.t. $\forall i \in \{1, \ldots, k\}$, $\{x_j : h_i(x_j) \neq h_0(x_j)\} = \{x_i\}$.



Theorem: $\sup_{P_X} \sup_{f^* \in \mathcal{H}} \theta = \sup_{P_X} \sup_{f^* \in \mathcal{H}} \varphi_c = \sup_{P_X} \sup_{f^* \in \mathcal{H}} \tilde{\theta} = \min\{\mathfrak{s}, \frac{1}{\epsilon}\} =: \mathfrak{s}_\epsilon$

Corollary:

Bounded noise # labels $\approx \mathfrak{s}_\epsilon d \log(\frac{1}{\epsilon})$

Agnostic ($\beta = R(f^*)$) # labels $\approx \mathfrak{s}_\beta d \frac{\beta^2}{\epsilon^2}$

Achieved by A^2

Distribution-Free Analysis

(Hanneke & Yang, 2015; Hanneke, 2016)

θ, φ, $\tilde{\theta}$ depend on f^*, P_X.

Can we do sample complexity analysis **without** distribution-dependence?

Definition: The **star number** $\mathfrak{s}$ is the largest k s.t. $\exists h_0, h_1, \ldots, h_k \in \mathcal{H}$, $\exists x_1, \ldots, x_k \in \mathcal{X}$ s.t. $\forall i \in \{1, \ldots, k\}$, $\{x_j : h_i(x_j) \neq h_0(x_j)\} = \{x_i\}$.

Theorem: $\sup_{P_X} \sup_{f^* \in \mathcal{H}} \theta = \sup_{P_X} \sup_{f^* \in \mathcal{H}} \varphi_c = \sup_{P_X} \sup_{f^* \in \mathcal{H}} \tilde{\theta} = \min\{\mathfrak{s}, \frac{1}{\epsilon}\} =: \mathfrak{s}_\epsilon$

Corollary:

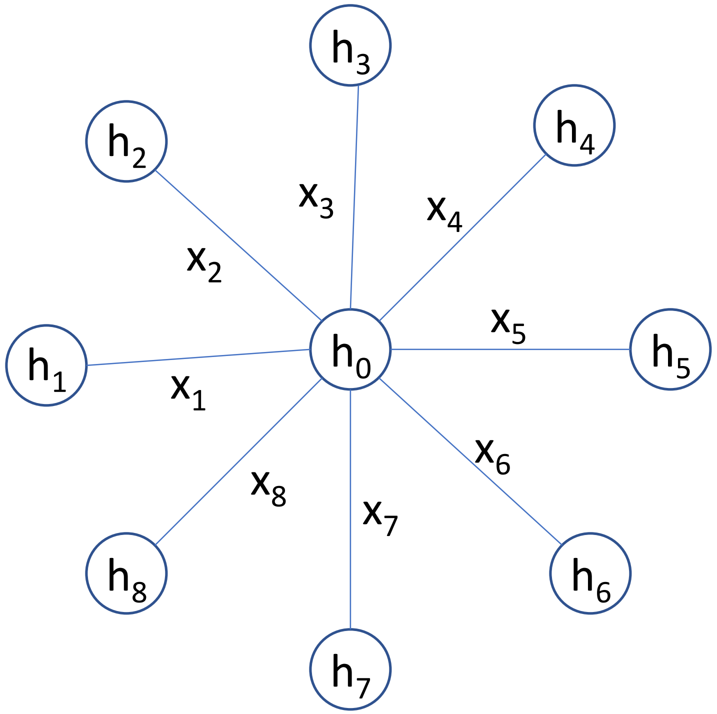
Bounded noise # labels $\approx \mathfrak{s}_\epsilon d \log(\frac{1}{\epsilon})$

Agnostic ($\beta = R(f^*)$) # labels $\approx \mathfrak{s}_\beta d \frac{\beta^2}{\epsilon^2}$

Achieved by A^2

Different alg., Bounded noise
labels $\approx \mathfrak{s}_{\epsilon/d} \log(\frac{1}{\epsilon})$

Near-matching **lower bound**:
$\mathfrak{s}_\epsilon + d \log(\frac{1}{\epsilon})$



Distribution-Free Analysis

(Hanneke & Yang, 2015; Hanneke, 2016)

θ, φ, $\tilde{\theta}$ depend on f^*, P_X.

Can we do sample complexity analysis **without** distribution-dependence?

Definition: The **star number** $\mathfrak{s}$ is the largest k s.t. $\exists h_0, h_1, \ldots, h_k \in \mathcal{H}$, $\exists x_1, \ldots, x_k \in \mathcal{X}$ s.t. $\forall i \in \{1, \ldots, k\}$, $\{x_j : h_i(x_j) \neq h_0(x_j)\} = \{x_i\}$.

Theorem: $\sup_{P_X} \sup_{f^* \in \mathcal{H}} \theta = \sup_{P_X} \sup_{f^* \in \mathcal{H}} \varphi_c = \sup_{P_X} \sup_{f^* \in \mathcal{H}} \tilde{\theta} = \min\{\mathfrak{s}, \frac{1}{\epsilon}\} =: \mathfrak{s}_\epsilon$

Corollary:

Bounded noise # labels $\approx \mathfrak{s}_\epsilon d \log(\frac{1}{\epsilon})$

Agnostic ($\beta = R(f^*)$) # labels $\approx \mathfrak{s}_\beta d \frac{\beta^2}{\epsilon^2}$

Achieved by A^2

Different alg., Bounded noise

labels $\approx \mathfrak{s}_{\epsilon/d} \log(\frac{1}{\epsilon})$

Near-matching **lower bound**:

$\mathfrak{s}_\epsilon + d \log(\frac{1}{\epsilon})$

Open Question:

Agnostic ($\beta = R(f^*)$)

labels

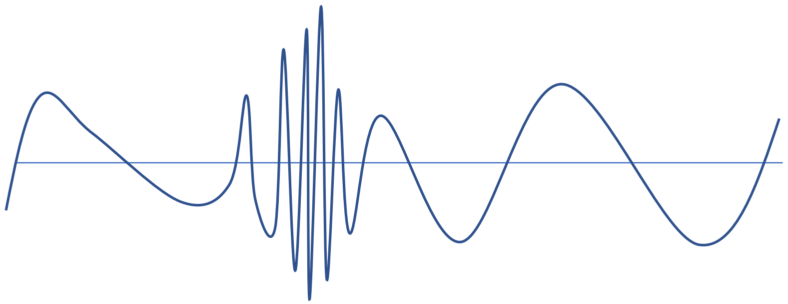
$\approx d\frac{\beta^2}{\epsilon^2} + \mathfrak{s}_{\epsilon/d} \log(\frac{1}{\epsilon})$ **?**

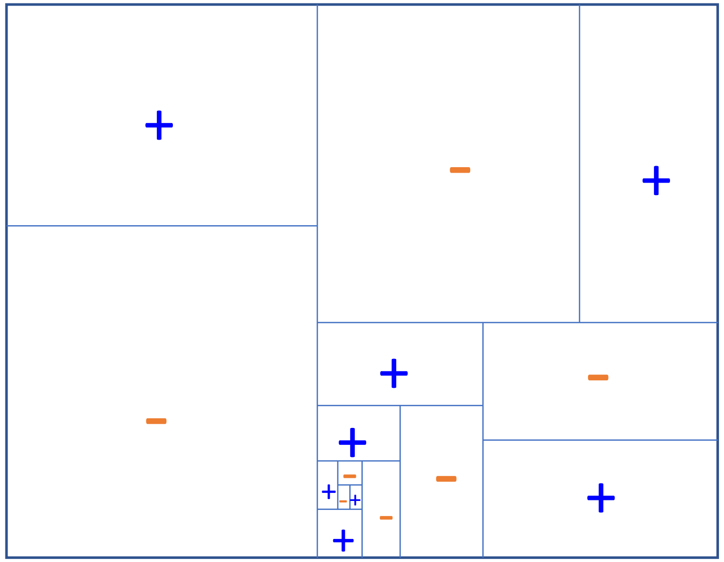
lower bound:

$d\frac{\beta^2}{\epsilon^2} + \mathfrak{s}_\epsilon + d \log(\frac{1}{\epsilon})$

Adapting to Heterogeneous Noise

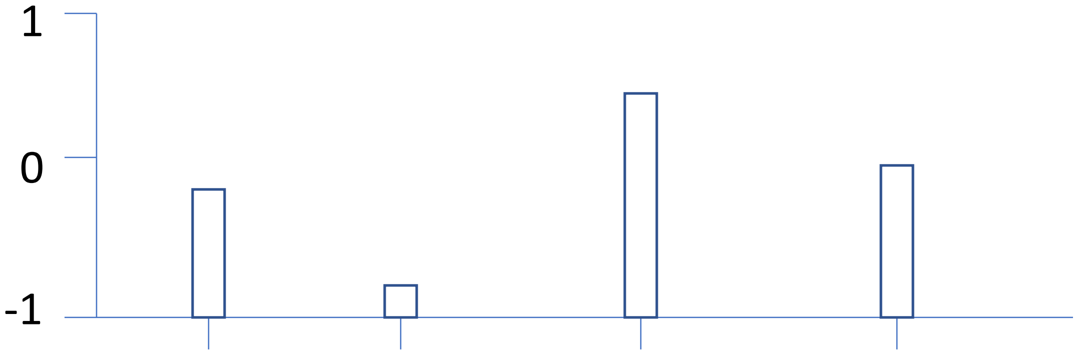
So far: Active learning for spatial heterogeneity of **opt function**:





Also consider: Spatial heterogeneity of **noise**:

$$\eta(x) := \mathbb{E}[Y|X=x]$$



Active Learning with TicToc

(Hanneke & Yang, 2015)

Algorithm: $\mathbb{A}(n)$
Input: Label budget n
Output: Classifier $\hat{f}_n$.

1. $\mathbb{L} \leftarrow \{\}$
2. For $m = 1, 2, \ldots$
3. $\quad X_{s_m} \leftarrow \textsc{GetSeed}(\mathbb{L}, m)$
4. $\quad \mathcal{L}_m \leftarrow \textsc{TicToc}(X_{s_m}, m)$
5. $\quad$ if $\mathcal{L}_m$ exists, $\mathbb{L} \leftarrow \mathbb{L} \cup \{(s_m, \mathcal{L}_m)\}$
6. $\quad$ If we've made n queries
7. $\qquad$ Return $\hat{f}_n \leftarrow \textsc{Learn}(\mathbb{L})$

An active learning alg. (e.g. A^2)

Main new part

A passive learning alg.

Active Learning with TicToc

(Hanneke & Yang, 2015)

Algorithm: $\mathbb{A}(n)$
Input: Label budget n
Output: Classifier $\hat{f}_n$.

1. $\mathbb{L} \leftarrow \{\}$
2. For $m = 1, 2, \ldots$
3. $\quad X_{s_m} \leftarrow \textsc{GetSeed}(\mathbb{L}, m)$
4. $\quad \mathcal{L}_m \leftarrow \textsc{TicToc}(X_{s_m}, m)$
5. $\quad$ if $\mathcal{L}_m$ exists, $\mathbb{L} \leftarrow \mathbb{L} \cup \{(s_m, \mathcal{L}_m)\}$
6. $\quad$ If we've made n queries
7. $\qquad$ Return $\hat{f}_n \leftarrow \textsc{Learn}(\mathbb{L})$

Denote $\eta(x) = \mathbb{E}[Y|X = x]$
Suppose f^* is the **global** optimal function: $f^*(x) = \text{sign}(\eta(x))$

$\textsc{TicToc}(\boldsymbol{X}, \boldsymbol{m})$:
Query X (or nearby) to try to guess $f^*(X)$
If can figure it out, return that label
If can't figure it out by $\boldsymbol{\tau_m}$ queries give up (don't return a label)

Focus queries on less-noisy points.

Double advantage:

- Focusing on the points we actually care about:

$$R(f|x) - R(f^\star|x) = |\eta(x)|\mathbb{I}[f(x) \neq f^\star(x)]$$

(small $|\eta(x)| \Rightarrow$ not much effect on $R(f|x)$ if $f(x) = f^*(x)$ or not).

- And those points require fewer queries to determine $f^\star(X_i)$!

$\sim \frac{1}{\eta(X_i)^2}$ queries
to determine $f^\star(X_i)$.

Active Learning with TicToc

(Hanneke & Yang, 2015)

Algorithm: $\mathbb{A}(n)$
Input: Label budget n
Output: Classifier $\hat{f}_n$.
1. $\mathbb{L} \leftarrow \{\}$
2. For $m = 1, 2, \ldots$
3. $\quad X_{s_m} \leftarrow \textsc{GetSeed}(\mathbb{L}, m)$
4. $\quad \mathcal{L}_m \leftarrow \textsc{TicToc}(X_{s_m}, m)$
5. $\quad$ if $\mathcal{L}_m$ exists, $\mathbb{L} \leftarrow \mathbb{L} \cup \{(s_m, \mathcal{L}_m)\}$
6. $\quad$ If we've made n queries
7. $\qquad$ Return $\hat{f}_n \leftarrow \textsc{Learn}(\mathbb{L})$

Theorem: Bounded noise: # labels $\approx \mathfrak{s}_{\epsilon/d} \log(\frac{1}{\epsilon})$

Denote $\eta(x) = \mathbb{E}[Y|X = x]$

Suppose f^* is the **global** optimal function: $f^*(x) = \text{sign}(\eta(x))$

$\textsc{TicToc}(\boldsymbol{X}, \boldsymbol{m})$:
Query X (or nearby) to try to guess $f^*(X)$
If can figure it out, return that label
If can't figure it out by $\boldsymbol{\tau_m}$ queries give up (don't return a label)

Focus queries on less-noisy points.

Double advantage:

- Focusing on the points we actually care about:

$$R(f|x) - R(f^\star|x) = |\eta(x)|\mathbb{I}[f(x) \neq f^\star(x)]$$

(small $|\eta(x)| \Rightarrow$ not much effect on $R(f|x)$ if $f(x) = f^*(x)$ or not).

- And those points require fewer queries to determine $f^\star(X_i)$!

$\sim \frac{1}{\eta(X_i)^2}$ queries to determine $f^\star(X_i)$.

Active Learning with TicToc

(Hanneke & Yang, 2015)

Algorithm: $\mathbb{A}(n)$
Input: Label budget n
Output: Classifier $\hat{f}_n$.
1. $\mathbb{L} \leftarrow \{\}$
2. For $m = 1, 2, \ldots$
3. $X_{s_m} \leftarrow$ GetSeed$(\mathbb{L}, m)$
4. $\mathcal{L}_m \leftarrow$ TicToc(X_{s_m}, m)
5. if $\mathcal{L}_m$ exists, $\mathbb{L} \leftarrow \mathbb{L} \cup \{(s_m, \mathcal{L}_m)\}$
6. If we've made n queries
7. Return $\hat{f}_n \leftarrow$ Learn$(\mathbb{L})$

Theorem: Agnostic ($\beta = R(f^*)$)
and suppose $f^* =$ global best:
\# labels

$$\approx d\frac{\beta^2}{\epsilon^2} + \mathfrak{s}_{\epsilon/d}\log(\tfrac{1}{\epsilon})$$

Confirms agnostic sample complexity conjecture but with extra assumption $f^* =$ global opt.

Near-match lower bound: $d\frac{\beta^2}{\epsilon^2} + \mathfrak{s}_\epsilon + d\log(\frac{1}{\epsilon})$

Denote $\eta(x) = \mathbb{E}[Y|X=x]$
Suppose f^* is the **global** optimal function: $f^*(x) = \text{sign}(\eta(x))$

TicToc(X, m):
Query X (or nearby) to try to guess $f^*(X)$
If can figure it out, return that label
If can't figure it out by τ_m queries give up (don't return a label)

Focus queries on less-noisy points.

Double advantage:

- Focusing on the points we actually care about:

  $$R(f|x) - R(f^\star|x) = |\eta(x)|\mathbb{I}[f(x) \neq f^\star(x)]$$

  (small $|\eta(x)| \Rightarrow$ not much effect on $R(f|x)$ if $f(x) = f^*(x)$ or not).

- And those points require fewer queries to determine $f^\star(X_i)$!

$\sim \frac{1}{\eta(X_i)^2}$ queries
to determine $f^\star(X_i)$.

Principles of Active Learning

1. Query in dense regions where $\hat{f}$ could disagree a lot with f^*

2. Query in regions with low noise

Tsybakov Noise

The alg. adapts to heterogeneity in the noise.

Let's try it with a model that explicitly describes heterogeneous noise:

Tsybakov Noise

Tsybakov Noise

(Tsybakov, 2004;
Mammen & Tsybakov 1999)

Denote $\eta(x) = \mathbb{E}[Y|X = x]$

Definition: (Tsybakov noise)
$f^\star(x) = \mathrm{sign}(\eta(x))$ and $\exists \alpha \in (0,1)$ s.t. $\forall \tau > 0$,

$$P_X(x : |\eta(x)| \leq \tau) \lesssim \tau^{\frac{\alpha}{1-\alpha}}.$$

Tsybakov Noise

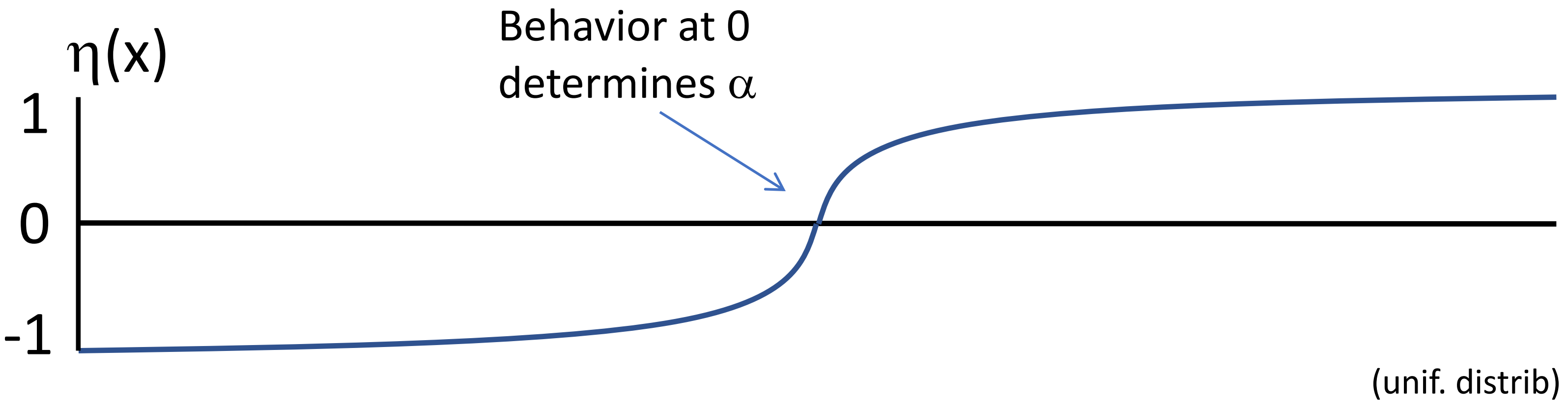
(Tsybakov, 2004;
Mammen & Tsybakov 1999)

Denote $\eta(x) = \mathbb{E}[Y|X = x]$

Definition: (Tsybakov noise)
$f^\star(x) = \mathrm{sign}(\eta(x))$ and $\exists \alpha \in (0,1)$ s.t. $\forall \tau > 0$,

$$P_X(x : |\eta(x)| \leq \tau) \lesssim \tau^{\frac{\alpha}{1-\alpha}}.$$

Example:
Thresholds



Tsybakov Noise

Denote $\eta(x) = \mathbb{E}[Y|X = x]$

Definition: (Tsybakov noise)
$f^\star(x) = \mathrm{sign}(\eta(x))$ and $\exists \alpha \in (0,1)$ s.t. $\forall \tau > 0$,

$$P_X(x : |\eta(x)| \leq \tau) \lesssim \tau^{\frac{\alpha}{1-\alpha}}.$$

Passive OPT: $\tilde{\Theta}\left(\frac{d}{\epsilon^{2-\alpha}}\right)$. (Massart & Nédélec, 2006)

Active OPT: $\begin{cases} \frac{d}{\epsilon^{2-2\alpha}} & \text{if } 0 < \alpha \leq 1/2 \\ \min\left\{\frac{d}{\epsilon^{2-2\alpha}}\left(\frac{\mathfrak{s}}{d}\right)^{2\alpha-1}, \frac{d}{\epsilon}\right\} & \text{if } 1/2 < \alpha < 1 \end{cases}.$ (Hanneke & Yang, 2015)
(roughly)

$$\boxed{\sim \begin{cases} \frac{1}{\varepsilon^{2-2\alpha}}, & \text{if } \mathfrak{s} < \infty \\ \frac{1}{\varepsilon}, & \text{if } \mathfrak{s} = \infty \end{cases}.}$$

Active Opt $\ll$ **Passive Opt.**
(always)

Conclusions

- Many proposals for going beyond Disagreement-based Active Learning

- Each exhibits improvements in certain cases

- We still don't know the **optimal agnostic active learning algorithm**

$$d\frac{\beta^2}{\epsilon^2} + \mathfrak{s}_{\epsilon/d} \log(\tfrac{1}{\epsilon})$$

Questions?

Further reading:

S. Dasgupta, A. Kalai, C. Monteleoni. Analysis of perceptron-based active learning. COLT 2005.

M. F. Balcan, A. Broder, T. Zhang. Margin based active learning. COLT 2007.

P. Awasthi, M. F. Balcan, P. Long. *Journal of the ACM*, 2017.

S. Hanneke. Theoretical Foundations of Active Learning. PhD Thesis, CMU, 2009.

S. Hanneke. Activized learning: Transforming passive to active with improved label complexity. *Journal of Machine Learning Research*, 2012.

C. Zhang, K. Chaudhuri. Beyond disagreement-based agnostic active learning. NeurIPS 2014.

R. M. Castro, R. D. Nowak. Minimax bounds for active learning. *IEEE Transactions on Information Theory*, 2008.

R. M. Castro, R.D. Nowak. Upper and lower error bounds for active learning. Allerton 2006.

S. Dasgupta. Coarse sample complexity bounds for active learning. NeurIPS 2005.

S. Hanneke, L. Yang. Minimax analysis of active learning. *Journal of Machine Learning Research*, 2015.

S. Hanneke. Refined error bounds for several learning algorithms. *Journal of Machine Learning Research*, 2016.

M. F. Balcan, S. Hanneke, J. Wortman Vaughan. The true sample complexity of active learning. *Machine Learning*, 2010.