Part 2: Theory of Active Learning General Case

- Disagreement-Based Agnostic Active Learning
- Disagreement Coefficient
- Sample Complexity Bounds

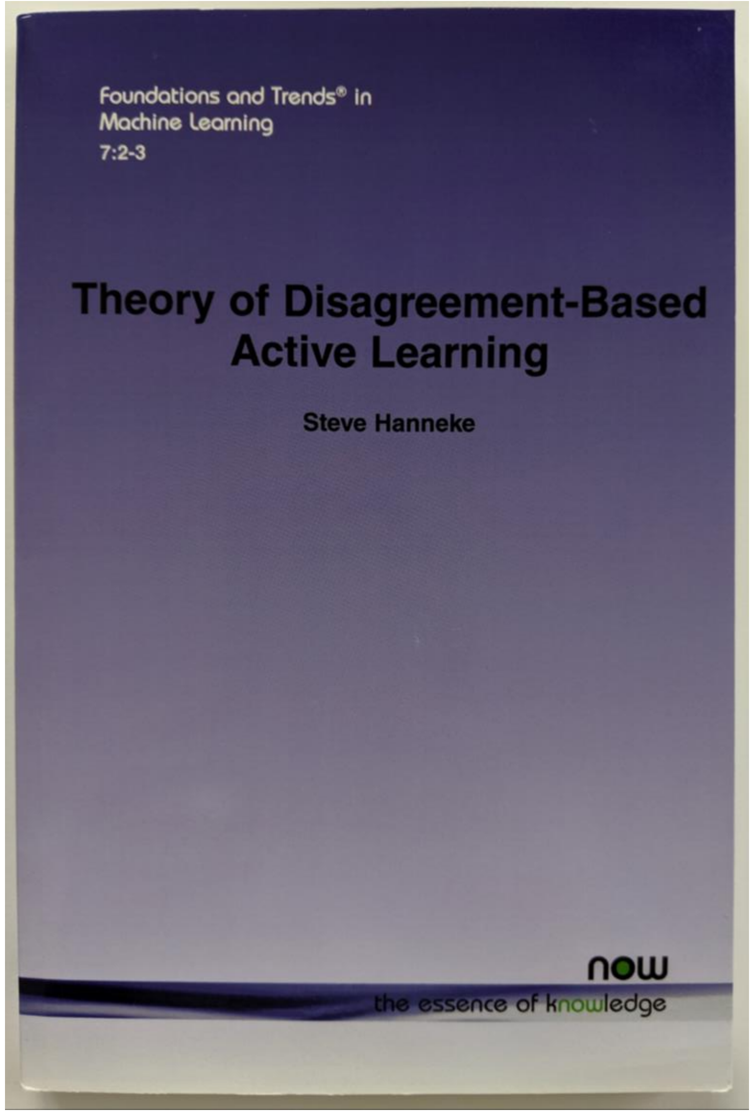
Tutorial on Active Learning: Theory to Practice

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

Agnostic Active Learning



Uniform Bernstein Inequality

Bernstein's inequality:

For m iid samples
$\forall f, f'$, w.p. $1-\delta$,
$$R(f) - R(f') \leq \hat{R}(f) - \hat{R}(f') + c\sqrt{\hat{P}(f \neq f')\frac{\log(1/\delta)}{m}} + \frac{\log(1/\delta)}{m}$$

Uniform Bernstein inequality:

VC dimension

w.p. $1-\delta$, $\forall f, f' \in \mathcal{H}$,
$$R(f) - R(f') \leq \hat{R}(f) - \hat{R}(f') + c\sqrt{\hat{P}(f \neq f')\frac{d\log(m/\delta)}{m}} + \frac{d\log(m/\delta)}{m}$$

Roughly:
$\forall f, f' \in \mathcal{H}$,
$$R(f) - R(f') \leq \hat{R}(f) - \hat{R}(f') + \sqrt{\hat{P}(f \neq f')\frac{d}{m}}$$

Agnostic Active Learning

Balcan, Beygelzimer, & Langford (2006)

Region of disagreement:

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

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 $\hat{f}$

Agnostic Active Learning

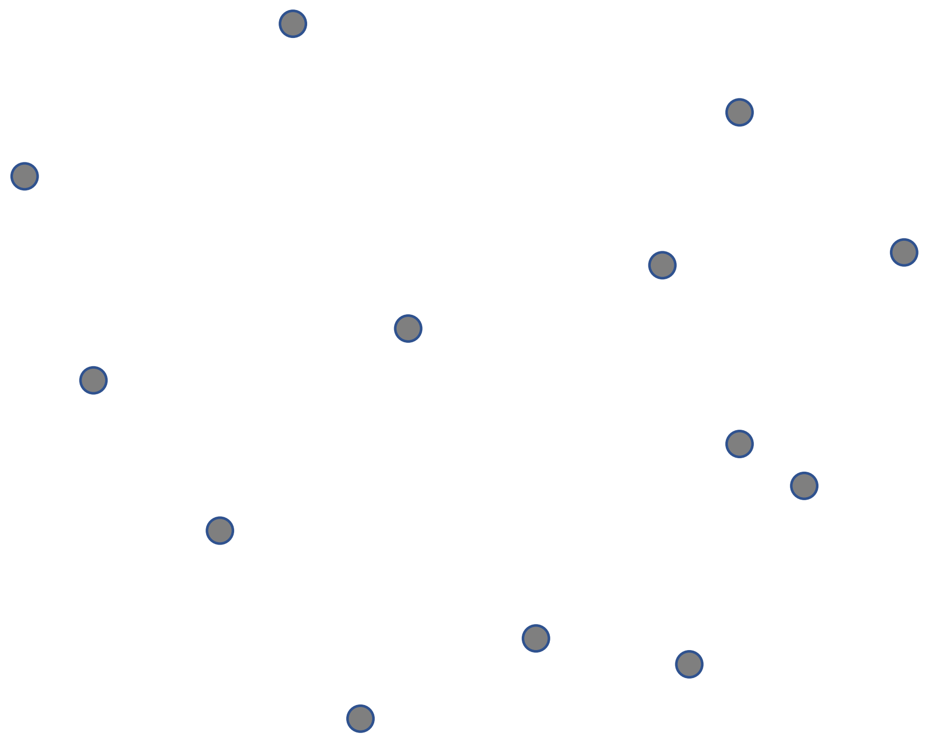
$$\mathrm{DIS}(\mathcal{H}) := \{x \in \mathcal{X} : \exists f, f' \in \mathcal{H}, f(x) \neq f'(x)\}$$

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 $\hat{f}$



Agnostic Active Learning

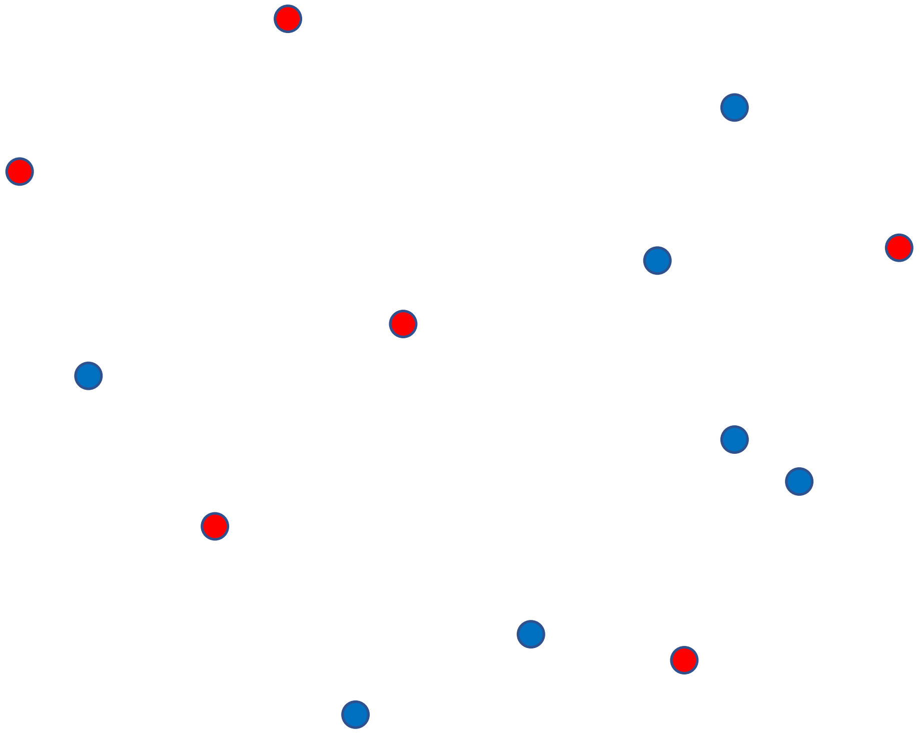
$$\mathrm{DIS}(\mathcal{H}) := \{x \in \mathcal{X} : \exists f, f' \in \mathcal{H}, f(x) \neq f'(x)\}$$

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 $\hat{f}$



Agnostic Active Learning

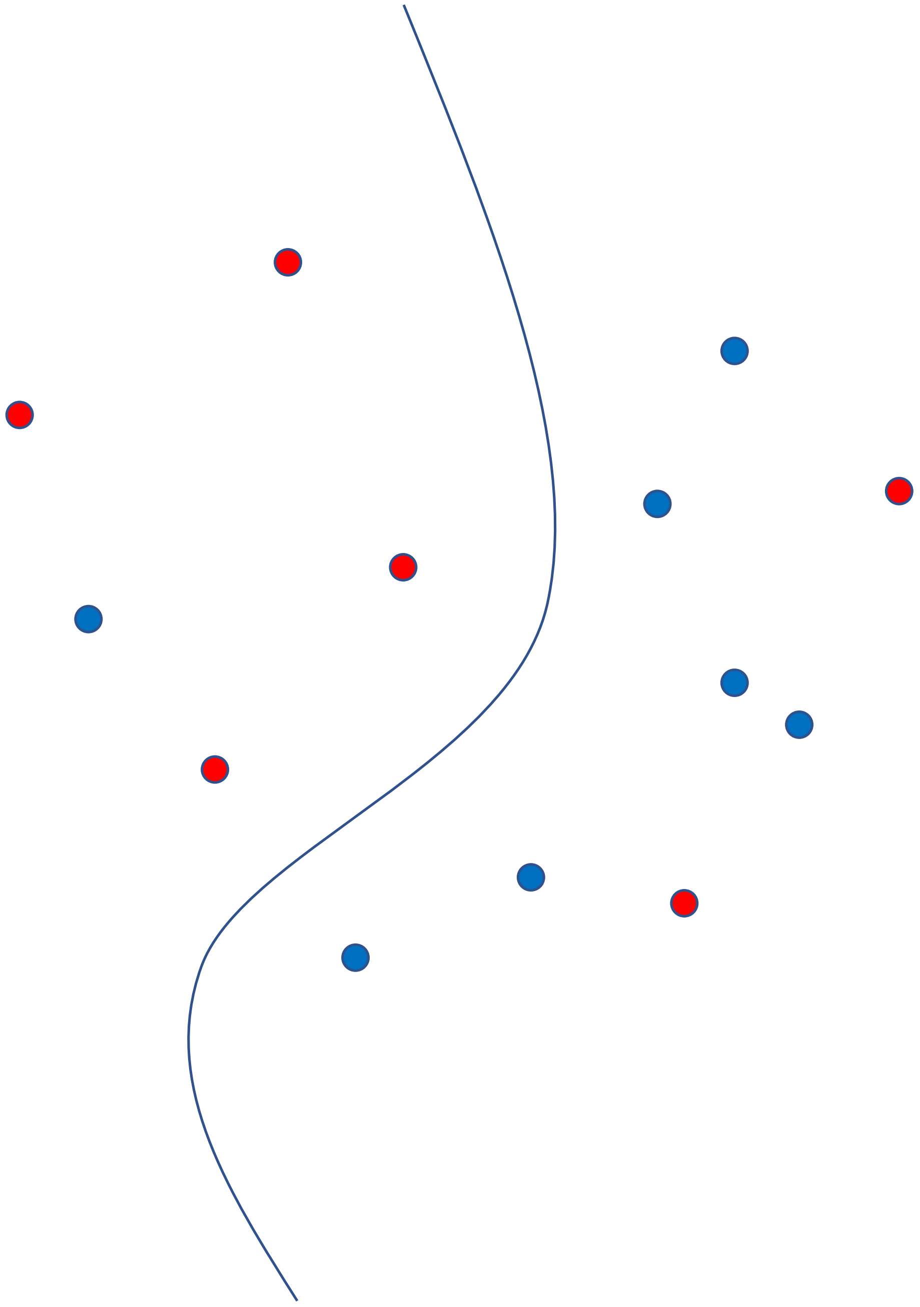
$$\mathrm{DIS}(\mathcal{H}) := \{x \in \mathcal{X} : \exists f, f' \in \mathcal{H}, f(x) \neq f'(x)\}$$

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 $\hat{f}$



Agnostic Active Learning

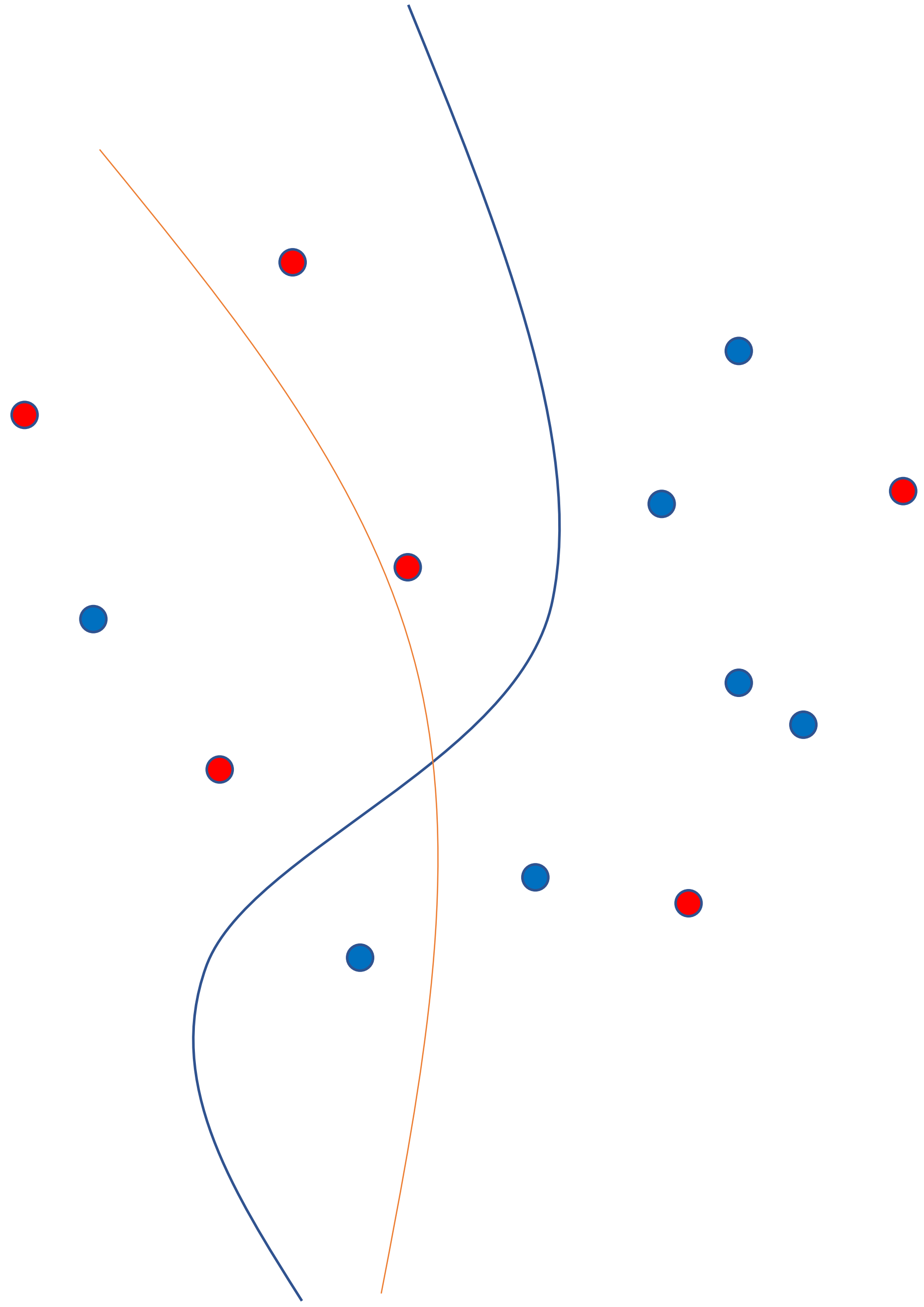
$$\mathrm{DIS}(\mathcal{H}) := \{x \in \mathcal{X} : \exists f, f' \in \mathcal{H}, f(x) \neq f'(x)\}$$

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 $\hat{f}$



Agnostic Active Learning

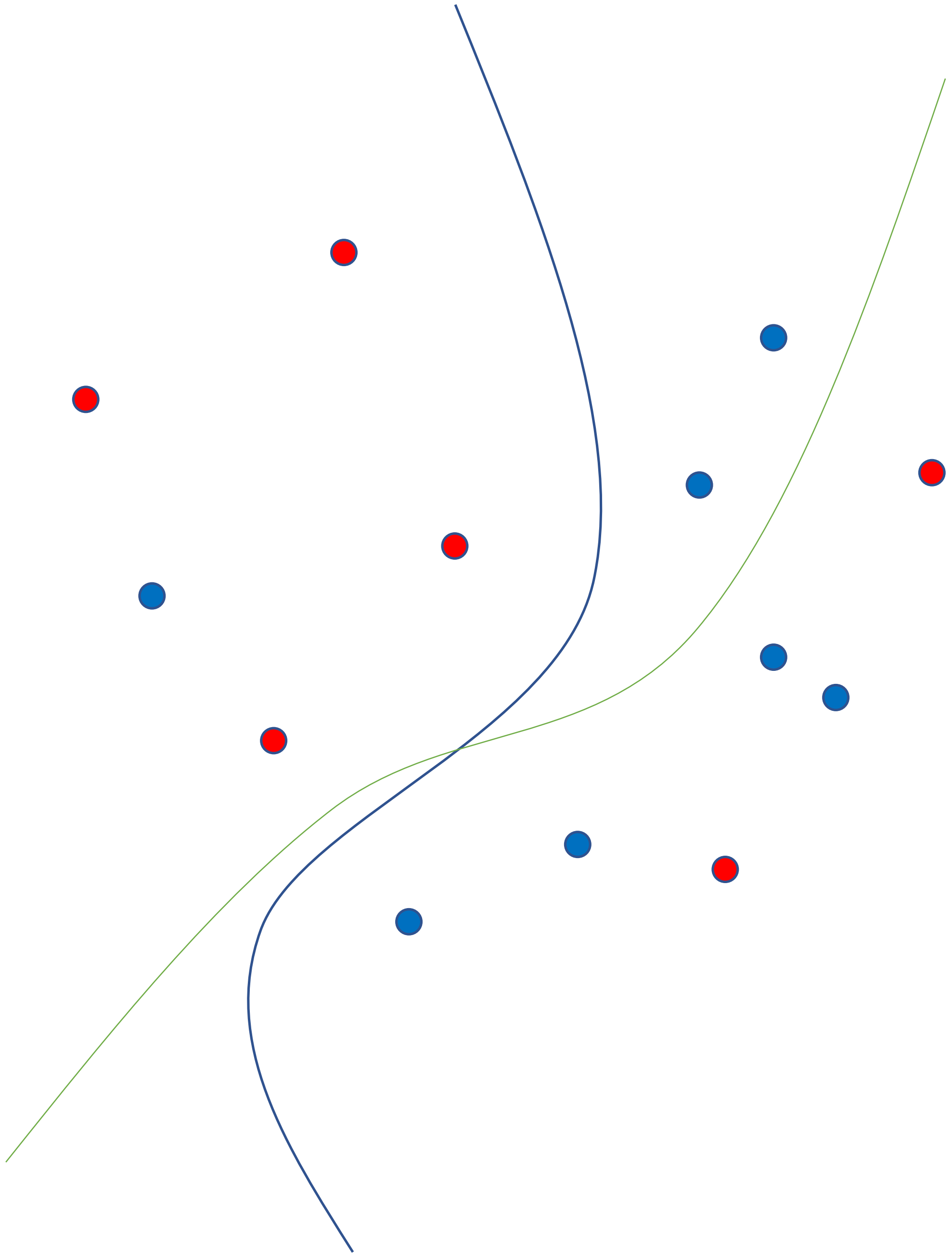
$$\mathrm{DIS}(\mathcal{H}) := \{x \in \mathcal{X} : \exists f, f' \in \mathcal{H}, f(x) \neq f'(x)\}$$

***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** $\hat{f} \leftarrow \underset{f \in \mathcal{H}}{\operatorname{argmin}} \hat{R}_Q(f)$
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 $\hat{f}$



Agnostic Active Learning

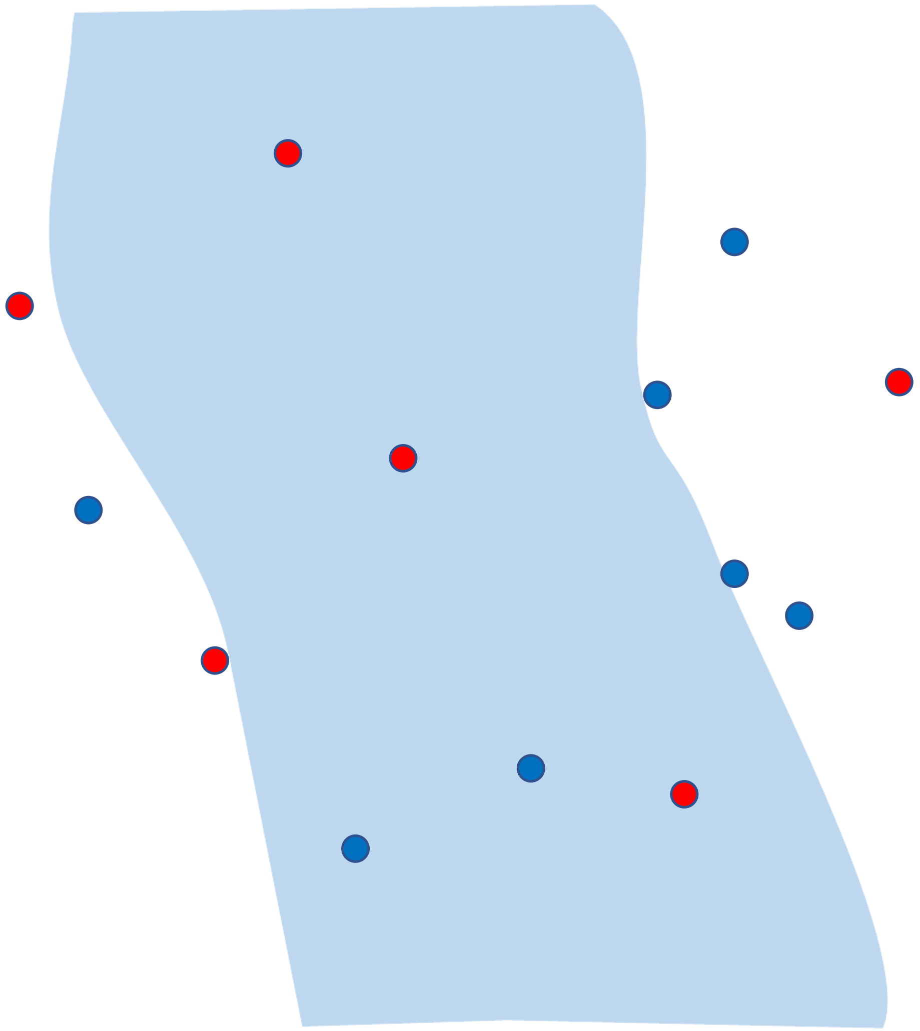
$$\mathrm{DIS}(\mathcal{H}) := \{x \in \mathcal{X} : \exists f, f' \in \mathcal{H}, f(x) \neq f'(x)\}$$

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** $\hat{f} \leftarrow \underset{f \in \mathcal{H}}{\operatorname{argmin}} \hat{R}_Q(f)$
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 $\hat{f}$



Agnostic Active Learning

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

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** $\hat{f} \leftarrow \underset{f \in \mathcal{H}}{\operatorname{argmin}} \hat{R}_Q(f)$
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 $\hat{f}$

Agnostic Active Learning

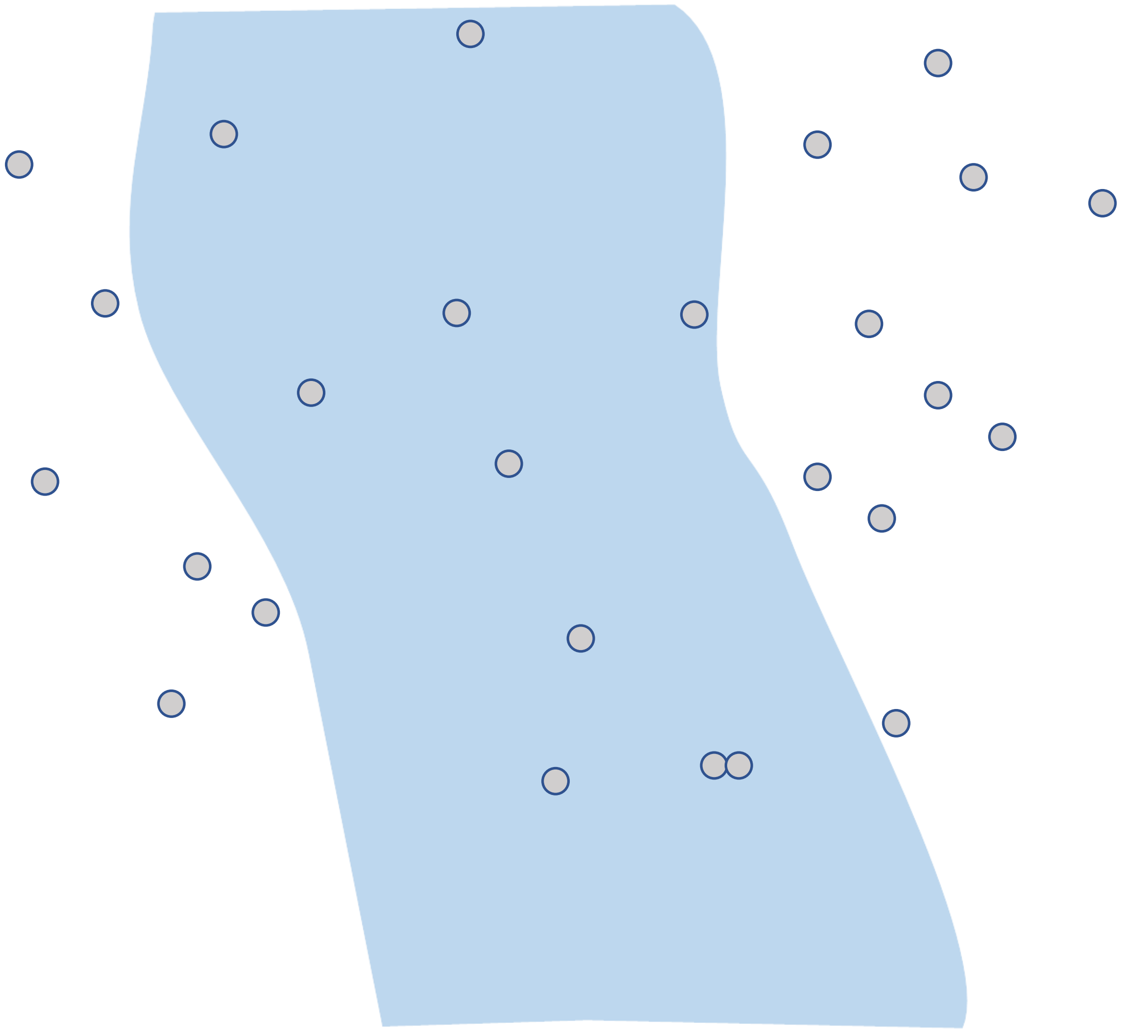
$$\mathrm{DIS}(\mathcal{H}) := \{x \in \mathcal{X} : \exists f, f' \in \mathcal{H}, f(x) \neq f'(x)\}$$

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** $\hat{f} \leftarrow \underset{f \in \mathcal{H}}{\operatorname{argmin}} \hat{R}_Q(f)$
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 $\hat{f}$



Agnostic Active Learning

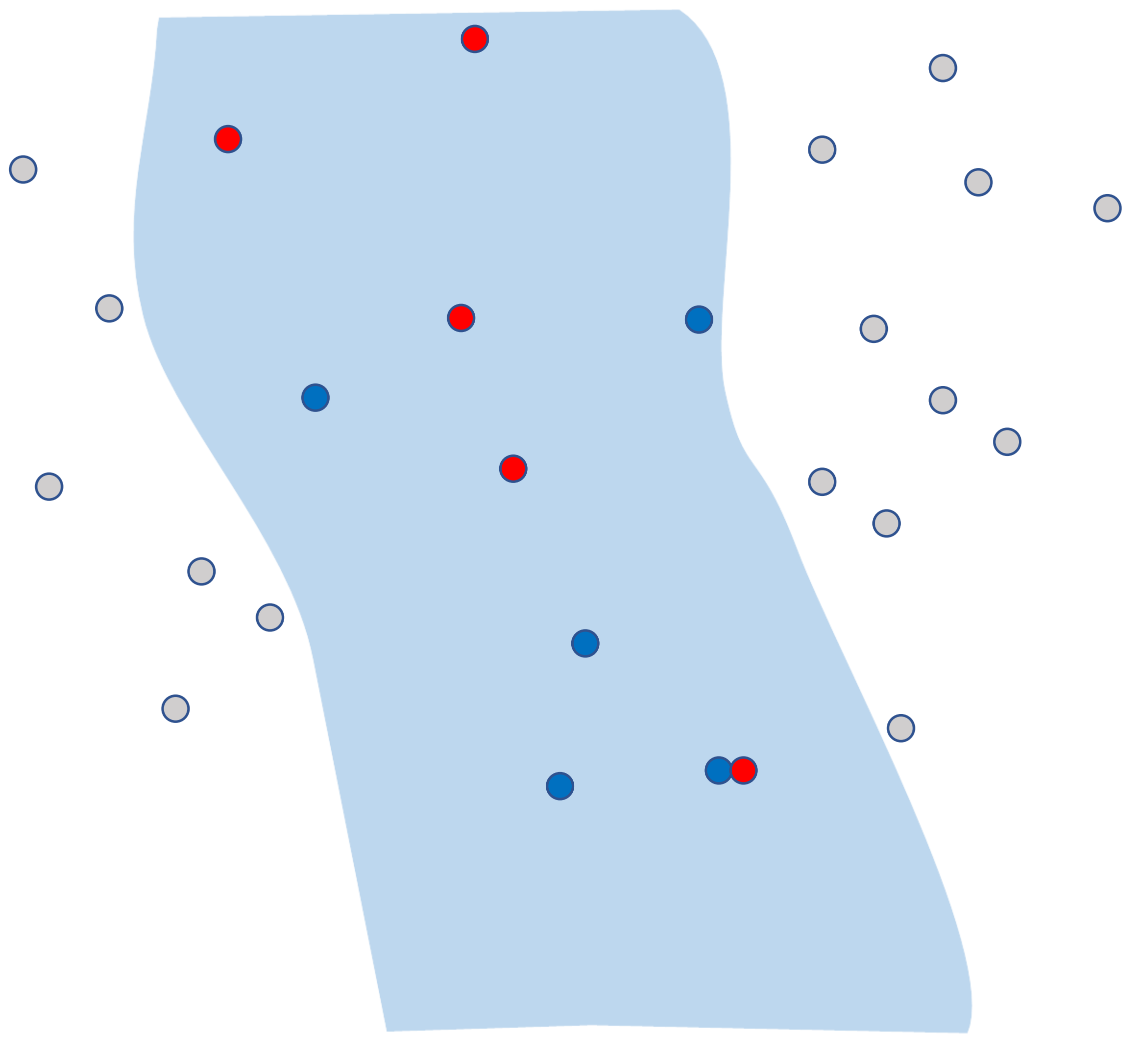
$$\mathrm{DIS}(\mathcal{H}) := \{x \in \mathcal{X} : \exists f, f' \in \mathcal{H}, f(x) \neq f'(x)\}$$

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 $\hat{f}$



Agnostic Active Learning

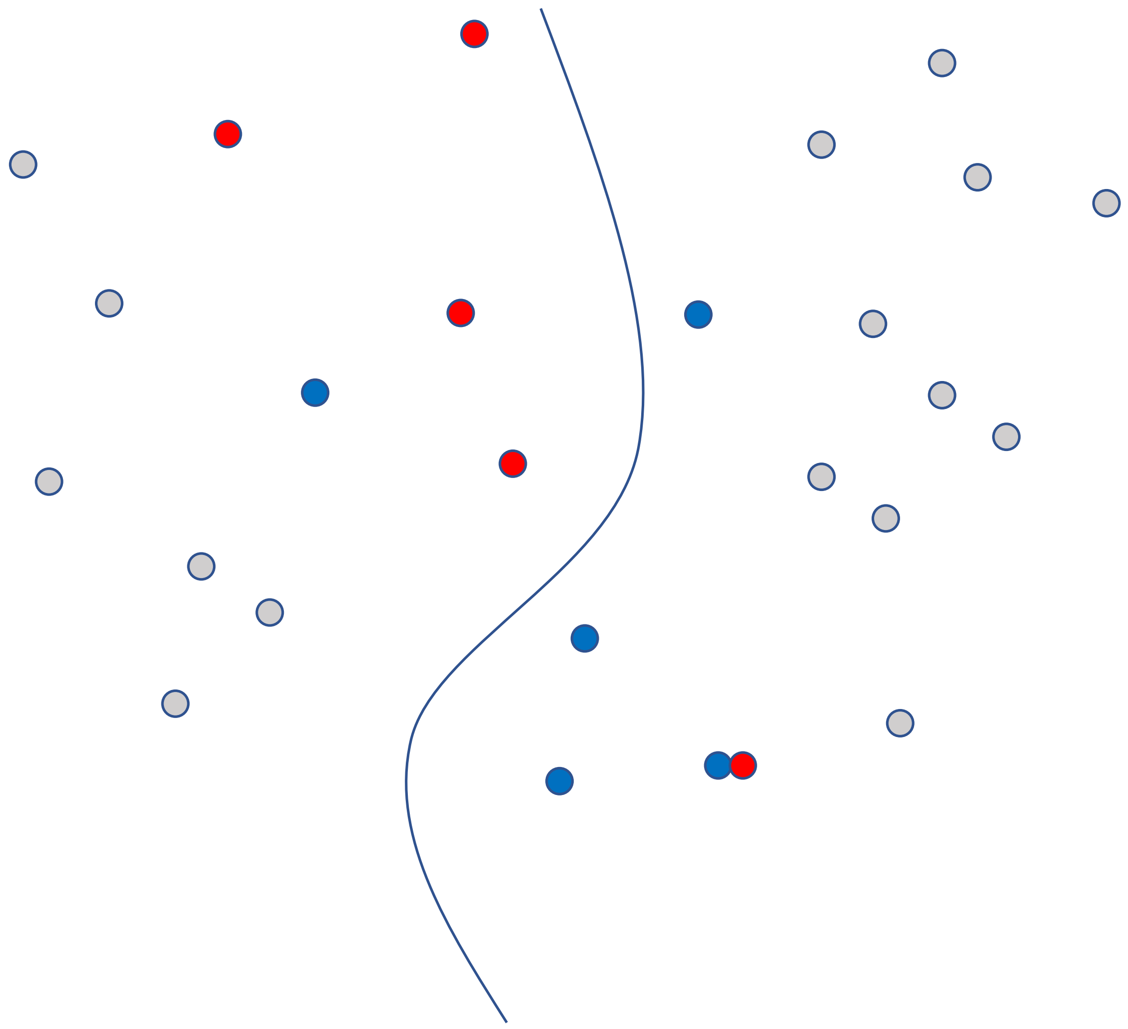
$$\mathrm{DIS}(\mathcal{H}) := \{x \in \mathcal{X} : \exists f, f' \in \mathcal{H}, f(x) \neq f'(x)\}$$

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 $\hat{f}$



Agnostic Active Learning

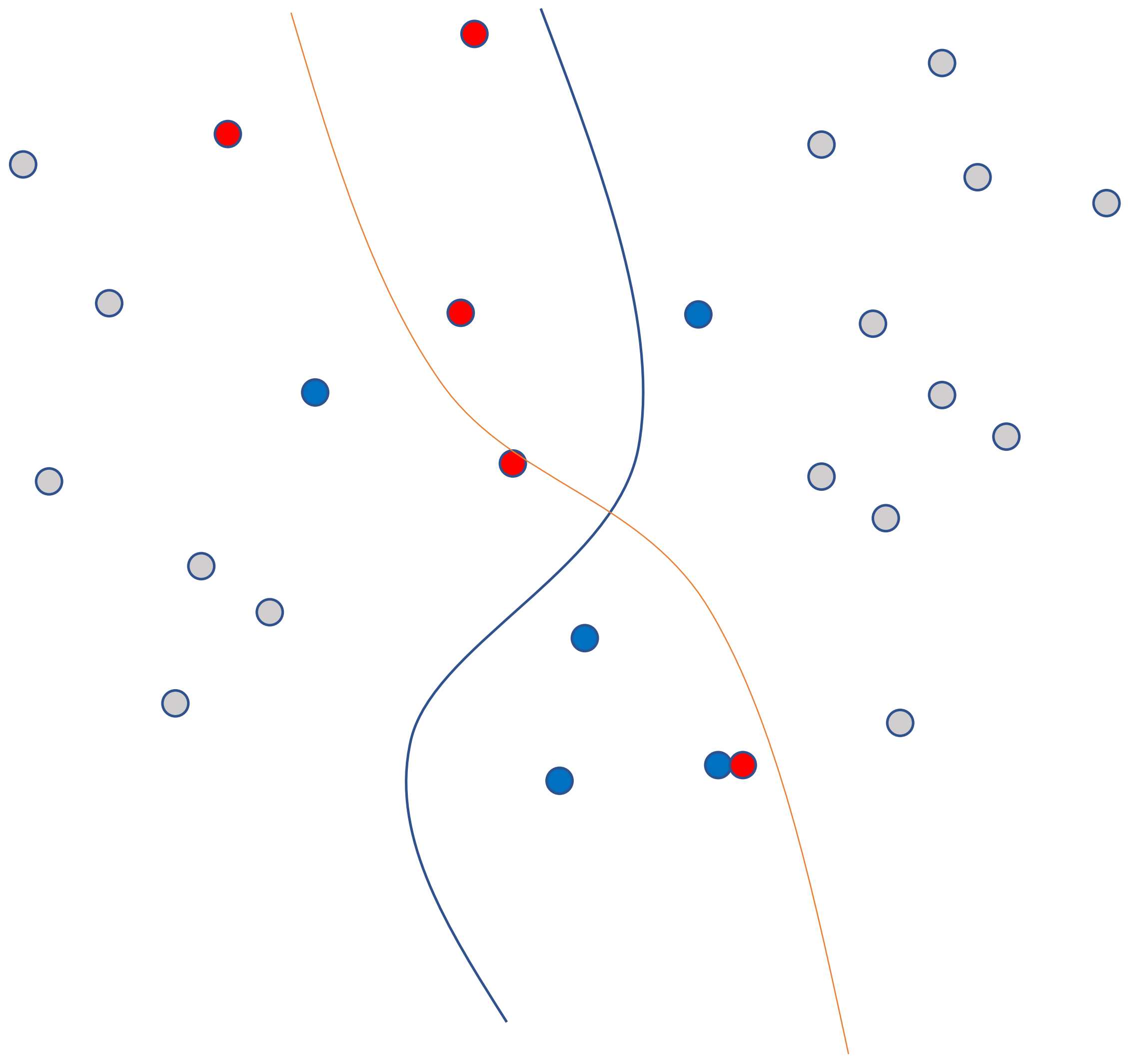
$$\mathrm{DIS}(\mathcal{H}) := \{x \in \mathcal{X} : \exists f, f' \in \mathcal{H}, f(x) \neq f'(x)\}$$

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** $\hat{f} \leftarrow \underset{f \in \mathcal{H}}{\operatorname{argmin}}\, \hat{R}_Q(f)$
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 $\hat{f}$



Agnostic Active Learning

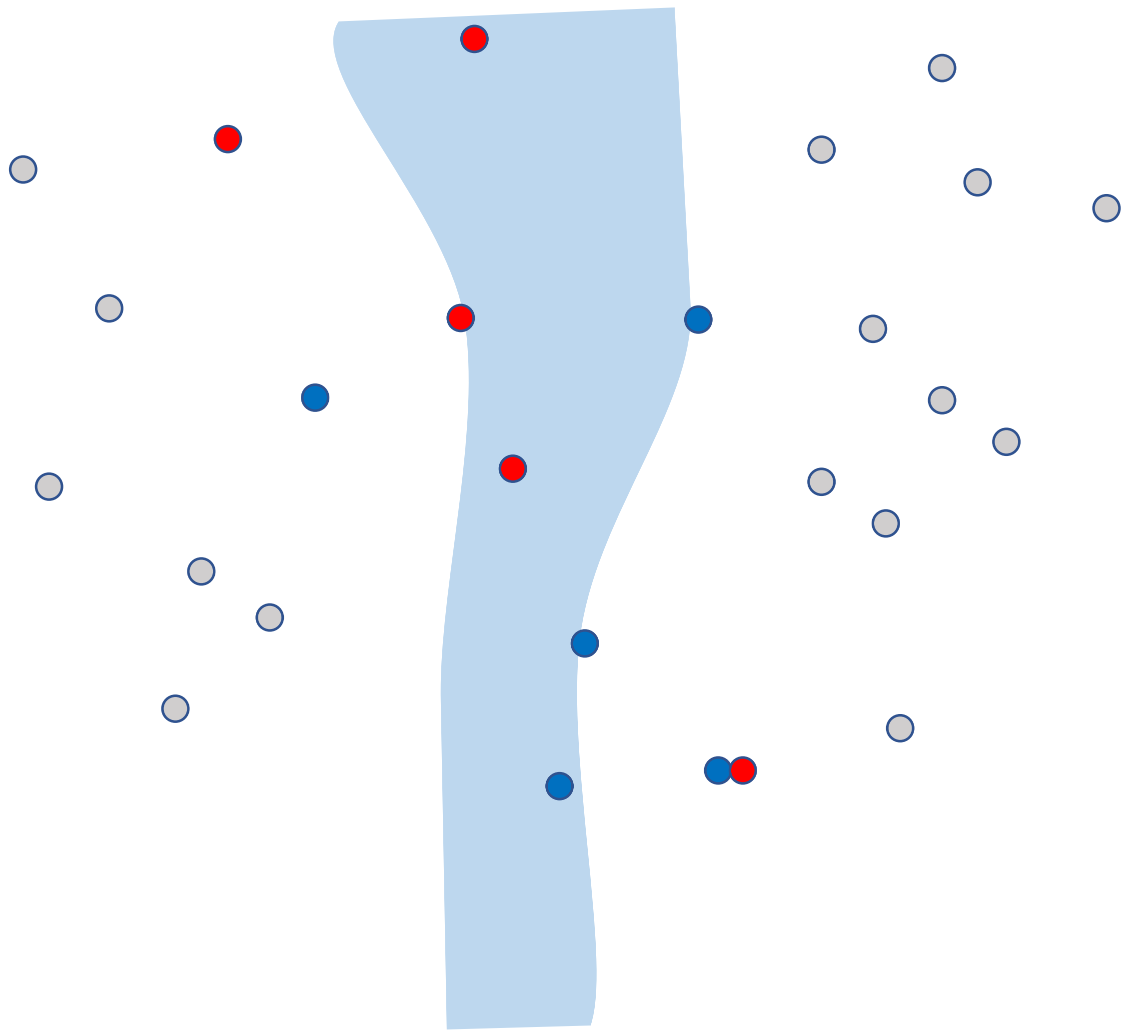
$$\mathrm{DIS}(\mathcal{H}) := \{x \in \mathcal{X} : \exists f, f' \in \mathcal{H}, f(x) \neq f'(x)\}$$

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 $\hat{f}$



Agnostic Active Learning

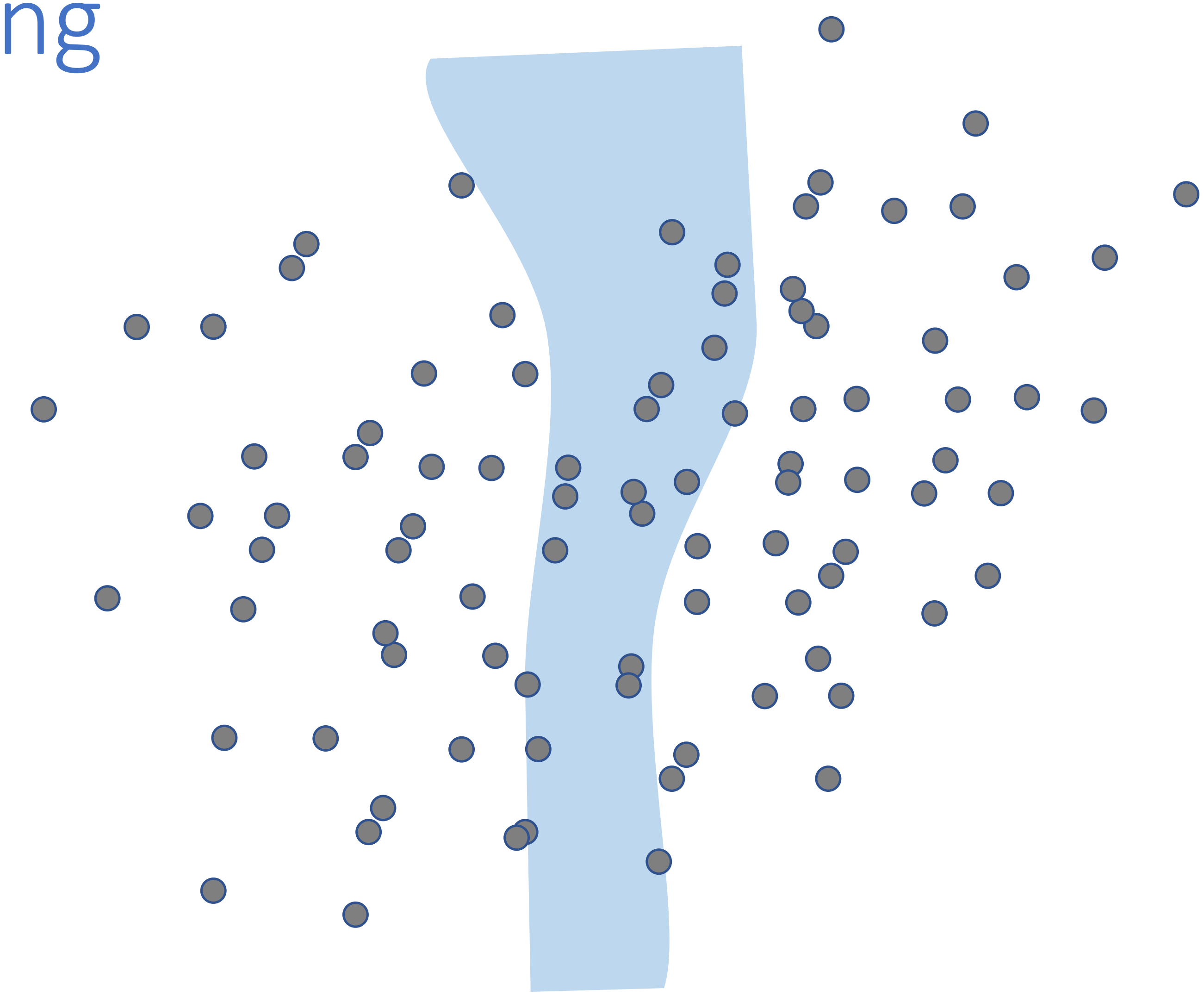
$$\text{DIS}(\mathcal{H}) := \{x \in \mathcal{X} : \exists f, f' \in \mathcal{H}, f(x) \neq f'(x)\}$$

A^2 (Agnostic Active)

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

1. **sample** 2^t unlabeled points S
2. **label** points in $Q = \text{DIS}(\mathcal{H}) \cap S$
3. **optimize** $\hat{f} \leftarrow \underset{f \in \mathcal{H}}{\text{argmin}}\, \hat{R}_Q(f)$
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 $\hat{f}$



Agnostic Active Learning

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

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** $\hat{f} \leftarrow \underset{f \in \mathcal{H}}{\operatorname{argmin}}\, \hat{R}_Q(f)$
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 $\hat{f}$

The point:

Any t with $f^* \in \mathcal{H}$ still,
$R(f^*|\mathrm{DIS}(\mathcal{H}))$ still **minimal** in $\mathcal{H}$

$\Rightarrow$

$$\hat{R}_Q(f^*) - \hat{R}_Q(\hat{f})$$
$$\leq R(f^*|\mathrm{DIS}(\mathcal{H})) - R(\hat{f}|\mathrm{DIS}(\mathcal{H})) + \sqrt{\hat{P}_Q(f^* \neq \hat{f}) \frac{d}{|Q|}}$$
$$\leq \sqrt{\hat{P}_Q(f^* \neq \hat{f}) \frac{d}{|Q|}}$$

$\Rightarrow$ **f^* never removed.**

Agnostic Active Learning

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

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** $\hat{f} \leftarrow \underset{f \in \mathcal{H}}{\operatorname{argmin}} \hat{R}_Q(f)$
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 $\hat{f}$

The point:

Any t with $f^* \in \mathcal{H}$ still,

$R(f^* | \mathrm{DIS}(\mathcal{H}))$ still **minimal** in $\mathcal{H}$

$\Rightarrow$

$$\hat{R}_Q(f^*) - \hat{R}_Q(\hat{f})$$
$$\leq R(f^* | \mathrm{DIS}(\mathcal{H})) - R(\hat{f} | \mathrm{DIS}(\mathcal{H})) + \sqrt{\hat{P}_Q(f^* \neq \hat{f}) \frac{d}{|Q|}}$$
$$\leq \sqrt{\hat{P}_Q(f^* \neq \hat{f}) \frac{d}{|Q|}}$$

$\Rightarrow$ **f^* never removed.**

Next: **How many labels does it use?**

Sample Complexity Analysis

Hanneke (2007,…)

Ball: $\mathrm{B}(f^*, r) := \{f \in \mathcal{H} : P_X(f \neq f^*) \leq r\}$

$\mathrm{DIS}(\mathrm{B}(f^*, r)) := \{x \in \mathcal{X} : \exists f, f' \in \mathrm{B}(f^*, r), f(x) \neq f'(x)\}$

Disagreement coefficient:

$$\theta = \sup_{r > \epsilon} \frac{P_X(\mathrm{DIS}(\mathrm{B}(f^*, r)))}{r}$$

Sample Complexity Analysis

Ball: $\mathrm{B}(f^*, r) := \{f \in \mathcal{H} : P_X(f \neq f^*) \leq r\}$

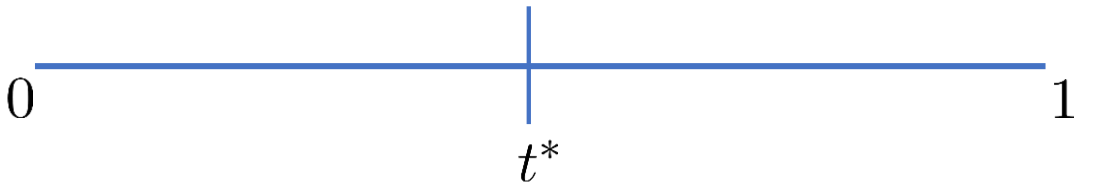
$\mathrm{DIS}(\mathrm{B}(f^*, r)) := \{x \in \mathcal{X} : \exists f, f' \in \mathrm{B}(f^*, r), f(x) \neq f'(x)\}$

Disagreement coefficient:

$$\theta = \sup_{r > \epsilon} \frac{P_X(\mathrm{DIS}(\mathrm{B}(f^*, r)))}{r}$$

Example: **Thresholds**, P_X Uniform$(0, 1)$

$f(x) = \mathbb{I}[x \geq t]$



Sample Complexity Analysis

Ball: $\mathrm{B}(f^*, r) := \{f \in \mathcal{H} : P_X(f \neq f^*) \leq r\}$

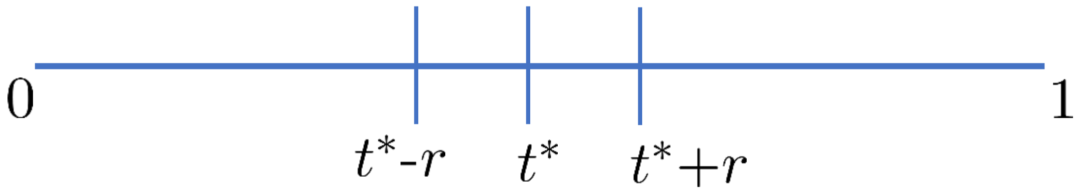
$\mathrm{DIS}(\mathrm{B}(f^*, r)) := \{x \in \mathcal{X} : \exists f, f' \in \mathrm{B}(f^*, r), f(x) \neq f'(x)\}$

Disagreement coefficient:

$$\theta = \sup_{r > \epsilon} \frac{P_X(\mathrm{DIS}(\mathrm{B}(f^*, r)))}{r}$$

Example: **Thresholds**, P_X Uniform$(0, 1)$

$f(x) = \mathbb{I}[x \geq t]$



$\mathrm{DIS}(\mathrm{B}(f^*, r)) = [t^* - r, t^* + r)$

$P_X(\mathrm{DIS}(\mathrm{B}(f^*, r))) = 2r$

$\theta = 2$

Sample Complexity Analysis

Ball: $\mathrm{B}(f^*, r) := \{f \in \mathcal{H} : P_X(f \neq f^*) \leq r\}$

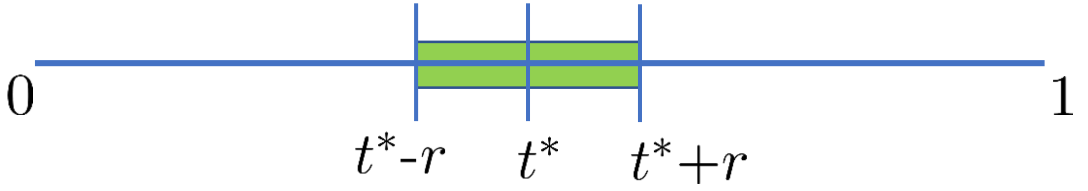
$\mathrm{DIS}(\mathrm{B}(f^*, r)) := \{x \in \mathcal{X} : \exists f, f' \in \mathrm{B}(f^*, r), f(x) \neq f'(x)\}$

Disagreement coefficient:

$$\theta = \sup_{r > \epsilon} \frac{P_X(\mathrm{DIS}(\mathrm{B}(f^*, r)))}{r}$$

Example: **Thresholds**, P_X Uniform$(0, 1)$

$f(x) = \mathbb{I}[x \geq t]$



$\mathrm{DIS}(\mathrm{B}(f^*, r)) = [t^* - r, t^* + r)$

$P_X(\mathrm{DIS}(\mathrm{B}(f^*, r))) = 2r$

$\Rightarrow \theta = 2$

Sample Complexity Analysis

Ball: $\mathrm{B}(f^*, r) := \{f \in \mathcal{H} : P_X(f \neq f^*) \leq r\}$

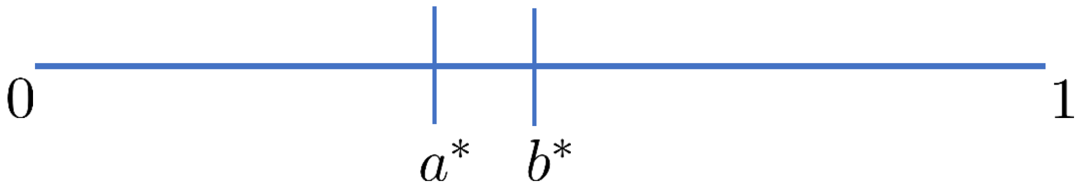
$\mathrm{DIS}(\mathrm{B}(f^*, r)) := \{x \in \mathcal{X} : \exists f, f' \in \mathrm{B}(f^*, r), f(x) \neq f'(x)\}$

Disagreement coefficient:

$$\theta = \sup_{r > \epsilon} \frac{P_X(\mathrm{DIS}(\mathrm{B}(f^*, r)))}{r}$$

Example: **Intervals**, P_X Uniform$(0, 1)$

$f(x) = \mathbb{I}[a \leq x \leq b]$



Sample Complexity Analysis

Ball: $\mathrm{B}(f^*, r) := \{f \in \mathcal{H} : P_X(f \neq f^*) \leq r\}$

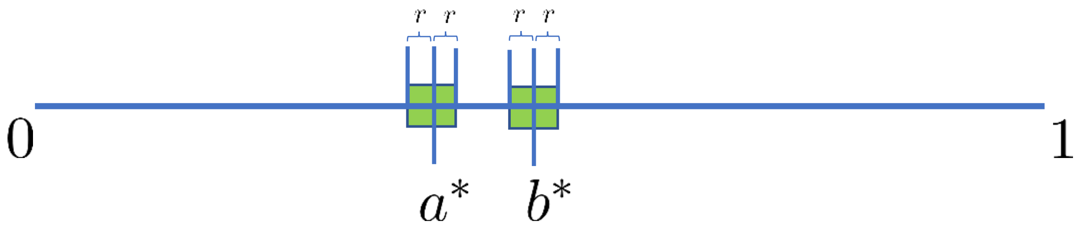
$\mathrm{DIS}(\mathrm{B}(f^*, r)) := \{x \in \mathcal{X} : \exists f, f' \in \mathrm{B}(f^*, r), f(x) \neq f'(x)\}$

Disagreement coefficient:

$$\theta = \sup_{r > \epsilon} \frac{P_X(\mathrm{DIS}(\mathrm{B}(f^*, r)))}{r}$$

Example: **Intervals**, P_X Uniform$(0,1)$

$f(x) = \mathbb{I}[a \leq x \leq b]$



$w^* := b^* - a^*$

If $\boldsymbol{r < w^*}$,

$\mathrm{DIS}(\mathrm{B}(f^*, r)) = [a^* - r, a^* + r) \cup (b^* - r, b^* + r]$

$P_X(\mathrm{DIS}(\mathrm{B}(f^*, r))) = 4r$

Sample Complexity Analysis

Ball: $\mathrm{B}(f^*, r) := \{f \in \mathcal{H} : P_X(f \neq f^*) \leq r\}$

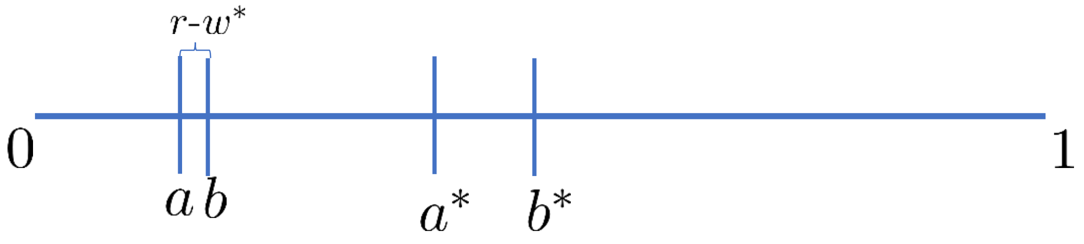
$\mathrm{DIS}(\mathrm{B}(f^*, r)) := \{x \in \mathcal{X} : \exists f, f' \in \mathrm{B}(f^*, r), f(x) \neq f'(x)\}$

Disagreement coefficient:

$$\theta = \sup_{r > \epsilon} \frac{P_X(\mathrm{DIS}(\mathrm{B}(f^*, r)))}{r}$$

Example: **Intervals**, P_X Uniform$(0,1)$

$f(x) = \mathbb{I}[a \leq x \leq b]$



$w^* := b^* - a^*$

If $\boldsymbol{r > w^*}$,

$\mathrm{DIS}(\mathrm{B}(f^*, r)) = \mathcal{X}$

$P_X(\mathrm{DIS}(\mathrm{B}(f^*, r))) = 1$

Sample Complexity Analysis

Ball: $\mathrm{B}(f^*, r) := \{f \in \mathcal{H} : P_X(f \neq f^*) \leq r\}$

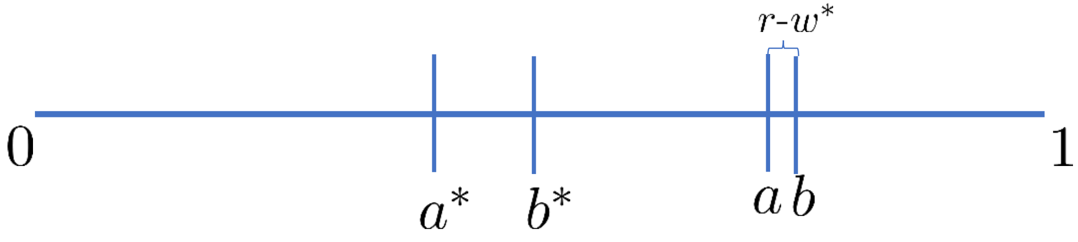
$\mathrm{DIS}(\mathrm{B}(f^*, r)) := \{x \in \mathcal{X} : \exists f, f' \in \mathrm{B}(f^*, r), f(x) \neq f'(x)\}$

Disagreement coefficient:

$$\theta = \sup_{r > \epsilon} \frac{P_X(\mathrm{DIS}(\mathrm{B}(f^*, r)))}{r}$$

Example: **Intervals**, P_X Uniform(0, 1)

$f(x) = \mathbb{I}[a \leq x \leq b]$



$w^* := b^* - a^*$

If $r > w^*$,

$\mathrm{DIS}(\mathrm{B}(f^*, r)) = \mathcal{X}$

$P_X(\mathrm{DIS}(\mathrm{B}(f^*, r))) = 1$

Sample Complexity Analysis

Ball: $\mathrm{B}(f^*, r) := \{f \in \mathcal{H} : P_X(f \neq f^*) \leq r\}$

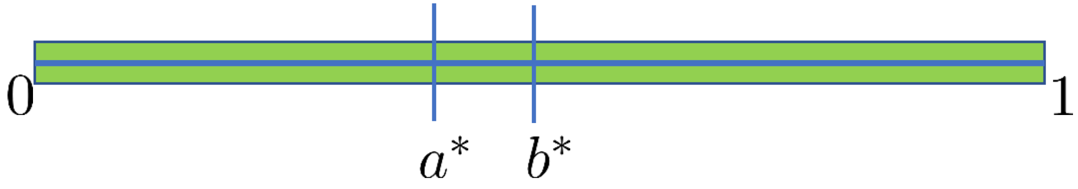
$\mathrm{DIS}(\mathrm{B}(f^*, r)) := \{x \in \mathcal{X} : \exists f, f' \in \mathrm{B}(f^*, r), f(x) \neq f'(x)\}$

Disagreement coefficient:

$$\theta = \sup_{r > \epsilon} \frac{P_X(\mathrm{DIS}(\mathrm{B}(f^*, r)))}{r}$$

Example: **Intervals**, P_X Uniform$(0, 1)$

$f(x) = \mathbb{I}[a \leq x \leq b]$



$w^* := b^* - a^*$

If $\boldsymbol{r > w^*}$,

$\mathrm{DIS}(\mathrm{B}(f^*, r)) = \mathcal{X}$

$P_X(\mathrm{DIS}(\mathrm{B}(f^*, r))) = 1$

Sample Complexity Analysis

Ball: $\mathrm{B}(f^*, r) := \{f \in \mathcal{H} : P_X(f \neq f^*) \leq r\}$

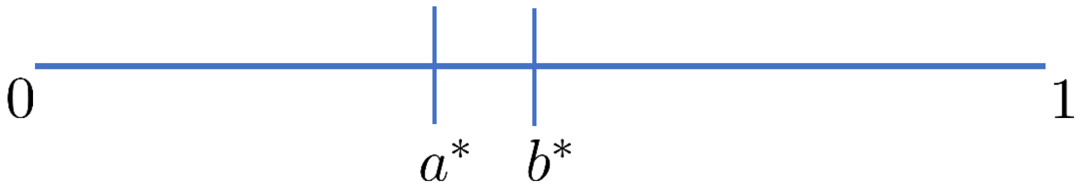
$\mathrm{DIS}(\mathrm{B}(f^*, r)) := \{x \in \mathcal{X} : \exists f, f' \in \mathrm{B}(f^*, r), f(x) \neq f'(x)\}$

Disagreement coefficient:

$$\theta = \sup_{r > \epsilon} \frac{P_X(\mathrm{DIS}(\mathrm{B}(f^*, r)))}{r}$$

Example: **Intervals**, P_X Uniform$(0, 1)$

$f(x) = \mathbb{I}[a \leq x \leq b]$



$w^* := b^* - a^*$

If $r < w^*$, $P_X(\mathrm{DIS}(\mathrm{B}(f^*, r))) = 4r$

If $r > w^*$, $P_X(\mathrm{DIS}(\mathrm{B}(f^*, r))) = 1$

$\Rightarrow \theta \leq \max\{4, \frac{1}{w^*}\}$

Sample Complexity Analysis

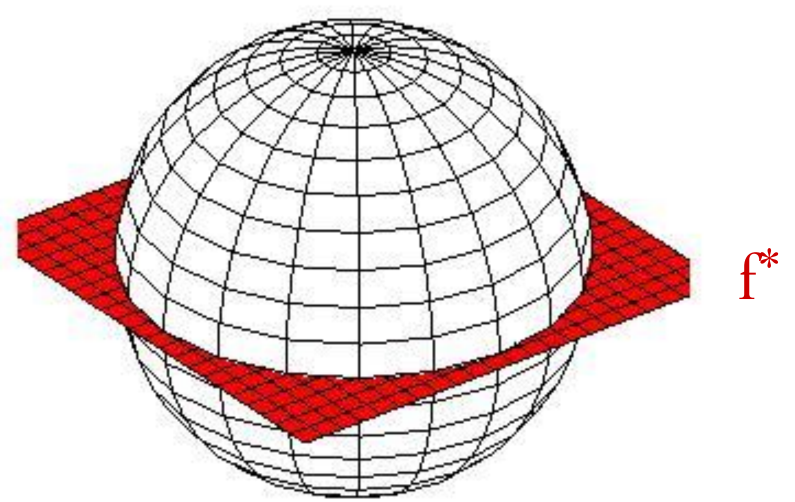
Ball: $\mathrm{B}(f^*, r) := \{f \in \mathcal{H} : P_X(f \neq f^*) \leq r\}$

$\mathrm{DIS}(\mathrm{B}(f^*, r)) := \{x \in \mathcal{X} : \exists f, f' \in \mathrm{B}(f^*, r), f(x) \neq f'(x)\}$

Disagreement coefficient:

$$\theta = \sup_{r > \epsilon} \frac{P_X(\mathrm{DIS}(\mathrm{B}(f^*, r)))}{r}$$

Example: homog. linear separators (bias 0), n dimensions, uniform P_X on sphere.



Sample Complexity Analysis

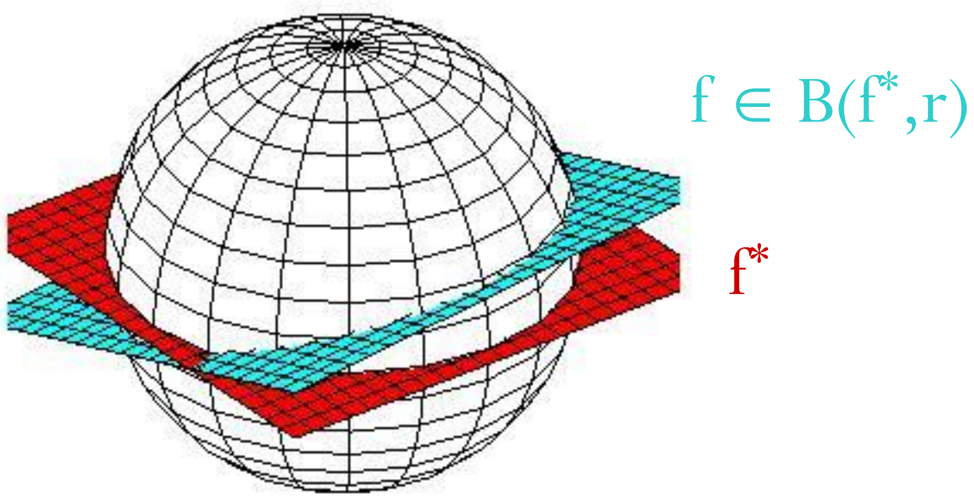
Ball: $\mathrm{B}(f^*, r) := \{f \in \mathcal{H} : P_X(f \neq f^*) \leq r\}$

$\mathrm{DIS}(\mathrm{B}(f^*, r)) := \{x \in \mathcal{X} : \exists f, f' \in \mathrm{B}(f^*, r), f(x) \neq f'(x)\}$

Disagreement coefficient:

$$\theta = \sup_{r > \epsilon} \frac{P_X(\mathrm{DIS}(\mathrm{B}(f^*, r)))}{r}$$

Example: homog. linear separators (bias 0), n dimensions, uniform P_X on sphere.



Sample Complexity Analysis

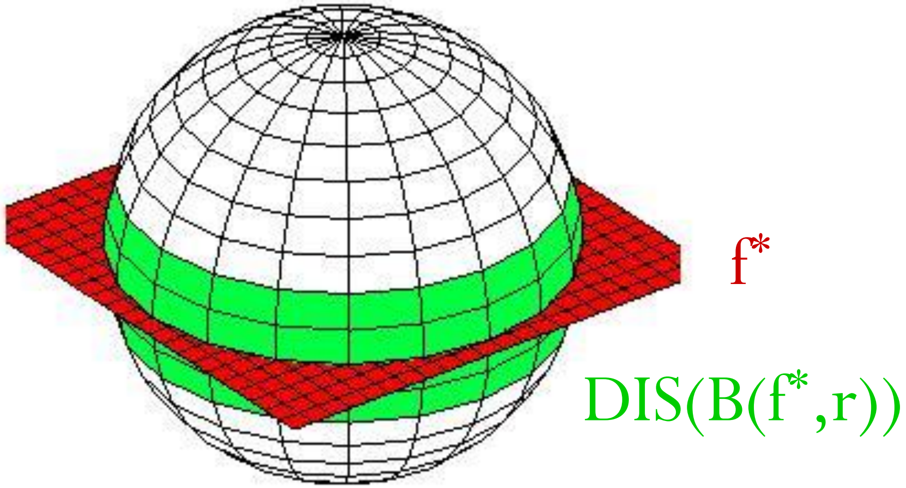
Ball: $\mathrm{B}(f^*, r) := \{f \in \mathcal{H} : P_X(f \neq f^*) \leq r\}$

$\mathrm{DIS}(\mathrm{B}(f^*, r)) := \{x \in \mathcal{X} : \exists f, f' \in \mathrm{B}(f^*, r), f(x) \neq f'(x)\}$

Disagreement coefficient:

$$\theta = \sup_{r > \epsilon} \frac{P_X(\mathrm{DIS}(\mathrm{B}(f^*, r)))}{r}$$

Example: homog. linear separators (bias 0), n dimensions, uniform P_X on sphere.



Sample Complexity Analysis

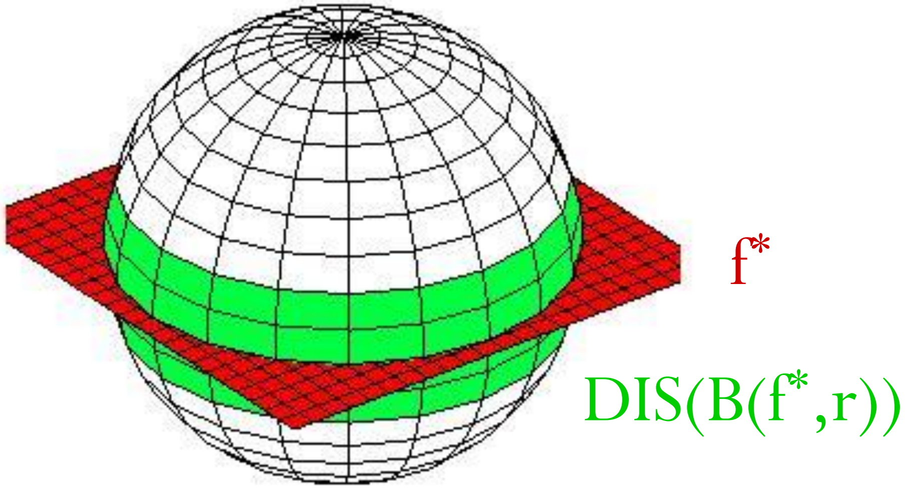
Ball: $\mathrm{B}(f^*, r) := \{f \in \mathcal{H} : P_X(f \neq f^*) \leq r\}$

$\mathrm{DIS}(\mathrm{B}(f^*, r)) := \{x \in \mathcal{X} : \exists f, f' \in \mathrm{B}(f^*, r), f(x) \neq f'(x)\}$

Disagreement coefficient:

$$\theta = \sup_{r > \epsilon} \frac{P_X(\mathrm{DIS}(\mathrm{B}(f^*, r)))}{r}$$

Example: homog. linear separators (bias 0), n dimensions, uniform P_X on sphere.



Some geometry $\Rightarrow$ for small r,

$P_X(DIS(B(f^*, r))) \propto \sqrt{n} r$.

$\Rightarrow \qquad \boldsymbol{\theta \propto \sqrt{n}}$.

Sample Complexity Analysis

Bounded Noise assumption: (aka Massart noise)

$$\exists \beta < 1/2 \text{ s.t. } P(Y \neq f^*(X)|X) \leq \beta \text{ everywhere}$$

	Sample Complexity: $R(\hat{f}) \leq R(f^*) + \epsilon$	Excess Error: n labels
Passive	$\frac{d}{\epsilon}$	$\frac{d}{n}$
Active	$d\theta \log(\frac{1}{\epsilon})$	$e^{-n/d\theta}$

Sample Complexity Analysis

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

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 $\hat{f}$

Theorem: $P(Y \neq f^*(X)|X) \leq \beta$. $R(\hat{f}) \leq R(f^*) + \epsilon$ with

$$\#\text{ labels} \approx d\theta \log(\tfrac{1}{\epsilon}).$$

Proof Sketch:

Round t, all $f \in \mathcal{H}$ **agree** on pts in $S \setminus Q$

Roughly, that means Step 4 only keeps f with
$R(f) - R(f^*) \lesssim \sqrt{P_X(f \neq f^*) \frac{d}{2^t}}$

$\Rightarrow$ surviving f after round t have $R(f) - R(f^*) \lesssim \frac{d}{2^t}$
$\Rightarrow t \gtrsim \log(\frac{d}{\epsilon})$ suffices

Also $\Rightarrow$ after round $t-1$, $\mathcal{H} \subseteq \mathrm{B}(f^*, d/2^{t-1})$

$\Rightarrow |Q| \lesssim P_X(\mathrm{DIS}(\mathrm{B}(f^*, d/2^{t-1})))|S| \leq \theta \frac{d}{2^{t-1}} |S| = \theta d 2$

$$\sum_{t=1}^{\log(d/\epsilon)} \theta d = \theta d \log(\tfrac{d}{\epsilon}) \qquad \square$$

Sample Complexity Analysis

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

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 $\hat{f}$

Bounded noise:

$$R(f) - R(f^*) = \int_{f \neq f^*} (P(Y = f^*(X)|X) - P(Y \neq f^*(X)|X))\mathrm{d}P_X$$
$$\geq (1 - 2\beta)P_X(f \neq f^*)$$

Theorem: $P(Y \neq f^*(X)|X) \leq \beta$. $R(\hat{f}) \leq R(f^*) + \epsilon$ with

$$\#\text{ labels} \approx d\theta \log(\tfrac{1}{\epsilon}).$$

Proof Sketch:

Round t, all $f \in \mathcal{H}$ **agree** on pts in $S \setminus Q$

Roughly, that means Step 4 only keeps f with
$R(f) - R(f^*) \lesssim \sqrt{P_X(f \neq f^*)\frac{d}{2^t}}$

$\Rightarrow$ surviving f after round t have $R(f) - R(f^*) \lesssim \frac{d}{2^t}$

$\Rightarrow t \gtrsim \log(\frac{d}{\epsilon})$ suffices

Also $\Rightarrow$ after round $t-1$, $\mathcal{H} \subseteq \mathrm{B}(f^*, d/2^{t-1})$

$\Rightarrow |Q| \lesssim P_X(\mathrm{DIS}(\mathrm{B}(f^*, d/2^{t-1})))|S| \leq \theta\frac{d}{2^{t-1}}|S| = \theta d 2$

$$\sum_{t=1}^{\log(d/\epsilon)} \theta d = \theta d \log(\tfrac{d}{\epsilon}) \qquad \square$$

Sample Complexity Analysis

Agnostic Learning: (no assumptions)

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

	Sample Complexity: $R(\hat{f}) \leq R(f^*) + \epsilon$	Excess Error: n labels
Passive	$d\frac{\beta}{\epsilon^2}$	$\sqrt{\frac{d\beta}{n}}$
Active	$d\theta\frac{\beta^2}{\epsilon^2}$	$\sqrt{\frac{d\beta^2\theta}{n}}$

Sample Complexity Analysis

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

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** $\hat{f} \leftarrow \underset{f \in \mathcal{H}}{\operatorname{argmin}}\, \hat{R}_Q(f)$
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 $\hat{f}$

Theorem: $\beta = R(f^*)$. $R(\hat{f}) \leq R(f^*) + \epsilon$ with

$$\#\text{ labels} \approx d\theta\frac{\beta^2}{\epsilon^2}.$$

Proof Sketch:
Round t, all $f \in \mathcal{H}$ **agree** on pts in $S \setminus Q$

Roughly, that means Step 4 only keeps f with
$R(f) - R(f^*) \lesssim \sqrt{P_X(f \neq f^*)\frac{d}{2^t}}$

$\Rightarrow$ surviving f after round t have $R(f) - R(f^*) \lesssim \sqrt{\beta\frac{d}{2^t}} + \frac{d}{2^t}$

(Roughly) $\sqrt{\beta\frac{d}{2^t}}$

$\Rightarrow t \gtrsim \log(d\frac{\beta}{\epsilon^2})$ suffices

Also $\Rightarrow$ after round $t-1$, $\mathcal{H} \subseteq \mathrm{B}\left(f^*, 2\beta + \sqrt{\beta\frac{d}{2^{t-1}}}\right) \subseteq \mathrm{B}(f^*, 3\beta)$ (for large t)

$\Rightarrow |Q| \lesssim P_X(\mathrm{DIS}(\mathrm{B}(f^*, 3\beta)))|S| \lesssim \theta\beta|S| = \theta\beta 2^t$

$$\sum_{t=1}^{\log(d\beta/\epsilon^2)} \theta\beta 2^t \sim \theta d\frac{\beta^2}{\epsilon^2}$$

$\square$

Sample Complexity Analysis

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

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** $\hat{f} \leftarrow \underset{f \in \mathcal{H}}{\operatorname{argmin}} \hat{R}_Q(f)$
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 $\hat{f}$

$\boxed{P_X(f \neq f^*) \leq R(f) + R(f^*) = 2\beta + R(f) - R(f^*)}$

Theorem: $\beta = R(f^*)$. $R(\hat{f}) \leq R(f^*) + \epsilon$ with

$$\#\text{ labels} \approx d\theta \frac{\beta^2}{\epsilon^2}.$$

Proof Sketch:

Round t, all $f \in \mathcal{H}$ **agree** on pts in $S \setminus Q$

Roughly, that means Step 4 only keeps f with
$R(f) - R(f^*) \lesssim \sqrt{P_X(f \neq f^*) \frac{d}{2^t}}$

$\Rightarrow$ surviving f after round t have $R(f) - R(f^*) \lesssim \sqrt{\beta \frac{d}{2^t}} + \frac{d}{2^t}$

(Roughly) $\sqrt{\beta \frac{d}{2^t}}$

$\Rightarrow t \gtrsim \log(d \frac{\beta}{\epsilon^2})$ suffices

Also $\Rightarrow$ after round $t-1$, $\mathcal{H} \subseteq \mathrm{B}\left(f^*, 2\beta + \sqrt{\beta \frac{d}{2^{t-1}}}\right) \subseteq \mathrm{B}(f^*, 3\beta)$ (for large t)

$\Rightarrow |Q| \lesssim P_X(\mathrm{DIS}(\mathrm{B}(f^*, 3\beta)))|S| \lesssim \theta\beta|S| = \theta\beta 2^t$

$$\sum_{t=1}^{\log(d\beta/\epsilon^2)} \theta\beta 2^t \sim \theta d \frac{\beta^2}{\epsilon^2}$$

$\square$

Sample Complexity Analysis

When is θ small?

- Linear separators, P_X has a density,
  f^* boundary intersects interior of support
  $\Rightarrow$ **$\boldsymbol{\theta}$ bounded**

- Linear separators, P_X has a density
  $\Rightarrow \boldsymbol{\theta \ll \frac{1}{\epsilon}}$

- $\mathcal{H}$ smoothly-parametrized model,
  P_X "regular" density w/ compact support,
  other technical conditions on $\mathcal{H}$
  $\Rightarrow$ **$\boldsymbol{\theta \propto \#}$ parameters for $\boldsymbol{\mathcal{H}}$**

- $\cdots$

Sample Complexity Analysis

When is θ small?

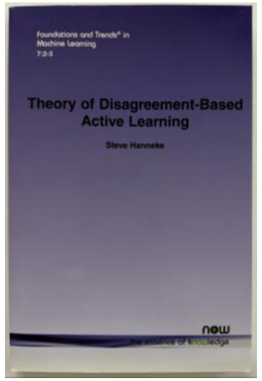
- Linear separators, P_X has a density,
  f^* boundary intersects interior of support
  $\Rightarrow \boldsymbol{\theta}$ **bounded**

- Linear separators, P_X has a density
  $\Rightarrow \boldsymbol{\theta \ll \frac{1}{\epsilon}}$

- $\mathcal{H}$ smoothly-parametrized model,
  P_X "regular" density w/ compact support,
  other technical conditions on $\mathcal{H}$
  $\Rightarrow \boldsymbol{\theta \propto \#}$ **parameters for** $\boldsymbol{\mathcal{H}}$

- $\ldots$

Lots more



Stopping Criterion

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

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 $\hat{f}$

Stopping criteria:

- Any-time
- Label budget
- Run out of unlabeled data
- Check $\max_{f \in \mathcal{H}} \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}} < \epsilon$

Simpler Agnostic Active Learning

Hsu (2010,...)

$Q \leftarrow \{\}$

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

1. **sample** a random point x
2. **optimize** $\forall y,\ \hat{f}_y \leftarrow \underset{f \in \mathcal{H}: f(x)=y}{\operatorname{argmin}} \hat{R}_Q(f)$
3. if $|\hat{R}_Q(\hat{f}_+) - \hat{R}_Q(\hat{f}_-)| \leq \sqrt{\hat{P}_Q(\hat{f}_- \neq \hat{f}_+)\frac{d}{|Q|}}$

   then **label** x, add it to Q

output $\hat{f} = \underset{f \in \mathcal{H}}{\operatorname{argmin}}\, \hat{R}_Q(f)$

- Roughly same sample complexity as A^2.
- Can implement as a **reduction** to ERM.
- In practice, replace ERM with any passive learner.

Surrogate Loss

Hanneke & Yang (2012)

$Q \leftarrow \{\}$

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

1. **sample** a random point x
2. **optimize** $\forall y,\ \hat{f}_y \leftarrow \underset{f \in \mathcal{H}: f(x)=y}{\operatorname{argmin}} \hat{R}^{\ell}_Q(f)$
3. if $|\hat{R}_Q(\hat{f}_+) - \hat{R}_Q(\hat{f}_-)| \leq \sqrt{\hat{P}_Q(\hat{f}_- \neq \hat{f}_+)\frac{d}{|Q|}}$

   then **label** x, add it to Q

output $\hat{f} = \underset{f \in \mathcal{H}}{\operatorname{argmin}} \hat{R}_Q(f)$

- Roughly same sample complexity as A^2.
- Can implement as a **reduction** to ERM.
- In practice, replace ERM with any passive learner.

Consider learner that minimizes a **surrogate loss**
$\ell : \mathbb{R} \times \{-1, +1\} \rightarrow \mathbb{R}_+$
(e.g., hinge loss, squared loss, exponential loss, . . .)

Now $\mathcal{H}$ is **real-valued** functions
$\hat{R}^{\ell}_Q(f) = \frac{1}{|Q|} \sum_{(x,y) \in Q} \ell(f(x), y)$

Theorem: Bounded noise, plus strong assumptions on $\mathcal{H}$,ℓ,P
still get $R(\hat{f}) \leq R(f^*) + \epsilon$ with # labels

$$\approx \theta d \log(\tfrac{1}{\epsilon})$$

Importance-Weighted Active Learning

Beygelzimer, Dasgupta, Langford (2009)

$Q \leftarrow \{\}$

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

1. **sample** a random point x
2. **set** sampling probability p_x
3. **flip** coin with prob p_x of heads
4. if heads, **label** x, add to Q with weight $1/p_x$

output $\hat{f} = \underset{f \in \mathcal{H}}{\operatorname{argmin}} \hat{R}_Q(f)$ (weighted loss)

Use importance weights to stay **unbiased**:
$\mathbb{E}[\hat{R}_Q(f)] = R(f)$

Now Q set of triples (x, y, w)

$$\hat{R}_Q(f) = \frac{1}{|Q|} \sum_{(x,y,w) \in Q} w\mathbb{I}[f(x) \neq y]$$

- **Any** choice of Step 2 (setting p_x) is fine (just p_x not too small, else high variance)

- Can set p_x in a way to recover A^2 sample complexity
$p_x = \mathbb{I}\Big[\ |\hat{R}_Q(\hat{f}_+) - \hat{R}_Q(\hat{f}_-)| \leq \sqrt{\hat{P}_Q(\hat{f}_+ \neq \hat{f}_-)\frac{d}{|Q|}}\ \Big]$

Importance-Weighted Active Learning

Beygelzimer, Dasgupta, Langford (2009)

$Q \leftarrow \{\}$

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

1. **sample** a random point x
2. **set** sampling probability p_x
3. **flip** coin with prob p_x of heads
4. if heads, **label** x, add to Q with weight $1/p_x$

output $\hat{f} = \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$ (weighted loss)

Use importance weights to stay **unbiased**:
$\mathbb{E}[\hat{R}_Q(f)] = R(f)$

Now Q set of triples (x, y, w)

$$\hat{R}_Q(f) = \frac{1}{|Q|} \sum_{(x,y,w) \in Q} w \mathbb{I}[f(x) \neq y]$$

- **Any** choice of Step 2 (setting p_x) is fine (just p_x not too small, else high variance)
- Can set p_x in a way to recover A^2 sample complexity
  $p_x = \mathbb{I}\left[\, |\hat{R}_Q(\hat{f}_+) - \hat{R}_Q(\hat{f}_-)| \leq \sqrt{\hat{P}_Q(\hat{f}_+ \neq \hat{f}_-)\frac{d}{|Q|}}\,\right]$
- In practice, replace ERM with any passive learner (e.g., ERM with a surrogate loss)
- (approx) implementation in **Vowpal Wabbit** library

Questions?

Further reading:

D. Cohn, L. Atlas, R. Ladner. Improving generalization with active learning. *Machine Learning*, 1994

M. F. Balcan, A. Beygelzimer, J. Langford. Agnostic active learning. *Journal of Computer and System Sciences*, 2009.

S. Hanneke. A bound on the label complexity of agnostic active learning. ICML 2007.

S. Dasgupta, D. Hsu, C. Monteleoni. A general agnostic active learning algorithm. NeurIPS 2007.

S. Hanneke. Rates of convergence in active learning. *The Annals of Statistics*, 2011.

A. Beygelzimer, S. Dasgupta, J. Langford. Importance weighted active learning. ICML 2009.

A. Beygelzimer, D. Hsu, J. Langford, T. Zhang. Agnostic active learning without constraints. NeurIPS 2010.

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

D. Hsu. Algorithms for Active Learning. PhD Thesis, UCSD, 2010.

Y. Wiener, S. Hanneke, R. El-Yaniv. A compression technique for analyzing disagreement-based active learning. *Journal of Machine Learning Research*, 2015.

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

E. Friedman. Active learning for smooth problems. COLT 2009.

S. Mahalanabis. Subset and Sample Selection for Graphical Models: Gaussian Processes, Ising Models and Gaussian Mixture Models. PhD Thesis, University of Rochester, 2012.

S. Hanneke. Theory of Disagreement-Based Active Learning. *Foundations and Trends in Machine Learning*, 2014.

S. Hanneke, L. Yang. Surrogate losses in passive and active learning. arXiv:1207.3772.