DS-GA 3001.007 Introduction to Machine Learning

Lecture 3

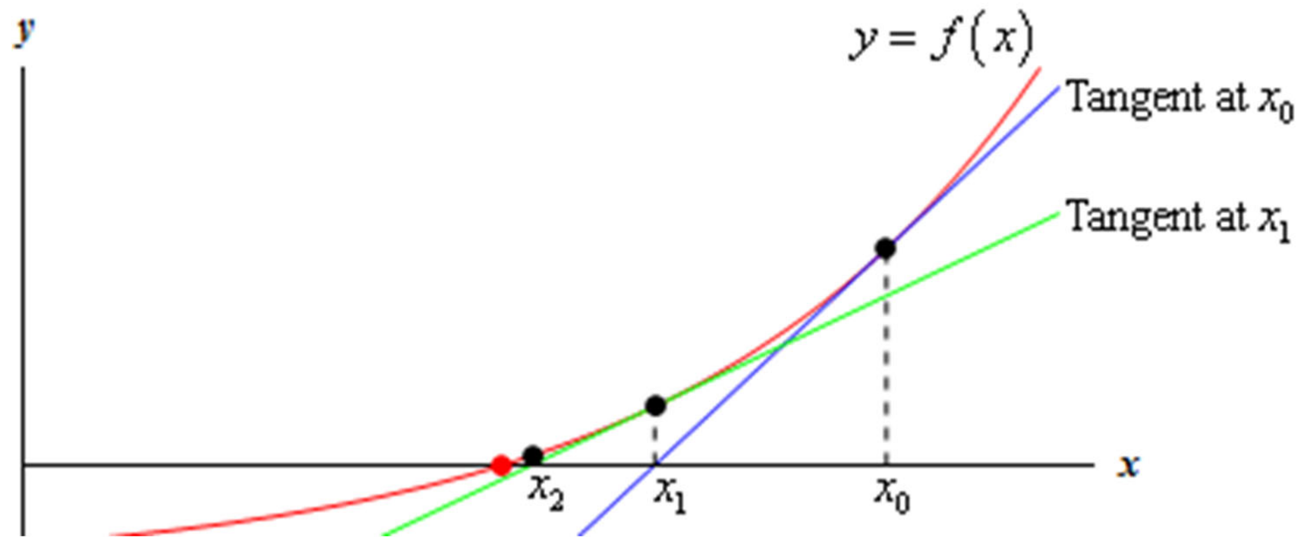
Agenda

- Review
  - Margin
- Lesson
  - Empirical Risk Minimization
- Demo
  - Pocket Algorithm

Reminders

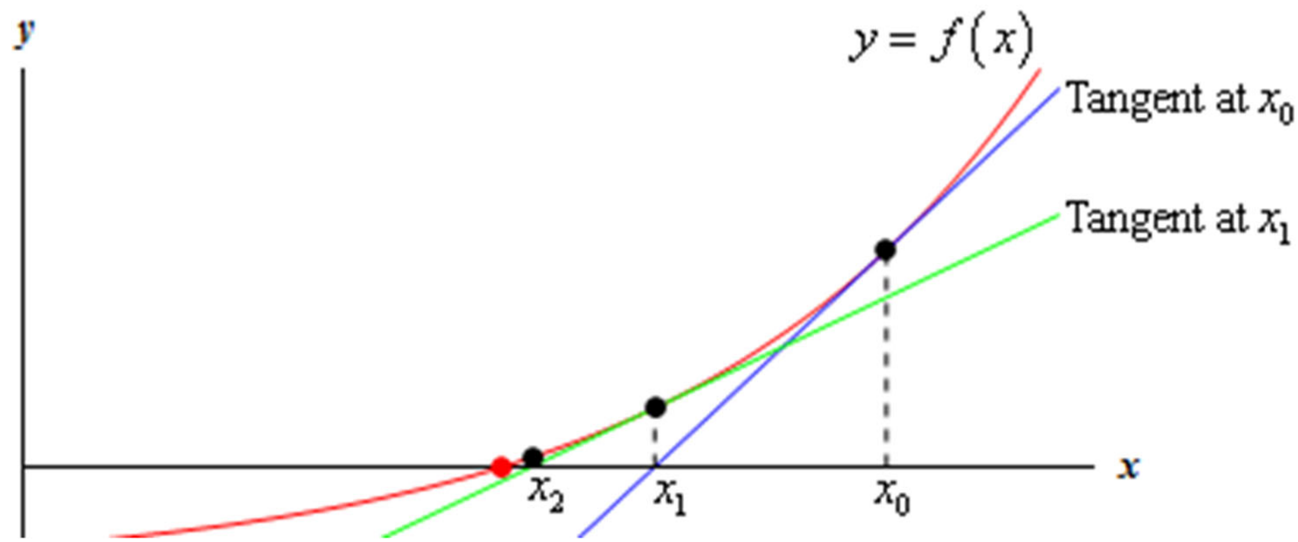
- Schedule
  - Office Hours
- Assignments
  - Homework Submission
- Materials
  - Links
- Surveys
  - Please complete Survey 2

Review



Review

- Loss Function l(x)
  - Set f(x) = l'(x)
  - Find r such that f(r) = 0

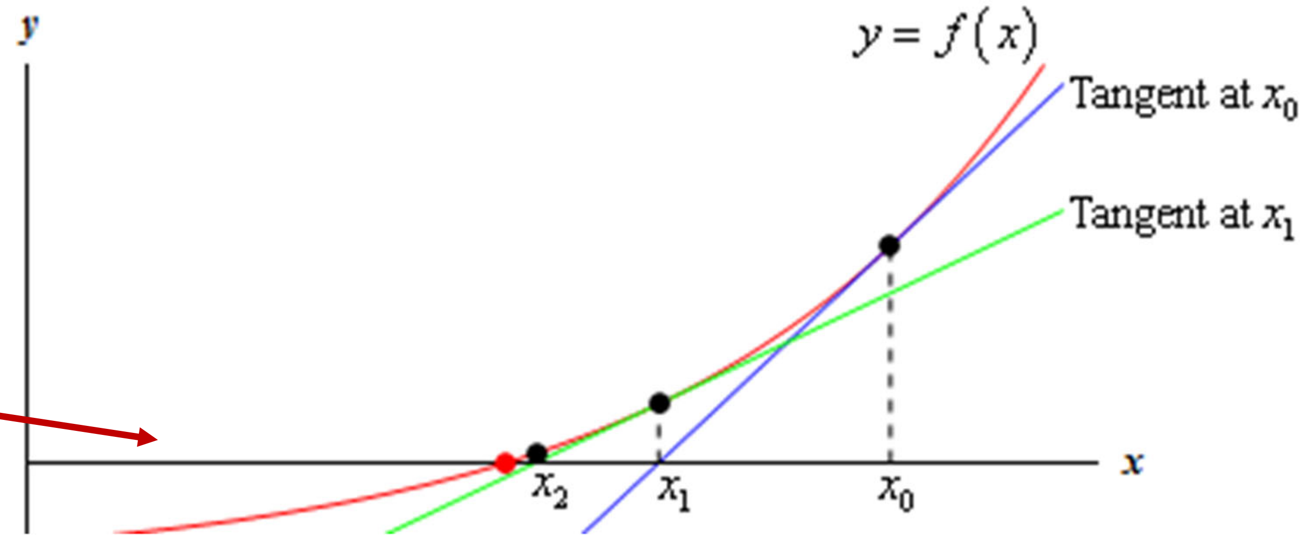


Review

- Loss Function l(x)
  - Set f(x) = l’(x)
  - Find r such that f(r) = 0

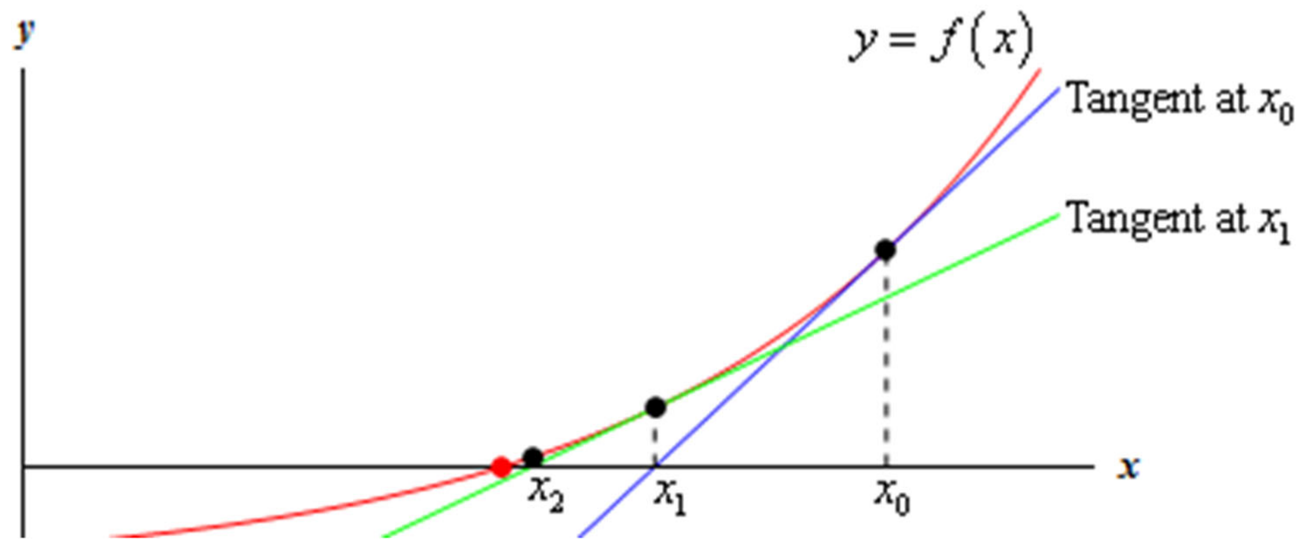
Hypothesis space

r is optimal predictor



Review

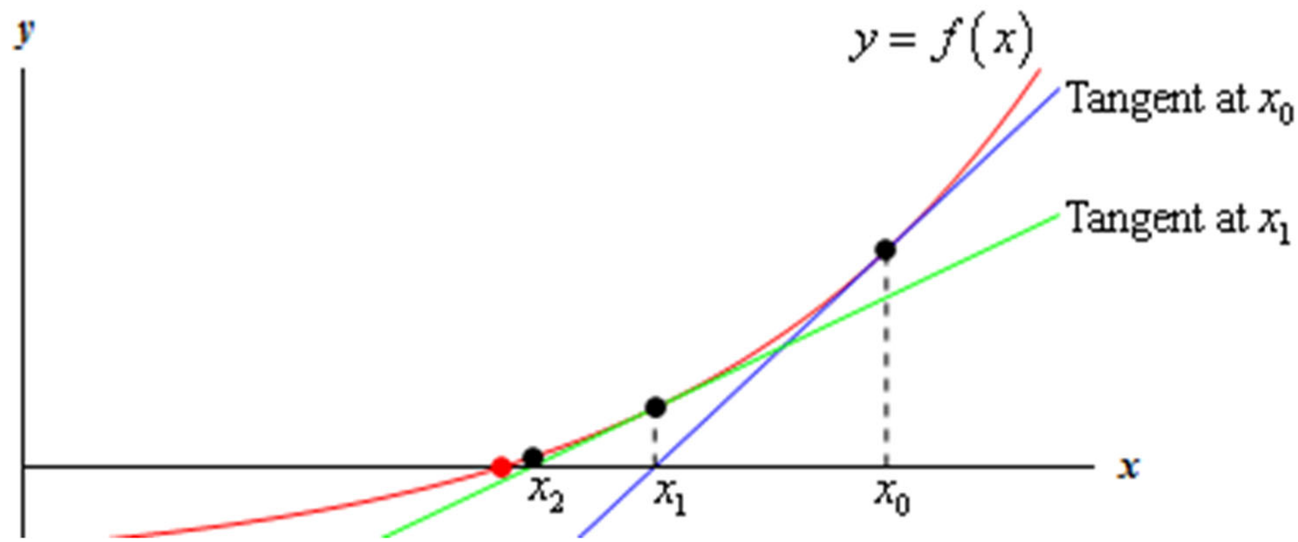
- Optimization
  - Set g(x) = x – (f(x) / f’(x))
  - Take x_{t+1} = g(x_t)



Review

- Optimization
  - Set g(x) = x – m (f(x) / f'(x))
  - Take x_{t+1} = g(x_t)

Regularization



Review

- Computational Complexity
  - Time to Run
  - Amount of Storage
  - Informed by Mistake Bound

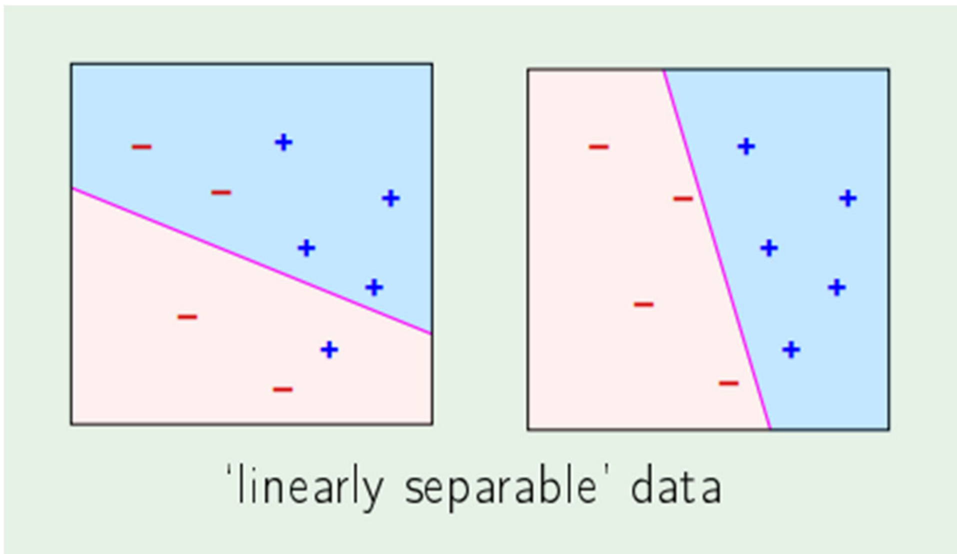
Review

- Computational Complexity
  - Time to Run
  - Amount of Storage
  - Informed by Mistake Bound
- Sample complexity
  - Amount of training data needed to learn successfully

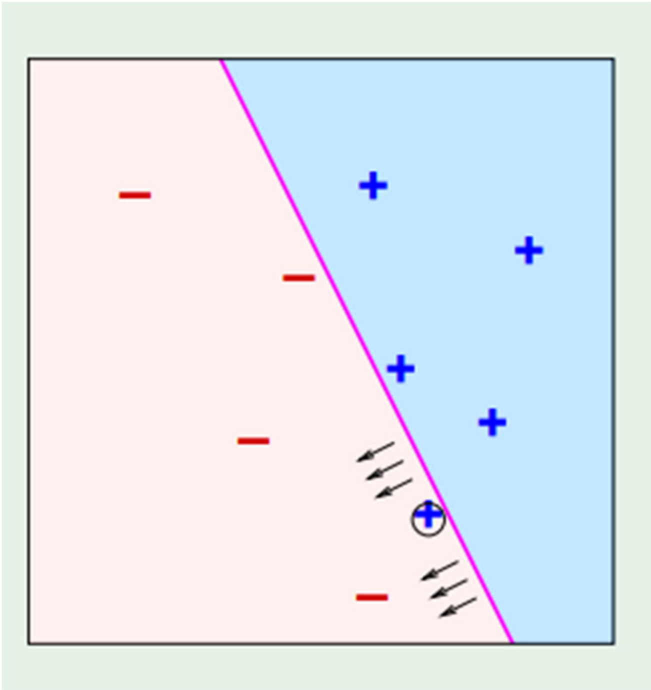
Review

- Computational Complexity
  - Time to Run
  - Amount of Storage
  - Informed by Mistake Bound
- Sample complexity
  - Amount of training data needed to learn successfully
  - Depends on the size of the hypothesis space

Review



Review



Review

- Step 1 (Input)

$$e \mapsto \mathbf{f}(e) = (h_1(e), \ldots, h_N(e)) = (x_1, \ldots, x_N)$$

Review

- Step 1 (Input)

$$e \mapsto \mathbf{f}(e) = (h_1(e), \ldots, h_N(e)) = (x_1, \ldots, x_N)$$

- Step 2 (Combine)

$$\langle \mathbf{w}, \mathbf{f}(e) \rangle = \sum_{i=1}^{N} w_i x_i$$

Review

- Step 1 (Input)

$$e \mapsto \mathbf{f}(e) = (h_1(e), \ldots, h_N(e)) = (x_1, \ldots, x_N)$$

- Step 2 (Combine)

$$\langle \mathbf{w}, \mathbf{f}(e) \rangle = \sum_{i=1}^{N} w_i x_i$$

- Step 3 (Output)

$$\mathrm{sign}(z) = \begin{cases} 1 & \text{if } z \geq 0 \\ -1 & \text{if } z < 0 \end{cases}$$

Review

x

"free money"

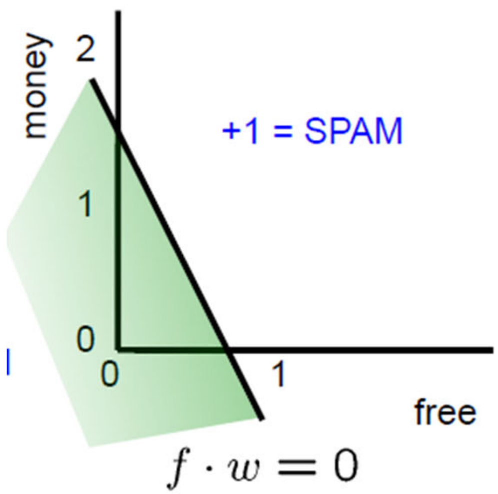
$f(x)$

```
BIAS  :  1
free  :  1
money :  1
...
```

w

```
BIAS  : -3
free  :  4
money :  2
...
```

$$
\begin{aligned}
&(1)(-3) \quad + \\
&(1)(4) \quad + \\
&(1)(2) \quad + \\
&\dots \\
&= 3
\end{aligned}
$$



Review

- Hypothesis

$$\operatorname{sign}\left(\langle \mathbf{w}, \mathbf{f}(e)\rangle - \text{threshold}\right) = \begin{cases} 1 & \text{then spam} \\ -1 & \text{then not spam} \end{cases}$$

- Step 2 (Combine)

$$f_{N+1}(e) \equiv 1$$

$$w_{N+1} = -\text{threshold}$$

- Step 3 (Output)

$$\operatorname{sign}\left(\langle \mathbf{w}, \mathbf{f}(e)\rangle\right) = \begin{cases} 1 & \text{then spam} \\ -1 & \text{then not spam} \end{cases}$$

Review

input: A training set $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$
initialize: $\mathbf{w}^{(1)} = (0, \ldots, 0)$
for $t = 1, 2, \ldots$
if $(\exists\, i \text{ s.t. } y_i \langle \mathbf{w}^{(t)}, \mathbf{x}_i \rangle \leq 0)$ then
$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} + y_i \mathbf{x}_i$
else
output $\mathbf{w}^{(t)}$

Review

$\mathbf{w}_1 \leftarrow \mathbf{w}_0$ $\quad \triangleright$ typically $\mathbf{w}_0 = \mathbf{0}$

for $t \leftarrow 1$ **to** T **do**

  Receive$(\mathbf{x}_t)$

  $\widehat{y}_t \leftarrow \text{sgn}(\mathbf{w}_t \cdot \mathbf{x}_t)$

  Receive(y_t)

  if $(\widehat{y}_t \neq y_t)$ **then**

    $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t +$ $\eta y_t \mathbf{x}_t, \eta > 0.$

  else $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t$

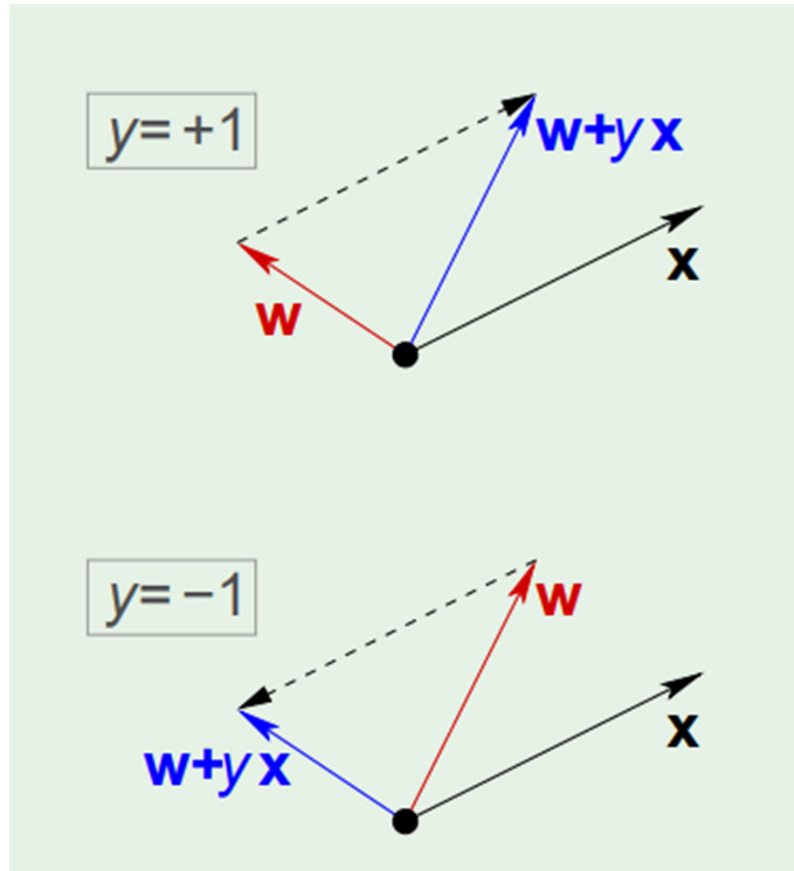
return $\mathbf{w}_{T+1}$

Review

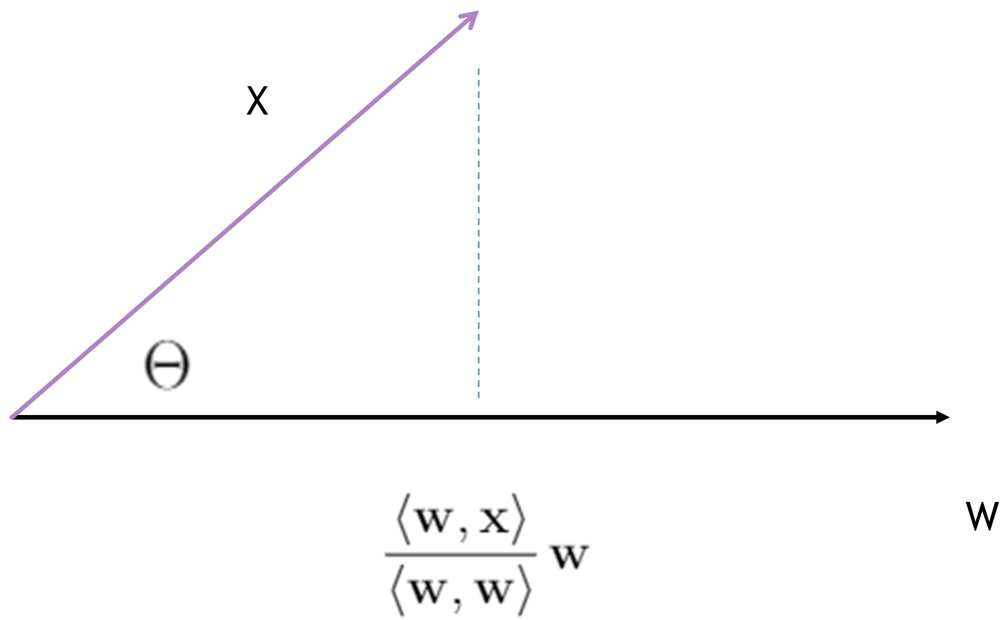
y \ Sign$\langle \mathbf{w}, \mathbf{x} \rangle$	$+1$	-1
$+1$	$+1$	-1
-1	-1	$+1$

Review

$$y \langle \mathbf{w_t}, \mathbf{x} \rangle$$
$$y \langle \mathbf{w}_{t+1}, \mathbf{x} \rangle = y \langle \mathbf{w}_t, \mathbf{x} \rangle + |\mathbf{x}|^2$$



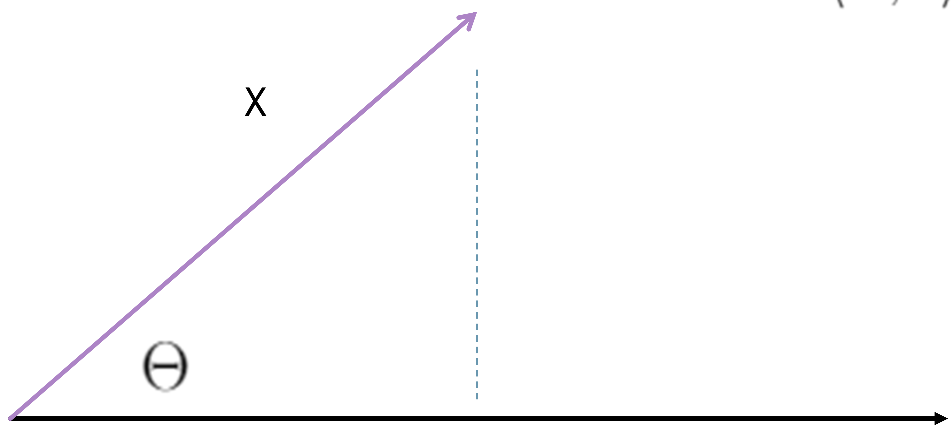
Review



X
Θ
W
$$\frac{\langle \mathbf{w}, \mathbf{x} \rangle}{\langle \mathbf{w}, \mathbf{w} \rangle} \mathbf{w}$$

Review

$$\langle \mathbf{w}, \mathbf{x} \rangle = |\mathbf{w}| \cdot |\mathbf{x}| \cdot \cos\Theta$$
$$\leq |\mathbf{w}| \cdot |\mathbf{x}|$$



$$\frac{\langle \mathbf{w}, \mathbf{x} \rangle}{\langle \mathbf{w}, \mathbf{w} \rangle} \mathbf{w}$$

W

Review

- Assume that the data is linearly separable. Set

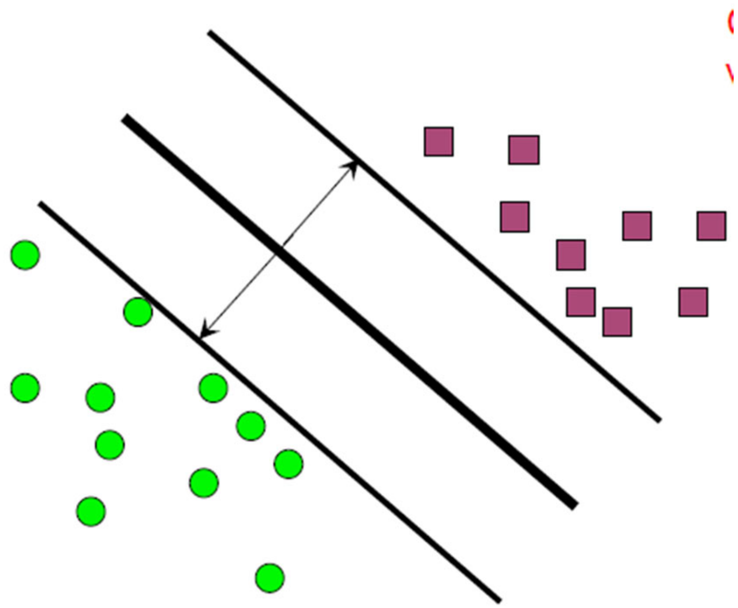
$$\rho = \min_{1 \leq t \leq T} \frac{y_i \langle \mathbf{w}, \mathbf{x}_t \rangle}{|\mathbf{v}|}$$

$$r = \max_{1 \leq i \leq T} |\mathbf{x}_i|$$

- The maximum number of mistakes made by the perceptron algorithm is

$$\frac{r^2}{\rho^2}$$

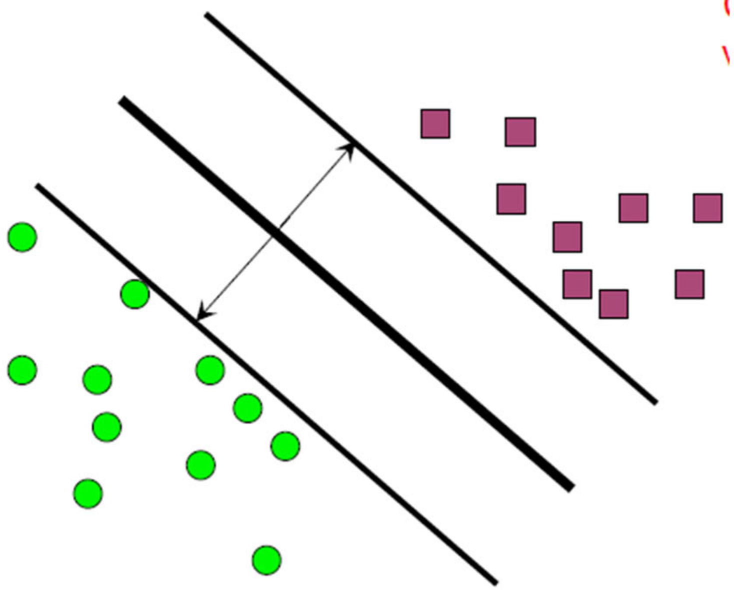
Review



Review

The distance between x and the plane defined by (w, b) is

$$\frac{|\langle w, x \rangle + b|}{||w||}$$



Review

- Assume that the data is linearly separable. Set

$$\rho = \min_{1 \le t \le T} \frac{y_i \langle \mathbf{w}, \mathbf{x}_t \rangle}{|\mathbf{v}|}$$

$$r = \max_{1 \le i \le T} |\mathbf{x}_i|$$

- The maximum number of mistakes made by the perceptron algorithm is

$$\frac{r^2}{\rho^2}$$

- Set

$$I \subset T \text{ with } \#I = M$$

Review

$$M\rho \leq \frac{\mathbf{v} \cdot \sum_{t \in I} y_t \mathbf{x}_t}{\|\mathbf{v}\|} \leq \left\| \sum_{t \in I} y_t \mathbf{x}_t \right\| \qquad \text{(Cauchy-Schwarz inequality)}$$

Review

$$
\begin{aligned}
M\rho \leq \frac{\mathbf{v} \cdot \sum_{t \in I} y_t \mathbf{x}_t}{\|\mathbf{v}\|} &\leq \left\| \sum_{t \in I} y_t \mathbf{x}_t \right\| && \text{(Cauchy-Schwarz inequality)} \\
&= \left\| \sum_{t \in I} (\mathbf{w}_{t+1} - \mathbf{w}_t) \right\| && \text{(definition of updates)}
\end{aligned}
$$

Review

$$
\begin{aligned}
M\rho \le \frac{\mathbf{v}\cdot\sum_{t\in I} y_t\mathbf{x}_t}{\|\mathbf{v}\|} \le \Big\|\sum_{t\in I} y_t\mathbf{x}_t\Big\| & \qquad \text{(Cauchy-Schwarz inequality)}\\
= \Big\|\sum_{t\in I}(\mathbf{w}_{t+1}-\mathbf{w}_t)\Big\| & \qquad \text{(definition of updates)}\\
= \|\mathbf{w}_{T+1}\| & \qquad \text{(telescoping sum, } \mathbf{w}_0 = 0\text{)}
\end{aligned}
$$

Review

$$
\begin{aligned}
M\rho \le \frac{\mathbf{v}\cdot\sum_{t\in I} y_t\mathbf{x}_t}{\|\mathbf{v}\|} &\le \Big\|\sum_{t\in I} y_t\mathbf{x}_t\Big\| && \text{(Cauchy-Schwarz inequality)}\\
&= \Big\|\sum_{t\in I}(\mathbf{w}_{t+1}-\mathbf{w}_t)\Big\| && \text{(definition of updates)}\\
&= \|\mathbf{w}_{T+1}\| && \text{(telescoping sum, } \mathbf{w}_0=0)\\
&= \sqrt{\sum_{t\in I}\|\mathbf{w}_{t+1}\|^2-\|\mathbf{w}_t\|^2} && \text{(telescoping sum, } \mathbf{w}_0=0)
\end{aligned}
$$

Review

$$
\begin{aligned}
M\rho \le \frac{\mathbf{v}\cdot\sum_{t\in I} y_t\mathbf{x}_t}{\|\mathbf{v}\|} &\le \left\|\sum_{t\in I} y_t\mathbf{x}_t\right\| && \text{(Cauchy-Schwarz inequality)}\\
&= \left\|\sum_{t\in I}(\mathbf{w}_{t+1}-\mathbf{w}_t)\right\| && \text{(definition of updates)}\\
&= \|\mathbf{w}_{T+1}\| && (\text{telescoping sum, } \mathbf{w}_0 = 0)\\
&= \sqrt{\sum_{t\in I}\|\mathbf{w}_{t+1}\|^2-\|\mathbf{w}_t\|^2} && (\text{telescoping sum, } \mathbf{w}_0 = 0)\\
&= \sqrt{\sum_{t\in I}\|\mathbf{w}_t+y_t\mathbf{x}_t\|^2-\|\mathbf{w}_t\|^2} && \text{(definition of updates)}
\end{aligned}
$$

Review

$$
\begin{aligned}
M\rho \leq \frac{\mathbf{v} \cdot \sum_{t \in I} y_t \mathbf{x}_t}{\|\mathbf{v}\|} &\leq \left\| \sum_{t \in I} y_t \mathbf{x}_t \right\| && \text{(Cauchy-Schwarz inequality)} \\
&= \left\| \sum_{t \in I} (\mathbf{w}_{t+1} - \mathbf{w}_t) \right\| && \text{(definition of updates)} \\
&= \|\mathbf{w}_{T+1}\| && \text{(telescoping sum, } \mathbf{w}_0 = 0) \\
&= \sqrt{\sum_{t \in I} \|\mathbf{w}_{t+1}\|^2 - \|\mathbf{w}_t\|^2} && \text{(telescoping sum, } \mathbf{w}_0 = 0) \\
&= \sqrt{\sum_{t \in I} \|\mathbf{w}_t + y_t \mathbf{x}_t\|^2 - \|\mathbf{w}_t\|^2} && \text{(definition of updates)} \\
&= \sqrt{\sum_{t \in I} 2\, \underbrace{y_t \mathbf{w}_t \cdot \mathbf{x}_t}_{\leq 0} + \|\mathbf{x}_t\|^2}
\end{aligned}
$$

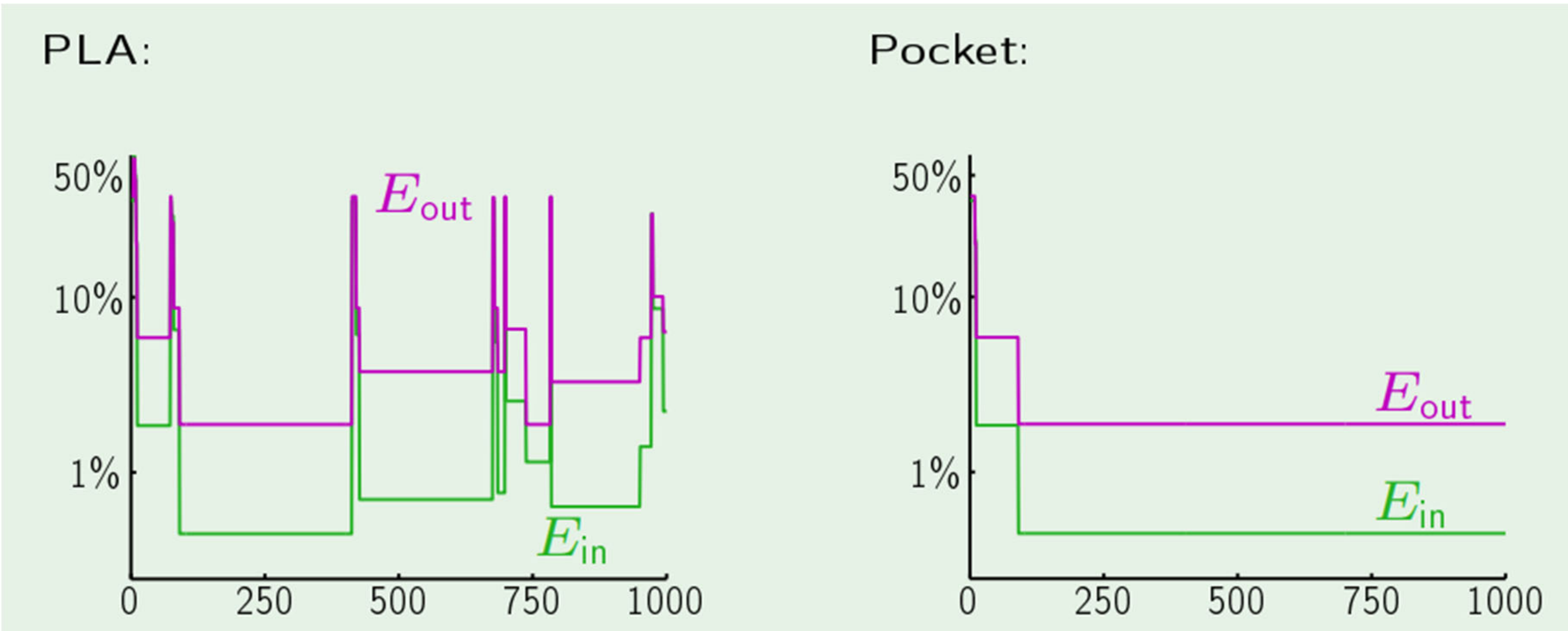
Review

$$
\begin{aligned}
M\rho \le \frac{\mathbf{v}\cdot\sum_{t\in I} y_t\mathbf{x}_t}{\|\mathbf{v}\|} &\le \Big\|\sum_{t\in I} y_t\mathbf{x}_t\Big\| && \text{(Cauchy-Schwarz inequality)}\\
&= \Big\|\sum_{t\in I}(\mathbf{w}_{t+1}-\mathbf{w}_t)\Big\| && \text{(definition of updates)}\\
&= \|\mathbf{w}_{T+1}\| && \text{(telescoping sum, } \mathbf{w}_0 = 0)\\
&= \sqrt{\sum_{t\in I}\|\mathbf{w}_{t+1}\|^2-\|\mathbf{w}_t\|^2} && \text{(telescoping sum, } \mathbf{w}_0 = 0)\\
&= \sqrt{\sum_{t\in I}\|\mathbf{w}_t+y_t\mathbf{x}_t\|^2-\|\mathbf{w}_t\|^2} && \text{(definition of updates)}\\
&= \sqrt{\sum_{t\in I} 2\,\underbrace{y_t\mathbf{w}_t\cdot\mathbf{x}_t}_{\le 0}+\|\mathbf{x}_t\|^2}\\
&\le \sqrt{\sum_{t\in I}\|\mathbf{x}_t\|^2} \le \sqrt{Mr^2}.
\end{aligned}
$$

Review

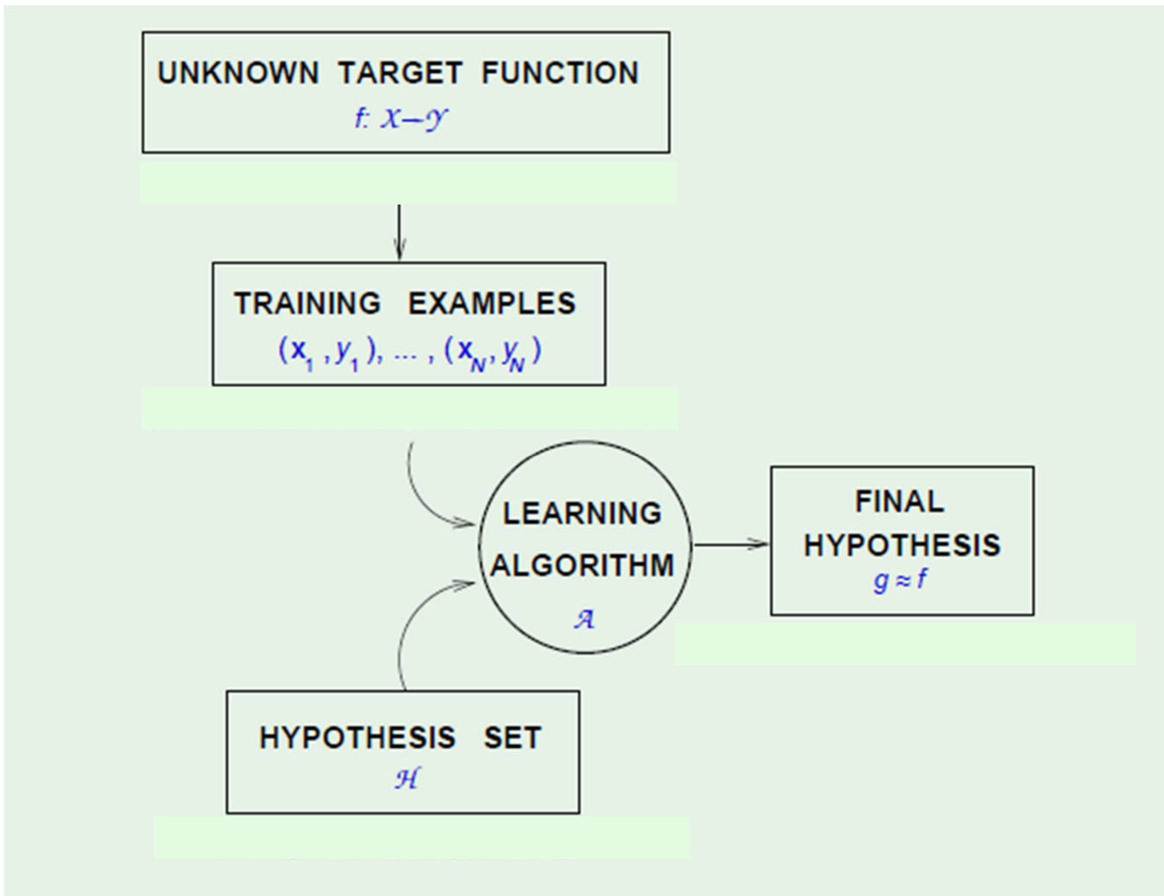
- Pocket Algorithm
  1. Set the weight **w** to **w_0** of PLA
  2. For t = 0,…,T-1 do
     1. Run PLA for one update to obtain **w_{t+1}**
     2. Count number of misclassfications
     3. If **w_{t+1}** is better than **w_{t}** then set **w** to **w_{t+1}**

Demo



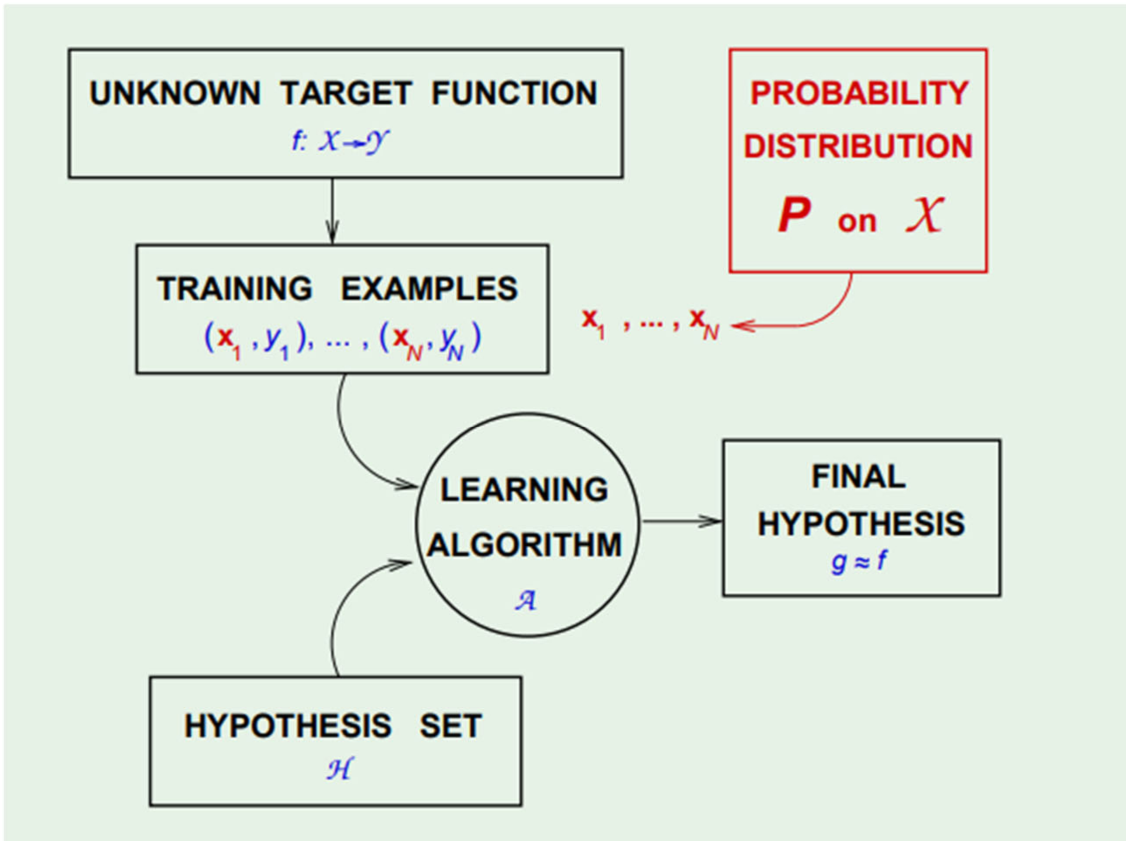
PLA:
50%
10%
1%
E_{out}
E_{in}
0
250
500
750
1000
Pocket:
50%
10%
1%
E_{out}
E_{in}
0
250
500
750
1000

Lesson



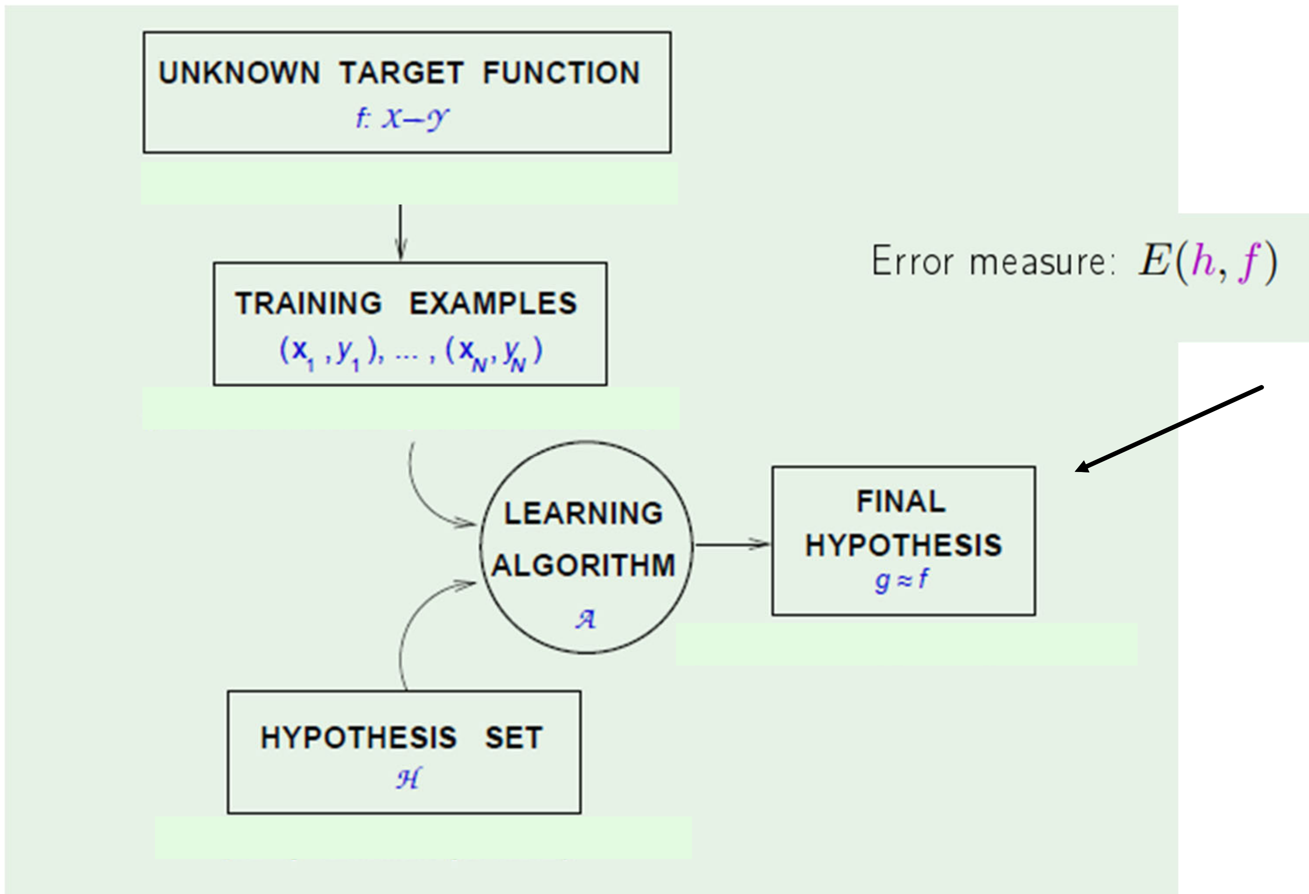
UNKNOWN TARGET FUNCTION
$f: \mathcal{X} \to \mathcal{Y}$
TRAINING EXAMPLES
$(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)$
LEARNING ALGORITHM
$\mathcal{A}$
FINAL HYPOTHESIS
$g \approx f$
HYPOTHESIS SET
$\mathcal{H}$

Lesson



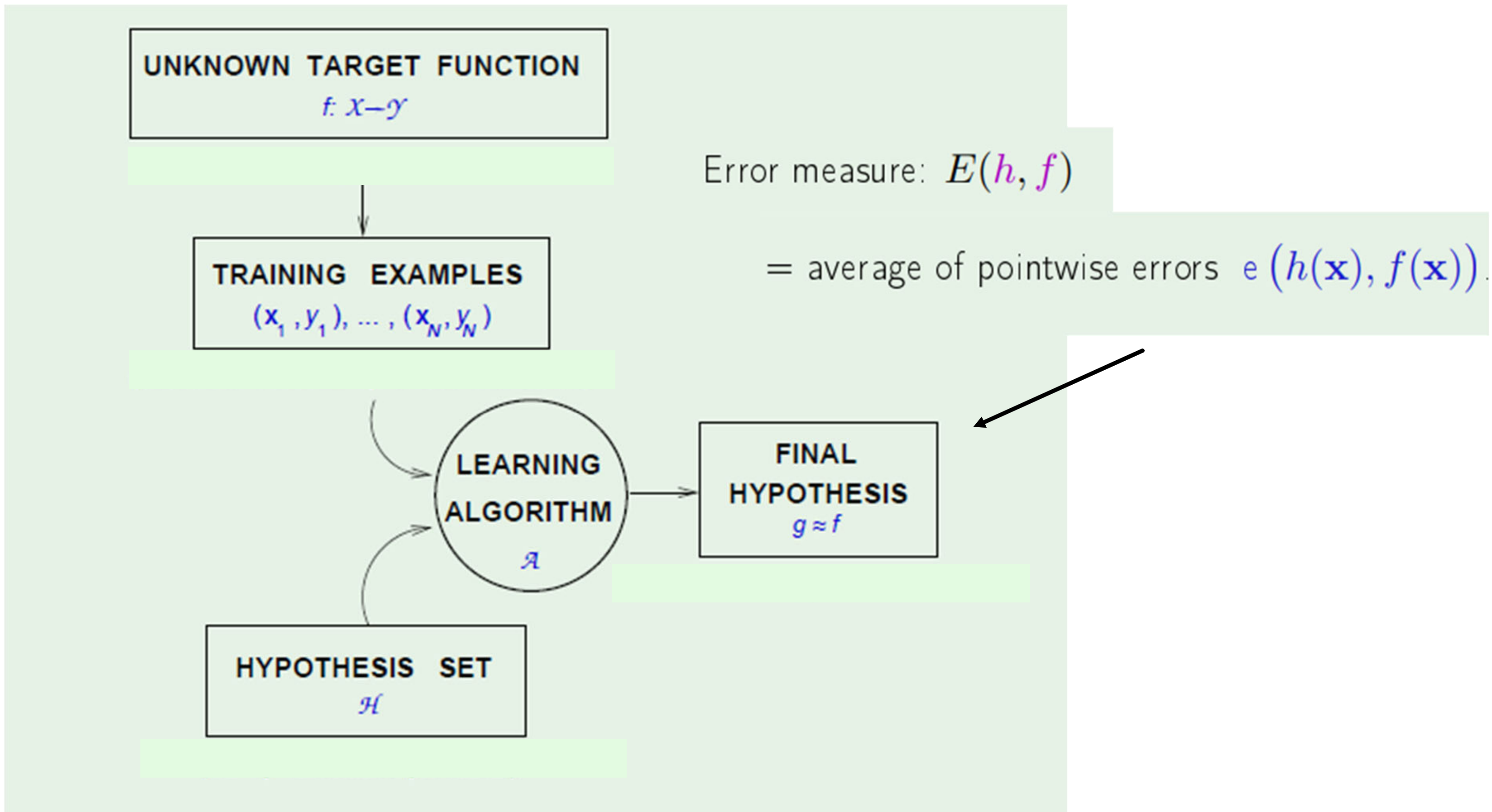
UNKNOWN TARGET FUNCTION
$f: \mathcal{X} \to \mathcal{Y}$
PROBABILITY DISTRIBUTION
P on $\mathcal{X}$
TRAINING EXAMPLES
$(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)$
$\mathbf{x}_1, \ldots, \mathbf{x}_N$
LEARNING ALGORITHM
$\mathcal{A}$
FINAL HYPOTHESIS
$g \approx f$
HYPOTHESIS SET
$\mathcal{H}$

Lesson



UNKNOWN TARGET FUNCTION
$f: \mathcal{X} \to \mathcal{Y}$
TRAINING EXAMPLES
$(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)$
LEARNING ALGORITHM
$\mathcal{A}$
FINAL HYPOTHESIS
$g \approx f$
HYPOTHESIS SET
$\mathcal{H}$
Error measure: $E(h, f)$

Lesson



UNKNOWN TARGET FUNCTION
$f: \mathcal{X} \to \mathcal{Y}$
Error measure: $E(h, f)$
$=$ average of pointwise errors $\mathrm{e}\left(h(\mathbf{x}), f(\mathbf{x})\right)$.
TRAINING EXAMPLES
$(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)$
LEARNING ALGORITHM
$\mathcal{A}$
FINAL HYPOTHESIS
$g \approx f$
HYPOTHESIS SET
$\mathcal{H}$

Lesson

In-sample error:

$$E_{\text{in}}(h) = \frac{1}{N} \sum_{n=1}^{N} \mathrm{e}\left(h(\mathbf{x}_n), f(\mathbf{x}_n)\right)$$

Out-of-sample error:

$$E_{\text{out}}(h) = \mathbb{E}_{\mathbf{x}}\left[\mathrm{e}\left(h(\mathbf{x}), f(\mathbf{x})\right)\right]$$

Lesson

		f	
		$+1$	-1
h	$+1$	no error	*false accept*
	-1	*false reject*	no error

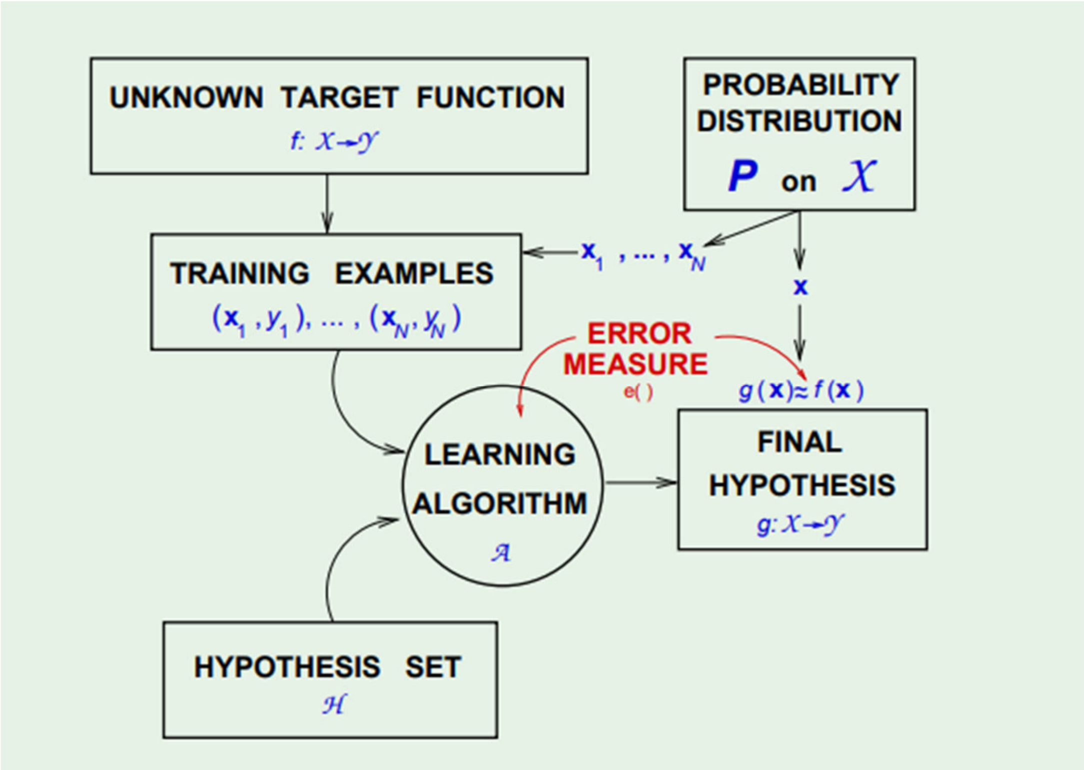
Lesson

		f	
		$+1$	-1
h	$+1$	0	1000
	-1	1	0

Lesson

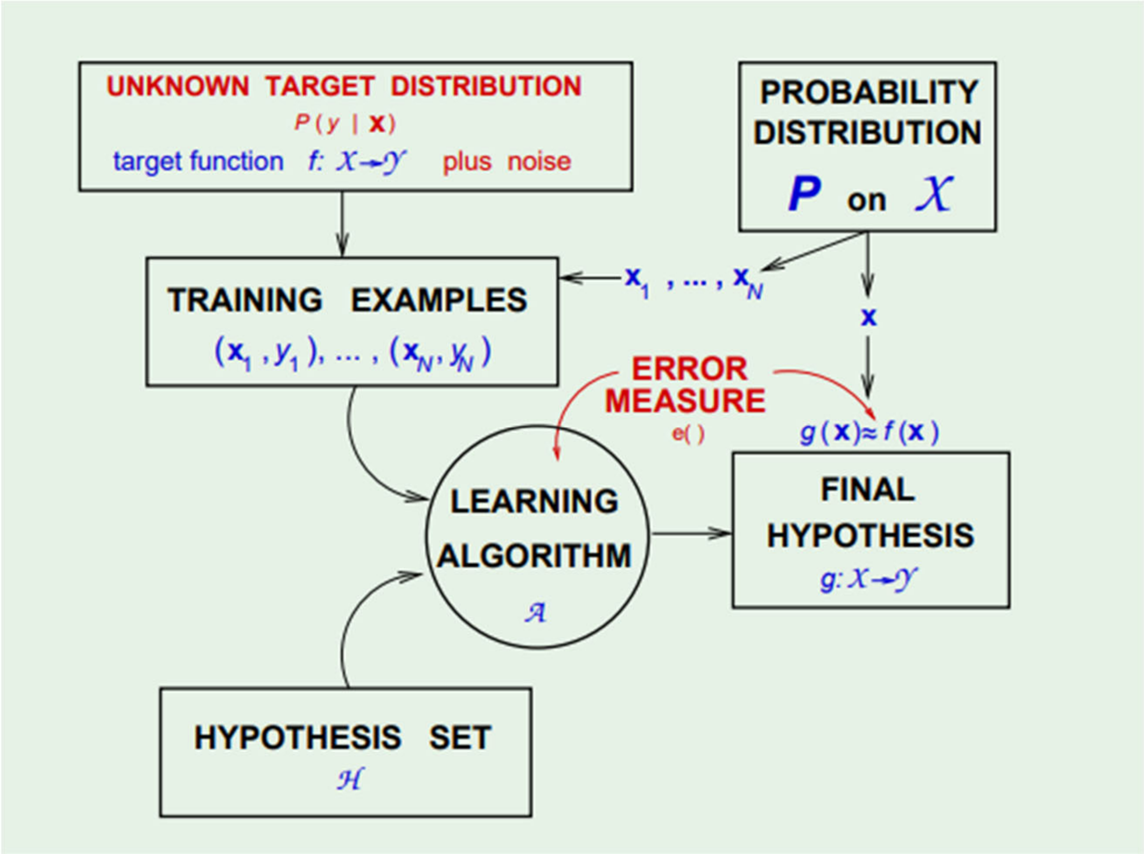
		f	
		$+1$	-1
h	$+1$	0	1
	-1	10	0

Lesson



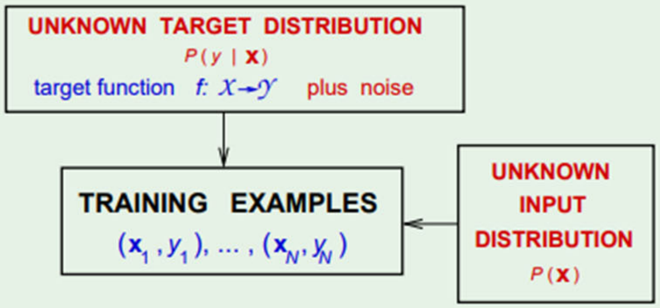
UNKNOWN TARGET FUNCTION
f: X → Y
PROBABILITY DISTRIBUTION
P on X
TRAINING EXAMPLES
(x1, y1), ... , (xN, yN)
x1, ... , xN
x
ERROR MEASURE
e( )
g(x) ≈ f(x)
LEARNING ALGORITHM
A
FINAL HYPOTHESIS
g: X → Y
HYPOTHESIS SET
H

Lesson



UNKNOWN TARGET DISTRIBUTION
P (y | x)
target function f: X → Y plus noise
PROBABILITY DISTRIBUTION
P on X
TRAINING EXAMPLES
(x1 , y1), ... , (xN, yN)
x1 , ... , xN
x
ERROR MEASURE
e( )
g (x) ≈ f (x)
LEARNING ALGORITHM
A
FINAL HYPOTHESIS
g: X → Y
HYPOTHESIS SET
H

Lesson



The target distribution $P(y \mid \mathbf{x})$
is what we are trying to learn

The input distribution $P(\mathbf{x})$
quantifies relative importance of $\mathbf{x}$

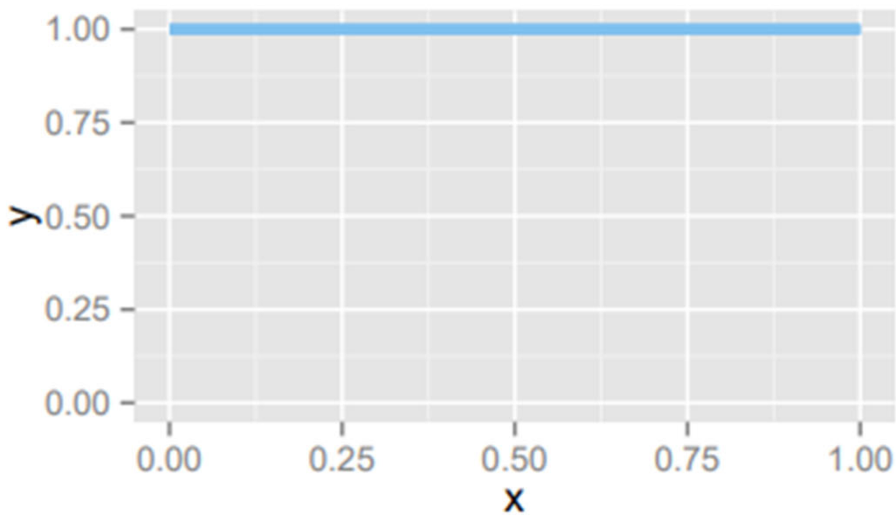
Merging $P(\mathbf{x})P(y|\mathbf{x})$ as $P(\mathbf{x}, y)$
mixes the two concepts

Lesson

1. Can we make sure that $E_{\text{out}}(g)$ is close enough to $E_{\text{in}}(g)$?
2. Can we make $E_{\text{in}}(g)$ small enough?

Lesson

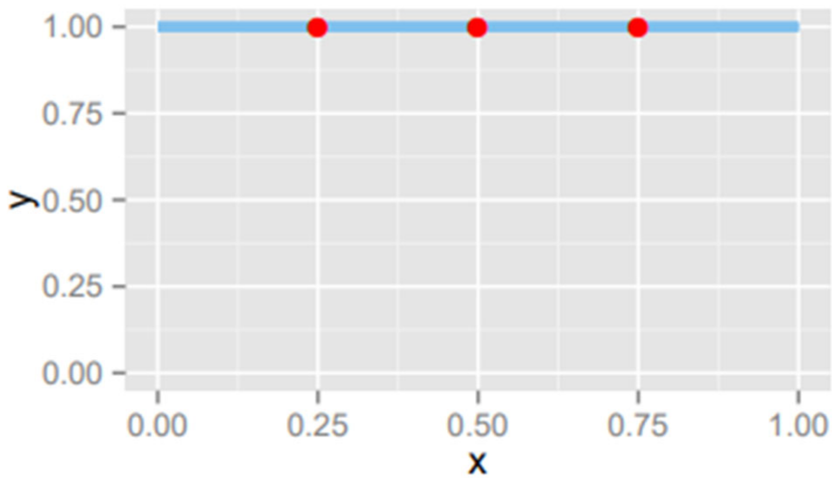
$P_{\mathcal{X}} = \text{Uniform}[0,1]$, $Y \equiv 1$ (i.e. Y is always 1).



$\mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$.

Lesson

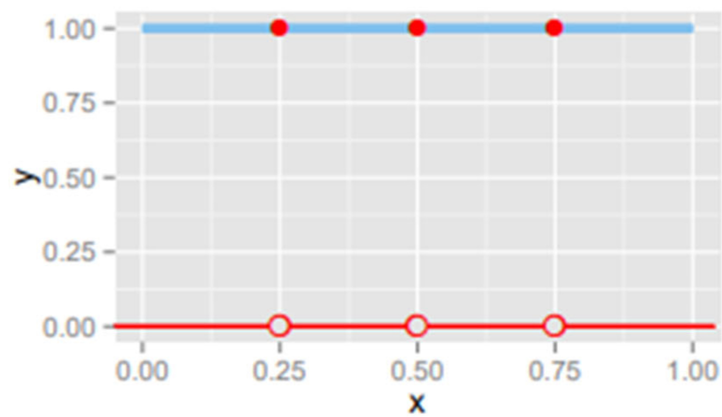
$P_{\mathcal{X}} = \text{Uniform}[0,1]$, $Y \equiv 1$ (i.e. Y is always 1).



A sample of size 3 from $\mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$.

Lesson

$P_{\mathcal{X}} = \text{Uniform}[0,1]$, $Y \equiv 1$ (i.e. Y is always 1).

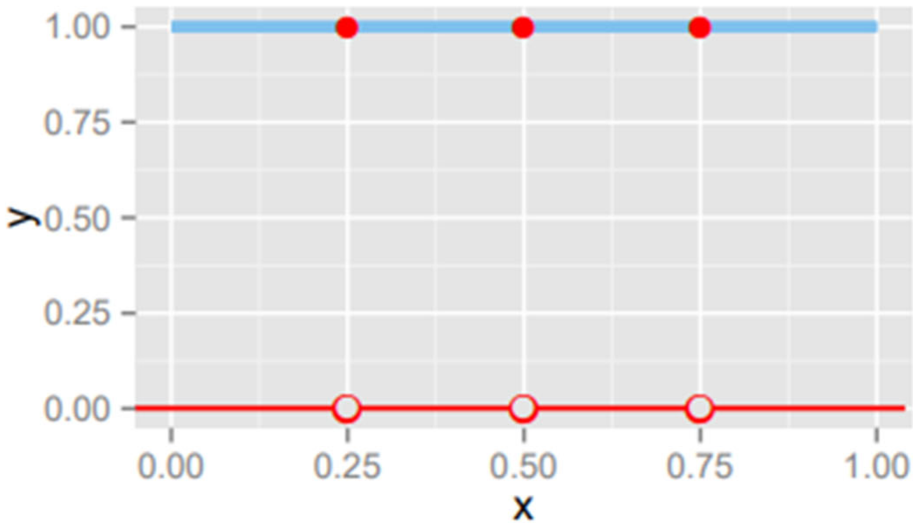


A proposed prediction function:

$$\hat{f}(x) = 1(x \in \{0.25, 0.5, 0.75\}) = \begin{cases} 1 & \text{if } x \in \{0.25, .5, .75\} \\ 0 & \text{otherwise} \end{cases}$$

Lesson

$P_{\mathcal{X}} = \text{Uniform}[0,1]$, $Y \equiv 1$ (i.e. Y is always 1).



Under square loss or 0/1 loss: $\hat{f}$ has Empirical Risk $= 0$ and Risk $= 1$.

Lesson

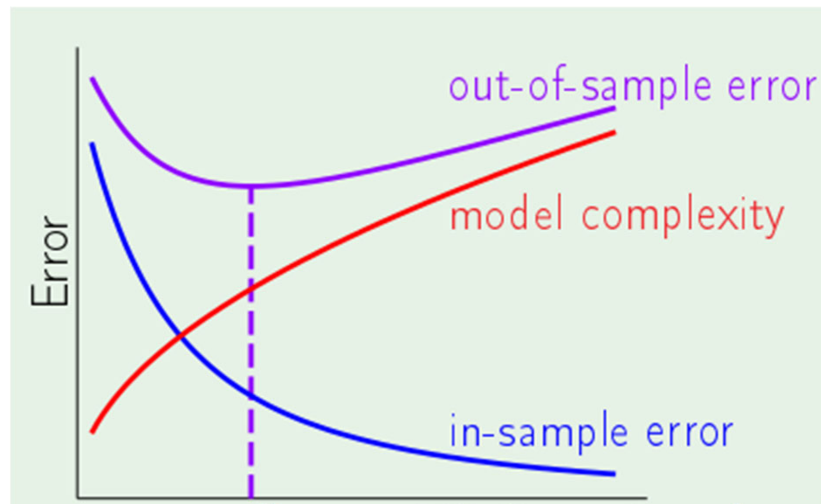
1. Can we make sure that $E_{\text{out}}(g)$ is close enough to $E_{\text{in}}(g)$?
2. Can we make $E_{\text{in}}(g)$ small enough?

Model complexity	$\uparrow$	E_{in}	$\downarrow$
Model complexity	$\uparrow$	$E_{\text{out}} - E_{\text{in}}$	$\uparrow$

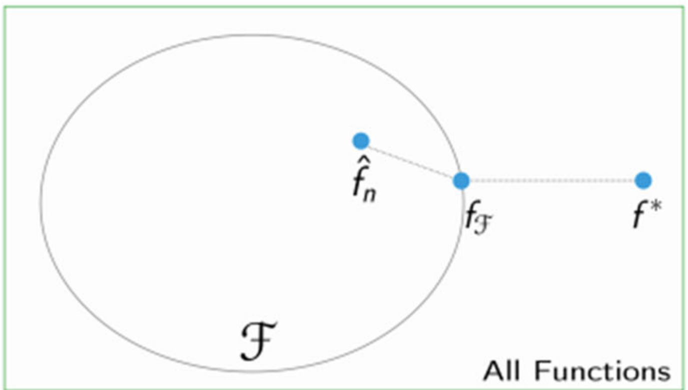
Lesson

1. Can we make sure that $E_{\text{out}}(g)$ is close enough to $E_{\text{in}}(g)$?
2. Can we make $E_{\text{in}}(g)$ small enough?

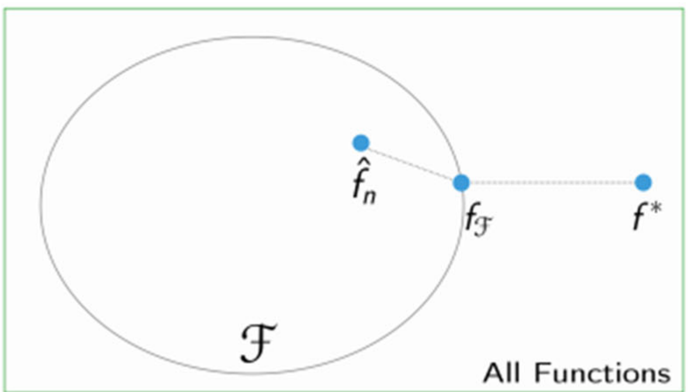
Model complexity	$\uparrow$	E_{in}	$\downarrow$
Model complexity	$\uparrow$	$E_{\text{out}} - E_{\text{in}}$	$\uparrow$



Lesson

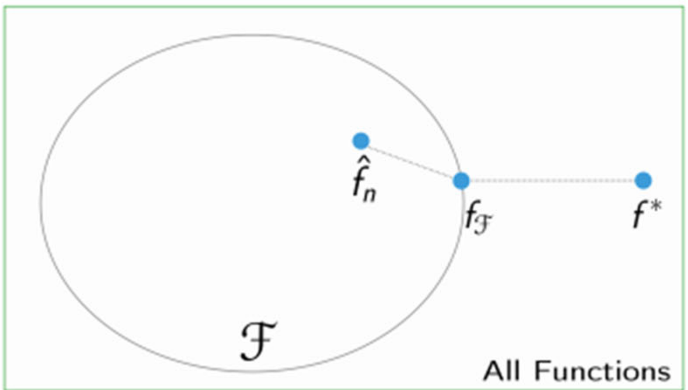


Lesson



$$f^* = \arg\min_f \mathbb{E}\ell(f(x), y)$$

Lesson

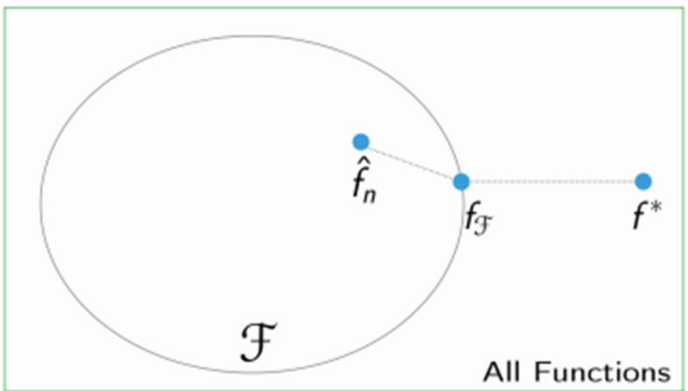


$$f^* = \arg\min_f \mathbb{E}\ell(f(x), y)$$

$$f_{\mathcal{F}} = \arg\min_{f \in \mathcal{F}} \mathbb{E}\ell(f(x), y))$$

n

Lesson

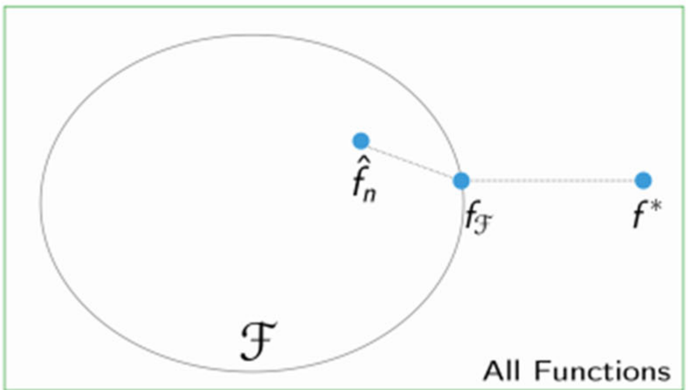


$$f^* = \arg\min_{f} \mathbb{E}\ell(f(x), y)$$

$$f_{\mathcal{F}} = \arg\min_{f \in \mathcal{F}} \mathbb{E}\ell(f(x), y))$$

$$\hat{f}_n = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i)$$

Lesson



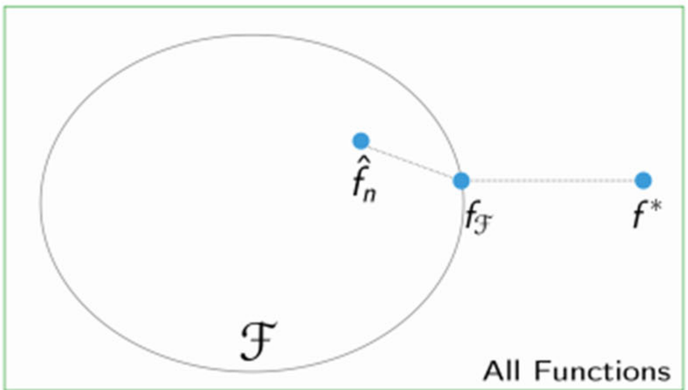
$$f^* = \arg\min_{f} \mathbb{E}\ell(f(x), y)$$

$$f_{\mathcal{F}} = \arg\min_{f \in \mathcal{F}} \mathbb{E}\ell(f(x), y))$$

$$\hat{f}_n = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i)$$

- **Approximation Error** (of $\mathcal{F}$) $= \ R(f_{\mathcal{F}}) - R(f^*)$

Lesson



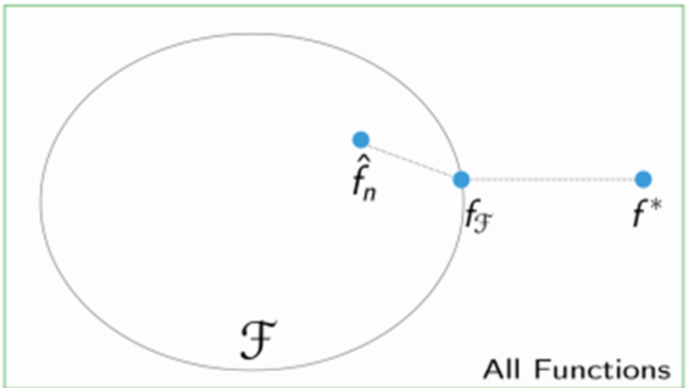
$$f^* = \arg\min_{f} \mathbb{E}\ell(f(x), y)$$

$$f_{\mathcal{F}} = \arg\min_{f \in \mathcal{F}} \mathbb{E}\ell(f(x), y))$$

$$\hat{f}_n = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i)$$

- **Approximation Error** (of $\mathcal{F}$) $= \ R(f_{\mathcal{F}}) - R(f^*)$
- **Estimation error** (of $\hat{f}_n$ in $\mathcal{F}$) $= \ R(\hat{f}_n) - R(f_{\mathcal{F}})$

Lesson



$$f^* = \arg\min_{f} \mathbb{E}\ell(f(x), y)$$

$$f_{\mathcal{F}} = \arg\min_{f \in \mathcal{F}} \mathbb{E}\ell(f(x), y))$$

$$\hat{f}_n = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i)$$

- **Approximation Error** (of $\mathcal{F}$) $= \; R(f_{\mathcal{F}}) - R(f^*)$
- **Estimation error** (of $\hat{f}_n$ in $\mathcal{F}$) $= \; R(\hat{f}_n) - R(f_{\mathcal{F}})$

$$\begin{aligned} \text{Excess Risk}(\hat{f}_n) &= R(\hat{f}_n) - R(f^*) \\ &= \underbrace{R(\hat{f}_n) - R(f_{\mathcal{F}})}_{\text{estimation error}} + \underbrace{R(f_{\mathcal{F}}) - R(f^*)}_{\text{approximation error}} . \end{aligned}$$

Lesson

$$f^* = \arg\min_f R(f)$$

Lesson

$$f^* = \arg\min_f R(f) = \arg\min_f \mathbb{E}[\mathbf{1}(f(X) \neq Y)]$$

Lesson

$$f^* = \arg\min_f R(f) = \arg\min_f \mathbb{E}[\mathbf{1}(f(X) \neq Y)] = \arg\min_f P(f(X) \neq Y),$$

Lesson

$$f^* = \arg\min_f R(f) = \arg\min_f \mathbb{E}[\mathbf{1}(f(X) \neq Y)] = \arg\min_f P(f(X) \neq Y),$$

where $(X, Y) \sim P_{\mathcal{X}\times\mathcal{Y}}$.

Lesson

$$f^* = \arg\min_f R(f) = \arg\min_f \mathbb{E}[\mathbf{1}(f(X) \neq Y)] = \arg\min_f P(f(X) \neq Y),$$

where $(X, Y) \sim P_{\mathcal{X}\times\mathcal{Y}}$. Let

$$f_1(x) = \arg\max_y P(Y = y \mid X = x),$$

Lesson

$$f^* = \arg\min_f R(f) = \arg\min_f \mathbb{E}[\mathbf{1}(f(X) \neq Y)] = \arg\min_f P(f(X) \neq Y),$$

where $(X, Y) \sim P_{\mathcal{X} \times \mathcal{Y}}$. Let

$$f_1(x) = \arg\max_y P(Y = y \mid X = x),$$

$$\begin{aligned} P(f_1(X) \neq Y) &= \textstyle\sum_x P(f_1(x) \neq Y | X = x) P(X = x) \\ &= \textstyle\sum_x (1 - P(f_1(x) = Y | X = x)) P(X = x) \end{aligned}$$

Lesson

$$f^* = \arg\min_f R(f) = \arg\min_f \mathbb{E}[\mathbf{1}(f(X) \neq Y)] = \arg\min_f P(f(X) \neq Y),$$

where $(X, Y) \sim P_{\mathcal{X}\times\mathcal{Y}}$. Let

$$f_1(x) = \arg\max_y P(Y = y \mid X = x),$$

$$\begin{array}{rcll}
P(f_1(X) \neq Y) & = & \sum_x P(f_1(x) \neq Y|X = x)P(X = x) & \\
& = & \sum_x (1 - P(f_1(x) = Y|X = x))P(X = x) & \\
& \leq & \sum_x (1 - P(f_2(x) = Y|X = x))P(X = x) & \text{(Defn of } f_1)
\end{array}$$

Lesson

$$f^* = \arg\min_f R(f) = \arg\min_f \mathbb{E}[\mathbf{1}(f(X) \neq Y)] = \arg\min_f P(f(X) \neq Y),$$

where $(X, Y) \sim P_{\mathcal{X}\times\mathcal{Y}}$. Let

$$f_1(x) = \arg\max_y P(Y = y \mid X = x),$$

$$\begin{aligned}
P(f_1(X) \neq Y) &= \textstyle\sum_x P(f_1(x) \neq Y | X = x) P(X = x) \\
&= \textstyle\sum_x (1 - P(f_1(x) = Y | X = x)) P(X = x) \\
&\leq \textstyle\sum_x (1 - P(f_2(x) = Y | X = x)) P(X = x) \quad \text{(Defn of } f_1) \\
&= \textstyle\sum_x P(f_2(x) \neq Y | X = x) P(X = x) \\
&= P(f_2(X) \neq Y).
\end{aligned}$$

Lesson

$$f^* = \arg\min_f R(f) = \arg\min_f \mathbb{E}[\mathbf{1}(f(X) \neq Y)] = \arg\min_f P(f(X) \neq Y),$$

where $(X, Y) \sim P_{\mathcal{X}\times\mathcal{Y}}$. Let

$$f_1(x) = \arg\max_y P(Y = y \mid X = x),$$

$$\begin{aligned}
P(f_1(X) \neq Y) &= \textstyle\sum_x P(f_1(x) \neq Y|X = x)P(X = x) \\
&= \textstyle\sum_x (1 - P(f_1(x) = Y|X = x))P(X = x) \\
&\leq \textstyle\sum_x (1 - P(f_2(x) = Y|X = x))P(X = x) \quad \text{(Defn of } f_1\text{)} \\
&= \textstyle\sum_x P(f_2(x) \neq Y|X = x)P(X = x) \\
&= P(f_2(X) \neq Y).
\end{aligned}$$

Thus $f^* = f_1$.