



DS-GA 3001.007

Introduction to Machine Learning

Lecture 4

Minimizing Empirical Risk - Gradient Descent

Reminders

- ▶ Survey 2
 - ▶ Please respond on Qualtrics by October 7

Reminders

- ▶ Survey 2

- ▶ Please respond on Qualtrics by October 7

- ▶ Homework 2

- ▶ Please submit to Gradescope by October 3

- ▶ Contact Ravi and Raghav through Messages

Reminders

- ▶ Survey 2
 - ▶ Please respond on Qualtrics by October 7
- ▶ Homework 2
 - ▶ Please submit to Gradescope by October 3
 - ▶ Contact Ravi and Raghav through Messages
- ▶ Project

Reminders

- ▶ Survey 2
 - ▶ Please respond on Qualtrics by October 7
- ▶ Homework 2
 - ▶ Please submit to Gradescope by October 3
 - ▶ Contact Ravi and Raghav through Messages
- ▶ Project
- ▶ Final
 - ▶ The final exam is scheduled for December 18 12-1:50pm.

Agenda

- ▶ Review
- ▶ Lesson
- ▶ Demo



Agenda

- ▶ Review
 - ▶ In Sample and Out of Sample
 - ▶ Estimation Error
 - ▶ Approximation Error
 - ▶ Optimization Error
- ▶ Lesson
- ▶ Demo



Agenda

- ▶ Review
- ▶ Lesson
 - ▶ Bound difference between
In sample and Out of Sample
 - ▶ Minimize In Sample
- ▶ Demo

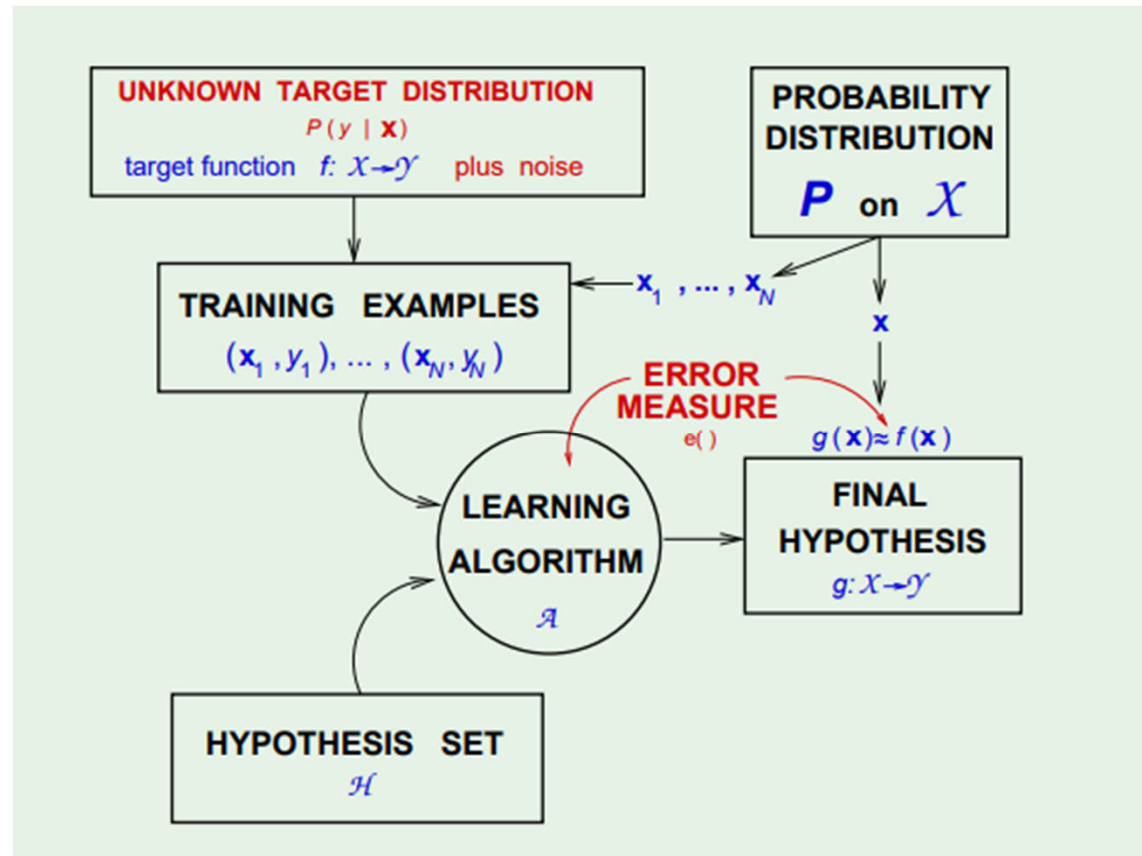


Agenda

- ▶ Review
- ▶ Lesson
- ▶ Demo



Hypotheses and Loss Functions



Hypotheses and Loss Functions

In-sample error:

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

Out-of-sample error:

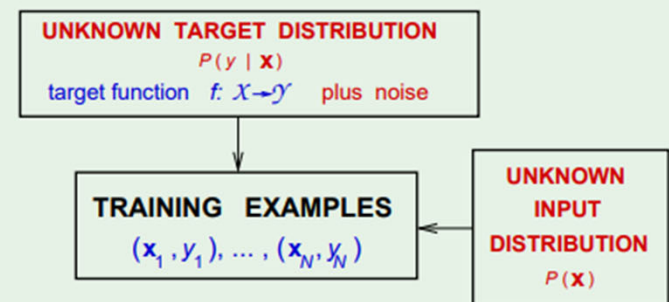
$$E_{\text{out}}(h) = \mathbb{E}_{\mathbf{x}}[e(h(\mathbf{x}), f(\mathbf{x}))]$$

Hypotheses and Loss Functions

The target distribution $P(y | \mathbf{x})$
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



Question

$$\begin{aligned} E[(c-y)^2] &= E[(c-E[y]+E[y]-y)^2] \\ &= E[(c-E[y])^2] + E[(t-E[y])^2] + 2E[c-E[y]]E[y-E[y]] \\ &= \text{bias} + \text{vars} + \\ &\text{want to be 0} \end{aligned}$$

Let $\mathcal{X} = \{1, \dots, 10\}$, let $\mathcal{Y} = \{1, \dots, 10\}$, and let $A = \mathcal{Y}$. Suppose the data generating distribution, P , has marginal $X \sim \text{Unif}\{1, \dots, 10\}$ and conditional distribution $Y|X = x \sim \text{Unif}\{1, \dots, x\}$. For each loss function below give a Bayes decision function.

mean (a) $\ell(a, y) = (a - y)^2$, regression

(b) $\ell(a, y) = |a - y|$,

mod (c) $\ell(a, y) = \mathbf{1}(a \neq y)$. if a is equal to y then the loss function will be 1, else 0 classification

l is for the loss function

there are three different loss functions

$p(y \geq c) = p(y < c)$ c: median

Hypotheses and Loss Functions

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?

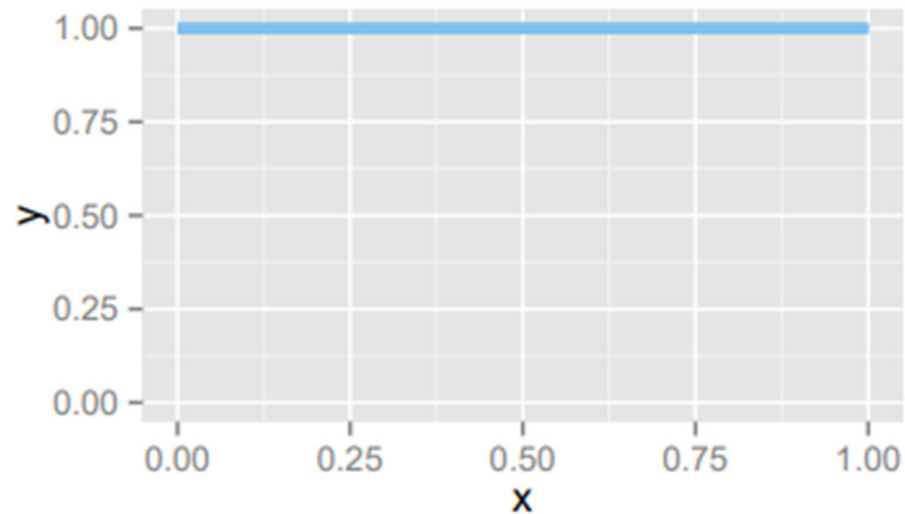
Hypotheses and Loss Functions

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	↑	E_{in}	↓
Model complexity	↑	$E_{\text{out}} - E_{\text{in}}$	↑

Overfitting and Underfitting

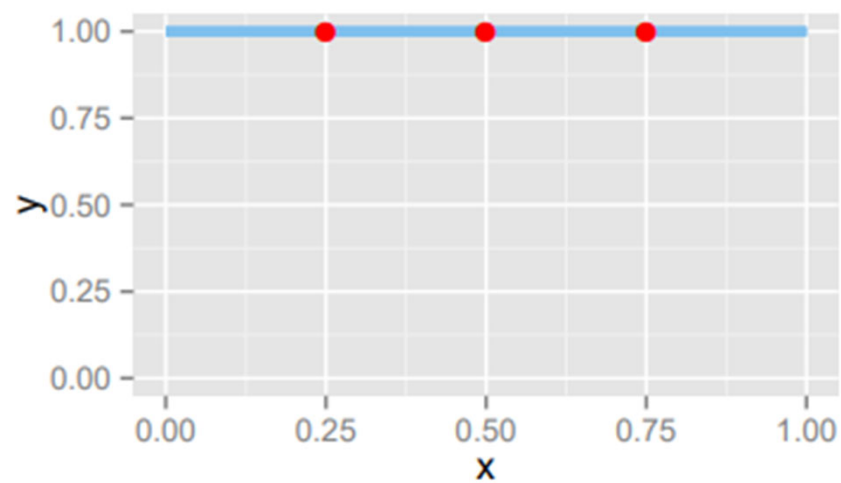
$P_{\mathcal{X}} = \text{Uniform}[0,1]$, $Y \equiv 1$ (i.e. Y is always 1).



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

Overfitting and Underfitting

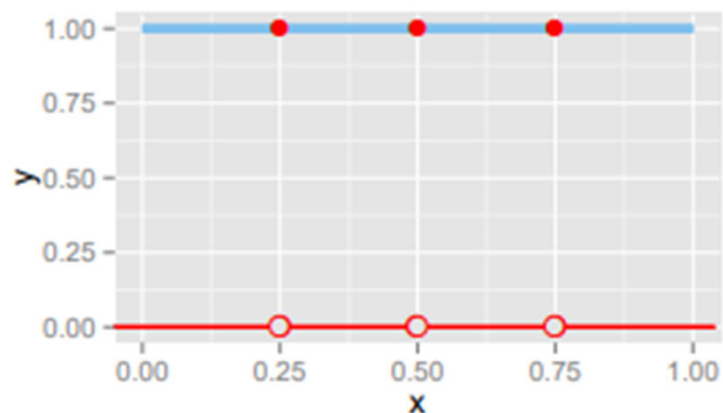
$P_{\mathcal{X}} = \text{Uniform}[0,1]$, $Y \equiv 1$ (i.e. Y is always 1).



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

Overfitting and Underfitting

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

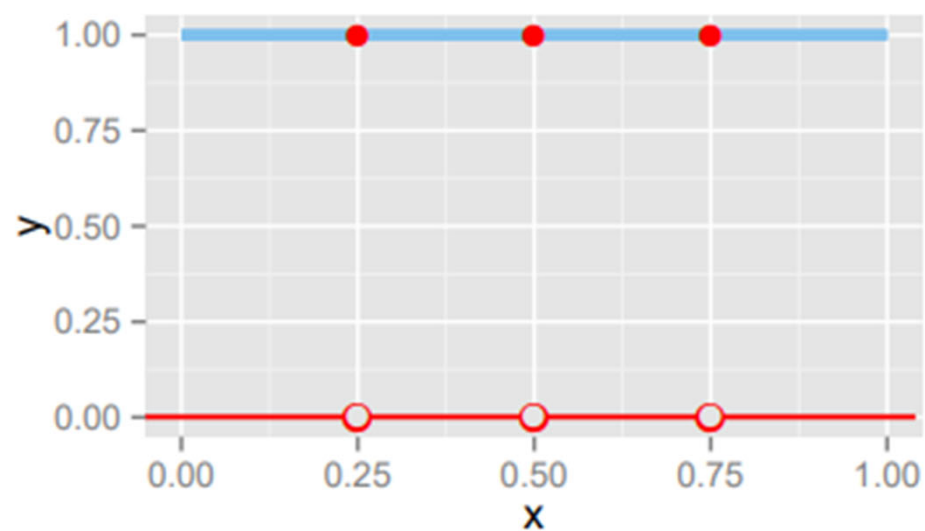


A proposed prediction function:

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

Overfitting and Underfitting

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



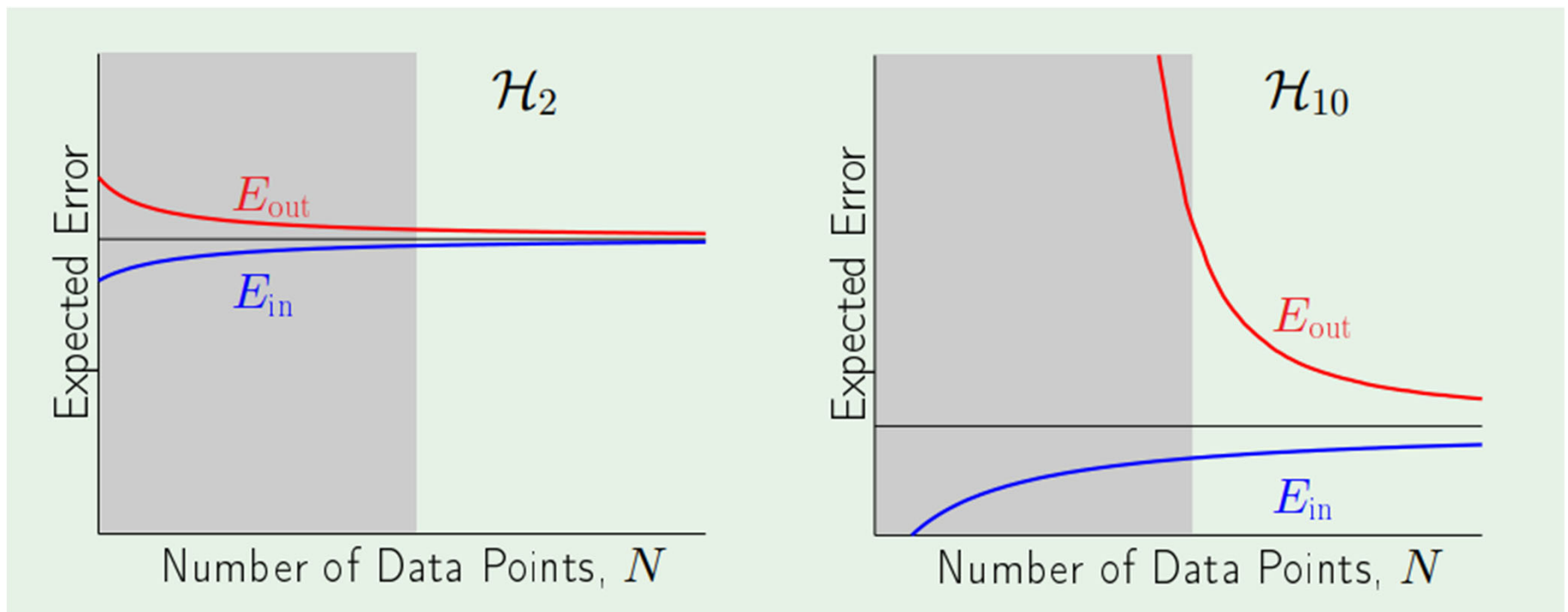
Under square loss or 0/1 loss: \hat{f} has Empirical Risk = 0 and Risk = 1.

Hypotheses and Loss Functions

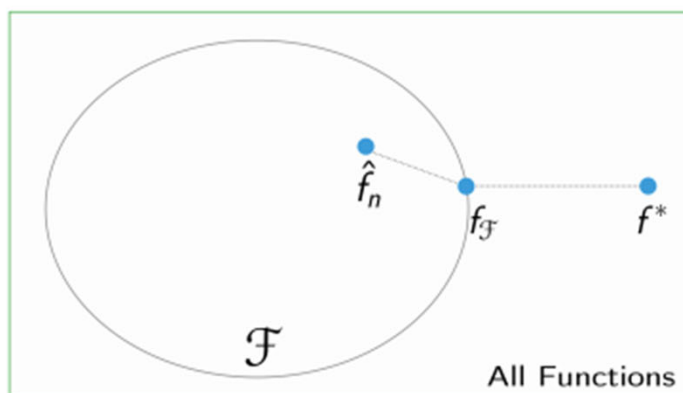
1. Can we make sure that $E_{\text{out}}(g)$ is close enough to $E_{\text{in}}(g)$?
2. Can we make $E_{\text{in}}(g)$ small enough?

Sample	complexity	↑	E_{in}	↑
Sample	complexity	↑	$E_{\text{out}} - E_{\text{in}}$	↓

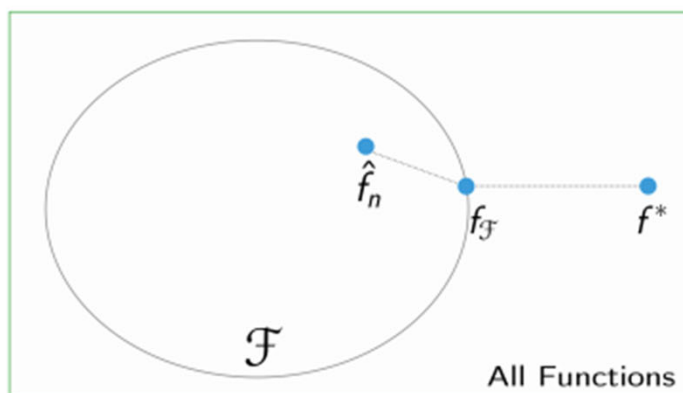
Overfitting and Underfitting



Finding Accurate Hypotheses

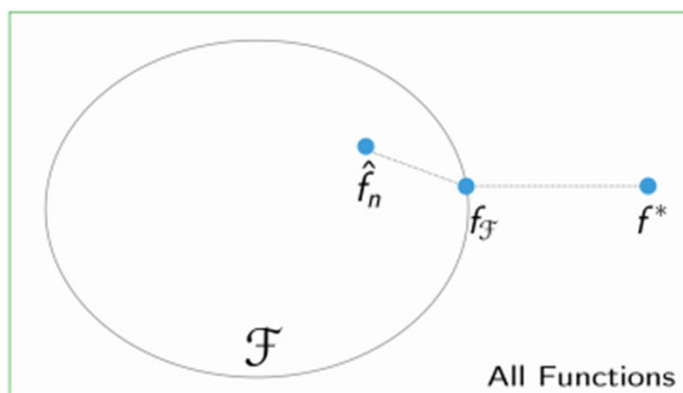


Finding Accurate Hypotheses



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

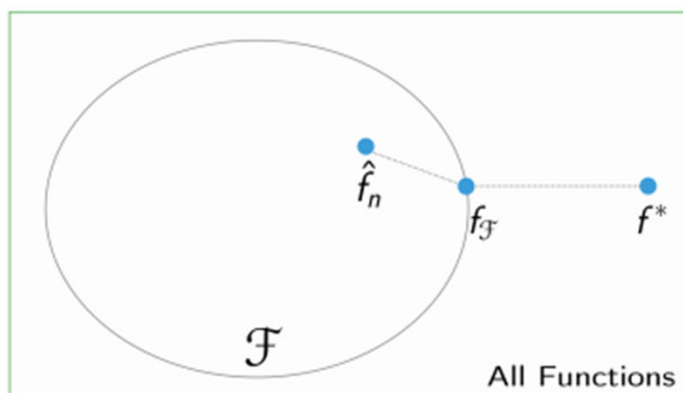
Finding Accurate Hypotheses



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

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

Finding Accurate Hypotheses



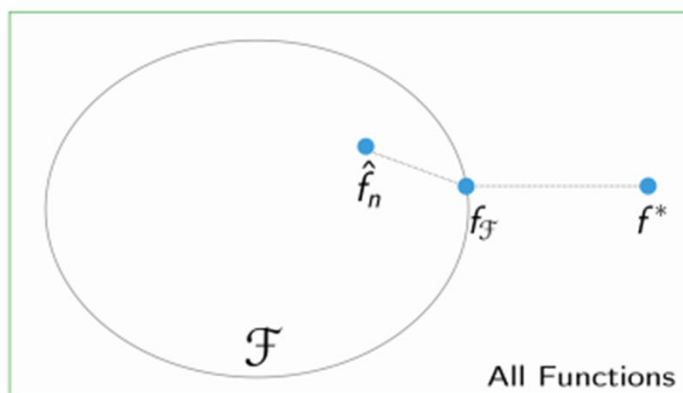
$$f^* = \arg \min_f \mathbb{E} \ell(f(x), y)$$

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

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

the data we have
ERM

Finding Accurate Hypotheses



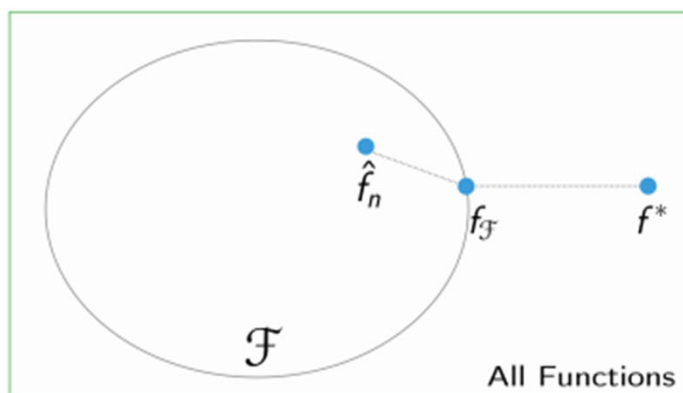
$$f^* = \arg \min_f \mathbb{E} \ell(f(x), y)$$

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

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

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

Finding Accurate Hypotheses



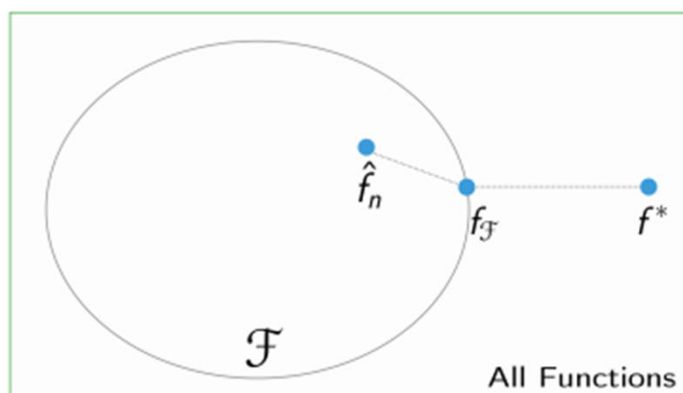
$$f^* = \arg \min_f \mathbb{E} \ell(f(x), y)$$

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

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

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

Finding Accurate Hypotheses



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

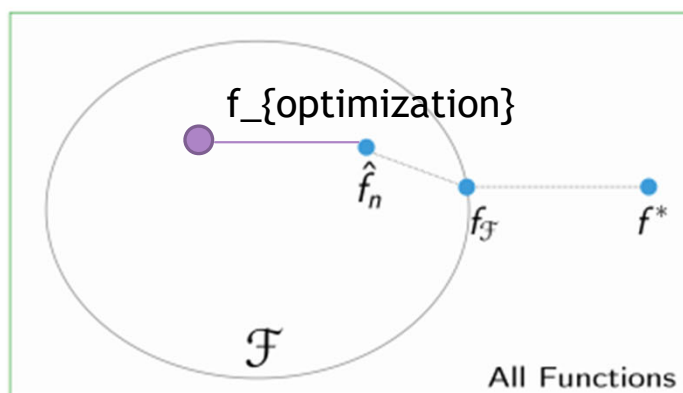
$$f_{\mathcal{F}} = \arg \min_{f \in \mathcal{F}} \mathbb{E} \ell(f(x), y)$$

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

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

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

Finding Accurate Hypotheses



can we compute it?

random or not?

positive or negative?

how change the model complexity?

how change with the sample complexity?

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

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

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

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

• Estimation error (of \hat{f}_n in \mathcal{F}) = $R(\hat{f}_n) - R(f_{\mathcal{F}})$

positive

$$\text{Excess Risk}(\hat{f}_n) = R(\hat{f}_n) - R(f^*)$$

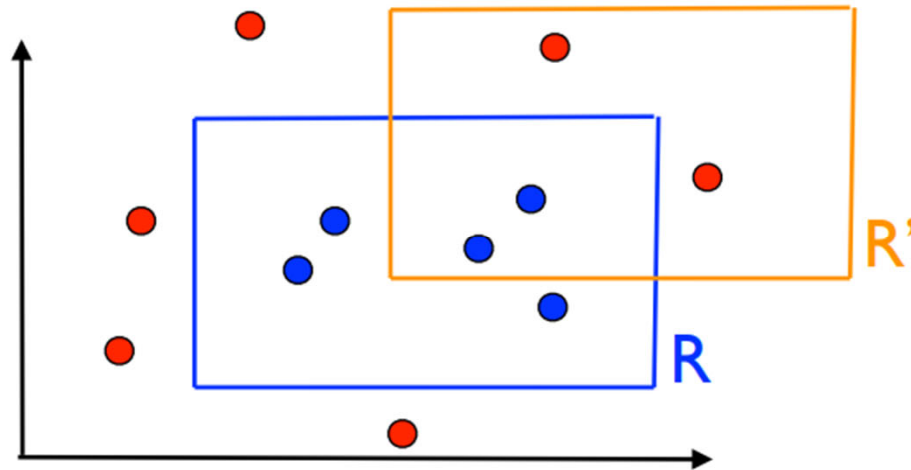
$$= \underbrace{R(\hat{f}_n) - R(f_{\mathcal{F}})}_{\text{estimation error}} + \underbrace{R(f_{\mathcal{F}}) - R(f^*)}_{\text{approximation error}}.$$

Question

- ▶ Review
- ▶ Lesson
- ▶ Demo

Bound Difference In Sample and Out of Sample

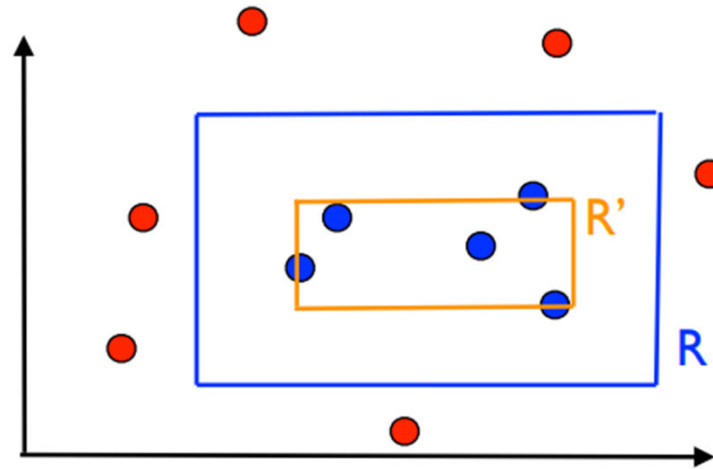
- **Problem:** learn unknown axis-aligned rectangle R using as small a labeled sample as possible.



- **Hypothesis:** rectangle R' . In general, there may be false positive and false negative points.

Bound Difference In Sample and Out of Sample

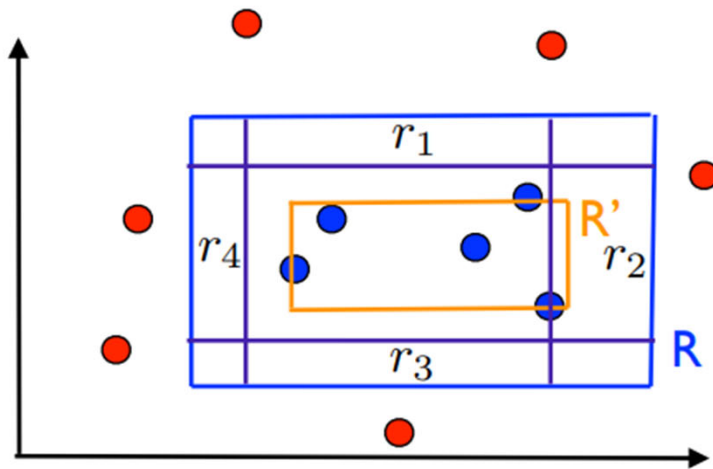
- **Simple method:** choose tightest consistent rectangle R' for a large enough sample. How large a sample?



- What is the probability that $R(R') > \epsilon$?

Bound Difference In Sample and Out of Sample

- Fix $\epsilon > 0$ and assume $\Pr_D[R] > \epsilon$ (otherwise the result is trivial).
- Let r_1, r_2, r_3, r_4 be four smallest rectangles along the sides of R such that $\Pr_D[r_i] \geq \frac{\epsilon}{4}$.



$$\begin{aligned}
 R &= [l, r] \times [b, t] \\
 r_4 &= [l, s_4] \times [b, t] \\
 s_4 &= \inf\{s: \Pr[l, s] \times [b, t] \geq \frac{\epsilon}{4}\} \\
 \Pr_D[l, s_4] \times [b, t] &< \frac{\epsilon}{4}
 \end{aligned}$$

Bound Difference In Sample and Out of Sample

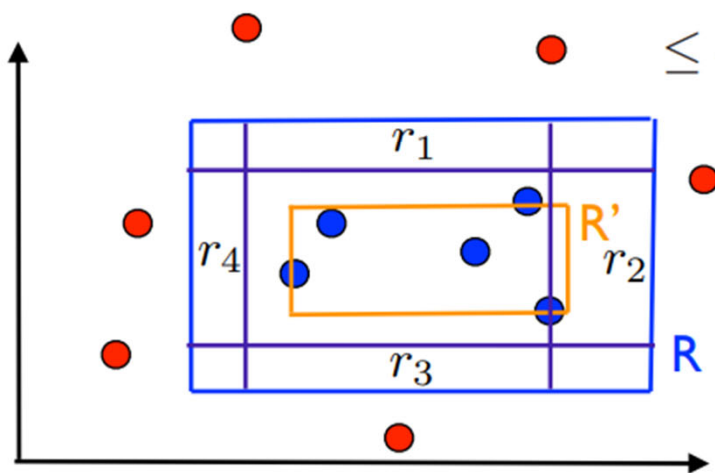
- Errors can only occur in $R - R'$. Thus (geometry),

$$R(R') > \epsilon \Rightarrow R' \text{ misses at least one region } r_i.$$

- Therefore, $\Pr[R(R') > \epsilon] \leq \Pr[\cup_{i=1}^4 \{R' \text{ misses } r_i\}]$

$$\leq \sum_{i=1}^4 \Pr[\{R' \text{ misses } r_i\}]$$

$$\leq 4(1 - \frac{\epsilon}{4})^m \leq 4e^{-\frac{m\epsilon}{4}}.$$



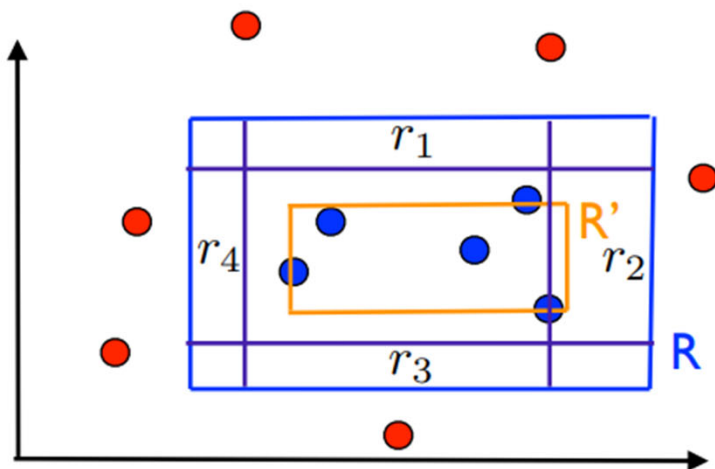
Bound Difference In Sample and Out of Sample

- Set $\delta > 0$ to match the upper bound:

$$4e^{-\frac{m\epsilon}{4}} \leq \delta \Leftrightarrow m \geq \frac{4}{\epsilon} \log \frac{4}{\delta}.$$

- Then, for $m \geq \frac{4}{\epsilon} \log \frac{4}{\delta}$, with probability at least $1 - \delta$,

$$R(R') \leq \epsilon.$$



Minimize In Sample Error

PERCEPTRON(\mathbf{w}_0)

```
1   $\mathbf{w}_1 \leftarrow \mathbf{w}_0$      $\triangleright$  typically  $\mathbf{w}_0 = \mathbf{0}$ 
2  for  $t \leftarrow 1$  to  $T$  do
3      RECEIVE( $\mathbf{x}_t$ )
4       $\hat{y}_t \leftarrow \text{sgn}(\mathbf{w}_t \cdot \mathbf{x}_t)$ 
5      RECEIVE( $y_t$ )
6      if ( $\hat{y}_t \neq y_t$ ) then
7           $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t + y_t \mathbf{x}_t$      $\triangleright$  more generally  $\eta y_t \mathbf{x}_t, \eta > 0$ .
8      else  $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t$ 
9  return  $\mathbf{w}_{T+1}$ 
```

Minimize In Sample Error

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

Minimize In Sample Error

► Take

$$F(\mathbf{w}) = \frac{1}{T} \sum_{t=1}^T \max \left(0, -y_t(\mathbf{w} \cdot \mathbf{x}_t) \right)$$

Minimize In Sample Error

► Take

$$F(\mathbf{w}) = \frac{1}{T} \sum_{t=1}^T \max \left(0, -y_t(\mathbf{w} \cdot \mathbf{x}_t) \right)$$

► Set

$$\tilde{F}(\mathbf{w}, \mathbf{x}) = \max \left(0, -f(\mathbf{x})(\mathbf{w} \cdot \mathbf{x}) \right)$$

Minimize In Sample Error

- Take

$$F(\mathbf{w}) = \frac{1}{T} \sum_{t=1}^T \max \left(0, -y_t(\mathbf{w} \cdot \mathbf{x}_t) \right)$$

- Set

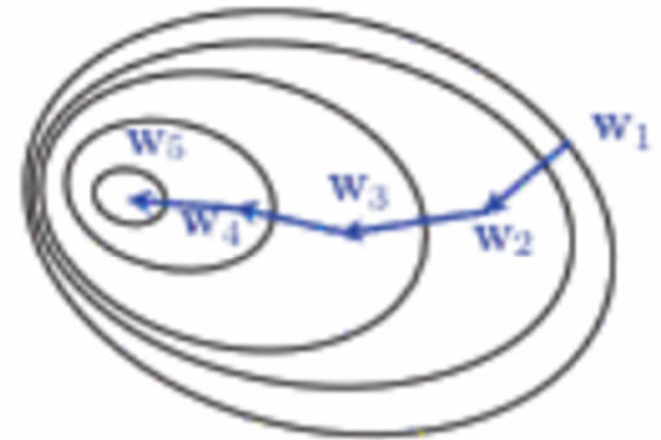
$$\tilde{F}(\mathbf{w}, \mathbf{x}) = \max \left(0, -f(\mathbf{x})(\mathbf{w} \cdot \mathbf{x}) \right)$$

- Update Guess

$$\mathbf{w}_{t+1} \leftarrow \begin{cases} \mathbf{w}_t - \eta \nabla_{\mathbf{w}} \tilde{F}(\mathbf{w}_t, \mathbf{x}_t) & \text{if } \mathbf{w} \mapsto \tilde{F}(\mathbf{w}, \mathbf{x}_t) \text{ differentiable at } \mathbf{w}_t \\ \mathbf{w}_t & \text{otherwise,} \end{cases}$$

Minimize In Sample Error

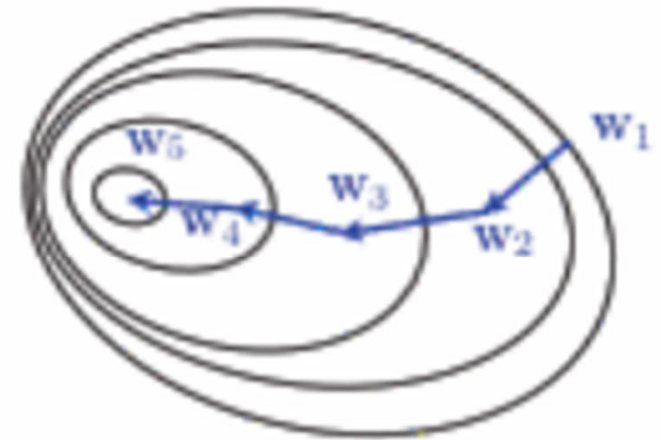
$$\mathbf{w}_{t+1} \leftarrow \begin{cases} \mathbf{w}_t + \eta y_t \mathbf{x}_t & \text{if } y_t(\mathbf{w} \cdot \mathbf{x}_t) < 0; \\ \mathbf{w}_t & \text{if } y_t(\mathbf{w} \cdot \mathbf{x}_t) > 0; \\ \mathbf{w}_t & \text{otherwise,} \end{cases}$$



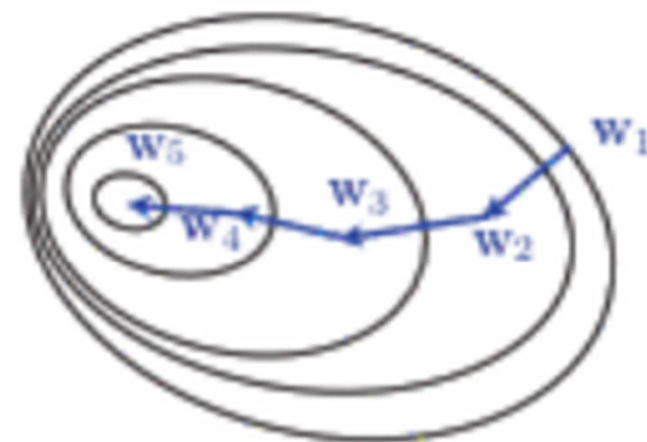
Minimize In Sample Error

$$\mathbf{w}_{t+1} \leftarrow \begin{cases} \mathbf{w}_t + \eta y_t \mathbf{x}_t & \text{if } y_t(\mathbf{w} \cdot \mathbf{x}_t) < 0; \\ \mathbf{w}_t & \text{if } y_t(\mathbf{w} \cdot \mathbf{x}_t) > 0; \\ \mathbf{w}_t & \text{otherwise,} \end{cases}$$

$$\nabla_{\mathbf{w}} \tilde{F}(\mathbf{w}, \mathbf{x}_t) = -y_t \mathbf{x}_t \quad \text{if } y_t(\mathbf{w} \cdot \mathbf{x}_t) < 0$$



Minimize In Sample Error



$$\mathbf{w}_{t+1} \leftarrow \begin{cases} \mathbf{w}_t + \eta y_t \mathbf{x}_t & \text{if } y_t(\mathbf{w} \cdot \mathbf{x}_t) < 0; \\ \mathbf{w}_t & \text{if } y_t(\mathbf{w} \cdot \mathbf{x}_t) > 0; \\ \mathbf{w}_t & \text{otherwise,} \end{cases}$$

$$\nabla_{\mathbf{w}} \tilde{F}(\mathbf{w}, \mathbf{x}_t) = -y_t \mathbf{x}_t \quad \text{if } y_t(\mathbf{w} \cdot \mathbf{x}_t) < 0$$

$$\nabla_{\mathbf{w}} \tilde{F}(\mathbf{w}, \mathbf{x}_t) = 0 \quad \text{if } y_t(\mathbf{w} \cdot \mathbf{x}_t) > 0.$$

Minimize In Sample Error

- Use derivatives to approximate a function by a linear function

$$f(\mathbf{w}) \approx f(\mathbf{w}^{(l)}) + \langle \mathbf{w} - \mathbf{w}^{(l)}, \nabla f(\mathbf{w}^{(l)}) \rangle$$

Minimize In Sample Error

- Use derivatives to approximate a function by a linear function

$$f(\mathbf{w}) \approx f(\mathbf{w}^{(t)}) + \langle \mathbf{w} - \mathbf{w}^{(t)}, \nabla f(\mathbf{w}^{(t)}) \rangle$$

- Approximation inaccurate for far away points

$$\mathbf{w}^{(t+1)} = \operatorname{argmin}_{\mathbf{w}} \frac{1}{2} \|\mathbf{w} - \mathbf{w}^{(t)}\|^2 + \eta \left(f(\mathbf{w}^{(t)}) + \langle \mathbf{w} - \mathbf{w}^{(t)}, \nabla f(\mathbf{w}^{(t)}) \rangle \right).$$

Minimize In Sample Error

- Use derivatives to approximate a function by a linear function

$$f(\mathbf{w}) \approx f(\mathbf{w}^{(t)}) + \langle \mathbf{w} - \mathbf{w}^{(t)}, \nabla f(\mathbf{w}^{(t)}) \rangle$$

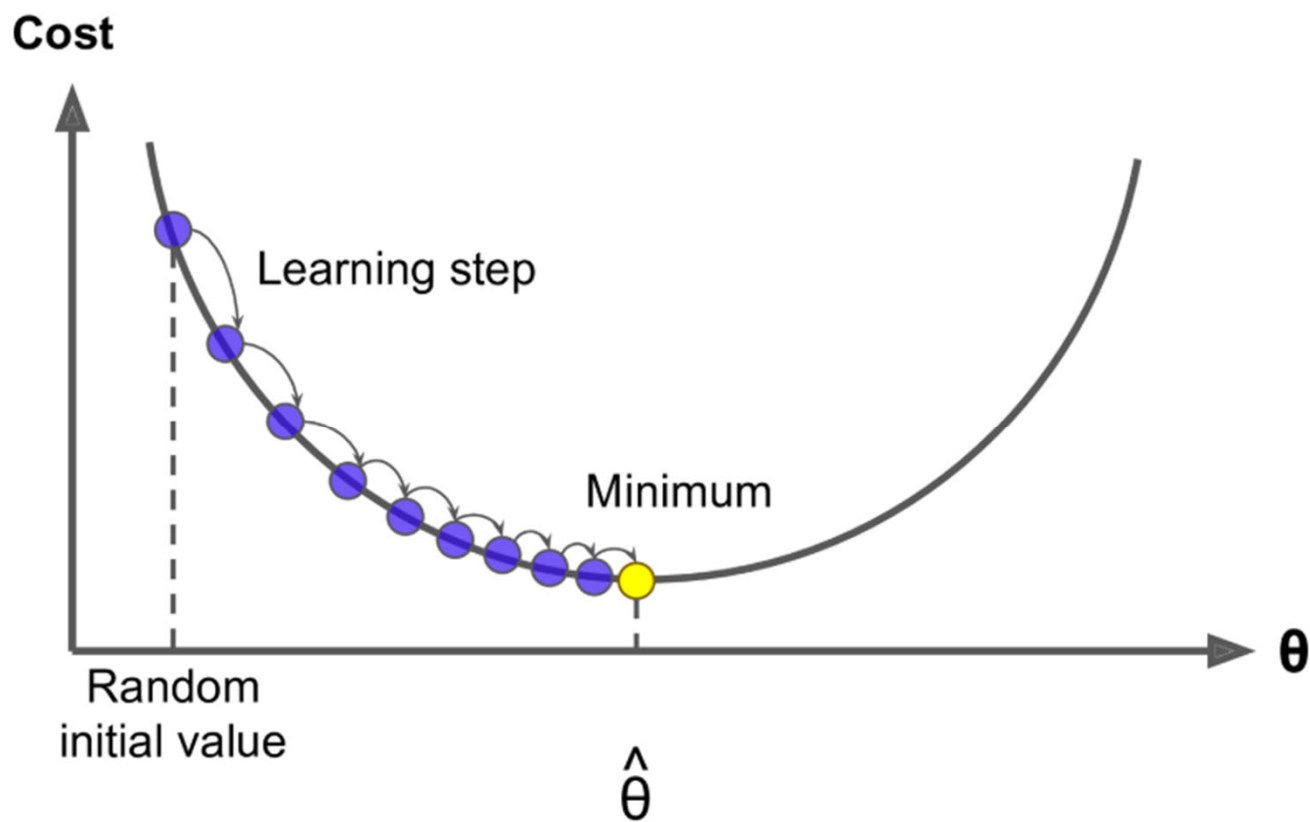
- Approximation inaccurate for far away points

$$\mathbf{w}^{(t+1)} = \operatorname{argmin}_{\mathbf{w}} \frac{1}{2} \|\mathbf{w} - \mathbf{w}^{(t)}\|^2 + \eta \left(f(\mathbf{w}^{(t)}) + \langle \mathbf{w} - \mathbf{w}^{(t)}, \nabla f(\mathbf{w}^{(t)}) \rangle \right).$$

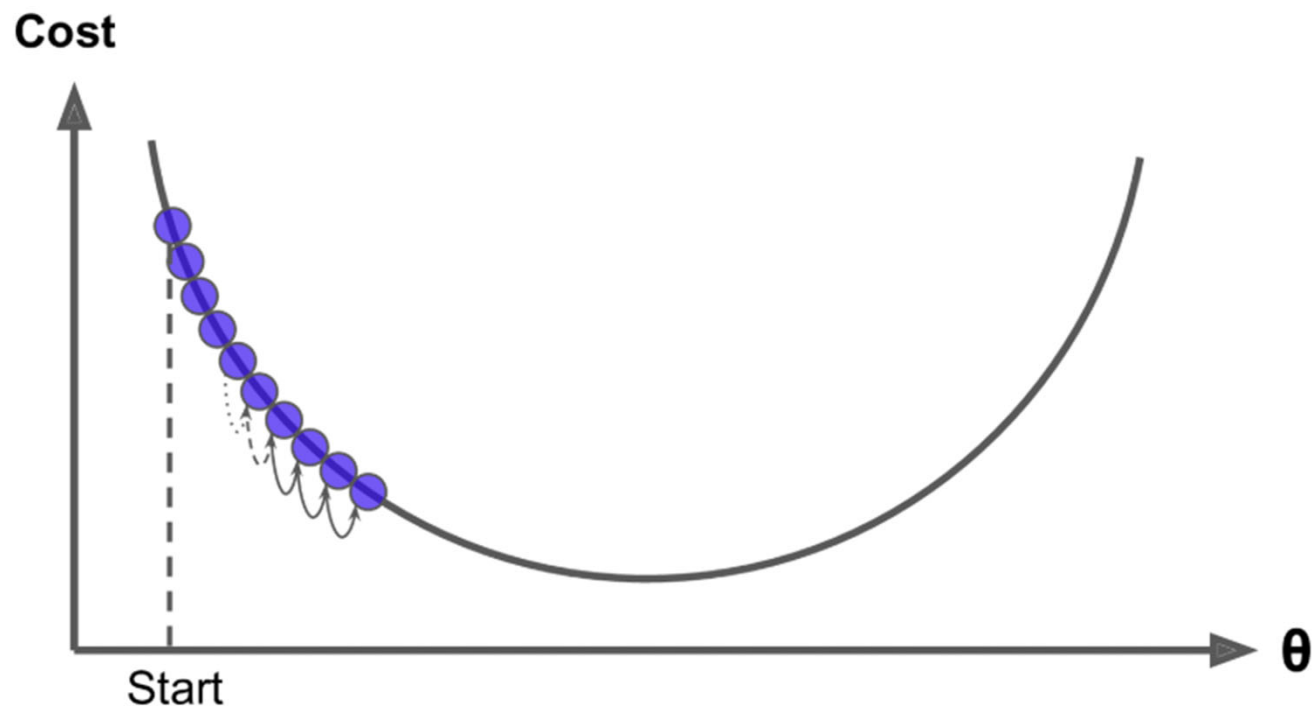
- Learning Rate controls the trade-off by determining the step size

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta \nabla f(\mathbf{w}^{(t)}),$$

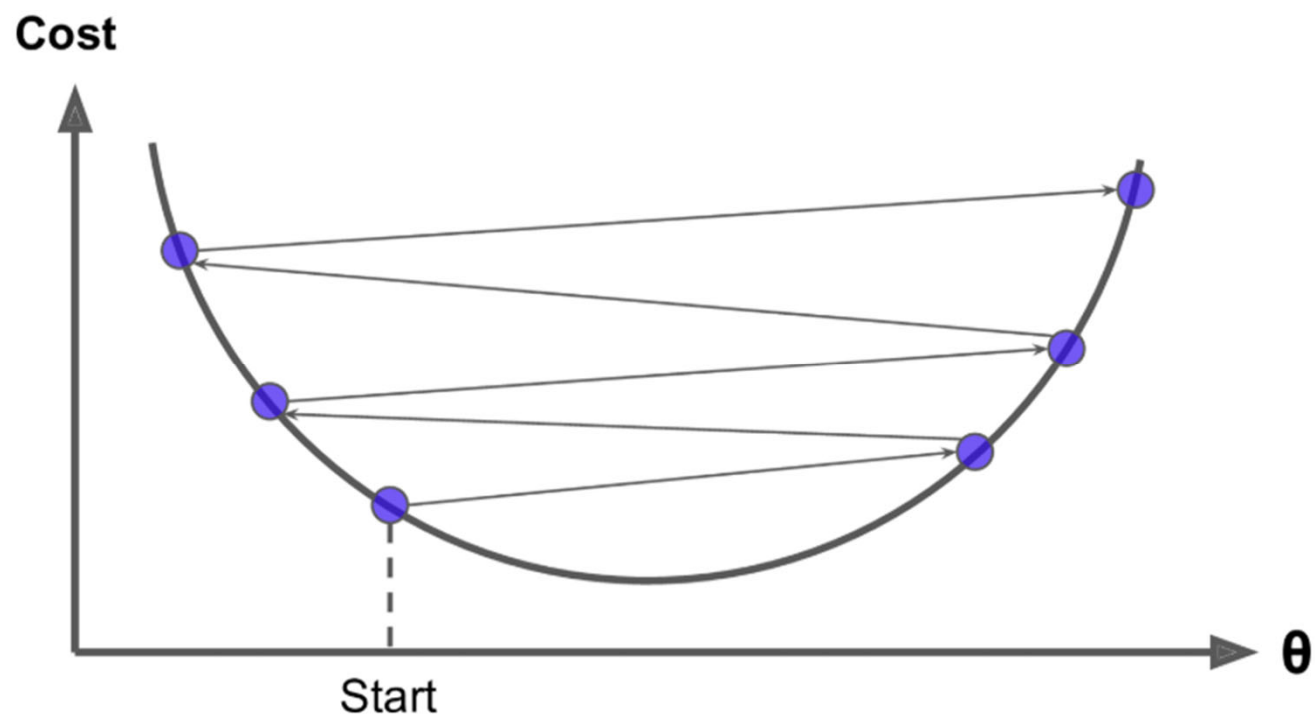
Minimize In Sample Error



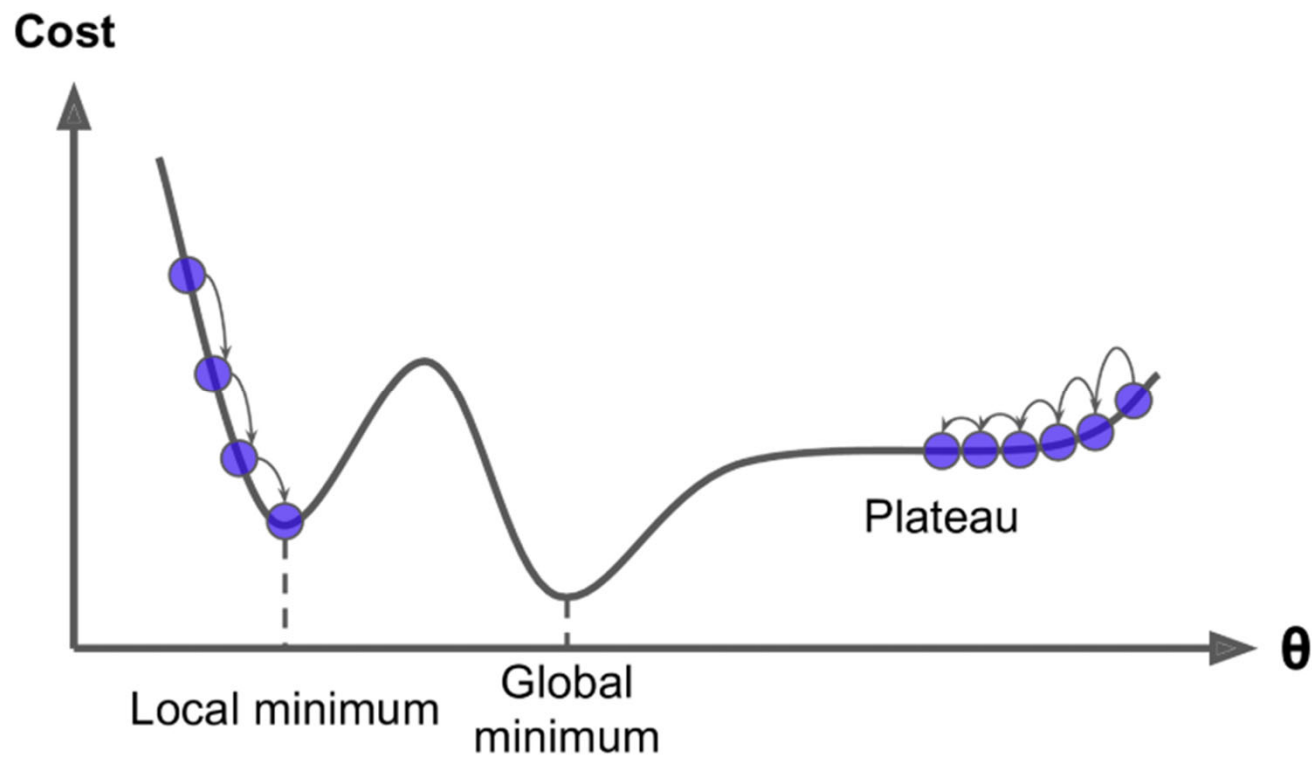
Minimize In Sample Error



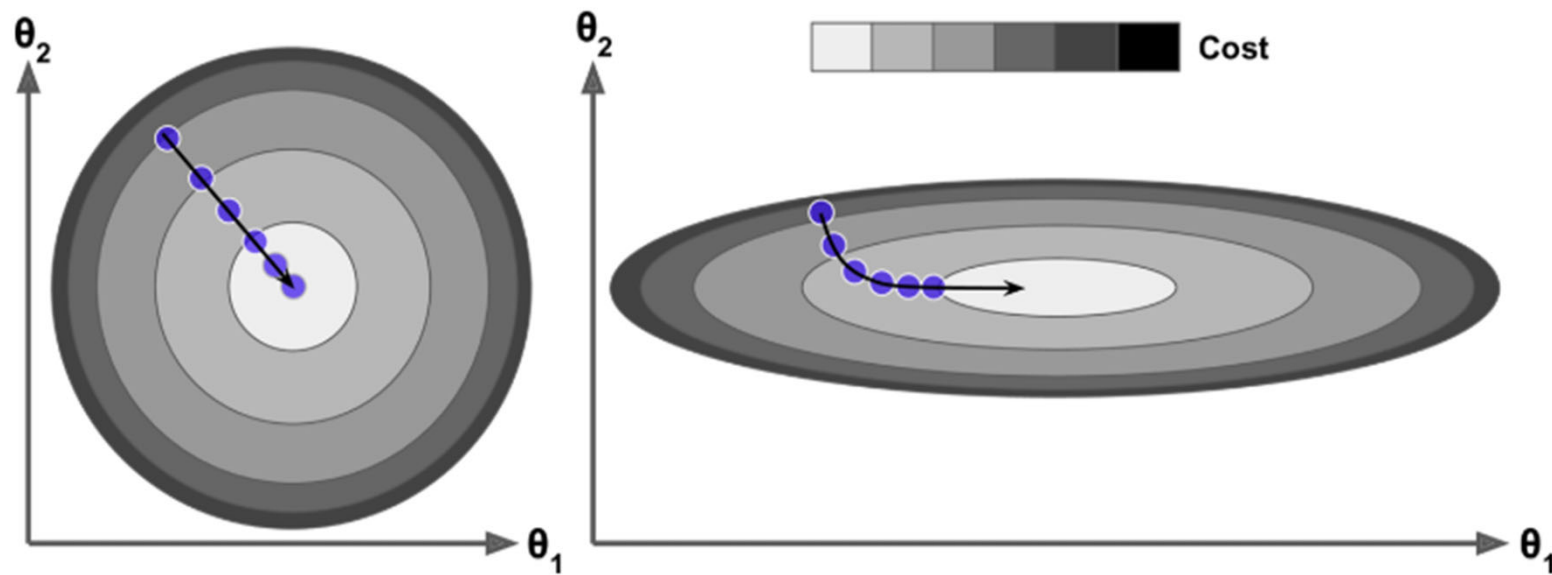
Minimize In Sample Error



Minimize In Sample Error



Minimize In Sample Error



Take-Aways

- ▶ What are excess risk, approximation error, estimation error, and optimization error.
- ▶ For nested hypothesis spaces, say $H_1 \subset H_2$. Explain how we would expect the approximation error and estimation error to change between them?
- ▶ Why could optimization error be negative but estimation error can never be negative?

Take-Aways

- ▶ Write the empirical risk for a particular loss function over a particular hypothesis space, such as for square loss over a hypothesis space of linear functions.
- ▶ Compare and contrast gradient descent and stochastic gradient descent.