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

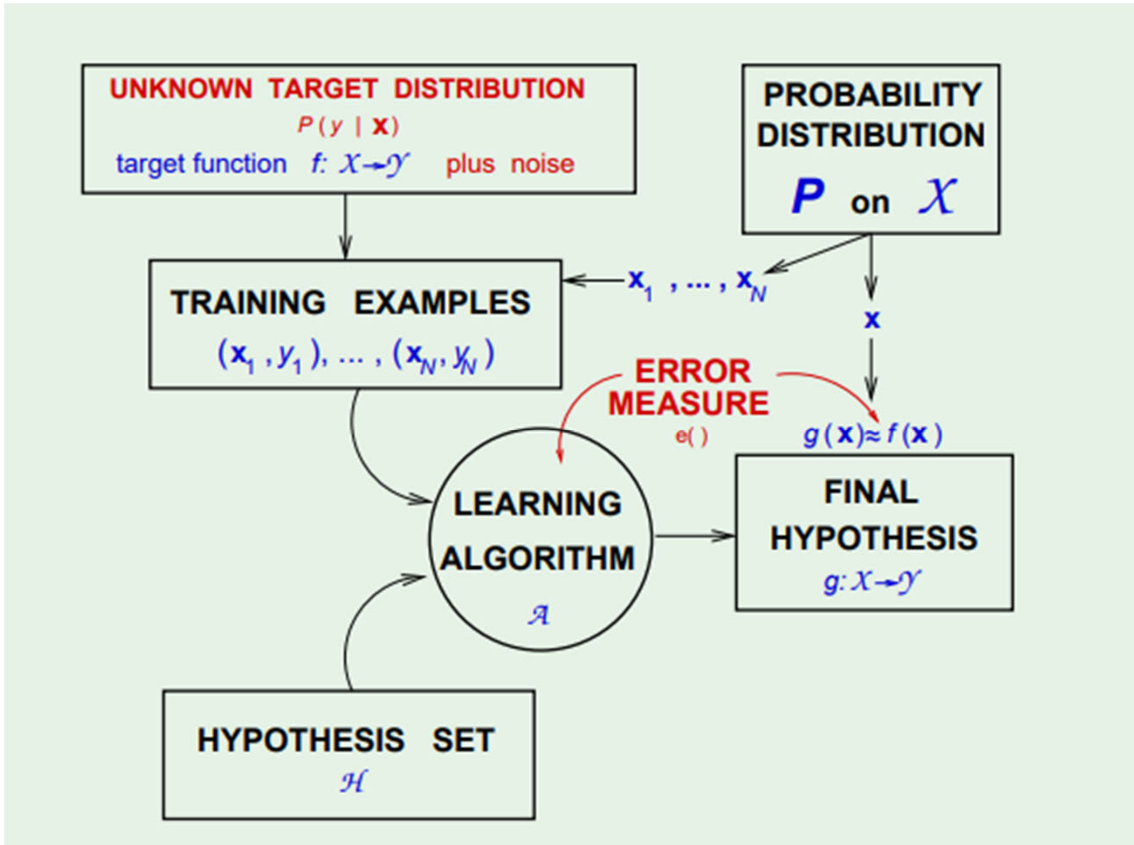
Agenda

- 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:

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

Out-of-sample error:

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

Hypotheses and Loss Functions

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

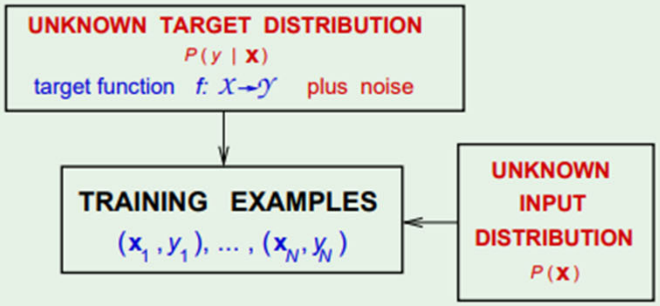
is what we are trying to learn

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



Question

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

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

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

(c) $\ell(a, y) = \mathbf{1}(a \neq y)$.

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?

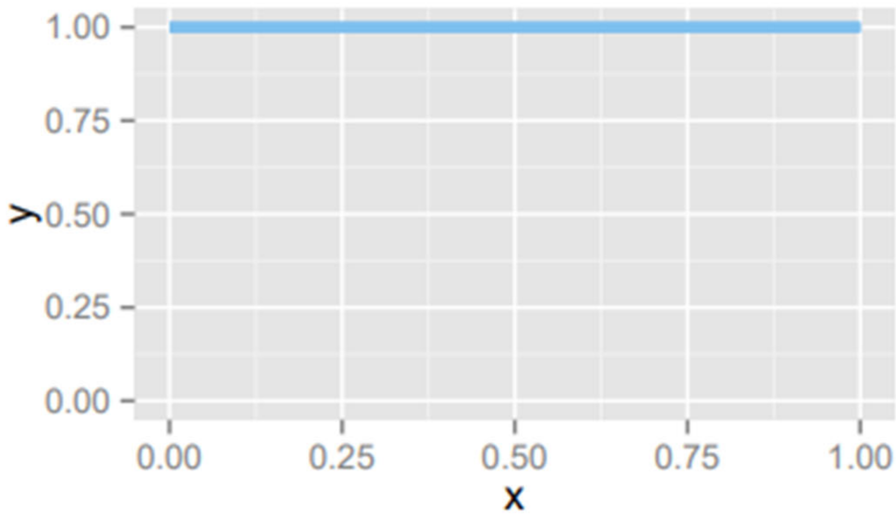
Hypotheses and Loss Functions

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

Overfitting and Underfitting

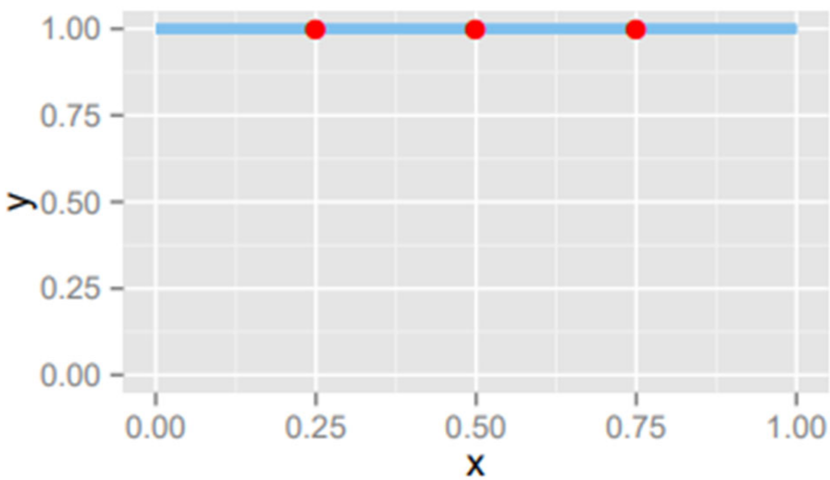
$P_{\mathcal{X}} = \text{Uniform}[0,1]$, $Y \equiv 1$ (i.e. Y is always 1).



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

Overfitting and Underfitting

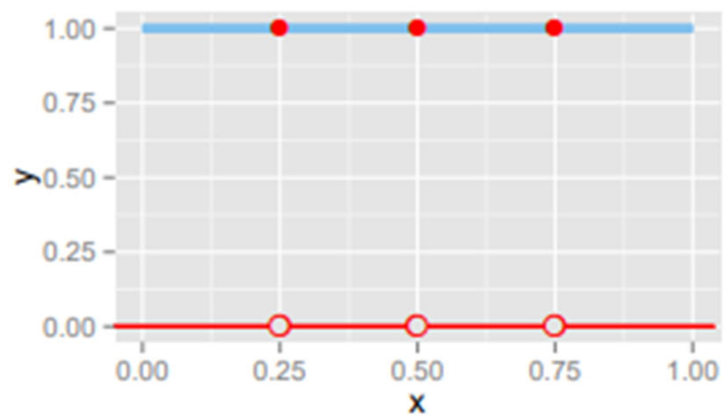
$P_{\mathcal{X}} = \text{Uniform}[0,1]$, $Y \equiv 1$ (i.e. Y is always 1).



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).

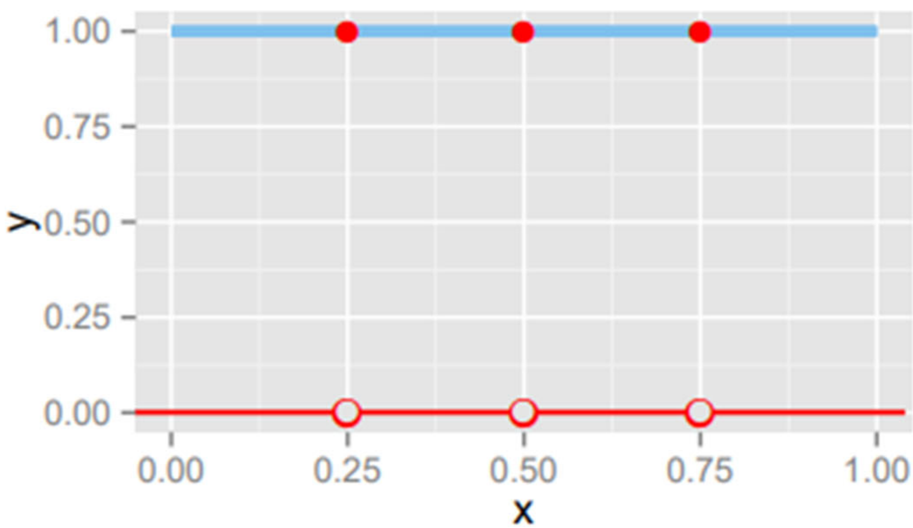


A proposed prediction function:

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

Overfitting and Underfitting

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



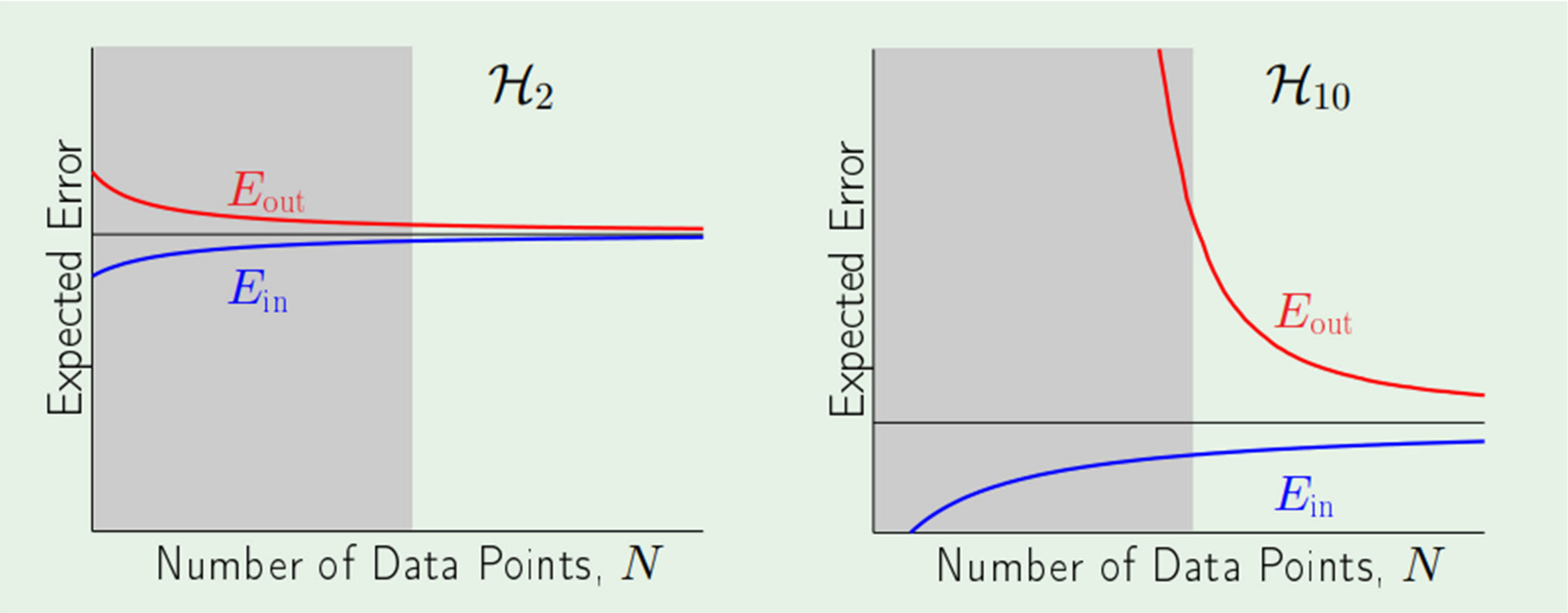
Under square loss or 0/1 loss: $\hat{f}$ has Empirical Risk $= 0$ and Risk $= 1$.

Hypotheses and Loss Functions

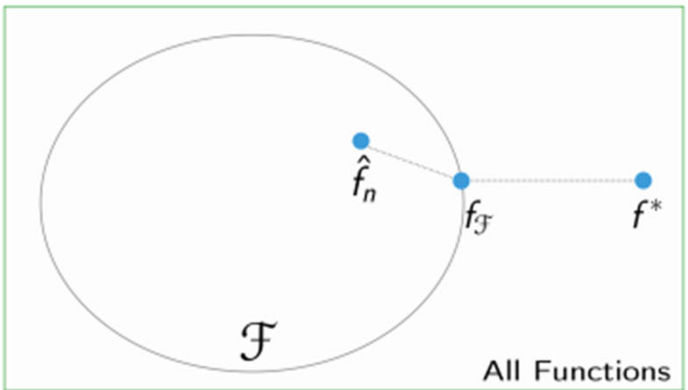
1. Can we make sure that $E_{\text{out}}(g)$ is close enough to $E_{\text{in}}(g)$?
2. Can we make $E_{\text{in}}(g)$ small enough?

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

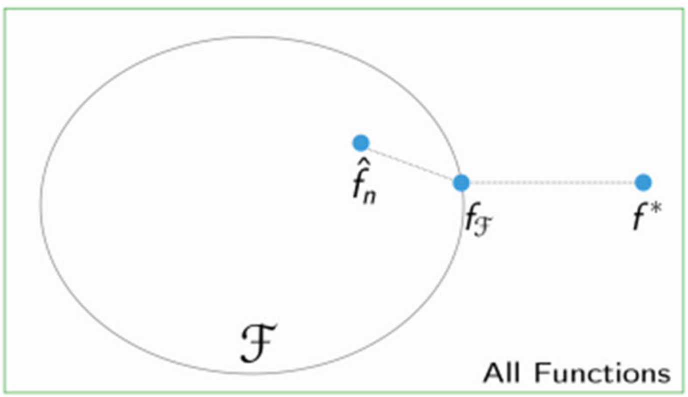
Overfitting and Underfitting



Finding Accurate Hypotheses

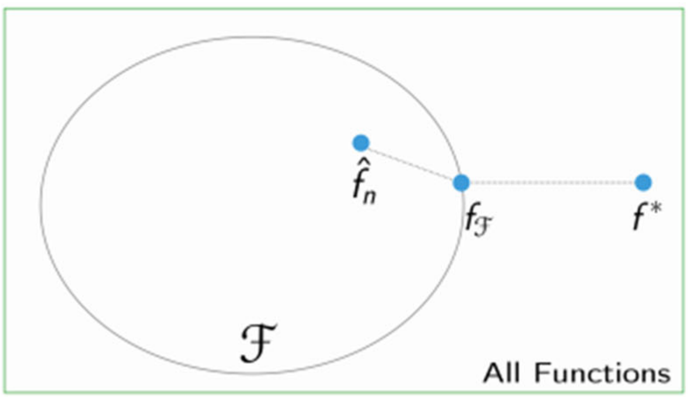


Finding Accurate Hypotheses



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

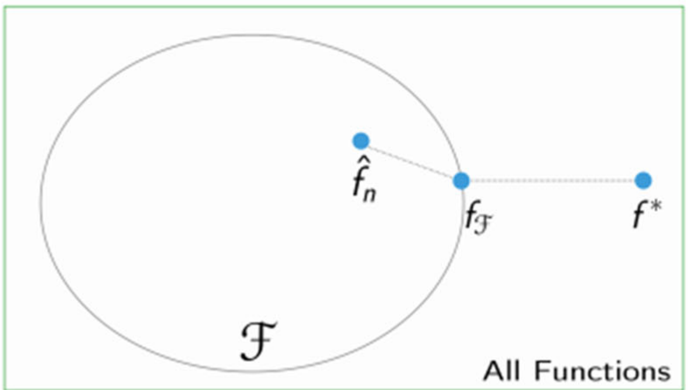
Finding Accurate Hypotheses



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

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

Finding Accurate Hypotheses

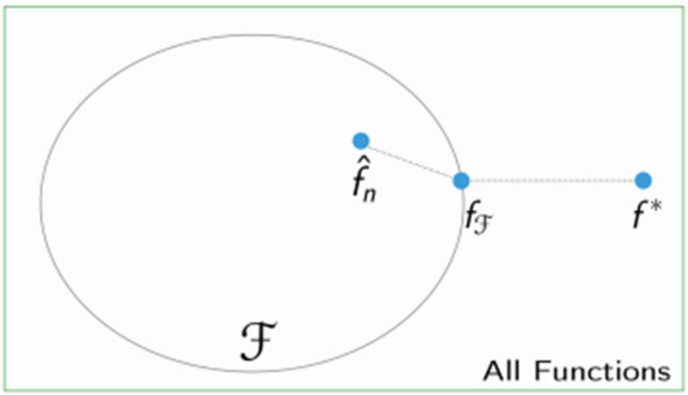


$$f^* = \arg\min_{f} \mathbb{E}\ell(f(x), y)$$

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

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

Finding Accurate Hypotheses



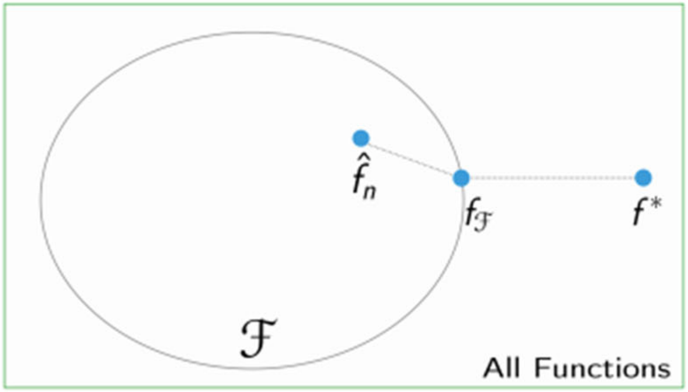
$$f^* = \arg\min_{f} \mathbb{E}\ell(f(x), y)$$

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

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

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

Finding Accurate Hypotheses



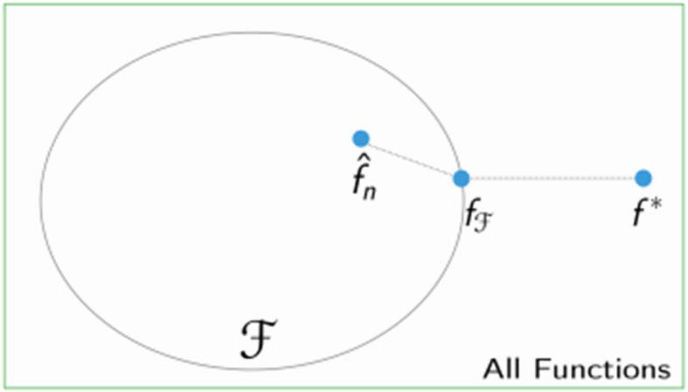
$$f^* = \arg\min_{f} \mathbb{E}\ell(f(x), y)$$

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

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

- **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



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

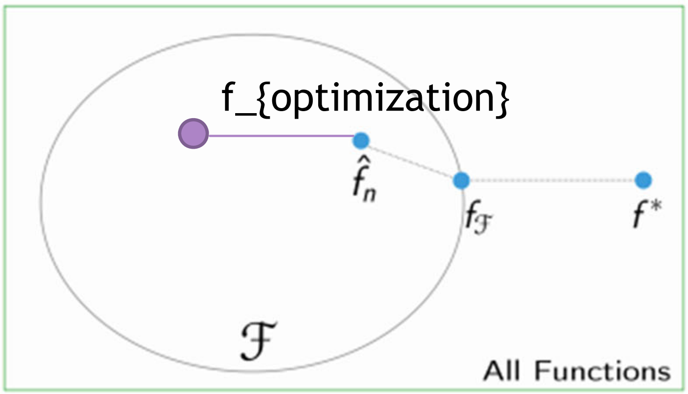
$$f_{\mathcal{F}} = \arg\min_{f \in \mathcal{F}} \mathbb{E}\ell(f(x), y))$$

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

- **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



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

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

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

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

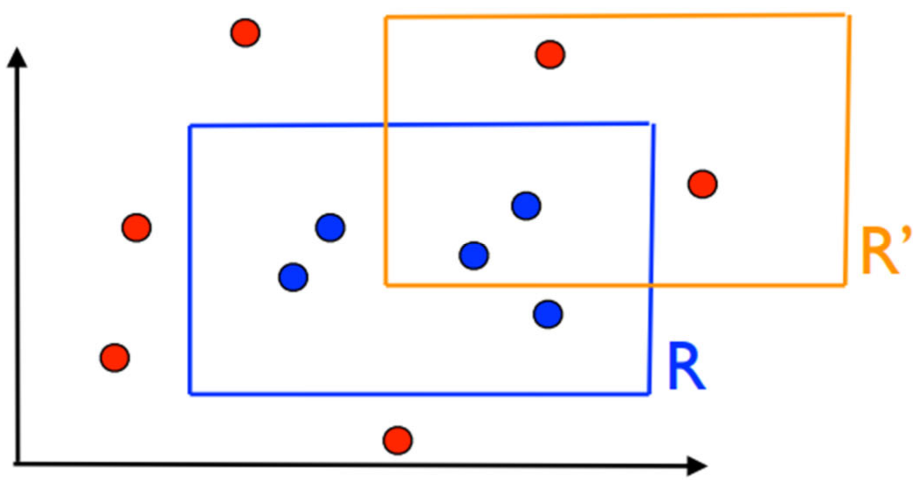
$$\begin{aligned} \text{Excess Risk}(\hat{f}_n) &= R(\hat{f}_n) - R(f^*) \\ &= \underbrace{R(\hat{f}_n) - R(f_{\mathcal{F}})}_{\text{estimation error}} + \underbrace{R(f_{\mathcal{F}}) - R(f^*)}_{\text{approximation error}}. \end{aligned}$$

Question

- Review
- Lesson
- Demo

Bound Difference In Sample and Out of Sample

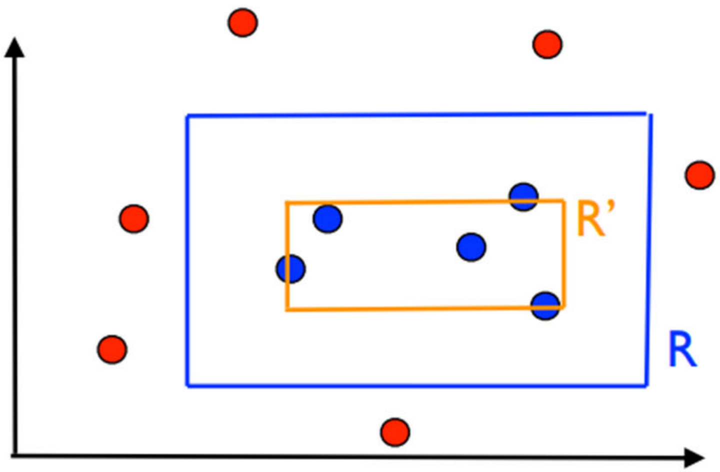
- Problem: learn unknown axis-aligned rectangle R using as small a labeled sample as possible.



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

Bound Difference In Sample and Out of Sample

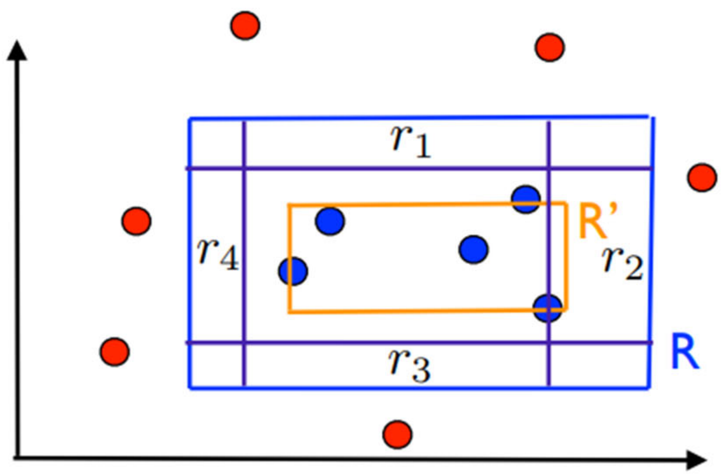
- Simple method: choose tightest consistent rectangle R' for a large enough sample. How large a sample?



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

Bound Difference In Sample and Out of Sample

- Fix $\epsilon > 0$ and assume $\Pr_D[\mathrm{R}] > \epsilon$ (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 $\Pr_D[r_i] \geq \frac{\epsilon}{4}$.



$$\mathrm{R} = [l, r] \times [b, t]$$
$$r_4 = [l, s_4] \times [b, t]$$
$$s_4 = \inf\{s \colon \Pr\left[[l, s] \times [b, t]\right] \geq \tfrac{\epsilon}{4}\}$$
$$\Pr_D\left[[l, s_4[\times [b, t]\right] < \tfrac{\epsilon}{4}$$

Bound Difference In Sample and Out of Sample

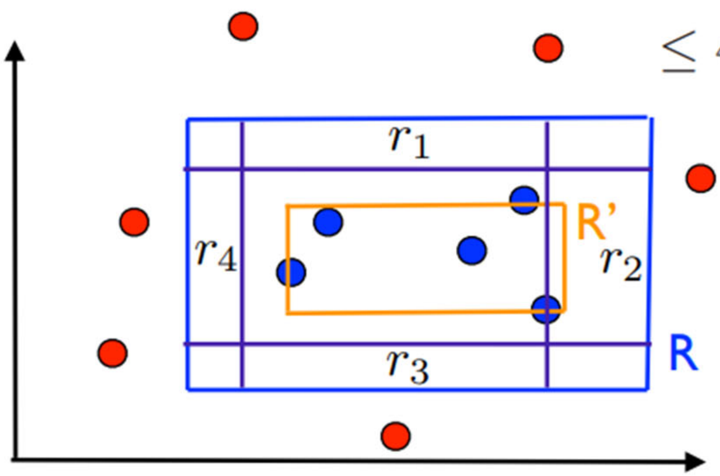
- Errors can only occur in $\mathrm{R}-\mathrm{R}'$. Thus (geometry),

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

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

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

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



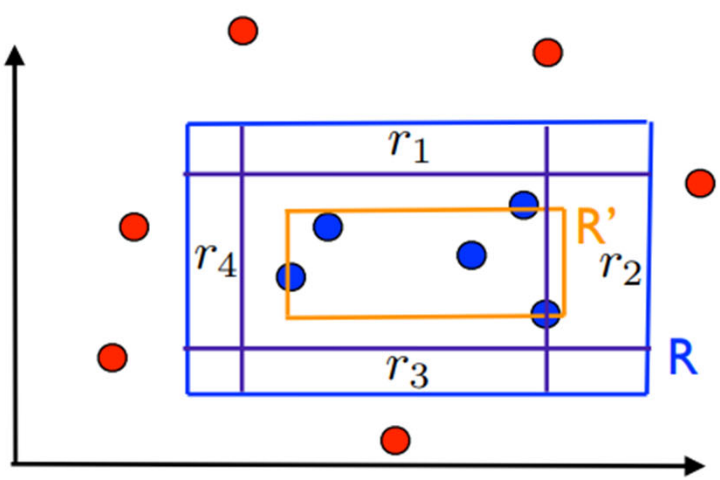
Bound Difference In Sample and Out of Sample

- 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 $1-\delta$,

$$R(\mathrm{R}') \leq \epsilon.$$



Minimize In Sample Error

$\textsc{Perceptron}(\mathbf{w}_0)$

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

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

$\quad\textsc{Receive}(\mathbf{x}_t)$

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

$\quad\textsc{Receive}(y_t)$

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

$\qquad\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t + y_t\mathbf{x}_t$ $\quad\triangleright$ more generally $\eta y_t \mathbf{x}_t, \eta > 0.$

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

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

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

Minimize In Sample Error

- Take

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

- Set

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

Minimize In Sample Error

- Take

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

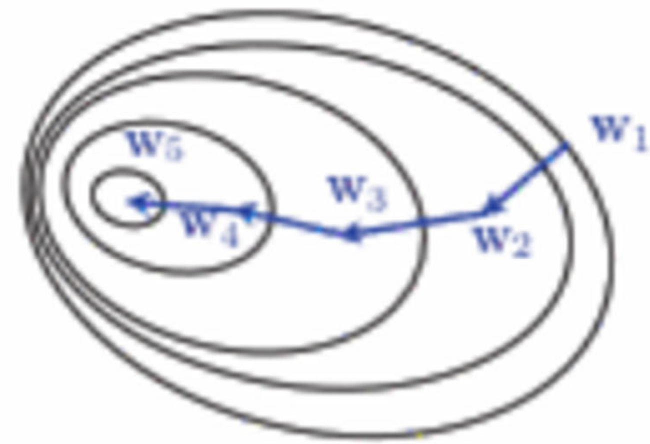
- Set

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

- Update Guess

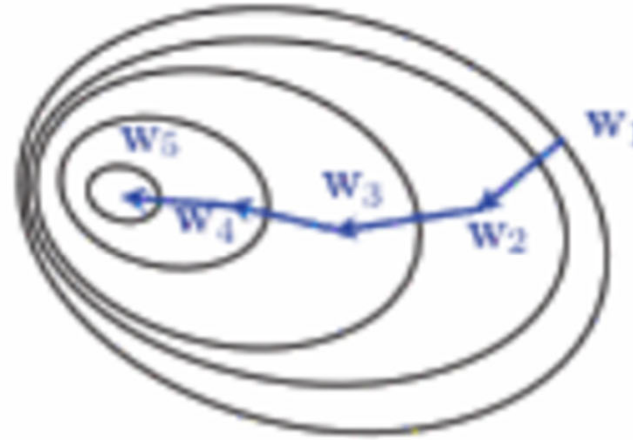
$$\mathbf{w}_{t+1} \leftarrow \begin{cases} \mathbf{w}_t - \eta\nabla_{\mathbf{w}}\widetilde{F}(\mathbf{w}_t,\mathbf{x}_t) & \text{if } \mathbf{w}\mapsto\widetilde{F}(\mathbf{w},\mathbf{x}_t) \text{ differentiable at } \mathbf{w}_t \\ \mathbf{w}_t & \text{otherwise,} \end{cases}$$

Minimize In Sample Error



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

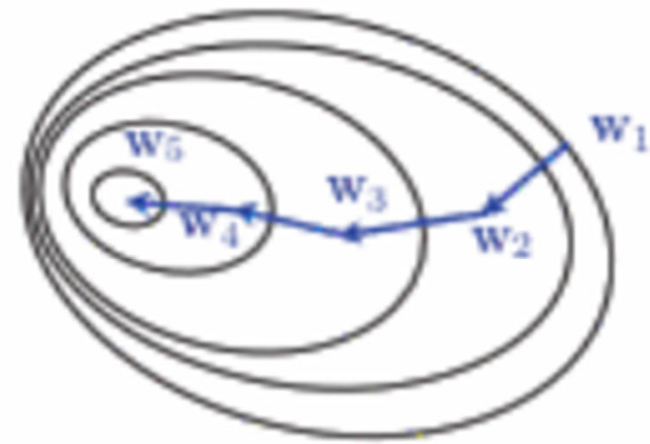
Minimize In Sample Error



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

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

Minimize In Sample Error



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

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

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

Minimize In Sample Error

- Use derivatives to approximate a function by a linear function

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

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

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

Minimize In Sample Error

- Use derivatives to approximate a function by a linear function

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

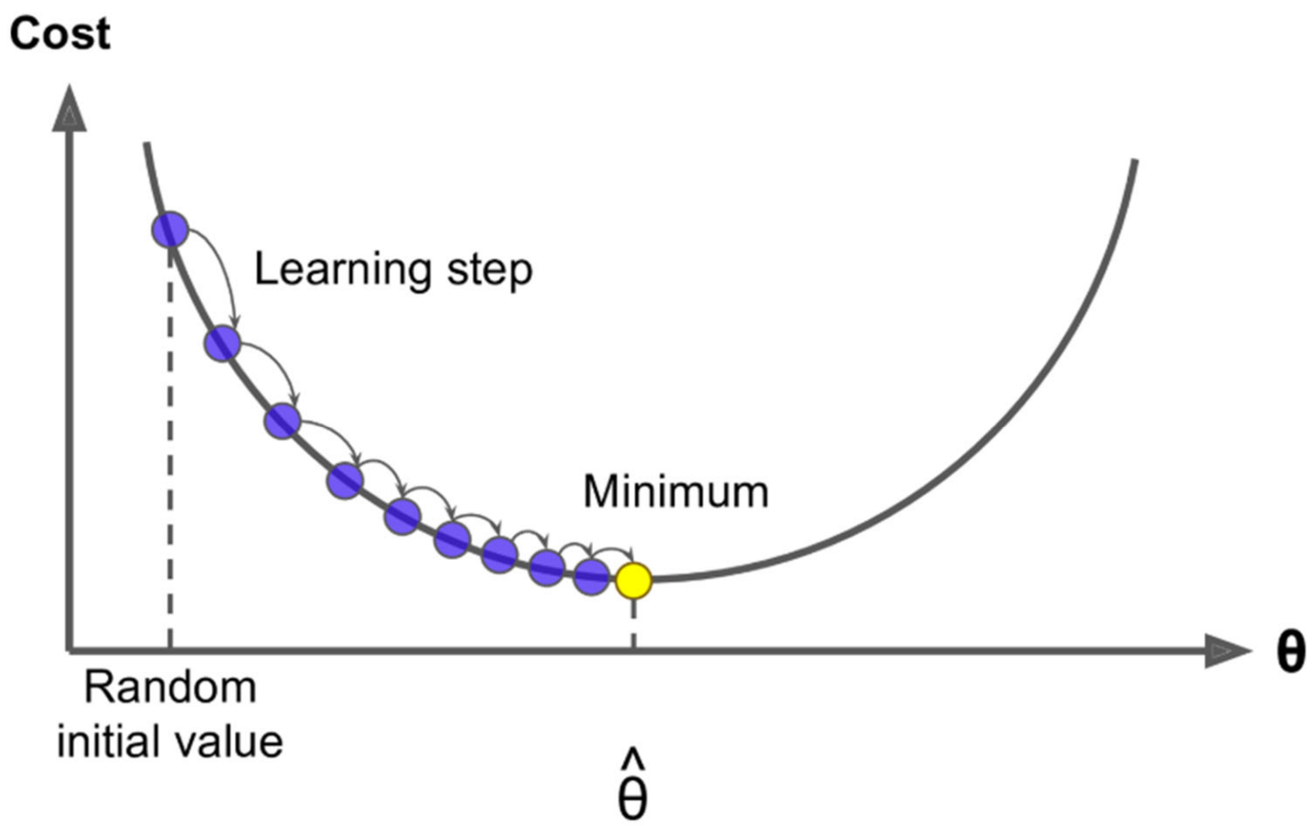
- Approximation inaccurate for far away points

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

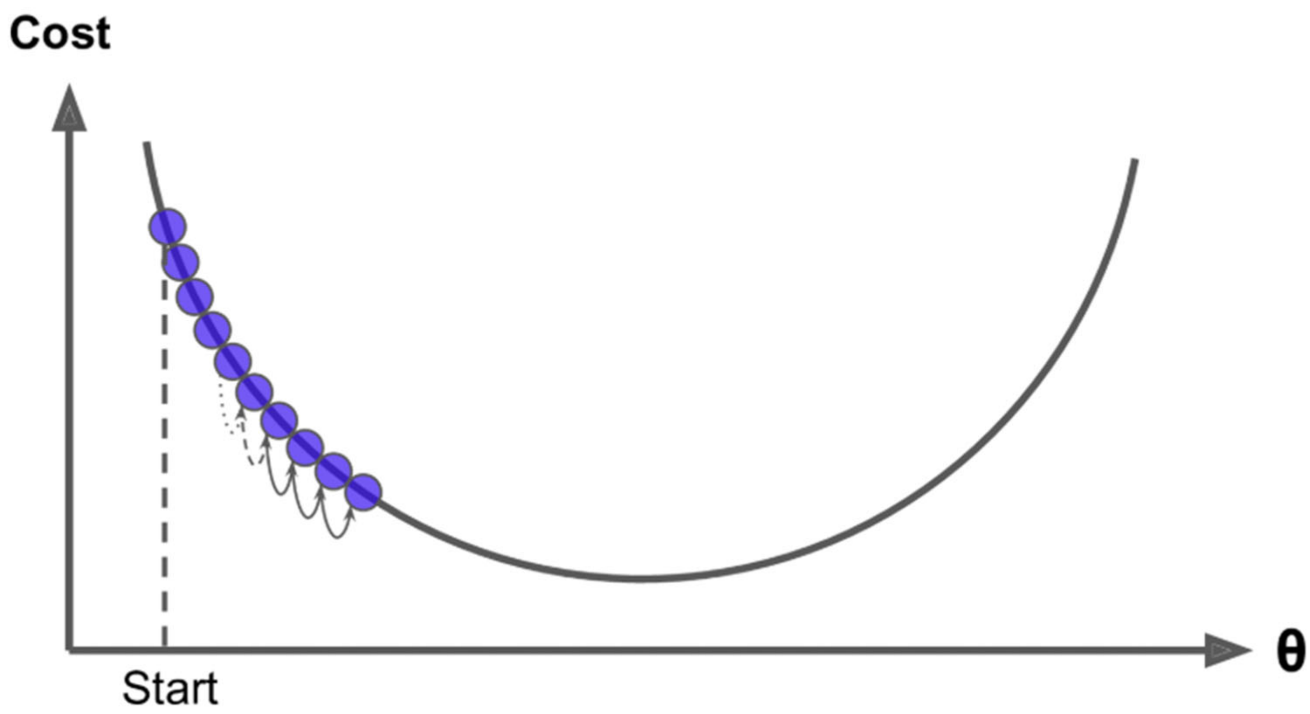
- Learning Rate controls the trade-off by determining the step size

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

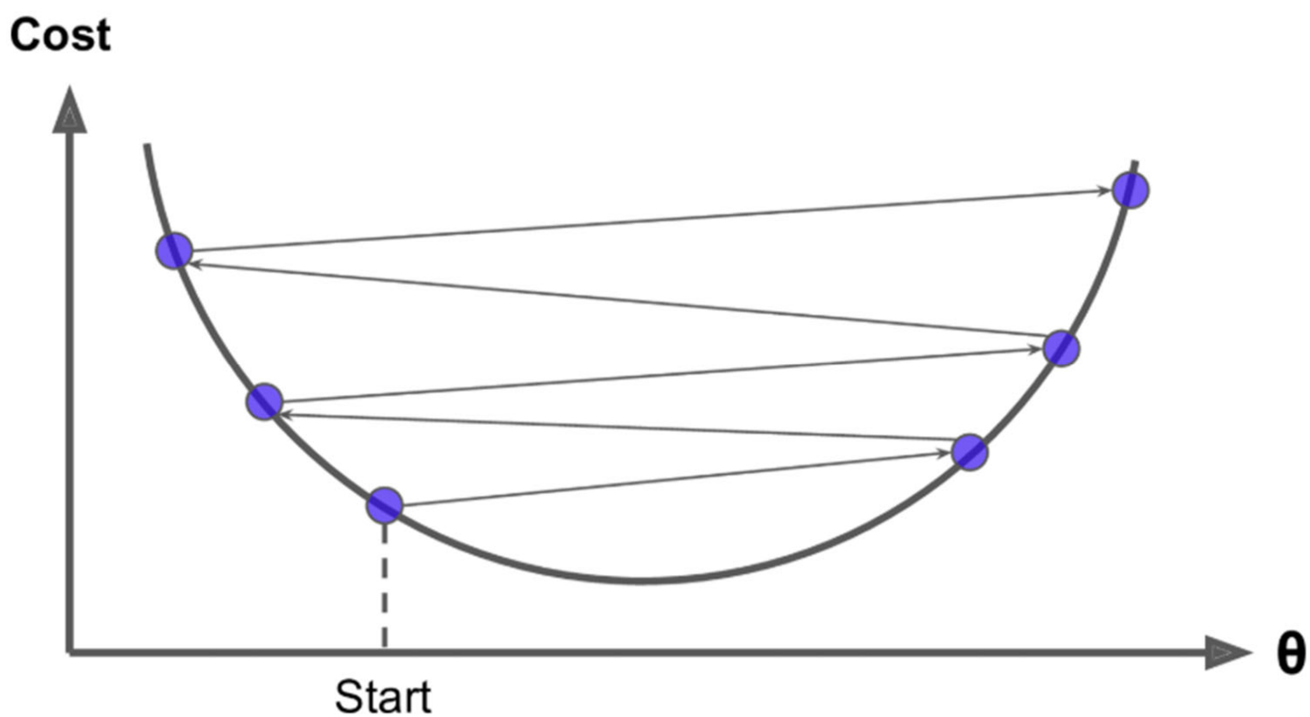
Minimize In Sample Error



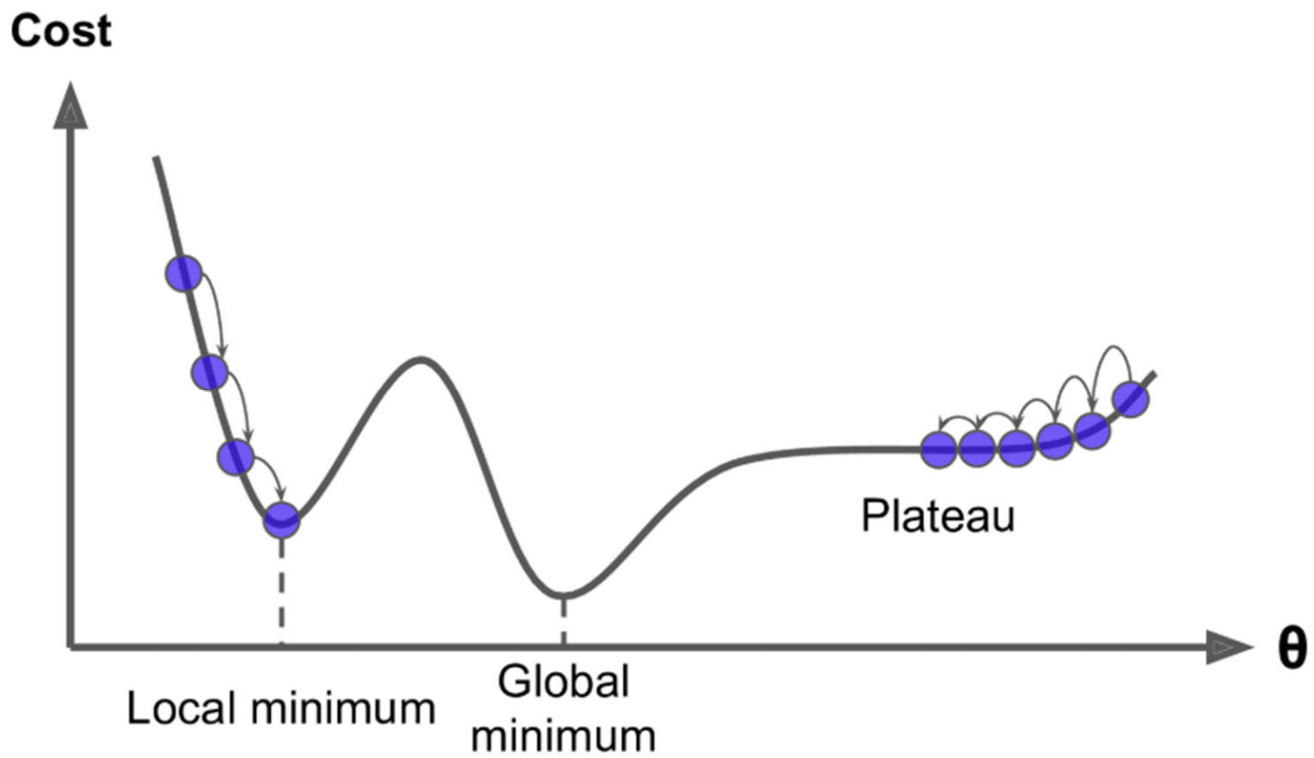
Minimize In Sample Error



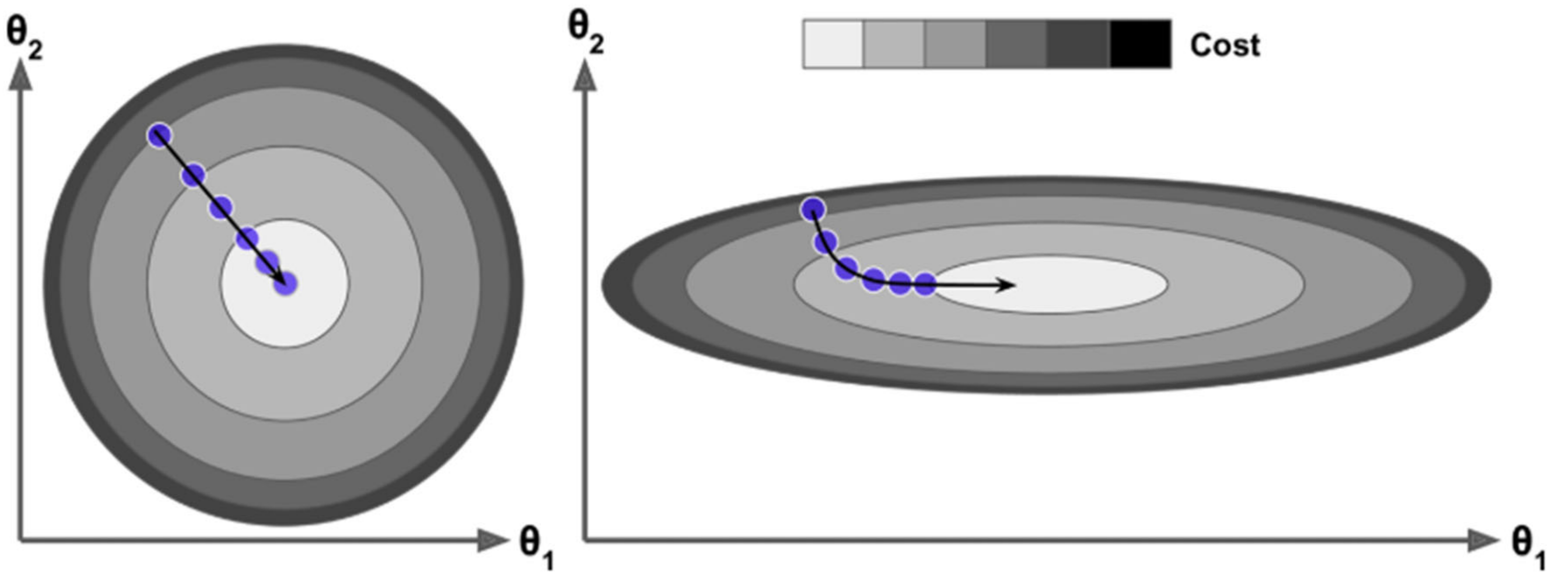
Minimize In Sample Error



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 H1 H2. 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.