

COMP5211: Machine Learning

Lecture 7

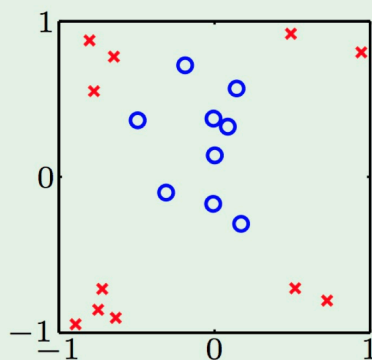
Minhao Cheng

From last lecture

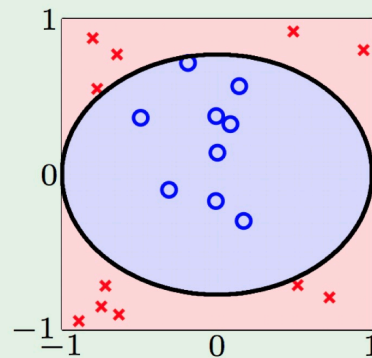
Linear hypotheses

- Up to now: linear hypotheses
 - Perception, Linear regression, Logistic regression, ...
- Many problems are not linearly separable

Data:



Hypothesis:

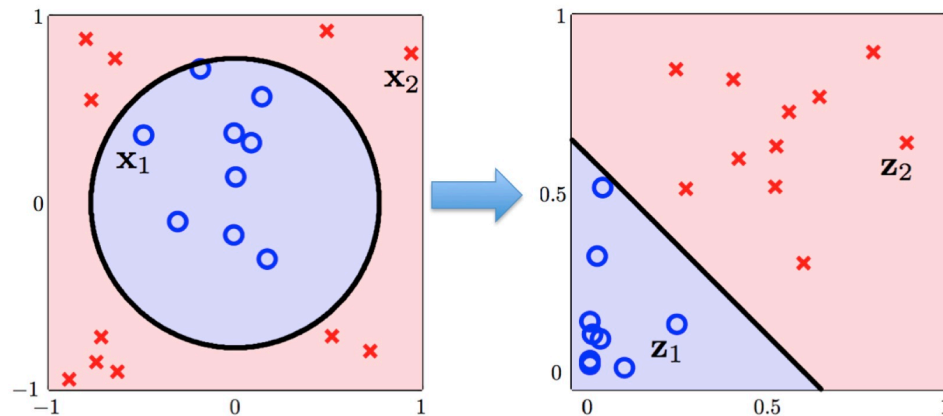


Nonlinear transformation

Circular Separable and Linear Separable

$$h(x) = \text{sign}(\underbrace{0.6}_{\tilde{w}_0} \cdot \underbrace{1}_{\tilde{z}_0} + \underbrace{(-1)}_{\tilde{w}_1} \cdot \underbrace{x_1^2}_{\tilde{z}_1} + \underbrace{(-1)}_{\tilde{w}_2} \cdot \underbrace{x_2^2}_{\tilde{z}_2})$$

- $= \text{sign}(\tilde{w}^T \tilde{z})$
- $\{(x_n, y_n)\}$ circular separable \Rightarrow
 $\{(z_n, y_n)\}$ linear separable
- $x \in \mathcal{X} \rightarrow x \in \mathcal{Z}$ (using a
nonlinear transformation ϕ)



Nonlinear Transformation

Definition

- Define nonlinear transformation
 - $\phi(\mathbf{x}) = (1, x_1^2, x_2^2) = (z_0, z_1, z_2) = \mathbf{z}$
- Linear hypotheses in \mathcal{Z} -space:
 - $\text{sign}(\tilde{h}(\mathbf{z})) = \text{sign}(\tilde{h}(\phi(\mathbf{x}))) = \text{sign}(w^T \phi(\mathbf{x}))$
- Line in \mathcal{Z} -space \Leftrightarrow some quadratic curves in \mathcal{X} -space

Nonlinear Transformation

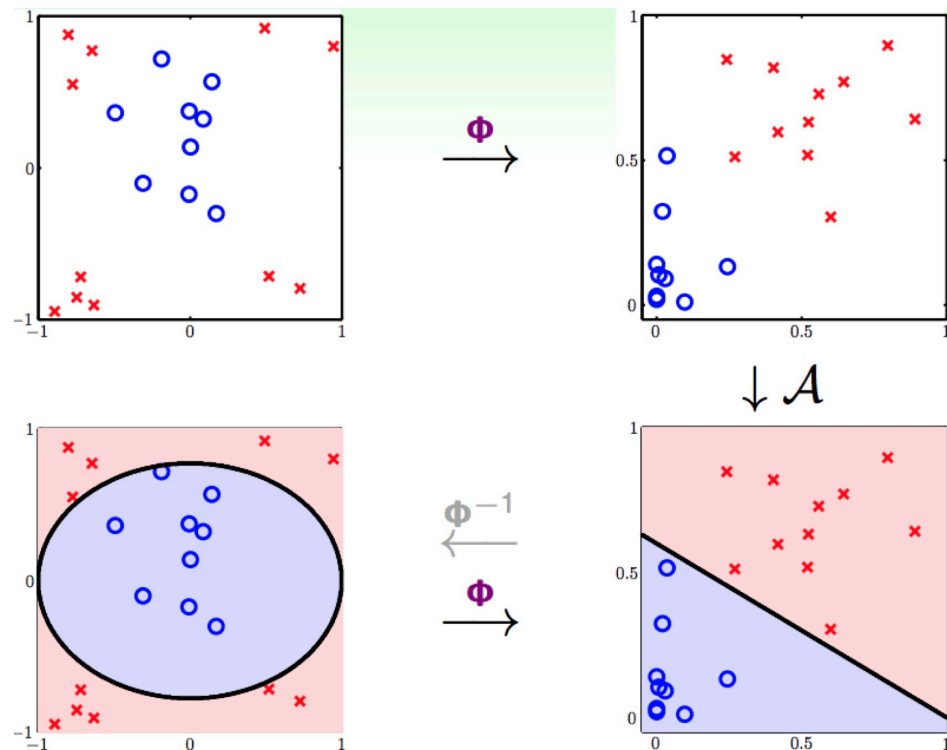
General Quadratic Hypothesis Set

- A “bigger ” \mathcal{Z} -space:
 - $\phi_2(\mathbf{x}) = (1, x_1, x_2, x_1^2, x_1x_2, x_2^2)$
- Linear in \mathcal{Z} -space \Leftrightarrow quadratic hypotheses in \mathcal{X} -space
- The hypotheses space:
 - $\mathcal{H}_{\phi_2} = \{h(x) : h(x) = \tilde{w}^T \phi_2(x) \text{ for some } \tilde{w}\}$ (quadratic hypotheses)
- Also include linear model as a degenerate case

Nonlinear transformation

Learning a good quadratic function

- Transform original data $\{x_n, y_n\}$ to $\{z_n = \phi(x_n), y_n\}$
- Solve a linear problem on $\{z_n, y_n\}$ using your favorite algorithm \mathcal{A} to get a good model \tilde{w}
- Return the model $h(x) = \text{sign}(\tilde{w}^T \phi(x))$



Nonlinear transformation

Polynomial mappings

- Can now freely do quadratic classification, quadratic regression
- Can easily extend to any degree of polynomial mappings

- E.g.,

$$\phi(x) = (x_1, x_2, x_3, x_1x_2, x_1x_3, x_2x_3, x_1x_2^2, x_1x_3^2, x_1x_2^2, x_2^2x_3, x_2^2x_3, x_1^3, x_2^3, x_3^3)$$



Nonlinear Transformation

The price we pay: computational complexity

- Q -th order polynomial transform:

$$\phi(x) = (1, x_1, x_2, \dots, x_d,$$

$$x_1^2, x_1x_2, \dots, x_d^2, \dots, x_d^2,$$

...

- $x_1^Q, x_1^{Q-1}x_2, \dots, x_d^Q)$

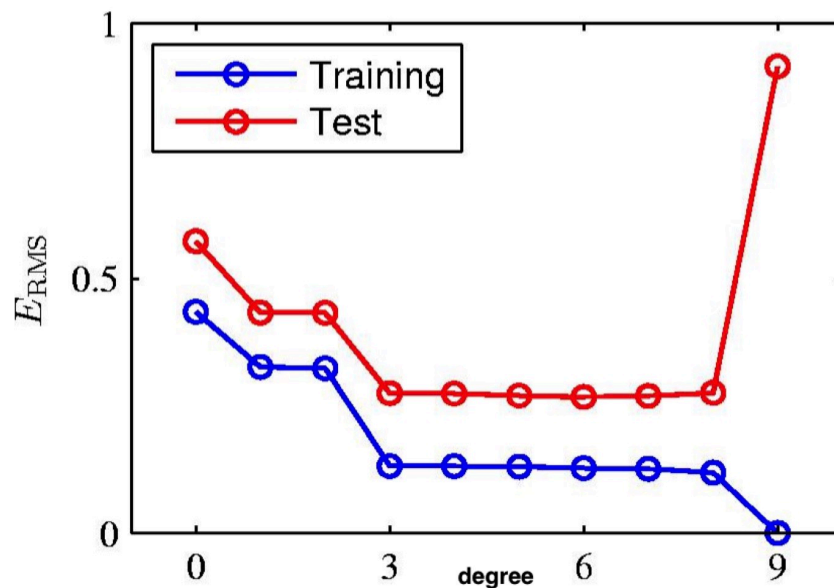
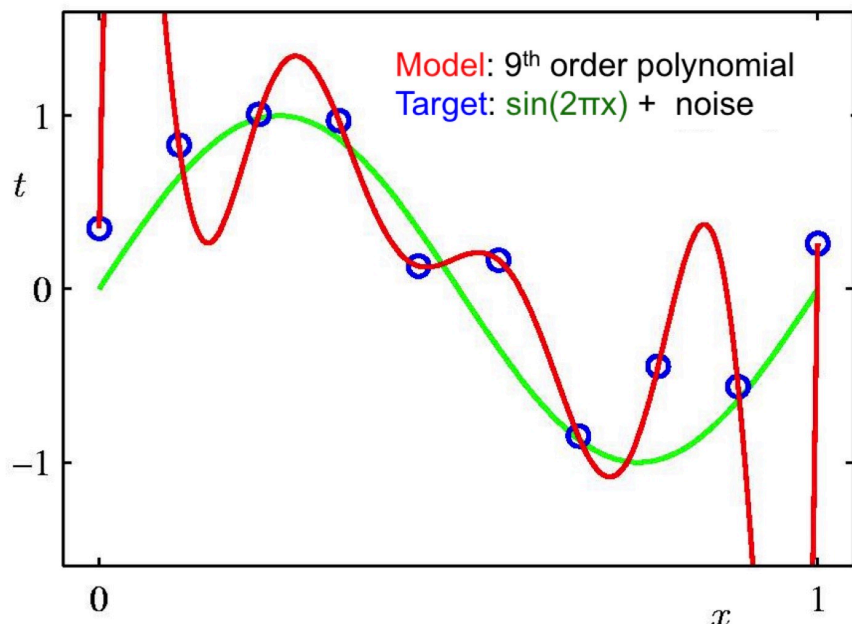
- $O(d^Q)$ dimensional vector \Rightarrow High computational cost

- Kernel method

Nonlinear Transformation

The price we pay: overfitting

- **Overfitting**: the model has low training error but high prediction error



Theory of Generalization

Training versus testing

- Machine learning pipeline:
 - Training phase:
 - Obtain the best model h by minimizing training error
 - Test (inference) phase:
 - For any incoming test data x
 - Make prediction by $h(x)$
 - Measure the performance of h : test error

Theory of Generalization

Training versus testing

- Does low training error imply low test error?
 - They can be totally different if
 - train distribution \neq test distribution

Theory of Generalization

Training versus testing

- Does low **training error** imply low **test error**?
 - They can be totally different if
 - **train distribution** \neq **test distribution**
 - Even under the same distribution, they can be very different:
 - Because h is picked to minimize **training error**, not **test error**

Theory of Generalization

Formal definition

- Assume training and test data are both sampled from D
- The ideal function (for generating labels) is $f : \mathcal{X} \rightarrow \mathcal{Y}$
- Training error: Sample x_1, \dots, x_N from D and

$$E_{tr}(h) = \frac{1}{N} \sum_{n=1}^N e(h(x_n), f(x_n))$$

- h is determined by x_1, \dots, x_n

- Test error: Sample x_1, \dots, x_N from D and

$$E_{te}(h) = \frac{1}{M} \sum_{m=1}^M e(h(x_m), f(x_m))$$

- h is independent to x_1, \dots, x_n

Theory of Generalization

Formal definition

- Assume training and test data are both sampled from D
- The ideal function (for generating labels) is $f : f(x) \rightarrow y$
- Training error: Sample x_1, \dots, x_N from D and

$$• E_{tr}(h) = \frac{1}{N} \sum_{n=1}^N e(h(x_n), f(x_n))$$

- h is determined by x_1, \dots, x_n

- Test error: Sample x_1, \dots, x_N from D and

$$• E_{te}(h) = \frac{1}{M} \sum_{m=1}^M e(h(x_m), f(x_m))$$

- h is independent to x_1, \dots, x_n

- Generalization error = Test error = Expected performance on D :

$$• E(h) = \mathbb{E}_{x \sim D}[e(h(x), f(x))] = E_{te}(h)$$

Theory of Generalization

The 2 questions of learning

- $E(h) \approx 0$ is achieved through:
 - $E(h) \approx E_{tr}(h)$ and $E_{tr}(h) \approx 0$

Theory of Generalization

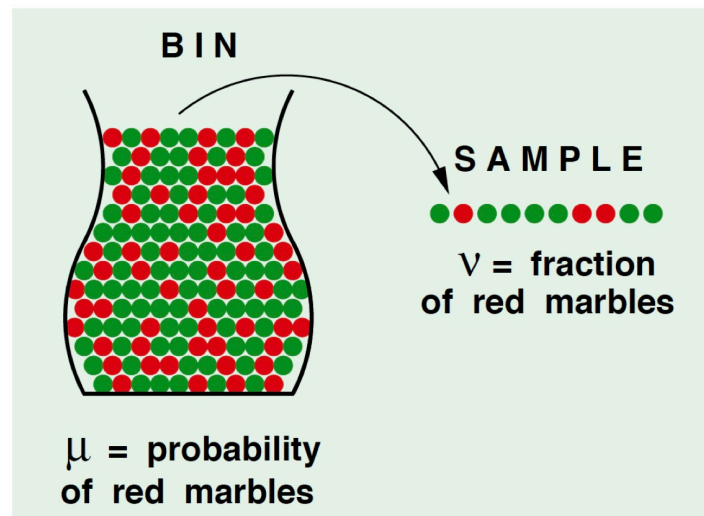
The 2 questions of learning

- $E(h) \approx 0$ is achieved through:
 - $E(h) \approx E_{tr}(h)$ and $E_{tr}(h) \approx 0$
- Learning is split into 2 questions:
 - Can we make sure that $E(h) \approx E_{tr}(h)$?
 - Today's focus
 - Can we make $E_{tr}(h)$ small?
 - Optimization

Theory of Generalization

Bound $\|E(h) - E_{tr}(h)\|$

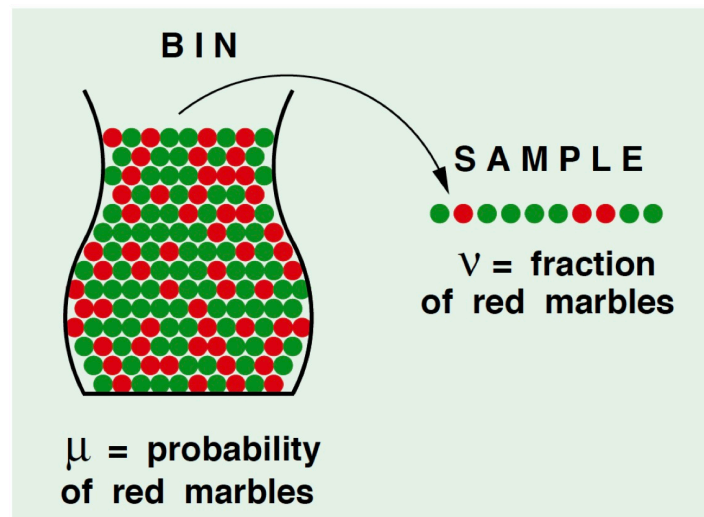
- Consider a bin with **red** and **green** marbles
 - $P[\text{picking a red mable}] = \mu$
 - $P[\text{picking a green mable}] = 1 - \mu$
- The value of μ is unknown to us
- How to infer μ ?
 - Pick N marbles independently
 - ν : the traction of **red** marble



Theory of Generalization

Inferring with probability

- Do we **know** μ
 - No
 - Sample can be mostly **green** while bin is mostly **red**
- Can we say something about μ ?
 - Yes
 - ν is “probably” close to μ



Theory of Generalization

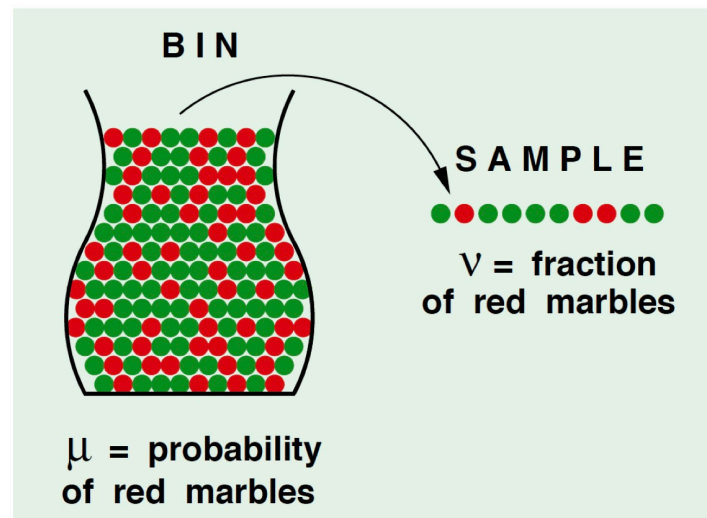
Hoeffding's inequality

- In big sample (large N), ν (sample mean) is probably close to μ :
 - $p[|\nu - \mu| > \epsilon] \leq 2e^{-2\epsilon^2 N}$
 - This is called **Hoeffding's inequality**
- The statement “ $\mu = \nu$ ” is **probably approximately correct** (PAC)

Theory of Generalization

Hoeffding's inequality

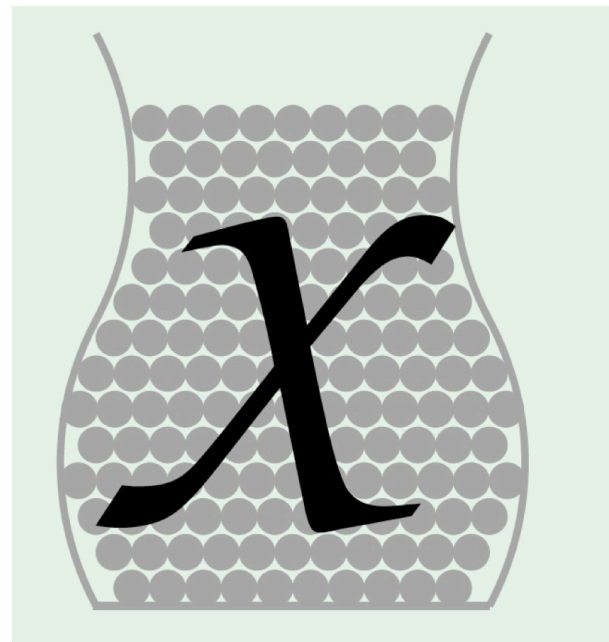
- $p[|\nu - \mu| > \epsilon] \leq 2e^{-2\epsilon^2 N}$
 - Valid for all N and $\epsilon > 0$
 - Does not depend on μ (no need to know μ)
 - Larger sample size N or looser gap $\epsilon \Rightarrow$ higher probability for $\mu \approx \nu$



Theory of Generalization

Connection to Learning

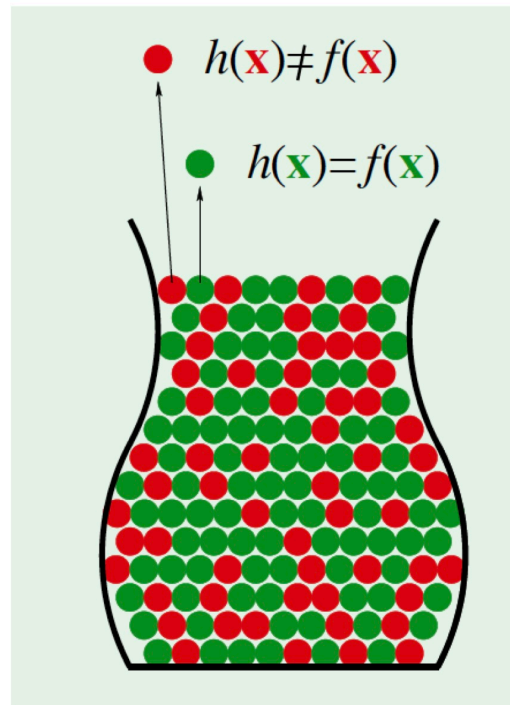
- How to connect this to learning?
 - Each marble (uncolored) is a data point $x \in \mathcal{X}$



Theory of Generalization

Connection to Learning

- How to connect this to learning?
 - Each marble (uncolored) is a data point $x \in \mathcal{X}$
 - **Red** marble: $h(x) \neq f(x)$
 - **Green** marble: $h(x) = f(x)$



Theory of Generalization

Connection to Learning

- Given a function h
- If we randomly draw x_1, \dots, x_n (independent to h):
 - $E(h) = \mathbb{E}_{x \sim D}[h(x) \neq f(x)] \Leftrightarrow \mu$ (generalization error, unknown)
 - $\frac{1}{N} \sum_{n=1}^N [h(x_n) \neq y_n] \Leftrightarrow \nu$ (error on sampled data, known)

Theory of Generalization

Connection to Learning

- Given a function h
- If we randomly draw x_1, \dots, x_n (independent to h):
 - $E(h) = \mathbb{E}_{x \sim D}[h(x) \neq f(x)] \Leftrightarrow \mu$ (generalization error, unknown)
 - $\frac{1}{N} \sum_{n=1}^N [h(x_n) \neq y_n] \Leftrightarrow \nu$ (error on sampled data, known)
- Based on Hoeffding's inequality:
 - $p[|\nu - \mu| > \epsilon] \leq 2e^{-2\epsilon^2 N}$
- “ $\mu = \nu$ ” Is probably approximately correct (PAC)

Theory of Generalization

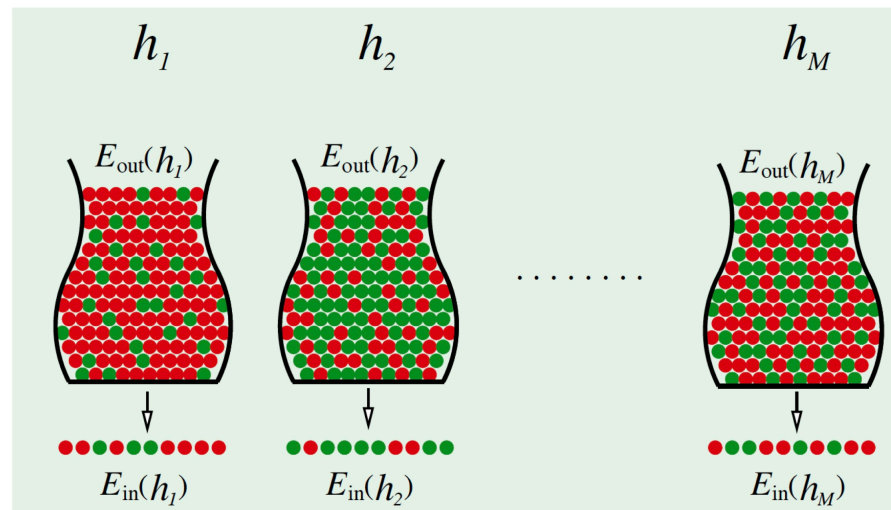
Connection to Learning

- Given a function h
- If we randomly draw x_1, \dots, x_n (independent to h):
 - $E(h) = \mathbb{E}_{x \sim D}[h(x) \neq f(x)] \Leftrightarrow \mu$ (generalization error, unknown)
 - $\frac{1}{N} \sum_{n=1}^N [h(x_n) \neq y_n] \Leftrightarrow \nu$ (error on sampled data, known)
- Based on Hoeffding's inequality:
 - $p[|\nu - \mu| > \epsilon] \leq 2e^{-2\epsilon^2 N}$
- “ $\mu = \nu$ ” Is probably approximately correct (PAC)
- However, this can only “verify” the error of a hypothesis:
 - h and x_1, \dots, x_N must be independent

Theory of Generalization

Apply to multiple bins (hypothesis)

- Can we apply to multiple hypothesis?
- Color in each bin depends on different hypothesis
 - **Bingo** when getting all **green balls**?



Theory of Generalization

Coin game

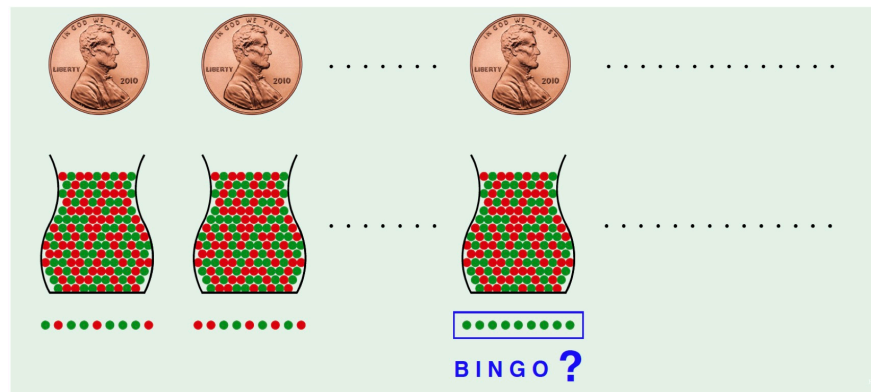
- If you have 150 **fair** coins, flip each coin 5 times, and **one of them gets 5 heads**. Is this coin (g) special?

- No. The probability of exiting at least one of the coin results in 5 heads is

$$1 - \left(\frac{31}{32}\right)^{150} > 99\%$$

- Because: there can exist some h such that E and E_{tr} are far away if **M is large**.

M -> number of hypothesis



Theory of Generalization

A simple solution

- For each particular h ,

- $P[|E_{tr}(h) - E(h)| > \epsilon] \leq 2e^{-2\epsilon^2 N}$

- If we have a hypothesis set \mathcal{H} , we want to derive the bound for $P[\sup_{h \in \mathcal{H}} |E_{tr}(h) - E(h)| > \epsilon]$

- $P[|E_{tr}(h_1) - E(h_1)| > \epsilon] \text{ or } \dots \text{ or } P[|E_{tr}(h_{|\mathcal{H}|}) - E(h_{|\mathcal{H}|})| > \epsilon]$

- $\leq \sum_{m=1}^{|\mathcal{H}|} P[|E_{tr}(h_m) - E(h_m)| > \epsilon] \leq 2|\mathcal{H}|e^{-2\epsilon^2 N}$

- Because of union bound inequality $P(\bigcup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} P(A_i)$

Theory of generalization

When is learning successful?

- When our learning algorithm \mathcal{A} picks the hypothesis g :
 - $P[\text{SUP}_{h \in \mathcal{H}} | E_{tr}(h) - E(h) | > \epsilon] \leq 2 |\mathcal{H}| e^{-2\epsilon^2 N}$
- If $|\mathcal{H}|$ is small and N is large enough:
 - If \mathcal{A} finds $E_{tr}(g) \approx 0 \Rightarrow E(g) \approx 0$ (Learning is successful!)

Theory of Generalization

Feasibility of Learning

- $P[|E_{tr}(g) - E(g)| > \epsilon] \leq 2|\mathcal{H}|e^{-2\epsilon^2N}$
 - Two questions:
 - 1. Can we make sure $E(g) \approx E_{tr}(g)$?
 - 2. Can we make sure $E_{tr}(g) \approx 0$?

Theory of Generalization

Feasibility of Learning

- $P[|E_{tr}(g) - E(g)| > \epsilon] \leq 2|\mathcal{H}|e^{-2\epsilon^2N}$
 - Two questions:
 - 1. Can we make sure $E(g) \approx E_{tr}(g)$?
 - 2. Can we make sure $E_{tr}(g) \approx 0$?
- $|\mathcal{H}|$: complexity of model
 - Small $|\mathcal{H}|$: 1 holds, but 2 may not hold (too few choices) (under-fitting)

Theory of Generalization

Feasibility of Learning

- $P[|E_{tr}(g) - E(g)| > \epsilon] \leq 2 |\mathcal{H}| e^{-2\epsilon^2 N}$
 - Two questions:
 - 1. Can we make sure $E(g) \approx E_{tr}(g)$?
 - 2. Can we make sure $E_{tr}(g) \approx 0$?
- $|\mathcal{H}|$: complexity of model
 - Small $|\mathcal{H}|$: 1 holds, but 2 may not hold (too few choices) (under-fitting)
 - Large $|\mathcal{H}|$: 1 doesn't hold, but 2 may hold (over-fitting)

Theory of Generalization

Feasibility of Learning

- Currently we only know

- $P[\text{SUP}_{h \in \mathcal{H}} | E_{tr}(h) - E(h) | > \epsilon] \leq 2 |\mathcal{H}| e^{-2\epsilon^2 N}$

Theory of Generalization

Feasibility of Learning

- Currently we only know

- $P[\text{SUP}_{h \in \mathcal{H}} |E_{tr}(h) - E(h)| > \epsilon] \leq 2 |\mathcal{H}| e^{-2\epsilon^2 N}$

- What if $|\mathcal{H}| = \infty$?

- (e.g. linear hyperplanes)

Theory of Generalization

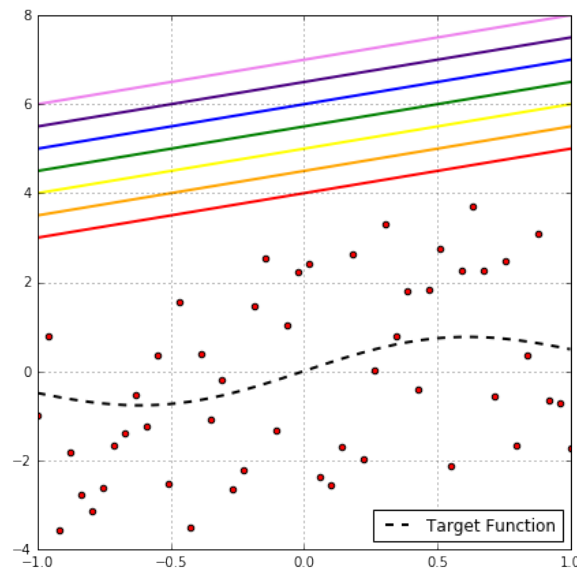
Deduce the dimension

- Why do we need to consider every possible hypothesis?
 - $P[\text{SUP}_{h \in \mathcal{H}} | E_{tr}(h) - E(h) | > \epsilon]$
 - If we omit one hypothesis, we might miss the biggest gap
- $P[\text{SUP}_{h \in \mathcal{H}} | E_{tr}(h) - E(h) | > \epsilon] \leq 2 | \mathcal{H} | e^{-2\epsilon^2 N}$
 - from the union bound, which assume the event is independent

Theory of Generalization

Deduce the dimension

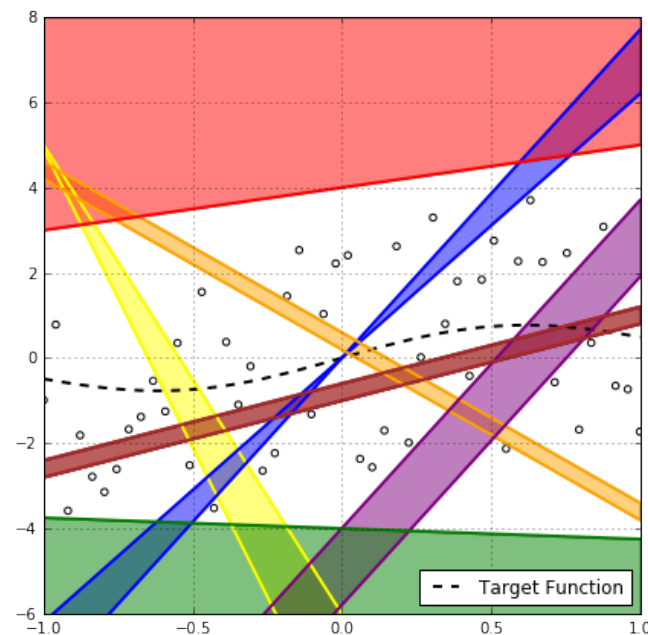
- Why do we need to consider every possible hypothesis?
- $P[\text{SUP}_{h \in \mathcal{H}} | E_{tr}(h) - E(h) | > \epsilon]$
- If we omit one hypothesis, we might miss the biggest gap
- However, are the events of each hypothesis having a big generalization gap are likely to be independent?



Theory of Generalization

Deduce the dimension

- Why do we need to consider every possible hypothesis?
- $P[\text{SUP}_{h \in \mathcal{H}} | E_{tr}(h) - E(h) | > \epsilon]$
- If we omit one hypothesis, we might miss the biggest gap
- However, are the events of each hypothesis having a big generalization gap are likely to be independent?
 - No



Theory of Generalization

Symmetrization lemma

- Imagine we have the ghost dataset S' with also size N :

$$\bullet P[\text{SUP}_{h \in \mathcal{H}} |E_{tr}(h) - E(h)| > \epsilon] \leq 2P[\text{SUP}_{h \in \mathcal{H}} |E_{tr}(h) - E'_{tr}(h)| > \frac{\epsilon}{2}]$$

Theory of Generalization

Growth function

- Imagine we have the ghost dataset S' with also size N :

- $P[\text{SUP}_{h \in \mathcal{H}} |E_{tr}(h) - E(h)| > \epsilon] \leq 2P[\text{SUP}_{h \in \mathcal{H}} |E_{tr}(h) - E'_{tr}(h)| > \frac{\epsilon}{2}]$

- By union bound:

- $P[\text{SUP}_{h \in \mathcal{H}_{S \cup S'}} |E_{tr}(h) - E'_{tr}(h)| > \frac{\epsilon}{2}] \leq |\mathcal{H}_{S \cup S'}| P[|E_{tr}(h) - E'_{tr}(h)| > \frac{\epsilon}{2}]$

Theory of Generalization

Growth function

- Imagine we have the ghost dataset S' with also size N :

- $P[\text{SUP}_{h \in \mathcal{H}} |E_{tr}(h) - E(h)| > \epsilon] \leq 2P[\text{SUP}_{h \in \mathcal{H}} |E_{tr}(h) - E'_{tr}(h)| > \frac{\epsilon}{2}]$

- By union bound:

- $P[\text{SUP}_{h \in \mathcal{H}_{S \cup S'}} |E_{tr}(h) - E'_{tr}(h)| > \frac{\epsilon}{2}] \leq |\mathcal{H}_{S \cup S'}| P[|E_{tr}(h) - E'_{tr}(h)| > \frac{\epsilon}{2}]$

- How to bound $|\mathcal{H}_{S \cup S'}|$

Theory of Generalization

Growth function

- For binary classification $\{+1, -1\}$, for a dataset with N samples,
 - The max number of distinct labellings is 2^N
- Growth function $\Delta_{\mathcal{H}}(N)$: The max number of distinct labellings on a dataset S of size N by a hypothesis space \mathcal{H}
- So,

$$\bullet P[\text{SUP}_{h \in \mathcal{H}_{S \cup S'}} |E_{tr}(h) - E'_{tr}(h)| > \frac{\epsilon}{2}] \leq \Delta_{\mathcal{H}}(2N)P[|E_{tr}(h) - E'_{tr}(h)| > \frac{\epsilon}{2}]$$

- And $\Delta_{\mathcal{H}}(N) \leq 2^m$