Teoria degli Algoritmi

Corso di Laurea Magistrale in Matematica Applicata

a.a. 2020-21

SAPIENZA
UNIVERSITÀ DI ROMA

Gabriele Tolomei

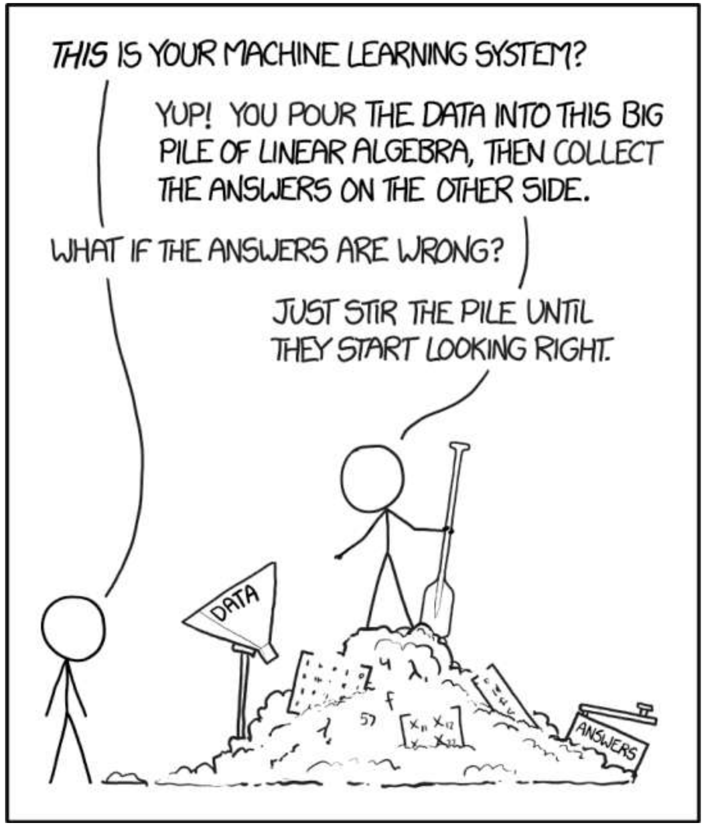
Dipartimento di Informatica

Sapienza Università di Roma

tolomei@di.uniroma1.it

How Much Data Do We Need?

In general, the more data we have the better we learn



source: https://xkcd.com/1838/

Is Learning Feasible, After All?

- Learning an unknown target function seems impossible!

Is Learning Feasible, After All?

- Learning an unknown target function seems impossible!
- We only dispose of a finite data sample (i.e., the training set) where we know the value of the unknown function

Is Learning Feasible, After All?

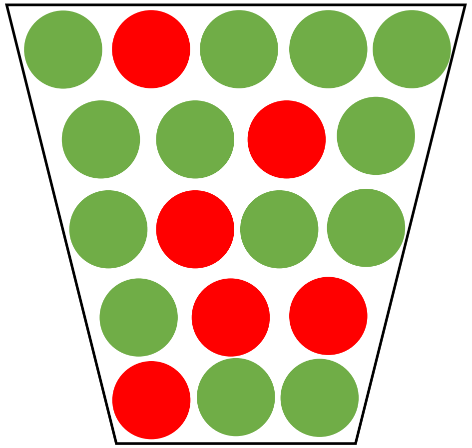
- Learning an unknown target function seems impossible!
- We only dispose of a finite data sample (i.e., the training set) where we know the value of the unknown function
- Outside of that, the function may take on any value!

Is Learning Feasible, After All?

- Learning an unknown target function seems impossible!
- We only dispose of a finite data sample (i.e., the training set) where we know the value of the unknown function
- Outside of that, the function may take on any value!
- **Question:** Can we use our finite sample to learn something outside of it?

A Related Experiment

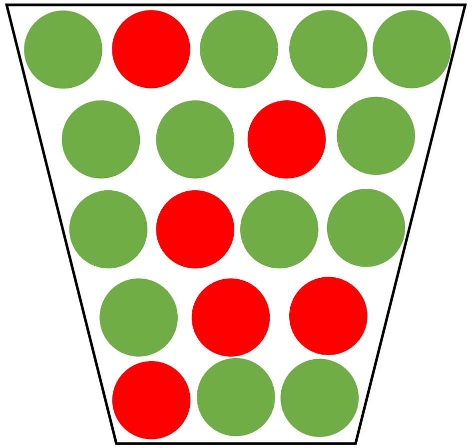
Consider a bin with red and green marbles



A Related Experiment

Consider a bin with red and green marbles

Let p be the probability of picking a red marble

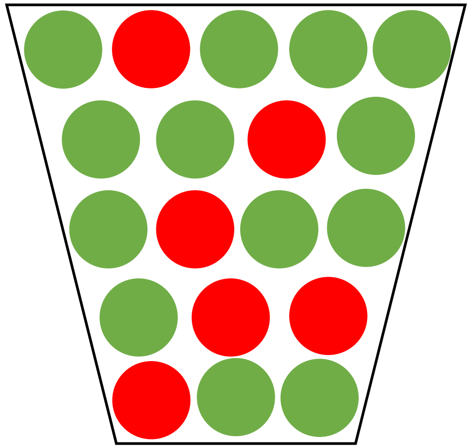


A Related Experiment

Consider a bin with red and green marbles

Let p be the probability of picking a red marble

Of course, $q = 1\text{-}p$ is the probability of picking a green marble



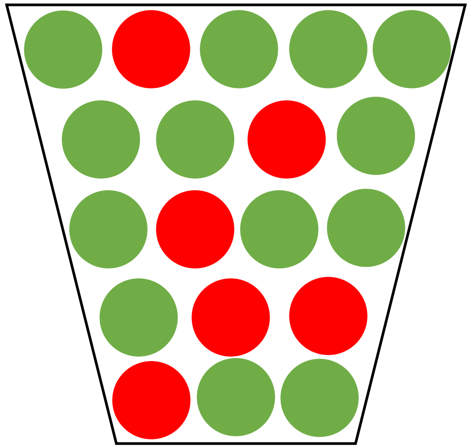
A Related Experiment

Consider a bin with red and green marbles

Let p be the probability of picking a red marble

Of course, $q = 1\text{-}p$ is the probability of picking a green marble

The value of p is fixed constant yet unknown to us



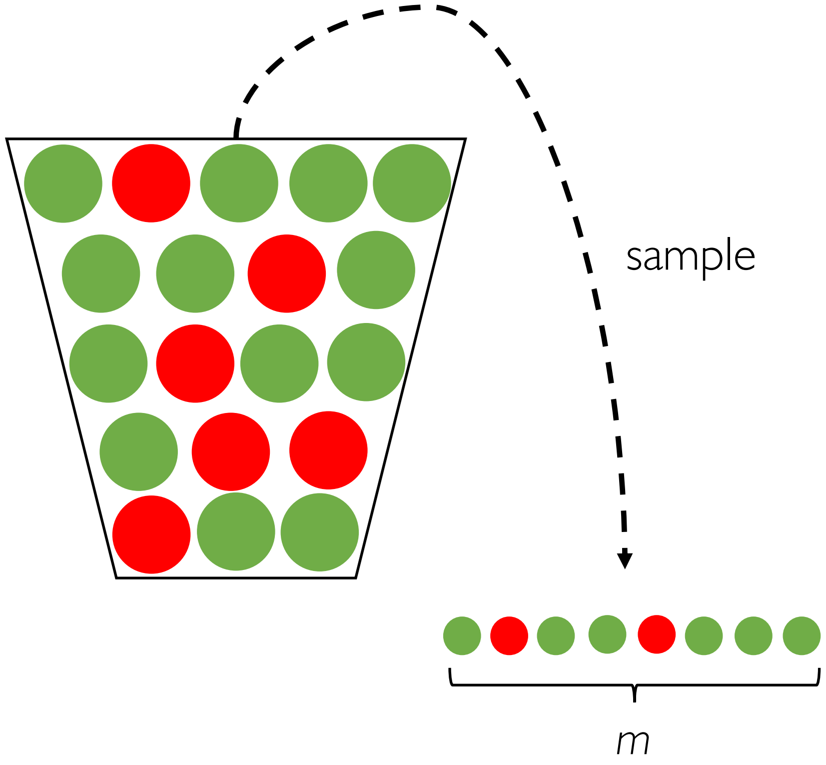
A Related Experiment

Consider a bin with red and green marbles

Let p be the probability of picking a red marble

Of course, $q = 1\text{-}p$ is the probability of picking a green marble

The value of p is fixed constant yet unknown to us



Suppose we extract a random sample of size m from the bin and we count how many red marbles we got, call it p' (sample frequency)

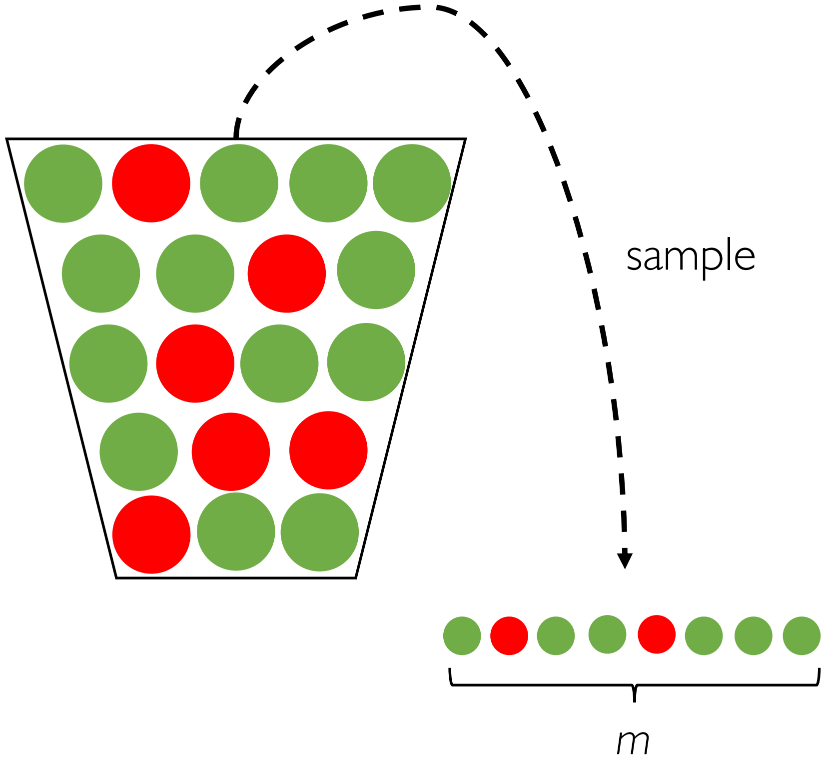
A Related Experiment

Consider a bin with red and green marbles

Let p be the probability of picking a red marble

Of course, $q = 1-p$ is the probability of picking a green marble

The value of p is fixed constant yet unknown to us

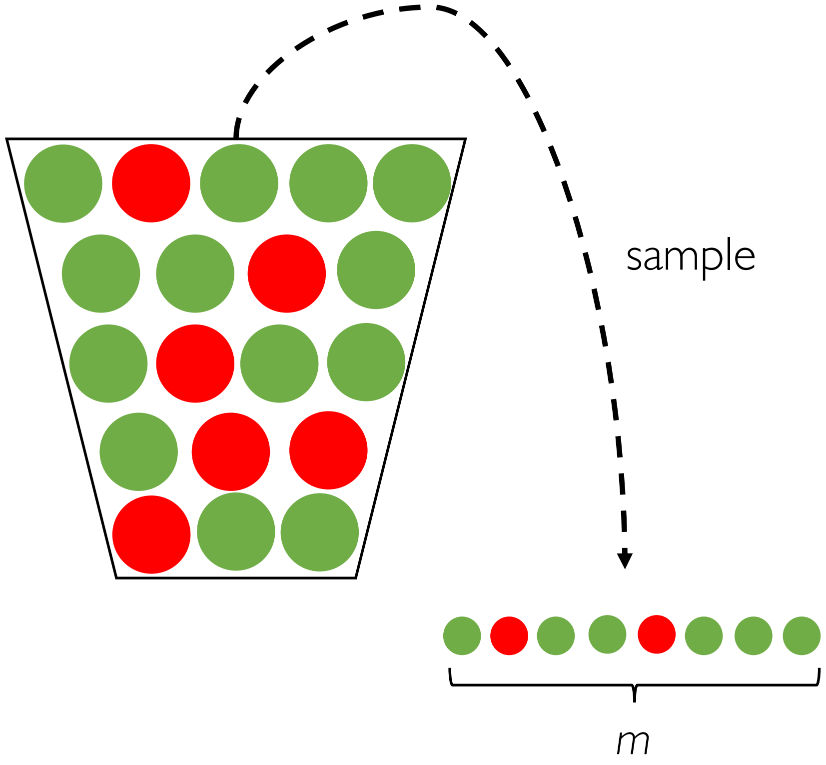


Suppose we extract a random sample of size m from the bin and we count how many red marbles we got, call it p' (sample frequency)

Note:
The bin can be considered either infinite or the sampling being done with replacement

A Related Experiment

Does p' say something about p?



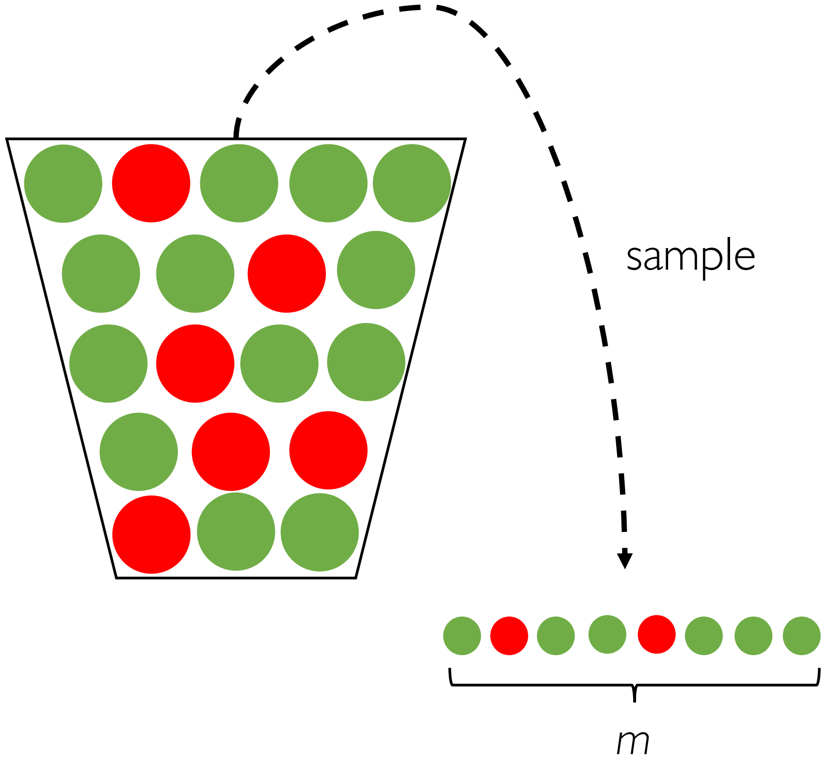
A Related Experiment

Does p' say something about p?

Short Answer: NO!

Our sample could be made of all green marbles even though the bin mostly contains red ones

possible



A Related Experiment

Does p' say something about p?

Short Answer: NO!

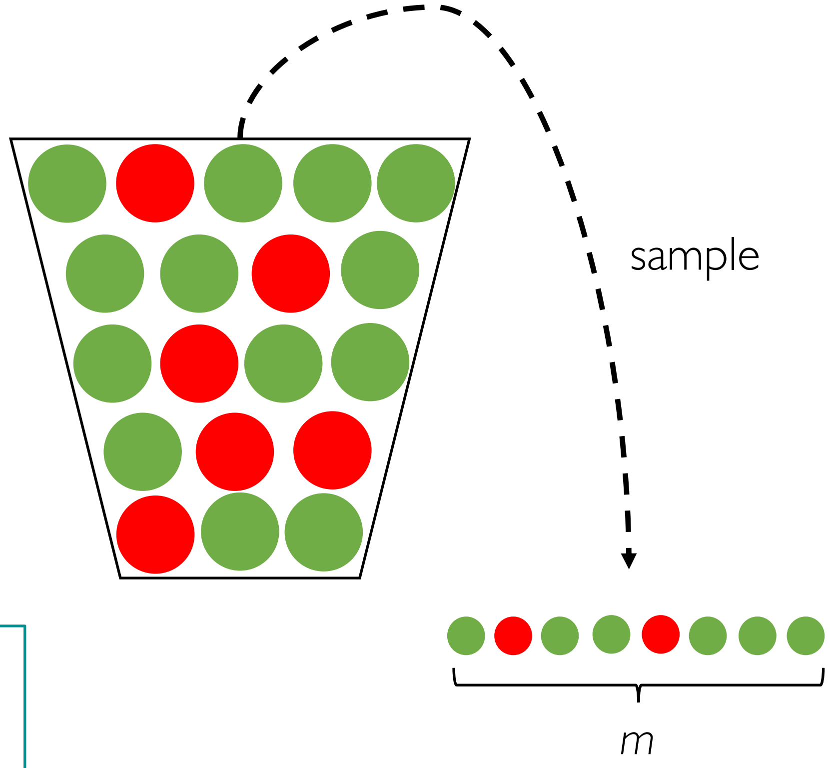
Our sample could be made of all green marbles even though the bin mostly contains red ones

possible

Long Answer: YES!

If the sample is "big enough" (m is "large"), sample frequency p' is likely close to the true bin frequency p

probable



A Related Experiment

Does p' say something about p?

Short Answer: NO!

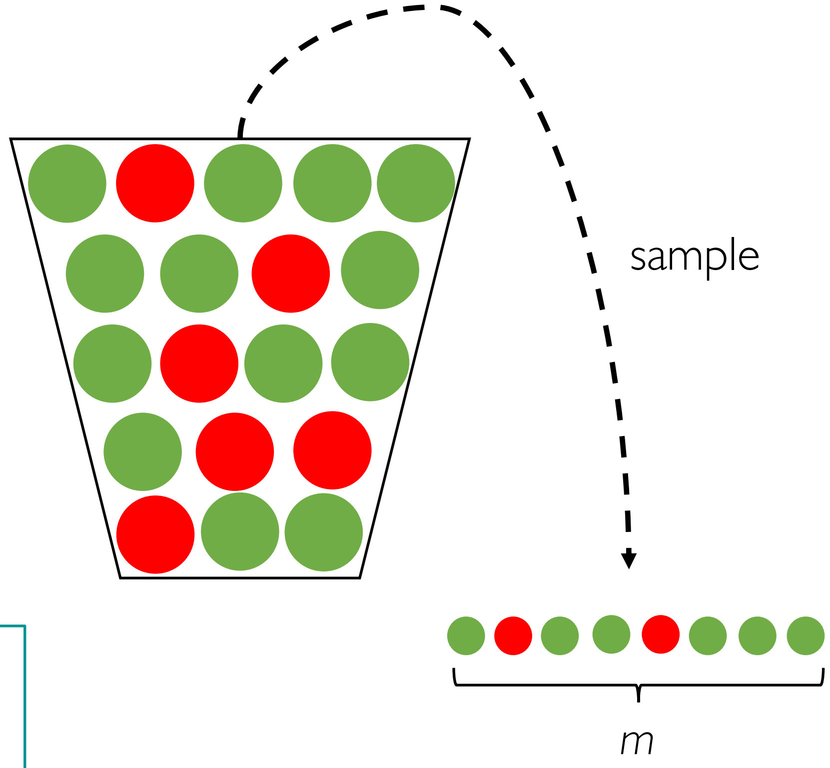
Our sample could be made of all green marbles even though the bin mostly contains red ones

possible

Long Answer: YES!

If the sample is "big enough" (m is "large"), sample frequency p' is likely close to the true bin frequency p

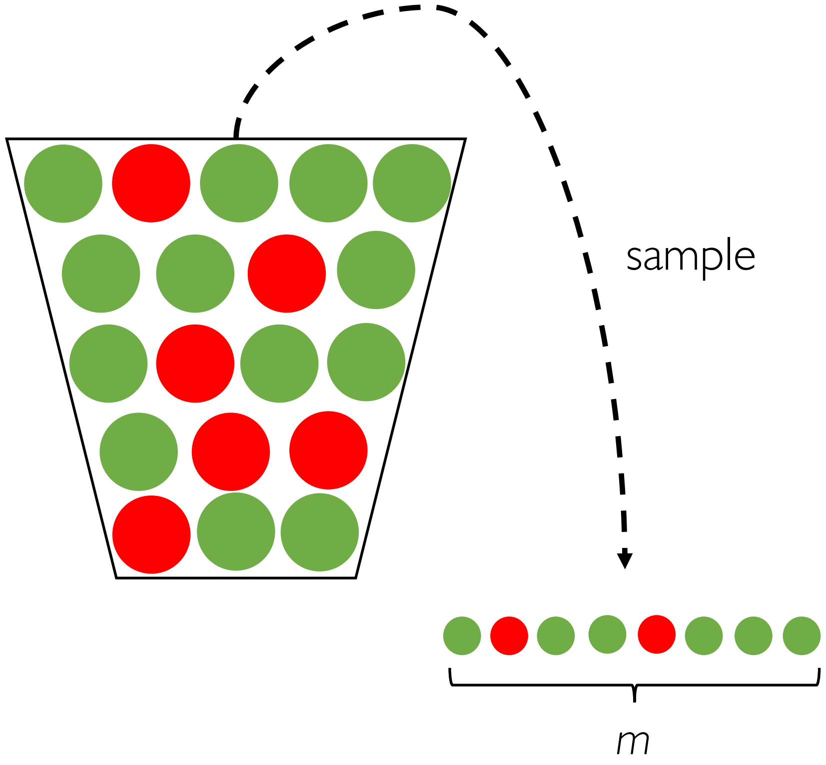
probable



But what does p' say about p, exactly?

A Related Experiment

In a big sample (large *m*), *p*' is probably close to *p* (within *epsilon*)

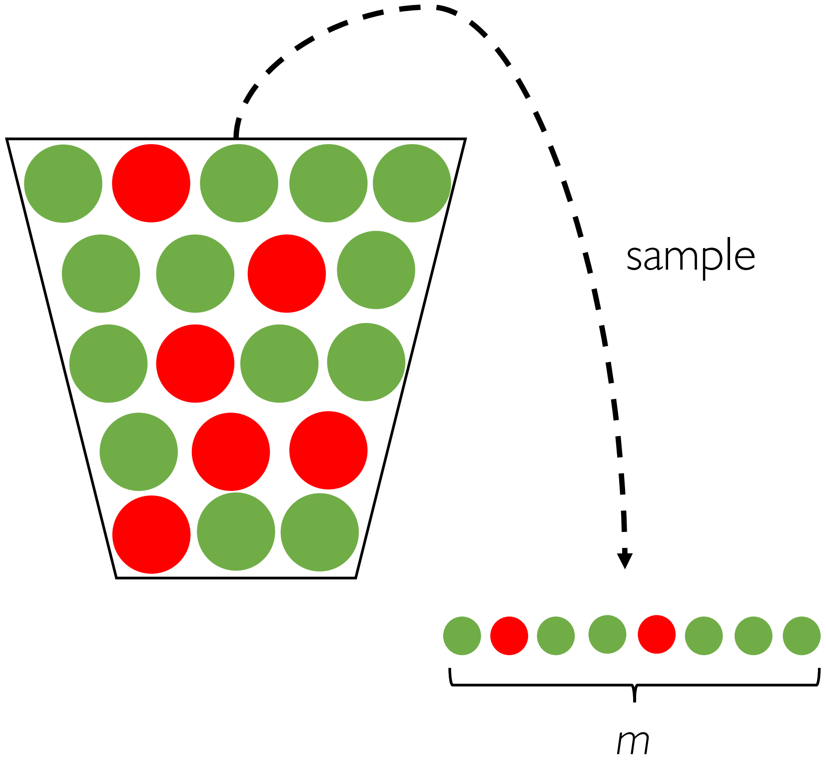


A Related Experiment

Hoeffding's Inequality

In a big sample (large *m*), *p*' is probably close to *p* (within *epsilon*)

$$P(|p' - p| > \epsilon) \leq 2e^{-2m\epsilon^2}$$

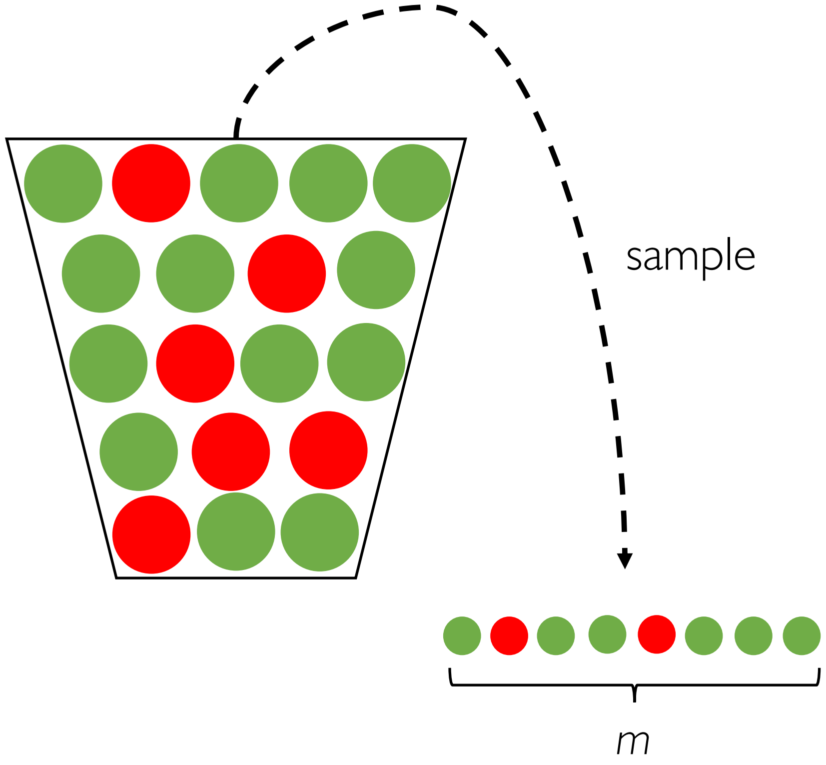


A Related Experiment

Hoeffding's Inequality

In a big sample (large *m*), *p*' is probably close to *p* (within *epsilon*)

$$\underbrace{P(|p' - p| > \epsilon)}_{\text{"bad event"}} \leq 2e^{-2m\epsilon^2}$$



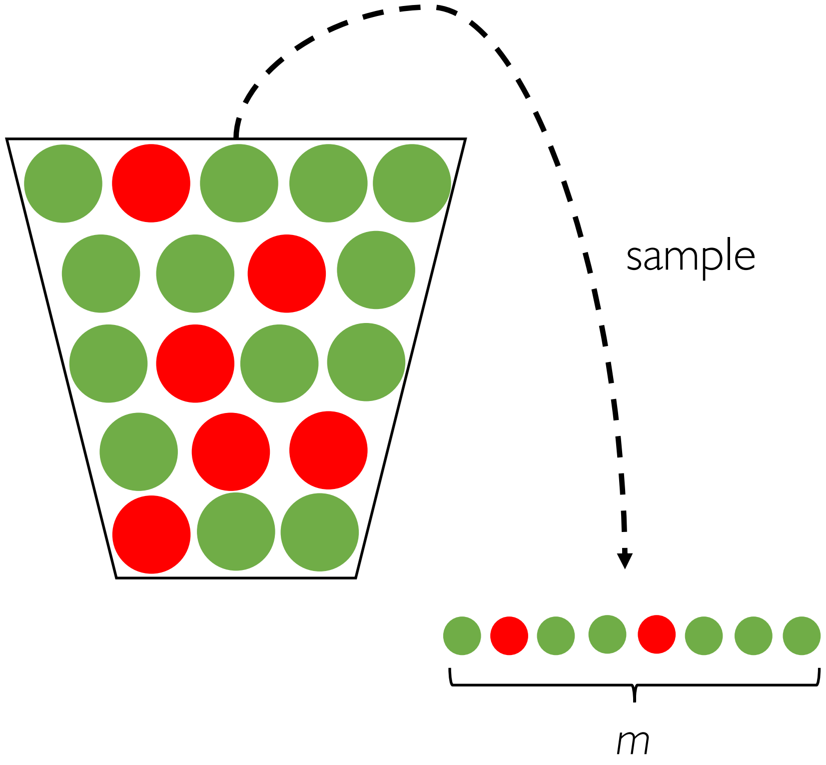
The "bad event" is *p*' deviating more than *epsilon* from the true *p*

A Related Experiment

Hoeffding's Inequality

In a big sample (large *m*), *p*' is **probably close** to *p* (within *epsilon*)

$$P(\underbrace{|p' - p| > \epsilon}_{\text{"bad event"}}) \leq 2e^{-2m\epsilon^2}$$



The "bad event" is *p*' deviating more than *epsilon* from the true *p*

We want the probability of such bad event to be small!

Hoeffding's Inequality

$$P(|p' - p| > \epsilon) \leq 2e^{\boxed{-2m}\epsilon^2}$$

The presence of *m* as a negative exponent contributes to keep the right-hand expression small (as *m* increases)

Hoeffding's Inequality

$$P(|p' - p| > \epsilon) \leq 2e^{-2m\boxed{\epsilon^2}}$$

Wait! *m* is multiplied by *epsilon* squared and therefore its effect as negative exponent gets diluted as *epsilon* gets smaller (i.e., the closer we want *p*' to the real *p*)

Hoeffding's Inequality

$$P(|p' - p| > \epsilon) \leq 2e^{-2m\epsilon^2}$$

There is a tradeoff between the sample size (*m*), the tolerance (*epsilon*), and the bound

Hoeffding's Inequality

$$P(|p' - p| > \epsilon) \leq 2e^{-2m\epsilon^2}$$

There is a tradeoff between the sample size (*m*), the tolerance (*epsilon*), and the bound

$$p' \overset{\text{PAC}}{=} p$$

The statement above is Probably Approximately Correct

Hoeffding's Inequality

- Belongs to a large class of mathematical laws called "the laws of large numbers"

Hoeffding's Inequality

- Belongs to a large class of mathematical laws called "the laws of large numbers"
- It is valid for every positive integer m and every *epsilon* > 0

Hoeffding's Inequality

- Belongs to a large class of mathematical laws called "the laws of large numbers"
- It is valid for every positive integer *m* and every *epsilon* > 0
- Bound does not depend on *p* because it is an unknown quantity, and *epsilon* just represents our tolerance

Connection to the Learning Problem

- How does PAC statement guaranteed by the Hoeffding's inequality help us to determine whether learning is feasible?

Connection to the Learning Problem

- How does PAC statement guaranteed by the Hoeffding's inequality help us to determine whether learning is feasible?
- We must make a link between the bin example and learning

Connection to the Learning Problem

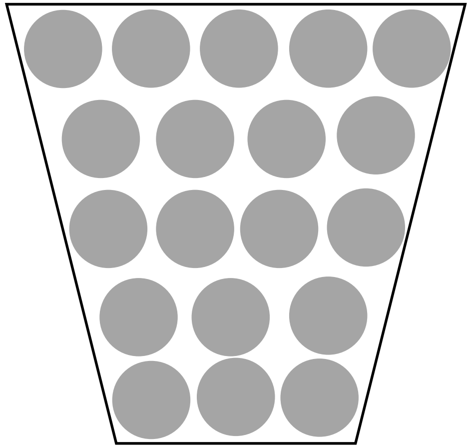
- How does PAC statement guaranteed by the Hoeffding's inequality help us to determine whether learning is feasible?
- We must make a link between the bin example and learning
- In the bin example, the unknown quantity we want to estimate is a single number p, i.e., the frequency of red marbles in the bin

Connection to the Learning Problem

- How does PAC statement guaranteed by the Hoeffding's inequality help us to determine whether learning is feasible?
- We must make a link between the bin example and learning
- In the bin example, the unknown quantity we want to estimate is a single number p, i.e., the frequency of red marbles in the bin
- In the learning problem, the unknown quantity we want to estimate is a full-fledged target function f: X → Y

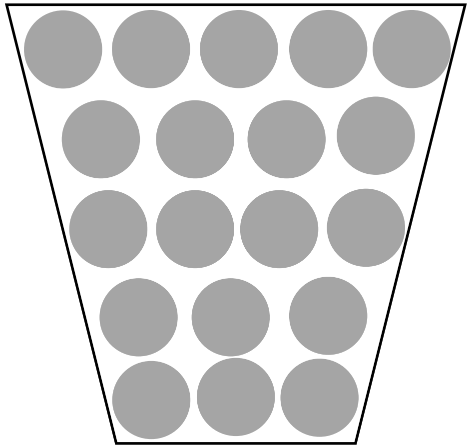
Connection to the Learning Problem

- The bin can be seen as the whole input space X
- Each marble is a single data point x in X



Connection to the Learning Problem

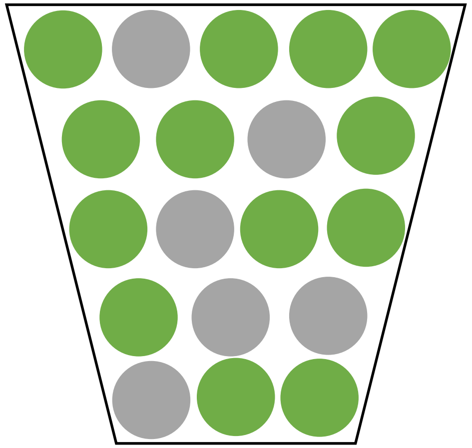
- The bin can be seen as the whole input space X
- Each marble is a single data point x in X



How do we color each marble?

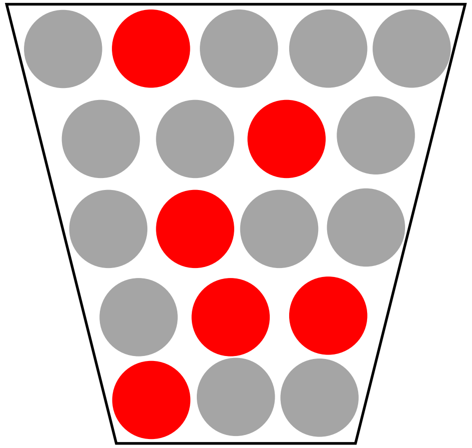
Connection to the Learning Problem

green marbles: correspond to data points where a given hypothesis h agrees with the true unknown target function f

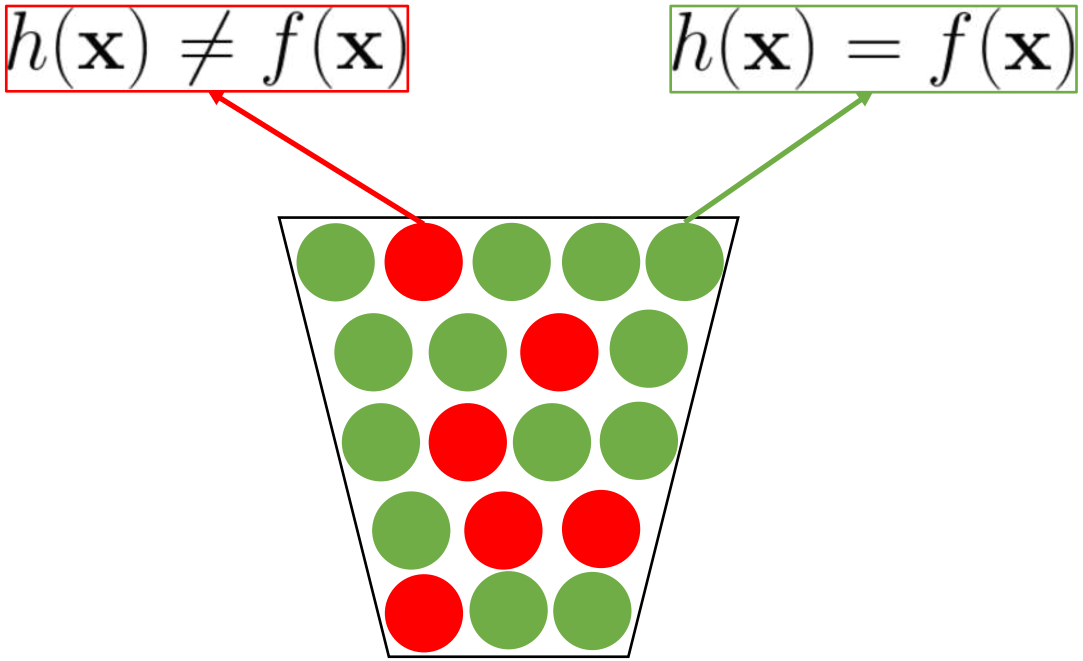


Connection to the Learning Problem

red marbles: correspond to data points where a given hypothesis h disagrees with the true unknown target function f

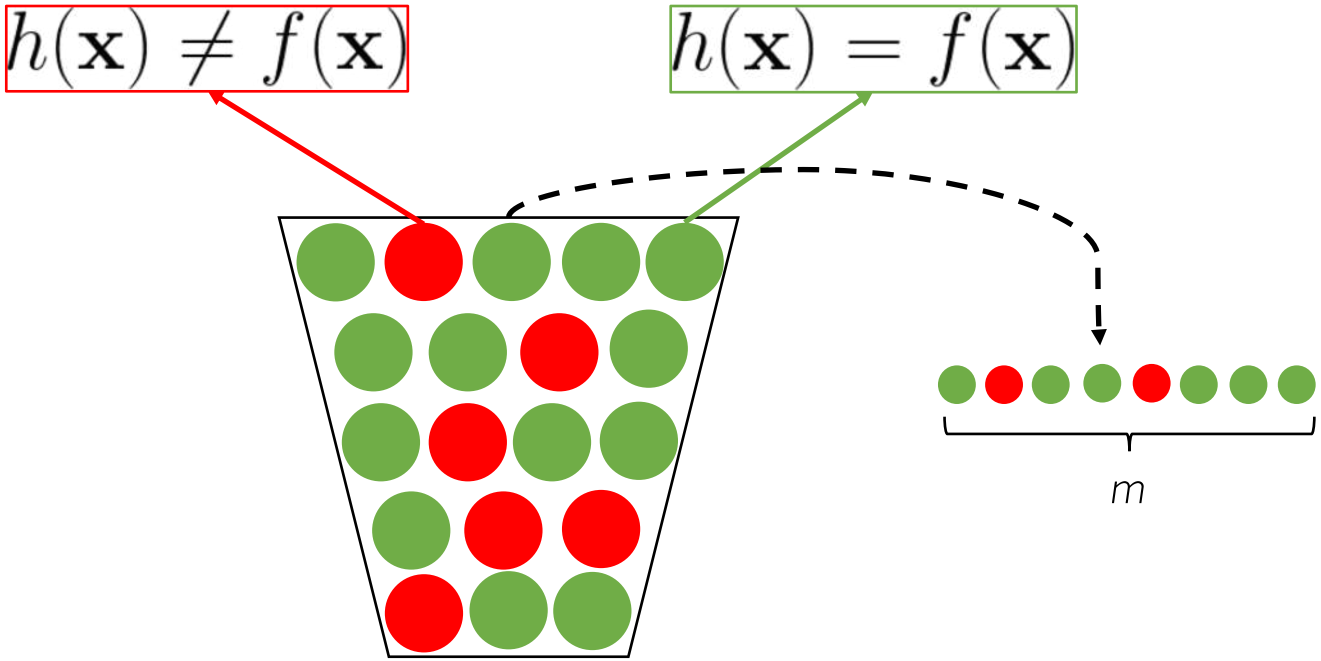


Connection to the Learning Problem



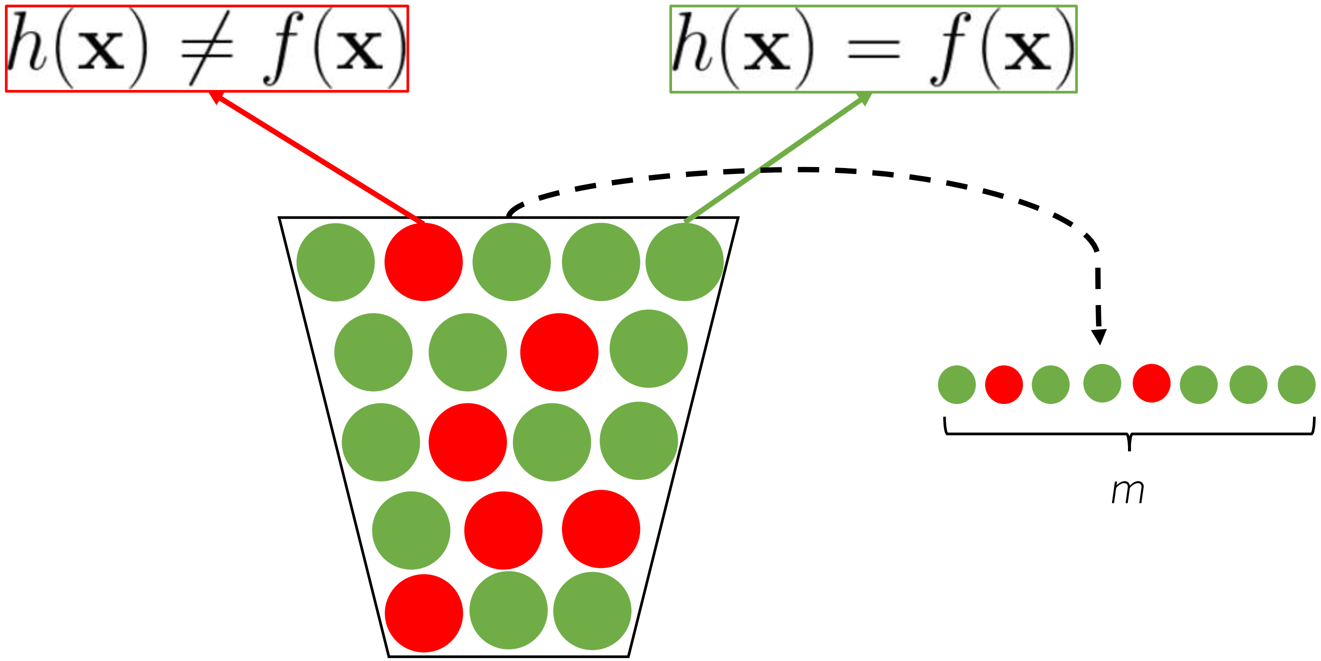
We introduce a probability distribution P over the input space X
There is no need to know what P is and no restriction on P

Are We Done?



For this specific h, p' (in-sample error) is PAC equivalent to p (out-of-sample error)

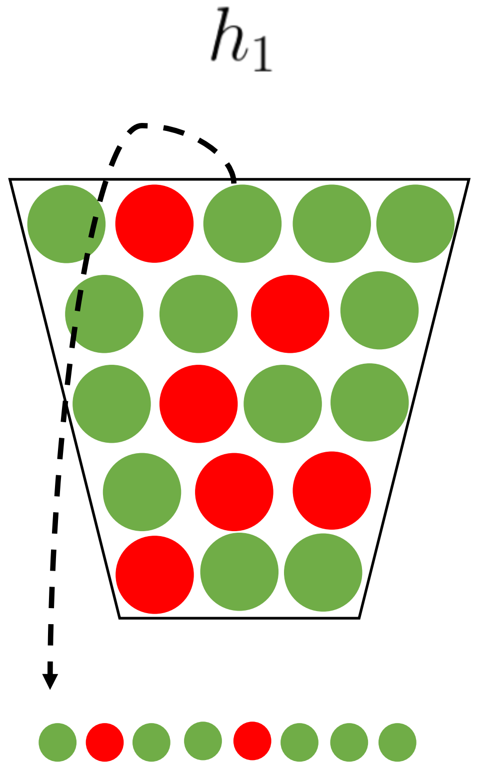
Are We Done?



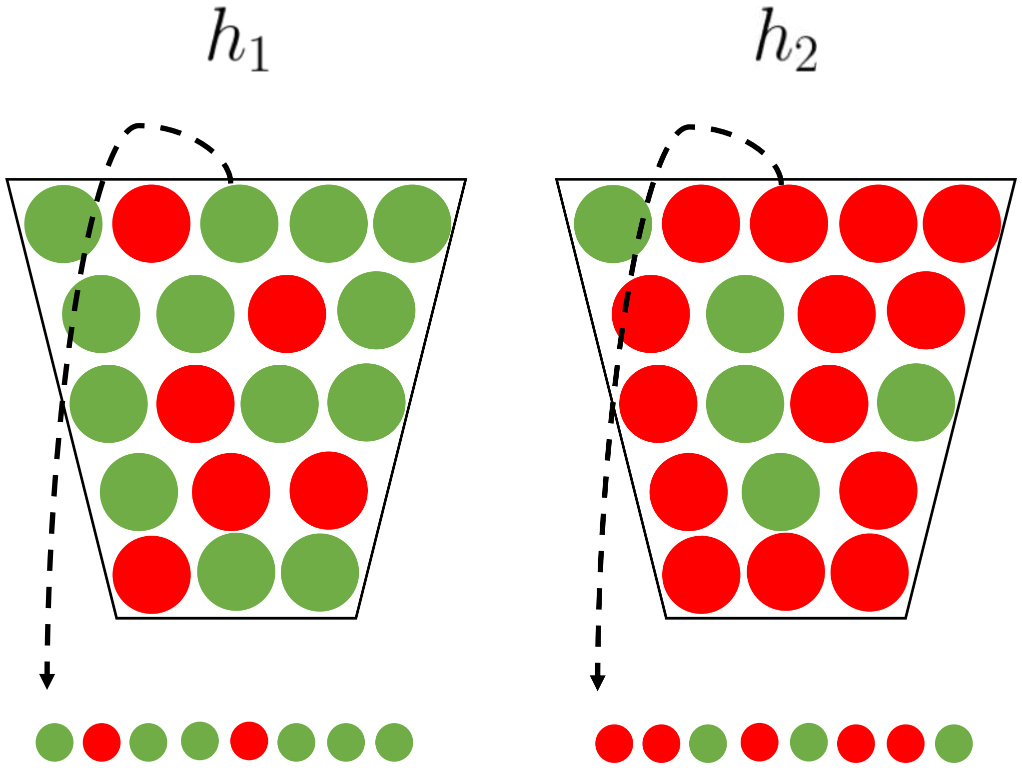
The problem here is that we fixed h

We have verified h rather than learning it from many different hypotheses

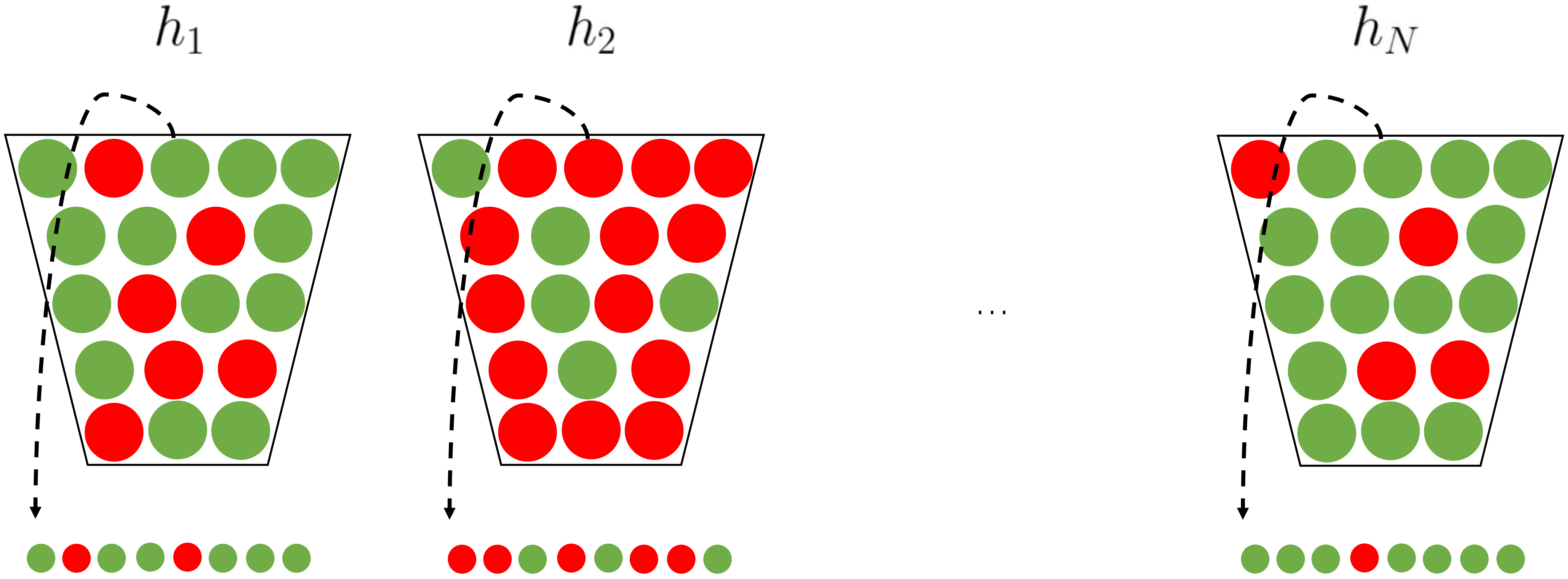
Generalize to Multiple Bins



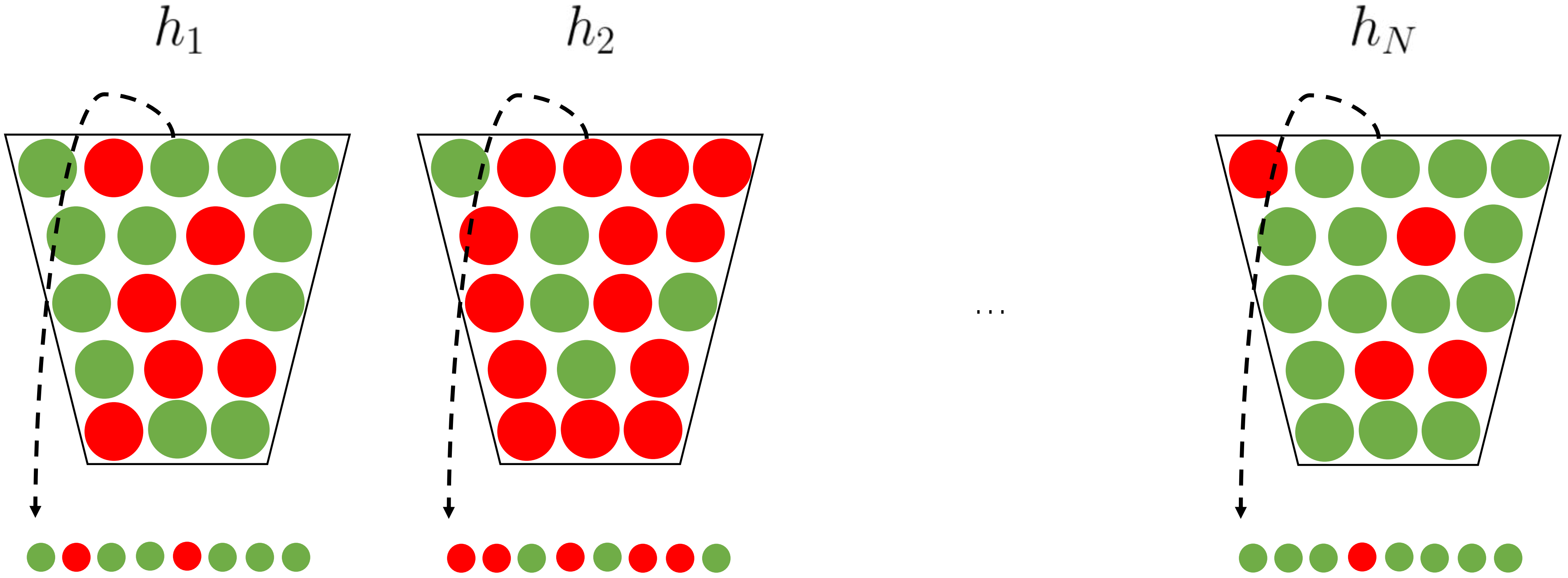
Generalize to Multiple Bins



Generalize to Multiple Bins

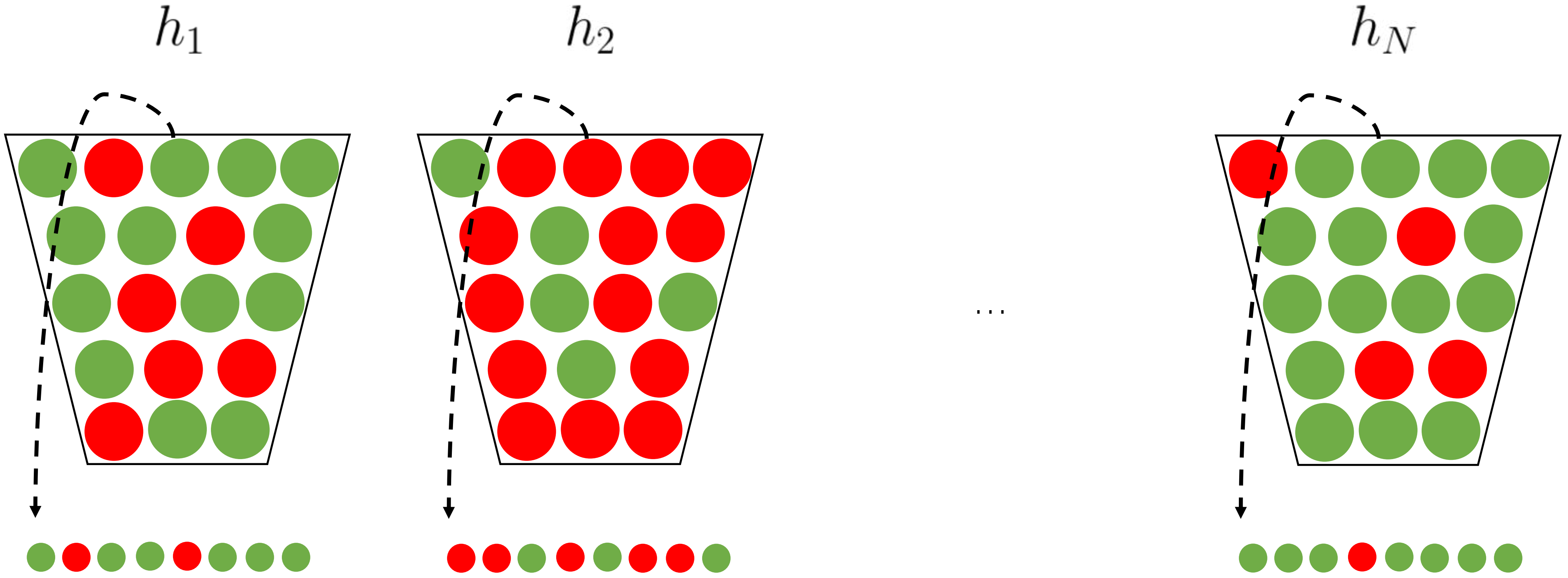


Generalize to Multiple Bins



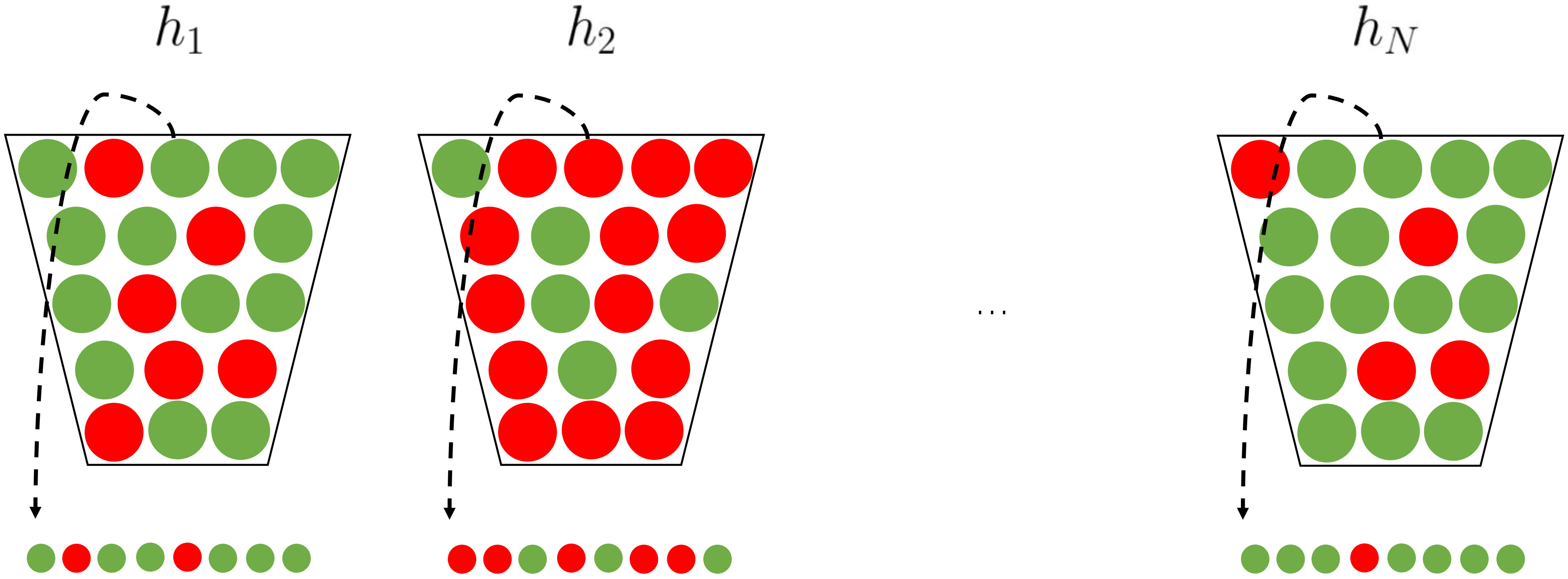
We must inspect samples generated under every hypothesis and pick the most "favorable" one

Generalize to Multiple Bins



Intuitively, this means scanning through all the *N* samples and selecting the one with the smallest value of p' (sample frequency of red marbles)

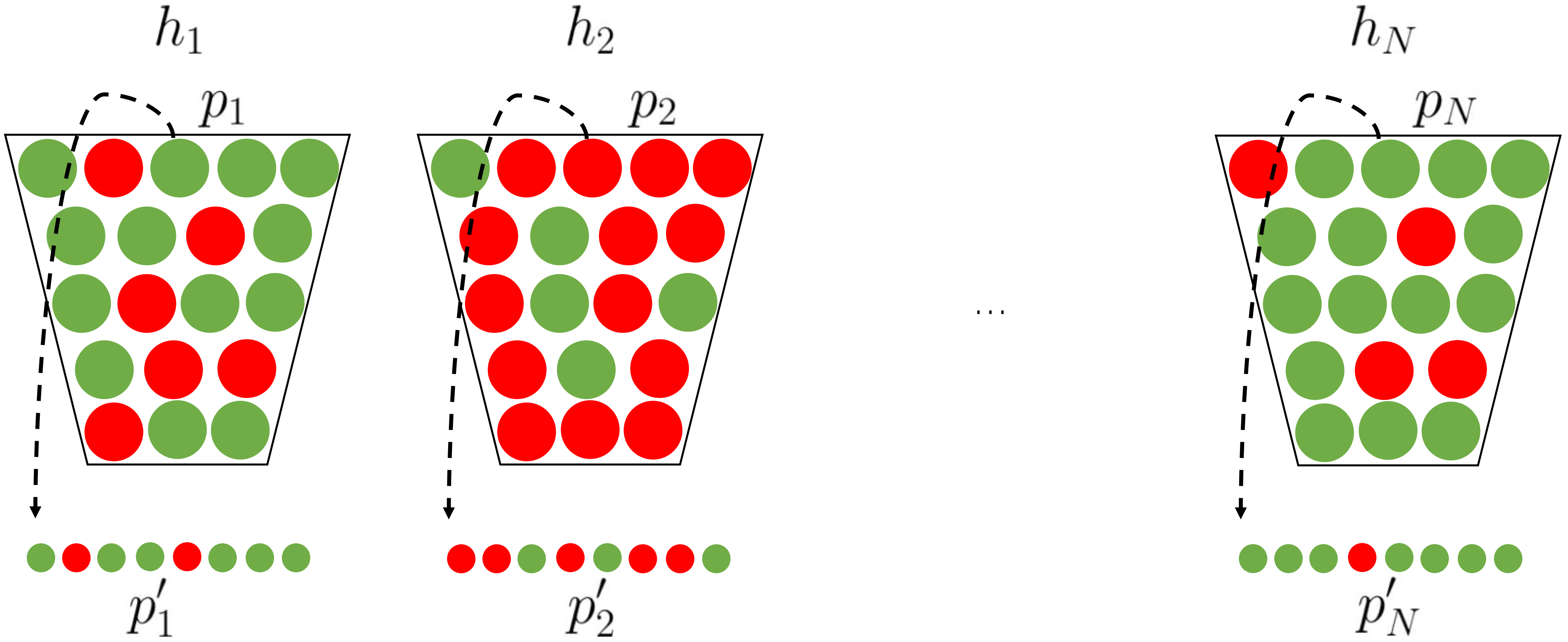
Generalize to Multiple Bins



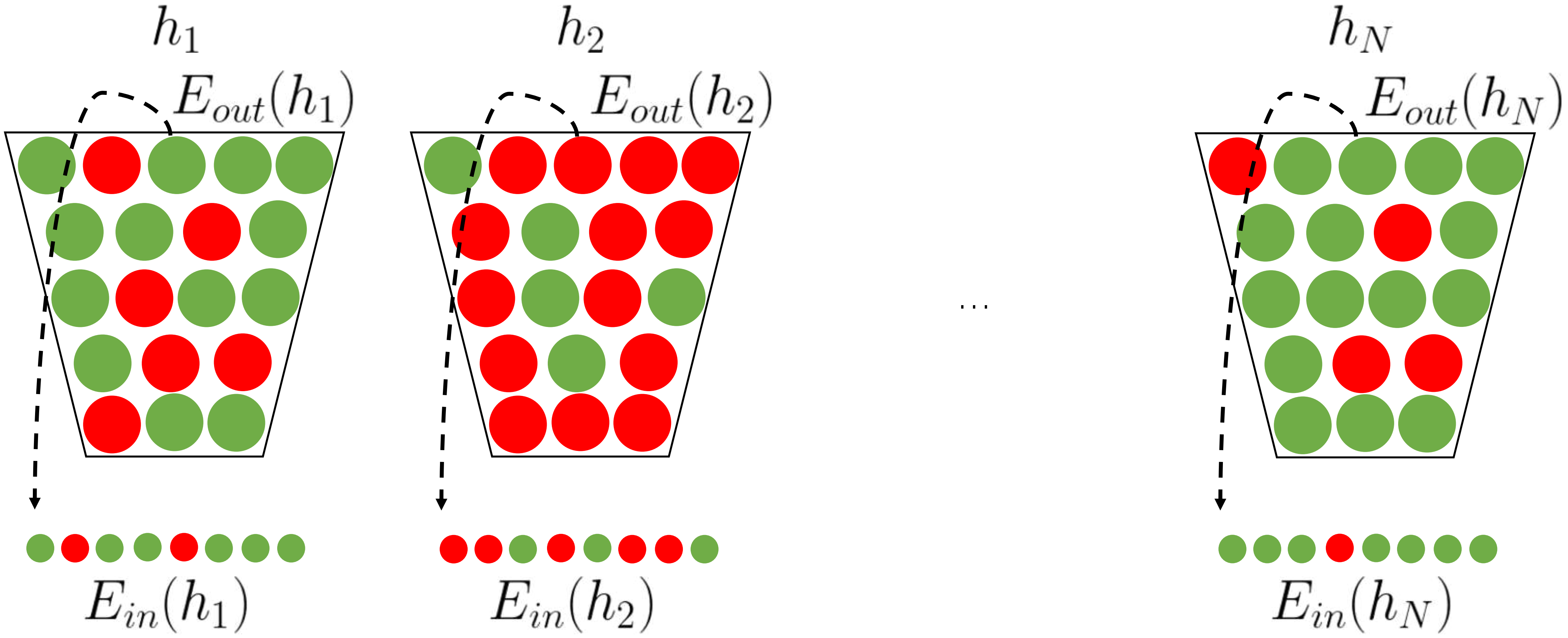
Note that p' is the in-sample error and it depends on a specific h

In other words, we will have a different in-sample error for every h

Generalize to Multiple Bins



Generalize to Multiple Bins



Generalize to Multiple Bins

$$P[|p' - p| > \epsilon] \leq 2e^{-2m\epsilon^2}$$

Generalize to Multiple Bins

$$P[|p' - p| > \epsilon] \leq 2e^{-2m\epsilon^2}$$

Hoeffding's inequality using our new notation

$$\Downarrow$$

$$P[|E_{in}(h) - E_{out}(h)| > \epsilon] \leq 2e^{-2m\epsilon^2}$$

Generalize to Multiple Bins

$$P[|p' - p| > \epsilon] \leq 2e^{-2m\epsilon^2}$$

Hoeffding's inequality using our new notation

$$\Downarrow$$

$$P[|E_{in}(h) - E_{out}(h)| > \epsilon] \leq 2e^{-2m\epsilon^2}$$

The probability that, for a given *h*, the in-sample error deviates from the true out-of-sample error by more than *epsilon* is less than or equal to a hopefully small quantity

Generalize to Multiple Bins

WAIT!

Hoeffding's inequality does NOT apply to "multiple bins"

Generalize to Multiple Bins

WAIT!

Hoeffding's inequality does NOT apply to "multiple bins"

In other words, we cannot rely on the same bound when we move from a single bin to multiple bins

Generalize to Multiple Bins

WAIT!

Hoeffding's inequality does NOT apply to "multiple bins"

In other words, we cannot rely on the same bound when we move from a single bin to multiple bins

Why?

Coin Analogy

Suppose we are given a fair coin and we toss it a few times

Coin Analogy

Suppose we are given a fair coin and we toss it a few times

$$X_i = \begin{cases} 1 & \text{if head comes up at the } i\text{-th toss} \\ 0 & \text{otherwise} \end{cases}$$

Coin Analogy

Suppose we are given a fair coin and we toss it a few times

$$X_i = \begin{cases} 1 & \text{if head comes up at the } i\text{-th toss} \\ 0 & \text{otherwise} \end{cases}$$

$$X_i \sim \text{Bernoulli}(p), p = 1/2$$

Coin Analogy

Suppose we are given a fair coin and we toss it a few times

$$X_i = \begin{cases} 1 & \text{if head comes up at the } i\text{-th toss} \\ 0 & \text{otherwise} \end{cases}$$

$$X_i \sim \text{Bernoulli}(p), p = 1/2$$

The coin is fair!

p is the parameter of the Bernoulli distribution
(i.e., the probability of "success", e.g., getting a head)

Coin Analogy

Q1: What is the probability that we will get 10 heads after 10 tosses?
(i.e., a very unlucky sample of 10 red marbles…)

Coin Analogy

Q1: What is the probability that we will get 10 heads after 10 tosses? (i.e., a very unlucky sample of 10 red marbles…)

A1: Each X_i is independent from each other and has the same p

Coin Analogy

Q1: What is the probability that we will get 10 heads after 10 tosses?
(i.e., a very unlucky sample of 10 red marbles…)

A1: Each X_i is independent from each other and has the same p

$$P(X_1 = 1, \ldots, X_{10} = 1) = \prod_{i=1}^{10} P(X_i = 1)$$

Coin Analogy

Q1: What is the probability that we will get 10 heads after 10 tosses? (i.e., a very unlucky sample of 10 red marbles…)

A1: Each X_i is independent from each other and has the same p

$$P(X_1 = 1, \ldots, X_{10} = 1) = \prod_{i=1}^{10} P(X_i = 1)$$

$$= \prod_{i=1}^{10} p = \left(\frac{1}{2}\right)^{10} \approx 0.1\%$$

Coin Analogy

Q1: What is the probability that we will get 10 heads after 10 tosses? (i.e., a very unlucky sample of 10 red marbles…)

A1: Each X_i is independent from each other and has the same p

$$P(X_1 = 1, \ldots, X_{10} = 1) = \prod_{i=1}^{10} P(X_i = 1)$$

$$= \prod_{i=1}^{10} p = \boxed{\left(\frac{1}{2}\right)^{10} \approx 0.1\%}$$

A very rare event!

Coin Analogy

We can also see this as an example of a binomial random variable

$$Y = X_1 + \ldots + X_{10}$$
$$Y \sim \text{Binomial}(n, p) = \text{Binomial}(10, 1/2)$$

$$P(Y = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

$$P(Y = 10) = \binom{10}{10} p^{10} (1 - p)^{10-10} = p^{10} = \left(\frac{1}{2}\right)^{10}$$

Coin Analogy

Q2: Suppose now we have 1,000 fair coins and we toss each of them 10 times, what is the probability that some coin will get 10 heads?

Coin Analogy

Q2: Suppose now we have 1,000 fair coins and we toss each of them 10 times, what is the probability that some coin will get 10 heads?

We can compute the probability q that a given coin does not come up with 10 heads

Coin Analogy

Q2: Suppose now we have 1,000 fair coins and we toss each of them 10 times, what is the probability that some coin will get 10 heads?

We can compute the probability q that a given coin does not come up with 10 heads

$$q = 1 - p^{10} = 1 - \left(\frac{1}{2}\right)^{10} \approx 99.9\%$$

Coin Analogy

Q2: Suppose now we have 1,000 fair coins and we toss each of them 10 times, what is the probability that some coin will get 10 heads?

Since coin tosses are i.i.d. events, the probability that no coins (out of 1,000 coins) gets 10 heads is:

$$q^{1000}$$

Coin Analogy

Q2: Suppose now we have 1,000 fair coins and we toss each of them 10 times, what is the probability that some coin will get 10 heads?

A2: The probability that at least one coin comes up 10 heads is:

$$1 - q^{1000} = 1 - \left[1 - \left(\frac{1}{2}\right)^{10}\right]^{1000} =$$

Coin Analogy

Q2: Suppose now we have 1,000 fair coins and we toss each of them 10 times, what is the probability that some coin will get 10 heads?

A2: The probability that at least one coin comes up 10 heads is:

$$1 - q^{1000} = 1 - \left[1 - \left(\frac{1}{2}\right)^{10}\right]^{1000} = 1 - \left(\frac{1023}{1024}\right)^{1000} \approx 62.4\%$$

Coin Analogy

Q2: Suppose now we have 1,000 fair coins and we toss each of them 10 times, what is the probability that some coin will get 10 heads?

A2: The probability that at least one coin comes up 10 heads is:

$$1 - q^{1000} = 1 - \left[1 - \left(\frac{1}{2}\right)^{10}\right]^{1000} = \boxed{1 - \left(\frac{1023}{1024}\right)^{1000} \approx 62.4\%}$$

Not rare at all!

Coin Analogy

- We can formulate the problem above as a sequence of $n = 1{,}000$ repeated experiments (i.e., one for each coin)

Coin Analogy

- We can formulate the problem above as a sequence of $n = 1{,}000$ repeated experiments (i.e., one for each coin)
- Each experiment is itself a sequence of 10 Bernoulli trials, where the probability of "success" is equal to getting 10 heads ($p = 2^{-10}$)

Coin Analogy

- We can formulate the problem above as a sequence of n = 1,000 repeated experiments (i.e., one for each coin)
- Each experiment is itself a sequence of 10 Bernoulli trials, where the probability of "success" is equal to getting 10 heads ($p = 2^{-10}$)
- The total number of success is given by another random variable Z

$$Z \sim \text{Binomial}(n, p),\ \ n = 1,000; p = \left(\frac{1}{2}\right)^{10}$$

Coin Analogy

$$Z \sim \text{Binomial}(n, p),\ \ n = 1,000; p = \left(\frac{1}{2}\right)^{10}$$

We therefore ask the following:

$$P(Z \geq 1) = 1 - P(Z = 0) =$$

$$= 1 - \binom{n}{0} p^0 (1-p)^{n-0} = 1 - (1-p)^{1000}$$

How Does this Relate to Learning?

- The number of coins represent the number of hypotheses

How Does this Relate to Learning?

- The number of coins represent the number of hypotheses
- In the coin example, each hypothesis is the same (as the coins are fair)

How Does this Relate to Learning?

- The number of coins represent the number of hypotheses
- In the coin example, each hypothesis is the same (as the coins are fair)
- It is actually likely that we pick an unlucky sample, even though the true out-of-sample error is 1/2 (fair coin)

How Does this Relate to Learning?

- The number of coins represent the number of hypotheses
- In the coin example, each hypothesis is the same (as the coins are fair)
- It is actually likely that we pick an unlucky sample, even though the true out-of-sample error is 1/2 (fair coin)
- Plain "vanilla" Hoeffding's inequality bound doesn't apply anymore when we have multiple hypotheses

A New Bound

Let's go back to our 1,000 coins example

$$B_i = \begin{cases} 1 & \text{if coin } i \text{ comes with 10 heads} \\ 0 & \text{otherwise} \end{cases}$$

$$B_i \sim \text{Bernoulli}(p), \;\; p = \left(\frac{1}{2}\right)^{10}$$

$$C = B_1 + B_2 \ldots + B_{1000}$$

A New Bound

Let's go back to our 1,000 coins example

$$C = B_1 + B_2 \ldots + B_{1000}$$

A New Bound

Let's go back to our 1,000 coins example

$$C = B_1 + B_2 \ldots + B_{1000}$$

$$P(C \geq 1) = P(B_1 = 1 \textbf{ or } B_2 = 1 \textbf{ or } \cdots \textbf{ or } B_{1000} = 1)$$

A New Bound

Let's go back to our 1,000 coins example

$$C = B_1 + B_2 \ldots + B_{1000}$$

$$P(C \geq 1) = P(B_1 = 1 \textbf{ or } B_2 = 1 \textbf{ or } \cdots \textbf{ or } B_{1000} = 1)$$

$$= P\left(\bigcup_{i=1}^{1000} B_i\right)$$

A New Bound

Let's go back to our 1,000 coins example

$$C = B_1 + B_2 \ldots + B_{1000}$$

$$P(C \geq 1) = P(B_1 = 1 \textbf{ or } B_2 = 1 \textbf{ or } \cdots \textbf{ or } B_{1000} = 1)$$

$$= P\left(\bigcup_{i=1}^{1000} B_i\right) \leq \sum_{i=1}^{1000} P(B_i = 1) = 1000 \cdot \left(\frac{1}{2}\right)^{10} \approx 97.7\%$$

A New Bound

Let's go back to our 1,000 coins example

$$C = B_1 + B_2 \ldots + B_{1000}$$

$$P(C \geq 1) = P(B_1 = 1 \textbf{ or } B_2 = 1 \textbf{ or } \cdots \textbf{ or } B_{1000} = 1)$$

$$= \boxed{P\left(\bigcup_{i=1}^{1000} B_i\right) \leq \sum_{i=1}^{1000} P(B_i = 1)} = 1000 \cdot \left(\frac{1}{2}\right)^{10} \approx 97.7\%$$

Boole's inequality (a.k.a. Union Bound)

A New Bound

$$P(C \geq 1) = P\left(\bigcup_{i=1}^{1000} B_i\right) \leq \sum_{i=1}^{1000} P(B_i = 1) = 1000 \cdot \left(\frac{1}{2}\right)^{10} \approx 97.7\%$$

Boole's inequality (a.k.a. Union Bound)

Note that this bound is not so tight!

A New Bound

$$P(C \geq 1) = P\left(\bigcup_{i=1}^{1000} B_i\right) \leq \sum_{i=1}^{1000} P(B_i = 1) = 1000 \cdot \left(\frac{1}{2}\right)^{10} \approx 97.7\%$$

Boole's inequality (a.k.a. Union Bound)

Note that this bound is not so tight!

If we considered $n = 1{,}024$ coins we would obtain the trivial bound

$$\boxed{P(C \geq 1) \leq 1}$$

Union Bound for the Learning Problem

$$X_i = \begin{cases} 1 & \text{if } |E_{in}(h_i) - E_{out}(h_i)| > \epsilon \\ 0 & \text{otherwise} \end{cases}$$

Union Bound for the Learning Problem

$$X_i = \begin{cases} 1 & \text{if } |E_{in}(h_i) - E_{out}(h_i)| > \epsilon \\ 0 & \text{otherwise} \end{cases}$$

$$\begin{aligned}
&P(|E_{in}(h^*) - E_{out}(h^*)| > \epsilon) \leq \\
&P(X_1 = 1 \textbf{ or } X_2 = 1 \textbf{ or } \cdots \textbf{ or } X_N = 1) = \\
&P(|E_{in}(h_1) - E_{out}(h_1)| > \epsilon) + \\
&P(|E_{in}(h_2) - E_{out}(h_2)| > \epsilon) + \\
&\ldots \\
&P(|E_{in}(h_N) - E_{out}(h_N)| > \epsilon)
\end{aligned}$$

Union Bound for the Learning Problem

$$X_i = \begin{cases} 1 & \text{if } |E_{in}(h_i) - E_{out}(h_i)| > \epsilon \\ 0 & \text{otherwise} \end{cases}$$

$$P(|E_{in}(h^*) - E_{out}(h^*)| > \epsilon) \leq$$
$$P(X_1 = 1 \textbf{ or } X_2 = 1 \textbf{ or } \cdots \textbf{ or } X_N = 1) =$$
$$\left.\begin{array}{l} P(|E_{in}(h_1) - E_{out}(h_1)| > \epsilon) + \\ P(|E_{in}(h_2) - E_{out}(h_2)| > \epsilon) + \\ \ldots \\ P(|E_{in}(h_N) - E_{out}(h_N)| > \epsilon) \end{array}\right.$$

Assuming a finite set of N hypotheses

Union Bound for the Learning Problem

$$X_i = \begin{cases} 1 & \text{if } |E_{in}(h_i) - E_{out}(h_i)| > \epsilon \\ 0 & \text{otherwise} \end{cases}$$

$$\begin{aligned} & P(|E_{in}(h^*) - E_{out}(h^*)| > \epsilon) \leq \\ & P(X_1 = 1 \textbf{ or } X_2 = 1 \textbf{ or } \cdots \textbf{ or } X_N = 1) = \\ & P(|E_{in}(h_1) - E_{out}(h_1)| > \epsilon) + \\ & P(|E_{in}(h_2) - E_{out}(h_2)| > \epsilon) + \\ & \cdots \\ & P(|E_{in}(h_N) - E_{out}(h_N)| > \epsilon) \end{aligned}$$

We can apply Hoeffding's inequality to each of them

Union Bound for the Learning Problem

$$X_i = \begin{cases} 1 & \text{if } |E_{in}(h_i) - E_{out}(h_i)| > \epsilon \\ 0 & \text{otherwise} \end{cases}$$

$$\begin{aligned}
&P(|E_{in}(h^*) - E_{out}(h^*)| > \epsilon) \leq \\
&P(X_1 = 1 \textbf{ or } X_2 = 1 \textbf{ or } \cdots \textbf{ or } X_N = 1) = \\
&P(|E_{in}(h_1) - E_{out}(h_1)| > \epsilon) + \\
&P(|E_{in}(h_2) - E_{out}(h_2)| > \epsilon) + \\
&\ldots \\
&P(|E_{in}(h_N) - E_{out}(h_N)| > \epsilon) \boxed{\leq \sum_{i=1}^{N} 2e^{-2m\epsilon^2} = 2Ne^{-2m\epsilon^2}}
\end{aligned}$$

Final Considerations

- h^* represents our chosen hypothesis (among a set of N possible ones)

Final Considerations

- h^* represents our chosen hypothesis (among a set of N possible ones)
- Thanks to the Boole's inequality we can give an adjusted bound to the probability of picking a "bad" hypothesis

Final Considerations

- h^* represents our chosen hypothesis (among a set of N possible ones)
- Thanks to the Boole's inequality we can give an adjusted bound to the probability of picking a "bad" hypothesis
- A bad hypothesis is one whose in-sample performance deviates from out-of-sample performance by more than a tolerance *epsilon*

Final Considerations

- h^* represents our chosen hypothesis (among a set of N possible ones)
- Thanks to the Boole's inequality we can give an adjusted bound to the probability of picking a "bad" hypothesis
- A bad hypothesis is one whose in-sample performance deviates from out-of-sample performance by more than a tolerance *epsilon*
- Note, though, that the bound we came up with is not tight at all as it assumes the worst case scenario
    - Each event of choosing a "bad" hypothesis is disjoint from each other

Final Considerations

- What about infinite hypothesis space?

Final Considerations

- What about infinite hypothesis space?
- After all, if we focus on, say, linear functions the hypothesis space is infinite

Final Considerations

- What about infinite hypothesis space?
- After all, if we focus on, say, linear functions the hypothesis space is infinite
- The mutual independence assumption we made for the worst case scenario would make the generalization gap unbound!

Final Considerations

- What about infinite hypothesis space?
- After all, if we focus on, say, linear functions the hypothesis space is infinite
- The mutual independence assumption we made for the worst case scenario would make the generalization gap unbound!
- But this is definitely too pessimistic!

Final Considerations

- What about infinite hypothesis space?
- After all, if we focus on, say, linear functions the hypothesis space is infinite
- The mutual independence assumption we made for the worst case scenario would make the generalization gap unbound!
- But this is definitely too pessimistic!

Take-home message
Learning is feasible in a **probabilistic sense**!