MACHINE LEARNING

LEARNING THEORY

AGENDA

01 Preliminaries

Risk minimization, union bound, and Hoeffding inequality.

02 The case of finite $\mathcal{H}$

Uniform convergence, sample complexity, error bound, bias-variance tradeoff.

03 The case of infinite $\mathcal{H}$

VC dimension

AI

PRELIMINARIES

LEARNING THEORY

PRELIMINARIES

We want to answer three main questions:

1. **Can we make formal the bias/variance tradeoff?**

2. **Why should doing well on the training set tell us anything about generalization error?**

3. **Are there conditions under which we can actually prove that learning algorithms will work well?**

LEARNING THEORY

PRELIMINARIES

We are going to **define** a **binary classifier**:

$$\boldsymbol{h_w(x) = g\left(w^T x\right)}$$

$$\boldsymbol{g(z) = 1\{z \geq 0\}}$$

We establish our **training set as**:

$$\boldsymbol{S = \left\{\left(x^{(i)}, y^{(i)}\right)\right\}_{i=1}^{m}, S \sim_{iid} D}$$

LEARNING THEORY

PRELIMINARIES

We are going to define the **training error** $\widehat{\mathcal{E}}_S$ of a hypothesis $\boldsymbol{h_w}$ in a **simple way:**

$$\widehat{\mathcal{E}}_S(\boldsymbol{h_w}) = \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}\{\boldsymbol{h_w}(\boldsymbol{x}^{(i)}) \neq \boldsymbol{y}^{(i)}\}$$

Fraction of data points where the hypothesis is wrong.

The **training error** is also called **RISK.**

LEARNING THEORY

PRELIMINARIES

As always, our **objective** consists in **minimizing** the **risk** (**training error**):

$$\hat{\boldsymbol{w}} = \underset{\boldsymbol{w}}{\mathbf{argmin}}\, \widehat{\mathcal{E}}_{\boldsymbol{S}}(\boldsymbol{h}_{\boldsymbol{w}})$$

Minimizing this **expression** deals with a **non-convex optimization problem**.
Logistic regression and **SVM** are **convex approximations** to this **problem**.

LEARNING THEORY

PRELIMINARIES

We are going to **change** the **problem.** Now, the **objective** will reside in **choosing** the **hypothesis** function $\boldsymbol{h_w}$ **instead** of the **parameters** $\boldsymbol{w}$.

Thus, we define the **hypothesis class** $\mathcal{H}$ as the **class** of all **linear classifiers** that the **algorithm** is **choosing from**.

$$\mathcal{H} = \{\boldsymbol{h_w}: \boldsymbol{h_w}(\boldsymbol{x}) = \mathbf{1}\{\boldsymbol{w}^T\boldsymbol{x} \geq \mathbf{0}\}, \boldsymbol{w} \in \mathbb{R}^{n+1}\}$$

Therefore, **empirical risk minimization** is **redefined** as:

$$\widehat{\boldsymbol{h_w}} = \underset{\boldsymbol{h_w} \in \mathcal{H}}{\mathbf{argmin}}\, \widehat{\mathcal{E}}_S(\boldsymbol{h_w})$$

NOTE: the **hypothesis class** $\mathcal{H}$ can represent any set of functions.

LEARNING THEORY

PRELIMINARIES

Let us remember that the **main goal** resides in the **generalization error not** in the **training error**. The **generalization error** would be **defined as**:

$$\mathcal{E}\,(h_w) = P_{(x,y)\sim D}(h_w(x) \neq y)$$

Probability that, if we now **draw** a **new example** $(\boldsymbol{x}, \boldsymbol{y})$ **from** the **distribution D,** $\boldsymbol{h}$ will **misclassify** it.

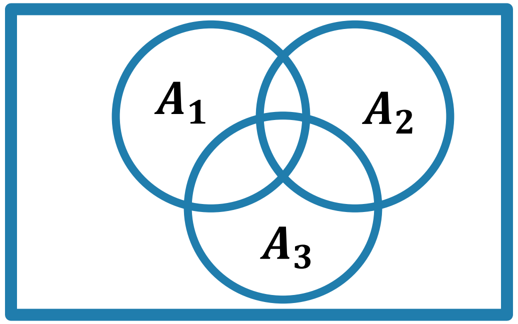
We want to **make** an **estimation** $\hat{\mathcal{E}}_{\boldsymbol{S}}$ **(training error)** to get **close** to the **generalization error** $\mathcal{E}\,(\boldsymbol{h_w})$.

LEARNING THEORY

PRELIMINARIES

To **reduce** the **generalization error** we will **need two lemmas**:

1. **Union bound:** Let $A_1, A_2, \ldots, A_k$ be k **different events** (that may not be independent). Then $P(A_1 \cup \cdots \cup A_k) \leq P(A_1) + \cdots + P(A_k)$.



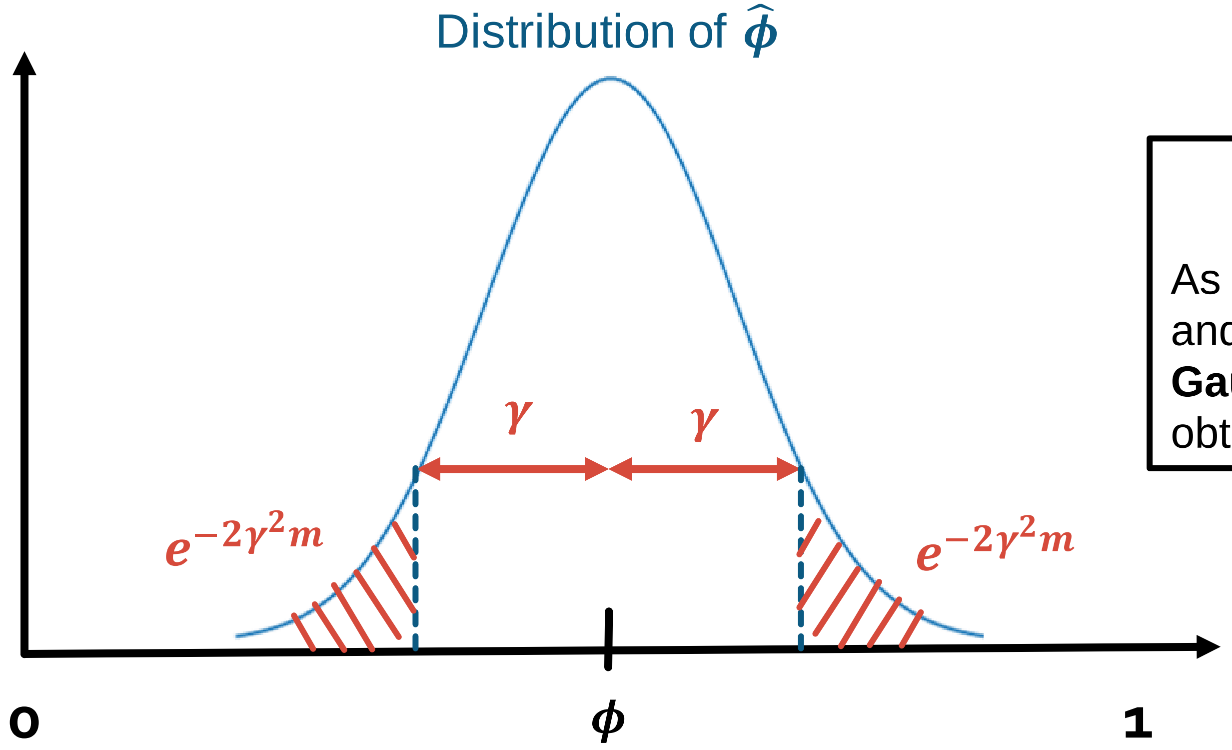
2. **Hoeffding inequality:** let $Z1, \ldots, Zn$ be n **independent** and **identically distributed (iid) random variables** drawn from a **Bernoulli**(φ) **distribution** with **mean** ϕ. Therefore, $P(Z_i = 1) = \phi$ and $P(Z_i = 0) = 1 - \phi$. Let $\hat{\phi} = \frac{1}{m}\sum_{i=1}^{m} z_i$ and let any $\gamma > 0$ be fixed. Then

$$P\left(\left|\phi - \hat{\phi}\right| > \gamma\right) \leq 2e^{-2\gamma^2 m}$$

LEARNING THEORY

PRELIMINARIES

The **Hoeffding inequality** says that if we take $\hat{\phi}$ (the **average** of m **Bernoulli(ϕ) random variables**) to be our **estimate** of ϕ**,** then the **probability** of **our being far** from the **true value** is **small,** so long as m is **large.**



Central Limit Theorem

As you take more m samples and **average them**, a **Gaussian distribution** is obtained.

THE CASE OF FINITE $\mathcal{H}$

LEARNING THEORY

THE CASE OF FINITE $\mathcal{H}$

Let us consider that we have a **finite hypothesis class** $\mathcal{H} = \{h_1, \ldots, h_k\}$ consisting of $\boldsymbol{k}$ **hypotheses** or **functions** mapping from χ to $\{0,1\}$.

Risk minimization will **choose** the **hypothesis** with the **lowest training error**.

We are going to **prove** that:

1. $\hat{\mathcal{E}}_{\boldsymbol{S}} \approx \mathcal{E}$.

2. There is an upper-bound to $\hat{\mathcal{E}}_{\boldsymbol{S}}$.

Therefore, if we **minimize** the **training error,** the **generalization** error **will decrease** as well.

LEARNING THEORY

THE CASE OF FINITE $\mathcal{H}$

We are going to take a **fixed hypothesis** $\boldsymbol{h_j} \in \mathcal{H}$ and will consider a **Bernoulli** random **variable** $\boldsymbol{Z_i \underset{iid}{\sim} D}$ which **misclassifies** an **example** $\boldsymbol{Z_i = 1\{h_j(x^{(i)}) \neq y^{(i)}\} \in \{0, 1\}}$.

The **probability** that, from a **fixed hypothesis** $\boldsymbol{h_j}$, we **misclassify** an **example** is the **expected value (mean of the distribution)** or **generalization error**:

$$P(\boldsymbol{Z_i} = 1) = \mathcal{E}(\boldsymbol{h_j})$$

On the other hand, we know that the **training error** is computed as the **fraction** of **misclassified examples (mean of the sample)**:

$$\hat{\mathcal{E}}_S(\boldsymbol{h_j}) = \frac{\mathbf{1}}{\boldsymbol{m}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{m}} \boldsymbol{Z_i} = \frac{\mathbf{1}}{\boldsymbol{m}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{m}} \mathbf{1}\{\boldsymbol{h_j}(\boldsymbol{x^{(i)}}) \neq \boldsymbol{y^{(i)}}\}$$

LEARNING THEORY

THE CASE OF FINITE $\mathcal{H}$

We will use the **Hoeffding inequality** to look at the **difference between** the **generalization** and **training errors**:

$$P\left(\left|\varepsilon\left(h_j\right) - \hat{\varepsilon}\left(h_j\right)\right| > \gamma\right) \leq 2e^{-2\gamma^2 m}$$

We have **proved** that for a **fixed hypothesis** h_j and a **large training set**, the **training error** will **approximate** the **generalization error** with a **high probability**.

Now, let us **prove** this statement for **all** $h \in \mathcal{H}$.

THE CASE OF FINITE $\mathcal{H}$

Let us think of $\boldsymbol{A_j}$ as an event $|\mathcal{E}(\boldsymbol{h_j}) - \hat{\mathcal{E}}(\boldsymbol{h_j})| > \boldsymbol{\gamma}$, thus $\boldsymbol{P}(\boldsymbol{A_j}) \leq \boldsymbol{2e^{-2\gamma^2 m}}$. Using the union bound lemma we have:

$$\boldsymbol{P}\left(\exists \boldsymbol{h_j} \in \mathcal{H} / |\mathcal{E}(\boldsymbol{h_j}) - \hat{\mathcal{E}}(\boldsymbol{h_j})| > \boldsymbol{\gamma}\right) = \boldsymbol{P}\left(\boldsymbol{A_1} \bigcup \cdots \bigcup \boldsymbol{A_k}\right)$$

$$\boldsymbol{P}\left(\exists \boldsymbol{h} \in \mathcal{H} / |\mathcal{E}(\boldsymbol{h_j}) - \hat{\mathcal{E}}(\boldsymbol{h_j})| > \boldsymbol{\gamma}\right) \leq \sum_{i=1}^{k} \boldsymbol{P}(\boldsymbol{A_i})$$

$$\boldsymbol{P}\left(\exists \boldsymbol{h} \in \mathcal{H} / |\mathcal{E}(\boldsymbol{h_j}) - \hat{\mathcal{E}}(\boldsymbol{h_j})| > \boldsymbol{\gamma}\right) \leq \sum_{i=1}^{k} \boldsymbol{2e^{-2\gamma^2 m}}$$

$$\boldsymbol{P}\left(\exists \boldsymbol{h} \in \mathcal{H} / |\mathcal{E}(\boldsymbol{h_j}) - \hat{\mathcal{E}}(\boldsymbol{h_j})| > \boldsymbol{\gamma}\right) \leq \boldsymbol{2ke^{-2\gamma^2 m}}$$

LEARNING THEORY

THE CASE OF FINITE $\mathcal{H}$

The **probability** that such **hypothesis does not exist** is defined as:

$$P\left(\neg\exists h_j \in \mathcal{H} / \left|\mathcal{E}(h_j) - \hat{\mathcal{E}}(h_j)\right| > \gamma\right) = 1 - P\left(\exists h_j \in \mathcal{H} / \left|\mathcal{E}(h_j) - \hat{\mathcal{E}}(h_j)\right| > \gamma\right)$$

$$P\left(\neg\exists h_j \in \mathcal{H} / \left|\mathcal{E}(h_j) - \hat{\mathcal{E}}(h_j)\right| > \gamma\right) = P\left(\forall h_j \in \mathcal{H} / \left|\mathcal{E}(h_j) - \hat{\mathcal{E}}(h_j)\right| \leq \gamma\right)$$

$$P\left(\neg\exists h_j \in \mathcal{H} / \left|\mathcal{E}(h_j) - \hat{\mathcal{E}}(h_j)\right| > \gamma\right) \geq 1 - 2ke^{-2\gamma^2 m}$$

We have proved that with **probability at least $1 - 2ke^{-2\gamma^2 m}$**, the **generalization error** $\mathcal{E}(h_j)$ will be **within** a **distance** γ of the **training error** $\hat{\mathcal{E}}(h_j)$ for all $h \in \mathcal{H}$.

UNIFORM CONVERGENCE

(Holds for all hypotheses)

LEARNING THEORY

THE CASE OF FINITE $\mathcal{H}$

There are **three quantities** of **interest**: $\boldsymbol{m}$**,** $\boldsymbol{\gamma}$, and the **probability** of **error** $\boldsymbol{\delta}$. We can **bound one** in terms of the other two.

Given $\boldsymbol{\delta}$ and $\boldsymbol{\gamma}$, we can obtain the **size** of the **training set** $\boldsymbol{m}$ for which the **training error** will be **within** $\boldsymbol{\gamma}$ of the **generalization error** with at least probability $\mathbf{1}-\boldsymbol{\delta}$.

$$\boldsymbol{\delta} = \mathbf{2}\boldsymbol{k}\boldsymbol{e}^{-\mathbf{2}\boldsymbol{\gamma}^{\mathbf{2}}\boldsymbol{m}}$$

$$\boldsymbol{m} \geq -\frac{\mathbf{1}}{\mathbf{2}\boldsymbol{\gamma}^{\mathbf{2}}}\log\left(\frac{\delta}{2k}\right)$$

$$\boldsymbol{m} \geq \frac{\mathbf{1}}{\mathbf{2}\boldsymbol{\gamma}^{\mathbf{2}}}\log\left(\left(\frac{\delta}{2k}\right)^{-1}\right)$$

$$\boldsymbol{m} \geq \frac{\mathbf{1}}{\mathbf{2}\boldsymbol{\gamma}^{\mathbf{2}}}\log\left(\frac{2k}{\delta}\right)$$

LEARNING THEORY

THE CASE OF FINITE $\mathcal{H}$

Therefore, we need a **training set size** of $\boldsymbol{m} \geq \frac{1}{2\gamma^2} \boldsymbol{log}\left(\frac{2k}{\delta}\right)$ to guarantee that with **probability** of **at least 1** $- \boldsymbol{\delta}$, we have that the **training error** $\hat{\mathcal{E}}(\boldsymbol{h_j})$ is within $\boldsymbol{\gamma}$ of the **generalization error** $\mathcal{E}(\boldsymbol{h_j})$. Formally

$$\left|\mathcal{E}(\boldsymbol{h_j}) - \hat{\mathcal{E}}(\boldsymbol{h_j})\right| \leq \boldsymbol{\gamma}, \forall \boldsymbol{h} \in \mathcal{H}$$

SAMPLE COMPLEXITY

(Number of training examples needed to achieve a certain bounding error)

NOTE: we can see that **even** if we **increase** the **number** of **hypotheses** $\boldsymbol{k}$ in the class $\mathcal{H}$, the **number** of **training examples** $\boldsymbol{m}$ needed will **remain small**.

LEARNING THEORY

THE CASE OF FINITE $\mathcal{H}$

If we hold $\boldsymbol{m}$ and $\boldsymbol{\delta}$ **fixed**, we can get the **following**:

$$\boldsymbol{\delta} = \boldsymbol{2ke^{-2\gamma^2 m}}$$

$$\boldsymbol{\gamma} = \sqrt{\frac{\mathbf{1}}{\mathbf{2}\boldsymbol{m}}\log\left(\frac{2k}{\delta}\right)}$$

We want to make $\boldsymbol{\gamma}$ the **upper bound error,** thus

$$\left|\varepsilon(\boldsymbol{h_j}) - \hat{\varepsilon}(\boldsymbol{h_j})\right| \leq \boldsymbol{\gamma}$$

$$\left|\varepsilon(\boldsymbol{h_j}) - \hat{\varepsilon}(\boldsymbol{h_j})\right| \leq \sqrt{\frac{\mathbf{1}}{\mathbf{2}\boldsymbol{m}}\log\left(\frac{2k}{\delta}\right)}$$

LEARNING THEORY

THE CASE OF FINITE $\mathcal{H}$

Let us assume that the **uniform convergence** $\forall \boldsymbol{h} \in \mathcal{H}, \left|\mathcal{E}(\boldsymbol{h_j}) - \widehat{\mathcal{E}}(\boldsymbol{h_j})\right| \leq \boldsymbol{\gamma}$ **holds.**

Is there something that we can **prove** about the **generalization error** $\mathcal{E}$ using the **estimated hypothesis** $\widehat{\boldsymbol{h}}$ with **empirical risk minimization**? Remembering that:

$$\widehat{\boldsymbol{h}} = \underset{h \in \mathcal{H}}{\mathbf{argmin}}\, \widehat{\mathcal{E}}(\boldsymbol{h})$$

Now, we are going to define the **best hypothesis** as the hypothesis that **minimizes** the **generalization error**:

$$\boldsymbol{h}^* = \underset{h \in \mathcal{H}}{\boldsymbol{argmin}}\, \mathcal{E}(\boldsymbol{h})$$

LEARNING THEORY

THE CASE OF FINITE $\mathcal{H}$

Starting with the **uniform convergence assumption** we have:

$$\left|\varepsilon(\boldsymbol{h_j}) - \hat{\varepsilon}(\boldsymbol{h_j})\right| \le \boldsymbol{\gamma}$$

$$\varepsilon(\hat{\boldsymbol{h}}) - \hat{\varepsilon}(\hat{\boldsymbol{h}}) \le \boldsymbol{\gamma}$$

$$\varepsilon(\hat{\boldsymbol{h}}) \le \hat{\varepsilon}(\hat{\boldsymbol{h}}) + \boldsymbol{\gamma}$$

Because we obtained $\hat{\boldsymbol{h}}$ with **empirical risk minimization**, there is **no** other **hypothesis** with **less training error** than $\hat{\boldsymbol{h}}$, thus $\hat{\varepsilon}(\hat{\boldsymbol{h}}) \le \hat{\varepsilon}(\boldsymbol{h}^*)$ and the **inequality remains true:**

$$\varepsilon(\hat{\boldsymbol{h}}) \le \hat{\varepsilon}(\boldsymbol{h}^*) + \boldsymbol{\gamma}$$

$$\varepsilon(\hat{\boldsymbol{h}}) \le \hat{\varepsilon}(\boldsymbol{h}^*) + \boldsymbol{2\gamma}$$

LEARNING THEORY

THE CASE OF FINITE $\mathcal{H}$

Theorem:

Let $|\boldsymbol{H}| = \boldsymbol{k}$, and let any $\boldsymbol{n}, \boldsymbol{\delta}$ be fixed. Then with **probability at least $\mathbf{1} - \boldsymbol{\delta}$**, we have that

$$\varepsilon(\hat{\boldsymbol{h}}) \leq \hat{\varepsilon}(\boldsymbol{h}^*) + \mathbf{2}\boldsymbol{\gamma}$$

$$\varepsilon(\hat{\boldsymbol{h}}) \leq \min_{h \in \mathcal{H}} \varepsilon(\boldsymbol{h}) + \mathbf{2}\sqrt{\frac{\mathbf{1}}{\mathbf{2}\boldsymbol{m}} \boldsymbol{log}\left(\frac{\mathbf{2}\boldsymbol{k}}{\boldsymbol{\delta}}\right)}$$

Thus, our **generalization error** of the **hypothesis** obtained with **ERM** $\varepsilon(\hat{\boldsymbol{h}})$ will be at most $\mathbf{2}\boldsymbol{\gamma}$ **higher** than the **training error** of the **best possible hypothesis** $\hat{\varepsilon}(\boldsymbol{h}^*)$.

LEARNING THEORY

THE CASE OF FINITE $\mathcal{H}$

By **analyzing** the **theorem** we can see the following:

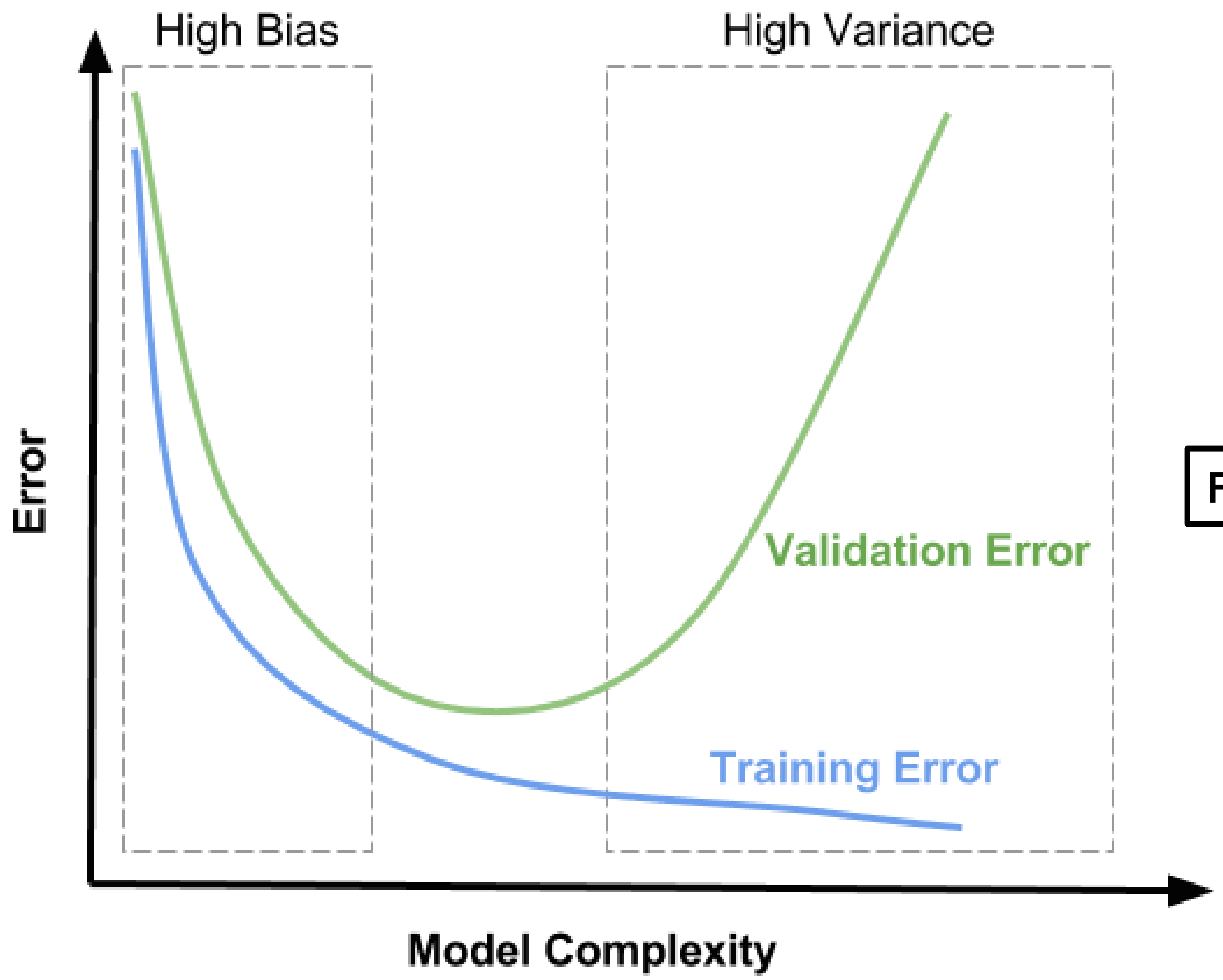
$$\varepsilon(\hat{h}) \leq \min_{h \in \mathcal{H}} \varepsilon(h) + 2\sqrt{\frac{1}{2m} \log\left(\frac{2k}{\delta}\right)}$$

If we switch to a **larger hypothesis class** $\mathcal{H}' \supseteq \mathcal{H}$ (i.e. quadratic), then $\min_{h \in \mathcal{H}} \varepsilon(h)$ will **decrease** because we have a larger set of hypothesis for which we can obtain the minimum, thus we reduce the bias.

On the other hand, k will **become larger** resulting in an **increase** of the **second term**, thus increasing the variance.

LEARNING THEORY

THE CASE OF FINITE $\mathcal{H}$



Fixed m

LEARNING THEORY

THE CASE OF FINITE $\mathcal{H}$

By looking at the **training** set **size** $\boldsymbol{m}$**,** we can obtain the following **complexity bound**:

Corollary: Let $|\mathcal{H}| = \boldsymbol{k}$, and let any $\boldsymbol{\delta}, \boldsymbol{\gamma}$ **be fixed**. Then for $\boldsymbol{\varepsilon}(\hat{\boldsymbol{h}}) \leq \min_{h \in H} \boldsymbol{\varepsilon}(\boldsymbol{h}) + \mathbf{2}\boldsymbol{\gamma}$ to hold with **probability at least** $\mathbf{1} - \boldsymbol{\delta}$, it suffices that:

$$\boldsymbol{m} > \frac{\mathbf{1}}{\mathbf{2}\boldsymbol{\gamma}^{\mathbf{2}}} \log\left(\frac{2k}{\delta}\right)$$

$$\boldsymbol{m} = \boldsymbol{O}\left(\frac{\mathbf{1}}{\boldsymbol{\gamma}^{\mathbf{2}}} \log\left(\frac{k}{\delta}\right)\right)$$

AI

THE CASE OF INFINITE $\mathcal{H}$

LEARNING THEORY

THE CASE OF INFINITE $\mathcal{H}$

Many **hypothesis classes** contain an **infinite** number of **functions** (i.e. **any function parametrized** by **real** numbers). We want to **prove** the **previous results** for this **infinite space** of **functions**.

We are going to make some **statements** that are **not correct** at all but **will help** with the **understanding** of the **proof**.

Let us say that the class $\boldsymbol{\mathcal{H}}$ is parametrized by $\boldsymbol{d}$ **real** numbers. Because we are constrained by computers that **use 64 bits** to represent floating-point numbers, we have **at most** $\boldsymbol{k = 2^{64d}}$ **different hypotheses**.

LEARNING THEORY

THE CASE OF INFINITE $\mathcal{H}$

To **hold** the **theorem** $\mathcal{E}(\widehat{\boldsymbol{h}}) \leq \widehat{\mathcal{E}}(\boldsymbol{h}^{*}) + \mathbf{2}\boldsymbol{\gamma}$ as **valid** with at least **probability** $\mathbf{1} - \boldsymbol{\delta}$ **,** we need to suffice the corollary:

$$\boldsymbol{m} = \boldsymbol{O}\left(\frac{\mathbf{1}}{\boldsymbol{\gamma}^{\mathbf{2}}} \log\left(\frac{k}{\delta}\right)\right)$$

$$\boldsymbol{m} = \boldsymbol{O}\left(\frac{\mathbf{1}}{\boldsymbol{\gamma}^{\mathbf{2}}} \log\left(\frac{\mathbf{2}^{\mathbf{64}\boldsymbol{d}}}{\delta}\right)\right) = \boldsymbol{O}\left(\frac{\boldsymbol{d}}{\boldsymbol{\gamma}^{\mathbf{2}}} \log\left(\frac{\mathbf{1}}{\delta}\right)\right)$$

Therefore, the **number** of **training examples needed** is **at most linear** in the **parameters** of the **model** $\boldsymbol{d}$.

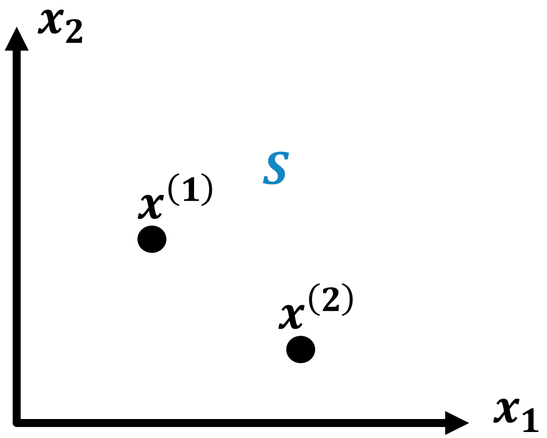
LEARNING THEORY

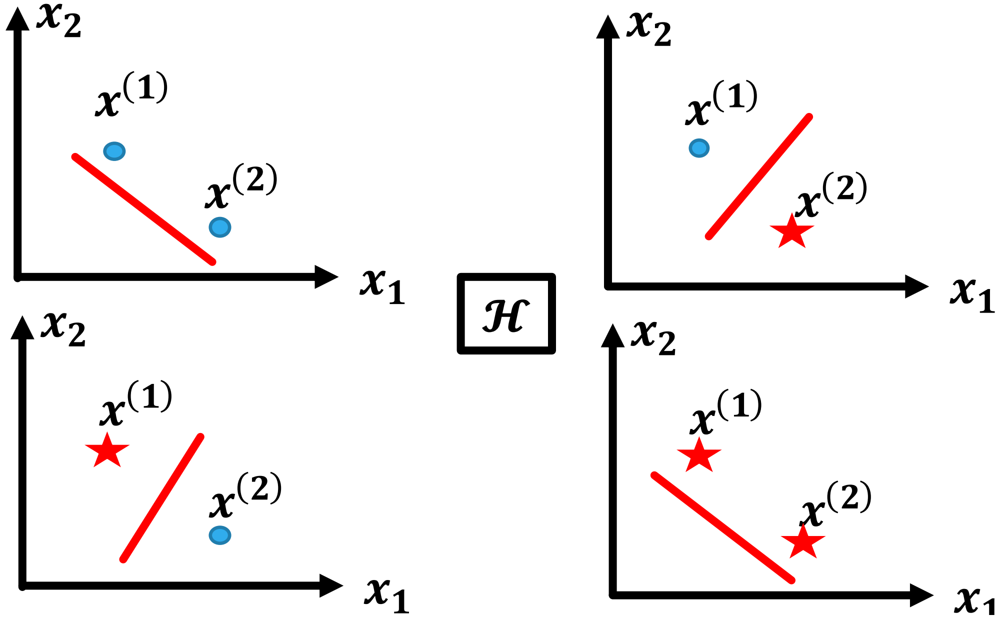
THE CASE OF INFINITE $\mathcal{H}$

We will **introduce** a **definition** to **make** the **proof** for **infinite classes** $\mathcal{H}$**.**

DEFINITION: given a set $S = \{x^{(1)}, \ldots, x^{(D)}\}$ of **points** $x^{(i)} \in \chi$**,** we say that $\mathcal{H}$ **SHATTERS** S if $\mathcal{H}$ can realize any labeling on S. That is, if **for any set** of **labels** $\{y^{(1)} \ldots, y^{(D)}\}$**,** there **exists some** $h \in \mathcal{H}$ so that $h(x^{(i)}) = y^{(i)}$ **for all** $i = 1, \ldots D$.

EXAMPLE:





SHATTER: for any possible labeling of these points, we can find a linear classifier that obtains "zero training error" on them.

LEARNING THEORY

THE CASE OF INFINITE $\mathcal{H}$

DEFINITION:

The **Vapnik-Chervonenkis dimension** of $\mathcal{H}$, $\left(VC(\mathcal{H})\right)$ is the **size** of the **largest set** shattered by $\mathcal{H}$.

EXAMPLE:

If $\mathcal{H} = \{\textbf{linear classifiers in 2D}\}$**,** therefore $VC(\mathcal{H}) = \mathbf{3}$**.** There is **no set of size 4** that it could **shatter**.

In a **general form** we have that if $\mathcal{H} = \{\textbf{linear classifiers in n Dimensions}\}$**,** therefore $VC(\mathcal{H}) = n + \mathbf{1}$.

LEARNING THEORY

THE CASE OF INFINITE $\mathcal{H}$

THEOREM:

Let $\boldsymbol{\mathcal{H}}$ be given and let $\boldsymbol{D} = \boldsymbol{VC}(\boldsymbol{\mathcal{H}})$. Then with **probability** at **least** $\mathbf{1} - \boldsymbol{\delta}$, we have that for all $\boldsymbol{h} \in \boldsymbol{\mathcal{H}}$,

$$\left|\mathcal{E}(\boldsymbol{h}) - \hat{\mathcal{E}}(\boldsymbol{h})\right| \leq \boldsymbol{O}\left(\sqrt{\frac{\boldsymbol{D}}{\boldsymbol{m}}\log\left(\frac{\boldsymbol{m}}{\boldsymbol{D}}\right) + \frac{\mathbf{1}}{\boldsymbol{m}}\log\left(\frac{1}{\delta}\right)}\right)$$

With probability at least $1 - \delta$, we also have that:

$$\hat{\mathcal{E}}(\boldsymbol{h}) \leq \mathcal{E}(\boldsymbol{h}^*) + \boldsymbol{O}\left(\sqrt{\frac{\boldsymbol{D}}{\boldsymbol{m}}\log\left(\frac{\boldsymbol{m}}{\boldsymbol{D}}\right) + \frac{\mathbf{1}}{\boldsymbol{m}}\log\left(\frac{1}{\delta}\right)}\right)$$

If a hypothesis class has finite VC dimension, then uniform convergence occurs as m becomes large.

LEARNING THEORY

THE CASE OF INFINITE $\mathcal{H}$

COROLLARY:

For $|\mathcal{E}(\boldsymbol{h}) - \hat{\mathcal{E}}(\boldsymbol{h}) \leq)| \leq \boldsymbol{\gamma}$ to hold for all $\boldsymbol{h} \in \boldsymbol{\mathcal{H}}$ (and hence $\mathcal{E}(\hat{\boldsymbol{h}}) \leq \mathcal{E}(\boldsymbol{h}^*) + \boldsymbol{2\gamma}$) with probability at least $\boldsymbol{1} - \boldsymbol{\delta}$, it suffices that $\boldsymbol{n} = \boldsymbol{O}_{\gamma,\delta}(\boldsymbol{D})$.

Thus, **sample complexity** is **upper-bounded** by the **VC dimension**. Also, for "**most**" **hypothesis** classes, the **VC dimension** is also **roughly linear** in the **number** of **parameters**.

We **conclude** that for a **given hypothesis class** $\boldsymbol{\mathcal{H}}$, the **number** of **training examples needed** to **achieve generalization error close** to that of the **optimal classifier** is usually **roughly linear** in the **number** of **parameters** of $\boldsymbol{\mathcal{H}}$.