THE FUNDAMENTAL THEOREM OF STATISTICAL LEARNING

LUOQIAN (OCEAN) WANG

Abstract. This paper explores the theoretical foundations of binary classification within the framework of Probably Approximately Correct (PAC) learning. We present the equivalence between PAC learnability, the uniform convergence property, and the finiteness of Vapnik–Chervonenkis (VC) dimension, and therefore provide a self-contained proof of the Fundamental Theorem of Statistical Learning.

Contents

1. Empirical Risk Minimization

In a binary classification problem with respect to 0–1 loss, a learning algorithm

$$A\colon \bigcup_{m=1}^{\infty} (\mathcal{X}\times\mathcal{Y})^m \to \mathcal{Y}^{\mathcal{X}}$$

receives a training sample S with m instances, drawn from some unknown distribution $\mathcal{D}$ and labeled by some unknown target function f, where $\mathcal{Y}=\{0,1\}$. Then the algorithm A outputs a hypothesis $h_s\colon \mathcal{X}\to\mathcal{Y}$, denoted by $A(S)$, the performance of which is measured by generalization error $L_{\mathcal{D},f}(h_S)$.

Definition 1.1 (Generalization Error)**.** The generalization error (or true error or true risk) of a hypothesis $h\colon \mathcal{X}\to\mathcal{Y}$ with respect to a distribution $\mathcal{D}$ and a target function f is defined as

$$L_{\mathcal{D},f}(h) = \mathbb{E}_{x\sim\mathcal{D}}[\ell(h(x),f(x))] = \mathbb{P}_{x\sim\mathcal{D}}[h(x)\neq f(x)] = \mathcal{D}(\{x\colon h(x)\neq f(x)\}).$$

Since the distribution $\mathcal{D}$ and the target function f are both unknown, a learning algorithm cannot directly compute the generalization error. Instead, given a training sample S, the algorithm can evaluate the empirical error.

Date: September 2025.

Definition 1.2 (Empirical Error)**.** The empirical error (or empirical risk) of a hypothesis h over the training sample $S = \{(x_i, y_i)\}_{i=1}^m$ is defined as

$$L_S(h) = \frac{1}{m}\sum_{i=1}^{m} \ell(h(x_i), y_i) = \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}\{h(x_i) \neq y_i\}.$$

The Empirical Risk Minimization (ERM) rule aims to minimize the empirical error over a given sample. Any learning algorithm that follows the ERM rule is called an $\mathrm{ERM}_{\mathcal{H}}$ learner. Given a predefined hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$, the $\mathrm{ERM}_{\mathcal{H}}$ learner selects its output hypothesis from $\mathcal{H}$ by minimizing the empirical risk. The output hypothesis, if it exists, is called an empirical risk minimizer, denoted by $A(S) = \mathrm{ERM}_{\mathcal{H}}(S) \in \arg\min_{h\in\mathcal{H}} L_S(h)$. Intuitively, restricting the search space of hypotheses from $\mathcal{Y}^{\mathcal{X}}$ to a hypothesis class $\mathcal{H}$ introduces an inductive bias (prior knowledge) and helps control overfitting. Without such a restriction, a hypothesis could perfectly fit the training sample S but fail to generalize to unseen data from the underlying distribution $\mathcal{D}$, leading to a large generalization error.

Our current task is to examine this intuition rigorously: by appropriately restricting the hypothesis class, we can ensure that, with high probability, the $\mathrm{ERM}_{\mathcal{H}}$ learner selects a hypothesis with a small generalization error, provided the training sample is sufficiently large. A natural way to impose such a restriction is to upper bound the size of the hypothesis class, which motivates us to begin the analysis with finite hypothesis classes.

Assumption 1.3 (Realizability Assumption)**.** There exists a hypothesis $h \in \mathcal{H}$ such that the generalization error $L_{\mathcal{D},f}(h) = 0$. As a result, the empirical error $L_S(h) = 0$.

Assumption 1.4 (i.i.d. Assumption)**.** The instances in the training sample S are drawn independently and identically from some distribution $\mathcal{D}$ over $\mathcal{X}$ and labeled by some target function f.

Problem 1.5. *Let $\mathcal{X}$ be any set, $\mathcal{Y} = \{0, 1\}$, and $\mathcal{H}$ be a finite set of functions from $\mathcal{X}$ to $\mathcal{Y}$. If Assumption 1.3 and Assumption 1.4 hold, then the* $\mathrm{ERM}_{\mathcal{H}}$ *learner is guaranteed with a high probability to output a hypothesis h_S with a small* generalization error $L_{\mathcal{D},f}(h_S)$*, given a sufficiently large size m of the training sample S.*

Proof. We aim to upper bound the probability that the $\mathrm{ERM}_{\mathcal{H}}$ learner fails, i.e., returns a hypothesis with high generalization error, when there are m elements in the training sample S. Let $S|_x = (x_1, \ldots, x_m)$ be the sequence of inputs in S. Let $h_S = \mathrm{ERM}_{\mathcal{H}}(S)$ denotes the hypothesis selected by the $\mathrm{ERM}_{\mathcal{H}}$ learner over S. If $L_{\mathcal{D},f}(h_S) > \epsilon$, then the $\mathrm{ERM}_{\mathcal{H}}$ learner fails. That is, we aim to upper bound

$$\mathbb{P}_{S\sim\mathcal{D}^m}[L_{\mathcal{D},f}(h_S) > \epsilon] = \mathcal{D}^m(\{S|_x \colon L_{\mathcal{D},f}(h_S) > \epsilon\}).$$

Let $\mathcal{H}_B = \{h \in \mathcal{H} \colon L_{\mathcal{D},f}(h) > \epsilon\}$ be the set of bad hypotheses. Let $M = \{S|_x \colon \exists h \in \mathcal{H}_B \text{ such that } L_S(h) = 0\}$ be the set of inputs from the misleading samples. If the $\mathrm{ERM}_{\mathcal{H}}$ learner fails (i.e., $L_{\mathcal{D},f}(h_S) > \epsilon$), then $h_S \in \mathcal{H}_B$. By Assumption 1.3, there exists some $h^* \in \mathcal{H}$ with $L_{\mathcal{D},f}(h^*) = 0$, so $L_S(h^*) = 0$. Since h_S is selected to minimize empirical risk, we have $L_S(h_S) \leq L_S(h^*) = 0$. Hence, any sample S for which the $\mathrm{ERM}_{\mathcal{H}}$ learner fails must be a misleading sample. In other words, we have $\{S|_x \colon L_{\mathcal{D},f}(h_S) > \epsilon\} \subseteq M$. It suffices to upper bound

$$\mathcal{D}^m(M) = \mathcal{D}^m\left(\bigcup_{h\in\mathcal{H}_B} \{S|_x \colon L_S(h) = 0\}\right) \geq \mathcal{D}^m(\{S|_x \colon L_{\mathcal{D},f}(h_S) > \epsilon\}).$$

Fix $h \in \mathcal{H}_B$. Then we have $\mathcal{D}^m(\{S|_x \colon L_S(h) = 0\}) = \mathcal{D}^m(\{S|_x \colon \forall i, h(x_i) = y_i\}) = \prod_{i=1}^m \mathcal{D}(\{x_i \colon h(x_i) = y_i\}) = (1 - L_{\mathcal{D},f}(h))^m \leq (1-\epsilon)^m$, where the second equality follows from Assumption 1.4, and the last inequality uses the definition of bad hypotheses.

Apply Lemma 3 in Appendix over all $h \in \mathcal{H}_B$. Then we have $\mathcal{D}^m(M) \leq \sum_{h\in\mathcal{H}_B} \mathcal{D}^m(\{S|_x \colon L_S(h) = 0\}) \leq |\mathcal{H}_B| \cdot (1-\epsilon)^m \leq |\mathcal{H}_B| \cdot e^{-\epsilon m} \leq |\mathcal{H}| \cdot e^{-\epsilon m}$, where the second last inequality follows from the bound $1 - \epsilon \leq e^{-\epsilon}$ for all $\epsilon \in [0,1]$.

Therefore, we have $\mathcal{D}^m(\{S|_x \colon L_{\mathcal{D},f}(\mathrm{ERM}_{\mathcal{H}}(S)) > \epsilon\}) \leq |\mathcal{H}| \cdot e^{-\epsilon m}$. To guarantee that the probability of the failure is at most δ, set $|\mathcal{H}| \cdot e^{-\epsilon m} \leq \delta$. Then we have $m \geq \frac{1}{\epsilon} \ln\left(\frac{|\mathcal{H}|}{\delta}\right)$. For any $\epsilon, \delta \in (0,1)$, if the training sample size satisfies $m \geq \frac{1}{\epsilon} \ln\left(\frac{|\mathcal{H}|}{\delta}\right)$, then the $\mathrm{ERM}_{\mathcal{H}}$ learner over a finite hypothesis class will be probably (at least $1-\delta$) approximately correct (up to an error of ϵ). □

2. PAC Learning Framework

Now we can formally present Probably Approximately Correct (PAC) learning framework, first introduced by Valiant [1].

Definition 2.1 (PAC Learnability)**.** A hypothesis class $\mathcal{H}$ is PAC learnable if there exist a learning algorithm A and a function $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ such that the following holds: For every $\epsilon, \delta \in (0,1)$, every probability distribution $\mathcal{D}$ over $\mathcal{X}$, and every target function $f : \mathcal{X} \to \{0,1\}$, if Assumption 1.3 holds, then for any training sample S of size $m \geq m_{\mathcal{H}}(\epsilon,\delta)$ drawn i.i.d. from $\mathcal{D}$ and labeled by f, the algorithm outputs a hypothesis $h = A(S)$ satisfying $\mathbb{P}_{S\sim\mathcal{D}^m}[L_{\mathcal{D},f}(h) > \epsilon] \leq \delta$.

Here, the accuracy parameter ϵ determines how far the output hypothesis can be from the optimal one. The confidence parameter δ indicates how likely the output hypothesis is to meet the accuracy requirement. The sample complexity of the hypothesis class $m_{\mathcal{H}}(\epsilon,\delta)$ is the minimal number of instances in a sample required to guarantee a probably approximately correct solution. The result in Problem 1.5 shows that $m \geq \frac{1}{\epsilon} \ln\left(\frac{|\mathcal{H}|}{\delta}\right)$ is the sufficient condition for finite hypothesis class to be PAC learnable. Hence we have the following upper bound of the sample complexity $m_{\mathcal{H}}(\epsilon,\delta)$ for some finite hypothesis class $\mathcal{H}$.

Corollary 2.2. *Every finite hypothesis class is PAC learnable with sample complexity* $m_{\mathcal{H}}(\epsilon,\delta) \leq \left\lceil \frac{1}{\epsilon} \ln\left(\frac{|\mathcal{H}|}{\delta}\right) \right\rceil$.

The PAC learning framework implicitly assumes a deterministic, noise-free data generation process: instances are drawn according to a distribution $\mathcal{D}$ over $\mathcal{X}$, and their labels are generated by some fixed target function $f \colon \mathcal{X} \to \mathcal{Y}$, so that for every $x \in \mathcal{X}$, the label is uniquely determined by $f(x)$. However, in practical scenarios, the labeling process may be noisy or stochastic, and therefore the observed labels need not be consistent with any single deterministic target function. Furthermore, under Assumption 1.3, the PAC learning framework provides a distribution-free guarantee on the number of needed instances in the training sample. However, such an assumption is too strong for practical applications where the true target function may not be contained in $\mathcal{H}$.

These considerations motivate the relaxation of the noise-free assumption and the realizability assumption: First, let $\mathcal{D}$ be a joint probability distribution over $\mathcal{X} \times \mathcal{Y}$. The distribution $\mathcal{D}$ can be decomposed into two components: a marginal distribution $\mathcal{D}_{\mathcal{X}}$ over $\mathcal{X}$, and a conditional distribution $\mathcal{D}((x, y)|x)$ over $\mathcal{Y}$ given each $x \in \mathcal{X}$. Accordingly, the generalization error of a hypothesis $h \colon \mathcal{X} \to \mathcal{Y}$ with respect to a joint distribution $\mathcal{D}$ is now defined as $L_{\mathcal{D}}(h) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(h(x), y)] = \mathbb{P}_{(x,y)\sim\mathcal{D}}[h(x) \neq y] = \mathcal{D}(\{(x, y) \colon h(x) \neq y\})$. Now, we can present agnostic PAC learning framework.

Definition 2.3 (Agnostic PAC Learnability)**.** A hypothesis class $\mathcal{H}$ is agnostically PAC learnable if there exist a learning algorithm A and a function $m_{\mathcal{H}} : (0, 1)^2 \to \mathbb{N}$ such that the following holds: For every $\epsilon, \delta \in (0, 1)$, and for every joint probability distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, for any training sample S of size $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ drawn i.i.d. from $\mathcal{D}$, the algorithm outputs a hypothesis $h = A(S)$ satisfying $\mathbb{P}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}(h) > \inf_{h'\in\mathcal{H}} L_{\mathcal{D}}(h') + \epsilon\right] \leq \delta$.

3. Uniform Convergence Property

So far we have established that any finite hypothesis class is PAC learnable in the realizable setting. We now extend this result to the agnostic PAC learning framework. Recall that the $\mathrm{ERM}_{\mathcal{H}}$ learner outputs a hypothesis h in $\mathcal{H}$ that minimizes $L_S(h)$ and gives small $L_{\mathcal{D}}(h)$ with high probability. It suffices to require that $L_S(h)$ is a reliable estimate of $L_{\mathcal{D}}(h)$ for every hypothesis in the class. This idea is formalized by the following definition:

Definition 3.1 (ϵ-Representative Sample)**.** A training sample S is ϵ-representative with respect to a domain $\mathcal{X} \times \mathcal{Y}$, a hypothesis class $\mathcal{H}$, a loss function ℓ, and a distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, if for every $h \in \mathcal{H}$, it holds that $|L_S(h) - L_{\mathcal{D}}(h)| \leq \epsilon$.

The following lemma will link the property of the sample to the performance of the $\mathrm{ERM}_{\mathcal{H}}$ learner.

Lemma 3.2. *Assume that a training sample S is $(\frac{\epsilon}{2})$-representative with respect to $\mathcal{X} \times \mathcal{Y}$, $\mathcal{H}$, ℓ, and $\mathcal{D}$. Then any empirical risk minimizer $h_S \in \arg\min_{h\in\mathcal{H}} L_S(h)$, if it exists, satisfies $L_{\mathcal{D}}(h_S) \leq \inf_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + \epsilon$.*

Proof. For any $h \in \mathcal{H}$, $L_{\mathcal{D}}(h_S) \leq L_S(h_S) + \frac{\epsilon}{2} \leq L_S(h) + \frac{\epsilon}{2} \leq L_{\mathcal{D}}(h) + \frac{\epsilon}{2} + \frac{\epsilon}{2} = L_{\mathcal{D}}(h) + \epsilon$. The first and the third inequality hold since S is $(\epsilon/2)$-representative. The second inequality holds since $h_S \in \arg\min_{h\in\mathcal{H}} L_S(h)$. □

Lemma 3.2 implies that to ensure the $\mathrm{ERM}_{\mathcal{H}}$ learner is a successful agnostic PAC learner, it suffices to guarantee that the training sample is $(\frac{\epsilon}{2})$-representative. This idea is formalized by the uniform convergence property of the hypothesis class.

Definition 3.3 (Uniform Convergence Property)**.** A hypothesis class $\mathcal{H}$ has the uniform convergence property with respect to $\mathcal{X} \times \mathcal{Y}$ and ℓ if there exists a function $m^{\mathrm{UC}}_{\mathcal{H}} : (0, 1)^2 \to \mathbb{N}$ such that for every $\epsilon, \delta \in (0, 1)$ and for every probability distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, if a sample S is drawn i.i.d. from $\mathcal{D}$ with size $m \geq m^{\mathrm{UC}}_{\mathcal{H}}(\epsilon, \delta)$, then with probability of at least $1 - \delta$, S is ϵ-representative.

The sample complexity of uniform convergence $m_{\mathcal{H}}^{\mathrm{UC}}$ is the minimal number of instances in the sample required to guarantee the uniform convergence property. In other words, $m_{\mathcal{H}}^{\mathrm{UC}}$ works for all hypotheses in $\mathcal{H}$, and the uniform convergence property is a sufficient but not necessarily a minimal condition for learnability. A direct corollary of Lemma 3.2 and Definition 3.3 is the following:

Corollary 3.4. *If a hypothesis class $\mathcal{H}$ has the uniform convergence property with sample complexity $m_{\mathcal{H}}^{UC}$, then $\mathcal{H}$ is agnostically PAC learnable with sample complexity $m_{\mathcal{H}}(\epsilon, \delta) \leq m_{\mathcal{H}}^{UC}(\epsilon/2, \delta)$. Furthermore, the* $\mathrm{ERM}_{\mathcal{H}}$ *learner is a successful agnostic PAC learner for $\mathcal{H}$.*

Previously, we conclude that every finite hypothesis class is PAC learnable in Corollary 2.2, yet what remains to be proven is that every finite hypothesis class is agnostically PAC learnable.

Theorem 3.5. *Every finite hypothesis class $\mathcal{H}$ is agnostically PAC learnable.*

Proof. By Corollary 3.4, it suffices to show that the uniform convergence property holds for $\mathcal{H}$. Formally, we need to show that for any distribution $\mathcal{D}$ and for any $\epsilon, \delta \in (0,1)$, there exists a sample size m such that when a sample $S = (z_1, \ldots, z_m)$ is drawn i.i.d. from $\mathcal{D}^m$, we have $\mathcal{D}^m(\{S\colon \forall h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| \leq \epsilon\}) \geq 1 - \delta$. Equivalently, we want to show that the following holds:

$$\mathcal{D}^m(\{S\colon \exists h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}) \leq \delta.$$

Since $\{S\colon \exists h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\} = \bigcup_{h \in \mathcal{H}} \{S\colon |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}$, by Lemma 3 in Appendix, we have

$$\mathcal{D}^m(\bigcup_{h \in \mathcal{H}} \{S\colon |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}) \leq \sum_{h \in \mathcal{H}} \mathcal{D}^m(\{S\colon |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}).$$

Fix a hypothesis $h \in \mathcal{H}$. The empirical risk $L_S(h) = \frac{1}{m}\sum_{i=1}^m \ell(h(x_i), y_i)$ is the average of m i.i.d. random variables $\theta_i = \ell(h(x_i), y_i)$, each with expected value $\mathbb{E}_{(x_i,y_i)\sim\mathcal{D}}[\theta_i] = \mathbb{E}_{(x_i,y_i)\sim\mathcal{D}}\big[\ell(h(x_i), y_i)\big] = L_{\mathcal{D}}(h)$. Assume that the range of ℓ is $[0,1]$ and therefore $\theta_i \in [0,1]$. Apply Lemma 7 in Appendix.[1] Then we have $\mathcal{D}^m(\{S\colon |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}) \leq 2\exp\left(-2m\epsilon^2\right)$. Finally, we have

$$\mathcal{D}^m(\{S : \exists h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}) \leq \sum_{h \in \mathcal{H}} 2\exp\left(-2m\epsilon^2\right) = 2|\mathcal{H}|\exp\left(-2m\epsilon^2\right).$$

Set $2|\mathcal{H}|\exp\left(-2m\epsilon^2\right) \leq \delta$. Then we have

$$m \geq \frac{1}{2\epsilon^2}\ln\left(\frac{2|\mathcal{H}|}{\delta}\right).$$

Therefore, for any m that satisfies the above condition, with probability at least $1 - \delta$ over the choice of an i.i.d. sample S of size m, we have that for all $h \in \mathcal{H}$, $|L_S(h) - L_{\mathcal{D}}(h)| \leq \epsilon$. □

[1] Note that, by the linearity of expectation, we have

$$\mathbb{E}_{S\sim\mathcal{D}^m}[L_S(h)] = \mathbb{E}_{S\sim\mathcal{D}^m}\left[\frac{1}{m}\sum_{i=1}^m \theta_i\right] = \frac{1}{m}\sum_{i=1}^m \mathbb{E}_{(x_i,y_i)\sim\mathcal{D}}[\theta_i] = \frac{1}{m}\sum_{i=1}^m L_{\mathcal{D}}(h) = L_{\mathcal{D}}(h).$$

This shows that the empirical risk is an unbiased estimator of the true risk, which motivates the application of Hoeffding's inequality to bound $|L_S(h) - L_{\mathcal{D}}(h)|$.

Corollary 3.6. *Assume $\mathcal{H}$ is a finite hypothesis class. Let $\ell\colon \mathcal{Y} \times \mathcal{Y} \to [0,1]$ be a loss function. Then $\mathcal{H}$ enjoys the uniform convergence property with sample complexity*

$$m_{\mathcal{H}}^{UC}(\epsilon, \delta) \leq \left\lceil \frac{1}{2\epsilon^2} \ln\left(\frac{2|\mathcal{H}|}{\delta}\right) \right\rceil .$$

Furthermore, the class is agnostically PAC learnable using the $\mathrm{ERM}_{\mathcal{H}}$ *learner with sample complexity*

$$m_{\mathcal{H}}(\epsilon, \delta) \leq m_{\mathcal{H}}^{UC}(\epsilon/2, \delta) \leq \left\lceil \frac{2}{\epsilon^2} \ln\left(\frac{2|\mathcal{H}|}{\delta}\right) \right\rceil .$$

4. The No-Free-Lunch Theorem

In our analysis so far, we have restricted the search space of the target function to a predefined hypothesis class $\mathcal{H}$, which serves as prior knowledge. This naturally raises the question: is such prior knowledge truly necessary?

To formalize this question, we can consider the notion of a universal learner. Specifically, does there exist a single learning algorithm A that, without access to a specific hypothesis class $\mathcal{H}$, can successfully learn any task that is learnable by some other algorithm equipped with the corresponding prior knowledge? The No-Free-Lunch Theorem shows that, for any learning algorithm A, there always exist learning tasks on which A fails to succeed.

Theorem 4.1 (The No-Free-Lunch Theorem)**.** *Let A be any learning algorithm for the binary classification task with respect to 0–1 loss over $\mathcal{X}$. Let m be the size of a training sample S smaller than $|\mathcal{X}|/2$. Then there exists a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$ such that:*

- **(i)** *There exists a function $f\colon \mathcal{X} \to \{0,1\}$ with $L_{\mathcal{D}}(f) = 0$.*
- **(ii)** *With a probability of at least $1/7$ over the choice of $S \sim \mathcal{D}^m$, we have: $L_{\mathcal{D}}(A(S)) \geq 1/8$.*

Proof. Let C be a subset of $\mathcal{X}$ with the size $2m$. Let $\mathcal{H}$ denote the set of all possible functions from C to $\{0,1\}$. Let T be the size of $\mathcal{H}$, where $\mathcal{H} = \{f_1, \cdots, f_T\}$. Therefore, $T = 2^{2m}$. For each $f_i \in \mathcal{H}$, define a distribution $\mathcal{D}_i$ over $\mathcal{C} \times \{0,1\}$ as the following:

$$\mathcal{D}_i(\{(x,y)\}) = \begin{cases} 1/2m & \text{if } x \in C \text{ and } y = f_i(x) \\ 0 & \text{otherwise} \end{cases} .$$

Therefore, for the function f_i, we have $L_{\mathcal{D}_i}(f_i) = 0$ on distribution $\mathcal{D}_i$.
Next, we aim to prove that for any learning algorithm A, we have $\mathcal{D}^m(S\colon L_{\mathcal{D}}(A(S)) \geq 1/8) \geq 1/7$. It suffices to show that $\mathbb{E}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S))] \geq \frac{1}{4}$.[2] Therefore, we can show that

$$\max_{i\in[T]} \mathbb{E}_{S\sim\mathcal{D}_i^m}[L_{\mathcal{D}_i}(A(S))] \geq \frac{1}{4}.$$

First, sample m instances from C with replacement. There are $k = (2m)^m$ possible sequences of m instances from C, denoted as $S_1, S_2, \cdots, S_k$. If we denote S_j as

[2] By Lemma 5 in Appendix, we have

$$\mathcal{D}^m(S\colon L_{\mathcal{D}}(A(S)) \geq 1/8) = \frac{\mathbb{E}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S))] - 1/8}{1 - 1/8} \geq \frac{1}{7}.$$

$(x_1, \cdots, x_m)$, then $S_j^i = ((x_1, f_i(x_1)), \cdots, (x_m, f_i(x_m)))$ represents the sequence S_j with the labeling function f_i. Then we have

$$\begin{aligned}\max_{i\in[T]} \mathbb{E}_{S\sim\mathcal{D}_i^m}[L_{\mathcal{D}_i}(A(S))] = \max_{i\in[T]} \frac{1}{k}\sum_{j=1}^{k} L_{\mathcal{D}_i}(A(S_j^i)) &\geq \frac{1}{T}\sum_{i=1}^{T}\frac{1}{k}\sum_{j=1}^{k} L_{\mathcal{D}_i}(A(S_j^i)) \\ &= \frac{1}{k}\sum_{j=1}^{k}\frac{1}{T}\sum_{i=1}^{T} L_{\mathcal{D}_i}(A(S_j^i)) \\ &\geq \min_{j\in[k]} \frac{1}{T}\sum_{i=1}^{T} L_{\mathcal{D}_i}(A(S_j^i)).\end{aligned}$$[3]

Let $j^* \in \arg\min_{j\in[k]} \frac{1}{T}\sum_{i=1}^{T} L_{\mathcal{D}_i}(A(S_j^i))$. Let $S_{j^*} = (x_1, \cdots, x_m)$. Let $\{v_1, \cdots, v_p\}$ be the set of unsampled instances from C. For every function $h\colon C \to \{0,1\}$ and $i \in [T]$, we have

$$\begin{aligned}L_{\mathcal{D}_i}(h) &= \frac{1}{2m}\sum_{x\in C} \mathbf{1}\{h(x) \neq f_i(x)\} \\ &\geq \frac{1}{2m}\sum_{r=1}^{p} \mathbf{1}\{h(v_r) \neq f_i(v_r)\} \\ &\geq \frac{1}{2p}\sum_{r=1}^{p} \mathbf{1}\{h(v_r) \neq f_i(v_r)\}.\end{aligned}$$

Therefore, we have:

$$\begin{aligned}\frac{1}{T}\sum_{i=1}^{T} L_{\mathcal{D}_i}(A(S_{j^*}^i)) &\geq \frac{1}{T}\sum_{i=1}^{T}\frac{1}{2p}\sum_{r=1}^{p} \mathbf{1}\{A(S_{j^*}^i)(v_r) \neq f_i(v_r)\} \\ &= \frac{1}{2p}\sum_{r=1}^{p}\frac{1}{T}\sum_{i=1}^{T} \mathbf{1}\{A(S_{j^*}^i)(v_r) \neq f_i(v_r)\} \\ &\geq \frac{1}{2}\min_{r\in[p]}\frac{1}{T}\sum_{i=1}^{T} \mathbf{1}\{A(S_{j^*}^i)(v_r) \neq f_i(v_r)\}\end{aligned}$$

Let $r^* \in \arg\min_{r\in[p]} \frac{1}{T}\sum_{i=1}^{T} \mathbf{1}\{A(S_{j^*}^i)(v_r) \neq f_i(v_r)\}$. Since v_{r^*} is not included in the training sequence S_{j^*}, the algorithm A has no information about its true label. Because $\{f_1, \ldots, f_T\}$ enumerates all possible functions $C \to \{0,1\}$, exactly half of the functions assign v_{r^*} the label 0 and half assign it the label 1. In other words, there always exist pairs $a, b \in [T]$ such that $\mathbf{1}\{A(S_{j^*}^i)(v_{r^*}) \neq f_a(v_{r^*})\} + \mathbf{1}\{A(S_{j^*}^i)(v_{r^*}) \neq f_b(v_{r^*})\} = 1$. Therefore, we have

$$\frac{1}{T}\sum_{i=1}^{T} \mathbf{1}\{A(S_{j^*}^i)(v_{r^*}) \neq f_i(v_{r^*})\} = \frac{1}{2}.$$

[3] Let I be a finite set and let $\{a_i\}_{i\in I} \subset \mathbb{R}$. Then we have

$$\min_{i\in I} a_i \;\leq\; \frac{1}{|I|}\sum_{i\in I} a_i \;\leq\; \max_{i\in I} a_i.$$

Finally, we have:

$$\begin{aligned}\max_{i\in[T]} \mathbb{E}_{S\sim\mathcal{D}_i^m}[L_{\mathcal{D}_i}(A(S))] &\geq \min_{j\in[k]} \frac{1}{T}\sum_{i=1}^{T} L_{\mathcal{D}_i}(A(S_j^i)) \\ &= \frac{1}{T}\sum_{i=1}^{T} L_{\mathcal{D}_i}(A(S_{j^*}^i)) \\ &\geq \frac{1}{2}\min_{r\in[p]} \frac{1}{T}\sum_{i=1}^{T} \mathbf{1}\{A(S_{j^*}^i)(v_r) \neq f_i(v_r)\} \\ &= \frac{1}{2}\cdot\frac{1}{T}\sum_{i=1}^{T} \mathbf{1}\{A(S_{j^*}^i)(v_{r^*}) \neq f_i(v_{r^*})\} \\ &= \frac{1}{2}\cdot\frac{1}{2} = \frac{1}{4}.\end{aligned}$$

□

Corollary 4.2. *Let $\mathcal{H}$ be the class of all functions from $\mathcal{X}$ to $\{0,1\}$. Then, $\mathcal{H}$ is not PAC learnable.*

Proof. Let us prove by contradiction. Assume $\mathcal{H}$ is PAC learnable. Then, by Definition 2.1, there exists a learning algorithm A and a function $m_{\mathcal{H}}$ such that for every $\epsilon \in (0, 1/8)$, $\delta \in (0, 1/7)$, for every distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$, and for some $f \in \mathcal{H}$, $L_{\mathcal{D}}(f) = 0$, then if running the algorithm on S with $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ i.i.d. instances generated by $\mathcal{D}$, we have $\mathbb{P}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S)) > \epsilon] \leq \delta$. Given that $\mathcal{X}$ is infinite, we have $|\mathcal{X}| > 2m$ for any finite m. Applying Theorem 4.1, we conclude that for for any learning algorithm, including A, we have $\mathbb{P}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S)) \geq 1/8] \geq 1/7$. Therefore, with probability at least $1/7 > \delta$, we have $L_{\mathcal{D}}(A(S)) \geq 1/8 > \epsilon$. This contradicts the assumed PAC learnability of $\mathcal{H}$. □

5. VC Dimension

Previously, we showed that all finite hypothesis classes are PAC learnable. On the other hand, we showed a case that the class of all functions over an infinite domain, which is an infinite class, is not. A further question we shall discuss is what distinguishes the learnable classes from the unlearnable ones.

First, we will show that the finiteness of the hypothesis classes is not a necessary condition for learnability since some infinite hypothesis classes are PAC learnable.

Example 5.1 (Class of Threshold Functions)**.** Let $\mathcal{H} = \{h_a \colon a \in \mathbb{R}\}$ be the set of threshold functions over $\mathbb{R}$, where $h_a \colon \mathbb{R} \to \{0,1\}$ is a function such that $h_a(x) = \mathbf{1}\{x < a\}$. Then $\mathcal{H}$ is PAC learnable with sample complexity of $m_{\mathcal{H}}(\epsilon, \delta) \leq \lceil \frac{1}{\epsilon} \ln\left(\frac{2}{\delta}\right) \rceil$.

Proof. Let a^* be the optimal threshold such that $h^*(x) = \mathbf{1}\{x < a^*\}$ satisfies $L_{\mathcal{D}}(h^*) = 0$. Let $\mathcal{D}_x$ be the marginal distribution over $\mathcal{X}$. Let a_0 and a_1 satisfy $a_0 < a^* < a_1$ and $\mathbb{P}_{x\sim\mathcal{D}_x}[x \in (a_0, a^*)] = \mathbb{P}_{x\sim\mathcal{D}_x}[x \in (a^*, a_1)] = \epsilon$. This guarantees that if the output hypothesis is chosen from (a_0, a_1), then it will have an error at most ϵ.

Given a training sample $S \sim \mathcal{D}^m$, let $b_0 = \max\{x \colon (x, 1) \in S\}$ (set $b_0 = -\infty$ if there is no positive instance) and let $b_1 = \min\{x \colon (x, 0) \in S\}$ (set $b_1 = \infty$ if there is no negative instance). Let $b_S \in (b_0, b_1)$ be the threshold corresponding to h_S, the output hypothesis by the $\text{ERM}_{\mathcal{H}}$ learner.

A sufficient condition for $L_{\mathcal{D}}(h_S) \leq \epsilon$ is that $b_0 \geq a_0$ and $b_1 \leq a_1$. Then we have $\mathbb{P}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(h_S) \leq \epsilon] \geq \mathbb{P}_{S\sim\mathcal{D}^m}[b_0 \geq a_0 \wedge b_1 \leq a_1]$. Equivalently, we have $\mathbb{P}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(h_S) > \epsilon] \leq \mathbb{P}_{S\sim\mathcal{D}^m}[b_0 < a_0 \vee b_1 > a_1]$. By Lemma 3 in Appendix, we have $\mathbb{P}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(h_S) > \epsilon] \leq \mathbb{P}_{S\sim\mathcal{D}^m}[b_0 < a_0] + \mathbb{P}_{S\sim\mathcal{D}^m}[b_1 > a_1]$. The event $b_0 < a_0$ means that all instances in S fall outside (a_0, a^*), thus leading to

$$\mathbb{P}_{S\sim\mathcal{D}^m}[b_0 < a_0] = \mathbb{P}_{S\sim\mathcal{D}^m}[\forall (x,y) \in S, x \notin (a_0, a^*)] = (1-\epsilon)^m \leq e^{-\epsilon m}.$$

For $m \geq \frac{1}{\epsilon}\ln\left(\frac{2}{\delta}\right)$, we have $\mathbb{P}_{S\sim\mathcal{D}^m}[b_0 < a_0] \leq \frac{\delta}{2}$. Similarly, we have $\mathbb{P}_{S\sim\mathcal{D}^m}[b_1 > a_1] \leq \frac{\delta}{2}$. Therefore, we have

$$\mathbb{P}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(h_S) > \epsilon] \leq \frac{\delta}{2} + \frac{\delta}{2} = \delta.$$

Then $\mathcal{H}$ is PAC learnable with sample complexity of $m_{\mathcal{H}}(\epsilon, \delta) \leq \lceil \frac{1}{\epsilon}\ln\left(\frac{2}{\delta}\right)\rceil$. □

To motivate the definition of the Vapnik–Chervonenkis (VC) dimension, which gives the correct characterization of learnability, recall how we proved Theorem 4.1: we constructed a distribution concentrated entirely on a finite subset $C \subset \mathcal{X}$, under which the learning algorithm fails. This occurs when the hypothesis class $\mathcal{H}$ is able to realize all possible labelings of the set C. This observation suggests that learnability does not depend on the total number of hypotheses in $\mathcal{H}$, but rather on how expressive $\mathcal{H}$ is when restricted to finite samples. If $\mathcal{H}$ is sufficiently rich to fit every possible labeling on some large finite set C, then it is vulnerable to such worst-case constructions. On the other hand, if the expressiveness of $\mathcal{H}$ is uniformly limited over all finite subsets of $\mathcal{X}$, then these situations can be ruled out. This intuition naturally leads to the following two definitions and a corollary related to Theorem 4.1.

Definition 5.2 (Restriction of $\mathcal{H}$ to C)**.** Let $\mathcal{H}$ be a hypothesis class of functions from $\mathcal{X}$ to $\{0,1\}$ and let C be a finite subset of $\mathcal{X}$. The restriction of $\mathcal{H}$ to C, denoted by $\mathcal{H}_C$, is the set of functions from C to $\{0,1\}$ that can be obtained from $\mathcal{H}$. Formally, we have $\mathcal{H}_C = \{f\colon C \to \{0,1\} \mid \exists h \in \mathcal{H} \text{ such that } f(c) = h(c) \text{ for all } c \in C\}$.

Definition 5.3 (Shattering)**.** A hypothesis class $\mathcal{H}$ shatters a finite subset $C \subset \mathcal{X}$ if the restriction of $\mathcal{H}$ to C is the set of all functions from C to $\{0,1\}$. In other words, we have $|\mathcal{H}_C| = 2^{|C|}$.

Corollary 5.4. *Let $\mathcal{H}$ be a hypothesis class of functions from $\mathcal{X}$ to $\{0,1\}$. Let m be the size of the training sample S. Assume that there exists a set $C \subset \mathcal{X}$ of size $2m$ that is shattered by $\mathcal{H}$. Then for any learning algorithm A, there exist a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$ and a hypothesis $h \in \mathcal{H}$ such that $L_{\mathcal{D}}(h) = 0$ but $\mathbb{P}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S)) \geq 1/8] \geq 1/7$.*

Now we can formally define VC dimension:

Definition 5.5 (VC Dimension)**.** The VC dimension of a hypothesis class $\mathcal{H}$, denoted by $\text{VCdim}(\mathcal{H})$, is the maximal size of a subset $C \subset \mathcal{X}$ that can be shattered by $\mathcal{H}$. If $\mathcal{H}$ can shatter sets of arbitrarily large size, then $\mathcal{H}$ has infinite VC dimension.

Example 5.6. Let $\mathcal{H}$ be a finite hypothesis class. For any subset $C \subset \mathcal{X}$ that is shattered by $\mathcal{H}$, we have $|\mathcal{H}| \geq |\mathcal{H}_C| = 2^{|C|}$, which implies $|C| \leq \log_2(|\mathcal{H}|)$.

Since the VC dimension is the maximal size of a shattered subset, it follows that $\mathrm{VCdim}(\mathcal{H}) \leq \log_2(|\mathcal{H}|)$.

From Example 5.6, we see that every finite hypothesis class has finite VC dimension. Recall from Corollary 2.2 that every finite hypothesis class is PAC learnable. Now we want to show that the finiteness of the VC dimension precisely characterizes learnability.

Theorem 5.7. *Let $\mathcal{H}$ be a class of infinite VC dimension. Then $\mathcal{H}$ is not PAC learnable. Equivalently, if $\mathcal{H}$ is PAC learnable, then it has finite VC dimension.*

Proof. Given that $\mathcal{H}$ has an infinite VC dimension, from Definition 5.5, for any training sample S with size m, there exists a shattered subset $C \subset \mathcal{X}$ of size $2m$. By Corollary 5.4, for any learning algorithm A, there exist a distribution $\mathcal{D}$ and a hypothesis $h \in \mathcal{H}$ with $L_{\mathcal{D}}(h) = 0$ such that $\mathbb{P}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S)) \geq 1/8] \geq 1/7$. Thus no algorithm can guarantee $L_{\mathcal{D}}(A(S)) \leq \epsilon$ with probability $1-\delta$ for $\epsilon < 1/8$ and $\delta < 1/7$. Hence $\mathcal{H}$ is not PAC learnable. □

To show that VC dimension characterizes learnability, we will show that the converse is also true within the following theorem.

Theorem 5.8 (The Fundamental Theorem of Statistical Learning)**.** *Let $\mathcal{H}$ be a hypothesis class of functions from a domain $\mathcal{X}$ to $\{0,1\}$ and let the loss function be the 0–1 loss. Then the following statements are equivalent.*

- **(i)** *$\mathcal{H}$ has the uniform convergence property.*
- **(ii)** *The $\mathrm{ERM}_{\mathcal{H}}$ learner is a successful agnostic PAC learner for $\mathcal{H}$.*
- **(iii)** *$\mathcal{H}$ is agnostic PAC learnable.*
- **(iv)** *$\mathcal{H}$ is PAC learnable.*
- **(v)** *The $\mathrm{ERM}_{\mathcal{H}}$ learner is a successful PAC learner for $\mathcal{H}$.*
- **(vi)** *$\mathcal{H}$ has a finite VC dimension.*

Proof. It suffices to show **(vi)** ⇒ **(i)**, given the following relations we have set up previously.

$$\begin{tikzcd}
& \textbf{(i)} \arrow[dr, "\text{Corollary 3.4}"] & \\
\textbf{(vi)} & & \textbf{(ii)} \arrow[dll, "\text{Definition 2.1}"', "\text{Definition 2.3}"] \arrow[d, "\text{Definition 2.3}"'] \\
\textbf{(v)} \arrow[u, "\text{Theorem 5.7}"] \arrow[dr, "\text{Definition 2.1}"] & & \textbf{(iii)} \arrow[dl, "\text{Definition 2.1}"', "\text{Definition 2.3}"] \\
& \textbf{(iv)} &
\end{tikzcd}$$

□

To complete the proof of Theorem 5.8, recall from Corollary 3.6 that finite hypothesis classes enjoy the uniform convergence property. Hence we shall complete the proof by extending this property to classes with finite VC dimension. The connection between finite VC dimension and uniform convergence was established by Vapnik and Chervonenkis [2] and later refined by Blumer et al. [3]. First, we introduce the following definition, lemma, and auxiliary theorem.

Definition 5.9 (Growth Function)**.** Let $\mathcal{H}$ be a hypothesis class. Then the growth function of $\mathcal{H}$, denoted by $\tau_{\mathcal{H}} \colon \mathbb{N} \to \mathbb{N}$, is defined as

$$\tau_{\mathcal{H}}(m) = \max_{\substack{C \subset \mathcal{X} \\ |C| = m}} |\mathcal{H}_C| \, .$$

Based on Definition 5.3, if $\mathrm{VCdim}(\mathcal{H}) = d$, then for any $m \le d$ we have $\tau_{\mathcal{H}}(m) = 2^m$. The following lemma will indicate the behavior of the growth function when m is larger than the VC dimension of the hypothesis class $\mathcal{H}$.

Lemma 5.10 (Sauer's Lemma [4])**.** *Let $\mathcal{H}$ be a hypothesis class with* $\mathrm{VCdim}(\mathcal{H}) \le d < \infty$*. Then, for all m, we have $\tau_{\mathcal{H}}(m) \le \sum_{i=0}^{d} \binom{m}{i}$. Furthermore, if $m \ge d + 2$ then $\tau_{\mathcal{H}}(m) \le (\frac{em}{d})^d$.*

Proof. Since $\mathrm{VCdim}(\mathcal{H}) \le d$, we have $|\{B \subseteq C \colon \mathcal{H} \text{ shatters } B\}| \le \sum_{i=0}^{d} \binom{m}{i}$. It suffices to prove that for any $C = \{c_1, \dots, c_m\}$, we have, for all $\mathcal{H}$,

$$|\mathcal{H}_C| \le |\{B \subseteq C \colon \mathcal{H} \text{ shatters } B\}|.$$

Let us prove by induction.

Base Case: For $m = 1$, we want to show $|\mathcal{H}_{\{c_1\}}| \le |\{B \subseteq \{c_1\} \colon \mathcal{H} \text{ shatters } B\}|$. First, the empty set $\emptyset$ is always shattered by any hypothesis class. Second, the singleton $\{c_1\}$ is shattered if and only if there exist hypotheses that assign both labels 0 and 1 to c_1. If $|\mathcal{H}_{\{c_1\}}| = 1$, then the number of shattered subsets ($\{\emptyset\}$) is 1; If $|\mathcal{H}_{\{c_1\}}| = 2$, then the number of shattered subsets ($\{\emptyset, \{c_1\}\}$) is 2.

Inductive Step: Assume that it holds for all sets of size $m - 1$. We want to show that it holds for all sets of size m.

First, fix $\mathcal{H}$ and $C = \{c_1, \dots, c_m\}$. To apply the inductive hypothesis, we can construct a subset of C with size $m - 1$ as $C' = \{c_2, \dots, c_m\}$.

Now, define the sets Y_0 and Y_1:

$$Y_0 = \{(y_2, \dots, y_m) : (0, y_2, \dots, y_m) \in \mathcal{H}_C \vee (1, y_2, \dots, y_m) \in \mathcal{H}_C\}$$

$$Y_1 = \{(y_2, \dots, y_m) : (0, y_2, \dots, y_m) \in \mathcal{H}_C \wedge (1, y_2, \dots, y_m) \in \mathcal{H}_C\}$$

From the disjuctive nature of Y_0, it is easy to see $Y_1 \subseteq Y_0$. If $(y_2, \dots, y_m) \in Y_0 \setminus Y_1$, then only one labeling exists on C (either $c_1 = 0$ or $c_1 = 1$). On the other hand, if $(y_2, \dots, y_m) \in Y_1$, then two labelings exist (both $c_1 = 0$ and $c_1 = 1$). Let us denote $Y_0' = Y_0 \setminus Y_1$. Then we have $|\mathcal{H}_C| = |Y_0'| + 2|Y_1|$. Since $|Y_0'| = |Y_0| - |Y_1|$, we have $|\mathcal{H}_C| = (|Y_0| - |Y_1|) + 2|Y_1| = |Y_0| + |Y_1|$.

Furthermore, since the construction of Y_0 follows the definition of restriction of $\mathcal{H}$ to C', we have $Y_0 = \mathcal{H}_{C'}$. Now, apply the inductive hypothesis (on $\mathcal{H}$ and C'). Then we obtain

$$\begin{aligned} |Y_0| = |\mathcal{H}_{C'}| &\le |\{B \subseteq C' \colon \mathcal{H} \text{ shatters } B\}| \\ &= |\{B \subseteq C \colon c_1 \notin B \wedge \mathcal{H} \text{ shatters } B\}|. \end{aligned}$$

Next, define $\mathcal{H}' \subseteq \mathcal{H}$ such that $\mathcal{H}'$ contains pairs of hypotheses that agree on C' and differ on c_1:

$$\mathcal{H}' = \{h \in \mathcal{H} \colon \exists h' \in \mathcal{H} \text{ s.t. } (1-h'(c_1), h'(c_2), \ldots, h'(c_m)) = (h(c_1), h(c_2), \ldots, h(c_m))\}.$$

Therefore, we can easily conclude that $\mathcal{H}'$ shatters a set $B \subseteq C'$ iff it also shatters the set $B \cup \{c_1\}$.

Furthermore, since the construction of Y_1 follows the definition of restriction of $\mathcal{H}'$ to C', we have $Y_1 = \mathcal{H}'_{C'}$. Now, apply the inductive hypothesis (on $\mathcal{H}'$ and C'). Then we obtain

$$\begin{aligned}|Y_1| = |\mathcal{H}'_{C'}| &\leq |\{B \subseteq C' \colon \mathcal{H}' \text{ shatters } B\}| \\ &= |\{B \subseteq C' \colon \mathcal{H}' \text{ shatters } B \cup \{c_1\}\}| \\ &= |\{B \subseteq C \colon c_1 \in B \wedge \mathcal{H}' \text{ shatters } B\}| \\ &\leq |\{B \subseteq C : c_1 \in B \wedge \mathcal{H} \text{ shatters } B\}|.\end{aligned}$$

Finally, we obtain

$$\begin{aligned}|\mathcal{H}_C| = |Y_0| + |Y_1| &\leq |\{B \subseteq C : c_1 \notin B \wedge \mathcal{H} \text{ shatters } B\}| + |\{B \subseteq C : c_1 \in B \wedge \mathcal{H} \text{ shatters } B\}| \\ &= |\{B \subseteq C \colon \mathcal{H} \text{ shatters } B\}|,\end{aligned}$$

completing the induction.

The proof that $\tau_{\mathcal{H}}(m) \leq (\frac{em}{d})^d$ holds for $m \geq d+2$ is provided in Lemma 8 in Appendix.

□

Theorem 5.11. *Let $\mathcal{H}$ be a hypothesis class and let $\tau_{\mathcal{H}}$ be its growth function. Then for every distribution $\mathcal{D}$ and every $\delta \in (0,1)$, with probability at least $1-\delta$ over the choice of $S \sim \mathcal{D}^m$, we have*

$$\sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)| \leq \frac{4 + \sqrt{\ln(\tau_{\mathcal{H}}(2m))}}{\delta\sqrt{2m}}.$$

Proof. It suffices to prove the following expectation bound:

$$\mathbb{E}_{S \sim \mathcal{D}^m}\left[\sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)|\right] \leq \frac{4 + \sqrt{\ln(\tau_{\mathcal{H}}(2m))}}{\sqrt{2m}}. \tag{5.12}$$

After that, apply Lemma 4 in Appendix to the non-negative random variable $Z = \sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)|$. Hence we have, for any $a > 0$,

$$\mathbb{P}_{S \sim \mathcal{D}^m}[Z \leq a] \geq 1 - \frac{\mathbb{E}_{S \sim \mathcal{D}^m}[Z]}{a}.$$

Let $a = \frac{\mathbb{E}_{S \sim \mathcal{D}^m}[Z]}{\delta}$. Then we have

$$\mathbb{P}_{S \sim \mathcal{D}^m}\left[Z \leq \frac{\mathbb{E}_{S \sim \mathcal{D}^m}[Z]}{\delta}\right] \geq 1 - \delta.$$

Finally, we can get the desirable result

$$\mathbb{P}_{S \sim \mathcal{D}^m}\left[\sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)| \leq \frac{4 + \sqrt{\ln(\tau_{\mathcal{H}}(2m))}}{\delta\sqrt{2m}}\right] \geq 1 - \delta.$$

To show (5.12), we introduce an independent ghost sample $S' = ((x_1', y_1'), \dots, (x_m', y_m')) \sim \mathcal{D}^m$. Since $L_{\mathcal{D}}(h) = \mathbb{E}_{S' \sim \mathcal{D}^m}[L_{S'}(h)]$, we have

$$\begin{aligned}
\mathbb{E}_{S\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)|\right] &= \mathbb{E}_{S\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}} |\mathbb{E}_{S'\sim\mathcal{D}^m}[L_{S'}(h)] - L_S(h)|\right] \\
&\le \mathbb{E}_{S\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}} \mathbb{E}_{S'\sim\mathcal{D}^m}\left[|L_{S'}(h) - L_S(h)|\right]\right] \quad \text{(triangle inequality)} \\
&\le \mathbb{E}_{S,S'\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}} |L_{S'}(h) - L_S(h)|\right] \quad \text{(supremum of expectation)} \\
&= \mathbb{E}_{S,S'\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}} \frac{1}{m}\left|\sum_{i=1}^m \big(\ell(h(x_i'), y_i') - \ell(h(x_i), y_i)\big)\right|\right].
\end{aligned}$$

Due to the fact that both S and S' are composed of independent elements from $\mathcal{D}$, the following holds: For every $\sigma \in \{\pm 1\}^m$, we have

$$\mathbb{E}_{S,S'\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}} \frac{1}{m}\left|\sum_{i=1}^m \big(\ell(h(x_i'), y_i') - \ell(h(x_i), y_i)\big)\right|\right] = \mathbb{E}_{S,S'\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}} \frac{1}{m}\left|\sum_{i=1}^m \sigma_i\big(\ell(h(x_i'), y_i') - \ell(h(x_i), y_i)\big)\right|\right].$$

Then let $\sigma \sim U_{\pm}^m$ be a uniform random variable with distribution over $\{\pm 1\}$. Take its expectation of σ and apply linearity of expectation. Then we have

$$\begin{aligned}
&\mathbb{E}_{S,S'\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}} \frac{1}{m}\left|\sum_{i=1}^m \big(\ell(h(x_i'), y_i') - \ell(h(x_i), y_i)\big)\right|\right] \\
&= \mathbb{E}_{\sigma\sim U_{\pm}^m}\mathbb{E}_{S,S'\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}} \frac{1}{m}\left|\sum_{i=1}^m \sigma_i\big(\ell(h(x_i'), y_i') - \ell(h(x_i), y_i)\big)\right|\right] \\
&= \mathbb{E}_{S,S'\sim\mathcal{D}^m}\mathbb{E}_{\sigma\sim U_{\pm}^m}\left[\sup_{h\in\mathcal{H}} \frac{1}{m}\left|\sum_{i=1}^m \sigma_i\big(\ell(h(x_i'), y_i') - \ell(h(x_i), y_i)\big)\right|\right].
\end{aligned}$$

Fix samples S and S', and let $C = \{x_1, \dots, x_m, x_1', \dots, x_m'\}$ be the set of at most $2m$ distinct input points.

Since the supremum depends only on the behavior on C, we have

$$\sup_{h\in\mathcal{H}} \frac{1}{m}\left|\sum_{i=1}^m \sigma_i\big(\ell(h(x_i'), y_i') - \ell(h(x_i), y_i)\big)\right| = \max_{h\in\mathcal{H}_C} \frac{1}{m}\left|\sum_{i=1}^m \sigma_i\big(\ell(h(x_i'), y_i') - \ell(h(x_i), y_i)\big)\right|.$$

Fix $h \in \mathcal{H}_C$. Let $\theta_h = \frac{1}{m}\sum_{i=1}^m \sigma_i\big(\ell(h(x_i'), y_i') - \ell(h(x_i), y_i)\big)$, which is an average of independent variables, each of which is bounded such that $\sigma_i\big(\ell(h(x_i'), y_i') - \ell(h(x_i), y_i)\big) \in [-1, 1]$. Since $\mathbb{E}[\sigma_i] = 0$ and σ_i are independent, we have $\mathbb{E}[\theta_h] = 0$.

Apply Lemma 7 in Appendix. Then, for any $\rho > 0$, we have

$$\mathbb{P}[|\theta_h| > \rho] \le 2\exp(-2m\rho^2).$$

Apply Lemma 3 in Appendix. Then we have

$$\mathbb{P}\left[\max_{h\in\mathcal{H}_C} |\theta_h| > \rho\right] \le 2|\mathcal{H}_C|\exp(-2m\rho^2). \tag{5.13}$$

Let $X = \max_{h\in\mathcal{H}_C} |\theta_h|$. Hence $X \le 1$. Since the probability bound in (5.13) will be larger than 1 when $|\mathcal{H}_C|$ is very small, we set a critical point $t_0 = \sqrt{\frac{\ln(2|\mathcal{H}_C|)}{2m}}$

such that $2|\mathcal{H}_C|e^{-2mt_0^2} = 1$. Hence, we have $2|\mathcal{H}_C| = e^{2mt_0^2}$. Then the expectation of X can be bounded as the following:

$$\begin{aligned}
\mathbb{E}[X] &= \int_0^\infty \mathbb{P}[X > t]\,dt = \int_0^1 \mathbb{P}[X > t]\,dt \\
&\leq \int_0^{t_0} 1\,dt + \int_{t_0}^1 2|\mathcal{H}_C|\exp(-2mt^2)\,dt \\
&= t_0 + e^{2mt_0^2}\int_{t_0}^1 e^{-2mt^2}\,dt \leq t_0 + e^{2mt_0^2}\int_{t_0}^\infty e^{-2mt^2}\,dt \\
&\leq t_0 + \frac{e^{2mt_0^2}}{\sqrt{2m}}\int_{\sqrt{2m}\,t_0}^\infty e^{-u^2}\,du \quad (\text{let } u = t\sqrt{2m}) \\
&\leq t_0 + \frac{e^{2mt_0^2}}{\sqrt{2m}} \cdot \frac{e^{-2mt_0^2}}{\sqrt{2m}\,t_0} \quad (\text{Gaussian tail bound}).
\end{aligned}$$

Since $2|\mathcal{H}_C|e^{-2mt_0^2} = 1$, we have

$$\mathbb{E}[X] \leq t_0 + \frac{1}{2mt_0} = \frac{\sqrt{\ln(2|\mathcal{H}_C|)} + \frac{1}{\sqrt{\ln(2|\mathcal{H}_C|)}}}{\sqrt{2m}} \leq \frac{4 + \sqrt{\ln(|\mathcal{H}_C|)}}{\sqrt{2m}}.$$

Given $|\mathcal{H}_C| \leq \tau_{\mathcal{H}}(2m)$, we obtain

$$\mathbb{E}_{S,S'\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}} \frac{1}{m}\left|\sum_{i=1}^m \big(\ell(h(x_i'), y_i') - \ell(h(x_i), y_i)\big)\right|\right] \leq \frac{4 + \sqrt{\ln(\tau_{\mathcal{H}}(2m))}}{\sqrt{2m}}.$$

□

Theorem 5.14 (**(vi)** ⇒ **(i)** in Theorem 5.8)**.** *Let $\mathcal{H}$ be a hypothesis class with* $\mathrm{VCdim}(\mathcal{H}) = d < \infty$. *Then, for every $\epsilon, \delta \in (0,1)$, the sample complexity for uniform convergence satisfies*

$$m_{\mathcal{H}}^{UC}(\epsilon,\delta) \leq \left\lceil \frac{8d}{(\delta\epsilon)^2}\ln\left(\frac{4d}{(\delta\epsilon)^2}\right) + \frac{4d\ln\left(\frac{2e}{d}\right)}{(\delta\epsilon)^2}\right\rceil.$$

Proof. By Theorem 5.11, we have

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}} |L_S(h) - L_{\mathcal{D}}(h)| \leq \frac{4 + \sqrt{\ln(\tau_{\mathcal{H}}(2m))}}{\delta\sqrt{2m}}\right] \geq 1 - \delta.$$

Since $\tau_{\mathcal{H}}(2m) \leq \left(\frac{2em}{d}\right)^d$ for $2m \geq d + 2$ by Lemma 5.10, it suffices to show

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}} |L_S(h) - L_{\mathcal{D}}(h)| \leq \frac{4 + \sqrt{d\ln\left(\frac{2em}{d}\right)}}{\delta\sqrt{2m}}\right] \geq 1 - \delta.$$

Since $\sqrt{d\ln\left(\frac{2em}{d}\right)} \geq 4$ for $m \geq \frac{d}{2e}\exp\left(\frac{16}{d}\right)$, it suffices to show

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}} |L_S(h) - L_{\mathcal{D}}(h)| \leq \frac{2\sqrt{d\ln\left(\frac{2em}{d}\right)}}{\delta\sqrt{2m}} = \frac{1}{\delta}\sqrt{\frac{2d\ln\left(\frac{2em}{d}\right)}{m}}\right] \geq 1 - \delta.$$

To meet the assumption of Definition 3.3, we should guarantee $\sup_{h\in\mathcal{H}} |L_S(h) - L_{\mathcal{D}}(h)| \le \epsilon$ with probability at least $1-\delta$. Therefore, we set

$$\frac{1}{\delta}\sqrt{\frac{2d\ln\left(\frac{2em}{d}\right)}{m}} \le \epsilon.$$

Then we have

$$m \ge \frac{2d\ln\left(\frac{2em}{d}\right)}{(\delta\epsilon)^2} = \frac{2d}{(\delta\epsilon)^2}\ln(m) + \frac{2d\ln\left(\frac{2e}{d}\right)}{(\delta\epsilon)^2}.$$

Let $a = \frac{2d}{(\delta\epsilon)^2}$ and $b = \frac{2d\ln\left(\frac{2e}{d}\right)}{(\delta\epsilon)^2}$. Then, by Lemma 10 in Appendix, the sufficient condition for m would be

$$m \ge \frac{8d}{(\delta\epsilon)^2}\ln\left(\frac{4d}{(\delta\epsilon)^2}\right) + \frac{4d\ln\left(\frac{2e}{d}\right)}{(\delta\epsilon)^2}.$$

□

Note that the derived upper bound for $m_{\mathcal{H}}^{\mathrm{UC}}(\epsilon,\delta)$ is not the tightest. However, the further quantitative analysis will not be discussed in this paper.

Acknowledgments

It is a pleasure to thank my mentor Marko Medvedev for his support, guidance, and recommendations in this domain of knowledge. Furthermore, I am grateful to Professor Peter May for reviewing and commenting on this work and for organizing the UChicago Mathematics REU 2025, an exceptionally invaluable research experience for me. Additionally, I would like to acknowledge the textbook *Understanding Machine Learning: From Theory to Algorithms* by Shalev-Shwartz and Ben-David [5], highly recommended by my mentor, which served as the primary reference for the proofs and derivations in this work. Finally, I want to express my admiration for my favorite philosopher Gilbert Harman. It was his reframing of the problem of induction as a problem of finding reliable inductive methods—which can be fruitfully investigated in statistical learning theory—that first motivated me to explore this domain of knowledge [6].

References

[1] L. G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984. DOI: https://doi.org/10.1145/1968.1972.

[2] V. N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–280, 1971. DOI: https://doi.org/10.1137/1116025.

[3] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Learnability and the Vapnik–Chervonenkis dimension. *Journal of the ACM*, 36(4):929–965, 1989. DOI: https://doi.org/10.1145/76359.76371.

[4] N. Sauer. On the density of families of sets. *Journal of Combinatorial Theory, Series A*, 13(1):145–147, 1972. DOI: https://doi.org/10.1016/0097-3165(72)90019-2.

[5] S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms.* Cambridge University Press, 2014.

[6] G. Harman and S. R. Kulkarni. *Reliable Reasoning: Induction and Statistical Learning Theory.* The MIT Press, 2007.

Appendix

Axiom 1 (Additivity of Probability)**.** *For a finite sequence of mutually exclusive events $A_1, \dots, A_n$ in a probability space $\mathbb{P}$, the following identity holds:*

$$\mathbb{P}\left[\bigcup_{i=1}^{n} A_i\right] = \sum_{i=1}^{n} \mathbb{P}[A_i].$$

Proposition 2 (Inclusion-Exclusion Principle)**.** *Let A and B be any two events in a probability space $\mathbb{P}$. Then the following identity holds:*

$$\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B] - \mathbb{P}[A \cap B].$$

Proof. We decompose the union $A \cup B$ into the mutually exclusive union $A \cup (B \setminus A)$ where $B \setminus A = B \cap A^c$. By Axiom 1, we have $\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B \setminus A]$. Similarly, we decompose the event B into the mutually exclusive union $(A \cap B) \cup (B \setminus A)$. By Axiom 1, we have $\mathbb{P}[B] = \mathbb{P}[A \cap B] + \mathbb{P}[B \setminus A]$. Therefore, we have $\mathbb{P}[B \setminus A] = \mathbb{P}[B] - \mathbb{P}[A \cap B]$. We conclude $\mathbb{P}[A \cup B] = \mathbb{P}[A] + (\mathbb{P}[B] - \mathbb{P}[A \cap B]) = \mathbb{P}[A] + \mathbb{P}[B] - \mathbb{P}[A \cap B]$. □

Lemma 3 (Union Bound)**.** *For a finite sequence of events $A_1, \dots, A_n$ in a probability space $\mathbb{P}$, the following inequality holds:*

$$\mathbb{P}\left[\bigcup_{i=1}^{n} A_i\right] \leq \sum_{i=1}^{n} \mathbb{P}[A_i].$$

Proof. Let us prove by induction. For $n = 1$, the inequality is trivial. For $n = 2$, by Proposition 2, we have $\mathbb{P}[A_1 \cup A_2] = \mathbb{P}[A_1] + \mathbb{P}[A_2] - \mathbb{P}[A_1 \cap A_2] \leq \mathbb{P}[A_1] + \mathbb{P}[A_2]$. Assume it holds for some $k \geq 2$, i.e. $\mathbb{P}[\cup_{i=1}^{k} A_i] \leq \sum_{i=1}^{k} \mathbb{P}[A_i]$. For $n = k + 1$, we have

$$\mathbb{P}\left[\bigcup_{i=1}^{k+1} A_i\right] = \mathbb{P}\big[\big(\cup_{i=1}^{k} A_i\big) \cup A_{k+1}\big] \leq \mathbb{P}\big[\cup_{i=1}^{k} A_i\big] + \mathbb{P}[A_{k+1}] \leq \sum_{i=1}^{k+1} \mathbb{P}[A_i].$$

Thus the inequality holds for all $n \geq 1$. □

Lemma 4 (Markov's Inequality)**.** *Let Z be a non-negative random variable with the probability density function $f_Z(z)$, and let $a > 0$. Then, the following inequality holds:*

$$\mathbb{P}[Z \geq a] \leq \frac{\mathbb{E}[Z]}{a}.$$

Proof. For $z \geq a$, we have $1 \leq z/a$. Hence, we have

$$\mathbb{P}[Z \geq a] = \int_a^\infty f_Z(z)\,dz \leq \int_a^\infty \tfrac{z}{a} f_Z(z)\,dz \leq \tfrac{1}{a} \int_0^\infty z f_Z(z)\,dz = \tfrac{\mathbb{E}[Z]}{a}.$$

□

The following lemma is a modified version of Lemma 4.

Lemma 5 (Modified Markov's Inequality)**.** *Let Z be a random variable that takes values in $[0, 1]$. Assume $\mathbb{E}[Z] = \mu$. Then, for any $a \in (0, 1)$, we have*

$$\mathbb{P}[Z \geq 1 - a] \geq \frac{\mu - (1 - a)}{a}.$$

Consequently, for any $b \in (0,1)$,

$$\mathbb{P}[Z \geq b] \geq \frac{\mu - b}{1 - b} \geq \mu - b.$$

Proof. Let $Y = 1 - Z$. Then, $\mathbb{E}[Y] = 1 - \mu$. By Lemma 4, we have $\mathbb{P}[Z < 1 - a] = \mathbb{P}[Y > a] \leq \frac{1-\mu}{a}$. Hence, $\mathbb{P}[Z \geq 1 - a] \geq 1 - \frac{1-\mu}{a} = \frac{\mu-(1-a)}{a}$. With $b = 1 - a$, it follows that $\mathbb{P}[Z \geq b] \geq \frac{\mu-b}{1-b} \geq \mu - b$. □

Lemma 6 (Hoeffding's Lemma)**.** *Let* X *take values in* $[a,b]$ *with* $\mathbb{E}[X] = 0$. *Then for every* $\lambda > 0$, *we have*

$$\mathbb{E}[e^{\lambda X}] \leq e^{\frac{\lambda^2(b-a)^2}{8}}.$$

Proof. Since $f(x) = e^{\lambda x}$ is a convex function, we have that for every $\alpha \in (0,1)$ and $x \in [a,b]$, $f(x) \leq \alpha f(a) + (1-\alpha)f(b) = \alpha e^{\lambda a} + (1-\alpha)e^{\lambda b}$. Setting $\alpha = \frac{b-x}{b-a} \in [0,1]$ (which implies $1 - \alpha = \frac{x-a}{b-a}$), we obtain $e^{\lambda x} \leq \frac{b-x}{b-a}e^{\lambda a} + \frac{x-a}{b-a}e^{\lambda b}$.

Taking the expectation of both sides, and using the linearity of expectation and the fact that $\mathbb{E}[X] = 0$, we obtain $\mathbb{E}[e^{\lambda X}] \leq \mathbb{E}\left[\frac{b-X}{b-a}e^{\lambda a} + \frac{X-a}{b-a}e^{\lambda b}\right] = \frac{b}{b-a}e^{\lambda a} - \frac{a}{b-a}e^{\lambda b}$.

Let $p = \frac{-a}{b-a}$. Then we have $\mathbb{E}[e^{\lambda X}] \leq (1-p)e^{-\lambda(b-a)p} + pe^{-\lambda(b-a)(p-1)}$.

Let $h = \lambda(b-a)$. Then we have $\mathbb{E}[e^{\lambda X}] \leq (1-p)e^{-hp} + pe^{-h(p-1)} = e^{-hp}(1-p+pe^h)$.

Let $L(h) = -hp + \ln(1-p+pe^h)$. Then we have $\mathbb{E}[e^{\lambda X}] \leq e^{L(h)}$.

Define the function $g(h) = L(h) - \frac{h^2}{8} = -hp + \ln(1-p+pe^h) - \frac{h^2}{8}$ for $h > 0$.

Observe $g(0) = 0, g'(0) = 0$, and $g''(h) = \frac{(1-p)pe^h}{[(1-p)+pe^h]^2} - \frac{1}{4} \leq 0$. It follows $g(h) \leq 0$ for all $h > 0$. Hence we obtain $L(h) \leq \frac{h^2}{8} = \frac{\lambda^2(b-a)^2}{8}$.

We conclude $\mathbb{E}[e^{\lambda X}] \leq e^{L(h)} \leq e^{\frac{h^2}{8}} = e^{\frac{\lambda^2(b-a)^2}{8}}$. □

Lemma 7 (Hoeffding's Inequality)**.** *Let* $Z_1, \ldots, Z_m$ *be a sequence of i.i.d. random variables and let* $\bar{Z} = \frac{1}{m}\sum_{i=1}^m Z_i$. *Assume that* $\mathbb{E}[Z_i] = \mu$ *and* $\mathbb{P}[a \leq Z_i \leq b] = 1$ *for every* i. *Then, for any* $\epsilon > 0$,

$$\mathbb{P}[\left|\bar{Z} - \mu\right| > \epsilon] \leq 2\exp\left(-\frac{2m\epsilon^2}{(b-a)^2}\right).$$

Proof. Define random variables $X_i = Z_i - \mathbb{E}[Z_i] = Z_i - \mu$, and let $\bar{X} = \frac{1}{m}\sum_{i=1}^m X_i = \bar{Z} - \mu$. Let $c = b - a$. Then $X_i \in [a-\mu, b-\mu]$ implies that the range of X_i is at most c. First, we bound the one-sided tail probability $\mathbb{P}\left[\bar{X} \geq \epsilon\right]$. For any $\lambda > 0$, using the monotonicity of the exponential function and Lemma 4, we have $\mathbb{P}\left[\bar{X} \geq \epsilon\right] = \mathbb{P}\left[e^{\lambda\bar{X}} \geq e^{\lambda\epsilon}\right] \leq e^{-\lambda\epsilon}\,\mathbb{E}\left[e^{\lambda\bar{X}}\right]$.

By the independence of X_i, we have $\mathbb{E}\left[e^{\lambda\bar{X}}\right] = \mathbb{E}\left[\prod_{i=1}^m e^{\lambda X_i/m}\right] = \prod_{i=1}^m \mathbb{E}\left[e^{\lambda X_i/m}\right]$.

Applying Lemma 6 to each X_i (which satisfies $\mathbb{E}[X_i] = 0$ and $X_i \in [a-\mu, b-\mu]$, in the range $c = b - a$), we have $\mathbb{E}\left[e^{\lambda X_i/m}\right] \leq \exp\left(\frac{\lambda^2c^2}{8m^2}\right)$. Therefore, we obtain $\mathbb{E}\left[e^{\lambda\bar{X}}\right] \leq \prod_{i=1}^m \exp\left(\frac{\lambda^2c^2}{8m^2}\right) = \exp\left(\frac{\lambda^2c^2}{8m}\right)$. Substituting back, we have $\mathbb{P}\left[\bar{X} \geq \epsilon\right] \leq \exp\left(-\lambda\epsilon + \frac{\lambda^2c^2}{8m}\right)$.

This bound holds for any $\lambda > 0$. To obtain the tightest bound, we minimize the exponent over λ. Let $g(\lambda) = -\lambda\epsilon + \frac{\lambda^2c^2}{8m}$. The function $g(\lambda)$ is minimized at

$\lambda^* = \frac{4m\epsilon}{c^2}$. Then the minimum value is $g(\lambda^*) = -\frac{4m\epsilon^2}{c^2} + \frac{(4m\epsilon/c^2)^2c^2}{8m} = -\frac{2m\epsilon^2}{c^2}$. Therefore, $\mathbb{P}\left[\bar{X} \geq \epsilon\right] \leq \exp\left(-\frac{2m\epsilon^2}{c^2}\right)$.

The same argument applies to $-\bar{X}$, yielding $\mathbb{P}\left[\bar{X} \leq -\epsilon\right] \leq \exp\left(-\frac{2m\epsilon^2}{c^2}\right)$. Therefore, we have $\mathbb{P}\left[|\bar{X}| > \epsilon\right] = \mathbb{P}\left[\bar{X} > \epsilon\right] + \mathbb{P}\left[\bar{X} < -\epsilon\right] \leq 2\exp\left(-\frac{2m\epsilon^2}{c^2}\right)$.

Recalling that $\bar{X} = \bar{Z} - \mu$ and $c = b - a$, we conclude that $\mathbb{P}\left[\left|\bar{Z} - \mu\right| > \epsilon\right] \leq 2\exp\left(-\frac{2m\epsilon^2}{(b-a)^2}\right)$. □

Lemma 8. *Let m, d be positive integers such that $d \leq m - 2$. Then we have*

$$\sum_{k=0}^{d} \binom{m}{k} \leq \left(\frac{em}{d}\right)^d.$$

Proof. Let us prove by induction. For $d = 1$, the LHS is $\sum_{k=0}^{1}\binom{m}{k} = 1 + m$, while the RHS is em. Since $1 + m \leq em$ for all $m \geq 1$, the inequality holds. Now, assume $\sum_{k=0}^{d}\binom{m}{k} \leq \left(\frac{em}{d}\right)^d$ for some $d \geq 1$. Then

$$\sum_{k=0}^{d+1}\binom{m}{k} \leq \left(\tfrac{em}{d}\right)^d + \binom{m}{d+1} = \left(\tfrac{em}{d}\right)^d\left[1 + \left(\tfrac{d}{em}\right)^d\binom{m}{d+1}\right].$$

Since $\binom{m}{d+1} = \frac{m(m-1)\cdots(m-d)}{(d+1)!} \leq \frac{m^{d+1}}{(d+1)d!}$, we get

$$\left(\tfrac{d}{em}\right)^d\binom{m}{d+1} \leq \tfrac{1}{d+1}\cdot\tfrac{d^d}{e^d d!}.$$

Using Stirling's bound $d! \geq \sqrt{2\pi d}\,(d/e)^d$, it follows that $\frac{d^d}{e^d d!} \leq \frac{1}{\sqrt{2\pi d}}$. Hence, we have

$$\left(\tfrac{d}{em}\right)^d\binom{m}{d+1} \leq \tfrac{1}{(d+1)\sqrt{2\pi d}}.$$

Therefore, we have

$$\sum_{k=0}^{d+1}\binom{m}{k} \leq \left(\tfrac{em}{d}\right)^d\left(1 + \tfrac{1}{(d+1)\sqrt{2\pi d}}\right) \leq \left(\tfrac{em}{d}\right)^d \cdot \tfrac{2d+3}{2(d+1)}.$$

Since $d \leq m - 2$, we have $\frac{2d+3}{2(d+1)} \leq \frac{m}{d+1}$, we have

$$\sum_{k=0}^{d+1}\binom{m}{k} \leq \left(\tfrac{em}{d}\right)^d \cdot \tfrac{m}{d+1}.$$

Meanwhile, since $\left(\frac{d}{d+1}\right)^d \geq 1/e$, we have

$$\left(\tfrac{em}{d+1}\right)^{d+1} = \left(\tfrac{em}{d}\right)^d \cdot \tfrac{em}{d+1} \cdot \left(\tfrac{d}{d+1}\right)^d \geq \left(\tfrac{em}{d}\right)^d \cdot \tfrac{m}{d+1}.$$

Finally, we obtain

$$\sum_{k=0}^{d+1}\binom{m}{k} \leq \left(\tfrac{em}{d+1}\right)^{d+1},$$

completing the induction. □

Lemma 9. *Let $a > 0$. Then $x \geq 2a\ln(a) \Rightarrow x \geq a\ln(x)$.*

Proof. If $a \in (0, \sqrt{e})$, then $\ln(x) \leq x/2 < x/a$ for all $x > 0$, so $x \geq a\ln(x)$ holds trivially. Assume $a \geq \sqrt{e}$. Define $f(x) = x - a\ln(x)$. Since $f'(x) = 1 - a/x$, f is strictly increasing for $x > a$. Hence, for $x \geq 2a\ln(a) > a$, it suffices to verify $f(2a\ln(a)) \geq 0$. We have $f(2a\ln(a)) = 2a\ln(a) - a\ln(2a\ln(a)) = a[\ln(a) - \ln(2\ln(a))]$. Hence $f(2a\ln(a)) \geq 0$ if and only if $a \geq 2\ln(a)$. Let $g(a) = a - 2\ln(a)$. Then $g'(a) = 1 - 2/a$, so g attains its minimum at $a = 2$, where $g(2) = 2 - 2\ln 2 > 0$. Thus $a > 2\ln(a)$ for all $a > 0$, implying $f(2a\ln(a)) > 0$. □

Lemma 10. *Let $a \geq 1$ and $b > 0$. Then $x \geq 4a\ln(2a) + 2b \Rightarrow x \geq a\ln(x) + b$.*

Proof. From $x \geq 4a\ln(2a) + 2b$, we have $x \geq 2b$ and $x \geq 4a\ln(2a)$. By Lemma 9, the latter implies $x \geq 2a\ln(x)$. Adding these, we have $x \geq a\ln(x) + b$. □