# Overfitting has a limitation

## A model-independent generalization gap bound based on Rényi entropy

Atsushi SUZUKI
Jing WANG

## This paper and slides are available at:

- arXiv: https://arxiv.org/abs/2506.00182
- Slides: https://ash-suzuki.github.io/material/renyi_entropy_overfitting_limit_slides.pdf

# Outline

# Introduction

# Introduction: Do large model and large data set have future potential?

Recent AIs using large-scale neural networks (NNs) have succeeded with extremely high-dimensional parameters, trained on large-scale data (e.g., ChatGPT [37, 8, 1, 23], Gemini [15, 16], LLaMA [48, 11], Claude [2], Qwen [4, 51, 52], DeepSeek [31, 21], Hunyuan [41], PaLM-E [10], etc.).

# Introduction: Do large model and large data set have future potential?

Recent AIs using large-scale neural networks (NNs) have succeeded with extremely high-dimensional parameters, trained on large-scale data (e.g., ChatGPT [37, 8, 1, 23], Gemini [15, 16], LLaMA [48, 11], Claude [2], Qwen [4, 51, 52], DeepSeek [31, 21], Hunyuan [41], PaLM-E [10], etc.).

Will machine learning continue to succeed by using extremely large machine learning models on even larger datasets in the future?

## Introduction: A large generalization gap can be a potential issue.

Many existing analytical results suggest that the **generalization gap** worsens as the scale of the machine learning model increases.

## Introduction: A large generalization gap can be a potential issue.

Many existing analytical results suggest that the **generalization gap** worsens as the scale of the machine learning model increases.

Here, **Generalization gap** = **Performance badness in reality**
- **Performance badness on the training data**.

E.g., Rademacher complexity [26, 27, 6] analyses on NNs
[35, 5, 49, 19, 20, 29, 22, 9] have provided upper bounds on the generalization gap.

## Introduction: A large generalization gap can be a potential issue.

Many existing analytical results suggest that the **generalization gap** worsens as the scale of the machine learning model increases.

Here, **Generalization gap** = **Performance badness in reality**
- **Performance badness on the training data**.

E.g., Rademacher complexity [26, 27, 6] analyses on NNs
[35, 5, 49, 19, 20, 29, 22, 9] have provided upper bounds on the generalization gap.

Better bounds have been given for compressible NNs using Rademacher complexity theories [3, 43, 44] and PAC Bayes theories [54, 32], but they still depend on the scale of the NN, and it is not trivial under what circumstances NNs can be efficiently compressed.

## Introduction: Motivation for a model-independent analysis

The above-mentioned upper bounds on generalization gap strongly depend on the model's construction. This makes it difficult to state success of extremely large-scale models.

## Introduction: Motivation for a model-independent analysis

The above-mentioned upper bounds on generalization gap strongly depend on the model's construction. This makes it difficult to state success of extremely large-scale models.

If a **model-independent** generalization gap theory could be developed, it would encourage the introduction of extremely large models (which might include completely novel NN layers, or might not even be NNs).

## Introduction: Motivation for a model-independent analysis

The above-mentioned upper bounds on generalization gap strongly depend on the model's construction. This makes it difficult to state success of extremely large-scale models.

If a **model-independent** generalization gap theory could be developed, it would encourage the introduction of extremely large models (which might include completely novel NN layers, or might not even be NNs).

**But, in the first place, is it possible?**

## Introduction: Motivation for a model-independent analysis

The above-mentioned upper bounds on generalization gap strongly depend on the model's construction. This makes it difficult to state success of extremely large-scale models.

If a **model-independent** generalization gap theory could be developed, it would encourage the introduction of extremely large models (which might include completely novel NN layers, or might not even be NNs).

**But, in the first place, is it possible?** Yes, if we focus on **the unevenness of a distribution**.

## Introduction: Motivation for a model-independent analysis

The above-mentioned upper bounds on generalization gap strongly depend on the model's construction. This makes it difficult to state success of extremely large-scale models.

If a **model-independent** generalization gap theory could be developed, it would encourage the introduction of extremely large models (which might include completely novel NN layers, or might not even be NNs).

**But, in the first place, is it possible?** Yes, if we focus on **the unevenness of a distribution**.

For example, if the true distribution of the data were concentrated at a single point, the generalization gap would be zero regardless of the machine learning model's construction.

## Introduction: Distribution matters!

Experimental results have shown that discussing models only cannot explain real generalization gap phenomena.

## Introduction: Distribution matters!

Experimental results have shown that discussing models only cannot explain real generalization gap phenomena.

It has been known that in classification problems, the generalization gap on real data is small, whereas if the same model is applied to random labels, the empirical risk can be made small while the expected risk is naturally large, leading to an extremely large generalization gap, even when using the same model [53].

## Introduction: Distribution matters!

Experimental results have shown that discussing models only cannot explain real generalization gap phenomena.

It has been known that in classification problems, the generalization gap on real data is small, whereas if the same model is applied to random labels, the empirical risk can be made small while the expected risk is naturally large, leading to an extremely large generalization gap, even when using the same model [53]. This cannot be explained in principle by focusing only on the model's construction.

## Introduction: Distribution matters!

Experimental results have shown that discussing models only cannot explain real generalization gap phenomena.

It has been known that in classification problems, the generalization gap on real data is small, whereas if the same model is applied to random labels, the empirical risk can be made small while the expected risk is naturally large, leading to an extremely large generalization gap, even when using the same model [53]. This cannot be explained in principle by focusing only on the model's construction. These observations suggest the we need to discuss the distribution in generalization gap analysis.

## Introduction: Our research derives a model-independent generalization gap upper bound!

Our research: there exists an **upper bound on the generalization gap determined solely by the Rényi entropy (an unevenness metric)** of the data-generating distribution!

## Introduction: Our research derives a model-independent generalization gap upper bound!

Our research: there exists an **upper bound on the generalization gap determined solely by the Rényi entropy (an unevenness metric)** of the data-generating distribution!

**TLDR; Uneven distribution leads a small generalization gap.**

## Introduction: Our research derives a model-independent generalization gap upper bound!

Our research: there exists an **upper bound on the generalization gap determined solely by the Rényi entropy (an unevenness metric)** of the data-generating distribution!

**TLDR; Uneven distribution leads a small generalization gap.**

(Provided that we use a machine learning algorithm whose hypothesis is determined by the histogram of the training data (a.k.a. a symmetric algorithm), such as training error minimization by exhaustive search or gradient methods.)

## Introduction: Main Contributions

The main contributions of this research are as follows:

1. We derived a novel generalization gap upper bound that depends only on Rényi entropy, holding under the sole assumption that the algorithm is symmetric and independent of the specific construction of the machine learning model, and showed with a concrete example that it is not vacuous.

## Introduction: Main Contributions

The main contributions of this research are as follows:

1. We derived a novel generalization gap upper bound that depends only on Rényi entropy, holding under the sole assumption that the algorithm is symmetric and independent of the specific construction of the machine learning model, and showed with a concrete example that it is not vacuous.
2. We successfully explained the phenomenon where the generalization gap deteriorates by randomizing labels even when using the same machine learning model, from the perspective of an increase in Rényi entropy.

## Introduction: Main Contributions

The main contributions of this research are as follows:

1. We derived a novel generalization gap upper bound that depends only on Rényi entropy, holding under the sole assumption that the algorithm is symmetric and independent of the specific construction of the machine learning model, and showed with a concrete example that it is not vacuous.
2. We successfully explained the phenomenon where the generalization gap deteriorates by randomizing labels even when using the same machine learning model, from the perspective of an increase in Rényi entropy.
3. We derived a novel no-free-lunch theorem for non-uniform distributions, showing that the exponential of Rényi entropy governs the data length required for learning, and that the aforementioned generalization gap upper bound is tight.

## Is it really possible?

Is it really possible in principle to provide a meaningful upper bound on the generalization gap without imposing assumptions on the model class?

## Is it really possible?

Is it really possible in principle to provide a meaningful upper bound on the generalization gap without imposing assumptions on the model class?

**Observation**: since the empirical distribution of the training data is sparse relative to the true distribution, the difference between performances on the empirical distribution and on the true distribution can be arbitrarily large!

## Is it really possible?

Is it really possible in principle to provide a meaningful upper bound on the generalization gap without imposing assumptions on the model class?

**Observation**: since the empirical distribution of the training data is sparse relative to the true distribution, the difference between performances on the empirical distribution and on the true distribution can be arbitrarily large!

The above observation is correct for a data space such as a real vector space and assume a continuous distribution over it. However, we assume:

**Assumption**: The data space is an **at most countably infinite set**, i.e, there exists an injection from the data space to the set of natural numbers.

## What happens for an at most countable data space?

If the data space is at most countable, the probability distribution is represented by a probability mass function (rather than a probability density function).

## What happens for an at most countable data space?

If the data space is at most countable, the probability distribution is represented by a probability mass function (rather than a probability density function).

Thus, if the training data length is sufficiently large, **the support of its empirical distribution can cover the data points generated from the true distribution with high probability**! Hence, there is a constraint between the performances on the training data and on the true distribution.

## What happens for an at most countable data space?

If the data space is at most countable, the probability distribution is represented by a probability mass function (rather than a probability density function).

Thus, if the training data length is sufficiently large, **the support of its empirical distribution can cover the data points generated from the true distribution with high probability**! Hence, there is a constraint between the performances on the training data and on the true distribution.

Recall that the probability of a generated data point being in the training data set is zero if the true distribution is continuous!

## Not too a strong assumption?

However, if an assumption is too strong, it is meaningless. Is the assumption that the data space is at most countable not too strong?

For example, we know from the diagonal argument that a real vector space of one or more dimensions is not a countable set. Is this acceptable?

## Not too a strong assumption?

However, if an assumption is too strong, it is meaningless. Is the assumption that the data space is at most countable not too strong?

For example, we know from the diagonal argument that a real vector space of one or more dimensions is not a countable set. Is this acceptable?

Mathematically, countability is a strong assumption. However, our interest lies in computers, not in real vector spaces. **When dealing with computers, the countability of the data space is a "vacuous assumption."** We can safely assume it at any time!

## Not too a strong assumption?

However, if an assumption is too strong, it is meaningless. Is the assumption that the data space is at most countable not too strong?

For example, we know from the diagonal argument that a real vector space of one or more dimensions is not a countable set. Is this acceptable?

Mathematically, countability is a strong assumption. However, our interest lies in computers, not in real vector spaces. **When dealing with computers, the countability of the data space is a "vacuous assumption."** We can safely assume it at any time!

This is because the set of all values that can be input into a computer is a countable set. More specifically, any input to a computer is a finite binary string; while the collection of such strings is infinite if there is no length limit, it is at most countably infinite!

# Notation

## Notation

- The set of all non-negative integers is denoted by $\mathbb{N}$. Note that $0 \in \mathbb{N}$.

## Notation

- The set of all non-negative integers is denoted by $\mathbb{N}$. Note that $0 \in \mathbb{N}$.
- The set of all real numbers is denoted by $\mathbb{R}$.

## Notation

- The set of all non-negative integers is denoted by $\mathbb{N}$. Note that $0 \in \mathbb{N}$.
- The set of all real numbers is denoted by $\mathbb{R}$.
- When $\mathcal{X}$ and $\mathcal{Y}$ are sets, $\mathcal{X} \times \mathcal{Y}$ denotes the Cartesian product of $\mathcal{X}$ and $\mathcal{Y}$, and $\mathcal{Y}^{\mathcal{X}}$ denotes the set of all maps from $X$ to $Y$:

$$\mathcal{Y}^{\mathcal{X}} = \{f \mid f : \mathcal{X} \to \mathcal{Y}\}.$$

## Notation

- The set of all non-negative integers is denoted by $\mathbb{N}$. Note that $0 \in \mathbb{N}$.
- The set of all real numbers is denoted by $\mathbb{R}$.
- When $\mathcal{X}$ and $\mathcal{Y}$ are sets, $\mathcal{X} \times \mathcal{Y}$ denotes the Cartesian product of $\mathcal{X}$ and $\mathcal{Y}$, and $\mathcal{Y}^{\mathcal{X}}$ denotes the set of all maps from $X$ to $Y$:

$$\mathcal{Y}^{\mathcal{X}} = \{f \mid f : \mathcal{X} \to \mathcal{Y}\}.$$

- For a set $\mathcal{X}$ and $n \in \mathbb{N}$, the Cartesian product of $n$ copies of $\mathcal{X}$ is denoted by $\mathcal{X}^n$.

## Notation (2)

- When the generating distribution of a random variable $Z$ is "$Q$", it is written as $Z \sim Q$.

## Notation (2)

- When the generating distribution of a random variable $Z$ is "$Q$", it is written as $Z \sim Q$.
- $Q^n$ denotes the $n$-fold product measure of $Q$. That is, $Q^n$ is the distribution followed by a sequence of random variables $\boldsymbol{X} := (X_1, X_2, \ldots, X_n)$ composed of $n$ independent random variables $X_1, X_2, \ldots, X_n \sim Q$.

## Notation (2)

- When the generating distribution of a random variable $Z$ is "$Q$", it is written as $Z \sim Q$.
- $Q^n$ denotes the $n$-fold product measure of $Q$. That is, $Q^n$ is the distribution followed by a sequence of random variables $\boldsymbol{X} := (X_1, X_2, \ldots, X_n)$ composed of $n$ independent random variables $X_1, X_2, \ldots, X_n \sim Q$.
- For a random variable $Z \sim Q$ on a set $\mathcal{Z}$ and a real-valued function $\phi : \mathcal{Z} \to \mathbb{R}$ on $\mathcal{Z}$, the expected value of $\phi(Z)$ is written as $\mathbb{E}_{Z \sim Q} \phi(Z)$.

## Notation (2)

- When the generating distribution of a random variable $Z$ is "$Q$", it is written as $Z \sim Q$.
- $Q^n$ denotes the $n$-fold product measure of $Q$. That is, $Q^n$ is the distribution followed by a sequence of random variables $\boldsymbol{X} := (X_1, X_2, \ldots, X_n)$ composed of $n$ independent random variables $X_1, X_2, \ldots, X_n \sim Q$.
- For a random variable $Z \sim Q$ on a set $\mathcal{Z}$ and a real-valued function $\phi : \mathcal{Z} \to \mathbb{R}$ on $\mathcal{Z}$, the expected value of $\phi(Z)$ is written as $\mathbb{E}_{Z \sim Q} \phi(Z)$.
- The probability that an event $A(Z)$ depending on $Z$ occurs is written as $\mathbb{P}_{Z \sim Q}(A(Z))$.

## Notation (3)

- All logarithms in this paper are natural logarithms $\ln$.

## Notation (3)

- All logarithms in this paper are natural logarithms $\ln$.
- All the distributions to appear in this paper are discrete ones on an at most countable set, since computers can handle those sets only.

## Notation (3)

- All logarithms in this paper are natural logarithms $\ln$.
- All the distributions to appear in this paper are discrete ones on an at most countable set, since computers can handle those sets only.
- Hence, we identify probability mass functions with probability measures. That is, when a probability measure $Q$ on an at most countable set $\mathcal{A}$ is given, $Q(\{a\})$ for $a \in \mathcal{A}$ is simply written as $Q(a)$, and $Q$ is regarded as a probability mass function.

# Preliminaries

## Summary of the Preliminaries

- **Expected risk** $\mathrm{Risk}_{(\ell,Q)}(h)$: the performance badness of hypothesis $h$ on the real application setting $Q$.
- **Empirical risk** $\mathrm{EmpRisk}_{(\ell,\boldsymbol{z})}(h)$: the performance badness of hypothesis $h$ on the training data sequence $\boldsymbol{z}$.
- **Generalization gap** = expected risk - empirical risk
- **Rényi entropy** $H_\alpha(Q)$: An unevenness metric for distribution $Q$.
  - A large $\alpha$ effectively ignores small probability masses.

## Problem Setting

Let $\mathcal{Z}$ be a **countable** data space, $\mathcal{H}_{\text{all}}$ be the whole hypothesis set, and $\ell : \mathcal{H}_{\text{all}} \times \mathcal{Z} \to \mathbb{R}$ be a loss function defined on $\mathcal{Z}$ and $\mathcal{H}_{\text{all}}$.

## Problem Setting

Let $\mathcal{Z}$ be a **countable** data space, $\mathcal{H}_{\text{all}}$ be the whole hypothesis set, and $\ell : \mathcal{H}_{\text{all}} \times \mathcal{Z} \to \mathbb{R}$ be a loss function defined on $\mathcal{Z}$ and $\mathcal{H}_{\text{all}}$.

**Example (Classification Problem)**

In the case of a classification problem, the data space is given by the Cartesian product of the input data space $\mathcal{X}$ and the output data space $\mathcal{Y}$, i.e., $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$.

## Problem Setting

Let $\mathcal{Z}$ be a **countable** data space, $\mathcal{H}_{\text{all}}$ be the whole hypothesis set, and $\ell : \mathcal{H}_{\text{all}} \times \mathcal{Z} \to \mathbb{R}$ be a loss function defined on $\mathcal{Z}$ and $\mathcal{H}_{\text{all}}$.

**Example (Classification Problem)**

In the case of a classification problem, the data space is given by the Cartesian product of the input data space $\mathcal{X}$ and the output data space $\mathcal{Y}$, i.e., $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$.

For a deterministic classification problem, the whole hypothesis set is the set of all maps from $\mathcal{X}$ to $\mathcal{Y}$, i.e., $\mathcal{H}_{\text{all}} = \mathcal{Y}^{\mathcal{X}}$.

# Loss Function Example on Classification

**Example (Classification Problem (continued))**

Then, the 0-1 loss

$$\ell_{0\text{-}1} : \mathcal{Y}^{\mathcal{X}} \times (\mathcal{X} \times \mathcal{Y}) \to \mathbb{R}$$

is defined as

$$\ell_{0\text{-}1}(f, (x, y)) := \mathbb{1}(y \neq f(x)) := \begin{cases} 1 & \text{if } y \neq f(x), \\ 0 & \text{otherwise,} \end{cases}$$

where $f \in \mathcal{Y}^{\mathcal{X}}$, $x \in \mathcal{X}$, and $y \in \mathcal{Y}$.

# Risk, Generalization Gap, and Overfitting

**Definition (Expected Risk and Empirical Risk)**

Let $\mathcal{Z}$ be a **countable** data space, $\mathcal{H}_{\text{all}}$ be the whole hypothesis set, and $\ell : \mathcal{H}_{\text{all}} \times \mathcal{Z} \to \mathbb{R}$ be a loss function defined on $\mathcal{Z}$ and $\mathcal{H}_{\text{all}}$.

## Risk, Generalization Gap, and Overfitting

**Definition (Expected Risk and Empirical Risk)**

Let $\mathcal{Z}$ be a **countable** data space, $\mathcal{H}_{\text{all}}$ be the whole hypothesis set, and $\ell : \mathcal{H}_{\text{all}} \times \mathcal{Z} \to \mathbb{R}$ be a loss function defined on $\mathcal{Z}$ and $\mathcal{H}_{\text{all}}$.

Also, let $Q$ be a (discrete) probability measure on $\mathcal{Z}$, and consider a data sequence of length $n \in \mathbb{N}$, $\boldsymbol{z} := (z_1, z_2, \ldots, z_n) \in \mathcal{Z}^n$.

## Risk, Generalization Gap, and Overfitting

**Definition (Expected Risk and Empirical Risk)**

Let $\mathcal{Z}$ be a **countable** data space, $\mathcal{H}_{\mathrm{all}}$ be the whole hypothesis set, and $\ell : \mathcal{H}_{\mathrm{all}} \times \mathcal{Z} \to \mathbb{R}$ be a loss function defined on $\mathcal{Z}$ and $\mathcal{H}_{\mathrm{all}}$.

Also, let $Q$ be a (discrete) probability measure on $\mathcal{Z}$, and consider a data sequence of length $n \in \mathbb{N}$, $\boldsymbol{z} := (z_1, z_2, \ldots, z_n) \in \mathcal{Z}^n$.

At this time, the **expected risk function** $\mathrm{Risk}_{(\ell,Q)} : \mathcal{H}_{\mathrm{all}} \to \mathbb{R}$ on $Q$ and the **empirical risk function** $\mathrm{EmpRisk}_{(\ell,\boldsymbol{z})} : \mathcal{H}_{\mathrm{all}} \to \mathbb{R}$ on $\boldsymbol{z}$ are defined respectively as follows:

$$\mathrm{Risk}_{(\ell,Q)}(h) := \mathbb{E}_{Z\sim Q}\ell(h, Z), \qquad \mathrm{EmpRisk}_{(\ell,\boldsymbol{z})}(h) := \frac{1}{n}\sum_{i=1}^{n}\ell(h, z_i).$$

**Definition (Generalization Gap)**

Furthermore, the **generalization gap function** $\mathrm{GenGap}_{(\ell,Q,\boldsymbol{z})} : \mathcal{H}_{\mathrm{all}} \to \mathbb{R}$ on $Q$ and $\boldsymbol{z}$ is defined by

$$\mathrm{GenGap}_{(\ell,Q,\boldsymbol{z})}(h) := \mathrm{Risk}_{(\ell,Q)}(h) - \mathrm{EmpRisk}_{(\ell,\boldsymbol{z})}(h).$$

When clear from the context, $\ell$ is omitted.

**Remark**

The loss $\ell(h, z)$ quantifies how bad the hypothesis $h \in \mathcal{H}_{\text{all}}$ is on the data point $z \in \mathcal{Z}$.

# Risk, Generalization Gap, and Overfitting

**Remark**

The loss $\ell(h, z)$ quantifies how bad the hypothesis $h \in \mathcal{H}_{\text{all}}$ is on the data point $z \in \mathcal{Z}$.

Therefore, using the loss function $\ell$ and the true data generating distribution $Q$, the goal of machine learning can be formulated as finding $h \in \mathcal{H}_{\text{all}}$ that minimizes the expected risk $\text{Risk}_{(\ell,Q)}(h)$ as much as possible.

# Risk, Generalization Gap, and Overfitting

**Remark**

The loss $\ell(h, z)$ quantifies how bad the hypothesis $h \in \mathcal{H}_{\text{all}}$ is on the data point $z \in \mathcal{Z}$.

Therefore, using the loss function $\ell$ and the true data generating distribution $Q$, the goal of machine learning can be formulated as finding $h \in \mathcal{H}_{\text{all}}$ that minimizes the expected risk $\text{Risk}_{(\ell,Q)}(h)$ as much as possible.

What is important is that the true data generating distribution $Q$ is unknown, so $\text{Risk}_{(\ell,Q)}(h)$ cannot be directly calculated.

**Remark (continued)**

On the other hand, $\mathrm{EmpRisk}_{(\ell,\boldsymbol{z})}(h)$ can be calculated on the training data sequence $\boldsymbol{z} \in \mathcal{Z}^n$.

**Remark (continued)**

On the other hand, $\mathrm{EmpRisk}_{(\ell,\boldsymbol{z})}(h)$ can be calculated on the training data sequence $\boldsymbol{z} \in \mathcal{Z}^n$.

Therefore, for the output $h$ of a machine learning algorithm, when the empirical risk $\mathrm{EmpRisk}_{(\ell,\boldsymbol{z})}(h)$ is calculated, we are interested in how much it differs from the expected risk $\mathrm{Risk}_{(\ell,Q)}(h)$, i.e., the generalization gap $\mathrm{GenGap}_{(\ell,Q,\boldsymbol{z})}(h)$.

**Remark (continued)**

On the other hand, $\mathrm{EmpRisk}_{(\ell,\boldsymbol{z})}(h)$ can be calculated on the training data sequence $\boldsymbol{z} \in \mathcal{Z}^n$.

Therefore, for the output $h$ of a machine learning algorithm, when the empirical risk $\mathrm{EmpRisk}_{(\ell,\boldsymbol{z})}(h)$ is calculated, we are interested in how much it differs from the expected risk $\mathrm{Risk}_{(\ell,Q)}(h)$, i.e., the generalization gap $\mathrm{GenGap}_{(\ell,Q,\boldsymbol{z})}(h)$.

This is why the evaluation of generalization gap is important in the field of machine learning. The phenomenon where the generalization gap becomes large is called **overfitting**.

# Example: Classification Problem

**Example (Classification Problem (continued))**

In classification with the 0-1 loss $\ell = \ell_{0\text{-}1}$, the expected risk of $f \in \mathcal{H}_{\text{all}} = \mathcal{Y}^{\mathcal{X}}$ is

$$\text{Risk}_{(\ell,Q)}(f) = \mathbb{E}_{Z\sim Q}\ell(f, Z) = \mathbb{P}_{(X,Y)\sim Q}(Y \neq f(X)),$$

which is the misclassification rate of $f$ in the true distribution, so this is exactly what we want to minimize in a classification problem.

# Example: Classification Problem

**Example (Classification Problem (continued))**

In classification with the 0-1 loss $\ell = \ell_{0\text{-}1}$, the expected risk of $f \in \mathcal{H}_{\text{all}} = \mathcal{Y}^{\mathcal{X}}$ is

$$\text{Risk}_{(\ell,Q)}(f) = \mathbb{E}_{Z\sim Q}\ell(f, Z) = \mathbb{P}_{(X,Y)\sim Q}(Y \neq f(X)),$$

which is the misclassification rate of $f$ in the true distribution, so this is exactly what we want to minimize in a classification problem.

Considering a natural language chatbot, both the input data set and the output data set can be infinite. They are sets of finite-length strings

$$\Sigma^* := \Sigma^0 \cup \Sigma^1 \cup \Sigma^2 \ldots,$$

where $\Sigma$ is a character set (e.g., all ASCII characters) and is a finite set, and for $l \in \mathbb{N}$, $\Sigma^l$ is the set of all strings of length $l$.

# Model and Learning Algorithm

**Definition (Model and Learning Algorithm)**

A subset $\mathcal{H} \subset \mathcal{H}_{\text{all}}$ of the whole hypothesis set is called a model.

# Model and Learning Algorithm

**Definition (Model and Learning Algorithm)**

A subset $\mathcal{H} \subset \mathcal{H}_{\text{all}}$ of the whole hypothesis set is called a model.

A map

$$\mathfrak{A} : \mathcal{Z}^* \to \mathcal{H},$$

from the set of finite data sequences

$$\mathcal{Z}^* := \mathcal{Z}^0 \cup \mathcal{Z}^1 \cup \mathcal{Z}^2 \cup \dots$$

to the model $\mathcal{H}$ is called a learning algorithm.

## Motivation: Symmetric Algorithms

The empirical risk function is determined by the histogram of the data sequence and does not depend on the order of appearance of each data point.

## Motivation: Symmetric Algorithms

The empirical risk function is determined by the histogram of the data sequence and does not depend on the order of appearance of each data point.

In other words, in a typical setting, we are not interested in the order of the training data points.

## Motivation: Symmetric Algorithms

The empirical risk function is determined by the histogram of the data sequence and does not depend on the order of appearance of each data point.

In other words, in a typical setting, we are not interested in the order of the training data points. This means that information about the order of the data sequence can be disregarded.

## Motivation: Symmetric Algorithms

The empirical risk function is determined by the histogram of the data sequence and does not depend on the order of appearance of each data point.

In other words, in a typical setting, we are not interested in the order of the training data points. This means that information about the order of the data sequence can be disregarded.

Therefore, when considering algorithms, it is natural to consider algorithms whose output is determined by the histogram of the data sequence and does not depend on the order of appearance of each data point. Such algorithms are called symmetric algorithms (e.g., [36]).

# Symmetric Maps

**Definition (Symmetry of a Map)**

For a data space $\mathcal{Z}$, a map

$$\phi : \mathcal{Z}^* \to \mathcal{T}$$

from the set of finite data sequences

$$\mathcal{Z}^* := \mathcal{Z}^0 \cup \mathcal{Z}^1 \cup \mathcal{Z}^2 \cup \dots$$

to some set $\mathcal{T}$ is symmetric if, for any permutation $\sigma \in \mathfrak{S}_n$ of $n$ elements,

$$\phi(z_1, z_2, \dots, z_n) = \phi(z_{\sigma(1)}, z_{\sigma(2)}, \dots, z_{\sigma(n)})$$

holds.

## Symmetric Maps (2)

**Definition (Symmetry of a Map (continued))**

In other words, $\phi$ is symmetric means that $\phi(\boldsymbol{z})$ is determined solely by the histogram of $\boldsymbol{z}$ and does not depend on the order of appearance of the data.

# Examples of Symmetric Maps

**Example (Important symmetric maps in machine learning)**

1. *Empirical risk function*: When a hypothesis $h \in \mathcal{H}_{\text{all}}$ is fixed, the empirical risk considered as a function of the data sequence,

$$\text{EmpRisk}_{(\ell,\cdot)}(h) : \mathcal{Z}^* \to \mathbb{R},$$

   is a symmetric map (real-valued function). This can be seen from the fact that the empirical risk depends only on the histogram of the data, not on its order.

# Examples of Symmetric Maps (2)

**Example (Important symmetric maps in machine learning (continued))**

2. *Gradient of empirical risk*: When hypotheses are identified with elements of a real vector space (i.e., parameterized by real vectors), its gradient in that real vector space,

$$\nabla \mathrm{EmpRisk}_{(\ell,\boldsymbol{z})}(h),$$

is a symmetric map (real vector-valued function).

# Symmetric Learning Algorithms

**Definition (Symmetry of a Learning Algorithm)**

A learning algorithm

$$\mathfrak{A} : \mathcal{Z}^* \to \mathcal{H} \subset \mathcal{H}_{\text{all}}$$

is symmetric if $\mathfrak{A}$ is symmetric as a map in the sense of Definition (Symmetry of a Map).

# Examples of Symmetric Learning Algorithms (1)

**Example (Examples of Symmetric Learning Algorithms)**

As a simple observation, if each step of a learning algorithm depends on the data sequence only through symmetric functions, then the learning algorithm is symmetric.

# Examples of Symmetric Learning Algorithms (1)

**Example (Examples of Symmetric Learning Algorithms)**

As a simple observation, if each step of a learning algorithm depends on the data sequence only through symmetric functions, then the learning algorithm is symmetric. Important examples are listed below.

1. *Empirical risk minimization by exhaustive search*: This can be written as

$$\mathfrak{A}(\boldsymbol{z}) = \operatorname*{argmin}_{h \in \mathcal{H}} \operatorname{EmpRisk}_{(\boldsymbol{z})}(h).$$

# Examples of Symmetric Learning Algorithms (2)

**Example (Examples of Symmetric Learning Algorithms (continued))**

1. *Empirical risk minimization by exhaustive search*: (continued)
   The fact that this empirical risk minimization is a symmetric learning algorithm follows from the fact that the empirical risk is a symmetric function with respect to the data sequence.

# Examples of Symmetric Learning Algorithms (3)

**Example (Examples of Symmetric Learning Algorithms (continued))**

2. *Gradient method with a fixed initial point*:
   This is a general term for methods where the initial hypothesis is $h_0 \in \mathcal{H}$, the hypothesis $h_t \in \mathcal{H}$ at step $t$ is selected depending on the history of past empirical risk gradients $(\nabla \mathrm{EmpRisk}_{(\ell,\boldsymbol{z})}(h_\tau))_{\tau=0}^{t-1}$ and the history of past selected hypotheses $(h_\tau)_{\tau=0}^{t-1}$, and the stopping condition also depends only on these.

# Examples of Symmetric Learning Algorithms (4)

**Example (Examples of Symmetric Learning Algorithms (continued))**

Note that this formulation allows the use of gradient information for $\tau < t - 1$, so it includes algorithms that use auxiliary variables in practice (e.g., Nesterov's accelerated gradient method [34], BFGS method [14, 18, 40]).

# Examples of Symmetric Learning Algorithms (4)

**Example (Examples of Symmetric Learning Algorithms (continued))**

Note that this formulation allows the use of gradient information for $\tau < t - 1$, so it includes algorithms that use auxiliary variables in practice (e.g., Nesterov's accelerated gradient method [34], BFGS method [14, 18, 40]).

Gradient methods are symmetric learning algorithms because the gradient of the empirical risk is a symmetric (real vector-valued) map with respect to the data sequence.

# Future Work: Stochastic Symmetric Algorithms

**Remark (Discussing stochastic symmetric algorithms is important future work)**

In this paper, we only consider **deterministic** symmetric methods, but do not consider stochastic symmetric algorithms.

# Future Work: Stochastic Symmetric Algorithms

**Remark (Discussing stochastic symmetric algorithms is important future work)**

In this paper, we only consider **deterministic** symmetric methods, but do not consider stochastic symmetric algorithms.

Since stochastic symmetric algorithms include algorithms widely used in modern machine learning, including stochastic gradient descent and Adam [25], extending this paper's discussion to those algorithms is important future work.

## Motivation: Unevenness of Distributions

The unevenness of a distribution has a large impact on generalization gap.

## Motivation: Unevenness of Distributions

The unevenness of a distribution has a large impact on generalization gap. To give an extreme example, no matter how large-scale a machine learning model is used, if the data distribution degenerates to a single point, the generalization gap is zero.

## Motivation: Unevenness of Distributions

The unevenness of a distribution has a large impact on generalization gap. To give an extreme example, no matter how large-scale a machine learning model is used, if the data distribution degenerates to a single point, the generalization gap is zero.

Even if not so extreme, there is an intuition that if the data is skewed, the generalization gap will be small.

## Motivation: Unevenness of Distributions

The unevenness of a distribution has a large impact on generalization gap. To give an extreme example, no matter how large-scale a machine learning model is used, if the data distribution degenerates to a single point, the generalization gap is zero.

Even if not so extreme, there is an intuition that if the data is skewed, the generalization gap will be small.

As an example, as already mentioned, even in practical deep learning models, there are known cases where replacing part of the data with uniform random numbers causes a sharp increase in generalization gap [53].

## Motivation: Unevenness of Distributions

The unevenness of a distribution has a large impact on generalization gap. To give an extreme example, no matter how large-scale a machine learning model is used, if the data distribution degenerates to a single point, the generalization gap is zero.

Even if not so extreme, there is an intuition that if the data is skewed, the generalization gap will be small.

As an example, as already mentioned, even in practical deep learning models, there are known cases where replacing part of the data with uniform random numbers causes a sharp increase in generalization gap [53].

This section introduces Rényi entropy as an indicator to quantify the unevenness of a distribution.

## Definition of Rényi Entropy

**Definition**

Let $\alpha \in [0, +\infty]$. The $\alpha$-**Rényi entropy** $H_\alpha(Q) \in [0, +\infty]$ of a discrete probability distribution $Q$ defined on an at most countable set $\mathcal{Z}$ is defined as follows:

$$H_\alpha(Q) = \begin{cases} \displaystyle\sum_{z \in \mathcal{Z}} Q(z) \ln \frac{1}{Q(z)}, & \text{if } \alpha = 1, \\ \ln\left|\operatorname{supp}(Q)\right|, & \text{if } \alpha = 0, \\ \displaystyle -\ln\left(\max_{z \in \mathcal{Z}} Q(z)\right), & \text{if } \alpha = \infty, \\ \displaystyle \frac{1}{1-\alpha} \ln\left(\sum_{z \in \mathcal{Z}} Q(z)^\alpha\right), & \text{otherwise,} \end{cases}$$

where $\operatorname{supp}(Q) := \{z \in \mathcal{Z} \mid Q(z) > 0\}$.

# Meaning of Rényi Entropy

**Remark (Meaning of Rényi Entropy)**

$H_\alpha(Q)$ represents, in some sense, the "unevenness" or "effective support size" (logarithm thereof) of the distribution $Q$.

# Meaning of Rényi Entropy

**Remark (Meaning of Rényi Entropy)**

$H_\alpha(Q)$ represents, in some sense, the "unevenness" or "effective support size" (logarithm thereof) of the distribution $Q$.

This can also be understood from the following observations:

1. For any fixed $\alpha \in [0, +\infty]$, $H_\alpha(Q)$ takes its minimum value of $0$ if and only if $Q$ is a point measure (i.e., $\exists z \in \mathcal{Z}, Q(z) = 1$).

# Meaning of Rényi Entropy

**Remark (Meaning of Rényi Entropy)**

$H_\alpha(Q)$ represents, in some sense, the "unevenness" or "effective support size" (logarithm thereof) of the distribution $Q$.

This can also be understood from the following observations:

1. For any fixed $\alpha \in [0, +\infty]$, $H_\alpha(Q)$ takes its minimum value of $0$ if and only if $Q$ is a point measure (i.e., $\exists z \in \mathcal{Z}, Q(z) = 1$).
2. If the support set $\mathcal{Z}$ is finite, then for any fixed $\alpha \in [0, +\infty]$, $H_\alpha(Q)$ takes its maximum value $\log |\mathcal{Z}|$ if and only if $Q$ is a uniform distribution on $\mathcal{Z}$.

## Monotonicity in $\alpha$

**Remark (Meaning of Rényi Entropy (2))**

Note that, for a fixed probability distribution $Q$, $H_\alpha(Q)$ is continuous and monotonically non-increasing with respect to $\alpha$.

## Monotonicity in $\alpha$

**Remark (Meaning of Rényi Entropy (2))**

Note that, for a fixed probability distribution $Q$, $H_\alpha(Q)$ is continuous and monotonically non-increasing with respect to $\alpha$.

This is because as $\alpha$ increases, the weights of elements with small probability mass are reduced, effectively ignoring them.

## Summary of the Preliminaries

- **Expected risk** $\mathrm{Risk}_{(\ell,Q)}(h)$: the performance badness of hypothesis $h$ on the real application setting $Q$.
- **Empirical risk** $\mathrm{EmpRisk}_{(\ell,\boldsymbol{z})}(h)$: the performance badness of hypothesis $h$ on the training data sequence $\boldsymbol{z}$.
- **Generalization gap** = expected risk - empirical risk
- **Rényi entropy** $H_\alpha(Q)$: An unevenness metric for distribution $Q$.
  - A large $\alpha$ effectively ignores small probability masses.

# Generalization Gap Bound by Rényi Entropy

## Main Theorem Overview

This section clarifies the relation between the generalization gap and Rényi entropy and the training data size.

## Main Theorem Overview

This section clarifies the relation between the generalization gap and Rényi entropy and the training data size.

The section consists of the following contents.

- The dependency of the Generalization gap on the Rényi entropy and the training data size.
- The sufficient condition about the training data size to achieve the aimed generalization gap standard given the Rényi entropy. (Obtained by the above formula.)
- Specific formulae on typical distributions.

# Main Theorem: Setup

**Theorem (Generalization Gap Bound by Rényi Entropy)**

*Fix a whole hypothesis set $\mathcal{H}_{\mathrm{all}}$ and a loss function $\ell : \mathcal{H}_{\mathrm{all}} \times \mathcal{Z} \to \mathbb{R}$ defined on a data space $\mathcal{Z}$ which is an at most countable set.*

# Main Theorem: Setup

**Theorem (Generalization Gap Bound by Rényi Entropy)**

*Fix a whole hypothesis set $\mathcal{H}_{\mathrm{all}}$ and a loss function $\ell : \mathcal{H}_{\mathrm{all}} \times \mathcal{Z} \to \mathbb{R}$ defined on a data space $\mathcal{Z}$ which is an at most countable set.*

*Define*

$$\mathrm{DI}(\ell) := \sup_{h \in \mathcal{H}_{\mathrm{all}},\, z \in \mathcal{Z}} \ell(h, z) - \inf_{h \in \mathcal{H}_{\mathrm{all}},\, z \in \mathcal{Z}} \ell(h, z) \in [0, +\infty]$$

*("$\mathrm{DI}$" means the diameter of the image).*

# Main Theorem: Setup (2)

**Theorem (Generalization Gap Bound by Rényi Entropy (continued))**

*Let $\mathcal{Z}^* := \mathcal{Z}^0 \cup \mathcal{Z}^1 \cup \mathcal{Z}^2 \cup \dots$ be the set of all finite-length data sequences, and let $\mathfrak{A} : \mathcal{Z}^* \to \mathcal{H}_{\text{all}}$ be a symmetric machine learning algorithm in the sense of Definition (Symmetry of a Learning Algorithm).*

# Main Theorem: Setup (2)

**Theorem (Generalization Gap Bound by Rényi Entropy (continued))**

*Let $\mathcal{Z}^* := \mathcal{Z}^0 \cup \mathcal{Z}^1 \cup \mathcal{Z}^2 \cup \dots$ be the set of all finite-length data sequences, and let $\mathfrak{A} : \mathcal{Z}^* \to \mathcal{H}_{\text{all}}$ be a symmetric machine learning algorithm in the sense of Definition (Symmetry of a Learning Algorithm).*

*Let $Q$ be a probability distribution on $\mathcal{Z}$, and for $\alpha \in [0, 1]$, define $\kappa_{(Q,\alpha)} : \mathbb{N} \to \mathbb{R}$ by*

$$\kappa_{(Q,\alpha)}(n) := n^{\alpha} \exp\left((1-\alpha) H_{\alpha}(Q)\right),$$

*and define $\kappa^*_{(Q)} : \mathbb{N} \to \mathbb{R}$ by*

$$\kappa^*_{(Q)}(n) := \min_{\alpha \in [0,1]} \kappa_{(Q,\alpha)}(n).$$

## Main Theorem: Statement

**Theorem (Generalization Gap Bound by Rényi Entropy (conclusion))**

*When* $n \in \mathbb{N}_{>0}$ *and* $\boldsymbol{Z} = (Z_1, Z_2, \ldots, Z_n) \sim Q^n$*, i.e.,* $Z_1, Z_2, \ldots, Z_n \sim Q$ *independently, for any* $\delta_1, \delta_2, \delta_3 > 0$*, the following holds with probability at least* $1 - (\delta_1 + \delta_2 + \delta_3)$*:*

$$\begin{aligned}&\mathrm{GenGap}_{(\ell,Q,\boldsymbol{Z})}\left(\mathfrak{A}(\boldsymbol{Z})\right)\\ &\leq \mathrm{DI}(\ell)\sqrt{\frac{\left(\kappa^*_{(Q)}(n) + \sqrt{\frac{n}{2}\ln\frac{2}{\delta_3}}\right)\left(3\ln n + \ln(2\pi) + \ln\frac{1}{\delta_2}\right) + \ln\frac{1}{\delta_1}}{2n}}.\end{aligned}$$

**Recall**: $\kappa^*_{(Q)}(n)$ depends on the Rényi entropy $H_\alpha(Q)$. The above inequality shows the relation between the generalization gap and the Rényi entropy.

## Model Independence

**Remark (Theorem (Generalization Gap Bound by Rényi Entropy) is model-independent.)**

Theorem (Generalization Gap Bound by Rényi Entropy) holds regardless of the construction of each hypothesis $h$, the structure of the hypothesis set $\mathcal{H}$, or the relationship between the hypothesis and the loss function $\ell$.

## Model Independence

**Remark (Theorem (Generalization Gap Bound by Rényi Entropy) is model-independent.)**

Theorem (Generalization Gap Bound by Rényi Entropy) holds regardless of the construction of each hypothesis $h$, the structure of the hypothesis set $\mathcal{H}$, or the relationship between the hypothesis and the loss function $\ell$.

No matter how complex a function an individual $h$ is, no matter how many parameters $\mathcal{H}$ has or how complex a model it is constructed with, and no matter how discontinuously $\ell$ behaves with respect to $h$ or $z$, Theorem (Generalization Gap Bound by Rényi Entropy) holds.

## Model Independence

**Remark (Theorem (Generalization Gap Bound by Rényi Entropy) is model-independent.)**

Theorem (Generalization Gap Bound by Rényi Entropy) holds regardless of the construction of each hypothesis $h$, the structure of the hypothesis set $\mathcal{H}$, or the relationship between the hypothesis and the loss function $\ell$.

No matter how complex a function an individual $h$ is, no matter how many parameters $\mathcal{H}$ has or how complex a model it is constructed with, and no matter how discontinuously $\ell$ behaves with respect to $h$ or $z$, Theorem (Generalization Gap Bound by Rényi Entropy) holds.

In that sense, Theorem (Generalization Gap Bound by Rényi Entropy) is model-independent.

## Rough Behavior of the Bound

**Remark (Rough behavior of the generalization gap upper bound)**

Let

$$\alpha^* := \operatorname*{argmin}_{\alpha \in [0,1]} \kappa_{(Q,\alpha)}(n).$$

In this case, the upper bound of Theorem (Generalization Gap Bound by Rényi Entropy) is roughly of the order

$$\sqrt{n^{\alpha^*-1} \exp\left((1-\alpha^*) H_{\alpha^*}(Q)\right)}.$$

## Rough Behavior of the Bound

**Remark (Rough behavior of the generalization gap upper bound)**

Let

$$\alpha^* := \operatorname*{argmin}_{\alpha \in [0,1]} \kappa_{(Q,\alpha)}(n).$$

In this case, the upper bound of Theorem (Generalization Gap Bound by Rényi Entropy) is roughly of the order

$$\sqrt{n^{\alpha^*-1} \exp\left((1-\alpha^*) H_{\alpha^*}(Q)\right)}.$$

If we ignore the dependence of $\alpha^*$ on $n$, the upper bound is exponential w.r.t. the Rényi entropy.

## Rough Behavior of the Bound

**Remark (Rough behavior of the generalization gap upper bound)**

Let

$$\alpha^* := \operatorname*{argmin}_{\alpha \in [0,1]} \kappa_{(Q,\alpha)}(n).$$

In this case, the upper bound of Theorem (Generalization Gap Bound by Rényi Entropy) is roughly of the order

$$\sqrt{n^{\alpha^*-1} \exp\left((1-\alpha^*) H_{\alpha^*}(Q)\right)}.$$

If we ignore the dependence of $\alpha^*$ on $n$, the upper bound is exponential w.r.t. the Rényi entropy.

A more detailed discussion will be provided later.

# Trade-off Regarding $\alpha$

**Remark (Trade-off regarding $\alpha$)**

To minimize the right-hand side, one should minimize

$$n^{\alpha} \exp\left((1-\alpha)H_{\alpha}(Q)\right)$$

with respect to $\alpha$.

## Trade-off Regarding $\alpha$

**Remark (Trade-off regarding $\alpha$)**

To minimize the right-hand side, one should minimize

$$n^{\alpha} \exp\left((1-\alpha)H_{\alpha}(Q)\right)$$

with respect to $\alpha$.

Since Rényi entropy is a non-increasing function of $\alpha$, $\exp\left((1-\alpha)H_{\alpha}(Q)\right)$ is a decreasing function in the range $\alpha \in [0,1]$.

## Trade-off Regarding $\alpha$

**Remark (Trade-off regarding $\alpha$)**

To minimize the right-hand side, one should minimize

$$n^{\alpha} \exp\left((1-\alpha)H_{\alpha}(Q)\right)$$

with respect to $\alpha$.

Since Rényi entropy is a non-increasing function of $\alpha$, $\exp\left((1-\alpha)H_{\alpha}(Q)\right)$ is a decreasing function in the range $\alpha \in [0,1]$.

On the other hand, $n^{\alpha}$ is an increasing function of $\alpha$. To obtain a good upper bound, it is necessary to determine a good $\alpha$ within this trade-off.

## Trade-off Regarding $\alpha$ (2)

**Remark (Trade-off regarding $\alpha$ (continued))**

As an extreme case, if we consider $\alpha = 1$, then

$$n^{\alpha} \exp\left((1-\alpha)H_{\alpha}(Q)\right) = n.$$

In this case, the right-hand side becomes $O(\ln n)$, which is a vacuous bound that does not converge to $0$ even if $n$ is increased.

## Trade-off Regarding $\alpha$ (2)

**Remark (Trade-off regarding $\alpha$ (continued))**

As an extreme case, if we consider $\alpha = 1$, then

$$n^{\alpha} \exp\left((1-\alpha)H_{\alpha}(Q)\right) = n.$$

In this case, the right-hand side becomes $O(\ln n)$, which is a vacuous bound that does not converge to $0$ even if $n$ is increased.

Therefore, an appropriate choice of $\alpha$ is essential.

# Case of Diverging Rényi Entropy

**Remark (Case where Rényi entropy diverges)**

There exist distributions $Q$ for which Rényi entropy $H_\alpha(Q)$ always diverges in the range $\alpha \in [0, 1]$.

## Case of Diverging Rényi Entropy

**Remark (Case where Rényi entropy diverges)**

There exist distributions $Q$ for which Rényi entropy $H_\alpha(Q)$ always diverges in the range $\alpha \in [0, 1]$.

This is equivalent to the divergence of Shannon entropy $H_1(Q)$.

# Example of Diverging Entropy

**Remark (Case where Rényi entropy diverges (continued))**

For example, a probability distribution on $\mathcal{Z} = \mathbb{N}$ with

$$Q(k) := \frac{1}{C(k+2)(\ln(k+2))^2},$$

where

$$C := \sum_{k'=0}^{+\infty} \frac{1}{(k'+2)(\ln(k'+2))^2} < +\infty,$$

is such an example.

## Example of Diverging Entropy

**Remark (Case where Rényi entropy diverges (continued))**

For example, a probability distribution on $\mathcal{Z} = \mathbb{N}$ with

$$Q(k) := \frac{1}{C(k+2)(\ln(k+2))^2},$$

where

$$C := \sum_{k'=0}^{+\infty} \frac{1}{(k'+2)(\ln(k'+2))^2} < +\infty,$$

is such an example.

If Rényi entropy $H_\alpha(Q)$ always diverges in the range $\alpha \in [0, 1]$, the upper bound of Theorem (Generalization Gap Bound by Rényi Entropy) is vacuous.

## Pathological Cases

**Remark (Case where Rényi entropy diverges (3))**

However, this is a pathological case, and in such cases, as will be discussed later, it includes cases where learning from finite-length training data is known to be impossible in the sense of the no-free-lunch theorem.

## Pathological Cases

**Remark (Case where Rényi entropy diverges (3))**

However, this is a pathological case, and in such cases, as will be discussed later, it includes cases where learning from finite-length training data is known to be impossible in the sense of the no-free-lunch theorem.

Also, as will be discussed later, the upper bound of Theorem (Generalization Gap Bound by Rényi Entropy) is usually not vacuous even when the tail probability of $Q$ decays according to a power law.

## Motivation: Data Length Conditions

The previous section provided an upper bound on the generalization gap when the data length is fixed.

## Motivation: Data Length Conditions

The previous section provided an upper bound on the generalization gap when the data length is fixed.

Conversely, we are often interested in the sufficient condition for the data length to achieve a target generalization gap.

## Motivation: Data Length Conditions

The previous section provided an upper bound on the generalization gap when the data length is fixed.

Conversely, we are often interested in the sufficient condition for the data length to achieve a target generalization gap.

Essentially, this involves solving Theorem (Generalization Gap Bound by Rényi Entropy) for $n$, but expressing the sufficient condition for data length using elementary functions is a somewhat tedious task because it involves the inverse function of a product of a polynomial and a logarithmic function.

# Theorem: Sufficient Data Length

**Theorem (Sufficient condition for data length determined by Rényi entropy)**

*Assume the same situation as in Theorem (Generalization Gap Bound by Rényi Entropy).*

# Theorem: Sufficient Data Length

**Theorem (Sufficient condition for data length determined by Rényi entropy)**

*Assume the same situation as in Theorem (Generalization Gap Bound by Rényi Entropy).*

*That is, fix a whole hypothesis set $\mathcal{H}_{\mathrm{all}}$, a loss function $\ell : \mathcal{H}_{\mathrm{all}} \times \mathcal{Z} \to \mathbb{R}$ defined on a data space $\mathcal{Z}$ which is an at most countable set, define $\mathrm{DI}(\ell)$ similarly, and let $\mathfrak{A} : \mathcal{Z}^* \to \mathcal{H}_{\mathrm{all}}$ be a symmetric machine learning algorithm in the sense of Definition (Symmetry of a Learning Algorithm).*

## Theorem: Sufficient Data Length (2)

**Theorem (Sufficient condition for data length determined by Rényi entropy (continued))**

*For a (discrete) probability measure $Q$ on $\mathcal{Z}$, define the extended real-valued functions*

$$\nu_{(Q,\alpha)} : (0,1] \to [0,+\infty), \quad \widetilde{\nu}_{(Q,\alpha)} : (0,1]^2 \to [0,+\infty)$$

*as follows:*

$$\nu_{(Q,\alpha)}(\varepsilon) := \left( \frac{24 H_\alpha(Q) \ln \frac{12}{\varepsilon^2(1-\alpha)}}{\varepsilon^2} \right)^{\frac{1}{1-\alpha}} \exp\left(H_\alpha(Q)\right),$$

$$\widetilde{\nu}_{(Q,\alpha)}(\delta,\varepsilon) := \left( \frac{36 \ln \frac{6\pi}{\delta}}{\varepsilon^2} \right)^{\frac{1}{1-\alpha}} \exp\left(H_\alpha(Q)\right).$$

# Theorem: Sufficient Data Length (3)

**Theorem (Sufficient condition for data length determined by Rényi entropy (continued))**

*Also, define*

$$\omega : (0,1]^2 \to [0,+\infty)$$

*as follows:*

$$\omega(\delta,\varepsilon) = \max\left\{ \frac{324 \ln \frac{3}{\delta}}{\varepsilon^4} \left( \ln \frac{9\sqrt{2 \ln \frac{3}{\delta}}}{\varepsilon^2} \right)^2, \frac{3}{2\varepsilon^2} \ln \frac{3}{\delta} \right\}.$$

# Theorem: Sufficient Data Length (4)

**Theorem (Sufficient condition for data length determined by Rényi entropy (conclusion))**

*Fix any* $(\delta, \varepsilon) \in (0,1)^2$.

# Theorem: Sufficient Data Length (4)

**Theorem (Sufficient condition for data length determined by Rényi entropy (conclusion))**

*Fix any* $(\delta, \varepsilon) \in (0,1)^2$.

*If for some* $\alpha \in [0,1]$,

$$n > \max\left\{\nu_{(Q,\alpha)}(\varepsilon), \widetilde{\nu}_{(Q,\alpha)}(\delta,\varepsilon), \omega(\delta,\varepsilon)\right\}$$

*holds, then, when* $\boldsymbol{Z} = (Z_1, Z_2, \ldots, Z_n) \sim Q^n$, *with probability at least* $1-\delta$,

$$\mathrm{GenGap}_{(\ell,Q,\boldsymbol{Z})}\left(\mathfrak{A}(\boldsymbol{Z})\right) < \mathrm{DI}(\ell)\,\varepsilon.$$

## Model Independence of Data-Length Result

**Remark (Theorem (Sufficient condition for data length determined by Rényi entropy) is also model-independent.)**

Theorem (Sufficient condition for data length determined by Rényi entropy) is model-independent in the same sense as stated in
Remark (Theorem (Generalization Gap Bound by Rényi Entropy) is model-independent).

**Remark (The main term is $\nu_{(Q,\alpha)}$)**

In large-scale problems, $H_\alpha(Q)$ usually becomes large, but $\omega(\delta, \varepsilon)$ does not depend on $H_\alpha(Q)$.

## Main Term and Rényi Entropy

**Remark (The main term is $\nu_{(Q,\alpha)}$)**

In large-scale problems, $H_\alpha(Q)$ usually becomes large, but $\omega(\delta, \varepsilon)$ does not depend on $H_\alpha(Q)$.

Also, if $H_\alpha(Q) \gg \ln \frac{1}{\delta}$, then $\nu_{(Q,\alpha)}(\varepsilon) \gg \widetilde{\nu}_{(Q,\alpha)}(\delta, \varepsilon)$.

## Main Term and Rényi Entropy

**Remark (The main term is $\nu_{(Q,\alpha)}$)**

In large-scale problems, $H_\alpha(Q)$ usually becomes large, but $\omega(\delta,\varepsilon)$ does not depend on $H_\alpha(Q)$.

Also, if $H_\alpha(Q) \gg \ln \frac{1}{\delta}$, then $\nu_{(Q,\alpha)}(\varepsilon) \gg \widetilde{\nu}_{(Q,\alpha)}(\delta,\varepsilon)$.

Therefore, $\nu_{(Q,\alpha)}$ is the main term.

## Main Term and Rényi Entropy

**Remark (The main term is $\nu_{(Q,\alpha)}$)**

In large-scale problems, $H_\alpha(Q)$ usually becomes large, but $\omega(\delta,\varepsilon)$ does not depend on $H_\alpha(Q)$.

Also, if $H_\alpha(Q) \gg \ln \frac{1}{\delta}$, then $\nu_{(Q,\alpha)}(\varepsilon) \gg \widetilde{\nu}_{(Q,\alpha)}(\delta,\varepsilon)$.

Therefore, $\nu_{(Q,\alpha)}$ is the main term. The specific form of $\nu_{(Q,\alpha)}$ implies that the data length should be at least proportional to $\exp(H_\alpha(Q))$, the exponential of the Rényi entropy.

# Table: Specific Distributions

| Type of Distribution | Condition | Generalization Gap | Sufficient Data Length |
|---|---|---|---|
| Distribution on a finite set | $\lvert\mathcal{Z}\rvert < +\infty$ | $O\left(\sqrt{\lvert\mathcal{Z}\rvert \frac{\ln n}{2n}}\right)$ | $O\left(\lvert\mathcal{Z}\rvert \ln \lvert\mathcal{Z}\rvert \cdot \frac{1}{\varepsilon^2} \ln \frac{1}{\varepsilon^2}\right)$ |
| Exponentially decaying distribution | $\exists C > 0,$ $r \in (0,1),$ $q_j \leq C r^j$ | $O\left(\sqrt{\frac{eC}{\ln \frac{1}{r}} \cdot \frac{(\ln n)^2}{2n}}\right)$ | $O\left(\frac{1}{\ln \frac{1}{r}} \cdot \frac{1}{\varepsilon^2} \left(\ln \frac{1}{\varepsilon}\right)^2\right)$ |
| Power-law decaying distribution | $\exists C > 0,$ $\gamma > 1,$ $q_j \leq C(j+1)^{-\gamma}$ | $O\left(\sqrt{\frac{(\ln n)^2}{(\gamma-1) n^{\frac{\gamma-1}{\gamma}}}}\right)$ | $O\left(\left(\frac{\gamma^2}{(\gamma-1)^3} \cdot \frac{1}{\varepsilon^2} \left(\ln \frac{1}{\varepsilon}\right)^2\right)^{\frac{\gamma}{\gamma-1}}\right)$ |

## Specific Probability Distributions

Let's see how generalization gap is suppressed through specific probability distributions.

## Specific Probability Distributions

Let's see how generalization gap is suppressed through specific probability distributions.

First, let's look at the relatively trivial case where $\mathcal{Z}$ is a finite set, and that is the only assumption. In this case, which includes the uniform distribution, the Rényi entropy is finite, so a meaningful generalization gap upper bound can be obtained.

## Tail Behavior and Gap

Next, we discuss cases where $\mathcal{Z}$ may be a countably infinite set.

## Tail Behavior and Gap

Next, we discuss cases where $\mathcal{Z}$ may be a countably infinite set.

Theorem (Generalization Gap Bound by Rényi Entropy) asserted that the generalization gap becomes smaller if the unevenness of the data distribution is larger.

## Tail Behavior and Gap

Next, we discuss cases where $\mathcal{Z}$ may be a countably infinite set.

Theorem (Generalization Gap Bound by Rényi Entropy) asserted that the generalization gap becomes smaller if the unevenness of the data distribution is larger.

In other words, the faster the tail of the probability distribution decays, the smaller the generalization gap.

## Exponential vs Power-law Decay

Here, we compare the case where the tail of the probability distribution decays exponentially and the case where it decays according to a power law, and see that the generalization gap upper bound is smaller for exponential decay, but the upper bound of Theorem (Generalization Gap Bound by Rényi Entropy) is not vacuous, i.e., converges to $0$ as $n \to +\infty$ even for power-law decay.

# Exponential vs Power-law Decay

Here, we compare the case where the tail of the probability distribution decays exponentially and the case where it decays according to a power law, and see that the generalization gap upper bound is smaller for exponential decay, but the upper bound of Theorem (Generalization Gap Bound by Rényi Entropy) is not vacuous, i.e., converges to $0$ as $n \to +\infty$ even for power-law decay.

Phenomena with power-law decaying distributions, such as Zipf's law [55], frequently appear especially in natural languages [30, 12, 13, 28, 38, 45, 46, 47].

# Exponential vs Power-law Decay

Here, we compare the case where the tail of the probability distribution decays exponentially and the case where it decays according to a power law, and see that the generalization gap upper bound is smaller for exponential decay, but the upper bound of Theorem (Generalization Gap Bound by Rényi Entropy) is not vacuous, i.e., converges to $0$ as $n \to +\infty$ even for power-law decay.

Phenomena with power-law decaying distributions, such as Zipf's law [55], frequently appear especially in natural languages [30, 12, 13, 28, 38, 45, 46, 47].

Therefore, whether machine learning generalizes for phenomena following these distributions is an important problem.

# Why Random Labels Worsen Generalization Gap

## Motivating Phenomenon

It is known that deep learning models used in practical image recognition have low generalization gap on original data (both training error rate and test error rate are low), but if the data labels are randomized, the generalization gap becomes extremely large (training error rate is low, but test error rate is high) [53].

## Motivating Phenomenon

It is known that deep learning models used in practical image recognition have low generalization gap on original data (both training error rate and test error rate are low), but if the data labels are randomized, the generalization gap becomes extremely large (training error rate is low, but test error rate is high) [53].

This phenomenon cannot be explained in principle by theories that focus only on the function class represented by the model.

## Motivating Phenomenon

It is known that deep learning models used in practical image recognition have low generalization gap on original data (both training error rate and test error rate are low), but if the data labels are randomized, the generalization gap becomes extremely large (training error rate is low, but test error rate is high) [53].

This phenomenon cannot be explained in principle by theories that focus only on the function class represented by the model.

This section provides a direct explanation for this phenomenon from the perspective of an increase in Rényi entropy.

## Effect of Random Labels on Entropy

More specifically, replacing a part of the data with uniform random numbers increases the Rényi entropy.

## Effect of Random Labels on Entropy

More specifically, replacing a part of the data with uniform random numbers increases the Rényi entropy.

Quantitatively, the following holds.

# Proposition: Deterministic vs Uniform Labels

**Proposition (Deterministic label vs uniform random label)**

*Let a random variable $X$ on $\mathcal{X}$ follow a probability distribution $Q$.*

# Proposition: Deterministic vs Uniform Labels

**Proposition (Deterministic label vs uniform random label)**

*Let a random variable $X$ on $\mathcal{X}$ follow a probability distribution $Q$.*

*Let a random variable $Y$ on a finite set $\mathcal{Y}$ be given by $Y = f(X)$ using a deterministic function $f$.*

# Proposition: Deterministic vs Uniform Labels

**Proposition (Deterministic label vs uniform random label)**

*Let a random variable $X$ on $\mathcal{X}$ follow a probability distribution $Q$.*

*Let a random variable $Y$ on a finite set $\mathcal{Y}$ be given by $Y = f(X)$ using a deterministic function $f$.*

*Let a random variable $Y'$ on $\mathcal{Y}$ follow a uniform distribution on $\mathcal{Y}$ independently of $X$.*

# Proposition: Deterministic vs Uniform Labels

**Proposition (Deterministic label vs uniform random label)**

*Let a random variable $X$ on $\mathcal{X}$ follow a probability distribution $Q$.*

*Let a random variable $Y$ on a finite set $\mathcal{Y}$ be given by $Y = f(X)$ using a deterministic function $f$.*

*Let a random variable $Y'$ on $\mathcal{Y}$ follow a uniform distribution on $\mathcal{Y}$ independently of $X$.*

*Then, for any $\alpha \in [0, +\infty]$,*

$$H_\alpha(X, Y') = H_\alpha(X, Y) + \ln |\mathcal{Y}|.$$

## Entropy Increase and Generalization Gap

When Rényi entropy increases additively, there is an exponential effect on the generalization gap.

# Entropy Increase and Generalization Gap

When Rényi entropy increases additively, there is an exponential effect on the generalization gap.

The following theorem formalizes this.

## Theorem: Entropy Increase and Gap

**Theorem (Deterioration of generalization gap caused by an increase in Rényi entropy)**

*Suppose that for two probability distributions $Q$ and $Q'$, there exists some $C \geq 0$ such that*

$$\forall \alpha \in [0,1], \quad H_\alpha(Q') \geq H_\alpha(Q) + C.$$

*Then, for any $n \in \mathbb{N}$,*

$$\kappa^*_{(Q')}(n) \geq \exp\left((1-\alpha'^*)C\right)\ \kappa^*_{(Q)}(n),$$

*where*

$$\alpha'^* := \operatorname*{argmin}_{\alpha \in [0,1]} \exp\left((1-\alpha)H_\alpha(Q')\right) n^\alpha.$$

# Theorem: Entropy Increase and Gap (2)

**Theorem (Deterioration of generalization gap caused by an increase in Rényi entropy (continued))**

*Also, for any* $\alpha \in [0, 1]$ *and any* $(\delta, \varepsilon) \in (0, 1]^2$*,*

$$\max\{\nu_{(Q',\alpha)}(\varepsilon),\, \widetilde{\nu}_{(Q',\alpha)}(\delta, \varepsilon)\} \geq \exp(C)\, \max\{\nu_{(Q,\alpha)}(\varepsilon),\, \widetilde{\nu}_{(Q,\alpha)}(\delta, \varepsilon)\}.$$

## Explanation via Rényi Entropy

**Remark (Deterioration of generalization gap can be explained by the increase in Rényi entropy)**

According to Theorem (Generalization Gap Bound by Rényi Entropy), the main term of the upper bound on generalization gap was

$$O\left(\sqrt{\kappa^*_{(Q)}(n)/n \cdot \ln n}\right).$$

## Explanation via Rényi Entropy

**Remark (Deterioration of generalization gap can be explained by the increase in Rényi entropy)**

According to Theorem (Generalization Gap Bound by Rényi Entropy), the main term of the upper bound on generalization gap was

$$O\left(\sqrt{\kappa^*_{(Q)}(n)/n \cdot \ln n}\right).$$

Therefore, the generalization gap for the probability distribution $Q'$ is roughly $\sqrt{\exp(C)^{1-\alpha^*}}$ times worse than for $Q$.

## Explanation via Rényi Entropy

**Remark (Deterioration of generalization gap can be explained by the increase in Rényi entropy)**

According to Theorem (Generalization Gap Bound by Rényi Entropy), the main term of the upper bound on generalization gap was

$$O\left(\sqrt{\kappa^*_{(Q)}(n)/n \cdot \ln n}\right).$$

Therefore, the generalization gap for the probability distribution $Q'$ is roughly $\sqrt{\exp(C)^{1-\alpha^*}}$ times worse than for $Q$.

Considering the example in Proposition (Deterministic label vs uniform random label), $C = \ln|\mathcal{Y}|$, so in the case of uniform labels $(X, Y')$, the generalization gap is $\sqrt{|\mathcal{Y}|^{1-\alpha^*}}$ times worse than in the case of deterministic labels $(X, Y)$.

## Explanation via Rényi Entropy (2)

**Remark (Deterioration of generalization gap can be explained by the increase in Rényi entropy (continued))**

When $\alpha'^* = 1$, the inequality $\kappa^*_{(Q')}(n) \geq \exp\left((1-\alpha'^*)C\right)\kappa^*_{(Q)}(n)$ is meaningless, but such cases are when Theorem (Generalization Gap Bound by Rényi Entropy) gives a vacuous upper bound, and as we will see in a later example, such cases are rare.

## Explanation via Rényi Entropy (2)

**Remark (Deterioration of generalization gap can be explained by the increase in Rényi entropy (continued))**

When $\alpha'^* = 1$, the inequality $\kappa^*_{(Q')}(n) \geq \exp\left((1-\alpha'^*)C\right)\kappa^*_{(Q)}(n)$ is meaningless, but such cases are when Theorem (Generalization Gap Bound by Rényi Entropy) gives a vacuous upper bound, and as we will see in a later example, such cases are rare.

Furthermore, the sufficient data length $n$ to make the generalization gap less than or equal to $\mathrm{DI}(\ell)\varepsilon$ is effectively given by $\max\left\{\nu_{(Q',\alpha)}(\varepsilon), \widetilde{\nu}_{(Q',\alpha)}(\delta,\varepsilon)\right\}$.

## Explanation via Rényi Entropy (3)

**Remark (Deterioration of generalization gap can be explained by the increase in Rényi entropy (continued))**

Therefore, applying the above theorem, the sufficient condition for data length in the case of probability distribution $Q'$ is $\exp(C)$ times worse than for $Q$.

## Explanation via Rényi Entropy (3)

**Remark (Deterioration of generalization gap can be explained by the increase in Rényi entropy (continued))**

Therefore, applying the above theorem, the sufficient condition for data length in the case of probability distribution $Q'$ is $\exp(C)$ times worse than for $Q$.

Considering the example in Proposition (Deterministic label vs uniform random label) again, in the case of uniform labels $(X, Y')$, the sufficient condition for data length is $|\mathcal{Y}|$ times worse than in the case of deterministic labels $(X, Y)$.

## Explanation via Rényi Entropy (3)

**Remark (Deterioration of generalization gap can be explained by the increase in Rényi entropy (continued))**

Therefore, applying the above theorem, the sufficient condition for data length in the case of probability distribution $Q'$ is $\exp(C)$ times worse than for $Q$.

Considering the example in Proposition (Deterministic label vs uniform random label) again, in the case of uniform labels $(X, Y')$, the sufficient condition for data length is $|\mathcal{Y}|$ times worse than in the case of deterministic labels $(X, Y)$.

This is why the generalization gap deteriorated when the labels were replaced with random labels in [53].

# Rényi Entropy Version of No-free-lunch Theorem

## Motivation of constructing a no-free-lunch theorem

The generalization gap upper bound provides a sufficient condition for a machine learning to succeed. It says that a data size of approximately the exponential of the Rényi entropy is needed.

## Motivation of constructing a no-free-lunch theorem

The generalization gap upper bound provides a sufficient condition for a machine learning to succeed. It says that a data size of approximately the exponential of the Rényi entropy is needed.

This section, we show that the data size is in a sense approximately necessary condition for a machine learning to succeed.

## Motivation of constructing a no-free-lunch theorem

The generalization gap upper bound provides a sufficient condition for a machine learning to succeed. It says that a data size of approximately the exponential of the Rényi entropy is needed.

This section, we show that the data size is in a sense approximately necessary condition for a machine learning to succeed.

We achieve it by constructing a no-free-lunch theorem.

## Classical No-free-lunch Theorem

The No-free-lunch theorem in the context of machine learning (e.g., [39]) formulates a certain theoretical limitation of machine learning, especially supervised learning.

# Classical No-free-lunch Theorem

The No-free-lunch theorem in the context of machine learning (e.g., [39]) formulates a certain theoretical limitation of machine learning, especially supervised learning.

Specifically, it means that even if information that the input-output relationship is a deterministic function is given, any machine learning algorithm will fail in the worst case regarding the input distribution and input-output relationship if there is not enough training data of a length corresponding to the size of the input data space.

# A More Specific Version

The following is a more specific version in [42].

# Theorem: No-free-lunch (Uniform Case)

**Theorem (No-free-lunch theorem)**

*Consider a learning problem from a domain set $\mathcal{X}$ to a codomain set $\mathcal{Y}$ such that $|\mathcal{Y}| \geq 1$, i.e., $\mathcal{Y} \neq \emptyset$.*

# Theorem: No-free-lunch (Uniform Case)

**Theorem (No-free-lunch theorem)**

*Consider a learning problem from a domain set $\mathcal{X}$ to a codomain set $\mathcal{Y}$ such that $|\mathcal{Y}| \geq 1$, i.e., $\mathcal{Y} \neq \emptyset$.*

*For a probability measure $Q$ on $\mathcal{X}$, a ground truth map $f_0 : \mathcal{X} \to \mathcal{Y}$, denote the 0-1 risk of a hypothesis map $f : \mathcal{X} \to \mathcal{Y}$ on $Q$ and $f_0$ by*

$$\mathrm{Risk}_{(\ell_{0\text{-}1},\, Q\circ(\mathrm{id}_{\mathcal{X}}, f_0))}(f),$$

*which is defined by*

$$\mathrm{Risk}_{(Q\circ(\mathrm{id}_{\mathcal{X}}, f_0)^{-1}, \ell_{0\text{-}1})}(f) = \mathbb{P}_{X\sim Q}(f(X) \neq f_0(X)).$$

# Theorem: No-free-lunch (Uniform Case, 2)

**Theorem (No-free-lunch theorem (continued))**

*Then, for any map (learning algorithm)*

$$\mathfrak{A} : (\mathcal{X} \times \mathcal{Y})^* \to (\mathcal{X} \to \mathcal{Y}),$$

*any nonnegative integer (training data size)* $n$ *that satisfies* $n \leq \frac{1}{2}|\mathcal{X}|$*, any finite positive integer* $p$ *satisfying* $1 \leq p \leq |\mathcal{Y}|$*, and any* $\varepsilon \in (0, 1)$*, there exist a ground truth map* $f_0 : \mathcal{X} \to \mathcal{Y}$ *and a finite subset* $\underline{\mathcal{X}} \subset \mathcal{X}$ *such that both the following inequalities hold.*

# Theorem: No-free-lunch (Uniform Case, 3)

**Theorem (No-free-lunch theorem (continued))**

$$\mathbb{E}_{\boldsymbol{Z}\sim(Q\circ(\mathrm{id}_{\mathcal{X}},f_0)^{-1})^n}\mathrm{Risk}_{(\ell_{0\text{-}1},\,\mathrm{Uniform}(\underline{\mathcal{X}})\circ(\mathrm{id}_{\mathcal{X}},f_0))}\left(\mathfrak{A}(\boldsymbol{Z})\right)\geq\mu_{\mathrm{err}}:=\frac{p-1}{2p},$$

$$\mathbb{P}_{\boldsymbol{Z}\sim(Q\circ(\mathrm{id}_{\mathcal{X}},f_0)^{-1})^n}\left(\mathrm{Risk}_{(\ell_{0\text{-}1},\,\mathrm{Uniform}(\underline{\mathcal{X}})\circ(\mathrm{id}_{\mathcal{X}},f_0))}\left(\mathfrak{A}(\boldsymbol{Z})\right)\geq\varepsilon\right)\geq\delta:=\frac{\mu_{\mathrm{err}}-\varepsilon}{1-\varepsilon}=\frac{p-1-2p\varepsilon}{2p-2p\varepsilon}.$$

## Interpretation of the Classical NFL

**Remark**

We are interested in the cases where $|\mathcal{Y}| \geq 2$ and we can take $p$ so that $p \geq 2$.

# Interpretation of the Classical NFL

**Remark**

We are interested in the cases where $|\mathcal{Y}| \geq 2$ and we can take $p$ so that $p \geq 2$.

If $p \geq 2$, then $\mu_{\text{err}} \geq \frac{1}{4}$ and $\delta \geq \frac{1-4\varepsilon}{4-4\varepsilon}$.

## Interpretation of the Classical NFL

**Remark**

We are interested in the cases where $|\mathcal{Y}| \geq 2$ and we can take $p$ so that $p \geq 2$.

If $p \geq 2$, then $\mu_{\text{err}} \geq \frac{1}{4}$ and $\delta \geq \frac{1-4\varepsilon}{4-4\varepsilon}$.

Moreover, if $\varepsilon = \frac{1}{8}$, then $\delta \geq \frac{1}{7}$.

## Limitations of Uniform-case NFL

The statement (and the original proof in [39]) says that the worst distribution is the uniform distribution, in which the training data size should be at least half of the data space size.

## Limitations of Uniform-case NFL

The statement (and the original proof in [39]) says that the worst distribution is the uniform distribution, in which the training data size should be at least half of the data space size.

It has often been pointed out that the situation where the input distribution is uniform is unlikely to apply to real data, and thus has little implication for real-world machine learning [17, 50].

## Consistency with Rényi-based Bounds

On the other hand, we have seen in Theorem (Sufficient condition for data length determined by Rényi entropy) that the sufficient condition about the training data length is almost of the order of the exponential of the Rényi entropy (recall Remark (The main term is $\nu_{(Q,\alpha)}$)), which can be much smaller than the data space cardinality.

## Consistency with Rényi-based Bounds

On the other hand, we have seen in Theorem (Sufficient condition for data length determined by Rényi entropy) that the sufficient condition about the training data length is almost of the order of the exponential of the Rényi entropy (recall Remark (The main term is $\nu_{(Q,\alpha)}$)), which can be much smaller than the data space cardinality.

It implies that if we know that Rényi entropy is small, then the original no-free-lunch theorem no longer holds since the uniform distribution is no longer allowed.

## Two Questions

Now, we have two questions.

## Two Questions

Now, we have two questions.

1. Is there a no-free-lunch theorem where the distribution is uneven, or its Rényi entropy has an upper limit?

## Two Questions

Now, we have two questions.

1. Is there a no-free-lunch theorem where the distribution is uneven, or its Rényi entropy has an upper limit?
2. If yes, is it consistent with Theorem (Sufficient condition for data length determined by Rényi entropy)?

## Two Questions

Now, we have two questions.

1. Is there a no-free-lunch theorem where the distribution is uneven, or its Rényi entropy has an upper limit?
2. If yes, is it consistent with Theorem (Sufficient condition for data length determined by Rényi entropy)?

The answers are yes for both.

## Rényi-Entropy Version of NFL

**Theorem (No-free-lunch theorem: the Rényi entropy version)**

*Consider the same setting as in Theorem (No-free-lunch theorem).*

## Rényi-Entropy Version of NFL

**Theorem (No-free-lunch theorem: the Rényi entropy version)**

*Consider the same setting as in Theorem (No-free-lunch theorem).*

*Then, for any map (learning algorithm)*

$$\mathfrak{A} : (\mathcal{X} \times \mathcal{Y})^* \to (\mathcal{X} \to \mathcal{Y}),$$

*any nonnegative integer (training data size)* $n$ *that satisfies* $n \leq n_0$*, any finite positive integer* $p$ *satisfying* $1 \leq p \leq |\mathcal{Y}|$*, and any* $\varepsilon \in (0, 1)$*, there exist a ground truth map* $f_0 : \mathcal{X} \to \mathcal{Y}$ *and a distribution* $Q$ *on* $\mathcal{X}$ *satisfying*

$$2 \exp\left(H_\alpha(Q)\right) \leq n_0 \quad \textit{for all } \alpha \in [0, 1]$$

*such that both the inequalities of Theorem (No-free-lunch theorem) hold.*

# Sketch of Proof Idea

**Proof idea (high level, no details).**

## Sketch of Proof Idea

**Proof idea (high level, no details).**

Consider $\mathcal{X}' \subset \mathcal{X}$ such that $|\underline{\mathcal{X}}| = 2n_0$. Then $Q = \mathrm{Uniform}(\mathcal{X}')$ satisfies $\exp(H_\alpha(Q)) = 2n_0$, and since $\frac{1}{2}|\mathcal{X}'| = 2n_0$, we obtain Theorem (No-free-lunch theorem: the Rényi entropy version) by applying Theorem (No-free-lunch theorem) with $\mathcal{X} = \underline{\mathcal{X}} = \mathcal{X}'$.

## Sketch of Proof Idea

**Proof idea (high level, no details).**

Consider $\mathcal{X}' \subset \mathcal{X}$ such that $|\underline{\mathcal{X}}| = 2n_0$. Then $Q = \mathrm{Uniform}(\mathcal{X}')$ satisfies $\exp(H_\alpha(Q)) = 2n_0$, and since $\frac{1}{2}|\mathcal{X}'| = 2n_0$, we obtain Theorem (No-free-lunch theorem: the Rényi entropy version) by applying Theorem (No-free-lunch theorem) with $\mathcal{X} = \underline{\mathcal{X}} = \mathcal{X}'$.

(Full proof is given in the appendix in the paper; omitted here.)

# Tightness of Our Bounds

**Remark**

Theorem (No-free-lunch theorem: the Rényi entropy version) essentially states that if an upper bound on $\exp(H_\alpha(Q))$ is given, learning will fail in the worst case if the training data length is not at least half of that upper bound.

# Tightness of Our Bounds

**Remark**

Theorem (No-free-lunch theorem: the Rényi entropy version) essentially states that if an upper bound on $\exp(H_\alpha(Q))$ is given, learning will fail in the worst case if the training data length is not at least half of that upper bound.

We remark that Theorem (Sufficient condition for data length determined by Rényi entropy) has stated that the sufficient condition with respect to the training data for a good generalisation was also almost proportional to $\exp(H_\alpha(Q))$, as stated in Remark (The main term is $\nu_{(Q,\alpha)}$).

## Tightness of Our Bounds

**Remark**

Theorem (No-free-lunch theorem: the Rényi entropy version) essentially states that if an upper bound on $\exp(H_\alpha(Q))$ is given, learning will fail in the worst case if the training data length is not at least half of that upper bound.

We remark that Theorem (Sufficient condition for data length determined by Rényi entropy) has stated that the sufficient condition with respect to the training data for a good generalisation was also almost proportional to $\exp(H_\alpha(Q))$, as stated in Remark (The main term is $\nu_{(Q,\alpha)}$).

In this sense, Theorem (Sufficient condition for data length determined by Rényi entropy) is tight with respect to the dependency on $H_\alpha(Q)$.

## Independence of $\alpha$

**Remark (On $\alpha$ in Theorem (No-free-lunch theorem: the Rényi entropy version))**

Note that the theorem statement itself does not depend on $\alpha$.

## Independence of $\alpha$

**Remark (On $\alpha$ in Theorem (No-free-lunch theorem: the Rényi entropy version))**

Note that the theorem statement itself does not depend on $\alpha$.

This is because the constructed worst case is a uniform distribution, and the Rényi entropy of a uniform distribution does not depend on the order $\alpha$.

# Conclusion

## Conclusion

- For a symmetric learning algorithm, there exists a probabilistic upper bound on the generalization gap determined by Rényi entropy (unevenness of the distribution), which does not depend on the specific construction or scale of the model.

## Conclusion

- For a symmetric learning algorithm, there exists a probabilistic upper bound on the generalization gap determined by Rényi entropy (unevenness of the distribution), which does not depend on the specific construction or scale of the model.
- It explains high generalization gap of settings with uniformly and independently distributed labels.

## Conclusion

- For a symmetric learning algorithm, there exists a probabilistic upper bound on the generalization gap determined by Rényi entropy (unevenness of the distribution), which does not depend on the specific construction or scale of the model.
- It explains high generalization gap of settings with uniformly and independently distributed labels.
- The novel no-free-lunch theorem shows that the data size implied by the above upper bound is in a sense necessary.

## Conclusion

- For a symmetric learning algorithm, there exists a probabilistic upper bound on the generalization gap determined by Rényi entropy (unevenness of the distribution), which does not depend on the specific construction or scale of the model.
- It explains high generalization gap of settings with uniformly and independently distributed labels.
- The novel no-free-lunch theorem shows that the data size implied by the above upper bound is in a sense necessary.

**Thank you for listening!**

## References i


[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al.
**GPT-4 technical report.**
arXiv preprint arXiv:2303.08774, 2023.

[2] Anthropic.
**The Claude 3 model family: Opus, Sonnet, Haiku, 2024.**

## References ii


**[3]** Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang.
**Stronger generalization bounds for deep nets via a compression approach.**
In International conference on machine learning, pages 254–263. PMLR, 2018.

[4] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al.
**Qwen technical report.**
arXiv preprint arXiv:2309.16609, 2023.

## References iii


[5] Peter L. Bartlett, Dylan J. Foster, and Matus J. Telgarsky.
**Spectrally-normalized margin bounds for neural networks.**
In Advances in Neural Information Processing Systems 30, pages 6240–6250, 2017.

[6] Peter L Bartlett and Shahar Mendelson.
**Rademacher and gaussian complexities: Risk bounds and structural results.**
Journal of Machine Learning Research, 3(Nov):463–482, 2002.

## References iv


[7] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal.
**Reconciling modern machine-learning practice and the classical bias–variance trade-off.**
Proceedings of the National Academy of Sciences, 116(32):15849–15854, 2019.

[8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al.
**Language models are few-shot learners.**
Advances in neural information processing systems, 33:1877–1901, 2020.

## References v


[9] Amit Daniely and Elad Granot.
**Generalization bounds for neural networks via approximate description length.**
In Advances in Neural Information Processing Systems 32, pages 11700–11710, 2019.

[10] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al.
**Palm-e: An embodied multimodal language model.**
In International Conference on Machine Learning, pages 8469–8488. PMLR, 2023.

## References vi


[11] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al.
**The llama 3 herd of models.**
arXiv preprint arXiv:2407.21783, 2024.

[12] Werner Ebeling and Alexander Neiman.
**Long-range correlations between letters and sentences in texts.**
Physica A: Statistical Mechanics and its Applications, 215(3):233–241, 1995.

[13] Werner Ebeling and Thorsten Pöschel.
**Entropy and long-range correlations in literary english.**
Europhysics Letters, 26(4):241, 1994.

[14] Roger Fletcher.
**A new approach to variable metric algorithms.**
The computer journal, 13(3):317–322, 1970.

[15] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al.
**Gemini: a family of highly capable multimodal models.**
arXiv preprint arXiv:2312.11805, 2023.

## References viii


[16] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al.
**Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context.**
arXiv preprint arXiv:2403.05530, 2024.

[17] Micah Goldblum, Marc Finzi, Keefer Rowan, and Andrew Gordon Wilson.
**Position: the no free lunch theorem, kolmogorov complexity, and the role of inductive biases in machine learning.**
In Proceedings of the 41st International Conference on Machine Learning. JMLR.org, 2024.

## References ix


[18] Donald Goldfarb.
**A family of variable-metric methods derived by variational means.**
Mathematics of computation, 24(109):23–26, 1970.

[19] Noah Golowich, Alexander Rakhlin, and Ohad Shamir.
**Size-independent sample complexity of neural networks.**
In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, Proceedings of the 31st Conference On Learning Theory, volume 75 of Proceedings of Machine Learning Research, pages 297–299. PMLR, 2018.

[20] Noah Golowich, Alexander Rakhlin, and Ohad Shamir.
**Size-independent sample complexity of neural networks.**
Information and Inference: A Journal of the IMA, 9(2):473–504, 2020.

## References x


[21] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al.
**Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.**
arXiv preprint arXiv:2501.12948, 2025.

[22] Nick Harvey, Christopher Liaw, and Abbas Mehrabian.
**Nearly-tight VC-dimension bounds for piecewise linear neural networks.**

In Satyen Kale and Ohad Shamir, editors, Proceedings of the 2017 Conference on Learning Theory, volume 65 of Proceedings of Machine Learning Research, pages 1064–1068. PMLR, 2017.

## References xi


[23] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al.
**Gpt-4o system card.**
arXiv preprint arXiv:2410.21276, 2024.

[24] Jiantao Jiao, Kartik Venkat, Yanjun Han, and Tsachy Weissman.
**Minimax estimation of functionals of discrete distributions.**
IEEE Transactions on Information Theory, 61(5):2835–2885, 2015.

[25] Diederik P Kingma.
**Adam: A method for stochastic optimization.**
arXiv preprint arXiv:1412.6980, 2014.

[26] Vladimir Koltchinskii and Dmitriy Panchenko.
**Rademacher processes and bounding the risk of function learning.**
In High dimensional probability II, pages 443–457. Springer, 2000.

[27] Vladimir Koltchinskii and Dmitry Panchenko.
**Empirical margin distributions and bounding the generalization error of combined classifiers.**
Annals of statistics, 30(1):1–50, 2002.

[28] Wentian Li.
**Mutual information functions of natural language texts.**
Technical Report 89-10-008, Santa Fe Institute, 1989.

## References xiii


[29] Xingguo Li, Junwei Lu, Zhaoran Wang, Jarvis Haupt, and Tuo Zhao.
**On tighter generalization bound for deep neural networks: CNNs, ResNets, and beyond.**
arXiv preprint arXiv:1806.05159, 2018.

[30] Henry Wanjune Lin and Max Erik Tegmark.
**Critical behavior in physics and probabilistic formal languages.**
Entropy, 19(7):299, June 2017.

[31] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al.
**Deepseek-v3 technical report.**
arXiv preprint arXiv:2412.19437, 2024.

## References xiv


[32] Sanae Lotfi, Marc Finzi, Sanyam Kapoor, Andres Potapczynski, Micah Goldblum, and Andrew G Wilson.
**Pac-bayes compression bounds so tight that they can explain generalization.**
Advances in Neural Information Processing Systems, 35:31459–31473, 2022.

[33] Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever.
**Deep double descent: Where bigger models and more data hurt.**
Journal of Statistical Mechanics: Theory and Experiment, 2021(12):124003, 2021.

[34] Yurii Nesterov.
**A method for solving the convex programming problem with convergence rate o (1/k2).**
In Dokl akad nauk Sssr, volume 269, page 543, 1983.

[35] Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro.
**Norm-based capacity control in neural networks.**
In Proceedings of The 28th Conference on Learning Theory, volume 40 of Proceedings of Machine Learning Research, pages 1376–1401. PMLR, 2015.

[36] Konstantinos E Nikolakakis, Farzin Haddadpour, Amin Karbasi, and Dionysios S Kalogerias.
**Beyond lipschitz: Sharp generalization and excess risk bounds for full-batch gd.**
arXiv preprint arXiv:2204.12446, 2022.

[37] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al.
**Language models are unsupervised multitask learners.**
OpenAI blog, 1(8):9, 2019.

[38] Timothy Sainburg, Brendan Theilman, Mark Thielk, and Timothy Q. Gentner.
**Parallels in the sequential organization of birdsong and human speech.**
Nature Communications, 10(1):3636, 2019.

[39] Shai Shalev-Shwartz and Shai Ben-David.
**Understanding machine learning: From theory to algorithms.**
Cambridge university press, 2014.

[40] David F Shanno.
**Conditioning of quasi-newton methods for function minimization.**
Mathematics of computation, 24(111):647–656, 1970.

## References xviii


[41] Xingwu Sun, Yanfeng Chen, Yiqing Huang, Ruobing Xie, Jiaqi Zhu, Kai Zhang, Shuaipeng Li, Zhen Yang, Jonny Han, Xiaobo Shu, et al.
**Hunyuan-large: An open-source moe model with 52 billion activated parameters by tencent.**
arXiv preprint arXiv:2411.02265, 2024.

[42] Atsushi Suzuki, Yulan He, Feng Tian, and Zhongyuan Wang.
**Hallucinations are inevitable but can be made statistically negligible. the "innate" inevitability of hallucinations cannot explain practical llm issues.**
arXiv preprint arXiv:2502.12187, 2025.

[43] Taiji Suzuki, Hiroshi Abe, Tomoya Murata, Shingo Horiuchi, Kotaro Ito, Tokuma Wachi, So Hirai, Masatoshi Yukishima, and Tomoaki Nishimura.
**Spectral pruning: Compressing deep neural networks via spectral analysis and its generalization error.**
arXiv preprint arXiv:1808.08558, 2018.

[44] Taiji Suzuki, Hiroshi Abe, and Tomoaki Nishimura.
**Compression based bound for non-compressed network: unified generalization error analysis of large compressible deep neural network.**
In International Conference on Learning Representations, 2020.

## References xx


[45] Shuntaro Takahashi and Kumiko Tanaka-Ishii.
**Do neural nets learn statistical laws behind natural language?**
PLoS ONE, 12(12):e0189326, 2017.

[46] Shuntaro Takahashi and Kumiko Tanaka-Ishii.
**Evaluating computational language models with scaling properties of natural language.**
Computational Linguistics, 45(3):481–513, 2019.

[47] Kumiko Tanaka-Ishii and Armin Bunde.
**Long-range memory in literary texts: On the universal clustering of the rare words.**
PLoS ONE, 11(11):e0164658, 2016.

[48] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al.
**Llama: Open and efficient foundation language models.**
arXiv preprint arXiv:2302.13971, 2023.

[49] Colin Wei and Tengyu Ma.
**Data-dependent sample complexity of deep neural networks via Lipschitz augmentation.**
In Advances in Neural Information Processing Systems 32, pages 9603–9613, 2019.

## References xxii


[50] Andrew Gordon Wilson.
**Deep learning is not so mysterious or different.**
arXiv preprint arXiv:2503.02113, 2025.

[51] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al.
**Qwen2 technical report.**
arXiv preprint arXiv:2407.10671, 2024.

[52] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al.
**Qwen2.5 technical report.**
arXiv preprint arXiv:2412.15115, 2024.

[53] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals.
**Understanding deep learning requires rethinking generalization.**
In International Conference on Learning Representations, 2017.

[54] Wenda Zhou, Victor Veitch, Morgane Austern, Ryan P Adams, and Peter Orbanz.
**Non-vacuous generalization bounds at the imagenet scale: a pac-bayesian compression approach.**
In International Conference on Learning Representations, 2019.

[55] George Kingsley Zipf.
**Human behavior and the principle of least effort: An introduction to human ecology.**
Addison-Wesley Press, Cambridge, MA, 1949.

## Conclusion (3): Future Directions

One interesting future work direction is to extend our framework to stochastic symmetric algorithms, including stochastic gradient descent method and its variants, as already discussed in Remark (Discussing stochastic symmetric algorithms is important future work).

## Conclusion (3): Future Directions

One interesting future work direction is to extend our framework to stochastic symmetric algorithms, including stochastic gradient descent method and its variants, as already discussed in Remark (Discussing stochastic symmetric algorithms is important future work).

Other future directions and limitations are discussed in Appendix.

## Conclusion (3): Future Directions

One interesting future work direction is to extend our framework to stochastic symmetric algorithms, including stochastic gradient descent method and its variants, as already discussed in Remark (Discussing stochastic symmetric algorithms is important future work).

Other future directions and limitations are discussed in Appendix.

While there is room for extension, the current version of our model-independent generalization gap bounds successfully justifies the use of even larger machine learning models in the future for real-world problems where the data distribution often deviates significantly from uniform.

# Limitations, discussions, and future work

## Limitations and Future Work Overview

We now discuss limitations, additional interpretations, and directions for future work.

## Limitations and Future Work Overview

We now discuss limitations, additional interpretations, and directions for future work.

These correspond to Section **??** in the paper.

## Limitation: Diverging Rényi Entropy

As stated in Remark (Case where Rényi entropy diverges), if Rényi entropy diverges, Theorems (Generalization Gap Bound by Rényi Entropy) and (Sufficient condition for data length determined by Rényi entropy) give vacuous upper bounds.

## Limitation: Diverging Rényi Entropy

As stated in Remark (Case where Rényi entropy diverges), if Rényi entropy diverges, Theorems (Generalization Gap Bound by Rényi Entropy) and (Sufficient condition for data length determined by Rényi entropy) give vacuous upper bounds.

However, as also stated in Remark (Case where Rényi entropy diverges), such cases are pathological, and since the no-free-lunch theorem discussed in Section **??** applies unconditionally, such cases are inherently unlearnable without additional assumptions.

## Can We Measure Rényi Entropy in Practice?

Can we explain why existing deep learning and other large-scale machine learning models are successful by measuring the Rényi entropy in the environments where they succeed, using the theorems of this research?

## Can We Measure Rényi Entropy in Practice?

Can we explain why existing deep learning and other large-scale machine learning models are successful by measuring the Rényi entropy in the environments where they succeed, using the theorems of this research?

The answer, unfortunately, is **no** in practical terms.

## Difficulty of Estimating Rényi Entropy

To reliably estimate the Rényi entropy of a probability distribution, a data size that overwhelmingly exceeds the number of elements in the data space is naturally required [24].

## Difficulty of Estimating Rényi Entropy

To reliably estimate the Rényi entropy of a probability distribution, a data size that overwhelmingly exceeds the number of elements in the data space is naturally required [24].

This is equivalent to or greater than the data size sufficient for the success of machine learning, as suggested by Remarks (Theorem (Generalization Gap Bound by Rényi Entropy) is model-independent) and (Theorem (Sufficient condition for data length determined by Rényi entropy) is also model-independent).

## Difficulty of Estimating Rényi Entropy

To reliably estimate the Rényi entropy of a probability distribution, a data size that overwhelmingly exceeds the number of elements in the data space is naturally required [24].

This is equivalent to or greater than the data size sufficient for the success of machine learning, as suggested by Remarks (Theorem (Generalization Gap Bound by Rényi Entropy) is model-independent) and (Theorem (Sufficient condition for data length determined by Rényi entropy) is also model-independent).

In other words, it is difficult to explain the success of actual large-scale machine learning models using the theory of this research.

## General Remark on Learning Theory

However, this is not a drawback specific to this research.

## General Remark on Learning Theory

However, this is not a drawback specific to this research.

There are many attempts to explain the success of deep learning by assuming the true hypothesis class, but in applications where large-scale machine learning models are successful, estimating the true hypothesis class is usually more difficult than the success of the machine learning model itself.

## General Remark on Learning Theory

However, this is not a drawback specific to this research.

There are many attempts to explain the success of deep learning by assuming the true hypothesis class, but in applications where large-scale machine learning models are successful, estimating the true hypothesis class is usually more difficult than the success of the machine learning model itself.

And, due to the existence of the no-free-lunch theorem, the success of machine learning cannot be explained without making assumptions about the true hypothesis class or the class of distributions.

## Role of This Theory

Due to these circumstances, in general, learning theories for large-scale machine learning models should be regarded not as explaining actual applications, but as showing one possible scenario for the future success of large-scale machine learning models.

## Role of This Theory

Due to these circumstances, in general, learning theories for large-scale machine learning models should be regarded not as explaining actual applications, but as showing one possible scenario for the future success of large-scale machine learning models.

This paper consists only of mathematical results, which is inevitable.

## Double Descent Phenomenon

The phenomenon known as double descent [7], where the generalization gap first increases and then decreases again as the scale of the machine learning model is increased, is known.

## Double Descent Phenomenon

The phenomenon known as double descent [7], where the generalization gap first increases and then decreases again as the scale of the machine learning model is increased, is known.

However, the magnitude of the effect of the double descent phenomenon is known to depend, for example, on the number of training epochs [33], and thus depends on the specific configuration of the learning algorithm.

## Double Descent Phenomenon

The phenomenon known as double descent [7], where the generalization gap first increases and then decreases again as the scale of the machine learning model is increased, is known.

However, the magnitude of the effect of the double descent phenomenon is known to depend, for example, on the number of training epochs [33], and thus depends on the specific configuration of the learning algorithm.

For this reason, the double descent phenomenon cannot be explained in principle within the framework of this paper.

## Double Descent and Our Setting

However, experimental results from double descent research also show that when the model scale becomes sufficiently large, the generalization gap is stable with respect to changes in model scale (entering the so-called modern regime).

## Double Descent and Our Setting

However, experimental results from double descent research also show that when the model scale becomes sufficiently large, the generalization gap is stable with respect to changes in model scale (entering the so-called modern regime).

Therefore, for the motivation of this paper, which is to understand the conditions for the success of extremely large models on large-scale data, double descent is not a direct problem.

## Consistency with Double Descent Experiments

The test error in the deteriorating part during double descent is also known experimentally to decrease with the number of data points in regions with a certain amount of data or more (e.g., Figure 11 in [33]).

## Consistency with Double Descent Experiments

The test error in the deteriorating part during double descent is also known experimentally to decrease with the number of data points in regions with a certain amount of data or more (e.g., Figure 11 in [33]).

Although the theory of this paper does not directly explain the double descent phenomenon, it is not inconsistent with related experimental results.

## Stronger Conclusions under Stronger Assumptions

As already stated in Remarks (Theorem (Generalization Gap Bound by Rényi Entropy) is model-independent) and (Theorem (Sufficient condition for data length determined by Rényi entropy) is also model-independent),
Theorems (Generalization Gap Bound by Rényi Entropy) and (Sufficient condition for data length determined by Rényi entropy) are model-independent.

## Stronger Conclusions under Stronger Assumptions

As already stated in Remarks (Theorem (Generalization Gap Bound by Rényi Entropy) is model-independent) and (Theorem (Sufficient condition for data length determined by Rényi entropy) is also model-independent),
Theorems (Generalization Gap Bound by Rényi Entropy) and (Sufficient condition for data length determined by Rényi entropy) are model-independent.

The absence of assumptions about the model is an advantage in terms of wide applicability.

## Stronger Conclusions under Stronger Assumptions

As already stated in Remarks (Theorem (Generalization Gap Bound by Rényi Entropy) is model-independent) and (Theorem (Sufficient condition for data length determined by Rényi entropy) is also model-independent),
Theorems (Generalization Gap Bound by Rényi Entropy) and (Sufficient condition for data length determined by Rényi entropy) are model-independent.

The absence of assumptions about the model is an advantage in terms of wide applicability.

On the other hand, as a general principle of theoretical analysis, the fewer assumptions a theorem has, the weaker its conclusion.

## Two Types of Learning Theory

As stated in the previous section, it is impossible to know the appropriate class containing the true hypothesis or the appropriate class containing the true distribution in actual applications, and it is also impossible to know the appropriate class of models corresponding to them.

## Two Types of Learning Theory

As stated in the previous section, it is impossible to know the appropriate class containing the true hypothesis or the appropriate class containing the true distribution in actual applications, and it is also impossible to know the appropriate class of models corresponding to them.

Therefore, both:

- creating theories with wide applicability at the cost of weaker conclusions, and
- creating theories that provide strong conclusions at the risk of not being theoretically applicable to actual applications

are important, and it is not the case that only one is important.

## Positioning of This Work

This research belongs to the former category in the sense that it makes no assumptions about the model.

## Positioning of This Work

This research belongs to the former category in the sense that it makes no assumptions about the model.

On the other hand, the direction of trying to obtain stronger conclusions by also placing some assumptions on the smoothness as a function of the model or its information-theoretic complexity is an interesting avenue for future work.