# Introduction to Machine Learning

**Changlong Wu & Wojciech Szpankowski**

Center for Science of Information
Purdue University

October 18, 2024

# Overview

- **What is Machine Learning?**
  - Inductive inference, generalizability, relation to other fields
- **Basic Concepts in Learning Theory**
  - Hypothesis space, concept classes
  - Realizable vs. agnostic learning
- **The Online Learning Framework**
  - Learning from expert advice
  - The halving algorithm
  - Exponentially Weighted Average algorithm

**Recommend textbooks**:
1. "*Understanding Machine Learning: From Theory to Algorithms*", by S. Shalev-Shwartz and S. Ben-David
2. "*Prediction, Learning, and Games*", by N. Cesa-Bianchi and G. Lugosi

# What is Machine Learning?

*"Machine Learning is the process of programming computers to automatically convert experience (training data) into expertise or knowledge (a model) that can perform tasks with broader generalization."*

- **Core objective**: Build models that *generalize* from a limited dataset to unseen data
  - A successful learner should be able to predict on new examples
- **Generalizability** is the key feature that *distinguishes* a learning system from one that simply memorizes the training data
  - Learning should be able to extract common patterns from the data
- The **core problem** of machine learning is to understand when *generalization* is possible and how to achieve it in an automatic and efficient way
  - This automatic procedure is referred to as learning rules

# Types of Learning Paradigms

Depending on how the data are *generated* and how one *leverages* the learned model, learning can be *roughly* classified into the following categories:

- **Supervised vs. Unsupervised**: *Supervised* learning uses training data with human annotation (such as labels) that is missing in test data, while *unsupervised* learning makes no distinction between training and test data
- **Passive vs. Active**: *Passive* learning simply observes data provided by the environment, while *active* learning interacts with the environment to acquire specific information to improve learning
- **Online vs. Batch**: In *online* learning, the learner *makes decisions* and *updates* model continuously with new data, whereas in *batch* learning, it processes all data at once before applying the acquired expertise

These paradigms are not mutually exclusive and can interact in complex ways.

# Machine Learning vs. traditional Statistics?

- **Assumptions on Data Models**:
  - The primary goal of machine learning is to make predictions on unseen data, with *minimal* assumptions on the *ground truth* data generation mechanism
  - Statistics primarily focuses on inferences (parameters, properties, etc.) of certain prescribed data models, such as the *Gaussian* distribution
- **Modeling of Hypotheses**:
  - Machine learning typically uses complex models, such as neural networks, to capture patterns in data
  - Statistics focuses on simpler models, such as linear regression
- **Algorithmic Consideration**:
  - Machine learning focuses heavily on computational efficiency and often optimizes models on large datasets
  - Statistics tends to prioritize analytical solutions, relying on simple data models where computational complexity is typically less emphasized

# Overview

- **What is Machine Learning?**
  - Inductive inference, generalizability, relation to other fields
- **Basic Concepts in Learning Theory**
  - Hypothesis space, concept classes
  - Realizable vs. agnostic learning
- **The Online Learning Framework**
  - Learning from expert advice
  - The halving algorithm
  - Exponentially Weighted Average algorithm

# Basic Concepts in Learning Theory

Let $\mathcal{X}$ be an instance space (or feature space), and $\mathcal{Y}$ be a label space (or outcome space). A prediction rule (or model) is defined as a function

$$h : \mathcal{X} \to \mathcal{Y}.$$

We denote $\mathcal{Y}^{\mathcal{X}}$ as the class of all predictors from $\mathcal{X} \to \mathcal{Y}$.

- A learning rule is a function

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

  which takes a training set as input and outputs a predictor from $\mathcal{X} \to \mathcal{Y}$.
- A hypothesis class $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$ is a set of predictors that the learning rule $\Phi$ explores during training.
  - E.g., a class of functions represented by a neural network architecture correspond to different weights.
- A concept class $\mathcal{C} \subset \mathcal{Y}^{\mathcal{X}}$ is the set of all possible target predictors that describe the true relationships in the data.
  - Typically depends on learner's prior knowledge on the learning target.

# Basic Concepts in Learning Theory

Let $\mathcal{C}$ be a concept class and $\mathcal{H}$ be a hypothesis class for a particular learning problem.

- We say the problem is **realizable** if $\mathcal{C} \subset \mathcal{H}$, i.e., every target predictor must be within the hypothesis class
- The problem is **agnostic** if the concept class $\mathcal{C}$ is completely unconstrained, in other words, we take $\mathcal{C} := \mathcal{Y}^{\mathcal{X}}$
  - Note that, there can also be intermediate scenarios between the *realizable* and *agnostic* learning paradigms
- We will only consider the *realizable* vs. *agnostic* dichotomy in our entire lectures, so that we do not explicitly refer to the concept class
  - Therefore, our following discussions will focus only on the hypothesis classes
- We will also sometimes relax the outputs of the learner $\Phi$ to be outside of the hypothesis class $\mathcal{H}$, a scenario called improper learning
  - We refer the case when the outputs of $\Phi$ are restricted to $\mathcal{H}$ as proper learning

# Overview

- **What is Machine Learning?**
  - Inductive inference, generalizability, relation to other fields
- **Basic Concepts in Learning Theory**
  - Hypothesis space, concept classes
  - Realizable vs. agnostic learning
- **The Online Learning Framework**
  - Learning from expert advice
  - The halving algorithm
  - Exponentially Weighted Average algorithm

# The Online Learning Game

For $t = 1, 2, \cdots, T$

1. Nature/Environment presents an instance $\mathbf{x}_t \in \mathcal{X}$
2. Learner predicts a label $\hat{y}_t \in \mathcal{Y}$
3. Nature reveals true label $y_t \in \mathcal{Y}$
4. Learner suffers loss $\ell(\hat{y}_t, y_t)$, for certain function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$

**Goal**: Finding a learning rule $\Phi$ that minimizes the risk

$$\text{risk}_T(\Phi) := \sum_{t=1}^{T} \ell(\hat{y}_t, y_t)$$

# Cover's Impossibility Result

Take $\mathcal{Y} := \{0, 1\}$ and let $\ell(\hat{y}, y) := 1\{\hat{y} \neq y\}$. Then, $\text{risk}_t(\Phi)$ reduces to the number of mistakes made by $\Phi$ in predicting the $y_t$'s.

Let $\Phi$ be any learning rule. Consider the following simple strategy for Nature:

- ⋆ At each time step $t$, after the learner makes the prediction $\hat{y}_t$, Nature *adversarially* chooses $y_t \in \mathcal{Y}$ such that $y_t \neq \hat{y}_t$.

The number of mistakes made by the learner equals $T$, i.e., the learner errs at every step. (This fact is attributed to T. M. Cover in a 1965 paper.)

# Cover's Impossibility Result

Take $\mathcal{Y} := \{0, 1\}$ and let $\ell(\hat{y}, y) := 1\{\hat{y} \neq y\}$. Then, $\text{risk}_t(\Phi)$ reduces to the number of mistakes made by $\Phi$ in predicting the $y_t$'s.

Let $\Phi$ be any learning rule. Consider the following simple strategy for Nature:

- ⋆ At each time step $t$, after the learner makes the prediction $\hat{y}_t$, Nature *adversarially* chooses $y_t \in \mathcal{Y}$ such that $y_t \neq \hat{y}_t$.

The number of mistakes made by the learner equals $T$, i.e., the learner errs at every step. (This fact is attributed to T. M. Cover in a 1965 paper.)

**Corollary**: Any learning rule $\Phi$ cannot achieve a mistake bound better than $T$.

# Cover's Impossibility Result

Take $\mathcal{Y} := \{0, 1\}$ and let $\ell(\hat{y}, y) := 1\{\hat{y} \neq y\}$. Then, $\text{risk}_t(\Phi)$ reduces to the number of mistakes made by $\Phi$ in predicting the $y_t$'s.

Let $\Phi$ be any learning rule. Consider the following simple strategy for Nature:

- ⋆ At each time step $t$, after the learner makes the prediction $\hat{y}_t$, Nature *adversarially* chooses $y_t \in \mathcal{Y}$ such that $y_t \neq \hat{y}_t$.

The number of mistakes made by the learner equals $T$, i.e., the learner errs at every step. (This fact is attributed to T. M. Cover in a 1965 paper.)

**Corollary**: Any learning rule $\Phi$ cannot achieve a mistake bound better than $T$.

**What's the catch?**

## Cover's Impossibility Result

Take $\mathcal{Y} := \{0, 1\}$ and let $\ell(\hat{y}, y) := 1\{\hat{y} \neq y\}$. Then, $\text{risk}_t(\Phi)$ reduces to the number of mistakes made by $\Phi$ in predicting the $y_t$'s.

Let $\Phi$ be any learning rule. Consider the following simple strategy for Nature:

- ⋆ At each time step $t$, after the learner makes the prediction $\hat{y}_t$, Nature *adversarially* chooses $y_t \in \mathcal{Y}$ such that $y_t \neq \hat{y}_t$.

The number of mistakes made by the learner equals $T$, i.e., the learner errs at every step. (This fact is attributed to T. M. Cover in a 1965 paper.)

**Corollary**: Any learning rule $\Phi$ cannot achieve a mistake bound better than $T$.

**What's the catch?** No prior knowledge about the learning target was used!

## Incorporating Prior Knowledge: Realizable Case

Let $\mathcal{H} := \{h_1, \cdots, h_K\} \subset \mathcal{Y}^{\mathcal{X}}$ be a hypothesis class, and assume that Nature's strategy is realizable, i.e., there exists an $h \in \mathcal{H}$ such that

$$\text{For all } t \leq T,\ h(\mathbf{x}_t) = y_t.$$

# Incorporating Prior Knowledge: Realizable Case

Let $\mathcal{H} := \{h_1, \cdots, h_K\} \subset \mathcal{Y}^{\mathcal{X}}$ be a hypothesis class, and assume that Nature's strategy is realizable, i.e., there exists an $h \in \mathcal{H}$ such that

$$\text{For all } t \leq T, \; h(\mathbf{x}_t) = y_t.$$

**The Consistent Predictor**:

1. At each time step $t$, find any consistent hypothesis $\hat{h}_t \in \mathcal{H}$ (which must exist due to realizability) such that:

$$\sum_{i=1}^{t-1} 1\{\hat{h}_t(\mathbf{x}_i) \neq y_i\} = 0.$$

2. Make the prediction: $\hat{y}_t = \hat{h}_t(\mathbf{x}_t)$.

# Incorporating Prior Knowledge: Realizable Case

Let $\mathcal{H} := \{h_1, \cdots, h_K\} \subset \mathcal{Y}^{\mathcal{X}}$ be a hypothesis class, and assume that Nature's strategy is realizable, i.e., there exists an $h \in \mathcal{H}$ such that

$$\text{For all } t \leq T, \; h(\mathbf{x}_t) = y_t.$$

**The Consistent Predictor**:

1. At each time step $t$, find any consistent hypothesis $\hat{h}_t \in \mathcal{H}$ (which must exist due to realizability) such that:

$$\sum_{i=1}^{t-1} 1\{\hat{h}_t(\mathbf{x}_i) \neq y_i\} = 0.$$

2. Make the prediction: $\hat{y}_t = \hat{h}_t(\mathbf{x}_t)$.

**How many mistakes will we make?**

# Incorporating Prior Knowledge: Realizable Case

Let $\mathcal{H} := \{h_1, \cdots, h_K\} \subset \mathcal{Y}^{\mathcal{X}}$ be a hypothesis class, and assume that Nature's strategy is realizable, i.e., there exists an $h \in \mathcal{H}$ such that

$$\text{For all } t \leq T, \; h(\mathbf{x}_t) = y_t.$$

**The Consistent Predictor**:

1. At each time step $t$, find any consistent hypothesis $\hat{h}_t \in \mathcal{H}$ (which must exist due to realizability) such that:

$$\sum_{i=1}^{t-1} 1\{\hat{h}_t(\mathbf{x}_i) \neq y_i\} = 0.$$

2. Make the prediction: $\hat{y}_t = \hat{h}_t(\mathbf{x}_t)$.

**How many mistakes will we make?** Each mistake will eliminate at least one function from $\mathcal{H}$, so the total number of mistakes is upper bounded by $|\mathcal{H}|$...

# Incorporating Prior Knowledge: Realizable Case

Let $\mathcal{H} := \{h_1, \cdots, h_K\} \subset \mathcal{Y}^{\mathcal{X}}$ be a hypothesis class, and assume that Nature's strategy is realizable, i.e., there exists an $h \in \mathcal{H}$ such that

$$\text{For all } t \leq T,\ h(\mathbf{x}_t) = y_t.$$

**The Consistent Predictor**:

1. At each time step $t$, find any consistent hypothesis $\hat{h}_t \in \mathcal{H}$ (which must exist due to realizability) such that:

$$\sum_{i=1}^{t-1} 1\{\hat{h}_t(\mathbf{x}_i) \neq y_i\} = 0.$$

2. Make the prediction: $\hat{y}_t = \hat{h}_t(\mathbf{x}_t)$.

**How many mistakes will we make?** Each mistake will eliminate at least one function from $\mathcal{H}$, so the total number of mistakes is upper bounded by $|\mathcal{H}|$...

In fact, the consistent predictor cannot do better than $|\mathcal{H}|$ mistakes.

# Proving the $|\mathcal{H}|$ Lower Bound

Consider the following hypothesis class:

| | $\mathbf{x}_1$ | $\mathbf{x}_2$ | $\mathbf{x}_3$ | $\mathbf{x}_4$ | $\cdots$ |
|---|---|---|---|---|---|
| $h_0$ | 0 | 0 | 0 | 0 | $\cdots$ |
| $h_1$ | 1 | 0 | 0 | 0 | $\cdots$ |
| $h_2$ | 0 | 1 | 0 | 0 | $\cdots$ |
| $h_3$ | 0 | 0 | 1 | 0 | $\cdots$ |
| $h_4$ | 0 | 0 | 0 | 1 | $\cdots$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\ddots$ |

# Proving the $|\mathcal{H}|$ Lower Bound

Consider the following hypothesis class:

| | $\mathbf{x}_1$ | $\mathbf{x}_2$ | $\mathbf{x}_3$ | $\mathbf{x}_4$ | $\cdots$ |
|---|---|---|---|---|---|
| $h_0$ | 0 | 0 | 0 | 0 | $\cdots$ |
| $h_1$ | 1 | 0 | 0 | 0 | $\cdots$ |
| $h_2$ | 0 | 1 | 0 | 0 | $\cdots$ |
| $h_3$ | 0 | 0 | 1 | 0 | $\cdots$ |
| $h_4$ | 0 | 0 | 0 | 1 | $\cdots$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\ddots$ |

- Assume that $h_0$ is the ground truth predictor.

# Proving the $|\mathcal{H}|$ Lower Bound

Consider the following hypothesis class:

| | $\mathbf{x}_1$ | $\mathbf{x}_2$ | $\mathbf{x}_3$ | $\mathbf{x}_4$ | $\cdots$ |
|---|---|---|---|---|---|
| $h_0$ | 0 | 0 | 0 | 0 | $\cdots$ |
| $h_1$ | 1 | 0 | 0 | 0 | $\cdots$ |
| $h_2$ | 0 | 1 | 0 | 0 | $\cdots$ |
| $h_3$ | 0 | 0 | 1 | 0 | $\cdots$ |
| $h_4$ | 0 | 0 | 0 | 1 | $\cdots$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\ddots$ |

- Assume that $h_0$ is the ground truth predictor.
- At each time step $t$, both $h_t$ and $h_0$ are consistent with the prior data.

# Proving the $|\mathcal{H}|$ Lower Bound

Consider the following hypothesis class:

| | $\mathbf{x}_1$ | $\mathbf{x}_2$ | $\mathbf{x}_3$ | $\mathbf{x}_4$ | $\cdots$ |
|---|---|---|---|---|---|
| $h_0$ | 0 | 0 | 0 | 0 | $\cdots$ |
| $h_1$ | 1 | 0 | 0 | 0 | $\cdots$ |
| $h_2$ | 0 | 1 | 0 | 0 | $\cdots$ |
| $h_3$ | 0 | 0 | 1 | 0 | $\cdots$ |
| $h_4$ | 0 | 0 | 0 | 1 | $\cdots$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\ddots$ |

- Assume that $h_0$ is the ground truth predictor.
- At each time step $t$, both $h_t$ and $h_0$ are consistent with the prior data.
- Consider a consistent predictor that always selects $h_t$ to make predictions at step $t$, which will incur at least $|\mathcal{H}|$ mistakes.

# Proving the $|\mathcal{H}|$ Lower Bound

Consider the following hypothesis class:

| | $\mathbf{x}_1$ | $\mathbf{x}_2$ | $\mathbf{x}_3$ | $\mathbf{x}_4$ | $\cdots$ |
|---|---|---|---|---|---|
| $h_0$ | 0 | 0 | 0 | 0 | $\cdots$ |
| $h_1$ | 1 | 0 | 0 | 0 | $\cdots$ |
| $h_2$ | 0 | 1 | 0 | 0 | $\cdots$ |
| $h_3$ | 0 | 0 | 1 | 0 | $\cdots$ |
| $h_4$ | 0 | 0 | 0 | 1 | $\cdots$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\ddots$ |

- Assume that $h_0$ is the ground truth predictor.
- At each time step $t$, both $h_t$ and $h_0$ are consistent with the prior data.
- Consider a consistent predictor that always selects $h_t$ to make predictions at step $t$, which will incur at least $|\mathcal{H}|$ mistakes.

**Corollary**: In the worst-case scenario, a consistent predictor cannot achieve a mistake bound better than $|\mathcal{H}|$.

# Incorporating Prior Knowledge: Realizable Case

How can we go beyond the $|\mathcal{H}|$ mistake bound barrier?

# Incorporating Prior Knowledge: Realizable Case

How can we go beyond the $|\mathcal{H}|$ mistake bound barrier?

- ⋆ Find a smarter way to eliminate the hypothesis in $\mathcal{H}$...

# Incorporating Prior Knowledge: Realizable Case

How can we go beyond the $|\mathcal{H}|$ mistake bound barrier?

- ⋆ Find a smarter way to eliminate the hypothesis in $\mathcal{H}$...

**The halving predictor**:

1. Maintain a running hypothesis class $\mathcal{H}^{(t)}$ with $\mathcal{H}^{(0)} := \mathcal{H}$
2. At each time step $t$, we define for $y \in \{0, 1\}$

$$\mathcal{H}_y^{(t)} = \{h \in \mathcal{H}^{(t-1)} : h(\mathbf{x}_t) = y\}.$$

3. Predict $\hat{y}_t = \arg\max_{y \in \{0,1\}} \{|\mathcal{H}_0^{(t)}|, |\mathcal{H}_1^{(t)}|\}$
4. Let $y_t$ be true label, update $\mathcal{H}^{(t)} = \mathcal{H}_{y_t}^{(t)}$

# Incorporating Prior Knowledge: Realizable Case

How can we go beyond the $|\mathcal{H}|$ mistake bound barrier?

- ⋆ Find a smarter way to eliminate the hypothesis in $\mathcal{H}$...

**The halving predictor**:

1. Maintain a running hypothesis class $\mathcal{H}^{(t)}$ with $\mathcal{H}^{(0)} := \mathcal{H}$
2. At each time step $t$, we define for $y \in \{0, 1\}$

$$\mathcal{H}_y^{(t)} = \{h \in \mathcal{H}^{(t-1)} : h(\mathbf{x}_t) = y\}.$$

3. Predict $\hat{y}_t = \arg\max_{y \in \{0,1\}} \{|\mathcal{H}_0^{(t)}|, |\mathcal{H}_1^{(t)}|\}$
4. Let $y_t$ be true label, update $\mathcal{H}^{(t)} = \mathcal{H}_{y_t}^{(t)}$

**How many mistakes do we make?**

- ✓ Every time a mistake happen (i.e., $\hat{y}_t \neq y_t$), we have $|\mathcal{H}^{(t)}| \leq |\mathcal{H}^{(t-1)}|/2$
- ✓ Total number of mistakes upper bounded by $\log |\mathcal{H}|$ (an exponential improvement over the $|\mathcal{H}|$ bound!)

# Incorporating Prior Knowledge: Agnostic Case

Both the consistent and halving predictors rely heavily on the assumption that the data is realizable, i.e., there exists $h \in \mathcal{H}$ that is consistent with all the data...

# Incorporating Prior Knowledge: Agnostic Case

Both the consistent and halving predictors rely heavily on the assumption that the data is realizable, i.e., there exists $h \in \mathcal{H}$ that is consistent with all the data...

A single mismatch between the true data and the best hypothesis in $\mathcal{H}$ will cause both predictors to catastrophically fail (prove it!).

# Incorporating Prior Knowledge: Agnostic Case

Both the consistent and halving predictors rely heavily on the assumption that the data is realizable, i.e., there exists $h \in \mathcal{H}$ that is consistent with all the data...

A single mismatch between the true data and the best hypothesis in $\mathcal{H}$ will cause both predictors to catastrophically fail (prove it!).

**Can we develop an algorithm that is robust to potential *noise*?**

# Incorporating Prior Knowledge: Agnostic Case

Both the consistent and halving predictors rely heavily on the assumption that the data is realizable, i.e., there exists $h \in \mathcal{H}$ that is consistent with all the data...

A single mismatch between the true data and the best hypothesis in $\mathcal{H}$ will cause both predictors to catastrophically fail (prove it!).

**Can we develop an algorithm that is robust to potential *noise*?**

- Clearly, an absolute mistake bound is not very informative.

# Incorporating Prior Knowledge: Agnostic Case

Both the consistent and halving predictors rely heavily on the assumption that the data is realizable, i.e., there exists $h \in \mathcal{H}$ that is consistent with all the data...

A single mismatch between the true data and the best hypothesis in $\mathcal{H}$ will cause both predictors to catastrophically fail (prove it!).

**Can we develop an algorithm that is robust to potential *noise*?**

- Clearly, an absolute mistake bound is not very informative.
- Instead, we consider guarantees relative to the minimal mistakes achievable by a hypothesis in $\mathcal{H}$.

# Incorporating Prior Knowledge: Agnostic Case

Both the consistent and halving predictors rely heavily on the assumption that the data is realizable, i.e., there exists $h \in \mathcal{H}$ that is consistent with all the data...

A single mismatch between the true data and the best hypothesis in $\mathcal{H}$ will cause both predictors to catastrophically fail (prove it!).

**Can we develop an algorithm that is robust to potential *noise*?**

- Clearly, an absolute mistake bound is not very informative.
- Instead, we consider guarantees relative to the minimal mistakes achievable by a hypothesis in $\mathcal{H}$.
- Let $\widehat{M}_T := \sum_{t=1}^{T} 1\{\hat{y}_t \neq y_t\}$ be the number of mistakes made by a predictor $\Phi$,

# Incorporating Prior Knowledge: Agnostic Case

Both the consistent and halving predictors rely heavily on the assumption that the data is realizable, i.e., there exists $h \in \mathcal{H}$ that is consistent with all the data...

A single mismatch between the true data and the best hypothesis in $\mathcal{H}$ will cause both predictors to catastrophically fail (prove it!).

**Can we develop an algorithm that is robust to potential *noise*?**

- Clearly, an absolute mistake bound is not very informative.
- Instead, we consider guarantees relative to the minimal mistakes achievable by a hypothesis in $\mathcal{H}$.
- Let $\widehat{M}_T := \sum_{t=1}^{T} 1\{\hat{y}_t \neq y_t\}$ be the number of mistakes made by a predictor $\Phi$, and $M_T^* := \inf_{h \in \mathcal{H}} \sum_{t=1}^{T} 1\{h(\mathbf{x}_t) \neq y_t\}$ be the minimal number of mistakes achievable by any hypothesis in $\mathcal{H}$.

# Incorporating Prior Knowledge: Agnostic Case

Both the consistent and halving predictors rely heavily on the assumption that the data is realizable, i.e., there exists $h \in \mathcal{H}$ that is consistent with all the data...

A single mismatch between the true data and the best hypothesis in $\mathcal{H}$ will cause both predictors to catastrophically fail (prove it!).

**Can we develop an algorithm that is robust to potential *noise*?**

- Clearly, an absolute mistake bound is not very informative.
- Instead, we consider guarantees relative to the minimal mistakes achievable by a hypothesis in $\mathcal{H}$.
- Let $\widehat{M}_T := \sum_{t=1}^{T} 1\{\hat{y}_t \neq y_t\}$ be the number of mistakes made by a predictor $\Phi$, and $M_T^* := \inf_{h \in \mathcal{H}} \sum_{t=1}^{T} 1\{h(\mathbf{x}_t) \neq y_t\}$ be the minimal number of mistakes achievable by any hypothesis in $\mathcal{H}$.
- We define the $\alpha$-agnostic regret as (for $\alpha > 0$):

$$\text{reg}_T^{\alpha}(\Phi, \mathcal{H}) := \widehat{M}_T - \alpha M_T^*.$$

# The Exponential Weighted Average Algorithm

Let $\mathcal{H} := \{h_1, \cdots, h_K\}$ be any finite hypothesis class of size $K$.

**The (deterministic) Exponential Weighted Average (EWA) Algorithm**:

1. Maintain a weight vector $\mathbf{w}^{(t)} \in \mathbb{R}^K$, initially $\mathbf{w}^{(0)} = (1, \cdots, 1)$.
2. At each step $t$, compute the weighted average:

$$\hat{p}_t = \sum_{k=1}^{K} \frac{\mathbf{w}_k^{(t-1)}}{\sum_{k=1}^{K} \mathbf{w}_k^{(t-1)}} h_k(\mathbf{x}_t).$$

3. Make prediction $\hat{y}_t = 1\{\hat{p}_t \geq \frac{1}{2}\}$, i.e., we predict the weighted-majority.
4. Update $\mathbf{w}_k^{(t)} = \mathbf{w}_k^{(t-1)}$ if $h_k(\mathbf{x}_t) = y_t$; and $\mathbf{w}_k^{(t)} = (1-\eta)\mathbf{w}_k^{(t-1)}$ if $h_k(\mathbf{x}_t) \neq y_t$, where $\eta \leq 1$ is a tunable parameter.

# The Exponential Weighted Average Algorithm

Let $\mathcal{H} := \{h_1, \cdots, h_K\}$ be any finite hypothesis class of size $K$.

**The (deterministic) Exponential Weighted Average (EWA) Algorithm**:

1. Maintain a weight vector $\mathbf{w}^{(t)} \in \mathbb{R}^K$, initially $\mathbf{w}^{(0)} = (1, \cdots, 1)$.
2. At each step $t$, compute the weighted average:

$$\hat{p}_t = \sum_{k=1}^{K} \frac{\mathbf{w}_k^{(t-1)}}{\sum_{k=1}^{K} \mathbf{w}_k^{(t-1)}} h_k(\mathbf{x}_t).$$

3. Make prediction $\hat{y}_t = 1\{\hat{p}_t \geq \frac{1}{2}\}$, i.e., we predict the weighted-majority.
4. Update $\mathbf{w}_k^{(t)} = \mathbf{w}_k^{(t-1)}$ if $h_k(\mathbf{x}_t) = y_t$; and $\mathbf{w}_k^{(t)} = (1-\eta)\mathbf{w}_k^{(t-1)}$ if $h_k(\mathbf{x}_t) \neq y_t$, where $\eta \leq 1$ is a tunable parameter.

**Theorem 1**: Regardless of how Nature generates the data, the (deterministic) EWA algorithm $\Phi$ enjoys the mistake bound:

$$\widehat{M}_T \leq 2(1+\eta)M_T^* + \frac{2\ln(|\mathcal{H}|)}{\eta}$$

# The Exponential Weighted Average Algorithm

Let $\mathcal{H} := \{h_1, \cdots, h_K\}$ be any finite hypothesis class of size $K$.

**The (deterministic) Exponential Weighted Average (EWA) Algorithm**:

1. Maintain a weight vector $\mathbf{w}^{(t)} \in \mathbb{R}^K$, initially $\mathbf{w}^{(0)} = (1, \cdots, 1)$.
2. At each step $t$, compute the weighted average:

$$\hat{p}_t = \sum_{k=1}^{K} \frac{\mathbf{w}_k^{(t-1)}}{\sum_{k=1}^{K} \mathbf{w}_k^{(t-1)}} h_k(\mathbf{x}_t).$$

3. Make prediction $\hat{y}_t = 1\{\hat{p}_t \geq \frac{1}{2}\}$, i.e., we predict the weighted-majority.
4. Update $\mathbf{w}_k^{(t)} = \mathbf{w}_k^{(t-1)}$ if $h_k(\mathbf{x}_t) = y_t$; and $\mathbf{w}_k^{(t)} = (1-\eta)\mathbf{w}_k^{(t-1)}$ if $h_k(\mathbf{x}_t) \neq y_t$, where $\eta \leq 1$ is a tunable parameter.

**Theorem 1**: Regardless of how Nature generates the data, the (deterministic) EWA algorithm $\Phi$ enjoys the mistake bound:

$$\widehat{M}_T \leq 2(1+\eta)M_T^* + \frac{2\ln(|\mathcal{H}|)}{\eta} \Rightarrow \text{reg}_T^2(\Phi, \mathcal{H}) \leq O(\sqrt{M_T^* \log|\mathcal{H}|}).$$

# Proof of the Regret Bound

For any $t$, we define the potential $W^{(t)} = \sum_{k=1}^{K} \mathbf{w}_k^{(t)}$ with $W^{(0)} = K$.

# Proof of the Regret Bound

For any $t$, we define the potential $W^{(t)} = \sum_{k=1}^{K} \mathbf{w}_k^{(t)}$ with $W^{(0)} = K$.

For any time step $t$, we denote $I_t := \{k \in [K] : h_k(\mathbf{x}_t) = y_t\}$ and $J_t := [K] \backslash I_t$.

# Proof of the Regret Bound

For any $t$, we define the potential $W^{(t)} = \sum_{k=1}^{K} \mathbf{w}_k^{(t)}$ with $W^{(0)} = K$.

For any time step $t$, we denote $I_t := \{k \in [K] : h_k(\mathbf{x}_t) = y_t\}$ and $J_t := [K] \backslash I_t$.

If $\hat{y}_t \neq y_t$, then $\sum_{k \in J_t} \mathbf{w}_k^{(t-1)} \geq \sum_{k \in I_t} \mathbf{w}_k^{(t-1)}$ due to the weighted-majority.

# Proof of the Regret Bound

For any $t$, we define the potential $W^{(t)} = \sum_{k=1}^{K} \mathbf{w}_k^{(t)}$ with $W^{(0)} = K$.

For any time step $t$, we denote $I_t := \{k \in [K] : h_k(\mathbf{x}_t) = y_t\}$ and $J_t := [K] \setminus I_t$.

If $\hat{y}_t \neq y_t$, then $\sum_{k \in J_t} \mathbf{w}_k^{(t-1)} \geq \sum_{k \in I_t} \mathbf{w}_k^{(t-1)}$ due to the weighted-majority.

Therefore, for any step $t$ where a mistake occurs, we have:

$$W^{(t)} = \sum_{k=1}^{K} w_k^{(t)} = (1-\eta) \underbrace{\sum_{k \in J_t} \mathbf{w}_k^{(t-1)}}_{A} + \underbrace{\sum_{k \in I_t} \mathbf{w}_k^{(t-1)}}_{B} \overset{(\bigstar)}{\leq} \left( \frac{1-\eta}{2} + \frac{1}{2} \right) W^{(t-1)},$$

## Proof of the Regret Bound

For any $t$, we define the potential $W^{(t)} = \sum_{k=1}^{K} \mathbf{w}_k^{(t)}$ with $W^{(0)} = K$.

For any time step $t$, we denote $I_t := \{k \in [K] : h_k(\mathbf{x}_t) = y_t\}$ and $J_t := [K] \backslash I_t$.

If $\hat{y}_t \neq y_t$, then $\sum_{k \in J_t} \mathbf{w}_k^{(t-1)} \geq \sum_{k \in I_t} \mathbf{w}_k^{(t-1)}$ due to the weighted-majority.

Therefore, for any step $t$ where a mistake occurs, we have:

$$W^{(t)} = \sum_{k=1}^{K} w_k^{(t)} = (1-\eta) \underbrace{\sum_{k \in J_t} \mathbf{w}_k^{(t-1)}}_{A} + \underbrace{\sum_{k \in I_t} \mathbf{w}_k^{(t-1)}}_{B} \overset{(\bigstar)}{\leq} \left( \frac{1-\eta}{2} + \frac{1}{2} \right) W^{(t-1)},$$

where step ($\bigstar$) follows from $A + B = W^{(t-1)}$, $A \geq B$, and $(1-\eta) \leq 1$.

# Proof of the Regret Bound

For any $t$, we define the potential $W^{(t)} = \sum_{k=1}^{K} \mathbf{w}_k^{(t)}$ with $W^{(0)} = K$.

For any time step $t$, we denote $I_t := \{k \in [K] : h_k(\mathbf{x}_t) = y_t\}$ and $J_t := [K] \backslash I_t$.

If $\hat{y}_t \neq y_t$, then $\sum_{k \in J_t} \mathbf{w}_k^{(t-1)} \geq \sum_{k \in I_t} \mathbf{w}_k^{(t-1)}$ due to the weighted-majority.

Therefore, for any step $t$ where a mistake occurs, we have:

$$W^{(t)} = \sum_{k=1}^{K} w_k^{(t)} = (1-\eta) \underbrace{\sum_{k \in J_t} \mathbf{w}_k^{(t-1)}}_{A} + \underbrace{\sum_{k \in I_t} \mathbf{w}_k^{(t-1)}}_{B} \stackrel{(\bigstar)}{\leq} \left( \frac{1-\eta}{2} + \frac{1}{2} \right) W^{(t-1)},$$

where step ($\bigstar$) follows from $A + B = W^{(t-1)}$, $A \geq B$, and $(1-\eta) \leq 1$.

Applying these inequalities for all $T$ steps, we get

$$(1-\eta)^{M_T^*} \leq W^{(T)} \leq W^{(0)} \left(1 - \frac{\eta}{2}\right)^{\widehat{M}_T} \leq K \cdot \left(1 - \frac{\eta}{2}\right)^{\widehat{M}_T}.$$

# Proof of the Regret Bound

For any $t$, we define the potential $W^{(t)} = \sum_{k=1}^{K} \mathbf{w}_k^{(t)}$ with $W^{(0)} = K$.

For any time step $t$, we denote $I_t := \{k \in [K] : h_k(\mathbf{x}_t) = y_t\}$ and $J_t := [K] \backslash I_t$.

If $\hat{y}_t \neq y_t$, then $\sum_{k \in J_t} \mathbf{w}_k^{(t-1)} \geq \sum_{k \in I_t} \mathbf{w}_k^{(t-1)}$ due to the weighted-majority.

Therefore, for any step $t$ where a mistake occurs, we have:

$$W^{(t)} = \sum_{k=1}^{K} w_k^{(t)} = (1-\eta) \underbrace{\sum_{k \in J_t} \mathbf{w}_k^{(t-1)}}_{A} + \underbrace{\sum_{k \in I_t} \mathbf{w}_k^{(t-1)}}_{B} \overset{(\bigstar)}{\leq} \left(\frac{1-\eta}{2} + \frac{1}{2}\right) W^{(t-1)},$$

where step ($\bigstar$) follows from $A + B = W^{(t-1)}$, $A \geq B$, and $(1-\eta) \leq 1$.

Applying these inequalities for all $T$ steps, we get

$$(1-\eta)^{M_T^*} \leq W^{(T)} \leq W^{(0)} \left(1 - \frac{\eta}{2}\right)^{\widehat{M}_T} \leq K \cdot \left(1 - \frac{\eta}{2}\right)^{\widehat{M}_T}.$$

Taking the natural logarithm $\ln$ on both sides and noting that for all $\eta < \frac{1}{2}$, $\ln(1-\eta) \geq -\eta - \eta^2$ and $\ln(1-\eta/2) \leq -\frac{\eta}{2}$, we complete the proof.

# Minimizing $\alpha$ in Regret

We have shown that the (deterministic) EWA algorithm achieves sub-linear $\alpha$-agnostic regret for $\alpha = 2$.

# Minimizing $\alpha$ in Regret

We have shown that the (deterministic) EWA algorithm achieves sub-linear $\alpha$-agnostic regret for $\alpha = 2$.

**Can we do better for smaller $\alpha$?**

# Minimizing $\alpha$ in Regret

We have shown that the (deterministic) EWA algorithm achieves sub-linear $\alpha$-agnostic regret for $\alpha = 2$.

**Can we do better for smaller $\alpha$?**

- Not really! In fact, $\alpha = 2$ is the minimal value required to achieve sub-linear $\alpha$-agnostic regret for *any* deterministic predictor.
- To see this, consider the hypothesis class $\mathcal{H} := \{h_0, h_1\}$, where $h_0$ labels every instance as $0$ and $h_1$ labels every instance as $1$.

# Minimizing $\alpha$ in Regret

We have shown that the (deterministic) EWA algorithm achieves sub-linear $\alpha$-agnostic regret for $\alpha = 2$.

**Can we do better for smaller $\alpha$?**

- Not really! In fact, $\alpha = 2$ is the minimal value required to achieve sub-linear $\alpha$-agnostic regret for *any* deterministic predictor.
- To see this, consider the hypothesis class $\mathcal{H} := \{h_0, h_1\}$, where $h_0$ labels every instance as $0$ and $h_1$ labels every instance as $1$.
  - An adversary, as in Cover's impossibility result, can force $\widehat{M}_T = T$.

# Minimizing $\alpha$ in Regret

We have shown that the (deterministic) EWA algorithm achieves sub-linear $\alpha$-agnostic regret for $\alpha = 2$.

**Can we do better for smaller $\alpha$?**

- Not really! In fact, $\alpha = 2$ is the minimal value required to achieve sub-linear $\alpha$-agnostic regret for *any* deterministic predictor.
- To see this, consider the hypothesis class $\mathcal{H} := \{h_0, h_1\}$, where $h_0$ labels every instance as $0$ and $h_1$ labels every instance as $1$.
  - An adversary, as in Cover's impossibility result, can force $\widehat{M}_T = T$.
  - Moreover, the minimal achievable mistake $M_T^*$ is upper-bounded by $T/2$.

# Minimizing $\alpha$ in Regret

We have shown that the (deterministic) EWA algorithm achieves sub-linear $\alpha$-agnostic regret for $\alpha = 2$.

**Can we do better for smaller $\alpha$?**

- Not really! In fact, $\alpha = 2$ is the minimal value required to achieve sub-linear $\alpha$-agnostic regret for *any* deterministic predictor.
- To see this, consider the hypothesis class $\mathcal{H} := \{h_0, h_1\}$, where $h_0$ labels every instance as $0$ and $h_1$ labels every instance as $1$.
  - An adversary, as in Cover's impossibility result, can force $\widehat{M}_T = T$.
  - Moreover, the minimal achievable mistake $M_T^*$ is upper-bounded by $T/2$.
  - Therefore, any $\alpha$-agnostic sub-linear regret must have $\alpha \geq 2$.

## Minimizing $\alpha$ in Regret

We have shown that the (deterministic) EWA algorithm achieves sub-linear $\alpha$-agnostic regret for $\alpha = 2$.

**Can we do better for smaller $\alpha$?**

- Not really! In fact, $\alpha = 2$ is the minimal value required to achieve sub-linear $\alpha$-agnostic regret for *any* deterministic predictor.
- To see this, consider the hypothesis class $\mathcal{H} := \{h_0, h_1\}$, where $h_0$ labels every instance as $0$ and $h_1$ labels every instance as $1$.
  - An adversary, as in Cover's impossibility result, can force $\widehat{M}_T = T$.
  - Moreover, the minimal achievable mistake $M_T^*$ is upper-bounded by $T/2$.
  - Therefore, any $\alpha$-agnostic sub-linear regret must have $\alpha \geq 2$.

**Homework**: Consider the empirical risk minimization (ERM) predictor:

$$\text{Predicts } \hat{y}_t = \hat{h}_t(\mathbf{x}_t) \text{ such that } \hat{h}_t = \arg\min_{h \in \mathcal{H}} \sum_{i=1}^{t-1} 1\{h(\mathbf{x}_i) \neq y_i\}.$$

Show that the ERM predictor achieves $\widehat{M}_T \leq (M_T^* + 1) \cdot |\mathcal{H}|$, and this is optimal for certain classes $\mathcal{H}$. (Hint: each hypothesis contributes $\leq M_T^* + 1$ mistakes.)

# Achieving $\alpha = 1$ via Randomized predictors

We now show that for randomized predictors, one can indeed achieves the $\alpha$-agnostic regret with $\alpha = 1$.

## Achieving $\alpha = 1$ via Randomized predictors

We now show that for randomized predictors, one can indeed achieves the $\alpha$-agnostic regret with $\alpha = 1$. Let $\mathcal{H} = \{h_1, \cdots, h_K\}$.

**The (randomized) EWA predictor**:

1. Maintain a weight vector $\mathbf{w}^{(t)} \in \mathbb{R}^K$, initially $\mathbf{w}^{(0)} = (1, \cdots, 1)$.
2. At each step $t$, sample $\hat{k}_t \sim \tilde{p}_t$ and predict $\hat{y}_t := h_{\hat{k}_t}(\mathbf{x}_t)$ where

$$\forall k \in [K], \ \tilde{p}_t[k] = \frac{\mathbf{w}_k^{(t-1)}}{\sum_{k=1}^K \mathbf{w}_k^{(t-1)}}.$$

3. Update $\mathbf{w}_k^{(t)} = \mathbf{w}_k^{(t-1)} e^{-\eta 1\{h_k(\mathbf{x}_t) \neq y_t\}}$, where $\eta < 1$ is tunable.

# Achieving $\alpha = 1$ via Randomized predictors

We now show that for randomized predictors, one can indeed achieves the $\alpha$-agnostic regret with $\alpha = 1$. Let $\mathcal{H} = \{h_1, \cdots, h_K\}$.

**The (randomized) EWA predictor**:

1. Maintain a weight vector $\mathbf{w}^{(t)} \in \mathbb{R}^K$, initially $\mathbf{w}^{(0)} = (1, \cdots, 1)$.
2. At each step $t$, sample $\hat{k}_t \sim \tilde{p}_t$ and predict $\hat{y}_t := h_{\hat{k}_t}(\mathbf{x}_t)$ where

$$\forall k \in [K], \ \tilde{p}_t[k] = \frac{\mathbf{w}_k^{(t-1)}}{\sum_{k=1}^K \mathbf{w}_k^{(t-1)}}.$$

3. Update $\mathbf{w}_k^{(t)} = \mathbf{w}_k^{(t-1)} e^{-\eta 1\{h_k(\mathbf{x}_t) \neq y_t\}}$, where $\eta < 1$ is tunable.

**Theorem 2**: Regardless of how Nature generates the data, as long as the selection is independent to the internal randomness of the predictor, we have

$$\mathbb{E}_{\hat{y}^T}\left[\sum_{t=1}^{T} 1\{\hat{y}_t \neq y_t\}\right] \leq M_T^* + \frac{\ln(|\mathcal{H}|)}{\eta} + \frac{\eta T}{8}$$

# Achieving $\alpha = 1$ via Randomized predictors

We now show that for randomized predictors, one can indeed achieves the $\alpha$-agnostic regret with $\alpha = 1$. Let $\mathcal{H} = \{h_1, \cdots, h_K\}$.

**The (randomized) EWA predictor**:

1. Maintain a weight vector $\mathbf{w}^{(t)} \in \mathbb{R}^K$, initially $\mathbf{w}^{(0)} = (1, \cdots, 1)$.
2. At each step $t$, sample $\hat{k}_t \sim \tilde{p}_t$ and predict $\hat{y}_t := h_{\hat{k}_t}(\mathbf{x}_t)$ where

$$\forall k \in [K], \ \tilde{p}_t[k] = \frac{\mathbf{w}_k^{(t-1)}}{\sum_{k=1}^K \mathbf{w}_k^{(t-1)}}.$$

3. Update $\mathbf{w}_k^{(t)} = \mathbf{w}_k^{(t-1)} e^{-\eta 1\{h_k(\mathbf{x}_t) \neq y_t\}}$, where $\eta < 1$ is tunable.

**Theorem 2**: Regardless of how Nature generates the data, as long as the selection is independent to the internal randomness of the predictor, we have

$$\mathbb{E}_{\hat{y}^T}\left[\sum_{t=1}^T 1\{\hat{y}_t \neq y_t\}\right] \leq M_T^* + \frac{\ln(|\mathcal{H}|)}{\eta} + \frac{\eta T}{8} \Rightarrow \text{reg}_T^1 \leq O(\sqrt{T \log |\mathcal{H}|}).$$

# Preparing for the Proof: Hoeffding's Lemma

**Hoeffding's Lemma**: Let $X$ be a random variable with $a \leq X \leq b$. Then for any $s \in \mathbb{R}$, we have

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

# Preparing for the Proof: Hoeffding's Lemma

**Hoeffding's Lemma**: Let $X$ be a random variable with $a \leq X \leq b$. Then for any $s \in \mathbb{R}$, we have

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

**Sketch of Proof**: Note that $\ln \mathbb{E}[e^{sX}] = s\mathbb{E}[X] + \ln \mathbb{E}[e^{s(X-\mathbb{E}[X])}]$, so we only need to consider the case where $\mathbb{E}[X] = 0$.

# Preparing for the Proof: Hoeffding's Lemma

**Hoeffding's Lemma**: Let $X$ be a random variable with $a \leq X \leq b$. Then for any $s \in \mathbb{R}$, we have

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

**Sketch of Proof**: Note that $\ln \mathbb{E}[e^{sX}] = s\mathbb{E}[X] + \ln \mathbb{E}[e^{s(X-\mathbb{E}[X])}]$, so we only need to consider the case where $\mathbb{E}[X] = 0$. Observe that for all $a \leq x \leq b$, we have

$$e^{sx} \leq \frac{x-a}{b-a}e^{sb} + \frac{b-x}{b-a}e^{sa},$$

by Jensen's inequality and the convexity of $e^{sx}$ over $x$ (verify it!).

# Preparing for the Proof: Hoeffding's Lemma

**Hoeffding's Lemma**: Let $X$ be a random variable with $a \leq X \leq b$. Then for any $s \in \mathbb{R}$, we have

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

**Sketch of Proof**: Note that $\ln \mathbb{E}[e^{sX}] = s\mathbb{E}[X] + \ln \mathbb{E}[e^{s(X-\mathbb{E}[X])}]$, so we only need to consider the case where $\mathbb{E}[X] = 0$. Observe that for all $a \leq x \leq b$, we have

$$e^{sx} \leq \frac{x-a}{b-a}e^{sb} + \frac{b-x}{b-a}e^{sa},$$

by Jensen's inequality and the convexity of $e^{sx}$ over $x$ (verify it!). Taking expectation over $x \sim X$ on both sides and using $\mathbb{E}[X] = 0$, the right-hand side can be expressed as a function of $s$.

# Preparing for the Proof: Hoeffding's Lemma

**Hoeffding's Lemma**: Let $X$ be a random variable with $a \leq X \leq b$. Then for any $s \in \mathbb{R}$, we have

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

**Sketch of Proof**: Note that $\ln \mathbb{E}[e^{sX}] = s\mathbb{E}[X] + \ln \mathbb{E}[e^{s(X-\mathbb{E}[X])}]$, so we only need to consider the case where $\mathbb{E}[X] = 0$. Observe that for all $a \leq x \leq b$, we have

$$e^{sx} \leq \frac{x-a}{b-a}e^{sb} + \frac{b-x}{b-a}e^{sa},$$

by Jensen's inequality and the convexity of $e^{sx}$ over $x$ (verify it!). Taking expectation over $x \sim X$ on both sides and using $\mathbb{E}[X] = 0$, the right-hand side can be expressed as a function of $s$. The lemma follows by Taylor expansion of this function up to the second order (verify it!). □

# Proving the Regret Bound of Randomized EWA

We again define the potential $W^{(t)} = \sum_{k=1}^{K} \mathbf{w}_k^{(t)}$.

# Proving the Regret Bound of Randomized EWA

We again define the potential $W^{(t)} = \sum_{k=1}^{K} \mathbf{w}_k^{(t)}$. Observe that:

$$\ln \frac{W^{(t)}}{W^{(t-1)}} = \ln \sum_{k=1}^{K} \frac{\mathbf{w}_k^{(t-1)}}{W^{(t-1)}} e^{-\eta 1\{h_k(\mathbf{x}_t) \neq y_t\}}$$

$$\overset{(\bigstar)}{\leq} -\eta \sum_{k=1}^{K} \frac{\mathbf{w}_k^{(t-1)}}{W^{(t-1)}} 1\{h_k(\mathbf{x}_t) \neq y_t\} + \frac{\eta^2}{8}$$

$$\overset{(\star\star)}{=} -\eta \mathbb{E}_{\hat{y}_t}[1\{\hat{y}_t \neq y_t\}] + \frac{\eta^2}{8},$$

# Proving the Regret Bound of Randomized EWA

We again define the potential $W^{(t)} = \sum_{k=1}^{K} \mathbf{w}_k^{(t)}$. Observe that:

$$
\begin{aligned}
\ln \frac{W^{(t)}}{W^{(t-1)}} &= \ln \sum_{k=1}^{K} \frac{\mathbf{w}_k^{(t-1)}}{W^{(t-1)}} e^{-\eta 1\{h_k(\mathbf{x}_t) \neq y_t\}} \\
&\overset{(\bigstar)}{\leq} -\eta \sum_{k=1}^{K} \frac{\mathbf{w}_k^{(t-1)}}{W^{(t-1)}} 1\{h_k(\mathbf{x}_t) \neq y_t\} + \frac{\eta^2}{8} \\
&\overset{(\star\star)}{=} -\eta \mathbb{E}_{\hat{y}_t}[1\{\hat{y}_t \neq y_t\}] + \frac{\eta^2}{8},
\end{aligned}
$$

where ($\bigstar$) follows by Hoeffding's lemma (verify it!) and ($\star\star$) follows from the definition of $\hat{y}_t$.

# Proving the Regret Bound of Randomized EWA

We again define the potential $W^{(t)} = \sum_{k=1}^{K} \mathbf{w}_k^{(t)}$. Observe that:

$$\ln \frac{W^{(t)}}{W^{(t-1)}} = \ln \sum_{k=1}^{K} \frac{\mathbf{w}_k^{(t-1)}}{W^{(t-1)}} e^{-\eta 1\{h_k(\mathbf{x}_t) \neq y_t\}}$$

$$\overset{(\bigstar)}{\leq} -\eta \sum_{k=1}^{K} \frac{\mathbf{w}_k^{(t-1)}}{W^{(t-1)}} 1\{h_k(\mathbf{x}_t) \neq y_t\} + \frac{\eta^2}{8}$$

$$\overset{(\star\star)}{=} -\eta \mathbb{E}_{\hat{y}_t}[1\{\hat{y}_t \neq y_t\}] + \frac{\eta^2}{8},$$

where $(\bigstar)$ follows by Hoeffding's lemma (verify it!) and $(\star\star)$ follows from the definition of $\hat{y}_t$. Summing from $t = 1$ to $T$, we get:

$$-\eta M_T^* \leq \ln W^{(T)} \leq -\eta \mathbb{E}_{\hat{y}^T} \left[ \sum_{t=1}^{T} 1\{\hat{y}_t \neq y_t\} \right] + \frac{\eta^2 T}{8} + \ln |\mathcal{H}|.$$

# Proving the Regret Bound of Randomized EWA

We again define the potential $W^{(t)} = \sum_{k=1}^{K} \mathbf{w}_k^{(t)}$. Observe that:

$$\ln \frac{W^{(t)}}{W^{(t-1)}} = \ln \sum_{k=1}^{K} \frac{\mathbf{w}_k^{(t-1)}}{W^{(t-1)}} e^{-\eta 1\{h_k(\mathbf{x}_t) \neq y_t\}}$$

$$\overset{(\bigstar)}{\leq} -\eta \sum_{k=1}^{K} \frac{\mathbf{w}_k^{(t-1)}}{W^{(t-1)}} 1\{h_k(\mathbf{x}_t) \neq y_t\} + \frac{\eta^2}{8}$$

$$\overset{(\star\star)}{=} -\eta \mathbb{E}_{\hat{y}_t}[1\{\hat{y}_t \neq y_t\}] + \frac{\eta^2}{8},$$

where ($\bigstar$) follows by Hoeffding's lemma (verify it!) and ($\star\star$) follows from the definition of $\hat{y}_t$. Summing from $t = 1$ to $T$, we get:

$$-\eta M_T^* \leq \ln W^{(T)} \leq -\eta \mathbb{E}_{\hat{y}^T} \left[ \sum_{t=1}^{T} 1\{\hat{y}_t \neq y_t\} \right] + \frac{\eta^2 T}{8} + \ln |\mathcal{H}|.$$

The regret bound follows by rearranging the inequality.

# EWA Algorithm for General Losses

Let $\mathcal{Y} = [0,1]$ and $\mathcal{H} \subset [0,1]^{\mathcal{X}}$ be a finite hypothesis class of size $K$. Let $\ell : \mathcal{Y} \times \mathcal{Y} \to [0,1]$ be a loss function that is convex in its first argument.

# EWA Algorithm for General Losses

Let $\mathcal{Y} = [0,1]$ and $\mathcal{H} \subset [0,1]^{\mathcal{X}}$ be a finite hypothesis class of size $K$. Let $\ell : \mathcal{Y} \times \mathcal{Y} \to [0,1]$ be a loss function that is convex in its first argument.

**The (generalized) EWA predictor**:

1. Maintain a weight vector $\mathbf{w}^{(t)} \in \mathbb{R}^K$, initially $\mathbf{w}^{(0)} = (1, \cdots, 1)$.
2. At each step $t$, predict $\hat{y}_t := \sum_{k=1}^{K} \tilde{p}_t[k] \cdot h_k(\mathbf{x}_t)$, where

$$\forall k \in [K], \;\; \tilde{p}_t[k] = \frac{\mathbf{w}_k^{(t-1)}}{\sum_{k=1}^{K} \mathbf{w}_k^{(t-1)}}.$$

3. Update $\mathbf{w}_k^{(t)} = \mathbf{w}_k^{(t-1)} e^{-\eta \ell(h_k(\mathbf{x}_t), y_t)}$, where $\eta < 1$ is a tunable parameter.

# EWA Algorithm for General Losses

Let $\mathcal{Y} = [0,1]$ and $\mathcal{H} \subset [0,1]^{\mathcal{X}}$ be a finite hypothesis class of size $K$. Let $\ell : \mathcal{Y} \times \mathcal{Y} \to [0,1]$ be a loss function that is convex in its first argument.

**The (generalized) EWA predictor**:

1. Maintain a weight vector $\mathbf{w}^{(t)} \in \mathbb{R}^K$, initially $\mathbf{w}^{(0)} = (1, \cdots, 1)$.
2. At each step $t$, predict $\hat{y}_t := \sum_{k=1}^{K} \tilde{p}_t[k] \cdot h_k(\mathbf{x}_t)$, where

$$\forall k \in [K], \;\; \tilde{p}_t[k] = \frac{\mathbf{w}_k^{(t-1)}}{\sum_{k=1}^{K} \mathbf{w}_k^{(t-1)}}.$$

3. Update $\mathbf{w}_k^{(t)} = \mathbf{w}_k^{(t-1)} e^{-\eta \ell(h_k(\mathbf{x}_t), y_t)}$, where $\eta < 1$ is a tunable parameter.

**Homework**: Show that, regardless of how Nature generates the data $\mathbf{x}^T, y^T$, the (generalized) EWA algorithm enjoys the following risk bound:

$$\sum_{t=1}^{T} \ell(\hat{y}_t, y_t) \le \inf_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell(h(\mathbf{x}_t), y_t) + \frac{\ln(|\mathcal{H}|)}{\eta} + \frac{\eta T}{8}.$$

(**Hint**: apply Jensen's inequality at step $(\star\star)$ using the convexity of $\ell$.)

# Concluding Remarks

- In this lecture, we only introduced the online learning framework very informally. For example, we did not explicitly define how Nature's strategies are selected, which will be covered in the upcoming lectures.
- Throughout the entire lectures, we will focus solely on online learning with non-structured experts (i.e., with *general* hypothesis classes).
- There is also a rich body of literature dealing with structured experts, such as the Online Convex Optimization (OCO) framework, which we unfortunately have to omit due to time constraints.
  - We refer interested readers to the book: "*Introduction to Online Convex Optimization*" by Elad Hazan.