

Learning from Dependent Observations

Ingo Steinwart

June 1, 2006

General Learning Goal

- ▶ X space of input samples
- ▶ Y space of labels, usually $Y \subset \mathbb{R}$.

General Learning Goal

- ▶ X space of input samples
 Y space of labels, usually $Y \subset \mathbb{R}$.
- ▶ $L : X \times Y \times \mathbb{R} \rightarrow [0, \infty)$ loss function.

General Learning Goal

- ▶ X space of input samples
 Y space of labels, usually $Y \subset \mathbb{R}$.
- ▶ $L : X \times Y \times \mathbb{R} \rightarrow [0, \infty)$ loss function.
- ▶ Already observed samples

$$T_{\text{past}} = ((x_1, y_1), \dots, (x_n, y_n)) \in (X \times Y)^n$$

General Learning Goal

- ▶ X space of input samples
 Y space of labels, usually $Y \subset \mathbb{R}$.
- ▶ $L : X \times Y \times \mathbb{R} \rightarrow [0, \infty)$ loss function.
- ▶ Already observed samples

$$T_{\text{past}} = ((x_1, y_1), \dots, (x_n, y_n)) \in (X \times Y)^n$$

- ▶ Future, unknown samples (of unknown length m)

$$T_{\text{future}} = ((x_{n+1}, y_{n+1}), \dots, (x_{n+m}, y_{n+m})) \in (X \times Y)^m$$

General Learning Goal

- ▶ X space of input samples
 Y space of labels, usually $Y \subset \mathbb{R}$.
- ▶ $L : X \times Y \times \mathbb{R} \rightarrow [0, \infty)$ loss function.
- ▶ Already observed samples

$$T_{\text{past}} = ((x_1, y_1), \dots, (x_n, y_n)) \in (X \times Y)^n$$

- ▶ Future, unknown samples (of unknown length m)

$$T_{\text{future}} = ((x_{n+1}, y_{n+1}), \dots, (x_{n+m}, y_{n+m})) \in (X \times Y)^m$$

- ▶ **Goal:**

With the help of T_{past} find a function $f : X \rightarrow \mathbb{R}$ such that

$$\mathcal{R}_{L, T_{\text{future}}}(f) := \frac{1}{m} \sum_{i=n+1}^{n+m} L(x_i, y_i, f(x_i))$$

is small.

Reformulation for i.i.d. Observations

► **Classical Learning Theory Assumption:**

T_{past} and T_{future} are i.i.d. samples from an unknown probability measure P on $X \times Y$.

Reformulation for i.i.d. Observations

▶ **Classical Learning Theory Assumption:**

T_{past} and T_{future} are i.i.d. samples from an unknown probability measure P on $X \times Y$.

▶ **Consequence:**

For fixed $f : X \rightarrow \mathbb{R}$ the law of large numbers (LLN) shows

$$\lim_{m \rightarrow \infty} \mathcal{R}_{L, T_{\text{future}}}(f) = \int_{X \times Y} L(x, y, f(x)) dP(x, y)$$

Reformulation for i.i.d. Observations

► **Classical Learning Theory Assumption:**

T_{past} and T_{future} are i.i.d. samples from an unknown probability measure P on $X \times Y$.

► **Consequence:**

For fixed $f : X \rightarrow \mathbb{R}$ the law of large numbers (LLN) shows

$$\lim_{m \rightarrow \infty} \mathcal{R}_{L, T_{\text{future}}}(f) = \int_{X \times Y} L(x, y, f(x)) dP(x, y)$$

► **Reformulated Goal:**

With the help of T_{past} find a function $f : X \rightarrow \mathbb{R}$ such that

$$\mathcal{R}_{L, P}(f) := \int_{X \times Y} L(x, y, f(x)) dP(x, y)$$

is as small as possible.

Traditional ERM Approach

- ▶ **Penalized ERM algorithm:**

Traditional ERM Approach

- ▶ **Penalized ERM algorithm:**
 - ▶ Fix a function class \mathcal{F} of functions $f : X \rightarrow \mathbb{R}$.

Traditional ERM Approach

▶ Penalized ERM algorithm:

- ▶ Fix a function class \mathcal{F} of functions $f : X \rightarrow \mathbb{R}$.
- ▶ Fix a penalty function $\text{pen} : (0, \infty) \times \mathcal{F} \rightarrow [0, \infty]$.

Traditional ERM Approach

▶ Penalized ERM algorithm:

- ▶ Fix a function class \mathcal{F} of functions $f : X \rightarrow \mathbb{R}$.
- ▶ Fix a penalty function $\text{pen} : (0, \infty) \times \mathcal{F} \rightarrow [0, \infty]$.
- ▶ Find

$$f_{T_{\text{past}}, \lambda} \in \underset{f \in \mathcal{F}}{\text{argmin}} \left(\text{pen}(\lambda, f) + \mathcal{R}_{L, T_{\text{past}}}(f) \right)$$

Traditional ERM Approach

▶ Penalized ERM algorithm:

- ▶ Fix a function class \mathcal{F} of functions $f : X \rightarrow \mathbb{R}$.
- ▶ Fix a penalty function $\text{pen} : (0, \infty) \times \mathcal{F} \rightarrow [0, \infty]$.
- ▶ Find

$$f_{T_{\text{past}}, \lambda} \in \underset{f \in \mathcal{F}}{\text{argmin}} \left(\text{pen}(\lambda, f) + \mathcal{R}_{L, T_{\text{past}}}(f) \right)$$

▶ Statistical Learning Theory:

Traditional ERM Approach

▶ Penalized ERM algorithm:

- ▶ Fix a function class \mathcal{F} of functions $f : X \rightarrow \mathbb{R}$.
- ▶ Fix a penalty function $\text{pen} : (0, \infty) \times \mathcal{F} \rightarrow [0, \infty]$.
- ▶ Find

$$f_{T_{\text{past}}, \lambda} \in \underset{f \in \mathcal{F}}{\text{argmin}} \left(\text{pen}(\lambda, f) + \mathcal{R}_{L, T_{\text{past}}}(f) \right)$$

▶ Statistical Learning Theory:

- ▶ Under certain assumptions on \mathcal{F} and λ_n we have

$$\lim_{n \rightarrow \infty} \mathcal{R}_{L, P}(f_{T_{\text{past}}, \lambda_n}) = \mathcal{R}_{L, P}^* := \inf \{ \mathcal{R}_{L, P}(f) \mid f : X \rightarrow \mathbb{R} \}.$$

Traditional ERM Approach

▶ Penalized ERM algorithm:

- ▶ Fix a function class \mathcal{F} of functions $f : X \rightarrow \mathbb{R}$.
- ▶ Fix a penalty function $\text{pen} : (0, \infty) \times \mathcal{F} \rightarrow [0, \infty]$.
- ▶ Find

$$f_{T_{\text{past}}, \lambda} \in \underset{f \in \mathcal{F}}{\text{argmin}} \left(\text{pen}(\lambda, f) + \mathcal{R}_{L, T_{\text{past}}}(f) \right)$$

▶ Statistical Learning Theory:

- ▶ Under certain assumptions on \mathcal{F} and λ_n we have

$$\lim_{n \rightarrow \infty} \mathcal{R}_{L, P}(f_{T_{\text{past}}, \lambda_n}) = \mathcal{R}_{L, P}^* := \inf \{ \mathcal{R}_{L, P}(f) \mid f : X \rightarrow \mathbb{R} \}.$$

- ▶ Examples include support vector machines, regularized boosting, structural risk minimization, ...

Traditional ERM Approach

▶ Penalized ERM algorithm:

- ▶ Fix a function class \mathcal{F} of functions $f : X \rightarrow \mathbb{R}$.
- ▶ Fix a penalty function $\text{pen} : (0, \infty) \times \mathcal{F} \rightarrow [0, \infty]$.
- ▶ Find

$$f_{T_{\text{past}}, \lambda} \in \underset{f \in \mathcal{F}}{\text{argmin}} \left(\text{pen}(\lambda, f) + \mathcal{R}_{L, T_{\text{past}}}(f) \right)$$

▶ Statistical Learning Theory:

- ▶ Under certain assumptions on \mathcal{F} and λ_n we have

$$\lim_{n \rightarrow \infty} \mathcal{R}_{L, P}(f_{T_{\text{past}}, \lambda_n}) = \mathcal{R}_{L, P}^* := \inf \{ \mathcal{R}_{L, P}(f) \mid f : X \rightarrow \mathbb{R} \}.$$

- ▶ Examples include support vector machines, regularized boosting, structural risk minimization, ...
- ▶ In many cases convergence rates are possible under additional assumptions.

The Link between Past and Future:

- ▶ Statistical learning theory (often) ensures

$$\lim_{n \rightarrow \infty} \mathcal{R}_{L,P}(f_{T_{\text{past}}, \lambda_n}) = \mathcal{R}_{L,P}^*.$$

The Link between Past and Future:

- ▶ Statistical learning theory (often) ensures

$$\lim_{n \rightarrow \infty} \mathcal{R}_{L,P}(f_{T_{\text{past}}, \lambda_n}) = \mathcal{R}_{L,P}^*.$$

- ▶ The law of large numbers ensures

$$\lim_{m \rightarrow \infty} \mathcal{R}_{L, T_{\text{future}}}(f_{T_{\text{past}}, \lambda_n}) = \mathcal{R}_{L,P}(f_{T_{\text{past}}, \lambda_n}).$$

The Link between Past and Future:

- ▶ Statistical learning theory (often) ensures

$$\lim_{n \rightarrow \infty} \mathcal{R}_{L,P}(f_{T_{\text{past}}, \lambda_n}) = \mathcal{R}_{L,P}^*.$$

- ▶ The law of large numbers ensures

$$\lim_{m \rightarrow \infty} \mathcal{R}_{L, T_{\text{future}}}(f_{T_{\text{past}}, \lambda_n}) = \mathcal{R}_{L,P}(f_{T_{\text{past}}, \lambda_n}).$$

- ▶ **Consequence:**

The Link between Past and Future:

- ▶ Statistical learning theory (often) ensures

$$\lim_{n \rightarrow \infty} \mathcal{R}_{L,P}(f_{T_{\text{past}}, \lambda_n}) = \mathcal{R}_{L,P}^*.$$

- ▶ The law of large numbers ensures

$$\lim_{m \rightarrow \infty} \mathcal{R}_{L, T_{\text{future}}}(f_{T_{\text{past}}, \lambda_n}) = \mathcal{R}_{L,P}(f_{T_{\text{past}}, \lambda_n}).$$

- ▶ **Consequence:**

- ▶ The distribution P and its risk $\mathcal{R}_{L,P}(\cdot)$ serves us as a bridge between past and future observations.

The Link between Past and Future:

- ▶ Statistical learning theory (often) ensures

$$\lim_{n \rightarrow \infty} \mathcal{R}_{L,P}(f_{T_{\text{past}}, \lambda_n}) = \mathcal{R}_{L,P}^*.$$

- ▶ The law of large numbers ensures

$$\lim_{m \rightarrow \infty} \mathcal{R}_{L, T_{\text{future}}}(f_{T_{\text{past}}, \lambda_n}) = \mathcal{R}_{L,P}(f_{T_{\text{past}}, \lambda_n}).$$

- ▶ **Consequence:**

- ▶ The distribution P and its risk $\mathcal{R}_{L,P}(\cdot)$ serves us as a bridge between past and future observations.
- ▶ Convergence rates are used to describe how well we can generalize from the past to the future.

Why not i.i.d.?

Often the samples are

- ▶ inherently temporal in nature (system diagnosis, market prediction, ...)

Why not i.i.d.?

Often the samples are

- ▶ inherently temporal in nature (system diagnosis, market prediction, ...)
- ▶ collected from different sources

Why not i.i.d.?

Often the samples are

- ▶ inherently temporal in nature (system diagnosis, market prediction, ...)
- ▶ collected from different sources

In the worst case the samples are neither independently nor identically distributed.

General assumptions for the non i.i.d. case

► **Basic assumption:**

All samples are realizations from a stochastic process

$$\mathcal{Z} := (Z_i)_{i \geq 1},$$

where $Z_i : \Omega \rightarrow X \times Y$ and (Ω, μ) is a probability space.

General assumptions for the non i.i.d. case

► **Basic assumption:**

All samples are realizations from a stochastic process

$$\mathcal{Z} := (Z_i)_{i \geq 1},$$

where $Z_i : \Omega \rightarrow X \times Y$ and (Ω, μ) is a probability space.

► **More formally:**

There is an $\omega \in \Omega$ sampled from μ such that

$$\begin{aligned} T_{\text{past}} &= (Z_1(\omega), \dots, Z_n(\omega)) \\ T_{\text{future}} &= (Z_{n+1}(\omega), \dots, Z_{n+m}(\omega)) \end{aligned}$$

General assumptions for the non i.i.d. case

► **Basic assumption:**

All samples are realizations from a stochastic process

$$\mathcal{Z} := (Z_i)_{i \geq 1},$$

where $Z_i : \Omega \rightarrow X \times Y$ and (Ω, μ) is a probability space.

► **More formally:**

There is an $\omega \in \Omega$ sampled from μ such that

$$\begin{aligned} T_{\text{past}} &= (Z_1(\omega), \dots, Z_n(\omega)) \\ T_{\text{future}} &= (Z_{n+1}(\omega), \dots, Z_{n+m}(\omega)) \end{aligned}$$

► **Fundamental Question:**

Can we generalize from past to future observations?

Identically Distributed Processes

A simple example

► **Assumption:**

All Z_i are identically distributed, i.e.

$$\mu_{Z_i} = \mu_{Z_j}, \quad i, j \geq 1$$

Identically Distributed Processes

A simple example

► **Assumption:**

All Z_i are identically distributed, i.e.

$$\mu_{Z_i} = \mu_{Z_j}, \quad i, j \geq 1$$

► **Question:**

Does the risk

$$\mathcal{R}_{L, \mu_{Z_1}}(f)$$

can serve us as a bridge between past and future?

Identically Distributed Processes

A simple example

► Assumption:

All Z_i are identically distributed, i.e.

$$\mu_{Z_i} = \mu_{Z_j}, \quad i, j \geq 1$$

► Question:

Does the risk

$$\mathcal{R}_{L, \mu_{Z_1}}(f)$$

can serve us as a bridge between past and future?

► Answer:

No, in general there is no LLN and hence we **cannot** guarantee

$$\lim_{m \rightarrow \infty} \mathcal{R}_{L, T_{\text{future}}}(f) = \mathcal{R}_{L, \mu_{Z_1}}(f).$$

Towards the general idea

► **Question:**

What distribution or risk can serve us as a bridge between past and future?

Towards the general idea

▶ **Question:**

What distribution or risk can serve us as a bridge between past and future?

▶ **Observation:**

If there is a distribution P on $X \times Y$ which can serve us as a bridge then we need at least

Towards the general idea

▶ **Question:**

What distribution or risk can serve us as a bridge between past and future?

▶ **Observation:**

If there is a distribution P on $X \times Y$ which can serve us as a bridge then we need at least

- ▶ a “law of large numbers”, i.e.

$$\lim_{m \rightarrow \infty} \mathcal{R}_{L, T_{\text{future}}}^m(f) = \mathcal{R}_{L, P}(f).$$

Towards the general idea

▶ **Question:**

What distribution or risk can serve us as a bridge between past and future?

▶ **Observation:**

If there is a distribution P on $X \times Y$ which can serve us as a bridge then we need at least

- ▶ a “law of large numbers”, i.e.

$$\lim_{m \rightarrow \infty} \mathcal{R}_{L, T_{\text{future}}} (f) = \mathcal{R}_{L, P}(f).$$

- ▶ a “new statistical learning theory” which ensures

$$\lim_{n \rightarrow \infty} \mathcal{R}_{L, P}(f_{T_{\text{past}}, \lambda_n}) = \mathcal{R}_{L, P}^*.$$

Towards the general idea

► **Question:**

What distribution or risk can serve us as a bridge between past and future?

► **Observation:**

If there is a distribution P on $X \times Y$ which can serve us as a bridge then we need at least

- a “law of large numbers”, i.e.

$$\lim_{m \rightarrow \infty} \mathcal{R}_{L, T_{\text{future}}} (f) = \mathcal{R}_{L, P}(f).$$

- a “new statistical learning theory” which ensures

$$\lim_{n \rightarrow \infty} \mathcal{R}_{L, P}(f_{T_{\text{past}}, \lambda_n}) = \mathcal{R}_{L, P}^*.$$

- In this case $\mathcal{R}_{L, P}(\cdot)$ can be used to redefine the learning goal.

Outline:

In the rest of this talk we will:

Outline:

In the rest of this talk we will:

- ▶ Consider general LLNs.

Outline:

In the rest of this talk we will:

- ▶ Consider general LLNs.
- ▶ Use these LLN's to define distribution that can serve us as a bridge between past and future observations.

Outline:

In the rest of this talk we will:

- ▶ Consider general LLNs.
- ▶ Use these LLN's to define distribution that can serve us as a bridge between past and future observations.
- ▶ Define consistency.

Outline:

In the rest of this talk we will:

- ▶ Consider general LLNs.
- ▶ Use these LLN's to define distribution that can serve us as a bridge between past and future observations.
- ▶ Define consistency.
- ▶ Establish “a sort of” consistency for general penalized ERM methods.

Outline:

In the rest of this talk we will:

- ▶ Consider general LLNs.
- ▶ Use these LLN's to define distribution that can serve us as a bridge between past and future observations.
- ▶ Define consistency.
- ▶ Establish “a sort of” consistency for general penalized ERM methods.
- ▶ Recall α -mixing processes.

Outline:

In the rest of this talk we will:

- ▶ Consider general LLNs.
- ▶ Use these LLN's to define distribution that can serve us as a bridge between past and future observations.
- ▶ Define consistency.
- ▶ Establish “a sort of” consistency for general penalized ERM methods.
- ▶ Recall α -mixing processes.
- ▶ Establish consistency of SVMs for α -mixing processes.

Assumptions in this Part of the Talk:

In this part of the talk we always assume that:

Assumptions in this Part of the Talk:

In this part of the talk we always assume that:

- ▶ (Ω, μ) is a probability space.

Assumptions in this Part of the Talk:

In this part of the talk we always assume that:

- ▶ (Ω, μ) is a probability space.
- ▶ $\mathcal{Z} := (Z_j)_{j \geq 1}$ is a stochastic process with $Z_j : \Omega \rightarrow X \times Y$.

Assumptions in this Part of the Talk:

In this part of the talk we always assume that:

- ▶ (Ω, μ) is a probability space.
- ▶ $\mathcal{Z} := (Z_i)_{i \geq 1}$ is a stochastic process with $Z_i : \Omega \rightarrow X \times Y$.
- ▶ All functions and sets are measurable.

Assumptions in this Part of the Talk:

In this part of the talk we always assume that:

- ▶ (Ω, μ) is a probability space.
- ▶ $\mathcal{Z} := (Z_i)_{i \geq 1}$ is a stochastic process with $Z_i : \Omega \rightarrow X \times Y$.
- ▶ All functions and sets are measurable.

Moreover, for a function $g : X \times Y \rightarrow \mathbb{R}$ we call the process $g \circ \mathcal{Z} := (g \circ Z_i)_{i \geq 1}$ an *image* of \mathcal{Z} .

Laws of Large Numbers:

- ▶ \mathcal{Z} satisfies weak law of large numbers for events (WLLNE):
for all $B \subset X \times Y$ there exists a $c_B \in \mathbb{R}$ such that

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \mathbf{1}_B \circ Z_i = c_B \quad \text{in probability } \mu. \quad (1)$$

Laws of Large Numbers:

- ▶ \mathcal{Z} satisfies weak law of large numbers for events (WLLNE):
for all $B \subset X \times Y$ there exists a $c_B \in \mathbb{R}$ such that

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \mathbf{1}_B \circ Z_i = c_B \quad \text{in probability } \mu. \quad (1)$$

In other words: for all $B \subset X \times Y$ there exists a $c_B \in \mathbb{R}$ such that for all $\varepsilon > 0$ we have

$$\lim_{n \rightarrow \infty} \mu \left(\left\{ \omega \in \Omega : \left| \frac{1}{n} \sum_{i=1}^n \mathbf{1}_B \circ Z_i(\omega) - c_B \right| > \varepsilon \right\} \right) = 0.$$

Laws of Large Numbers:

- ▶ \mathcal{Z} satisfies weak law of large numbers for events (WLLNE):
for all $B \subset X \times Y$ there exists a $c_B \in \mathbb{R}$ such that

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \mathbf{1}_B \circ Z_i = c_B \quad \text{in probability } \mu. \quad (1)$$

- ▶ \mathcal{Z} satisfies strong law of large numbers for events (SLLNE):
(1) holds μ -almost surely.

Laws of Large Numbers:

- ▶ \mathcal{Z} satisfies weak law of large numbers for events (WLLNE):
for all $B \subset X \times Y$ there exists a $c_B \in \mathbb{R}$ such that

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \mathbf{1}_B \circ Z_i = c_B \quad \text{in probability } \mu. \quad (1)$$

- ▶ \mathcal{Z} satisfies strong law of large numbers for events (SLLNE):
(1) holds μ -almost surely.
- ▶ \mathcal{Z} is asymptotically mean stationary (AMS):
for all $B \subset X \times Y$ there exists a $P(B) \in \mathbb{R}$ such that

$$P(B) = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \mathbb{E}_\mu \mathbf{1}_B \circ Z_i$$

Properties Related to Laws of Large Numbers:

- ▶ If \mathcal{Z} is AMS then P is a probability measure on $X \times Y$.

Properties Related to Laws of Large Numbers:

- ▶ If \mathcal{Z} is AMS then P is a probability measure on $X \times Y$.

Proof:

$$P_n := \frac{1}{n} \sum_{i=1}^n \mu_{Z_i}, \quad n \geq 1$$

are probability measures.

Properties Related to Laws of Large Numbers:

- ▶ If \mathcal{Z} is AMS then P is a probability measure on $X \times Y$.

Proof:

$$P_n := \frac{1}{n} \sum_{i=1}^n \mu_{Z_i}, \quad n \geq 1$$

are probability measures. AMS means

$$P(B) = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{\mu} \mathbf{1}_B \circ Z_i = \lim_{n \rightarrow \infty} P_n(B).$$

Properties Related to Laws of Large Numbers:

- ▶ If \mathcal{Z} is AMS then P is a probability measure on $X \times Y$.

Proof:

$$P_n := \frac{1}{n} \sum_{i=1}^n \mu_{Z_i}, \quad n \geq 1$$

are probability measures. AMS means

$$P(B) = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{\mu} \mathbf{1}_B \circ Z_i = \lim_{n \rightarrow \infty} P_n(B).$$

Vitali-Hahn-Saks then shows that P is probability measure.

Properties Related to Laws of Large Numbers:

- ▶ If \mathcal{Z} is AMS then P is a probability measure on $X \times Y$.
 P is called the **asymptotical mean distribution** of \mathcal{Z} .

Properties Related to Laws of Large Numbers:

- ▶ If \mathcal{Z} is AMS then P is a probability measure on $X \times Y$.
 P is called the **asymptotical mean distribution** of \mathcal{Z} .
- ▶ If \mathcal{Z} satisfies WLLNE then it is AMS and we have

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \mathbf{1}_B \circ Z_i = P(B) \quad \text{in probability } \mu.$$

Properties Related to Laws of Large Numbers:

- ▶ If \mathcal{Z} is AMS then P is a probability measure on $X \times Y$.
 P is called the **asymptotical mean distribution** of \mathcal{Z} .
- ▶ If \mathcal{Z} satisfies WLLNE then it is AMS and we have

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \mathbf{1}_B \circ Z_i = P(B) \quad \text{in probability } \mu.$$

Proof (sketch): WLLNE means

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \mathbf{1}_B \circ Z_i = c_B \quad \text{in probability } \mu.$$

Properties Related to Laws of Large Numbers:

- ▶ If \mathcal{Z} is AMS then P is a probability measure on $X \times Y$.
 P is called the **asymptotical mean distribution** of \mathcal{Z} .
- ▶ If \mathcal{Z} satisfies WLLNE then it is AMS and we have

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \mathbf{1}_B \circ Z_i = P(B) \quad \text{in probability } \mu.$$

Proof (sketch): WLLNE means

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \mathbf{1}_B \circ Z_i = c_B \quad \text{in probability } \mu.$$

Obviously, we have $\mathbf{1}_B \circ Z_i \in [0, 1]$ and hence $c_B \in [0, 1]$.

Properties Related to Laws of Large Numbers:

- ▶ If \mathcal{Z} is AMS then P is a probability measure on $X \times Y$.
 P is called the **asymptotical mean distribution** of \mathcal{Z} .
- ▶ If \mathcal{Z} satisfies WLLNE then it is AMS and we have

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \mathbf{1}_B \circ Z_i = P(B) \quad \text{in probability } \mu.$$

Proof (sketch): WLLNE means

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \mathbf{1}_B \circ Z_i = c_B \quad \text{in probability } \mu.$$

Obviously, we have $\mathbf{1}_B \circ Z_i \in [0, 1]$ and hence $c_B \in [0, 1]$.
Therefore convergence also holds in $L_1(\mu)$.

Properties Related to Laws of Large Numbers:

- ▶ If \mathcal{Z} is AMS then P is a probability measure on $X \times Y$.
 P is called the **asymptotical mean distribution** of \mathcal{Z} .
- ▶ If \mathcal{Z} satisfies WLLNE then it is AMS and we have

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \mathbf{1}_B \circ Z_i = P(B) \quad \text{in probability } \mu.$$

Convergence also holds in $L_1(\mu)$, and hence we have

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \mathbb{E}_\mu \mathbf{1}_B \circ Z_i = \lim_{n \rightarrow \infty} \int_{\Omega} \left| \frac{1}{n} \sum_{i=1}^n \mathbf{1}_B \circ Z_i \right| d\mu = \mathbb{E}_\mu |\mathbf{1}_B|.$$

Properties Related to Laws of Large Numbers:

- ▶ If \mathcal{Z} is AMS then P is a probability measure on $X \times Y$. P is called the **asymptotical mean distribution** of \mathcal{Z} .
- ▶ If \mathcal{Z} satisfies WLLNE then it is AMS and for all bounded $f : X \times Y \rightarrow \mathbb{R}$ we have

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n f \circ Z_i = \mathbb{E}_P f \quad \text{in probability } \mu. \quad (2)$$

Properties Related to Laws of Large Numbers:

- ▶ If \mathcal{Z} is AMS then P is a probability measure on $X \times Y$. P is called the **asymptotical mean distribution** of \mathcal{Z} .
- ▶ If \mathcal{Z} satisfies WLLNE then it is AMS and for all bounded $f : X \times Y \rightarrow \mathbb{R}$ we have

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n f \circ Z_i = \mathbb{E}_P f \quad \text{in probability } \mu. \quad (2)$$

- ▶ Analogous results for SLLNE.

Properties Related to Laws of Large Numbers:

- ▶ If \mathcal{Z} is AMS then P is a probability measure on $X \times Y$. P is called the **asymptotical mean distribution** of \mathcal{Z} .
- ▶ If \mathcal{Z} satisfies WLLNE then it is AMS and for all bounded $f : X \times Y \rightarrow \mathbb{R}$ we have

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n f \circ Z_i = \mathbb{E}_P f \quad \text{in probability } \mu. \quad (2)$$

- ▶ Analogous results for SLLNE.
- ▶ \mathcal{Z} satisfies **weak law of large numbers (WLLN)** if (2) holds for all $f \in L_1(P)$.

Properties Related to Laws of Large Numbers:

- ▶ If \mathcal{Z} is AMS then P is a probability measure on $X \times Y$. P is called the **asymptotical mean distribution** of \mathcal{Z} .
- ▶ If \mathcal{Z} satisfies WLLNE then it is AMS and for all bounded $f : X \times Y \rightarrow \mathbb{R}$ we have

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n f \circ Z_i = \mathbb{E}_P f \quad \text{in probability } \mu. \quad (2)$$

- ▶ Analogous results for SLLNE.
- ▶ \mathcal{Z} satisfies **weak law of large numbers (WLLN)** if (2) holds for all $f \in L_1(P)$.
- ▶ \mathcal{Z} satisfies **strong law of large numbers (SLLN)** if convergence in (2) is almost sure for all $f \in L_1(P)$.

Uncorrelated and independent processes:

- ▶ “Classical” weak law of large numbers.

If all image processes $\mathbf{1}_B \circ \mathcal{Z}$ are uncorrelated then:
 \mathcal{Z} satisfies WLLNE $\Leftrightarrow \mathcal{Z}$ is AMS.

Uncorrelated and independent processes:

- ▶ “Classical” weak law of large numbers.

If all image processes $\mathbf{1}_B \circ \mathcal{Z}$ are uncorrelated then:
 \mathcal{Z} satisfies WLLNE $\Leftrightarrow \mathcal{Z}$ is AMS.

- ▶ Classical strong law of large numbers.

\mathcal{Z} i.i.d. $\Rightarrow \mathcal{Z}$ satisfies SLLN.

Uncorrelated and independent processes:

- ▶ “Classical” weak law of large numbers.
If all image processes $\mathbf{1}_B \circ \mathcal{Z}$ are uncorrelated then:
 \mathcal{Z} satisfies WLLNE $\Leftrightarrow \mathcal{Z}$ is AMS.
- ▶ Classical strong law of large numbers.
 \mathcal{Z} i.i.d. $\Rightarrow \mathcal{Z}$ satisfies SLLN.
- ▶ If all image processes $\mathbf{1}_B \circ \mathcal{Z}$ are *independent* we have:
 \mathcal{Z} satisfies SLLNE $\Leftrightarrow \mathcal{Z}$ is AMS.

Uncorrelated and independent processes:

- ▶ “Classical” weak law of large numbers.
If all image processes $\mathbf{1}_B \circ \mathcal{Z}$ are uncorrelated then:
 \mathcal{Z} satisfies WLLNE $\Leftrightarrow \mathcal{Z}$ is AMS.
- ▶ Classical strong law of large numbers.
 \mathcal{Z} i.i.d. $\Rightarrow \mathcal{Z}$ satisfies SLLN.
- ▶ If all image processes $\mathbf{1}_B \circ \mathcal{Z}$ are *independent* we have:
 \mathcal{Z} satisfies SLLNE $\Leftrightarrow \mathcal{Z}$ is AMS.
- ▶ Etemadi, 1981:
 \mathcal{Z} identically distributed and pairwise independent
 $\Rightarrow \mathcal{Z}$ satisfies SLLN.
- ▶ Further generalizations (e.g. backward martingales) ...

Other Processes satisfying an LLN:

▶ **Birkhoff's theorem**

\mathcal{Z} stationary and ergodic $\Rightarrow \mathcal{Z}$ satisfies SLLN.

Other Processes satisfying an LLN:

- ▶ **Birkhoff's theorem**
 \mathcal{Z} stationary and ergodic $\Rightarrow \mathcal{Z}$ satisfies SLLN.
- ▶ \mathcal{Z} homogeneous Markov chain, then:
Doebelin condition \Rightarrow SLLN
- ▶ ...

Loss functions:

- ▶ $L : X \times Y \times \mathbb{R} \rightarrow [0, \infty)$ is called **loss function**. We say that

Loss functions:

- ▶ $L : X \times Y \times \mathbb{R} \rightarrow [0, \infty)$ is called **loss function**. We say that
 - ▶ continuous/convex if $L(x, y, \cdot) : \mathbb{R} \rightarrow [0, \infty)$ is so for all x, y .

Loss functions:

- ▶ $L : X \times Y \times \mathbb{R} \rightarrow [0, \infty)$ is called **loss function**. We say that
 - ▶ continuous/convex if $L(x, y, \cdot) : \mathbb{R} \rightarrow [0, \infty)$ is so for all x, y .
 - ▶ locally bounded if $L|_{X \times Y \times A}$ is bounded for all bounded $A \subset \mathbb{R}$.

Loss functions:

- ▶ $L : X \times Y \times \mathbb{R} \rightarrow [0, \infty)$ is called **loss function**. We say that
 - ▶ continuous/convex if $L(x, y, \cdot) : \mathbb{R} \rightarrow [0, \infty)$ is so for all x, y .
 - ▶ locally bounded if $L|_{X \times Y \times A}$ is bounded for all bounded $A \subset \mathbb{R}$.
 - ▶ **margin-based** if $Y = \{-1, 1\}$ and $L(x, y, t) = \varphi(yt)$.

Loss functions:

- ▶ $L : X \times Y \times \mathbb{R} \rightarrow [0, \infty)$ is called **loss function**. We say that
 - ▶ continuous/convex if $L(x, y, \cdot) : \mathbb{R} \rightarrow [0, \infty)$ is so for all x, y .
 - ▶ locally bounded if $L|_{X \times Y \times A}$ is bounded for all bounded $A \subset \mathbb{R}$.
 - ▶ **margin-based** if $Y = \{-1, 1\}$ and $L(x, y, t) = \varphi(yt)$.
 - ▶ **distance-based** if $Y = \mathbb{R}$ and $L(x, y, t) = \psi(y - t)$.

Loss functions:

- ▶ $L : X \times Y \times \mathbb{R} \rightarrow [0, \infty)$ is called **loss function**. We say that
 - ▶ continuous/convex if $L(x, y, \cdot) : \mathbb{R} \rightarrow [0, \infty)$ is so for all x, y .
 - ▶ locally bounded if $L|_{X \times Y \times A}$ is bounded for all bounded $A \subset \mathbb{R}$.
 - ▶ **margin-based** if $Y = \{-1, 1\}$ and $L(x, y, t) = \varphi(yt)$.
 - ▶ **distance-based** if $Y = \mathbb{R}$ and $L(x, y, t) = \psi(y - t)$.
- ▶ **Risk** of a function $f : X \rightarrow \mathbb{R}$ is

$$\mathcal{R}_{L,P}(f) := \int_{X \times Y} L(x, y, f(x)) dP(x, y).$$

Loss functions:

- ▶ $L : X \times Y \times \mathbb{R} \rightarrow [0, \infty)$ is called **loss function**. We say that
 - ▶ continuous/convex if $L(x, y, \cdot) : \mathbb{R} \rightarrow [0, \infty)$ is so for all x, y .
 - ▶ locally bounded if $L|_{X \times Y \times A}$ is bounded for all bounded $A \subset \mathbb{R}$.
 - ▶ **margin-based** if $Y = \{-1, 1\}$ and $L(x, y, t) = \varphi(yt)$.
 - ▶ **distance-based** if $Y = \mathbb{R}$ and $L(x, y, t) = \psi(y - t)$.
- ▶ **Risk** of a function $f : X \rightarrow \mathbb{R}$ is

$$\mathcal{R}_{L,P}(f) := \int_{X \times Y} L(x, y, f(x)) dP(x, y).$$

- ▶ **Bayes Risk** is

$$\mathcal{R}_{L,P}^* := \inf \{ \mathcal{R}_{L,P}(f) \mid f : X \rightarrow \mathbb{R} \}.$$

Consistency:

- ▶ $\mathcal{Z} = (X_i, Y_i)_{i \geq 1}$ satisfies WLLNE with asymptotic mean P and L is locally bounded loss. Then for all bounded $f : X \rightarrow \mathbb{R}$ and $n_0 \geq 0$

$$\mathcal{R}_{L,P}(f) = \lim_{n \rightarrow \infty} \frac{1}{n - n_0} \sum_{i=n_0+1}^n L(X_i, Y_i, f(X_i)) \quad \text{in probability } \mu.$$

Consistency:

- ▶ $\mathcal{Z} = (X_i, Y_i)_{i \geq 1}$ satisfies WLLNE with asymptotic mean P and L is locally bounded loss. Then for all bounded $f : X \rightarrow \mathbb{R}$ and $n_0 \geq 0$

$$\mathcal{R}_{L,P}(f) = \lim_{n \rightarrow \infty} \frac{1}{n - n_0} \sum_{i=n_0+1}^n L(X_i, Y_i, f(X_i)) \quad \text{in probability } \mu.$$

$\mathcal{R}_{L,P}(f)$ approximates future empirical error.

Consistency:

- ▶ $\mathcal{Z} = (X_i, Y_i)_{i \geq 1}$ satisfies WLLNE with asymptotic mean P and L is locally bounded loss. Then for all bounded $f : X \rightarrow \mathbb{R}$ and $n_0 \geq 0$

$$\mathcal{R}_{L,P}(f) = \lim_{n \rightarrow \infty} \frac{1}{n - n_0} \sum_{i=n_0+1}^n L(X_i, Y_i, f(X_i)) \quad \text{in probability } \mu.$$

$\mathcal{R}_{L,P}(f)$ approximates future empirical error.

- ▶ Learning method \mathcal{L} constructs to every $T \in (X \times Y)^n$ an $f_T : X \rightarrow \mathbb{R}$.

Consistency:

- ▶ $\mathcal{Z} = (X_i, Y_i)_{i \geq 1}$ satisfies WLLNE with asymptotic mean P and L is locally bounded loss. Then for all bounded $f : X \rightarrow \mathbb{R}$ and $n_0 \geq 0$

$$\mathcal{R}_{L,P}(f) = \lim_{n \rightarrow \infty} \frac{1}{n - n_0} \sum_{i=n_0+1}^n L(X_i, Y_i, f(X_i)) \quad \text{in probability } \mu.$$

$\mathcal{R}_{L,P}(f)$ approximates future empirical error.

- ▶ Learning method \mathcal{L} constructs to every $T \in (X \times Y)^n$ an $f_T : X \rightarrow \mathbb{R}$.
- ▶ **Consistency**
 \mathcal{L} is L -risk consistent for \mathcal{Z} if

$$\lim_{n \rightarrow \infty} \mathcal{R}_{L,P}(f_{T_n}) = \mathcal{R}_{L,P}^* \quad \text{in probability } \mu,$$

where $T_n := ((X_1, Y_1), \dots, (X_n, Y_n))$ for $n \geq 1$.

Consistency of penalized ERM algorithms I

- ▶ **Penalized ERM algorithm:**

Consistency of penalized ERM algorithms I

- ▶ **Penalized ERM algorithm:**
 - Fix a function class \mathcal{F} of functions $f : X \rightarrow \mathbb{R}$.

Consistency of penalized ERM algorithms I

- ▶ **Penalized ERM algorithm:**
 - Fix a function class \mathcal{F} of functions $f : X \rightarrow \mathbb{R}$.
 - Fix a penalty function $\text{pen} : (0, \infty) \times \mathcal{F} \rightarrow [0, \infty]$.

Consistency of penalized ERM algorithms I

► **Penalized ERM algorithm:**

$$f_{T_{\text{past}}, \lambda} \in \operatorname{argmin}_{f \in \mathcal{F}} \left(\operatorname{pen}(\lambda, f) + \mathcal{R}_{L, T_{\text{past}}}(f) \right)$$

Consistency of penalized ERM algorithms I

► **Penalized ERM algorithm:**

$$f_{T_{\text{past}},\lambda} \in \operatorname{argmin}_{f \in \mathcal{F}} \left(\operatorname{pen}(\lambda, f) + \mathcal{R}_{L, T_{\text{past}}}(f) \right)$$

Assumptions:

Consistency of penalized ERM algorithms I

► **Penalized ERM algorithm:**

$$f_{T_{\text{past}},\lambda} \in \operatorname{argmin}_{f \in \mathcal{F}} \left(\operatorname{pen}(\lambda, f) + \mathcal{R}_{L, T_{\text{past}}}(f) \right)$$

Assumptions:

- X compact metric space, $Y \subset \mathbb{R}$ bounded.

Consistency of penalized ERM algorithms I

► Penalized ERM algorithm:

$$f_{T_{\text{past}},\lambda} \in \operatorname{argmin}_{f \in \mathcal{F}} \left(\operatorname{pen}(\lambda, f) + \mathcal{R}_{L, T_{\text{past}}}(f) \right)$$

Assumptions:

- X compact metric space, $Y \subset \mathbb{R}$ bounded.
- L locally Lipschitz continuous and locally bounded loss.

Consistency of penalized ERM algorithms I

► **Penalized ERM algorithm:**

$$f_{T_{\text{past}}, \lambda} \in \operatorname{argmin}_{f \in \mathcal{F}} \left(\operatorname{pen}(\lambda, f) + \mathcal{R}_{L, T_{\text{past}}}(f) \right)$$

Assumptions:

- X compact metric space, $Y \subset \mathbb{R}$ bounded.
- L locally Lipschitz continuous and locally bounded loss.
- $\lim_{\lambda \rightarrow 0} \operatorname{pen}(\lambda, f) = 0$ and $\operatorname{pen}(\lambda, 0) = 0$.
 $\operatorname{pen}(\cdot, f) : (0, \infty) \rightarrow [0, \infty)$ increasing for all $f \in \mathcal{F}$.

Consistency of penalized ERM algorithms I

► **Penalized ERM algorithm:**

$$f_{T_{\text{past}}, \lambda} \in \operatorname{argmin}_{f \in \mathcal{F}} \left(\operatorname{pen}(\lambda, f) + \mathcal{R}_{L, T_{\text{past}}}(f) \right)$$

Assumptions:

- X compact metric space, $Y \subset \mathbb{R}$ bounded.
- L locally Lipschitz continuous and locally bounded loss.
- $\lim_{\lambda \rightarrow 0} \operatorname{pen}(\lambda, f) = 0$ and $\operatorname{pen}(\lambda, 0) = 0$.
 $\operatorname{pen}(\cdot, f) : (0, \infty) \rightarrow [0, \infty)$ increasing for all $f \in \mathcal{F}$.
- $\mathcal{R}_{L, Q}(0) \leq 1$ for all distributions Q on $X \times Y$.

Consistency of penalized ERM algorithms I

► Penalized ERM algorithm:

$$f_{T_{\text{past}}, \lambda} \in \operatorname{argmin}_{f \in \mathcal{F}} \left(\operatorname{pen}(\lambda, f) + \mathcal{R}_{L, T_{\text{past}}}(f) \right)$$

Assumptions:

- X compact metric space, $Y \subset \mathbb{R}$ bounded.
- L locally Lipschitz continuous and locally bounded loss.
- $\lim_{\lambda \rightarrow 0} \operatorname{pen}(\lambda, f) = 0$ and $\operatorname{pen}(\lambda, 0) = 0$.
 $\operatorname{pen}(\cdot, f) : (0, \infty) \rightarrow [0, \infty)$ increasing for all $f \in \mathcal{F}$.
- $\mathcal{R}_{L, Q}(0) \leq 1$ for all distributions Q on $X \times Y$.
- $\mathcal{F}_\lambda := \{f \in \mathcal{F} : \operatorname{pen}(\lambda, f) \leq 1\}$ is pre-compact subset in $C(X)$.

Complexity assumption

Consistency of penalized ERM algorithms I

► **Penalized ERM algorithm:**

$$f_{T_{\text{past}}, \lambda} \in \operatorname{argmin}_{f \in \mathcal{F}} \left(\operatorname{pen}(\lambda, f) + \mathcal{R}_{L, T_{\text{past}}}(f) \right)$$

Assumptions:

- X compact metric space, $Y \subset \mathbb{R}$ bounded.
- L locally Lipschitz continuous and locally bounded loss.
- $\lim_{\lambda \rightarrow 0} \operatorname{pen}(\lambda, f) = 0$ and $\operatorname{pen}(\lambda, 0) = 0$.
 $\operatorname{pen}(\cdot, f) : (0, \infty) \rightarrow [0, \infty)$ increasing for all $f \in \mathcal{F}$.
- $\mathcal{R}_{L, Q}(0) \leq 1$ for all distributions Q on $X \times Y$.
- $\mathcal{F}_\lambda := \{f \in \mathcal{F} : \operatorname{pen}(\lambda, f) \leq 1\}$ is pre-compact subset in $C(X)$.
- $\inf_{f \in \mathcal{F}} \mathcal{R}_{L, P}(f) = \mathcal{R}_{L, P}^*$.

Approximation assumption

Consistency of penalized ERM algorithms II

- ▶ **Consistency of Penalized ERM algorithm:**

Consistency of penalized ERM algorithms II

- ▶ **Consistency of Penalized ERM algorithm:**
 - ▶ Let \mathcal{L} be the penalized ERM algorithm described above.

Consistency of penalized ERM algorithms II

- ▶ **Consistency of Penalized ERM algorithm:**
 - ▶ Let \mathcal{L} be the penalized ERM algorithm described above.
 - ▶ $\mathcal{Z} = (X_i, Y_i)_{i \geq 1}$ stochastic process that satisfies WLLNE.

Consistency of penalized ERM algorithms II

► Consistency of Penalized ERM algorithm:

- Let \mathcal{L} be the penalized ERM algorithm described above.
- $\mathcal{Z} = (X_i, Y_i)_{i \geq 1}$ stochastic process that satisfies WLLNE.

Then there exists a positive null sequence (γ_n) such that for all nullsequences (λ_n) with $\lambda_n \geq \gamma_n$ the penalized ERM with regularization sequence (λ_n) is L -risk consistent for \mathcal{Z} .

Consistency of penalized ERM algorithms II

► Consistency of Penalized ERM algorithm:

- Let \mathcal{L} be the penalized ERM algorithm described above.
- $\mathcal{Z} = (X_i, Y_i)_{i \geq 1}$ stochastic process that satisfies WLLNE.

Then there exists a positive null sequence (γ_n) such that for all nullsequences (λ_n) with $\lambda_n \geq \gamma_n$ the penalized ERM with regularization sequence (λ_n) is L -risk consistent for \mathcal{Z} . I.e.

$$\lim_{n \rightarrow \infty} \mathcal{R}_{L,P}(f_{T_n}, \lambda_n) = \mathcal{R}_{L,P}^* \quad \text{in probability } \mu,$$

where $T_n := ((X_1, Y_1), \dots, (X_n, Y_n))$ for $n \geq 1$.

Idea of the Proof I

Preparations:

Idea of the Proof I

Preparations:

▶ $f_{T,\lambda_n} \in \mathcal{F}_{\gamma_n}$

Idea of the Proof I

Preparations:

- ▶ $f_{T,\lambda_n} \in \mathcal{F}_{\gamma_n}$
- ▶ For $\mathcal{R}_{L,P,\lambda}^* := \inf_{f \in \mathcal{F}} \text{pen}(\lambda, f) + \mathcal{R}_{L,P}(f)$ we have

$$\lim_{\lambda \rightarrow 0} \mathcal{R}_{L,P,\lambda}^* = \mathcal{R}_{L,P}^*.$$

Idea of the Proof I

Preparations:

- ▶ $f_{T,\lambda_n} \in \mathcal{F}_{\gamma_n}$
- ▶ For $\mathcal{R}_{L,P,\lambda}^* := \inf_{f \in \mathcal{F}} \text{pen}(\lambda, f) + \mathcal{R}_{L,P}(f)$ we have

$$\lim_{\lambda \rightarrow 0} \mathcal{R}_{L,P,\lambda}^* = \mathcal{R}_{L,P}^*.$$

- ▶ There exists a finite ε -net \mathcal{G}_λ of \mathcal{F}_λ .

Idea of the Proof I

Preparations:

- ▶ $f_{T,\lambda_n} \in \mathcal{F}_{\gamma_n}$
- ▶ For $\mathcal{R}_{L,P,\lambda}^* := \inf_{f \in \mathcal{F}} \text{pen}(\lambda, f) + \mathcal{R}_{L,P}(f)$ we have

$$\lim_{\lambda \rightarrow 0} \mathcal{R}_{L,P,\lambda}^* = \mathcal{R}_{L,P}^*.$$

- ▶ There exists a finite ε -net \mathcal{G}_λ of \mathcal{F}_λ .
- ▶ Standard ε -argument shows

$$\sup_{f \in \mathcal{F}_\lambda} |\mathcal{R}_{L,T_n}(f) - \mathcal{R}_{L,P}(f)| \leq 2\varepsilon + \sup_{g \in \mathcal{G}_\lambda} |\mathcal{R}_{L,T_n}(g) - \mathcal{R}_{L,P}(g)|$$

Idea of the Proof I

Preparations:

- ▶ $f_{T,\lambda_n} \in \mathcal{F}_{\gamma_n}$
- ▶ For $\mathcal{R}_{L,P,\lambda}^* := \inf_{f \in \mathcal{F}} \text{pen}(\lambda, f) + \mathcal{R}_{L,P}(f)$ we have

$$\lim_{\lambda \rightarrow 0} \mathcal{R}_{L,P,\lambda}^* = \mathcal{R}_{L,P}^*.$$

- ▶ There exists a finite ε -net \mathcal{G}_λ of \mathcal{F}_λ .
- ▶ Standard ε -argument shows

$$\sup_{f \in \mathcal{F}_\lambda} |\mathcal{R}_{L,T_n}(f) - \mathcal{R}_{L,P}(f)| \leq 2\varepsilon + \sup_{g \in \mathcal{G}_\lambda} |\mathcal{R}_{L,T_n}(g) - \mathcal{R}_{L,P}(g)|$$

- ▶ WLLNE and union bound gives

$$\lim_{n \rightarrow \infty} \sup_{g \in \mathcal{G}_\lambda} |\mathcal{R}_{L,T_n}(g) - \mathcal{R}_{L,P}(g)| = 0 \quad \text{in probability } \mu.$$

Idea of the Proof II

- ▶ Putting everything together by standard argument shows:

Idea of the Proof II

- ▶ Putting everything together by standard argument shows:
for all $\varepsilon > 0$ and $\gamma > 0$ we have

$$\lim_{n \rightarrow \infty} \mu \left(\left\{ \omega \in \Omega : \sup_{\lambda \geq \gamma} \left| \mathcal{R}_{L,P}(f_{T_n(\omega),\lambda}) - \mathcal{R}_{L,P,\lambda}^* \right| \geq \varepsilon \right\} \right) = 0$$

Idea of the Proof II

- ▶ Putting everything together by standard argument shows:
for all $\varepsilon > 0$ and $\gamma > 0$ we have

$$\lim_{n \rightarrow \infty} \mu \left(\left\{ \omega \in \Omega : \sup_{\lambda \geq \gamma} \left| \mathcal{R}_{L,P}(f_{T_n(\omega), \lambda}) - \mathcal{R}_{L,P,\lambda}^* \right| \geq \varepsilon \right\} \right) = 0$$

- ▶ Now the assertion follows from a **Selection Lemma**:
Let $F : (0, \infty) \times \mathbb{N} \rightarrow [0, \infty)$ be a function with
 $\lim_{n \rightarrow \infty} F(\gamma, n) = 0$ for all $\gamma > 0$. Then there exists a
sequence $(\gamma_n) \subset (0, 1]$ with

$$\lim_{n \rightarrow \infty} \gamma_n = 0$$

and

$$\lim_{n \rightarrow \infty} F(\gamma_n, n) = 0.$$

Consistency of penalized ERM algorithms II

► **Consistency of Penalized ERM algorithm:**

Then there exists a positive null sequence (γ_n) such that for all nullsequences (λ_n) with $\lambda_n \geq \gamma_n$ the penalized ERM with regularization sequence (λ_n) is L -risk consistent for \mathcal{Z} .

Consistency of penalized ERM algorithms II

► **Consistency of Penalized ERM algorithm:**

Then there exists a positive null sequence (γ_n) such that for all nullsequences (λ_n) with $\lambda_n \geq \gamma_n$ the penalized ERM with regularization sequence (λ_n) is L -risk consistent for \mathcal{Z} .

► **Nobel, 1999:**

There exists no learning method that is least square or classification consistent for all stationary ergodic processes.

Consistency of penalized ERM algorithms II

- ▶ **Consistency of Penalized ERM algorithm:**

Then there exists a positive null sequence (γ_n) such that for all nullsequences (λ_n) with $\lambda_n \geq \gamma_n$ the penalized ERM with regularization sequence (λ_n) is L -risk consistent for \mathcal{Z} .

- ▶ **Nobel, 1999:**

There exists no learning method that is least square or classification consistent for all stationary ergodic processes.

- ▶ **Consequences**

Consistency of penalized ERM algorithms II

- ▶ **Consistency of Penalized ERM algorithm:**

Then there exists a positive null sequence (γ_n) such that for all nullsequences (λ_n) with $\lambda_n \geq \gamma_n$ the penalized ERM with regularization sequence (λ_n) is L -risk consistent for \mathcal{Z} .

- ▶ **Nobel, 1999:**

There exists no learning method that is least square or classification consistent for all stationary ergodic processes.

- ▶ **Consequences**

- ▶ There exists no universal sequence (γ_n) .

Consistency of penalized ERM algorithms II

► Consistency of Penalized ERM algorithm:

Then there exists a positive null sequence (γ_n) such that for all nullsequences (λ_n) with $\lambda_n \geq \gamma_n$ the penalized ERM with regularization sequence (λ_n) is L -risk consistent for \mathcal{Z} .

► Nobel, 1999:

There exists no learning method that is least square or classification consistent for all stationary ergodic processes.

► Consequences

- There exists no universal sequence (γ_n) .
- In order to find “universal” sequences for *classes* of processes we need suitable concentration inequalities.

Mixing Coefficients

- ▶ \mathcal{A} and \mathcal{B} be two σ -algebras on Ω .

Mixing Coefficients

- ▶ \mathcal{A} and \mathcal{B} be two σ -algebras on Ω .
- ▶ μ be a probability measure on $\sigma(\mathcal{A} \cup \mathcal{B})$.

Mixing Coefficients

- ▶ \mathcal{A} and \mathcal{B} be two σ -algebras on Ω .
- ▶ μ be a probability measure on $\sigma(\mathcal{A} \cup \mathcal{B})$.
- ▶ **α -mixing coefficient**

$$\alpha(\mathcal{A}, \mathcal{B}, \mu) := \sup_{\substack{A \in \mathcal{A} \\ B \in \mathcal{B}}} |\mu(A \cap B) - \mu(A)\mu(B)|$$

Mixing Coefficients

- ▶ \mathcal{A} and \mathcal{B} be two σ -algebras on Ω .
- ▶ μ be a probability measure on $\sigma(\mathcal{A} \cup \mathcal{B})$.
- ▶ **α -mixing coefficient**

$$\alpha(\mathcal{A}, \mathcal{B}, \mu) := \sup_{\substack{A \in \mathcal{A} \\ B \in \mathcal{B}}} |\mu(A \cap B) - \mu(A)\mu(B)|$$

- ▶ **β -mixing coefficient**

$$\beta(\mathcal{A}, \mathcal{B}, \mu) := \mathbb{E}_{\mu} \sup_{B \in \mathcal{B}} |\mu(B) - \mathbb{E}_{\mu}(B|\mathcal{A})|$$

Mixing Coefficients

- ▶ \mathcal{A} and \mathcal{B} be two σ -algebras on Ω .
- ▶ μ be a probability measure on $\sigma(\mathcal{A} \cup \mathcal{B})$.
- ▶ **α -mixing coefficient**

$$\alpha(\mathcal{A}, \mathcal{B}, \mu) := \sup_{\substack{A \in \mathcal{A} \\ B \in \mathcal{B}}} |\mu(A \cap B) - \mu(A)\mu(B)|$$

- ▶ **β -mixing coefficient**

$$\beta(\mathcal{A}, \mathcal{B}, \mu) := \mathbb{E}_\mu \sup_{B \in \mathcal{B}} |\mu(B) - \mathbb{E}_\mu(B|\mathcal{A})|$$

- ▶ **φ -mixing coefficient**

$$\varphi(\mathcal{A}, \mathcal{B}, \mu) := \sup_{\substack{A \in \mathcal{A} \\ B \in \mathcal{B}}} \left| \frac{\mu(A \cap B) - \mu(A)\mu(B)}{\mu(A)} \right|$$

Mixing Coefficients

- ▶ \mathcal{A} and \mathcal{B} be two σ -algebras on Ω .
- ▶ μ be a probability measure on $\sigma(\mathcal{A} \cup \mathcal{B})$.
- ▶ **α -mixing coefficient**

$$\alpha(\mathcal{A}, \mathcal{B}, \mu) := \sup_{\substack{A \in \mathcal{A} \\ B \in \mathcal{B}}} |\mu(A \cap B) - \mu(A)\mu(B)|$$

- ▶ **β -mixing coefficient**

$$\beta(\mathcal{A}, \mathcal{B}, \mu) := \mathbb{E}_\mu \sup_{B \in \mathcal{B}} |\mu(B) - \mathbb{E}_\mu(B|\mathcal{A})|$$

- ▶ **φ -mixing coefficient**

$$\varphi(\mathcal{A}, \mathcal{B}, \mu) := \sup_{\substack{A \in \mathcal{A} \\ B \in \mathcal{B}}} \left| \frac{\mu(A \cap B) - \mu(A)\mu(B)}{\mu(A)} \right|$$

- ▶ $2\alpha(\mathcal{A}, \mathcal{B}, \mu) \leq \beta(\mathcal{A}, \mathcal{B}, \mu) \leq \varphi(\mathcal{A}, \mathcal{B}, \mu)$

Mixing Coefficients for processes

- ▶ \mathcal{Z} stochastic process

Mixing Coefficients for processes

- ▶ \mathcal{Z} stochastic process
- ▶ ξ one of the above mixing coefficients.

Mixing Coefficients for processes

- ▶ \mathcal{Z} stochastic process
- ▶ ξ one of the above mixing coefficients.
- ▶ ξ -bi-mixing coefficient of \mathcal{Z} :

$$\xi(\mathcal{Z}, \mu, i, j) := \xi(\sigma(Z_i), \sigma(Z_j), \mu), \quad i, j \geq 1.$$

Mixing Coefficients for processes

- ▶ \mathcal{Z} stochastic process
- ▶ ξ one of the above mixing coefficients.
- ▶ ξ -bi-mixing coefficient of \mathcal{Z} :

$$\xi(\mathcal{Z}, \mu, i, j) := \xi(\sigma(Z_i), \sigma(Z_j), \mu), \quad i, j \geq 1.$$

- ▶ ξ -mixing coefficient of \mathcal{Z} :

$$\xi(\mathcal{Z}, \mu, n) := \sup_{i \geq 1} \xi(\sigma(Z_i), \sigma(Z_{i+n}), \mu), \quad n \geq 1.$$

Mixing Coefficients for processes

- ▶ \mathcal{Z} stochastic process
- ▶ ξ one of the above mixing coefficients.
- ▶ ξ -bi-mixing coefficient of \mathcal{Z} :

$$\xi(\mathcal{Z}, \mu, i, j) := \xi(\sigma(Z_i), \sigma(Z_j), \mu), \quad i, j \geq 1.$$

- ▶ ξ -mixing coefficient of \mathcal{Z} :

$$\xi(\mathcal{Z}, \mu, n) := \sup_{i \geq 1} \xi(\sigma(Z_i), \sigma(Z_{i+n}), \mu), \quad n \geq 1.$$

- ▶ classical ξ -mixing coefficient of \mathcal{Z} :

$$\xi(\mathcal{Z}, \mu, n) := \sup_{i \geq 1} \xi(\sigma(Z_1, \dots, Z_i), \sigma(Z_{i+n}, Z_{i+1+n}, \dots), \mu), \quad n \geq 1.$$

Remarks on Mixing Coefficients for processes

- ▶ If \mathcal{Z} stationary, homogenous Markov chain then

$$\xi(\mathcal{Z}, \mu, n) = \bar{\xi}(\mathcal{Z}, \mu, n)$$

Remarks on Mixing Coefficients for processes

- ▶ If \mathcal{Z} stationary, homogenous Markov chain then

$$\xi(\mathcal{Z}, \mu, n) = \bar{\xi}(\mathcal{Z}, \mu, n)$$

- ▶ For Markov chains we often have exponential decay.

Remarks on Mixing Coefficients for processes

- ▶ If \mathcal{Z} stationary, homogenous Markov chain then

$$\xi(\mathcal{Z}, \mu, n) = \bar{\xi}(\mathcal{Z}, \mu, n)$$

- ▶ For Markov chains we often have exponential decay.
- ▶ For many other types of processes mixing properties are known. Examples include ARMA, GARCH, ...

Remarks on Mixing Coefficients for processes

- ▶ If \mathcal{Z} stationary, homogenous Markov chain then

$$\xi(\mathcal{Z}, \mu, n) = \bar{\xi}(\mathcal{Z}, \mu, n)$$

- ▶ For Markov chains we often have exponential decay.
- ▶ For many other types of processes mixing properties are known. Examples include ARMA, GARCH, ...
- ▶ There exist stationary processes with an arbitrarily slow decay of mixing coefficients.

Remarks on Mixing Coefficients for processes

- ▶ If \mathcal{Z} stationary, homogenous Markov chain then

$$\xi(\mathcal{Z}, \mu, n) = \bar{\xi}(\mathcal{Z}, \mu, n)$$

- ▶ For Markov chains we often have exponential decay.
- ▶ For many other types of processes mixing properties are known. Examples include ARMA, GARCH, ...
- ▶ There exist stationary processes with an arbitrarily slow decay of mixing coefficients.
- ▶ If we have

$$\lim_{n \rightarrow \infty} \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^{i-1} \alpha(\mathcal{Z}, \mu, i, j) = 0$$

then \mathcal{Z} satisfies WLLNE $\Leftrightarrow \mathcal{Z}$ is AMS.

Support Vector Machines:

- ▶ Support vector machines (SVMs) solve the problem

$$\arg \min_{f \in H} \lambda \|f\|_H^2 + \frac{1}{n} \sum_{i=1}^n L(x_i, y_i, f(x_i)) , \quad (3)$$

where

Support Vector Machines:

- ▶ Support vector machines (SVMs) solve the problem

$$\arg \min_{f \in H} \lambda \|f\|_H^2 + \frac{1}{n} \sum_{i=1}^n L(x_i, y_i, f(x_i)) , \quad (3)$$

where

- ▶ H is a RKHS,

Support Vector Machines:

- ▶ Support vector machines (SVMs) solve the problem

$$\arg \min_{f \in H} \lambda \|f\|_H^2 + \frac{1}{n} \sum_{i=1}^n L(x_i, y_i, f(x_i)) , \quad (3)$$

where

- ▶ H is a RKHS,
- ▶ $T = ((x_1, y_1), \dots, (x_n, y_n)) \in (X \times Y)^n$ is a training set,

Support Vector Machines:

- ▶ Support vector machines (SVMs) solve the problem

$$\arg \min_{f \in H} \lambda \|f\|_H^2 + \frac{1}{n} \sum_{i=1}^n L(x_i, y_i, f(x_i)) , \quad (3)$$

where

- ▶ H is a RKHS,
- ▶ $T = ((x_1, y_1), \dots, (x_n, y_n)) \in (X \times Y)^n$ is a training set,
- ▶ $\lambda > 0$ is a *free* regularization parameter,

Support Vector Machines:

- ▶ Support vector machines (SVMs) solve the problem

$$\arg \min_{f \in H} \lambda \|f\|_H^2 + \frac{1}{n} \sum_{i=1}^n L(x_i, y_i, f(x_i)) , \quad (3)$$

where

- ▶ H is a RKHS,
- ▶ $T = ((x_1, y_1), \dots, (x_n, y_n)) \in (X \times Y)^n$ is a training set,
- ▶ $\lambda > 0$ is a *free* regularization parameter,
- ▶ L is a loss function.

Support Vector Machines:

- ▶ Support vector machines (SVMs) solve the problem

$$\arg \min_{f \in H} \lambda \|f\|_H^2 + \frac{1}{n} \sum_{i=1}^n L(x_i, y_i, f(x_i)) , \quad (3)$$

where

- ▶ H is a RKHS,
 - ▶ $T = ((x_1, y_1), \dots, (x_n, y_n)) \in (X \times Y)^n$ is a training set,
 - ▶ $\lambda > 0$ is a *free* regularization parameter,
 - ▶ L is a loss function.
- ▶ For simplicity we only consider:

Support Vector Machines:

- ▶ Support vector machines (SVMs) solve the problem

$$\arg \min_{f \in H} \lambda \|f\|_H^2 + \frac{1}{n} \sum_{i=1}^n L(x_i, y_i, f(x_i)) , \quad (3)$$

where

- ▶ H is a RKHS,
 - ▶ $T = ((x_1, y_1), \dots, (x_n, y_n)) \in (X \times Y)^n$ is a training set,
 - ▶ $\lambda > 0$ is a *free* regularization parameter,
 - ▶ L is a loss function.
- ▶ For simplicity we only consider:
 - ▶ H RKHS of Gaussian kernel over \mathbb{R}^d .

Support Vector Machines:

- ▶ Support vector machines (SVMs) solve the problem

$$\arg \min_{f \in H} \lambda \|f\|_H^2 + \frac{1}{n} \sum_{i=1}^n L(x_i, y_i, f(x_i)) , \quad (3)$$

where

- ▶ H is a RKHS,
 - ▶ $T = ((x_1, y_1), \dots, (x_n, y_n)) \in (X \times Y)^n$ is a training set,
 - ▶ $\lambda > 0$ is a *free* regularization parameter,
 - ▶ L is a loss function.
- ▶ For simplicity we only consider:
 - ▶ H RKHS of Gaussian kernel over \mathbb{R}^d .
 - ▶ L hinge loss, i.e. $Y = \{-1, 1\}$ and $L(x, y, t) = \max\{0, 1 - yt\}$.

Consistency for Mixing Coefficients for processes

Main Theorem (simplified)

Assume that there are constants $C > 0$ and $\alpha \in (0, 1]$ with

$$\left| \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{\mu} f \circ Z_i - \mathbb{E}_P f \right| \leq C \|f\|_{\infty} n^{-\alpha}$$
$$\frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^{i-1} \alpha(\mathcal{Z}, \mu, i, j) \leq C n^{-\alpha}$$

Consistency for Mixing Coefficients for processes

Main Theorem (simplified)

Assume that there are constants $C > 0$ and $\alpha \in (0, 1]$ with

$$\left| \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{\mu} f \circ Z_i - \mathbb{E}_P f \right| \leq C \|f\|_{\infty} n^{-\alpha}$$

$$\frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^{i-1} \alpha(\mathcal{Z}, \mu, i, j) \leq C n^{-\alpha}$$

Then the SVM with regularization (λ_n) is classification consistent if $\lambda_n \rightarrow 0$ and $\lambda_n^2 n^{\alpha} \rightarrow \infty$.

Idea of the proof I

► **Stability:**

For all $\lambda > 0$ there exists a function $h_\lambda : X \times Y \rightarrow [-1, 1]$ such that for all T we have

$$\|f_{P,\lambda} - f_{T,\lambda}\|_H \leq \frac{1}{\lambda} \|\mathbb{E}_P h_\lambda \Phi - \mathbb{E}_T h_\lambda \Phi\|_H,$$

where $\Phi : X \rightarrow H$ is canonical feature map.

Idea of the proof I

► **Stability:**

For all $\lambda > 0$ there exists a function $h_\lambda : X \times Y \rightarrow [-1, 1]$ such that for all T we have

$$\|f_{P,\lambda} - f_{T,\lambda}\|_H \leq \frac{1}{\lambda} \|\mathbb{E}_P h_\lambda \Phi - \mathbb{E}_T h_\lambda \Phi\|_H,$$

where $\Phi : X \rightarrow H$ is canonical feature map.

► **Markov inequality:**

$$\begin{aligned} & \mu \left(\left\{ \omega \in \Omega : \|\mathbb{E}_{T_n(\omega)} h_\lambda \Phi - \mathbb{E}_{P_n} h_\lambda \Phi\|_H \geq \varepsilon \lambda_n \right\} \right) \\ & \leq \frac{1}{\varepsilon^2 \lambda_n^2} \mathbb{E}_{\omega \sim \mu} \|\mathbb{E}_{T_n(\omega)} h_\lambda \Phi - \mathbb{E}_{P_n} h_\lambda \Phi\|_H^2 \end{aligned}$$

Idea of the proof II

- ▶ Write

$$g_{n,i} := (h_{\lambda_n} \Phi) \circ (X_i, Y_i) - \mathbb{E}_\mu(h_{\lambda_n} \Phi) \circ (X_i, Y_i)$$

Idea of the proof II

- Write

$$g_{n,i} := (h_{\lambda_n} \Phi) \circ (X_i, Y_i) - \mathbb{E}_\mu(h_{\lambda_n} \Phi) \circ (X_i, Y_i)$$

- Hilbert space norm:**

$$\begin{aligned} & \mathbb{E}_{\omega \sim \mu} \|\mathbb{E}_{T_n(\omega)} h_n \Phi - \mathbb{E}_{P_n} h_n \Phi\|_H^2 \\ = & n^{-2} \sum_{i=1}^n \|g_{n,i}\|_\infty^2 + 2n^{-2} \sum_{i=1}^n \sum_{j=1}^{i-1} \mathbb{E}_\mu \langle g_{n,i}, g_{n,j} \rangle \\ \leq & 2n^{-1} + 2n^{-2} \sum_{i=1}^n \sum_{j=1}^{i-1} \mathbb{E}_\mu \langle g_{n,i}, g_{n,j} \rangle \end{aligned}$$

Idea of the proof II

- Write

$$g_{n,i} := (h_{\lambda_n} \Phi) \circ (X_i, Y_i) - \mathbb{E}_\mu(h_{\lambda_n} \Phi) \circ (X_i, Y_i)$$

- Hilbert space norm:**

$$\begin{aligned} & \mathbb{E}_{\omega \sim \mu} \|\mathbb{E}_{T_n(\omega)} h_n \Phi - \mathbb{E}_{P_n} h_n \Phi\|_H^2 \\ = & n^{-2} \sum_{i=1}^n \|g_{n,i}\|_\infty^2 + 2n^{-2} \sum_{i=1}^n \sum_{j=1}^{i-1} \mathbb{E}_\mu \langle g_{n,i}, g_{n,j} \rangle \\ \leq & 2n^{-1} + 2n^{-2} \sum_{i=1}^n \sum_{j=1}^{i-1} \mathbb{E}_\mu \langle g_{n,i}, g_{n,j} \rangle \end{aligned}$$

- Grothendieck inequality unsures:**

$$\mathbb{E}_\mu \langle g_{n,i}, g_{n,j} \rangle \leq c \cdot \alpha(\mathcal{Z}, \mu, i, j)$$

Remarks

- ▶ (The few) previous results require stronger conditions in terms of classical mixing conditions.

Remarks

- ▶ (The few) previous results require stronger conditions in terms of classical mixing conditions.
- ▶ Polynomial mixing condition can be weakened.

Remarks

- ▶ (The few) previous results require stronger conditions in terms of classical mixing conditions.
- ▶ Polynomial mixing condition can be weakened.
- ▶ Reasonable rates are possible by employing stronger concentration inequalities.

Remarks

- ▶ (The few) previous results require stronger conditions in terms of classical mixing conditions.
- ▶ Polynomial mixing condition can be weakened.
- ▶ Reasonable rates are possible by employing stronger concentration inequalities.
- ▶ A similar result holds for other loss functions and bounded noise.

Remarks

- ▶ (The few) previous results require stronger conditions in terms of classical mixing conditions.
- ▶ Polynomial mixing condition can be weakened.
- ▶ Reasonable rates are possible by employing stronger concentration inequalities.
- ▶ A similar result holds for other loss functions and bounded noise.
- ▶ A similar result holds for distance-based losses of growth-type $p \in [1, 2]$ if p -th moment of noise is finite.