Two sided empirical process

$$\xi^l = \sup_{\alpha \in \Lambda} \left| \int q(z,\alpha) \, dF(z) - \frac{1}{l} \sum_{i=1}^{l} q(z_i, \alpha) \right|$$

Uniform two-sided convergence

$$\lim_{l \to \infty} \mathbb{P} \left\{ \sup_{\alpha \in \Lambda} \left| \int q(z,\alpha) \, dF(z) - \frac{1}{l} \sum_{i=1}^{l} q(z_i, \alpha) \right| > \varepsilon \right\} = 0$$

One-sided empirical process

$$\xi_+^l = \sup_{\alpha \in \Lambda} \left( \int q(z,\alpha) \, dF(z) - \frac{1}{l} \sum_{i=1}^{l} q(z_i, \alpha) \right)$$

Uniform one-sided convergence

$$\lim_{l \to \infty} \mathbb{P} \left( \sup_{\alpha \in \Lambda} \left( \int q(z,\alpha) \, dF(z) - \frac{1}{l} \sum_{i=1}^{l} q(z_i, \alpha) \right) > \varepsilon \right) = 0$$

\(\forall \varepsilon > 0\)
Random entropy \[ H^\wedge(Z_1, \ldots, Z_e) = \ln N^\wedge(Z_1, \ldots, Z_e) \]

Entropy of the set of indicator functions \( \Omega(Z, \alpha) \) \( \alpha \in \Lambda \) on samples of size \( e \)

\[ H^\wedge(\mathcal{E}) = \mathbb{E} \ln N^\wedge(Z_1, \ldots, Z_e) \]

- The general case and VC-entropy

**Definition.** Let \( \Lambda \subseteq \Omega(Z, \alpha) \subseteq B \), \( \alpha \in \Lambda \) be a set of bounded loss functions. Using this set of functions and the training set \( Z_1, \ldots, Z_e \) one can construct the following set of \( e \) dimensional vectors

\[ q(\alpha) = (\Omega(Z_1, \alpha), \ldots, \Omega(Z_e, \alpha)) \alpha \in \Lambda \]

This set of vectors belongs to the \( e \) dimensional cube and has a finite minimal \( \varepsilon \)-net in the metric \( C^0 \) (or \( L_p \)). Let \( N = N^\wedge(\varepsilon; Z_1, \ldots, Z_e) \) be the number of elements of the minimal \( \varepsilon \)-net of this set of vectors \( q(\alpha), \alpha \in \Lambda \).
Random VC entropy

\[ H^A(\varepsilon; z, \ldots, z\varepsilon) = \ln N^A(\varepsilon; z, \ldots, z\varepsilon) \]

VC entropy

\[ H^A(\varepsilon; z) = \prod H^A(\varepsilon; z, \ldots, z\varepsilon) \]

Conditions for uniform two-sided convergence

**Theorem** For uniform two-sided convergence it is necessary and sufficient that

\[ \lim_{\varepsilon \to \infty} \frac{H^A(\varepsilon; \varepsilon)}{\varepsilon} = 0 \quad \forall \varepsilon > 0 \]

Conditions for uniform one-sided convergence

**Theorem** In order for uniform one-sided convergence of empirical means to their expectations to hold for the set of totally bounded functions \( Q(\varepsilon, x) \), it is necessary and sufficient that for any positive \( s, \eta \) and \( \varepsilon \) there exists a set of functions \( Q^*(\varepsilon, x^*) \), \( x^* \in \Lambda^* \) satisfying

\[ Q(\varepsilon, x) - Q^*(\varepsilon, x^*) \geq 0 \quad \forall x \]

\[ \int (Q(\varepsilon, x) - Q^*(\varepsilon, x^*)) dF(x) \leq s \]
such that
\[
\lim_{l \to \infty} \frac{H^* (\varepsilon, l)}{l} < \eta
\]
holds for the \( \varepsilon \)-entropy of the set \( \Omega^* (\varepsilon, x), x \in \Gamma^* \)
on samples of size \( l \).

**Question** What if \( \lim_{l \to \infty} \frac{H^* (\varepsilon_0, l)}{l} \neq 0 \) for some \( \varepsilon_0 \)?

ERM does not hold \( \Rightarrow \) non-consistent
But Why non-consistent?

Theory of Non-falsifiability in Philosophy

Kant's problem of Demarcation

(i). Deductive: General \( \Rightarrow \) Particular
(ii). Inductive: Particular \( \Rightarrow \) General

Kant's problem: What is the difference between the cases of a justified inductive step, and those for which the inductive step is not justified?

All Natural Science is a result of inductive inference.
So Kant's problem = Is there a formal way to distinguish true theories and false theories?
A necessary condition for justifiability of a theory is the feasibility of its falsification.

\textit{falsification = existence of a collection of particular assertions which cannot be explained by the given theory although they fall into its domain.}

If a given theory can be falsified \( \Rightarrow \) \textit{scientific}
If a given theory cannot be falsified \( \Rightarrow \) \textit{non-scientific}

\( > \) Complete \( \textit{Non-falsifiability} \)

We know for indicator functions
\[
H^\uparrow(l) = \mathbb{E} \ln N^\uparrow(z_1, \ldots, z_l), \quad N^\uparrow(z_1, \ldots, z_l) \leq 2^l
\]

If \( \lim_{l \to \infty} \frac{H^\uparrow(l)}{l} = \ln 2 \) (maximal entropy)

Then for almost all samples \( z_1, \ldots, z_l \) we have
\[
N^\uparrow(z_1, \ldots, z_l) = 2^l
\]

That is, almost all sample \( z_1, \ldots, z_l \) of arbitrary size \( l \) can be separated in all possible ways by functions
of this set, so minimum of empirical risk for this machine is 0.

Such a machine can give a general explanation for almost any data.

Non-falsifiable machine → Not scientific.

Minimal value of the empirical risk is 0 independent of the expected risk.

That is why we need variance and bias!

> Partial Non-falsifiability

Theorem For the set of indicator functions \( \lambda(z, \alpha) \), \( x \in A \) let the convergence

\[
\lim_{d \to \infty} \frac{h^d}{d} = c > 0
\]

be valid. Then there exists a subset \( z^* \) of the set \( z \) for which \( P(z^*) = \alpha(c) \neq 0 \) and for almost any training set \( z_1, \ldots, z_e \) we have \( z_1^*, \ldots, z_k^* = \{z_1, \ldots, z_e\} \cup z^* \) and any sequence of binary values \( s_1, \ldots, s_k \) \( s_i \in \{0, 1\} \).
there exists a function $Q(z, x^*)$ for which
define $y_i = Q(z_i, x^*)$ for $i = 1, 2, \ldots, k$
holds true.

So if the conditions of uniform two-sided convergence fail, then the exists some subspace of the input space where the learning machine is non-falsifiable.

There is also a more sophisticated "potentially non-falsifiable" related to general VC entropy that I will skip here.

Three milestones in learning theory.

Consider the case of indicator functions $Q(z, x)$

VC entropy $H^V(l) = \frac{1}{l} \ln N^V(z_1, \ldots, Z_2)$

Annealed VC entropy $H_{\text{ann}}^V(l) = \bar{H} \ln N^V(z_1, \ldots, Z_2)$

Growth function $G^V(l) = \ln \sup_{z_1 \ldots Z_l} N^V(z_1, \ldots, Z_2)$

We have $H^V(l) \leq H_{\text{ann}}^V(l) \leq G^V(l)$
First milestone in learning theory

\[
\lim_{d \to \infty} \frac{H^d(l)}{d} = 0 \implies \text{Consistency of the ERM principle}
\]

Second milestone in learning theory

\[
\lim_{d \to \infty} \frac{H_{\text{ann}}^d(l)}{d} = 0 \implies \text{fast convergence of } R(l^d) \text{ towards } R(x_0)
\]

\[
P \left( R(l^d) - R(x_0) > \varepsilon \right) < e^{-c\varepsilon^2 l} \quad (c > 0)
\]

Third milestone in learning theory

\[
\lim_{d \to \infty} \frac{G^d(l)}{d} = 0 \implies \text{fast convergence of } R(l^d) \text{ towards } R(x_0) \text{ independent of the choice of the probability measure.}
\]