

Symmetrization (where we start seeing ghost samples)

In words...

The symmetrization Lemma is one of the major tools needed to derive [VC error bounds](#). It is based on the introduction of a **ghost sample**, i.e., an additional but virtual sample of data. With this ghost sample, we can bound the probability of having a large error between the empirical and expected risks via the probability of having a large error between the empirical risk computed on the training sample and the one computed on the ghost sample.

Therefore, after applying the symmetrization Lemma, one is left with the task of bounding a quantity that only depends on the output of classifiers on two samples, i.e., for which [capacity measures based on the projection of the function class on the sample](#) can be used.

In pictures...

In maths...

Consider a **ghost sample**

$$D' = \{(X'_i, Y'_i)\}_{i=1}^N$$

of N independent and identically distributed copies of (X, Y) . Then, the symmetrization Lemma states that

For $N\epsilon^2 \geq 2$,

$$P^N \left\{ \sup_{f \in \mathcal{F}} |R_{\text{emp}}(f) - R(f)| \geq \epsilon \right\} \leq 2P^{2N} \left\{ \sup_{f \in \mathcal{F}} |R_{\text{emp}}(f) - R'_{\text{emp}}(f)| \geq \frac{\epsilon}{2} \right\},$$

and

$$P^N \left\{ \sup_{f \in \mathcal{F}} (R(f) - R_{\text{emp}}(f)) \geq \epsilon \right\} \leq 2P^{2N} \left\{ \sup_{f \in \mathcal{F}} (R'_{\text{emp}}(f) - R_{\text{emp}}(f)) \geq \frac{\epsilon}{2} \right\},$$

where $R'_{\text{emp}}(f) = \frac{1}{N} \sum_{i=1}^N \mathbf{1}_{f(X'_i) \neq Y'_i}$ is the empirical risk computed on the ghost sample D' .

Here, P^N refers to the product probability over the N random pairs of the training sample $D = \{(X_i, Y_i)\}_{i=1}^N$, while P^{2N} refers to the product probability over the $2N$ random pairs of the sample DD' obtained by concatenation of the training and ghost samples.

Proof of the first inequality (with absolute values): Let f^* denote the function for which the supremum is obtained:

$$|R_{\text{emp}}(f^*) - R(f^*)| = \sup_{f \in \mathcal{F}} |R_{\text{emp}}(f) - R(f)|.$$

This allows us to rewrite the quantity of interest as

$$P^N \left\{ \sup_{f \in \mathcal{F}} |R_{\text{emp}}(f) - R(f)| \geq \epsilon \right\} = P^N \left\{ |R_{\text{emp}}(f^*) - R(f^*)| \geq \epsilon \right\}.$$

Given that f^* is a particular function of \mathcal{F} , we also have $\sup_{f \in \mathcal{F}} |R_{\text{emp}}(f) - R'_{\text{emp}}(f)| \geq |R_{\text{emp}}(f^*) - R'_{\text{emp}}(f^*)|$, which implies

$$P^{2N} \left\{ \sup_{f \in \mathcal{F}} |R_{\text{emp}}(f) - R'_{\text{emp}}(f)| \geq \frac{\epsilon}{2} \right\} \geq P^{2N} \left\{ |R_{\text{emp}}(f^*) - R'_{\text{emp}}(f^*)| \geq \frac{\epsilon}{2} \right\},$$

To bound the right-hand side of this inequality, we use the fact that, for any $(a, b, c) \in \mathbb{R}^3$, $P(|a - b| \geq \epsilon/2) \geq P(|a - c| \geq \epsilon, |b - c| \leq \epsilon/2)$, since $\{|a - c| \geq \epsilon \text{ and } |b - c| \leq \epsilon/2\}$ characterizes one of the particular cases in which $|a - b| \geq \epsilon/2$.

With $a = R_{\text{emp}}(f^*)$, $b = R'_{\text{emp}}(f^*)$ and $c = R(f^*)$, this leads to

the ghost sample D' , hence the two events are **independent** and the probability can be decomposed as a product:

$$P^{2N} \left\{ |R_{emp}(f^*) - R'_{emp}(f^*)| \geq \frac{\epsilon}{2} \right\} \geq P^N \left\{ |R_{emp}(f^*) - R(f^*)| \geq \epsilon \right\} P^N \left\{ |R'_{emp}(f^*) - R(f^*)| \leq \frac{\epsilon}{2} \right\}.$$

Using the **inversion principle**, the last probability can be rewritten as

$$P^N \left\{ |R'_{emp}(f^*) - R(f^*)| \leq \frac{\epsilon}{2} \right\} = 1 - P^N \left\{ |R'_{emp}(f^*) - R(f^*)| > \frac{\epsilon}{2} \right\}$$

In addition, with **Chebyshev's inequality** and the **bound on the variance of the bounded random variable** $\mathbf{1}_{f(X) \neq Y} \in [0, 1]$, we have

$$P^N \left\{ |R'_{emp}(f^*) - R(f^*)| > \frac{\epsilon}{2} \right\} \leq \frac{4 \text{Var}(\mathbf{1}_{f(X) \neq Y})}{N\epsilon^2} \leq \frac{1}{N\epsilon^2} \leq \frac{1}{2},$$

where we used the assumption $N\epsilon^2 \geq 2$ for the last inequality.

Gathering these results, we have

$$P^N \left\{ |R'_{emp}(f^*) - R(f^*)| \leq \frac{\epsilon}{2} \right\} \geq 1 - \frac{1}{2} = \frac{1}{2}$$

and

$$P^{2N} \left\{ |R_{emp}(f^*) - R'_{emp}(f^*)| \geq \frac{\epsilon}{2} \right\} \geq \frac{1}{2} P^N \left\{ |R_{emp}(f^*) - R(f^*)| \geq \epsilon \right\},$$

which leads to the desired result

$$P^{2N} \left\{ \sup_{f \in \mathcal{F}} |R_{emp}(f) - R'_{emp}(f)| \geq \frac{\epsilon}{2} \right\} \geq \frac{1}{2} P^N \left\{ \sup_{f \in \mathcal{F}} |R_{emp}(f) - R(f)| \geq \epsilon \right\}.$$

Proof of the second inequality (without absolute values): We can reproduce the same steps to prove the result without absolute values with the particular choice of f^* such that

$$R(f^*) - R_{emp}(f^*) = \sup_{f \in \mathcal{F}} (R(f) - R_{emp}(f)).$$

However, we must pay attention to a few details. In particular, we now consider the difference $R'_{emp}(f^*) - R_{emp}(f^*)$ instead of $|R_{emp}(f^*) - R'_{emp}(f^*)|$ (note the reversed order of the terms). For instance, this means that we use the fact

$$(c - a) \geq \epsilon \wedge (b - c) \geq \frac{-\epsilon}{2} \Rightarrow b - a \geq \frac{\epsilon}{2},$$

and thus that $P(b - a \geq \epsilon/2) \geq P(c - a \geq \epsilon, b - c \geq -\epsilon/2)$, with $a = R_{emp}(f^*)$, $b = R'_{emp}(f^*)$ and $c = R(f^*)$ to obtain

$$P^{2N} \left\{ R'_{emp}(f^*) - R_{emp}(f^*) \geq \frac{\epsilon}{2} \right\} \geq P^N \left\{ R(f^*) - R_{emp}(f^*) \geq \epsilon \right\} P^N \left\{ R'_{emp}(f^*) - R(f^*) \geq \frac{-\epsilon}{2} \right\},$$

where we used the independence of the two events to factorize the right-hand side. Then, the last of the three probabilities above can be lower bounded as

$$\begin{aligned} P^N \left\{ R'_{emp}(f^*) - R(f^*) \geq \frac{-\epsilon}{2} \right\} &= P^N \left\{ R'_{emp}(f^*) - R(f^*) \in \left[\frac{-\epsilon}{2}, \frac{\epsilon}{2} \right] \right\} + P^N \left\{ R'_{emp}(f^*) - R(f^*) > \frac{\epsilon}{2} \right\} \\ &\geq P^N \left\{ |R'_{emp}(f^*) - R(f^*)| \leq \frac{\epsilon}{2} \right\} \\ &\geq \frac{1}{2} \end{aligned}$$

where the last line is obtained via **Chebyshev's inequality**, as in the previous proof. The rest of the proof is similar.