CHAPTER 1

The Basic Conformal Prediction Framework

Vladimir Vovk

Computer Learning Research Centre, Department of Computer Science,
Royal Holloway, University of London, United Kingdom

CHAPTER OUTLINE HEAD

The aim of this chapter is to give a gentle introduction to the method of conformal prediction. It will define conformal predictors and discuss their properties, leaving various extensions of conformal prediction for Chapter 2.

1.1 The Basic Setting and Assumptions

In the bulk of this chapter we consider the basic setting where we are given a training set of examples, and our goal is to predict a new example. We will assume that the examples are elements of an *example space* $\mathbf{Z}$ (formally, this is assumed to be a measurable space, i.e., a set equipped with a σ-algebra). We always assume that $\mathbf{Z}$ contains more than one element, $|\mathbf{Z}| > 1$, and that each singleton is measurable. The examples in the training set will usually be denoted $z_1, \dots, z_l$ and the example

Conformal Prediction for Reliable Machine Learning. http://dx.doi.org/10.1016/B978-0-12-398537-8.00001-8

to be predicted, (*test example*) z_{l+1}. Mathematically the training set is a sequence, $(z_1, \dots, z_l)$, not a set.

The basic setting might look restrictive, but later in this chapter we will see that it covers the standard problems of classification (Section 1.6) and regression (Section 1.7); we will also see that the algorithms developed for our basic setting can be applied in the online (Section 1.8) and batch (Section 2.4) modes of prediction.

We will make two main kinds of assumptions about the way the examples z_i, $i = 1, \dots, l+1$, are generated. Let us fix the size $l \geq 1$ of the training set for now. Under the *randomness assumption*, the $l+1$ examples are generated independently from the same unknown probability distribution Q on $\mathbf{Z}$. Under the *exchangeability assumption*, the sequence $(z_1, \dots, z_{l+1})$ is generated from a probability distribution P on $\mathbf{Z}^{l+1}$ that is *exchangeable*: for any permutation π of the set $\{1, \dots, l+1\}$, the distribution of the permuted sequence $(z_{\pi(1)}, \dots, z_{\pi(l+1)})$ is the same as the distribution P of the original sequence $(z_1, \dots, z_{l+1})$. It is clear that the randomness assumption implies the exchangeability assumption, and in Section 1.5 we will see that the exchangeability assumption is much weaker. (On the other hand, in the online mode of prediction the difference between the two assumptions almost disappears, as we will see in Section 1.8.)

The randomness assumption is a standard assumption in machine learning. Methods of conformal prediction, however, usually work for the weaker exchangeability assumption. In some important cases even the exchangeability assumption can be weakened; see, for example, Chapters 8 and 9 of [365] dealing with online compression modeling.

1.2 Set and Confidence Predictors

In this book we are concerned with reliable machine learning, and so consider prediction algorithms that output a set of elements of $\mathbf{Z}$ as their prediction; such a set is called a *prediction set* (or a *set prediction*). The statement implicit in a prediction set is that it contains the test example z_{l+1}, and the prediction set is regarded as erroneous if and only if it fails to contain z_{l+1}. We will be looking for a compromise between reliability and informativeness of the prediction sets output by our algorithms; an example of prediction sets we try to avoid is the whole of $\mathbf{Z}$; it is absolutely reliable but not informative.

A *set predictor* is a function Γ that maps any sequence $(z_1, \dots, z_l) \in \mathbf{Z}^l$ to a set $\Gamma(z_1, \dots, z_l) \subseteq \mathbf{Z}$ and satisfies the following measurability condition: the set

$$\{(z_1, \dots, z_{l+1}) \mid z_{l+1} \in \Gamma(z_1, \dots, z_l)\} \tag{1.1}$$

is measurable in $\mathbf{Z}^{l+1}$.

We will often consider nested families of set predictors depending on a parameter $\epsilon \in [0, 1]$, which we call the *significance level*, reflecting the required reliability of prediction. Our parameterization of reliability will be such that smaller values of ϵ correspond to greater reliability. (This is just a convention: e.g., if we used the *confidence level* $1 - \epsilon$ as the parameter, larger values of the parameter would correspond to greater reliability.)

Formally, a *confidence predictor* is a family $(\Gamma^\epsilon \mid \epsilon \in [0, 1])$ of set predictors that is nested in the following sense: whenever $0 \leq \epsilon_1 \leq \epsilon_2 \leq 1$,

$$\Gamma^{\epsilon_1}(z_1, \dots, z_l) \supseteq \Gamma^{\epsilon_2}(z_1, \dots, z_l). \tag{1.2}$$

1.2.1 Validity and Efficiency of Set and Confidence Predictors

The two main indicators of the quality of set and confidence predictors are what we call their validity (how reliable they are) and efficiency (how informative they are).[1] We say that a set predictor Γ is *exactly valid at a significance level* $\epsilon \in [0, 1]$ if, under any power probability distribution $P = Q^{l+1}$ on $\mathbf{Z}^{l+1}$, the probability of the event $z_{l+1} \notin \Gamma(z_1, \dots, z_l)$ that Γ makes an error is ϵ. However, it is obvious that the property of exact validity is impossible to achieve unless ϵ is either 0 or 1:

Proposition 1.1. *At any level* $\epsilon \in (0, 1)$*, no set predictor is exactly valid.*

Proof. Let Q be a probability distribution on $\mathbf{Z}$ that is concentrated at one point. Then any set predictor makes a mistake with probability either 0 or 1. □

In Section 1.8 we will see that exact validity can be achieved using randomization.

The requirement that can be achieved (even trivially) is that of "conservative validity." A set predictor Γ is said to be *conservatively valid* (or simply *valid*) *at a significance level* $\epsilon \in [0, 1]$ if, under any power probability distribution $P = Q^{l+1}$ on $\mathbf{Z}^{l+1}$, the probability of $z_{l+1} \notin \Gamma^\epsilon(z_1, \dots, z_l)$ does not exceed ϵ. The trivial way to achieve this, for any $\epsilon \in [0, 1]$, is to set $\Gamma(z_1, \dots, z_l) := \mathbf{Z}$ for all $z_1, \dots, z_l$. A confidence predictor $(\Gamma^\epsilon \mid \epsilon \in [0, 1])$ is (*conservatively*) *valid* if each of its constituent set predictors Γ^ϵ is valid at the significance level ϵ. Conformal predictors will provide nontrivial conservatively, and in some sense almost exactly, valid confidence predictors. In the following chapter we will discuss other notions of validity.

By the efficiency of set and confidence predictors we mean the smallness of the prediction sets they output. This is a vague notion, but in any case it can be meaningful only if we impose some restrictions on the predictors that we consider. Without restrictions, the trivial set predictor $\Gamma(z_1, \dots, z_l) := \emptyset, \forall z_1, \dots, z_l$, and the trivial confidence predictor $\Gamma^\epsilon(z_1, \dots, z_l) := \emptyset, \forall z_1, \dots, z_l, \epsilon$, are the most efficient ones. We will be looking for the most efficient confidence predictors in the class of valid confidence predictors; different notions of validity (including "conditional validity" considered in the next chapter) and different formalizations of the notion of efficiency will lead to different solutions to this problem.

[1] Only validity and efficiency will be used as technical terms.

1.3 Conformal Prediction

Let $n \in \mathbb{N}$, where $\mathbb{N} := \{1, 2, \ldots\}$ is the set of natural numbers. A *(non)conformity n-measure* is a measurable function A that assigns to every sequence $(z_1, \ldots, z_n)$ of n examples a sequence $(\alpha_1, \ldots, \alpha_n)$ of n real numbers that is equivariant with respect to permutations: for any permutation π of $\{1, \ldots, n\}$,

$$(\alpha_1, \ldots, \alpha_n) = A(z_1, \ldots, z_n) \Longrightarrow (\alpha_{\pi(1)}, \ldots, \alpha_{\pi(n)}) = A(z_{\pi(1)}, \ldots, z_{\pi(n)}). \tag{1.3}$$

Let $n = l + 1$. The *conformal predictor* determined by A as a nonconformity measure is defined by

$$\Gamma^{\epsilon}(z_1, \ldots, z_l) := \left\{z \mid p^z > \epsilon\right\}, \tag{1.4}$$

where for each $z \in \mathbf{Z}$ the corresponding *p-value* p^z is defined by

$$p^z := \frac{\left|\left\{i = 1, \ldots, l+1 \mid \alpha_i^z \geq \alpha_{l+1}^z\right\}\right|}{l+1} \tag{1.5}$$

and the corresponding sequence of *nonconformity scores* is defined by

$$(\alpha_1^z, \ldots, \alpha_{l+1}^z) := A(z_1, \ldots, z_l, z). \tag{1.6}$$

Similarly, the conformal predictor determined by A as a conformity measure is defined by (1.4)–(1.6) with $\alpha_i^z \geq \alpha_{l+1}^z$ in (1.5) replaced by $\alpha_i^z \leq \alpha_{l+1}^z$ (in which case (1.6) are referred to as *conformity scores*); this is not really a new notion as the conformal predictor determined by A as a conformity measure is the same thing as the conformal predictor determined by $-A$ as a nonconformity measure.

Remark 1.1. It is easy to see that the prediction set (1.4) output by the conformal predictor Γ depends on ϵ only via

$$[\epsilon]_l := \frac{\lfloor \epsilon(l+1) \rfloor}{l+1}.$$

In other words, $\Gamma^{\epsilon_1} = \Gamma^{\epsilon_2}$ when ϵ_1 and ϵ_2 are *l-equivalent*, in the sense $[\epsilon_1]_l = [\epsilon_2]_l$. Notice that $[\epsilon]_l$ is the smallest value that is l-equivalent to ϵ.

Proposition 1.2. *If examples $z_1, \ldots, z_{l+1}$ are generated from an exchangeable probability distribution on $\mathbf{Z}^{l+1}$, the probability of error, $z_{l+1} \notin \Gamma^{\epsilon}(z_1, \ldots, z_l)$, will not exceed ϵ for any $\epsilon \in [0, 1]$ and any conformal predictor Γ.*

In view of Remark 1.1, we can replace "will not exceed ϵ" in Proposition 1.2 by "will not exceed $[\epsilon]_l$." This proposition was first proved in [364] and [297]; we reproduce the simple argument under simplifying assumptions.

Proof sketch of Proposition 1.2. Let $(\alpha_1, \ldots, \alpha_{z+1}) := A(z_1, \ldots, z_{l+1})$, where A is the nonconformity measure determining Γ. An error is made if and only if α_{l+1} is among the $\lfloor \epsilon(l+1) \rfloor$ largest elements in the sequence $(\alpha_1, \ldots, \alpha_{l+1})$. Because of the assumption of exchangeability, the distribution of $(z_1, \ldots, z_{l+1})$, and so the distribution of $(\alpha_1, \ldots, \alpha_{l+1})$, is invariant under permutations; in particular, all permutations of $(\alpha_1, \dots, \alpha_{l+1})$ are equiprobable. For simplicity we assume that the probability of each permutation is positive and that all αs are different. A random permutation moves one of the $\lfloor \epsilon(l+1) \rfloor$ largest αs to the $(l+1)$th position with probability $[\epsilon]_l$, which is therefore the probability of error. □

1.3.1 The Binary Case

We first consider a toy conformal predictor assuming that $\mathbf{Z} = \{0, 1\}$. The simplest nontrivial example of a nonconformity measure is $A(z_1, \dots, z_n) := (z_1, \dots, z_n)$. The conformal predictor determined by A is

$$\Gamma^{\epsilon}(z_1, \dots, z_l) = \begin{cases} \emptyset & \text{if } \epsilon = 1 \\ \{0\} & \text{if } \frac{k+1}{l+1} \le \epsilon < 1 \\ \{0, 1\} & \text{if } \epsilon < \frac{k+1}{l+1}, \end{cases} \tag{1.7}$$

where l is the size of the training set and k is the number of 1s in it. In other words, we can make a confident prediction that $z_{l+1} = 0$ only if the allowed probability ϵ of error is $\frac{k+1}{l+1}$ or more. For large l this agrees with Laplace's rule of succession, which says that $z_{l+1} = 1$ with probability $\frac{k+1}{l+2}$, and with many other estimates of the probability of success in Bernoulli trials (see Section 2.8 for further details).

The usual justification of Laplace's rule of succession is Bayesian: $\frac{k+1}{l+2}$ is the mean of the posterior distribution of the parameter $p \in [0, 1]$ (the probability of success, represented by 1) of the Bernoulli model with the uniform prior on p after observing k 1s and $n - k$ 0s. Avoiding such assumptions (at least to some degree) can make our conclusions more robust.

1.3.2 The Gaussian Case

Figure 1.1 is an empirical illustration of using conformal prediction to make our conclusions more robust by partially avoiding Bayesian assumptions. (It is similar to Figure 1 in [374], as corrected in [362, Figure 11.1]. For other empirical studies of this kind, see, e.g., [365, Section 10.3].) Four observations z_1, z_2, z_3, z_4 are generated from the statistical model $\{N(\theta, 1) \mid \theta \in \mathbb{R}\}$, where $\mathbb{R}$ is the set of real numbers and $N(\theta, 1)$ is the Gaussian distribution with mean θ and variance 1. In the top plot we take $\theta = 1$, in the middle $\theta = 10$, and in the bottom $\theta = 100$. We consider the Gaussian prior $N(0, 1)$ on the parameter θ, so that the Bayesian assumption can be regarded as satisfied for the top plot, violated for the middle plot, and grossly violated for the bottom plot. We take 80% as the confidence level.

The dashed lines are the Bayes prediction intervals and the solid lines are the prediction intervals output by a conformal predictor based on the Bayesian assumption (for details, see later). The performance of both kinds of prediction intervals depends on the assumption, but in very different ways. When the Bayesian assumption is satisfied (the top plot), both prediction intervals give reasonable results; in particular, they cover the four observations. When it becomes violated (middle plot), the conformal prediction interval becomes wider in order to cover the four observations whereas

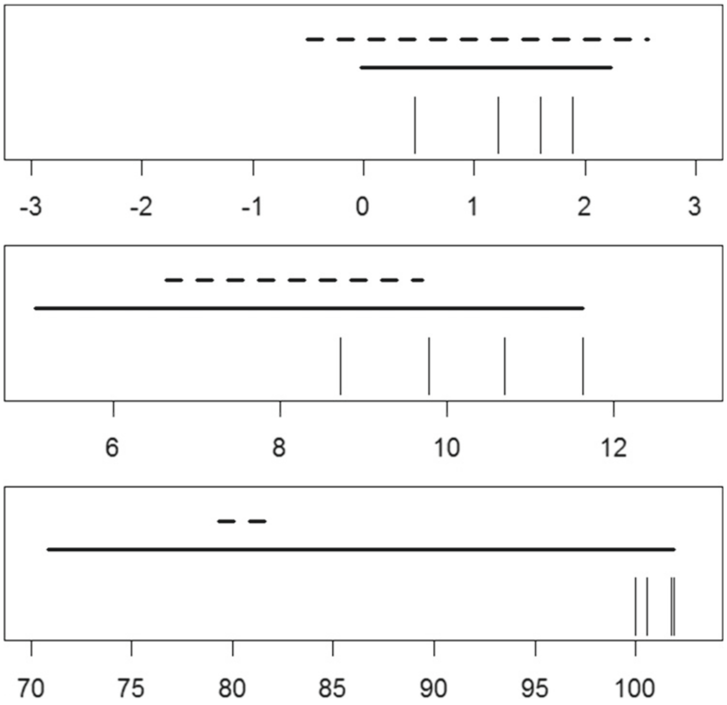

FIGURE 1.1

In the top plot, the four observation are generated from $N(1, 1)$; in the middle plot, from $N(10, 1)$; and in the bottom plot, from $N(100, 1)$. The solid lines are the prediction intervals output by the conformal predictor and the dashed lines are the Bayes prediction intervals.

the Bayes prediction interval ceases to be valid and covers clearly only one observation. And when it becomes grossly violated (bottom plot), the Bayes prediction interval becomes grossly invalid and does not cover any observations; the conformal prediction interval becomes long but still valid (in accordance with Proposition 1.2).

If we measure the efficiency of a prediction interval by its length, we can see that there is a certain symmetry between Bayes and conformal prediction intervals: as the Bayesian assumption becomes less and less satisfied, the Bayes prediction intervals lose their validity while maintaining their efficiency, and the conformal prediction intervals lose their efficiency while maintaining their validity. (In the top plot, the Bayes prediction interval happens to be wider, but for a random seed of the random number generator the Bayes prediction intervals are shorter in approximately 54% of cases.) Validity, however, is more important than efficiency, and efficiency is a meaningful notion only in the presence of validity.

These are the details of the Bayes and conformal predictors used earlier. The conditional distribution of z_5 given $z_1, \ldots, z_4$ under the Bayesian assumption is $N\left(\frac{4}{5}\bar{z}, \frac{6}{5}\right)$, where $\overline{z} := \frac{1}{4}(z_1 + \cdots + z_4)$ is the mean of the given observations. This gives the Bayes prediction interval $[0.8\overline{z} - 1.2c, 0.8\overline{z} + 1.2c]$, where c is the upper 10% quantile of the standard Gaussian distribution. The conformal predictor is based on the nonconformity measure that maps $(z_1, \ldots, z_5) \in \mathbb{R}^5$ to $(\alpha_1, \ldots, \alpha_5)$, where the nonconformity score α_i of z_i is defined as $\alpha_i := \left|z_i - \frac{5}{6}\overline{z}\right|$, $\frac{5}{6}\overline{z}$ being the mean of the posterior distribution of θ after observing $z_1, \ldots, z_5$, and $\overline{z} := \frac{1}{5}(z_1 + \cdots + z_5)$ being the mean of the five observations including z_5. This conformal predictor always outputs prediction sets that are intervals; in general, the *prediction interval* output by a conformal predictor is defined to be the convex hull of the prediction set (1.4).

The dependence of the validity of prediction intervals on the Bayesian assumption (which rarely has justifications other than mathematical convenience) is particularly serious in nonparametric statistics. At least in parametric statistics there are several results about the decreasing importance of the prior as the amount of data grows. But in nonparametric statistics, “the prior can swamp the data, no matter how much data you have” ([78, Section 4]), and using Bayes prediction intervals becomes even more problematic.

1.4 Efficiency in the Case of Prediction without Objects

In this section we describe a recent result by Lei, Robins, and Wasserman [200], who propose an asymptotically efficient conformal predictor. The example space $\mathbf{Z}$ is now simply the Euclidean space $\mathbb{R}^d$. We let Λ stand for the Lebesgue measure on $\mathbb{R}^d$.

Let Q be the data-generating distribution on $\mathbf{Z}$; we will assume that the examples $z_1, \ldots, z_{l+1}$ are generated from the power probability distribution Q^{l+1}. Let $\epsilon \in (0, 1)$ be a given significance level. Both Q and ϵ will be fixed throughout this section (except for Remark 1.2).

A natural notion of efficiency of a prediction set is its closeness to the “oracle” prediction set

$$C^{\text{or}} := \arg\min_{C} \Lambda(C), \tag{1.8}$$

where C ranges over the measurable subsets of $\mathbf{Z}$ such that $Q(C) \geq 1 - \epsilon$. We will be interested in the case where the arg min is attained on an essentially unique set C (in particular, this will be implied by the assumptions of Theorem 1.1). Lei et al. [200] construct a conformal predictor Γ such that Γ^ϵ is close to C^{or} for a big training set. The closeness of Γ^ϵ and C^{or} will be measured by the Lebesgue measure $\Lambda(\Gamma^\epsilon \bigtriangleup C^{\text{or}})$ of their symmetric difference.

Lei et al.’s conformal predictor enjoys properties of validity and efficiency. As for any conformal predictor, the property of validity does not require any assumptions apart from exchangeability. The property of efficiency, however, will be stated under the following assumptions about the data generating distribution Q:

1. The data-generating distribution has a differentiable density q.
2. The gradient of q is bounded.

3. There exists t such that $Q(\{z \mid q(z) \geq t\}) = 1 - \epsilon$.
4. The gradient of q is bounded away from 0 in a neighbourhood of the set $\{z \mid q(z) = t\}$.

It is clear that under Assumption 2 the arg min in (1.8) is indeed attained at some C; for example, at $\{z \mid q(z) > t\}$ and at $\{z \mid q(z) \geq t\}$, where t is defined in Assumption 3. For concreteness, let us define

$$C^{\text{or}} = \{z \mid q(z) \geq t\};$$

this specific choice does not matter, since $\Lambda(C \bigtriangleup C^{\text{or}}) = 0$ for any C at which the arg min is attained.

Theorem 1.1 ([200, Theorem 3.3]). *Suppose Assumptions 1–4 hold. There exists a conformal predictor* Γ *(independent of* ϵ*) such that for any* $\lambda > 0$ *there exists* B *such that, as* $l \to \infty$,

$$\mathbb{P}\left(\Lambda(\Gamma^{\epsilon}(z_1, \dots, z_l) \bigtriangleup C^{\text{or}}) \geq B\left(\frac{\log l}{l}\right)^{\frac{1}{d+2}}\right) = O\left(l^{-\lambda}\right). \tag{1.9}$$

Remark 1.2. It remains an open problem whether the rate $(\log l/l)^{1/(d+2)}$ in (1.9), or its corollary

$$\mathbb{E}\left(\Lambda(\Gamma^{\epsilon}(z_1, \dots, z_l) \bigtriangleup C^{\text{or}})\right) = O\left(\frac{\log l}{l}\right)^{\frac{1}{d+2}}, \tag{1.10}$$

is optimal. Rigollet and Vert [286] establish a lower bound matching (1.10): for a wide class $\mathcal{P}$ of densities, there exists a constant C such that, for any sample size l and any set predictor Γ (not required to satisfy any conditions of validity),

$$\sup_{q \in \mathcal{P}} \mathbb{E}\left(\Lambda(\Gamma(z_1, \dots, z_l) \bigtriangleup C^{\text{or}})\right) \geq C\left(\frac{\log l}{l}\right)^{\frac{1}{d+2}}, \tag{1.11}$$

(this is a special case of Rigollet and Vert's Theorem 5.1). However, there are at least two reasons why (1.11) does not prove the optimality of (1.10): first, the class $\mathcal{P}$ in (1.11) consists of data-generating distributions Q satisfying Assumptions 1 and 2 but not necessarily Assumption 4 (and the probability distributions used in the proof do not satisfy Assumption 4); and second, Rigollet and Vert prove (1.11) only in a special symmetric case (that would correspond to $\epsilon = 1/2$ if Assumption 4 were satisfied). It would be ideal to establish an optimal rate for each value ϵ of the significance level.

The conformity measure on which Lei et al.'s conformal predictor is based is easy to describe. Let $K : \mathbf{Z} \to \mathbb{R}$ be a symmetric ($K(z) = K(-z)$ for all z) continuous function that is concentrated on $[-1, 1]^d$ and integrates to 1. (It is interesting that the condition that K be nonnegative is not required.) Choose any sequence $h_1, h_2, \dots$ of positive numbers satisfying

$$h_n \asymp \left(\frac{\log n}{n}\right)^{\frac{1}{d+2}},$$

where $\asymp$ stands for the asymptotic coincidence to within a constant factor. For a given sequence $(z_1, \dots, z_n)$ of examples define the function

$$\hat{p}_n(z) := \frac{1}{nh_n^d} \sum_{i=1}^{n} K\left(\frac{z - z_i}{h_n}\right) \tag{1.12}$$

(this is the usual kernel density estimate). The conformity score α_i of each example z_i is defined by $\alpha_i := \hat{p}_n(z_i)$. The conformal predictor Γ determined by this conformity measure will satisfy Theorem 1.1.

One disadvantage of the prediction set $\Gamma^\epsilon(z_1, \dots, z_l)$ is that it may be difficult to compute. A more computationally efficient version $\Gamma^+(z_1, \dots, z_l)$ can be defined as follows. For a given training set $(z_1, \dots, z_l)$ define the function

$$\hat{p}_l(z) := \frac{1}{lh_l^d} \sum_{i=1}^{l} K\left(\frac{z - z_i}{h_l}\right) \tag{1.13}$$

(in analogy with (1.12)) and set

$$\Gamma^+(z_1, \dots, z_l) := \left\{z \mid \hat{p}_l(z) \geq \hat{p}_l(z_{(k)}) - (lh_l^d)^{-1}(\sup K - \inf K)\right\}, \tag{1.14}$$

where $z_{(\cdot)}$ refers to the reordering of the z_i such that $\hat{p}_l(z_{(1)}) \leq \cdots \leq \hat{p}_l(z_{(l)})$ and $k := \lfloor \epsilon(l+1) \rfloor$. The set predictor Γ^+ is not a conformal predictor but it is guaranteed to satisfy $\Gamma^\epsilon(z_1, \dots, z_l) \subseteq \Gamma^+(z_1, \dots, z_l)$ (and so is conservatively valid) and Theorem 1.1 continues to hold when Γ^ϵ is replaced by Γ^+.

Lei et al.'s paper [200] contains more general results. For example, if we impose strong smoothness conditions on the density q of the data-generating distribution Q, the convergence rate in Theorem 1.1 becomes much faster: namely, we can replace the exponent $1/(d+2)$ by an exponent as close to $1/2$ as we wish. Even faster rates of convergence can be achieved if we replace Assumption 4 by an assumption of a faster change of the density q when moving away from the set $\{q = t\}$.

The material of this section is closely connected to the problem of anomaly detection (see Chapter 4). Now we interpret ϵ as our chosen tolerance level for anomalies and regard a new observation z as anomalous in view of the known observations $z_1, \dots, z_l$ if $z \notin \Gamma^\epsilon(z_1, \dots, z_l)$. A more computationally efficient procedure is to regard z as anomalous in view of $z_1, \dots, z_l$ if $z \notin \Gamma^+(z_1, \dots, z_l)$, where Γ^+ (depending on ϵ) is as defined earlier. In both cases the probability of a false alarm does not exceed ϵ.

1.5 Universality of Conformal Predictors

In this section we will see that the conformal predictors are universal in the sense of being the only way to achieve validity (in a suitable sense). This notion of universality applies to the whole class of conformal predictors (in Section 1.8 we will discuss individual conformal predictors that are universal in the sense of having the best possible asymptotic efficiency among asymptotically valid set predictors, and Lei et al.'s conformal predictors discussed in the previous section are also universal in a certain sense).

Let us say that a confidence predictor Γ is *invariant* if

$$\Gamma^{\epsilon}(z_1, \ldots, z_l) = \Gamma^{\epsilon}(z_{\pi(1)}, \ldots, z_{\pi(l)})$$

for any permutation π of the indices $\{1, \ldots, l\}$. Under the exchangeability assumption, this is a very natural class of confidence predictors (by the sufficiency principle; see, e.g., [63]).

For simplicity we will consider a stronger notion of validity than usual: let us say that a confidence predictor Γ is *(conservatively) valid under (the assumption of) exchangeability* at a significance level $\epsilon \in [0, 1]$ if, under any exchangeable probability distribution P on $\mathbf{Z}^{l+1}$, the probability of error $z_{l+1} \notin \Gamma^{\epsilon}(z_1, \ldots, z_l)$ does not exceed ϵ.

We will say that a confidence predictor Γ_2 is *at least as good as* another confidence predictor Γ_1 if, for any significance level ϵ,

$$\Gamma_2^{\epsilon}(z_1, \ldots, z_l) \subseteq \Gamma_1^{\epsilon}(z_1, \ldots, z_l)$$

holds for almost all $z_1, \ldots, z_l$. It turns out that any invariant confidence predictor that is conservatively valid under exchangeability is a conformal predictor or can be improved to become a conformal predictor.

Proposition 1.3 ([252]). *Let Γ_1 be an invariant confidence predictor that is conservatively valid under exchangeability. Then there is a conformal predictor Γ_2 that is at least as good as Γ_1.*

Proof. The conformity measure that determines Γ_2 is defined as follows. Let $(z_1, \ldots, z_{l+1}) \in \mathbf{Z}^{l+1}$. The conformity score of z_i is defined as

$$\begin{aligned} \alpha_i &:= \inf\left\{\epsilon \mid z_i \notin \Gamma_1^{\epsilon}(z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_{l+1})\right\} \\ &= \sup\left\{\epsilon \mid z_i \in \Gamma_1^{\epsilon}(z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_{l+1})\right\}. \end{aligned}$$

Without loss of generality we can assume that the inf is always attained (cf. [365], Proposition 2.11).

Let check that Γ_2 is at least as good as Γ_1. We are required to prove that $z_{l+1} \in \Gamma_1^{\epsilon}(z_1, \ldots, z_l)$ whenever $z_{l+1} \in \Gamma_2^{\epsilon}(z_1, \ldots, z_l)$. Fix a data sequence $(z_1, \ldots, z_{l+1})$ and a significance level ϵ, and suppose $z_{l+1} \notin \Gamma_1^{\epsilon}(z_1, \ldots, z_l)$. Since Γ_1 is conservatively valid under exchangeability,

$$z_i \notin \Gamma_1^{\epsilon}(z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_{l+1})$$

for at most $\lfloor \epsilon(l+1) \rfloor z_i$s (this follows from the definition of conservative validity under exchangeability applied to the uniform distribution on the multiset of all $(l+1)!$

permutations of the data sequence $(z_1, \dots, z_{l+1})$). We can see that α_{l+1} is among the $\lfloor \epsilon(l+1) \rfloor$ smallest α_is. Therefore, $z_{l+1} \notin \Gamma_2^{\epsilon}(z_1, \dots, z_l)$. □

For more sophisticated versions of Proposition 1.3, see [253] and [365, Section 2.4]. In particular, Theorem 2.6 of [365] is the analogue of Proposition 1.3 for the standard notion of (conservative) validity.

1.6 Structured Case and Classification

Starting from this section we consider the case where each example consists of two components, $z_i = (x_i, y_i)$; the first component x_i is called an *object* and the second component y_i a *label*. The example space $\mathbf{Z}$ is the Cartesian product $\mathbf{Z} = \mathbf{X} \times \mathbf{Y}$ of the *object space* $\mathbf{X}$ and the *label space* $\mathbf{Y}$; we always assume $|\mathbf{Y}| > 1$. This "structured" case covers two fundamental machine learning problems: *classification*, in which $\mathbf{Y}$ is a finite set (with the discrete σ-algebra), and *regression*, in which $\mathbf{Y} = \mathbb{R}$ is the set of real numbers. (Of course, this is a very primitive structure on the examples; both objects and labels can themselves have a nontrivial structure.)

In the structured case $\mathbf{Z} = \mathbf{X} \times \mathbf{Y}$ the prediction algorithm is usually allowed to output its prediction for the label y_{l+1} after being fed with the object x_{l+1}. In this case it is natural to rewrite a confidence predictor $(\Gamma^{\epsilon} \mid \epsilon \in [0, 1])$ as the function

$$\Gamma^{\epsilon}(z_1, \dots, z_l, x_{l+1}) := \left\{ y \mid (x_{l+1}, y) \in \Gamma^{\epsilon}(z_1, \dots, z_l) \right\}. \tag{1.15}$$

In the structured case we will usually refer to the right-hand side of (1.15) (rather than $\Gamma^{\epsilon}(z_1, \dots, z_l)$) as the *prediction set*. The analogue of (1.4) in the structured case is

$$\Gamma^{\epsilon}(z_1, \dots, z_l, x_{l+1}) := \left\{ y \in \mathbf{Y} \mid p^{(x_{l+1}, y)} > \epsilon \right\}. \tag{1.16}$$

In the rest of this section we will concentrate on the problem of classification, $|\mathbf{Y}| < \infty$, starting from two simple examples of nonconformity measures suitable for classification. Given a sequence of examples $z_1, \dots, z_n$, where $z_i = (x_i, y_i)$ for all i, the corresponding nonconformity scores $(\alpha_1, \dots, \alpha_n)$ can be computed as follows:

1. In the spirit of the 1-nearest neighbour algorithm, we can set

$$\alpha_i := \frac{\min_{j=1,\dots,n : j \neq i \ \& \ y_j = y_i} \Delta(x_i, x_j)}{\min_{j=1,\dots,n : y_j \neq y_i} \Delta(x_i, x_j)}, \tag{1.17}$$

   where Δ is a metric on $\mathbf{X}$. (Intuitively, an example conforms to the sequence if it is close to the other examples with the same label and far from the examples with a different label.)
2. Suppose, additionally, that $|\mathbf{Y}| = 2$ (our classification problem is *binary*). Train a support vector machine on $z_1, \dots, z_n$ and use the corresponding Lagrange multipliers $\alpha_1, \dots, \alpha_n$ as nonconformity scores.

These conformity measures determine conformal predictors by the formula (1.16).

Reporting prediction sets (1.16) at a given significance level ϵ (or, preferably, at several significance levels ϵ) is just one way of presenting the prediction produced by the conformal predictor. Another way is to report the *point prediction* $\hat{y} \in \arg\max_y p^{x_{l+1},y}$ (let us assume that $\left|\arg\max_y p^{x_{l+1},y}\right| = 1$ for simplicity), the *credibility* $p^{x_{l+1},\hat{y}}$, and the *confidence* $1 - \max_{y \neq \hat{y}} p^{x_{l+1},y}$. A high (i.e., close to 1) confidence means that there is no likely alternative to the point prediction, and a low (i.e., close to 0) credibility means that even the point prediction is unlikely (reflecting the fact that the known data $z_1, \ldots, z_l, x_{l+1}$ are very unusual under the exchangeability assumption, which can happen, albeit with a low probability, even if the data sequence $z_1, \ldots, z_{l+1}$ has been really generated from an exchangeable probability distribution).

A more direct representation of the prediction is as the *p-value function* mapping $y \in \mathbf{Y}$ to $p^{x_{l+1},y}$. The notion of a p-value function was introduced by Miettinen [230] in the context of confidence intervals (rather than prediction sets). It has been widely used in epidemiology; see, for example, [25, Section 6], for further references and a discussion of several alternative terms to "p-value function." The notion of a p-value function is also applicable in the unstructured case (cf. (1.5) earlier) and in the case of regression discussed in the next section.

1.7 Regression

In regression problems, a very natural nonconformity measure is

$$\alpha_i := \Delta\left(y_i, f(x_i)\right), \tag{1.18}$$

where $\Delta : \mathbf{Y}^2 \to \mathbb{R}$ is a measure of difference between two labels (usually a metric) and $f : \mathbf{X} \to \mathbf{Y}$ is a prediction rule (for predicting the label given the object) found from $((x_1, y_1) \ldots, (x_n, y_n))$ as the training set; to simplify notation, we suppressed the dependence of f on $((x_1, y_1) \ldots, (x_n, y_n))$ in (1.18).

In the case where $\Delta(y, y') := \left|y - y'\right|$ and f is found using the ridge regression procedure, the conformal predictor determined by the nonconformity measure (1.18) is called the *ridge regression confidence machine* (RRCM). This assumes that objects are vectors in a Euclidean space, $\mathbf{X} \subseteq \mathbb{R}^d$. The explicit representation of this nonconformity measure is

$$\alpha_i := \left|y_i - x_i'(X'X + aI)^{-1}X'Y\right|, \tag{1.19}$$

where $a \geq 0$ is the parameter of the algorithm (with $a = 0$ corresponding to the least squares algorithm), X is the $n \times d$ object matrix whose rows are $x_1', \ldots, x_n'$, Y is the label vector $(y_1, \ldots, y_n)'$, I is the unit $d \times d$ matrix, and $'$ stands for matrix transposition.

The prediction set output by the RRCM at a given significance level can be computed efficiently; in fact, it is not difficult to show that for a fixed dimension d the RRCM can be implemented with running time $O(l \log l)$.

The definition of the RRCM is somewhat arbitrary: for example, instead of using the *residuals* (1.19) as nonconformity scores we could use *deleted residuals* defined as in (1.19) but with X and Y not including the ith object and label, respectively:

$$\begin{aligned} X &:= (x_1, \ldots, x_{i-1}, x_{i+1}, x_n)', \\ Y &:= (y_1, \ldots, y_{i-1}, y_{i+1}, y_n)'. \end{aligned}$$

The conformal predictor determined by this nonconformity measure is called the *deleted RRCM*. In the case $a = 0$, an explicit description of the deleted RRCM is given in [365, p. 34]. However, the most natural modification of the RRCM is the "studentized" version, which is in some sense half-way between the RRCM and the deleted RRCM ([365, p. 35], where only the case $a = 0$ is considered). An interesting direction of further research would be to extend the definitions of deleted and studentized RRCM to the case $a \neq 0$ and to study the three versions empirically on benchmark datasets.

The definition of RRCM is ultimately based on fitting linear functions to the data. In nonlinear cases, we can use the kernelized version of the RRCM, in which the nonconformity measure is defined as

$$\alpha_i := \left| y_i - Y'(K + aI)^{-1} k \right|,$$

where Y and a are as in (1.19), K is the $n \times n$ matrix $K_{i,j} := \mathcal{K}(x_i, x_j)$, I is the $n \times n$ unit matrix, k is the vector $k_i := \mathcal{K}(x, x_i)$ in $\mathbb{R}^n$, and $\mathcal{K}$ is a given kernel (another parameter of the algorithm). The kernelized RRCM is also computationally efficient. In the online prediction protocol (discussed in the next section), the computations at step n of the online protocol can be performed in time $O(n^2)$.

For details of the RRCM and its various modifications, see [365, Section 2.3]. The R package `PredictiveRegression` (available from the CRAN web page) includes a program implementing the RRCM.

1.8 Additional Properties of Validity and Efficiency in the Online Framework

The property of validity of conformal predictors can be stated in an especially strong form in the following *online framework*. The examples $z_1, z_2, \ldots$ (which may be structured, $z_i = (x_i, y_i)$) arrive one by one, and before observing z_n (or y_n in the structured case) the prediction algorithm outputs a prediction set; as usual, we say that the algorithm makes an error if the prediction set fails to contain z_n.

The *smoothed conformal predictor* determined by a nonconformity measure A is defined in the same way as the conformal predictor except that the p-values (1.5) are replaced by the *smoothed p-values*

$$p^z := \frac{\left|\left\{i = 1, \ldots, n \mid \alpha_i^z > \alpha_n^z\right\}\right| + \tau \left|\left\{i = 1, \ldots, n \mid \alpha_i^z = \alpha_n^z\right\}\right|}{n}, \qquad (1.20)$$

where τ is a random variable generated from the uniform distribution on [0, 1] (the same value of τ can be used for all z). In other words, the prediction made by the smoothed conformal predictor determined by A at step n is defined by the right-hand side of (1.4), where p^z are defined by (1.20) and the nonconformity scores α_i^z are defined by (1.6) (with $n-1$ in place of l). We will sometimes refer to τ as the *tie-breaking random variable*.

For smoothed conformal predictors, the probability of error is exactly ϵ. Moreover, when used in the online mode, smoothed conformal predictors make errors at different steps independently (assuming that the tie-breaking random variables τ at different steps are independent between themselves and of the examples). This is spelled out in the following theorem, where 1_E stands for the indicator function of E.

Theorem 1.2. *Let $N \in \mathbb{N}$ and suppose that examples $z_1, \ldots, z_N$ are generated from an exchangeable distribution on $\mathbf{Z}^N$. For any nonconformity measure A and any significance level ϵ, the smoothed conformal predictor Γ^ϵ determined by A at significance level ϵ makes errors with probability ϵ independently at different steps when applied in the online mode. In other words, the random variables $1_{z_n \notin \Gamma^\epsilon(z_1,\ldots,z_{n-1})}$, $n = 1, \ldots, N$, are independent Bernoulli variables with parameter ϵ.*

Theorem 1.2 is proved in [365] (Theorem 8.2 and Section 8.7) in the general framework of online compression models; for a proof in our current context of exchangeability, see, for example, [359, Theorem 2]. Rényi's "lemme fondamental" ([284, Lemma 2]) is a predecessor of Theorem 1.2 (and in fact implies a version of Theorem 1.2 for nonsmoothed predictors when each nonconformity score α_i depends only on the corresponding example z_i and this dependence is continuous in a suitable sense).

Theorem 1.2 immediately implies that if $z_1, z_2, \ldots$ is an infinite sequence of examples generated from an exchangeable distribution on $\mathbf{Z}^\infty$, at each significance level ϵ any smoothed conformal predictor will still make errors with probability ϵ independently at different steps.

Corollary 1.1. *Suppose examples $z_1, z_2, \ldots$ are generated independently from the same distribution Q on $\mathbf{Z}$. For any nonconformity measure A and any significance level ϵ, the smoothed conformal predictor Γ^ϵ determined by A at significance level ϵ makes errors with probability ϵ independently at different steps when applied in the online mode. In other words, the random variables $1_{z_n \notin \Gamma^\epsilon(z_1,\ldots,z_{n-1})}$, $n = 1, 2, \ldots$, are independent Bernoulli variables with parameter ϵ.*

In this infinite-horizon case, the strong law of large numbers implies that the limiting relative frequency of errors will be ϵ. As a smoothed conformal predictor makes an error whenever the corresponding conformal predictor makes an error, for conformal predictors the limiting relative frequency of errors (in the sense of upper limit) will not exceed ϵ.

In Theorem 1.2 we make the assumption of exchangeability rather than randomness to make it stronger: it is very easy to give examples of exchangeable distributions on $\mathbf{Z}^N$ that are not of the form Q^N (e.g., in the case $\mathbf{Z} = \{0, 1\}$, the uniform probability distribution on the set of all binary sequences of length n containing exactly k 1s, for some $k \in \{1, \ldots, N-1\}$). On the other hand, in the infinite-horizon case (which is the standard setting for the online mode of prediction) the difference between the exchangeability and randomness assumptions essentially disappears: by de Finetti's theorem, each exchangeable probability distribution is a mixture of power probability distributions Q^∞, provided $\mathbf{Z}$ is a Borel space. In particular, using the assumption of randomness rather than exchangeability in Corollary 1.1 hardly weakens it: the two forms are equivalent when $\mathbf{Z}$ is a Borel space.

1.8.1 Asymptotically Efficient Conformal Predictors

Let us say that a prediction set (1.15) is *multiple* if it contains more than one label. In this subsection we consider randomized set predictors, whose prediction at each step depends on an additional random input that is independent of the examples. If Γ is such a set predictor, we let $\Gamma_n := \Gamma(z_1, \ldots, z_{n-1}, x_n)$ stand for the prediction set output at step n of the online protocol and

$$\mathrm{Mult}_n(\Gamma) := \sum_{i=1}^{n} 1_{|\Gamma_n|>1}$$

stand for the cumulative number of multiple predictions over the first n steps. We will also use the notation

$$\mathrm{Err}_n(\Gamma) := \sum_{i=1}^{n} 1_{y_n \notin \Gamma_n}$$

for the number of errors and

$$\mathrm{Emp}_n(\Gamma) := \sum_{i=1}^{n} 1_{|\Gamma_n|=0}$$

for the number of empty predictions made by Γ over the first n steps.

The number of multiple predictions $\mathrm{Mult}_n(\Gamma)$ is a natural measure of efficiency of Γ: the fewer the number of multiple predictions the more efficient the predictor. Of course, we compare only the efficiency of valid set predictors. It turns out (see, e.g., [365, Theorem 3.1]) that there exists a conformal predictor (explicit and computationally efficient) which is at least as asymptotically efficient as any other set predictor that is valid in a weak asymptotic sense. Here we will only state this result; for details, see Chapter 3 of [365].

There is a conformal predictor Γ that is universal in the sense of satisfying the following two conditions. Let us fix a significance level ϵ (Γ will satisfy these conditions for any ϵ).

Let us say that a randomized set predictor Γ' is *asymptotically conservative* at the significance level ϵ for a probability distribution Q on $\mathbf{Z}$ if

$$\limsup_{n\to\infty} \frac{\mathrm{Err}_n(\Gamma')}{n} \le \epsilon \quad \text{a.s.}$$

under the power distribution Q^∞ on $\mathbf{Z}^\infty$. Since Γ is a conformal predictor, Γ^ϵ is automatically asymptotically conservative for any probability distribution Q on $\mathbf{Z}$.

Let us say that Γ' is *asymptotically optimal* at the significance level ϵ for a probability distribution Q on $\mathbf{Z}$ if, for any randomized set predictor Γ'' that is asymptotically conservative for Q,

$$\limsup_{n\to\infty} \frac{\mathrm{Mult}_n(\Gamma')}{n} \le \liminf_{n\to\infty} \frac{\mathrm{Mult}_n(\Gamma'')}{n} \quad \text{a.s.} \tag{1.21}$$

under the power distribution Q^∞ on $\mathbf{Z}^\infty$. The first nontrivial condition that Γ^ϵ satisfies is that it is asymptotically optimal for any Q.

The final condition that Γ^ϵ satisfies is that, for any probability distribution Q on $\mathbf{Z}$ and any randomized set predictor Γ' that is asymptotically conservative and asymptotically optimal for Q,

$$\liminf_{n\to\infty} \frac{\mathrm{Emp}_n(\Gamma^\epsilon)}{n} \ge \limsup_{n\to\infty} \frac{\mathrm{Emp}_n(\Gamma')}{n} \quad \text{a.s.} \tag{1.22}$$

under Q^∞.

The condition (1.21) is very natural, but (1.22) might appear less so. Empty predictions (always leading to an error) provide a warning that the object whose label is being predicted is untypical (very different from the previously observed objects), and we would like to be warned as often as possible once we have a guarantee that the long-run frequency of errors will not exceed ϵ.

An asymptotically efficient conformal predictor, satisfying the properties (1.21) and (1.22), is explicitly constructed in [365] using nearest neighbors as the underlying algorithm; the number K_n of nearest neighbors at step n is slowly growing to infinity. The computations at step n take time $O(\log n)$.

Asymptotically efficient conformal predictors discussed in this section have been designed for classification problems, and it would be very interesting to carry out a similar construction in the case of regression. This construction would be different from the one given in [201] (and discussed in Section 2.6); for a further discussion, see Remark 2.3 in Chapter 2.

The criterion of efficiency stated earlier (beating any other predictor in the sense of (1.21) and, if (1.21) holds as an equality, (1.22)) is only one of several natural criteria of efficiency. For a detailed discussion, see [363], which also defines an important class of criteria of efficiency called proper (the criterion (1.21)–(1.22) does not belong to this class). Constructing asymptotically efficient conformal predictors under such criteria is another interesting direction of research. Both for the criterion (1.21)–(1.22) and for proper criteria of efficiency, the rates of convergence in (1.21)–(1.22) or their analogues are of great interest.

Acknowledgments

I am grateful to Sasha Tsybakov for his advice. The empirical studies described in this chapter used the R language and the R package `PredictiveRegression`; I am grateful to Ilia Nouretdinov for cowriting (and writing the first version of) the package. This work was supported in part by the Cyprus Research Promotion Foundation (TPE/ORIZO/0609(BIE)/24) and EPSRC (EP/K033344/1).