Notes on Conformal Prediction and Testing

Kee Siong Ng

February 8, 2026

1 Split Conformal Prediction

Conformal prediction [VGS05, AB21] is arguably the most elegant and practical technique for improving the robustness of how we use predictions from (black-box) machine-learning models. The classification setting is simplest to explain so that's where we will start. Suppose we have a model $\hat{f} : X \to [0,1]^K$ that, given say an image $x \in X$, provides a vector $\hat{f}(x) \in [0,1]^K$ where the i-th entry denotes the probability that x belongs to class $i \in \{1,\ldots,K\}$. The goal of conformal prediction is to construct, given an arbitrary pair $(x,y) \in X \times \{1,\ldots,K\}$ drawn from the same (unknown) distribution as the training data for $\hat{f}$, a prediction set $C(x) \subseteq \{1,\ldots,K\}$ using $\hat{f}$ such that

$$1 - \alpha \leq \Pr(y \in C(x)) \tag{1}$$

for a user-specified error rate $\alpha \in [0,1]$. So, for example, if α is set at 0.1, then we want the prediction set to contain the true label with probability at least 0.9. In other words, what we are seeking to do in (1) is to calibrate the (raw) probability estimates of the predictions of $\hat{f}$.

There are several ways to achieve (1). The simplest procedure is what is known as split conformation prediction, whereby the available labelled dataset $D = \{(x_i, y_i)\}_{i=1}^{N}$ is randomly split into a training set D_t of size $N - n$ and a validation set D_v of size n. The model $\hat{f}$ is obtained by training on D_t. We then define a conformal score $s : X \times \{1,\ldots,K\} \to \mathbb{R}$ by

$$s(x,y) = 1 - \hat{f}(x)_y \tag{2}$$

and use it to compute $\hat{q}$, the $\lceil (1-\alpha)(n+1) \rceil / n$ quantile of the set

$$\{s(x,y) : (x,y) \in D_v\}.$$

Here, $\lceil \cdot \rceil$ is the ceiling function and $\hat{q}$ is the empirical estimate of the $1-\alpha$ quantile adjusted for the size of the validation set. The conformal score (2) is a number in $[0,1]$, and it is close to 0 when the model $\hat{f}$ assigns high probability to a class prediction, and close to 1 otherwise.

Finally, given any new test data point $x \in X$, the prediction set $C(x)$ is defined to be

$$C(x) = \{y : s(x, y) \leq \hat{q}\}. \tag{3}$$

Thus, the size of $C(x)$ provides a measure of the model's confidence on the true classification of x, with higher confidence correlated with smaller size. Remarkably, the prediction set $C(x)$ as constructed satisfies (1) in the following more precise sense

$$1 - \alpha \leq \Pr(y \in C(x)) \leq 1 - \alpha + \frac{1}{n+1},$$

regardless of what the model class for $\hat{f}$ is and what the underlying (unknown) probability distribution $\mathcal{D}$ on $X \times \{1, \ldots, K\}$ is, as long as the training, validation, and test data are all independently and identically distributed according to $\mathcal{D}$. The reason the coverage formula holds is because, under the i.i.d. assumption which implies the training, validation and test data are all exchangable, the conformal score of the test data point has equal probability of falling anywhere in the ordered set of conformal scores for the validation set. Although coverage is distribution-free, the size of the conformal sets depends heavily on how informative the base predictor is. This is where the sample complexity [AB99, BM02] of the model class from which the base predictor is obtained matters.

In terms of applications, conformal prediction is well-suited for multi-label classification problems, for example in image-classification problems where there could be multiple objects in a scene. Importantly, conformal prediction can be used to improve model robustness in a principled model-agnostic way, whereby we only use a model's predictions when it is sufficiently confident. (See [AB21, §5.5] for details.) Through appropriately designed conformal score functions, conformal prediction schemes have been generalised to other problem classes like regression, outlier detection, and time-series prediction, including in setups where the underlying data distribution can shift / drift in different ways [GWDR21]. More recently, conformal prediction is being investigated for quantifying output uncertainty and controlling hallucination in LLMs [KLG+23, CGC24].

Split conformal prediction has also been generalised to a general method of risk control called Learn-then-Test [ABC+25]. The general setup is as follows. Given a model $\hat{f} : X \to Y$ trained using data generated i.i.d from an unknown distribution $\mathcal{D}$ on $X \times Y$, we construct a family of predictors $\hat{f}_\lambda : X \to Y'$, indexed by λ in some set Λ, where Y' is an arbitrary space related to Y. (E.g., in the case of prediction sets for classification problems, $Y' = 2^Y$.) We then allow the user to choose a way to measure the risk associated with each predictor $\hat{f}_\lambda$, which typically takes the following form

$$R(\hat{f}_\lambda) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\Big[l(\hat{f}_\lambda(x), y)\Big], \tag{4}$$

where $l(\cdot, \cdot)$ is a loss function and the risk score measures the expected loss with respect to the unknown underlying distribution $\mathcal{D}$.

Definition 1. We say a predictor $\hat{f}_\lambda$ is a (α, δ)-risk-controlling prediction if, with probability at $1-\delta$, we have $R(\hat{f}_\lambda) \leq \alpha$.

In practice, the goal of risk control, given user-specified (α, δ), is to use a validation set to estimate (4) for each possible $\hat{f}_\lambda$ to find one that satisfies Definition 1. The definition covers many special cases. For example, in the classification setting where $\hat{f}_\lambda$ is defined as in (3), we recover (1) by setting the loss function to be $l(\hat{f}_\lambda(x), y) = \mathbb{1}\left[y \notin \hat{f}_\lambda(x)\right]$ since

$$R(\hat{f}_\lambda) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathbb{1}(y \notin \hat{f}_\lambda(x))\right] = P(y \notin \hat{f}_\lambda(x)) = 1 - P(y \in \hat{f}_\lambda(x)).$$

Many other interesting setups like multi-label classification with controlled false-discovery rate, and simultaneous guarantees on out-of-distribution detection and coverage are covered in [AB21, ABC+25].

2 Online Conformal Prediction

We now provide a brief discussion of conformal prediction in the online learning setting, following the treatment in [GV07]. In online learning, data arrives sequentially and we have to make a prediction at every time step. In particular, at time t, we would have observed a sequence of data

$$[(x_1, y_1), (x_2, y_2), \ldots, (x_{t-1}, y_{t-1})],$$

where $x_i \in X, y_i \in Y$ and we are asked to predict the label of a new object x_t. Building on the general result in algorithm information theory that a sequence is random iff it is unpredictable [Sch24], a general approach to solve the above prediction problem is to use a randomness test to measure, for each $y_t \in Y$, how unusual the possible continuation of the sequence

$$[(x_1, y_1), (x_2, y_2), \ldots, (x_{t-1}, y_{t-1}), (x_t, y_t)]$$

is and then use the randomness test scores to rule out unlikely labels.

Formally, a (computable) function $T : (X \times Y)^* \to [0, 1]$ is a randomness test if for all $\epsilon \in (0, 1)$, all $t \in \{1, 2, \ldots\}$, and all probability distributions $\mathcal{D}$ on $X \times Y$, we have

$$\mathcal{D}^t\{z \in (X \times Y)^t : T(z) \leq \epsilon\} \leq \epsilon.$$

In other words, under the assumption that a sequence z is drawn i.i.d from an (unknown) distribution $\mathcal{D}$, if $T(z) \leq 0.01$, then the probability of seeing z is at most 0.01 regardless of what $\mathcal{D}$ is.

There are different ways to design practical randomness tests. The general approach taken in [VGS05] starts with a nonconformity function that maps every data sequence $[(x_1, y_1), \ldots, (x_t, y_t)]$ to a sequence of nonconformity scores $[\alpha_1, \ldots, \alpha_t]$ in such a way that interchanging any two data points (x_i, y_i) and (x_j, y_j) leads to the interchange of α_i and α_j with everything else staying the

same. Nonconformity scores are so named because each of the score α_i for data point (x_i, y_i) is supposed to capture how unusual that data point is with respect to the other data points in the data sequence, and the higher the nonconformity score is, the more unusual a data point is.

Armed with nonconformity scores, we can then construct a randomness test using the so-called p-value associated with each possible y_t defined as follows

$$p_{y_t} = \frac{|\{i \in \{1, \ldots, t\} : \alpha_i \geq \alpha_t\}|}{t}, \tag{5}$$

which is the proportion of the nonconformity scores that are at least as large as the last value α_t corresponding to (x_t, y_t). Given such a randomness test, the conformal predictor is defined as the predictor that, at time t, given past data $[(x_1, y_1), \ldots, (x_{t-1}, y_{t-1})]$, a new object x_t, and a desired confidence level $1 - \epsilon \in (0, 1)$, outputs the prediction set

$$\Gamma^{\epsilon}([(x_1, y_1), \ldots, (x_{t-1}, y_{t-1})], x_t) = \{y_t \in Y : p_{y_t} > \epsilon\}. \tag{6}$$

Under the assumption that data are generated i.i.d from an unknown distribution $\mathcal{D}$, which implies the exchangeability assumption on the nonconformity scores, the conformal predictor comes with the guarantee that the actual value of y_t for x_t drawn from $\mathcal{D}$ is in the prediction set (6) with probability at least $1 - \epsilon$.

There are a variety of ways of designing nonconformity scores. We have seen how nonconformality score can be defined with respect to a given classifier $\hat{f}$ in (2). We can similarly define the nonconformality score for a regression model $\hat{h}$ using the residuals

$$\alpha_i = s(x_i, y_i) = |y_i - \hat{h}(x_i)|,$$

as is done in [NMV01]. (It is worth noting that monotonic transformations of a nonconformity score will not change the output of conformal prediction, since it is only the rank order of nonconformity scores that matter.) If $\hat{f}$ and $\hat{h}$ are not black boxes but have some known structure, we can sometimes exploit that structure to construct the nonconformity scores. For example, given a support vector machine, we can use the Lagrange multipliers associated with each training data as the nonconformity scores.

3 Conformal Testing

Conformal prediction has been extended beyond exchangeable data to also work on a class of online compression models [Vov06] that includes variable-order Markov models and various statistical models based on exponential family distributions. It turns out there is also something between exchangeability and online compression models where conformal prediction works well and can be computed efficiently: martingales. The details of the latter can be found in [Vov25, VNG25]. These estimators work with so-called e-values rather than p-values; e-values can always be obtained from p-values and the 'e' stands for

expectations. An important application for conformal e-testing is that it can be used to detect and quantify model misspecification when using Bayesian mixture estimators like Context Tree Weighting [WST95], variations of which have been used successfully in various approximations of the AIXI model [HQC24], including [VNH+11, YZWN22, YZNH24, NYZCC25]. It maybe possible to improve the robustness of these universal reinforcement learning algorithms using conformal prediction at relatively low computational cost. More speculatively, this line of research could result in general methods for dealing with the problematic behaviour of Bayesian predictors under model misspecification [GVO17, vEGM+15], by putting together in complementary ways arguably some of the most useful estimators from the Bayesian and Frequentist approach to probability theory. These questions will be explored in greater detail in a separate paper.

References

[AB99] Martin Anthony and Peter L Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.

[AB21] Anastasios N Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv:2107.07511*, 2021.

[ABC+25] Anastasios N Angelopoulos, Stephen Bates, Emmanuel J Candès, Michael I Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641–1662, 2025.

[BM02] Peter L Bartlett and Shahar Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

[CGC24] John Cherian, Isaac Gibbs, and Emmanuel Candes. Large language model validity via enhanced conformal prediction methods. *Advances in Neural Information Processing Systems*, 37:114812–114842, 2024.

[GV07] Alexander Gammerman and Vladimir Vovk. Hedging predictions in machine learning. *The Computer Journal*, 50(2):151–163, 2007.

[GVO17] Peter Grünwald and Thijs Van Ommen. Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Analysis*, 12:1069–1103, 2017.

[GWDR21] Asaf Gendler, Tsui-Wei Weng, Luca Daniel, and Yaniv Romano. Adversarially robust conformal prediction. In *International Conference on Learning Representations*. PMLR, 2021.

[HQC24] Marcus Hutter, David Quarel, and Elliot Catt. *An Introduction to Universal Artificial Intelligence*. CRC Press, 2024.

[KLG+23] Bhawesh Kumar, Charlie Lu, Gauri Gupta, Anil Palepu, David Bellamy, Ramesh Raskar, and Andrew Beam. Conformal prediction with large language models for multi-choice question answering. *arXiv preprint arXiv:2305.18404*, 2023.

[NMV01] Ilia Nouretdinov, Thomas Melluish, and Volodya Vovk. Ridge regression confidence machine. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 385–392, 2001.

[NYZCC25] Kee Siong Ng, Samuel Yang-Zhao, and Timothy Cadogan-Cowper. The problem of social cost in multi-agent general reinforcement learning: Survey and synthesis. *arXiv:2412.02091*, 2025.

[Sch24] Lenhart K Schubert. Predictability and randomness. *arXiv:2401.13066*, 2024.

[vEGM+15] Tim van Erven, Peter Grunwald, Nishant A Mehta, Mark Reid, and Robert Williamson. Fast rates in statistical and online learning. *Journal of Machine Learning Research*, 2015.

[VGS05] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

[VNG25] Vladimir Vovk, Ilia Nouretdinov, and Alex Gammerman. Conformal e-testing. *Pattern Recognition*, page 111841, 2025.

[VNH+11] Joel Veness, Kee Siong Ng, Marcus Hutter, William Uther, and David Silver. A Monte-Carlo AIXI approximation. *Journal of Artificial Intelligence Research*, 40:95–142, 2011.

[Vov06] Vladimir Vovk. Well-calibrated predictions from on-line compression models. *Theoretical computer science*, 364(1):10–26, 2006.

[Vov25] Vladimir Vovk. Conformal e-prediction. *Pattern Recognition*, page 111674, 2025.

[WST95] Frans MJ Willems, Yuri M Shtarkov, and Tjalling J Tjalkens. The context-tree weighting method: Basic properties. *IEEE Transactions on Information Theory*, 41(3):653–664, 1995.

[YZNH24] Samuel Yang-Zhao, Kee Siong Ng, and Marcus Hutter. Dynamic knowledge injection for AIXI agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38(15), pages 16388–16397, 2024.

[YZWN22] Samuel Yang-Zhao, Tianyu Wang, and Kee Siong Ng. A direct approximation of AIXI using logical state abstractions. *Advances in Neural Information Processing Systems*, 35:36640–36653, 2022.