Ellipsoidal distribution-free set

Dmitriy Klyushin [1,*,†], Andrii Tymoshenko[1,*,†]

[1]Taras Shevchenko National University of Kyiv, 60 Volodymyrska Street, 01033, Kyiv, Ukraine

Abstract

This paper introduces a distribution-free approach based on the Hill's assumption and the Petunin ellipsoids. Several distributions are used to generate points and build ellipsoids, which are then used to check if test points with same distribution are created inside largest ellipsoid. As a result, a new prediction set is constructed in the form of Petunin ellipsoid, while confidence level refers to the number of points. The method described here works effectively for chosen distributions. Moreover, statistical analysis of the quantity of points inside is performed. This method is a useful tool for solving many urgent problems of machine learning, e.g. generalization of training samples, effective cross-validation etc.

Keywords

data mining, prediction set, Petunin ellipsoid, outlier detection

1. Introduction

Construction of prediction sets is a popular problem in machine learning, often placed together with neural networks. In recent years many international scientists have discussed prediction sets. For example, Adam Khakhar, Stephen Mell and Osbert Bastani [1] used a trained code generation model in algorithm that leverages an abstract syntax tree based on programming language to create a set of programs with high confidence about the correct program. Another example by Soroush H. Zargarbashi, Mohammad Sadegh Akhondzadeh and Aleksandar Bojchevski [2] derive provably robust sets by defining bounds for the worst-case change related to conformity scores.

Another important idea is to find distribution-free classification and sets. Here we can mention a work by Chirag Gupta, Aleksandr Podkopaev and Aaditya Ramdas [3] where they study calibration and prediction sets combined with confidence intervals. Their research is dedicated to binary classification in case of distribution-free sets. Based on demonstrated theorems, confidence intervals for binned probabilities allow to perform distribution free calibration.

In [4] Hongxiang Qiu, Edgar Dobriban and Eric Tchetgen Tchetgen offer a novel flexible distribution-free method named PredSet-1Step for constructing prediction sets where asymptotic coverage is guaranteed under unknown covariate shift.

As for [5] A.N. Angelopoulos, S. Bates, J. Malik and M.I. Jordan present an algorithm which changes a chosen classifier to determine a predictive set, where the true label is inside with probability set by user. This simple and fast algorithm reminds of Platt scaling but results into a formal finite-sample coverage for every model and dataset.

Approach described in [6] by analysis of a holdout set allows to choose the size of the prediction sets and leads to explicit finite-sample guarantees. As a result, simple, distribution-free and

Information Technology and Implementation (IT&I-2024), November 20-21, 2024, Kyiv, Ukraine

* Corresponding author.

† These authors contributed equally.

dmytroklyushin@knu.ua (D. Klyushin); andriitymoshenko@knu.ua (A. Tymoshenko)

0000-0003-4554-1049 (D. Klyushin); 0000-0002-8884-9054 (A. Tymoshenko)

rigorous error control is obtained for many tasks, demonstrated on five large-scale machine learning problems.

Some more works related to predictions offer various approaches: prediction based on language models [7], neural networks compared to calibrated predictions [8], distribution-free uncertainty quantification and conformal prediction [9], conformal risk control [10], conformal predictors applied for medical imaging [11], confident prediction in case distributions shift [12], conformal prediction robust to label noise [13], conformal prediction via probabilistic circuits [14].

Not only predictions attract modern scientists. Some related topics are also worth being mentioned: uncertainty quantification is performed over graph using conformalized graph neural networks [15], adversarial robustness applied to randomly smoothed classifiers [16], randomized smoothing for graphs and images [17], adversarially trained smoothed classifiers [18].

The purpose of our paper is to describe a method to construct ellipsoidal prediction set for a set of the randomly generated points, based on chosen distribution. The main tools for our forecast are predictive sets represented by ellipses, constructed using generated points. Test points are generated with same distribution, the more ellipses contain a point – the higher probability of belonging to same class can be expected. Consider the problem of creating conformal prediction based on points $x_1, x_2, ..., x_m \in \mathbb{R}^d$. The aim is to find a prediction set $E(x_1, x_2, ..., x_m) \subset \mathbb{R}^d$ resulting into probability $p(x_{m+1} \in E) \geq 1-\alpha$, where $0 < \alpha < 1$ is a chosen significance level, so that $p(x_{m+1} \in E)$ is the confidence level of the predictive set.

2. Hill`s Assumption $A_{(m)}$

As $x_1, x_2, ..., x_m$ we denote a sample drawn from a population generated according to absolutely continuous distribution *F*. Next, we arrange it in the increasing order and create the variance series $x_{(1)} \leq x_{(2)} \leq ... \leq x_{(m)}$, where $x_{(i)}$ is *i*-th order statistics. The resulting order statistics $x_{(1)}, x_{(2)}, ..., x_{(m)}$ are dependent. The distribution $F_k(x)$ of the *k*-th order statistics $x_{(k)}$ can be calculated as $F_k(u) = \sum_{i=k}^{m} C_m^i \left[F(u)\right]^i \left[1-F(u)\right]^{m-1}$, where $F(u) = p(x \leq u)$.

The Hill's assumptions $A_{(m)}$ [19] states that if $x_{(i)}$ is chosen from the population according to distribution F then

$$p\left(x_{m+1} \in \left(x_{(i)}, x_{(j)}\right)\right) = \frac{j-i}{m+1}, \; j > i. \tag{1}$$

The Hill's assumption $A_{(n)}$ was proven in papers of Yu.I. Petunin [20] and by several other scientists. Let us recall the proof for random variables ξ and η. If they are independent, then

$$p(\xi < \eta) = \int_{-\infty}^{\infty} F_{\xi}(u) dF_{\eta}(u), \tag{2}$$

where $F_{\xi}(u)$ and $F_{\eta}(u)$ denote the distribution functions of ξ and η, respectively. The probability density of *i*-th order statistics is:

$$f_k(u) = \sum_{i=k}^{m} C_m^i \left[F(u)\right]^i \left[1-F(u)\right]^{m-i} = \sum_{i=k}^{m} G_i(u). \tag{3}$$

Hence, $F_k'(u) = \sum_{i=k}^{m} G_i'(u)$,

$$G_k'(u) = C_m^k \left[k\left(F(u)\right)^{k-1} \left(1-F(u)\right)^{m-k} f(u) - \left(F(u)\right)^k (m-k) \left(1-F(u)\right)^{m-k-1} f(u) \right] \tag{4}$$

$$G'_{k+1}(u)=\left[C_m^{k+1}\left(F(u)\right)^{k+1}\left(1-F(u)\right)^{m-k-1}\right]'=$$
$$=C_m^{k+1}\left[(k+1)\left(F(u)\right)^{k}\left(1-F(u)\right)^{m-k-1}f(u)-\left(F(u)\right)^{k+1}\left(1-F(u)\right)^{m-k-2}(m-k-1)f(u)\right].$$

The second term is compensated by the first term:

$$-C_m^k(m-k)+C_m^{k+1}(k+1)=-\frac{m!(m-k)}{(m-k)!k!}+\frac{m!(k+1)}{(m-k-1)!(k+1)!}=-\frac{m!}{(m-k-1)!k!}+\frac{m!}{(m-k-1)!k!}=0.$$

The last term of the previous sum is equal to zero

$$\left(F(u)\right)^{m}\left(1-F(u)\right)^{0}(m-m)f(u)=0\ .$$

Thus,

$$f_k(u)=mC_{m-1}^{k-1}\left[F(u)\right]^{k-1}\left[1-F(u)\right]^{m-k}f(u).$$

Let us find $p\left(x<x^{(i)}\right)$ and $p\left(x<x^{(j)}\right)$. Using the above equations, we have

$$p\left(x<x^{(i)}\right)=\int_{-\infty}^{\infty}F(u)dF_i(u)=mC_{m-1}^{i-1}\int_{-\infty}^{\infty}\left[F(u)\right]^{i}\left[1-F(u)\right]^{m-i}dF(u)=mC_{m-1}^{i-1}\int_0^1 v^i(1-v)^{m-1}dv.$$

It is proven that,

$$\int_0^1 x^{p-1}(1-x)^{q-1}dx=\frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q-1)}=\frac{(p-1)!(q-1)!}{(p+q-1)!}\ .$$

We can apply this equation as

$$\int_0^1 x^{j+1-1}(1-x)^{m-i+1-1}dx=B(i+1,m-i+1)=\frac{\Gamma(i+1)\Gamma(m-i+1)}{\Gamma(i+1+m+1-i)}=\frac{\Gamma(i+1)\Gamma(m-i+1)}{\Gamma(m+2)}=\frac{i!(m-i)!}{(m+1)!}$$

$$p\left(x_{m+1}<x_{(j)}\right)=mC_{m-1}^{i-1}\frac{j!(m-j)!}{(m+1)!}=\frac{m(m-1)!(j-1)!\,j!(m-j)!}{(m-j)!(j-1)!m(m+1)(m-1)!}=\frac{j}{m+1}.$$

Previous equation was obtained by multiplying numerator and denominator by (j − 1). So, we get

$$p\left(x_{m+1}\in\left(x_{(i)},x_{(j)}\right)\right)=p\left(x_{m+1}<x_j\right)-p\left(x_{m+1}<x_i\right)=\frac{j}{m+1}-\frac{i}{m+1}=\frac{j-i}{m+1}.$$

So, in case a random variable x is independent from $x_1,x_2,...,x_m$ and it is chosen by sampling from the same population based on distribution $F(u)$, then

$$p\left(x\in\left(x_{(1)},x_{(m)}\right)\right)=\frac{m-1}{m+1}\ .$$

Remark 1. The confidence level of the tolerance interval $\left(x_{(1)},x_{(m)}\right)$ is $\frac{m-1}{m+1}$, thus for $m\geq 39$ the confidence level of this interval is less than 0.05.

3. Petunin Ellipsoids

The algorithm for construction of the ellipsoid containing the set as m random points with the probability $\frac{m-1}{m+1}$ was proposed by Yu. I. Petunin. Statistical and geometrical properties of the Petunin ellipsoids were investigated in [21].

Here we applied the Petunin's algorithm for two-dimensional case. First, we find two points farthest from each other $\vec{x}_k$ and $\vec{x}_l$ of the set $M_n=\{\vec{x}_1,...,\vec{x}_m\}$. Connect them with a line (next mentioned as diameter), then project all the points to the hyperplane orthogonal to this line. To simplify this, we can rotate all objects together around line center to make it horizontal. But then we will need to rotate it back in the end (Figure 1).

Figure 1: The furthest from each other points

Next, we need to find the farthest points from the line. Construct lines parallel to the diameter through these points. Create lines that are orthogonal to diameter and pass through the farthest from each other points. As a result, we obtain a rectangle, which covers the given set of points and lies on a two-dimensional plane (Figure 2).

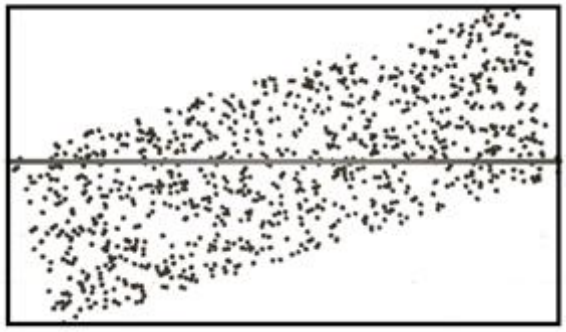

Figure 2: Rectangle covering the images of the points

By dividing the shorter side length by the longer side length, we can obtain the shrinking coefficient. Translate, rotate and shrink mentioned above rectangle to construct a square covering the given points.

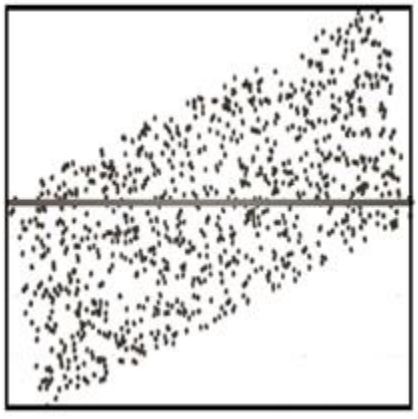

Figure 3: Square covering the images of the points

Find its center and all distances from the center of the square to every image of point. Then, we need to find maximal distance and create a circle with center same as square center and radius that is equal to maximum distance from the center to the images of points.

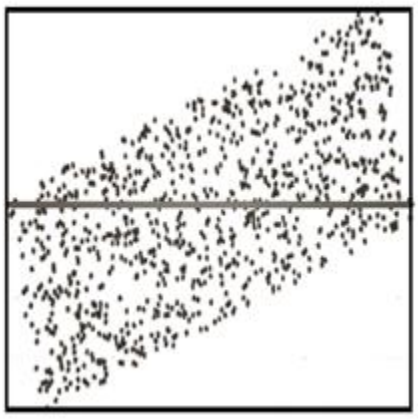

Figure 4: The circle covering the images of the points

Perform inverse transformations of this circle. The result is the Petunin ellipse (Figure 5).

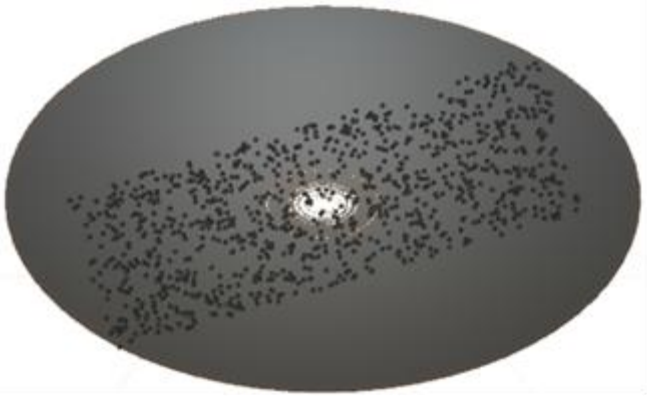

Figure 5: Petunin ellipse covering the initial points

In high-dimensional case, we can construct a minimum volume axis-aligned orthogonal parallelepiped that contains images $\vec{x}_1',...,\vec{x}_m'$ of initial points. Perform shrinking transformation from the orthogonal parallelepiped to a hypercube. Find its center and distances from it to $\vec{x}_1',...,\vec{x}_m'$. Next find the maximum distance. After that, construct a hypersphere with the center and radius that is equal to the maximum distance from the center to the images of the points. Perform inverse transformations (translation, rotation and stretching) and obtain the Petunin ellipsoid E_m. Hill's assumption $A_{(m)}$ is true, so $P(x_{m+1} \in E) = \frac{m-1}{m+1}$.

Since at the last stage of construction of the Petunin ellipsoid we obtain the concentric spheres with one unique point at the surface, using the Petunin ellipsoids we can arrange the points by their statistical depth. The median point of the set (the most probable point) is the point nearest to the center of the Petunin ellipsoid and the outlier is the point at the boundary of the Petunin ellipsoid.

4. Numerical results

In this section testing results are described. First, we generate a chosen distribution-based set of 1000 points and build ellipses through each point. Then we generate 1000 more points with the

same distribution and check the number of points inside the largest ellipse. Statistical characteristics of these results are shown below for three different distributions.

4.1. Normal distribution

The first test was performed on normal distribution testing, 3 to 1. We generate 12 random numbers from 0 to 1, calculate their sum and subtract 6. Then we modify values by multiplying the result and adding value so that horizontal and vertical coordinate values can be generated in proportion 3 to 1.

Expected probability 0.99762
Mean 0.99762
Mean Deviation 0.001873
Mode 0.996477
Median 0.998
Standard Deviation 0.0022867
Variance 0.0000052279
Kurtosis 3.24629
Skewness -0.892779
Range 0.01
Maximum 1
Minimum 0.989
Geometric Mean 0.9976174
Harmonic Mean 0.9976148

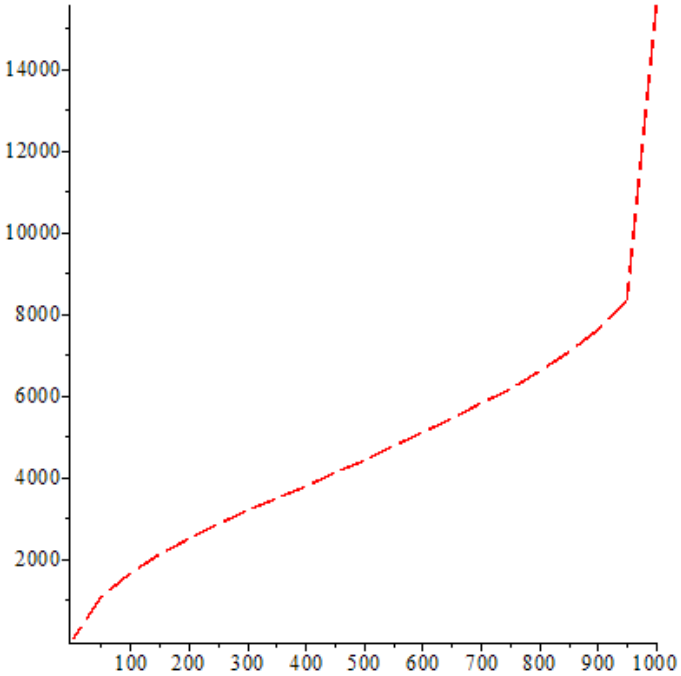

Figure 6: Average ellipse areas – Normal distribution, 100 tests

As we can see, the largest ellipse contains almost all points and areas slowly grow until last 50 points defining largest ellipses.

4.2. Exponential distribution

Exponential with parameters -17, -50 as multipliers for logarithm from random value from 0 to 1.

Expected probability 0.998158
Mean 0.998158
Mean Deviation 0.00138457
Mode 0.99943
Median 0.999
Standard Deviation 0.0018747
Variance 0.00000351465
Kurtosis 5.98476568
Skewness -1.5228786
Range 0.01
Maximum 1
Minimum 0.989999
Geometric Mean 0.998
Harmonic Mean 0.998

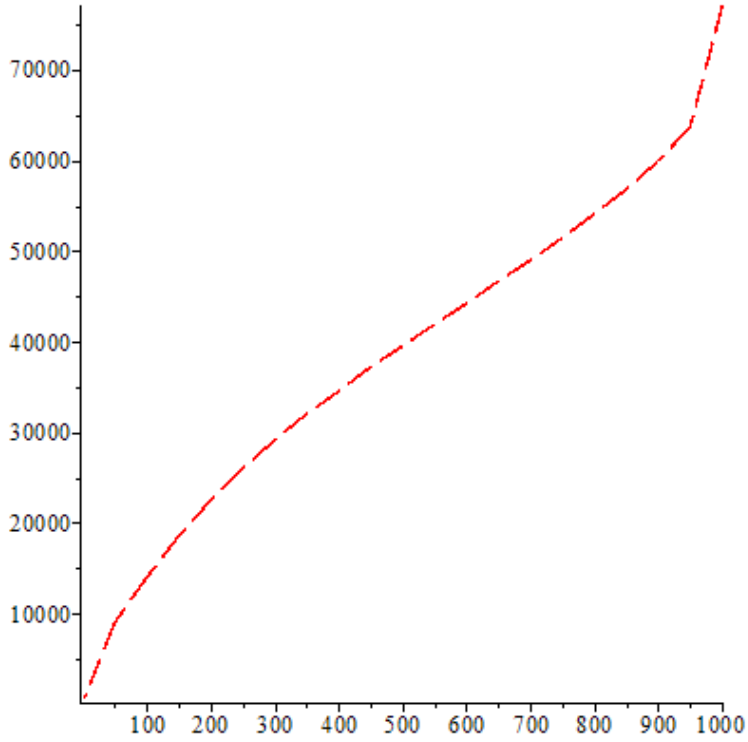

Figure 7: Average ellipse areas – Exponential distribution, 100 tests

Here the largest ellipse contains almost all points, but ellipse areas grow much faster.

4.3. Gamma distribution

Gamma distribution was used here based on pseudorandom numbers with parameters 50 and 90 for horizontal and vertical values respectively.

Expected probability 0.997760000000000

Mean 0.997760000000000
Mean Deviation 0.0017616
Mode 0.998438
Median 0.998
Standard Deviation 0.002399
Variance 0.0000057599
Kurtosis 5.95855
Skewness -1.65869
Range 0.012
Maximum 1
Minimum 0.987999
Geometric Mean 0.997757
Harmonic Mean 0.997754

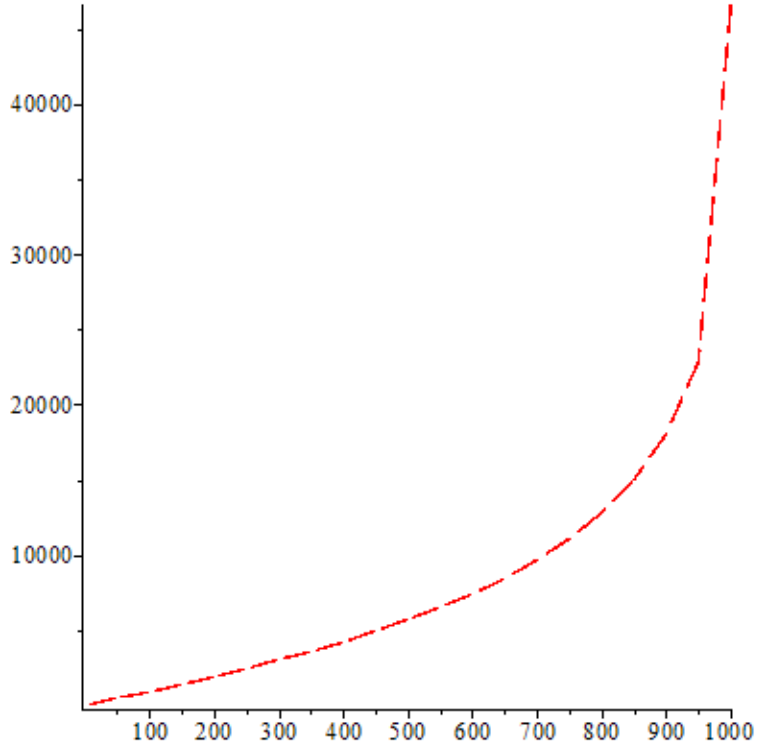

Figure 8: Average ellipse areas – Gamma distribution, 100 tests

In this test ellipse contains almost all points and ellipse areas increase very fast after ellipse based on 800 points.

More tests were performed for these distributions with other parameters and results were alike. Ellipse areas increased smoothly at first, but for the last 100-200 most distant points faster increase was reported. As for accuracy, we expected values to be approximately 0.998 and received alike results.

In our previous work, we demonstrated that Hill's assumption is also true for uniform distribution with rectangular area covered.

5. Conclusion

Constructing Petunin ellipsoid is a useful approach for arranging data and detecting anomalies using statistical depth. According to obtained results, the algorithm leads to effective prediction

sets based on Petunin ellipsoid. The confidence level reached is theoretically precise for tested distributions. It allows us to compute statistical depth based on every point and detect outliers of the set. Experimental results approved theoretical properties of the Petunin ellipses.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] A. Khakhar, S. Mell, O. Bastani, PAC Prediction Sets for Large Language Models of Code. 2023. doi:10.48550/arXiv.2302.08703

[2] S. H. Zargarbashi, M. S. Akhondzadeh, A. Bojchevski, Robust Yet Efficient Conformal Prediction Sets. 2024. doi:10.48550/arXiv.2407.09165.

[3] C. Gupta, A. Podkopaev, A. Ramdas, Distribution-free binary classification: prediction sets, confidence intervals and calibration, 2020. URL: https://proceedings.neurips.cc/paper/2020/file/26d88423fc6da243ffddf161ca712757-Paper.pdf.

[4] H. Qiu, E. Dobriban, E. Tchetgen, Prediction sets adaptive to unknown covariate shift, Journal of the Royal Statistical Society Series B: Statistical Methodology 85(5) (2023) 1680–1705. doi:10.1093/jrsssb/qkad069.

[5] A. Angelopoulos, S. Bates, J. Malik, M. I. Jordan, Uncertainty sets for image classifiers using conformal prediction, 2020. URL: https://arxiv.org/abs/2009.14193.

[6] S. Bates, A. Angelopoulos, L. Lei, J. Malik, M. I. Jordan, Distribution-free, risk-controlling prediction sets, 2021. URL: https://arxiv.org/abs/2101.02703.

[7] T. Liu, Y. Jiang, N. Monath, R. Cotterell, M. Sachan, Autoregressive structured prediction with language models, 2022. URL: https://arxiv.org/abs/2210.14698.

[8] S. Park, O. Bastani, N. Matni, I. Lee, Pac confidence sets for deep neural networks via calibrated prediction, 2020. URL: https://arxiv.org/abs/2001.00106.

[9] A. N. Angelopoulos, S. Bates, A gentle introduction to conformal prediction and distribution-free uncertainty quantification, 2021. URL: https://arxiv.org/abs/2107.07511.

[10] A. N. Angelopoulos, S. Bates, A. Fisch, L. Lei, T. Schuster, Conformal risk control. ArXiv, abs/2208.02814, 2022. URL: https://api.semanticscholar.org/CorpusID:251320513.

[11] C. Lu, A., Lemay, K. Chang, K. Hobel, J. Kalpathy-Cramer, Fair conformal predictors for applications in medical imaging, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, 2022, pp. 12008–12016.

[12] M. Cauchois, S. Gupta, A. Aliand, J.C. Duchi, Robust validation: Confident predictions even when distributions shift. arXiv preprint arXiv:2008.04267, 2020.

[13] B.-S. Einbinder, S. Bates, A. N. Angelopoulos, A. Gendler, Y. Romano, Conformal prediction is robust to label noise. ArXiv, abs/2209.14295, 2022. URL: https://api.semanticscholar.org/CorpusID:262091979.

[14] M. Kang, N. M. Gurel, L. Li, B. Li, COLEP: Certifiably robust learning-reasoning conformal prediction via probabilistic circuits. In The Twelfth International Conference on Learning Representations, 2024. URL: https://openreview.net/forum?id=XN6ZPINdSg.

[15] K. Huang, Y. Jin, E. Candes, J. Leskovec, Uncertainty quantification over graph with conformalized graph neural networks, 2023. URL: https://arxiv.org/pdf/2305.14535.

[16] G.-H. Lee, Y. Yuan, S. Chang, T. Jaakkola, Tight certificates of adversarial robustness for randomly smoothed classifiers. Advances in Neural Information Processing Systems, 32, 2019.

[17] A. Bojchevski, J. Gasteiger, S. Gunnemann, Efficient robustness certificates for discrete data: Sparsity-aware randomized smoothing for graphs, images and more. In: International Conference on Machine Learning, pp. 1003–1013. PMLR, 2020.

[18] H. Li, J. Salman, I. Razenshteyn, P. Zhang, H. Zhang, S. Bubeck, G. Yang, Provably robust deep learning via adversarially trained smoothed classifiers. Advances in Neural Information Processing Systems, volume 32, 2019.
[19] B. M. Hill, Posterior distribution of percentiles: Bayes' theorem for sampling from a population, Journal of the American Statistical Association 63 (1968) 677–691.
[20] I. Madreimov, Yu. I. Petunin, Characterization of the uniform distribution using order statistics, Theory of Probability and Mathematiical Statistics 27 (1983) 105–110.
[21] S. I. Lyashko, B. V. Rublev, Minimal Ellipsoids and Maximal Simplexes in 3D Euclidean Space, Cybernetics and Systems Analysis 39 (2003) 831–834. doi:10.1023/B:CASA.0000020224.83374.d7.