Fact Sheet 1: Dvoretzky–Kieffer–Wolfowitz inequality
2019/20 ICFP Master (first year)

Botao Li, Valentina Ros, Victor Dagard, Werner Krauth

Introduction and example of DKW inequality

I. INTRODUCTION

In this factsheet, we discuss one of the fundamental (and quite recent) achievements in non-parametric statistics, the DKW inequality, which dates back to 1956[1, 2] but was proven in a practically useful (tight and non-asymptotic) version only in 1990 [3]. From a finite number of samples of an (unknown) distribution, it computes a region (a "corridor") that contains the entire CDF with high probability. The bound is entirely independent of F. The DKW inequality translates generation-old tools-of-the-trade into rigorous mathematics.

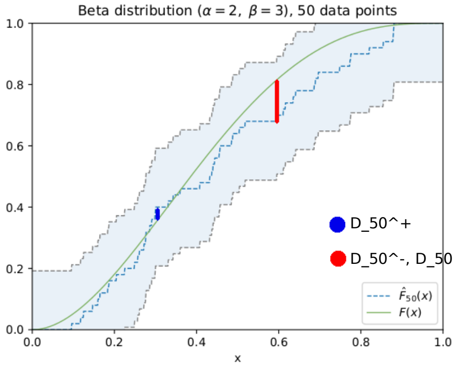

FIG. 1. A CDF $F(x)$ and an empirical CDF $\hat{F}$ for $n = 50$. The statistics D_{50}^+, D_{50}^-, and D_{50} are indicated. The colored area is obtained by shifting $\hat{F}$ up and down by a distance ϵ. With a properly chosen ϵ, it contains the entire CDF with high probability.

II. DEFINITIONS, STATEMENT OF RESULT

Let $\xi_1, ..., \xi_n$ be independent, identically distributed real-valued random variables with a continuous cumulative distribution function (CDF) F. Define the empirical CDF

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{(\xi_i \leq x)}$$

(see Fig. 1). We define (see [3]):

$$D_n^+ = \sqrt{n} \sup_x \left[\hat{F}_n(x) - F(x) \right] ; \quad D_n^- = \sqrt{n} \sup_x \left[F(x) - \hat{F}_n(x) \right] ; \quad D_n = \sqrt{n} \sup_x |F(x) - \hat{F}_n(x)| \qquad (1)$$

(see Fig. 1).

The Dvoretzky–Kieffer–Wolfowitz(DKW) inequality [1, 2] is

$$\mathbb{P}\left\{D_n^+ > \epsilon\right\} = \mathbb{P}\left\{D_n^- > \epsilon\right\} \leq \mathrm{e}^{-2\epsilon^2}, \qquad \text{when } \epsilon \geq \min\left(\sqrt{(\log 2)/2}, \gamma n^{-1/6}\right) \tag{2}$$

where $\gamma = 1.0841$. Since $n \in \mathbb{Z}_{>0}$, the condition for the DKW inequality to be valid indicates $\mathrm{e}^{-2\epsilon^2} \leq 1/2$. Since

$$\mathbb{P}\left(D_n > \epsilon\right) \leq 2\mathbb{P}\left(D_n^+ > \epsilon\right) \tag{3}$$

there is a two-sided constraint which is valid for all values of ϵ:

$$\mathbb{P}\left[\sup_x \left|\hat{F}_n(x) - F(x)\right| > \epsilon\right] \leq 2\mathrm{e}^{-2n\epsilon^2} \tag{4}$$

Shifting the empirical CDF up and down by ϵ constructs a region which has more than $1 - 2\exp\left(-2n\epsilon^2\right)$ chance of entirely covering $F(x)$ (see Fig. 2). The DKW inequality does not depend on the CDF $F(x)$.

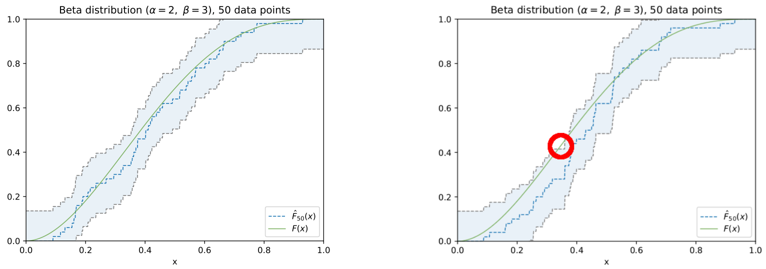

FIG. 2. The colored region is constructed by shifting the empirical CDF up and down by ϵ. Here, $\exp\left(-2n\epsilon^2\right) = 0.32$. (a): The colored region covers the CDF for all x. (b): The colored region does not cover the CDF for all x. The DKW inequality bounds the probability with which this can happen. This bound is independent of the distribution F.

The DKW inequality was originally proposed with an unspecified constant instead of the 1 on the right-hand side of eq. (2). This tight constant, i.e. $C = 1$, was proven only in 1990 [3]. When the number of data points $\to \infty$, the inequality in eq. (2) becomes tight.

The DKW inequality thus provides a "corridor", which replaces error bars. It teaches us that we should look at cumulative distribution functions F (for which bounds exist), rather than at histograms of the probability density function f.. Finally, DKW illustrates that the best way to compare two random variables is to study their double-side Kolmogorov distance, D_n (see eq. (1)). In fact, the Kolmogorov distance is the statistic given by the Kolmogorov–Smirnov test, which measures the difference between two distributions. The DKW inequality provides the confidence interval of the Kolmogorov–Smirnov test, which is valid regardless of the distribution and of n.

III. EXAMPLES

The probability density function of beta distribution, which is defined on $[0, 1]$, is

$$f(x;\alpha,\beta) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{1-\beta}$$

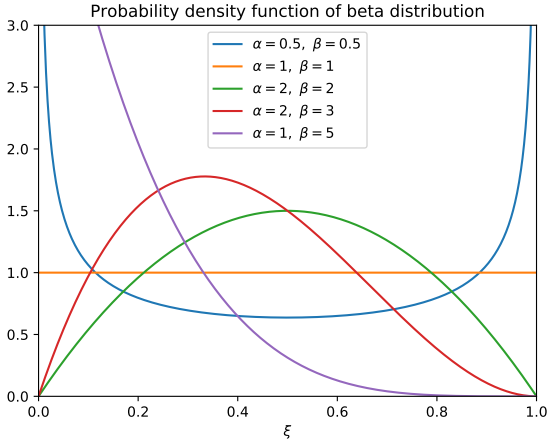

FIG. 3. Probability density function of beta distribution. The PDF plotted in red is used to generate data points.

depending on the value of parameters α and β, the beta distribution can have various properties (Fig. 3). Beta distributions appear in physics, for example, in the 1D hard-sphere system.

5, and 5000 data points are generated by beta distribution ($\alpha = 3$ and $\beta = 2$). $F(x)$ and $\hat{F}_n(x)$ are plotted in Fig. 4. (These are just enumerations. It is impossible to make statistical statement from them.) The colored region is the region indicated by eq. (4). The value of ϵ is tuned so that $2\exp\left(-2n\epsilon^2\right) = 0.05$. This means, if infinite number of data sets are generated, the region constructed from 5% of them will not cover the CDF. As the number of data points increases, the region become thinner and thinner (while leaving the value of ϵ unchanged). And when there are a considerable number of data points, DKW inequality can give a good constraint on the probability function without any assumption of it.

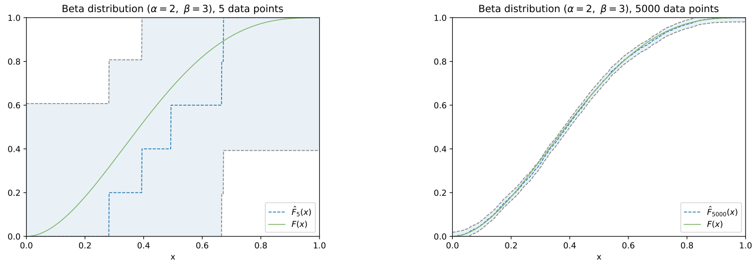

FIG. 4. Examples of the DKW interval corridor for $n = 5$ and $n = 5000$.

Using the same distribution, 9 sets of 50 data points were generated (see Fig. 5). Here, ϵ is tuned so that $\exp\left(-2n\epsilon^2\right) = 0.32$, which means the probability of failing to cover the CDF is less than 32%. The regions in Fig. 5 and are already quite thin. For 1000 data sets, 30.7% of them fail to cover the CDF. When the same

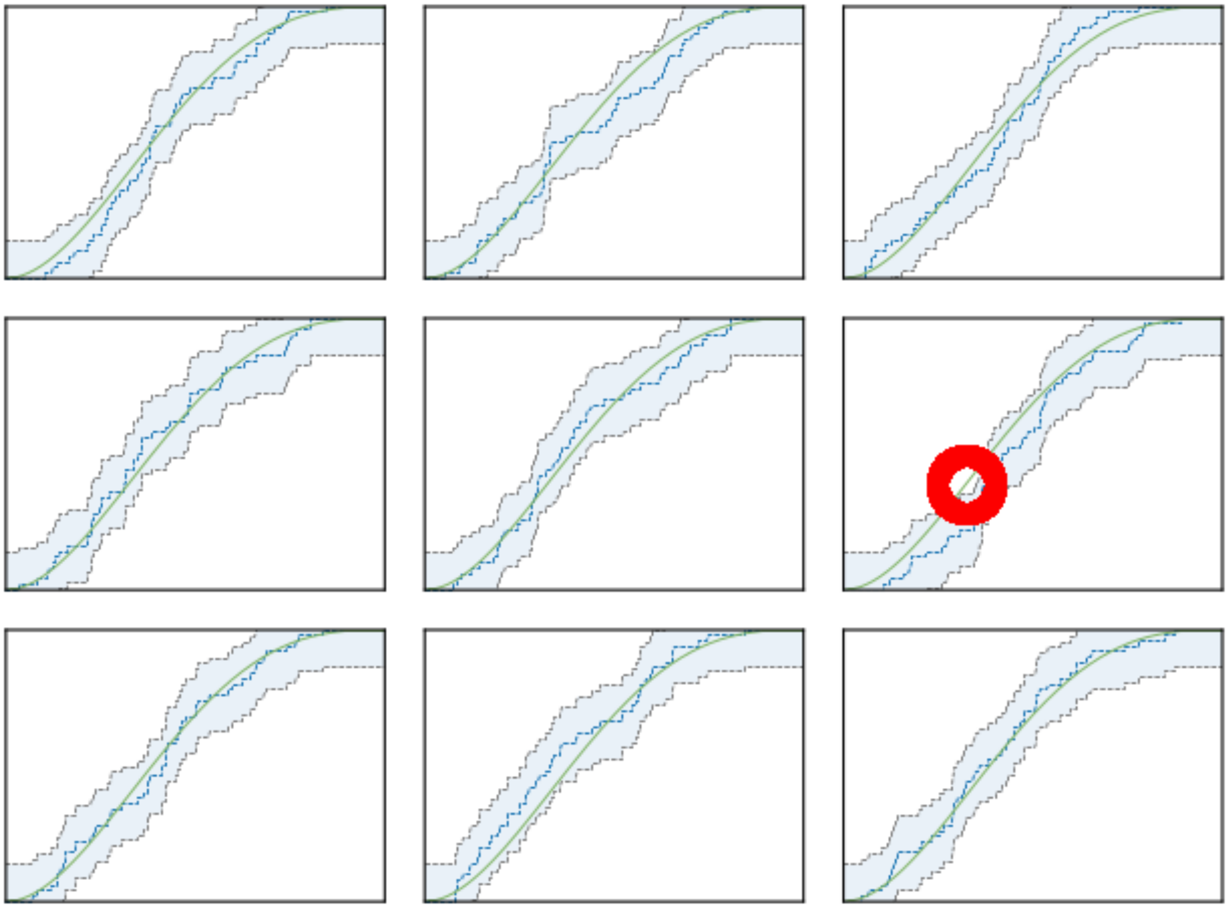

FIG. 5. Empirical probability function of 9 sets of data points, as well as their constraints on the CDF. The data set represented by the plot in middle right violates $\sup_x \left[|\hat{F}_n(\xi) - F(\xi)|\right] < \epsilon$, but the other do not.

test is applied to the Lévy distribution[1], the result is 29.1%, which indicates that the DKW inequality is valid for arbitrary distributions (the difference between the two numbers is statistically not significant). Due to the asymptotic results given by Kolmogorov[4] and Smirnov[5], this rate can never reach 32%, if the number of data points is large enough.

[1] The PDF of Lévy distribution is $f(x;\mu,c) = \sqrt{\frac{c}{2\pi}} \frac{e^{-\frac{c}{2(x-\mu)}}}{(x-\mu)^{3/2}}$. In the test, $c = 1$ and $\mu = 0$. Lévy distribution has a 'fat tail', i.e. the random variable which follows Lévy distribution can have large fluctuations.

[1] A. Dvoretzky, J. Kiefer, J. Wolfowitz, *et al.*, “Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator,” *The Annals of Mathematical Statistics*, vol. 27, no. 3, pp. 642–669, 1956.

[2] L. Wasserman, *All of Statistics.* New York: Springer, 2004.

[3] P. Massart, “The Tight Constant in the Dvoretzky-Kiefer-Wolfowitz Inequality,” *The Annals of Probability*, vol. 18, no. 3, pp. 1269–1283, 1990.

[4] A. Kolmogorov, “Sulla determinazione empirica di una lgge di distribuzione,” *Inst. Ital. Attuari, Giorn.*, vol. 4, pp. 83–91, 1933.

[5] N. V. Smirnov, “Approximate laws of distribution of random variables from empirical data,” *Uspekhi Matematicheskikh Nauk*, no. 10, pp. 179–206, 1944.