**Kolmogorov–Smirnov test**

To measure the goodness-of-fit statistics of random samples against some theoretical probability distribution function, the classical one-dimensional Kolmogorov-Smirnov test is a common option, that is used in [2] [3]. The one-dimensional Kolmogorov-Smirnov test (KS test) is a non-parametric test of the similarity of continuous one-dimensional probability distributions. The one-sample KS test tests whether a sample could come from a specific distribution, answering the question “what is the probability that a given data set would exist if it were drawn randomly from a given distribution?”.

The one-dimensional KStest with continuous data is the most commonly applied variation of the Kolmogorov-Smirnov test because it is distribution-free, meaning it is independent of direction of ordering of the data, and that every individual data point is used. The KS test is mostly used in one-dimensional cases because it is challenging to adapt the KS test to multi-dimensional cases due to the “curse of dimensionality”, because when there are d dimensions involved there are 2d – 1 different combinations of defining a cumulative distribution function [99].

There is another variation of the KS test, which test if two samples are from the same distribution, called the two-sided KS test. The question the two-sided KS test answers is “what is the possibility that two sets of data were drawn from the same probability distribution?”. Basically, checking how similar two sets of data or two distributions are. The two-sided KS test is the goodness-of-fit measure that [2] and [3] used to evaluate the generated images, which is why the two-sided KS test will be used as an evaluation metric in this paper.

Why is the two-sided KS test relevant for galaxy generation? The two-sided KS test can compare and test two distributions, which provide the opportunity to test the training data distribution and the generated data distribution. Using the two-sided KS test in this manner will provide a measure of how similar the original and generated data are, by measuring how likely it is that both distributions come from the same distribution.

The two-sided KS test statistic D is a popular goodness-of-fit measure between two distributions. The statistic D is defined as:

A black text on a white background

Description automatically generated

Figure 1. From paper 98

Where n is the number of independent observations and G and G^are two cumulative distribution functions (cdfs) [98]. D quantifies the distance between the two distributions as the largest absolute difference between the two distributions across all values x, which is what the sup (the supremum of the set of distances) function does.

By the Glivenko-Cantelli theorem, in the one-sample scenario, D converges to 0 as the number of samples n goes to infinity if the sample comes from the given distribution. And in practive, the KS test statistic require many data points.

The KS test output is the statistic D and a p-value measuring the probability of the null hypothesis. The null hypothesis under the two-sample KS test is that the two distributions come from the same distribution, that is that the two distributions that is tested are not significantly different. Usually, the null hypothesis is rejected if the p-value is less than the significance level 0.05. However, for large samples the null hypothesis is rejected if

A square root of a mathematical equation

Description automatically generated

Figure 2. wikipedia KStest

Where alpha is the significance level, often set to 0.05. Here, m and n are the sizes of the two distributions [143]. One question regarding this rejection is how many samples is regarded as large?

The open and free python library scipy, which is commonly used for scientific data processing and methods, has its own two-sample KS test called “scipy.stats.ks\_2samp”. The two-sample KS test from the scipy library will attempt to compute the exact p-value unless both sample sizes are less than 10 000. If both sample sizes are equal to or above 10 000, the asymptotic method will be used. That means the scipy library defines “large sample size” as at least 10 000, which is extremely far from the dataset size, where each variable has num\_samples \* row \* col = 38348 \* 29 \* 10 = 11 120 920, which is over 1112 times as large as scipys definition of large sample size. [142]

This is in fact regarded as a shortcoming of the KS test. The KS test is developed to be sensitive to all differences between the two distributions [142]. The test is sensitive to both location and shape in the distributions. I have conducted some tests regarding sensitivity and deviance threshold in terms of changes in results, these can be found in notebook 9 – KStest check. My analysis shows that changing all variable values of 30 out of 38 348 merger trees to the mean nonzero value, compared to only changing 5 images results to a drop in p-value from 0.9999 (when only changing 5 merger trees) to 0.00867 (when changing 30 merger trees). Even though the KS statistic is 0.000115 and 0.0007 respectively. This drastically drop in p-value with a minor change in the distribution illustrate the sensitivity of the KS test in terms of the p-value.

There are multiple sources that note the weaknesses of the KS test sources, one of those is [4], which states they only use the KS test due to its popularity, but they suggest that there may be better options depending on the circumstances because of the shortcomings of the KS test.