A Proposed High Dimensional Kolmogorov-Smirnov Distance

Alex Hagen*, Jan Strube*, Isabel Haide†, James Kahn†, Shane Jackson*, Connor Hainje* - *Pacific Northwest National Laboratory; †Karlsruhe Institute of Technology

Machine Learning and the Physical Sciences: Workshop at the 34th Conference on Neural Information Processing Systems (NeurIPS) - Vancouver, Canada - December 11, 2020

Abstract

We present a high dimensional test statistic inspired directly by the Kolmogorov–Smirnov (KS) test statistic [1, 2] and Press' extension of the KS test to two dimensions [3]. We call this the ddKS statistic. To preclude the high computational cost associated with working in higher dimensions, we present an implementation using tensor primitives. This allows parallel computation on CPU or GPU. We explore the behavior of the test statistic in comparing two three-dimensional samples, and use a standard statistical method - the permutation method - to explore its significance. We show that, while the Kullback–Leibler divergence is a good choice for general distribution comparison, ddKS has properties that make it more desirable for surrogate model training and validation than the former.

Motivation

- Comparison of distributions, especially with strong statistical guarantees, is important throughout physical sciences and surrogate modeling
- Statistical comparison in higher dimensions than 1 is often overlooked

Test Statistics

- Numerical summaries of data values to set thresholds for hypothesis testing
- Use cases:
  - One-sample tests (data is compared to given probability distribution)
  - Two-sample tests (determine if two data sets are drawn from the same distribution)
- Two-sample tests gain even more importance e.g. through rise of generative models in machine learning
- As number of data samples increases, fast computation of statistical tests is invaluable for most analyses

One Dimensional

- Popular statistical tests (e.g. integrated mean squared error or Earth Mover's Distance) only used in one-dimensional space
- Scaling to higher dimensions often paired with high time cost
- One-dimensional tests cannot identify covariances between variables
- Most test statistics require assumptions/approximations of underlying distribution
- The Kolmogorov-Smirnov test:
  - Also one-dimensional, but non-parametric
  - Defined as maximum difference between two cumulative distribution functions (CDF)

$$D_n = \sup_x |F_{1,n}(x) - F_{2,m}(x)| \tag{1}$$

- KS one of the most general non-parametric tests, using both shape and position of CDFs

Definition

- We take the case of the two sample test of N samples between predicted X_p and true X_t, each of dimension d
- We seek to test the null hypothesis H_0, that the two samples come from the same distribution. Statistical significance p is then compared to action level $\alpha = 0.05$, and if $p \leq \alpha$, H_0 can be rejected.

The ddKS Test Statistic

- ddKS compares cumulative distribution function between two distributions
- Use membership in orthants partitioned at each point in X_p and X_t as surrogate for full CDF
- Region membership calculated in 2^d sized vector - $x_i \in X_p$ and $V_j^p(x_i), V_j^t(x_i)$ is jth component of the membership vector
- ddKS is then defined as

$$D_p = \max_{i,j} |V_j^t(x_i) - V_j^p(x_i)| \tag{2}$$

High Dimensional CDF

- Calculating a CDF in high dimensions is ambiguous
- We use a method inspired by Press [3], calculating the membership of the orthants surrounding each test point.

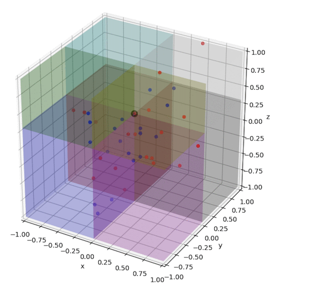

Figure 1: One point at a time is selected (black) to divide the space into octants. The number of points from the two data sets (red, blue) in each octant are counted and compared. Animation at youtu.be/NcUta4lBJPU

Permutation Test

- Allows for the calculation of statistical significance using any distance or divergence measure
- Calculate test statistic D_p for predicted X_p and true X_t
- Randomly mix X_p and X_t to produce two new distributions made of approximately half the samples from both, recalculating D_p for the two new distributions (labelled D_0)
- Repeat M times to produce $D_{0,i}, i \in [1, M]$, with M large enough to approximate D_p under the Null hypothesis
- p-value is the fraction of $D_{0,i}$ greater than D_p

$$p = \frac{N_{D_p < D_{0,i}}}{N} \tag{3}$$

- To account for binomial statistics of $N_{D_p<D_{0,i}}$ use expectation value

$$\langle p \rangle = \frac{1 + N_{D_p<D_{0,i}}}{2 + N} \tag{4}$$

Considered Test Statistics

- Because of the permutation test, we can use any distance or divergence as a test statistic. To show ddKS's utility for physical sciences, we compare it to two other test statistics:
  - **The one dimensional KS test**: We calculate the KS test statistic on each dimension individually, summing the results to create a pseudo-multi-dimensional test statistic. We indicate this as `ks-1d` on figures
  - We also calculate the diagonal distance of each point in each pairwise dimension using the l_2 norm, subsequently summing each dimension's KS test statistic as above. We indicate this as `ks-diag` on figures
  - **The Kullback-Leibler (KL) Divergence**: We calculate the KL divergence between an estimated probability density function of the two distributions. We use a histogram using Scott's [4] rule, sizing the number of bins by $\propto N^{\frac{d}{d+2}}$, to estimate probability density. We indicate this as `kldiv-hist` on figures
  - We also calculate a lower resolution probability density using only 3 bins in each dimension, subsequently calculating the KL Divergence as above. We indicate this as `kldiv-hist25` on figures

Implementation

- Loop based implementation possible: loop through every point in one distribution, counting how many points fall in each surrounding orthant, this implementation was prohibitively slow to calculate during testing ($\mathcal{O}(N^2)$) for all N

Tensor Primitive Based Computation

- By using `pytorch` tensor primitives, implicit parallelism can be used for small N, reducing time complexity to $\mathcal{O}(1)$ and enabling GPU calculation
- Trade time for memory complexity by constructing tensors ($\mathbb{P}, \mathbb{Q}, \mathbb{T}, \mathbb{U}$) from X_p and X_t where $\mathbb{P}[i,j,k] = X_p[i,j]$ for all k
- Build tensors of partition comparison by performing elementwise operations, e.g.

$$\mathbb{G}_P = \mathbb{P} \geq \mathbb{Q}, \tag{5}$$

- Each point is surrounded by 2^d orthants. We construct a membership tensor $\mathbb{M}$ by using a positional encoding function

$$S(x, f) = (-1)^{\lfloor 4fx \rfloor} \tag{6}$$

with $f = 2^{-j-2}$ and $x \in [0, 2^d - 1]$, shown for 3 dimensions in Figure 2.

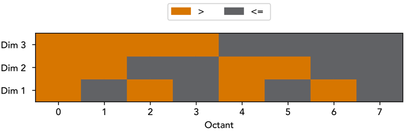

Figure 2: Positional Encoding function S for 3 dimensions

- Then, we fill the membership tensor

$$\mathbb{M}[i,j] = \sum_{k=1}^{N} \prod_{j=1}^{d} \left(\mathbb{G}[l,j,k] \cdot S[i,k] + \left| \frac{S[i,k] - 1}{2} \right| \right) \tag{7}$$

- Calculate ddKS divergence from each distribution to the other

$$D_1 = \max |\mathbb{M}_1 - \mathbb{M}_2| \tag{8}$$

$$D_2 = \max |\mathbb{M}_3 - \mathbb{M}_4| \tag{9}$$

- Finally, average to calculate the final metric

$$D = \frac{D_1 + D_2}{2} \tag{10}$$

Time Complexity

- ddKS and KL Divergence are both $\mathcal{O}(N^2)$ at large N
- `pytorch`'s implicit parallelization makes all metrics $\mathcal{O}(1)$ at small N (except loop based implementations - not shown on figure 3)

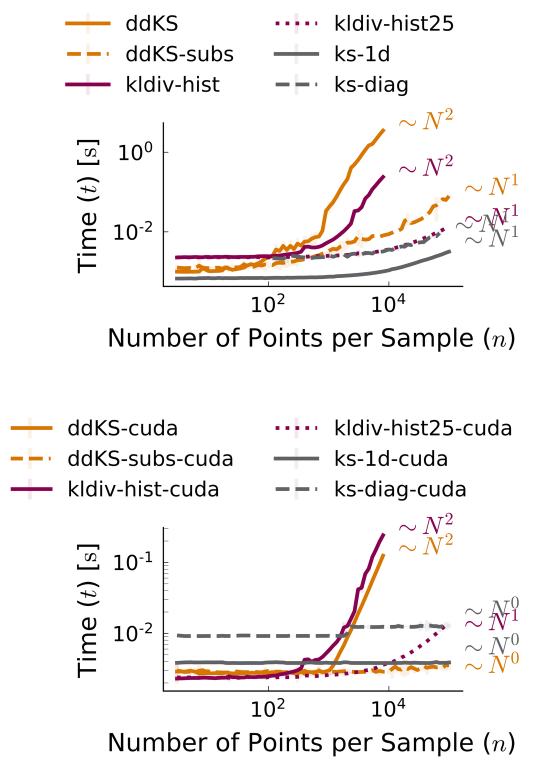

Figure 3: Time to evaluate versus number of points for metrics considered. Time is recorded for permutation tests using 100 permutations, therefore the number to evaluate the test statistic once is $\leq 100\times$ that recorded on this chart. Estimated time complexities as $N \to \infty$ are printed to the right side of each line

Higher Dimensions

- A test set for higher dimensions was constructed. Both samples were filled with 50% background from a uniform distribution from -100 to 100. Then, a hypersphere of dimension d was constructed, the radius 50 and 45 in each respective distribution.
- ddKS is able to reject the null hypothesis for every trial up to dimension 3 at $\alpha \leq 0.05$
- ddKS is able to sometimes reject the null hypothesis in dimension $4 \leq d \leq 8$
- Above dimension $d \geq 9$, ddKS is no longer able to reject the null hypothesis
- KL Divergence is able to reject the null hypothesis for $\alpha \leq 0.05$ sometimes in dimensions $d \in [1, 3, 4, 5]$
- Above dimension $d \geq 5$, KL Divergence requires $\geq 32\text{GB}$ of memory, including requesting 374PiB of memory for dimension $d = 10$

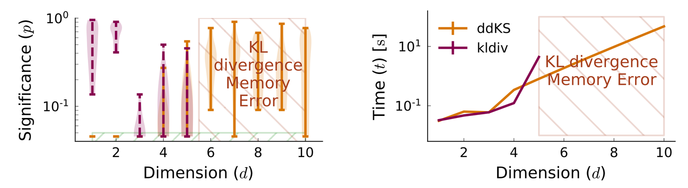

Figure 4: Time complexity and significance with increasing dimensions for ddKS and KL Divergence using Scott's rules for constructing the histogram. Each permutation test was performed with 20 permutations, and each trial was performed 10 times. The green region indicates regions where H_0 could be rejected to $\alpha \leq 0.05$

Accelerated Computations

Subsampling

- High cost incurred by calculating membership vectors of regions centered at every
- Uniformly sampling less than N points from each distribution as centers reduces complexity
  - Faster by a constant factor if a fixed proportion of N points are subsampled
  - $\mathcal{O}(N)$ for constant number of subsampled points
- Tests show similar statistical efficiency to full ddKS, described in the Behavior section
- Expected to have lower statistical efficiency for distributions with differences only in the tails

Voxel Based

- Spatially decompose space into d-dimensional voxels and fill with both sample sets
- Divide space into orthants using each non-empty voxel as an origin
- Approximate D by finding the largest difference in orthant occupation
- $\mathcal{O}(NV)$ or $\mathcal{O}(V^2)$ scaling for N data points and V voxels
- Tests show $\sim \mathcal{O}(2^d N)$ and similar behavior to full ddKS, with decreased performance on "Background Included" data
- Pairwise comparisons within close voxels should be implemented to improve performance on background included data

Radius Based

- Select 2^d origins corresponding to the corners of the entire sample space
- For each origin, sort the data points according to distance from origin ($\mathcal{O}(2^d N \log N)$ operation)
- Approximate D by comparing sample membership to each origin between X_p and X_t for each test point ($\mathcal{O}(2^d N)$ operation)
- Tests show $\sim \mathcal{O}(2^d N)$ time complexity, and comparable behavior to full ddKS
- Current implementation is loop based, and as such is slower than ddKS until $N > 10,000$. Rewrite in C++ would increase speed.

Data

- Two pathological datasets were created: one to illustrate the problem with using one dimensional test statistics, the other to demonstrate an oft-encountered detection physics problem: comparison of signals in varied background

Cherenkov Cone

- A dataset mimicking data collected in Cherenkov cone detectors was constructed
- A charged particle traveling at speed faster than light in quartz enters a quartz medium, and emits photons at φ from the track, uniformly distributed azimuthally around the track
- An ideal detection plane collects the photons location and time of arrival
- Compared with photons emitted isotropically from the top plane of the quartz medium, the single dimensional distributions of detection location and time look identical, however their full distribution is clearly not identical

Background Included

- A cone was generated as above, but time in a very large quartz medium
- Volumetric radiological contamination of the quartz was simulated, and photons emitted isotropically, uniformly distributed within the quartz volume are also detected
- Comparison between two different "cone"s is then difficult because of the multiple scales of the distribution. Two datasets including cones with $\varphi = 15°$ and $\varphi = 20°$ were simulated.
- Comparing the two distributions, the single dimensional distributions of detection location and time look very similar, however their full distribution is not identical

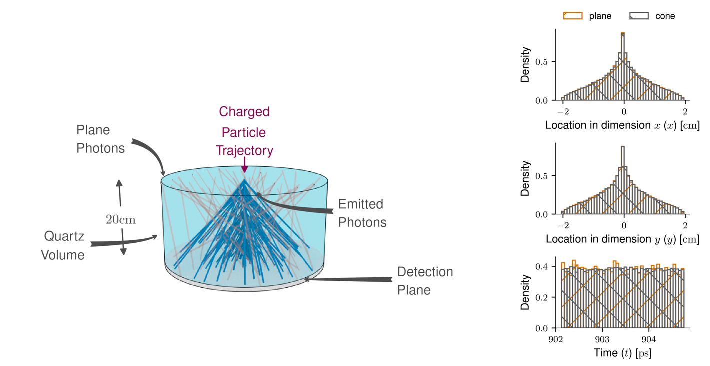

Figure 5: Dataset constructed mimicking photon emission during Cherenkov process. Histograms of detection position and time (silver and copper hatched regions) overlap almost exactly.

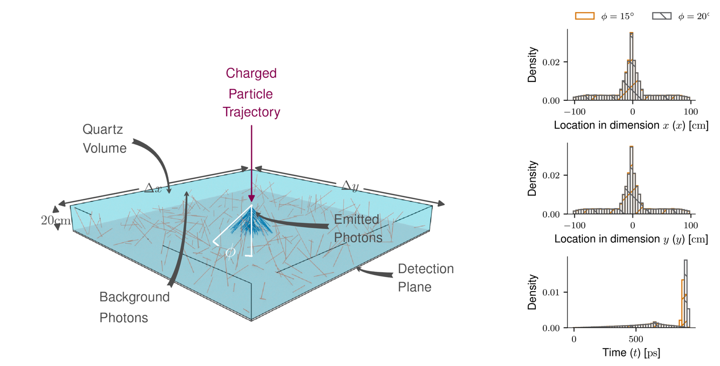

Figure 6: Dataset constructed mimicking photon emission during Cherenkov process with a volumetric background. Histograms of detection position and time (silver and copper hatched regions) overlap closely.

Behavior

Cherenkov Cone

- ddKS, and ddKS using subsampling all reject the null hypothesis to $\alpha \leq 0.05$ by 5 points per sample
- KL Divergence and KL Divergence (~25 bin) reject the null hypothesis by 15-20 points per sample
- One dimensional KS tests cannot reject the null hypothesis to $\alpha \leq 0.05$ until $>$ 20 points per sample

Background Included

- ddKS is able to reject the null hypothesis to $\alpha \leq 0.05$ by ~ 100 points per sample
- KL Divergence rejects the null hypothesis to $\alpha \leq 0.05$ by ~ 125 points per sample
- ddKS using subsampling is able to reject the null hypothesis to $\alpha \leq 0.05$ by between 125 and 1000 points per sample
- KL Divergence (~25 bin) is never able to reject the null hypothesis to $\alpha \leq 0.05$

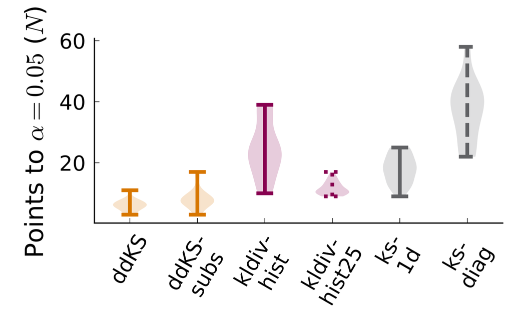

Figure 7: P-Value versus number of points for KL, 1d KS, ddKS tests on the comparison between a Cherenkov cone and a volume source. Each permutation test was performed using 100 permutations, and trials were repeated 25 times.

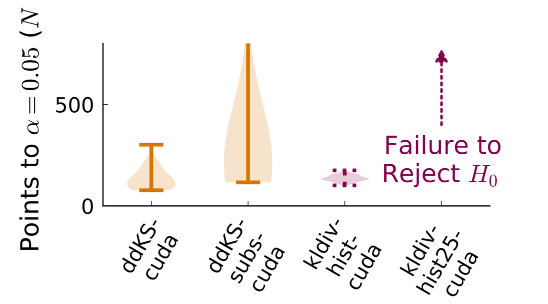

Figure 8: P-Value versus number of points for KL, 1d KS, ddKS tests on the comparison between two Cherenkov cones with θ of 15^{deg} and 20^{deg} in a wide background of volumetric photon emissions. Each permutation test was performed using 100 permutations, and trials were repeated 25 times.

Conclusions

- In general, we find ddKS to be a useful test statistic for high dimensional data, out-performing one dimensional metrics and KL divergence on the scientific data sets we explored

Applications

- ddKS is a metric, which suggests its use as a loss function for high dimensional data - in particular in scientific applications
- Surrogate modeling (replacing computational expensive simulators of scientific data with ML applications) is growing in popularity. ddKS is useful as uncertainty quantification or loss function for these surrogate models.
- ddKS could place statistical significance on predictions from other ML applications with high dimensional latent spaces

Future Work

- Investigation of theoretical ways to determine significance would reduce computational cost by avoiding the permutation test
- Further development of accelerated methods (including acceleration of dimension related time complexity) will make ddKS more usable
- High dimensional CDF approximations in other measures (such as Earth Mover's Distance) could be used as a loss function for generative modeling

References

[1] Andrey Kolmogorov. "Sulla determinazione empirica di una lgge di distribuzione". In: *Inst. Ital. Attuari, Giorn.* 4 (1933), pp. 83–91.

[2] N. Smirnov. "Table for Estimating the Goodness of Fit of Empirical Distributions". In: *Ann. Math. Statist.* 19.2 (June 1948), pp. 279–281. DOI: `10.1214/aoms/1177730256`. URL: `https://doi.org/10.1214/aoms/1177730256`.

[3] William H Press and Saul A Teukolsky. "Kolmogorov-Smirnov Test for Two-Dimensional Data". In: *Citation: Computers in Physics* 2 (1988), p. 74. DOI: `10.1063/1.4822753`. URL: `https://doi.org/10.1063/1.4822753`.

[4] David W. Scott and Stephan R. Sain. "Multi-dimensional Density Estimation". In: *Handbook of Statistics vol 23 Data Mining and Computational Statistics* August 2004 (2004), pp. 1–39. ISSN: 01697161. DOI: `10.1016/S0169-7161(04)24009-3`.

Acknowledgements

This work was supported by the U.S. Department of Energy under contract DE-AC06-76RLO1830 at PNNL with collaboration from the Karlsruhe Institute of Technology (KIT). The motivation and datasets investigated were inspired by the needs of the Belle II experiment.

For additional information, contact:
Alexander Hagen | (509) 375-2403 | alexander.hagen@pnnl.gov