Accelerated Computation of a High-Dimensional Kolmogorov-Smirnov Distance

Isabel Haide[†], Connor Hainje[*], Alex Hagen[*], James Kahn[†‡], Shane Jackson[*], Jan Strube[*] - [†]*Karlsruhe Institute of Technology (KIT);* [*]*Pacific Northwest National Laboratory;* [‡] *Helmholtz AI, KIT*

20th International Workshop on Advanced Computing and Analysis Techniques in Physics Research - Daejeon, South Korea - December 01, 2021

arXiv: 2106.13706
GitHub: pnnl/DDKS

Abstract

Surrogate modeling and data-model convergence are important in any field utilizing probabilistic modeling, including High Energy Physics and Nuclear Physics. However, demonstrating that the model produces samples from the same underlying distribution as the true source can be problematic if the data is many-dimensional. The 1-D and multi-dimensional Kolmogorov-Smirnov test (ddKS) is a statistically powerful nonparametric test which can be implemented as a one- or two-sample test. We have developed three algorithms, one exact and two approximate, for the multi-dimensional Kolmogorov-Smirnov test proposed by Fasano. We apply ddKS to the comparison of photon distributions in the Belle II time-of-propagation detector using the collaboration's Geant4 simulation and our own neural network surrogate model. Additionally, we have derived an analytic form for the statistical significance of ddKS. Our approximations reduce the input time complexity from quadratic to log-linear (vdKS) and reduce the dimensional time complexity from exponential to linear (rdKS). The approximation methods maintain the statistical power of the exact method requiring tens of data points to indicate differences between most sampled distributions.

Motivation

- Comparison of distributions, especially with strong statistical guarantees, is important throughout physical sciences and surrogate modeling.
- Statistical comparison in high dimensions 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 is 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:
  - Is also one-dimensional, but non-parametric.
  - Defined as maximum difference between two cumulative distribution functions (CDF):

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

- KS is 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

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

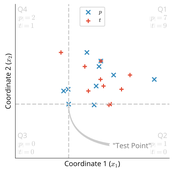


Figure 1: ddKS compares points to a test point along basis vectors of the space, assigning membership to quadrants based on their relationship to the test point.

Permutation Test

- Allows 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 $D_{0,i}$).
- 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 three other test statistics:
  - **One dimensional Kolmogorov-Smirnov test (OnedKS):** We compare our ddKS against one dimensional test statistics by formulating a combined one dimensional KS test in all dimensions. To do so, we take the maximum of the KS statistic in any dimension.
  - **Hotelling's T2 test (Hotelling-T2):** We compare ddKS against a mean-only high dimensional test first published by Hotelling [1].
  - **Kullback-Leibler Divergence (KLDiv):** We compare ddKS to a modern distribution distance, the Kullback-Leibler Divergence [2]. To calculate KLDiv, an estimate of the underlying probability density of each sample is required. We perform this estimate by taking the d-dimensional histogram with constant bin size and bin density defined by Scott's suggestions in [3].

Implementation

- Naive, loop-based implementation: loop through every point in one distribution, count how many fall in each surrounding orthant. Prohibitively slow for all N points ($\mathcal{O}\left(N^2\right)$) .

Tensor Primitive Based Computation

For **small** N:

- Using `PyTorch` tensor primitives: implicit parallelism, reduces time complexity to $\mathcal{O}(1)$, enables GPU calculation.

For **large** N:

- Trade time for memory complexity.

Algorithm:

1. Construct 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.
2. Build tensors of partition comparison by performing elementwise operations, e.g.

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

3. Each point surrounded by 2^d orthants. Construct membership tensor $\mathbb{M}$ using the positional encoding

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

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

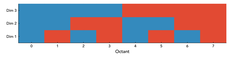


Figure 2: Positional Encoding function S for 3 dimensions

4. Fill the membership tensor

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

5. Calculate ddKS divergence from each distribution to the other.

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

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

6. Finally, average to calculate the final metric

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

Accelerated Computations

Voxel Based

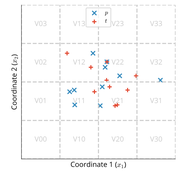


Figure 3: vdKS initially separates the hyperspace into voxels, computing membership within each. For voxels with high membership, ddKS can be performed on that voxel's membership for higher fidelity.

- Voxel based pairwise approximation ddKS (vdKS) divides the space into hypervoxels and counts the membership for each class in each voxel.
- Approximates ddKS in $\mathcal{O}\left(2^d N k\right)$ (where k is the number of voxels).

Radius Based

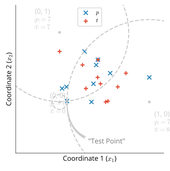


Figure 4: rdKS uses chosen "corners" for comparison, assigning membership for a point to a given quadrant if the euclidean distance between that point and the "corner" is smaller than between the corner and a test point.

- In rdKS $d+1$ corner points are identified and, for each point, the sample points are sorted by their distance from each corner.
- rdKS approximates ddKS in $\mathcal{O}\left((d+1)\, N \log N\right)$, thus providing a good method for larger dimensions.

Time Complexity

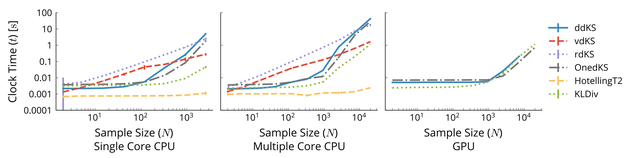


Figure 5: Time to compute a single test statistic for ddKS accelerated methods and selected other methods from the literature. A lower time for computation is better, however there is a tradeoff between time complexity as visualized here, and statistical power.

- Tensor primitive based computation uses implicit parallelization in `PyTorch`'s tensor primitive operations; reduces computational complexity to $\mathcal{O}(1)$ until the memory or core count is exhausted.

Analytical Significance Calculation

- We derived a formula for the significance for the two-sample ddKS test by assuming that each element of $\mathbb{M}[i,j]$ can be thought of as the result of N binomial trials. $\lambda_{i,k}$ is the rate corresponding to each entry.
- The analytical significance is then given by

$$S\left(D, N_P, N_T, \vec{\lambda}\right) = 1.0 - \prod_{i=0}^{2^d} \prod_{k=0}^{N} p_{<\delta}\left(D, N_P, N_T, \lambda_{i,k}\right). \tag{11}$$

- The analytical significance closely matches the significance given by the permutation test.

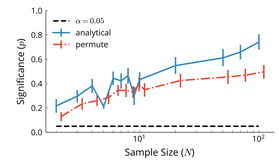
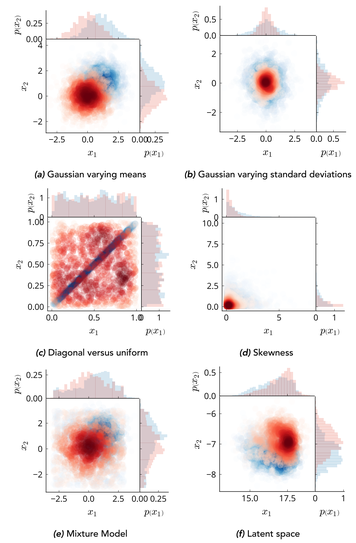


Figure 6: Statistical significance with which we can reject H_0 with increasing sample size N on a given distribution in 3 dimensions, repeated 100 times.

Behavior

Datasets



(a) Gaussian varying means
(b) Gaussian varying standard deviations
(c) Diagonal versus uniform
(d) Skewness
(e) Mixture Model
(f) Latent space

Figure 7: Illustration of two-dimensional version of all datasets tested. Red and blue show the different distribution members.

Results

- We test ddKS by calculating the minimum sample size to correctly reject H_0 given the default parameters of each dataset.
- We compare this to the **KL divergence**, the **Hotelling's T2 test** and the **one-dimensional KS test** (fig. 8).
- We also compare the three ddKS accelerated computation methods, ddKS, vdKS and rdKS (fig. 9).

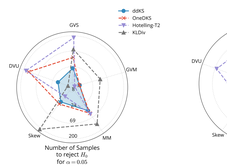


Figure 8: ddKS vs. other test statistics. Samples required to reject H_0 (shown radially, log scale) for each dataset around circumference (closer to center is better). ddKS reliably shows low sample sizes; other methods fail on one or more datasets.

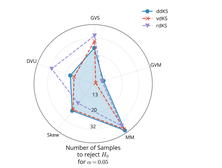


Figure 9: ddKS vs. vdKS vs. rdKS Samples required to reject H_0 (shown radially, log scale) for each dataset around circumference (closer to center is better). All xdKS methods show similar performance: able to discriminate all datasets investigated with small sample sizes.

Belle II application

- The Belle II TOP detector consists of 16 quartz bars centered around the beam pipe to measure Cherenkov photons created by crossing particles for particle identification. [4]
- Current TOP Monte Carlo simulation (Geant4) is **largest time contribution** to overall Belle II detector simulation. Any faster surrogate models must be validated against this.
- Both photon and particle identification uncertainty can be analysed with ddKS.
- Behavior of photon detection values is difficult to learn by a fast simulation. **ddKS can be used as an evaluation metric.**
- A fast analytical calculation is being developed, taking $(78.9 \pm 6.9)\,\frac{\mu\mathrm{s}}{\mathrm{photon}}$, $2000\times$ faster than Monte Carlo ($(142.1 \pm 8.1)\,\frac{\mu\mathrm{s}}{\mathrm{photon}}$).
- ddKS shows significant differences $D > 0.7$ between faster surrogate model and Monte Carlo (repeated simulations give $D < 0.05$).

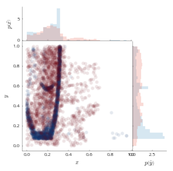


Figure 10: Surrogate versus monte carlo simulated photons for a predefined $\vec{x}$ and $\vec{y}$

- ddKS can also be used to profile distribution changes as a function of input parameters, locating interesting regions in phase space.
- Figure 11 shows large distribution differences as ϑ and ψ change, indicating the importance of initial angle on final detection pixel and time.

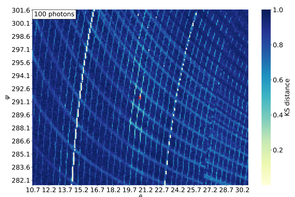


Figure 11: ddKS distances of photon detection values. Photons are generated starting at the same origin, with momentum in the angle interval $\vartheta = 20.7^\circ \pm 10^\circ$ and $\psi = 291.6^\circ \pm 10^\circ$. Detection values of every point in the ϑ-ψ grid are compared with ddKS to those of the origin marked by the red x (100 photons per point). [5]

Conclusions and Applications

- In general, we have shown ddKS to be a useful test statistic for high dimensional data, **out-performing** one dimensional metrics and KL divergence on the scientific data sets explored.
- **ddKS is a metric**, which suggests its use as a loss function for high dimensional data problems - in particular in scientific applications.
- Surrogate modeling (replacing computational expensive simulators of scientific data with ML solutions) is growing in popularity. ddKS is useful as uncertainty quantification or a loss function for these surrogate models.
- ddKS could place statistical significance on predictions from other ML applications with high dimensional latent spaces.

References

[1] Harold Hotelling. "The Generalization of Student's Ratio". In: *The Annals of Mathematical Statistics* 2.3 (1931), pp. 360–378. ISSN: 0003-4851. DOI: 10.1214/aoms/1177732979. URL: http://projecteuclid.org/euclid.aoms/1177732979.

[2] Solomon Kullback and Richard A Leibler. "On information and sufficiency". In: *The annals of mathematical statistics* 22.1 (1951), pp. 79–86.

[3] D. W. Scott and S. Sain. "Multi-dimensional Density Estimation". In: 2004.

[4] T Abe et al. *Belle II Technical Design Report*. Tech. rep. Comments: Edited by: Z. Doležal and S. Uno. 2010. arXiv: 1011.0352. URL: http://cds.cern.ch/record/1304162.

[5] Isabel Haide. "Fast Simulation and Validation of the Time of Propagation Detector at the Belle II Experiment". MA thesis. Karlsruhe Institute of Technology (KIT), 2021.

Acknowledgements

This work was supported by the U.S. Department of Energy under contract DE-AC06-76RLO1830 at PNNL in collaboration with the Karlsruhe Institute of Technology (KIT). The motivation and datasets investigated were inspired by the needs of the Belle II experiment. James Kahn was funded by Helmholtz Association's Initiative and Networking Fund through Helmholtz AI. Isabel Haide's work was supported by the Federal Ministry of Education and Research of Germany (BMBF).

For additional information, contact:
Isabel Haide | isabel.haide@kit.edu

12/01/21 | PNNL-SA-168608