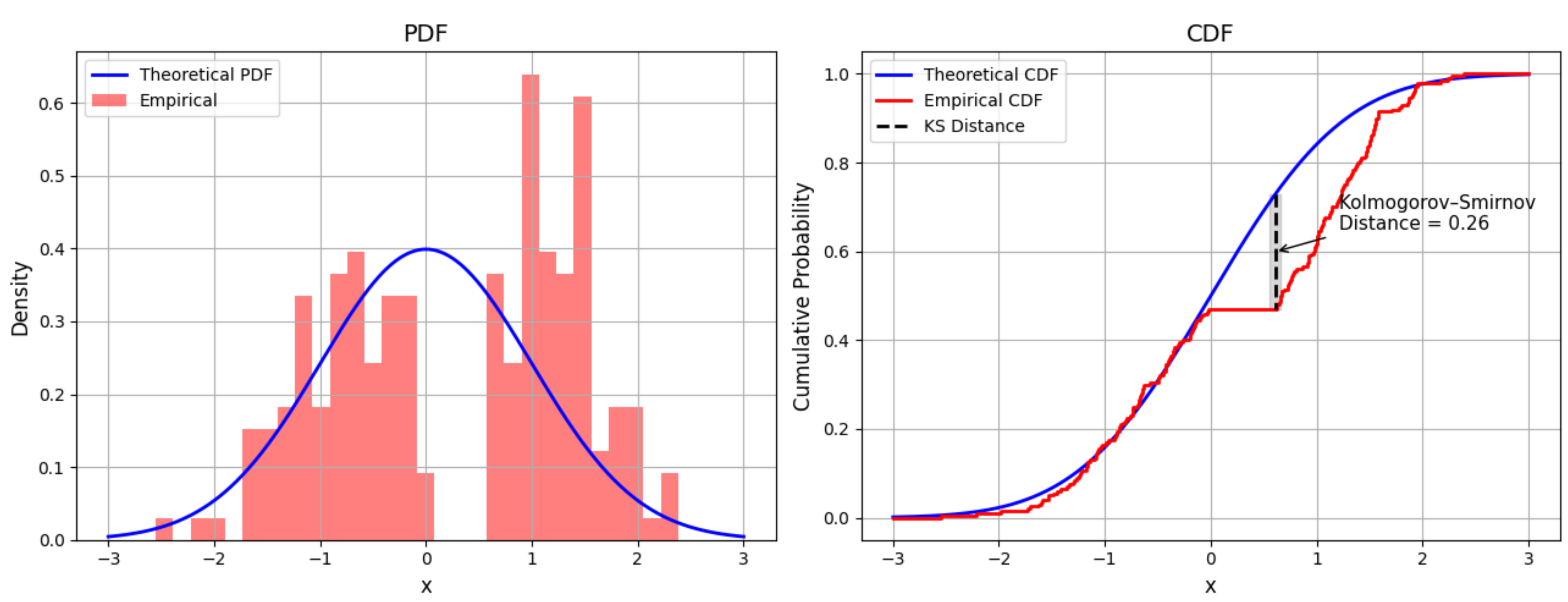
A Smarter Way to Compare Multi-Dimensional Data Distributions

Foad Namjoo, Peter Jacobs, Jeff M. Phillips

Supported by NSF IIS-1816149

KOLMOGOROV—SMIRNOV DISTANCE

KS Distance is the maximum vertical distance between and



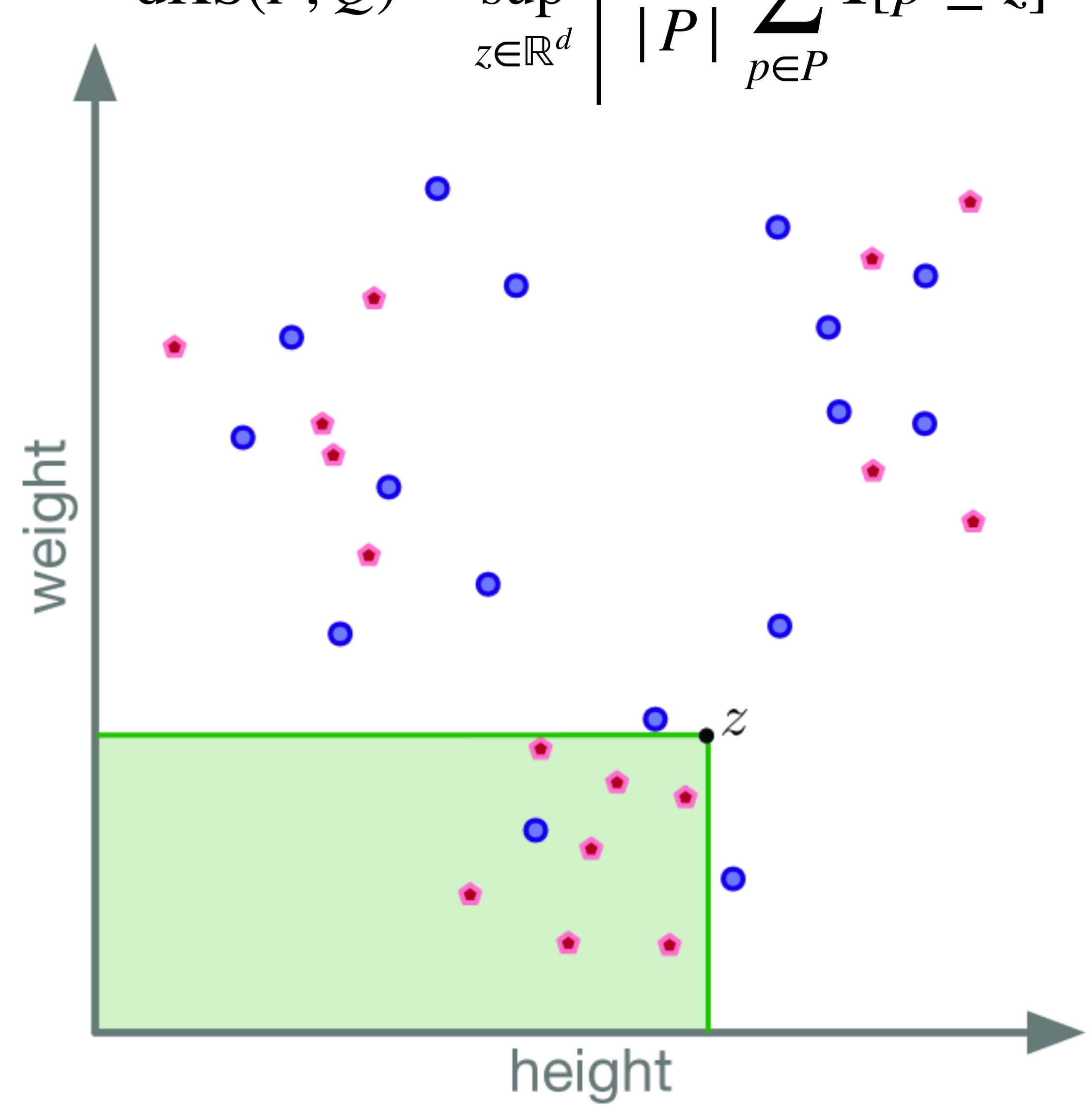
$$D_{KS}(F,G) = \sup_x |F(x) - G(x)|$$

1. The KS test works great in 1D—**but** breaks down in higher D.
2. Existing fixes (e.g., mdKS) are unstable and lack theoretical support.

OUR SOLUTION: dKS

d-dimensional Kolmogorov–Smirnov (dKS) distance formula:

$$\mathrm{dKS}(P,Q) = \sup_{z\in\mathbb{R}^d} \left| \frac{1}{|P|}\sum_{p\in P} \mathbf{1}[p \preceq z] - \frac{1}{|Q|}\sum_{q\in Q} \mathbf{1}[q \preceq z] \right|$$



ADVANTAGE OF dKS

- Comparing multi-dimensional distributions in stats, AI, ML, etc.
- **Invariant to unit** scaling (e.g., cm vs inches)
- Sample Complexity:

  Stable to sampling error
- An Integral Probability Metric (IPM)
  Satisfies **triangle inequality**
- **Efficient** to compute:
  - d = 2: as fast as in 1D
  - d = 3, 4: near-linear in sample size

ALGORITHM ()

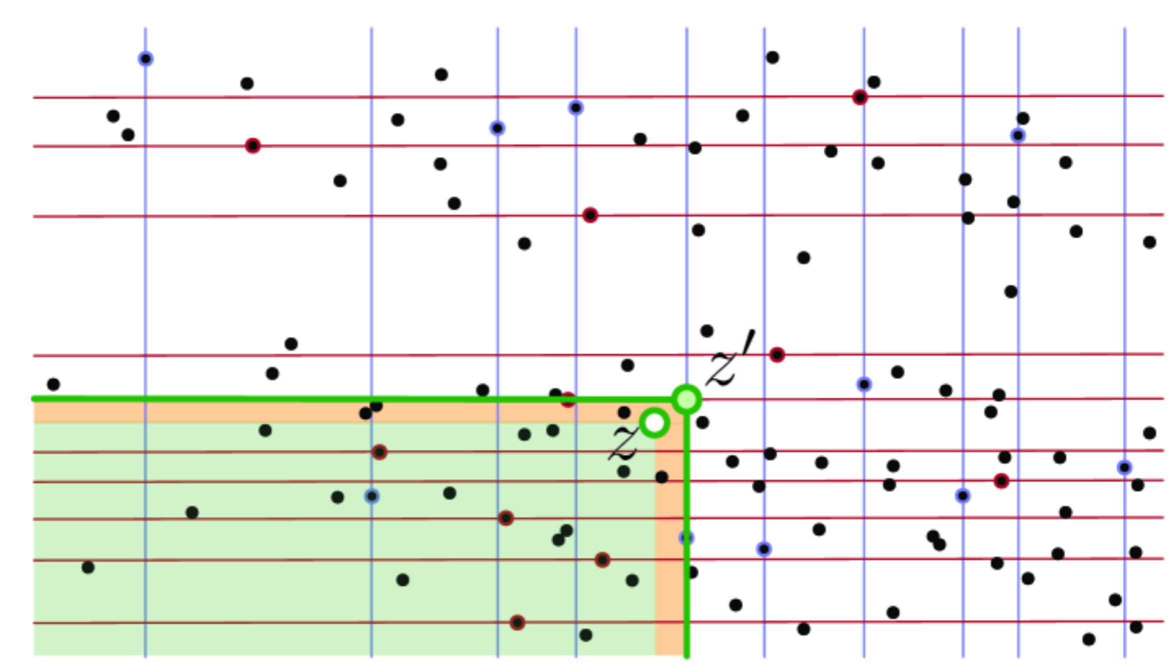
1- Exact Algorithm

Draw i.i.d. points:

2- Grid construction

Sort the union
Select evenly spaced points in each dimension

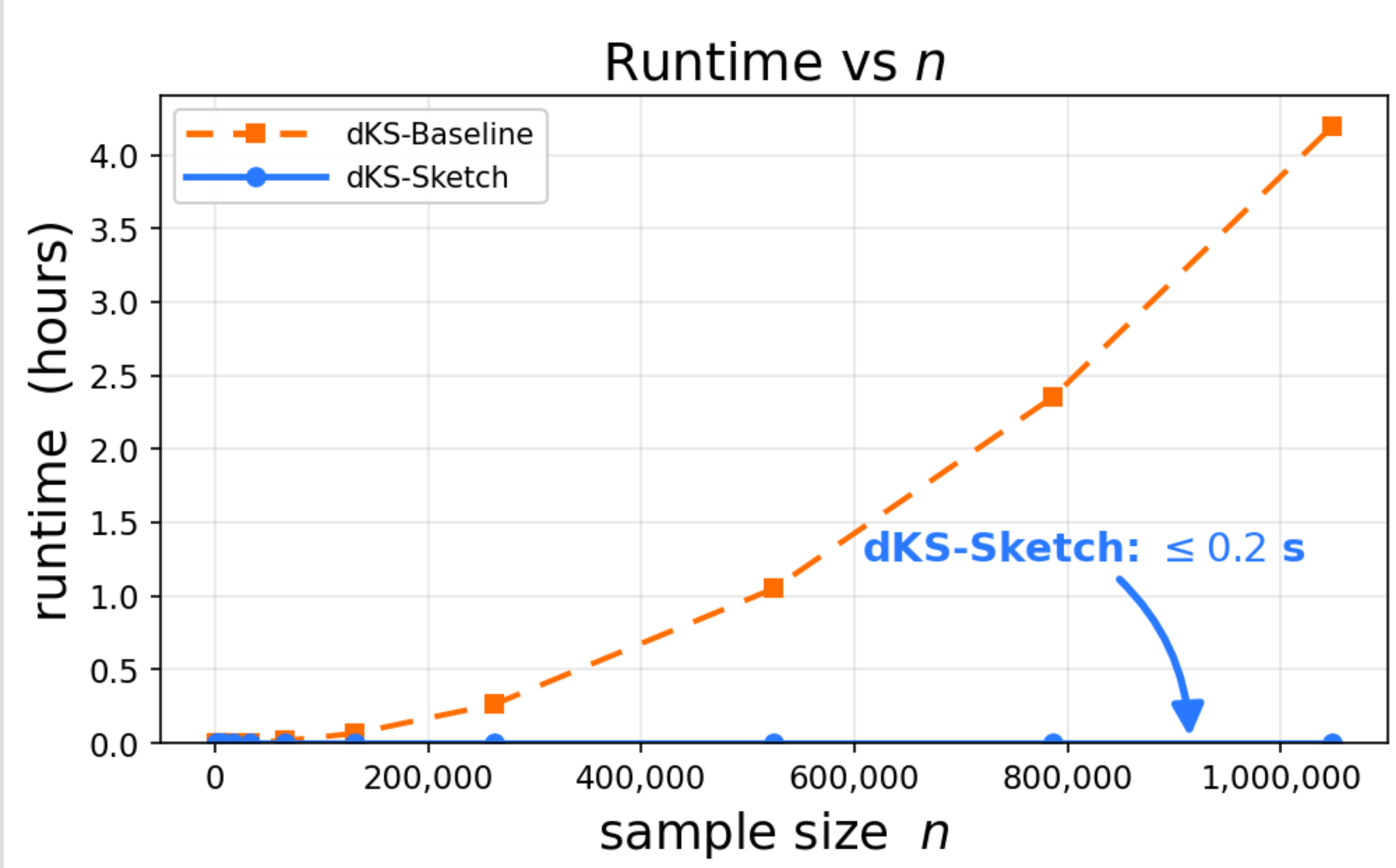
3- Round points to grid cells



4- Compute dKS

Evaluate the maximum empirical CDF difference over the induced grid rectangles

Guarantee: This induces error , on the same order as sampling noise. But is much faster than checking all rectangles.



STATISTICAL TEST

dKS enables accurate two-sample tests with faster runtimes and stronger guarantees than mdKS or Peacock.

Citation	Test Statistic	Stabile	Precise Test	(n,δ) Test Runtime	
[25, 35]	Peacock	Yes	No	$O(n^d)$	
[10, 19]	mdKS	No	No	$O(n \log n)$	
this paper	dKS	Yes	Yes	$O(n \log n \log(1/\delta))$	$d = 2$
this paper	dKS	Yes	Yes	$O(n \log^3 n \log(1/\delta))$	$d = 3$
this paper	dKS	Yes	Yes	$O(n \log^{21+\zeta} n \log(1/\delta))$	$d = 4$