DS 298: Random Variates in Computation

Naman Pesricha
namanp@iisc.ac.in
SR - 24115

Work Assignment - 1

Sample from a truncated normal distribution $N\left(\frac{1}{2}, \frac{1}{36}\right)$, the arcsine distribution (a beta distribution with $\alpha = \beta = \frac{1}{2}$), and the uniform distribution with the limits of the random variables as $(0, 1)$. These empirical distributions (histograms) can be generated using inbuilt functions of Python or Matlab. Write your description of observations and conclusions including the plots and tables in it.

1. Plot the variation of Kolmogorov-Smirnov (K-S) statistic with the number of samples n for the above 3 distributions, by comparing the corresponding empirical and reference distributions as n varies from 10^2 to 10^5 samples. Note that when the samples are drawn from a distribution, the empirical CDF can be directly generated without an associated PDF. Average the K-S statistic over multiple trials for a smooth plot.

2. Generate a K-S statistic comparison table (again averaging over multiple trials) in the form of a 3×3 symmetric confusion matrix for each of the sample sizes 10^2, 10^3 and 10^4, where now each empirical distribution is compared with all the three reference distributions given.

3. Repeat the construction of the above confusion matrix using the Bhattacharya and the Hellinger distances using 10^4 samples. In these cases, the required distance integral can be replaced by a point-wise summation over the histogram of the empirical density, and its normalization using the point-wise sum of the reference density (which is 1 in the case of an exact integration).

Note: Use logarithmic scale in an axis/plot wherever appropriate. Submit the descriptive response and the code as separate files, all zipped into a single folder identified by your name in full, on the MS Teams channel for the class.

---

1 Introduction

This assignment investigates statistical distance measures applied to empirical distributions sampled from three distinct reference distributions on the interval $(0, 1)$:

1. **Truncated normal distribution**: $\mathcal{N}(\mu = 0.5, \sigma^2 = 1/36)$ truncated to $(0, 1)$.
2. **Arcsine distribution**: A special case of the Beta distribution with parameters $\alpha = \beta = 1/2$, exhibiting a U-shaped probability density function (PDF).
3. **Uniform distribution**: $\mathcal{U}(0, 1)$ with constant density over the interval.

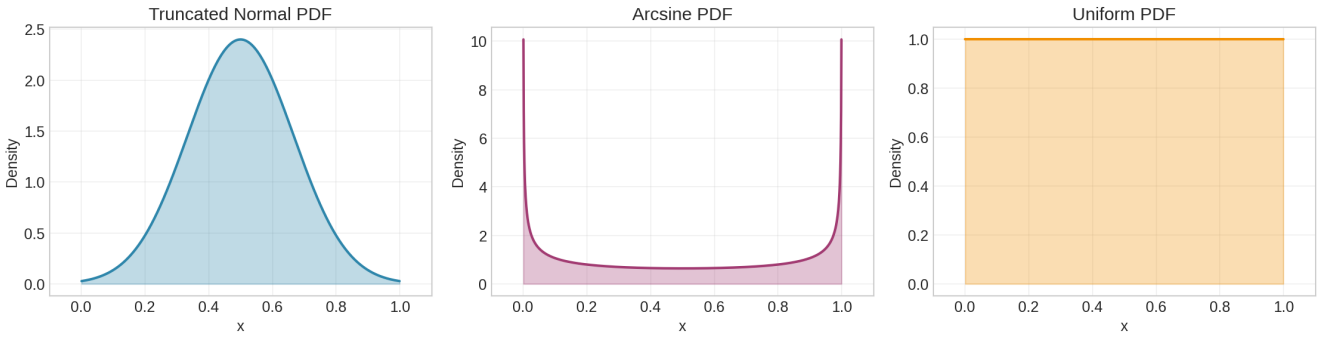

Figure 1: Probability density functions of the three reference distributions on $(0, 1)$. Left: Truncated normal $\mathcal{N}(0.5, 1/36)$ showing unimodal concentration at the center. Center: Arcsine distribution (Beta$(\frac{1}{2}, \frac{1}{2})$) concentrates probability mass near the boundaries. Right: Uniform distribution $\mathcal{U}(0, 1)$ with constant density.

The three distance measures under investigation are:

- **Kolmogorov-Smirnov (K-S) statistic**: Measures the maximum absolute deviation between empirical and reference cumulative distribution functions (CDFs):

$$D_n = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \tag{1}$$

where F_n is the empirical CDF and F is the reference CDF.

- **Bhattacharyya distance**:

$$D_B(p, q) = -\ln\left(\int \sqrt{p(x)q(x)}\, dx\right) \tag{2}$$

- **Hellinger distance**: A bounded metric derived from the Bhattacharyya coefficient $BC(p, q) = \int \sqrt{p(x)q(x)}\, dx$:

$$H(p, q) = \sqrt{1 - BC(p, q)} \tag{3}$$

with $H(p, q) \in [0, 1]$.

2 Methodology

2.1 Sampling and Empirical Distribution Estimation

For each distribution, we generated independent samples using SciPy's statistical distribution modules:

- Truncated normal: `scipy.stats.truncnorm` with parameters $a = (0 - \mu)/\sigma$, $b = (1 - \mu)/\sigma$
- Arcsine: `scipy.stats.beta(0.5, 0.5)`
- Uniform: `scipy.stats.uniform(0, 1)`

All experiments used a fixed random seed (0) for reproducibility.

2.2 Distance Computation

- **K-S statistic**: Computed using `scipy.stats.kstest`, which implements the exact Kolmogorov distribution for finite samples.
- **Bhattacharyya & Hellinger distances**: We first conducted a convergence analysis to determine optimal histogram resolution for PDF estimation. This analysis revealed that 69 bins minimize error for Bhattacharyya distance estimation, while 183 bins are optimal for Hellinger distance (see Figure 6). Using these optimal bin counts, we computed discrete approximations:

$$BC(p, q) \approx \sum_{i=1}^{k} \sqrt{p_i q_i} \tag{4}$$

$$D_B(p, q) = -\ln(BC(p, q)) \tag{5}$$

$$H(p, q) = \sqrt{1 - BC(p, q)} \tag{6}$$

where p_i, q_i are normalized histogram counts forming discrete probability distributions, and k is the optimal number of bins.

2.3 Experimental Design

- **Problem 1**: Sample sizes $n \in [10^2, 10^5]$ (logarithmically spaced, 30 points). Each point averaged over 100 independent trials. Plotted on log-log scale with theoretical $1/\sqrt{n}$ reference.
- **Problem 2**: K-S confusion matrices for $n = 10^2, 10^3, 10^4$, each entry averaged over 100 trials. Additionally, we computed true K-S distances between reference distributions for comparison.
- **Problem 3**: Bhattacharyya and Hellinger confusion matrices for $n = 10^4$, averaged over 100 trials using optimal bin counts determined through convergence analysis.

3 Results and Analysis

3.1 Empirical Distribution Convergence

Figure 2 illustrates the convergence of empirical histograms to their corresponding analytical probability density functions as the sample size increases from $n = 100$ to $n = 10{,}000$. As n increases, sampling variability decreases and the empirical densities increasingly align with the reference PDFs, particularly in regions of high curvature and near distribution boundaries. Each row corresponds to a different sample size ($n = 100$, $n = 1{,}000$, and $n = 10{,}000$ from top to bottom), while each column represents a different reference distribution (Truncated Normal, Arcsine, and Uniform). Independent axis scaling is used for each subplot to preserve distribution-specific density magnitudes and clearly highlight convergence behavior.

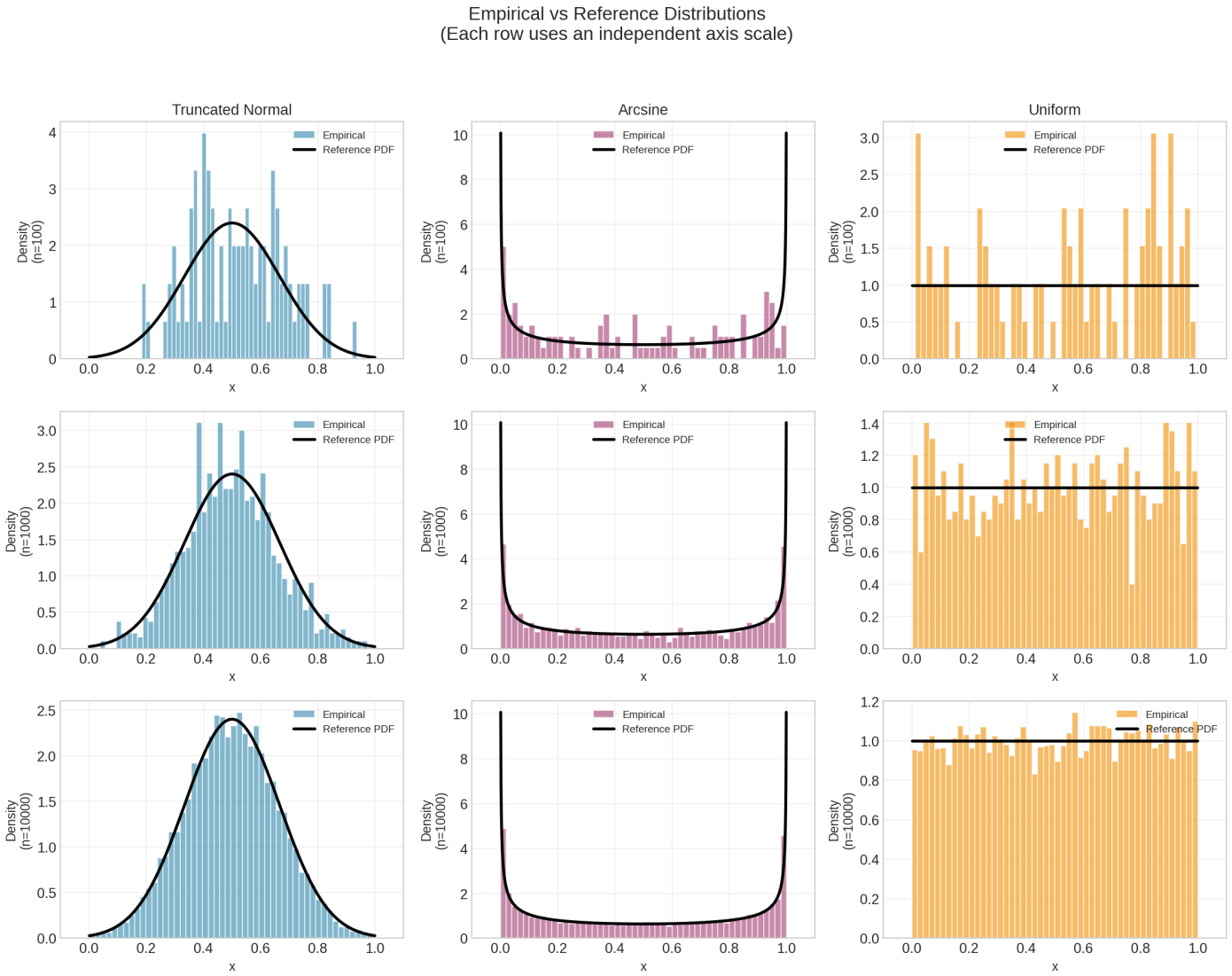

Figure 2: Empirical histograms and analytical PDFs for three distributions at increasing sample sizes. Rows correspond to sample sizes $n = 100$, $n = 1{,}000$, and $n = 10{,}000$ (top to bottom), while columns correspond to the Truncated Normal, Arcsine, and Uniform distributions (left to right). Each subplot uses an independent axis scale.

3.2 Problem 1: K-S Statistic Variation with Sample Size

Figure 3 shows the Kolmogorov-Smirnov statistic as a function of sample size n for all three distributions, with n ranging from 10^2 to 10^5. Each curve represents the mean K-S statistic computed over 100 independent trials when comparing empirical samples to their matching reference distribution. Shaded regions indicate ± 1 standard deviation across trials. The dashed black line shows the theoretical $1/\sqrt{n}$ reference decay.

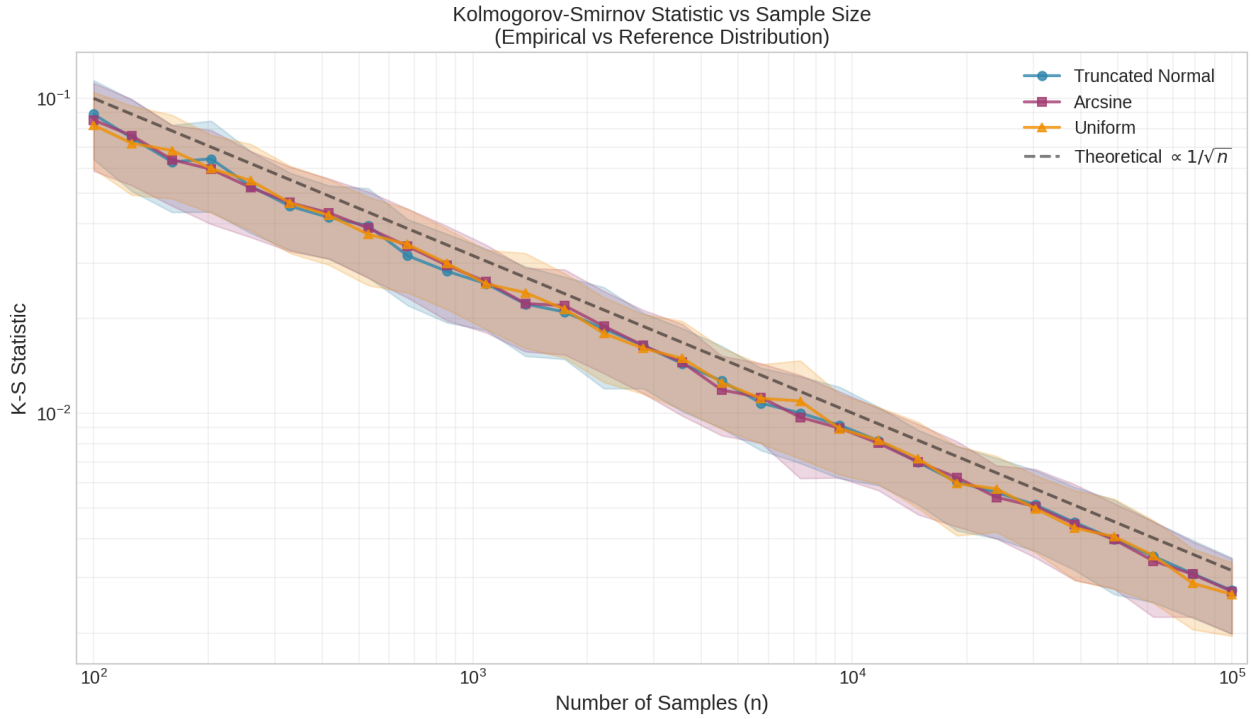

Figure 3: Kolmogorov-Smirnov statistic versus sample size on log-log scale. Curves show mean values over 100 trials for matching empirical and reference distributions. Shaded bands represent ± 1 standard deviation. The dashed black line shows theoretical $O(1/\sqrt{n})$ decay.

Key observations:

- All three distributions exhibit decay proportional to $n^{-1/2}$, confirming the asymptotic behavior predicted by the inequality covered in class. The empirical convergence rate matches the theoretical $1/\sqrt{n}$ scaling exactly.
- Despite differences in PDF geometry (unimodal, U-shaped, flat), all distributions converge at the same asymptotic rate, confirming that K-S convergence depends primarily on sample size rather than distribution shape for these distributions.

3.3 Problem 2: K-S Confusion Matrices

Figure 4 presents the theoretical K-S distances between reference distributions, computed directly from analytical CDFs. This serves as the ground truth for maximum distinguishability.

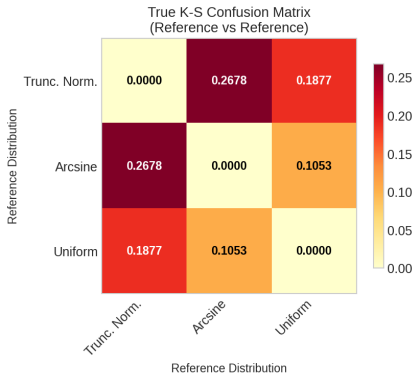

Figure 4: True K-S confusion matrix computed from reference vs reference distributions. This represents the theoretical maximum distinguishability between distributions.

Figure 5 presents 3×3 confusion matrices comparing each empirical distribution against all three reference distributions at sample sizes $n = 10^2$, 10^3, and 10^4. Matrix entries represent mean K-S statistics over 100 trials. Rows correspond to empirical distributions; columns correspond to reference distributions.

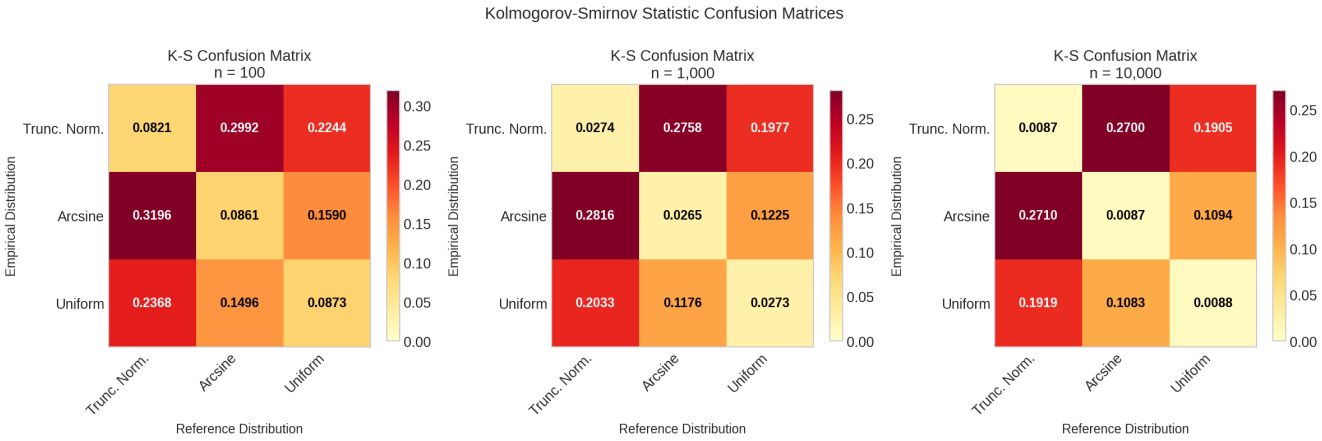

Figure 5: K-S statistic confusion matrices for sample sizes $n = 10^2$, 10^3, and 10^4. Rows: empirical distributions; columns: reference distributions. Darker shades indicate larger distances. Values shown are means over 100 trials.

Observations:

- For all sample sizes, the diagonal entries of the K-S confusion matrices are consistently smaller than the off-diagonal entries, indicating that the K-S statistic correctly identifies matching empirical–reference distribution pairs by minimizing the supremum difference between their cumulative distribution functions.
- As the sample size increases from $n = 10^2$ to $n = 10^4$, the diagonal elements decrease monotonically and approach zero, reflecting the uniform convergence of the empirical CDF $F_n(x)$ to the true reference CDF $F(x)$.
- At n = 10,000, the empirical confusion matrix closely matches the true K-S confusion matrix computed using reference distributions alone, confirming the consistency of the empirical K-S estimator.
- Comparisons involving the arcsine distribution yield the largest K-S values, reflecting its boundary-concentrated CDF, while the truncated normal and uniform distributions exhibit smaller mutual distances due to their more similar cumulative behavior.
- The matrix formed by the empirical K-S statistics is not symmetric, since in general the K–S statistic between an empirical distribution derived from samples of F and a reference distribution G is not equal to the K–S statistic between an empirical distribution derived from samples of G and a reference distribution F, i.e.,

$$D(F_n, G) \neq D(G_n, F),$$

where F_n and G_n denote the empirical CDFs constructed from samples drawn from F and G, respectively.

3.4 Problem 3: Bhattacharyya and Hellinger Distance Confusion Matrices

Before presenting the final distance matrices, Figure 6 shows our convergence analysis for optimal histogram bin selection. This rigorous approach ensures minimal estimation error in our distance computations.

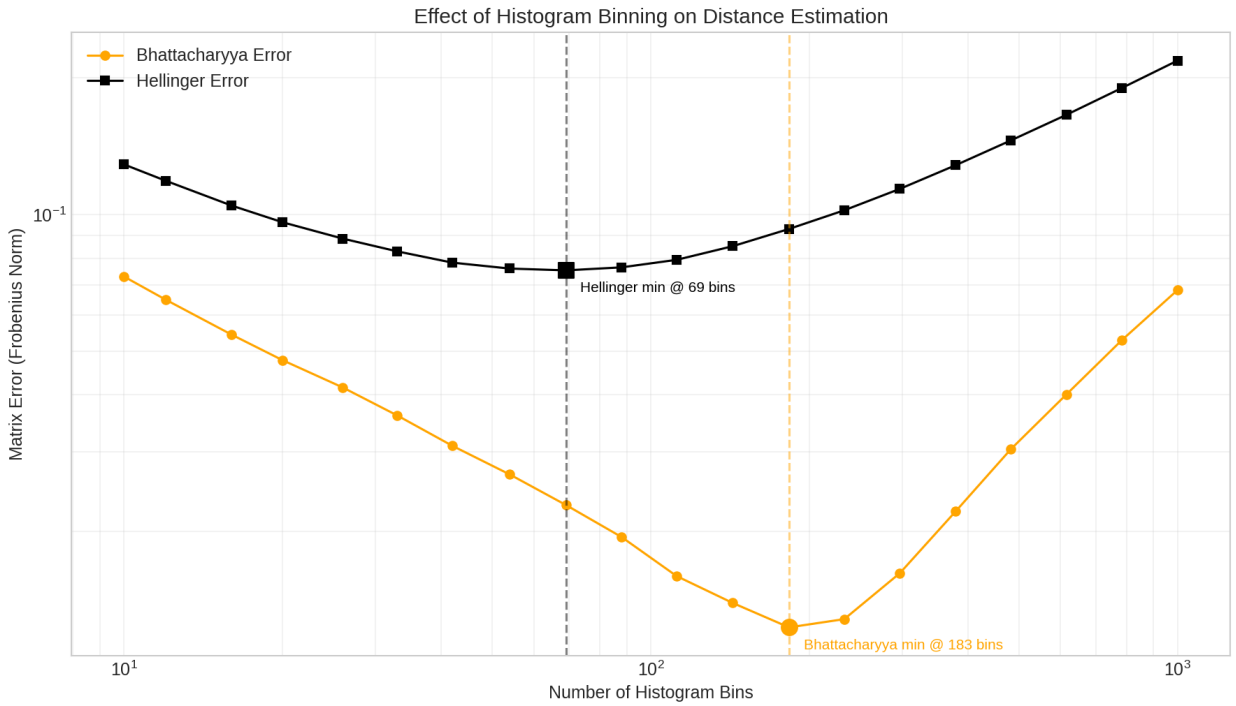

Figure 6: Convergence analysis for histogram bin selection. Matrix error (Frobenius norm) between empirical and true distance matrices as a function of bin count. Optimal bin counts minimize estimation error: 69 bins for Bhattacharyya, 183 bins for Hellinger distances.

Figure 7 shows the theoretical Bhattacharyya and Hellinger distances computed directly from reference PDFs, providing ground truth for maximum distinguishability.

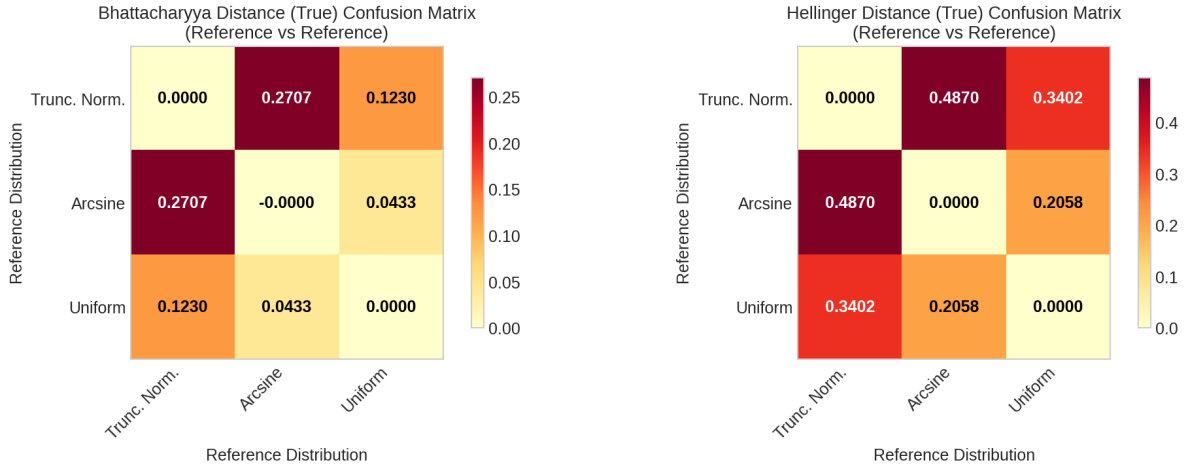

Figure 7: True Bhattacharyya (left) and Hellinger (right) distance confusion matrices computed from reference vs reference distributions.

Figure 8 shows the empirical confusion matrices for Bhattacharyya and Hellinger distances at $n = 10^4$ samples, using the optimal bin counts determined in Figure 6. Both metrics were averaged over 100 trials.

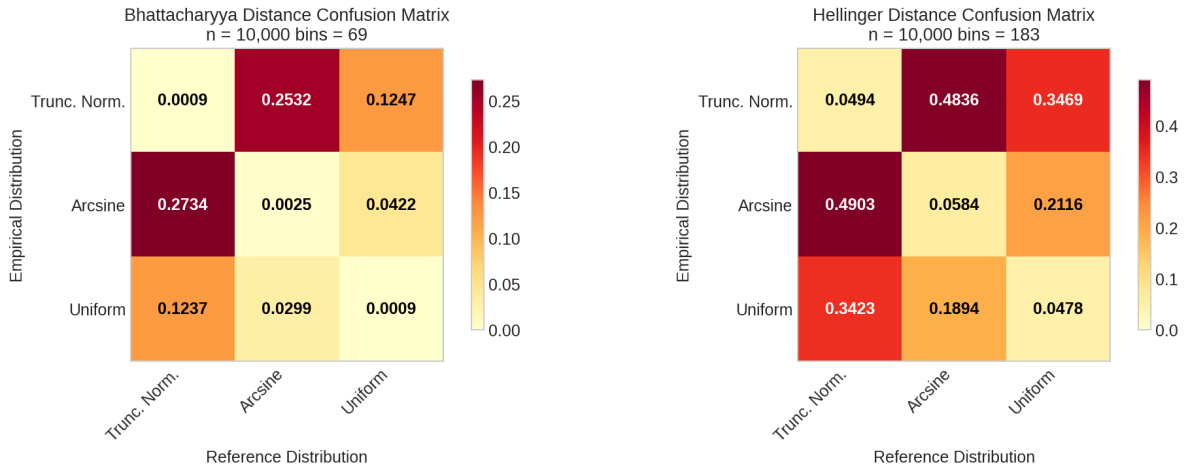

Figure 8: Bhattacharyya (left) and Hellinger (right) distance confusion matrices for $n = 10^4$. Rows: empirical distributions; columns: reference distributions. Values shown are means over 100 trials using optimal bin counts (69 for Bhattacharyya, 183 for Hellinger).

Observations:

- Both Bhattacharyya and Hellinger distance confusion matrices computed using 10^4 samples exhibit strong diagonal dominance, indicating correct identification of matching empirical–reference distribution pairs.
- The empirical Bhattacharyya and Hellinger distance matrices closely match their corresponding true reference matrices, with off-diagonal entries converging to theoretical values, confirming the consistency of histogram-based, point-wise summation as an accurate numerical approximation of the underlying distance integrals.
- Small nonzero diagonal entries in the empirical matrices arise from finite-sample effects and histogram discretization, and vanish in the true distance matrices computed from exact reference densities.
- The Bhattacharyya distance is more sensitive to histogram bin resolution trade-off as compared to the Hellinger distance as evidenced by the sharper error minimum in Figure 6.