# Differentially Private Distribution Estimation Using Functional Approximation

**Ye Tao, Anand D. Sarwate**

**ECE Department, Rutgers University, Piscataway, New Jersey 08854**

## Motivation

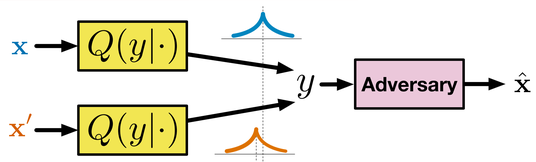
Cumulative distribution functions (CDFs) and empirical CDFs (eCDFs) are widely used to summarize distributions. The Dvoretzky-Kiefer-Wolfowitz (DKW) inequality shows the eCDF is a good approximation of the true.

If the underlying data is sensitive, we can use privacy-preserving approximations of the CDF. We want to:

- Balance privacy and utility/accuracy.
- Be compatible with federated and online updating/learning.

**Goal: Design a differentially private CDF approximation that allows for continuous updating.**

## Differential Privacy

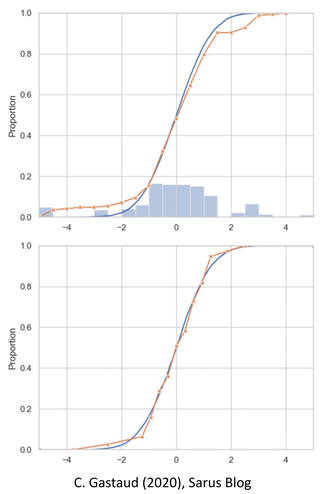



Privacy comes from a randomized map $Q(y|x)$. We say $Q$ guarantees $(\epsilon, \delta)$-differential privacy [1-3] if

$$Q(\mathcal{S}|x) \leq e^{\epsilon} Q(\mathcal{S}|x') + \delta$$

for all measurable subsets $\mathcal{S} \subseteq \mathcal{Y}$ and all $x, x' \in \mathcal{X}$ with $x \sim x'$ differ in a single $x_i$.

## Prior work on private CDF estimation




C. Gastaud (2020), Sarus Blog

Given scalar data $\mathcal{D} = \{x_i \in [A, B] : i \in [n]\}$, estimate the eCDF

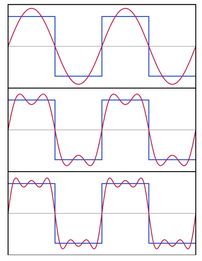
$$\hat{F}_X(t) = \frac{1}{n} \sum_{i=1}^{n} 1(x_i \leq t).$$

Two standard algorithms:

- **Histogram Queries (HQ)** bins the data, computes a differentially private histogram, and interpolates a CDF.
- **Adaptive Quantiles (AQ)** queries splits of data to sequentially estimate values of CDF (cf. binary search).

HQ and AQ methods have limitations in various scenarios.

## CDF estimation as signal approximation



The first three partial sums of the Fourier series for a square wave. Image sourced from Wikipedia.

**Approach: function approximation to estimate the CDF**

The eCDF is a very structured signal: piecewise constant, bounded, monotonic….

Signal processing has lots of techniques for structured signal approximation (Fourier Series, wavelets, polynomials).

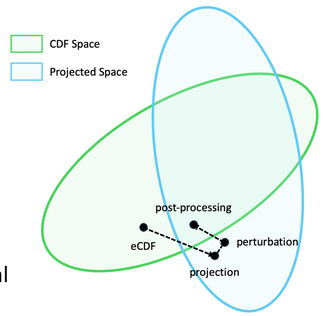
How should we introduce privacy?

## Polynomial Projection (PP) with Postprocessing

**Step 1**: Estimate the empirical CDF using a series of polynomial functions

$$\hat{F}_X(t) \approx \sum c_k P_k(t).$$

- Choose the polynomial space $\mathcal{P}$ spanned by the first $K+1$ Legendre polynomials $\{P_0, P_1, \ldots, P_K\}$, where each $P_k$ is a polynomial of degree of $k$.
- Obtain the optimal CDF estimate in the polynomial space by projecting the empirical CDF onto this space.




**Step 2**: Add noise to the coefficients for privacy protection.

- These coefficients are associated with the moments of the samples $\frac{1}{n}\sum x_i^m$, providing a straightforward approach for sensitivity computation.

**Step 3**: Isotonic regression is used as a post-processing method to ensure the estimated CDF is non-decreasing.

## Prior work on private CDF estimation

**Theorem (Upper Bound for $\| F^* - \tilde{F} \|_2$).**

Let $F^*$ be the true CDF for a random variable with $x \in [-1,1]$. If $\check{F}$ is the optimal approximation of $F^*$ in the polynomial space $\mathcal{P}$ and $\| F^* - \check{F} \|_2 \leq \alpha$, then with probability at least

$$1 - 2\exp\left(-\frac{N(\eta-\alpha)^2}{16}\right) - 2(K+1)\exp\left(-\frac{(\eta-\alpha)^2}{4(K+1)^4\sigma^2}\right),$$

we have $\| F^* - \tilde{F} \|_2 \leq \eta$ for $\eta > \alpha > 0$.

If the space is well-chosen, implying that $\check{F}$ represents $F^*$ well, then the DP $\tilde{F}$ can also approximate $F^*$ well.

Work supported by the USA NIH under Award 2R01DA040487: COINSTAC 2.0 (PI: V. Calhoun)

## Results

**Result 1**: PP outperforms HQ and is comparable to AQ, even better than it for certain distributions, such as Beta distribution in centralized settings.

**Result 2**: PP outperforms both HQ and AQ in various scenarios, such as decentralized settings and those involving newly collected data.

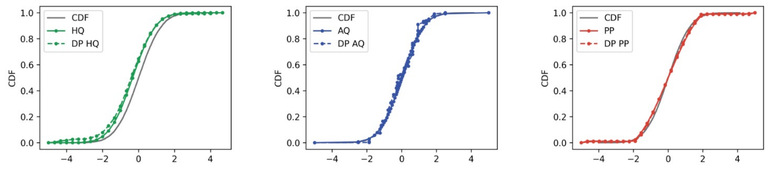



Figure 1. Apply different methods with Gaussian mechanism to normal distribution $\mathcal{N}(0,1)$ using the following parameters: $N = 10^4, \epsilon = 0.1, \delta = N^{-3/2}, K = 6$. The bin number for HQ is set to 30, and the number of iterations for AQ is 50.

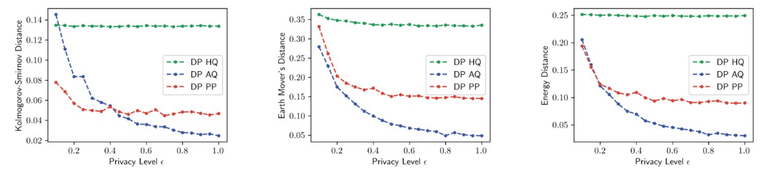



Figure 2. Comparison of distances between different DP CDF methods and the true CDF on normal distribution $\mathcal{N}(0,1)$. Experiment was run 50 times with $N = 10^4, \delta = N^{-3/2}, K = 6$. The bin number in HQ was set to 30, and the number of iterations in AQ was 50.

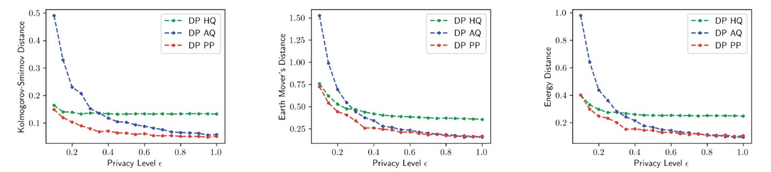



Figure 3. The experiment was run 50 times in decentralized setting on normal distribution $\mathcal{N}(0,1)$ with $N = 10^4, \delta = N^{-3/2}, K = 6$, and 10 sites . The bin number in HQ was set to 30, and the number of iterations in AQ was 50.

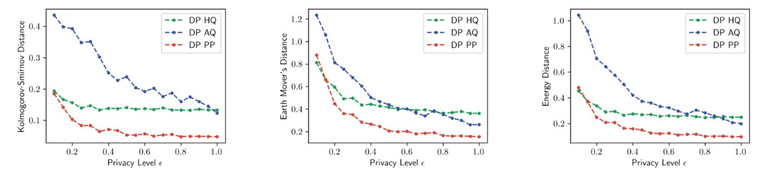



Figure 4. The experiment was run 50 times on distribution $\mathcal{N}(0,1)$ with $N = 10^4, \delta = N^{-3/2}, K = 6$. The CDF was updated after every 1000 new data points, for a total of 10 rounds of updates. Bin number in HQ was 30, and number of iterations in AQ was 50.

## Looking ahead

**Further/better approximation guarantees…**

**… empirical approaches using dictionary learning**

[1] Cynthia Dwork and Aaron Roth, "The algorithmic foundations of differential privacy," Foundations and Trends® in Theoretical Computer Science, vol. 9, no. 3–4, pp. 211–407, 2014.
[2] Peter Kairouz, Sewoong Oh, and Pramod Viswanath, "The composition theorem for differential privacy," in International Conference on Machine Learning. PMLR, 2015, pp. 1376–1385.
[3] Ilya Mironov, "Renyi differential privacy," in 2017 IEEE 30th Computer Security Foundations Symposium (CSF). IEEE, 2017, pp. 263–275.
[4] D. G. Luenberger, Optimization by vector space methods. John Wiley & Sons, 1997.