THE UNIVERSITY OF SYDNEY

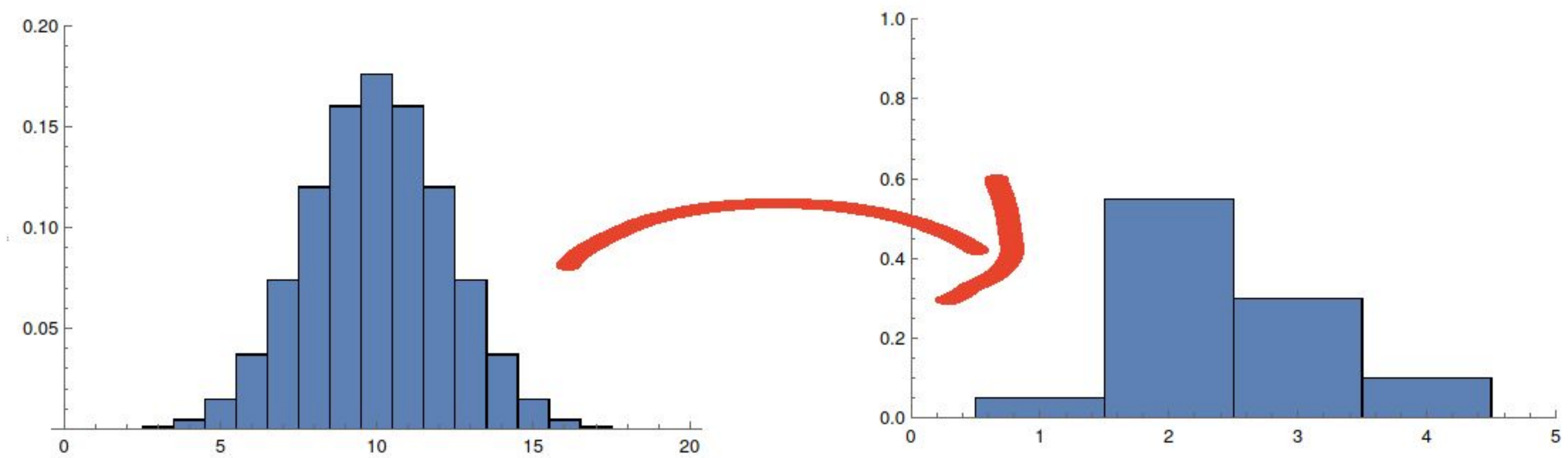
Dimensionality Reduction on the Simplex

Domain Compression Made Practical

Alex Tan, Clément Canonne

What is domain compression?

A **randomized algorithm/hash** that **compresses the support size** (i.e. number of elements) of **discrete probability distributions**



$$\mathcal{A} : \Delta_k \to \Delta_L \quad \text{for } L \ll k$$

Here Δ_k is the set of discrete probability distributions over {1, 2, ..., k}

In particular, we study a very simple hashing scheme:

- **each element of the original distribution is hashed to an element in the compressed distribution uniformly at random**
- as it turns out this simple method still **preserves some desirable statistical properties** of the distributions being compressed!

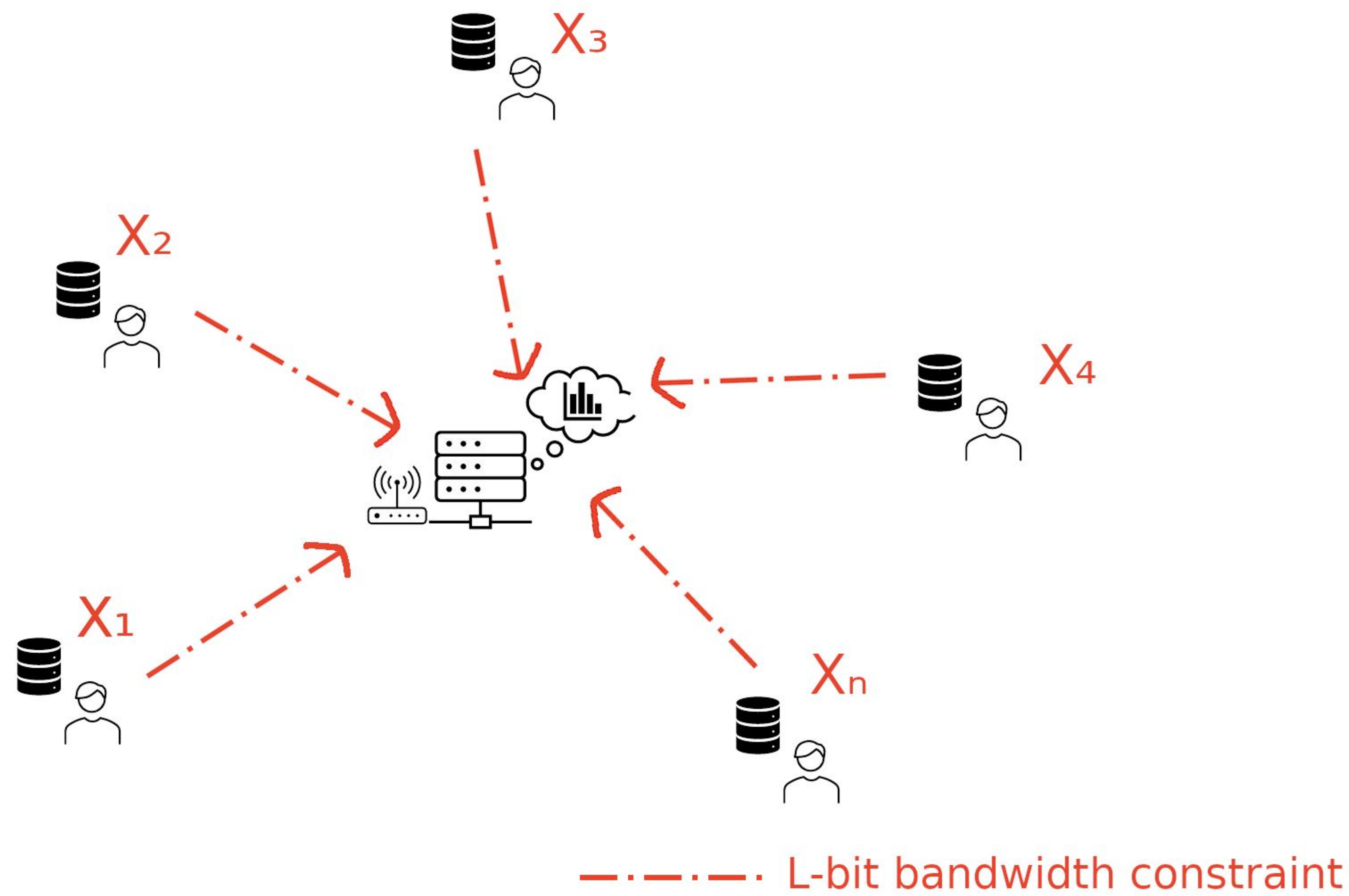
Why perform domain compression?

Applications:

- **communication bandwidth constraints**
- local differential privacy
- shuffle differential privacy
- quantum algorithms

Why perform domain compression? (cont.)

Example: statistical inference with **communication bandwidth constraints**



- Each user has **one independent sample** X_i from a fixed but unknown distribution p over {1, 2, ..., k}
- Central server **wants to use samples X_i to test a hypothesis** e.g. is the unknown distribution the uniform distribution?
- But each user can only send L < $\log_2 k$ bits of information to the server! The **full sample X_i cannot be sent**!
- Solution: **use domain compression**, and send compressed samples $A(X_i)$
- Server can **reconstruct i.i.d. samples from p using compressed samples $A(X_i)$** then perform inference using these new samples!

THE UNIVERSITY OF SYDNEY

Dimensionality Reduction on the Simplex

Domain Compression Made Practical

Alex Tan, Clément Canonne

Aim

$$\mathcal{A} : \Delta_k \to \Delta_L$$

This algorithm has the property that **distances between any two probability distributions are *probably approximately* preserved** i.e. there exist constants α>0, δ>0 such that

$$\mathbb{P}[\|\mathcal{A}(p) - \mathcal{A}(q)\|_2 > \alpha\|p - q\|_2] \geq 1 - \delta$$

for any two probability distributions p, q $\in \Delta_k$. We want to **improve upon these constants**.

Results

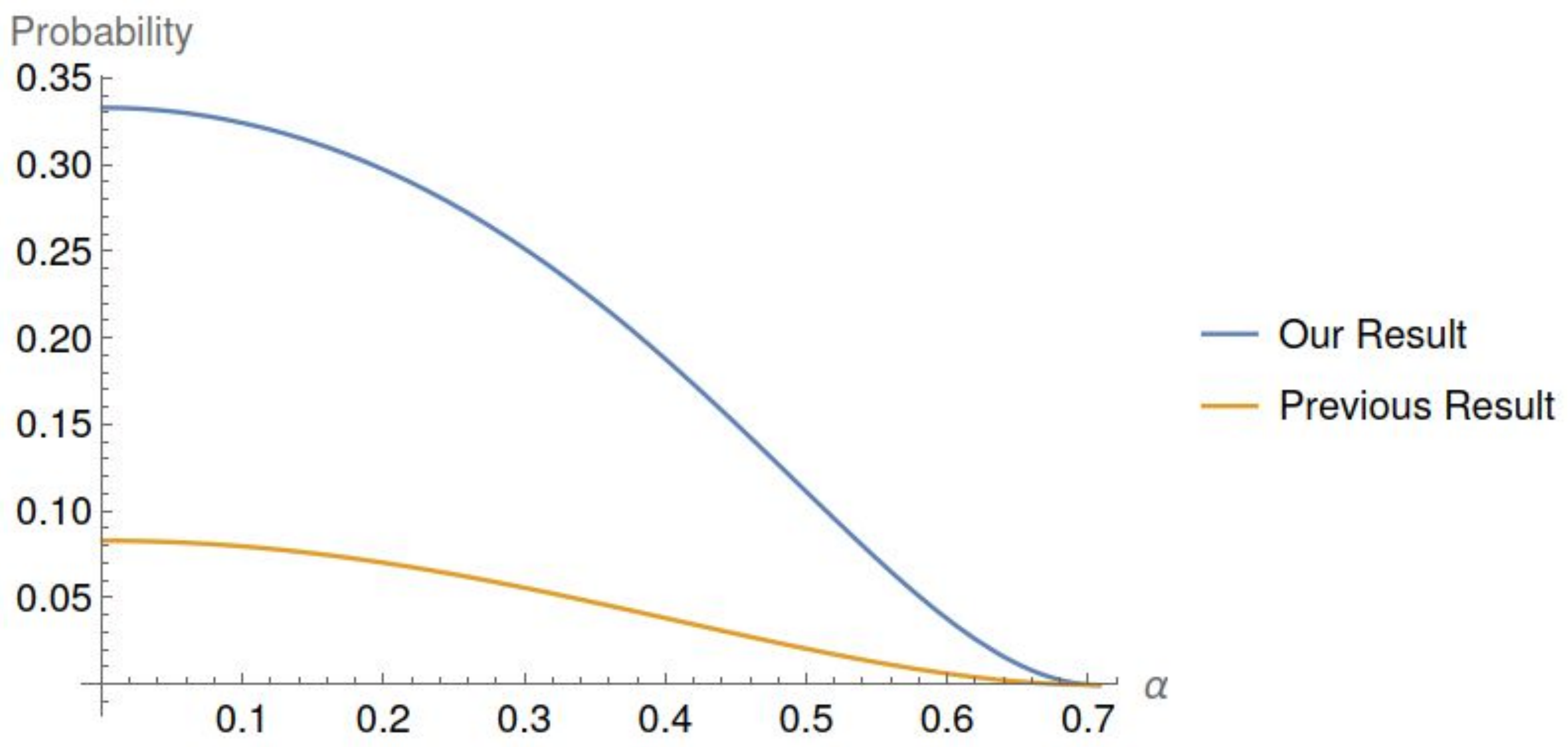
Using the **Paley-Zygmund inequality**, we prove that:

$$\mathbb{P}[\|\mathcal{A}(p) - \mathcal{A}(q)\|_2 > \alpha\|p - q\|_2] \geq 1 - \frac{2(L-1)}{L^2(\alpha^2-1)^2 + 2L\alpha^2 - 1}$$

In particular, for the binary case (L=2) this implies:

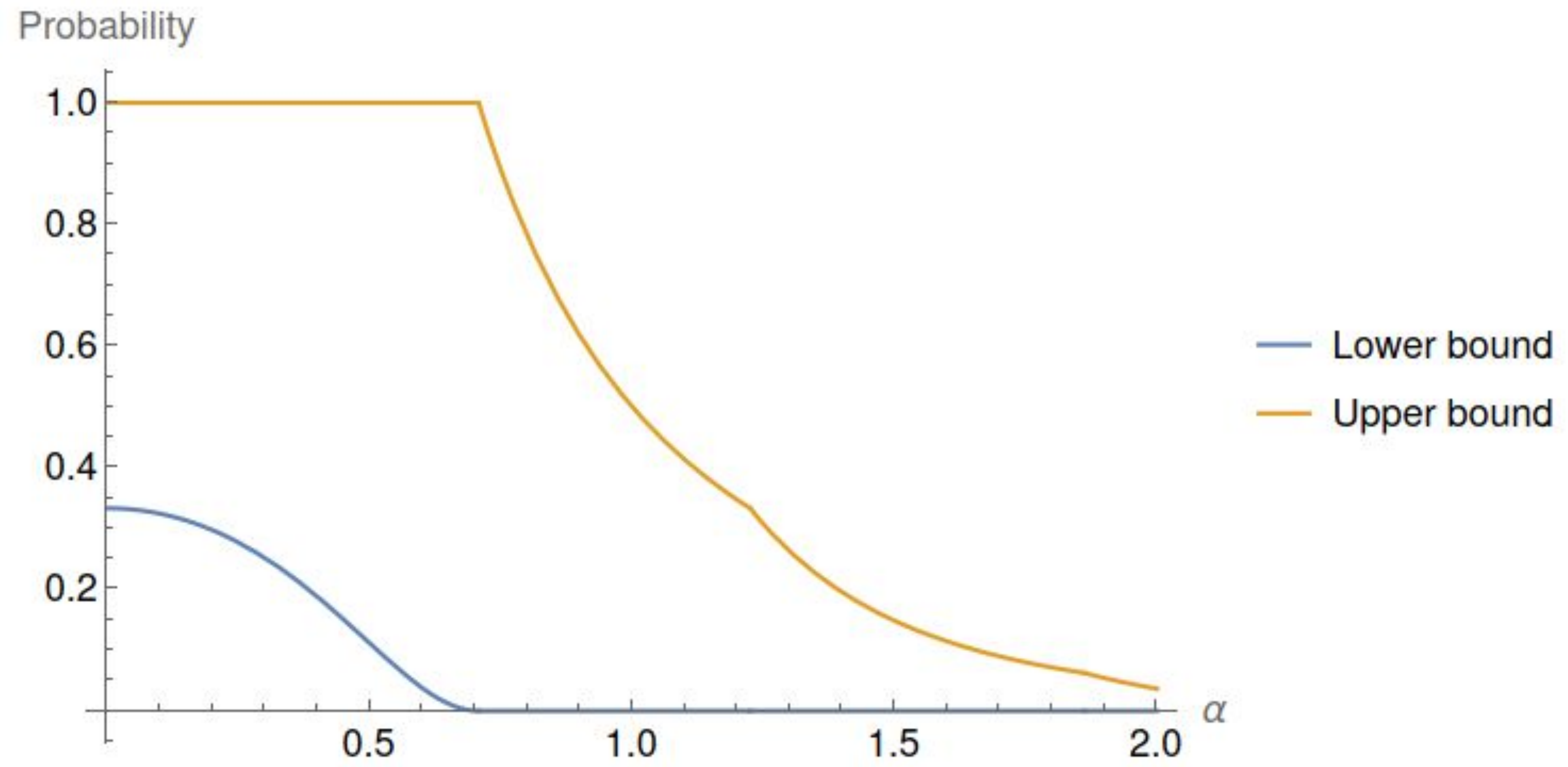
$$\mathbb{P}[\|\mathcal{A}(p) - \mathcal{A}(q)\|_2 > \alpha\|p - q\|_2] \geq 1 - \frac{2}{4\alpha^4 - 4\alpha^2 + 3}$$

This is an **improvement over previous results**.



Results (cont.)

For the binary case, we additionally prove an **upper bound** on the probability i.e. this tells us the **best we can hope for** using this compression scheme.



$$\mathbb{P}[\|\mathcal{A}(p) - \mathcal{A}(q)\|_2 > \alpha\|p - q\|_2] \leq \min\left(\frac{1}{2\alpha^2}, \frac{3}{4\alpha^4}, 2\exp(-\alpha^2)\right)$$

Proof involves using a cocktail of **concentration inequalities** (Markov's inequality, Chebyshev's inequality, Hoeffding's inequality)

Future work

- prove a similar result for **total variation distance** (L_1 norm) instead of L_2 norm
- are these bounds **tight**?

References

1. Canonne, Clément L. "Topics and Techniques in Distribution Testing: A Biased but Representative Sample." *Foundations and Trends® in Communications and Information Theory* 19.6 (2022): 1032-1198.
2. Acharya, Jayadev, et al. "Domain compression and its application to randomness-optimal distributed goodness-of-fit." Conference on Learning Theory. PMLR, 2020.
3. Acharya, Jayadev, Clément L. Canonne, and Himanshu Tyagi. "Inference under information constraints II: Communication constraints and shared randomness." IEEE Transactions on Information Theory 66.12 (2020): 7856-7877.