Lecture 6 (2-6-26)

Today: Linear Queries

Factorization Mechanisms

Restate linear queries setup from Tuesday.

General Question:

Given target "workload" of queries, what's the best subset to answer?

① Write dataset in normalized histogram representation

$$x = (x_1, \dots, x_n) \in \mathcal{U}^n$$

↳ say finite $|\mathcal{U}| = m$

$$\Rightarrow$$

h_x is vector in $\mathbb{R}^{\mathcal{U}}$

$$(h_x)_u = \frac{\#\{i : x_i = u\}}{n}$$

② For $x \sim x'$, $\|h_x - h_{x'}\|_1 \leq \frac{2}{n}$

③ For any predicate $\phi : \mathcal{U} \to \{0,1\}$, can write

$$f(x) = \frac{1}{n} \sum_{i=1}^{n} \phi(x_i) = \langle v_\phi, h_x \rangle.$$

$$v_\phi = (\phi(u_1), \dots \phi(u_m))$$

④ Workload $f_1, \ldots, f_k$, $F(x)$

$$F(x) = F h_x$$

F: $k \times m$, h_x: $m \times 1$

⑤ Write $F = RM$

$$M_{R,M}(x) = R(M h_x + Z)$$

R: reconstruct, M: measure, Z: Gaussian noise

$$= RM h_x + RZ$$

$$= F h_x + RZ$$

What is the error of a given factorization?

Q1: How much noise for privacy?

Q2: How much error?

Ans: $\gamma_2(F) = \min_{R,M} \|R\|_F \cdot \|M\|_{1 \to 2}$

optimal in some regimes