

# Locally Private Sampling with Public Data

Behnoosh Zamanlooy\* Mario Diaz Shahab Asoodeh

zamanlob@mcmaster.ca

## Contributions

- We introduce a locally private sampler that leverages public information within the minimax privacy-utility tradeoff framework.
- For discrete distributions, we fully characterize the privacy-utility tradeoff and propose an algorithm that achieves it.
- We demonstrate the significantly improved performance of our approach through comprehensive benchmarks against the baseline method proposed by [1], using both synthetic and real-world datasets.

## Locally Private Sampling [1]



**Definition.** The sampler is said to be  $\epsilon$ -LDP if

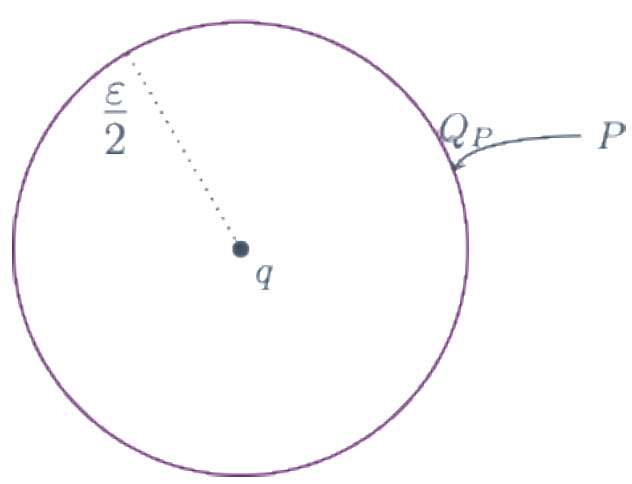
$$\forall P \text{ and } P' \in \Delta(\mathcal{X}) : \sup_{\text{event } A} \frac{Q_P(A)}{Q_{P'}(A)} \leq e^\epsilon.$$

## Background

### LDP Samplers through Relative Mollifiers [1]

Pick an arbitrary distribution,  $q$ , as a reference point and:

$$\mathcal{M}_{\epsilon, q} := \left\{ \tilde{q} \in \Delta(\mathcal{X}) : \sup_{x \in \mathcal{X}} \max \left\{ \frac{q(x)}{\tilde{q}(x)}, \frac{\tilde{q}(x)}{q(x)} \right\} \leq e^{\epsilon/2} \right\}$$



Objective:  $Q_P^* := \operatorname{argmin}_{P' \in \mathcal{M}_{\epsilon, q}} \text{KL}(P \| P')$

Solution:

$$Q_P^*(x) = \min \left\{ \max \left\{ \frac{q(x)}{e^{\epsilon/2}}, \frac{P(x)}{r_P} \right\}, e^{\epsilon/2} q(x) \right\}$$

where  $r_P$  is the normalizing constant.

### Minimax Optimal Samplers [2]

Objective:  $\inf_Q \sup_{P \in \Delta(\mathcal{X})} D_f(p \| Q_P)$

Solution (non-linear sampler):

$$Q_P^*(x) = \max \left( \frac{1}{r_P} P(x), \frac{1}{e^\epsilon + k - 1} \right)$$

where  $k$  is the support size and  $r_P$  is the normalizing constant.

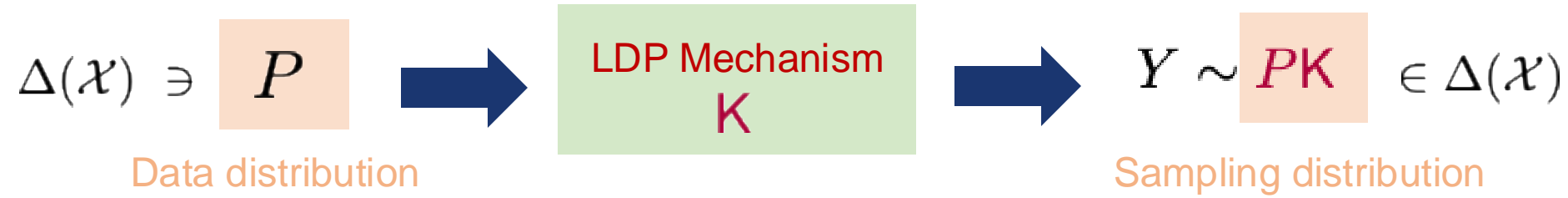
Solution (linear sampler):

$$Q_P^*(x) = \frac{e^\epsilon - 1}{e^\epsilon + k - 1} P(x) + \frac{k}{e^\epsilon + k - 1} U(x)$$

where  $k$  is the support size of the  $P$  and  $U$  is the uniform distribution.

## Our Setting

**Observation:** If LDP sampler is linear in  $P$ , then it can be characterized as a classical LDP mechanism  $K$ .



$P_{\text{pub}} \sim \begin{cases} \text{Demographic Information} \\ \text{Historical Data} \end{cases}$

**Question.** Given public prior  $P_{\text{pub}}$  on individuals and linear samplers, how to design a sampler that doesn't perturb  $P$  if  $P \approx P_{\text{pub}}$ ?

$$P_{\text{pub}} K = P_{\text{pub}}$$

## Main Question

Find  $K_{P_{\text{pub}}, \epsilon}$  that solves:

$$\Gamma_f(P_{\text{pub}}, \epsilon) := \inf_{\substack{\epsilon\text{-LDP Mechanism } K \\ \text{with fixed point } P_{\text{pub}}}} \sup_{p \in \Delta(\mathcal{X})} D_f(p \| pK)$$

## Optimal Utility

**Theorem.** Define  $\alpha := \min_{x \in \mathcal{X}} P_{\text{pub}}(x)$ . Then

$$\Gamma_f(q, \epsilon) = \frac{1 - \alpha}{e^\epsilon \alpha + 1 - \alpha} f(0) + \frac{e^\epsilon \alpha}{e^\epsilon \alpha + 1 - \alpha} f\left(\frac{e^\epsilon \alpha + 1 - \alpha}{e^\epsilon \alpha}\right).$$

**Sketch of Proof.** Follows from the joint convexity of  $f$ -divergences and the fact that the supremum is attained at Dirac distributions.

## Optimal Mechanism

**Lemma.** Take  $P_{\text{pub}} \sim \text{Ber}(\alpha)$  with  $\alpha \leq \frac{1}{2}$ , then:

$$K_{P_{\text{pub}}, \epsilon} = \frac{1}{e^\epsilon \alpha + 1 - \alpha} \begin{bmatrix} e^\epsilon \alpha & 1 - \alpha \\ \alpha & (e^\epsilon - 1)\alpha + 1 - \alpha \end{bmatrix}.$$

**Algorithm 1**  $K_{\text{with\_prior}}$ : Algorithm to compute the optimal mechanism  $K_{q, \epsilon}$

**Require:**  $q$  - Increasingly Sorted Public Prior,  $\epsilon$  - Privacy Parameter

**Ensure:**  $\epsilon$ -LDP mechanism  $K_{q, \epsilon}$  with  $qK_{q, \epsilon} = q$

- $n \leftarrow$  length of  $q$
- if**  $n = 2$  **then**
- return** The optimal binary mechanism
- else**
- $K_{q, \epsilon} \leftarrow$  zeros( $n, n$ )
- $d \leftarrow (e^\epsilon \cdot q_{\min} + 1 - q_{\min})$
- $(K_{q, \epsilon})_{11} \leftarrow \frac{e^\epsilon \cdot q_{\min}}{d}$
- for**  $i \leftarrow 2$  **to**  $n$  **do**
- $(K_{q, \epsilon})_{i1} \leftarrow \frac{q_{\min}}{d}$
- end for**
- $\bar{q} \leftarrow \frac{1}{\sum_{i=2}^n q_i} [q_2, \dots, q_n]$
- $K_{\bar{q}, \epsilon} \leftarrow K_{\text{with\_prior}}(\bar{q}, \epsilon)$
- $m \leftarrow 1 - \frac{q_{\min}}{d}$
- $K_{q, \epsilon}[2 : (n), 2 : (n)] \leftarrow m \cdot K_{\bar{q}, \epsilon}$
- for**  $i \leftarrow 2$  **to**  $n$  **do**
- $(K_{q, \epsilon})_{1i} \leftarrow \frac{q_i}{d}$
- end for**
- return**  $K_{q, \epsilon}$
- end if**

$$\bar{q} = \frac{1}{\sum_{i=2}^n q_i} [q_2, \dots, q_n]$$

$$K_{q, \epsilon} = \begin{bmatrix} \frac{e^\epsilon q_{\min}}{e^\epsilon q_{\min} + 1 - q_{\min}} & \frac{q_2}{e^\epsilon q_{\min} + 1 - q_{\min}} & \dots & \frac{q_n}{e^\epsilon q_{\min} + 1 - q_{\min}} \\ \frac{q_{\min}}{e^\epsilon q_{\min} + 1 - q_{\min}} & m K_{\bar{q}, \epsilon} & & \end{bmatrix}$$

## Optimality of Algorithm

**Theorem.** The mechanism described by Algorithm 1 is an  $\epsilon$ -LDP mechanism with fixed point  $P_{\text{pub}}$  and is **independent** of the selected  $f$ -divergence.

## Synthetic Data

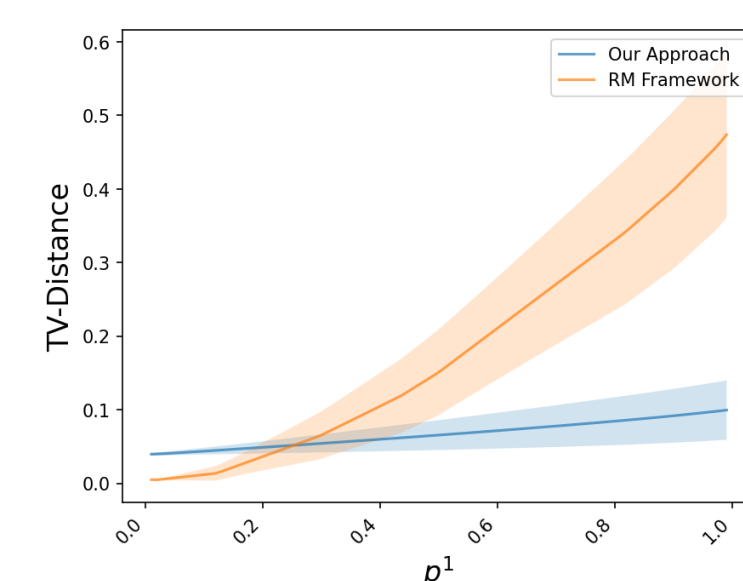
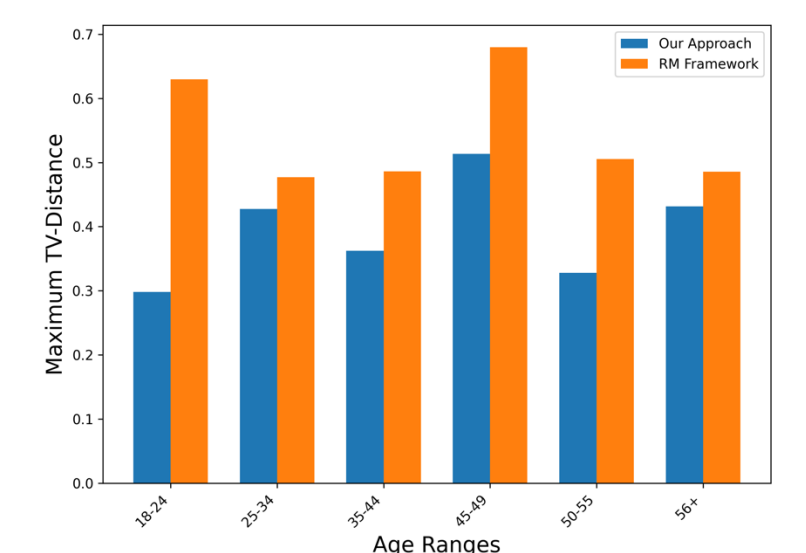
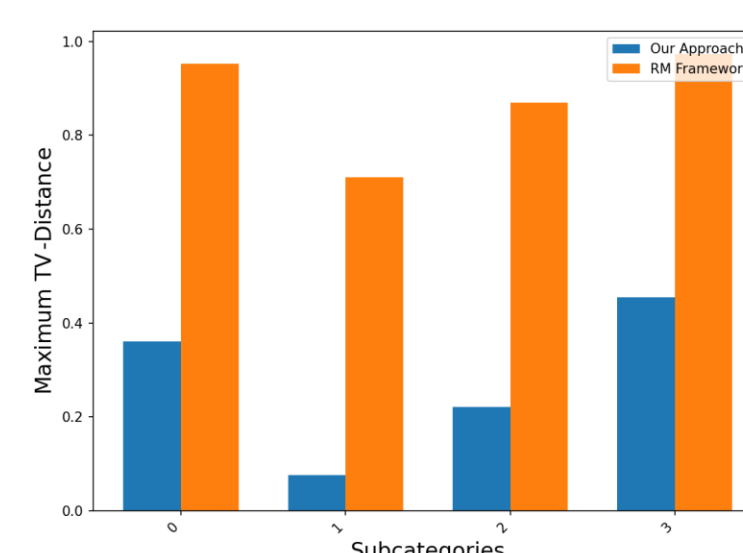


Figure 1. Comparison of our approach with the RM framework when  $P = [p^1, (1-p^1)/n, (1-p^1)/n, \dots, (1-p^1)/n]$  with  $P_{\text{pub}}(x) \propto U(0, 1)$  and  $\epsilon = 8$ .

## Click Rate and MovieLens Datasets



- Private distribution:** Each user's click-rate history per website with at least 100 clicks within anonymous category C18.
- Public prior:** Average of users' private distributions in each subcategory.

- Private distribution:** Each user's rating history per genre in (MovieLens 1M) with at least 20 ratings.
- Public prior:** (MovieLens 100K) is used to infer users' popular genre by age range.