Bayesian Frequency Estimation Under Local Differential Privacy With an Adaptive Randomized Response Mechanism

Sinan Yıldırım
Joint work with Soner Aydın (PhD candidate, SU)

January 27, 2025

Data analysis vs Privacy

Sensitive data set of n individuals: $X_1, \ldots, X_n$

Two conflicting interests:

1. We want to work with sensitive data sets
   - to perform inference about a population.
   - for optimization
   - etc.

2. Individuals contributing to data sets with their sensitive information want to preserve their privacy.

A significant amount of research is devoted to developing useful methods for data analysis while protecting data privacy.

An outline

This talk: Introduce **AdOBEst-LDP**: A framework for efficient parameter estimation under privacy constraints.

- Local differential privacy
- Randomized response mechanisms
- Posterior sampling
- Some theory

Local Privacy

Individual with *sensitive* information $X \in \mathcal{X}$.

X is shared as Y through some mechanism.

Data privacy: main question

How should Y be shared so that

- privacy of each individual is protected, and
- the shared information Y is useful.

Some extreme solutions(?)

- **Full transparency:** Share $Y = X$.
  - Very useful, but not private.
- **Full secrecy:** Toss a coin and share the outcome.
  - Very private, but not useful.

Local differential privacy

Uses a randomized mechanism to generate Y from X.

Local Differentila Privacy (LDP)

A randomized mechanism $M : \mathcal{X} \to \mathcal{Y}$ satisfies ϵ-LDP if:

$$e^{-\epsilon} \leq \frac{\Pr(M(x) = y)}{\Pr(M(x') = y)} \leq e^{\epsilon}, \quad \forall x, x' \in \mathcal{X}, y \in \mathcal{Y}.$$

- Smaller ϵ implies stronger privacy guarantees.
- LDP operates on individual data points, unlike global DP, which operates on datasets.

Categorical data

Sensitive individual data: $X \in [K] := \{1, \ldots, K\}$.

Randomized response $Y \in [K]$ using a mechanism M.

Requirement for ϵ-LDP:

$$e^{-\epsilon} \leq \frac{\Pr(M(x) = y)}{\Pr(M(x') = y)} \leq e^{\epsilon}, \quad \forall x, x', y \in [K].$$

Standard randomized response (SRR) mechanism

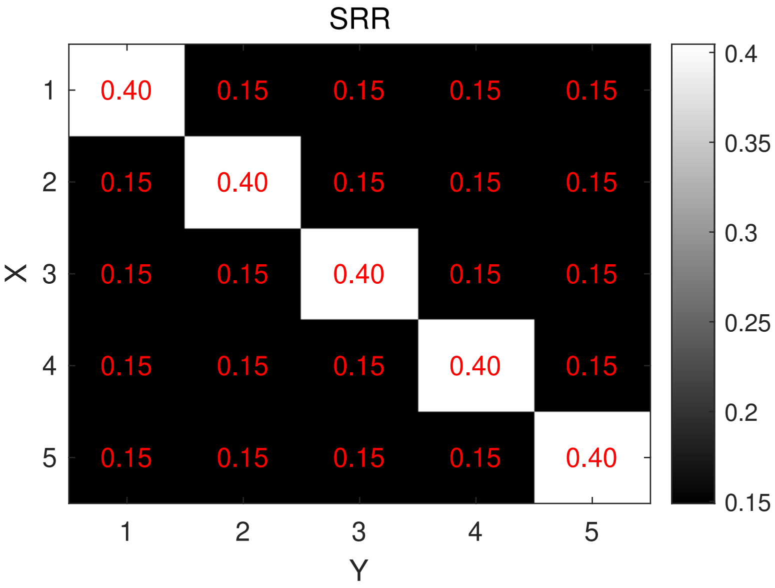
SRR

Return $Y = X$ with probability $e^\epsilon/(e^\epsilon + K - 1)$, else return any other element at random.

As a general mechanism on a finite set Ω:

$$\mathrm{SRR}(X; \Omega, \epsilon) = \begin{cases} X & \text{w.p. } e^\epsilon/(e^\epsilon + |\Omega| - 1) \\ \sim \mathrm{Uniform}(\Omega/\{X\}) & \text{else} \end{cases}.$$

Transition matrix for SRR



What to do with randomized responses?

- **Sensitive data** from n individuals from a population:

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Categorical}(\theta_1 \ldots, \theta_K).$$

$(\Pr(X_i = k) = \theta_k)$

- **Observations:** Randomized responses are collected.

$$Y_1 = M(X_1), \ldots, Y_n = M(X_n)$$

- **Goal:** Estimate $\theta = (\theta_1, \ldots, \theta_K)$ from $Y_1, \ldots, Y_n$ as accurately as possible, while maintaining ϵ-DP.

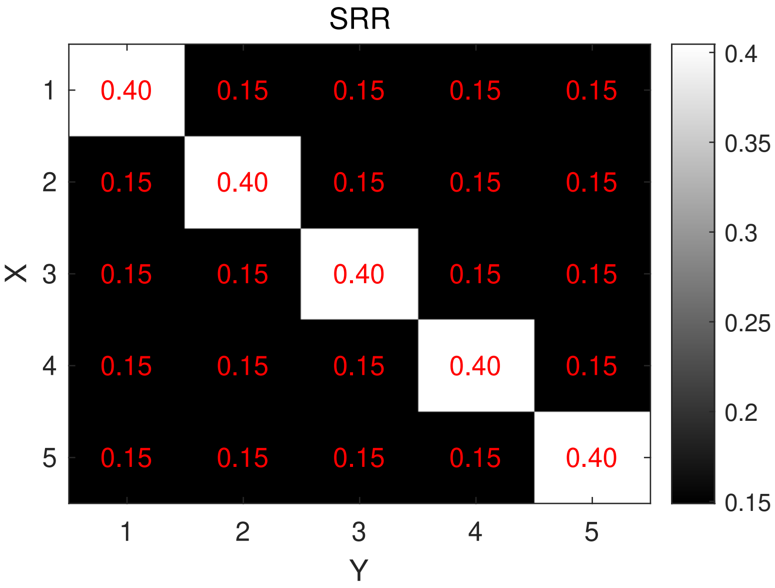
Can we increase utility adaptively?

An ϵ-LDP mechanism is not unique; SRR is just one of them.

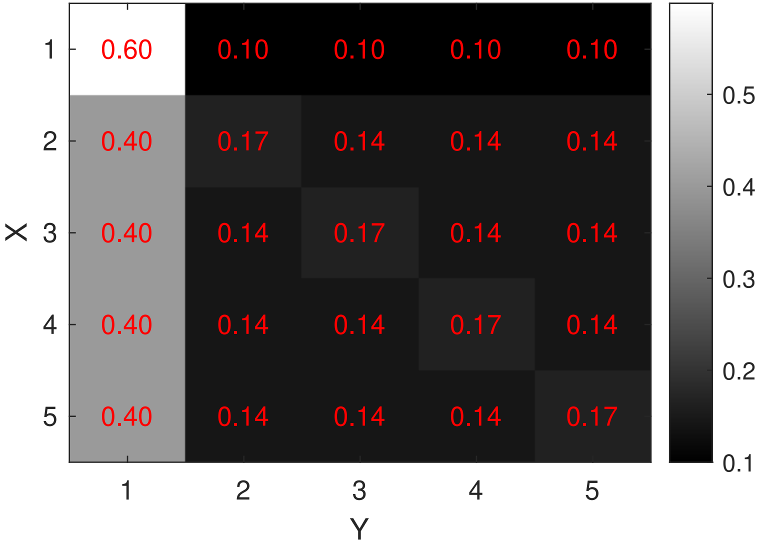
We have freedom over the mechanism to generate the response Y_i (under the ϵ-DP constraint).

Research question: Can we design a randomized mechanism adaptable to current knowledge of θ?

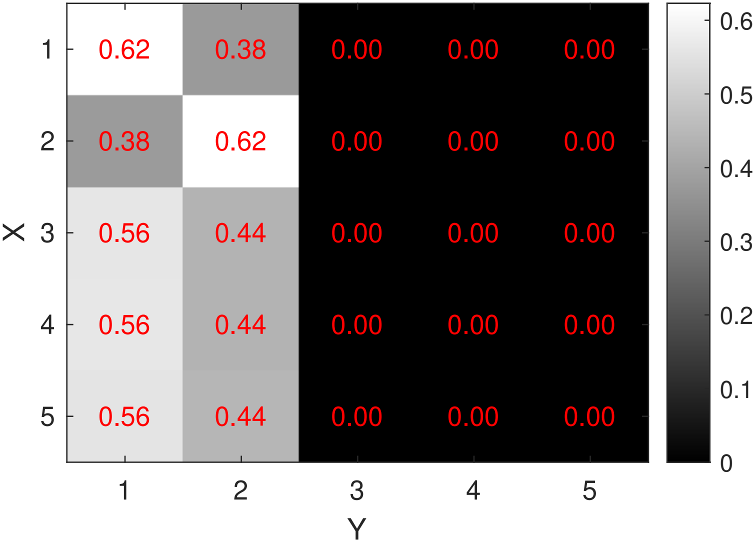
Some ϵ-LDP mechanisms



Some ϵ-LDP mechanisms



Some ϵ-LDP mechanisms



Main idea with an example

Suppose there are 20 political parties,

Only 4 parties (1, 2, 3, 4) are estimated to constitute %95 of the votes.

A naive mechanism based on this estimate:

- If the user's party $X_i \in \{1, \ldots, 4\}$; apply SRR on $\{1, \ldots, 4\}$;
- Otherwise, return a random element from $\{5, 6, \ldots, 20\}$.

With prob. 0.95, we will receive $Y = X$ with probability $e^{\epsilon}/(3 + e^{\epsilon})$ (in contrast to $\epsilon^{\epsilon}/(19 + e^{\epsilon})$).

RRRR: Randomly restricted randomized response

Randomizes responses over a high-probability subset S (mostly!)

Algorithm 1: $\texttt{RRRR}(X; S, \epsilon)$

Input: Input $X \in [K]$, subset $S \subset [K]$, privacy parameters $\epsilon_1, \epsilon_2 > 0$

Output: Randomized response $Y \in [K]$

if $X \in S$ **then**

  Draw $R \sim \text{Uniform}(S^c)$.

  Set $Y = \texttt{SRR}(X; S \cup \{R\}, \epsilon_1)$.

else

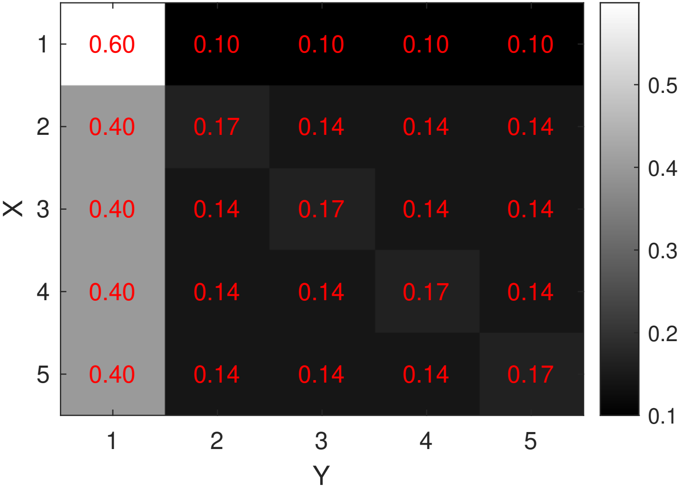
  Set $R = \texttt{SRR}(X; S^c, \epsilon_2)$.

  Set $Y = \texttt{SRR}(R; S \cup \{R\}, \epsilon_1)$.

return Y

Transition matrix for RRRR

RRRR designed for $\theta = (0.80, 0.05, 0.05, 0.05, 0.05)$



LDP of RRRR

LDP of RRRR

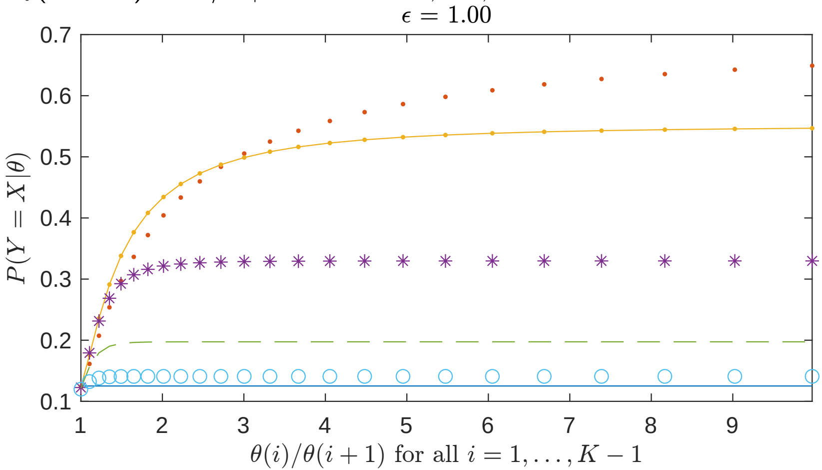
RRRR is ϵ-LDP if $\epsilon_1 \leq \epsilon$ and

$$\epsilon_2 = \begin{cases} \min\left\{\epsilon, \ln \frac{|S^c|-1}{e^{\epsilon_1-\epsilon}|S^c|-1}\right\} & \text{for } \epsilon-\epsilon_1 < \ln|S^c| \text{ and } |S|>0 \\ \epsilon & \text{else} \end{cases}.$$

With $|S| = 0$ and $\epsilon_2 = \epsilon$, RRRR reduces to SRR.

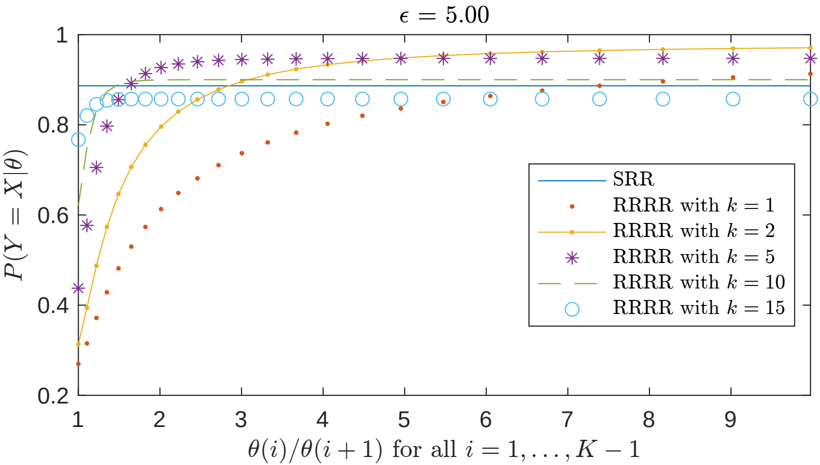
Illustration

$\mathbb{P}_\theta(Y = X)$ vs θ_i/θ_{i+1} for all $i = 1, \ldots, K-1$ with $K = 20$. $\epsilon = 1$



Illustration

$\mathbb{P}_\theta(Y = X)$ vs θ_i/θ_{i+1} for all $i = 1, \ldots, K-1$ with $K = 20$. $\epsilon = 5$



Subset selection in RRRR

$U(\theta, S, \epsilon)$: utility of $Y = \mathtt{RRRR}(X; S, \epsilon)$ when $X \sim \text{Category}(\theta)$.

$$S_\theta^* = \arg \max_{S \subset \{0,\ldots,K\}} U(\theta, S, \epsilon).$$

There are $2^K - 1$ choices for S, one must confine the search space.

Subset selection in RRRR

RRRR becomes most relevant when the set S is a high-probability set.

Consider the alternatives

$$S_{\theta,k} := \{\sigma_\theta(1), \sigma_\theta(2), \ldots, \sigma_\theta(k)\}, \quad k = 1, \ldots, K.$$

where σ_θ is such that $\theta_{\sigma_\theta(1)} \geq \ldots \geq \theta_{\sigma_\theta(K)}$.

Then the subset selection problem can be formulated as finding

$$k^* = \arg \max_{k \in \{0,\ldots,K-1\}} U(\theta, S_{k,\theta}, \epsilon).$$

Utility Functions for Subset Selection

1. **Fisher Information**

$$U_1(\theta, S, \epsilon) = -\mathrm{Tr}\big(F^{-1}(\theta; S, \epsilon)\big),$$

   where F is the Fisher Information Matrix.

2. **Entropy of Randomized Response**

$$U_2(\theta, S, \epsilon) = -\sum_{y \in Y} \Pr(Y = y|\theta) \log \Pr(Y = y|\theta).$$

3. **Total Variation Distance - 1**

$$U_3(\theta, S, \epsilon) = \mathbb{E}[\mathrm{TV}(\Pr(X|Y, \theta), \Pr(X|\theta))].$$

Utility Functions for Subset Selection

4. **Total variation distance**

$$U_4(\theta, S, \epsilon) = -\mathrm{TV}(\Pr(Y|\theta), \Pr(X|\theta))$$

   where F is the Fisher Information Matrix.

5. **Expected mean squared error**

$$U_5(\theta, S, \epsilon) = -\arg\min_{\widehat{e_X}} \mathbb{E}_\theta \left[\|e_X - \widehat{e_X}(Y)\|^2 \right].$$

6. **Probability of honest response**

$$U_6(\theta, S, \epsilon) = \Pr(Y = X|S, \theta).$$

Overall algorithm: AdOBEst-LDP

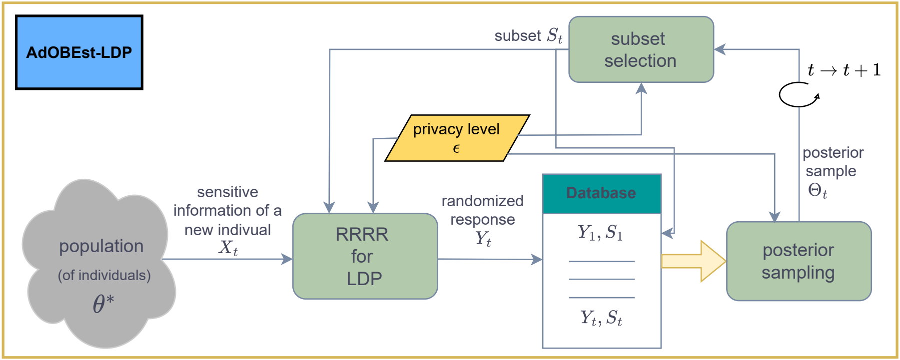
Algorithm 2: AdOBEst-LDP: Adaptive Online Bayesian Estimation with LDP

Initialization: Start with an initial estimator $\Theta_0 = \theta_{\text{init}}$.

for $t = 1, 2, \ldots$ **do**

- **Step 1: Subset selection in RRRR:** Based on Θ_{t-1}, determine the subset S_t for RRRR.
- **Step 2: LDP response generation** The sensitive information X_t of individual t is shared as $Y_t = \mathtt{RRRR}(X_t; S_t, \epsilon)$.
- **Step 3:** Draw a sample Θ_t from the posterior distribution given $Y_{1:t}$.

AdOBEst-LDP


AdOBEst-LDP
subset S_t
subset selection
$t \to t+1$
privacy level ϵ
posterior sample Θ_t
sensitive information of a new indivual X_t
randomized response Y_t
Database
Y_1, S_1
Y_t, S_t
population
(of individuals)
θ^*
RRRR for LDP
posterior sampling

Posterior Sampling: Stochastic Gradient Langevin Dynamics

Goal: Sampling θ from the posterior:

$$\pi(\theta | Y_{1:n}, S_{1:n}) \propto \eta(\theta) \prod_{t=1}^{n} \Pr(Y_t | \theta, S_t).$$

Solution: Use SGLD for scalable, approximate sampling:

- Latent variables $\phi_i \sim \text{Gamma}(\rho_i, 1)$ such that $\theta_i = \phi_i / \sum_j \phi_j$.
- Perform updates with minibatches of size m:

$$\phi^{(r)} = \left| \phi^{(r-1)} + \frac{\gamma_n}{2} \left(\nabla_\phi \ln p(\phi^{(r-1)}) + \frac{n}{m} \sum_{i \in u} \nabla_\phi \ln \Pr(y_i | \phi^{(r-1)}) \right) + \gamma_n W_r \right|.$$

where $W_j \sim \mathcal{N}(0, I)$.
Reflection ensures positivity.

Theoretical results

- Given $Y_{1:n}$ and $S_{1:n}$, the posterior distribution

$$\Pi(A|Y_{1:n}, S_{1:n}) := \frac{\int_A \eta(\theta) \prod_{t=1}^n P_{S_t,\epsilon}(Y_t|\theta)\mathrm{d}\theta}{\int_\Delta \eta(\theta) \prod_{t=1}^n P_{S_t,\epsilon}(Y_t|\theta)\mathrm{d}\theta}.$$

- $Q(\cdot|Y_{1:n}, S_{1:n}, \Theta_{n-1})$: posterior sampling for Θ_n.
- S^*_θ: best subset at θ so that $S_t = S^*_{\Theta_{t-1}}$.

The joint law of $S_{1:n}$, $Y_{1:n}$:

$$P_{\theta^*}(S_{1:n}, Y_{1:n}) := \prod_{t=1}^n P_{S_t,\epsilon}(Y_t|\theta^*) \left[\int_\Delta \mathbb{I}(S_t = S_{k^*,\theta_{t-1}}) Q(\mathrm{d}\theta_{t-1}|Y_{1:t-1}, S_{1:t-1}, \theta_{t-2})\right],$$

Does $\Pi(\cdot|Y_{1:n}, S_{1:n})$ converge to θ^*?

Convergence of the posterior distribution

Regularity assumption on the prior

There exist finite positive constants $d > 0$ and $B > 0$ such that $\eta(\theta)/\eta(\theta') < B$ for all $\theta, \theta' \in \Delta$ whenever $\|\theta' - \theta^*\| < d$.

Theorem

There exists a constant $c > 0$ such that, for any $0 < a < 1$ and the sequence of sets

$$\Omega_n = \{\theta \in \Delta : \|\theta - \theta^*\|^2 \leq cn^{-a}\},$$

the sequence of probabilities

$$\lim_{n\to\infty} \Pi(\Omega_n | Y_{1:n}, S_{1:n}) \overset{P_{\theta^*}}{\to} 1,$$

regardless of the choice of Q.

Probability of best subset selection

Let $S^* := S^*_{\theta^*}$ be the best subset at θ^*. How often is it selected?

Assumptions

- The components of θ^* are strictly ordered.
- Given any $S \subset [K]$ and $\epsilon > 0$, $U(\theta, S, \epsilon)$ is a continuous function of θ with respect to the L_2-norm.
- The best subset S_{θ^*} is unique.

Theorem

If Θ_ts are generated by exact sampling,

$$\lim_{n\to\infty} P_{\theta^*}(S_n = S^*) \to 1.$$

$$\lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} E_{\theta^*}\left[\mathbb{I}(S_t = S^*)\right] = 1.$$

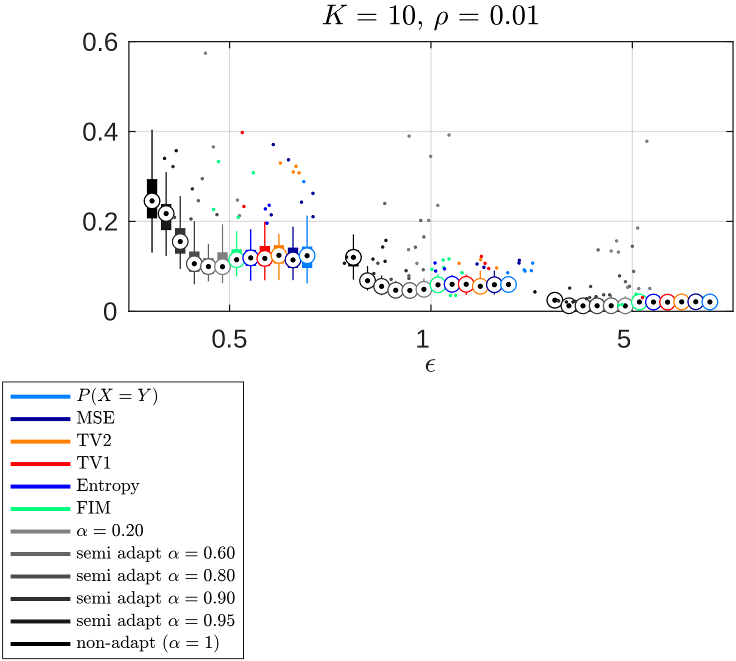
Numerical Experiments

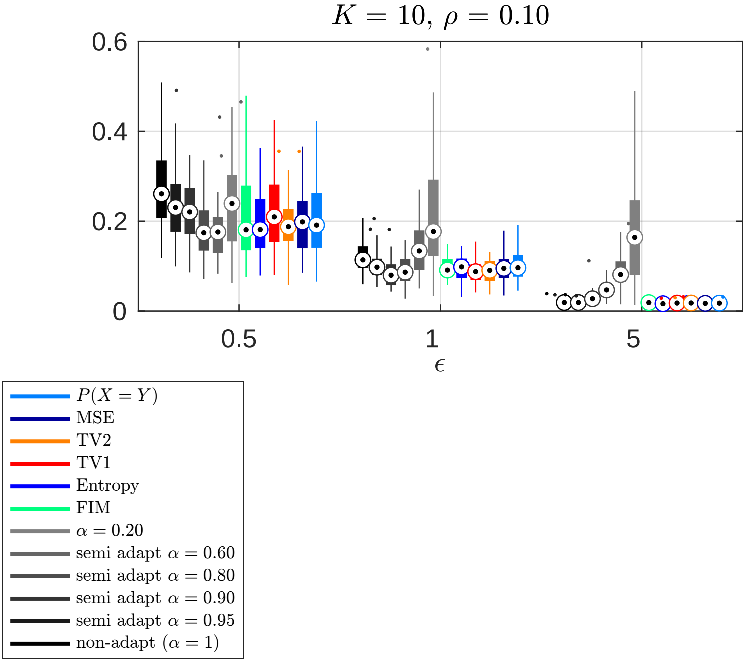
AdOBEst-LDP was tested with varying parameters:

- Privacy levels $\epsilon \in \{0.5, 1, 5\}$.
- Population distributions with uneven components (e.g., Dirichlet hyperparameter $\rho \in \{0.01, 0.1, 1\}$).

Performance metric:

$$\frac{1}{2} \sum_{k=1}^{K} |\theta_k - \hat{\theta}|.$$


$K = 10,\ \rho = 0.01$
0.6
0.4
0.2
0
0.5
1
5
ϵ
$P(X = Y)$
MSE
TV2
TV1
Entropy
FIM
$\alpha = 0.20$
semi adapt $\alpha = 0.60$
semi adapt $\alpha = 0.80$
semi adapt $\alpha = 0.90$
semi adapt $\alpha = 0.95$
non-adapt ($\alpha = 1$)


$K = 10,\ \rho = 0.10$
0.6
0.4
0.2
0
0.5
1
5
ϵ
$P(X = Y)$
MSE
TV2
TV1
Entropy
FIM
$\alpha = 0.20$
semi adapt $\alpha = 0.60$
semi adapt $\alpha = 0.80$
semi adapt $\alpha = 0.90$
semi adapt $\alpha = 0.95$
non-adapt ($\alpha = 1$)

Performance evaluation

Key Findings:

- Adaptive methods outperform non-adaptive counterparts, especially
  - at high privacy levels ($\epsilon < 1$)
  - non-even distribution ($\rho \ll 1$).
- Utility functions yield robust performance across settings.
- (Semi-adaptive approaches are computationally cheaper but require careful tuning.)

Key Takeaways

- AdOBEst-LDP: A new framework for Bayesian frequency estimation via adaptive LDP.
- SGLD makes the approach scalable.
- Several utility functions provide flexibility.

Future Work

- Extending to non-categorical data distributions.
- Investigating alternative utility functions for subset selection.
- Enhancing scalability for very large population sizes.