Estimating the True Distribution of Data Collected with Randomized Response

carlos.pinzon@inria.fr

Carlos Pinzón[1], Ehab ElSalamouny[1,5], Lucas Massot[3], Alexis Miller[4], Héber H. Arcolezi[2], Catuscia Palamidessi[1]

[1]INRIA Saclay, France; [2]INRIA Grenoble, France; [3]École Polytechnique, France; [4]Ecole Normale Supérieure de Lyon, France; [5]Suez Canal University, Egypt

Code and paper

Introduction

Local Differential Privacy (LDP) provides privacy guarantees for data collection without requiring trust in the data collector.

Randomized Response (RR) is a fundamental LDP mechanism:

- Each user reports their true value with probability p or a random value otherwise
- Satisfies ϵ-LDP with $\epsilon = \log(p/q)$ where $q = (1-p)/(K-1)$
- Used by major tech companies for telemetry data collection
- Optimal for small domains: $K < 3e^{\epsilon} + 2$

Problem: The standard debiasing rule can produce *invalid* histograms with negative values. What is the best fix?

Problem Formulation

- N users, each with secret value $x_u \in \{1, \ldots, K\}$
- True distribution (unknown): $\theta_i = \frac{|\{u : x_u = i\}|}{N}$. $\quad \theta \in \Delta := \{\theta \in \mathbb{R}^K : \sum_i \theta_i = 1, \theta_i \geq 0\}$
- Each user applies RR mechanism: $y_u = \mathcal{M}(x_u)$
- Observed histogram: $\phi_i = \frac{|\{u : y_u = i\}|}{N}$. $\quad \phi \in \Delta$
- **Goal:** Estimate θ from ϕ

RR Mechanism:

$$\Pr(\mathcal{M}(x) = y) = \begin{cases} p \text{ if } y = x & p = \dfrac{e^{\epsilon}}{e^{\epsilon} + K - 1} \\ q \text{ otherwise} & q = \dfrac{1}{e^{\epsilon} + K - 1} \end{cases}$$

Existing Estimators

Method	Complexity	Properties
`Inv`	$O(K)$	Unbiased but invalid
`InvN`	$O(K)$	Valid, simple workaround
`InvP`	$O(K \log K)$	Valid, closest to `Inv`
`IBU`	$O(K\, N_{\text{iters}})$	Valid, MLE as $N_{\text{iters}} \to \infty$
`MLE*`	$\mathbf{O(K \log K)}$	**Valid, exact MLE**

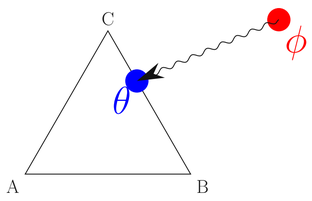
Linear Inversion (`Inv`) — debiasing rule:

$$\hat{\theta}_i^{\texttt{Inv}} = \frac{\phi_i - q}{p - q}$$

- Unbiased: $E[\hat{\theta}^{\texttt{Inv}}] = \theta$
- Can have $\hat{\theta}_i < 0$ (invalid!)
- Probability of invalidity is non-negligible for any N

Workarounds:

- `InvN`: Set negatives to zero and normalize
- `InvP`: Project to simplex $\hat{\theta}^{\texttt{InvP}} = \arg\min_{\theta \in \Delta} \|\theta - \hat{\theta}^{\texttt{Inv}}\|$



Iterative Bayesian Update (`IBU`):

- Converges to MLE as iterations $\to \infty$
- Requires many iterations in practice
- No theory on stopping conditions

Our Contribution: Closed Formula for MLE

Key Idea: Apply threshold transformation to ϕ, then use `Inv`

For threshold τ, define:

$$\phi_i^{\tau} = \begin{cases} q & \text{if } \phi_i < \tau \\ c_\tau\, \phi_i & \text{otherwise} \end{cases} \quad \text{where} \quad c_\tau = \frac{1 - \sum_{\phi_i < \tau} q}{\sum_{\phi_i \geq \tau} \phi_i}$$

Our Estimator:

$$\texttt{MLE*}(\phi) = \texttt{Inv}(\phi^{\tau^*})$$
$$\tau^* = \min\{\tau \mid \forall i,\ \phi_i < \tau \vee\ c_\tau\, \phi_i \geq q\} \ldots \text{ can be found in } O(K \log K)$$

Theorem 1. *The MLE for the RR mechanism is* ***unique*** *and given by* `MLE*`.

Proof Sketch:

1. Using Lagrange multipliers: $\hat{\theta}_i = 0$ or $\hat{\theta}_i = \frac{\phi_i}{\lambda} - \frac{q}{p-q}$
2. MLE is monotonic: $\hat{\theta}_i \leq \hat{\theta}_j \Leftrightarrow \phi_i \leq \phi_j$
3. Zeros occur in smallest components of ϕ
4. Find optimal number of zeros via threshold τ^*

Theoretical Comparison

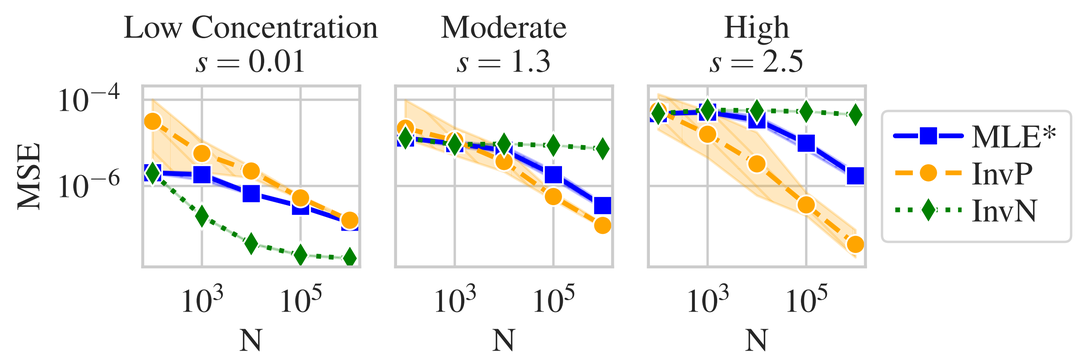
Property	`Inv`	`InvN`	`InvP`	`IBU`	`MLE*`
Valid	✗	✓	✓	✓	✓
Unbiased	✓	✗	✗	✗	✗
Is the MLE	✗	✗	✗	Only if $N_{\text{iters}} \to \infty$	✓
Complexity	$O(K)$	$O(K)$	$O(K \log K)$	$O(K N_{\text{iters}})$	$O(K \log K)$
Consistent	✓	✓	✓	✓	✓

Key Insights:

- All estimators are consistent: $\text{MSE} = O(K/N) \to 0$ as $N \to \infty$
- Valid estimators are necessarily biased (cannot be both valid and unbiased)
- `MLE*` achieves exact MLE with complexity $O(K \log K)$ vs. $O(K\, N_{\text{iters}})$ for `IBU`

Experimental Results

Setup: Varied $\epsilon \in \{1, \ldots, 10\}$, $N \in \{10^2, \ldots, 10^6\}$, $K \in \{50, \ldots, 5000\}$, Zipf skewness $s \in \{0.01, 1.3, 2.5\}$. For $\epsilon = 4.0$, $K = 10000$:



Key Finding: `MLE*` is a **robust default estimator**

- `InvP` and `InvN` alternate as best depending on distribution
- `MLE*` is never the worst estimator
- Always achieves lowest negative log-likelihood (by construction)
- Safe choice when true distribution is unknown

Real-World Data: Tested on Kosarak (clickstream) and ACSIncome (census) datasets with K up to 1.4M. `MLE*` remains consistently robust (non-worst) across all distributions.

Conclusion

- **Exact and mathematically proven formula** for the MLE of Randomized Response
- **Efficient algorithm:** $O(K \log K)$ complexity, significantly faster than iterative `IBU`
- **Robust performance:** Consistently close to best estimator across diverse data distributions
- **Practical impact:** Reliable default choice for practitioners deploying LDP systems

Future Work: Extension to other mechanisms: (1) arbitrary channel matrix and (2) longitudinal protocols (RAPPOR, Local Hashing)