Simple Binary Hypothesis Testing:
Locally Private and Communication-Efficient

Ankit Pensia

ITA 2023
Joint Work With

Amir Asadi  Varun Jog  Po-Ling Loh
Outline

- Motivation
  - Problem Statement
  - Our Results
    - Statistical
    - Computational
  - Proof Sketch
- Conclusion
• Let $p$ and $q$ be two known distributions over \{1, ..., $k$\}

Problem (Simple Hypothesis Testing):
Input: i.i.d. samples from either $p$ or $q$
Simple Hypothesis Testing: Centralized

- Let $p$ and $q$ be two known distributions over $\{1, ..., k\}$

Problem (Simple Hypothesis Testing):
- Input: i.i.d. samples from either $p$ or $q$
- Output: whether they came from $p$ or $q$
• Let $p$ and $q$ be two known distributions over $\{1, \ldots, k\}$

Problem (Simple Hypothesis Testing):
- Input: i.i.d. samples from either $p$ or $q$
- Output: whether they came from $p$ or $q$

• Arguably, the simplest statistical problem
• Let $p$ and $q$ be two known distributions over $\{1, \ldots, k\}$

Problem (Simple Hypothesis Testing):
- Input: i.i.d. samples from either $p$ or $q$
- Output: whether they came from $p$ or $q$

• Arguably, the simplest statistical problem
  • Optimal test: Likelihood ratio test
Let $p$ and $q$ be two known distributions over $\{1, \ldots, k\}$

**Problem (Simple Hypothesis Testing):**
- **Input:** i.i.d. samples from either $p$ or $q$
- **Output:** whether they came from $p$ or $q$

Arguably, the simplest statistical problem
- Optimal test: Likelihood ratio test

Requires access to $X_i$’s
Let $p$ and $q$ be two known distributions over $\{1, \ldots, k\}$

**Problem (Simple Hypothesis Testing):**
- Input: i.i.d. samples from either $p$ or $q$
- Output: whether they came from $p$ or $q$

- Arguably, the simplest statistical problem
  - Optimal test: Likelihood ratio test
- Data is distributed these days
  - Limited communication bandwidth
  - Privacy concerns

Requires access to $X_i$'s
Let $p$ and $q$ be two known distributions over $\{1, \ldots, k\}$. 

**Problem (Simple Hypothesis Testing):**
- **Input:** i.i.d. samples from either $p$ or $q$
- **Output:** whether they came from $p$ or $q$

Arguably, the simplest statistical problem
- Optimal test: Likelihood ratio test
- Data is distributed these days
- Limited communication bandwidth
- Privacy concerns

Requires access to $X_i$’s

Requires quantizing/privatizing $X_i$’s
Simple Hypothesis Testing: Decentralized

- Let $p$ and $q$ be two known distributions over $\{1, \ldots, k\}$

**Problem (Simple Hypothesis Testing):**
- Input: i.i.d. samples from either $p$ or $q$
- Output: whether they came from $p$ or $q$
Simple Hypothesis Testing: Decentralized

• Let $p$ and $q$ be two known distributions over $\{1, \ldots, k\}$

Problem (Simple Hypothesis Testing):

Input: i.i.d. samples from either $p$ or $q$
Output: whether they came from $p$ or $q$

• \(\Delta\): captures communication and/or privacy
Simple Hypothesis Testing: Decentralized

- Let $p$ and $q$ be two known distributions over $\{1, ..., k\}$

Problem (Decentralized Simple Hypothesis Testing):
- Input: modified samples from either $p$ or $q$
- Output: whether they came from $p$ or $q$

- ☑: captures communication and/or privacy
Simple Hypothesis Testing: Decentralized

- Let $p$ and $q$ be two known distributions over $\{1, \ldots, k\}$

Problem (Decentralized Simple Hypothesis Testing):
- Input: modified samples from either $p$ or $q$
- Output: whether they came from $p$ or $q$

- \(\square\): captures communication and/or privacy

Simple Hypothesis Testing: Decentralized

- Let $p$ and $q$ be two known distributions over $\{1, \ldots, k\}$

Problem (Decentralized Simple Hypothesis Testing):
- Input: modified samples from either $p$ or $q$
- Output: whether they came from $p$ or $q$

- $\text{Ξ}$: captures communication and/or privacy

How do we perform decentralized hypothesis testing?

Outline

- Motivation
- **Problem Statement**
- Our Results
  - Statistical
  - Computational
- Proof Sketch
- Conclusion
Privacy Model and Communication Constraints

- Local Differential Privacy (LDP)
  - Everyone releases a randomized version of data
  - Channel \( \epsilon \)-LDP if:
    \[
    \frac{\Pr(Y_i=y|X_i=x)}{\Pr(Y_i=y|X_i=x')} \leq e^\epsilon \quad \text{for all} \quad x, x', y
    \]
  - Non-interactive (private-coin): \( Y_i \)'s are independent

Can’t reliably distinguish between \( x \) and \( x' \) using values of \( Y_i \)
Privacy Model and Communication Constraints

- Local Differential Privacy (LDP)
  - Everyone releases a randomized version of data
  - Channel $\epsilon$-LDP if:
    \[
    \frac{\Pr(Y_i = y | X_i = x)}{\Pr(Y_i = y | X_i = x')} \leq e^\epsilon \text{ for all } x, x', y
    \]
  - Non-interactive (private-coin): $Y_i$’s are independent

- Communication-constraints
  - $Y_i \in \{1, \ldots, \ell\}$ for some $\ell \ll k$
Privacy Model and Communication Constraints

### Local Differential Privacy (LDP)
- Everyone releases a randomized version of data.
- Channel $\mathcal{C}$ is $\epsilon$-LDP if:
  \[ \frac{\Pr(Y_i = y | X_i = x)}{\Pr(Y_i = y | X_i = x')} \leq e^{\epsilon} \quad \text{for all } x, x', y \]
- Non-interactive (private-coin): $Y_i$’s are independent.

### Communication Constraints
- $Y_i \in \{1, \ldots, \ell\}$ for some $\ell \ll k$

Today’s focus: Privacy (LDP)
Questions of Interest

Problem (Decentralized Simple Hypothesis Testing):
Input: modified samples from either p or q
Output: whether they came from p or q
Questions of Interest

**Goal:** Design the test and channels so that the probability of error $\leq 0.1$.

Problem (Decentralized Simple Hypothesis Testing):
Input: modified samples from either p or q
Output: whether they came from p or q
Questions of Interest

**Goal:** Design the test and channels so that the probability of error $\leq 0.1$

**Sample Complexity:** Minimum $n$ to achieve above goal

Problem (Decentralized Simple Hypothesis Testing):
Input: modified samples from either $p$ or $q$
Output: whether they came from $p$ or $q$
Questions of Interest

**Goal:** Design the test and channels so that the probability of error $\leq 0.1$

**Sample Complexity:** Minimum $n$ to achieve above goal

$n^* :=$ Sample complexity (no constraints)

$n^*(\epsilon) :=$ Sample complexity with channels satisfying $\epsilon$-LDP

Problem (Decentralized Simple Hypothesis Testing):
- **Input:** modified samples from either p or q
- **Output:** whether they came from p or q
Questions of Interest

**Goal:** Design the test and channels so that the probability of error $\leq 0.1$

**Sample Complexity:** Minimum $n$ to achieve above goal

$n^* :=$ Sample complexity (no constraints)

$n^*(\epsilon) :=$ Sample complexity with channels satisfying $\epsilon$-LDP

**Questions:**
Questions of Interest

Goal: Design the test and channels so that the probability of error $\leq 0.1$.

Sample Complexity: Minimum $n$ to achieve above goal.

$n^*$ := Sample complexity (no constraints)

$n^*(\epsilon)$ := Sample complexity with channels satisfying $\epsilon$-LDP

Questions:

1. (Statistical) How much does sample complexity change? $n^*(\epsilon)$ vs. $n^*$.
Questions of Interest

**Goal:** Design the test and channels so that the probability of error $\leq 0.1$

**Sample Complexity:** Minimum $n$ to achieve above goal

$n^* :=$ Sample complexity (no constraints)

$n^*(\epsilon) :=$ Sample complexity with channels satisfying $\epsilon$-LDP

**Questions:**

1. (Statistical) How much does sample complexity change?

2. (Computational) How to find (near)-optimal channels fast?
Statistical Cost of Privacy: Existing Results

- Sample Complexity

Sample Complexity $n^*(\epsilon)$

$\epsilon$ (Privacy parameter)
Sample Complexity

Sample Complexity

\( n^* (\epsilon) \)

High-privacy

\( \epsilon \)

(Privacy parameter)

No privacy
Statistical Cost of Privacy: Existing Results

- Sample Complexity

\[ n^* (\epsilon) \]

\[ \frac{1}{d_h^2(p, q)} \]: sample complexity without any constraints

\[ d_h^2: \text{Hellinger divergence} \]
\[ d_{TV}: \text{Total variation distance} \]

Statistical Cost of Privacy: Existing Results

- Sample Complexity

\[ n^*(\epsilon) = \frac{1}{\epsilon^2 d_{TV}^2(p,q)} \]

\(d_{TV}^2\): Total variation distance

\(d_{h}^2\): Hellinger divergence

Statistical Cost of Privacy: Existing Results

- Sample Complexity

\[
\frac{1}{\varepsilon^2 d_{TV}^2(p, q)}
\]

Statistical Cost of Privacy: Existing Results

• Sample Complexity

\[ n^*(\epsilon) \]

*Sample Complexity without any constraints*

*Optimal sample complexity in high-privacy*

\[ \frac{1}{\epsilon^2 d_{TV}^2(p, q)} \]

\[ \frac{1}{d_{TV}^2(p, q)} \]

\[ d_{TV}^2(p, q) \]: Total variation distance

\[ d_{h}^2 \]: Hellinger divergence

---

Statistical Cost of Privacy: Existing Results

- Sample Complexity

\[
\text{Sample Complexity} \quad n^*(\epsilon) = \frac{1}{\epsilon^2 d^2_{TV}(p, q)}
\]

Optimal sample complexity in high-privacy

\[
\frac{1}{d^2_{TV}(p, q)}
\]

Existing lower bound

\[
\frac{1}{\epsilon^2 d^2_{TV}(p, q)}
\]


\( d^2_n \): Hellinger divergence

\( d_{TV} \): Total variation distance
Sample Complexity

\[ n^*(\epsilon) \]

\[ \frac{1}{\epsilon^2 d^2_{TV}(p,q)} \]

\[ \frac{1}{d^2_{TV}(p,q)} \]

\[ \frac{1}{e\epsilon d^2_{TV}(p,q)} \]

High-privacy

No privacy

\( \epsilon \)

(Privacy parameter)

Optimal sample complexity in high-privacy

[PAJL23]: Existing lower bound is tight for Bernoulli distributions


Statistical Cost of Privacy: Existing Results

- Sample Complexity

Sample Complexity \( n^*(\epsilon) \)

Optimal sample complexity in high-privacy

\[ \frac{1}{\epsilon^2 d_{TV}^2(p, q)} \]

Existing lower bound

\[ e\epsilon \frac{d_{TV}^2(p, q)}{2} \]

[PAJL23]: Existing lower bound is tight for Bernoulli distributions

What about general distributions?


Our Results: Minimax Optimal Sample Complexity

**Theorem** [PAJL23] There exist ternary distributions $p$ and $q$ with larger sample complexities.
Our Results: Minimax Optimal Sample Complexity

Theorem [PAJL23] There exist ternary distributions $p$ and $q$ with larger sample complexities.
Our Results: Minimax Optimal Sample Complexity

Theorem [PAJL23] There exist ternary distributions $p$ and $q$ with larger sample complexities.

Theorem [PAJL23] There is an efficient algorithm with nearly-matching upper bounds for all distributions.
Our Results: Minimax Optimal Sample Complexity

**Theorem** [PAJL23] There exist ternary distributions $p$ and $q$ with larger sample complexities.

**Theorem** [PAJL23] There is an efficient algorithm with nearly-matching upper bounds for all distributions.
Our Results: Minimax Optimal Sample Complexity

**Theorem[PJL23]** There exist ternary distributions $p$ and $q$ with larger sample complexities.

Overall, a satisfying story for minimax optimality

**Theorem[PJL23]** There is an efficient algorithm with nearly-matching upper bounds for all distributions.
Our Results: Minimax Optimal Sample Complexity

**Theorem** [PAJL23] There exist ternary distributions $p$ and $q$ with larger sample complexities.

Overall, a satisfying story for minimax optimality

What about instance-optimality? Are there efficient algorithms?

**Theorem** [PAJL23] There is an efficient algorithm with nearly-matching upper bounds for all distributions.
Outline

- Motivation
- Problem Statement
- Our Results
  - Statistical
  - Computational
- Proof Sketch
- Conclusion
Recall we need to map the original data $X_i \rightarrow Y_i$
Computational Cost of Privacy

- Recall we need to map the original data $X_i \rightarrow Y_i$
- Performance depends on the channel
Recall we need to map the original data $X_i \rightarrow Y_i$

Performance depends on the channel

- Once the channel is fixed, perform likelihood ratio test

Computational Cost of Privacy
Computational Cost of Privacy

- Recall we need to map the original data $X_i \rightarrow Y_i$
- Performance depends on the channel
  - Once the channel is fixed, perform likelihood ratio test
- Prior work on finding the optimal channel
Computational Cost of Privacy

• Recall we need to map the original data $X_i \rightarrow Y_i$
• Performance depends on the channel
  • Once the channel is fixed, perform likelihood ratio test
• Prior work on finding the optimal channel
  • $\epsilon \ll 1$: Well-understood
Computational Cost of Privacy

- Recall we need to map the original data $X_i \rightarrow Y_i$
- Performance depends on the channel
  - Once the channel is fixed, perform likelihood ratio test
- Prior work on finding the optimal channel
  - $\epsilon \ll 1$: Well-understood
  - $\epsilon \gg 1$: No polynomial-time algorithm
Recall we need to map the original data $X_i \rightarrow Y_i$

- Performance depends on the channel
  - Once the channel is fixed, perform likelihood ratio test

- Prior work on finding the optimal channel
  - $\epsilon \ll 1$: Well-understood
  - $\epsilon \gg 1$: No polynomial-time algorithm
    - [KOV14] gave an exponential-time algorithm

Computational Cost of Privacy

• Recall we need to map the original data $X_i \rightarrow Y_i$

• Performance depends on the channel
  • Once the channel is fixed, perform likelihood ratio test

• Prior work on finding the optimal channel
  • $\epsilon \ll 1$: Well-understood
  • $\epsilon \gg 1$: No polynomial-time algorithm
    • [KOV14] gave an exponential-time algorithm

Can we efficiently find the (near)-optimal channel?

Our Results: Computational Cost of Privacy

**Theorem [PAJL23]** Given any two distributions $p$ and $q$ on $[k]$ and $\epsilon$, ...
Our Results: Computational Cost of Privacy

**Theorem** [PAJL23] Given any two distributions $p$ and $q$ on $[k]$ and $\epsilon$, there is a **linear-time algorithm** to find an $\epsilon$-LDP channel.
Our Results: Computational Cost of Privacy

**Theorem [PAJL23]** Given any two distributions $p$ and $q$ on $[k]$ and $\epsilon$, there is a **linear-time algorithm** to find an $\epsilon$-LDP channel whose sample complexity is **near-optimal**.
Our Results: Computational Cost of Privacy

**Theorem (PAJL23)** Given any two distributions $p$ and $q$ on $[k]$ and $\epsilon$, there is a **linear-time algorithm** to find an $\epsilon$-LDP channel whose sample complexity is **near-optimal**.

- More broadly, consider the optimization problem
Our Results: Computational Cost of Privacy

**Theorem** [PAJL23] Given any two distributions $p$ and $q$ on $[k]$ and $\epsilon$, there is a **linear-time algorithm** to find an $\epsilon$-LDP channel whose sample complexity is **near-optimal**.

• More broadly, consider the optimization problem

$$g(p, q)$$

$g$: a (quasi)-convex objective
Our Results: Computational Cost of Privacy

**Theorem** [PAJL23] Given any two distributions $p$ and $q$ on $[k]$ and $\epsilon$, there is a **linear-time algorithm** to find an $\epsilon$-LDP channel whose sample complexity is **near-optimal**.

- More broadly, consider the optimization problem

$$
\max_{\epsilon \in \mathcal{P}(\epsilon, \ell)} g(p, q)
$$

$\mathcal{P}(\epsilon, \ell)$: All $\epsilon$-LDP channels of output size $\ell$  

$g$: a (quasi)-convex objective
Our Results: Computational Cost of Privacy

**Theorem [PAJL23]** Given any two distributions $p$ and $q$ on $[k]$ and $\epsilon$, there is a **linear-time algorithm** to find an $\epsilon$-LDP channel whose sample complexity is **near-optimal**.

- More broadly, consider the optimization problem

$$\max_{\epsilon \in \mathcal{P}(\epsilon, \ell)} g(p, q)$$

$\mathcal{P}(\epsilon, \ell)$: All $\epsilon$-LDP channels of output size $\ell$

$g$: a (quasi)-convex objective

Recall: maximizing a convex objective is usually hard!
**Our Results: Computational Cost of Privacy**

**Theorem [PAJL23]** Given any two distributions $p$ and $q$ on $[k]$ and $\epsilon$, there is a **linear-time algorithm** to find an $\epsilon$-LDP channel whose sample complexity is **near-optimal**.

- More broadly, consider the optimization problem

  $\max_{\mathcal{P}(\epsilon, \ell)} g(p, q)$

  $\mathcal{P}(\epsilon, \ell)$: All $\epsilon$-LDP channels of output size $\ell$

  $g$: a (quasi)-convex objective

Recall: maximizing a convex objective is usually hard!

**Theorem [PAJL23]** There is a $\text{poly}(k^{2\ell^2})$-time algorithm to find the optimum.
Outline

- Motivation
- Problem Statement
- Our Results
  - Statistical
  - Computational
  - **Proof Sketch**
- Conclusion
Proof Sketch: Exponential Search to Linear

• Say, we want to find the optimal binary channel $T^*$

$$\max_{T \in \mathcal{P}(\epsilon, 2)} g(T_p, T_q)$$
Say, we want to find the optimal binary channel $\mathbf{T}^*$

Can show that optimal $\mathbf{T}^*$ is of the form:

Proof Sketch: Exponential Search to Linear

$$\max_{\mathbf{T} \in \mathcal{P}(\epsilon, 2)} g(\mathbf{T}_p, \mathbf{T}_q)$$
• Say, we want to find the optimal binary channel $\mathbf{T}^*$
• Can show that optimal $\mathbf{T}^*$ is of the form:
  • First, use a binary deterministic channel $\mathbf{T}'$ to partition $[k]$ into two sets

Proof Sketch: Exponential Search to Linear

$$\max_{\mathbf{T} \in \mathcal{P}(\epsilon,2)} g(\mathbf{T}_p, \mathbf{T}_q)$$
Proof Sketch: Exponential Search to Linear

- Say, we want to find the optimal binary channel $T^*$
- Can show that optimal $T^*$ is of the form:
  - First, use a binary deterministic channel $T'$ to partition $[k]$ into two sets
  - Ensure privacy using the randomized response channel (BSC)
• Say, we want to find the optimal binary channel $T^*$
• Can show that optimal $T^*$ is of the form:
  • First, use a binary deterministic channel $T'$ to partition $[k]$ into two sets
  • Ensure privacy using the randomized response channel (BSC)
• But the number of possible partitions: $2^k$
Proof Sketch: Exponential Search to Linear

• Say, we want to find the optimal binary channel $T^*$
• Can show that optimal $T^*$ is of the form:
  • First, use a binary deterministic channel $T'$ to partition $[k]$ into two sets
  • Ensure privacy using the randomized response channel (BSC)
• But the number of possible partitions: $2^k$
• Can we use $p$ and $q$ to reduce our search space?
Proof Sketch: Exponential Search to Linear

• Say, we want to find the optimal binary channel $\mathbf{T}^*$
• Can show that optimal $\mathbf{T}^*$ is of the form:
  • First, use a binary deterministic channel $\mathbf{T}'$ to partition $[k]$ into two sets
  • Ensure privacy using the randomized response channel (BSC)
• But the number of possible partitions: $2^k$
• Can we use $p$ and $q$ to reduce our search space?
• Our answer: yes!
Say, we want to find the optimal binary channel $T^*$
Can show that optimal $T^*$ is of the form:
  • First, use a binary deterministic channel $T'$ to partition $[k]$ into two sets
  • Ensure privacy using the randomized response channel (BSC)
But the number of possible partitions: $2^k$
Can we use $p$ and $q$ to reduce our search space?
Our answer: yes!
  • Optimal partition must respect the likelihood ratios of $p$ and $q
Outline

- Motivation
- Problem Statement
- Our Results
  - Statistical
  - Computational
- Proof Sketch
- Conclusion
Conclusion and Future Directions

• Derived minmax-optimal sample complexities under privacy
  • No longer depends only on TV distance and Hellinger
Conclusion and Future Directions

• Derived minmax-optimal sample complexities under privacy
  • No longer depends only on TV distance and Hellinger
• Computationally and Communication-efficient algorithms
Conclusion and Future Directions

• Derived minmax-optimal sample complexities under privacy
  • No longer depends only on TV distance and Hellinger
• Computationally and Communication-efficient algorithms

• Open problems:
Conclusion and Future Directions

• Derived minmax-optimal sample complexities under privacy
  • No longer depends only on TV distance and Hellinger

• Computationally and Communication-efficient algorithms

• Open problems:
  • Role of interactivity
Conclusion and Future Directions

• Derived minmax-optimal sample complexities under privacy
  • No longer depends only on TV distance and Hellinger
• Computationally and Communication-efficient algorithms

• Open problems:
  • Role of interactivity
  • Characterization of instance-optimal sample complexity
    • Looking beyond TV distance and Hellinger divergence
Conclusion and Future Directions

- Derived minmax-optimal sample complexities under privacy
  - No longer depends only on TV distance and Hellinger
- Computationally and Communication-efficient algorithms

- Open problems:
  - Role of interactivity
  - Characterization of instance-optimal sample complexity
    - Looking beyond TV distance and Hellinger divergence
  - M-ary hypothesis testing, optimally
Conclusion and Future Directions

• Derived minmax-optimal sample complexities under privacy
  • No longer depends only on TV distance and Hellinger
• Computationally and Communication-efficient algorithms

• Open problems:
  • Role of interactivity
  • Characterization of instance-optimal sample complexity
    • Looking beyond TV distance and Hellinger divergence
  • M-ary hypothesis testing, optimally

Thank you!