Hadamard Response: Local Private Distribution Estimation

Jayadev Acharya, Ziteng Sun, Huanyu Zhang
Cornell University
Information Theory and Applications (ITA), 2018

Based on:

Distribution Estimation

• $p$: unknown discrete distribution over $k$ elements
• $\alpha$: accuracy
• Input: independent samples $X_1, X_2, ..., X_n$ from $p$
• Output: $\hat{p}$ such that w.p. at least 0.9:

$$d(p, \hat{p}) \leq \alpha$$

• We consider $\ell_1, \ell_2$ distances
Sample Complexity

Sample Complexity: Least $n$ to estimate $p$

To estimate to $\ell_1 \leq \alpha$:

$$\Theta \left( \frac{k}{\alpha^2} \right)$$

Empirical distribution works
Distribution Estimation with Privacy

• Samples are sensitive

• Drug abuse
  • Learn underlying drug usage behavior (for policy design)
  • Maintain privacy of users

• Internet
  • Distribution of web traffic to websites
  • Maintain browsing of a particular user private
Model

- $X_1 \ldots X_n$ stored over $n$ users
- User $i$ transmits $Z_i$ to data collector/server
- Server has to learn $p$
- Without privacy: send $X_i$
Local Differential Privacy (LDP)

- $Q$: a channel with input $[k]$ and output $\mathcal{Z}$

$\varepsilon$-LDP [DuchiWainwrightJordan’12, ErlingssonPihurKorolova’14]:

$$\frac{Q(z|x)}{Q(z|x')} \leq e^{\varepsilon}$$

User $i$ passes $X_i$ through $Q$, send output $Z_i$
Randomized Response (RR)

[Warner’65, KairouzBonawitzRamadge’14]: \( Z = [k] \)

\[
Q_\varepsilon(z|x) = \begin{cases} 
\frac{e^\varepsilon}{e^\varepsilon + k - 1}, & z = x \\
\frac{1}{e^\varepsilon + k - 1}, & z \neq x
\end{cases}
\]

Optimal only in the low privacy regime \( (\varepsilon > \log k) \)
RAPPOR

[DuchiWainwrightJordan’12, ErlingssonPihurKorolova’14]: $\mathcal{Z} = \{0,1\}^k$.

- One hot encoding: $x \rightarrow e_x$ (basis vector with $x$th entry 1)
- Flip each entry in $e_x$ with probability $\frac{1}{e^{\varepsilon/2} + 1}$

$e_x, e_x'$ differ in at most two positions

- Optimal only for $\varepsilon \leq 1$, and $\varepsilon > 2 \log k$
Subset Selection (SS)

[WangHuangWangNieXuYangLiQiao’16, YeBarg’17]:

\[ \mathcal{Z}: \text{strings in } \{0,1\}^k \text{ with Hamming weight } \frac{k}{e^{\varepsilon+1}} \]

Optimal in all regimes
Sample Complexity

<table>
<thead>
<tr>
<th>$\varepsilon$</th>
<th>RR</th>
<th>RAPPOR</th>
<th>SS</th>
<th>HR</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0,1)</td>
<td>$\frac{k^3}{\varepsilon^2 \alpha^2}$</td>
<td>$\frac{k^2}{\varepsilon^2 \alpha^2}$</td>
<td>$\frac{k^2}{\varepsilon^2 \alpha^2}$</td>
<td>$\frac{k^2}{\varepsilon^2 \alpha^2}$</td>
</tr>
<tr>
<td>(1, $\log k$)</td>
<td>$\frac{k^3}{e^{2\varepsilon} \alpha^2}$</td>
<td>$\frac{k^2}{e^{\varepsilon/2} \alpha^2}$</td>
<td>$\frac{k^2}{e^{\varepsilon} \alpha^2}$</td>
<td>$\frac{k^2}{e^{\varepsilon} \alpha^2}$</td>
</tr>
</tbody>
</table>

For constant $\varepsilon$, say $\varepsilon = 1$,

$$\frac{k}{\alpha^2} \rightarrow \frac{k^2}{\alpha^2}$$
Other Resources

Computational Complexity:

What is the encoding/decoding time?
Impractical if high running time, even if sample optimal

Communication Complexity:

How much communication to server?

Many papers considering these resources, including today on both!
## Resources for $\varepsilon \in (0, 1)$

<table>
<thead>
<tr>
<th></th>
<th>RR</th>
<th>RAPPOR</th>
<th>SS</th>
<th>HR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Communication</td>
<td>$\log k$</td>
<td>$k$</td>
<td>$k$</td>
<td>$\log k$</td>
</tr>
<tr>
<td>Decoding time</td>
<td>$n$</td>
<td>$n \cdot k$</td>
<td>$n \cdot k$</td>
<td>$n$</td>
</tr>
<tr>
<td>Samples</td>
<td>$\frac{k^3}{\varepsilon^2 \alpha^2}$</td>
<td>$\frac{k^2}{\varepsilon^2 \alpha^2}$</td>
<td>$\frac{k^2}{\varepsilon^2 \alpha^2}$</td>
<td>$\frac{k^2}{\varepsilon^2 \alpha^2}$</td>
</tr>
</tbody>
</table>

**How to claim bounds on time and communication?**

**Faithful implementation:**

- Communication $\geq H(Z)$ bits.
- Decoding Time $\geq n \cdot H(Z)$
Communication requirements

<table>
<thead>
<tr>
<th></th>
<th>RR</th>
<th>RAPPOR</th>
<th>SS</th>
<th>HR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Communication</td>
<td>( \log k )</td>
<td>( \log k + \frac{k}{e^{\varepsilon/2}} )</td>
<td>( \log k + \frac{k}{e^{\varepsilon}} )</td>
<td>( \log k )</td>
</tr>
</tbody>
</table>

All these are entropy bounds!!
Other Resources

Large domain:
- Browsing patterns of internet users
- Distribution of product purchases of Target

Communication:
- Handheld devices with low uplink capacity
- Low battery power, 4G data
General encoding matrices

$M$: $\pm 1$ matrix of size $k \times K$

$h$: #1’s in each row

$$Q_\varepsilon(z|x) = \begin{cases} \frac{e^\varepsilon}{e^\varepsilon + K - h}, & M(x, z) = +1 \\ \frac{1}{e^\varepsilon + K - h}, & M(x, z) = -1 \end{cases}$$

$$\begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & \text{OR} & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \end{pmatrix}$$
Hadamard Matrix

$H_m$: $m \times m$ matrix

$H_1 = [1]$, and for other $m$:

\[
H_m = \begin{bmatrix}
H_{m/2} & H_{m/2} \\
H_{m/2} & -H_{m/2}
\end{bmatrix}.
\]

- The first row & column has $m$ ‘1’s
- Every other row & column has $\frac{m}{2}$ ‘1’s
- Hamming distance between any two rows is $\frac{m}{2}$
- Matrix vector multiplication real fast!
(b, B)-Hadamard Matrix

b, B: powers of 2, and K = b \cdot B

\[
H^K_b = \begin{pmatrix}
H_b & P_b & \cdots & P_b \\
P_b & H_b & \cdots & P_b \\
& & \ddots & \ddots \\
P_b & P_b & \cdots & H_b
\end{pmatrix}
\]

P_b: b \times b matrix with all entries ‘-1’

B = 1, H^K_b = H_b

b = 1, H^K_b = Identity matrix
Encoding Matrix

Rows of $H^b_K$ have different number of 1’s

- Delete the first row of each embedded $H_b$
- The first $k$ rows is the encoding matrix $M$
Selecting the parameters

\( B \): largest power of 2 less than \( \min\{ e^\epsilon, 2k \} \)

\( b \): smallest power of 2 larger than \( \left\lceil \frac{k}{B} \right\rceil \)

\[
K = B \cdot b \leq 4k
\]

Communication: \( \log K \leq \log k + 2 \) bits.
Key arguments

Large Hamming distance -> Sample Optimality
Fast Hadamard Transform -> Fast Decoding
L2 error plots ($k = 1000$, Geo(0.8))

(a) $\varepsilon = 0.5$

(b) $\varepsilon = 2$
L2 error plots ($k = 1000$, Geo(0.8))

(c) $\varepsilon = 5$

(d) $\varepsilon = 7$
Running time Geo(0.8)

(a) $k = 100$

(b) $k = 1000$
Running time Geo(0.8)

(c) $k = 5000$

(d) $k = 10000$
Thank You

Details in the paper online!