Communication Complexity in Locally Private Distribution Estimation and Heavy Hitters

ICML 2019, Long Beach

June 11th, 2019

Jayadev Acharya, Cornell University

Ziteng Sun, Cornell University

Distribution Learning

- $[k] = \{0, 1, 2, ..., k-1\}$, a discrete set of size k.
- p : an **unknown** distribution over $[k]$.
- n users, user i has an independent $X_i \sim p$.
- Estimator $\hat{p} : [k]^n \to$ a distribution over $[k]$.

Goal: For all p, with probability at least 2/3

$$\ell_1(\hat{p}, p) = \sum_{x \in [k]} |\hat{p}(x) - p(x)| \leq \alpha.$$

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

Frequency/ Heavy Hitter Estimation

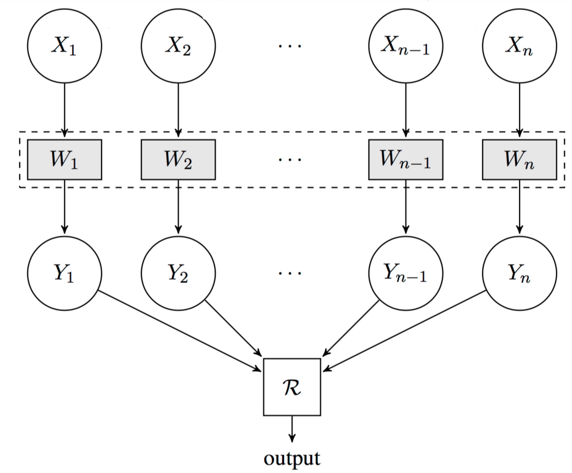
- $[k] = \{0, 1, 2, ..., k-1\}$ is a discrete set of size k.
- n users, user i has a data point $X_i \in [k]$.
- No distribution assumption.
- $\forall x \in [k], N_x = \sum_i \mathbf{1}\{X_i = x\}$.

Goal: For all X^n, with probability at least 2/3

$$\ell_\infty(\hat{p}, p) = \max_{x \in [k]} \left| \hat{p}(x) - \frac{N_x}{n} \right| \leq \beta.$$

Simultaneous Message Passing (SMP) Protocal

Each user sends a message $Y_i = W_i(X_i) \in \mathcal{Y}$



Resources to Consider

- **Privacy.** Data may contain sensitive information.
- **Communication.** How many bits are communicated from each user?
- **Shared Randomness.** Is **shared randomness** available among users?
- **Symmetry.** Are the channels symmetric?

Local Differential Privacy (LDP)

[Warner, 1965, Dwork et al., 2006, Kasiviswanathan et al., 2011, Erlingsson et al., 2014]

W is *ε-LDP* if for all $x, x' \in \mathcal{X}$, and $y \in \mathcal{Y}$,

$$\sup_{y \in \mathcal{Y}} \frac{W(y|x)}{W(y|x')} \leq e^{\varepsilon}.$$

We will focus on the case of high privacy. ($\varepsilon = O(1)$)

Private and Shared Randomness

Private-coin protocols:

$U_1, U_2, ..., U_n$ independent

W_i is decided by U_i.

Public-coin protocols:

U: random bits generated at $\mathcal{R}$, available to all players.

W_i : determined by U.

0.5 round of interaction.

Symmetric, Private-coin Schemes

Distribution Learning

Theorem

[Acharya et al., 2019] Hadamard Response, which is a symmetric scheme without shared randomness, achieves the following sample complexity with only $\log k$ *bits of communication from each user:*

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

Heavy Hitter Estimation Algorithms

[Bassily and Smith, 2015, Bassily et al., 2017, Hsu et al., 2012, Wang and Blocki, 2017, Bun et al., 2018, Zhu et al., 2019] : Finding the **heavy hitters** under LDP constraints. Sample complexity:

$$n = \Theta\left(\frac{\log k}{\alpha^2 \varepsilon^2}\right)$$

Require **interaction** or **shared randomness**.

Optimality of HR for Heavy Hitter Estimation

Theorem

[Acharya and Sun, 2019] To estimate each of the frequencies up to ℓ_∞ accuracy α, HR uses

$$n = O\left(\frac{\log k}{\alpha^2 \varepsilon^2}\right).$$

samples.

Communication Lower Bound for Symmetric Schemes

Theorem

[Acharya and Sun, 2019] Without shared randomness, any optimal symmetric schemes for distribution learning/ frequency estimation must require at least $\log k$ *bits of communication.*

Communication Lower Bound for Symmetric Schemes

Theorem

[Acharya and Sun, 2019] Without shared randomness, any optimal symmetric schemes for distribution learning/ frequency estimation must require at least $\log k$ *bits of communication.*

Question: What if we allow asymmetric schemes, or schemes with shared randomness?

One-bit Suffices for Schemes with Shared-Randomness

Theorem

[Bassily and Smith, 2015] In the regime where $\varepsilon = O(1)$, for any locally private algorithm, using **shared-randomness***, there exists a locally private scheme with only one-bit communication which has the same privacy guarantee and the same performance, up to constant factors.*

One-bit Suffices for Schemes with Shared-Randomness

Theorem

[Bassily and Smith, 2015] In the regime where $\varepsilon = O(1)$, for any locally private algorithm, using **shared-randomness***, there exists a locally private scheme with only one-bit communication which has the same privacy guarantee and the same performance, up to constant factors.*

Question: Is **shared-randomness** necessary to reduce communication from users?

Optimal One-bit Scheme without Shared Randomness

For distribution learning,

NO!

Theorem

[Acharya and Sun, 2019] There exists a private-coin scheme with only one bit communication from each user that achieve optimal performance for distribution learning.

One Bit is not Enough for Heavy Hitter Estimation

For heavy hitter estimation,

YES!

Theorem

[Acharya and Sun, 2019] Any optimal private-coin schemes for frequency estimation must require at least $\min\{\log k, \log n\}$ *bits of communication.*

Summary of Results

Randomness \ Communication	$O(1)$ bits	$O(\log k)$ bits
Symmetric, Private Randomness	∞ (Acharya & Sun, 2019)	$\Theta\left(\frac{k^2}{\alpha^2\varepsilon^2}\right)$ (Acharya et al., 2019)
Private Randomness	$\Theta\left(\frac{k^2}{\alpha^2\varepsilon^2}\right)$ (Acharya & Sun, 2019)	$\Theta\left(\frac{k^2}{\alpha^2\varepsilon^2}\right)$
Public Randomness	$\Theta\left(\frac{k^2}{\alpha^2\varepsilon^2}\right)$	$\Theta\left(\frac{k^2}{\alpha^2\varepsilon^2}\right)$

Table 3. Sample Complexity for distribution learning under different communication budget and available randomness.

Randomness \ Communication	$O(1)$ bits	$O(\log k)$ bits
Symmetric, Private Randomness	∞	$\Theta\left(\frac{\log k}{\alpha^2\varepsilon^2}\right)$ (Acharya & Sun, 2019)
Private Randomness	∞ (Acharya & Sun, 2019)	$\Theta\left(\frac{\log k}{\alpha^2\varepsilon^2}\right)$
Public Randomness	$\Theta\left(\frac{\log k}{\alpha^2\varepsilon^2}\right)$ (Bassily & Smith, 2015)	$\Theta\left(\frac{\log k}{\alpha^2\varepsilon^2}\right)$

Table 4. Sample Complexity for frequency estimation under different communication budget and available randomness.

The End

Paper available on arXiv:

https://arxiv.org/abs/1905.11888.

06:30 – 09:00 PM, Pacific Ballroom

#177

Acharya, J. and Sun, Z. (2019).
Communication complexity in locally private distribution estimation and heavy hitters.
In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 51–60, Long Beach, California, USA. PMLR.

Acharya, J., Sun, Z., and Zhang, H. (2019).
Hadamard response: Estimating distributions privately, efficiently, and with little communication.
In Chaudhuri, K. and Sugiyama, M., editors, *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pages 1120–1129. PMLR.

Bassily, R., Nissim, K., Stemmer, U., and Thakurta, A. G. (2017).

Practical locally private heavy hitters.
In *Advances in Neural Information Processing Systems*, pages 2285–2293.

Bassily, R. and Smith, A. (2015).
Local, private, efficient protocols for succinct histograms.
In *STOC*, pages 127–135. ACM.

Bun, M., Nelson, J., and Stemmer, U. (2018).
Heavy hitters and the structure of local privacy.
In *Proceedings of the 35th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, pages 435–447. ACM.

Dwork, C., Mcsherry, F., Nissim, K., and Smith, A. (2006).
Calibrating noise to sensitivity in private data analysis.

In *In Proceedings of the 3rd Theory of Cryptography Conference*.

Erlingsson, Ú., Pihur, V., and Korolova, A. (2014).
Rappor: Randomized aggregatable privacy-preserving ordinal response.
In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pages 1054–1067. ACM.

Hsu, J., Khanna, S., and Roth, A. (2012).
Distributed private heavy hitters.
In *International Colloquium on Automata, Languages, and Programming*, pages 461–472. Springer.

Kasiviswanathan, S. P., Lee, H. K., Nissim, K., Raskhodnikova, S., and Smith, A. (2011).
What can we learn privately?

SIAM Journal on Computing, 40(3):793–826.

Wang, T. and Blocki, J. (2017).
Locally differentially private protocols for frequency estimation.
In *Proceedings of the 26th USENIX Security Symposium*.

Warner, S. L. (1965).
Randomized response: A survey technique for eliminating evasive answer bias.
Journal of the American Statistical Association, 60(309):63–69.

Zhu, W., Kairouz, P., Sun, H., McMahan, B., and Li, W. (2019).
Federated heavy hitters discovery with differential privacy.
arXiv preprint arXiv:1902.08534.