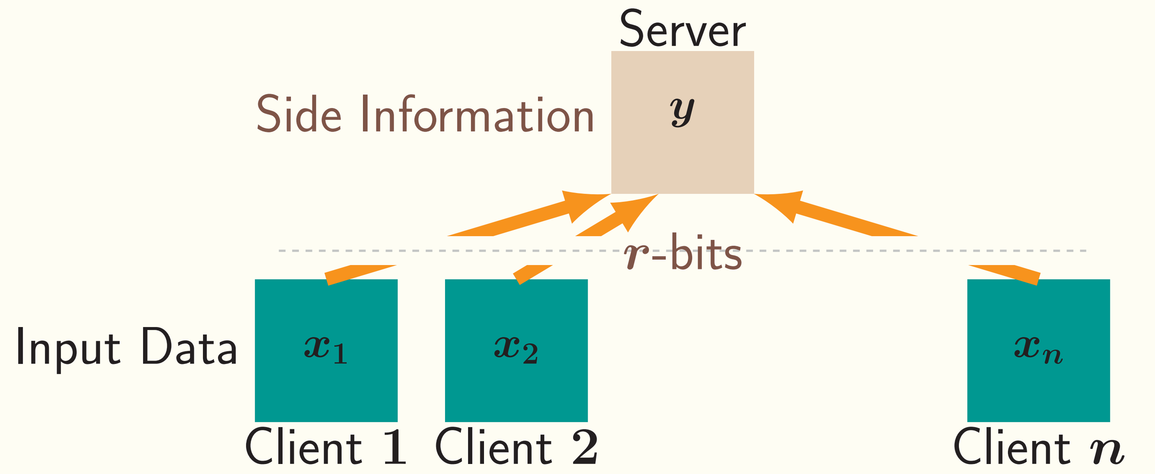
Wyner-Ziv Estimators: Efficient Distributed Mean Estimation with Side Information

Prathamesh Mayekar, Indian Institute of Science
Ananda Theertha Suresh, Google
Himanshu Tyagi, Indian Institute of Science

Setup



Assumptions: $\|x_i - y\|_2 \leq \Delta$, for all $i \in \{1, \ldots, n\}$.

Server's Goal: **Estimate Sample Mean** $\bar{x}$: $= \frac{1}{n}\sum_{i=1}^{n} x_i$.

Each client can send only $r (\leq d)$ bits.

Two settings:

1. The known setting, where Δ is known to everyone;
2. The unknown setting, where Δ is unknown to everyone.

Application: Important subroutine in several distributed learning scenarios (e.g. gradient aggregation in Federated Learning).

Prior Work

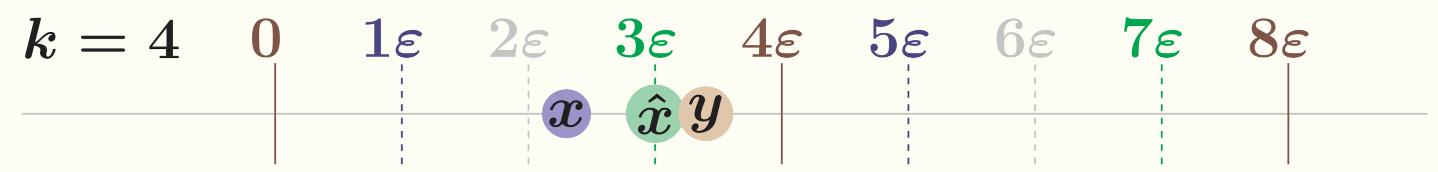
1. The no side information case [1]:
   - $\|x_i\|_2 \leq 1$, for all $i \in [n]$, and no side information.
   - For any $r \in [d]$, $MSE \approx \Theta\left(\frac{d}{nr}\right)$.
2. The known setting [2]:
   - Focuses on the high precision regime of $r \geq d$.
   - Supoptimal results in the low precision regime ($r \leq d$).
   - Algorithm is computationally expensive.

Our Contributions

1. In the known Δ setting, $MSE \approx \Theta\left(\Delta^2 \cdot \frac{d}{nr}\right)$.
   - Our results hold for x_i, $i \in [n]$, and y lying anywhere in $\mathbb{R}^d$.
2. In the unknown Δ setting, $MSE \approx O\left(\Delta \cdot \frac{d}{nr}\right)$.
   - Our results hold for x_i, $i \in [n]$, and y lying anywhere in the unit Euclidean ball.
3. Our algorithms are nearly linear time.

Building block in the known setting: Modulo Quantizer

Modulo Quantizer [2], [3]: Parameters k & ε.
Scalar input x and side-information y such that $|x - y| \leq \Delta$.



Encoder

1. $z_u = \lceil x/\varepsilon \rceil$, $z_l = \lfloor x/\varepsilon \rfloor$.
2. $\tilde{z} = \begin{cases} z_u, & w.p.\ x/\varepsilon - z_l \\ z_l, & w.p.\ z_u - x/\varepsilon. \end{cases}$
3. **Output:** $\tilde{z}$ mod k

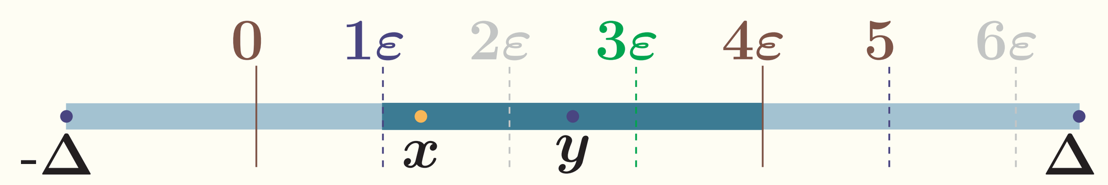
Decoder

1. Input $w \in \{0, \ldots, k-1\}$.
2. **Output:** $\hat{x}$ = Point closest to y in $\{(zk + w) \cdot \varepsilon : z \in \mathbb{Z}\}$

Suppose $k \cdot \varepsilon \gtrsim \Delta$. Then, $\mathbb{E}[\hat{x}] = x$ and $|\hat{x} - x| \leq \varepsilon$.

RMQ: Modulo Quantizer +Random Rotation

Input and Side Information: x and y such that $\|x - y\|_2 \leq \Delta$.



1. Rotate x and y using randomized Hadamard transform.
   - Each coordinate of $x - y$ is subgaussian with a variance factor $\frac{\Delta^2}{d}$.
2. For each coordinate, use Modulo Quantizer.
3. Bias-MSE Tradeoff:

3.1 Need the grid size ε to be small for a smaller MSE.
3.2 Error Event $|x - y| \gtrsim k\varepsilon$ induces bias.
3.3 Optimize over k and ε to minimize overall MSE.

Putting it all together

- Wyner-Ziv Estimator in the known setting: For each client i
  1. Sample $\approx r$ coordinates using public randomness (between client and server).
  2. Send encoded values of RMQ for those coordinates.
- Sample mean estimator: Average of the decoded estimates.
- Leads to the first main result.

Key Idea in Unknown Setting: Correlated Sampling Idea [4]

Let $x, y \in [0, 1]$ and $U \sim \mathtt{Unif}[0, 1]$.
Two different 1-bit estimators of x:

1. $\mathbb{1}_{\{U \leq x\}}$.
   - $\mathbb{E}\left[\mathbb{1}_{\{U \leq x\}}\right] = x$.
   - $\mathtt{Var}(\mathbb{1}_{\{U \leq x\}}) = x - x^2$.
2. $\hat{X} = \mathbb{1}_{\{U \leq x\}} - \mathbb{1}_{\{U \leq y\}} + y$.
   - $\mathbb{E}\left[\hat{X}\right] = x$.
   - $\mathtt{Var}(\hat{X}) = |x - y| - (x - y)^2$.

Possibility of distance-dependent bounds without its knowledge!

RDAQ: Correlated sampling with multpe scales + Random Rot.

Input and Side Information: x and y s.t. $\max\{\|x\|_2, \|y\|_2\} \leq \Delta$.

1. Rotate x and y using randomized Hadamard transform.
2. Correlated sampling + Tetration idea of RATQ [5].
   - $M_{i+1}^2 \approx e^{M_i^2} (tetration)$.
   - Use indep rvs $\{U(i)\}_{i \in [h]}$, where $U(i) \sim \mathtt{Unif}[-M_i, M_i]$.
   - $\hat{X}_i = 2M_i \left(\mathbb{1}_{\{U(i) \leq x(i)\}} - \mathbb{1}_{\{U(i) \leq y(i)\}}\right) + y$.
   - Use the smallest interval containing x and y.

Subsampled version of RDAQ gives the second main result.

References

1. Suresh, A. T., Felix, X. Y., Kumar, S., & McMahan, H. B. (2017, July). Distributed mean estimation with limited communication. In International Conference on Machine Learning (pp. 3329-3337). PMLR.
2. Davies, P., Gurunathan, V., Moshrefi, N., Ashkboos, S., & Alistarh, D. (2020). Distributed Variance Reduction with Optimal Communication. arXiv preprint arXiv:2002.09268.
3. Forney, G. D. (1988). Coset codes. I. Introduction and geometrical classification. IEEE Transactions on Information Theory, 34(5), 1123-1151.
4. Holenstein, T. (2007, June). Parallel repetition: simplifications and the no-signaling case. In Proceedings of the thirty-ninth annual ACM symposium on Theory of computing (pp. 411-419).
5. Mayekar, P., & Tyagi, H. (2020, June). RATQ: A universal fixed-length quantizer for stochastic optimization. In International Conference on Artificial Intelligence and Statistics (pp. 1399-1409). PMLR.