Practical Interactive Scheme for Extremum Computation in Distributed Networks

Solmaz Torabi

Dept. of Electrical and Computer Engineering
Drexel University

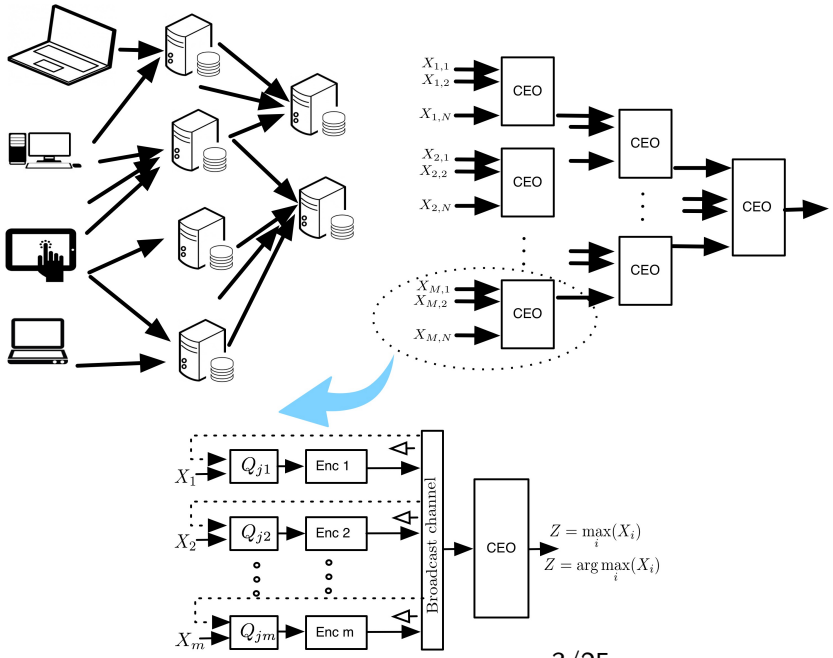
Joint work with:
Jie Ren, Dr. John M. Walsh

Outline

- Overview and problem formulation
- Computing the fundamental limit
- Designing optimal dynamic program
- Experimental results

Overview

- Fundamental limits of interprocessor communication for massively parallel computing.
- Designs to approach the limits.
- Computing a simple function, when communicating interactively over a network.



Problem setup

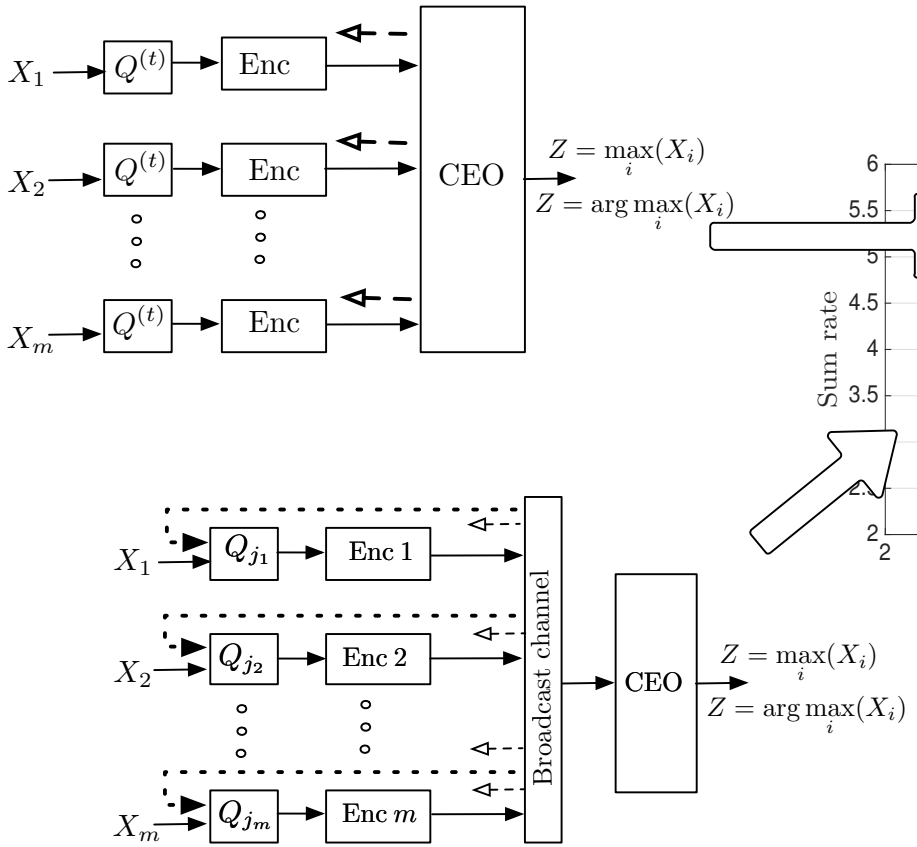
- Each source terminal $j \in [m]$ observes a sequence of discrete random variables $\boldsymbol{X}_j^n = (X_j(1), ..., X_j(n)) \in \mathcal{X}_j^n \Rightarrow$ assumption: iid
- Assumption $\Rightarrow$ each $X_j(i)$ is independent and identically distributed (in both $j \in [m]$ and $i \in [n]$)
- wlog $\Rightarrow$ Assume uniform discrete distribution with support set $\mathcal{X} = \{1, ..., L\}$.
- The sink terminal wishes to compute an extremum function $f : \mathcal{X}_1 \times ... \times \mathcal{X}_m \to \mathcal{Z}$ elementwise.
- Obtain the the sequence $\boldsymbol{Z}^n = (Z(1), ..., Z(n))$
- Elementwise $\Rightarrow Z(i) = f(X_1(i), X_2(i), ..., X_m(i))$

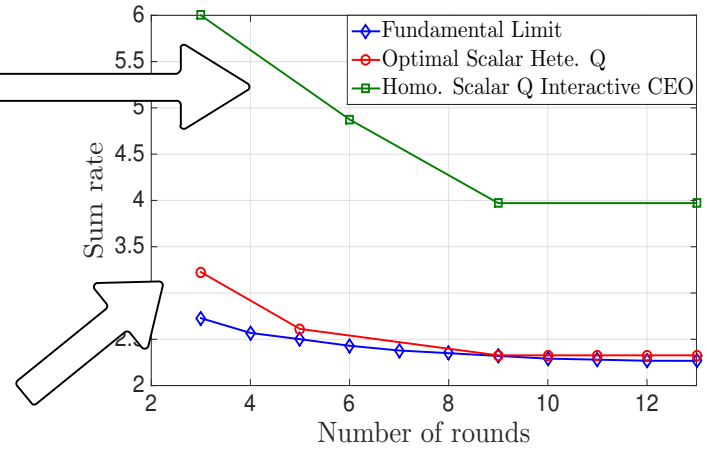
Problem setup

- The function f under study will either specify the
  - Max $\Rightarrow Z(i) = f_M(X_{1:m}(i)) = \max_{j\in[m]} X_j(i)$
  - Argmax $\Rightarrow Z(i) = f_A(X_{1:m}(i))$ such that $X_{f_A(X_{1:m}(i))} = f_M(X_{1:m}(i))$,
- To enable the sink to determine one of these functions, the sources interact with each other and also with the sink in an interactive communication model
- An optimal dynamic program achieving the minimum expected rate and expected rate delay tradeoff over all scalar quantization schemes is presented

Interaction Model

- Relatively simple interactive scheme built from several rounds of scalar quantization followed by Huffman coding in [1] [1].





[1] Boyle, B., Ren, J., Walsh, J., Weber, S. (2015). Interactive Scalar Quantization for Distributed Resource Allocation.

Problem setup

- a t-message distributed source code consists of block encoding functions $\psi_i : \mathcal{X}_j^n \times \bigotimes_{l=1}^{i-1} \mathcal{M}_l \to \mathcal{M}_i$ for $i \in [t]$ with $j = ((i-1) \bmod m) + 1$,
- a decoding function $\phi : \bigotimes_{l=1}^{t} \mathcal{M}_l \to \mathcal{Z}$
- For each t, $R_t = (1/n) \log |\mathcal{M}_t|$ is called the t-th bock coding rate and the associated sum rate is $R_{sum,t} = \sum_{i=1}^{t} R_i$.
- This minimum sum rate for all finite t is the solution to the optimization $R_{sum,t} = \min_{p_{U^t|X^m} \in \mathcal{C}} I(X^m; U^t)$
- $\mathcal{C}$ requires the message random variables $U_i, i \leq t$ to obey the Markov conditions, and $\boldsymbol{U}^t$ determine the function $f(X_1, \ldots, X_m)$.

Outline

- Overview and problem formulation
- Computing the fundamental limit
- Designing optimal dynamic program
- Experimental results

Minimum Sum-rate for Computing Extrema

- The rate reduction functional is defined as [2]

$$\rho_t = H(X^m) - R_{sum,t} = \max_{p_{u^t|X^m} \in \mathcal{C}} H(X^m|U^t)$$

- The reduction in the minimum rate for function computation in comparison to the rate required for source reproduction.
- Rate reduction at zero message is

$$\rho_0 = \begin{cases} H(X^m) & \text{if } p_{X^m} \in \mathcal{P} \\ -\infty & \text{o.w.} \end{cases}$$

[2] Ma, N., Ishwar, P., Gupta, P. (2012). Interactive source coding for function computation in collocated networks. Information Theory, IEEE Transactions on 9725

Zero Sum Rate distribution

Lemma

- The set of distributions that result in the zero message for computing the max can be found by

$$\mathcal{P}_M = \Big\{ \prod_{i=1}^{m} p_{X_i} \Big| \exists i \in [m], \exists a \in \mathcal{X} : p_i(a) = 1, \forall j \neq i, \forall b > a, p_j(b) = 0, \text{and } \forall k \in [m], \sum_{x \in \mathcal{X}} p_k(x) = 1 \Big\}.$$

- The set of distributions that result in the zero message for computing the arg max is

$$\mathcal{P}_A = \Big\{ \prod_{i=1}^{m} p_{X_i} \Big| \exists i \in [m], \forall j \neq i, \mathcal{S}_j < \mathcal{S}_i, \forall k \in [m], \sum_{x \in \mathcal{S}_k} p_k(x) = 1 \Big\}.$$

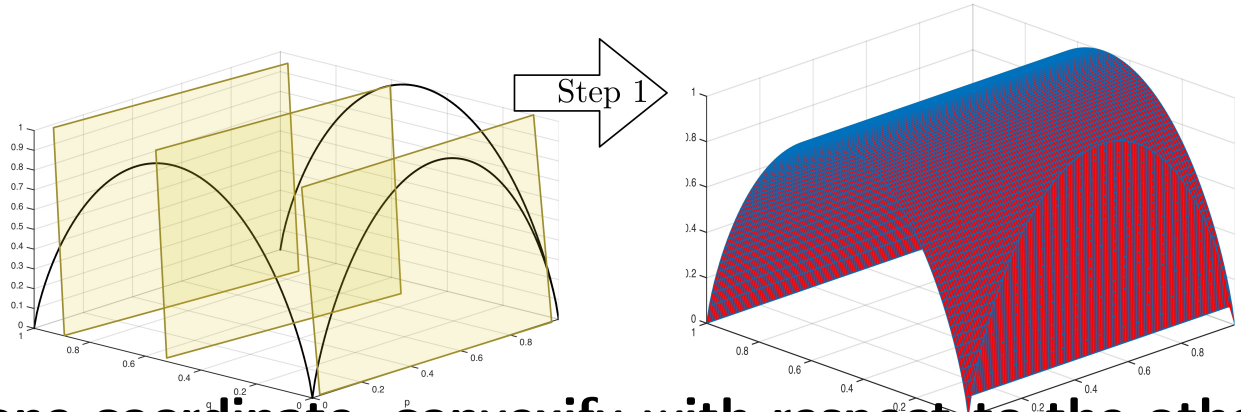
Rate reduction functional computation

Connection between the rate reduction of the original problem and the rate reduction of the subproblems

$$\rho_t(P_{X^m}) = \max_{p_{U^t|X^m} \in \mathcal{C}} H(X^m|U^t) = \max_{P_{U_1|X_1}} \left\{ \sum_{u_1} P_{U_1}(u_1) \rho_{t-1}(P_{X^m|U_1}(.|u_1)) \right\}$$

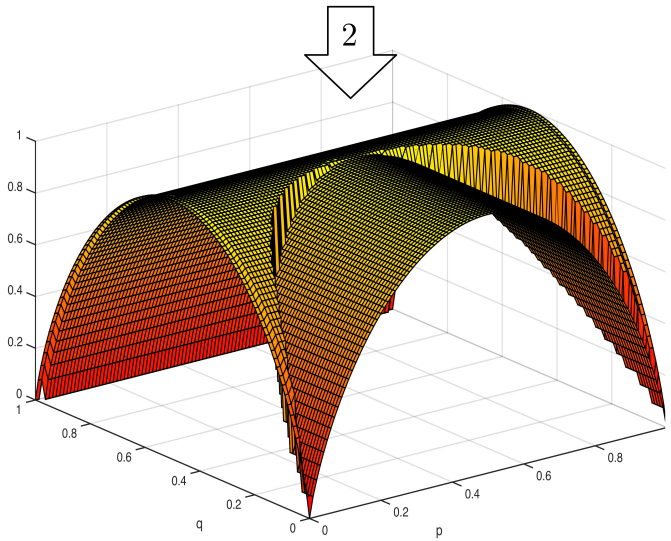
- Begin by choosing the first message, which corresponds to choosing the auxiliary r.v U_1.
- For each realization $U_1 = u_1$, the remaining part becomes $t-1$ message subproblem

Rate Reduction Functional



- **Fix one coordinate, convexify with respect to the other coordinate[2]**

- Step 0: Derive ρ_0, using $\mathcal{P}_M, \mathcal{P}_A$
- Step 1: Fix q, convexify with respect to $p \Rightarrow \rho_1$
- Step 2: Fix p, convexify with respect to $q \Rightarrow \rho_2$
- Continue Until it's convex with respect to all coordinates.



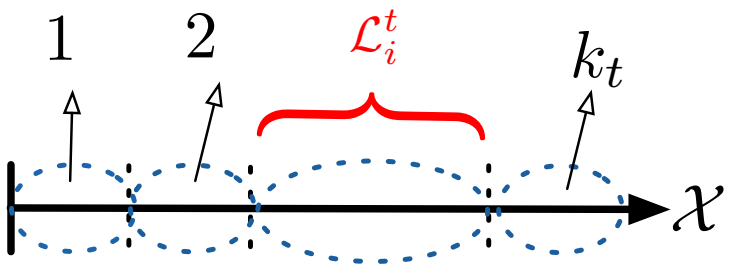
Outline

- Overview and problem formulation
- Computing the fundamental limit
- Designing optimal dynamic program
- Experimental results

Interactive Scalar Quantization

- Quantization function $\Rightarrow Q_t : \mathcal{X}_i \times \bigotimes_{l=1}^{t-1} \mathcal{Q}_l \rightarrow \mathcal{Q}_t$ for user $i = ((t-1) \bmod m + 1)$,
- Q_t partitions the support set L^t into k_t intervals.

- $k_t \Rightarrow$ Number of bins
- Bin sizes $\Rightarrow q_t = (n_1^t, ..., n_{k_t}^t)$,
- $\psi : \bigotimes_{l=1}^{T} \mathcal{Q}_l \rightarrow \mathcal{Z}$.
- The interval associated with the quantized message at round t is $\mathcal{L}^t = \mathcal{L}^t_{Q_t(X_i)} \Rightarrow$ The interval in which X_t lies in



k_t – bin quantizer

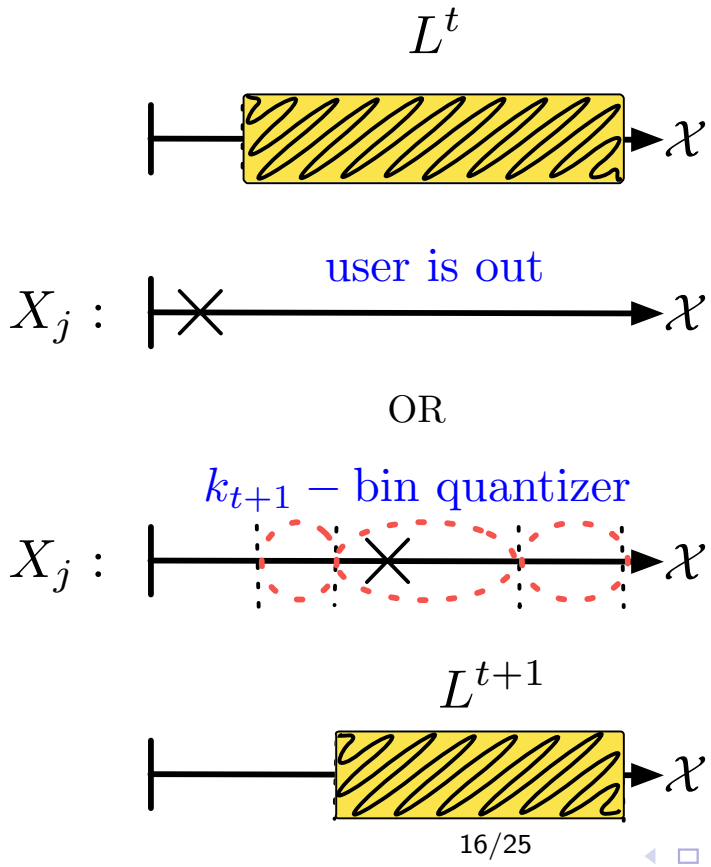
Optimal Interactive Scalar Quantization

- The optimal quantizers to utilize in this scheme in terms of minimizing
  - expected sum-rate
  - expected delay
  - weighted combination of the pair

  can be selected through a dynamic programming framework.
- At each round t, the users keep track of the public knowledge about the state of the system via $s^t = (N^t, L^t, \mathcal{N}^t, \mathcal{A}^t)$
- $L^t \Rightarrow$ remaining possible support for the maximum at time t
- $\mathcal{A}_i^t \Rightarrow$ the most recent public interval that user i was announced to be in at time t.
- $N^t \Rightarrow$ the number of active users at time t

Update Rule

- If $x_j \notin L^t$, $\Rightarrow$ this user simply sends no message and incurs no rate penalty in doing so, decreasing N^t accordingly.
- If $x_j \in L^t \Rightarrow X_j$ utilizes a k_{t+1}-bin quantizer $(n_1^{t+1}, ..., n_{k_{t+1}}^{t+1})$ to quantize his message, then Huffman encodes the result $Q_{t+1}(X_j)$.



Update Rule

- The list of active users at time $t+1$

$$\mathcal{N}^{t+1} = \begin{cases} \mathcal{N}^t \backslash \{j\} & x_j \notin L^t \\ \{i \in \mathcal{N}^t \big| \big| \mathcal{A}_i^t \cup \mathcal{L}^{t+1} \big| > 1\} & x_j \in L^t \end{cases}$$

- The most recent public interval that user i at time $t+1$

$$\mathcal{A}_i^{t+1} = \begin{cases} \mathcal{L}^{t+1} & i = j, x_j \in L^t \\ \mathcal{A}_i^t & \text{otherwise} \end{cases}$$

- The remaining possible support for the maximum at time $t+1$

$$L^{t+1} = \begin{cases} L^t \backslash \{\mathcal{L}_1^{t+1} \cup \ldots \cup \mathcal{L}_{Q_{t+1}-1}^{t+1}\} & x_j \in L^t \\ L^t & x_j \notin L^t \end{cases}$$

- The number of active users at time $t+1$

$$N^{t+1} = |\mathcal{N}^{t+1}|$$

Dynamic Program

- If at round t the state of the system is $s^t = (N^t, L^t, \mathcal{N}^t, \mathcal{A}^t)$, then the next online user X_j selects its quantization function Q_t such that the combination

$$C(s^t) = \min_{Q_t} \left\{(1-\lambda)R(q_t, s^t) + \lambda D(q_t, s^t) + \mathbb{E}(C(s^{t+1})|s^t, Q_t)\right\},$$

- The set of terminating states is smaller when the CEO wishes to determine the max

$$\mathcal{S}_A^* = \{s^t | N^t = 1 \text{ or } L^t = 1 \text{ or } |\mathcal{N}^t| = 1 \text{ or } |\mathcal{A}^t| = 1\}$$
$$\mathcal{S}_M^* = \{s^t |\ L^t = 1\}.$$

Reduced Complexity Quantization Strategies

- **Theorem** For a fixed $|\mathcal{X}| = L$ the expected sum rate required in an interactive scheme to compute the max is $\Theta(1)$ in the number of users m.
  - Scheme: send $T = m$ messages initiated by node 1
  - The sources have a block of N observations.
  - $t = 1 \Rightarrow$ The first user utilizes L-bin quantizer and entropy encodes the index associated with the observation $\Rightarrow$ broadcasts its message $\Rightarrow N \log L$ expected bit
  - $t = 2 \Rightarrow X_2$ quantizes only those instances that are greater than the first received message using a trivial L^2-bin quantizer.
  - the support set in each block is shrinking to $L_n^2, n \in [N]$.
  - Let Y_2 be a random variable indicating the number of instances that X_2 has a lower value than the received message.
  - $\mathbb{E}(R_2) = \log L\,(N - \mathbb{E}(Y_2))$
  - so $\mathbb{E}[R_2] = N\frac{L-1}{2L}\log L$ expected number bit at $t = 2$.

Reduced Complexity Quantization Strategies

- **Lemma** If there are $m = 2$ users, and the aim is to compute the arg max or max of the users' values, the k-level quantizer $\mathcal{U}(L) = \{\mathbf{1}_{k-1} \oplus (L - k + 1) : k = 2, ..., L\}$, has lower cost than any other quantizer obtained by permutation $\mathcal{U}_s(L) = \{\mathbf{1}_s \oplus (L - k + 1) \oplus \mathbf{1}_{k-s+1}\}$.

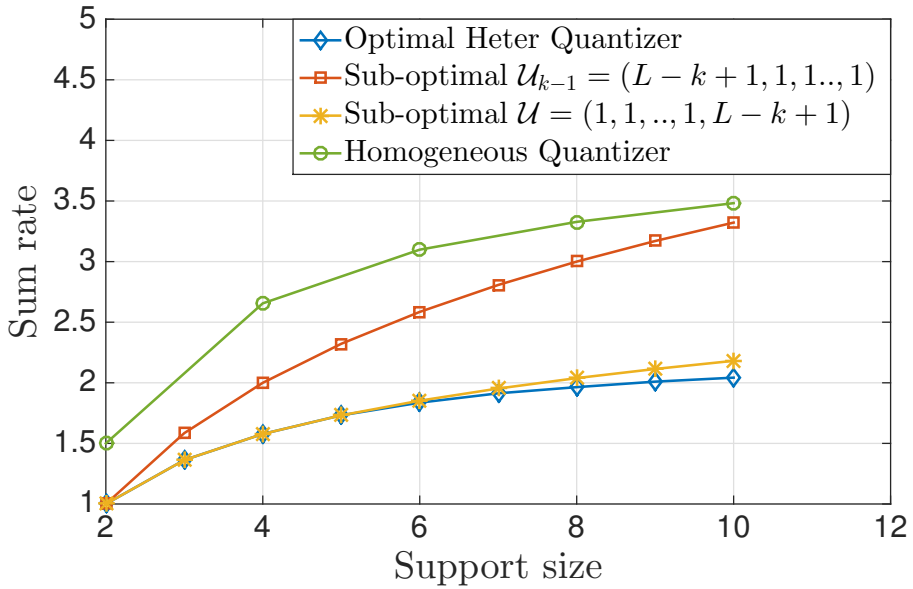
Outline

maybe NOK didn't have too much revenue past few years.

- Overview and problem formulation
- Computing the fundamental limit
- Designing optimal dynamic program
- Experimental results

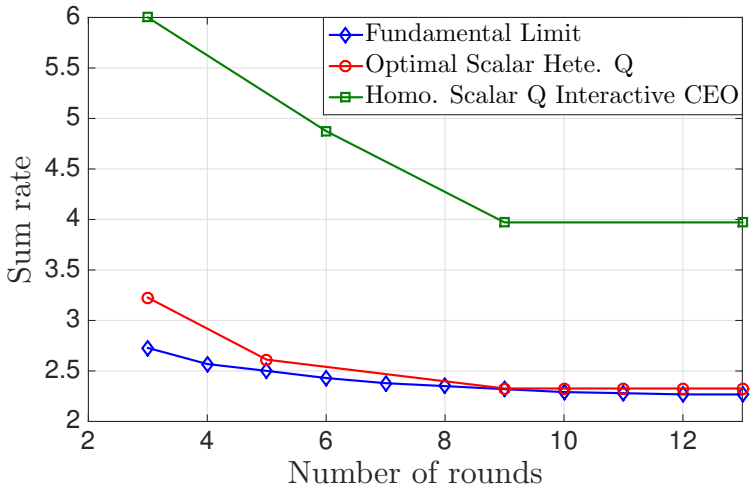
Numerical Results

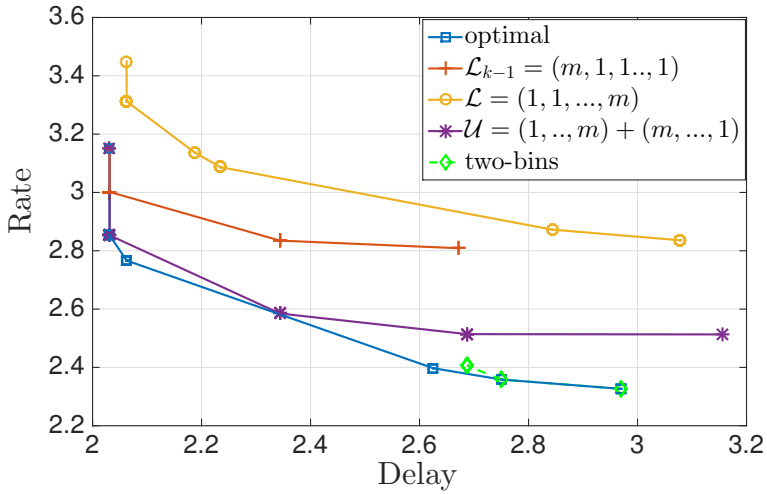
- Sum rate vs. the support s
- Fixed number of users, and for a varying support size $\Rightarrow \lambda = 0$
- Searching over $\mathcal{U}$ reduces the size of the search space



Numerical Results

- Number of users $N = 3$, and size of the support set $L = 4$





- Optimized encoding based on scalar quantization is shown to achieve sum-rate very close to the fundamental limit
- Interactive strategy in collocated network outperforms parallel message transmission
- Extending the family $\mathcal{U}$ to $\mathcal{L} = \mathcal{U} \cup \mathcal{U}_{k-1}$ results in reasonable rate-delay trade-offs as compared to the fundamental limit
- As we get to the larger delays, two-bin quantizers become optimal

Refrences

B. Boyle, J. Ren, J. Walsh, and S. Weber, "Interactive scalar quantization for distributed resource allocation," 2015.

N. Ma, P. Ishwar, and P. Gupta, "Interactive source coding for function computation in collocated networks," *IEEE Trans. Inf. Theory*, vol. 58, no. 7, pp. 4289–4305, 2012.

Thank You!

Question?