Correlated Quantization for Faster Nonconvex Distributed Optimization

Based on a joint work by

Andrei Panferov, Yury Demidovich,

Ahmad Rammal, Peter Richtárik

KAUST: King Abdullah University of Science and Technology

UAI 2025



Problem formulation

Parameters of the model

$$\min_{x \in \mathbb{R}^d} \left\{ f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x) \right\}$$

Non-convex loss

Clients

Goal: finiding an approximately stationary point of the nonconvex problem – a (random) vector $\hat{\boldsymbol{x}} \in \mathbb{R}^{\boldsymbol{d}}$ s.t.

$$\mathbb{E}[\|\nabla f(\hat{\boldsymbol{x}})\|^2] \leq \varepsilon^2,$$

all while minimizing the amount of communication between the n clients and the server

Communication Complexity in Distributed Training

- Key effectiveness metric: **communication complexity**

Number of **communication rounds** to find $\hat{\boldsymbol{x}}$

$\times$

Amount of **data exchanged per round**

- Assumption (standard in literature): **client-to-server** communication is **a bottleneck**

Communication Complexity Reduction

Reduce number of communication rounds

- Momentum
- Acceleration
- Local Training

Reduce amount of data exchanged per round

- Compression

Most of the common compression techniques: sparsification and quantization

- Sparsification methods reduce communication by only selecting an important sparse subset of the vectors to broadcast at each step
- Quantization methods quantize each component through randomized rounding to a discrete set of values, preserving the statistical properties of the original vector

🚀 Contributions

- We **extend the analysis** of the SOTA distributed optimization method MARINA beyond **independent quantizers**
- Prove better **communication complexity** of MARINA with **Correlated Quantizers (CQ)** in the **zero-Hessian-variance regime**
- Compare against strong **independent quantizer baselines**
- ✅ **Experiments** validate the theory

🚀 Contributions

- We **compare two distributed algorithms** using correlated quantizers: MARINA and DCGD

- In the **zero-Hessian-variance regime**, MARINA shows **significantly lower communication complexity**

- This makes MARINA the **superior choice** in this setting

- ✅ Our **experimental results** confirm the theoretical findings

🚀 Contributions

- We show that **Correlated Quantizers (CQ)** achieve **much lower Mean Squared Error (MSE)** —
  by a factor of n compared to independent quantizers on **homogeneous data**
- We also provide **insights into why CQ are especially effective** when combined with MARINA in the **zero-Hessian-variance regime**
- These findings highlight the **theoretical and practical benefits** of using CQ in distributed optimization

MARINA: SOTA method

Input: initial point $x^0 \in \mathbb{R}^d$, rate $\gamma > 0$, probability $p \in (0, 1]$, number of iterations T

$g^0 = \nabla f(x^0)$

for $t = 0, 1, \ldots, T-1$ **do**

  Sample $c_t \sim \text{Bern}(p)$

  Broadcast g^t to all workers

  for $i = 1, \ldots, n$ **in parallel do**

    $x^{t+1} = x^t - \gamma g^t$

    $g_i^{t+1} = \nabla f_i(x^{t+1})$ if $c_t = 1$, and $g_i^{t+1} = g_i^t + \mathcal{Q}_i(\nabla f_i(x^{t+1}) - \nabla f_i(x^t))$ otherwise

  end for

  $g^{t+1} = \frac{1}{n}\sum_{i=1}^n g_i^{t+1}$

end for

Output: $\hat{x}^T$ uniformly from $\{x^t\}_{t=0}^{T-1}$

E. Gorbunov, K. Burlachenko, Z. Li, P. Richtárik. MARINA: Faster non-convex distributed learning with compression ICML21

Given n vectors $a_1, \ldots, a_n \in \mathbb{R}^d$, **Mean Square Error (MSE)** associated with the set of randomized compressors $\{\mathcal{Q}_i\}_{i=1}^n$ is the quantity $\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \mathcal{Q}_i(a_i) - \frac{1}{n}\sum_{i=1}^n a_i\right\|^2\right]$.

- Theoretical complexity of MARINA grows with MSE
- Crucial to identify compressors with low MSE
- Typically, there exists a trade-off between MSE and communication cost

Zero-Hessian-Variance Regime

Let $L_{\pm} \geq 0$ be the smallest constant such that

$$\frac{1}{n}\sum_{i=1}^{n}\left\|\nabla f_i(x) - \nabla f_i(y)\right\|^2 - \left\|\nabla f(x) - \nabla f(y)\right\|^2 \leq L_{\pm}^2 \left\|x - y\right\|^2, \quad x, y \in \mathbb{R}^d.$$

The quantity $L_{\pm}^2$ is called Hessian variance.

- $L_{\pm} = 0$ extends the case where the clients are homogeneous or nearly homogeneous
- Achieving the zero-Hessian-variance regime in practice can be challenging
- Practical problems can indeed have $L_{\pm}$ values very close to zero

Correlated Quantizers

One-dimensional

Input: $a_1, a_2, \ldots, a_n, l, r \in \mathbb{R}; \forall i \in [n], a_i \in [l, r]$
Generate π, a random permutation of $\{0, 1, \ldots, n-1\}$
for $i = 1$ to n **do**
  $y_i = \frac{a_i - l}{r - l}$.
  $U_i = \frac{\pi_i}{n} + \gamma_i$, where γ_i has a continuous uniform distribution $\mathrm{U}[0, 1/n)$.
  $Q_i(a_i) = (r - l)1_{U_i < y_i}$.
end for
Output: $\frac{1}{n}\sum_{i=1}^{n} Q_i(a_i)$.

A. Suresh, Z. Sun, J. Ro, F. Yu.
Correlated quantization for distributed mean estimation and optimization
ICML22

Multi-dimensional

Assume that each a_i is a d-dimensional vector and that Q_i quantizes each coordiante independently

MSE of quantizers on homogeneous data

Assume $a_i = a,\ l = -\|a\|,\ r = \|a\|$

Theorem: MSE of CQ

CQ $\{\mathcal{Q}_i\}_{i=1}^n$ are individually unbiased and the MSE of quantizers $\{\mathcal{Q}_i\}_{i=1}^n$ associated with the set of vectors $\{a_i\}_{i=1}^n$ can be bounded from above in the following way:

$$\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n a_i - \mathcal{Q}_i(a_i)\right\|^2\right] \leq \frac{d\|a\|^2}{n^2}.$$

Independent Quantizers (IQ)

One-dimensional

$Q_i(a_i) = r$ with probability $\frac{a_i - l}{r - l}$,

and $Q_i(a_i) = l$ otherwise

Multi-dimensional

Assume that each a_i is a d-dimensional vector and that Q_i quantizes each coordiante independently

In contrast, the upper bound on MSE of IQ is $\frac{d\|a\|^2}{n}$

Communication Complexity of MARINA

Let $L_{\pm} = 0$. Denote by $\mathcal{C}_{\text{cor}}$ the communication complexity per client in MARINA with CQ. Similarly, denote by $\mathcal{C}_{\text{ind}}$ the communication complexity per client in MARINA with IQ. Then

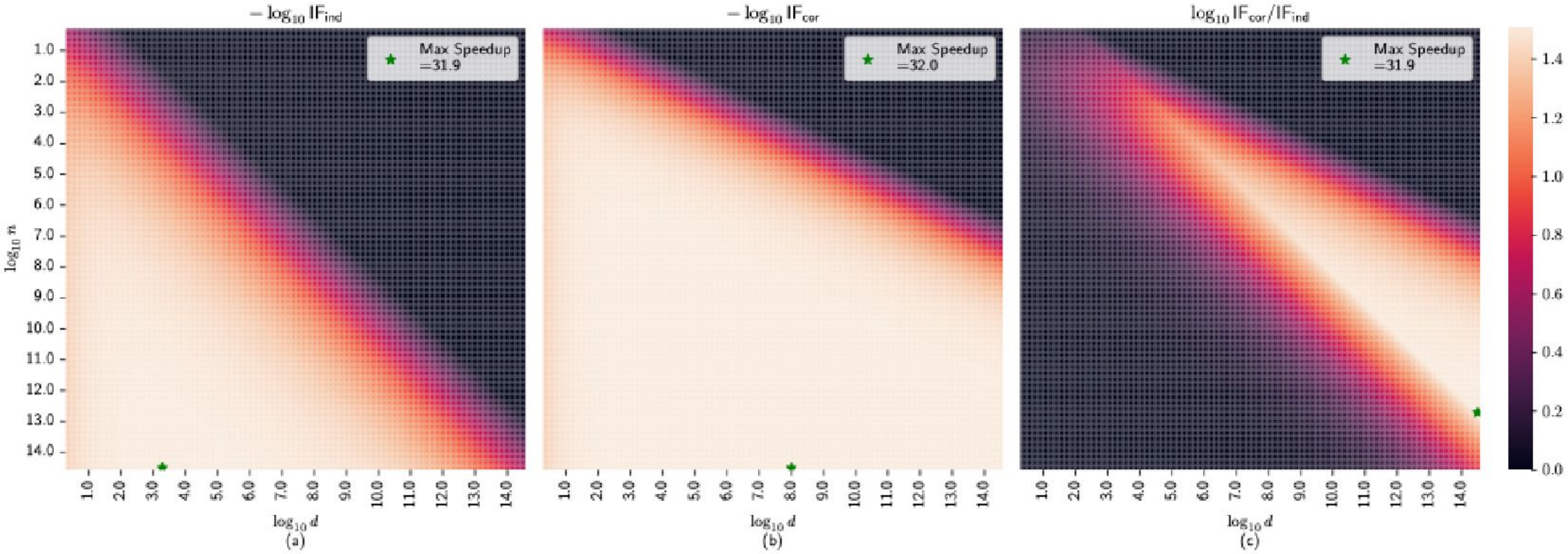
$$\frac{\mathcal{C}_{\text{ind}}}{\mathcal{C}_{\text{cor}}} = \frac{1 + \sqrt{\frac{(1-p)}{p}\frac{d}{4n}}}{1 + \sqrt{\frac{(1-p)}{p}\frac{d}{4n^2}}}.$$

That is, $\forall p \in [0,1]$, $\mathcal{C}_{\text{cor}} \leq \mathcal{C}_{\text{ind}}$. In particular, we show that $\mathcal{C}_{\text{cor}} = \mathcal{O}\left(\frac{\Delta^0 L}{\varepsilon^2} \min\left\{d, 1 + \frac{d}{n}\right\}\right)$ and $\mathcal{C}_{\text{ind}} = \mathcal{O}\left(\frac{\Delta^0 L}{\varepsilon^2} \min\left\{d, 1 + \frac{d}{\sqrt{n}}\right\}\right)$.

- Experiments suggest that when $d = n \gg 1$, the complexity ratio is approximately 7.29
- The ratio can reach up to 32

An Improvement Factor (IF) is a ratio of complexities of MARINA and GD

Speedup of MARINA with CQ



Logarithmic speedup of MARINA with CQ/IQ over GD.

(c): Logarithmic speedup of MARINA+CQ compared to MARINA+IQ

(a) MARINA+IQ defaults to GD when $n \ll d$ and achieves the best possible speedup of ×32 (owing to the compressor's 1-bit per coordinate behavior) when $n \gg d$.

(b) CQ are distinguished by $d = n^2$

(c) CQ surpass IQ by up to a factor of ×32 when $\sqrt{d} < n < d$.

Correlated Quantizers in Zero-Hessian-Variance Regime

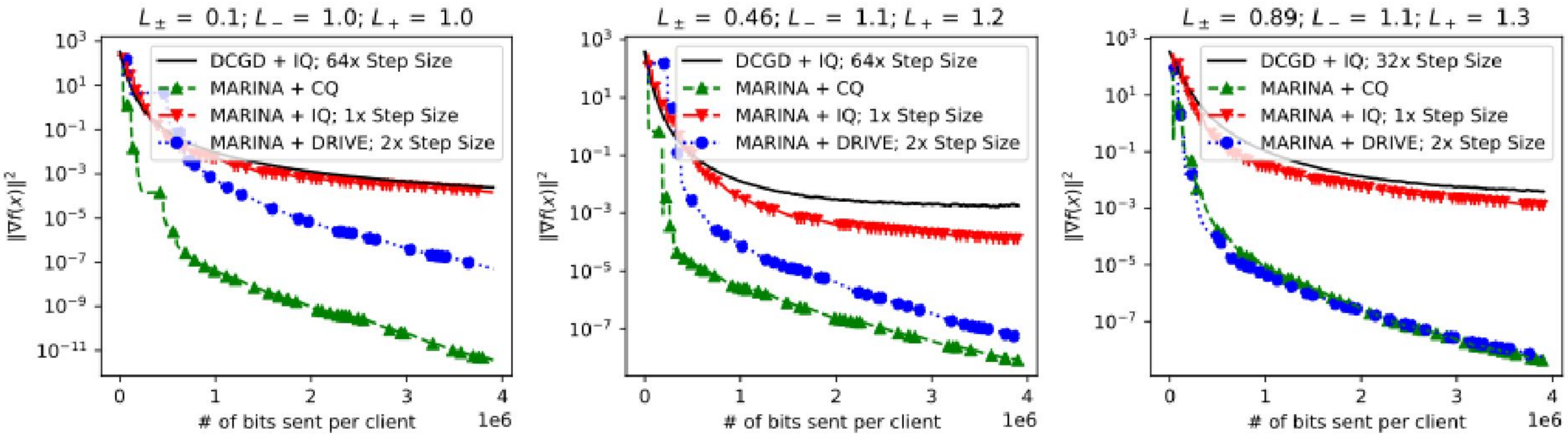
Table 1: Comparison of communication complexities of different distributed methods combined with different quantizers in the nonconvex regime with homogeneous clients (see Section 3.2), when $d \leq n$. In the homogeneous scenario, $L_- = L_+ = L$ and $L_\pm = 0$. Notation: $\Delta^0 = f(x^0) - f^*$. Abbreviations: CQ = "Correlated Quantizers", ISCC = "Importance Sampling Combinatorial Compressors", IQ = "Independent Quantizers".

Method	Quantizer	Communication Complexity	Correlated Compressors	Reference
DCGD	IQ, Def. 6	$\mathcal{O}\left(\frac{\Delta^0 dL}{\varepsilon^2}\right)$	✗	Suresh et al. [2022]
DCGD	CQ, Def. 7	$\mathcal{O}\left(\frac{\Delta^0 dL}{\varepsilon^2}\right)$	✓	Suresh et al. [2022]
MARINA	$\mathcal{D}_{nat}^{q,k}$, Def. 3	$\mathcal{O}\left(\frac{\Delta^0 L}{\varepsilon^2} \min\left\{d, 1 + \frac{d}{\sqrt{n}}\right\}\right)$	✗	Gorbunov et al. [2022]
MARINA	IQ, Def. 6	$\mathcal{O}\left(\frac{\Delta^0 L}{\varepsilon^2} \min\left\{d, 1 + \frac{d}{\sqrt{n}}\right\}\right)$	✗	Gorbunov et al. [2022]
MARINA	ISCC, Asm. 6	$\mathcal{O}\left(\frac{\Delta^0 d}{\varepsilon^2} \min\left\{L, \frac{L}{n} + \frac{\sqrt{\omega+1}L_{avg}}{\sqrt{n}}\right\}\right)$	✗	Corollary 4, this work
MARINA	CQ, Def. 7	$\mathcal{O}\left(\frac{\Delta^0 L}{\varepsilon^2} \min\left\{d, 1 + \frac{d}{n}\right\}\right)$	✓	Proposition 4, this work

Table 2: Comparison of important characteristics of different quantizers in the nonconvex zero-Hessian-variance regime and when $d \leq n$: bits sent per client and MSE (Mean Square Error, Section 3.1). Notation: $\mathcal{D}_{sta}^{2,k}$ – Standard Dithering, $\mathcal{D}_{sta}^{\infty,1}$ – Ternary Quantization, $\mathcal{D}_{nat}^{q,k}$ – Natural Dithering.

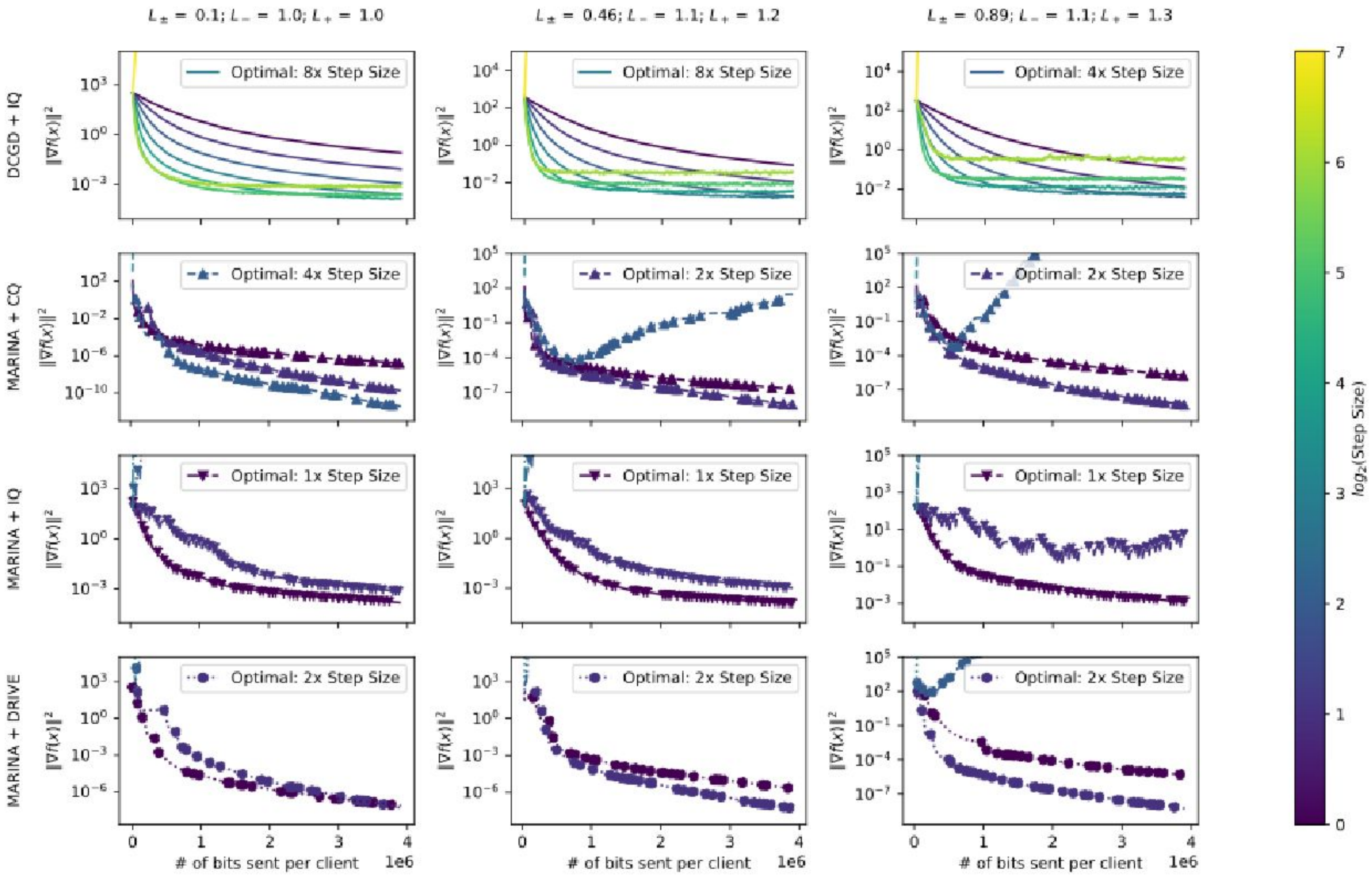
Quantizer	Bits Sent	MSE	Correlated?	Reference
$\mathcal{D}_{sta}^{2,k}$, Def. 2	$\mathcal{O}\left(k(k+\sqrt{d})\right)$	$\frac{\sqrt{d}}{nk}$	✗	Alistarh et al. [2017]
$\mathcal{D}_{sta}^{\infty,1}$, Def. 2	$31 + d\log_2 3$	$\frac{\sqrt{d}-1}{n}$	✗	Wen et al. [2017]
$\mathcal{D}_{nat}^{q,k}$, Def. 3	$31 + d\log_2(2k+1)$	$\frac{\sqrt{d}}{n2^{k-1}}$	✗	Gorbunov et al. [2022]
IQ, Def. 6	$32 + d$	$\frac{d\|a\|^2}{n}$, Cor. 1	✗	Gorbunov et al. [2022]
CQ, Def. 7	$32 + d$	$\frac{d\|a\|^2}{n^2}$, Cor. 2	✓	Suresh et al. [2022]
ISCC, Asm. 6	$\frac{\mathcal{O}(d)}{n}$	$\left(\frac{A}{n^2}\sum_{i=1}^n \frac{1}{w_i} - B\right)\|a\|^2$, Asm. 6	✗	Corollary 4, this work

Baseline Comparison on Quadratics



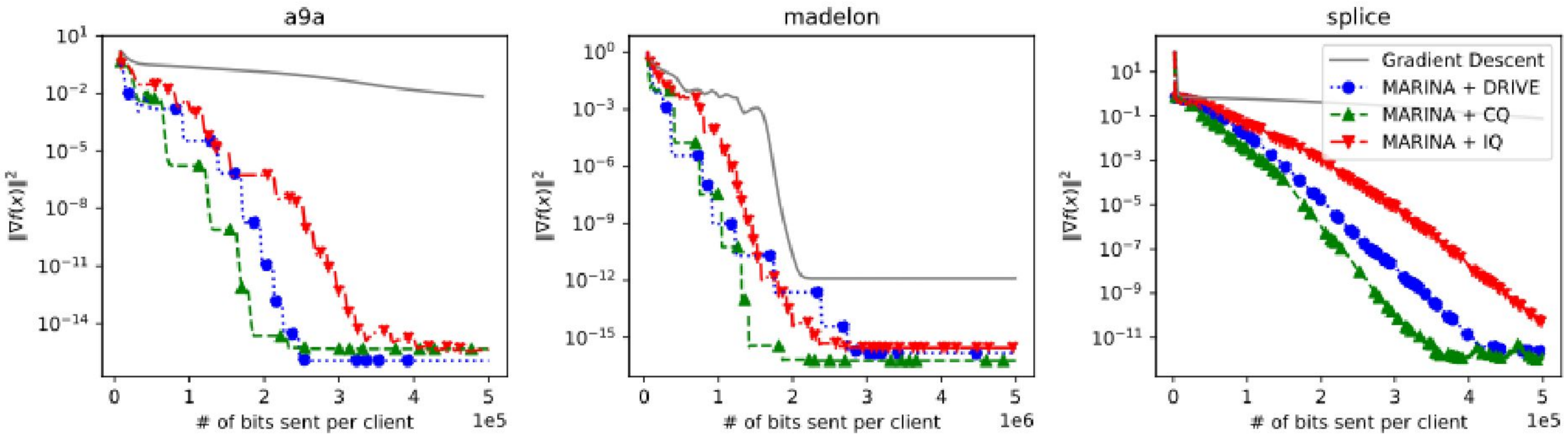
- Experiments on quadratic optimization tasks with varying smoothness constants
- Enables control over $L_\pm$ values
- $d = 1024,\ n = 128$, regularization $\lambda = 0.001$, noise scale $s \in \{0, 0.5, 1.0\}$
- CQ outperform IQ and are on par with DRIVE even in tasks where $L_\pm$ substantially deviates from 0
- Theory only for $L_\pm = 0$, no theoretical stepsize when $L_\pm > 0$

Baseline Comparison on Quadratics



- We increment the step size in multiples of 2 (2, 4, 8, ...) of the theoretically optimal step size.
- Our aim is to identify the step size that ensures the algorithm's best performance at 4*10^6 bits communicated from each client to the server (sufficiently large to demonstrate relative convergence between different algorithms).
- The convergence plots, as well as details about the selected optimal step sizes.

Non-Convex Logistic Regression

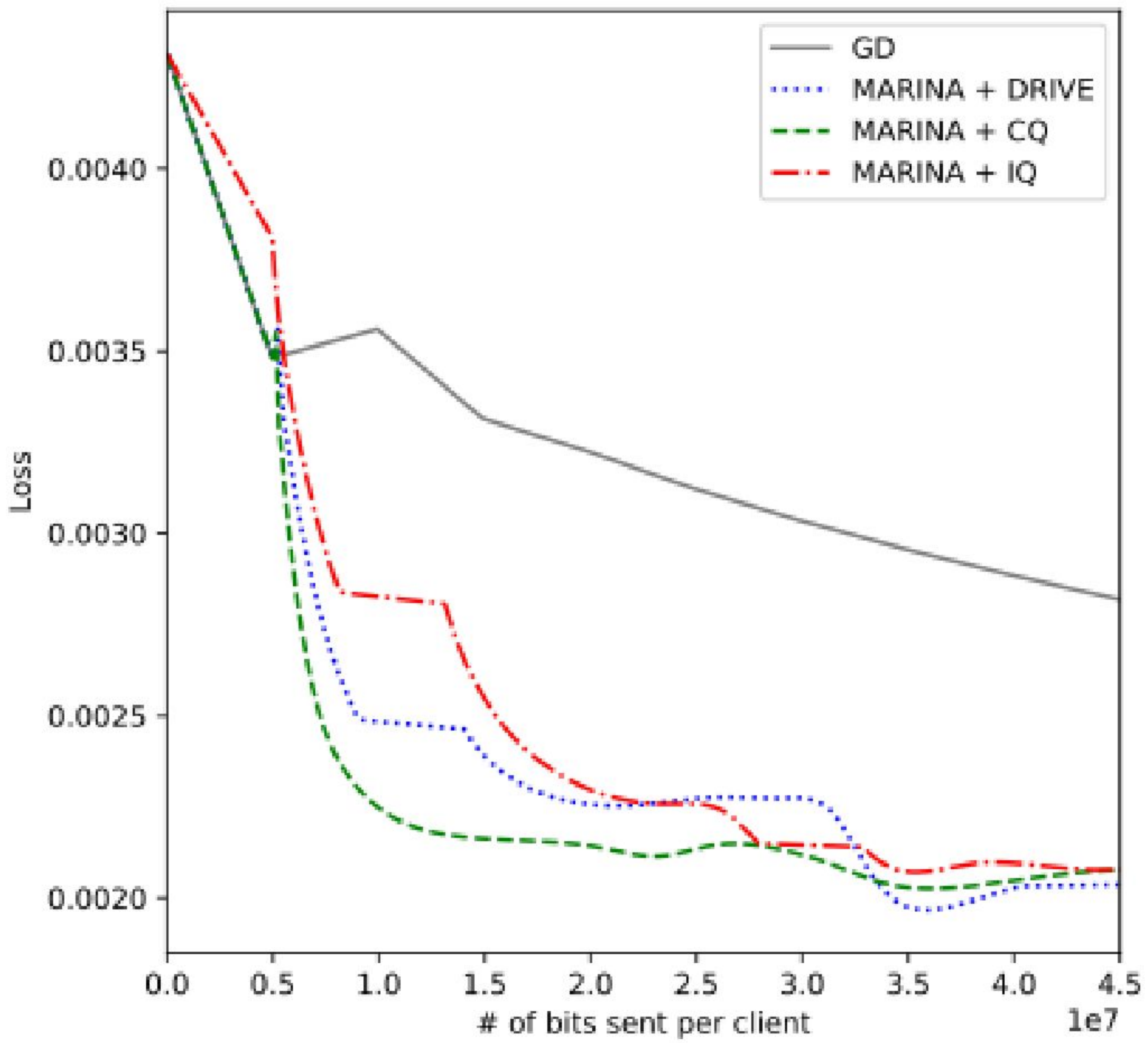


Dataset	$n = d$	N	$\lfloor N/d \rfloor$
a9a	123	32,561	264
madelon	500	2,000	4
splice	60	1,000	16

$$f(x) = \frac{1}{m}\sum_{k=1}^{m} \log\left(1 + \exp\left(-y_k a_k^T x\right)\right) + \lambda \sum_{j=1}^{d} \frac{x_j^2}{1 + x_j^2}$$

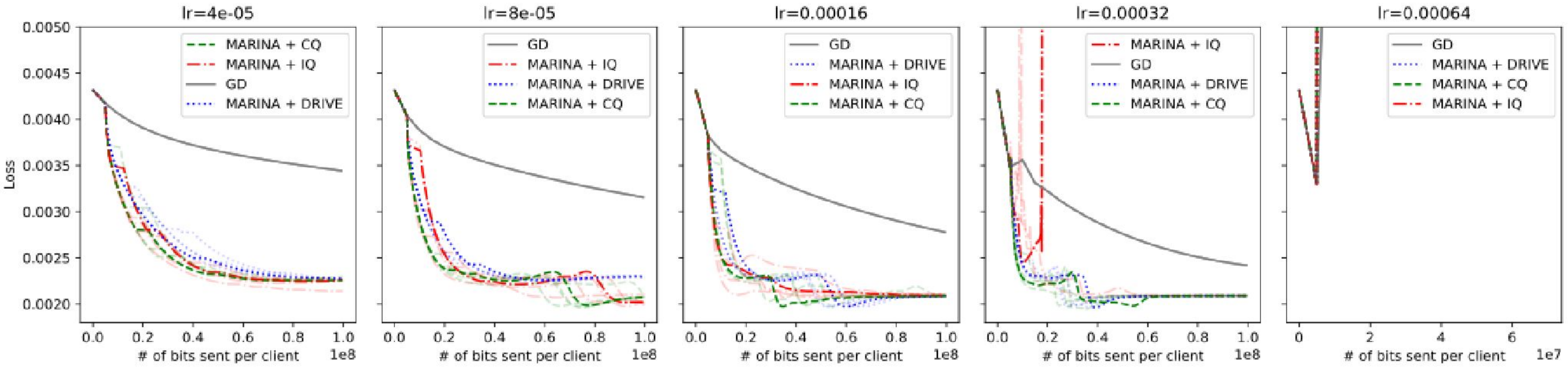
- LibSVM datasets are partitioned into $n = d$ uniform segments
- Included DGD, MARINA+DRIVE, MARINA+CQ, MARINA+IQ
- $L_{\pm} > 0$, calculation is infeasible
- Our approach is mostly dominant even in $L_{\pm} > 0$ case against a strong baseline MARINA+DRIVE.

Experiments with an MLP



- Experiments with an MLP classifier on the a9a dataset with 131 clients
- MARINA+CQ exhibits reduced complexity compared to MARINA+DRIVE, DCGD+IQ, MARINA+IQ.
- MARINA+CQ accommodates larger step sizes due to lower compression errors compared to MARINA+IQ, resulting in faster convergence in terms of loss.

Experiments with an MLP



- We provide the optimal stepsize selection procedure for the MLP classifier experiment on the a9a dataset, involving 131 clients.

- The largest step size was chosen such that the median of five optimization runs still converged.

Thank you! Questions?