

Quantized Langevin Stochastic Dynamics for Bayesian Federated Learning

Maxime Vono

Senior Researcher
Criteo AI Lab, Paris, France

Joint work with Aymeric Dieuleveut, Alain Durmus, Eric Moulines and Vincent Plassier

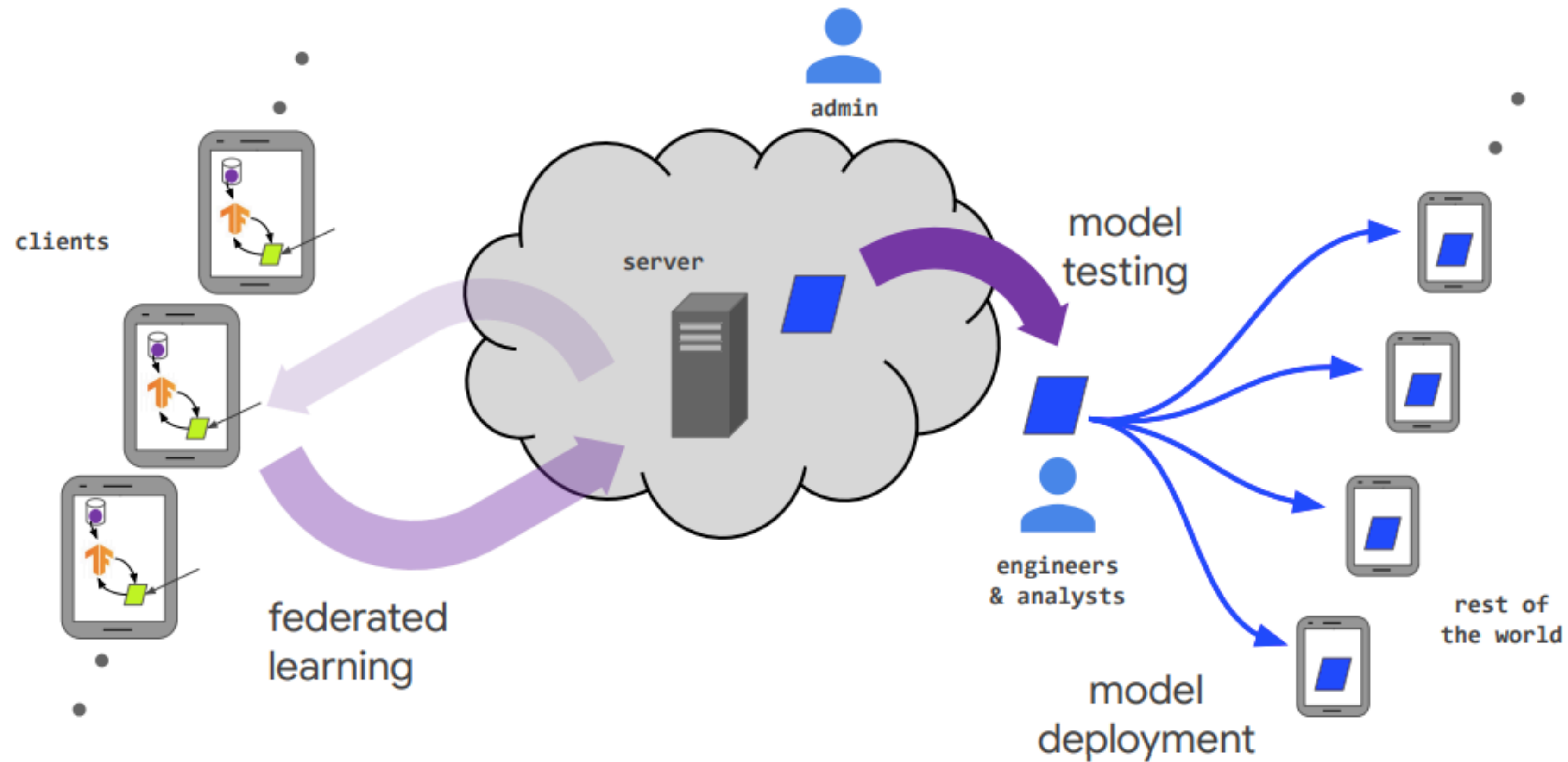
CRITEO
AI Lab



école
normale
supérieure
paris-saclay

Federated Learning (FL)

Users (clients) collaborate with each other via a central server without sharing their data



source: Kairouz et al. (2019), Advances and Open Problems in Federated Learning

FL is associated with several challenges

1. Clients/server architecture requires efficient **communication**
2. Local datasets might be large : **computational bottleneck**
3. Clients might be unreliable : **partial device participation**
4. Local datasets can be highly heterogenous: **data heterogeneity**
5. Clients want to protect their local data: **data privacy**

Bayesian FL

Instead of solving

$$\theta^{\star} = \arg \min_{\theta} \sum_{k=1}^K \frac{N_k}{N} f_k(\theta) , \text{ with } f_k = \frac{1}{N_k} \sum_{i \in D_k} U_{k,i}$$

Weight

Local empirical risk

we want to sample from

$$\pi(\theta \mid D) \propto \prod_{k=1}^K \pi_k(\theta \mid D_k) \propto \prod_{k=1}^K e^{-U_k(\theta)}$$

$$U_k = \sum_{i \in D_k} U_{k,i}$$

Local unnormalized posterior density
for each client

We focus here on the most popular sampling approach in ML : **Langevin dynamics**

Langevin dynamics

$$\pi(\theta \mid D) \propto \prod_{k=1}^K e^{-U_k(\theta)}$$

Posterior density

Langevin stochastic differential equation

$$d\theta_t = - \sum_{k=1}^K \nabla U_k(\theta_t) dt + \sqrt{2} dB_t$$

$(B_t)_{t \geq 0}$: Brownian motion

admits $\pi(\theta \mid D)$ as invariant measure

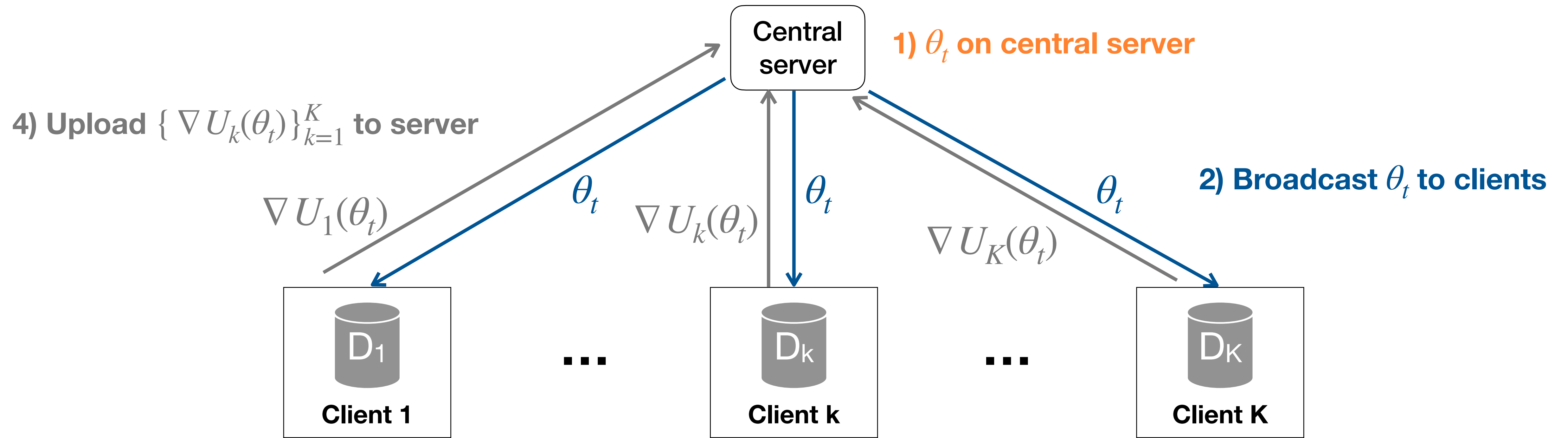
**Without Gaussian noise,
idem as gradient descent**

In practice, we consider its Euler discretization

$$\theta_{t+1} = \theta_t - \gamma \sum_{k=1}^K \nabla U_k(\theta_t) + \sqrt{2\gamma} Z_{t+1}, \quad Z_{t+1} \sim \mathcal{N}(0_d, I_d)$$

Langevin dynamics in FL context

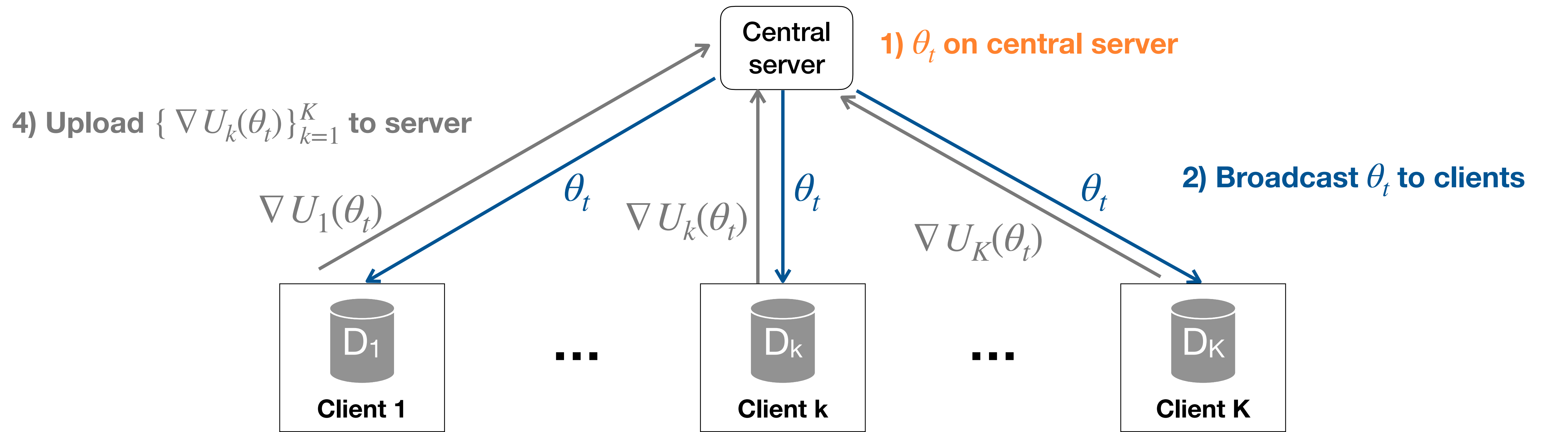
5) Update $\theta_{t+1} = \theta_t - \gamma \sum_{k=1}^K \nabla U_k(\theta_t) + \sqrt{2\gamma} Z_{t+1}$ on server



3) Each client k computes a local gradient at θ_t : $\nabla U_k(\theta_t)$

Langevin dynamics in FL context

5) Update $\theta_{t+1} = \theta_t - \gamma \sum_{k=1}^K \nabla U_k(\theta_t) + \sqrt{2\gamma}Z_{t+1}$ on server

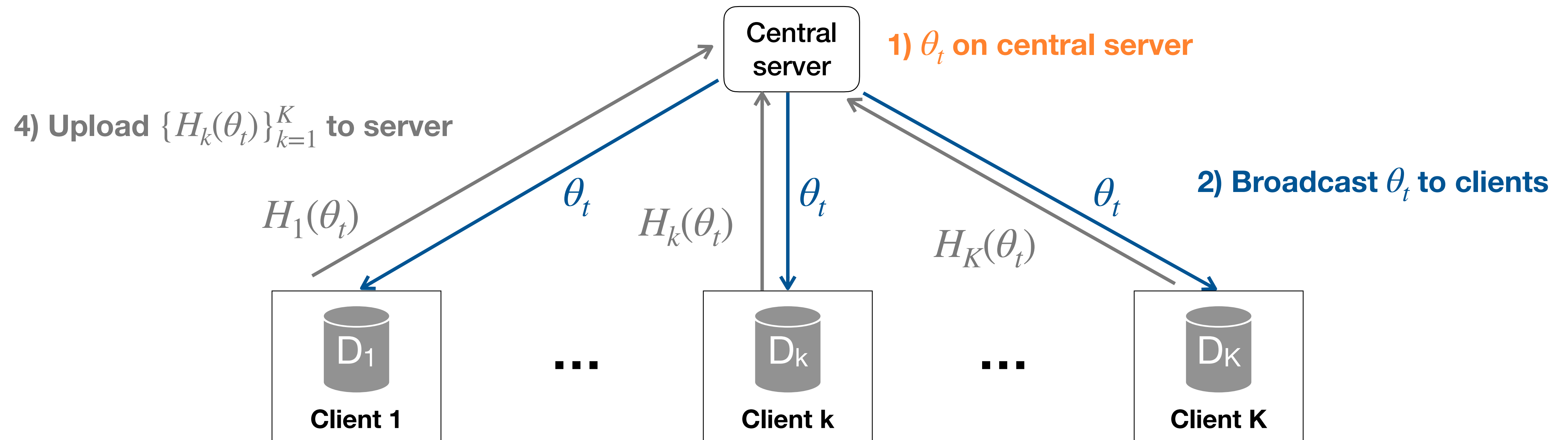


3) Each client k computes a local gradient at θ_t : $\nabla U_k(\theta_t)$

Local computational cost $O(N_k)$ might be large

Langevin dynamics in FL context

5) Update $\theta_{t+1} = \theta_t - \gamma \sum_{k=1}^K H_k(\theta_t) + \sqrt{2\gamma} Z_{t+1}$ on server



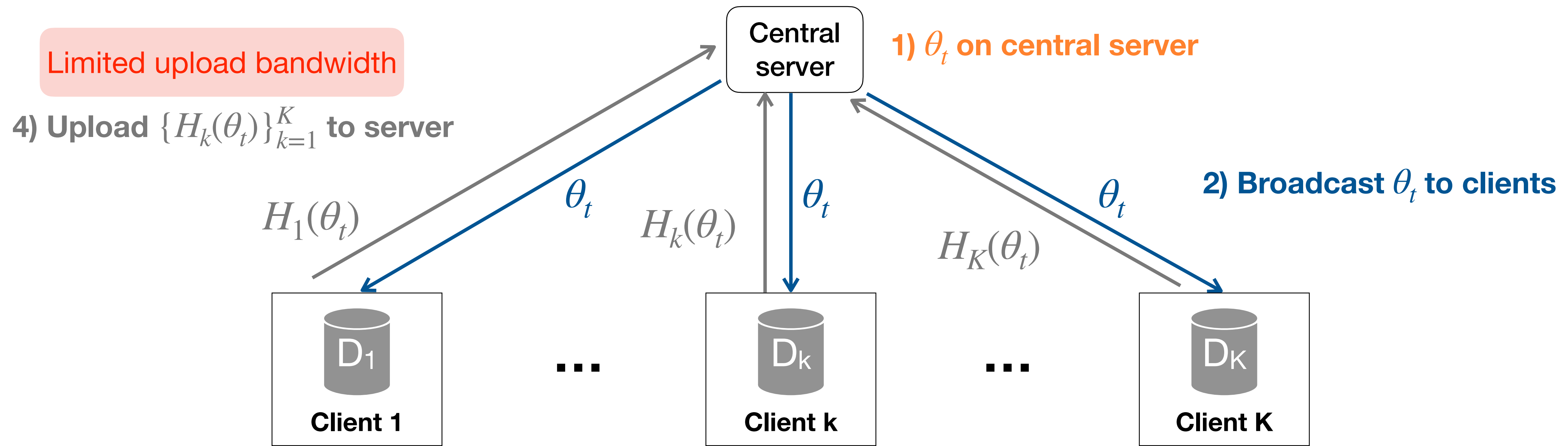
3) Each client k computes a local stochastic gradient at θ_t :

$$H_k(\theta_t) = \frac{N_k}{n_k} \sum_{i \in S_{k,t}} \nabla U_{k,i}(\theta_t)$$

Use stochastic gradients

Langevin dynamics in FL context

5) Update $\theta_{t+1} = \theta_t - \gamma \sum_{k=1}^K H_k(\theta_t) + \sqrt{2\gamma} Z_{t+1}$ on server



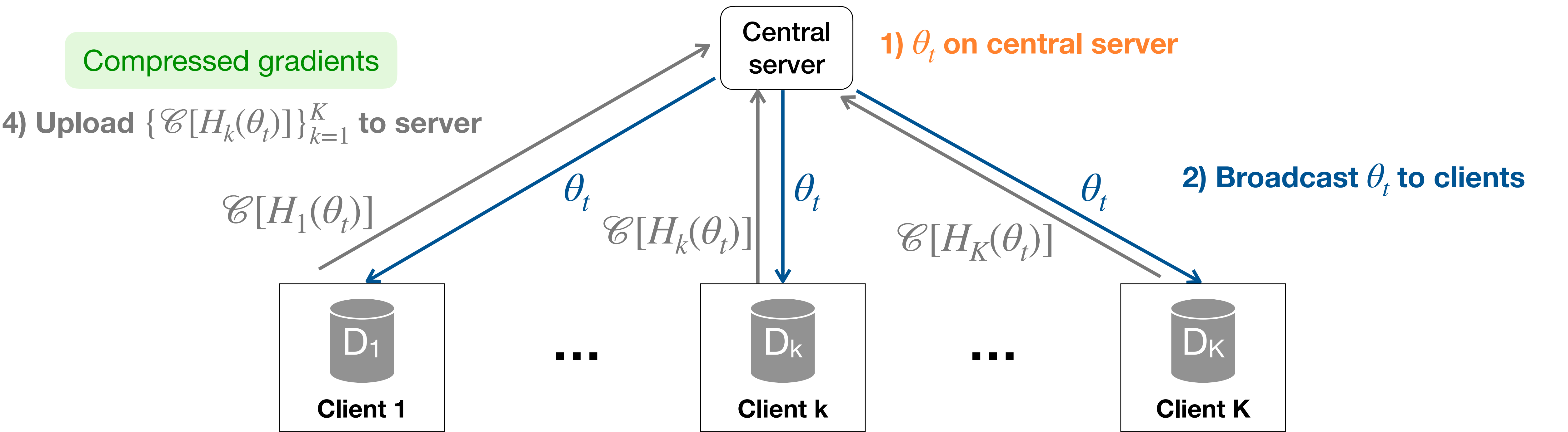
3) Each client k computes a local stochastic gradient at θ_t :

$$H_k(\theta_t) = \frac{N_k}{n_k} \sum_{i \in S_{k,t}} \nabla U_{k,i}(\theta_t)$$

Use stochastic gradients

Langevin dynamics in FL context

5) Update $\theta_{t+1} = \theta_t - \gamma \sum_{k=1}^K \mathcal{C}[H_k(\theta_t)] + \sqrt{2\gamma}Z_{t+1}$ on server

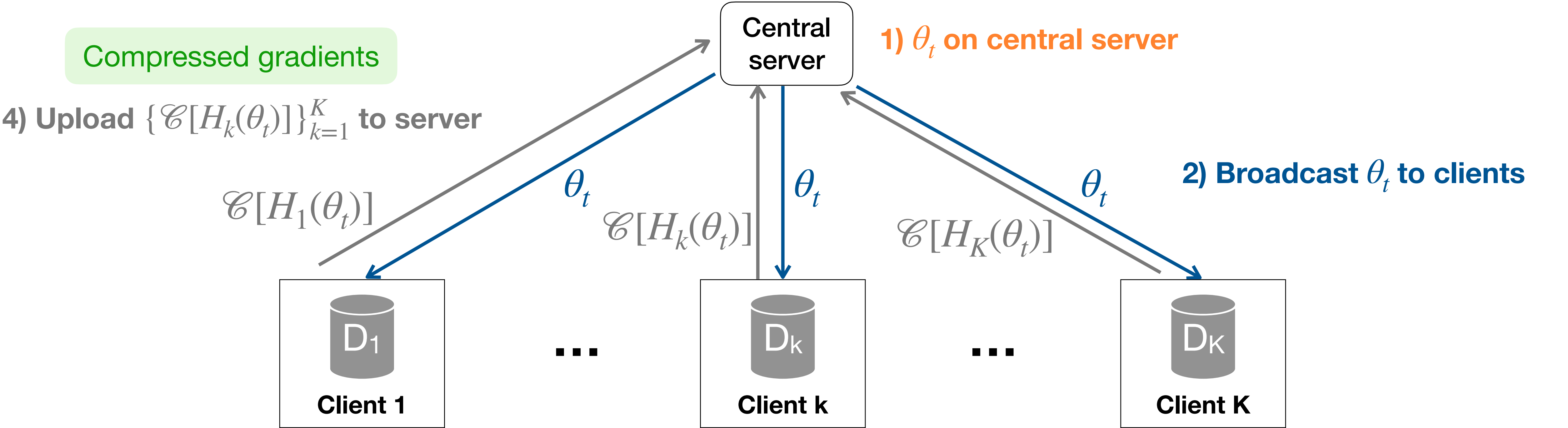


$$H_k(\theta_t) = \frac{N_k}{n_k} \sum_{i \in S_{k,t}} \nabla U_{k,i}(\theta_t)$$

Use stochastic gradients

Langevin dynamics in FL context

5) Update $\theta_{t+1} = \theta_t - \gamma \sum_{k=1}^K \mathcal{C}[H_k(\theta_t)] + \sqrt{2\gamma}Z_{t+1}$ on server



$$H_k(\theta_t) = \frac{N_k}{n_k} \sum_{i \in S_{k,t}} \nabla U_{k,i}(\theta_t)$$

Use stochastic gradients

Quantized Langevin stochastic dynamics (QLSD)

$$\theta_{t+1} = \theta_t - \gamma \sum_{k=1}^K \mathcal{C}[H_k(\theta_t)] + \sqrt{2\gamma} Z_{t+1} \quad Z_{t+1} \sim \mathcal{N}(0_d, \mathbf{I}_d)$$

2 ingredients:

1. Local stochastic gradients $H_k(\theta_t) = \frac{N_k}{n_k} \sum_{i \in S_{k,t}} \nabla U_{k,i}(\theta_t)$

N_k : number of data on client k
 $S_{k,t}$: random minibatch of size n_k

2. Unbiased compression operator $\mathcal{C}$

Without Gaussian noise,
similar to QSGD
[Alistarh et al. 2017]

Convergence analysis for QLSD

$$\theta_{t+1} = \theta_t - \gamma \sum_{k=1}^K \mathcal{C}[H_k(\theta_t)] + \sqrt{2\gamma} Z_{t+1} \quad Z_{t+1} \sim \mathcal{N}(0_d, \mathbf{I}_d)$$

Assumptions :

1. $\mathbb{E}[\mathcal{C}(v)] = v$ and $\mathbb{E}[\|\mathcal{C}(v) - v\|^2] \leq \omega \|v\|^2$

For $\mathcal{C}^{(s)}$: $\omega = \min(d/s^2, \sqrt{d}/s)$

2. Unbiasedness of $\sum_k H_k$ + bounds on the variance of H_k

3. U_k m - strongly convex and L -gradient Lipschitz

$$W_2^2(\text{Law}(\theta_t), \pi) \lesssim C \cdot (1 - \gamma m/2)^t + \mathcal{O}(\gamma) \left[N^2 + (\omega + 1)d + \omega \sum_{k=1}^K \|\nabla U_k(\theta^\star)\|^2 \right]$$

Convergence analysis for QLSD

Global dataset D of size $N = \sum_k N_k$

d : dimension of param. θ

$$W_2^2(\text{Law}(\theta_t), \pi) \lesssim C \cdot (1 - \gamma m/2)^t + O(\gamma) \left[N^2 + (\omega + 1)d + \omega \sum_{k=1}^K \|\nabla U_k(\theta^\star)\|^2 \right]$$

Convergence analysis for QLSD

Global dataset D of size $N = \sum_k N_k$

d : dimension of param. θ

$$W_2^2(\text{Law}(\theta_t), \pi) \lesssim C \cdot (1 - \gamma m/2)^t + O(\gamma) \left[N^2 + (\omega + 1)d + \omega \sum_{k=1}^K \|\nabla U_k(\theta^\star)\|^2 \right]$$

contracting term

Convergence analysis for QLSD

Global dataset D of size $N = \sum_k N_k$

d : dimension of param. θ

$$W_2^2(\text{Law}(\theta_t), \pi) \lesssim C \cdot (1 - \gamma m/2)^t + O(\gamma) \left[N^2 + (\omega + 1)d + \omega \sum_{k=1}^K \|\nabla U_k(\theta^\star)\|^2 \right]$$

contracting term

error term involving :

- discretization error
- variance of stochastic gradients
- compression impact

Convergence analysis for QLSD

Global dataset D of size $N = \sum_k N_k$

d : dimension of param. θ

$$W_2^2(\text{Law}(\theta_t), \pi) \lesssim C \cdot (1 - \gamma m/2)^t + O(\gamma) \left[N^2 + (\omega + 1)d + \omega \sum_{k=1}^K \|\nabla U_k(\theta^\star)\|^2 \right]$$

contracting term

error term involving :

- discretization error

- variance of stochastic gradients

- compression impact

Convergence analysis for QLSD

Global dataset D of size $N = \sum_k N_k$

d : dimension of param. θ

$$W_2^2(\text{Law}(\theta_t), \pi) \lesssim C \cdot (1 - \gamma m/2)^t + O(\gamma) \left[N^2 + (\omega + 1)d + \omega \sum_{k=1}^K \|\nabla U_k(\theta^\star)\|^2 \right]$$

contracting term

error term involving :

- discretization error
- variance of stochastic gradients
- compression impact

Convergence analysis for QLSD

Global dataset D of size $N = \sum_k N_k$

d : dimension of param. θ

$$W_2^2(\text{Law}(\theta_t), \pi) \lesssim C \cdot (1 - \gamma m/2)^t + O(\gamma) \left[N^2 + (\omega + 1)d + \omega \sum_{k=1}^K \|\nabla U_k(\theta^\star)\|^2 \right]$$

contracting term

error term involving :

2 terms slowing down convergence :

1. N^2 : variance of stochastic gradients

2. $\sum_{k=1}^K \|\nabla U_k(\theta^\star)\|^2$: data heterogeneity

- discretization error
- variance of stochastic gradients
- compression impact

Convergence analysis for QLSD

Global dataset D of size $N = \sum_k N_k$

d : dimension of param. θ

$$W_2^2(\text{Law}(\theta_t), \pi) \lesssim C \cdot (1 - \gamma m/2)^t + O(\gamma) \left[N^2 + (\omega + 1)d + \omega \sum_{k=1}^K \|\nabla U_k(\theta^\star)\|^2 \right]$$

contracting term

error term involving :

- discretization error
- variance of stochastic gradients
- compression impact

2 terms slowing down convergence :

1. N^2 : variance of stochastic gradients
2. $\sum_{k=1}^K \|\nabla U_k(\theta^\star)\|^2$: data heterogeneity

**Convergence improvement
via control variates**

Variance-reduced version of QLSD

$$W_2^2(\text{Law}(\theta_t), \pi) \lesssim C \cdot (1 - \gamma m/2)^t + O(\gamma) \left[N^2 + (\omega + 1)d + \omega \sum_{k=1}^K \|\nabla U_k(\theta^\star)\|^2 \right]$$

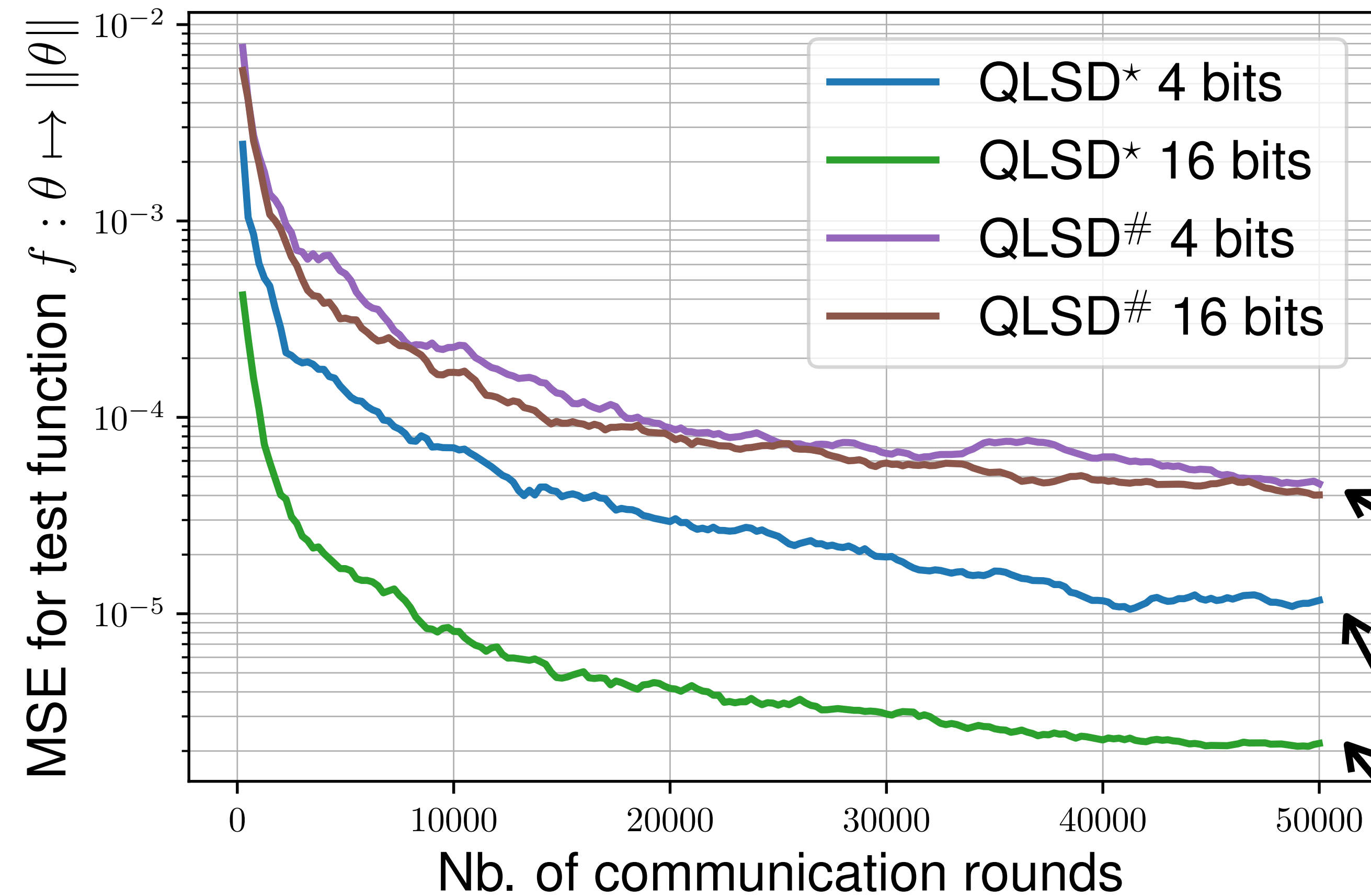
QLSD

versus

$$W_2^2(\text{Law}(\theta_t), \pi) \lesssim C \cdot (1 - \gamma m/2)^t + O(\gamma) (\omega + 1)Nd$$

QLSD with variance reduction

Numerical experiments - Toy Gaussian example



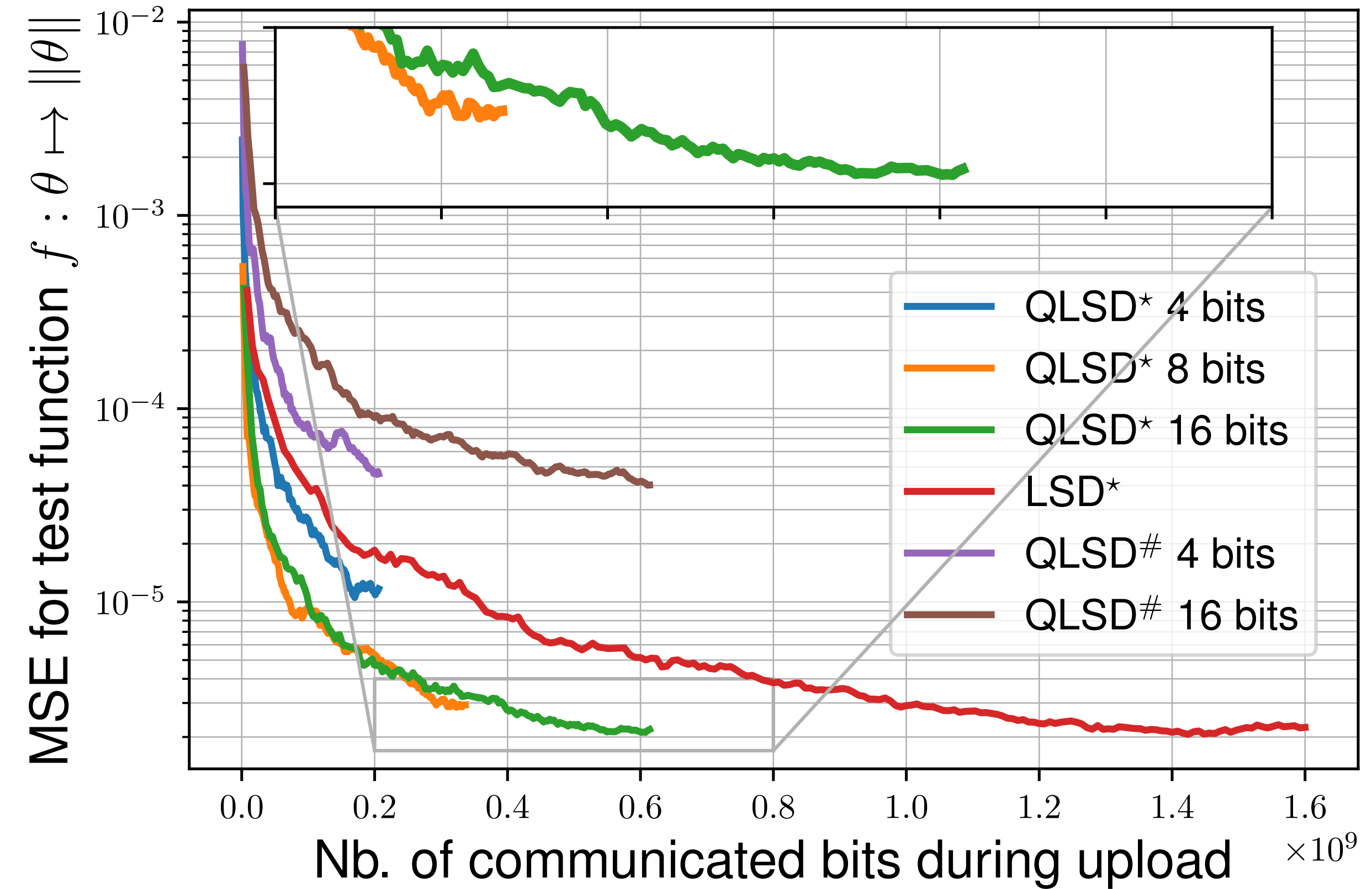
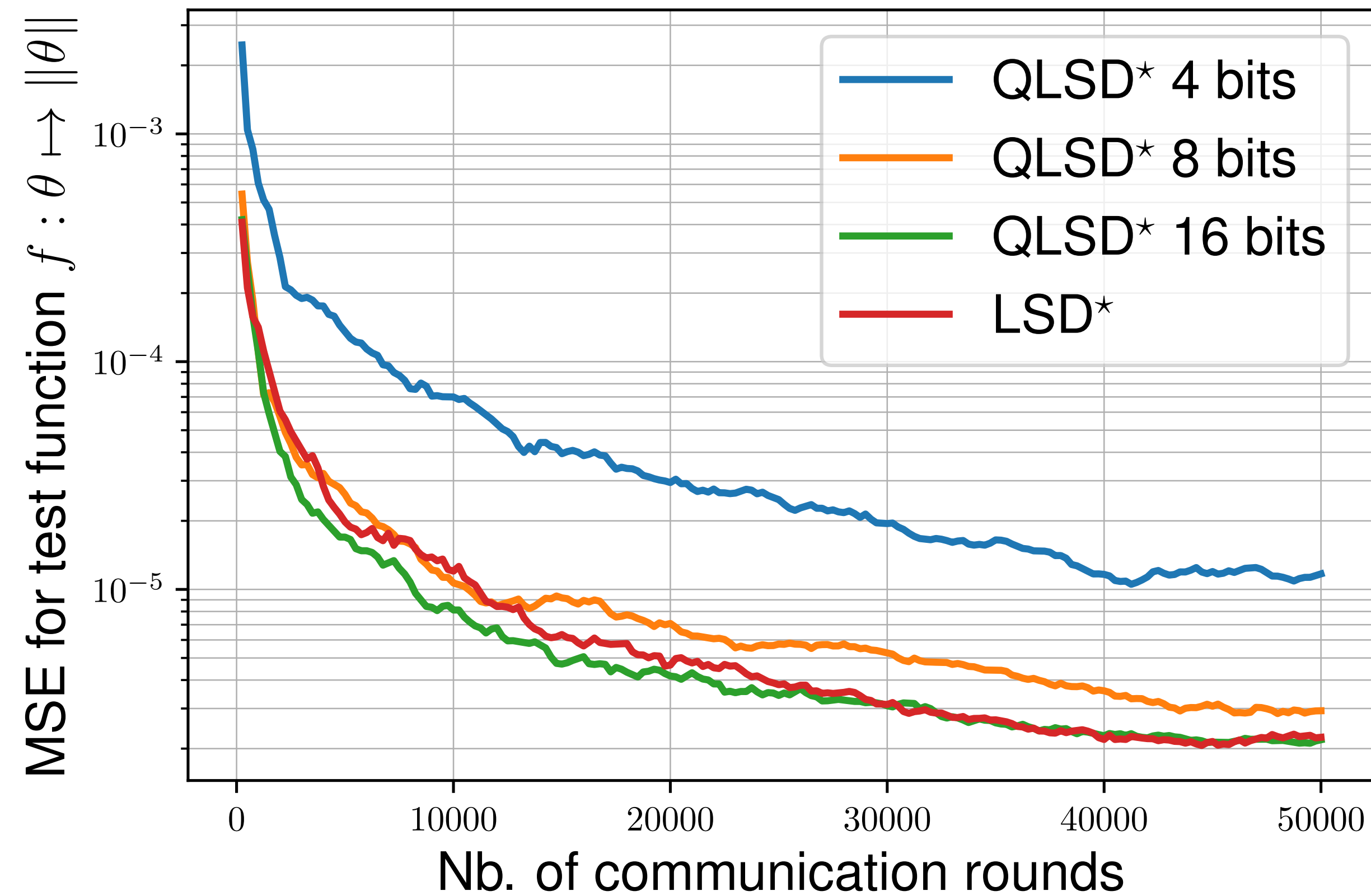
Minibatch size $n = N/10$

$$s \in \{2^4, 2^{16}\}$$

Large impact of the variance of stochastic gradients

Variance reduction helps

Numerical experiments - Toy Gaussian example



QLSD : 2.5 x more efficient than its non-compressed counterpart

Numerical experiments - Bayesian logistic regression

QLSD⁺⁺ : QLSD with SVRG variance reduction scheme + memory mechanism

HPD : highest posterior density region

QLSD ⁺⁺	99% HPD error	Rel. efficiency
4 bits	6.1e-3	7.6
8 bits	4.3e-3	6.7
16 bits	6.9e-4	3.1

FEMNIST dataset

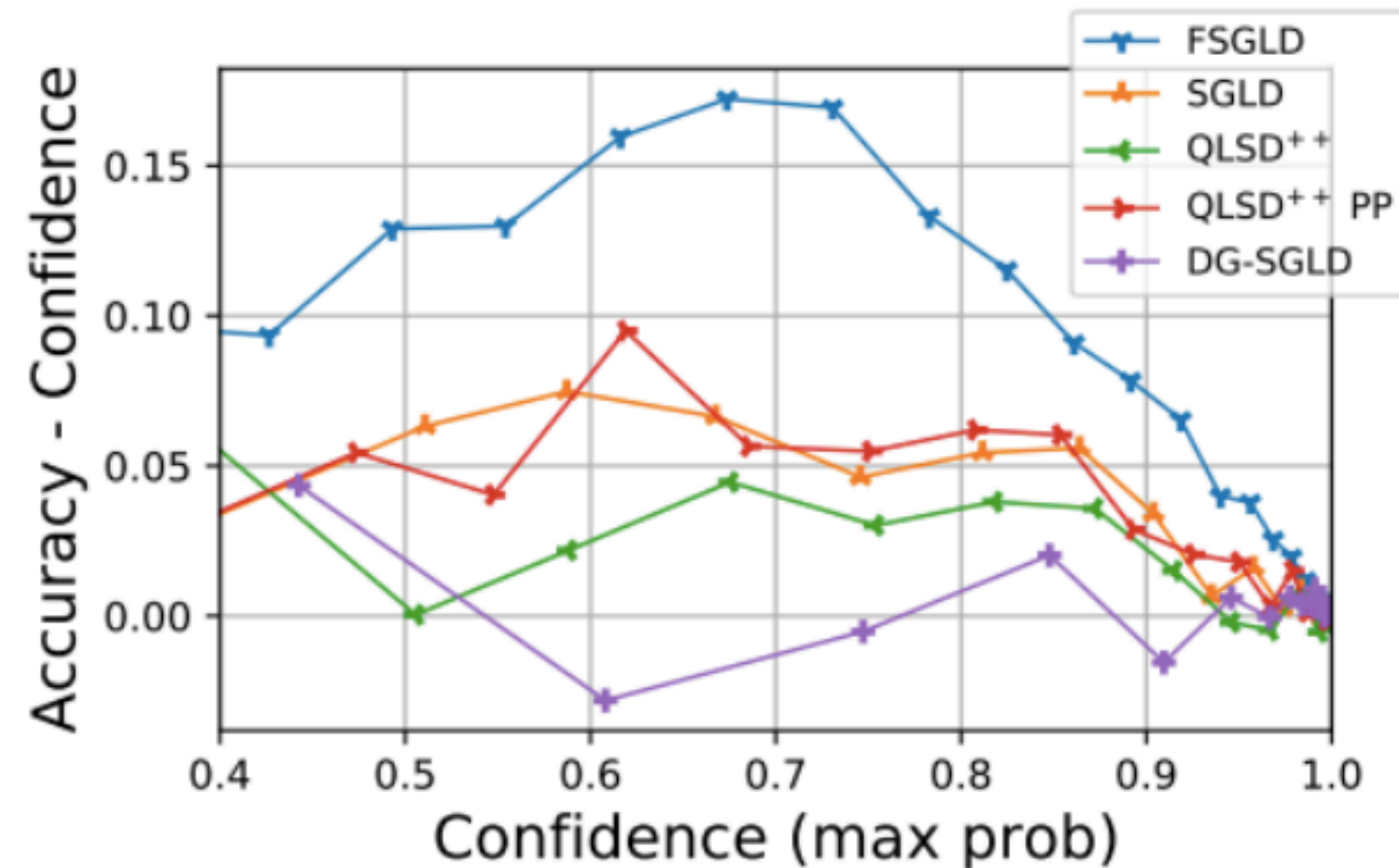
K = 50 clients

d = 784

QLSD : 3 to 7 x more efficient than its non-compressed counterpart

Numerical experiments - Bayesian neural networks

Method	HMC	SGLD	QLSD ⁺⁺	QLSD ⁺⁺ PP	FedBe-Dirichlet	FedBe-Gauss.	DG-SGLD	FSGLD
Accuracy	89.6	88.8	88.1	86.6	90.7	90.2	92.2	87.5
Agreement	0.94	0.91	0.90	0.90	0.90	0.89	0.91	0.91
TV	0.07	0.11	0.12	0.12	0.16	0.16	0.13	0.13



Competitive results compared to the multi-local-steps communication scheme