Randomized Subspace Embeddings for Learning Under Resource Constraints

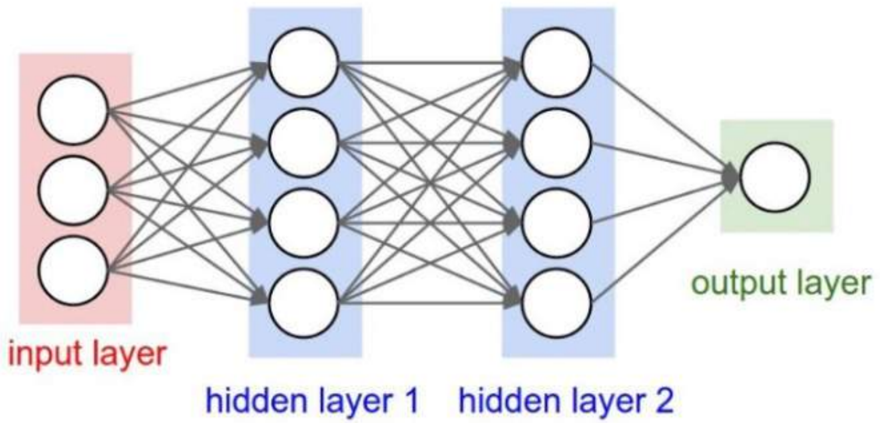
Rajarshi Saha
Electrical Engineering
Stanford University

Joint work with **Mert Pilanci** (Stanford) and **Andrea Goldsmith** (Princeton)

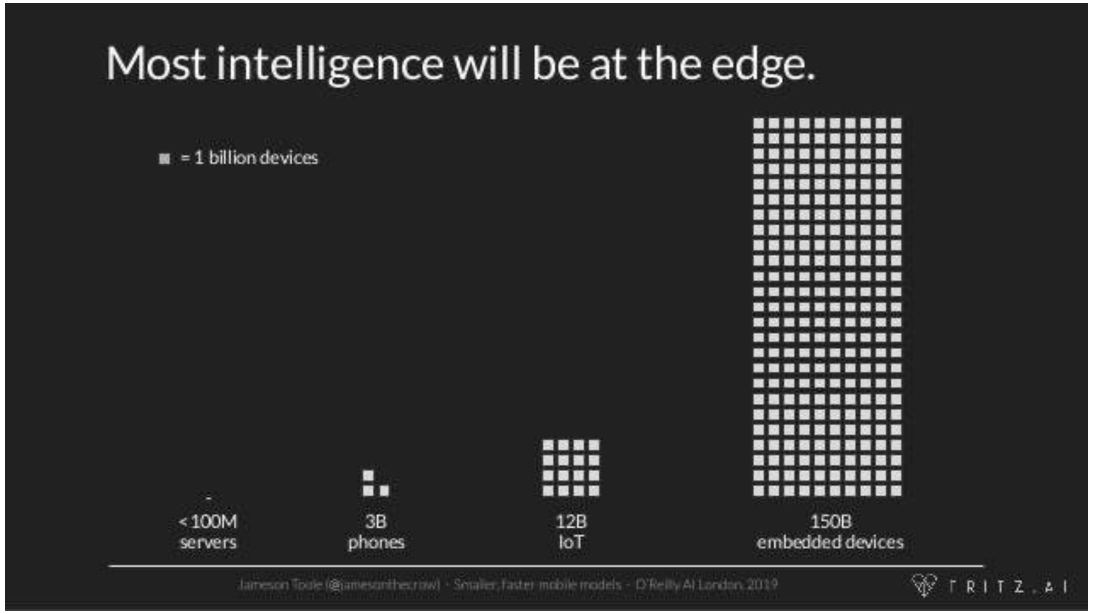
January 24, 2021

More data and Bigger models…

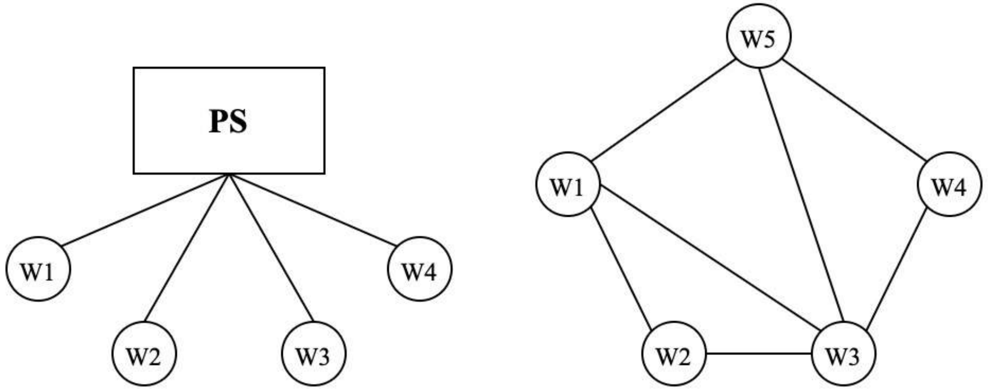
at the expense of Resources: Memory, Computation, Bandwidth, …



huggingface.co/blog/



Training with large distributed datasets

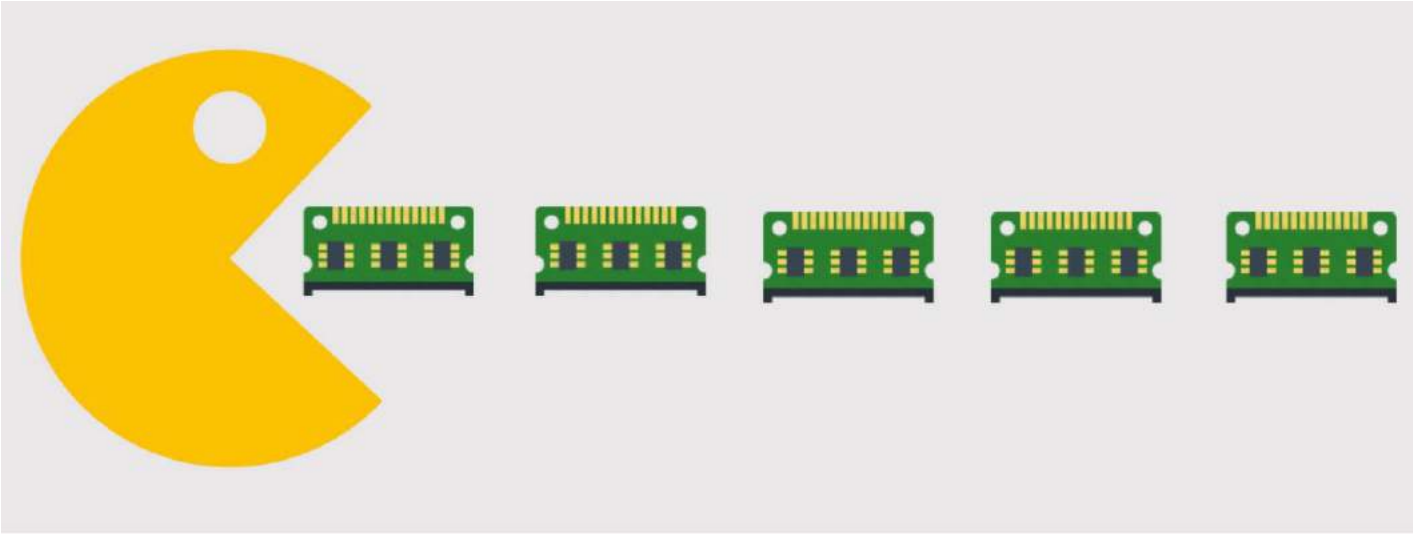


Two pertinent questions:

1. Given a fixed bandwidth allocated for distributed training purposes, what is the information-theoretic limit on how quickly you can train a model?

2. What is an efficient training algorithm that can train a model as fast as (or nearly as fast as) what those limits dictate?

Deploying large models at the edge



(Source: https://miro.medium.com/max/3512/1*d-ZbdlmPx4zRW0zK4QL49w.jpeg)

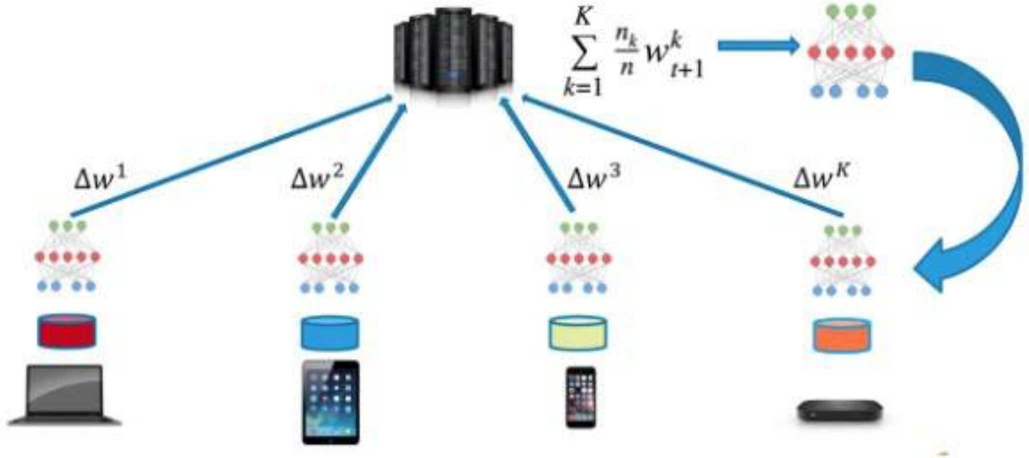
Two more pertinent questions:

1. Given a memory-constraint, what is the information-theoretic limit on the performance when you compress a model?

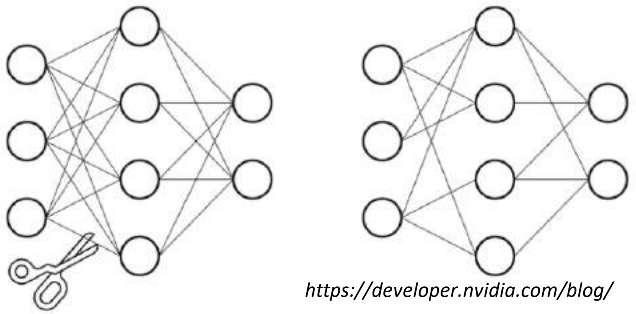
2. What are some efficient algorithms to compress a model so that the performance of the compressed model deteriorates as little as possible?

Vector Quantization

Distributed Learning under Network Bandwidth constraints: Quantize (pseudo) gradients.

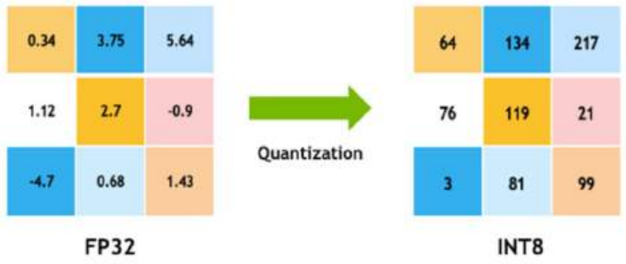


Federated Learning (Source: https://proandroiddev.com/federated-learning-e79e054c33ef)



https://developer.nvidia.com/blog/

Compress/Quantize a Model to deploy on Memory-constrained devices



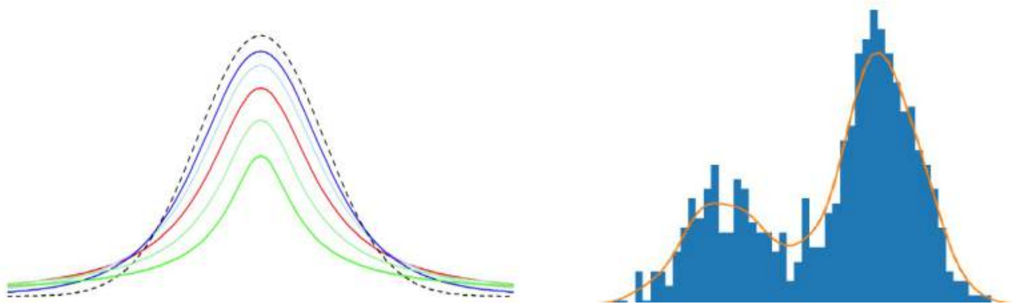
https://towardsdatascience.com/

1	1	2	3	5	8	13	21	34	55	89	144

Need a practical and efficient vector quantization scheme!

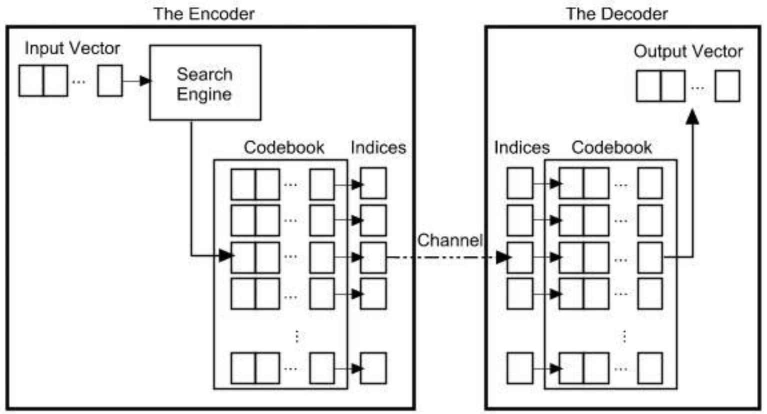
VQ for Learning: Challenges

- **VQ must be agnostic to any distributional information.**
  - Except for very well-structured problems with several assumptions, statistical information about the vector entries are not known.
  - **Fit a distribution?** Computationally intensive. Weights and gradients are constantly changing.



- **Universal Vector Quantization:** Do not want a complicated lattice. Ideally, complexity should be linear in dimension.

- **Lossy Source Coding:** Codebook should be easily available to decoder.



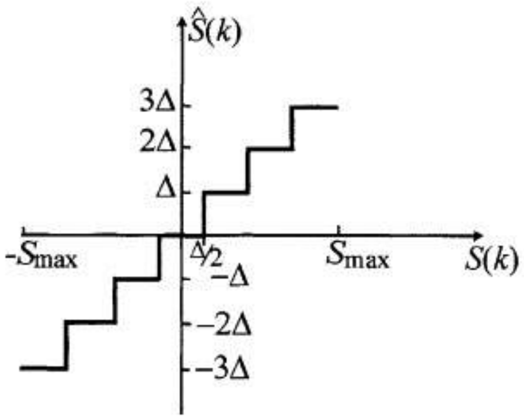
Given a bit-budget of B bits per dimension, how do we quantize a vector in R^d ?

The problem of bit-allocation

- B-bits per dimension $\Longrightarrow$ dB bits to quantize.
- **How to allocate dB bits to d coordinates?**
- **Is it worth designing a sophisticated bit allocation scheme?**
    - Vectors are constantly changing.
    - Hardware implementation of non-uniform quantizers is difficult.

$$\begin{bmatrix} 16 \\ 1 \\ 0.01 \\ \vdots \\ 5 \end{bmatrix}$$

Orders of magnitude difference.



Uniform Quantizers

$\|\mathbf{x}\|_\infty = 1$ and B bits per dimension

$\Longrightarrow 2^B$ points per coordinate given by $v_i = -1 + (2i-1)\Delta/2,\ i = 1, \ldots M,\ \Delta = 2/M.$

$$\mathrm{Q}(\mathbf{x}) = [x'_1, \ldots, x'_N]^\top;\ x'_j \triangleq \underset{y \in \{v_1, \ldots, v_M\}}{\arg\min.} |y - x_j|$$

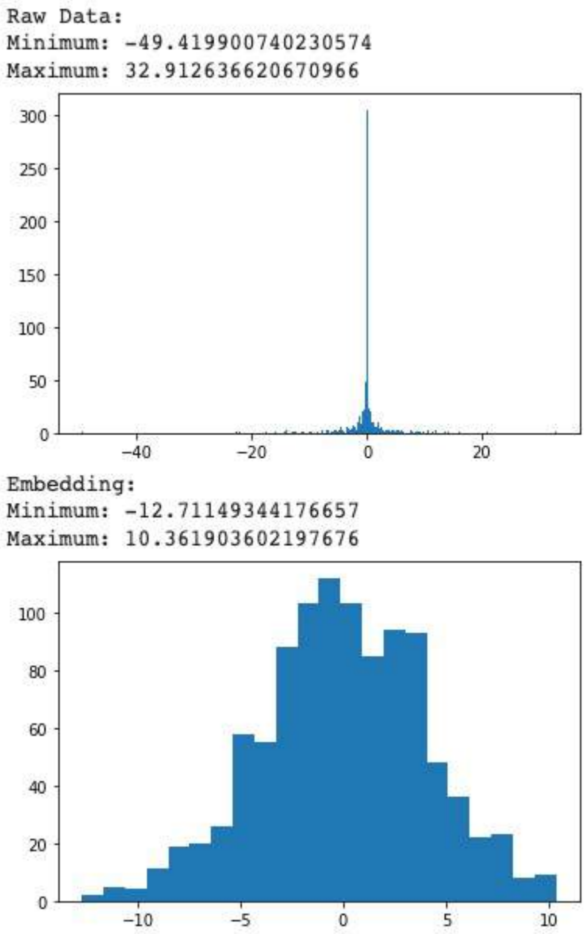
$$\sup_{\mathbf{x} \in B^d_\infty(1)} \|\mathrm{Q}(\mathbf{x}) - \mathbf{x}\|_2 = \frac{\Delta}{2}\sqrt{d}$$

How do Random Embeddings help?

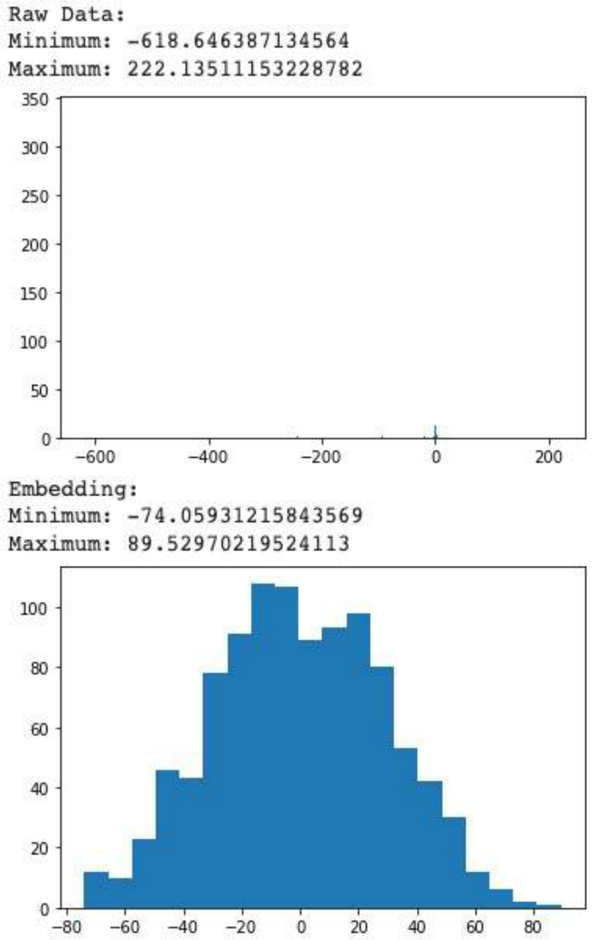
y: A vector in R^d whose coordinates can be arbitrarily large.

Random embedding from R^d to R^D

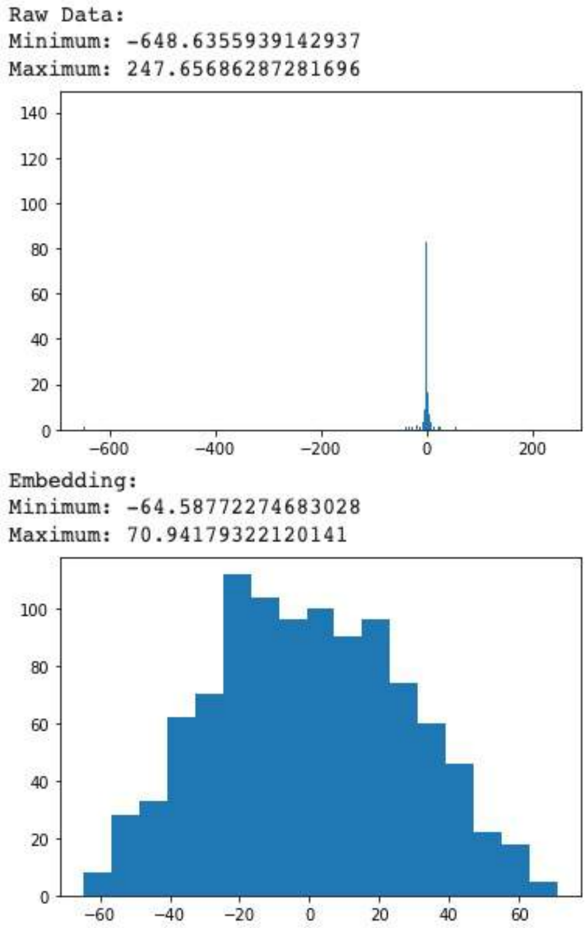
x: A vector in R^D whose coordinates are equalized.



Gaussian^3

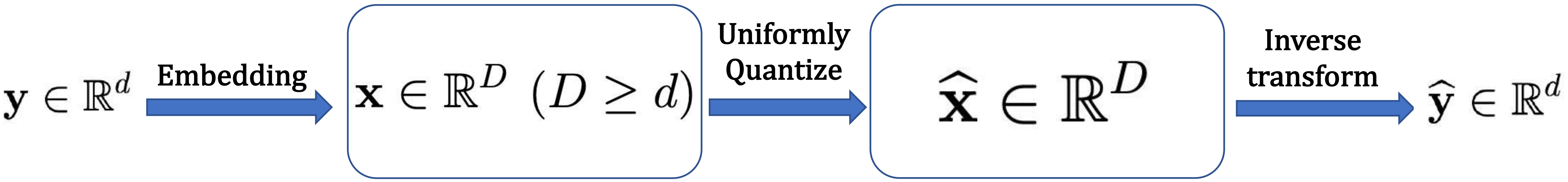


Gaussian^5



Student-t (df = 1)

Quantizing the Random Embeddings



With randomized embeddings

$$\sup_{\mathbf{x} \in B_\infty^d(1)} \|\mathrm{Q}(\mathbf{x}) - \mathbf{x}\|_2 = O(1)$$

(Computational complexity: $O(d^2)$)

$$\sup_{\mathbf{x} \in B_\infty^d(1)} \|\mathrm{Q}(\mathbf{x}) - \mathbf{x}\|_2 = O(\sqrt{\log d})$$

(Computational complexity: O(d log d))

Worst-case quantization error is dimension-independent or weak-logarithmic dependence!

Part 1
Model Compression

Compressing Linear Models

$$\mathbf{X} = \mathbf{W}\boldsymbol{\theta} + \mathbf{v}$$

Observations $\in \mathbb{R}^n$

Noise $\in \mathbb{R}^n$

Arbitrary measurement matrix $\in \mathbb{R}^{n \times d}$

Ground-truth model $\in \mathbb{R}^d$

Worker estimates model $\boldsymbol{\theta}$ and can send it to the server **using only** dB bits.

$$\widetilde{\boldsymbol{\theta}} := \arg\min_{\mathbf{s} \in \boldsymbol{\mathcal{S}}} \|\mathbf{X} - \mathbf{W}\mathbf{s}\|_2^2$$

$$R(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \mathbb{E}_{\mathbf{X}}\left[\frac{1}{d}\left\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}\right\|_2^2\right]$$

(Risk of any quantized model)

Information-Theoretic Limits

Definition

An (n, d, B)-**learning code** $\mathrm{Q} : \mathbb{R}^n \to \Theta$ is defined to be the composition of encoder and decoder mappings E and D, such that for any given data $\mathrm{X} \in \mathbb{R}^n$, $\mathrm{Q}(\mathrm{X}) \equiv \mathrm{D}(\mathrm{E}(\mathrm{X})) \in \Theta$.

Minimax risk:

$$\mathcal{R}_{\mathbf{W},B,\sigma.c} := \liminf_{d\to\infty} \inf_{\mathrm{Q}\in\mathcal{Q}_{n,d,B}} \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} R(\mathrm{Q}(\mathbf{X}), \boldsymbol{\theta})$$

Theorem

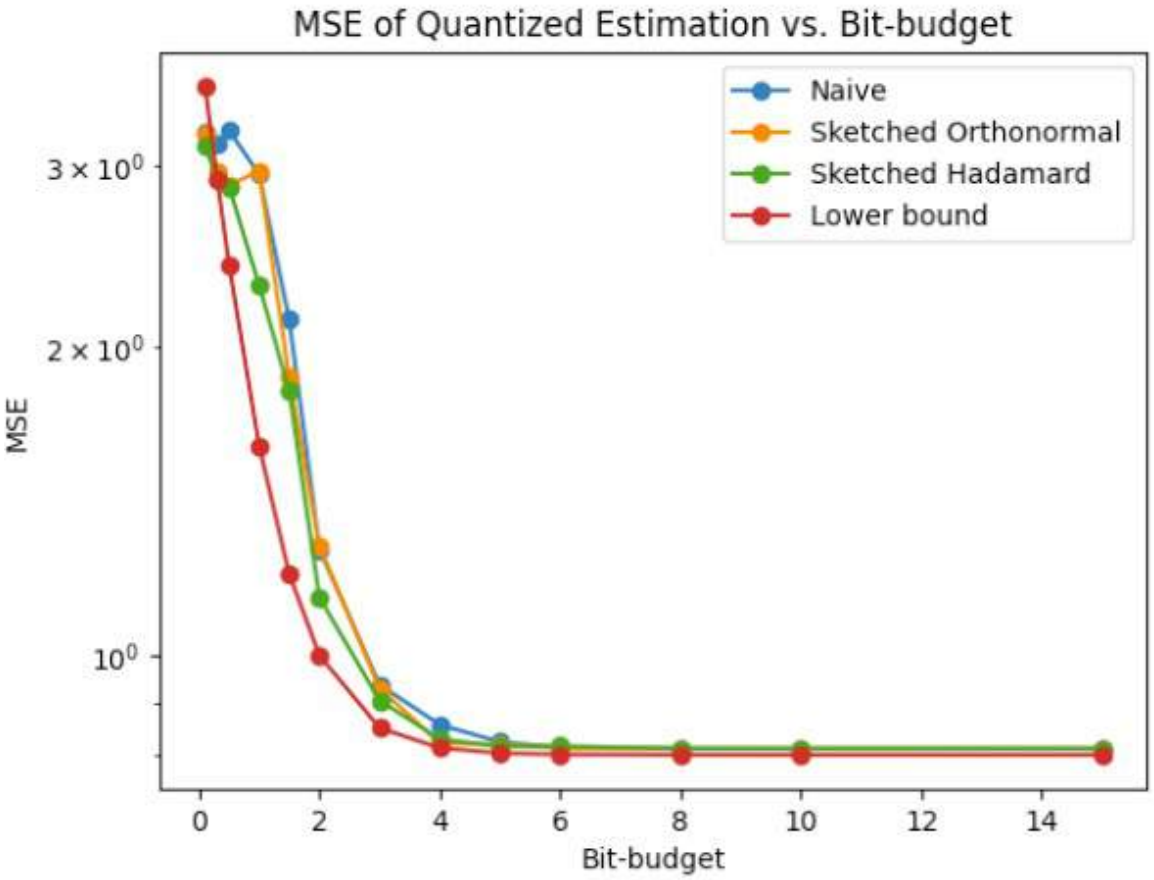
For $B > 0, \sigma > 0, c > 0$, and $\mathrm{W} \in \mathbb{R}^{n\times d}$ with minimum and maximum singular values as σ_m and σ_M respectively, the asymptotic minimax risk can be lower bounded as:

$$\mathcal{R}_{\mathrm{W},B,\sigma,c} \geq \frac{c^2\sigma^2}{\sigma^2 + c^2\sigma_M^2} + \frac{c^4\sigma_m^2}{\sigma^2 + c^2\sigma_m^2} \cdot 2^{-2B}.$$

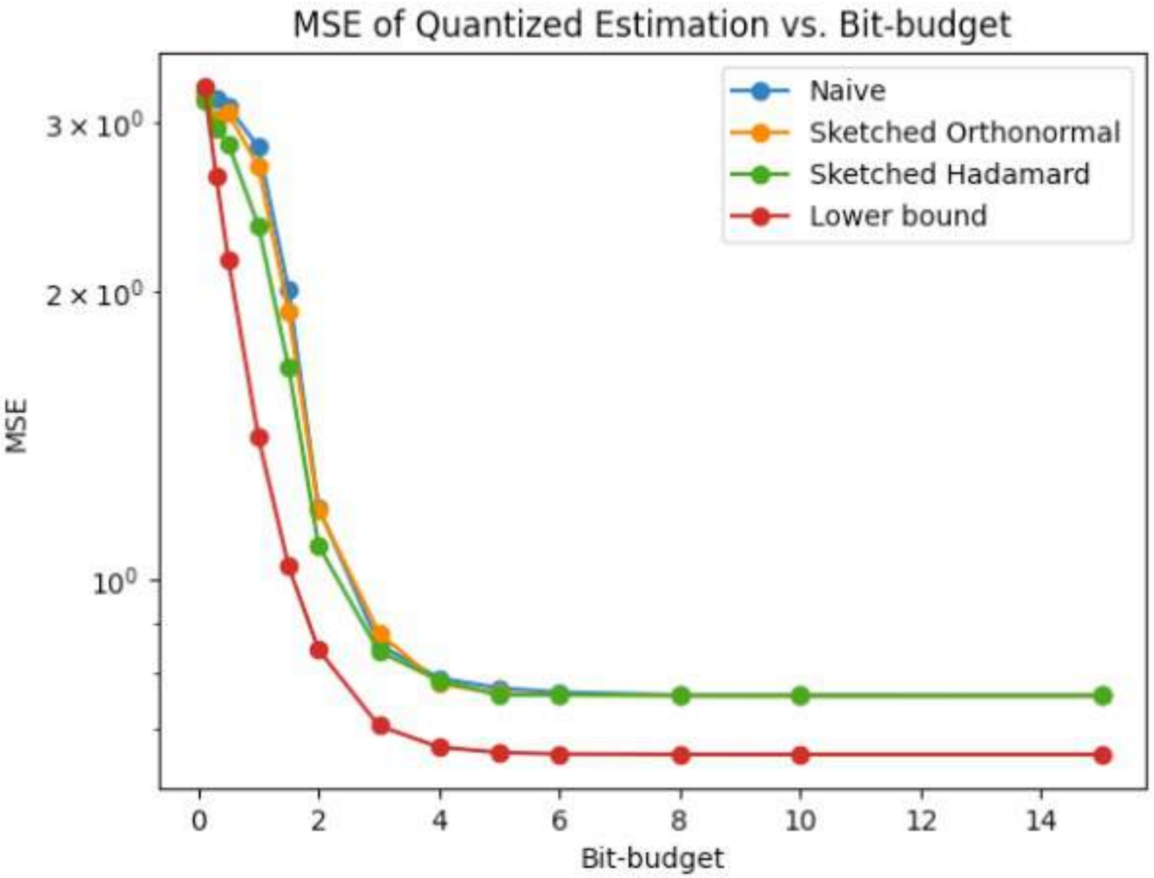
Optimally Compressing Linear Models

Learning Codes	Performance Guarantee (holds w.h.p.)	Computational Complexity	Remarks
Random Projections on the Unit Sphere	$R\left(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}\right) \leq \frac{c^2\sigma^2}{\sigma^2+c^2\sigma_{min}^2} + \frac{c^4\sigma_{max}^2}{\sigma^2+c^2\sigma_{max}^2}2^{-2B}$	**exp (d)**	Tight w.r.t. lower bound.
Democratic Quantized Estimation	$R\left(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}\right) \leq \frac{2c^2\sigma^2}{\sigma^2+c^2\sigma_{min}^2} + \frac{16K_u c^4\sigma_{max}^2}{\sigma^2+c^2\sigma_{max}^2}2^{-\frac{2B}{\lambda}}$	**O (d^2)**	Optimal within constant factors.
Near-Democratic Quantized Estimation	$R\left(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}\right) \leq \frac{2c^2\sigma^2}{\sigma^2+c^2\sigma_{min}^2} + \frac{32\sqrt{\log(2d)}c^4\sigma_{max}^2}{\sigma^2+c^2\sigma_{max}^2}2^{-\frac{2B}{\lambda}}$	**O (d · log d)**	Near linear-time; Mild logarithmic dependence.

How tight are the Lower and Upper bounds?

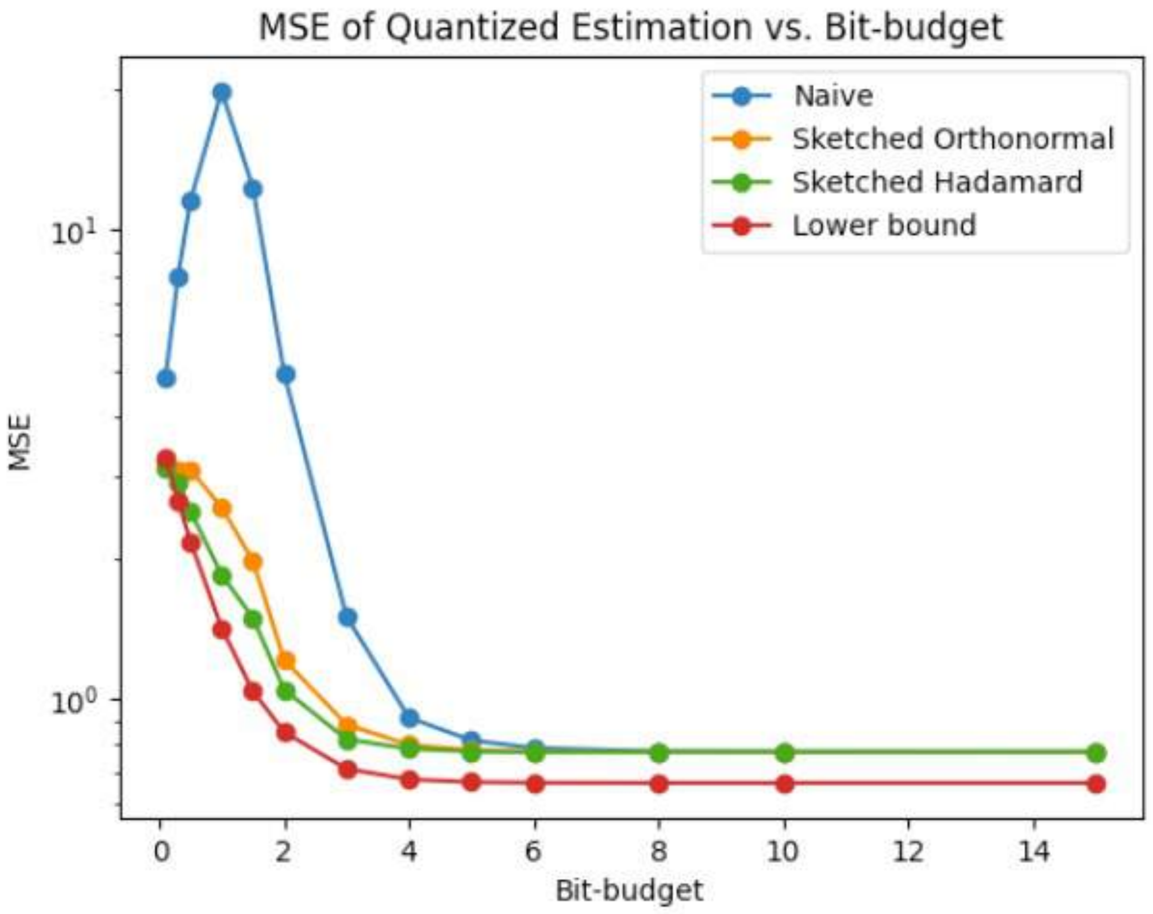


W: Identity, $\boldsymbol{\theta}$: Gaussian

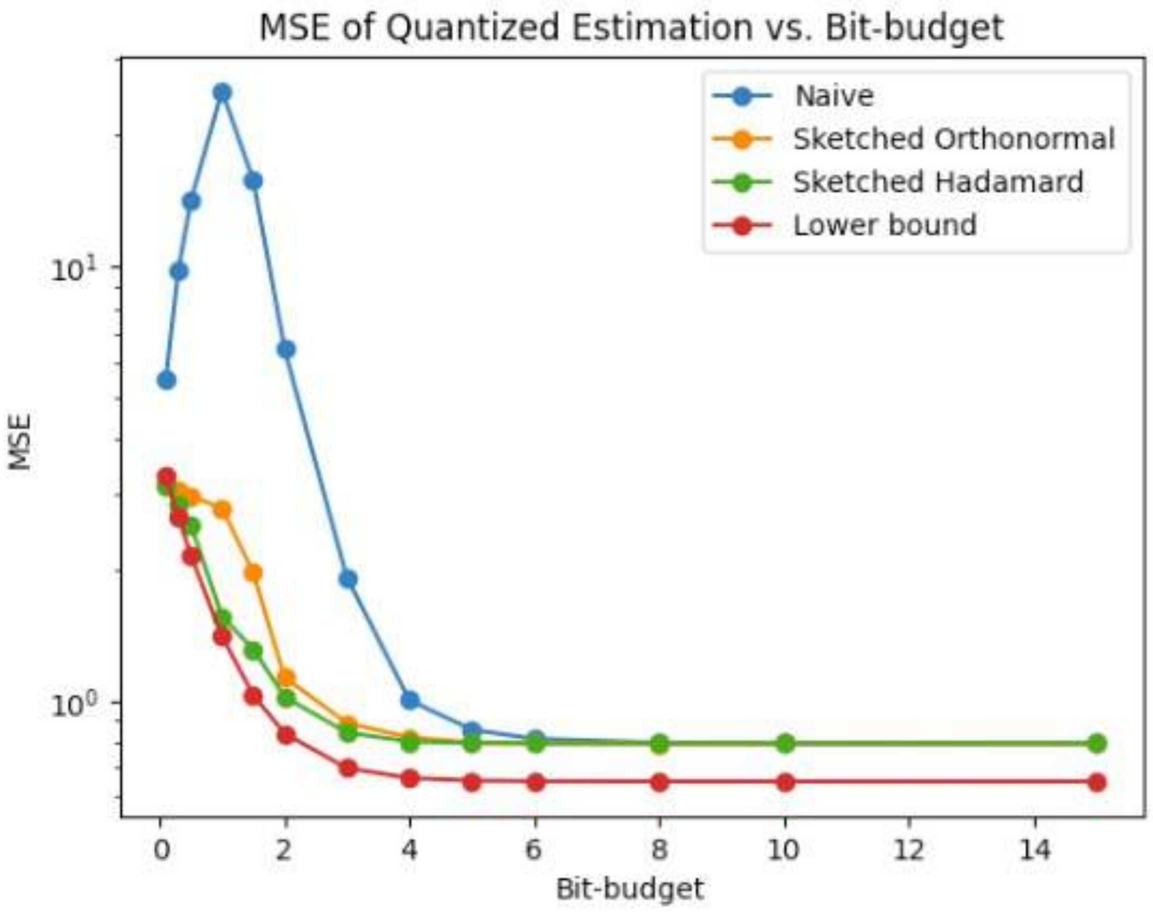


W: Perturbed orthonormal, $\boldsymbol{\theta}$: Gaussian

Compressing Heavy-Tailed Models



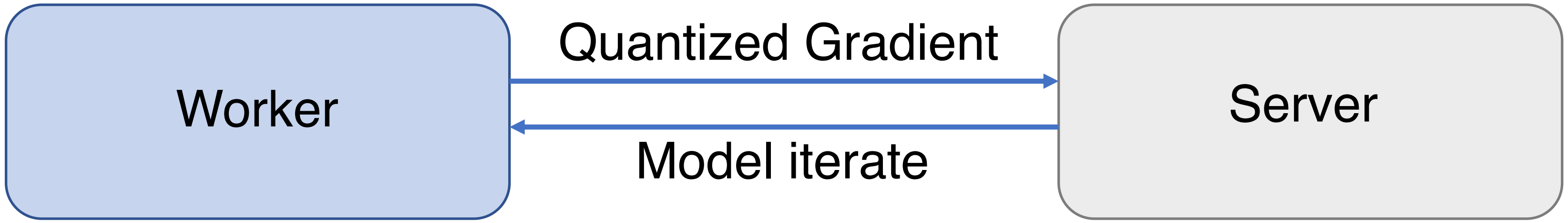
W: Perturbed orthonormal, $\boldsymbol{\theta}$: Gaussian[3]



W: Perturbed orthonormal, $\boldsymbol{\theta}$: Student-t (df = 1)

Part 2
Communication-Constrained Distributed Optimization

Iterative First-Order Optimization Protocols



- How to design **efficient algorithms** to achieve the optimal convergence rate when the worker can communicate to the server using **only dB bits per round**?

L - smooth and μ - strongly convex objectives

Minimax convergence rate:

$$C(B) \triangleq \inf_{\pi \in \Pi_B} \limsup_{T \to \infty} \sup_{f \in \mathcal{F}_{\mu,L,D}} \left(\frac{\left\| \mathbf{x}_T(\pi) - \mathbf{x}_f^* \right\|_2}{D} \right)^{\frac{1}{T}}$$

Information-theoretic limit
("Differentially Quantized Gradient Methods", Chung-Yi Lin and Victoria Kostina and Babak Hassibi, 2021)

$$C(B) \geq \max\{\sigma, 2^{-B}\}$$

Optimization Algorithm	Performance Guarantee	Computational Complexity	Remarks
DQ-PSGD	$\left(\frac{\mathbf{x}_T - \mathbf{x}_f^*}{D}\right)^{\frac{1}{T}} \leq \max\{\sigma, c_1 \cdot 2^{-B}\}$	**O (d^2)**	Optimal within constant factors.
Near DQ-PSGD	$\left(\frac{\mathbf{x}_T - \mathbf{x}_f^*}{D}\right)^{\frac{1}{T}} \leq \max\{\sigma, c_2\sqrt{\log d} \cdot 2^{-B}\}$	**O (d · log d)**	Near linear-time; Mild logarithmic dependence.

General convex and non-smooth objectives

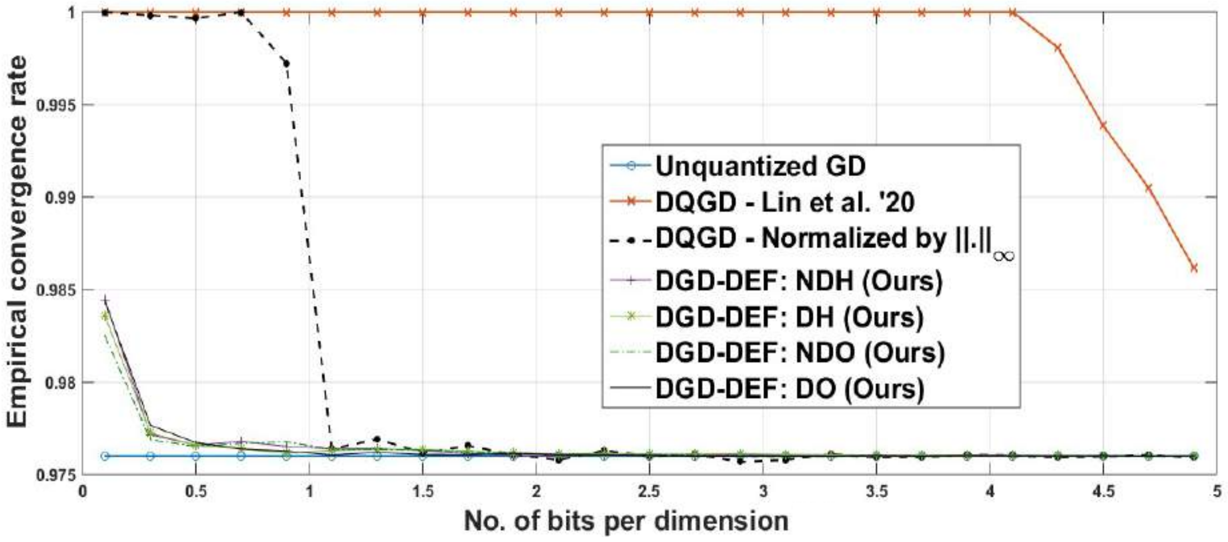
Minimax suboptimality gap: $$\mathcal{E}(T,B) \triangleq \inf_{\pi \in \Pi_{T,B}} \sup_{(f,\mathcal{O})} \mathbb{E} f(\mathbf{x}(\pi)) - f(\mathbf{x}^*)$$

Information-theoretic limit
("Limits on Gradient Compression for Stochastic Optimization" Prathamesh Mayekar and Himanshu Tyagi, 2020)

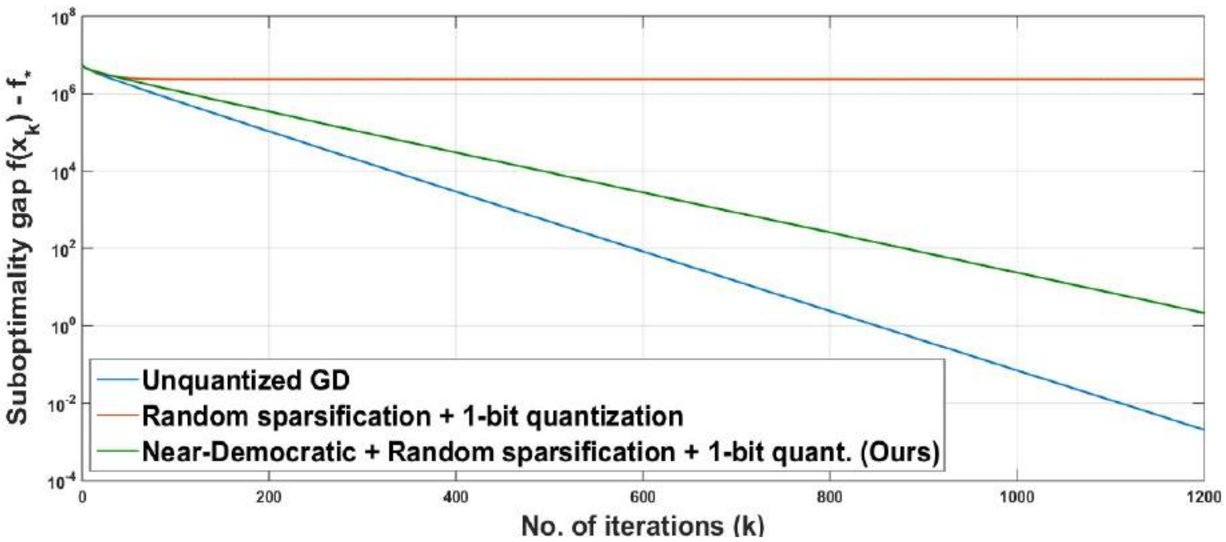
$$\mathcal{E}(T,B) \geq \frac{cD\sigma}{\sqrt{T}\sqrt{\min\{1,B\}}}$$

Optimization Algorithm	Performance Guarantee	Computational Complexity	Remarks
DQ-PSGD	$\mathcal{E}(T,B) \leq \frac{c_1 D\sigma}{\sqrt{T}\sqrt{\min\{1,B\}}}$	**O (d^2)**	Optimal within constant factors.
Near DQ-PSGD	$\mathcal{E}(T,B) \leq \frac{c_2 D\sigma\sqrt{\log d}}{\sqrt{T}\sqrt{\min\{1,B\}}}$	**O (d · log d)**	Near linear-time; Mild logarithmic dependence.

Numerical Results

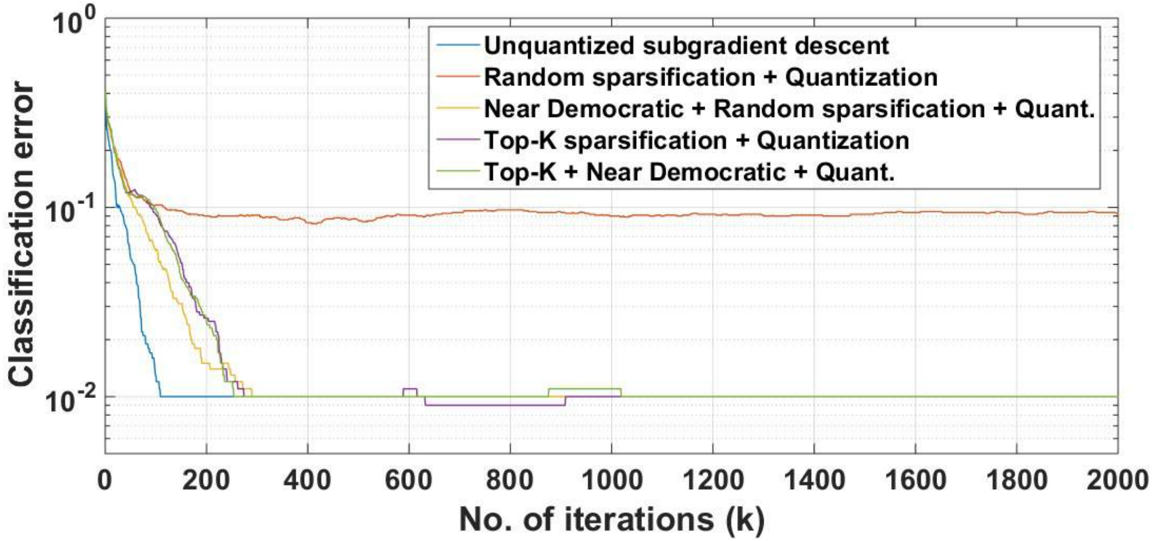


Least squares: Synthetic data

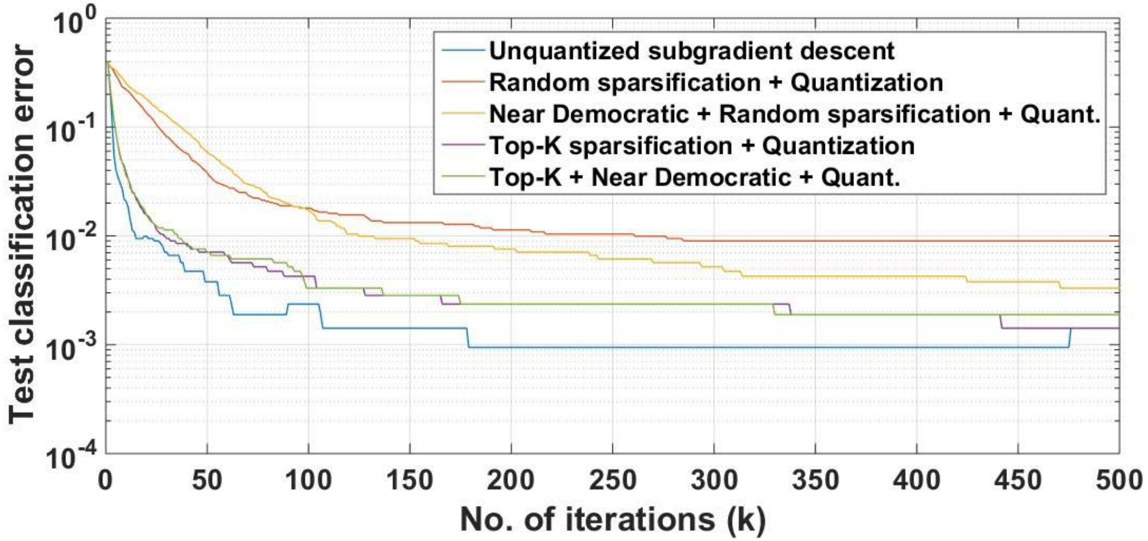


Least squares: MNIST

Numerical Results (contd..)

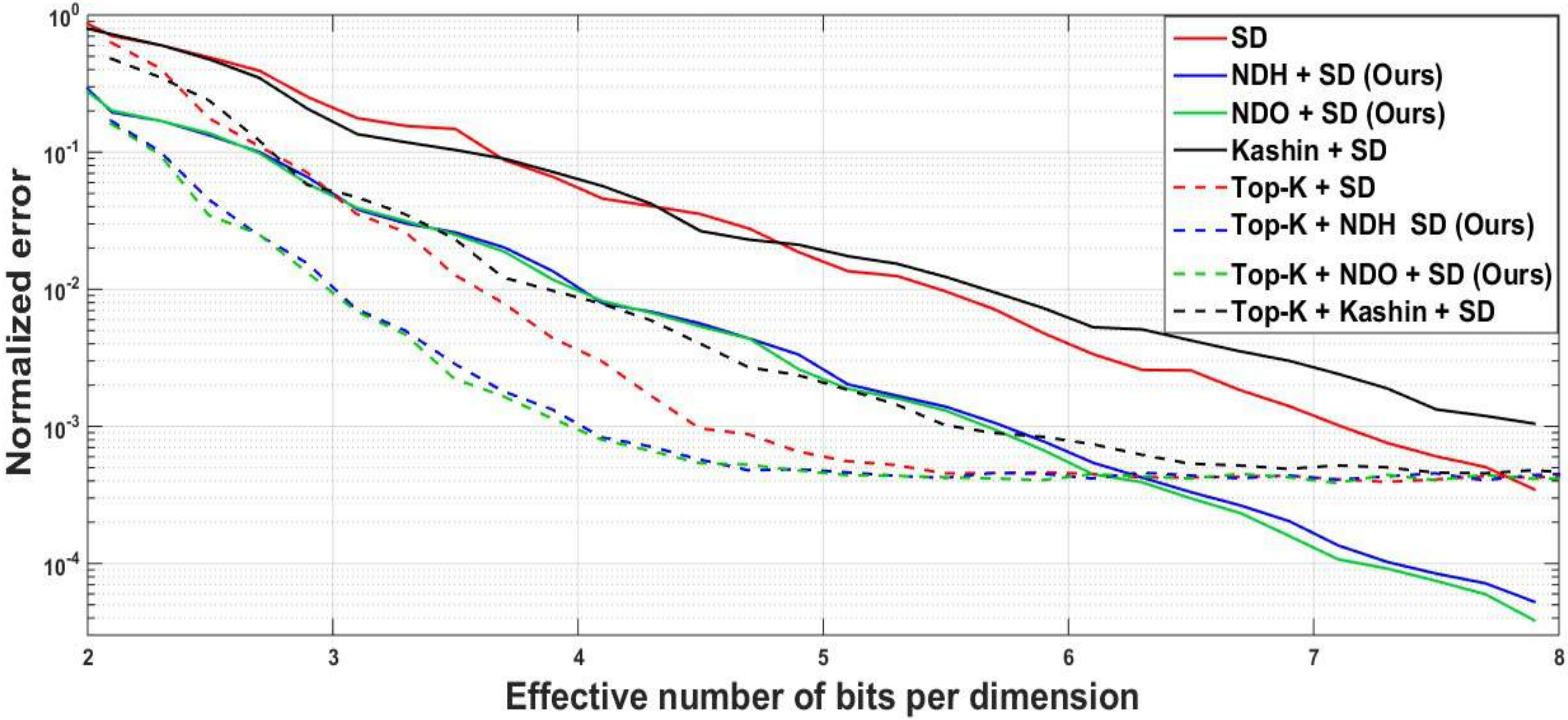


Support Vector Machine: Synthetic data



Support Vector Machine: MNIST

General stochastic compression schemes



Thank you!