Differentially Quantized Gradient Descent

Chung-Yi Lin   Victoria Kostina   Babak Hassibi

Department of Electrical Engineering
California Institute of Technology

IEEE International Symposium on Information Theory
May 28, 2021

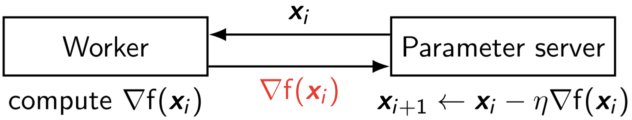
Unquantized gradient descent (GD)

$$\boxed{\text{Solve } \boldsymbol{x}^* = \arg\min_{\boldsymbol{x}\in\mathbb{R}^n} \mathrm{f}(\boldsymbol{x})}$$

- with GD:

$$\boldsymbol{x}_{i+1} \leftarrow \boldsymbol{x}_i - \eta \nabla \mathrm{f}(\boldsymbol{x}_i)$$

  - $\eta > 0$ is a constant stepsize.
- in distributed training



Exchanging the gradients carries high communication cost.

Previous works

Gradient quantization

- Stochastic gradient descent
  - Seide et al. 2014
  - Wen et al. 2017
  - Alistarh et al. 2017
  - Bernstein et al. 2018
  - Wu et al. 2018
  - Gandikota et al. 2019
  - Ramezani-Kebrya et al. 2019
  - Mayekar & Tyangi 2019, 2020
- GD
  - Luo & Tseng 1993
  - Friedlander & Schmidt 2012
  - Alistarh et al. 2016

Gradient sparsification

- Strom et al. 2015
- Aji & Heafield 2017
- Lin et al. 2018
- Wangni et al. 2018
- Wang et al. 2018
- Stich et al. 2018

Convergence lower bounds are scarse.

This work: the necessary and sufficient bit rate to achieve a target convergence rate.

Class of functions

$$\mathcal{F}_n(\mu, L, r) \triangleq \{\mathrm{f} \colon \mathbb{R}^n \to \mathbb{R} \mid \text{f satisfies the following.}\}$$

- f is L-smooth: $\|\nabla \mathrm{f}(\boldsymbol{v}) - \nabla \mathrm{f}(\boldsymbol{w})\| \leq L \|\boldsymbol{v} - \boldsymbol{w}\|$
- f is μ-strongly convex: $(\nabla \mathrm{f}(\boldsymbol{v}) - \nabla \mathrm{f}(\boldsymbol{w}))^{\mathrm{T}} (\boldsymbol{v} - \boldsymbol{w}) \geq \mu \|\boldsymbol{v} - \boldsymbol{w}\|^2$
- The minimizer $\|\boldsymbol{x}^*(\mathrm{f})\| \leq r$ for some $r > 0$

Unquantized gradient descent

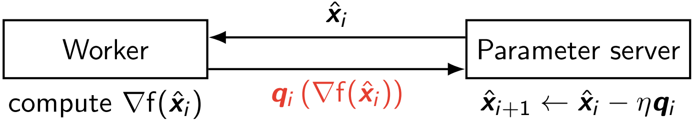
For *any* L-smooth and μ-strongly convex f, unquantized GD satisfies [Polyak, 1987]:

$$\|\boldsymbol{x}_T - \boldsymbol{x}^*(\mathrm{f})\| \leq \sigma^T \|\boldsymbol{x}_0 - \boldsymbol{x}^*(\mathrm{f})\|$$

- $\sigma \triangleq \frac{L-\mu}{L+\mu}$: contraction factor of GD.
- The bound is tight: $\exists$f s.t. "$=$" holds.

Naively Quantized Gradient Descent (NQ-GD)

NQ-GD [Friedlander & Schmidt 2012, Alistarh et al. 2016] directly quantizes the gradient at $\hat{\boldsymbol{x}}_i$:



Theorem: NQ-GD

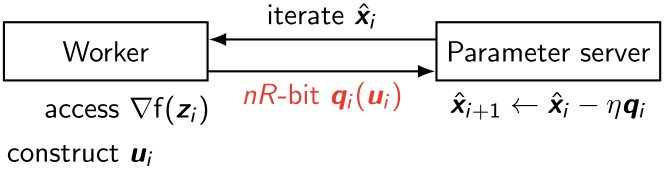
NQ-GD achieves the following contraction factor over $\mathcal{F}_n$

$$\sigma_{\text{NQ-GD}}(n, R) \leq \sigma + \rho_n 2^{-R}$$

ρ_n: covering efficiency of the quantizer

- uniform scalar quantizer: $\rho_n = \sqrt{n}$.
- $\rho_n \geq 1$.
- $\rho_n = 1 + o_n(1)$ is achievable with lattice quantizers [Rogers 1963].

Quantized gradient descent (QGD)



The worker, based on $\hat{\boldsymbol{x}}_i$ and $\boldsymbol{e}_0, \ldots, \boldsymbol{e}_{i-1}$ $(\boldsymbol{e}_\ell \triangleq \boldsymbol{q}_\ell - \boldsymbol{u}_\ell)$, decides:

- gradient query point $\boldsymbol{z}_i$
- quantizer's input $\boldsymbol{u}_i$.

Goals:

- characterize the tradeoff between how fast any QGD algorithm converges and R.
- propose an algorithm that achieves it.

Quantized Gradient Descent: worst-case contraction factor

For a QGD algorithm A operating at R bits per problem dimension, worst-case (over $\mathsf{f} \in \mathcal{F}_n(\mu, L, r)$) contraction factor:

$$\sigma_{\mathrm{A}}(n, R) \triangleq \sup_{\mathsf{f} \in \mathcal{F}_n} \limsup_{T \to \infty} \|\hat{\mathbf{x}}_T(R) - \mathbf{x}^*(\mathsf{f})\|^{\frac{1}{T}}$$

- unquantized GD: $\sigma_{\mathrm{GD}}(n, \infty) = \sigma = \frac{L-\mu}{L+\mu}$.

Converse

Theorem: converse

The contraction factor of any QGD algorithm A operating at R bits per problem dimension satisfies

$$\sigma_{\mathrm{A}}(n, R) \geq \max\left\{\sigma,\ 2^{-R}\right\}$$

Proof combines two converses:

- Reduction to unquantized GD: $\sigma_{\mathrm{A}}(n, R) \geq \sigma$;
- Volume division argument: $\sigma_{\mathrm{A}}(n, R) \geq 2^{-R}$.

Differentially Quantized Gradient Descent (DQ-GD)

Algorithm 1: DQ-GD

Initialize $\boldsymbol{e}_{-1} \leftarrow \boldsymbol{0}$

for $i = 0$ **to** $T-1$ **do**

  Worker:

    $\boldsymbol{z}_i \leftarrow \hat{\boldsymbol{x}}_i + \eta \boldsymbol{e}_{i-1}$

    $\boldsymbol{u}_i \leftarrow \nabla \mathrm{f}(\boldsymbol{z}_i) - \boldsymbol{e}_{i-1}$

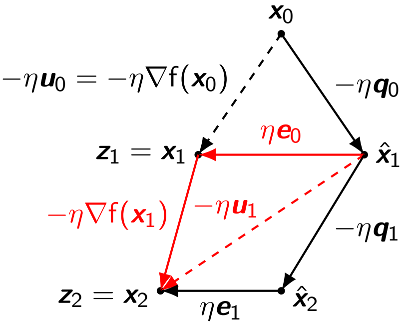
    $\boldsymbol{q}_i = \mathrm{q}_i(\boldsymbol{u}_i)$

    $\boldsymbol{e}_i \leftarrow \boldsymbol{q}_i - \boldsymbol{u}_i$

  Parameter server:

    $\hat{\boldsymbol{x}}_{i+1} \leftarrow \hat{\boldsymbol{x}}_i - \eta \boldsymbol{q}_i$

end

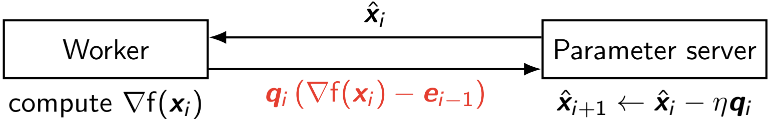


Differential quantization[1] directs the quantized trajectory to the unquantized trajectory.

[1] The idea of error compensation dates back to ΣΔ modulation [Gray, 1989].

Differentially Quantized Gradient Descent (DQ-GD)

DQ-GD computes the gradient at $\boldsymbol{x}_i$ and compensates the previous quantization error:



Theorem: DQ-GD

DQ-GD achieves the following contraction factor over $\mathcal{F}_n$

$$\sigma_{\text{DQ-GD}}(n, R) \leq \max\left\{\sigma,\ \rho_n 2^{-R}\right\}$$

- Since $\rho_n \to 1$ is achievable [Rogers, 1963], DQ-GD attains the converse as $n \to \infty$
- $R \geq \log_2 \rho_n/\sigma$: achieves the contraction factor σ of unquantized GD
- $R < \log_2 \rho_n/\sigma$: achieved contraction factor is only $\rho_n 2^{-R}$

Proof sketch

$$\boxed{\text{Induction: } \hat{\boldsymbol{x}}_i = \boldsymbol{x}_i - \eta \boldsymbol{e}_{i-1}}$$

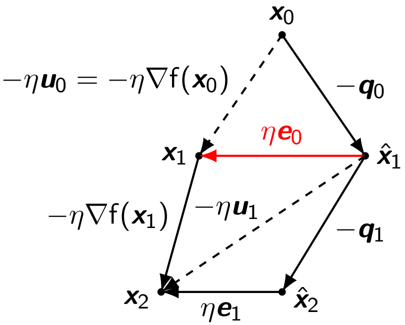
- $\|\hat{\boldsymbol{x}}_i - \boldsymbol{x}^*(\mathrm{f})\| \leq \|\boldsymbol{x}_i - \boldsymbol{x}^*(\mathrm{f})\| + \eta \|\boldsymbol{e}_{i-1}\|$
- 1st term: convergence of GD

$$\|\boldsymbol{x}_T - \boldsymbol{x}^*(\mathrm{f})\| \leq \sigma^T \|\boldsymbol{x}_0 - \boldsymbol{x}^*(\mathrm{f})\| .$$

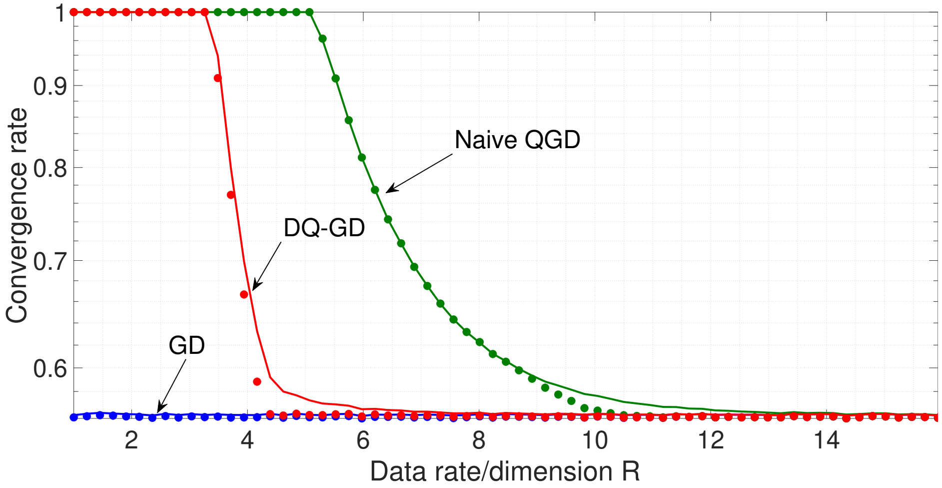
- 2nd term: choose the *dynamic range* r_i of the quantizer carefully so that

$$\sup_{\|\boldsymbol{u}_i\| \leq r_i} \|\boldsymbol{e}_i\| \leq \frac{\rho_n}{2^R} r_i$$

for each iteration i.

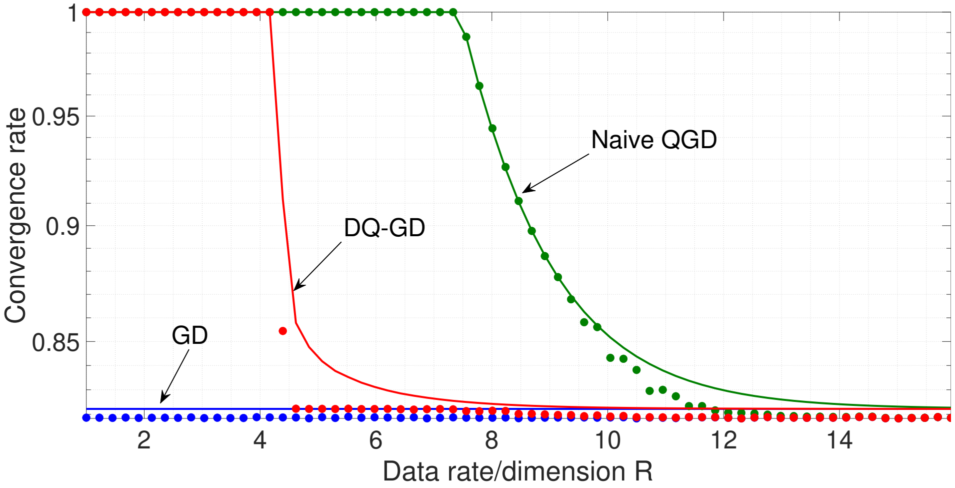


Least-squares problems: Gaussian ensemble



$f(\boldsymbol{x}) = \|\boldsymbol{y} - \mathbf{A}\boldsymbol{x}\|^2/2$. $\mathbf{A} \in \mathbb{R}^{1000\times 100}$ and $\boldsymbol{y} \in \mathbb{R}^{1000}$ iid standard normal entries. $\kappa(\mathbf{A}) \approx 1.8862$ on average.

Least-squares problems: Real-world matrix



$\mathbf{A} \in \mathbb{R}^{958 \times 292}$ with $\kappa(\mathbf{A}) \approx 3.2014$ (SuiteSparse matrix collection) and $\boldsymbol{y} \in \mathbb{R}^{958}$ iid standard normal entries.

Recap: main result

Optimal contraction factor over $f \in \mathcal{F}_n$ and $\mathrm{A} \in \mathrm{QGD}$

$$\lim_{n \to \infty} \inf_{\mathrm{QGD\ A}} \sigma_{\mathrm{A}}(n, R) = \max\left\{\sigma, 2^{-R}\right\}.$$

Phase transition

- $R \geq \log_2 1/\sigma$: contraction factor σ of unquantized GD is achievable
- $R < \log_2 1/\sigma$: only 2^{-R} is achievable

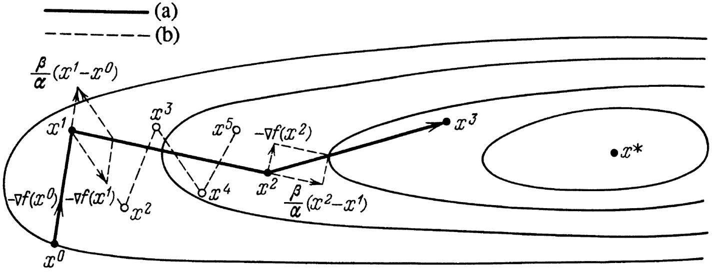
Extension: gradient methods with momentum

- Accelerated Gradient Descent [Nesterov, 1982]:

$$\boldsymbol{y}_{i+1} \leftarrow \boldsymbol{x}_i - \eta \nabla f(\boldsymbol{x}_i)$$
$$\boldsymbol{x}_{i+1} \leftarrow \boldsymbol{y}_{i+1} + \gamma (\boldsymbol{y}_{i+1} - \boldsymbol{y}_i)$$

- Heavy Ball Method [Polyak, 1987]:

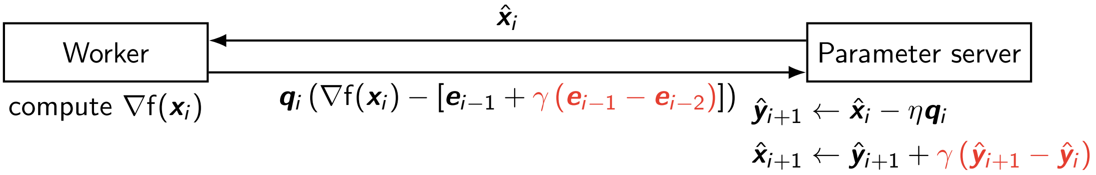
$$\boldsymbol{x}_{i+1} \leftarrow \boldsymbol{x}_i - \eta \nabla f(\boldsymbol{x}_i) + \gamma (\boldsymbol{x}_i - \boldsymbol{x}_{i-1})$$



(a) Heavy Ball Method
(b) Gradient Descent

Differentially Quantized Accelerated Gradient Descent (DQ-AGD)

DQ-AGD computes the gradient at $\boldsymbol{x}_i$ and compensates two past quantization errors:



Theorem: DQ-AGD

DQ-AGD achieves the following contraction factor over $\mathcal{F}_n$

$$\sigma_{\text{DQ-AGD}}(n, R) \leq \max\left\{\sigma_{\text{AGD}},\ \phi_{\text{DQ-AGD}}(n, R)\right\}$$

- σ_{AGD}: contraction factor of unquantized AGD
- $\phi_{\text{DQ-AGD}}(n, R)$: exponentially decreasing function of R

Differentially Quantized Heavy Ball Method (DQ-HB)

DQ-AGD computes the gradient at $\boldsymbol{x}_i$ and compensates two past quantization errors:

$\hat{\boldsymbol{x}}_i$

Worker

Parameter server

compute $\nabla \mathrm{f}(\boldsymbol{x}_i)$

$\boldsymbol{q}_i\left(\nabla \mathrm{f}(\boldsymbol{x}_i) - [\boldsymbol{e}_{i-1} + \gamma\left(\boldsymbol{e}_{i-1} - \boldsymbol{e}_{i-2}\right)]\right)$

$\hat{\boldsymbol{x}}_{i+1} \leftarrow \hat{\boldsymbol{x}}_i - \eta \boldsymbol{q}_i + \gamma\left(\hat{\boldsymbol{x}}_{i+1} - \hat{\boldsymbol{x}}_i\right)$

Theorem: DQ-HB

DQ-HB achieves the following contraction factor over $\mathrm{f} \in \mathcal{F}_n$ that are twice continuously differentiable:

$$\sigma_{\text{DQ-HB}}(n, R) \leq \max\left\{\sigma_{\text{HB}}, \phi_{\text{DQ-HB}}(n, R)\right\}$$

- σ_{HB}: contraction factor of unquantized HB
- $\phi_{\text{DQ-HB}}(n, R)$: exponentially decreasing function of R

Conclusion

- Introduced Differential Quantization.
- Differential Quantization is substantially better than naive quantization.
- If $R \geq R_{\mathrm{A}}$, differentially quantized $\mathrm{A} \in \{\mathrm{GD}, \mathrm{AGD}, \mathrm{HB}\}$ attains the contraction factor of the unquantized A.
- In the limit of $n \to \infty$, DQ-GD attains the optimal contraction factor within the class of QGD algorithms.
- Multiworker?