Layer-wise Quantization for Distributed Variational Inequalities

Anh Duc Nguyen
National University of Singapore
anh_duc@u.nus.edu

Ilia Markov
IST Austria
ilia.markov@ist.ac.at

Ali Ramezani-Kebrya
University of Oslo, Norwegian Centre for Knowledge-driven Machine Learning (Integreat), and Visual Intelligence Centre
ali@uio.no

Kimon Antonakopoulos
Laboratory for Information and Inference Systems (LIONS), EPFL
kimon.antonakopoulos@epfl.ch

Dan Alistarh
IST Austria & Neural Magic
dan.alistarh@ist.ac.at

Volkan Cevher
Laboratory for Information Inference Systems (LIONS), EPFL
volkan.cevher@epfl.ch

Abstract

We develop a general layer-wise and adaptive quantization framework with with error and code-length guarantees and applications to solving large-scale distributed variational inequality problems. We also propose Quantized and Generalized Optimistic Dual Averaging (QODA) which achieves the optimal rate of convergence for distributed monotone VIs under absolute noise. We empirically show that the adaptive layer-wise quantization achieves up to a 47% speedup in end-to-end training time for training Wasserstein GAN on 4 GPUs.

1 Introduction

For high-dimensional and non-convex settings with deep neural networks (DNNs), minimizing the empirical risk is a challenging optimization task due to non-convexity and lack of guarantees in terms of global optimality. Beyond empirical risk minimization, formulating the problems of training generative adversarial networks (GANs) [1] and equilibrium in more general and possibly non-zero-sum game-theoretic settings require more complicated mathematical frameworks. Variational inequality (VI) is a mathematical framework for modeling equilibrium problems [2–4], e.g., in applications such as robust adversarial reinforcement learning [5], auction theory [6], and adversarially robust learning [7]. For an operator $A : \mathbb{R}^d \to \mathbb{R}^d$, a VI finds some $\boldsymbol{x}^\star \in \mathbb{R}^d$ such that

$$\langle A(\boldsymbol{x}^\star), \boldsymbol{x} - \boldsymbol{x}^\star \rangle \geq 0 \quad \text{for all} \quad \boldsymbol{x} \in \mathbb{R}^d. \tag{VI}$$

In terms of implementation in a synchronous system with K nodes, first-order solvers for empirical risk minimization and VI-solvers are scaled by distributing computation among nodes, e.g., by partitioning the entire dataset in a cloud data center, followed by aggregation of local computations.[1] Nodes can be, e.g., hospitals and cellphones that train a global model or personalized models collaboratively in a federated learning setting.

In large-scale settings, communication costs for broadcasting huge stochastic gradients and dual vectors is the main performance bottleneck [8–11]. Several methods have been proposed to accelerate

[1] For simplicity, in the following, we use the term *node* to refer to client, FPGA, APU, CPU, GPU, worker.

Machine Learning and Compression Workshop, NeurIPS 2024.

large-scale training such as quantization, sparsification, and reducing the frequency of communication though local updates [10]. In particular, *unbiased quantization* is unique due to both enjoying strong theoretical guarantees along with providing communication efficiency on the fly, i.e., it converges under the same hyperparameters tuned for uncompressed variants while providing substantial savings in terms of communication costs [9, 11].

Popular DNNs including convolutional architectures, transformers, and vision transformers have various *types of layers* such as feedforward, residual, multi-head attention including self-attention and cross-attention, bias, and normalization layers [12–14]. Different types of layers learn different types of features. The current literature communication-efficient liteature does not rigorously take into account heterogeneity in terms of representation power, impact on the final learning outcome, and statistical heterogeneity across various layers of neural networks and across training for each layer. Recently, layer-wise and adaptive compression schemes have shown tremendous empirical success in accelerating training deep neural networks and transformers in large-scale settings [15, 16], which is yet to have strong theoretical guarantees and to handle statistical heterogeneity over the course of training. Hence, these layer-wise compression schemes suffer from a dearth of generalization and statistically rigorous argument to optimize the sequence of quantization and the number of sparsification levels for each layer.

1.1 Summary of Contributions

- We propose a theoretical framework for **layer-wise** and **adaptive** unbiased quantization schemes. We also establish tight variance and code-length bounds, which *generalize* those of global quantization frameworks [9, 17, 18].
- We propose Quantized Optimistic Dual Averaging (QODA) and establish joint convergence and communication guarantees with the competitive rate $\mathcal{O}(1/\sqrt{T})$ under absolute noise models. We obtain these bounds **without the restrictive almost sure boundedness assumption** of stochastic dual vectors in related VI works [4, 19, 20] including the SoTA distributed VI-solverQ-GenX [11].
- Empirically, we show that QODA with layer-wise compression improves accuracy compared to [11] and achieves up to a $\mathbf{47}\%$ speedup in terms of end-to-end training time in an application of training Wasserstein Generative Adversarial Network [21] on 4 GPUs.

2 Preliminaries

A detailed literature review is in Appendix A.1. A summary of commonly used notations in this paper is provided in Appendix A.2. Given an operator $A : \mathbb{R}^d \to \mathbb{R}^d$, consider the standard assumptions:

Assumption 2.1 (Monotonicity)**.** We have that for all $\boldsymbol{x}, \hat{\boldsymbol{x}} \in \mathbb{R}^d$, $\langle A(\boldsymbol{x}) - A(\hat{\boldsymbol{x}}), \boldsymbol{x} - \hat{\boldsymbol{x}} \rangle \geq 0$.

Assumption 2.2 (Solution Existence)**.** The solution set $\mathcal{X}^\star := \{\boldsymbol{x}^\star \in \mathbb{R}^d : \boldsymbol{x}^\star \text{solves (VI)}\} \neq \emptyset$.

Assumption 2.3 (L-Lipschitz)**.** Let $L \in \mathbb{R}^+$. Then an operator A is L-Lipschitz if

$$\|A(\boldsymbol{x}) - A(\boldsymbol{x}')\|_* \leq L\|\boldsymbol{x} - \boldsymbol{x}'\| \quad \forall\, \boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^d.$$

In this work, we consider methods that rely on a so-called *stochastic first-order oracle* [22]. This oracle, when called at $\boldsymbol{x}$, draws an i.i.d. sample ω from a complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and returns a *stochastic dual vector* $g(\boldsymbol{x}; \omega)$ given by:

$$g(\boldsymbol{x}; \omega) = A(\boldsymbol{x}) + U(\boldsymbol{x}; \omega), \tag{1}$$

where $U(\boldsymbol{x}; \omega)$ denotes the (possibly random) error in the measurement or noise. Next, we consider absolute noise profile formally defined as:

Assumption 2.4 (Absolute Noise)**.** Let $\boldsymbol{x} \in \mathbb{R}^d$ and $\omega \sim \mathbb{P}$. The oracle $g(\boldsymbol{x}; \omega)$ satisfies unbiasedness $\mathbb{E}[g(\boldsymbol{x}; \omega)] = A(\boldsymbol{x})$ and bounded absolute variance $\mathbb{E}\left[\|U(\boldsymbol{x}, \omega)\|_*^2\right] \leq \sigma^2$.

As the noise variance is independent of the value of the operator at the queried point, this type of randomness is *absolute*. Absolute noise is quite common in the (distributed) VI literature [23–25]. This noise profile is also known as the bounded variance assumption in stochastic optimization literature [26, 27].

Let $\mathcal{X} \subset \mathbb{R}^d$ denote a non-empty and compact test domain. The main measure to evaluate the quality of a candidate solution is the restricted gap function [28, 29] (more properties in Appendix A.3):

$$\text{GAP}_{\mathcal{X}}(\hat{\boldsymbol{x}}) = \sup_{\boldsymbol{x} \in \mathcal{X}} \langle A(\boldsymbol{x}), \hat{\boldsymbol{x}} - \boldsymbol{x} \rangle. \tag{GAP}$$

3 Quantized Optimistic Dual Averaging

Consider a distributed and synchronous setting with K nodes, along the lines of the standard setting for data-parallel SGD [9, 30]. Here, the nodes partition the entire dataset among themselves such that each node retains only local copy of the current parameter vector while having access to independent private stochastic dual vectors. In each iteration, each node receives stochastic dual vectors, aggregates them, computes an update, and broadcasts the compressed update to accelerate training. These compressed updates are decompressed before the next aggregation step at each node.

3.1 Adaptive Layer-wise Quantization

Let $V_{k,t}$ and $\hat{V}_{k,t}$ denote the uncompressed and compressed stochastic dual vector in node k at time t, respectively. Let $\boldsymbol{v} \in \mathbb{R}^d$ be a vector to be quantized. For $i = 1, \ldots, d$, let $u_i = |v_i|/\|\boldsymbol{v}\|_q$ be the normalized coordinate. At each time t, instead of a global sequence of quantization levels for all coordinates [9, 11], we consider a set $\mathbb{L}^{t,M}$ of M types of sequences $\{\boldsymbol{\ell}^{t,1}, \ldots, \boldsymbol{\ell}^{t,M}\}$ to be optimized with flexible and adjustable numbers of levels $\alpha_1, \ldots, \alpha_M$, respectively. We denote $\boldsymbol{\ell}^{t,m} \in \mathbb{L}^{t,M}$ the sequence of type m at time t, given by $[\ell_0, \ell_1^{t,m}, \ldots, \ell_{\alpha_m}^{t,m}, \ell_{\alpha_m+1}]^\top$, where $0 = \ell_0 < \ell_1^{t,m} < \cdots < \ell_{\alpha_m}^{t,m} < \ell_{\alpha_m+1} = 1$. Let $\mathbb{S}^{t,m}$ be the set of all normalized coordinates that use type m sequence $\boldsymbol{\ell}^{t,m}$ at time t. Let $\tau^{t,m}(u)$ denote the index of a level with respect to $u \in [0,1]$ such that $\ell_{\tau^{t,m}(u)}^{t,m} \le u < \ell_{\tau^{t,m}(u)+1}^{t,m}$. Let $\xi^{t,m}(u) = (u - \ell_{\tau^{t,m}(u)}^{t,m})/(\ell_{\tau^{t,m}(u)+1}^{t,m} - \ell_{\tau^{t,m}(u)}^{t,m})$ be the relative distance of u to the level $\tau^{t,m}(u)+1$. Define the random variable $q_{\boldsymbol{\ell}^{t,m}}(u) = \ell_{\tau^{t,m}(u)}^{t,m}$ with probability $1 - \xi^{t,m}(u)$, and $\ell_{\tau^{t,m}(u)+1}^{t,m}$ with probability $\xi^{t,m}(u)$.

We then define the random quantization of vector $\boldsymbol{v}$ as $Q_{\mathbb{L}^{t,M}}(\boldsymbol{v}) = [Q_{\mathbb{L}^{t,M}}(v_1), \ldots, Q_{\mathbb{L}^{t,M}}(v_d)]^\top$ where for $m = 1, 2, \ldots, M$, and any $u_i \in \mathbb{S}^{t,m}$, we have $Q_{\mathbb{L}^{t,M}}(v_i) = \|\boldsymbol{v}\|_q \cdot \text{sign}(v_i) \cdot q_{\boldsymbol{\ell}^{t,m}}(u_i)$. Let $\boldsymbol{q}_{\mathbb{L}^{t,M}} \sim \mathbb{P}_Q$ represent d variables $\{q_{\boldsymbol{\ell}^{t,m}}(u_i)\}_{i\in[d]}$ sampled independently for random quantization. As this scheme is unbiased, we can measure the quantization error by measuring the variance $\mathbb{E}_{\boldsymbol{q}_{\mathbb{L}^{t,M}}}[\|Q_{\mathbb{L}^{t,M}}(\boldsymbol{v}) - \boldsymbol{v}\|_2^2]$ given by

$$\|\boldsymbol{v}\|_q^2 \sum_{m=1}^{M} \sum_{u_i \in \mathbb{S}^{t,m}} \sigma_Q^2(u_i; \boldsymbol{\ell}^{t,m}), \tag{Var}$$

where $\sigma_Q^2(u_i; \ell^{t,m}) = \mathbb{E}[(q_{\boldsymbol{\ell}^{t,m}}(u_i) - u_i)^2] = (\ell_{\tau^{t,m}(u_i)+1}^{t,m} - u_i)(u_i - \ell_{\tau^{t,m}(u_i)}^{t,m})$ is the variance of quantization of a single coordinate $u_i \in \mathbb{S}^{t,m}$ with type m sequence $\boldsymbol{\ell}^{t,m}$. We can optimize M quantization sequences by minimizing the overall quantization variance

$$\min_{\mathbb{L}^{t,M} \in \mathcal{L}^{t,M}} \mathbb{E}_\omega \mathbb{E}_{\boldsymbol{q}_{\mathbb{L}^{t,M}}} \left[\|Q_{\mathbb{L}^{t,M}}(g(\boldsymbol{x}_t; \omega)) - A(\boldsymbol{x}_t)\|_2^2\right],$$

where $\mathcal{L}^{t,M} = \left\{\{\boldsymbol{\ell}^{t,1}, \ldots, \boldsymbol{\ell}^{t,M}\} : \forall m \in [M],\ \forall j \in [\alpha_m], \ell_j^{t,m} \le \ell_{j+1}^{t,m}, \ell_0 = 0, \ell_{\alpha_m+1} = 1\right\}$, denoting the collection of all feasible sets of type m levels. Since random quantization and random samples are statistically independent, the above minimization is equivalent to

$$\min_{\mathbb{L}^{t,M} \in \mathcal{L}^{t,M}} \mathbb{E}_\omega \mathbb{E}_{\boldsymbol{q}_{\mathbb{L}^{t,M}}} \left[\|Q_{\mathbb{L}^{t,M}}(g(\boldsymbol{x}_t; \omega)) - g(\boldsymbol{x}_t; \omega)\|_2^2\right]. \tag{MQV}$$

3.2 Encoding

Coding schemes are applied on top of our layer-wise quantization to further reduce communication costs. We now introduce a practical coding protocol for layer-wise quantization that *require a fine-grained analysis* different from those for global quantization [9, 11, 17, 18]. For some $q \in \mathbb{Z}_+$, any vector $\boldsymbol{v} \in \mathbb{R}^d$ can be uniquely represented by a tuple $(\|\boldsymbol{v}\|_q, \boldsymbol{s}, \boldsymbol{u})$ where $\|\boldsymbol{v}\|_q$ is the L^q norm of $\boldsymbol{v}$, $\boldsymbol{s} := [\text{sign}(v_1), \ldots, \text{sign}(v_d)]^\top$ comprises of signs of each coordinate v_i, and $\boldsymbol{u} := [u_1, \ldots, u_d]^\top$, where $u_i = |v_i|/\|\boldsymbol{v}\|_q$, are the normalized coordinates. Note that $0 \le u_i \le 1$ for all $i \in [d]$.
Let $\mathcal{A}^{t,m} = \{\ell_0^{t,m}, \ell_1^{t,m}, \ldots, \ell_{\alpha_m}^{t,m}, \ell_{\alpha_m+1}^{t,m}\}$ be the collection of all the levels of the sequence $\boldsymbol{\ell}^{t,m}$. Let $\Omega^{t,M} = \bigcup_{m=1}^M \mathcal{A}^{t,m}$ be the collection of all the levels of M sequences at time t. The overall encoding, i.e., composition of coding and quantization, $\text{ENC}(\|\boldsymbol{v}\|_q, \boldsymbol{s}, \boldsymbol{q}_{\mathbb{L}^{t,M}}) : \mathbb{R}_+ \times \{\pm 1\}^d \times (\Omega^{t,M})^d \to \{0,1\}^*$ in Algorithm 1 uses a standard floating point encoding with C_q bits to represent the non-negative scalar $\|\boldsymbol{v}\|_q$, encodes the sign of each coordinate with one bit, and then utilizes an

integer encoding scheme $\Psi : (\Omega^{t,M})^d \to \{0,1\}^*$ to efficiently encode every quantized coordinate with the minimum expected code-length. To solve (MQV), we sample Z stochastic dual vectors $\{g(\boldsymbol{x}_t;\omega_1),\ldots,g(\boldsymbol{x}_t;\omega_Z)\}$. Let F_z denote the marginal cumulative distribution function (CDF) of normalized coordinates conditioned on observing $\|g(\boldsymbol{x}_t;\omega_z)\|_q$. By law of total expectation, for $\mathbb{L}^{t,M} \in \mathcal{L}^{t,M}$, (MQV) can be approximated by:

$$\min_{\mathbb{L}^{t,M}} \sum_{z=1}^{Z} \|g(\boldsymbol{x}_t;\omega_z)\|_q^2 \sum_{m=1}^{M}\sum_{i=0}^{\alpha_m} \int_{\ell_i^{t,m}}^{\ell_{i+1}^{t,m}} \sigma_Q^2(u;\boldsymbol{\ell}^{t,m})\,\mathrm{d}F_z(u) \text{ or } \min_{\mathbb{L}^{t,M}} \sum_{m=1}^{M}\sum_{i=0}^{\alpha_m} \int_{\ell_i^{t,m}}^{\ell_{i+1}^{t,m}} \sigma_Q^2(u;\boldsymbol{\ell}^{t,m})\,\mathrm{d}\tilde{F}(u), \quad (2)$$

where $\tilde{F}(u) = \sum_{z=1}^{Z} \lambda_z F_z(u)$ is the weighted sum of the conditional CDFs with

$$\lambda_z = \|g(\boldsymbol{x}_t;\omega_z)\|_q^2 / \sum_{z=1}^{Z} \|g(\boldsymbol{x}_t;\omega_z)\|_q^2. \quad (3)$$

To solve (MQV), we first sample Z stochastic dual vectors $\{g(\boldsymbol{x}_t;\omega_1),\ldots,g(\boldsymbol{x}_t;\omega_Z)\}$. Let F_z^m denote the marginal CDF of normalized coordinates of type m conditioned on observing $\|g(\boldsymbol{x}_t;\omega_z)\|_q$.

In our implementation (details in Section 5), we utilize L-GreCo [16] which executes a dynamic programming algorithm optimizing the total compression ratio while minimizing compression error. The decoding $\mathrm{DEC} : \{0,1\}^* \to \mathbb{R}^d$ first reads C_q bits to reconstruct $\|\boldsymbol{v}\|_q$, then applies decoding schemes $(\Psi^m)^{-1} : \{0,1\}^* \to \mathcal{A}^{t,m}$ to obtain normalized type m coordinates. The discussion for the choice of a specific lossless prefix code and more details on coding schemes are in Appendix C.1.

3.3 Optimistic Dual Averaging

Our described layer-wise quantization and the coding protocol are **general** with applications such as empirical risk minimization by training transformers [15, 16]. In this section, we will show one such application with our novel *Quantized Optimistic Dual Averaging (QODA)*, Algorithm 1, to efficiently solve distributed VI. Importantly, this optimistic approach **reduces one "extra" gradient step** that extra gradient methods and variants such as the baseline [11] take (by storing the gradient from the previous iteration, refer to line 9 and 16), thereby reducing the communication burden by half decoupled from acceleration due to quantization. At certain steps, every node calculates the sufficient statistics of a parametric distribution to estimate distribution of dual vectors in lines 3 to 5. Let $\hat{V}_{k,t} = Q(V_{k,t}) = Q(A_k(X_t) + U_k(X_t))$ denote the unbiased and quantized stochastic dual vectors for node $k \in [K]$ and iteration $t \in [T]$. The *optimistic dual averaging* updates in (4) appear in lines 10, 17 and 18. Layer-wise quantization with $Q_{\mathbb{L}^{t,M}}$ and the coding protocol are in lines 12 and 15. The loops are executed *in parallel* on the nodes.

Algorithm 1: Quantized Optimistic Dual Averaging

Input: Local training data; local copies of X_t, Y_t; update steps set $\mathcal{U}$; learning rates $\{\gamma_t\}, \{\eta_t\}$

1: **for** $t = 1$ **to** T **do**
2: **if** $t \in \mathcal{U}$ **then**
3: **for** $i = 1$ **to** K **do**
4: Estimate distributions of normalized dual vectors and update $\mathbb{L}^{t,M}$ (Appendix A.4)
5: Update M sequences of levels *in parallel*
6: **end for**
7: **end if**
8: **for** $i = 1$ **to** K **do**
9: Retrieve previously stored $\hat{V}_{k,t-1/2}$
10: $X_{t+1/2} \leftarrow X_t - \gamma_t \sum_{k=1}^{K} \hat{V}_{k,t-1/2}/K$
11: $V_{i,t+1/2} \leftarrow A_i(X_{t+1/2}) + U_i(X_{t+1/2})$
12: $d_{i,t} \leftarrow \mathrm{ENCODE}\left(Q_{\mathbb{L}^{t,M}}(V_{i,t+1/2}); \mathbb{L}^{t,M}\right)$
13: Broadcast $d_{i,t}$
14: Receive $d_{i,t}$ from each node i
15: $\hat{V}_{i,t+1/2} \leftarrow \mathrm{DECODE}(d_{i,t}; \mathbb{L}^{t,M})$
16: Store $\hat{V}_{k,t+1/2}$
17: $Y_{t+1} \leftarrow Y_t - \sum_{k=1}^{K} \hat{V}_{k,t+1/2}/K$
18: $X_{t+1} \leftarrow \eta_{t+1} Y_{t+1} + X_1$
19: **end for**
20: **end for**

$$X_{t+1/2} = X_t - \gamma_t \sum_{k=1}^{K} \hat{V}_{k,t-1/2}/K;\; Y_{t+1} = Y_t - \sum_{k=1}^{K} \hat{V}_{k,t+1/2}/K;\; X_{t+1} = X_1 + \eta_{t+1} Y_{t+1}. \quad (4)$$

In general, learning rates γ_t and η_t can be chosen such that they are non-increasing and $\gamma_t \geq \eta_t > 0$. We propose the following *adaptive* learning rate schedules for updates (4) and in Algorithm 1.

$$\eta_t = \gamma_t = \left(1 + \sum_{s=1}^{t-1}\sum_{k=1}^{K} \left\|\hat{V}_{k,s+1/2} - \hat{V}_{k,s-1/2}\right\|_*^2 / K^2\right)^{-1/2}. \quad (5)$$

4 Theoretical Guarantees

4.1 Quantization Bounds

We drop time index t for notation simplicity. Let $q \in \mathbb{Z}_+$. Let $\bar{\ell}^m = \max_{0\le j\le \alpha_m} \ell_{j+1}^m/\ell_j^m$, and $\bar{\ell}^M = \max_{1\le m\le M} \bar{\ell}^M$. Denote the largest level 1 across M types $\bar{\ell}_1^M = \max_{1\le m\le M} \ell_1^M$. Let $d_{th} = (2/\bar{\ell}_1^M)^{\min\{2,q\}}$. We now present the variance bounds for our layer-wise quantization schemes:

Theorem 4.1 (Quantization Variance Bound)**.** *Let $\boldsymbol{v} \in \mathbb{R}^d$ be a vector to be quantized with L^q normalization. With unbiased quantization of $\boldsymbol{v}$, i.e., $\mathbb{E}_{q_{\mathbb{L}^M}}[Q_{\mathbb{L}^M}(\boldsymbol{v})] = \boldsymbol{v}$, we have that*

$$\mathbb{E}_{q_{\mathbb{L}^M}}\left[\|Q_{\mathbb{L}^M}(\boldsymbol{v}) - \boldsymbol{v}\|_2^2\right] \le \varepsilon_Q \|\boldsymbol{v}\|_2^2, \tag{6}$$

where

$$\varepsilon_Q = (\bar{\ell}^M - 1)^2/(4\bar{\ell}^M) + (\bar{\ell}_1^M)^2 d^{2/\min\{q,2\}} \mathbb{1}\{d < d_{th}\}/4 + (\bar{\ell}_1^M d^{2/\min\{q,2\}} - 1) d^{2/\min\{q,2\}} \mathbb{1}\{d \ge d_{th}\}.$$

The proof is in Appendix B. For the special case of $M = 1$, our bound recovers [11, Theorem 1], matching the lower bound $\Omega(d)$ in the specific regime of large d and L^2 normalization. Furthermore, this bound, under $M = 1$, holds for general L^q normalization and arbitrary sequence of quantization levels in comparison to [9, Theorem 3.2] and [17, Theorem 4], which hold only for L^2 normalization with uniform and exponentially spaced levels, respectively. Next, we establish code-length bounds for the coding protocol. The guarantee for coding protocol is as follows with the proof in Appendix C.2

Theorem 4.2 (Code-length Bound)**.** *Let p_j^m denote the probability of occurrence of ℓ_j^m for $m \in [M]$ and $j \in [\alpha_m]$. Under the setting specified in Theorem 4.1, the expectation $\mathbb{E}_w \mathbb{E}_{\boldsymbol{q}_{\mathbb{L}^M}}\left[\mathrm{ENC}\left(Q_{\mathbb{L}^M}(g(\boldsymbol{x};\omega)); \mathbb{L}^M\right)\right]$ of the number of bits under the coding protocol is*

$$\mathbb{E}_\omega \mathbb{E}_{\boldsymbol{q}_{\mathbb{L}^M}}\left[\mathrm{ENC}\left(Q_{\mathbb{L}^M}(g(\boldsymbol{x};\omega)); \mathbb{L}^M\right)\right] = \mathcal{O}\left(\left(-\sum_{m=1}^M p_0^m - \sum_{m=1}^M \sum_{j=1}^{\alpha_m} p_j^m \log p_j^m\right) d\right). \tag{7}$$

For the special case of $M = 1$, our bound for the coding protocol in Theorem 4.2 recovers [11, Theorem 2]. Under the specific scenario of $M = 1$, L^2 normalization and $s = \sqrt{d}$ as in [9, Theorem 3.4], our bound for the coding protocol can be arbitrarily smaller than [9, Theorem 3.4] and [17, Theorem 5] depending on the probabilities $\{p_0, \ldots, p_{s+1}\}$. Under similar settings, we obtain that the expected $\mathcal{O}(Kd/\varepsilon)$ bits are required to reach an ε gap, matching the lower bound for convex optimization problems with finite-sum structures [31, 32].

4.2 Algorithm Complexity

Here, Algorithm 1 is executed for T iterations on K nodes with learning rates in (5). Denote the average square root variance bound $\widehat{\varepsilon_Q} = \sum_{m=1}^M \sum_{j=1}^{J^m} T_{m,j} \sqrt{\varepsilon_{Q,m,j}}/T$.

Under the absolute noise model, we can bound GAP of Algorithm 1 with the proof in Appendix D:

Theorem 4.3. *Suppose the iterates X_t of Algorithm 1 are updated with learning rate schedule in (5) for all $t = 1/2, 1, \ldots, T$. Let $\mathcal{X} \subset \mathbb{R}^d$ be a compact neighborhood of a VI solution and $D^2 := \sup_{\boldsymbol{p}\in\mathcal{X}} \|X_1 - \boldsymbol{p}\|_2^2$. Under Assumptions 2.1, 2.2, 2.3, and 2.4, we have*

$$\mathbb{E}\left[\mathrm{Gap}_{\mathcal{X}}\left(\sum_{t=1}^T X_{t+1/2}/T\right)\right] = \mathcal{O}\left(\left((LD + \|A(X_1)\|_2 + \sigma)\widehat{\varepsilon_Q} + \sigma\right) D^2 L^2/\sqrt{TK}\right).$$

This theorem show that increasing the number of processors K lead to faster convergence for monotone VIs, matching the asymptotic rates for T and K in [11] which requires an extra almost sure boundedness assumption. Under the absolute noise model and by setting the number of gradients per round to one, our results match the known lower bound for convex and smooth optimization $\Omega(1/\sqrt{TK})$ [23, Theorem 1]. Previously, [11, Theorem 3] matches this lower bound but with an extra assumption that the operator is almost sure bounded.

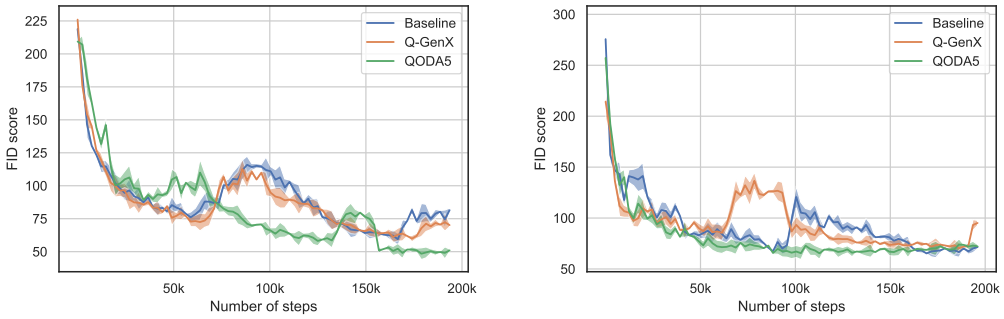

Figure 1: FID evolution during training. We compare basic Adam optimization against QODA-based extension of Adam with global (Q-GenX [11]) and layer-wise (L-GreCo) quantizations.

5 Numerical Experiments

We have implemented QODA in Algorithm 1 based on the codebase of [33] and train WGAN [21] on CIFAR10 and CIFAR100 [34]. To support efficient compression, we used the `torch_cgx` Pytorch extension [15]. Moreover, we adapt compression choices layer-wise, following the L-GreCo [16] algorithm. Specifically, L-GreCo periodically collects gradients statistics, then executes a dynamic programming algorithm optimizing the total compression ratio while minimizing compression error. In our experiments, we use 4 nodes, each with a single NVIDIA RTX 3090 GPU, in a multi-node Genesis Cloud environment. For the communication backend, we picked the best option for quantized and full-precision regimes: OpenMPI [35] and NCCL [36], respectively. The maximum bandwidth between nodes is estimated to be around 5 Gbit/second.

We follow the training recipe of Q-GenX [11], where authors set large batch size (1024) and keep all other hyperparameters as in the original codebase of [33]. For global and layer-wise compression, we use 5 bits (with bucket size 128), and run the L-GreCo adaptive compression algorithm every 10K optimization steps for both the generator and discriminator models. The convergence results are presented in Figure 1. The figure demonstrates that the adaptive QODA approach not only recovers the baseline accuracy but also improves convergence relative to Q-GenX [11].

In order to illustrate the impact of QODA on the wall-clock training time, we have benchmarked the training in three different communication setups. The first is the original 5 Gbps bandwidth, whereas the second and the third reduce this to half and 1/5 of this maximum bandwidth. We measured the time per training step for uncompressed and QODA 5-bit training. Note that time per step is similar for for both data sets. Table 1 shows that layer-wise quantization achieves up to a 47% improvement in terms of end-to-end training time.

Mode	1 Gbps	2.5 Gbps	5 Gbps
Baseline	291	265	251
QODA5	197	195	195
Speedup	1.47×	1.36×	1.28×

Table 1: Time per optimization step[2](in ms) for baseline and QODA5 with different inter-node bandwidths.

6 Conclusion

In brief, we introduce *optimism* in distributed VI with adaptive learning rates, develop layer-wise quantization with joint convergence and communication guarantees, and show improvements in end-to-end training time in a practical multi-node WGAN setting. We also establish tight variance and code-length bounds for a general layer-wise and adaptive family of compression schemes that generalize previous bounds for global quantization.

[2]The optimization step includes forward and backward times. More precisely, the backward step consists of backpropagation, compression, communication and de-compression.

7 Acknowledgment

This work was supported by 1) the Research Council of Norway through its Centres of Excellence scheme, Integreat - Norwegian Centre for knowledge-driven machine learning, project number 332645; 2) the Research Council of Norway through its Centre for Research-based Innovation funding scheme (Visual Intelligence under grant no. 309439), and Consortium Partners; 3) Swiss National Science Foundation (SNSF) under grant number 200021_205011; 4) Hasler Foundation Program: Hasler Responsible AI (project number 21043), Research was sponsored by the Army Research Office and was accomplished under Grant Number W911NF-24-1-0048,

References

[1] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[2] Francisco Facchinei and Jong-Shi Pang. *Finite-dimensional variational inequalities and complementarity problems*. Springer, 2003.

[3] Heinz H. Bauschke and Patrick L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, 2017.

[4] Kimon Antonakopoulos, Thomas Pethick, Ali Kavis, Panayotis Mertikopoulos, and Volkan Cevher. Sifting through the noise: Universal first-order methods for stochastic variational inequalities. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 13099–13111, 2021.

[5] Lerrel Pinto, James Davidson, Rahul Sukthankar, and Abhinav Gupta. Robust adversarial reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2017.

[6] Vasilis Syrgkanis, Alekh Agarwal, Haipeng Luo, and Robert E Schapire. Fast convergence of regularized learning in games. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

[7] Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially robust generalization requires more data. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[8] Nikko Strom. Scalable distributed DNN training using commodity GPU cloud computing. In *INTERSPEECH*, 2015.

[9] Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. QSGD: Communication-efficient SGD via gradient quantization and encoding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[10] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, R. G. L. D'Oliveira, S. E. Rouayheb, D. Evans, J. Gardner, Z. Garrett, A. Gascón, B. Ghazi, P. B. Gibbons, M. Gruteser, Z. Harchaoui, C. He, L. He, Z. Huo, B. Hutchinson, J. Hsu, M. Jaggi, T. Javidi, G. Joshi, M. Khodak, J. Konečný, A. Korolova, F. Koushanfar, S. Koyejo, T. Lepoint, Y. Liu, P., M. Mohri, R. Nock, A. Özgür, R. Pagh, M. Raykova, H. Qi, D. Ramage, R. Raskar, D. Song, W. Song, S. U. Stich, Z. Sun, A. T. Suresh, F. Tramèr, P. Vepakomma, J. Wang, L. Xiong, Z. Xu, Q. Yang, F. X. Yu, H. Yu, and S. Zhao. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.

[11] Ali Ramezani-Kebrya, Kimon Antonakopoulos, Igor Krawczuk, Justin Deschenaux, and Volkan Cevher. Distributed extra-gradient with optimal complexity and communication guarantees. In *International Conference on Learning Representations (ICLR)*, 2023.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

[15] Ilia Markov, Hamidreza Ramezanikebrya, and Dan Alistarh. Cgx: adaptive system support for communication-efficient deep learning. In *Proceedings of the 23rd ACM/IFIP International Middleware Conference*, pages 241–254, 2022.

[16] Ilia Markov, Kaveh Alim, Elias Frantar, and Dan Alistarh. L-greco: Layerwise-adaptive gradient compression for efficient data-parallel deep learning. *Proceedings of Machine Learning and Systems*, 6:312–324, 2024.

[17] Ali Ramezani-Kebrya, Fartash Faghri, Ilya Markov, Vitalii Aksenov, Dan Alistarh, and Daniel M. Roy. NUQSGD: Provably communication-efficient data-parallel SGD via nonuniform quantization. *Journal of Machine Learning Research (JMLR)*, 22(114):1–43, 2021.

[18] Fartash Faghri, Iman Tabrizian, Ilia Markov, Dan Alistarh, Daniel M. Roy, and Ali Ramezani-Kebrya. Adaptive gradient quantization for data-parallel SGD. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[19] Francis Bach and Kfir Y Levy. A universal algorithm for variational inequalities adaptive to smoothness and noise. In *Conference on learning theory*, pages 164–194. PMLR, 2019.

[20] Yu-Guan Hsieh, Kimon Antonakopoulos, and Panayotis Mertikopoulos. Adaptive learning in continuous games: Optimal regret bounds and convergence to nash equilibrium. In *Conference on Learning Theory*, pages 2388–2422. PMLR, 2021.

[21] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning (ICML)*, pages 214–223. PMLR, 2017.

[22] Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Kluwer Academic Publishers, 2004.

[23] Blake E. Woodworth, Brian Bullins, Ohad Shamir, and Nathan Srebro. The min-max complexity of distributed stochastic convex optimization with intermittent communication. In *Annual Conference Computational Learning Theory*, 2021. URL `https://api.semanticscholar.org/CorpusID:231749558`.

[24] Alina Ene and Huy Le Nguyen. Adaptive and universal algorithms for variational inequalities with optimal convergence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 6559–6567, 2022.

[25] Nazarii Tupitsa, Abdulla Jasem Almansoori, Yanlin Wu, Martin Takac, Karthik Nandakumar, Samuel Horváth, and Eduard Gorbunov. Byzantine-tolerant methods for distributed variational inequalities. *Advances in Neural Information Processing Systems*, 36, 2024.

[26] Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4): 1574–1609, 2009.

[27] Anatoli Juditsky, Arkadi Nemirovski, and Claire Tauvel. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58, 2011.

[28] Yurii Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical programming*, 120(1):221–259, 2009.

[29] Kimon Antonakopoulos, Veronica Belmega, and Panayotis Mertikopoulos. An adaptive mirror-prox method for variational inequalities with singular operators. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

[30] Jeffrey Dean, Greg S. Corrado, Rajat Monga, Kai Chen, Matthieu Devin, Mark Z. Mao, Marc'aurelio Ranzato, Andrew Senior, Paul Tucker, Ke Yang, Quoc Le, and Andrew Y. Ng. Large scale distributed deep networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2012.

[31] John N. Tsitsiklis and Zhi-Quan Luo. Communication complexity of convex optimization. *Journal of Complexity*, 3(3):231–243, 1987.

[32] Janne H Korhonen and Dan Alistarh. Towards tight communication lower bounds for distributed optimisation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[33] Gauthier Gidel, Hugo Berard, Gaëtan Vignoud, Pascal Vincent, and Simon Lacoste-Julien. A variational inequality perspective on generative adversarial networks. *arXiv preprint arXiv:1802.10551*, 2018.

[34] Alex Krizhevsky. Learning multiple layers of features from tiny images. *Technical report, University of Toronto*, 2009.

[35] Open MPI: Open Source High Performance Computing. `https://www.open-mpi.org/`, 2023.

[36] NVIDIA Collective Communication Library. `https://developer.nvidia.com/nccl`, 2023.

[37] Hongyi Wang, Scott Sievert, Shengchao Liu, Zachary Charles, Dimitris Papailiopoulos, and Stephen Wright. Atomo: Communication-efficient learning via atomic sparsification. *Advances in Neural Information Processing Systems (NeurIPS)*, 31, 2018.

[38] Maksim Makarenko, Elnur Gasanov, Rustem Islamov, Abdurakhmon Sadiev, and Peter Richtárik. Adaptive compression for communication-efficient distributed training. *arXiv preprint arXiv:2211.00188*, 2022.

[39] Jinrong Guo, Wantao Liu, Wang Wang, Jizhong Han, Ruixuan Li, Yijun Lu, and Songlin Hu. Accelerating distributed deep learning by adaptive gradient quantization. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020.

[40] Saurabh Agarwal, Hongyi Wang, Kangwook Lee, Shivaram Venkataraman, and Dimitris Papailiopoulos. Adaptive gradient communication via critical learning regime identification. In *Conference on Machine Learning and Systems (MLSys)*, 2021.

[41] Wei Wen, Cong Xu, Feng Yan, Chunpeng Wu, Yandan Wang, Yiran Chen, and Hai Li. TernGrad: Ternary gradients to reduce communication in distributed deep learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[42] Hantian Zhang, Jerry Li, Kaan Kara, Dan Alistarh, Ji Liu, and Ce Zhang. ZipML: Training linear models with end-to-end low precision, and a little bit of deep learning. In *International Conference on Machine Learning (ICML)*, 2017.

[43] Peter Davies, Vijaykrishna Gurunanthan, Niusha Moshrefi, Saleh Ashkboos, and Dan Alistarh. New bounds for distributed mean estimation and variance reduction. In *International Conference on Learning Representations (ICLR)*, 2021.

[44] Dmitry Kovalev, Aleksandr Beznosikov, Abdurakhmon Sadiev, Michael Persiianov, Peter Richtárik, and Alexander Gasnikov. Optimal algorithms for decentralized stochastic variational inequalities. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 31073–31088, 2022.

[45] Aleksandr Beznosikov, Eduard Gorbunov, Hugo Berard, and Nicolas Loizou. Stochastic gradient descent-ascent: Unified theory and new efficient methods. In *International Conference on Artificial Intelligence and Statistics*, pages 172–235. PMLR, 2023.

[46] Aleksandr Beznosikov, Peter Richtárik, Michael Diskin, Max Ryabinin, and Alexander Gasnikov. Distributed methods with compressed communication for solving variational inequalities, with theoretical guarantees. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 14013–14029, 2022.

[47] Aleksandr Beznosikov, Darina Dvinskikh, Andrei Semenov, and Alexander Gasnikov. Bregman proximal method for efficient communications under similarity, 2023.

[48] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research (JMLR)*, 12(7), 2011.

[49] Deming Yuan, Shengyuan Xu, Huanyu Zhao, and Lina Rong. Distributed dual averaging method for multi-agent optimization with quantized communication. *Systems & Control Letters*, 61(11):1053–1061, 2012.

[50] Konstantinos I Tsianos and Michael G Rabbat. Distributed dual averaging for convex optimization under communication delays. In *2012 American Control Conference (ACC)*, pages 1067–1072. IEEE, 2012.

[51] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, USA, 2006. ISBN 0471241954.

[52] David A. Huffman. A method for the construction of minimum-redundancy codes. *Proceedings of the IRE*, 40(9):1098–1101, 1952. doi: 10.1109/JRPROC.1952.273898.

[53] P. Elias. Universal codeword sets and representations of the integers. *IEEE Transactions on Information Theory*, 21(2):194–203, 1975. doi: 10.1109/TIT.1975.1055349.

[54] Kfir Y Levy, Alp Yurtsever, and Volkan Cevher. Online adaptive methods, universality and acceleration. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.

[55] H Brendan McMahan and Matthew Streeter. Adaptive bound optimization for online convex optimization. *arXiv preprint arXiv:1002.4908*, 2010.

Contents

A Addition Information

A.1 Literature Review

For empirical risk minimization, *adaptive quantization*, has been proposed to adapt quantization levels [18, 37, 38] and the number of quantization levels [39, 40] over the trajectory of optimization. All these quantization schemes are *global w.r.t. layers* and do not take into account heterogeneities in terms of representation power and impact on the final learning outcome across various layers of neural networks and across training for each layer. Markov et al. [15, 16] have empirically studied unbiased and *layer-wise quantization* where quantization parameters are updated across layers in a heuristic manner and have shown tremendous empirical success in training popular DNNs.

Unbiased quantization provides communication efficiency on the fly for empirical risk minimization, i.e., quantized variants of SGD converge under the same hyperparameters tuned for uncompressed variants while providing substantial savings in terms of communication costs [9, 15–18, 41, 42]. [43] has proposed lattice-based quantization for distributed mean estimation problem.

There is a line of research that focuses on designing *distributed methods for VI and saddle points problems*. Kovalev et al. [44] consider strongly monotone VI; Beznosikov et al. [45] concern with VI problems under co-coercivity assumptions. Assumptions such as strong monotonicity and co-coercivity are quite restrictive in ML applications. Beznosikov et al. [46, 47] consider VI problems with finite sum structure with an extra δ-similarity assumption in [47] . Several works [48–50] explore *dual averaging* for distributed finite-sum minimization in networks.

A.2 Notations

We use lower-case bold letters to denote vectors. $\mathbb{E}[\cdot]$ denotes the expectation operator. $\|\cdot\|_0$ and $\|\cdot\|_*$ are number of nonzero elements of a vector and dual norm, respectively. $|\cdot|$ denotes the length of a binary string, the length of a vector, and cardinality of a set. Sets are typeset in a calligraphic font. The base-2 logarithm is denoted by $\log$, and the set of binary strings is denoted by $\{0,1\}^*$. For any integer n, we use $[n]$ to denote the set $\{1,\dots,n\}$. $\mathbb{1}$ denotes the indicator function.

A.3 More Details on GAP

Several properties of (GAP) have been explored in the literature [28, 29]. In particular, the following classical results characterize the solutions of (VI) via zeros of (GAP).

Proposition A.1. *[28] Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a non-empty and convex set. Then, we have*

- *$GAP_{\mathcal{X}}(\hat{\boldsymbol{x}}) \geq 0$ for all $\hat{\boldsymbol{x}} \in \mathcal{X}$;*
- *If $GAP_{\mathcal{X}}(\hat{\boldsymbol{x}}) = 0$ and $\mathcal{X}$ contains a neighbourhood of $\hat{\boldsymbol{x}}$, then $\hat{\boldsymbol{x}}$ is a solution of (VI).*

A.4 Clarifications about Algorithm 1

One possible solution of efficiently estimating the distributions of dual vectors (line 4 in Algorithm 1) is to use a parametric model of density estimation such as modelling via truncated normal with efficiently computing sufficient statistics [18]. The set of update steps set U in Algorithm 1 is determined by the dynamics of distribution of of normalized dual vectors over the course of training. In Section 5, we use L-Greco [16] to update the levels.

B Proof of Quantization Variance Bound

Theorem 4.1 (Quantization Variance Bound)**.** *Let $\boldsymbol{v} \in \mathbb{R}^d$ be a vector to be quantized with L^q normalization. With unbiased quantization of $\boldsymbol{v}$, i.e., $\mathbb{E}_{q_{\mathbb{L}^M}}\left[Q_{\mathbb{L}^M}(\boldsymbol{v})\right] = \boldsymbol{v}$, we have that*

$$\mathbb{E}_{q_{\mathbb{L}^M}}\left[\|Q_{\mathbb{L}^M}(\boldsymbol{v}) - \boldsymbol{v}\|_2^2\right] \leq \varepsilon_Q \|\boldsymbol{v}\|_2^2, \tag{6}$$

where

$$\varepsilon_Q = (\bar{\ell}^M - 1)^2/(4\bar{\ell}^M) + (\bar{\ell}_1^M)^2 d^{2/\min\{q,2\}} \mathbb{1}\{d < d_{th}\}/4 + (\bar{\ell}_1^M d^{2/\min\{q,2\}} - 1) d^{2/\min\{q,2\}} \mathbb{1}\{d \geq d_{th}\}.$$

Proof. First let us remind ourselves of the notations in the main paper. Fix a time t. Let the normalized coordinates be $\boldsymbol{u}$. Let $\bar{\ell}^m = \max_{0\le j\le \alpha_m} \ell^m_{j+1}/\ell^m_j$, and $\bar{\ell}^M = \max_{1\le m\le M} \bar{\ell}^M$. Denote the largest level 1 among the M sequences $\bar{\ell}^M_1 = \max_{1\le m\le M} \ell^M_1$. Also let $d_{th} = (2/\bar{\ell}^M_1)^{\min\{2,q\}}$. Let $\mathcal{B}^m_j \coloneqq [\ell^m_j, \ell^m_{j+1}]$ for $m \in [M], j \in [\alpha_m]$.

Now, we can rewrite the equation (Var) for a fixed time t as follows

$$\begin{aligned}
\mathbb{E}_{q_{\mathbb{L}^M}}\left[\|Q_{\mathbb{L}^M}(\boldsymbol{v}) - \boldsymbol{v}\|_2^2\right] &= \|\boldsymbol{v}\|_q^2 \sum_{m=1}^{M} \sum_{u_i\in\mathbb{S}^m} \sigma_Q^2(u_i; \boldsymbol{\ell}^m) \\
&= \|\boldsymbol{v}\|_q^2 \sum_{m=1}^{M} \sum_{u_i\in\mathbb{S}^m} (\ell^m_{\tau^m(u_i)+1} - u_i)(u_i - \ell^m_{\tau^m(u_i)}) \\
&= \|\boldsymbol{v}\|_q^2 \sum_{m=1}^{M} \left(\sum_{u_i\in\mathcal{B}^m_0} (\ell^m_1 - u_i)u_i + \sum_{j=1}^{\alpha_m} \sum_{u_i\in\mathcal{B}^m_j} (\ell^m_{j+1} - u_i)(u_i - \ell^m_j) \right).
\end{aligned}$$

We now find the minimum k^m_j, satisfying $(\ell^m_{j+1} - u_i)(u_i - \ell^m_j) \le k^m_j u_i^2$ for $u_i \in \mathcal{B}^m_j$ for $m \in [M]$, $j \in [\alpha_m]$. Let $u_i = \ell^m_j \theta$ for $1 \le \theta \le \ell^m_{j+1}/\ell^m_j$. Then, we have

$$\begin{aligned}
k^m_j &= \max_{1\le\theta\le \ell^m_{j+1}/\ell^m_j} \frac{(\ell^m_{j+1} - u_i)(u_i - \ell^m_j)}{(\ell^m_j\theta)^2} \\
&= \max_{1\le\theta\le \ell^m_{j+1}/\ell^m_j} \frac{(\ell^m_{j+1}/\ell^m_j - \theta)(\theta - 1)}{\theta^2} \\
&= \frac{(\ell^m_{j+1}/\ell^m_j - 1)^2}{4(\ell^m_{j+1}/\ell^m_j)},
\end{aligned}$$

where the last equality follows from a simple differentiation with respect to θ. Since the function $(x-1)^2/(4x)$ is monotonically increasing function for $x > 1$, we obtain

$$\frac{(\ell^m_{j+1}/\ell^m_j - 1)^2}{4(\ell^m_{j+1}/\ell^m_j)} \le \frac{(\bar{\ell}^M - 1)^2}{4\bar{\ell}^M},$$

which leads to

$$\begin{aligned}
\sum_{j=1}^{\alpha_m} \sum_{u_i\in\mathcal{B}^m_j} (\ell^m_{j+1} - u_i)(u_i - \ell^m_j) &\le \sum_{j=1}^{\alpha_m} \sum_{u_i\in\mathcal{B}^m_j} k^m_j u_i^2 \\
&= \sum_{j=1}^{\alpha_m} \sum_{u_i\in\mathcal{B}^m} \frac{(\ell^m_{j+1}/\ell^m_j - 1)^2}{4(\ell^m_{j+1}/\ell^m_j)} u_i^2 \\
&\le \sum_{j=1}^{\alpha_m} \sum_{u_i\in\mathcal{B}^m} \frac{(\bar{\ell}^M - 1)^2}{4\bar{\ell}^M} u_i^2 \\
&= \frac{(\bar{\ell}^M - 1)^2}{4\bar{\ell}^M} \sum_{u_i\in\mathbb{S}^m/\mathcal{B}^m_0} u_i^2,
\end{aligned}$$

yielding

$$\begin{aligned}
\|\boldsymbol{v}\|_q^2 \sum_{m=1}^{M} \sum_{j=1}^{\alpha_m} \sum_{u_i\in\mathcal{B}^m_j} (\ell^m_{j+1} - u_i)(u_i - \ell^m_j) &\le \|\boldsymbol{v}\|_q^2 \sum_{m=1}^{M} \frac{(\bar{\ell}^M - 1)^2}{4\bar{\ell}^M} \sum_{u_i\in\mathbb{S}^m/\mathcal{B}^m_0} u_i^2 \\
&= \|\boldsymbol{v}\|_q^2 \frac{(\bar{\ell}^M - 1)^2}{4\bar{\ell}^M} \sum_{m=1}^{M} \sum_{u_i\in\mathbb{S}^m/\mathcal{B}^m_0} u_i^2 \\
&\le \|\boldsymbol{v}\|_q^2 \frac{(\bar{\ell}^M - 1)^2}{4\bar{\ell}^M} \frac{\|\boldsymbol{v}\|_2^2}{\|\boldsymbol{v}\|_q^2} \\
&= \frac{(\bar{\ell}^M - 1)^2}{4\bar{\ell}^M} \|\boldsymbol{v}\|_2^2.
\end{aligned}$$

Next, we attempt to bound $\sum_{m=1}^{M}\sum_{u_i\in\mathcal{B}_0^m}(\ell_1^m-u_i)u_i$ with these two known lemmas

Lemma B.1. *Let $\boldsymbol{v}\in\mathbb{R}^d$. Then, for all $0<p<q$, we have $\|\boldsymbol{v}\|_q\le\|\boldsymbol{v}\|_p\le d^{1/p-1/q}\|\boldsymbol{v}\|_q$. This holds even when $q<1$ and $\|\cdot\|$ is merely a seminorm.*

Lemma B.2. *[17, Lemma 15] Let $p\in(0,1)$ and $u\in\mathcal{B}_0$. Then we have $u(\ell_1-u)\le K_p\ell_1^{2-p}u^p$, where*

$$K_p=\frac{1/p}{2/p-1}\left(\frac{1/p-1}{2/p-1}\right)^{1-p}.$$

Now, from these two lemma, for any $0<p<1$ and $q\le 2$, we obtain that

$$\begin{aligned}
\|\boldsymbol{v}\|_q^2\sum_{m=1}^{M}\sum_{u_i\in\mathcal{B}_0^m}(\ell_1^m-u_i)u_i &\le \|\boldsymbol{v}\|_q^2\sum_{m=1}^{M}\sum_{u_i\in\mathcal{B}_0^m}K_p(\ell_1^m)^{2-p}u_i^p\\
&\le \|\boldsymbol{v}\|_q^2K_p(\bar{\ell}_1^M)^{2-p}\sum_{m=1}^{M}\sum_{u_i\in\mathcal{B}_0^m}u_i^p\\
&= \|\boldsymbol{v}\|_q^2K_p(\bar{\ell}_1^M)^{2-p}\sum_{m=1}^{M}\sum_{u_i\in\mathcal{B}_0^m}\frac{|v_i|^p}{\|\boldsymbol{v}\|_q^p}\\
&\le K_p(\bar{\ell}_1^M)^{2-p}\|\boldsymbol{v}\|_p^p\|\boldsymbol{v}\|_q^{2-p}\\
&\le K_p(\bar{\ell}_1^M)^{2-p}\|\boldsymbol{v}\|_2^pd^{1-p/2}\|\boldsymbol{v}\|_2^{2-p}\\
&= K_p(\bar{\ell}_1^M)^{2-p}d^{1-p/2}\|\boldsymbol{v}\|_2^2,
\end{aligned}$$

where the penultimate inequality holds due to the first given lemma and $\|\boldsymbol{v}\|_q\le\|\boldsymbol{v}\|_2$ for $q\ge 2$. Now combining the bounds, we obtain

$$\mathbb{E}_{q_{\mathbb{L}^M}}[\|Q_{\mathbb{L}^M}(\boldsymbol{v})-\boldsymbol{v}\|_2^2]\le\left(\frac{(\bar{\ell}^M-1)^2}{4\bar{\ell}^M}+K_p(\bar{\ell}_1^M)^{2-p}d^{1-p/2}\right)\|\boldsymbol{v}\|_2^2.$$

Moreover, if $q\ge 1$, note that $\|\boldsymbol{v}\|_q^{2-p}\le\|\boldsymbol{v}\|_2^{2-p}d^{\frac{2-p}{\min\{2,q\}}-\frac{2-p}{2}}$, yielding

$$\mathbb{E}_{q_{\mathbb{L}^M}}[\|Q_{\mathbb{L}^M}(\boldsymbol{v})-\boldsymbol{v}\|_2^2]\le\left(\frac{(\bar{\ell}^M-1)^2}{4\bar{\ell}^M}+K_p(\bar{\ell}_1^M)^{2-p}d^{\frac{2-p}{\min\{2,q\}}}\right)\|\boldsymbol{v}\|_2^2.$$

Now we can minimize ε_Q with finding the optimal p^* by minimizing

$$\lambda(p)=\frac{1/p}{2/p-1}\left(\frac{1/p-1}{2/p-1}\right)^{1-p}\upsilon^{1-p}=\frac{1}{2-p}\left(\frac{1-p}{2-p}\right)^{1-p}\upsilon^{1-p}=(2-p)^{p-2}(1-p)^{1-p}\upsilon^{1-p},$$

where $\upsilon=\bar{\ell}_1^Md^{\frac{1}{\min\{2,q\}}}$. This is equivalent to minimizing the log

$$\log\lambda(p)=(p-2)\log(2-p)+(1-p)\log(1-p)+(1-p)\log(\upsilon)$$

Setting the derivative of $\log\lambda(p)$ to zero, we have

$$-1+\log(2-p^*)+1-\log(1-p^*)+\log(\upsilon)=0,$$

yielding the optimal p^* to be

$$p^*=\begin{cases}\dfrac{\upsilon-2}{\upsilon-1}, & \upsilon\ge 2\quad\text{or}\quad d\ge d_{th}\\ 0, & \upsilon<2\quad\text{or}\quad d<d_{th}.\end{cases}$$

In brief, we have

$$\varepsilon_Q=\frac{(\bar{\ell}^M-1)^2}{4\bar{\ell}^M}+(\bar{\ell}_1^Md^{\frac{2}{\min\{q,2\}}}-1)d^{\frac{2}{\min\{q,2\}}}\mathbb{1}\{d\ge d_{th}\}+\frac{1}{4}(\bar{\ell}_1^M)^2d^{\frac{2}{\min\{q,2\}}}\mathbb{1}\{d<d_{th}\}.$$

■

C Coding Framework

C.1 Further Details on Coding Framework

The choice of a specific lossless prefix code for encoding $\boldsymbol{q}_{\mathbb{L}^{t,M}}$ relies on the extent to which the distribution of the discrete alphabet of levels is known. If we can estimate or know the distribution of the frequency of the discrete alphabet $\Omega^{t,M}$, we can apply the classical Huffman coding with an efficient encoding/decoding scheme and achieve the minimum expected code-length among methods encoding symbols separately [51, 52]. On the other hand, if we only know smaller values are more frequent than larger values without knowing the distribution of the discrete alphabet, we can consider Elias recursive coding (ERC) [53].

The decoding $\mathrm{DEC} : \{0,1\}^* \to \mathbb{R}^d$ first reads C_q bits to reconstruct $\|\boldsymbol{v}\|_q$, then applies decoding scheme $\Psi^{-1} : \{0,1\}^* \to (\Omega^{t,M})^d$ to obtain normalized coordinates.

Given quantization levels $\boldsymbol{\ell}^{t,m}$ and the marginal PDF of normalized coordinates, K nodes can construct the Huffman tree in parallel. A Huffman tree of a source with $s+2$ symbols can be constructed in time $\mathcal{O}(s)$ through sorting the symbols by the associated probabilities. It is well-known that Huffman codes minimize the expected code-length:

Theorem C.1. *[51, Theorems 5.4.1 and 5.8.1] Let Z denote a random source with a discrete alphabet $\mathcal{Z}$. The expected code-length of an optimal prefix code to compress Z is bounded by $H(Z) \leq \mathbb{E}[L] \leq H(Z)+1$ where $H(Z) \leq \log_2(|\mathcal{Z}|)$ is the entropy of Z in bits.*

C.2 Proof of Code Length Bound for Coding Protocol

Theorem 4.2 (Code-length Bound)**.** *Let p_j^m denote the probability of occurrence of ℓ_j^m for $m \in [M]$ and $j \in [\alpha_m]$. Under the setting specified in Theorem 4.1, the expectation $\mathbb{E}_w \mathbb{E}_{\boldsymbol{q}_{\mathbb{L}^M}}\left[\mathrm{ENC}\left(Q_{\mathbb{L}^M}(g(\boldsymbol{x};\omega)); \mathbb{L}^M\right)\right]$ of the number of bits under the coding protocol is*

$$\mathbb{E}_\omega \mathbb{E}_{\boldsymbol{q}_{\mathbb{L}^M}}\left[\mathrm{ENC}\left(Q_{\mathbb{L}^M}(g(\boldsymbol{x};\omega)); \mathbb{L}^M\right)\right] = \mathcal{O}\left(\Big(-\sum_{m=1}^{M} p_0^m - \sum_{m=1}^{M}\sum_{j=1}^{\alpha_m} p_j^m \log p_j^m\Big) d\right). \tag{7}$$

Proof. Following the Coding Protocol , we first use a constant C_q bits to represent the positive scalar $\|\boldsymbol{v}\|_q$ with a standard 32-bit floating point encoding. Then we use 1 bit to encode the sign of each nonzero entry of $\boldsymbol{u}$. Next, the probabilities associated with the symbols to be encoded, i.e., the levels in Ω^M, can be computed using the weighted sum of the conditional CDFs of normalized coordinates as follows.

Proposition C.2. *Let $j \in [\alpha_m]$, we have the probability p_j^m of occurrence of ℓ_j^m is*

$$p_j^m = Pr(\ell_j^m) = \int_{\ell_{j-1}^m}^{\ell_j^m} \frac{u-\ell_{j-1}^m}{\ell_j^m - \ell_{j-1}^m}\,\mathrm{d}\tilde{F}(u) + \int_{\ell_j^m}^{\ell_{j+1}^m} \frac{\ell_{j+1}^m - u}{\ell_{j+1}^m - \ell_j^m}\,\mathrm{d}\tilde{F}(u),$$

where $\tilde{F}(u)$ is the weighted sum of the conditional CDFs as defined in (2). Consequently we deduce

$$p_0^m = Pr(\ell_0^m) = \int_{\ell_0^m}^{\ell_1^m} \frac{\ell_1^m - u}{\ell_1^m - \ell_0^m}\,\mathrm{d}\tilde{F}(u) = \int_0^{\ell_1^m} \frac{\ell_1^m - u}{\ell_1^m}\,\mathrm{d}\tilde{F}(u),$$

$$p_{\alpha_m+1}^m = Pr(\ell_{\alpha_m+1}^m) = \int_{\ell_{\alpha_m}^m}^{\ell_{\alpha_m+1}^m} \frac{u-\ell_{\alpha_m}^m}{\ell_{\alpha_m+1}^m - \ell_{\alpha_m}^m}\,\mathrm{d}\tilde{F}(u) = \int_{\ell_{\alpha_m}^m}^{1} \frac{u-\ell_{\alpha_m}^m}{1-\ell_{\alpha_m}^m}\,\mathrm{d}\tilde{F}(u).$$

Then, we can get the expected number of non-zeros after quantization.

Lemma C.3. *For arbitrary $\boldsymbol{v} \in \mathbb{R}^d$, the expected number of non-zeros in $Q_{\mathbb{L}}^M(\boldsymbol{v})$ is*

$$\mathbb{E}\left[\|Q_{\mathbb{L}}^M(\boldsymbol{v})\|_0\right] = \left(1 - \sum_{m=1}^{M} p_0^m\right) d.$$

The optimal expected code-length for transmitting one random symbol is within one bit of the entropy of the source [51]. Hence, we can transmit entries of normalized $\boldsymbol{u}$ in at most $\left(\sum_{m=1}^{M} H(\boldsymbol{\ell}^m) + 1\right) d$, where $H(\boldsymbol{\ell}^m) = -\sum_{j=1}^{\alpha_m} p_j^m \log(p_j^m)$ is the entropy in bits.

In brief, we obtain

$$\mathbb{E}_w \mathbb{E}_{\boldsymbol{q}_{\mathbb{L}^M}} \left[\mathrm{ENC}\left(Q_{\mathbb{L}^M}(g(\boldsymbol{x};\omega)); \mathbb{L}^M \right) \right] = C_q + \left(1 - \sum_{m=1}^{M} p_0^m\right) d + \left(\sum_{m=1}^{M} H(\boldsymbol{\ell}^m) + 1\right) d.$$

■

C.3 Unbiased Compression under Absolute Noises

The following two lemmas show how additional noise due to compression affects the upper bounds under absolute noise Assumption 2.4. Let's keep in mind that $\boldsymbol{q}_{\mathbb{L}^M} \sim \mathbb{P}_Q$ represent d variables sampled independently for random quantization, and $\boldsymbol{q}_{\mathbb{L}^M}$ is independent of random sample $w \sim \mathbb{P}$.

Lemma C.4 (Unbiased Compression under Absolute Noise)**.** *Let $\boldsymbol{x} \in \mathcal{X}$ and $w \sim \mathbb{P}$. Suppose the oracle $g(\boldsymbol{x};\omega)$ satisfies Assumption 2.4. Suppose $Q_{\mathbb{L}^M}$ satisfies Theorem 4.1 and Theorem 4.2, then the compressed $Q_{\mathbb{L}^M}(g(\boldsymbol{x};\omega))$ satisfies Assumption 2.4 with*

$$\mathbb{E}\left[\|Q_{\mathbb{L}^M}(g(\boldsymbol{x};\omega)) - A(\boldsymbol{x})\|_2^2\right] \leq \varepsilon_Q(2L^2D^2 + 2\|A(X_1)\|_2^2 + \sigma^2) + \sigma^2.$$

Proof. The unbiasedness property immediately follows from the construction of the unbiased quantization $Q_{\mathbb{L}^M}$. Next, we note that that the maximum norm increase when compressing $Q_{\mathbb{L}^M}(g(\boldsymbol{x};\omega))$ occurs when each normalized coordinate of $g(\boldsymbol{x};\omega)$, $\{u_i\}_{i\in[d]}$, is mapped to the upper level $\ell^m_{\tau^m(u_i)+1}$ for some $m \in [M]$. We can show bounded absolute variance as follows

$$\begin{aligned}
\mathbb{E}_w \mathbb{E}_{\boldsymbol{q}_{\mathbb{L}^M}} \left[\|Q_{\mathbb{L}^M}(g(\boldsymbol{x};\omega)) - A(x)\|_2^2\right] &= \mathbb{E}_w \mathbb{E}_{\boldsymbol{q}_{\mathbb{L}^M}} \left[\|Q_{\mathbb{L}^M}(g(\boldsymbol{x};\omega)) - g(\boldsymbol{x};\omega) \right. \\
&\quad \left. + g(\boldsymbol{x};\omega) - A(x)\|_2^2\right] \\
&= \mathbb{E}_w \mathbb{E}_{\boldsymbol{q}_{\mathbb{L}^M}} \left[\|Q_{\mathbb{L}^M}(g(\boldsymbol{x};\omega)) - g(\boldsymbol{x};\omega)\|_2^2\right] \\
&\quad + \mathbb{E}_w \left[\|U(\boldsymbol{x};\omega)\|_2^2\right] \\
&\leq \varepsilon_Q \mathbb{E}_w \left[\|g(\boldsymbol{x};\omega)\|_2^2\right] + \sigma^2 \\
&= \varepsilon_Q \mathbb{E}_w \left[\|A(\boldsymbol{x}) + U(\boldsymbol{x};\omega)\|_2^2\right] + \sigma^2 \\
&= \varepsilon_Q \|A(\boldsymbol{x})\|_2^2 + \varepsilon_Q \mathbb{E}_w \left[\|U(\boldsymbol{x};\omega)\|_2^2\right] + \sigma^2 \\
&\leq \varepsilon_Q \|A(\boldsymbol{x})\|_2^2 + \varepsilon_Q \sigma^2 + \sigma^2,
\end{aligned}$$

where the second equality occurs due to unbiasedness of $\boldsymbol{q}_{\mathbb{L}^M}$, the third steps follos from Theorem 4.1, and the last inequality holds according to Assumption 2.4 for $g(\boldsymbol{x};\omega)$.

Now we note that in Theorem 4.3, $D^2 := \sup_{\boldsymbol{x}\in\mathcal{X}} \|X_1 - \boldsymbol{x}\|_2^2$, where $\mathcal{X} \subset \mathbb{R}^d$ is a compact neighborhood of a VI solution. Since A is L-Lipschitz (Assumption 2.3), we note that

$$\|A(X_1) - A(\boldsymbol{x})\|_2^2 \leq L^2 \|X_1 - \boldsymbol{x}\|_2^2 \leq L^2 D^2 \quad \forall\, \boldsymbol{x} \in \mathcal{X}.$$

Since X_1 is our initialization, $A(X_1)$ has a finite value, so $A(\boldsymbol{x})$ is bounded for all $\boldsymbol{x} \in \mathcal{X}$. Hence for the quantization in Algorithm 1, we can obtain

$$\|A(\boldsymbol{x})\|_2^2 \leq 2\|A(X_1) - A(\boldsymbol{x})\|_2^2 + 2\|A(X_1)\|_2^2 \leq 2L^2D^2 + 2\|A(X_1)\|_2^2,$$

which implies the desired conclusion. ■

D QODA Convergence Analysis

Proposition D.1 (Template Inequality)**.** *Suppose the iterates X_t of (4) are updated with non-increasing step-size schedule γ_t and η_t as in (5) for all $t = 1/2, 1, \ldots$. Then for any $X \in \mathbb{R}^d$,*

we have

$$\sum_{t=1}^{T}\left\langle \frac{1}{K}\sum_{k=1}^{K}\hat{V}_{k,t+1/2}, X_{t+1/2}-X\right\rangle$$
$$\leq \frac{\|X\|_*^2}{2\eta_{T+1}}+\sum_{t=1}^{T}\frac{\eta_t}{2K^2}\sum_{k=1}^{K}\left\|\hat{V}_{k,t+1/2}-\hat{V}_{k,t-1/2}\right\|_*^2-\sum_{t=1}^{T}\frac{\|X_t-X_{t+1/2}\|_*^2}{2\eta_t}.$$

Proof. First, decompose the LHS individual term $\frac{1}{K}\left\langle\sum_{k=1}^{K}\hat{V}_{k,t+1/2}, X_{t+1/2}-X\right\rangle$ into two terms as follows

$$\frac{1}{K}\left\langle\sum_{k=1}^{K}\hat{V}_{k,t+1/2}, X_{t+1/2}-X\right\rangle = A+B,$$

where

$$A=\frac{1}{K}\left\langle\sum_{k=1}^{K}\hat{V}_{k,t+1/2}, X_{t+1/2}-X_{t+1}\right\rangle,\ B=\frac{1}{K}\left\langle\sum_{k=1}^{K}\hat{V}_{k,t+1/2}, X_{t+1}-X\right\rangle.$$

From the update rule of 4 (with η_t), note that

$$\begin{aligned}
B&=\langle Y_t-Y_{t+1}, X_{t+1}-X\rangle\\
&=\left\langle Y_t-\frac{\eta_{t+1}}{\eta_t}Y_{t+1}, X_{t+1}-X\right\rangle+\left\langle\frac{\eta_{t+1}}{\eta_t}Y_{t+1}-Y_{t+1}, X_{t+1}-X\right\rangle\\
&=\frac{1}{\eta_t}\langle\eta_tY_t-\eta_{t+1}Y_{t+1}, X_{t+1}-X\rangle+\left(\frac{1}{\eta_{t+1}}-\frac{1}{\eta_t}\right)\langle-\eta_{t+1}Y_{t+1}, X_{t+1}-X\rangle\\
&=\frac{1}{\eta_t}\langle X_t-X_{t+1}, X_{t+1}-X\rangle+\left(\frac{1}{\eta_{t+1}}-\frac{1}{\eta_t}\right)\langle X_1-X_{t+1}, X_{t+1}-X\rangle\\
&=\frac{1}{2\eta_t}\left(\|X_t-X\|_*^2-\|X_t-X_{t+1}\|_*^2-\|X_{t+1}-X\|_*^2\right)\\
&+\left(\frac{1}{2\eta_{t+1}}-\frac{1}{2\eta_t}\right)\left(\|X_1-X\|_*^2-\|X_1-X_{t+1}\|_*^2-\|X_{t+1}-X\|_*^2\right)\\
&\leq\frac{1}{2\eta_t}\|X_t-X\|_*^2-\frac{1}{2\eta_t}\|X_t-X_{t+1}\|_*^2\\
&-\frac{1}{2\eta_{t+1}}\|X_{t+1}-X\|_*^2+\left(\frac{1}{2\eta_{t+1}}-\frac{1}{2\eta_t}\right)\|X_1-X\|_*^2,
\end{aligned}$$

the last inequality holds as the non-positive term $-\left(\frac{1}{2\eta_{t+1}}-\frac{1}{2\eta_t}\right)\|X_1-X_{t+1}\|_*^2$ is dropped. We can rearrange the above inequality as

$$\begin{aligned}
\frac{1}{2\eta_{t+1}}\|X_{t+1}-X\|_*^2&\leq\frac{1}{2\eta_t}\|X_t-X\|_*^2-\frac{1}{2\eta_t}\|X_t-X_{t+1}\|_*^2+\left(\frac{1}{2\eta_{t+1}}-\frac{1}{2\eta_t}\right)\|X\|_*^2-B\\
&=\frac{1}{2\eta_t}\|X_t-X\|_*^2-\frac{1}{2\eta_t}\|X_t-X_{t+1}\|_*^2+\left(\frac{1}{2\eta_{t+1}}-\frac{1}{2\eta_t}\right)\|X\|_*^2\\
&+\frac{1}{K}\left\langle\sum_{k=1}^{K}\hat{V}_{k,t+1/2}, X_{t+1/2}-X_{t+1}\right\rangle-\frac{1}{K}\left\langle\sum_{k=1}^{K}\hat{V}_{k,t+1/2}, X_{t+1/2}-X\right\rangle.
\end{aligned}\tag{*}$$

Next, also by the update rule (with γ_t), we have for any $X\in\mathbb{R}^d$

$$\begin{aligned}
\frac{\eta_t}{K}\left\langle\sum_{k=1}^{K}\hat{V}_{k,t-1/2}, X_{t+1/2}-X\right\rangle&\leq\frac{\gamma_t}{K}\left\langle\sum_{k=1}^{K}\hat{V}_{k,t-1/2}, X_{t+1/2}-X\right\rangle\\
&=\langle X_t-X_{t+1/2}, X_{t+1/2}-X\rangle\\
&=\frac{1}{2}\|X_t-X\|_*^2-\frac{1}{2}\|X_t-X_{t+1/2}\|_*^2-\frac{1}{2}\|X_{t+1/2}-X\|_*^2.
\end{aligned}$$

Substituting $X = X_{t+1}$ and dividing both sides of the inequality by η_t, we have

$$\frac{1}{K}\left\langle \sum_{k=1}^{K} \hat{V}_{k,t-1/2}, X_{t+1/2} - X_{t+1} \right\rangle$$
$$\leq \frac{1}{2\eta_t}\|X_t - X_{t+1}\|_*^2 - \frac{1}{2\eta_t}\|X_t - X_{t+1/2}\|_*^2 - \frac{1}{2\eta_t}\|X_{t+1/2} - X_{t+1}\|_*^2. \qquad (**)$$

Combining (*) with (**) and after some rearrangements, we obtain

$$\begin{aligned}\frac{1}{K}\left\langle \sum_{k=1}^{K} \hat{V}_{k,t+1/2}, X_{t+1/2} - X \right\rangle &\leq \frac{1}{2\eta_t}\|X_t - X\|_*^2 - \frac{1}{2\eta_{t+1}}\|X_{t+1} - X\|_*^2 \\ &+ \left(\frac{1}{2\eta_{t+1}} - \frac{1}{2\eta_t}\right)\|X_1 - X\|_*^2 \\ &+ \frac{1}{K}\left\langle \sum_{k=1}^{K} \hat{V}_{k,t+1/2} - \hat{V}_{k,t-1/2}, X_{t+1/2} - X_{t+1} \right\rangle \\ &- \frac{1}{2\eta_t}\|X_t - X_{t+1/2}\|_*^2 - \frac{1}{2\eta_t}\|X_{t+1/2} - X_{t+1}\|_*^2.\end{aligned}$$

Then, by summing the above expression over $t = 1, 2, \ldots, T$ and with some telescoping terms, we obtain

$$\begin{aligned}\sum_{t=1}^{T}\frac{1}{K}\left\langle \sum_{k=1}^{K} \hat{V}_{k,t+1/2}, X_{t+1/2} - X \right\rangle &\leq \frac{1}{2\eta_1}\|X_1 - X\|_*^2 - \frac{1}{2\eta_{T+1}}\|X_{T+1} - X\|_*^2 \\ &+ \left(\frac{1}{2\eta_{T+1}} - \frac{1}{2\eta_1}\right)\|X_1 - X\|_*^2 \\ &+ \sum_{t=1}^{T}\frac{1}{K}\left\langle \sum_{k=1}^{K}\left(\hat{V}_{k,t+1/2} - \hat{V}_{k,t-1/2}\right), X_{t+1/2} - X_{t+1} \right\rangle \\ &- \sum_{t=1}^{T}\frac{1}{2\eta_t}\|X_t - X_{t+1/2}\|_*^2 - \sum_{t=1}^{T}\frac{1}{2\eta_t}\|X_{t+1/2} - X_{t+1}\|_*^2.\end{aligned}$$

Next we consider the substitution $X_1 = 0$ which is just for notation simplicity and can be relaxed at the expense of obtaining a slightly more complicated expression. We can further drop the term $\frac{1}{2\eta_{T+1}}\|X_{T+1} - X\|_*^2$ to obtain

$$\begin{aligned}\frac{1}{K}\sum_{t=1}^{T}\left\langle \sum_{k=1}^{K} \hat{V}_{k,t+1/2}, X_{t+1/2} - X \right\rangle &\leq \frac{1}{2\eta_{T+1}}\|X\|_*^2 \\ &+ \frac{1}{K}\sum_{t=1}^{T}\left\langle \sum_{k=1}^{K}\left(\hat{V}_{k,t+1/2} - \hat{V}_{k,t-1/2}\right), X_{t+1/2} - X_{t+1} \right\rangle \\ &- \sum_{t=1}^{T}\frac{1}{2\eta_t}\|X_t - X_{t+1/2}\|_*^2 - \sum_{t=1}^{T}\frac{1}{2\eta_t}\|X_{t+1/2} - X_{t+1}\|_*^2.\end{aligned} \qquad (\dagger)$$

Note that by Cauchy-Schwarz and triangle inequalities, we have

$$\begin{aligned}&\frac{1}{K}\left\langle \sum_{k=1}^{K}\left(\hat{V}_{k,t+1/2} - \hat{V}_{k,t-1/2}\right), X_{t+1/2} - X_{t+1} \right\rangle \\ &= \frac{1}{K}\sum_{k=1}^{K}\left\langle \hat{V}_{k,t+1/2} - \hat{V}_{k,t-1/2}, X_{t+1/2} - X_{t+1} \right\rangle \\ &\leq \sum_{k=1}^{K}\left\|\hat{V}_{k,t+1/2} - \hat{V}_{k,t-1/2}\right\|_* \quad \left\|\frac{X_{t+1/2} - X_{t+1}}{K}\right\|_*.\end{aligned}$$

Combining with the AM-GM inequality of the form $xy \leq \frac{\eta_t}{2K^2}x^2 + \frac{K^2}{2\eta_t}y^2$, we deduce from (†) further that

$$\begin{aligned}&\frac{1}{K}\sum_{t=1}^{T}\left\langle\sum_{k=1}^{K}\left(\hat{V}_{k,t+1/2}-\hat{V}_{k,t-1/2}\right), X_{t+1/2}-X_{t+1}\right\rangle\\&\leq\sum_{t=1}^{T}\frac{\eta_t}{2K^2}\sum_{k=1}^{K}\left\|\hat{V}_{k,t+1/2}-\hat{V}_{k,t-1/2}\right\|_*^2+\sum_{t=1}^{T}\frac{1}{2\eta_t}\|X_{t+1/2}-X_{t+1}\|_*^2. \qquad (\dagger\dagger)\end{aligned}$$

Plugging (††) into (†), we obtain

$$\begin{aligned}\frac{1}{K}\sum_{t=1}^{T}\left\langle\sum_{k=1}^{K}\hat{V}_{k,t+1/2}, X_{t+1/2}-X\right\rangle\leq\frac{\|X\|_*^2}{2\eta_{T+1}}+\sum_{t=1}^{T}\sum_{k=1}^{K}\frac{\eta_t}{2K^2}\left\|\hat{V}_{k,t+1/2}-\hat{V}_{k,t-1/2}\right\|_*^2\\-\sum_{t=1}^{T}\frac{1}{2\eta_t}\|X_t-X_{t+1/2}\|_*^2,\end{aligned}$$

as desired. ■

We first introduce following two useful lemmas that will help to bound the (GAP):

Lemma D.2. *[54, 55] For all non-negative numbers $\alpha_1,\ldots,\alpha_t$, it holds that*

$$\sqrt{\sum_{t=1}^{T}\alpha_t}\leq\sum_{t=1}^{T}\frac{\alpha_t}{\sqrt{\sum_{i=1}^{t}\alpha_i}}\leq 2\sqrt{\sum_{t=1}^{T}\alpha_t}.$$

Lemma D.3. *[19] Let $\mathcal{C}\in\mathbb{R}^d$ be a convex set and $h:\mathcal{C}\to\mathbb{R}$ be a 1-strongly convex w.r.t. a norm $\|\cdot\|$. Assume that $h(\boldsymbol{x})-\min_{\boldsymbol{x}\in\mathcal{C}}h(\boldsymbol{x})\leq D^2/2$ for all $\boldsymbol{x}\in\mathcal{C}$. Then, for any martingale difference $(\boldsymbol{z}_t)_{t=1}^T\in\mathbb{R}^d$ and any $\boldsymbol{x}\in\mathcal{C}$, we have*

$$\mathbb{E}\left[\left\langle\sum_{t=1}^{T}\boldsymbol{z}_t,\boldsymbol{x}\right\rangle\right]\leq\frac{D^2}{2}\sqrt{\sum_{t=1}^{T}\mathbb{E}[\|\boldsymbol{z}_t\|^2]}. \qquad (8)$$

Now we state and prove the complexity of Algorithm 1 under absolute noise and fixed compression scheme.

Theorem 4.3. *Suppose the iterates X_t of Algorithm 1 are updated with learning rate schedule in (5) for all $t=1/2,1,\ldots,T$. Let $\mathcal{X}\subset\mathbb{R}^d$ be a compact neighborhood of a VI solution and $D^2:=\sup_{\boldsymbol{p}\in\mathcal{X}}\|X_1-\boldsymbol{p}\|_2^2$. Under Assumptions 2.1, 2.2, 2.3, and 2.4, we have*

$$\mathbb{E}\left[\mathrm{Gap}_{\mathcal{X}}\left(\sum_{t=1}^{T}X_{t+1/2}/T\right)\right]=\mathcal{O}\left(\left((LD+\|A(X_1)\|_2+\sigma)\widehat{\varepsilon_Q}+\sigma\right)D^2L^2/\sqrt{TK}\right).$$

Proof. Suppose first that no compression is applied, i.e., $\varepsilon_Q=0$. Using the result of the template inequality Proposition D.1, we can drop the negative term to obtain

$$\frac{1}{K}\sum_{t=1}^{T}\left\langle\sum_{k=1}^{K}\hat{V}_{k,t+1/2}, X_{t+1/2}-X\right\rangle\leq\frac{\|X\|_*^2}{2\eta_{T+1}}+\sum_{t=1}^{T}\sum_{k=1}^{K}\frac{\eta_t}{2K^2}\|\hat{V}_{k,t+1/2}-\hat{V}_{k,t-1/2}\|_*^2.$$

Next we can expand the LHS with the absolute noise model Assumption 2.4 as follows

$$\begin{aligned}
LHS &= \frac{1}{K}\sum_{t=1}^{T}\left\langle \sum_{k=1}^{K} A_k(X_{t+1/2}), X_{t+1/2} - X \right\rangle + \frac{1}{K}\sum_{t=1}^{T}\left\langle \sum_{k=1}^{K} U_k(X_{t+1/2}), X_{t+1/2} - X \right\rangle \\
&\geq \frac{1}{K}\sum_{t=1}^{T}\left\langle \sum_{k=1}^{K} A_k(X), X_{t+1/2} - X \right\rangle + \frac{1}{K}\sum_{t=1}^{T}\left\langle \sum_{k=1}^{K} U_k(X_{t+1/2}), X_{t+1/2} - X \right\rangle \\
&= \frac{1}{K}\left\langle \sum_{k=1}^{K} A_k(X), \sum_{t=1}^{T} X_{t+1/2} - \sum_{t=1}^{T} X \right\rangle + \frac{1}{K}\sum_{t=1}^{T}\left\langle \sum_{k=1}^{K} U_k(X_{t+1/2}), X_{t+1/2} - X \right\rangle \\
&= \frac{T}{K}\sum_{k=1}^{K}\left\langle A_k(X), \bar{X}_{T+1/2} - X \right\rangle + \frac{1}{K}\sum_{t=1}^{T}\left\langle \sum_{k=1}^{K} U_k(X_{t+1/2}), X_{t+1/2} - X \right\rangle,
\end{aligned}$$

where the second inequality follows from the monotonicity of A and $\bar{X}_{T+1/2} = \sum_{t=1}^{T} X_{t+1/2}/T$. Plugging this back to the result from template inequality with some rearrangement, we obtain

$$\begin{aligned}
\frac{1}{K}\sum_{k=1}^{K}\left\langle A_k(X), \bar{X}_{T+1/2} - X \right\rangle \leq \frac{1}{T}\Bigg(&\frac{\|X\|_*^2}{2\eta_{T+1}} + \sum_{t=1}^{T}\sum_{k=1}^{K}\frac{\eta_t}{2K^2}\|\hat{V}_{k,t+1/2} - \hat{V}_{k,t-1/2}\|_*^2 \\
&+ \frac{1}{K}\sum_{t=1}^{T}\left\langle \sum_{k=1}^{K} U_k(X_{t+1/2}), X - X_{t+1/2} \right\rangle\Bigg).
\end{aligned}$$

By taking the supremum over X, then dividing by T and then taking expectation on both sides, we get

$$\mathbb{E}\left[\sup_X \frac{1}{K}\sum_{k=1}^{K}\left\langle A_k(X), \bar{X}_{T+1/2} - X \right\rangle\right] \leq \frac{1}{T}(S_1 + S_2 + S_3),$$

where

$$S_1 = \mathbb{E}\left[\frac{D^2}{2\eta_{T+1}}\right],\ S_2 = \mathbb{E}\left[\sum_{t=1}^{T}\sum_{k=1}^{K}\frac{\eta_t}{2K^2}\|\hat{V}_{k,t+1/2} - \hat{V}_{k,t-1/2}\|_*^2\right],$$

$$S_3 = \mathbb{E}\left[\sup_X \frac{1}{K}\sum_{t=1}^{T}\left\langle \sum_{k=1}^{K} U_k(X_{t+1/2}), X - X_{t+1/2} \right\rangle\right].$$

Here we make an important observation that

$$\begin{aligned}
\mathbb{E}\left[\sum_{k=1}^{K}\left\|\hat{V}_{k,t+1/2} - \hat{V}_{k,t-1/2}\right\|_*^2\right] &\leq 2\mathbb{E}\left[\sum_{k=1}^{K}\left\|A_k(X_{t+1/2}) - A_k(X_{t-1/2})\right\|_*^2\right] \\
&\quad + 2\mathbb{E}\left[\sum_{k=1}^{K}\left\|U_k(X_{t+1/2}) - U_k(X_{t-1/2})\right\|_*^2\right] \\
&\leq 2\sum_{k=1}^{K} L^2\mathbb{E}\left[\left\|X_{t+1/2} - X_{t-1/2}\right\|_*^2\right] + 4K\sigma^2 \\
&\leq 2KL^2D^2 + 4K\sigma^2, \qquad (9)
\end{aligned}$$

where the second inequality comes from L-Lipschitzness the operator for the first summand and the absolute noise assumption for the second summand. Now we proceed to bound these terms one by one. For S_1, from the choice of learning rates $\eta_t \leq 1$, with Equation (9)we obtain

$$\begin{aligned}
S_1 &= D^2\mathbb{E}\left[\sqrt{1 + \sum_{t=1}^{T}\frac{1}{K^2}\sum_{k=1}^{K}\left\|\hat{V}_{k,t+1/2} - \hat{V}_{k,t-1/2}\right\|_*^2}\right] \\
&\leq D^2\sqrt{1 + \sum_{t=1}^{T}\mathbb{E}\left[\frac{1}{K^2}\sum_{k=1}^{K}\left\|\hat{V}_{k,t+1/2} - \hat{V}_{k,t-1/2}\right\|_*^2\right]} \\
&\leq D^2\sqrt{1 + \frac{2T(L^2D^2 + 2\sigma^2)}{K}}.
\end{aligned}$$

Next, we proceed to bound S_2

$$
\begin{aligned}
S_2 &= \mathbb{E}\left[\sum_{t=1}^{T}\sum_{k=1}^{K}\frac{\eta_t}{2K^2}\|\hat{V}_{k,t+1/2}-\hat{V}_{k,t-1/2}\|_*^2\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T}\sum_{k=1}^{K}\left(\frac{\eta_t}{2K^2}-\frac{\eta_{t+1}}{2K^2}\right)\|\hat{V}_{k,t+1/2}-\hat{V}_{k,t-1/2}\|_*^2\right] \\
&+ \mathbb{E}\left[\sum_{t=1}^{T}\sum_{k=1}^{K}\frac{\eta_{t+1}}{2K^2}\|\hat{V}_{k,t+1/2}-\hat{V}_{k,t-1/2}\|_*^2\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T}\left(\frac{\eta_t}{2K^2}-\frac{\eta_{t+1}}{2K^2}\right)(2KL^2D^2+4K\sigma^2)\right] \\
&+ \frac{1}{2}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{k=1}^{K}\frac{\|\hat{V}_{k,t+1/2}-\hat{V}_{k,t-1/2}\|_*^2/K^2}{\sqrt{1+\sum_{s=1}^{t}\sum_{k=1}^{K}\left\|\hat{V}_{k,s+1/2}-\hat{V}_{k,s-1/2}\right\|^2/K^2}}\right] \quad \text{(from Equation (9))} \\
&\leq 2L^2D^2+4\sigma^2+\frac{1}{2}\mathbb{E}\left[\sqrt{1+\frac{1}{K^2}\sum_{t=1}^{T}\sum_{k=1}^{K}\left\|\hat{V}_{k,t+1/2}-\hat{V}_{k,t-1/2}\right\|^2}\right] \quad \text{(from Lemma D.2)} \\
&\leq 2L^2D^2+4\sigma^2+\frac{1}{2}\sqrt{1+\frac{2T(L^2D^2+2\sigma^2)}{K}}.
\end{aligned}
$$

Lastly, let's consider S_3

$$
S_3 = \mathbb{E}\left[\sup_X \frac{1}{K}\sum_{t=1}^{T}\left\langle\sum_{k=1}^{K}U_k(X_{t+1/2}),X\right\rangle\right]-\mathbb{E}\left[\sup_X \frac{1}{K}\sum_{t=1}^{T}\left\langle\sum_{k=1}^{K}U_k(X_{t+1/2}),X_{t+1/2}\right\rangle\right]
$$

We can bound the first term with Lemma D.3 as follows

$$
\mathbb{E}\left[\sup_X \frac{1}{K}\sum_{t=1}^{T}\left\langle\sum_{k=1}^{K}U_k(X_{t+1/2}),X\right\rangle\right] \leq \frac{D^2}{2K}\sqrt{\mathbb{E}\left[\sum_{t=1}^{T}\sum_{k=1}^{K}\|U_{k,t+1/2}\|^2\right]} \leq \frac{D^2\sigma\sqrt{T}}{2\sqrt{K}}
$$

For the second term, we use law of total expectation

$$
\mathbb{E}\left[\sum_{t=1}^{T}\left\langle\sum_{k=1}^{K}U_k(X_{t+1/2}),X_{t+1/2}\right\rangle\right] = \mathbb{E}\left[\sum_{t=1}^{T}\sum_{k=1}^{K}\mathbb{E}\left[\left\langle U_k(X_{t+1/2}),X_{t+1/2}\right\rangle|X_{t+1/2}\right]\right] = 0,
$$

implying

$$
S_3 \leq \frac{D^2\sigma\sqrt{T}}{2\sqrt{K}}.
$$

Combining the bounds of S_1, S_2 and S_3, we finally obtain the complexity without compression as

$$
\begin{aligned}
&\mathbb{E}\left[\mathrm{Gap}_{\mathcal{X}}\left(\bar{X}_{t+1/2}\right)\right] \\
&= \mathbb{E}\left[\sup_X \frac{1}{K}\sum_{k=1}^{K}\left\langle A_k(X),\bar{X}_{T+1/2}-X\right\rangle\right] \leq \frac{1}{T}\mathcal{O}\left(\frac{\sqrt{T}D^2L^2}{\sqrt{K}}\right) = \mathcal{O}\left(\frac{D^2L^2}{\sqrt{TK}}\right).
\end{aligned}
$$

Now, we consider applying layer-wise compression to this bound. Firstly, recall that the average square root expected code-length bound is denoted as

$$
\widehat{\varepsilon_Q} = \sum_{m=1}^{M}\sum_{j=1}^{J^m}\frac{T_{m,j}\sqrt{\varepsilon_{Q,m,j}}}{T}.
$$

Finally, by applying compression bound Lemma C.4 along the ideas of [18, Theorem 4] and [11, Theorem 3], we get the desired result

$$
\mathbb{E}\left[\mathrm{Gap}_{\mathcal{X}}\left(\bar{X}_{t+1/2}\right)\right] = \mathcal{O}\left(\frac{((LD+\|A(X_1)\|_2+\sigma)\widehat{\varepsilon_Q}+\sigma)\,D^2L^2}{\sqrt{TK}}\right).
$$

■