Supplementary Document: Quantized Compressive Sampling of Stochastic Gradients for Efficient Communication in Distributed Training

Afshin Abdi, Faramarz Fekri*
School of Electrical and Computer Engineering
Georgia Institute of Technology, Atlanta, GA
{abdi,fekri}@gatech.edu

1 Dithered Quantization

It is well-known that the error in ordinary quantization, especially when the number of quantization levels is low, depends on the input signal and is not necessarily uniformly distributed. In *Dithered Quantization*, a (pseudo-)random signal called dither is added to the input signal prior to quantization. Adding this controlled perturbation can cause the statistical behavior of the quantization error be more desirable (Schuchman 1964; Gray and Stockham 1993; Gray and Neuhoff 1998).

Let $Q(\cdot)$ be an M-level uniform quantizer with quantization step size of Δ, i.e., $Q(v) = \Delta \lfloor v/\Delta \rceil$ and the output range of $Q(\cdot)$ is $\{-M, \dots, 0, \dots, M\}$.[1] The dithered quantizer is defined as follows;

Definition (Dithered Quantization)**.** For an input signal x, let u be a dither signal, independent of x. The dithered quantization of x is defined as $\tilde{x} = Q(x + u) - u$.

Remark 1. To transmit the dithered quantization of x, it is sufficient to send the index of the quantization bin that $x + u$ resides in, i.e., $\lfloor (x + u)/\Delta \rceil$. The receiver reproduces the (pseudo-)random sequence u using the same random number generator algorithm and seed number as the sender. It is then subtracted from $Q(x + u)$ to reconstruct $\tilde{x}$.

Theorem I ((Schuchman 1964))**.** *If 1) the quantizer does not overload, i.e., $|x+u| \leq \frac{M\Delta}{2}$ for all input signals x and dither u, and 2) The characteristic function of the dither signal, defined as $M_u(j\nu) = \mathbb{E}_u\left[e^{j\nu u}\right]$, satisfies $M_u(j\frac{2\pi l}{\Delta}) = 0$ for all $l \neq 0$, then the quantization error $e = x - \tilde{x}$ is uniform over $(-\Delta/2, \Delta/2]$ and it is independent of the signal x.*

It is common to consider $\mathcal{U}(-\Delta/2, \Delta/2)$ as the distribution of the random dither signal which can be easily verified that it satisfies the conditions of Thm. I.

*This material is based upon work supported by the National Science Foundation under Grant Number CPS-1837369 and by Sony Inc. under the Sony Faculty Research Award.

[1]Throughout the paper, we assume that all quantizers are centered around 0 (with the exception of sign-based quantization). This is the case also for ternary (Wen et al. 2017) and stochastic quantization (Alistarh et al. 2017).

In some cases, the receiver may not be able to reproduce the dither signal to subtract from $Q(x + u)$. Hence, quantization is simply defined as $\tilde{x}_h = Q(x + u)$. We refer to this approach as the *half-dithered quantization* as the dither signal is applied only to the quantization, not the reconstruction of x. In this case, the quantization error is not necessarily independent of the signal, however by an appropriate choice of the dither signal, the moments of the quantization error will be independent (Gray and Stockham 1993). For example, if the dither signal u is the sum of k independent random variables, each having uniform distribution $\mathcal{U}(-\Delta/2, \Delta/2)$, then the k-th moment of the quantization error, $\epsilon = x - \tilde{x}_h$, would be independent of the signal; $\mathbb{E}\left[\epsilon^k | x\right] = \mathbb{E}\left[\epsilon^k\right]$.

Remark 2 (1-Bit Dithered Quantization). Note that the output range of the dithered quantization is $\{-M, \dots, +M\}$. Hence, each value is represented by minimum of $\log_2(1 + 2M)$ bits (without applying any compression to the quantized sequence). Reducing the number of bits to only 1-bit while keeping the desired properties of the dithered quantizers can potentially reduce the transmission bits by almost 50% (from at least $\log_2 3 \approx 1.58$ bits to 1 bit).

Without loss of generality, assume that $|x| \leq 1/2$. We propose the following dithered 1-bit quantization:

$$q = \text{sign}(x + u) := \begin{cases} +1 & \text{if } u + x > 0 \\ -1 & \text{o.w.} \end{cases}, \qquad (1)$$

where $u \sim \mathcal{U}(-1/2, 1/2)$ is the random dither signal. The dequatized value is then given by

$$\tilde{x} = q - u. \qquad (2)$$

It is straightforward to show that this 1-bit dithered quantizer is unbiased and the quantization noise is uniformly distributed and independent of x;

$$\mathbb{E}[\tilde{x} - x] = 0, \quad \text{Var}[\tilde{x} - x] = \frac{1}{12}. \qquad (3)$$

Relationship with Ternary and Stochastic Quantizations Without loss of generality, assume that the vector $\boldsymbol{x}$ is normalized such that $|x_i| \leq 1$. Although the reconstruction of quantized values in our method is different from those in TernGrad and QSGD, we show that these quantizers can be considered as a special case of the half-dithered quantizer.

M-level Stochastic Quantization in (Alistarh et al. 2017) is defined as

$$Q^{(s)}(x_i) = \begin{cases} \operatorname{sign}(x_i)\,\frac{l}{M} & \text{with prob. } 1-d(x_i) \\ \operatorname{sign}(x_i)\,\frac{l+1}{M} & \text{with prob. } d(x_i) \end{cases} \quad (4)$$

where l is the quantization bin that $|x_i|$ resides in, i.e., $|x_i| \in [l/M, (l+1)/M]$ and $d(x_i) := M|x_i| - l$. The ternary quantizer of (Wen et al. 2017) can be considered as a special case of stochastic quantizer with $M = 1$.

Lemma. *Stochastic quantization is the same as* $(2M+1)$*-level half-dithered quantizer with step-size* $\Delta = \frac{1}{M}$ *and uniform dither* $u \sim \mathcal{U}(-\frac{1}{2M}, \frac{1}{2M})$.

Proof. Let $Q(\cdot)$ be a $2M+1$-level quantizer with step size $\Delta = 1/M$. Let $u \sim \mathcal{U}[-\Delta/2, \Delta/2]$ be the dither signal. Let $0 \le x \le 1$ be an arbitrary number. Assume that $l/M \le x < (l+1)/M$ and define $d = x - l/M$. Note that $0 \le d < \Delta$ and

$$P\left(Q(x+u) = \frac{l}{M}\right) = P\left(|x+u-l/M| \le \frac{\Delta}{2}\right)$$
$$= P\left(u \le \frac{\Delta}{2} - d\right) = 1 - \frac{d}{\Delta} = 1 - Md.$$

Similarly, $P(Q(x+u) = (l+1)/M) = Md$. Comparing with stochastic quantizer, we see that they both assign the quantization points with the same probability. The case $x < 0$ can be verified similarly. ■

In other words, stochastic quantizer adds a uniformly distributed dither to the input signal before quantization. However, the receiver *does not* subtract the dither from the quantized value. Therefore, the quantization error is not independent of the signal. It can be easily verified that although the quantization is unbiased, $\mathbb{E}\big[\boldsymbol{x} - Q^{(s)}(\boldsymbol{x})\big] = \mathbf{0}$, its variance depends on the value of the input signal and varies in $[0, 1/4M^2]$, depending on the input signal x;

$$\mathbb{E}\Big[([Q^{(s)}(\boldsymbol{x}) - \boldsymbol{x}]_i)^2\Big] = \frac{d(x_i)(1-d(x_i))}{M^2}.$$

On the other hand, the variance of the dithered quantization noise would be uniformly $1/12M^2$, independent of x. For example, if x is uniformly distributed over $[-1, 1]$, the average quantization variance of the stochastic quantizer would be $1/6M^2$, twice the variance of the dithered quantization.

2 Properties of the Mixing Matrix

Lemma II. *The random mixing matrix* $\boldsymbol{T}$ *has the following properties:*

$$\boldsymbol{T}\boldsymbol{T}^{\mathsf{T}} = \frac{n}{k}\boldsymbol{I}, \quad \mathbb{E}\big[\boldsymbol{T}^{\mathsf{T}}\boldsymbol{T}\big] = \boldsymbol{I}. \quad (5)$$

Proof. Recall that $\boldsymbol{T} = \frac{1}{\sqrt{k}}\boldsymbol{H}\boldsymbol{R}$. Hence,

$$\boldsymbol{T}\boldsymbol{T}^{\mathsf{T}} = \frac{1}{k}\boldsymbol{H}\boldsymbol{R}\boldsymbol{R}^{\mathsf{T}}\boldsymbol{H}^{\mathsf{T}} \overset{(a)}{=} \frac{1}{k}\boldsymbol{H}\boldsymbol{H}^{\mathsf{T}} \overset{(b)}{=} \frac{n}{k}\boldsymbol{I},$$

where (a) is due to the fact that $\boldsymbol{R} = \operatorname{diag}(\boldsymbol{r})$ and $r_i^2 = 1$, and (b) is a result of $\boldsymbol{H}$ being any k rows of Hadamard matrix $\boldsymbol{H}_n$ satisfying $\boldsymbol{H}_n\boldsymbol{H}_n^{\mathsf{T}} = n\boldsymbol{I}$.

For the second property,

$$\mathbb{E}\big[\boldsymbol{T}^{\mathsf{T}}\boldsymbol{T}\big] = \frac{1}{k}\,\mathbb{E}\big[\boldsymbol{R}\boldsymbol{H}\boldsymbol{H}^{\mathsf{T}}\boldsymbol{R}^{\mathsf{T}}\big].$$

Now, consider an arbitrary (i,j)-th element,

$$[\boldsymbol{R}^{\mathsf{T}}\boldsymbol{H}^{\mathsf{T}}\boldsymbol{H}\boldsymbol{R}]_{i,j} = r_i\, r_j\ \big[\boldsymbol{H}^{\mathsf{T}}\boldsymbol{H}\big]_{i,j}.$$

On the other hand, $\mathbb{E}[r_i\, r_j] = 1$ if $i = j$ and 0 if $i \ne j$. Moreover, since $H_{i,l} = \pm 1$, $\big[\boldsymbol{H}^{\mathsf{T}}\boldsymbol{H}\big]_{i,i} = \sum_{l=1}^{k} (\boldsymbol{H}_{i,l})^2 = k$. Therefore,

$$\mathbb{E}\big[[\boldsymbol{R}^{\mathsf{T}}\boldsymbol{H}^{\mathsf{T}}\boldsymbol{H}\boldsymbol{R}]_{i,j}\big] = \begin{cases} 0 & i \ne j \\ 1 & i = j \end{cases}$$

■

Note on Remark 1 of Paper Let $\boldsymbol{M}_n$ be an $n \times n$ random Rademacher matrix, i.e., each entry is an i.i.d. random variable $\mathrm{P}(+1) = \mathrm{P}(-1) = 0.5$. (Bourgain, Vu, and Wood 2010) has shown that

$$\mathrm{P}(\boldsymbol{M}_n \text{ is singular}) \le \left(\frac{1}{\sqrt{2}} + o(1)\right)^n. \quad (6)$$

Now, for a constant c, consider collection of $m = cn/k$ of the QCS mixing matrix, $\boldsymbol{T}_i = \boldsymbol{H}\boldsymbol{R}_i$ for $i = 1, \ldots, m$, where $\boldsymbol{R}_i = \operatorname{diag}(\boldsymbol{r}_i)$ and $\boldsymbol{r}_i$'s are i.i.d. Rademacher random variables. Note that each of $\boldsymbol{T}_i$'s are full row-rank and the probability of $\boldsymbol{T}_i$'s not covering the entire $\mathbb{R}^n$ would be less than the probability of $\boldsymbol{M}_n$ being singular.

3 Unbiased QCS

Proof of Thm. 2

For a fixed $\boldsymbol{g}$, we note that the randomness in $\widehat{\boldsymbol{g}}$ stems from the random mixing matrix $\boldsymbol{T}$ and dither signal $\boldsymbol{u}$, which are independent of each other and $\boldsymbol{g}$. Moreover, the quantization noise of $\boldsymbol{v}$ can be written as

$$\widehat{\boldsymbol{v}} = \boldsymbol{v} - \varsigma\boldsymbol{\varepsilon} = \boldsymbol{T}\boldsymbol{g} - \varsigma\boldsymbol{\varepsilon}, \quad (7)$$

where as a result of Thm. I, $\boldsymbol{\varepsilon}$ is an independent random variable and $\boldsymbol{\varepsilon} \sim \mathcal{U}(-1/2, 1/2)$.[2] Therefore, the quantization noise can be decomposed as

$$\boldsymbol{e} = \boldsymbol{g} - \widehat{\boldsymbol{g}} = \overbrace{(\boldsymbol{I} - \boldsymbol{T}^{\mathsf{T}}\boldsymbol{T})\boldsymbol{g}}^{\boldsymbol{e}_g} + \overbrace{\varsigma\boldsymbol{T}^{\mathsf{T}}\boldsymbol{\varepsilon}}^{\boldsymbol{e}_d}. \quad (8)$$

Unbiasedness.

$$\mathbb{E}[\boldsymbol{e}_g] = (\boldsymbol{I} - \mathbb{E}\big[\boldsymbol{T}^{\mathsf{T}}\boldsymbol{T}\big])\boldsymbol{g} \overset{(a)}{=} \mathbf{0},$$
$$\mathbb{E}[\boldsymbol{e}_d] = \mathbb{E}\big[\varsigma\boldsymbol{T}^{\mathsf{T}}\boldsymbol{\varepsilon}\big] \overset{(b)}{=} \mathbb{E}\big[\varsigma\boldsymbol{T}^{\mathsf{T}}\big]\,\mathbb{E}[\boldsymbol{\varepsilon}] = \mathbf{0},$$

where (a) is due to $\mathbb{E}\big[\boldsymbol{T}^{\mathsf{T}}\boldsymbol{T}\big] = \boldsymbol{I}$ (Lemma II) and (b) is because of independence of $\boldsymbol{\varepsilon}$ from $\varsigma\boldsymbol{T}$ and $\mathbb{E}[\boldsymbol{\varepsilon}] = \mathbf{0}$. This proves the unbiasedness of QCS.

[2] Note that this is not the case for ordinary quantization or stochastic quantization of (Alistarh et al. 2017).

Variance. Note that since $\boldsymbol{e}_g$ is a function of only $\boldsymbol{T}$ and $\boldsymbol{g}$, it is independent of $\boldsymbol{\varepsilon}$ and

$$\mathbb{E}\big[\boldsymbol{e}_g^{\mathsf{T}}\boldsymbol{e}_d\big] = \mathbb{E}\big[\varsigma \boldsymbol{e}_g^{\mathsf{T}}\boldsymbol{T}^{\mathsf{T}}\big]\,\mathbb{E}[\boldsymbol{\varepsilon}] = 0.$$

Therefore,

$$\mathbb{E}\big[\|\boldsymbol{e}\|_2^2\big] = \mathbb{E}\big[\|\boldsymbol{e}_g\|_2^2\big] + \mathbb{E}\big[\|\boldsymbol{e}_d\|_2^2\big]$$

For an arbitrary $\boldsymbol{g}$, note that

$$\begin{aligned}\mathbb{E}\big[\|\boldsymbol{T}\boldsymbol{g}\|_2^2\big] &= \boldsymbol{g}^{\mathsf{T}}\,\mathbb{E}\big[\boldsymbol{T}^{\mathsf{T}}\boldsymbol{T}\big]\boldsymbol{g} = \boldsymbol{g}^{\mathsf{T}}\boldsymbol{I}\boldsymbol{g} = \|\boldsymbol{g}\|_2^2\\ \mathbb{E}\big[\|\boldsymbol{T}^{\mathsf{T}}\boldsymbol{T}\boldsymbol{g}\|_2^2\big] &= \mathbb{E}\big[\boldsymbol{g}^{\mathsf{T}}\boldsymbol{T}^{\mathsf{T}}\boldsymbol{T}\boldsymbol{T}^{\mathsf{T}}\boldsymbol{T}\boldsymbol{g}\big]\\ &= \frac{n}{k}\,\mathbb{E}\big[\boldsymbol{g}^{\mathsf{T}}\boldsymbol{T}^{\mathsf{T}}\boldsymbol{T}\boldsymbol{g}\big] = \frac{n}{k}\|\boldsymbol{g}\|_2^2.\end{aligned}$$

Therefore,

$$\begin{aligned}\mathbb{E}\big[\|\boldsymbol{e}_g\|_2^2\big] &= \mathbb{E}\big[\|(\boldsymbol{I}-\boldsymbol{T}^{\mathsf{T}}\boldsymbol{T})\boldsymbol{g}\|_2^2\big]\\ &= \mathbb{E}\big[\|\boldsymbol{T}^{\mathsf{T}}\boldsymbol{T}\boldsymbol{g}\|_2^2\big] + \|\boldsymbol{g}\|_2^2 - 2\,\mathbb{E}\big[\|\boldsymbol{T}\boldsymbol{g}\|_2^2\big]\\ &= (\frac{n}{k}-1)\|\boldsymbol{g}\|_2^2.\end{aligned}$$

On the other hand,

$$\begin{aligned}\mathbb{E}\big[\|\boldsymbol{e}_d\|_2^2\big] &= \mathbb{E}\big[\|\varsigma\boldsymbol{T}^{\mathsf{T}}\boldsymbol{\varepsilon}\|_2^2\big] = \mathbb{E}_{\boldsymbol{T}}\big[\mathbb{E}_{\boldsymbol{\varepsilon}}\big[\|\varsigma\boldsymbol{T}^{\mathsf{T}}\boldsymbol{\varepsilon}\|_2^2|\boldsymbol{T}\big]\big]\\ &= \mathbb{E}_{\boldsymbol{T}}\big[\varsigma^2\,\mathbb{E}_{\boldsymbol{\varepsilon}}\big[\|\boldsymbol{T}^{\mathsf{T}}\boldsymbol{\varepsilon}\|_2^2|\boldsymbol{T}\big]\big] \overset{(c)}{=} \mathbb{E}_{\boldsymbol{T}}\left[\varsigma^2\frac{\|\boldsymbol{T}\|_F^2}{12}\right]\\ &\overset{(d)}{=} \frac{n}{12Q^2}\,\mathbb{E}_{\boldsymbol{T}}\big[\|\boldsymbol{T}\boldsymbol{g}\|_\infty^2\big],\end{aligned}$$

where (c) is due to $\boldsymbol{\varepsilon}$ being i.i.d. $\mathcal{U}(-1/2, 1/2)$ and (d) is because of $\|\boldsymbol{T}\|_F^2 = n$ and definition of $\varsigma = \|\boldsymbol{T}\boldsymbol{g}\|_\infty/Q$.

To bound $\mathbb{E}_{\boldsymbol{T}}\big[\|\boldsymbol{T}\boldsymbol{g}\|_\infty^2\big]$ we need the following lemma.

Lemma III. *Let $\boldsymbol{a} \in \mathbb{R}^n$ be fixed and $\boldsymbol{r}$ be an i.i.d. Rademacher random vector. Then for all $0 \le \lambda < 1/(2\|\boldsymbol{a}\|_2^2)$,*

$$\mathbb{E}_{\boldsymbol{r}}\left[e^{\lambda(\boldsymbol{a}^{\mathsf{T}}\boldsymbol{r})^2}\right] \le \frac{1}{\sqrt{1-2\lambda\|\boldsymbol{a}\|_2^2}} \tag{9}$$

Proof. Let $\omega \sim \mathcal{N}(0,1)$ be an independent normal random variable. Note that for a fixed $\boldsymbol{r}$,

$$e^{\lambda(\boldsymbol{a}^{\mathsf{T}}\boldsymbol{r})^2} = \mathbb{E}_\omega\Big[\exp\Big(\sqrt{2\lambda}(\boldsymbol{a}^{\mathsf{T}}\boldsymbol{r})\omega\Big)\Big].$$

Therefore,

$$\begin{aligned}\mathbb{E}_{\boldsymbol{r}}\left[e^{\lambda(\boldsymbol{a}^{\mathsf{T}}\boldsymbol{r})^2}\right] &= \mathbb{E}_{\boldsymbol{r}}\left[\mathbb{E}_\omega\left[\exp\Big(\sqrt{2\lambda}(\boldsymbol{a}^{\mathsf{T}}\boldsymbol{r})\omega\Big)\,|\boldsymbol{r}\right]\right]\\ &= \mathbb{E}_\omega\left[\mathbb{E}_{\boldsymbol{r}}\left[\exp\Big(\sqrt{2\lambda}(\boldsymbol{a}^{\mathsf{T}}\boldsymbol{r})\omega\Big)\,|\omega\right]\right]\\ &\overset{(e)}{=} \mathbb{E}_\omega\left[\prod_{i=1}^n \mathbb{E}_{r_i}\left[\exp\Big(\sqrt{2\lambda}\omega a_i r_i\Big)\,|\omega\right]\right]\\ &\overset{(f)}{\le} \mathbb{E}_\omega\left[\prod_{i=1}^n \exp\big(\lambda\omega^2 a_i^2\big)\right]\\ &= \mathbb{E}_\omega\big[\exp\big(\lambda\|\boldsymbol{a}\|_2^2\omega^2\big)\big] = \frac{1}{\sqrt{1-2\lambda\|\boldsymbol{a}\|_2^2}}\end{aligned}$$

where (e) is because of independence of r_i's, (f) is from Hoeffding's lemma, ∎

Using the above lemma, for $\boldsymbol{v} = \boldsymbol{T}\boldsymbol{g}$ and arbitrary $\lambda > 0$,

$$\begin{aligned}\mathbb{E}\big[\|\boldsymbol{v}\|_\infty^2\big] &\le \frac{1}{\lambda}\,\mathbb{E}\left[\log\left(\sum_{i=1}^k \exp\big(\lambda v_i^2\big)\right)\right]\\ &\le \frac{1}{\lambda}\log\mathbb{E}\left[\sum_{i=1}^k \exp\big(\lambda v_i^2\big)\right]\\ &= \frac{1}{\lambda}\log\left(\sum_{i=1}^k \mathbb{E}\big[\exp(\lambda v_i^2)\big]\right)\end{aligned}$$

Note that for an arbitrary fixed i, $\zeta_j := r_j H_{i,j}$ would be i.i.d. Rademacher random variables and $v_i = \sum_j (\frac{g_j}{\sqrt{k}})\zeta_j$. Therefore, by Lemma III

$$\mathbb{E}\big[\exp(\lambda v_i^2)\big] \le \frac{1}{\sqrt{1-2\lambda\|\boldsymbol{g}\|_2^2/k}},$$

and

$$\mathbb{E}\big[\|\boldsymbol{v}\|_\infty^2\big] \le \frac{1}{\lambda}\log\left(\frac{k}{\sqrt{1-2\lambda\|\boldsymbol{g}\|_2^2/k}}\right).$$

For $k \ge 2$, let $\lambda = \frac{k}{2\|\boldsymbol{g}\|_2^2}(1-k^{-\delta})$. Therefore,

$$\begin{aligned}\frac{1}{\lambda}\log\left(\frac{k}{\sqrt{1-2\lambda\|\boldsymbol{g}\|_2^2/k}}\right) &= \frac{2\|\boldsymbol{g}\|_2^2}{k}\,\frac{1}{1-k^{-\delta}}\log(k^{1+\delta/2})\\ &= \|\boldsymbol{g}\|_2^2\,\frac{\log(k)}{k}\,\frac{2+\delta}{1-k^{-\delta}}.\end{aligned} \tag{10}$$

Setting $\delta = 1$, results in the bound

$$\mathbb{E}\big[\|\boldsymbol{v}\|_\infty^2\big] \le 3\|\boldsymbol{g}\|_2^2\,\frac{\log(k)}{k-1}. \tag{11}$$

Summarizing the above results for $k \ge 2$, we have

$$\mathbb{E}\big[\|\widehat{\boldsymbol{g}}-\boldsymbol{g}\|_2^2\big] \le \left(\frac{n}{k} - 1 + \frac{n}{4Q^2}\,\frac{\log(k)}{k-1}\right)\|\boldsymbol{g}\|_2^2. \tag{12}$$

For $k = 1$, note that since v is a scalar, by the definition of the used dithered quantizer sending magnitude of v and its sign results in $\hat{v} = v$ and $e_d = 0$. Therefore,

$$\mathbb{E}\big[\|\widehat{\boldsymbol{g}}-\boldsymbol{g}\|_2^2\big] = (n-1)\|\boldsymbol{g}\|_2^2. \tag{13}$$

Lower Bound on the Compression Gain of Sparsification

Recall that the variance of the unbiased sparsification is given by ((5) of (Wangni et al. 2018))

$$\mathbb{E}\big[\|\boldsymbol{g}-\widehat{\boldsymbol{g}}\|_2^2\big] = \sum_{i=1}^n \frac{g_i^2}{p_i} - \|\boldsymbol{g}\|_2^2, \tag{14}$$

where $0 \le p_i \le 1$ and $\sum_i p_i = s$ controls the average transmission rate, i.e., the average compression gain is $\frac{n}{s}$. Finding the optimum value for p_i's for general $\boldsymbol{g}$ is not straightforward and hard to analyze. Hence, we assume that $\boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ and compute the expected value of the sparsification variance over $\boldsymbol{g}$.

Using the inequality $\frac{a^2}{x}+\frac{b^2}{y}\geq\frac{(a+b)^2}{x+y}$ for arbitrary positive numbers (Titu's Lemma), the variance of error can be lower bounded as

$$\mathbb{E}\big[\|\boldsymbol{g}-\widehat{\boldsymbol{g}}\|_2^2\big]\geq\frac{(\sum_i|g_i|)^2}{\sum_i p_i}-\|\boldsymbol{g}\|_2^2=\frac{\|\boldsymbol{g}\|_1^2}{s}-\|\boldsymbol{g}\|_2^2.$$

Therefore,

$$\gamma=\frac{\mathbb{E}\big[\|\boldsymbol{g}-\widehat{\boldsymbol{g}}\|_2^2\big]}{\|\boldsymbol{g}\|_2^2}\geq\frac{1}{s}\frac{\|\boldsymbol{g}\|_1^2}{\|\boldsymbol{g}\|_2^2}-1.$$

On the other hand, for Gaussian random vector, $\mathbb{E}\big[\|\boldsymbol{g}\|_1^2\big]=\sum_{i,j}\mathbb{E}[|g_i||g_j|]=n(1+(n-1)\frac{2}{\pi})$, $\mathbb{E}\big[\|\boldsymbol{g}\|_2^2\big]=n$, and $\mathbb{E}\Big[\frac{\|\boldsymbol{g}\|_1^2}{\|\boldsymbol{g}\|_2^2}\Big]$ is close to that ratio. Especially, for large n, $\|\boldsymbol{g}\|_1^2$ and $\|\boldsymbol{g}\|_2^2$ are concentrated around their mean. Therefore,

$$\bar{\gamma}=\mathbb{E}[\gamma]\gtrapprox\frac{1}{s}\big(\frac{2}{\pi}n+(1-\frac{2}{\pi})\big)-1\geq\frac{2}{\pi}\frac{n}{s}-1.$$

4 MMSE QCS

Proof of Lemma 3

For $\widehat{\boldsymbol{g}}=\alpha\boldsymbol{T}^\mathsf{T}\widehat{\boldsymbol{v}}$, using the same argument as for the unbiased QCS, it can be easily shown that

$$\begin{aligned}
&\mathbb{E}\big[\|\boldsymbol{g}-\widehat{\boldsymbol{g}}\|_2^2\big]=\mathbb{E}\big[\|\boldsymbol{g}-\alpha\boldsymbol{T}^\mathsf{T}\boldsymbol{T}\boldsymbol{g}+\alpha\varsigma\boldsymbol{T}^\mathsf{T}\boldsymbol{\varepsilon}\|_2^2\big]\\
&=\|\boldsymbol{g}\|_2^2(1-2\alpha+\alpha^2\frac{n}{k})+\alpha^2\frac{n}{12Q^2}\,\mathbb{E}\big[\|\boldsymbol{T}\boldsymbol{g}\|_\infty^2\big]\\
&\leq\|\boldsymbol{g}\|_2^2\big(1-2\alpha+\alpha^2\frac{n}{k}+\alpha^2\frac{n}{4Q^2}\,\frac{\log(k)}{k-1}\big)\\
&=\|\boldsymbol{g}\|_2^2\big(1-2\alpha+\alpha^2(\gamma+1)\big),
\end{aligned}$$

where $\gamma=\frac{n}{k}+\frac{n}{4Q^2}\,\frac{\log(k)}{k-1}$ for $k\geq 2$, and $\gamma=n-1$ for $k=1$.

Minimizing the upper bound of the error results in

$$\alpha=\frac{1}{1+\gamma},$$

and by substituting α, the minimum mean squared error is given by

$$\mathbb{E}\big[\|\boldsymbol{g}-\widehat{\boldsymbol{g}}\|_2^2\big]\leq\|\boldsymbol{g}\|_2^2\frac{\gamma}{1+\gamma}$$

5 Weighted Error Feedback

In weighted error feedback, instead of directly compressing the SGs, we add a weighted residue of compression from previous step to the SG and then compress it.

$$\boldsymbol{z}_t\leftarrow\boldsymbol{g}_t+\beta\boldsymbol{r}_t \tag{15a}$$
$$\widehat{\boldsymbol{z}}_t\leftarrow\textsc{Compress}(\boldsymbol{z}_t) \tag{15b}$$
$$\boldsymbol{e}_t\leftarrow\boldsymbol{z}_t-\widehat{\boldsymbol{z}}_t \tag{15c}$$
$$\boldsymbol{r}_{t+1}\leftarrow(1-\beta)\boldsymbol{r}_t+\boldsymbol{e}_t \tag{15d}$$

Proof of Lemma 4

Recall that $\boldsymbol{e}_t=\boldsymbol{z}_t-\widehat{\boldsymbol{z}}_t=(\boldsymbol{I}-\alpha\boldsymbol{T}_t^\mathsf{T}\boldsymbol{T}_t)\boldsymbol{z}_t-\alpha\varsigma\boldsymbol{T}_t^\mathsf{T}\boldsymbol{\varepsilon}$, where $\alpha=1$ for Unbiased-QCS and $\alpha=1/(1+\gamma)$ for MMSE-QCS, and $\boldsymbol{\varepsilon}$ is the scaled quantization noise of $\boldsymbol{T}\boldsymbol{z}_t$ (see, e.g., (8)). Therefore,

$$\begin{aligned}
\mathbb{E}\big[\|\boldsymbol{r}_{t+1}\|_2^2\big]&=(1-\beta)^2\,\mathbb{E}\big[\|\boldsymbol{r}_t\|_2^2\big]+\mathbb{E}\big[\|\boldsymbol{e}_t\|_2^2\big]\\
&\quad+2(1-\beta)\,\mathbb{E}\big[\boldsymbol{r}_t^\mathsf{T}\boldsymbol{e}_t\big].
\end{aligned}$$

On the other hand,

$$\begin{aligned}
\mathbb{E}\big[\boldsymbol{r}_t^\mathsf{T}\boldsymbol{e}_t\big]&=\mathbb{E}\big[\boldsymbol{r}_t^\mathsf{T}(\boldsymbol{I}-\alpha\boldsymbol{T}_t^\mathsf{T}\boldsymbol{T}_t)\boldsymbol{z}_t\big]-\alpha\,\mathbb{E}\big[\boldsymbol{r}_t^\mathsf{T}\varsigma\boldsymbol{T}_t^\mathsf{T}\boldsymbol{\varepsilon}\big]\\
&=\mathbb{E}\big[\mathbb{E}_{\boldsymbol{T}}\big[\boldsymbol{r}_t^\mathsf{T}(\boldsymbol{I}-\alpha\boldsymbol{T}_t^\mathsf{T}\boldsymbol{T}_t)\boldsymbol{z}_t|\boldsymbol{r}_t,\cdots\big]\big]\\
&\quad-\alpha\,\mathbb{E}\big[\mathbb{E}_{\boldsymbol{\varepsilon}}\big[\boldsymbol{r}_t^\mathsf{T}\varsigma\boldsymbol{T}_t^\mathsf{T}\boldsymbol{\varepsilon}|\boldsymbol{T}_t,\boldsymbol{r}_t,\cdots\big]\big]\\
&\overset{(a)}{=}\mathbb{E}\big[\mathbb{E}_{\boldsymbol{T}}\big[\boldsymbol{r}_t^\mathsf{T}(\boldsymbol{I}-\alpha\boldsymbol{T}_t^\mathsf{T}\boldsymbol{T}_t)\boldsymbol{z}_t|\boldsymbol{r}_t,\cdots\big]\big]\\
&\overset{(b)}{=}(1-\alpha)\,\mathbb{E}\big[\boldsymbol{r}_t^\mathsf{T}\boldsymbol{z}_t\big],
\end{aligned}$$

where (a) is because of $\boldsymbol{\varepsilon}$ being an independent zero-mean random vector and (b) due to $\mathbb{E}_{\boldsymbol{T}}\big[\boldsymbol{T}^\mathsf{T}\boldsymbol{T}\big]=\boldsymbol{I}$.

Therefore,

$$\begin{aligned}
\mathbb{E}\big[\|\boldsymbol{r}_{t+1}\|_2^2\big]&=(1-\beta)^2\,\mathbb{E}\big[\|\boldsymbol{r}_t\|_2^2\big]+\mathbb{E}\big[\|\boldsymbol{e}_t\|_2^2\big]\\
&\quad+2(1-\beta)(1-\alpha)\,\mathbb{E}\big[\boldsymbol{r}_t^\mathsf{T}\boldsymbol{z}_t\big].
\end{aligned}$$

First, we consider the Unbiased-QCS.

Lemma IV. *In Unbiased-QCS with error-feedback, residual signal and the stochastic gradients are uncorrelated, i.e.,* $\forall t,\tau$: $\mathbb{E}\big[\boldsymbol{g}_t^\mathsf{T}\boldsymbol{r}_\tau\big]=0$.

Proof. The proof is based on induction. For $\tau=0$, since $\boldsymbol{r}_\tau=\boldsymbol{0}$, the claim holds. Assume that the claim is true for $\tau-1$.

$$\begin{aligned}
\mathbb{E}\big[\boldsymbol{g}_t^\mathsf{T}\boldsymbol{r}_\tau\big]&=\mathbb{E}\big[\boldsymbol{g}_t^\mathsf{T}((1-\beta)\boldsymbol{r}_{\tau-1}+\boldsymbol{e}_\tau)\big]\\
&=(1-\beta)\,\mathbb{E}\big[\boldsymbol{g}_t^\mathsf{T}\boldsymbol{r}_{\tau-1}\big]+\mathbb{E}\big[\boldsymbol{g}_t^\mathsf{T}\boldsymbol{e}_\tau\big]\\
&=\mathbb{E}\big[\boldsymbol{g}_t^\mathsf{T}((\boldsymbol{I}-\boldsymbol{T}_\tau^\mathsf{T}\boldsymbol{T}_\tau)\boldsymbol{z}_\tau+\varsigma\boldsymbol{T}_\tau^\mathsf{T}\boldsymbol{\varepsilon}_\tau)\big]\\
&=\mathbb{E}\big[\mathbb{E}_{\boldsymbol{T}_\tau}\big[\boldsymbol{g}_t^\mathsf{T}(\boldsymbol{I}-\boldsymbol{T}_\tau^\mathsf{T}\boldsymbol{T}_\tau)\boldsymbol{z}_\tau\big|\boldsymbol{g}_t,\boldsymbol{z}_\tau\big]\big]+\\
&\quad\mathbb{E}\big[\mathbb{E}_{\boldsymbol{\varepsilon}_\tau}\big[\varsigma\boldsymbol{g}_t^\mathsf{T}\boldsymbol{T}_\tau^\mathsf{T}\boldsymbol{\varepsilon}_\tau\big|\boldsymbol{g}_t,\boldsymbol{T}_\tau\big]\big]=0.
\end{aligned}$$

This completes the proof. ∎

Since in unbiased QCS, $\alpha=1$, we have

$$\begin{aligned}
\mathbb{E}\big[\|\boldsymbol{r}_{t+1}\|_2^2\big]&=(1-\beta)^2\,\mathbb{E}\big[\|\boldsymbol{r}_t\|_2^2\big]+\mathbb{E}\big[\|\boldsymbol{e}_t\|_2^2\big]\\
&\overset{(c)}{\leq}(1-\beta)^2\,\mathbb{E}\big[\|\boldsymbol{r}_t\|_2^2\big]+\gamma\,\mathbb{E}\big[\|\boldsymbol{z}_t\|_2^2\big]\\
&=(1-\beta)^2\,\mathbb{E}\big[\|\boldsymbol{r}_t\|_2^2\big]+\gamma\big(\,\mathbb{E}\big[\|\boldsymbol{g}_t+\beta\boldsymbol{r}_t\|_2^2\big]\big)\\
&\overset{(d)}{=}\big((1-\beta)^2+\beta^2\gamma\big)\,\mathbb{E}\big[\|\boldsymbol{r}_t\|_2^2\big]+\gamma\,\mathbb{E}\big[\|\boldsymbol{g}_t\|_2^2\big]\\
&\overset{(e)}{\leq}\big((1-\beta)^2+\beta^2\gamma\big)\,\mathbb{E}\big[\|\boldsymbol{r}_t\|_2^2\big]+\gamma B \qquad (16)
\end{aligned}$$

where (c) is due to the fact that for all $\boldsymbol{z}$, $\mathbb{E}\big[\|\boldsymbol{z}-\widehat{\boldsymbol{z}}\|_2^2\big]\leq\gamma\|\boldsymbol{z}\|_2^2$, (d) is from Lemma IV and (e) is from boundedness of $\boldsymbol{g}$. The recursive equation (16) with $\boldsymbol{r}_0=\boldsymbol{0}$ implies that

$$\mathbb{E}\big[\|\boldsymbol{r}_t\|_2^2\big]\leq\frac{\gamma}{1-\big((1-\beta)^2+\beta^2\gamma\big)}B, \qquad (17)$$

for all β that $\beta < 1$ and $(1-\beta)^2 + \beta^2\gamma < 1$, hence, $\beta < \min(1, 2/(1+\gamma))$.

For MMSE-QCS, the stochastic gradients and residual signal might be correlated. However, for an arbitrary $c > 0$, their correlation can be bounded as

$$\begin{aligned} 0 \le \mathbb{E}\Big[\|\frac{1}{\sqrt{c}}\boldsymbol{g}_t \pm \sqrt{c}\boldsymbol{r}_\tau\|_2^2\Big] &= \frac{1}{c}\,\mathbb{E}\big[\|\boldsymbol{g}_t\|_2^2\big] + c\mathbb{E}\big[\|\boldsymbol{r}_\tau\|_2^2\big] \\ &\quad \pm 2\,\mathbb{E}\big[\boldsymbol{g}_t^{\mathsf{T}}\boldsymbol{r}_\tau\big] \\ \Rightarrow 2\big|\,\mathbb{E}\big[\boldsymbol{g}_t^{\mathsf{T}}\boldsymbol{r}_\tau\big]\big| &\le \frac{1}{c}\,\mathbb{E}\big[\|\boldsymbol{g}_t\|_2^2\big] + c\mathbb{E}\big[\|\boldsymbol{r}_\tau\|_2^2\big]. \end{aligned}$$

Therefore, noting that $\mathbb{E}\big[\|\boldsymbol{e}\|_2^2\big] \le (1-\alpha)\|\boldsymbol{z}\|_2^2$ and $\mathbb{E}\big[\|\boldsymbol{g}_t\|_2^2\big] \le B$,

$$\begin{aligned} \mathbb{E}\big[\|\boldsymbol{r}_{t+1}\|_2^2\big] &= (1-\beta)^2\,\mathbb{E}\big[\|\boldsymbol{r}_t\|_2^2\big] + \mathbb{E}\big[\|\boldsymbol{e}_t\|_2^2\big] \\ &\quad + 2(1-\beta)(1-\alpha)\,\mathbb{E}\big[\boldsymbol{r}_t^{\mathsf{T}}\boldsymbol{z}_t\big] \\ &\le (1-\beta)^2\,\mathbb{E}\big[\|\boldsymbol{r}_t\|_2^2\big] + (1-\alpha)\,\mathbb{E}\big[\|\boldsymbol{g}_t + \beta\boldsymbol{r}_t\|_2^2\big] + \\ &\quad 2(1-\beta)(1-\alpha)\,\mathbb{E}\big[\boldsymbol{r}_t^{\mathsf{T}}(\boldsymbol{g}_t + \beta\boldsymbol{r}_t)\big] \\ &\le \big(1-\alpha\beta(2-\beta)\big)\,\mathbb{E}\big[\|\boldsymbol{r}_t\|_2^2\big] + \\ &\quad (1-\alpha)(c\,\mathbb{E}\big[\|\boldsymbol{r}_t\|_2^2\big] + \frac{1}{c}B) + (1-\alpha)B \\ &= \big(1-\alpha\beta(2-\beta) + c(1-\alpha)\big)\,\mathbb{E}\big[\|\boldsymbol{r}_t\|_2^2\big] + \\ &\quad (1-\alpha)(1+\frac{1}{c})B. \end{aligned}$$

Therefore, if $\big|1-\alpha\beta(2-\beta)+c(1-\alpha)\big| < 1$,

$$\mathbb{E}\big[\|\boldsymbol{r}_t\|_2^2\big] \le \frac{1+1/c}{\alpha\beta(2-\beta) - c(1-\alpha)}(1-\alpha)B. \tag{18}$$

Minimizing w.r.t. c and substituting $\alpha = 1/(1+\gamma)$ results in

$$\begin{aligned} \mathbb{E}\big[\|\boldsymbol{r}_t\|_2^2\big] &\le \frac{1-\alpha}{\big(\sqrt{1-\alpha(1-\beta)^2} - \sqrt{1-\alpha}\big)^2}B \qquad (19) \\ &= \frac{\gamma}{\big(\sqrt{\gamma+1-(1-\beta)^2} - \sqrt{\gamma}\big)^2}B. \qquad (20) \end{aligned}$$

Note that if the SGs have bounded variance, i.e., $\mathbb{E}\big[\|\boldsymbol{g} - \nabla f\|_2^2\big] \le \sigma^2$, a similar approach can be used to bound $\mathbb{E}\big[\|\boldsymbol{r}_t\|_2^2\big]$ based on the σ^2 and the weighted average of $\|\nabla f(\boldsymbol{w}_{t-i})\|_2^2$ for $i = 0, \ldots, t$. This is specially helpful when analyzing the convergence of the training algorithm with error feedback under the assumption of bounded variance SG.

6 Convergence Analysis

Proof of Lemm 5

The proof follows the same line of argument as for ordinary SGD which is repeated here for the sake of completeness.

Recall that for Lipschitz-smooth function $f(\cdot)$, for arbitrary $\boldsymbol{w}$ and $\boldsymbol{\delta}$,

$$f(\boldsymbol{w}+\boldsymbol{\delta}) \le f(\boldsymbol{w}) + \boldsymbol{\delta}^{\mathsf{T}}\nabla f(\boldsymbol{w}) + \frac{L}{2}\|\boldsymbol{\delta}\|_2^2.$$

First, we consider Unbiased-QCS. Tt the t-th iteration of training, $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \mu\widehat{\boldsymbol{g}}_t$, where $\mathbb{E}[\widehat{\boldsymbol{g}}_t] = \nabla f(\boldsymbol{w}_t)$ and $\mathbb{E}\big[\|\widehat{\boldsymbol{g}}_t\|_2^2\big] \le (1+\gamma)\,\mathbb{E}\big[\|\boldsymbol{g}\|_2^2\big] \le (1+\gamma)B$. Hence,

$$\begin{aligned} \mathbb{E}[f(\boldsymbol{w}_{t+1})] &\le f(\boldsymbol{w}_t) + \langle\nabla f(\boldsymbol{w}_t), \mathbb{E}[\boldsymbol{w}_{t+1} - \boldsymbol{w}_t]\rangle \\ &\quad + \frac{L}{2}\,\mathbb{E}\big[\|\boldsymbol{w}_{t+1} - \boldsymbol{w}_t\|_2^2\big] \\ &\le f(\boldsymbol{w}_t) - \mu\|\nabla f(\boldsymbol{w}_t)\|_2^2 + \frac{L}{2}\mu^2(1+\gamma)B. \end{aligned}$$

Rearranging terms, taking expectation and summing from $t = 0$ to $T-1$, results in

$$\begin{aligned} \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\big[\|\nabla f(\boldsymbol{w}_t)\|_2^2\big] &\le \frac{f(\boldsymbol{w}_0) - \mathbb{E}[f(\boldsymbol{w}_T)]}{T\mu} + \frac{L}{2}\mu(1+\gamma)B \\ &\le \frac{f(\boldsymbol{w}_0) - f^*}{T\mu} + \frac{L}{2}\mu(1+\gamma)B. \end{aligned}$$

Setting $\mu = 1/\sqrt{T}$, results in

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\big[\|\nabla f(\boldsymbol{w}_t)\|_2^2\big] \le \frac{f(\boldsymbol{w}_0) - f^* + \frac{L}{2}(1+\gamma)B}{\sqrt{T}}.$$

Note that in the case that the stochastic gradients have bounded variance[3], i.e., $\mathbb{E}\big[\|\boldsymbol{g} - \nabla f\|_2^2\big] \le \sigma^2$, $\mathbb{E}\big[\|\widehat{\boldsymbol{g}}_t\|_2^2\big] \le (1+\gamma)\,\mathbb{E}\big[\|\boldsymbol{g}\|_2^2\big] \le (1+\gamma)(\sigma^2 + \|\nabla f\|_2^2)$ and we can modify the above argument as follows to bound the convergence rate,

$$\begin{aligned} \mathbb{E}[f(\boldsymbol{w}_{t+1})] - f(\boldsymbol{w}_t) &\le + \langle\nabla f(\boldsymbol{w}_t), \mathbb{E}[\boldsymbol{w}_{t+1} - \boldsymbol{w}_t]\rangle \\ &\quad + \frac{L}{2}\,\mathbb{E}\big[\|\boldsymbol{w}_{t+1} - \boldsymbol{w}_t\|_2^2\big] \\ &\le -\mu\|\nabla f(\boldsymbol{w}_t)\|_2^2 + \frac{L}{2}\mu^2(1+\gamma)(\sigma^2 + \|\nabla f(\boldsymbol{w}_t)\|_2^2) \\ &= -(\mu - \frac{L}{2}\mu^2(1+\gamma))\|\nabla f(\boldsymbol{w}_t)\|_2^2 + \frac{L}{2}\mu^2(1+\gamma)\sigma^2. \end{aligned}$$

Following same argument as before,

$$\begin{aligned} \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\big[\|\nabla f(\boldsymbol{w}_t)\|_2^2\big] &\le \frac{2(f(\boldsymbol{w}_0) - f^*)}{T(2\mu - L\mu^2(1+\gamma))} \\ &\quad + \frac{L\mu^2(1+\gamma)}{2\mu - L\mu^2(\gamma+1)}\sigma^2. \end{aligned}$$

It can be verified that if $T > 4L^2(\gamma+1)^2$, we can find $\mu \le 2/\sqrt{T}$ such that $2\mu - L\mu^2(\gamma+1) = 2/\sqrt{T}$. This simplifies the above equation to

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\big[\|\nabla f(\boldsymbol{w}_t)\|_2^2\big] \le \frac{f(\boldsymbol{w}_0) - f^*}{\sqrt{T}} + \frac{2L(1+\gamma)}{\sqrt{T}}\sigma^2.$$

The analysis for MMSE-QCS is straightforward. Note that $\widehat{\boldsymbol{g}}_{mmse} = \frac{1}{\gamma+1}\widehat{\boldsymbol{g}}_u$, where $\widehat{\boldsymbol{g}}_{mmse}$ is the MMSE-QCS quantized SG and $\widehat{\boldsymbol{g}}_u$ is the output of Unbiased-QCS. Therefore, training with MMSE-QCS and step-size μ would be the same as using Unbiased-QCS with step-size $\mu/(\gamma+1)$.

[3] In this case, it is not necessary to assume that the cost function has bounded gradient everywhere.

Remark 3. Note that since Unbiased-QCS has bounded variance and is unbiased, the compressed SG will be stochastic gradient itself with bounded variance. Hence, majority of the results can be readily applied to prove the convergence of Unbiased-QCS and MMSE-QCS under different conditions such as (Bottou 1998; Bubeck 2015).

Proof of Lemma 6

The proof is based on the ideas from (Karimireddy et al. 2019) and follows the similar arguments with slight modifications, which is repeated here for the sake of completeness.

Let $\widetilde{\boldsymbol{w}}_t = \boldsymbol{w}_t - \mu \boldsymbol{r}_t$. Note that since by Lemma **??** the residue signal has bounded variance, $\widetilde{\boldsymbol{w}}_t$ would be bounded. It can be easily verified that $\widetilde{\boldsymbol{w}}_{t+1} = \widetilde{\boldsymbol{w}}_t - \mu \boldsymbol{g}_t$. Hence, following similar argument as in (Karimireddy et al. 2019), for arbitrary $\rho > 0$

$$\begin{aligned}
&\mathbb{E}[f(\widetilde{\boldsymbol{w}}_{t+1})] - f(\widetilde{\boldsymbol{w}}_t) \\
&\leq \frac{L}{2}\,\mathbb{E}\big[\|\widetilde{\boldsymbol{w}}_{t+1} - \widetilde{\boldsymbol{w}}_t\|_2^2\big] + \langle \nabla f(\widetilde{\boldsymbol{w}}_t), \mathbb{E}[\widetilde{\boldsymbol{w}}_{t+1} - \widetilde{\boldsymbol{w}}_t]\rangle \\
&\leq \frac{L}{2}\mu^2 B - \mu(1-\rho)\|\nabla f(\boldsymbol{w}_t)\|_2^2 + \frac{1}{\rho}L^2\mu^3\,\mathbb{E}\big[\|\boldsymbol{r}_t\|_2^2\big].
\end{aligned}$$

On the other hand, from Lemma **??**, the residue is bounded as $\mathbb{E}\big[\|\boldsymbol{r}_t\|_2^2\big] \leq \eta B$ where η is a constant depending on the β (weight of error feedback) and γ, according to (**??**) or (**??**) of the main paper for Unbiased-QCS and MMSE-QCS. Therefore,

$$\begin{aligned}
\mu(1-\rho)\|\nabla f(\boldsymbol{w}_t)\|_2^2 \leq &\big(\frac{L}{2}\mu^2 + \frac{1}{\rho}L^2\mu^3\eta\big)B \\
&+ f(\widetilde{\boldsymbol{w}}_t) - \mathbb{E}[f(\widetilde{\boldsymbol{w}}_{t+1})].
\end{aligned}$$

Taking expectation, rearranging terms and noting that $\widetilde{\boldsymbol{w}}_0 = \boldsymbol{w}_0$ and $\mathbb{E}[f(\boldsymbol{w}_T)] \geq f^*$, we conclude that for $0 < \rho < 1$,

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\big[\|\nabla f(\boldsymbol{w}_t)\|_2^2\big] \leq \frac{f(\boldsymbol{w}_0) - f^*}{T\mu(1-\rho)} + \frac{LB}{1-\rho}\big(\frac{\mu}{2} + \frac{L\eta\mu^2}{\rho}\big).$$

Setting $\rho = 0.5$, gives the desired result.

For tighter analysis, let μ and ρ be such that $\mu(1-\rho) = 1/\sqrt{T}$ and $\frac{\mu}{1-\rho} = \frac{1+\epsilon}{\sqrt{T}}$ for arbitrary $\epsilon > 0$, i.e.,

$$\mu = \frac{\sqrt{1+\epsilon}}{\sqrt{T}}, \quad 1-\rho = \frac{1}{\sqrt{1+\epsilon}}.$$

Therefore,

$$\begin{aligned}
\min_t \mathbb{E}\big[\|\nabla f(\boldsymbol{w}_t)\|_2^2\big] &\leq \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\big[\|\nabla f(\boldsymbol{w}_t)\|_2^2\big] \\
&\leq \frac{f(\boldsymbol{w}_0) - f^* + \frac{L}{2}B}{\sqrt{T}} + L^2 B\frac{(1+\epsilon)^2}{\sqrt{1+\epsilon}-1}\frac{\eta}{T}.
\end{aligned}$$

References

Alistarh, D.; Grubic, D.; Li, J.; Tomioka, R.; and Vojnovic, M. 2017. QSGD: Communication-Efficient SGD via Gradient Quantization and Encoding. In *Advances in Neural Information Processing Systems*, 1707–1718.

Bottou, L. 1998. Online Learning and Stochastic Approximations, Revised 2018. *On-Line Learning in Neural Networks* 17(9):1–35.

Bourgain, J.; Vu, V. H.; and Wood, P. M. 2010. On the singularity probability of discrete random matrices. *Journal of Functional Analysis* 258(2):559–603.

Bubeck, S. 2015. Convex optimization: Algorithms and complexity. *Foundations and Trends®in Machine Learning* 8(3-4):231–357.

Gray, R. M., and Neuhoff, D. L. 1998. Quantization. *IEEE Transactions on Information Theory* 44(6):2325–2383.

Gray, R. M., and Stockham, T. G. 1993. Dithered quantizers. *IEEE Transactions on Information Theory* 39(3):805–812.

Karimireddy, S. P.; Rebjock, Q.; Stich, S. U.; and Jaggi, M. 2019. Error Feedback Fixes SignSGD and other Gradient Compression Schemes. *arXiv preprint arXiv:1901.09847*.

Schuchman, L. 1964. Dither signals and their effect on quantization noise. *IEEE Transactions on Communication Technology* 12(4):162–165.

Wangni, J.; Wang, J.; Liu, J.; and Zhang, T. 2018. Gradient sparsification for communication-efficient distributed optimization. In *Advances in Neural Information Processing Systems*, 1306–1316.

Wen, W.; Xu, C.; Yan, F.; Wu, C.; Wang, Y.; Chen, Y.; and Li, H. 2017. TernGrad: Ternary Gradients to Reduce Communication in Distributed Deep Learning. In *TernGrad: Ternary Gradients to Reduce Communication in Distributed Deep Learning*, 1509–1519. Curran Associates, Inc.