Quantization and Compression for Distributed Optimization

Master Semester Project

Cem Ates Musluoglu

January 11, 2019

Advisors: Prof. Martin Jaggi, Dr. Sebastian Stich

Machine Learning and Optimization Laboratory, EPFL

Abstract

In the context of distributed optimization, the large number of features for each sample of data implies a high dimensional gradient vector. Adding it to the numerical precision of each entry results in communication of the vectors being the bottleneck. For this purpose, various compression/quantization methods have been studied. This work studies subproblems of finding the best function for this purpose.

Introduction

We are interested in the problem of minimizing a sum of convex functions $f_1, \ldots, f_N$ which can be written as follows:

$$\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmin}} \frac{1}{N} \sum_{n=1}^{N} f_n(\mathbf{w}) = \underset{\mathbf{w}}{\operatorname{argmin}} f(\mathbf{w}) \tag{1.1}$$

with $f_n : \mathbb{R}^d \to \mathbb{R}$. Typically, we call the function f the cost, loss or objective function in the context of Machine Learning and N corresponds to the number of samples we have. The most common techniques used for minimizing the objective belong to the family of gradient descent algorithms, which we will review in the following section. To reduce the cost of computation per machine, the distributed variants of gradient descent have become a very popular choice.

However, the problem now becomes the communication of the gradient vectors between the machines and often turns into the bottleneck of the system. To reduce this cost and hence increase the parallelization of computation for optimization, various methods to compress/quantize the gradient vectors have been looked into. Some "aggressive" techniques include 1-bit SGD by Seide et al. [2014] and signSGD from Bernstein et al. [2018] which use very few values to represent the vectors and show that, under some settings, we can still achieve optimization. Other examples include the QSGD Alistarh et al. [2017] and TernGrad Wen et al. [2017] which rely on norms of the gradient vectors to quantize the gradient. Stich et al. [2018] define contraction operators and prove convergence for this general class of compression functions. Examples include randomly selecting entries from the original vector, or keeping only the largest components in absolute value.

We first look into previous work done in this area and some baselines, to introduce and further develop the context. Then, different subproblems are studied, namely, we compare different types of norms to construct functions similar to the one presented in Alistarh et al. [2017] to find whether we can achieve a better compression under some criteria. On the other hand, we also look at the approximation of vectors using what we call dictionaries by using projections or linear combinations, to analyse how good we can approximate a vector using a small number of dictionary vectors. For this we present a general framework consisting of orthogonal projections. We also discuss about overcomplete representations to have more flexibility. Overall, our aim is to search for better compression schemes to have less costly communication between the machines in the distributed optimization setting.

Baselines and Previous Works

This section provides useful background on the topic which will be helpful to state the problem and further develop our analysis. We will also look into the previous research in the area of quantization of gradient vectors.

2.1 The Optimization Problem

As stated above, we are interested in minimizing an objective function $f : \mathbb{R}^d \to \mathbb{R}$ which can be written as a sum of convex functions $f_i, i \in \{1, \ldots, N\}$. In the Machine Learning context, this represents a loss, usually the sum of errors for each data sample between our estimate and the real value. Regression is the process of determining the dependence between the input and the output and it corresponds to the way we estimate the output. In practice, we often use a linear model for its simplicity but it has also proven to be quite powerful and accurate, given that we do the necessary pre-processing beforehand. For example, in Machine Learning, we may add useful features to the original data to better adapt the classification or regression task. These features are often non-linear in the original data space but in the extended feature space, we still use a linear regression scheme.

The condition of the function being convex is important in the fact that it implies the existence of a global minimum. This is what gradient descent methods rely on to optimize the objective function. We now give an overview of some of these iterative methods.

Gradient Descent (GD): This is the most basic algorithm in this family; the iterates are given by:

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \gamma_t \nabla f(\mathbf{w}^{(t)}) \tag{2.1}$$

where the step-size can either be fixed or depending on the iteration which should be decided based on the specific problem. The convergence analysis of this method can be found in Bubeck et al. [2015] and we can achieve linear convergence rates $\mathcal{O}(c^t)$ in iterates, $0 < c < 1$, under the assumption that f is smooth and strongly convex, for which the definitions can be found in Appendix A. If the objective is L-smooth and μ-strongly convex, taking a constant step size $\gamma = \frac{2}{L+\mu}$ results in this constant c to be equal to $\frac{L-\mu}{L+\mu}$. If $\gamma = 1/L$, we achieve a slightly slower rate with $c = \sqrt{\frac{L-\mu}{L+\mu}}$.

As we can see, this method requires to compute the full gradient of f which is costly when the number of data samples N we have is large, which is usually the case for Machine Learning applications. For example, in the case of linear regression using mean squared error, the computation complexity of a single iteration of GD is $\mathcal{O}(Nd)$. In order to resolve this problem, another well-known algorithm exists and is presented next.

Stochastic Gradient Descent (SGD): The Stochastic Gradient Descent method uses the fact that the objective function can be written as a sum of functions that each depend on a single data sample. Hence, instead of computing the full gradient, at each iteration, we consider the gradient of one single f_i, the index being chosen uniformly at random.

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \gamma_t \nabla f_{i_t}(\mathbf{w}^{(t)}) \tag{2.2}$$

This scheme allows us to reduce the cost of computation by a factor of N per iteration compared to GD, which means that it is independent of the number of samples we have, this is a great improvement. Coming back to our previous example for GD, linear regression using mean squared error has a computation complexity in the order of $\mathcal{O}(d)$ by iteration. Another interesting property is that in expectation, the computed "stochastic gradients" will give the full gradient, since $i_t \sim \mathcal{U}\{1, \dots, N\}$ at each iteration t.

However, this reduction of computation comes at a cost. Indeed, we need stricter conditions for SGD to converge and the convergence is slower, compared to GD. For a strongly convex function we achieve a convergence rate of $\mathcal{O}(1/t)$, which is sublinear and slower compared to GD. An extended analysis and proofs can be found in Bubeck et al. [2015], Chapter 6 and the lectures of Cevher [2015], Lecture 7 and Jaggi [2018], Chapter 5. In the following Lemma, we give an upper bound on the expected value of the difference of the objective function evaluated at two consecutive iterates.

Lemma 2.1 (Cevher [2015]) *We consider an L-smooth convex function f which we want to minimize using SGD. The iterates then satisfy*

$$\begin{aligned} E[f(\mathbf{w}^{(t+1)}) - f(\mathbf{w}^{(t)})] \leq & (\gamma_t^2 L - \gamma_t) E[||\nabla f(\mathbf{w}^{(t)})||^2] \\ & + L\gamma_t^2 E[||\nabla f_{i_t}(\mathbf{w}^{(t)}) - \nabla f(\mathbf{w}^{(t)})||^2] \end{aligned}$$

Remark 2.2 *Throughout this work, we will use $||\cdot||$ to denote the ℓ_2 norm for simplicity.*

This lemma shows us that the variance of the stochastic gradients plays an important role in the convergence of SGD. Indeed, when we are close to a minimum, the norm of the full gradient should become nearly 0, hence the upper bound will be dominated by the value of $E[||\nabla f_{i_t}(\mathbf{w}^{(t)})||^2]$, this is why we should try to keep the variance as small as possible to ensure good convergence properties. The variance will be a key property in our study of compression of gradient vectors.

There have been several ways studied to reduce the variance of stochastic optimization methods. One of them is to use a decreasing step size per iteration which can be interpreted as a way of controlling the variance and can be seen from the result in 2.1, since, as stated above, the variance will be

dominant as we approach the minimum, hence a small step size will help reducing the variance. Another intuitive approach is called **Mini-Batch SGD** which consists of taking a set B of unique indices and computing the stochastic gradient using an average of the gradients of f_i's, $i \in B$. This technique is in between GD and SGD which are its two extremal cases. It allows us to have a trade-off between computation time and variance reduction. Indeed, this algorithm is $|B|$ times more costly computationally than SGD per iteration, however, having more samples in B reduces the variance. Mini-Batch SGD can be easily adapted to distributed settings where we use $|B|$ machines to compute the gradients. This implies the need of communicating these vectors between the machines or to a common server and is at the heart of the communication bottleneck in the distributed optimization setting. Other algorithms include SVRG (Johnson and Zhang [2013]), SAGA (Defazio et al. [2014]), SARAH (Nguyen et al. [2017]) which all use a constant step size and achieve a faster convergence rate (linear for strongly convex objectives) than SGD by reducing the variance of the stochastic gradients.

These examples show that there are several research on reducing the variance of the vectors and the importance of this property to achieve better convergence rates. This is why we will also be interested in having a small variance for our compression schemes. The next section introduces the context and some previous work on quantification/compression of gradient vectors.

2.2 Compression

To understand the need of compression and how communication can become a bottleneck, let us take the example $|B|$ machines having access to only one data sample at each iteration and each machine communicates its gradient vector to a centralized server computing the average of the gradients. Overall, this scheme corresponds to applying mini-batch SGD using a batch size $|B|$. We assume that each entry of the vectors can be described by 64 bits. Then, per iteration of the algorithm, we have $|B|$ vectors, each of d dimension, for which each of them needs 64 bits. Hence the communication cost of this scheme is $64|B|d$ bits per iteration. For example, assuming that we have a dataset of 100×100 images and we want to use linear and quadratic features, we need to communicate nearly 400 MB per gradient and there are $|B|$ of them per iteration. This quickly becomes very large considering the size of the datasets in modern applications. A more detailed analysis on the communication complexity in the distributed setting can be found in Arjevani and Shamir [2015].

To simplify the notation, we will call the vectors to compress $\mathbf{x} \in \mathbb{R}^d$ which represents the gradient vectors. The compression is represented by a func-

tion $Q : \mathbb{R}^d \to \mathbb{R}^d$ and $Q(\mathbf{x})$ represents the compressed vector. This function can either be deterministic or containing randomness, which both have their advantages that we will see in the following parts. Next, we define a useful property for Q, introduced in Stich et al. [2018].

Definition 2.3 *For $k \in]0, d]$, a function Q satisfies the k-contraction property if*

$$E[||\mathbf{x} - Q(\mathbf{x})||^2] \leq \left(1 - \frac{k}{d}\right) ||\mathbf{x}||^2$$

This property represents in a sense the fact that the compressed vector can be coded more efficiently than the original one, but as pointed out in Stich et al. [2018], it does not necessarily imply sparsity. It also links the compressed vector to the variance of the stochastic gradients by giving an upper bound. An important result is that using a compression function under this constraint implies that the compressed SGD algorithm using memory converges.

Theorem 2.4 (Stich et al. [2018]) *For a compression operator Q satisfying the k-contraction property, and assuming a μ-strongly convex and L-smooth objective with bounded gradients, the variant of SGD using the following updates converges.*

$$\mathbf{w}^{(\mathbf{t+1})} = \mathbf{w}^{(\mathbf{t})} - Q(\mathbf{m}^{(t)} + \gamma_t \nabla f_{i_t}(\mathbf{w}^{(\mathbf{t})})) = \mathbf{w}^{(\mathbf{t})} - \mathbf{g}^{(t)}$$

Where $\mathbf{m}^{(0)} = \mathbf{0}$, and for $t \geq 1$, $\mathbf{m}^{(t+1)} = \mathbf{m}^{(t)} + \gamma_t \nabla f_{i_t}(\mathbf{w}^{(\mathbf{t})}) - \mathbf{g}^{(t)}$. Moreover, the expected iterates and objective values satisfy

$$\begin{aligned}\gamma_t(E[f(\mathbf{w}^{(t)})] - f(\mathbf{w}^*)) \leq &\left(1 - \frac{\mu\gamma_t}{2}\right) E[||\mathbf{w}^{(t)} + \mathbf{m}^{(t)} - \mathbf{w}^*||^2] \\ &- E[||\mathbf{w}^{(t+1)} + \mathbf{m}^{(t+1)} - \mathbf{w}^*||^2] \\ &- \gamma_t^2 E[||\nabla f_{i_t}(\mathbf{w}^{(t)})||^2] + \gamma_t(\mu + 2L) E[||\mathbf{m}^{(t)}||^2]\end{aligned}$$

where $\mathbf{w}^*$ denotes the optimal value.

More detailed analysis and the proof of this theorem are presented in Stich et al. [2018]. However, an important note is that the last term $E[||\mathbf{m}^{(t)}||^2]$ can be bounded when we have a compression operator satisfying the contraction property, hence the convergence. With this result, we have a guarantee of convergence for contraction operators using SGD with memory. It can also be shown that in the general case, using classical SGD will generally result in slower convergence for operators satisfying the contraction property.

We now present some of the previous works in this domain. An idea of compression is to select quantization levels and to project each entry to one of the nearest level with a certain probability, the latter being greater the more the value is close to the given level. The function proposed by Alistarh

et al. [2017] uses this idea. Considering we have s levels, each component x_i of the vector is quantified by the following function:

$$\xi_i(\mathbf{x}, s) = \begin{cases} l_i/s & \text{with probability } 1 - p_i(\mathbf{x}, s) \\ (l_i + 1)/s & \text{otherwise} \end{cases} \tag{2.3}$$

where $l_i : |x_i|/||\mathbf{x}|| \in [l_i/s, (l_i + 1)/s]$ and $p_i(\mathbf{x}, s) = s|x_i|/||\mathbf{x}|| - l_i$. The final expression for the compression function is:

$$[Q_s(\mathbf{x})]_i = ||\mathbf{x}|| \cdot \text{sign}(x_i) \cdot \xi_i(\mathbf{x}, s) \tag{2.4}$$

Lemma 2.5 (Alistarh et al. [2017]) *The compression function Q_s as defined above is unbiased; $E[Q_s(\mathbf{x})] = \mathbf{x}$. Moreover, the variance satisfies $E[||\mathbf{x} - Q_s(\mathbf{x})||^2] \leq min(d/s^2, \sqrt{d}/s)||\mathbf{x}||^2$.*

A proof for the variance bound in a more general case will be shown in section 3.1.2. The result for the variance of this lemma shows that unless we choose a large enough number of levels, we cannot achieve a low variance. Moreover, we see here a first example of the trade-off between bias and variance reduction/contraction property. Indeed, both cannot be obtained easily for the general case and one of them should be dropped. An issue about the convergence of the fully quantized SVRG in the distributed setting has been pointed out and a solution been proposed by Künstner [2017]. In Zhang et al. [2016], a similar function, but this time with non-uniform quantization intervals has been studied. The paper presents an algorithm which outputs quantization levels which minimize the variance.

A very similar function using the ℓ_∞ norm instead of the ℓ_2 norm is presented in Wen et al. [2017] but with only 3 levels:

$$[Q(\mathbf{x})]_i = ||\mathbf{x}||_\infty \cdot \text{sign}(x_i) \cdot \xi_i(\mathbf{x}) \tag{2.5}$$

with $\xi_i(\mathbf{x}) = 1$ with probability $|x_i|/||\mathbf{x}||_\infty$ and 0 otherwise. These functions above have a convergence guarantee under the assumption of bounded gradients. A generalized framework for this model of functions is presented in Wang et al. [2018]. In section 3.1.2, we will study the variance bounds for this type of functions Q_s^p for any ℓ_p norm.

More aggressive schemes include the 1-Bit SGD (Seide et al. [2014]) and signSGD (Bernstein et al. [2018]). In the latter paper, the proposed compression function only keeps the sign of the gradient's entries. This method gives up on the unbiasedness to achieve better variance reduction and hence avoids the randomisation of the compression operator. Their experiments show that signSGD achieves a convergence rate similar to the one of SGD.

On the other hand, Stich et al. [2018] proposed a class of functions called k-contraction operators, which satisfy the contraction property with parameter k. They proved that in the setting of SGD with memory, convergence is guaranteed using these compression schemes. In particular, the paper proposes rand_k and top_k, the former choosing k entries from $\mathbf{x}$ randomly and the latter keeps only the k components of the vector having the largest absolute value. In section 3, we propose a framework which will allow to express these functions as orthogonal projections.

As a final remark before ending this section, we show that a randomized compression operator Q can either be made unbiased or to satisfy the contraction property by simply multiplying it by a constant, but satisfying both properties in general requires too much constraints.

Lemma 2.6 *We consider a compression operator Q satisfying $E[Q(\mathbf{x})] = \alpha\mathbf{x}$ and $E[||Q(\mathbf{x})||^2] \leq b||\mathbf{x}||^2$, α and b chosen such that it satisfies the contraction property with parameter k, i.e. $\exists k,\ 0 < k \leq d : 2\alpha - b > k/d$. Then, the unbiased operator $\tilde{Q} : \tilde{Q}(\mathbf{x}) = Q(\mathbf{x})/\alpha$ may not satisfy the contraction property for any $k \in]0, d]$. The converse is also true, namely, an unbiased compression operator $\tilde{Q}$ can be made to satisfy the contraction property by multiplying it with a constant: $Q : Q(\mathbf{x}) = \alpha\tilde{Q}(\mathbf{x})$. Then, Q is biased.*

Proof The variance bound can be expressed as:

$$\begin{aligned} E[||\mathbf{x} - Q(\mathbf{x})||^2] &= ||\mathbf{x}||^2 + E[||Q(\mathbf{x})||^2] - 2\alpha||\mathbf{x}||^2 \\ &\leq ||\mathbf{x}||^2(1 + b - 2\alpha) \\ &\leq \left(1 - \frac{k}{d}\right)||\mathbf{x}||^2 \end{aligned}$$

Then, for $\tilde{Q}$, we may write:

$$E[||\mathbf{x} - \tilde{Q}(\mathbf{x})||^2] \leq ||\mathbf{x}||^2\left(1 - \left(2 - \frac{b}{\alpha^2}\right)\right)$$

The initial assumption $1 \geq 2\alpha - b \geq k/d$ does not allow us to reach any conclusions about $2 - b/\alpha^2$, hence we cannot have a guarantee of contraction.□

Example 2.7 *Taking $Q : \mathbb{R}^d \to \mathbb{R}^d$ to be the rand_k operator as defined previously and originally in Stich et al. [2018], we have $E[Q(\mathbf{x})] = \frac{k}{d}\mathbf{x}$ and $E[||\mathbf{x} - Q(\mathbf{x})||^2 \leq \left(1 - \frac{k}{d}\right)||\mathbf{x}||^2$. On the other hand, the function $\tilde{Q} : \tilde{Q}(\mathbf{x}) = \frac{d}{k}Q(\mathbf{x})$ results in $E[\tilde{Q}(\mathbf{x})] = \mathbf{x}$ and $\left(\frac{d}{k} - 1\right)||\mathbf{x}||^2 \geq ||\mathbf{x}||^2$.*

In general, the unbiased operator either does not satisfy the contraction property at all or satisfies it for too constraining cases. As an example to backup this claim, we can look at the function defined by Alistarh et al. [2017]. Taking $s = 2$ which corresponds to having 5 levels in total considering both

positive values, negative ones and 0 would require the number of features to be no more than 4 to satisfy the contraction property, which is not possible in applications.

2.3 Notes on Greedy Algorithms

Greedy algorithms are general methods that rely on solving a subproblem at each stage to obtain a local optimum in the goal of finally obtaining (an approximation of) the global optimum of the overall problem. In section 3.2 and 3.3, we will consider a dictionary $\mathcal{D} = \{\mathbf{v}_j\}_{1 \leq j \leq L}, ||\mathbf{v}_j|| = 1$ and try to find good approximations of the gradient vectors using a linear combination of $M \leq L$ vectors $\mathbf{v}_j$. In particular, we want to solve, for a given vector $\mathbf{x}$

$$\hat{\mathbf{x}} = \underset{\lambda, \mathcal{D}}{\operatorname{argmin}} \left|\left| \mathbf{x} - \sum_{j=1}^{M} \lambda_j \mathbf{v}_j \right|\right|^2 \tag{2.6}$$

Matching Pursuit (MP): This problem has been studied in Mallat and Zhang [1993] where they propose the Matching Pursuit algorithm, which is a greedy method to solve it, with a computational complexity of $\mathcal{O}(Md^2)$. It is initially for functions but can be extended to vectors.

We denote by $R^j\mathbf{x}$ the residual at iteration j and start with $R^0\mathbf{x} = \mathbf{x}$. Then, for $1 \leq j \leq M$, the iterates are given by:

$$R^j\mathbf{x} = R^{j-1}\mathbf{x} - \langle R^{j-1}\mathbf{x}, \mathbf{v}_{(j)} \rangle \mathbf{v}_{(j)}$$

where we have

$$\mathbf{v}_{(j)} = \underset{\mathbf{v} \in \mathcal{D}}{\operatorname{argmax}} |\langle R^{j-1}\mathbf{x}, \mathbf{v} \rangle|$$

We will suppose that a maximum exists at each step. By construction of the algorithm, the approximation of the vector $\mathbf{x}$ at step M is $\hat{\mathbf{x}} = \mathbf{x} - R^M\mathbf{x} = \sum_{j=1}^{M} \langle R^{j-1}\mathbf{x}, \mathbf{v}_{(j)} \rangle \mathbf{v}_{(j)}$. An important observation is that for each iteration, the vectors $R^j\mathbf{x}$ and $\mathbf{v}_{(j)}$ are orthogonal to each other. This leads to the following result.

Lemma 2.8 *For $\mathbf{x} \in \mathbb{R}^d$ and the residual $R^M\mathbf{x}$ at the M-th step of the MP algorithm, we have*

$$||\mathbf{x} - \hat{\mathbf{x}}||^2 = ||R^M\mathbf{x}||^2 \leq ||\mathbf{x}||^2 (1 - I(\mathbf{x}, \mathcal{D})^2)^M$$

where $\hat{\mathbf{x}} = \mathbf{x} - R^M\mathbf{x} = \sum_{j=1}^{M} \langle R^{j-1}\mathbf{x}, \mathbf{v}_{(j)} \rangle \mathbf{v}_{(j)}$ is the approximation of $\mathbf{x}$ at step M and $I(\mathbf{x}, \mathcal{D}) = \underset{\mathbf{x} \in \mathbb{R}^d}{inf} \ \underset{\mathbf{v} \in \mathcal{D}}{sup} |\langle \mathbf{x}, \mathbf{v} \rangle| / ||\mathbf{x}||$.

Proof Using the updates, and knowing that $R^j\mathbf{x}$ and $\mathbf{v}_{(j)}$ are orthogonal, we can write

$$\begin{aligned}||R^M\mathbf{x}||^2 &= ||R^{M-1}\mathbf{x}||^2 - \langle R^{M-1}\mathbf{x}, \mathbf{v}_{(M)}\rangle^2 \\ &= ||R^{M-1}\mathbf{x}||^2 - \underset{\mathbf{v}\in\mathcal{D}}{sup}|\langle R^{M-1}\mathbf{x}, \mathbf{v}\rangle|^2 \\ &= ||R^{M-1}\mathbf{x}||^2 \left(1 - \underset{\mathbf{v}\in\mathcal{D}}{sup}\frac{|\langle R^{M-1}\mathbf{x}, \mathbf{v}\rangle|^2}{||R^{M-1}\mathbf{x}||^2}\right)\end{aligned}$$

Applying these steps recursively until arriving at $R^0\mathbf{x} = \mathbf{x}$, we obtain

$$\begin{aligned}||R^M\mathbf{x}||^2 &= ||\mathbf{x}||^2 \prod_{j=1}^{M}\left(1 - \underset{\mathbf{v}\in\mathcal{D}}{sup}\frac{|\langle R^{j-1}\mathbf{x}, \mathbf{v}\rangle|^2}{||R^{j-1}\mathbf{x}||^2}\right) \\ &\leq ||\mathbf{x}||^2 \left(1 - \underset{\mathbf{x}\in\mathbb{R}^d}{inf}\ \underset{\mathbf{v}\in\mathcal{D}}{sup}\frac{|\langle \mathbf{x}, \mathbf{v}\rangle|^2}{||\mathbf{x}||^2}\right)^M\end{aligned}$$

which gives the desired result. □

Moreover, we observe that the quantity $||\mathbf{x} - \hat{\mathbf{x}}||$ is hence always smaller than $||\mathbf{x}||$ because $0 \leq I(\mathbf{x}, \mathcal{D}) \leq 1$. This result allows us to see that the bound on the variance decays exponentially on M using MP. We will see a tighter bound while using an orthonormal dictionary in Section 3.2.

Considering that $\mathcal{D}$ contains no linearly dependent vector, the chosen vectors at each step will be different from each other. Moreover, if the dictionary contains orthonormal vectors, we have the following pleasing result.

Lemma 2.9 *Considering that the dictionary $\mathcal{D}$ contains orthonormal vectors, the M-step approximation will be* $\hat{\mathbf{x}} = \sum_{j=1}^{M}\langle \mathbf{x}, \mathbf{v}_{(j)}\rangle \mathbf{v}_{(j)}$

Proof The expansion coefficients are given by

$$\begin{aligned}\langle R^{j-1}\mathbf{x}, \mathbf{v}_{(j)}\rangle &= \langle R^{j-2}\mathbf{x} - \langle R^{j-2}\mathbf{x}, \mathbf{v}_{(j-1)}\rangle \mathbf{v}_{(j-1)}, \mathbf{v}_{(j)}\rangle \\ &= \langle R^{j-2}\mathbf{x}, \mathbf{v}_{(j)}\rangle\end{aligned}$$

Repeating this operation $j-2$ times, we obtain $\langle R^{j-1}\mathbf{x}, \mathbf{v}_{(j)}\rangle = \langle \mathbf{x}, \mathbf{v}_{(j)}\rangle$. □

This shows that for an orthonormal set of vectors, and allowing ourselves to approximate $\mathbf{x}$ using only M vectors, the MP algorithm achieves the minimum of the problem defined above. As shown in Mallat and Zhang [1993] and pointed out in Davis [1994], MP may not converge, in the sense that the residual becomes null, in a finite number of steps. In our case, we do not seek the least squares error to be zero, since we accept lossy compression, but for the sake of completeness, we present below an alternative algorithm solving this problem.

Orthogonal Matching Pursuit (OMP): The OMP is an algorithm proposed by Pati et al. [1993] and Davis [1994] in parallel. Its main difference with MP is that, at each step, the algorithm creates an orthonormal set of vectors using the Gram-Schmidt procedure. This brings a better convergence property, which ensures that for a finite dimensional space, the algorithm converges (the residual becomes zero) in a finite number of steps, which may not be the case for MP. However, this comes at a computational cost, which for the OMP, is in the order of $\mathcal{O}\left(\frac{M(M+1)}{2}d^2 + Md^2\right)$ because of the burden of constructing an orthonormal basis at each step.

We denote by $R^j\mathbf{x}$ the residual at iteration j and start with $R^0\mathbf{x} = \mathbf{x}$ and $\mathbf{u}_1 = \mathbf{v}_{(1)}$. Then, for $1 \leq j \leq M$, the iterates are given by:

$$\mathbf{u}_j = \mathbf{v}_{(j)} - \sum_{l=1}^{j} \frac{\langle \mathbf{v}_{(j)}, \mathbf{u}_l \rangle}{||\mathbf{u}_l||^2} \mathbf{u}_l$$

$$R^j\mathbf{x} = R^{j-1}\mathbf{x} - \frac{\langle R^{j-1}\mathbf{x}, \mathbf{u}_j \rangle}{||\mathbf{u}_j||^2} \mathbf{u}_j$$

where we have

$$\mathbf{v}_{(j)} = \underset{\mathbf{v} \in \mathcal{D}}{\operatorname{argmax}} |\langle R^{j-1}\mathbf{x}, \mathbf{v} \rangle|$$

OMP has similar properties to MP; $R^j\mathbf{x}$ and $\mathbf{u}_j$ are orthogonal to each other at each step. More detailed comparisons and analysis on greedy algorithms for vector/function approximation can be found in Davis et al. [1997], DeVore and Temlyakov [1996] and Barron et al. [2008].

In sections 3.2 and 3.3, we will come back to our analysis of this class of algorithms and see that they imply nice properties in our case of compression.

Main Contributions

In this section, we present the main analysis and results obtained in the study of compression function for distributed optimization. We concentrated on several subproblems and several functions are studied. The main goal was to find a better variance bound for these operators and we consider both biased and unbiased examples.

3.1 Analysis of Q_s^p compression functions

As introduced 2.2, Alistarh et al. [2017] and Wen et al. [2017] proposed quantification functions based on the ℓ_2 and ℓ_∞ norms respectively. It consists of quantizing each component of the gradient vectors to a level "normalized" by their respective norms and with a certain probability such that the resulting operator is unbiased. In the former paper, more than one level of quantization has been considered. Inspired by their scheme, we are interested to study it for any ℓ_p norm.

Definition 3.1 (ℓ_p norm) *We define in a vector space V a norm as a function that we denote $||\cdot||$ and which satisfies $||\mathbf{x}|| \geq 0$ with equality if and only if $\mathbf{x} = \mathbf{0}$, $||\alpha\mathbf{x}|| = |\alpha| \cdot ||\mathbf{x}||$ and $||\mathbf{x} + \mathbf{y}|| \leq ||\mathbf{x}|| + ||\mathbf{y}||$ with equality if and only if $\mathbf{y} = \alpha\mathbf{x}$, for some constant $\alpha \in \mathbb{C}$ and $\mathbf{x}, \mathbf{y} \in V$. In particular, if $V = \mathbb{R}^d$, we define the ℓ_p norm of $\mathbf{x}$ for $p \geq 1$, $p \in \mathbb{R}$ as:*

$$||\mathbf{x}||_p = \left(\sum_{i=1}^{d} |x_i|^p \right)^{1/p}$$

We note that we will not consider the case of pseudo-norms $p < 1$ because it fails to satisfy the triangle inequality property which we will need in our analysis.

We now define the general norm quantization function.

Definition 3.2 (Q_s^p) *For an ℓ_p norm and for $s \geq 1$, the Q_s^p compression scheme can be written as:*

$$[Q_s^p(\mathbf{x})]_i = ||\mathbf{x}||_p \cdot sign(x_i) \cdot \xi_i^p(\mathbf{x}, s) \tag{3.1}$$

where

$$\xi_i^p(\mathbf{x}, s) = \begin{cases} l_i/s & \text{with probability } 1 - s|x_i|/||\mathbf{x}||_p + l_i \\ (l_i + 1)/s & \text{otherwise} \end{cases}$$

and $0 \leq l_i < s$ is the integer such that $|x_i|/||\mathbf{x}||_p \in [l_i/s, (l_i + 1)/s]$. This constraint on l_i implies $0 \leq s|x_i|/||\mathbf{x}||_p - l_i \leq 1$, which is indeed coherent with a probability.

Before analyzing the variance bounds, it is important to see if the class of functions that we defined "makes sense" in our context, i.e. whether we achieve convergence. For this, we will need to make some assumptions on bounded gradients but the original papers proposing these functions have made similar ones. We first show unbiasedness of the operator.

Lemma 3.3 *For Q_s^p as defined above, $E[Q_s^p(\mathbf{x})] = \mathbf{x}$.*

The proof can be found in Appendix B.

3.1.1 Convergence

To show convergence, we will use a similar context to the one presented in Wen et al. [2017] which is itself based on results presented in Bottou [1998] and our proof will be adapted from the ones in these works. Namely, we assume SGD is used, but instead of the exact stochastic gradients $\mathbf{x}^{(t)} = \nabla f_{(t)}(\mathbf{w}^{(t)})$, we use its quantized variant $Q_s^p(\mathbf{x}^{(t)})$:

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \gamma_t Q_s^p(\mathbf{x}^{(t)}) \tag{3.2}$$

The following assumptions are necessary for convergence results.

Assumption 3.4 *We consider functions f that contain a single minimum $\mathbf{w}^*$ and that satisfy*

$$\inf_{||\mathbf{w}-\mathbf{w}^*||>\varepsilon} (\mathbf{w}-\mathbf{w}^*)^T \nabla f(\mathbf{w}) > 0, \forall \varepsilon > 0$$

If we only consider convex functions, this assumption is automatically satisfied. Moreover, the step size should have some constraints, they should not decrease too slowly nor too rapidly.

Assumption 3.5 (Robbins and Monro [1985]) *Considering as $\boldsymbol{\gamma}$ the (possibly infinite length) sequence of the step sizes, we assume $\boldsymbol{\gamma} \in \ell_2(\mathbb{N})$ but $\boldsymbol{\gamma} \notin \ell_1(\mathbb{N})$*

A detailed explanation of this assumption can be found in Robbins and Monro [1985]. Finally, our last assumption concerns a bound for the stochastic gradients.

Assumption 3.6 *There exists constants A and B such that the stochastic gradients $\mathbf{x} = \nabla f_i(\mathbf{w})$ for any $i \in \{1, \dots, N\}$ satisfy*

$$E[||\mathbf{x}||_p \cdot ||\mathbf{x}||_1] \leq A + B||\mathbf{w} - \mathbf{w}^*||^2$$

for $p \geq 1$.

Using the first two assumptions, we have the following theorem.

Lemma 3.7 (Bottou [1998]) *For A, B satisfying*

$$\begin{aligned} &E[||\mathbf{w}^{(t+1)} - \mathbf{w}^*||^2 - (1 + B\gamma_t^2)||\mathbf{w}^{(t)} - \mathbf{w}^*||^2 \big| \boldsymbol{\theta}_t] \\ \leq &- 2\gamma_t(\mathbf{w}^{(t)} - \mathbf{w}^*)^T \nabla f(\mathbf{w}^{(t)}) + A\gamma_t^2 \end{aligned}$$

f converges almost surely in probability towards its minimum:

$$Pr\left(\lim_{t \to +\infty} \mathbf{w}^{(t)} = \mathbf{w}^*\right) = 1$$

The set $\boldsymbol{\theta}_t = \{\mathbf{y}^{(t-1)}, \dots, \mathbf{y}^{(0)}; \mathbf{w}^{(t-1)}, \dots, \mathbf{w}^{(0)}; \gamma_{t-1}, \dots, \gamma_0\}$ represents the parameters $\mathbf{y}$ of the objective function (for example data samples), and all the previous computations made by the algorithm.

We will not develop the proof, which can be found in Bottou [1998]. We have now the full prerequisites needed for convergence.

Theorem 3.8 *The algorithm using the iterates*

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \gamma_t Q_s^p(\mathbf{x}^{(t)})$$

converges almost surely to the minimum of f.

Proof We show the case $s = 1$ which is the most "aggressive" compression scheme, hence can be generalized for $s > 1$. The iterations give

$$\begin{aligned} &\mathbf{w}^{(t+1)} - \mathbf{w}^{(t)} = -\gamma_t Q_1^p(\mathbf{x}^{(t)}) \\ &||\mathbf{w}^{(t+1)} - \mathbf{w}^* - \mathbf{w}^{(t)} + \mathbf{w}^*||^2 = \gamma_t^2 ||Q_1^p(\mathbf{x}^{(t)})||^2 \end{aligned}$$

The left hand side of the last equality can be developped as

$$\begin{aligned} &||\mathbf{w}^{(t+1)} - \mathbf{w}^*||^2 + ||\mathbf{w}^{(t)} - \mathbf{w}^*||^2 - 2(\mathbf{w}^{(t+1)} - \mathbf{w}^*)^T(\mathbf{w}^{(t)} - \mathbf{w}^*) \\ =&||\mathbf{w}^{(t+1)} - \mathbf{w}^*||^2 - ||\mathbf{w}^{(t)} - \mathbf{w}^*||^2 + 2(\mathbf{w}^{(t)} - \mathbf{w}^*)^T(\mathbf{w}^{(t)} - \mathbf{w}^{(t+1)}) \\ =&||\mathbf{w}^{(t+1)} - \mathbf{w}^*||^2 - ||\mathbf{w}^{(t)} - \mathbf{w}^*||^2 + 2(\mathbf{w}^{(t)} - \mathbf{w}^*)^T \gamma_t Q_1^p(\mathbf{x}^{(t)}) \end{aligned}$$

Then, we can take the expected value of difference of the two norms given the parameters

$$\begin{aligned} E[||\mathbf{w}^{(t+1)} - \mathbf{w}^*||^2 - ||\mathbf{w}^{(t)} - \mathbf{w}^*||^2 | \boldsymbol{\theta}_t] = &- 2(\mathbf{w}^{(t)} - \mathbf{w}^*)^T \gamma_t E[Q_1^p(\mathbf{x}^{(t)}) | \boldsymbol{\theta}_t] \\ &+ \gamma_t^2 E[||Q_1^p(\mathbf{x}^{(t)})||^2 | \boldsymbol{\theta}_t] \end{aligned}$$

Since Q_s^p is unbiased, we consider only the last term:

$$\begin{aligned} E[||Q_1^p(\mathbf{x}^{(t)})||^2 | \boldsymbol{\theta}_t] &= E[||\mathbf{x}^{(t)}||_p^2 \cdot ||\boldsymbol{\xi}(\mathbf{x}^{(t)}, 1)||^2 | \boldsymbol{\theta}_t] \\ &= E\left[||\mathbf{x}^{(t)}||_p^2 \cdot \sum_{i=1}^{d} \frac{|x_i|}{||\mathbf{x}^{(t)}||_p} \middle| \boldsymbol{\theta}_t\right] \\ &= E[||\mathbf{x}^{(t)}||_p \cdot ||\mathbf{x}^{(t)}||_1 | \boldsymbol{\theta}_t] \end{aligned}$$

Now, using our final assumption, we may write

$$E[||\mathbf{w}^{(t+1)} - \mathbf{w}^*||^2 - ||\mathbf{w}^{(t)} - \mathbf{w}^*||^2|\boldsymbol{\theta}_t] \leq -2(\mathbf{w}^{(t)} - \mathbf{w}^*)^T \gamma_t \nabla f(\mathbf{w}^{(t)}) + A\gamma_t^2 + B\gamma_t^2 E[||\mathbf{w}^{(t)} - \mathbf{w}^*||^2|\boldsymbol{\theta}_t]$$

which allow us to conclude the proof. □

Compared to classical SGD, and assuming bounded gradients, we have the additional term $B\gamma_t^2 E[||\mathbf{w}^{(t)} - \mathbf{w}^*||^2|\boldsymbol{\theta}_t]$. Considering a step size in $\mathcal{O}(1/t)$, $[||\mathbf{w}^{(t)} - \mathbf{w}^*||^2|\boldsymbol{\theta}_t]$ behaves as $\mathcal{O}(1/t)$. For strongly convex objectives, this does not hurt convergence, which will be in $\mathcal{O}(1/t)$ as for SGD, as is shown below.

Proof Jaggi [2018] Assuming μ-strongly convex objective f, bounded gradients such that $||\nabla f_{i_t}(\mathbf{w})^{(t)}||^2 \leq C$, and $\gamma_t = \mathcal{O}(1/t)$, we have for SGD

$$E[f(\mathbf{w}^{(t)}) - f(\mathbf{w}^*)] \leq \mathcal{O}(C/t) + \mathcal{O}(1)$$

Doing a very similar analysis as in Jaggi [2018], for the quantized case, under the previous assumptions, we find

$$E[f(\mathbf{w}^{(t)}) - f(\mathbf{w}^*)] \leq \mathcal{O}(A/t) + \mathcal{O}(B/t^2) + \mathcal{O}(1)$$ □

As a final remark, it is important to point out that this analysis can be extended to mini batch SGD. This shows us that we have guarantees of convergence under some assumptions.

3.1.2 Variance bounds

Earlier, in 2.2, we saw that we cannot have both unbiasedness and contraction, unless we consider cases that are so constraining that do not have any use in practice. The class of functions Q_s^p are unbiased, but in this section we will study for which p we get the best bounds on the variance. Alistarh et al. [2017] have proved it for the ℓ_2 case, and the proof for the general case will be adapted from theirs. First, we remind Hölder's inequality, which gives us a relationship between ℓ_p norms.

Lemma 3.9 (Hölder, vector case) *Let p and $q \geq 1$ such that $1/p + 1/q = 1$. Then, for vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$, we have the following inequality:*

$$\sum_{i=1}^{d} |u_i v_i| \leq ||\mathbf{u}||_p \cdot ||\mathbf{v}||_q$$

The proof follows from Hölder's for functions by taking the measure to be the Dirac delta. Another useful Lemma to study the variance of the compression function is given below and results directly from Hölder's inequality.

Corollary 3.10 *For $1 \leq p < q$ such that $1/p + 1/q = 1$ and $\mathbf{x} \in \mathbb{R}^d$*

$$||\mathbf{x}||_p \leq ||\mathbf{x}||_q \cdot d^{1/p-1/q}$$

The proof of this can be found in Appendix B. We now have all the required tools to write the bound.

Theorem 3.11 *For the class of function Q_s^p, $s \geq 1$, $p \geq 1$, we have the following bound on the variance of the compression*

$$\begin{aligned} E[||\mathbf{x} - Q_s^p(\mathbf{x})||^2] &\leq min\left(\frac{d}{s^2}, \frac{||\mathbf{x}||_1}{s||\mathbf{x}||_p}\right)||\mathbf{x}||_p^2 \\ &\leq min\left(\frac{d}{s^2}, \frac{d^{(p-1)/p}}{s}\right)\zeta(p)||\mathbf{x}||^2 \end{aligned}$$

where ζ is defined as

$$\zeta(p) = \begin{cases} d^{(2-p)/p} & \text{if } 1 \leq p \leq 2 \\ 1 & \text{if } p > 2 \end{cases}$$

Proof We start with

$$\begin{aligned} E\left[||\mathbf{x} - Q_s^p(\mathbf{x})||^2\right] &= E\left[||Q_s^p(\mathbf{x})||^2\right] - ||\mathbf{x}||^2 \\ &= \sum_{i=1}^{d} ||\mathbf{x}||_p^2 \cdot E[\xi_i^p(\mathbf{x}, s)^2] - ||\mathbf{x}||^2 \end{aligned}$$

Let us denote $|x_i|s/||\mathbf{x}||_p - l_i$ by $p(x_i)$. Then, we have

$$E\left[\xi_i^p(\mathbf{x}, s)^2\right] = E[\xi_i^p(\mathbf{x}, s)]^2 + E[(\xi_i^p(\mathbf{x}, s) - E[\xi_i^p(\mathbf{x}, s)])^2]$$

Using the definition of ξ, the right hand side can be written as

$$\begin{aligned} &E[\xi_i^p(\mathbf{x}, s)]^2 + E[(\xi_i^p(\mathbf{x}, s) - E[\xi_i^p(\mathbf{x}, s)])^2] \\ &= \frac{|x_i|^2}{||\mathbf{x}||_p^2} + p(x_i)\left(\frac{l_i + 1}{s} - \frac{|x_i|}{||\mathbf{x}||_p}\right)^2 + (1 - p(x_i))\left(\frac{l_i}{s} - \frac{|x_i|}{||\mathbf{x}||_p}\right)^2 \\ &= \left(\frac{l_i}{s} - \frac{|x_i|}{||\mathbf{x}||_p}\right)^2 + \frac{p(x_i)}{s^2} - \frac{2p(x_i)|x_i|}{s||\mathbf{x}||_p} + \frac{2p(x_i)l_i}{s^2} \\ &= \frac{p(x_i)^2}{s^2} + \frac{p(x_i)}{s^2} + \frac{2p(x_i)}{s^2}\left(l_i - \frac{|x_i|s}{||\mathbf{x}||_p}\right) \\ &= \frac{1}{s^2}p(x_i)(1 - p(x_i)) \end{aligned}$$

Replacing this in the above equation results in

$$\begin{aligned} E\left[\xi_i^p(\mathbf{x}, s)^2\right] &= \frac{|x_i|^2}{||\mathbf{x}||_p^2} + \frac{1}{s^2} p(x_i)(1 - p(x_i)) \\ &\leq \frac{|x_i|^2}{||\mathbf{x}||_p^2} + \frac{1}{s^2} p(x_i) \end{aligned}$$

Which gives us for the original computation

$$E\left[||\mathbf{x} - Q_s^p(\mathbf{x})||^2\right] \leq \sum_{i=1}^{d} \frac{||\mathbf{x}||_p^2}{s^2} p(x_i)$$

Since $p(x_i) \leq 1$ and $p(x_i) \leq |x_i| s / ||\mathbf{x}||_p$, we may write

$$\begin{aligned} E\left[||\mathbf{x} - Q_s^p(\mathbf{x})||^2\right] &\leq \min\left(\frac{d}{s^2}, \frac{||\mathbf{x}||_1}{s||\mathbf{x}||_p}\right) ||\mathbf{x}||_p^2 \\ &= \min\left(\frac{d}{s^2}, \frac{||\mathbf{x}||_1}{s||\mathbf{x}||_p}\right) \frac{||\mathbf{x}||_p^2}{||\mathbf{x}||^2} ||\mathbf{x}||^2 \end{aligned}$$

Using Corollary 3.10, for $1 \leq p < q$, we have $||\mathbf{x}||_p \leq ||\mathbf{x}||_q \cdot d^{1/p - 1/q}$. If $p \geq q$, we simply have $||\mathbf{x}||_p \leq ||\mathbf{x}||_q$. Then, we can write:

$$\begin{aligned} E\left[||\mathbf{x} - Q_s^p(\mathbf{x})||^2\right] &\leq \min\left(\frac{d}{s^2}, \frac{||\mathbf{x}||_1}{s||\mathbf{x}||_p}\right) \frac{||\mathbf{x}||_p^2}{||\mathbf{x}||^2} ||\mathbf{x}||^2 \\ &\leq \min\left(\frac{d}{s^2}, \frac{d^{(p-1)/p}}{s}\right) \zeta(p) ||\mathbf{x}||^2 \end{aligned}$$ □

From the results of this theorem, we see that the bound resembles to the one in Alistarh et al. [2017] and gives the same result for $p = 2$. In particular, if $1 \leq p \leq 2$, we have:

$$E\left[||\mathbf{x} - Q_s^p(\mathbf{x})||^2\right] \leq \min\left(\frac{d^{2/p}}{s^2}, \frac{d^{1/p}}{s}\right) ||\mathbf{x}||^2$$

and if $p > 2$:

$$E\left[||\mathbf{x} - Q_s^p(\mathbf{x})||^2\right] \leq \min\left(\frac{d}{s^2}, \frac{d^{(p-1)/p}}{s}\right) ||\mathbf{x}||^2$$

In the first case, if $d < s$, then the variance is upper bounded by $\frac{d^{2/p}}{s^2}||\mathbf{x}||^2$ which is minimized for $p = 2$. If $\sqrt{d} > s$, then the upper bound is $\frac{d^{1/p}}{s}||\mathbf{x}||_2$,

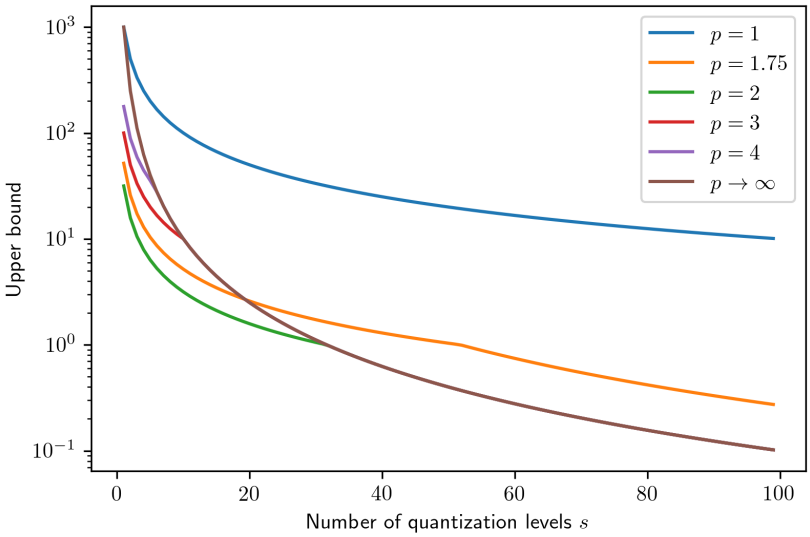

Figure 3.1: Comparison of the upper bound for various values of p. In this example, $d = 1000$.

which is again minimized by $p = 2$. Otherwise, it can also be shown that the variance is minimized for the 2-norm. For the second case, if $s < 1$, the variance is upper bounded by $\frac{d^{(p-1)/p}}{s}||\mathbf{x}||^2$, which is minimized for $p \to 2$. If $s > \sqrt{d}$, the variance is upper bounded by $\frac{d}{s^2}||\mathbf{x}||^2$, which does not depend on p. Otherwise, it can be shown that the lowest variance bound is obtained for $p \to 2$. Figure 3.1 shows an example of this upper bound for some values of p. Putting the computational costs aside, the lowest upper bound we can achieve is for $p = 2$ and is given by $\min\left(\frac{d}{s^2}, \frac{\sqrt{d}}{s}\right)||\mathbf{x}||^2$.

Moreover, the more quantization levels we have the better the bound but this means that we "compress" less, since the quantized vector will have more different values. Overall, we can see that $p = 2$ is generally the best choice to keep the variance lowest but this comes at a cost. Since the norm becomes larger as p becomes smaller, we need to have a more constraining assumption for the gradient bound from 3.6. Hence, p and s should be seen as parameters that need to be tuned in practice, based on the problem.

3.2 Orthogonal projections for compression, a Signal Processing approach

In this section we turn our interest to another compression scheme and present a general framework using orthogonal projections. In this case, we want to keep the contraction property, hence have a biased compression

operator. For this, the idea has been inspired by sampling and interpolation from signal processing. We can see the (stochastic) gradient vectors $\mathbf{x}$ as finite-length discrete sequences that we sample, therefore the resulting vector will have a smaller size and will be the one to be transmitted. At reception, it will be interpolated so as to have an approximation of the original vector, $Q(\mathbf{x})$, the compressed vector. The notation is largely based from Vetterli et al. [2014], where further detail on Signal Processing can be found. In this section, we will see that the approximations $Q(\mathbf{x})$ are satisfying the contraction property, therefore SGD with memory as presented in Theorem 2.4 converges for the choices of Q presented below.

3.2.1 Definition and Properties

We define an orthonormal set of vectors $\{\varphi_k\}_{1\leq k\leq M} \subset \mathcal{B}$, where $\mathcal{B}$ is an orthonormal basis for $\mathbb{R}^d$, with $\varphi_k \in \mathbb{R}^d$ and the matrix Φ of this set where the columns of Φ are the vectors φ_k. Consequently, Φ^*, the adjoint of Φ, corresponds to the sampling operator and Φ is the interpolation operator. In the case of Φ being a real matrix, $\Phi^* = \Phi^T$. Therefore, sampling followed by interpolation can be represented as the operator (matrix) product $\Phi\Phi^*$. As $\{\varphi_k\}_{1\leq k\leq M}$ is an orthonormal set, we note that $\Phi^*\Phi = I$

Lemma 3.12 *$\Phi\Phi^*$ is an orthogonal projection.*

Proof We need to verify that the operators squared give back themselves so that they are indeed projections. Moreover, for them to be orthogonal projection we need to verify that they are self-adjoint. We have $(\Phi\Phi^*)^2 = \Phi\Phi^*\Phi\Phi^* = \Phi\Phi^*$ because $\Phi^*\Phi = I$ hence we indeed have a projection. Moreover, $(\Phi\Phi^*)^* = \Phi\Phi^*$ and we have an orthogonal projection. □

We now define our compression function Q to be $Q(\mathbf{x}) = \Phi\Phi^*\mathbf{x}$. The previous Lemma leads to very interesting properties. We saw that the quantity $||\mathbf{x} - Q(\mathbf{x})||^2$ is of great importance in our analysis. Moreover, since Q is an orthogonal projection, it projects $\mathbf{x}$ into a subspace of $\mathbb{R}^d$, namely $\mathcal{R}(\Phi) = \{\mathbf{v} : \mathbf{v} = \Phi\mathbf{u}\}$, the range space of Φ.

Lemma 3.13 *For $\mathbf{x} \in \mathbb{R}^d$, $Q(\mathbf{x}) = \Phi\Phi^*\mathbf{x}$ is the unique orthogonal projection of $\mathbf{x}$ onto $\mathcal{R}(\Phi)$ that satisfies*

$$Q(\mathbf{x}) = \underset{\mathbf{v}\in\mathcal{R}(\Phi)}{argmin}||\mathbf{x} - \mathbf{v}||^2$$

The proof can be found in Vetterli et al. [2014] Pages 52 and 53 and is based on the Hilbert orthogonal projection theorem. Hence, by using an orthogonal projection, we have a first guarantee of optimality in $\mathcal{R}(\Phi)$. We can take our analysis even further so that we ask ourselves how good these orthogonal projections are in terms of the contraction property.

Theorem 3.14 *The orthogonal projections $Q(\mathbf{x})$ are guaranteed to satisfy the contraction property:*

$$||\mathbf{x} - Q(\mathbf{x})||^2 \leq ||\mathbf{x}||^2$$

Proof We have

$$\begin{aligned} ||\mathbf{x} - Q(\mathbf{x})||^2 &= ||\mathbf{x} - \Phi\Phi^*\mathbf{x}||^2 \\ &= ||(I - \Phi\Phi^*)\mathbf{x}||^2 \\ &\leq ||I - \Phi\Phi^*||_2^2 \cdot ||\mathbf{x}||^2 \\ &= \lambda_{max}\left((I - \Phi\Phi^*)^*(I - \Phi\Phi^*)\right)||\mathbf{x}||^2 \end{aligned}$$

where λ_{max} corresponds to the largest eigenvalue of the matrix. Moreover, we see that $(I - \Phi\Phi^*)^* = I - \Phi\Phi^*$, therefore, $(I - \Phi\Phi^*)^*(I - \Phi\Phi^*) = I - \Phi\Phi^*$. Using Lemma B.1, we can compute the maximum eigenvalue of this matrix and conclude that

$$\begin{aligned} ||\mathbf{x} - Q(\mathbf{x})||^2 &\leq (1 - \lambda_{min}(\Phi\Phi^*))||\mathbf{x}||^2 \\ &= ||\mathbf{x}||^2 \end{aligned}$$

□

Hence, choosing an orthogonal projection to project the original vector into a subspace gives a that norm of the difference will be smaller than the norm of $\mathbf{x}$. Moreover, Theorem 2.4 proposes an algorithm that guarantees convergence for this type of compression methods, hence we satisfy convergence in this framework.

Corollary 3.15 *Using a compression function Q as defined above, SGD with memory using the iterations described below (Cf. Theorem 2.4) converges.*

$$\mathbf{w}^{(\mathbf{t}+\mathbf{1})} = \mathbf{w}^{(\mathbf{t})} - Q(\mathbf{m}^{(t)} + \gamma_t \nabla f_{i_t}(\mathbf{w}^{(\mathbf{t})})) = \mathbf{w}^{(\mathbf{t})} - \mathbf{g}^{(t)}$$

Where $\mathbf{m}^{(0)} = \mathbf{0}$, and for $t \geq 1$, $\mathbf{m}^{(t+1)} = \mathbf{m}^{(t)} + \gamma_t \nabla f_{i_t}(\mathbf{w}^{(\mathbf{t})}) - \mathbf{g}^{(t)}$.

This result is however at the limit of satisfying the contraction property, it is an interesting result for such a general case of compression but not satisfying enough. To achieve better compression, we need to look at more specific cases.

3.2.2 $rand_M^{\mathcal{B}}$ and $top_M^{\mathcal{B}}$

There are two intuitive choices we decided to look into, one is to select M vectors φ_k from $\mathcal{B}$ uniformly at random and doing an orthogonal projection using these, and the other is to choose the vectors minimizing the variance. Choosing $\mathcal{B}$ to be the standard basis, the first option corresponds to the $rand_M$ operator, where we choose components of $\mathbf{x}$ randomly and keep only those, putting the others to 0. The latter is top_M, keeping only the entries

with the highest magnitude (taking account of the sign). Both are defined in Section 2 and originally in Stich et al. [2018]. We also define the notation $rand_M^{\mathcal{B}}$ and $top_M^{\mathcal{B}}$ to denote that the orthonormal basis in question may not be the standard one.

Lemma 3.16 *The random compression operator $rand_M^{\mathcal{B}}$ for an orthonormal basis $\mathcal{B}$ of $\mathbb{R}^d$ is a biased function with $E[rand_M^{\mathcal{B}}(\mathbf{x})] = \frac{M}{d}\mathbf{x}$, and satisfies the contraction property with parameter M. Moreover, $top_M^{\mathcal{B}}$ also verifies this property with parameter less than M.*

Proof We first concentrate on the results for $Q = rand_M^{\mathcal{B}}$. Let us consider the sets of indices $\mathcal{I}^j \subseteq [d]$, $|\mathcal{I}^j| = M$ which all contain a different combination of M indices out of d, there are $\binom{d}{M}$ such sets and $Pr(\text{The chosen set is } \mathcal{I}^j) = 1/\binom{d}{M}$. We denote by $(\Phi\Phi^*)^j_{rc}$ the component on the r-th row and c-th column of the projection matrix obtained using the j-th subset. Then,

$$(\Phi\Phi^*)^j_{rc} = \sum_{i_j \in \mathcal{I}^j} \varphi_{i_j}(r)\varphi_{i_j}(c)$$

Moreover,

$$E[\Phi\Phi^*] = \sum_{j=1}^{\binom{d}{l}} \frac{1}{\binom{d}{M}} (\Phi\Phi^*)^j$$

$$\Rightarrow E[(\Phi\Phi^*)_{rc}] = \sum_{j=1}^{\binom{d}{M}} \frac{1}{\binom{d}{M}} \sum_{i_j \in \mathcal{I}^j} \varphi_{i_j}(r)\varphi_{i_j}(c)$$

Each index appears $\binom{d}{M} \cdot \frac{M}{d}$ times, hence

$$E[(\Phi\Phi^*)_{rc}] = \frac{1}{\binom{d}{M}} \cdot \binom{d}{M} \cdot \frac{M}{d} \sum_{k=1}^{d} \varphi_k(r)\varphi_k(c)$$

The vectors $\tilde{\varphi}_m = [\varphi_1(m), \varphi_2(m), \ldots, \varphi_d(m)]^T$, $1 \leq m \leq d$ are orthonormal since $\{\varphi_k\}_{1 \leq k \leq d}$ is an orthonormal basis. Hence $E[(\Phi\Phi^*)_{rc}] = \frac{M}{d}\delta_{rc}$ which leads us to conclude:

$$E[rand_M^{\mathcal{B}}(\mathbf{x})] = E[\Phi\Phi^*]\mathbf{x} = \frac{M}{d}\mathbf{x}$$

As for the variance, using the result for the expectation allows us to write:

$$\begin{aligned}
E[||\mathbf{x} - rand_M^{\mathcal{B}}(\mathbf{x})||^2] &= ||\mathbf{x}||^2 + E[rand_M^{\mathcal{B}}(\mathbf{x})]^T\mathbf{x} \\
&\quad - \mathbf{x}^T E[rand_M^{\mathcal{B}}(\mathbf{x})] - E[rand_M^{\mathcal{B}}(\mathbf{x})^T rand_M^{\mathcal{B}}(\mathbf{x})] \\
&= ||\mathbf{x}||^2 - \frac{2M}{d}||\mathbf{x}||^2 + \mathbf{x}^T E[(\Phi\Phi^*)^T\Phi\Phi^*]\mathbf{x} \\
&= ||\mathbf{x}||^2 - \frac{2M}{d}||\mathbf{x}||^2 + \mathbf{x}^T E[\Phi\Phi^*]\mathbf{x} \\
&= ||\mathbf{x}||^2 \left(1 - \frac{M}{d}\right)
\end{aligned}$$

On the other hand, $top_M^{\mathcal{B}}$ chooses vectors so as to minimize the last quantity, hence, by construction $||\mathbf{x} - top_M^{\mathcal{B}}(\mathbf{x})||^2 \leq ||\mathbf{x} - rand_M^{\mathcal{B}}(\mathbf{x})||^2$, which implies that $top_M^{\mathcal{B}}$ satisfies the contraction property with parameter less than M. □

Given a basis $\mathcal{B}$, the best orthogonal projection we can do to minimize the variance using M vectors is hence given by the $top_M^{\mathcal{B}}$ operator. However, in contrast to $rand_M^{\mathcal{B}}$ for example, to compute it we need to find the optimal M vectors. To solve this problem, we may use a greedy algorithm as introduced in Section 2.3.

Lemma 3.17 *Using the Matching Pursuit algorithm and considering using a dictionary $\mathcal{D} = \mathcal{B}$, the M-th step of the algorithm for vector $\mathbf{x}$ will result in $top_M^{\mathcal{B}}(\mathbf{x})$.*

To see this, we use the result of Lemma 2.9, where we have shown that when we consider an orthonormal set of vectors for the MP algorithm, we achieve the optimal solution. This is a pleasing result as we are guaranteed optimality without requiring too much computations. Lemma 2.8 already allowed us to upper bound $||\mathbf{x} - \text{top}_M^{\mathcal{B}}(\mathbf{x})||^2$ with $\mathcal{I}(\mathbf{x}, \mathcal{B}) = 1/\sqrt{d}$, since we consider an orthonormal dictionary, where the bound decays in $\mathcal{O}\left((1 - 1/d)^M\right)$. However, we see from 3.16 that we have a tighter bound that follows directly from the contraction property.

3.2.3 Choosing among bases

Until now, we have considered an arbitrary orthonormal basis $\mathcal{B}$ and the previous results assume that it is given. The final question in this context to ask ourselves is how do we choose a good orthonormal basis. For this purpose, let us assume that we have some constraints or a prior knowledge on the vectors, which can be represented as $\mathbf{x} \in \mathcal{S} \subset \mathbb{R}^d$. Coming back to the initial definition, the problem can be described as solving $\min_{\Phi^*} \max_{\mathbf{x}\in\mathcal{S}} ||\Phi^*\mathbf{x}||^2/||\mathbf{x}||^2$, as will be shown below. This corresponds to minimizing the operator norm of Φ^* constrained on the set $\mathcal{S}$ over the possible orthonormal bases.

Proposition 3.18 *We assume that the vectors to be compressed satisfy some constraints/we have some prior knowledge of them, which is described by the set* $\mathcal{S} \in \mathbb{R}^d$*. The problem of choosing the optimal basis can be written as*

$$\min_{\Phi^*} \max_{\mathbf{x}\in\mathcal{S}} ||\Phi^*\mathbf{x}||^2/||\mathbf{x}||^2$$

If we do not have a prior knowledge of these vectors, i.e. $\mathcal{S} = \mathbb{R}^d$*, then we can take any orthonormal basis spanning* $\mathbb{R}^d$*.*

Proof To see that the problem can be written as such, we come back to the initial definition $||\mathbf{x} - Q(\mathbf{x})||^2 = ||\mathbf{x} - \Phi\Phi^*\mathbf{x}||^2$. Then, we have

$$||\mathbf{x} - \Phi\Phi^*\mathbf{x}||^2 = ||\mathbf{x}||^2 - \mathbf{x}^T\Phi\Phi^*\mathbf{x} = ||\mathbf{x}||^2 - ||\Phi^*\mathbf{x}||^2$$

Since we want to minimize this variance for any $\mathbf{x} \in \mathcal{S}$, this is equivalent to maximizing $||\Phi^*\mathbf{x}||^2/||\mathbf{x}||^2$, which is the operator norm of Φ^* on $\mathcal{S}$. Finally, minimizing this result over all various orthonormal bases gives the problem above. If $\mathcal{S} = \mathbb{R}^d$, then this corresponds to maximizing the spectral norm of Φ^*. However, $||\Phi^*||_2 = \sqrt{\lambda_{max}(\Phi\Phi^*)} = 1$, $\forall\Phi^*$, using Lemma B.1, since $\Phi\Phi^*$ is an orthogonal projection. Hence, we can take any orthonormal basis. □

If we assume that the vectors in $\mathcal{S}$ are sparse in the standard basis sense, i.e. contain a large number of entries close to zero, a natural choice would be to take the standard basis to represent these vectors. Unfortunately, without a good knowledge of $\mathcal{S}$, we cannot develop further this problem, but assume we have access to K vectors $\mathbf{x}_j$, $1 < j \leq K$ created by the same process in independent cases. In practice, this can correspond to keeping in memory previous stochastic gradient vectors. The method described below is based from an analysis in Gastpar et al. [2018], and further details can be found on these notes. As before, we are trying to find an approximation $\hat{\mathbf{x}}_j = Q(\mathbf{x}_j) = \Phi\mathbf{f}_j$ for each vector so as to minimize $||\mathbf{x}_j - \hat{\mathbf{x}}_j||^2$. In this notation, the columns of Φ correspond to the M' basis vectors $\varphi_i \in \mathbb{R}^d$ we want to find and $\mathbf{f}_j$ is an M dimensional vector (representing $\Phi^*\mathbf{x}_j$). A way to do this would be to minimize the sum $\sum_{j=1}^K ||\mathbf{x}_j - \hat{\mathbf{x}}_j||^2$ over the basis vectors and over $\mathbf{f}_j$'s. It can be shown that this sum is equal to $||X - \Phi F||_F^2$, where $||\cdot||_F$ is the Frobenius matrix norm, X is a $d \times K$ matrix containing $\mathbf{x}_j$'s in its columns and F is an $M \times K$ matrix with columns given by $\mathbf{f}_j$'s.

Theorem 3.19 (Eckart-Young) *Let* X *be a rank* r *matrix and its singular value decomposition given by* $\sum_{i=1}^r \sigma_i\mathbf{u}_i\mathbf{v}_i^T$*, with the singular values* σ_i *of* X *arranged in decreasing order. Then, for* $M' < r$*, the truncated sum*

$$\tilde{X} = \sum_{i=1}^{M'} \sigma_i\mathbf{u}_i\mathbf{v}_i^T$$

is a minimizer of $||X - A||_F$ *over the matrices with rank less than* M'*, and the minimum is* $||X - \tilde{X}||_F = \left(\sum_{i=M'+1}^{r} \sigma_i^2\right)^{1/2}$.

This important theorem has its origins in Eckart and Young [1936], where a further analysis can be found. The result in our context is summarized in the following Lemma.

Lemma 3.20 *Considering we have K vectors* $\mathbf{x}_j \in \mathbb{R}^d$ *which represent the columns of the matrix X, we want to find* $M' < d$ *vectors* $\varphi_i \in \mathbb{R}^d$ *to have a good representation of the* $\mathbf{x}_j$*, in the sense that we minimize* $\sum_{j=1}^{K} ||\mathbf{x}_j - \hat{\mathbf{x}}_j||^2$*, where* $\hat{\mathbf{x}}_j$*'s are a linear combination of these vectors. Then, taking* $\varphi_i = \mathbf{u}_i$*,* $1 \leq i \leq M'$ *achieves this objective, where* $\mathbf{u}_i$*'s are the left singular vectors of X.*

The result directly comes from Theorem 3.19. Hence the (pseudo-) basis consisting of the left singular vectors of X can be used to represent with minimum error the vectors $\mathbf{x}_j$. As a side note, this technique also corresponds to Principal Component Analysis applied to X. Taking $M' > M$ allow us to use this (pseudo-) basis to approximate new vectors, assuming that they are created by a similar process, using MP.

3.3 Study of extended/non-orthonormal sets

Until now, we only considered orthonormal bases, and using their properties, arrived at interesting results. Another idea is now to look at other sets, in particular to study what happens when we add some vectors to the orthonormal basis. As in the previous section, we want to find a compression operator Q which will compute the optimal M approximation of $\mathbf{x}$, but this time using an overcomplete dictionary. This section discusses the challenges of this extension and is intended more as an introduction for potential future development.

Formally, we consider a set of vectors $\mathcal{V} = \{\mathbf{v}_k\}_{1 \leq k \leq L'} \subset \mathbb{R}^d$ such that $||\mathbf{v}_k||^2 = 1, \forall k$ and the new dictionary will be $\mathcal{D} = \mathcal{B} \cup \mathcal{V}$, where $\mathcal{B}$ is an orthonormal basis. This represents an overcomplete representation of $\mathbb{R}^d$, as the orthonormal basis would be sufficient to express any vector of dimension d. The advantage to have a redundant dictionary is that it gives more flexibility and can better adapt to the problem.

This of course comes at a cost, first of all, we have a more complex problem, and in fact, it is shown in Davis et al. [1997] (Theorem 2.1) that finding the optimal M approximation given all possible dictionaries and a vector $\mathbf{x} \in \mathbb{R}^d$ is NP-hard. As a note, this does not imply that the problem is NP-hard for a given dictionary, i.e., given that we already have a dictionary, computing the optimal M approximation over this dictionary is not an intractable problem. Moreover, the results of greedy algorithms are also affected by this. For

a non-orthonormal set, MP does not guarantee convergence to the global optimum of Problem 2.6. If, on the other hand, we consider an extended dictionary as described above, then the performance of MP is still affected. Indeed, in DeVore and Temlyakov [1996], it is shown that the dictionary $\tilde{\mathcal{D}} = \mathcal{B} \cup \{\mathbf{u}\}$, for a suitably chosen vector $\mathbf{u}$, will lead to an increased approximation error of MP compared to the the dictionary containing only $\mathcal{B}$, using the same number of steps. It is shown in DeVore and Temlyakov [1996] Theorem 4.1 that the approximation error (in the least squares sense) is lower bounded by a term in $\mathcal{O}(M^{-1/2})$ when $\mathbf{x}$ is a linear combination of only two vectors in $\mathcal{B}$, whereas if we were to use MP only on $\mathcal{B}$, for this $\mathbf{x}$, we would have achieved a least squares error of 0, as long as we use $M > 2$ steps. In the more general case where we consider an arbitrary dictionary, we have an upper bound in $\mathcal{O}(M^{-1/6})$.

To tackle this problem, we can think of an algorithm that applies Matching Pursuit first on the basis $\mathcal{B}$ only, and then on $\mathcal{V}$ and takes the best result. This gives a guarantee of being at least as good as MP on an orthonormal basis and for a suitable $\mathcal{V}$ we have an even better result. Formally, the algorithm will choose the vectors in $\mathcal{V}$ if $\sum_{j=1}^{M} |\langle R_{\mathcal{V}}^{j-1}\mathbf{x}, \mathbf{v}_{(j)}\rangle|^2 \geq \sum_{j=1}^{M} |\langle \mathbf{x}, \varphi_{(j)}\rangle|^2$ for a given $\mathbf{x}$, where $R_{\mathcal{V}}^{j}\mathbf{x}$ is the residual of MP at step j using the dictionary $\mathcal{V}$. Unfortunately, for a general case, we cannot develop much further this analysis.

Similar studies have been done for the case where the dictionary contains a set of vectors $\{\mathbf{v}_i\}_{1\leq i\leq d'}$ that constitutes a frame for $\mathbb{R}^d$, with $d' > d$. Frames are overcomplete expansions of a space that satisfy $A||\mathbf{x}||^2 \leq \sum_{i=1}^{d'} |\langle \mathbf{x}, \mathbf{v}_i\rangle|^2 \leq B||\mathbf{x}||^2$, $\forall \mathbf{x} \in \mathbb{R}^d$, for some constants A and B. Approximations of functions using frames are analyzed in Adcock and Huybrechs [2016] and can be adapted to vectors. On the other hand, Engan et al. [1999] studies a method to design frame vectors that fit well the context.

Conclusions

This work gives a theoretical analysis on compression operators in the context of distributed optimization. For this purpose, we defined the class of functions we called Q_s^p based on ℓ_p norms of the gradient vectors. The study showed that in general, the ℓ_2 norm gives the lowest variance, considering that we have enough quantization levels, but we still can tune these parameters in practice to better adapt the problem.

As a second subproblem, we presented a general approximation framework that is based on orthogonal projections. They have the pleasing property of satisfying the contraction property which guarantees convergence when using Stochastic Gradient Descent with memory. In particular, we can achieve the optimal M approximation of the gradient vectors using the Matching Pursuit greedy algorithm given an orthonormal basis. Choosing on which basis to represent the vectors is on the other hand based more application dependent. For sparse vectors, the standard basis is a simple and effective choice. Another option would be to construct the basis by using the left singular vectors of a collection of gradient vectors. A further study can be based on a similar approach, where we can use the fact that usually the gradients of the loss function are linear combinations of the data vectors. Hence the space containing the gradients is spanned by the data vectors, and this property can be used to construct a basis for representing an approximation of the gradient vectors.

The study of extended sets has been unfortunately less rewarding but looks like a promising study case under some assumptions. There has already been some research on overcomplete representations using frames, and these can be used as a starting point.

Appendix A

Appendix for "Baselines and Previous Works"

Definition A.1 *A convex and differentiable function* $f : \mathbb{R}^d \to \mathbb{R}$ *is called L-smooth if* $\exists L > 0$:

$$||\nabla f(\mathbf{w}_1) - \nabla f(\mathbf{w}_2)||^2 \leq L||\mathbf{w}_1 - \mathbf{w}_2||^2$$
$$\iff f(\mathbf{w}_1) \leq f(\mathbf{w}_2) + \nabla f(\mathbf{w}_2)^T(\mathbf{w}_1 - \mathbf{w}_2) + \frac{L}{2}||\mathbf{w}_1 - \mathbf{w}_2||^2$$

$\forall \mathbf{w}_1, \mathbf{w}_2 \in \mathbb{R}^d$

Definition A.2 *A convex and differentiable function* $f : \mathbb{R}^d \to \mathbb{R}$ *is called* μ*-strongly convex if* $\exists \mu > 0$:

$$f(\mathbf{w}_1) \geq f(\mathbf{w}_2) + \nabla f(\mathbf{w}_2)^T(\mathbf{w}_1 - \mathbf{w}_2) + \frac{\mu}{2}||\mathbf{w}_1 - \mathbf{w}_2||^2$$

$\forall \mathbf{w}_1, \mathbf{w}_2 \in \mathbb{R}^d$

Proof 2.1 From the first order condition of convexity of f, we have:

$$f(\mathbf{w}^{(t+1)}) - f(\mathbf{w}^{(t)}) \leq \nabla f(\mathbf{w}^{(t+1)})^T(\mathbf{w}^{(t+1)} - \mathbf{w}^{(t)})$$

Moreover, the L-smoothness of f implies:

$$(\nabla f(\mathbf{w}^{(t+1)}) - \nabla f(\mathbf{w}^{(t)}))^T(\mathbf{w}^{(t+1)} - \mathbf{w}^{(t)}) \leq L||\mathbf{w}^{(t+1)} - \mathbf{w}^{(t)}||^2$$

Combining both of these relationships and taking the expectation, we write:

$$\begin{aligned}
&E[f(\mathbf{w}^{(t+1)}) - f(\mathbf{w}^{(t)})] \leq E[\nabla f(\mathbf{w}^{(t)})^T(\mathbf{w}^{(t+1)} - \mathbf{w}^{(t)})] + LE[||\mathbf{w}^{(t+1)} - \mathbf{w}^{(t)}||^2] \\
&= -\gamma_t E[||\nabla f(\mathbf{w}^{(t)})||^2] + L\gamma_t^2 E[||\nabla f_{i_t}(\mathbf{w}^{(t)})||^2] \\
&= (\gamma_t^2 L - \gamma_t)E[||\nabla f(\mathbf{w}^{(t)})||^2] + L\gamma_t^2 E[||\nabla f_{i_t}(\mathbf{w}^{(t)}) - \nabla f(\mathbf{w}^{(t)})||^2]
\end{aligned}$$

where, for the second equality we used the updates of SGD and for the last, the fact that for a random vector X, $E[||X - E[X]||^2] = E[||X||^2] - ||E[X]||^2$, which ends the proof. □

Appendix for "Main Contributions"

Proof 3.3 For each component of the quantized vector, we can write:

$$\begin{aligned}
&E[[Q_s^p(\mathbf{x})]_i] = ||\mathbf{x}||_p \cdot \text{sign}(x_i) \cdot E[\xi_i^p(\mathbf{x}, s)] \\
&= ||\mathbf{x}||_p \cdot \text{sign}(x_i) \cdot \left(\frac{l_i}{s} \left(1 - s \frac{|x_i|}{||\mathbf{x}||_p} + l_i \right) + \frac{(l_i + 1)}{s} \left(s \frac{|x_i|}{||\mathbf{x}||_p} - l_i \right) \right) \\
&= ||\mathbf{x}||_p \cdot \text{sign}(x_i) \cdot \frac{|x_i|}{||\mathbf{x}||_p} = x_i
\end{aligned}$$

Hence, generalizing this result for any component, $E[Q_s^p(\mathbf{x})] = \mathbf{x}$. □

Proof 3.10 Using Hölder's inequality and taking a vector $\mathbf{v} : v_i = |x_i|^p$, then :

$$\begin{aligned}
||\mathbf{v}||_1 = ||\mathbf{x}||_p^p &= \sum_{i=1}^{d} |x_i|^p \cdot 1 \\
&\leq \left(\sum_{i=1}^{d} (|x_i|^p)^{q/p} \right)^{p/q} \cdot \left(\sum_{i=1}^{d} 1^{q/(q-p)} \right)^{(q-p)/q} \\
&= \left(\sum_{i=1}^{d} (|x_i|^p)^{q/p} \right)^{p/q} \cdot d^{(q-p)/q}
\end{aligned}$$

□

Lemma B.1 *The eigenvalues λ of an orthogonal projection matrix P are either* 0 *or* 1.

Proof For an orthogonal projection P, we have:

$$\lambda v = Pv = P^2 v = P(\lambda v) = \lambda^2 v$$

Hence $\lambda^2 = \lambda$ and $\lambda \in \{0, 1\}$. □

Bibliography

Frank Seide, Hao Fu, Jasha Droppo, Gang Li, and Dong Yu. 1-Bit Stochastic Gradient Descent and Application to Data-Parallel Distributed Training of Speech DNNs. In *Interspeech 2014*, September 2014. URL https://www.microsoft.com/en-us/research/publication/1-bit-stochastic-gradient-descent-and-application-to-data-parallel-distributed-training-of-speech-dnns/.

Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Anima Anandkumar. signSGD: compressed optimisation for non-convex problems. *arXiv preprint arXiv:1802.04434*, 2018.

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*, pages 1709–1720, 2017.

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*, pages 1509–1519, 2017.

Sebastian U Stich, Jean-Baptiste Cordonnier, and Martin Jaggi. Sparsified sgd with memory. In *Advances in Neural Information Processing Systems*, pages 4452–4463, 2018.

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

Volkan Cevher. EPFL, EE-556, Mathematics of Data: from theory to computation. Lecture slides, 2015. URL https://lions.epfl.ch/mathematics_of_data.

Martin Jaggi. EPFL, CS-439, Optimization for Machine Learning. Lecture slides, 2018. URL https://github.com/epfml/OptML_course.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in neural information processing systems*, pages 315–323, 2013.

Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in neural information processing systems*, pages 1646–1654, 2014.

Lam M Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. Sarah: A novel method for machine learning problems using stochastic recursive gradient. *arXiv preprint arXiv:1703.00102*, 2017.

Yossi Arjevani and Ohad Shamir. Communication complexity of distributed convex learning and optimization. In *Advances in neural information processing systems*, pages 1756–1764, 2015.

Frederik Künstner. Fully quantized distributed gradient descent. 2017. URL http://infoscience.epfl.ch/record/234548.

Hantian Zhang, Jerry Li, Kaan Kara, Dan Alistarh, Ji Liu, and Ce Zhang. The zipml framework for training models with end-to-end low precision: The cans, the cannots, and a little bit of deep learning. *arXiv preprint arXiv:1611.05402*, 2016.

Hongyi Wang, Scott Sievert, Shengchao Liu, Zachary Charles, Dimitris Papailiopoulos, and Stephen Wright. Atomo: Communication-efficient learning via atomic sparsification. In *Advances in Neural Information Processing Systems*, pages 9872–9883, 2018.

Stéphane Mallat and Zhifeng Zhang. Matching pursuit with time-frequency dictionaries. Technical report, Courant Institute of Mathematical Sciences New York United States, 1993.

Geoffrey Davis. *Adaptive nonlinear approximations*. PhD thesis, New York University, Graduate School of Arts and Science, 1994.

Yagyensh Chandra Pati, Ramin Rezaiifar, and Perinkulam Sambamurthy Krishnaprasad. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Signals, Systems and Computers, 1993. 1993 Conference Record of The Twenty-Seventh Asilomar Conference on*, pages 40–44. IEEE, 1993.

Geoff Davis, Stephane Mallat, and Marco Avellaneda. Adaptive greedy approximations. *Constructive approximation*, 13(1):57–98, 1997.

Ronald A DeVore and Vladimir N Temlyakov. Some remarks on greedy algorithms. *Advances in computational Mathematics*, 5(1):173–187, 1996.

Andrew R Barron, Albert Cohen, Wolfgang Dahmen, Ronald A DeVore, et al. Approximation and learning by greedy algorithms. *The annals of statistics*, 36(1):64–94, 2008.

Léon Bottou. Online learning and stochastic approximations. *On-line learning in neural networks*, 17(9):142, 1998.

Herbert Robbins and Sutton Monro. A stochastic approximation method. In *Herbert Robbins Selected Papers*, pages 102–109. Springer, 1985.

Martin Vetterli, Jelena Kovacevic, and Vivek K Goyal. *Foundations of Signal Processing*. Cambridge University Press, 2014. URL http://www.fourierandwavelets.org/FSP_v1.1_2014.pdf.

Michael Gastpar, Emre Telatar, and Ruediger Urbanke. EPFL, COM-406, Information Theory and Signal Processing. Lecture notes, 2018. URL https://ipg.epfl.ch/teaching/page-146889-en-html/page-146918-en-html/page-147664-en-html/.

Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

Ben Adcock and Daan Huybrechs. Frames and numerical approximation. *arXiv preprint arXiv:1612.04464*, 2016.

Kjersti Engan, Sven Ole Aase, and J Hakon Husoy. Method of optimal directions for frame design. In *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, volume 5, pages 2443–2446. IEEE, 1999.