Gradient Descent with Compressed Iterates

Ahmed Khaled*
Cairo University
akregeb@gmail.com

Peter Richtárik
KAUST†
peter.richtarik@kaust.edu.sa

Abstract

We propose and analyze a new type of stochastic first order method: gradient descent with compressed iterates (GDCI). GDCI in each iteration first compresses the current iterate using a lossy randomized compression technique, and subsequently takes a gradient step. This method is a distillation of a key ingredient in the current practice of federated learning, where a model needs to be compressed by a mobile device before it is sent back to a server for aggregation. Our analysis provides a step towards closing the gap between the theory and practice of federated learning, and opens the possibility for many extensions.

1   Introduction

Federated learning is a machine learning setting where the goal is to learn a centralized model given access only to local optimization procedures distributed over many devices [18, 22, 36]. This situation is common in large-scale distributed optimization involving many edge devices, and common challenges include data heterogeneity [37], privacy [7], resource management [25, 33], and system heterogeneity as well as communication efficiency [16, 18]. The most commonly used optimization methods in federated learning are variants of distributed gradient decent, stochastic gradient and gradient-based methods such as Federated Averaging [18].

The training of high-dimensional federated learning models [16, 17] reduces to solving an optimization problem of the form

$$x_* = \underset{x \in \mathbb{R}^d}{\operatorname{argmin}} \left[f(x) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} f_i(x) \right],$$

where n is the number of consumer devices (e.g., mobile devices), d is the number of parameters/features of the model, and $f_i : \mathbb{R}^d \to \mathbb{R}$ is a loss function that depends on the private data stored on the ith device. The simplest benchmark method[3] for solving this problem is gradient descent, which performs updates of the form

$$x_{k+1} = \frac{1}{n} \sum_{i=1}^{n} \left(x_k - \gamma \nabla f_i(x_k) \right).$$

That is, all nodes in parallel first perform a single gradient descent step starting from x_k based on their local data, the resulting models are then communicated to a central machine/aggregator, which performs model averaging. The average model is subsequently communicated back to all devices, and the process is repeated until a model of a suitable quality is found.

Practical considerations of federated learning impose several constraints on the feasibility of this process. First, due to geographical and other reasons, model averaging is performed in practice

---

*Work done during an internship at KAUST.

†King Abdullah University of Science and Technology, Thuwal, Saudi Arabia

[3]Which is a starting point for the development of more advanced methods.

on a subset of nodes at a time only. Second, in a hope to address the communication bottleneck, each device is typically allowed to take multiple steps of gradient descent or stochastic gradient descent before aggregation takes place. Methods of this type are known as *local* methods in the literature [16, 22]. Third, in the large dimensional case, the models are typically compressed [5, 16] by the devices before they are communicated to the aggregator, and/or by the aggregator before the averaged model is pushed to the devices.

In distributed stochastic gradient methods, the cost of gradient communication between training nodes and the master node or parameter server has been observed to be a significant performance bottleneck. As a result, there are many algorithms designed with the goal of reducing communication in stochastic gradient methods: including SignSGD (1-bit quantization) [4], TernGrad (ternary quantization) [35], QSGD [1], DIANA (with arbitrary quantization) [11], ChocoSGD [15], and others, see e.g. [3, 21] and the references therein. Among compression operators used in quantized distributed stochastic gradient methods, compression operators satisfying Assumption 2 are ubiquitous and include natural compression [9], dithering [8, 27], natural dithering [9], sparsification [29], ternary quantization [35], and others. As an alternative to costly parameter server communication, decentralized methods can achieve better communication efficiency by using inter-node communication. [3, 19] and combinations of decentralization and gradient quantization have been studied in recent work, see e.g. [14, 15, 30]. Another line of work focused on local stochastic gradient methods that communicate only intermittently and average models, such as Local SGD [20, 28, 31, 32] and Federated Averaging [22], and combinations of such methods and update quantization (where the sum of gradients over an epoch is quantized) have also been explored in the literature [2, 12].

Gaps in theory of federated learning. There are considerable gaps in our theoretical understanding of federated learning algorithms which use these tricks. For instance, until very recently [13], no convergence results were known for the simplest of all local methods—local gradient descent—in the case when the functions f_i are allowed to be arbitrarily different, which is a requirement of any efficient federated learning method since data stored on devices of different users can be arbitrarily heterogeneous. Further, while there is ample work on non-local methods which communicate compressed gradients [1, 3, 11, 34, 35], including methods which perform variance-reduction to remove the variance introduced by compression [10, 23, 24], to the best of our knowledge there is little work on methods performing iterative model compression, and the only one we are aware of is the very recent work in [26] which is a distributed variant of SGD that quantizes iterate communication. To remove the iterate quantization variance, they do a relaxation over time in the iterates and no results are provided when averaging across time is not performed. Similar statements can be made about our understanding of other elements of current practice.

Iterative model compression. In this paper we focus on a single element behind efficient federated learning methods—iterative model compression—and analyze it in isolation. Surprisingly, we are not aware of any theoretical results in this area, even in the simplest of settings: the case of a single device ($n = 1$) with a smooth and strongly convex function.

Motivated by the desire to take step towards bridging the gap between theory and practice of federated learning, in this paper we study the algorithm

$$x_{k+1} = \mathcal{C}(x_k) - \gamma \nabla f(\mathcal{C}(x_k)), \tag{1}$$

where $\mathcal{C} : \mathbb{R}^d \to \mathbb{R}^d$ is a sufficiently well behaved unbiased stochastic compression operator (see Assumption 2 for the definitions). We call this method *gradient descent with compressed iterates (GDCI)*. The update in equation (1) captures the use of compressed iterates/models in place of full iterates on a single node. Clearly, this method should be understood if we are to tackle the more complex realm of distributed optimization for federated learning, including the $n > 1$ setting, partial participation and local variants. We believe that our work will be starting point of healthy research into iterative methods with compressed iterates. One of the difficulties in analyzing this method is the observation that $\nabla f(\mathcal{C}(x))$ is not an unbiased estimator of the gradient, even if $\mathcal{C}$ is unbiased.

2 Assumptions and Contributions

In this work we assume that f is smooth and strongly convex:

Assumption 1. The function $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth and μ-strongly convex: that is, there exists $L \geqslant \mu > 0$ such that

$$\mu \|x - y\| \leqslant \|\nabla f(x) - \nabla f(y)\| \leqslant L \|x - y\|$$

for all $x, y \in \mathbb{R}^d$. We define the condition number of f as $\kappa \overset{\text{def}}{=} \frac{L}{\mu}$.

We make the following assumptions on the compression operator:

Assumption 2. The compression operator $\mathcal{C} : \mathbb{R}^d \to \mathbb{R}$ is unbiased, i.e.,

$$\mathbb{E}\left[\mathcal{C}(x) \mid x\right] = x, \qquad \forall x \in \mathbb{R}^d, \tag{2}$$

and there exists $\omega \geqslant 0$ such that its variance is bounded as follows

$$\mathbb{E}\left[\|\mathcal{C}(x) - x\|^2\right] \leqslant \omega \|x\|^2, \qquad \forall x \in \mathbb{R}^d. \tag{3}$$

Our main contribution is to show that the iterates generated by GDCI (Algorithm (1)) converge linearly, at the same rate as gradient descent, to a neighbourhood of the solution x_* of size $\mathcal{O}(\kappa\omega)$, where $\kappa = L/\mu$ is the condition number of the solution.

Theorem 1. Suppose the Assumptions 1 and 2 hold. Suppose that GDCI is run with a constant stepsize $\gamma > 0$ such that $\gamma \leqslant \frac{1}{2L}$ and assume that the compression coefficient $\omega \geqslant 0$ satisfies

$$\frac{4\omega}{\mu} \leqslant \frac{1 - 2\gamma L}{2\gamma L^2 + \frac{2}{\gamma} + L - \mu}. \tag{4}$$

Then,

$$\mathbb{E}\left[\|x_k - x_*\|^2\right] \leqslant (1 - \gamma\mu)^k \|x_0 - x_*\|^2 + \frac{2\omega}{\mu}\left(4\gamma L^2 + \frac{4}{\gamma} + L - \mu\right)\|x_*\|^2. \tag{5}$$

The proof of Theorem 1 is provided in the supplementary material. The following corollary gives added insight:

Corollary 1. In Theorem 1, suppose that $\gamma = \frac{1}{4L}$ and that $\omega \leqslant \frac{1}{73\kappa}$, then the bound in Equation (4) is satisfied and substituting in (5) we have,

$$\mathbb{E}\left[\|x_k - x_*\|^2\right] \leqslant \left(1 - \frac{1}{4\kappa}\right)^k \|x_0 - x_*\|^2 + 2\omega\left(18\kappa - 1\right)\|x_*\|^2.$$

This is the same rate as gradient descent, but only to a $\mathcal{O}(\kappa\omega)$ neighbourhood (in squared distances) of the solution.

Note that if we want to set the neighbourhood to $\mathcal{O}(1)$, then we should have $\omega = \mathcal{O}\left(\kappa^{-1}\right)$. While this seems to be a pessimistic bound on the compression level possible, we note that in practice compression is done only intermittently (this could be modelled by an appropriate choice of $\mathcal{C}$; more on this below) or in a combination with averaging (which naturally reduces the variance associated with quantization). In practical situations where averaging is not performed, such as the quantization of server-to-client communication, high compression levels do not seem possible without serious deterioration of the accuracy of the solution [5], and our experiments also suggest that this is the case.

3 Experiments

To confirm our theoretical results, we experiment with a logistic regression problem:

$$\min_{w}\left\{f(w) = \frac{1}{n}\sum_{i=1}^{n}\log(1 + \exp\left(-b_i x_i^\top w\right)) + \frac{\mu}{2}\|w\|^2\right\}, \tag{6}$$

where $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$ are the data samples for $i \in [n]$. We consider the "a7a" and "a5a" datasets from the UCI Machine Learning repository [6] with $n = 16100$ for "a7a" and $n = 6414$ for "a5a" and $d = 123$ in both cases. We set the regularization parameter $\mu = 0.02$ and estimate $\kappa \simeq 161$ for the "a7a" dataset and $\kappa \simeq 65$ for the "a5a" dataset. We consider the random sparsification operator, where each coordinate is independently set to zero according to some given probability. That is, given $p \in (0, 1]$ we have for $c : \mathbb{R} \to \mathbb{R}$,

$$c(x) = \begin{cases} \frac{x}{p} & \text{with probability } p \\ 0 & \text{with probability } 1 - p \end{cases} \tag{7}$$
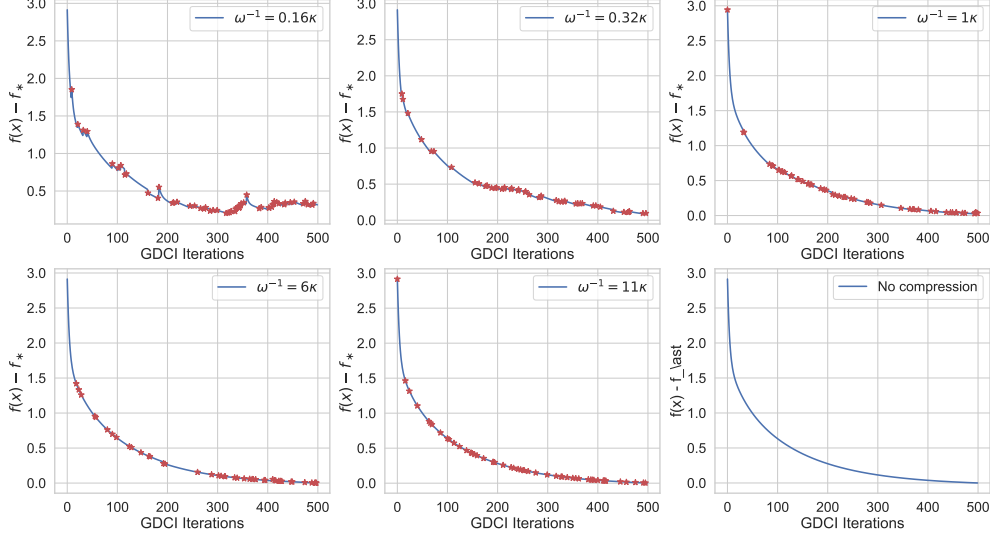
3

Figure 1: GDCI as ω varies for the "a7a" dataset. Red star indicates $\mathcal{C}$ was applied in that iteration.
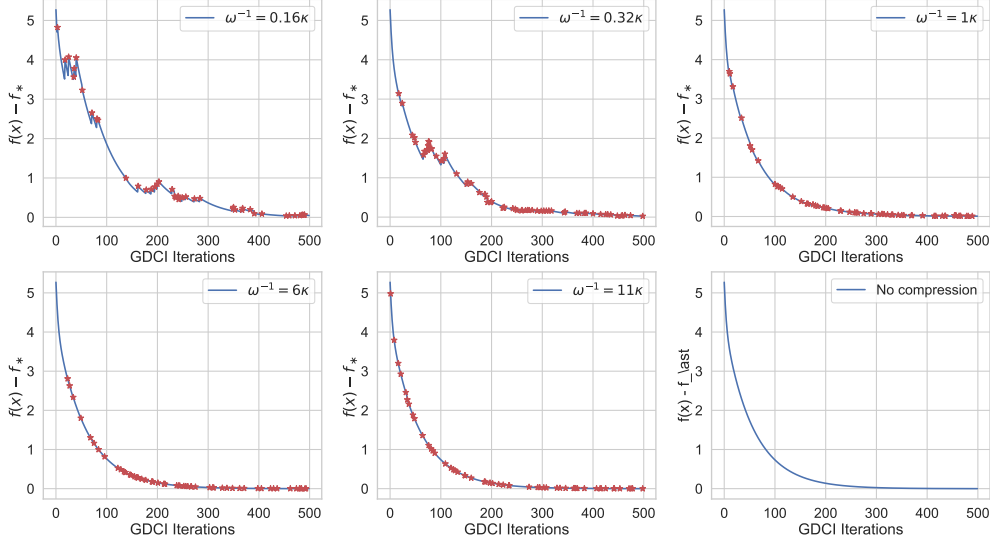

Figure 2: GDCI as ω varies for the "a5a" dataset. Red star indicates $\mathcal{C}$ was applied in that iteration.

and we define $\mathcal{C} : \mathbb{R}^d \rightarrow \mathbb{R}$ by $(\mathcal{C}(x))_i = c(x_i)$ for all $i \in [n]$ independently. Note that for this quantization operator $\mathcal{C}$ we have that Assumption 2 is satisfied with $\omega = 1 - p$.

To model intermittent quantization experimentally, we apply the quantization operator C with probability $1/10$ and keep the iterate as it is with probability $9/10$. We vary ω as $\frac{1}{\alpha\kappa}$ for various settings of α. The results are shown for the "a7a" dataset are shown in Figure 1 and for the "a5a" dataset in Figure 2.

The results of Figure 1 show that for ω small enough the effect on convergence is negligible, but the effect on the error at convergence becomes noticeable at $\omega \in \left\{ \frac{4}{25\kappa}, \frac{8}{25\kappa} \right\}$ and we have observed divergent behavior for larger values of ω. Similar behavior is observed for the plots in Figure 2.

References

[1] 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.

[2] Debraj Basu, Deepesh Data, Can Karakus, and Suhas Diggavi. Qsparse-local-SGD: Distributed SGD with Quantization, Sparsification, and Local Computations. *arXiv:1906.02367*, 2019.

[3] Tal Ben-Nun and Torsten Hoefler. Demystifying Parallel and Distributed Deep Learning: An In-Depth Concurrency Analysis. *arXiv:1802.09941*, 2018.

[4] Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Anima Anandkumar. signSGD: Compressed Optimisation for Non-Convex Problems. *arXiv:1802.04434*, 2018.

[5] Sebastian Caldas, Jakub Konečný, H. Brendan McMahan, and Ameet Talwalkar. Expanding the Reach of Federated Learning by Reducing Client Resource Requirements. *arXiv:1812.07210*, 2018.

[6] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

[7] Robin C. Geyer, Tassilo Klein, and Moin Nabi. Differentially Private Federated Learning: A Client Level Perspective. *arXiv:1712.07557*, 2017.

[8] W. M. Goodall. Television by Pulse Code Modulation*. *Bell System Technical Journal*, 30(1): 33–49, 1951.

[9] Samuel Horvath, Chen-Yu Ho, Ludovit Horvath, Atal Narayan Sahu, Marco Canini, and Peter Richtárik. Natural Compression for Distributed Deep Learning. *arxiv preprint arXiv:1905.10988*, 2019.

[10] Samuel Horváth, Dmitry Kovalev, Konstantin Mishchenko, Sebastian Stich, and Peter Richtárik. Stochastic distributed learning with gradient quantization and variance reduction. *arXiv*, 2019.

[11] Samuel Horváth, Dmitry Kovalev, Konstantin Mishchenko, Sebastian Stich, and Peter Richtárik. Stochastic Distributed Learning with Gradient Quantization and Variance Reduction. *arxiv preprint arXiv:1904.05115*, 2019.

[12] Peng Jiang and Gagan Agrawal. A Linear Speedup Analysis of Distributed Deep Learning with Sparse and Quantized Communication. *Advances in Neural Information Processing Systems 31*, (NeurIPS):2530–2541, 2018.

[13] Ahmed Khaled, Konstantin Mishchenko, and Peter Richtárik. First Analysis of Local GD on Heterogeneous Data. Technical report, KAUST, Saudi Arabia, 2019.

[14] Anastasia Koloskova, Tao Lin, Sebastian U. Stich, and Martin Jaggi. Decentralized Deep Learning with Arbitrary Communication Compression. *arxiv preprint arXiv:1907.09356*, 2019.

[15] Anastasia Koloskova, Sebastian U. Stich, and Martin Jaggi. Decentralized Stochastic Optimization and Gossip Algorithms with Compressed Communication. *arXiv:1902.00340*, 2019.

[16] Jakub Konečný, H. Brendan McMahan, Felix X. Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated Learning: Strategies for Improving Communication Efficiency. In *NIPS Private Multi-Party Machine Learning Workshop*, 2016.

[17] Jakub Konečný, H. Brendan McMahan, Daniel Ramage, and Peter Richtárik. Federated optimization: distributed machine learning for on-device intelligence. *arXiv:1610.02527*, 2016.

[18] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated Learning: Challenges, Methods, and Future Directions. *arxiv preprint arXiv:1908.07873*, 2019.

[19] Xiangru Lian, Ce Zhang, Huan Zhang, Cho-Jui Hsieh, Wei Zhang, and Ji Liu. Can Decentralized Algorithms Outperform Centralized Algorithms? A Case Study for Decentralized Parallel Stochastic Gradient Descent. *arXiv:1705.09056*, 2017.

[20] Tao Lin, Sebastian U. Stich, and Martin Jaggi. Don't Use Large Mini-Batches, Use Local SGD. *arXiv:1808.07217*, 2018.

[21] Yujun Lin, Song Han, Huizi Mao, Yu Wang, and William J. Dally. Deep Gradient Compression: Reducing the Communication Bandwidth for Distributed Training. *arXiv:1712.01887*, 2017.

[22] H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. *Proceedings of the 20 th International Conference on Artificial Intelligence and Statistics (AISTATS) 2017. JMLR: W&CP volume 54*, 2016.

[23] Konstantin Mishchenko, Eduard Gorbunov, Martin Takáč, and Peter Richtárik. Distributed learning with compressed gradient differences. *arXiv preprint arXiv:1901.09269*, 2019.

[24] Konstantin Mishchenko, Filip Hanzely, and Peter Richtárik. 99% of distributed optimization is a waste of time: The issue and how to fix it. *arXiv preprint arXiv:1901.09437*, 2019.

[25] T. Nishio and R. Yonetani. Client Selection for Federated Learning with Heterogeneous Resources in Mobile Edge. In *ICC 2019 - 2019 IEEE International Conference on Communications (ICC)*, pages 1–7, May 2019.

[26] Amirhossein Reisizadeh, Hossein Taheri, Aryan Mokhtari, Hamed Hassani, and Ramtin Pedarsani. Robust and Communication-Efficient Collaborative Learning. *arxiv preprint arXiv:1907.10595*, 2019.

[27] L. Roberts. Picture coding using pseudo-random noise. *IRE Transactions on Information Theory*, 8(2):145–154, February 1962.

[28] Sebastian U. Stich. Local SGD Converges Fast and Communicates Little. *arXiv:1805.09767*, 2018.

[29] Sebastian U. Stich, Jean-Baptiste Cordonnier, and Martin Jaggi. Sparsified SGD with Memory. *arXiv:1809.07599*, 2018.

[30] Hanlin Tang, Ce Zhang, Shaoduo Gan, Tong Zhang, and Ji Liu. Decentralization Meets Quantization. *arXiv:1803.06443*, 2018.

[31] Jianyu Wang and Gauri Joshi. Cooperative SGD: A Unified Framework for the Design and Analysis of Communication-Efficient SGD Algorithms. *arXiv:1808.07576*, 2018.

[32] Jianyu Wang and Gauri Joshi. Adaptive Communication Strategies to Achieve the Best Error-Runtime Trade-off in Local-Update SGD. *arxiv preprint arXiv:1810.08313*, 2018.

[33] Shiqiang Wang, Tiffany Tuor, Theodoros Salonidis, Kin K. Leung, Christian Makaya, Ting He, and Kevin Chan. When Edge Meets Learning: Adaptive Control for Resource-Constrained Distributed Machine Learning. *arXiv:1804.05271*, 2018.

[34] Jianqiao Wangni, Jialei Wang, Ji Liu, and Tong Zhang. Gradient Sparsification for Communication-Efficient Distributed Optimization. *arXiv:1710.09854*, 2017.

[35] 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. *arXiv:1705.07878*, 2017.

[36] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. Federated Machine Learning: Concept and Applications. *arXiv:1902.04885*, 2019.

[37] Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. Federated Learning with Non-IID Data. *arxiv preprint arXiv:1806.00582*, 2018.

Gradient Descent with Compressed Iterates Supplementary Material

4   Basic Inequalities

We will often use the bound

$$\|a + b\|^2 \leqslant 2\|a\|^2 + 2\|b\|^2.\tag{8}$$

If f is an L–smooth and convex function, then the following inequalities hold

$$\|\nabla f(x) - \nabla f(y)\| \leqslant L\|x - y\|,\tag{9}$$

$$f(x) \leqslant f(y) + \langle \nabla f(y), x - y \rangle + \frac{L}{2}\|x - y\|^2,\tag{10}$$

$$f(y) + \langle \nabla f(y), x - y \rangle + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|^2 \leqslant f(x).\tag{11}$$

If f is μ-strongly convex, then the following inequality holds

$$f(y) \geqslant f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|y - x\|^2, \quad \forall x, y \in \mathbb{R}^d.\tag{12}$$

We define $\delta(x) \stackrel{\text{def}}{=} \mathcal{C}(x) - x$.

5   Five Lemmas

In the first lemma we give an upper bound on the variance of the compression operator $\mathcal{C}$.

Lemma 1. Suppose that a compression operator $\mathcal{C} : \mathbb{R}^d \to \mathbb{R}^d$ satisfies Assumption 2, then

$$\mathbb{E}\left[\|\mathcal{C}(x) - x\|^2\right] \leqslant 2\alpha\left(f(x) - f(x_*)\right) + \beta,\tag{13}$$

with $\alpha = \frac{2\omega}{\mu}$ and $\beta = 2\omega\|x_*\|^2$.

Proof. First, note that $\|x\|^2 \leqslant 2\|x - x_*\|^2 + 2\|x_*\|^2$. If f is μ-strongly convex, then by (12) we have that $\|x - x_*\|^2 \leqslant \frac{2}{\mu}\left(f(x) - f(x_*)\right)$, and putting these inequalities together, we arrive at

$$\mathbb{E}\left[\|C(x) - x\|^2\right] \leqslant \omega\|x\|^2 \leqslant 2\omega\|x - x_*\|^2 + 2\omega\|x_*\|^2 \leqslant \frac{4\omega}{\mu}\left(f(x) - f(x_*) + 2\omega\|x_*\|^2\right).$$

∎

Our second lemma is an extension of several standard inequalities which trivially hold (for L-smooth and convex functions) in the case of no compression, i.e., $\delta(x) \equiv 0$, to a situation where a compression is applied. Indeed, notice that (14) is a generalization of (9), and the second inequality in (15) is a generalization of (10).

Lemma 2. If the compression operator $\mathcal{C}$ satisfies (2) and f is convex and L-smooth, then

$$\mathbb{E}\left[\|\nabla f(x + \delta(x)) - \nabla f(y)\|^2\right] \leqslant L^2\left(\|x - y\|^2 + \mathbb{E}\left[\|\delta(x)\|^2\right]\right), \quad \forall x, y \in \mathbb{R}^d.\tag{14}$$

And for all $x, y \in \mathbb{R}^d$ we also have,

$$f(x) \leqslant \mathbb{E}\left[f(x + \delta(x))\right] \leqslant f(y) + \langle \nabla f(y), x - y \rangle + \frac{L}{2}\|x - y\|^2 + \frac{L}{2}\mathbb{E}\left[\|\delta(x)\|^2\right].\tag{15}$$

Proof. Fix x and let $\delta = \delta(x)$. Inequality (14) follows from Lipschitz continuity of the gradient, applying expectation and using (2):

$$\mathbb{E}\left[\|\nabla f(x + \delta) - \nabla f(y)\|^2\right] \leqslant L^2\mathbb{E}\left[\|x + \delta - y\|^2\right] \stackrel{(2)}{=} L^2\left(\|x - y\|^2 + \mathbb{E}\left[\|\delta\|^2\right]\right).$$

The first inequality in (15) follows by applying Jensen's inequality and using (2). Since f is L–smooth, we have

$$
\begin{aligned}
\mathbb{E}\left[f(x+\delta)\right] &\leqslant \mathbb{E}\left[f(y)+\langle\nabla f(y),x+\delta-y\rangle+\frac{L}{2}\left\|x+\delta-y\right\|^2\right] \\
&\overset{(2)}{=} f(y)+\langle\nabla f(y),x-y\rangle+\frac{L}{2}\left\|x-y\right\|^2+\frac{L}{2}\mathbb{E}\left[\left\|\delta\right\|^2\right].
\end{aligned}
$$

∎

Lemma 3. If the compression operator $\mathcal{C}$ satisfies (2), then for all $x,y\in\mathbb{R}^d$

$$
\mathbb{E}\left[\left\|\frac{\delta(x)}{\gamma}-\nabla f(x+\delta(x))\right\|^2\right]\leqslant 2\left\|\nabla f(y)\right\|^2+2L^2\left\|x-y\right\|^2+2\left(L^2+\frac{1}{\gamma^2}\right)\mathbb{E}\left[\left\|\delta(x)\right\|^2\right].
\tag{16}
$$

Proof. Fix x, and let $\delta=\delta(x)$. Then for every $y\in\mathbb{R}^d$ we can write

$$
\begin{aligned}
\mathbb{E}\left[\left\|\frac{\delta}{\gamma}-\nabla f(x+\delta)\right\|^2\right] &= \mathbb{E}\left[\left\|\frac{\delta}{\gamma}-\nabla f(y)+\nabla f(y)-\nabla f(x+\delta)\right\|^2\right] \\
&\overset{(8)}{\leqslant} 2\mathbb{E}\left[\left\|\frac{\delta}{\gamma}-\nabla f(y)\right\|^2\right]+2\mathbb{E}\left[\left\|\nabla f(y)-\nabla f(x+\delta)\right\|^2\right] \\
&\overset{(14)}{\leqslant} 2\mathbb{E}\left[\frac{1}{\gamma^2}\left\|\delta\right\|^2-\frac{1}{\gamma}\langle\delta,\nabla f(y)\rangle+\left\|\nabla f(y)\right\|^2\right] \\
&\qquad +2L^2\left(\left\|x-y\right\|^2+\mathbb{E}\left[\left\|\delta\right\|^2\right]\right) \\
&\overset{(2)}{\leqslant} \frac{2}{\gamma^2}\mathbb{E}\left[\left\|\delta\right\|^2\right]+2\left\|\nabla f(y)\right\|^2 \\
&\qquad +2L^2\left(\left\|x-y\right\|^2+\mathbb{E}\left[\left\|\delta\right\|^2\right]\right).
\end{aligned}
$$

(17)

(18)

∎

The next lemma generalizes the strong convexity inequality (12). Indeed, (12) is recovered in the special case $\delta(x)\equiv 0$.

Lemma 4. Suppose Assumptions 2 and 1 hold. Then for all $x,y\in\mathbb{R}^d$,

$$
f(y)\geqslant f(x)+\langle\mathbb{E}\left[\nabla f(x+\delta)\right],y-x\rangle+\frac{\mu}{2}\left\|y-x\right\|^2-\frac{L-\mu}{2}\mathbb{E}\left[\left\|\delta(x)\right\|^2\right].
\tag{19}
$$

Proof. Fix x and let $\delta=\delta(x)$. Using (12) with $x\leftarrow x+\delta$, we get

$$
f(y)\geqslant f(x+\delta)+\langle\nabla f(x+\delta),y-x-\delta\rangle+\frac{\mu}{2}\left\|y-x-\delta\right\|^2.
$$

Applying expectation, we get

$$
\begin{aligned}
f(y) &\geqslant \mathbb{E}\left[f(x+\delta)\right]+\mathbb{E}\left[\langle\nabla f(x+\delta),y-x\rangle\right]-\mathbb{E}\left[\langle\nabla f(x+\delta),\delta\rangle\right] \\
&\qquad +\frac{\mu}{2}\left\|y-x\right\|^2+\frac{\mu}{2}\mathbb{E}\left[\left\|\delta\right\|^2\right].
\end{aligned}
\tag{20}
$$

The term $-\mathbb{E}\left[\langle\nabla f(x+\delta),\delta\rangle\right]$ can be estimated using L–smoothness and applying expectation as follows:

$$
\begin{aligned}
\mathbb{E}\left[-\langle\nabla f(x+\delta),\delta\rangle\right] &\geqslant \mathbb{E}\left[f(x)-f(x+\delta)-\frac{L}{2}\left\|\delta\right\|^2\right] \\
&= f(x)-\mathbb{E}\left[f(x+\delta)\right]-\frac{L}{2}\mathbb{E}\left[\left\|\delta\right\|^2\right].
\end{aligned}
$$

It remains to plug this inequality to (20).

∎

Lemma 5. Suppose that Assumptions 1 and 2 hold. Then

$$\mathbb{E}\left[\left\|\frac{\delta(x)}{\gamma} - \nabla f(x + \delta(x))\right\|^2\right] \leqslant 4A(f(x) - f(x_*)) + 2B, \quad \forall x \in \mathbb{R}^d, \qquad (21)$$

where $A = L + \left(L^2 + \frac{1}{\gamma^2}\right)\alpha$ and $B = 2\left(L^2 + \frac{1}{\gamma^2}\right)\beta$ and α, β are defined in Lemma 1.

Proof. Using (16) with $y = x$, we get

$$
\begin{aligned}
\mathbb{E}\left[\left\|\frac{\delta(x)}{\gamma} - \nabla f(x + \delta(x))\right\|^2\right] \quad &\overset{(16)}{\leqslant} \quad 2\|\nabla f(x)\|^2 + 2\left(L^2 + \frac{1}{\gamma^2}\right)\mathbb{E}\left[\|\delta(x)\|^2\right] \\
&\overset{(11)+(13)}{\leqslant} \quad 4L(f(x) - f(x_*)) + 2\left(L^2 + \frac{1}{\gamma^2}\right)(2\alpha(f(x) - f(x_*)) + \beta) \\
&= \quad 4\left(L + \left(L^2 + \frac{1}{\gamma^2}\right)\alpha\right)(f(x_k) - f(x_*)) + 2\left(L^2 + \frac{1}{\gamma^2}\right)\beta.
\end{aligned}
$$

∎

6 Proof of Theorem 1

Proof. Let $r_k = \|x_k - x_*\|^2$, $\delta_k = \delta(x_k)$ (hence $\mathcal{C}(x_k) = x_k + \delta_k$). Then

$$
\begin{aligned}
r_{k+1} &= \|\mathcal{C}(x_k) - \gamma\nabla f(\mathcal{C}(x_k)) - x_*\|^2 \\
&= \|x_k - x_* + \delta_k - \gamma\nabla f(x_k + \delta_k)\|^2 \\
&= r_k + 2\langle\delta_k - \gamma\nabla f(x_k + \delta_k), x_k - x_*\rangle + \|\delta_k - \gamma\nabla f(x_k + \delta_k)\|^2.
\end{aligned}
$$

Taking conditional expectation, we get

$$
\begin{aligned}
\mathbb{E}\left[r_{k+1} \mid x_k\right] &= r_k + 2\gamma\langle\mathbb{E}\left[\nabla f(x_k + \delta_k) \mid x_k\right], x_* - x_k\rangle + \mathbb{E}\left[\|\delta_k - \gamma\nabla f(x_k + \delta_k)\|^2 \mid x_k\right] \\
&\overset{(19)}{\leqslant} r_k + 2\gamma\left[f(x_*) - f(x_k) - \frac{\mu}{2}\|x_k - x_*\|^2 + \frac{L - \mu}{2}\mathbb{E}\left[\|\delta_k\|^2 \mid x_k\right]\right] \\
&\quad + \gamma^2\mathbb{E}\left[\left\|\frac{\delta_k}{\gamma} - \nabla f(x_k + \delta_k)\right\|^2 \mid x_k\right] \\
&= (1 - \gamma\mu)r_k - 2\gamma(f(x_k) - f(x_*)) + \gamma(L - \mu)\mathbb{E}\left[\|\delta_k\|^2 \mid x_k\right] \\
&\quad + \gamma^2\mathbb{E}\left[\left\|\frac{\delta_k}{\gamma} - \nabla f(x_k + \delta_k)\right\|^2 \mid x_k\right] \\
&\overset{(21)}{\leqslant} (1 - \gamma\mu)r_k - 2\gamma(f(x_k) - f(x_*)) + \gamma(L - \mu)\mathbb{E}\left[\|\delta_k\|^2 \mid x_k\right] \\
&\quad + 4\gamma^2 A(f(x_k) - f(x_*)) + 2\gamma^2 B \\
&= (1 - \gamma\mu)r_k + \gamma(4\gamma A - 2)(f(x_k) - f(x_*)) + 2\gamma^2 B + \gamma(L - \mu)\mathbb{E}\left[\|\delta_k\|^2 \mid x_k\right] \\
&\overset{(13)}{\leqslant} (1 - \gamma\mu)r_k + \gamma(4\gamma A - 2)(f(x_k) - f(x_*)) + 2\gamma^2 B \\
&\quad + \gamma(L - \mu)(2\alpha(f(x_k) - f(x_*)) + \beta) \\
&= (1 - \gamma\mu)r_k + 2\gamma(2\gamma A + \alpha(L - \mu) - 1)(f(x_k) - f(x_*)) + 2\gamma^2 B + \gamma(L - \mu)\beta,
\end{aligned}
$$

where α and β are as in Lemma 1 and A and B are defined Lemma 5. By assumption on α and γ, we have $2\gamma A + \alpha(L - \mu) \leqslant 1$, and hence $\mathbb{E}\left[r_{k+1} \mid x_k\right] \leqslant (1 - \gamma\mu)r_k + D$, where $D = 2\gamma^2 B + \gamma(L - \mu)\beta$. Taking expectation, unrolling the recurrence, and applying the tower property, we get

$$\mathbb{E}\left[r_k\right] \leqslant (1 - \gamma\mu)^k r_0 + \frac{D}{\gamma\mu}.$$

9

Writing out D yields the expression for the convergence rate in (5). For the bound on ω, we first write out the definition of A in $2\gamma A + \alpha(L - \mu) \leqslant 1$ we have,

$$2\gamma \left(L + \left(L^2 + \frac{1}{\gamma^2} \right) \alpha \right) + \alpha \left(L - \mu \right) \leqslant 1. \tag{22}$$

Rearranging terms in (22) we get that,

$$\alpha \leqslant \frac{1 - 2\gamma L}{2\gamma L^2 + \frac{2}{\gamma} + L - \mu}.$$

Using the fact that $\omega = \frac{\alpha\mu}{2}$ yields (4). ∎