INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2023IPPAX083

Thèse de doctorat

Bidirectional compression for federated learning in heterogeneous setting

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à l'École polytechnique

École doctorale n°626 École doctorale de l'Institut Polytechnique de Paris (EDIPP)
Spécialité de doctorat: Mathématiques et Informatique

Thèse présentée et soutenue à Palaiseau, le 18 septembre 2023, par

CONSTANTIN PHILIPPENKO

Composition du Jury :

Mikael Johansson Professeur, KTH Royal Institute of Technology	Rapporteur
Jérôme Malick Chercheur CNRS, Université Grenoble Alpes, LJK	Rapporteur
Manon Costa Professeur assistant, Institut de Mathématiques de Toulouse	Examinateur
Robert Gower Research scientist, Flatiron Institute	Examinateur
Martin Jaggi Professeur, EPFL	Examinateur
Kevin Scaman Research scientist, Inria Paris	Examinateur
Aymeric Dieuleveut Professeur assistant, École Polytechnique (UMR 7641)	Directeur de thèse

Invités:

Eric Moulines Professeur, École Polytechnique (UMR 7641)	Co-directeur de thèse
Laetitia Kameni IT R&D Lead, Accenture Labs, Sophia-Antipolis	

Remerciements

Au cours de ces trois années et demi de doctorat, j'ai eu la chance d'avoir deux excellents directeurs de thèse, Aymeric Dieuleveut et Eric Moulines.

Cher Aymeric, merci beaucoup pour cette étroite collaboration que nous avons eue tout au long de mes années de thèse et au cours desquelles j'ai énormément appris et progessé. Tu as été un superviseur remarquable : brillant, enthousiaste, exigent, pédagogue, patient, bienveillant, humain. Quand je me rappelle notre première rencontre à l'Indiana Café de Denfert-Rochereau, j'étais loin de m'imaginer l'aventure dans laquelle je m'engageais, ni les pas de géant que tu me ferais faire à tes côtés. Tu es toujours en retard à nos réunions et en train de déborder de trois ou quatre heures les horaires de fin, le tableau ou la feuille numérique sur lesquels tu écris sont toujours illisibles après dix minutes de travail avec des couleurs en bouquet désordonné mais – tonnerre ! – que c'est formateur, passionnant et stimulant !

Cher Eric, tu as été le déclencheur qui m'a permis d'obtenir cette thèse et les financements, c'est aussi toi qui m'a mis en contact avec Aymeric. Merci de t'être trouvé sur ma route un soir de février pour m'écouter exposer ma motivation à faire une thèse et pour m'avoir fait confiance, je t'en suis très reconnaissant. Ce fût un honneur d'être co-encadré par toi. Au cours de ces trois années au CMAP ou au Lagrange, ce furent des discussions à n'en pas terminer avec toi au cours desquels j'ai été impressionné par ton érudition scientifique et ta connaissance des rouages administratifs – toutes les deux indispensable aux succès des travaux de recherche. Merci.

Sans l'implication d'Accenture France, et en particulier sans les efforts de Laetitia Kameni et de Richard Vidal, cette thèse n'aurait pas existé. Chers Laetitia et Richard, je vous suis reconnaissant pour avoir permis de concrétiser mon projet de thèse. Vous m'avez suivi sur ces trois années et toujours encouragé, vous avez complètement joué le jeu même si vous ne voyiez pas toujours comment utiliser mes travaux pour votre équipe, et vous m'avez apporté une précieuse ouverture au monde de l'industrie. J'ai énormément apprécié le format de notre collaboration.

I must thank Mikael Johansson and Jérôme Malick for reviewing my manuscript. It is a real honour for me. For agreeing to be part of my Ph.D.'s defense jury, I would like also to express my sincere gratitude to Manon Costa, Robert Gower, Martin Jaggi and Kevin Kevin Scaman. Thank you for your time and your consideration. However, I must make a special mention of Jérôme Malick and Robert Gower, who were members of my Follow-up Committee and were asked to ensure that my thesis was on the right track; it was always enriching discussions for me.

Il est impossible de continuer mes remerciements sans mentionner les autres personnes sans qui cette thèse aurait été impossible. Je parle bien sûr de l'équipe administrative du CMAP et en particulier de Nasséra Naar, Alexandra Noiret, Alexandra Liot, Georgia Sant'Anna ... mais pas seulement ! L'équipe informatique du CMAP nous est tout autant indispensable ; leur bureau est toujours ouvert, ils sont ultra réactifs à nos mail, jamais en panne d'idées face à nos problèmes et toujours bienveillants, il s'agit bien sûr de Pierre Straebler et de Sylvain Ferrand.

À présent, j'aimerai mentionner les nombreux professeurs du CMAP membres de l'équipe SIMPAS avec qui j'ai eu souvent l'occasion de discuter : Josselin Garnier, Marylou Gabrié, Emmanuel Gobet, Rémi Flamary, El Mahdi El Mhamdi, Stephano de Marco, Alain Durmus, Erwan Le Pennec, Erwan Scornet. Mahdi, merci pour toutes ces discussions incroyables sur l'éthique de l'IA, ton enthousiasme, ton exigence, ta bienveillance sont contagieux ; je me souviens en particulier d'un débat que nous avons eu en février dans l'eau froide de la Méditerranée ! Erwan, merci pour tes retours toujours très encourageants sur le cours de MAP545 et sur le plan de mon introduction de thèse. Marylou, merci de dynamiser l'équipe SIMPAS en organisant les réunions mensuelles. Et je dois justement terminer avec le grand architecte de notre équipe : Marine Saux !

En commançant ma thèse, j'ai été le premier doctorant d'Aymeric, et aussi le seul à travailler avec lui pendant près d'un an. Cela a des avantages, par exemple nous avons pu énormément travailler ensemble au cours de cette période, favorisé par le premier confinement au cours duquel nous nous appelions très réguièrement. Mais c'est vrai que l'ambiance d'équipe me manquait, heureusement Baptiste Goujaud, Margaux Zaffran et Alexis Ayme sont vite arrivés pour mettre de l'ambiance dans la "Dieuleveut Team". Merci à vous trois pour votre soutien, votre aide, votre enthousiasme et pour toutes ces discussions au CMAP, au Turing, à Jussieu, au CIRM, à l'IHP, à Font-Romeu, à Hyères ! Baptiste, ce sont des discussions mathématiques à n'en pas terminer avec toi qui, lorsqu'on y rentre, nous happent dans un trou spatio-temporel. Merci pour ton regard critique et tes retours constructifs, en particulier sur mon introduction de thèse ; et surtout pour ton aide sur la preuve de la covariance pour laquelle je m'arrachais les cheveux ! Margaux, toujours de bonne humeur, toujours pleine d'énergie, toujours prête à rendre service ! Impossible de lister toutes les fois où tu proposes et donnes un coup de main, je vais donc me contenter de remercier ta relecture extrêment consciencieuse, méticuleuse et rigoureuse de mon introduction de thèse et ton aide pour répéter ma soutenance de doctorat, merci ! Enfin, je souhaite bon courage à ceux qui viennent d'arriver dans l'équipe : Renaud Gaucher, Damien Ferbach, Rémi Leluc et Mahmoud Hegazy.

À un moment où j'étais un peu en panne d'inspiration sur mon dernier chapitre de thèse et avec un sentiment de solitude, j'ai pu faire une sorte de récréation mathématique en collaborant sur un projet orchestré par Jean du Terrail et Mathieu Andreux. Cela m'a permis d'explorer un nouvel axe de recherche, de travailler en équipe et de collaborer avec Aurélien Bellet, Paul Mangold et Edwige Cyffers. Merci pour nos riches réflexions sur les formes d'hétérogénéités possibles.

Je vais bientôt entamer une nouvelle aventure à l'Inria pour prospecter une autre zone de ce champs de recherche qu'est l'optimisation. C'est une chance pour moi de pouvoir collaborer avec Laurent Massoulié et Kevin Scaman ; j'ai hâte d'explorer de nouveaux horizons mathématiques avec vous !

Au cours de cette thèse, j'ai eu la chance de participer à une aventure assez incroyable qui m'a transformée, il s'agit du *Congrès des Jeunes Chercheurs en Mathématiques Appliquées* organisé en octobre 2021 à l'Ecole polytechnique par une bande de doctorants très motivés : Thomas Belloti, Guillaume Bonnet, Apolline Louvet, Claire Ecotiere, Corentin Houpert, Baptiste Kerleguer, Pierre Lavigne, Clément Mantoux, Solange Pruilh, Louis Reboul, Dominik Stantejsky, Josué Tchouanti Fotso. Merci à vous d'avoir rendu possible ce congrès. Cela n'a pas été toujours facile, nous nous sommes écharpés par moments, mais nous avons tenu bon en gardant la boussole pointée sur nos objectifs. J'ai beaucoup appris à votre contact et je me suis découvert un nouvel interêt pour les structures associatives. Mais il faut signaler que nous étions épaulés par un efficace comité scientifique que je tiens à remercier également : Matthieu Aussal, Juliette Chevallier, Fedor Goncharov, Baptiste Kerleguer, Pierre Lavigne, Aude Sportisse, Amandine Véber. À nouveau, je dois faire une mention spéciale pour Matthieu qui nous as chuchoté l'idée d'organiser ce congrès pour les jeunes par les jeunes et qui nous a fait profiter de son expérience dans le domaine, c'était dans la salle à café du CMAP, nous sortions du confinement et nous avions faim d'interactions sociales ! Je n'oublie pas Oliver Goubet, le Président de la SMAI qui a accepté de soutenir notre projet et qui nous a fait confiance. Pour concrétiser cette conférence, nous avions évidemment besoin de l'accord du Président du CMAP, Thierry Bodineau ; merci pour ta bienveillance.

Pendant presque un an, un an de confinement, le bureau 2010 était passablement vide, et je me retrouvais seul à chaque fois que j'y allais. À la longue, c'était "longuet". Mais un matin, lorsque je suis arrivé dans le bureau, Manon était là. J'ai essayé de la mettre à l'aise comme j'ai pu ... mais je crois qu'elle m'a pris pour un ours. Par la suite, elle m'a avoué être allée vérifier auprès des bureaux voisins que je n'étais pas un fou furieux. Au prix d'un grand effort sur elle-même, elle a finalement réussi à s'adapter à mon humeur décalée et ma bonhommie. À la fin, entre deux séances de piscine, de théâtre ou de boxe, nous étions même devenus les psychologues attitrés l'un de l'autre ! Manon, je te dédicace un merci spécial pour avoir accepté de suivre mes "cours de boxe" !

Une thèse est une succession de hauts et de bas, mais fort heureusement, on peut toujours compter sur les doctorants du CMAP pour égayer nos idées entre deux confinements, deux preuves coriaces ou deux rer en panne par des discussions stratosphériques, des jeux de sociétés endiablés ou des clubs de lectures méditatifs : Mehdi Abou El Quassime, Naoufal Acharki, Leila Bassou, Dorinel Bastide, Thomas Bellotti, Guillaume Bonnet, Wassim Bouaziz, Guillaume Chennetier, Amin Dhaou, Clément Mantoux, Pierre Clavier, Benedicte Colnet, Claire Ecotière, Celia Escribe, Orso Forghieri, Armand Gissler, Louis Grenioux, Corentin Houpert, Tom Huix, Pablo Jimenez, Baptiste Kerleguer, Madeleine Kubasch, Paul Lartaud, Pierre Lavigne, Jessie Levillain, Kang Liu, Arthur Loison, Apolline Louvet, Ignacio Madrid-Canales, Maxence Noble, Gregoire Pacreau, Vincent Plassier, Solange Pruilh, Louis Reboul, Benjamin Riu, Emmanouil Sfendourakis, Dominik Stantejsky, Josué Tchouanti, Achille Thin, Antoine Van-Biesbroeck, Gabriel Victorino-Cardoso, Wanqing Wang.

Maroc 2022 m'a laissé un souvenir particulièrement fort. J'y ai vécu une expérience unique en compagnie de Vincent Plassier, Pablo Jimenez, Maxence Noble et Benjamin Riu. Je me souviens en particulier d'un restaurant sur les terrasses de Marrakech face à la grande place et savourant du chameau confit au citron, d'un retour en taxi où nous avons eu peur de nous faire détrousser et laisser sur la route de la Mort au milieu du désert, d'un bain turc dans la grande mosquée de Casablanca et d'une visite guidée de sa Médina avec Akram Benazzou. Merci à vous tous !

À tous mes professeurs qui ont façonné ma trajectoire intellectuelle en m'initiant à la quête du savoir ou en m'aidant à en passer les différentes étapes : Madame Hijazi, Messieurs Cailhol, Leborgne, Dusuel, Sollogoub, Meignen, Gaudoin, Strijov.

À mes amis des *JPO* qui permettent de faire vivre cette belle aventure : Antoine Doucet, Dimitri Goudkoff, Matthieu Jurconi, Pierre Rehbinder, Soline Renard, Dimitri Sollogoub et P. Serge Ciolkovitch ! À chacun de nos évènements, c'est une vraie joie de voir les fruits de notre initiative ... et une bouffé d'oxygène d'en profiter !

Дорогие Тётя Маша, Дядя Дима, Аня, Серёжа, Анна и Лиза! Знали ли вы, что именно в Москве мне открылась область машинного обучения? Эта диссертация, можно сказать, частично благодаря вам. Я провёл замечательные 6 месяцев в Москве, и в моей памяти остались яркие и тёплые воспоминания. Спасибо вам за всё!

Aux Doucet, ces fidèles amis générations après générations : Christian, Catherine, Paul, Géraud, Antoine, Jean, Nicolas, Marie, Catherine, et à présent aussi Lionel.

Au soutien fraternel de ma grande et bruyante famille : Katia, David, Vika, Gilbert, Natacha, Christian, Nastia, Omar, Artémis, Adam, Marie, Marie, Sacha, Roman, Vera, Vassiliok, Daniil, Tania, Ivan, Mélania, Sandra, Petia, Théophane, Lara, Alex, Nina, Séraphim, Anna, Maria, Lena, Vincent, Varvara, Kolia, Ambroise, Oxana, Kyp, Silouan, Dorothée, Iovan, Elie, Nicolas, Petia, Alexandra, Grigou, Antou, Marina et Sachok. Je vous aime et j'ai besoin de vous.

À tous mes frères et sœurs Xénia, Anne, Elisabeth, Matthieu, Daniel et Clément. Sans le constant soutien de ma fratrie pour me donner du pétillement et du rebond, je ne pas sais ce que je ferais. Vous êtes une des plus grandes richesses de ma vie.

Un mot spécial pour ma grande soeur Xioucha qui a toujours été là pour me soutenir et m'encourager à viser plus haut. Elle a ouvert la voie et je m'y suis engouffré à sa suite. C'est toi, qui en commençant une thèse, a planté une petite graine dans mon esprit, qui en grandissant, m'a montré qu'il fallait que je t'emboîte le pas.

Дорогие мои бабушки и дедушки! Хочу выразить вам глубокую благодарность! Без вас я бы не смог стать таким, каким я есть. Вы передали мне - несмотря на все те трудности, в которые вас бросили исторические обстоятельсва - корни, культуру, ценности, стремление к самосовершенствованию и интеллектуальное любопытство.

Avant de finir, un mot pour mes parents. C'est vous qui avaient été les forgerons de ma vie par ce que vous m'avez transmis. Vous avez été tout les deux des exemples pour moi, et vous l'êtes toujours. J'admire votre maximalisme, vos efforts, vos choix et j'essaye de les imiter. Au cours de mes études, quand c'était difficile, quand il fallait s'accrocher, quand il fallait travailler et se donner du mal, c'est votre exigence et votre parcours intellectuel qui m'a donné de la force pour persévérer et croire que, moi aussi, je pouvais parvenir à avoir la même exigence que vous. Pour cet exemple que vous avez été chaque jour de ma vie et que vous serez encore, pour les valeurs que vous m'avez transmises, pour l'identité que vous m'avez léguée, pour le goût de l'effort que vous m'avez fait cultiver, je vous suis infiniment reconnaissant.

À Hedwige, qui m'a donné sa lumière quand j'étais dans l'obscurité.

Слава в вышних Богу, и на земле мир, в человеках благоволение!

Abstract

The last two decades have witnessed an unprecedented increase in computational power, leading to a vast surge in the volume of available data. Datasets can include billions of observations, models can involve millions of parameters. As a consequence, machine learning algorithms have evolved to adapt to this new situation. Especially, stochastic algorithms, which were first introduced in the 1950s, have recently received renewed attention due to their ability to handle large-scale datasets with millions of parameters. These algorithms use first-order information to alleviate the computational cost associated with high-dimensional data. Additionally, they efficiently process a large number of observations by computing stochastic gradients. These methods are key to the remarkable progress in machine learning over the last two decades.

However, many modern applications now use a network of clients to store the data and compute the models: efficient learning in this framework is harder, especially under communication constraints. This is why, a new approach has been proposed, *federated learning*, which considers a *distributed* setting: the data is kept on the original server and a central server orchestrates the training process across multiple clients.

This thesis aims to address two fundamental aspects of federated learning. The first goal is to analyze the trade-offs of distributed learning with communication constraints, with the objective of reducing its energy cost and environmental footprint. The second goal is to tackle problems resulting from heterogeneity among clients, which hinders the convergence toward an optimal solution. This thesis focuses on bidirectional compression and summarizes my contributions to this field of research.

In our first contribution, we focus on the intertwined effect of compression and client (statistical) heterogeneity. We introduce a framework of algorithms, named `Artemis`, that tackles the problem of learning in a federated setting with communication constraints. To alleviate the communication cost, `Artemis` enables to compress the information sent in both directions (from the clients to the server and conversely) combined with a memory mechanism. We highlight the key impact of memory on convergence in the heterogeneous setting.

In our second contribution, we move the focus toward feedback loops to reduce the impact of compression. We introduce an algorithm, coined `MCM`; it builds upon `Artemis` and introduces a new paradigm that preserves the central model from down compression. This mechanism allows to carry out bidirectional compression while asymptotically achieving the rates of convergence of unidirectional compression.

In our third contribution, we go beyond the classical worst-case assumption on the variance of compressors and provide a fine-grained analysis of the impact of compression within the fundamental learning framework of least-squares regression. Within this setting, we highlight differences in convergence between several unbiased compression schemes having the same variance increase.

Overall, this thesis proposes contributions to the field of federated learning by addressing central challenges and proposing solutions for efficient and sustainable learning in a distributed and heterogeneous framework. This work aligns with a global effort to make the use of large-scale federated learning viable by minimizing its environmental impact. Although benefits are expected, at least with respect to energy concerns, cautiousness is still required, as a rebound effect could occur: having faster and less energy-consuming algorithms could lead to a sharp increase in their applications, reducing or even canceling out the gains made by progress in their design.

Key-words: Federated learning, optimization, bidirectional compression, heterogeneity.

Résumé

Les deux dernières décennies ont été marquées par une augmentation sans précédent de la puissance de calcul et du volume de données disponibles. Les ensembles de données peuvent comprendre des milliards d'observations et les modèles peuvent comporter des millions de paramètres. En conséquence, les algorithmes d'apprentissage automatique ont évolué pour s'adapter à cette nouvelle situation. En particulier, les algorithmes stochastiques, qui ont été introduits pour la première fois dans les années 1950, ont récemment bénéficié d'un regain d'attention en raison de leur capacité à traiter des ensembles de données à grande échelle comportant des millions de paramètres. Ces algorithmes utilisent des informations de premier ordre pour réduire les coûts de calcul associés aux données en haute dimension. En outre, ils traitent efficacement un grand nombre d'observations en calculant des gradients stochastiques. Ces méthodes sont la clé des remarquables progrès réalisés au cours des deux dernières décennies dans le domaine de l'apprentissage automatique.

Cependant, beaucoup d'applications modernes utilisent désormais des réseaux de clients pour stocker les données et calculer les modèles : l'apprentissage devient plus complexe, en particulier en raison des contraintes de communication. C'est pourquoi, une nouvelle approche a été proposé, l'*apprentissage fédéré*, où les données sont gardées sur leur support d'origine tandis qu'un serveur central est mis en place pour orchestrer l'entraînement.

Cette thèse vise à aborder deux aspects fondamentaux de l'apprentissage fédéré. Le premier objectif est d'analyser les compromis de l'apprentissage distribué sous contraintes de communication ; le but étant de réduire le coût énergétique et l'empreinte environnementale. Le second objectif est d'aborder les problèmes résultant de l'hétérogénéité des clients qui complexifie la convergence de l'algorithme vers une solution optimale. Cette thèse se concentre sur la compression bidirectionnelle et résume mes contributions à ce domaine de recherche.

Dans notre première contribution, nous nous concentrons sur l'effet entremêlé de la compression et de l'hétérogénéité (statistique) des clients. Nous introduisons un framework d'algorithmes, appelé `Artemis`, qui s'attaque au problème des coûts de communication de l'apprentissage fédéré. Pour les réduire, `Artemis` permet de compresser les informations envoyées dans les deux sens (des clients vers le serveur et inversement) en combinaison avec un mécanisme de mémoire. Dans le cas de clients hétérogènes, nous mettons en lumière l'impact clé de la mémoire sur la convergence.

Dans notre deuxième contribution, nous mettons l'accent sur les boucles de rétroaction afin de réduire l'impact de la compression. Nous introduisons un algorithme, `MCM`, qui s'appuie sur `Artemis` et propose un nouveau paradigme qui préserve le modèle central lors de la compression descendante. Ce mécanisme permet d'effectuer une compression bidirectionnelle tout en atteignant asymptotiquement des taux de convergence identiques à ceux de la compression unidirectionnelle.

Dans notre troisième contribution, nous allons au-delà de l'hypothèse classique du pire cas sur la variance et fournissons une analyse fine de l'impact de la compression dans le cadre de la régression des moindres carrés. Dans cette configuration, nous mettons en évidence les différences de convergence entre plusieurs schémas de compression sans biais ayant pourtant la même variance.

Cette thèse apporte des contributions au domaine de l'apprentissage fédéré en relevant des défis importants et en proposant des solutions pour un apprentissage efficace et durable dans un cadre distribué et hétérogène. Ce travail s'inscrit dans un effort global visant à rendre viable l'utilisation de l'apprentissage fédéré à grande échelle en minimisant son impact sur l'environnement. Bien que des bénéfices soient attendus, du moins en ce qui concerne les préoccupations énergétiques, la prudence reste indispensable, car un effet de rebond pourrait survenir : disposer d'algorithmes énergétiquement moins chers et plus rapides pourrait entraîner une forte augmentation de leurs applications, réduisant voire annulant les gains réalisés par le progrès de leur conception.

Mots-clés : Apprentissage fédéré, optimisation, compression bidirectionnelle, hétérogénéité.

Contents

Notations

$:=$	Defined as
$\mathbb{1}$	Indicator/characteristic function
$\mathbb{R}, \mathbb{N}, \mathbb{N}^*,$	Sets of real, natural number, and natural without zero
$\mathbb{R}^d$	Set of d-dimensional real-valued vectors
$\langle x, y \rangle$	Inner product of vectors $x, y \in \mathbb{R}^d$
$x \otimes y$	Kronecker product of vectors $x, y \in \mathbb{R}^d$
$x \odot y$	Element-wise product of vectors $x, y \in \mathbb{R}^d$
$x \wedge y$	Minimum of x and y in $\mathbb{R}^d$
$\|x\|$	Euclidean norm of vector $x \in \mathbb{R}^d$
$\overline{x}_{K-1} := \sum_{k=0}^{K-1} x_k/K$	Average of any sequence of vector $(x_k)_{k \in \{0,\dots,K-1\}}$
e_i	Vector in $\mathbb{R}^d$ with zero everywhere except at coordinate i
$\mathbb{R}^{n \times d}$	Set of real matrices of size $n \times d$
$\mathcal{S}_d(\mathbb{R})$	Set of real symmetric matrices of size $d \times d$
$\mathcal{S}_d^+(\mathbb{R}), \mathcal{S}_d^{++}(\mathbb{R})$	Set of real symmetric positive (semi)-definite matrices of size $d \times d$
$\mathcal{O}_d(\mathbb{R})$	Group of orthogonal matrices
I_d	Identity matrix of size $d \times d$
$A^\top$	Transpose of matrix A
$A^\dagger := A^\top (A A^\top)^{-1}$	Moore–Penrose pseudo-inverse of A in $\mathbb{R}^{d \times n}$ s.t. $AA^\top$ inversible
$\mathrm{Tr}\,(A)$	Trace of matrix A
$\mathrm{eig}(A)$	Set of eigenvalues of matrix A
$\|A\|^2 := \mathrm{Tr}\left(A^\top A\right)$	Frobenius norm for matrix A in $\mathbb{R}^{n \times d}$
$\|\|\|A\|\|\| := \sqrt{\max \mathrm{eig}(A^\top A)}$	Operator norm for matrix A in $\mathbb{R}^{n \times d}$
$A \preccurlyeq B$	$B - A$ positive semi-definite (p.s.d.)
$A^{1/2}$	The p.s.d. square root of any symmetric p.s.d. matrix A
J_r	The $d \times d$ diagonal matrix whose r first diagonal elements are equal to one and all other matrix's coefficients equal to zero
$\mathcal{B}(\mathbb{R}^n)$	Borel set of $\mathbb{R}^n$
$\mathbb{P}(A)$	Probability of an event A
$\mathbb{E}[X]$	Expectation of a random variable X
$\mathbb{V}[X]$	Variance of a random variable X

$X \sim P$	Random variable X has distribution P
$\mathrm{Unif}(\mathcal{X})$	Uniform distribution on set $\mathcal{X}$
$\mathrm{Bern}(p)$	Bernouilli distribution with parameter p
$\mathcal{N}(\mu, \Sigma)$	Gaussian distribution with mean μ and covariance matrix Σ
$\mathcal{P}_k(\{1, \dots, n\})$	Set of all subset of $\{1, \dots, n\}$ with k elements
$\mathcal{L}^p(\Omega, \mathcal{A}, \mathbb{P})$	Set of random vectors defined on the probability space $(\Omega, \mathcal{A}, \mathbb{P})$ such that $\mathbb{E}[\|X\|^p] < \infty$
$\mathscr{C}^p(\mathbb{R}^d)$	Set of p times continuously differentiable functions from $\mathbb{R}^d$ into $\mathbb{R}$
∇F	Gradient function of $F : \mathbb{R}^d \to \mathbb{R}$
$\nabla^2 F$	Hessian matrix of $F : \mathbb{R}^d \to \mathbb{R}$
$\mathcal{C}_{\mathrm{up}}, \mathcal{C}_{\mathrm{dwn}}$	Uplink and downlink compressors
w_*	Optimal model minimizing F (if there exists at least one)
w_k, w_k^i	Model held by the central server (resp. local client i in $\{1, \dots, N\}$) at iteration k in $\mathbb{N}$

Thesis outline

This summary assumes that the reader knows about optimization, federated learning and compression mechanisms. The reader may choose to read the introduction (Chapter 1) first, and then come back to this summary.

Chapter 1. The opening Chapter of this thesis provides an overview of the key areas that are necessary for the understanding of the subsequent chapters. Firstly, we introduce the general setting of statistical learning, including its historical development, mathematical formulation, and examples of real-life use cases. We then delve into convex optimization, which is the cornerstone of the theoretical findings presented in this thesis. We furthermore present the case of federated learning, which is the main motivation of this thesis, and more particularly the bidirectional compression setting which is the focus of Chapters 2 and 3. Next, we motivate our choice to analyze bidirectional compression, highlighting its relevance in federated learning. As a conclusion of this introductive Chapter, we provide a mathematical summary of the three Chapters of this manuscript and shed light on the key messages of this thesis.

Chapter 2. In this Chapter, we focus on the intertwined effect of compression and client (statistical) heterogeneity. We introduce a framework – `Artemis` – to tackle the problem of learning in a distributed or federated setting with communication constraints. Several clients perform the optimization process using a central server to aggregate their computations. To alleviate the communication cost, `Artemis` allows to compress the information sent in *both directions* (from the clients to the server and conversely) combined with a memory mechanism. It improves on existing algorithms that only consider unidirectional compression (to the server), or use very strong assumptions on the compression operator. We provide fast rates of convergence (linear up to a threshold) under weak assumptions on the stochastic gradients (noise's variance bounded only *at optimal point*) in non-i.i.d. setting, highlight the impact of memory for unidirectional and bidirectional compression, and analyze Polyak-Ruppert averaging. We use convergence in distribution to obtain a *lower bound* of the asymptotic variance that highlights the practical limits of compression. We provide experimental results to demonstrate the validity of our analysis.

Chapter 3. In this Chapter, we move the focus toward feedback loops to reduce the impact of compression. We develop a new approach to tackle communication constraints in distributed learning problems with a central server. We propose and analyze an algorithm that performs bidirectional compression and achieves asymptotically the same convergence rate as algorithms using only uplink (from the local clients to the central server) compression. This algorithm, `MCM`, is such

that the downlink compression *only impacts local models*, while the global model is preserved. As a result, and contrary to previous works, the gradients on local servers are computed on *perturbed models*. Consequently, convergence proofs are more challenging and require a precise control of this perturbation. To ensure it, `MCM` additionally combines model compression with a memory mechanism. This analysis opens new doors, e.g. incorporating worker dependent randomized-models and partial participation.

Chapter 4. In this Chapter, we go beyond the classical worst-case assumption on the variance of compressors and provide a fine-grained analysis of the impact of compression within the fundamental learning framework of least-squares regression (LSR). Within this setting, we underline differences in terms of convergence rates between several unbiased compression operators, that all satisfy the same condition on their variance, thus going beyond the classical worst-case analysis. To do so, we analyze a general stochastic approximation algorithm for minimizing quadratic functions relying on a random field. We consider weak assumptions on the random field, tailored to the analysis (specifically, expected Hölder regularity), and on the noise covariance, enabling the analysis of various randomizing mechanisms, including compression. We then extend our results to the case of federated learning.

More formally, we highlight the impact on the convergence of the covariance $\mathfrak{C}_{\text{ania}}$ of the *additive noise induced by the algorithm*. We demonstrate that despite the non-regularity of the stochastic field, the limit variance term depends on $\operatorname{Tr}\left(\mathfrak{C}_{\text{ania}} H_F^{-1}\right)/K$ (where H_F is the Hessian of the optimization problem and K the number of iterations) as opposed to the vanilla LSR case where it is $\sigma^2 \operatorname{Tr}\left(H_F H_F^{-1}\right)/K = \sigma^2 d/K$ [Bach and Moulines, 2013]. Then, we analyze the dependency of $\mathfrak{C}_{\text{ania}}$ on the compression strategy and ultimately its impact on convergence.

Chapter 5. This Chapter concludes the thesis by summarizing our contributions and describing possible extensions.

Publications and preprints related to this manuscript are listed below:

1. **Chapitre 2** is based on *Bidirectional compression in heterogeneous settings for distributed or federated learning with partial participation: tight convergence guarantees* [Philippenko and Dieuleveut, 2020].
2. **Chapitre 3** is based on our work *Preserved central model for faster bidirectional compression in distributed settings* [Philippenko and Dieuleveut, 2021] published at Neurips 2021.
3. **Chapitre 4** is based on our work *Convergence rates for distributed, compressed and averaged least-squares regression: application to federated learning* [Philippenko and Dieuleveut, 2023] submitted at JMLR.

In this thesis, I did not include my contribution to [du Terrail et al., 2022], published at Neurips 2022. This project was a collaborative effort involving researchers from various worldwide institutions. The aim was to establish a benchmark for cross-silo federated learning with natural partitioning. The resulting benchmark, called `FLamby`, is focused on healthcare applications and is available on this repository. My contribution to the project consisted of two parts. First, I implemented some classical federated algorithms, and second, I conducted an analysis of the heterogeneity of the Flamby datasets, which can be found in Appendix M of the corresponding article.

Vue d'ensemble de la thèse

Ce résumé suppose que le lecteur possède une bonne connaissance de l'optimisation, de l'apprentissage fédéré et des mécanismes de compression. Si nécessaire, le lecteur peut choisir de lire d'abord l'introduction (Chapitre 1), puis de revenir à ce résumé.

Chapter 1. Le premier Chapitre de cette thèse donne une vue d'ensemble des domaines clés nécessaires à la compréhension des chapitres subséquents. Tout d'abord, nous présentons le cadre général de l'apprentissage statistique. Nous y incorporons une courte fresque historique de son développement, une formulation mathématique du problème et quelques exemples d'applications réels. Nous nous penchons ensuite sur l'optimisation convexe, qui est la pierre angulaire des résultats théoriques présentés dans cette thèse. Nous présentons enfin le cas de l'apprentissage fédéré, qui est la principale motivation de cette thèse, et plus particulièrement le cadre de la compression bidirectionnelle. Nous motivons dans la section suivante notre choix d'analyser la compression bidirectionnelle et montrons sa pertinence dans le contexte de l'apprentissage fédéré. En conclusion de ce Chapitre introductif, nous fournissons un résumé mathématique des trois Chapitres de ce manuscrit et soulignons les messages clés de cette thèse.

Chapter 2. Dans ce Chapitre, nous présentons `Artemis`, un paradigme permettant d'aborder le problème de l'apprentissage distribué ou fédéré avec des contraintes de communication. Plusieurs clients effectuent un processus d'optimisation et communiquent avec un serveur central qui agrége le résultat de leurs calculs. Pour réduire les coûts de communication, `Artemis` compresse les informations envoyées dans les deux sens (des clients vers le serveur central et inversement) en utilisant un mécanisme de mémoire, améliorant ainsi les algorithmes existants qui prennent en compte uniquement la compression unidirectionnelle (vers le serveur central), ou bien qui utilisent des hypothèses fortes sur les opérateurs de compression. Cela nous permet de fournir des taux de convergence rapides (linéaires jusqu'à un seuil) sous des hypothèses faibles sur les gradients stochastiques (variance du bruit limitée seulement *au point optimal*) dans un contexte non i.i.d., et de mettre en évidence l'impact de la mémoire pour la compression unidirectionnelle et bidirectionnelle, en outre, nous analysons le scénario où nous utilisons la moyenne de Polyak-Ruppert. Enfin, nous utilisons la convergence en loi pour obtenir une borne inférieure de la variance asymptotique, ce qui met en évidence les limites pratiques de la compression. Nous fournissons des résultats expérimentaux pour démontrer la validité de notre analyse.

Chapter 3. Dans ce Chapitre, nous développons une nouvelle approche pour faire face aux contraintes de communication dans un problème d'apprentissage distribué utilisant un serveur

central. Nous proposons et analysons un algorithme qui effectue une compression bidirectionnelle en atteignant asymptotiquement un taux de convergence identique à ceux d'algorithmes utilisant uniquement la compression ascendante (des clients vers le serveur central). Cet algorithm, `MCM`, est tel que la compression de la liaison descendante *impacte seulement les modèles locaux*, tandis que le modèle global est lui préservé. Par conséquent, et contrairement aux travaux existants, les gradients sur les serveurs locaux sont calculés sur des *modèles perturbés*. Pour cette raison, les preuves de convergence sont plus difficiles à obtenir et nécessitent un contrôle précis de cette perturbation. Pour garantir la convergence, `MCM` ajoute un mécanisme de mémoire lors de l'étape de compression descendante. Cette analyse ouvre de nouvelles portes, par exemple l'incorporation de modèles aléatoires qui dépendent des clients, ou le cas de la participation partielle.

Chapter 4. Dans ce Chapitre, nous allons au-delà de l'hypothèse classique du pire cas sur la variance des compresseurs et fournissons une analyse fine de l'impact de la compression dans le cadre de la régression des moindres carrés (LSR). Nous soulignons les différences en termes de taux de convergence entre plusieurs opérateurs de compression sans biais, qui satisfont tous la même condition sur leur variance, allant ainsi au-delà de l'analyse classique du pire cas. Pour ce faire, nous donnons une analyse générale d'un problème d'approximation stochastique reposant sur un champ aléatoire. Nous considérons des hypothèses faibles sur le champ aléatoire (en particulier la régularité de Hölder en espérance) et sur la covariance du bruit, ce qui permet l'analyse de divers mécanismes de compression. Nous étendons ensuite nos résultats au cas de l'apprentissage fédéré.

Plus formellement, nous mettons en évidence l'impact sur la convergence de la covariance $\mathfrak{C}_{\text{ania}}$ du bruit additif induit par l'algorithm stochastique. Nous démontrons, que malgré la non-régularité du champ stochastique, le terme de variance limite dépend de $\operatorname{Tr}\left(\mathfrak{C}_{\text{ania}} H_F^{-1}\right)/K$ (où H_F est la Hessienne du problème d'optimisation et K le nombre d'itérations) par opposition au cas LSR canonique où il dépend de $\sigma^2 \operatorname{Tr}\left(H_F H_F^{-1}\right)/K = \sigma^2 d/K$ [Bach and Moulines, 2013]. Ensuite, nous mettons en lumière la façon dont la matrice $\mathfrak{C}_{\text{ania}}$ dépend du choix du compresseur et enfin, la façon dont elle impacte la convergence.

Chapter 5. Ce Chapitre conclut la thèse en résumant nos contributions et en décrivant des extensions possibles.

Les publications et pré-publications liées à ce manuscrit sont énumérées ci-dessous.

1. Le **Chapitre 2** se fonde sur notre travail *Bidirectional compression in heterogeneous settings for distributed or federated learning with partial participation: tight convergence guarantees* [Philippenko and Dieuleveut, 2020].
2. Le **Chapitre 3** est basé sur notre article *Preserved central model for faster bidirectional compression in distributed settings* [Philippenko and Dieuleveut, 2021] publié à Neurips 2021.
3. Le **Chapitre 4** utilise les résultat de notre travail *Convergence rates for distributed, compressed and averaged least-squares regression: application to federated learning* [Philippenko and Dieuleveut, 2023] soumis à JMLR.

Dans cette thèse, je n'ai pas inclus ma contribution à [du Terrail et al., 2022], publié à Neurips 2022. Ce projet était une collaboration impliquant des chercheurs de diverses institutions venant du monde entier. L'ojectif était de créer un benchmark pour l'apprentissage fédéré inter-silo avec des jeux de données naturellement partitionné entre différentes entités. Ce benchmark, appelé `FLamby`, est composé de données provenant du domaine de la santé ; il est disponible sur ce dépôt. Ma contribution au projet a consisté en deux parties : j'ai implémenté quelques algorithmes fédérés considérés comme classiques, et deuxièmement, j'ai effectué une analyse de l'hétérogénéité des sept jeus de données inclus dans `FLamby` qui peut être trouvée dans l'annexe M de l'article correspondant.

1 Introduction

"Luminous beings are we, not this crude matter."

Yoda to Luke, *Episode V: The Empire Strikes Back*, George Lucas.

The opening Chapter of this thesis provides an overview of the key areas that are necessary for the understanding of the subsequent Chapters. Firstly, we introduce the general setting of statistical learning, including its historical development, mathematical formulation, and examples of real-life use cases. We then delve into convex optimization, which is the cornerstone of the theoretical findings presented in this thesis. We furthermore present the case of federated learning, which is the main motivation of this thesis, and more particularly the bidirectional compression setting which is the focus of Chapters 2 and 3. Next, we motivate our choice to analyze bidirectional compression, highlighting its relevance in federated learning. As a conclusion of this introductive Chapter, we provide a mathematical summary of the three chapters of this manuscript and shed light on the key messages of this thesis.

Contents

1.1 Statistical learning

"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E." [Mitchell, 1997, see p. 2, chapter 1]

1.1.1 Historical overview

Statistical learning is the science of developing and analyzing methods for making predictions or decisions based on data. Its roots can be traced back to the 19th century, with early pioneers such as Legendre and Gauss, who introduced independently the method of least-squares regression [Legendre, 1806, Gauss, 1809], or Laplace, who introduced the concept of conditional probability [Laplace, 1820].

In the mid-20th century, the field of statistics experienced a surge of interest in machine learning and artificial intelligence, with the development of methods such as the perceptron algorithm [McCulloch and Pitts, 1943, Rosenblatt, 1958] and decision trees [Hunt et al., 1966]. However, progress in statistical learning was hindered by the limitations of computing power and the availability of large datasets, leading to a "AI winter" in the 1970s and 1980s.

The 1990s saw a resurgence of interest in statistical learning, with the development of new methods that could handle large datasets and complex models. These methods include support vector machines [Cortes and Vapnik, 1995], boosting [Freund and Schapire, 1996], random forests

[Breiman, 2001] and neural networks [LeCun et al., 1998, 1999]. The rise of the internet and the availability of massive amounts of data then contributed to the growth of statistical learning.

Today, statistical learning is an extremely dynamic field, with applications in many areas, including climate studies, finance, healthcare, robotics, social media, to name just a few. Researchers continue to develop methods for analyzing data and making predictions, including deep learning [LeCun et al., 2015], reinforcement learning [Sutton and Barto, 2018], explainable AI [Ribeiro et al., 2016] and distributed learning [Konečný et al., 2016, McMahan et al., 2017].

1.1.2 Supervised machine learning

In this thesis, we consider only supervised learning [Duda et al., 1973, Vapnik, 1982, 1999, Hastie et al., 2009] and formalize the problem as follows. Suppose we have access to a pair (x, y) in $(\mathcal{X} \times \mathcal{Y})$, where $\mathcal{X}$ and $\mathcal{Y}$ are supposed to be two measurable spaces. The vector x is the explanatory variable or *features*, and y is the variable of interest or *output*. The set $\mathcal{Y}$, which describes the outputs, can be either quantitative ($\mathcal{Y} \subset \mathbb{R}$), or categorical ($\mathcal{Y}$ is a finite set, typically $\{-1, 1\}$ if there are two possible categories). This leads to the two main tasks of supervised learning:

- *Regression*, when one predicts a quantitative outcome.
- *Classification*, when one predicts a categorical outcome, e.g., with $\mathcal{Y} = \{-1, 1\}$.

The aim of supervised machine learning is to find a *predictor* (a measurable function) $f : \mathcal{X} \to \mathcal{Y}$ which predicts an output y in $\mathcal{Y}$ for any new input x in $\mathcal{X}$. The set of possible predictors is denoted $\mathcal{F}(\mathcal{X}, \mathcal{Y})$.

Quality of the predictor. To measure the quality of a predictor we choose a *loss function* (or cost function) $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, supposed measurable, such that, intuitively, for any (y, y') in $\mathcal{Y}^2$, $\ell(y, y')$ is small if y and y' are similar, and large otherwise. For the two tasks above, classical loss functions are:

- for regression, the squared loss $\ell(y, y') = \frac{1}{2}(y - y')^2$,
- for classification, the logistic loss $\ell(y, y') = \log(1 + \exp(-yy'))$ (this model is know as *logistic regression*) or the hinge loss $\ell(y, y') = \max\{0, 1 - yy'\}$.

We then define the *risk* R of a predictor f in $\mathcal{F}(\mathcal{X}, \mathcal{Y})$ as the averaged loss under the distribution $\mathcal{D}$ of the observations:

$$R(f) := \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(f(x), y)]\,.$$

The learning process seeks to find the best predictor f_* that minimizes the risk R. Such a predictor is called the *Bayes predictor* when it exists. In most situations, the quality of a predictor f is in fact not measured w.r.t. its loss, but rather using the *excess risk* $R(f) - R(f_*)$. To approximate f_*, the learning process consists first in choosing an (often parametric) family $\mathcal{F} \subset \mathcal{F}(\mathcal{X}, \mathcal{Y})$ of predictors and then minimizing the risk over it, hence finding $f_{\mathcal{F}} = \arg\min_{f\in\mathcal{F}} R(f)$.

Empirical risk minimization. In practice the observations' distribution $\mathcal{D}$ is unknown and we only have access to a dataset D of cardinal K, composed of pairs $(x_k, y_k)_{k\in\{1,\dots,K\}}$ in $(\mathcal{X} \times \mathcal{Y})^K$. This is why we define the *empirical risk error* (ERM):

$$R_K(f) = \frac{1}{K}\sum_{k=1}^{K} \ell(f(x_k), y_k)\,. \tag{ERM}$$

One approach is to minimize it instead of the true risk R, which leads to considering $f_K := \arg\min_{f\in\mathcal{F}} R_K(f)$. One of the major pitfalls of such an approach is *overfitting*: it corresponds to a scenario where the empirical risk error is very low, but the excess risk is large. In this case, the true risk is also called *generalization error* as it measures how accurately the predictor f_K trained on a dataset D is able to predict output values for unseen data.

1.1.3 Risk decomposition

Trade-off between approximation and estimation errors. The starting point of the learning process is to choose a family $\mathcal{F}$ of candidate predictors and to find f_K that minimizes the empirical risk R_K. However, first, the optimal predictive function f_* is unlikely to belong to the family $\mathcal{F}$, and secondly, the goal is not to find the predictor f_K that minimizes R_K, but the predictor $f_\mathcal{F}$ that minimizes R over $\mathcal{F}$. This is why, it is useful to decompose the excess risk error as follows [e.g. Bottou and Bousquet, 2007]:

$$\mathcal{E} = R(f_K) - R(f_*) = \underbrace{R(f_K) - R(f_\mathcal{F})}_{\text{estimation error}} + \underbrace{R(f_\mathcal{F}) - R(f_*)}_{\text{approximation error}} .$$

The *approximation error* measures how closely functions in $\mathcal{F}$ can approximate the optimal solution f_*; it can be reduced by choosing a larger family of function $\mathcal{F}$. The *estimation error* measures the effect of minimizing the empirical risk R_K instead of the expected risk R. The estimation error depends on the number of training examples and on the capacity of the family of functions; to reduce it, one can increase the number of points or choose a smaller family of functions $\mathcal{F}$. It thus appears that large families of functions have smaller approximation errors (bias) but lead to higher estimation errors (variance), therefore leading to the well-known "*bias-variance*" trade-off.

Optimization error. After having chosen a family $\mathcal{F}$ of predictors, the next step of the learning process consists in approximating f_K. Since f_K is itself an approximation, there is no need to carry out a minimization with the highest possible accuracy. Instead, all algorithms run the minimization for several steps and then return a predictor $\widehat{f}_K$. Therefore, an additional term $R(\widehat{f}_K) - R(f_K)$ appears in the decomposition of the excess risk $\mathcal{E}' = R(\widehat{f}_K) - R(f_*)$ [Bottou and Bousquet, 2007]:

$$\mathcal{E}' = R(\widehat{f}_K) - R(f_*) = \underbrace{R(\widehat{f}_K) - R(f_K)}_{\text{optimization error}} + \underbrace{R(f_K) - R(f_\mathcal{F})}_{\text{estimation error}} + \underbrace{R(f_\mathcal{F}) - R(f_*)}_{\text{approximation error}} .$$

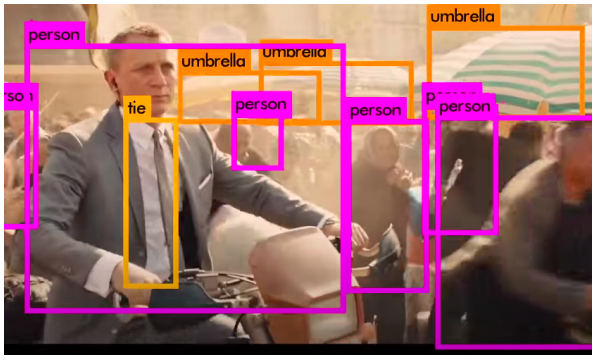
This *optimization error* reflects the approximation made during the optimization process, when minimizing the empirical risk. In this thesis, we provide bounds on the optimization error, either computed using the excess risk $\mathcal{E}_{\text{opt.}} := R(\widehat{f}_K) - R(f_K)$, either computed using the empirical excess risk $\mathcal{E}^K_{\text{opt.}} := R_K(\widehat{f}_K) - R_K(f_K)$. These bounds guarantee the convergence of our algorithms.

1.1.4 Least-squares regression

One of the simplest models of supervised learning is *least-squares regression* (LSR). Suppose that we have a dataset $(x_k, y_k)_{k\in\{1,\dots,K\}}$ in $(\mathcal{X}\times\mathbb{R})^K$ with K in $\mathbb{N}^*$. We consider the squared loss $\ell : (y, y') \mapsto \frac{1}{2}(y - y')^2$, and the parameterized family of functions $\mathcal{F}_\phi = \{f_w : z \mapsto \langle \phi(z), w\rangle \text{ with } w \in \mathbb{R}^d\}$ and $\phi : \mathcal{X} \to \mathbb{R}^d$ a function transforming the input features. It leads to the following ERM:

$$\arg\min_{f_w\in\mathcal{F}_\phi} R_K(f_w) = \arg\min_{w\in\mathbb{R}^d} \frac{1}{2K}\sum_{k=1}^{K}\Big(y_k - \phi(x_k)^\top w\Big)^2 .$$

Noting $y = (y_1, \dots, y_K)^\top$ and Φ in $\mathbb{R}^{K\times d}$ the matrix of inputs whose k-th row is $\phi(x_k)^\top$, we can rewrite the ERM as $\arg\min_{w\in\mathbb{R}^d} \frac{1}{2K}\|y - \Phi w\|^2$.



(a) Object detection using YoLo V2, tested on James Bond (Skyfall) [YoloV2, 2023].

(b) Generated with Dall-E: "An old castle on a cloud in a Miyazaki style" [Dall-E, 2023].

Figure 1.1: Some applications of machine learning.

When Φ has full column rank, there exists a unique closed-form solution to this problem given by the *ordinary least-square* (OLS) estimator: $\widehat{w} = (\Phi^\top \Phi)^{-1} \Phi^\top y$. Geometrically, the vector of prediction $\Phi\widehat{w} = \Phi(\Phi^\top \Phi)^{-1}\Phi^\top y$ is the orthogonal projection of y in $\mathbb{R}^K$ onto $\mathrm{Im}(\Phi) \subset \mathbb{R}^K$.

But in high dimension, this exact approach is computationally too expensive and cannot be applied. This is why, in practice, LSR is solved using stochastic gradient descent, which is often referred to as the *Least-Mean-Squares* (LMS) algorithm [Bershad, 1986, Macchi, 1995].

This framework will be the starting point of Chapter 4. We will consider a model where we have access to K in $\mathbb{N}^*$ i.i.d. observations $(x_k, y_k)_{k\in\{1,\dots,K\}} \sim \mathcal{D}^{\otimes K}$, such that there exists a well-defined model w_* in $\mathbb{R}^d$:

$$\forall k \in \{1,\dots,K\}, \quad y_k = \langle \phi(x_k), w_* \rangle + \varepsilon_k, \qquad \text{with } \varepsilon_k \sim \mathcal{N}(0,\sigma^2) \ , \tag{1.1}$$

for an i.i.d. sequence $\big((\varepsilon_k)_{k\in\{1,\dots,K\}}\big)$ independent from $\big((x_k)_{k\in\{1,\dots,K\}}\big)$. We assume that the inputs' second moment is bounded to define $\mathbb{E}[\phi(x_1)\otimes\phi(x_1)] = H$; it is the features' covariance.

1.1.5 Real-life application of machine learning.

Supervised machine learning has applications in a wide range of domains; we provide a selection of examples here.

Image recognition. It is one of the most popular applications of supervised learning [Felzenszwalb et al., 2009, Krizhevsky et al., 2009, LeCun et al., 2010, Girshick, 2015, Simonyan and Zisserman, 2015, Szegedy et al., 2015, Ren et al., 2015, Xiao et al., 2017, Redmon et al., 2016, He et al., 2016, Szegedy et al., 2016, Redmon and Farhadi, 2017]. From a dataset of labeled images, we train a neural network to distinguish a picture from another; once trained to recognize particular items, the model can be deployed in real-life applications. In Figure 1.1a, we give an example of object detection tested on Jame Bond[1] based on version 2 of Yolo [Redmon et al., 2016, Redmon and Farhadi, 2017], three known labels are being detected: ties, persons, and umbrellas.

Synthethic image generation. In the same vein, a recent line of work has focused on generating synthetic images [Mirza and Osindero, 2014, Radford et al., 2015, Salimans et al., 2016, Arjovsky et al., 2017, Karras et al., 2019, Goodfellow et al., 2020]. This field of research has enabled the creation of high-quality and realistic images that can be used for a wide range of applications, including art, entertainment, and scientific research. For instance, we provide on Figure 1.1b an example of an image generated using Dall-E V2 with the prompt: "An old castle on a cloud in a

[1] Screenshot from [YoloV2, 2023]

Miyazaki style". Note that usually these techniques are considered to be unsupervised learning, but they use a supervised loss as part of the training, hence their mention in this section.

Medicine. Supervised learning has also a major impact in medicine [Gulshan et al., 2016, Ting et al., 2017, Rajkomar et al., 2019, Esteva et al., 2019, Choi et al., 2016, Shickel et al., 2017, Miotto et al., 2016, Chen and Asch, 2017], for instance, to improve medical diagnosis, treatment planning, or patient outcomes. A recent line of research [Sheller et al., 2020, Rieke et al., 2020, Dayan et al., 2021, Pati et al., 2021, du Terrail et al., 2021, 2022] has additionally considered the case of private and sensitive datasets that are split across various clients. These clients cannot share their data but aim to collaborate with others to improve the training. This line of research has unlocked the ability to take advantage of these previously out-of-reach datasets and doing so, to improve the development of new clinical research.

Product recommendations. Recommender systems are ones of the most successful and widespread applications of machine learning technologies [Resnick and Varian, 1997, Huang et al., 2004, Ziegler et al., 2005, Bell et al., 2007, Koren, 2008, 2009, Zhou et al., 2010, Lü et al., 2012, Covington et al., 2016, Cheng et al., 2016, Zhang et al., 2019]. Many companies rely on this business model: Amazon, Google, Meta, Tiktok, Netflix, Criteo, to name just a few. The goal is to increase customer satisfaction or consumption by analyzing their interests and extrapolating relevant information from other people's behavior. Such a framework requires two kinds of recommender models: (1) a global one minimizing the empirical risk over all clients and (2) a personalized one that is adapted to each user's singularity.

Applications based on sensors. A wide variety of supervised tasks use datasets gathered from sensors: cameras, microphones, accelerometers, thermometers... Considering that sensors are today deployed at a large scale - for instance, in smartphones, buildings, cars, boats, drones, planes, satellites - a lot of initiatives have emerged to benefit from these new sources of data. It has led, for instance, to develop autonomous cars [Kanade et al., 1986, Campbell et al., 2010, Bitam et al., 2015, Contreras-Castillo et al., 2017, Hussain and Zeadally, 2018, Badue et al., 2021] or smart buildings [Morvaj et al., 2011, Albino et al., 2015, Ghayvat et al., 2015, Plageras et al., 2018, Brandi et al., 2020].

All of these examples share two characteristics. First, the data may originate from various sources, for instance, sensors, cameras, smartphones, hospitals, user accounts ... As a result, the datasets are inherently heterogeneous, raising significant challenges in developing a general model. Second, it might not be feasible to collect all the data from all these sources on a single server, requiring the design of a distributed learning process and leading to a high cost of communication.

In this thesis, we focus on supervised learning problems where the dataset is heterogeneously split across several clients. Our aim is to address the challenge of reducing the communication cost of the training while finding a global consensus among these *statistically* heterogeneous data sources.

Therefore, we present in Section 1.2 the framework of optimization for machine learning which allows finding a minimizer of the expected/empirical risk error. Then, we introduce the more specific setting of interest, namely distributed and heterogeneous learning, in Section 1.3.

1.2 Optimization for machine learning

The goal of the learning process is to find a solution to the (ERM) problem, which corresponds to a minimization problem. In order to solve it, we rely in this thesis on the widely studied *gradient descent* (GD) procedure. We denote $\mathscr{C}^p(\mathbb{R}^d)$ the set of p times continuously differentiable functions from $\mathbb{R}^d$ into $\mathbb{R}$.

1.2.1 Gradient descent

Let F a function from $\mathbb{R}^d$ to $\mathbb{R}$. The goal of optimization [see e.g. for introductory lectures, Nesterov, 2004, Boyd et al., 2004, Bubeck, 2015] is to find an *optimal point* w_* (not necessarily unique) that minimizes F:

$$w_* = \arg\min_{w \in \mathbb{R}^d} F(w)\,. \tag{1.2}$$

Solving this problem with an accuracy $\varepsilon > 0$ means finding an approximate solution $\widehat{w}$ after K in $\mathbb{N}^*$ iteration, such that the error $F(\widehat{w}) - F(w_*)$ is smaller than ε. The relationship between the number of iterations K is determined by the *complexity* function $\mathcal{T} : \mathbb{R} \mapsto \mathbb{N}$ s.t. $K = \mathcal{T}(\varepsilon)$; we say that the optimization error has a *worst-case complexity* of $O(\mathcal{T}(\varepsilon))$. In other words, it represents the maximum number of iterations K in $\mathbb{N}^*$ to reach a given precision ε. In the next chapters, we analyze algorithms through the lens of convergence rate analysis, and use the results to get insights into their practical performance.

In practice, to find the optimal point w_* minimizing the objective function, we use a method that is able to collect specific information about F depending on its regularity. The process of collecting this information is called an *oracle*. In accordance with the degree of smoothness of F, we can rely on different types of oracles. Let w in $\mathbb{R}^d$,

1. *zeroth-order* oracle returns the value $F(w)$;
2. *first-order* oracle returns the value $F(w)$ and the gradient $\nabla F(w)$, if F is differentiable;
3. *second-order* oracle returns the value $F(w)$, the gradient $\nabla F(w)$, and the Hessian $\nabla^2 F(w)$, if F is twice differentiable.

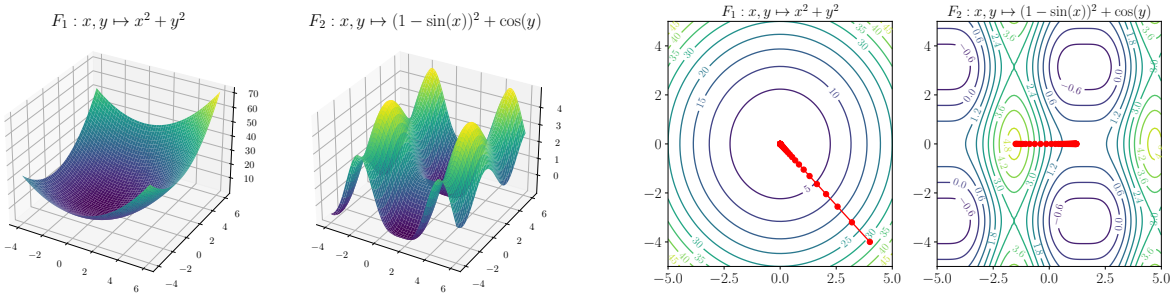
Exhaustive-search – an example of zeroth-order method. A naive approach to find the optimal point w_* by using a zeroth-order oracle is to build a grid over the search-space and then evaluate the function over each node of the grid. Of course, this approach is extremely costly. For instance, suppose we want to find the optimum on a grid $[0, 1]$; to achieve an ε-accuracy, we need $\lfloor 1/\varepsilon \rfloor$ evaluations of F, but in dimension d, we need $\lfloor 1/\varepsilon \rfloor^d$ evaluations! Thereby, this method is never used in practice, except to tune the hyper-parameters.

Gradient descent – an example of first-order method. For F in $\mathscr{C}^1(\mathbb{R}^d)$, first-order methods rely on the gradient ∇F and take advantage of the fact that gradients are orthogonal to the level sets, therefore pointing toward the steepest direction. At each step, the algorithm's complexity is thus $O(d)$ (cost of the gradient computation in $\mathbb{R}^d$). These algorithms start from a random point w_0 and take repeated steps in the opposite direction of the gradient. Mathematically, its leads to a sequence $(w_k)_{k>0}$ defined by:

$$w_k = w_{k-1} - \gamma \nabla F(w_{k-1})\,, \tag{GD}$$

with $\gamma > 0$ a *step-size* (also known as *learning rate*) that controls the update's magnitude. The choice of the step-size rate is fundamental and has been one of the most studied questions: taking γ too small slows down convergence and γ too big leads to divergence.

Newton method – an example of second-order method. For F in $\mathscr{C}^2(\mathbb{R}^d)$, second-order methods use the Hessian $\nabla^2 F$ to update the model, thus the complexity of the algorithm is now



(a) Left convex function, right non-convex function.

(b) Corresponding examples of gradient descent.

Figure 1.2: GD with a convex function and a non-convex function.

increased to $O(d^2)$ per iteration. Mathematically, the Newton method results to a sequence of model $(w_k)_{k>0}$ updated as follows:

$$w_k = w_{k-1} - \nabla^2 F(w_{k-1})^{-1} \nabla F(w_{k-1}) \,,$$

The intuition behind the second-order Newton method is to use the 2^{nd}-order Taylor approximation of the function to approximate it, and at each iteration to minimize this quadratic function.

The difficulty of the optimization process, i.e., of finding the optimal point w_*, depends on various properties that are verified or not by F. The knowledge of these properties has a key impact on the choice of the optimization algorithm and on its convergence rate.

- **Problem's dimensionality.** An evident difficulty comes from the problem's dimensionality. As the dimension d increases, the computational's complexity of each iteration grows, making it harder to retrieve the first and second oracles on the function F. In high-dimension, the prohibitive computational cost of Newton's method makes it unusable.
- **Dataset's size.** In supervised learning, we are optimizing not the expected risk R but the empirical risk R_K, the objective function F is hence defined by the dataset. It follows that the complexity of evaluating the function, the gradient, or the Hessian is proportional to the dataset size. This is why the cost of computing the oracle can become prohibitive if the dataset size increases exponentially. To circumvent this issue, we will consider in Subsection 1.2.3 (and throughout this thesis) an oracle g that computes an approximation of the gradient ∇F with a computational cost independent of the dataset size K.
- **Regularity of F.** The regularity of the objective function F is important in optimization as it ensures the success of learning and its speed, in particular, the (GD) method and Newton's method require respectively that F belongs to $\mathscr{C}^1(\mathbb{R}^d)$ and $\mathscr{C}^2(\mathbb{R}^d)$. In contrast, a non-regular function may have discontinuities, singularities, or oscillations that can cause optimization algorithms to get stuck or to converge slowly, see Figure 1.2 for illustration (more details are given below). This is why, regularity is a desirable property for objective functions in optimization; we provide additional details in Subsection 1.2.2.

1.2.2 Regularity assumption

An important regularity condition to guarantee the convergence of (GD) is the convexity of F.

Assumption 1.1 (Convexity)**.** *F is differentiable and convex, that is for all vectors z, z' in $\mathbb{R}^d$, it verifies:*

$$F(z') \geq F(z) + (z' - z)^T \nabla F(z) \,.$$

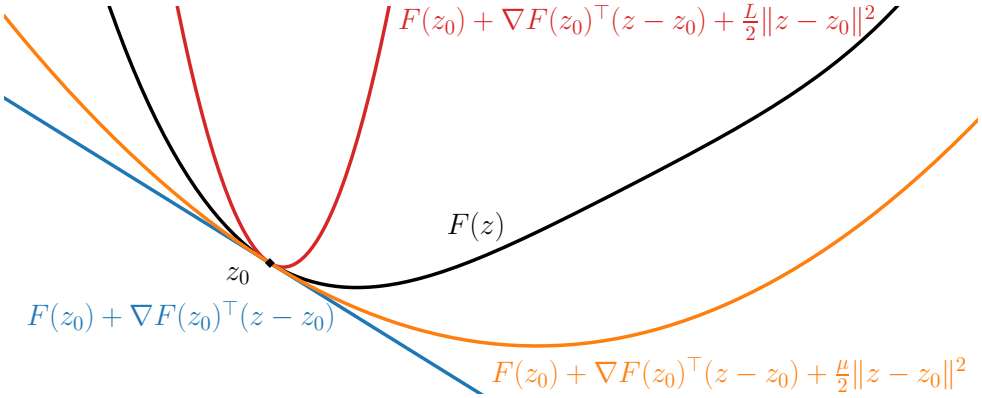


Figure 1.3: Illustration of the quadratic upper bound and lower bound provided respectively by L-smoothness (in red) and μ-strong-convexity (in orange) at a point z_0 for a function F.

The main reason why convex functions are attractive in optimization is that every stationary point is also a global minimum. This property ensures that the optimization process does not become trapped in a local optima or saddle point. Note first that it does not guarantee the existence of a global minimum (take for instance $F : z \mapsto z$) and secondly, that the existence of a global minimum does not ensure its uniqueness.

On Figure 1.2a we give two examples of functions: on the left $F_1 : x, y \mapsto x^2 + y^2$ is convex, on the right $F_2 : x, y \mapsto (1 - \sin(x))^2 + \cos(y)$ is not. Then on Figure 1.2b we run for each function a gradient descent procedure using $\gamma = 0.1$. Observe that for F_1, the gradient procedure converges quickly to the optimal point while for F_2 the gradient descent is stuck into a saddle point, illustrating the difficulty of finding a local minimum in this setting.

A stronger assumption on convexity is to assume that the function is strongly-convex, which in turn, guarantees the unicity of the optimal point w_*. Geometrically, strong-convexity can be interpreted as the possibility to lower bound F at any point with a quadratic function (see Figure 1.3).

Assumption 1.2 (Strong convexity)**.** *F is differentiable and μ-strongly convex (with $\mu \geq 0$), that is for all vectors z, z' in $\mathbb{R}^d$:*

$$F(z') \geq F(z) + (z' - z)^T \nabla F(z) + \frac{\mu}{2}\|z' - z\|_2^2 \,.$$

Note that we recover the convex case if $\mu = 0$.

Additionally, a classical desirable property of F is its L-smoothness which corresponds considering that F can be upper-bounded at any point by a quadratic function (see Figure 1.3).

Assumption 1.3 (Smoothness)**.** *F is in $\mathscr{C}^1(\mathbb{R}^d)$ and is L-smooth (with $L \geq 0$), that is for all vectors z, z' in $\mathbb{R}^d$:*

$$\|\nabla F(z) - \nabla F(z')\| \leq L\|z - z'\| \,.$$

Condition number κ. Two quantities are then of particular importance when minimizing F: the strongly-convexity constant μ and the smoothness constant L. These two coefficients directly impact the speed of convergence of algorithms based on gradient descent. This is why, under Assumptions 1.2 and 1.3, we define the *condition number* $\kappa = L/\mu$: the bigger κ, the slower the convergence towards the optimal point. Note also that if F is in $\mathscr{C}^2(\mathbb{R}^d)$, smoothness and strong-convexity are equivalent to having for any point z in $\mathbb{R}^d$: $\mu \mathrm{I}_d \preccurlyeq \nabla^2 F(z) \preccurlyeq L\mathrm{I}_d$.

These assumptions enable to give two theorems of convergence for (GD) in the smooth convex setting, and in the smooth strongly-convex setting.

Theorem 1.1 (Theorem 2.1.15 from Nesterov [2004]). *Consider Assumptions 1.2 and 1.3, the sequence of iterates $(w_k)_{k>0}$ produced by (GD) initialized at w_0 in $\mathbb{R}^d$ and using a step-size $\gamma = 1/L$ verifies for any K in $\mathbb{N}$:*

$$F(w_K) - F(w_*) \leq (1 - \kappa^{-1})^K \left(F(w_0) - F(w_*)\right) .$$

Taking $\gamma = 2/(\mu + L)$ leads to a slightly more powerful result:

$$F(w_K) - F(w_*) \leq \left(1 - 2(\kappa + 1)^{-1}\right)^{2K} \frac{L \|w_0 - w_*\|^2}{2} .$$

This theorem states that the sequence $(w_k)_{k>0}$ converges at an exponential rate to the optimal point w_*, we say that the convergence of (GD) is *linear*. And in the smooth convex setting, we obtain the following.

Theorem 1.2 (Corollary 2.1.2 from Nesterov [2004]). *Consider Assumptions 1.1 and 1.3, the sequence of iterates $(w_k)_{k>0}$ produced by (GD) initialized at w_0 in $\mathbb{R}^d$ and using a step-size $\gamma = 1/L$ verifies for any K in $\mathbb{N}$:*

$$F(w_K) - F(w_*) \leq \frac{2L\|w_0 - w_*\|^2}{K + 4} .$$

These two theorems show that the choice $\gamma = 1/L$ is very powerful as a unique algorithm is adapted to both convex and strongly-convex functions, although the convergence is faster in the latter case. Hence, this step-size does not require the knowledge of μ which can be arbitrarily small, as opposed to $\gamma = 2/(\mu + L)$. Moreover, in the latter case, if μ tends to zero, the algorithm does not converge, since κ tends to infinity.

1.2.3 Stochastic gradient descent

As mentioned earlier, one of the major difficulties of machine learning comes from the fact that the dataset size can be extremely large, making it impossible to evaluate the function F. *Stochastic gradient descent* (SGD) introduced by Robbins and Monro [1951], solves this issue and has become one of the most popular tools of machine learning because of its practical efficiency and its theoretical performance. In this manuscript, we focus on stochastic methods and we consider that at any iteration k in $\mathbb{N}^*$, we have access to an oracle g_k that evaluates an *unbiased* approximation of the true gradient ∇F (that is g_k is a random field), hence updating (GD) as following:

$$w_k = w_{k-1} - \gamma g_k(w_{k-1}) . \tag{SGD}$$

In addition to the condition number κ, two quantities dictate the behavior of SGD: the initial distance (bias) $\|w_0 - w_*\|^2$ and the variance σ^2 associated with the stochastic gradient, it leads to a "bias/variance decomposition" [see Hsu et al., 2012, Bach and Moulines, 2013]. Therefore, we consider the following assumption on the stochastic gradients.

Assumption 1.4 (Noise over stochastic gradients computation). *The noise over stochastic gradients is zero-centered and its variance is uniformly bounded by a constant $\sigma \in \mathbb{R}_+$, such that for all k in $\mathbb{N}$, for all z in $\mathbb{R}^d$ we have: $\mathbb{E}[\|g_k(z) - \nabla F(z)\|^2] \leq \sigma^2$.*

Using these stochastic oracles $(g_k)_{k\in\mathbb{N}^*}$ of the true gradient ∇F leads to the below upper bound.

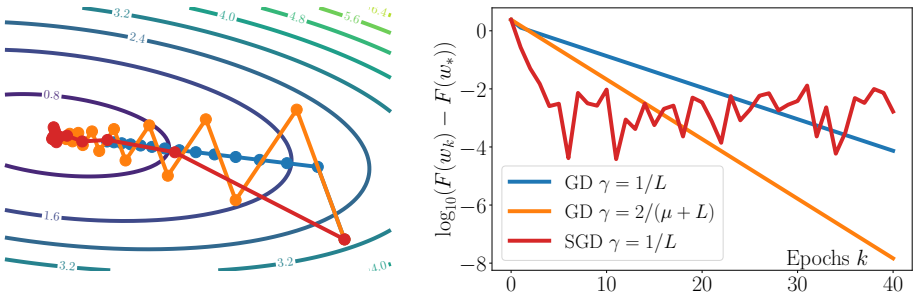


Figure 1.4: Left: level set of F and iterates $(w_k)_{k\in\mathbb{N}^*}$ resulting from the three gradient descents. Right: logarithm excess loss w.r.t. the number of epochs.

Theorem 1.3. *Consider Assumptions 1.2 to 1.4, the sequence of iterates $(w_k)_{k>0}$ produced by (SGD) initialized at w_0 in $\mathbb{R}^d$ and using a step-size γ in $\mathbb{R}_+^*$ verifies for any K in $\mathbb{N}$:*

$$\|w_K - w_*\|^2 \leq (1-\gamma\mu)^K \|w_0 - w_*\|^2 + \frac{\gamma\sigma^2}{\mu}\,.$$

We illustrate Theorems 1.1 and 1.3 on Figure 1.4: on the left we represent the iterates $(w_k)_{k\in\mathbb{N}^*}$ obtained after some iterations and on the right the logarithm excess loss. The objective function F is quadratic and the dataset is generated using Equation (1.1) with $w_* = (0, 0.6)^\top$, $\sigma^2 = 1$ and $H = \left(\begin{smallmatrix} 1 & 1 \\ 1 & 10 \end{smallmatrix}\right)$. We run GD with the two step-sizes proposed in Theorem 1.1: both converge linearly to the optimal point, but we observe the superiority of choosing $\gamma = \frac{2}{\mu+L}$ in the convex scenario. On the contrary, SGD with constant step-size first converges linearly, and then *saturates* at a level depending on σ^2. This is due to the oscillations of the iterates around the optimal point.

However, Assumption 1.4 is in fact very restrictive and is not verified even by the simple setting of LSR (presented in Subsection 1.1.4). This is why we sometimes instead only assume that the variance is bounded by a constant σ_*^2 *at the optimal point* w_* (hence requiring its existence which excludes non-convex setting). In Chapter 2, following the work of Gower et al. [2019], Dieuleveut et al. [2020], Assumption 1.5 will result in a linear convergence rate up to a threshold *proportional to* σ_*^2.

Assumption 1.5 (Noise over stochastic gradients computation at optimal points)**.** *The noise over stochastic gradients at the global optimal point is zero-centered and its variance is bounded by a constant $\sigma_* \in \mathbb{R}_+$, such that for all k in $\mathbb{N}$, we have: $E[\|g_k(w_*) - \nabla F(w_*)\|^2] \leq \sigma_*^2$.*

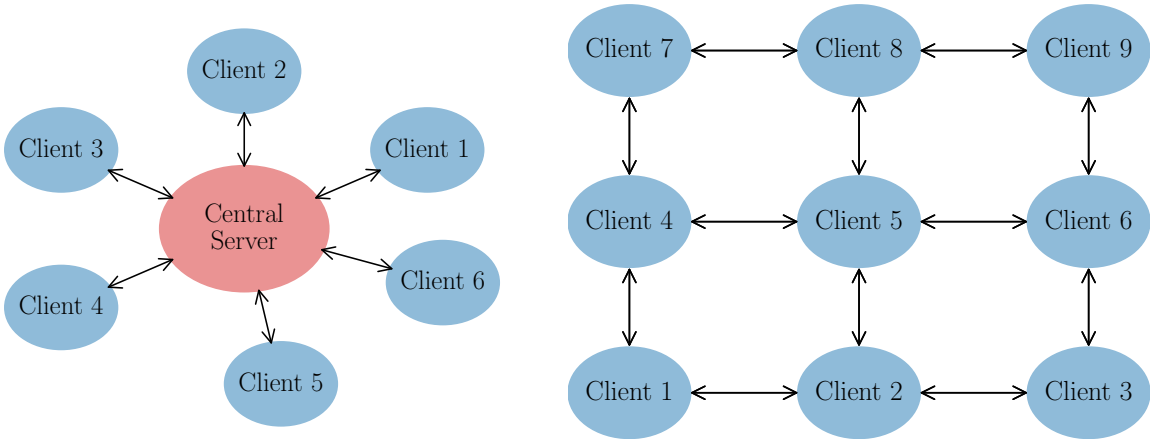
But then instead of assuming the smoothness of F, we will assume the cocoercivity [see Zhu and Marcotte, 1996, for more details about this hypothesis] of gradients, which implies the smoothness.

Assumption 1.6 (Cocoercivity of stochastic gradients (in quadratic mean))**.** *We suppose that for all k in $\mathbb{N}$, stochastic gradient functions g_k are L-cocoercive (with $L \geq 0$) in quadratic mean. That is, for k in $\mathbb{N}$, and for all vectors z, z' in $\mathbb{R}^d$, we have $\mathbb{E}[\|g_k(z) - g_k(z')\|^2] \leq L\langle \nabla F(z) - \nabla F(z'), z - z'\rangle$.*

Using Assumptions 1.5 and 1.6 enables to recover the same results as Theorem 1.3 but with σ^2 replaced by σ_*^2 [Gower et al., 2019].

Note that various methods have been proposed to reduce the variance induced by the stochastic gradient: Polyak-Ruppert averaging [Polyak and Juditsky, 1992], mini-batch or tail-averaging (see for instance [Jain et al., 2018b, Muecke et al., 2019]), variance-reduction methods [Johnson and Zhang, 2013, Schmidt and Roux, 2013, Defazio et al., 2014]. In particular, in Chapters 2 to 4, we provide theorems on the Polyak-Ruppert iterate $\overline{w}_{K-1} = \frac{1}{K}\sum_{k=0}^{K-1} w_k$. This iterate can be computed online, since for any K in $\mathbb{N}^*$, we have:

$$\bar{w}_K = \frac{1}{K+1} w_K + \frac{K}{K+1}\bar{w}_{K-1}\,.$$



(a) Topology of federated learning. (b) Decentralized topology, e.g. clients are on a grid.

Figure 1.5: Example of two topologies in distributed learning.

In this thesis, we design algorithms that are based on stochastic gradient descent. We provide theoretical analysis in strong-convex, convex, and non-convex scenarios and give theorems guaranteeing the convergence of the proposed algorithms.

We next introduce the setting of federated learning which is the main motivation of this thesis and present the challenge of reducing the cost of communication.

1.3 Federated learning

Federated learning (FL) is a machine learning setting where many clients (e.g. mobile devices or whole organizations) collaboratively train a model under the orchestration of a central server (e.g. service provider), while keeping the training data decentralized. [Kairouz et al., 2019, see p.4]

1.3.1 Framework

In Sections 1.1 and 1.2, we supposed that the learning process has permanent access to a dataset D and that at any moment, it can access any of its points. However, nowadays problems are often not valid for such an assumption, considering that in most situations data are generated from various sources (see examples given Subsection 1.1.5). In this setting, it is not always possible to centralize the data coming from these different sources on a single server and to exploit them. This can happen for various reasons.

- The datasets are too large to be stored on a single server.
- The cost/time of data communication is too high to send all at once.
- The datasets are private or sensitive and can not leave their source.
- The data is obtained in a streaming fashion and hence is constantly changing.

This is why a new approach has been developed in the last years and considers a *distributed* setting instead. In such a setting, the data is kept on its origin server. We call *client* (or *worker*) a server holding a dataset and participating in the training. Then, either a central server is put in

place to orchestrate the training, this is *federated learning* (FL), either clients communicate in a peers-to-peers fashion, this is *decentralized learning*. On Figure 1.5 we plot two possible topologies of distributed learning. On Figure 1.5a we present a federated network with a central server, and on Figure 1.5b we present an example of topology without a central server where clients are placed on a grid and can communicate only with their neighborhoods.

In this thesis, we consider the FL setting [Konečný et al., 2016, McMahan et al., 2017, Kairouz et al., 2019], where N clients communicate with a central server that aggregates all updates and then broadcasts back a message to all clients. Formally, we have a number of features $d \in \mathbb{N}^*$, and a convex cost function $F : \mathbb{R}^d \to \mathbb{R}$. We want to solve the following distributed convex optimization problem using stochastic gradient algorithms [Robbins and Monro, 1951, Bottou, 2010]:

$$\min_{w \in \mathbb{R}^d} F(w) \text{ with } F(w) = \frac{1}{N} \sum_{i=1}^{N} F_i(w) \,,$$

where $(F_i)_{i \in \{1,\dots,N\}}$ are the *local objective functions* on client i in $\{1, \dots, N\}$ (they could be either the local empirical risk or the local expected risk).

The starting point of the algorithms presented in this thesis is *Distributed SGD* where at each iteration $k \in \mathbb{N}^*$ and for each client i in $\{1, \dots, N\}$, a local gradient g_{k+1}^i is computed using a single batch and then sent to the central server which aggregates the information from all clients before updating the model. It results in the below update equation:

$$w_{k+1} = w_k - \gamma \frac{1}{N} \sum_{i=1}^{N} g_{k+1}^i(w_k) \,. \qquad \text{(Dist. SGD)}$$

In this distributed setting, the main challenge is the communication cost which has been identified as an important bottleneck [e.g. Strom, 2015, Kairouz et al., 2019]. Firstly, exchanging information can slow down the whole training process. Secondly, sharing high volumes of data can be problematic for the users in terms of bandwidth usage (for instance ResNet-50 proposed in He et al. [2016] has over 23 million parameters, and VGG16 designed by Simonyan and Zisserman [2015] has 138 million). Thirdly, the energy cost of the communication process is also significant and should be considered alongside other classical constraints in machine learning, as seen in studies such as Henderson et al. [2020] or Anthony et al. [2020]. Three approaches can be used to reduce the cost: (1) increase the number of local updates, (2) reduce the frequency of communication between the clients and the central server or (3) compress the information exchanged from the clients to the central server and conversely.

Algorithm 1: Federated Averaging (`FedAvg`)

Input: Initial model w_0, number of communication rounds K, local epochs E, learning rate γ, batch size b, and the proportion p of active workers at each round.
Output: Global model w_K
for $k = 1, 2, ..., K$ **do**
  $S_k \leftarrow$ (random set of $\lfloor pN \rfloor$ clients)
  for *each client* i *in* S_k **do**
    $\mathcal{B}_i \leftarrow$ (split dataset D_i in batch of size b)
    $w^i = w_{k-1}$
    for *each local epoch* $e = 1, 2, ..., E$ **do**
      for *each batch* B *in* $\mathcal{B}_i$ **do**
        $w^i = w^i - \gamma g_B^i(w^i)$
    $w_k^i = w^i$
  $w_k = \frac{1}{N} \sum_{i=1}^{N} n_i w_k^i$

1. Local updates. Local update algorithms involve performing multiple updates on each client before transmitting the final update to the central server; the server then applies these updates to its global state, and the process is repeated [McMahan et al., 2017, Karimireddy et al., 2020, 2021, Ghadikolaei et al., 2021, Koloskova et al., 2020, Lin et al., 2018, Stich, 2019, Malinovskiy et al., 2020,

Gao et al., 2021]. One of the most influential algorithms is `FedAvg` proposed by McMahan et al. [2017], we give its pseudo-code in Algorithm 1 because of its founding role. It selects at each round k in $\mathbb{N}^*$ a random subset $S_k \subset \{1, \dots, N\}$ of $\lfloor pN \rfloor$ clients with p in $]0; 1]$; each client i in S_k run E in $\mathbb{N}^*$ iterations of gradient descent by splitting their dataset D_i in a set $\mathcal{B}_i$ of batches of size b in $\mathbb{N}^*$. We note g_B^i the stochastic gradient computed on client i in $\{1, \dots, N\}$ when using the batch B in $\mathcal{B}_i$.

2. Partial participation (PP). In federated optimization, it is classical to consider that at each round, all participants are not used for the training [e.g. Zhao and Zhang, 2015, Csiba and Richtárik, 2018, Mishchenko et al., 2018, Chen et al., 2020, Yang et al., 2021, Wang et al., 2022, Luo et al., 2022, Wang and Xu, 2022, Jhunjhunwala et al., 2022, Eichner et al., 2019, Fraboni et al., 2021a, 2022, Yang et al., 2021, Rodio et al., 2023]. For instance, it is common to consider mobile devices to be ready to participate only when idle, charging and connected to a fast network, leading a client to be regularly switched on/off during the training process. This setting results in new constraints; for instance how to synchronize the training, how to design efficient sampling strategies, how to guarantee convergence to a global optimum ...

3. Compression. Compression is a critical aspect of distributed learning, as it addresses the three issues mentioned above that hinder the practical implementation of distributed learning. Therefore, compression for distributed learning is an important subject of research. It has led to a significant effort of the community to design the best algorithms, either by compressing only the uplink channel [Seide et al., 2014, Alistarh et al., 2018, Khirirat et al., 2018, Karimireddy et al., 2019, Wu et al., 2018, Chraibi et al., 2019, Mishchenko et al., 2019, Horvath et al., 2022, Reisizadeh et al., 2020, Gorbunov et al., 2020a, Li et al., 2020b, Haddadpour et al., 2021, Richtarik et al., 2021, Kovalev et al., 2021, Li and Richtárik, 2021], either both uplink and downlink channels [Tang et al., 2019, Liu et al., 2020, Zheng et al., 2019, Philippenko and Dieuleveut, 2020, 2021, Gorbunov et al., 2020b, Sattler et al., 2019, Horvath et al., 2022, Fatkhullin et al., 2021].

In this thesis, we have considered the last solution, and in particular, we have designed two algorithms doing bidirectional compression. Note that the scenario of partial participation is naturally covered by our analysis by being considered as a particular case of compression where is sent either the complete gradient, either 0. We give more details about compression and review related work in the next Subsection 1.3.2.

1.3.2 Compression

We consider a compressor $\mathcal{C} : \mathbb{R}^d \to \mathbb{R}^d$ as a random function that verify the below assumption:

Assumption 1.7 (Compression)**.** *There exists a constant $\omega \in \mathbb{R}_+^*$, such that the random operator $\mathcal{C}$ satisfies for all z in $\mathbb{R}^d$ the following two properties:*

$$\mathbb{E}[\mathcal{C}(z)] = z \qquad \textit{and} \qquad \mathbb{E}[\|\mathcal{C}(z) - z\|^2] \leq \omega \|z\|^2 \,.$$

In other words, the compression is *unbiased* and its variance is relatively bounded by a constant ω. Another common assumption, that is not considered in this thesis is that the compressor is *contractive*, i.e. for any z in $\mathbb{R}^d$, $\|\mathcal{C}(z) - z\|^2 < (1 - \delta)\|z\|^2$ with $\delta \in]0; 1[$ [almost surely or in expectation, see for instance Seide et al., 2014, Stich et al., 2018, Karimireddy et al., 2019, Ivkin et al., 2019, Koloskova et al., 2019, Gorbunov et al., 2020b, Beznosikov et al., 2020, Richtárik et al., 2022]. We use unbiased operators because they allow reducing by a factor N the variance, while the bias is kept independent of N: suppose we have N clients compressing independently the same vector z in $\mathbb{R}$ using compressors $(\mathcal{C}_i)_{i=1}^N$ that verify Assumption 1.7, then $\frac{1}{N}\sum_{i=1}^N \mathcal{C}_i(z)$ also verifies this assumption but with a constant ω/N.

In the following, we define several unbiased compression operators that verify Assumption 1.7. These operators are classical in the literature and will be considered throughout this thesis.

Definition 1.1 (Compression operators)**.** *Let z in $\mathbb{R}^d$.*

1. *1-**quantization** is defined as $\mathcal{C}_{\mathrm{q}}(z) = \|z\|\mathrm{sign}(z) \odot \chi \in \mathbb{R}^d$ with $\chi \sim (\mathrm{Bern}(|z|/\|z\|_2))_{i=1}^d$.*
2. ***Rand-**h is defined as $\mathcal{C}_{\mathrm{rd}h}(z) := \frac{d}{h} B(S) \odot z$ with $S \sim \mathrm{Unif}(\mathcal{P}_h(\{1, \dots, d\}))$ and $B(S)_i = \mathbb{1}_{i \in S}$.*
3. ***Sparsification** is defined as $\mathcal{C}_{\mathrm{s}}(z) = \frac{1}{p} B \odot z \in \mathbb{R}^d$ with $B \sim (\mathrm{Bern}(p))_{i=1}^d$.*
4. ***Partial participation** can also be seen as a technique of compression as it reduces the cost of communication. We define $\mathcal{C}_{\mathrm{PP}}(z) = \frac{1}{p} bz$ with $b \sim \mathrm{Bern}(p)$.*
5. ***Sketching**, also known as Random Projection, is defined as $\mathcal{C}_{\Phi}(z) = \frac{1}{p} \Phi^\dagger \Phi z$, where $h \ll d$ in $\mathbb{N}$, $p = h/d$ and $\Phi \in \mathbb{R}^{h \times d}$ is a random projection matrix into a lower-dimension space.*

While we consider only unidirectional compression (from clients to central server) in Chapter 4, we focus on bidirectional compression in Chapters 2 and 3. It consists in compressing communications in both directions between the central server and remote devices. We use two different compression operators, respectively $\mathcal{C}_{\mathrm{up}}$ and $\mathcal{C}_{\mathrm{dwn}}$, to compress the message in each direction. In its simplest form, Equation (Dist. SGD) becomes:

$$w_{k+1} = w_k - \gamma \mathcal{C}_{\mathrm{dwn}} \left(\frac{1}{N} \sum_{i=1}^N \mathcal{C}_{\mathrm{up}}(g_{k+1}^i(w_k)) \right). \tag{1.3}$$

From an abstract standpoint, we can introduce three quantities: (1) $g_k^i := g_{k+1}^i(w_k)$ the gradient computed at iteration k in $\mathbb{N}^*$ on client i in $\{1, \dots, N\}$, (2) $\widehat{g}_k^i$ the effective gradient shared by the client i to the central server, and (3) $\widehat{G}_k$ the effective gradient used to update the model w_k. In the context of bidirectional compression defined in Equation (1.3), we have: $\widehat{g}_k^i = \mathcal{C}_{\mathrm{up}}(g_k^i)$ and $\widehat{G}_k = \mathcal{C}_{\mathrm{dwn}}(\frac{1}{N} \sum_{i=1}^N \mathcal{C}_{\mathrm{up}}(g_k^i))$.

One of the main challenges of compression is to attenuate its inherent error that might lead the training process to diverge. For this purpose, two mechanisms have been introduced: (1) *error-feedback* (EF) by Seide et al. [2014], and (2) *memory* by Mishchenko et al. [2019]. We briefly describe hereafter how these choices affect $\widehat{g}_k^i$ and $\widehat{G}_k$.

Error-feedback. Error-feedback (or error-compensation) is a mechanism that accumulates errors of compression and corrects the gradient computation at each step. This approach was applied mainly to biased operators of compression (which excludes the compressors considered in this thesis) and has been successfully used in various works, for instance in [Seide et al., 2014, Stich et al., 2018, Zheng et al., 2019, Karimireddy et al., 2019, Tang et al., 2019, Zheng et al., 2019, Beznosikov et al., 2020, Liu et al., 2020, Stich and Karimireddy, 2020, Gorbunov et al., 2020b, Qian et al., 2021]. In the setting of unidirectional compression, it is mathematically described as:

$$\begin{cases} \widehat{g}_k^i = \mathcal{C}_{\mathrm{up}}(\gamma g_k^i + e_k^i)/\gamma \\ e_{k+1}^i = e_k^i + \gamma(g_k^i - \widehat{g}_k^i) \,. \end{cases}$$

In the context of double compression, it has been shown to improve convergence for a restrictive class of *contracting* compression operators (which are generally biased) by Zheng et al. [2019], Tang et al. [2019]. But for unbiased operators, it did not lead to any theoretical improvement [see Remark 2 in Sec. 4.1., Liu et al., 2020].

Memory. Memory has been introduced in `Diana` by [Mishchenko et al., 2019] for unbiased compressors. It consists in compressing the difference between the gradients and a *local memory*

term, making the compression error tends to zero, and thus improving the convergence. In the setting of unidirectional compression, it corresponds mathematically to having:

$$\begin{cases} \widehat{g}_k^i = \mathcal{C}_{\mathrm{up}}(g_k^i - h_{k-1}^i) + h_{k-1}^i \\ h_k^i = h_{k-1}^i + \alpha \mathcal{C}_{\mathrm{up}}(g_k^i - h_{k-1}^i)\,, \end{cases}$$

where α is the memory's learning rate. It corresponds to a client-wise control-variate, as already used by Schmidt et al. [2017] for variance-reduction.

This approach has later been analyzed in many papers. Horváth et al. [2022], Condat and Richtarik [2022] have added a variance-reduction approach. It has been combined with acceleration in the strongly-convex regime by Li et al. [2020b] and in the non-strongly-convex setting by Li and Richtárik [2021]. Gorbunov et al. [2021] has used it to develop `Marina` an algorithm using a biased stochastic estimator of the gradient in a non-convex setting. Memory and EF have been combined together in the unidirectional case by Gorbunov et al. [2020b], and for bidirectional compression by Liu et al. [2020]. Still in the bidirectional setting, Safaryan et al. [2021] has developed an algorithm based on Newton method to compress the Hessian shift. Zhao et al. [2021] have used successfully memory to reduce client-variance in the case of partial participation. Li et al. [2022b] has used memory to design an algorithm that applies compression directly to differentially-private stochastic gradients. The memory mechanism has also been applied for $\alpha = 1$ to biased compressors by Richtarik et al. [2021], Fatkhullin et al. [2021], Gruntkowska et al. [2022]. Memory-like ideas have also been used beyond ERM, for instance for Langevin Stochastic Dynamics by Vono et al. [2022], or for Expectation-Maximization algorithms by Dieuleveut et al. [2021].

Memory versus EF. Memory and EF are motivated by two different goals. EF is doing a retro-compensation of the past errors of compression accumulated over iterations in order to remove them, and thus $(e_k)_{k\in\mathbb{N}}$ tends to zero. Therefore, the quantity that is compressed corresponds to a *compensated gradient*. Memory (aka "control-variate") is put in place to compensate the clients' heterogeneity by learning the specificity of each client, thus all memories $(h_k^i)_{i=1}^N$ tends to $(\nabla F_i(w_*))_{i=1}^N$. Therefore, the quantity that is compressed with this mechanism corresponds to the *innovation* of the new iteration. In other words, EF leads to a "feedback loop" providing information from the past that helps to correct the drift induced by the error of compression, while on the other hand, memory reduces the variance induced by the statistical heterogeneity of clients; this last point being shown in Chapter 2.

The setting of clients' statical heterogeneity setting is introduced in the following Subsection 1.3.3.

1.3.3 Client statistical heterogeneity

The natural setting of federated learning [see e.g. Kairouz et al., 2019] is the case of *statistical* heterogeneous[2] clients, i.e., each client i in $\{1, \cdots, N\}$ holds its own data distribution $\mathcal{D}_i$ potentially different from the others. Below, we use the following assumption to quantify the heterogeneity of clients in the network; this assumption is considered in Chapter 2 for the convex setting.

Assumption 1.8 (Bounded gradient at w_*)**.** *There exists an optimal parameter w_* minimizing F (not necessarily unique) and a constant $B \in \mathbb{R}_+$, such that*

$$\frac{1}{N}\sum_{i=1}^N \|\nabla F_i(w_*)\|^2 = B^2\,.$$

[2]Note that it can be found in the literature that clients are said heterogeneous when they face variability in hardware (CPU, memory) and power (battery level). In this thesis are considered only *statistical* heterogeneity and not a *system* heterogeneity.

In fact, this is merely an assumption, but rather a definition of the constant B. In the streaming *i.i.d. setting* – $D_1 = \cdots = D_N$ and $F_1 = \cdots = F_N$ – the assumption is satisfied with $B = 0$. On Figure 1.6, we illustrate an example of three heterogeneous clients with a local objective function of the form $F_i : w \in \mathbb{R}^2 \mapsto (w - w_*^i)^\top H_F (w - w_*^i) + F_i(w_*^i)$, i.e. with the same Hessian H_F but with different optimal points $(w_*^i)_{i \in \{1,\dots,3\}}$. In green, we represent the level sets of the global objective function and its optimal point w_*. The gradient of F_1, F_2, F_3 evaluated at w_* are not null.

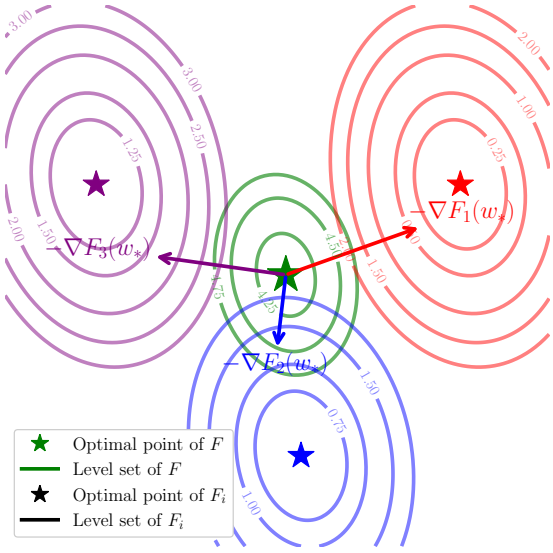


Figure 1.6: Illustration of heterogeneity on three clients, the objective functions are quadratic. We represent the optimal points, the level set, and the opposite gradient at the optimal point.

In this scenario, as analyzed by Li et al. [2019b], the simple `FedAvg` algorithm works very poorly and results to a model whose performance may vary significantly across the clients. Therefore, a lot of studies are investigating this problem in order to find efficient ways to handle it. To tackle this challenge, two strategies can be considered: (1) finding a global consensus between clients [Smith et al., 2017, Li et al., 2019a, Karimireddy et al., 2019, Hsu et al., 2019, Li et al., 2022a, Marfoq et al., 2022, Pillutla et al., 2022a,b, Laguel et al., 2020, 2021, du Terrail et al., 2022, Caldas et al., 2019, Li et al., 2020a, Mitra et al., 2021, Collins et al., 2021, Li et al., 2021, Mansour et al., 2020, Zhang et al., 2021] or (2) personalizing the model for each client [Deng et al., 2020, Grimberg et al., 2020, Beaussart et al., 2021, Even et al., 2022, Fallah et al., 2020, Li et al., 2021].

In this thesis, we focus on the first strategy and aim to design effective algorithms tackling the clients' heterogeneity. The challenge under this setting is to make the algorithm converge with the best possible limit variance.

The goal of this thesis is to focus simultaneously on two challenges of federated learning: reducing the cost of communication in a *heterogeneous* setting by doing *bidirectional compression*.

Next, in Section 1.4, we motivate our choice to analyze bidirectional compression rather than simply unidirectional. Many research papers assume that downlink speed is higher than uplink, therefore resulting in a lower communication cost that can be neglected. However, we show that this is not the case in practice and we highlight scenarios where downlink speed should not be ignored.

1.4 Motivation of using bidirectional compression

There are several reasons to consider downlink compression, and not simply compressing the uplink signal. First, the difference between upload and download speeds is not significant enough to ignore the impact of the downlink direction (see Subsection 1.4.1 for an analysis of bandwidth). Moreover, if we consider for instance a small number N of clients training a very heavy model – the size of deep learning models generally exceeds hundreds of MB [Dean et al., 2012, Huang et al., 2019] – the training speed will be limited by the exchange time of the updates, thus using downlink compression is key to accelerate the process. Secondly, in a different setting in which a network of smartphones collaborate to train a large-scale model in a federated framework, participants to the training would not be eager to download hundreds of MB for each update on their phone. Here again, downlink compression appears to be necessary.

In Subsection 1.4.1, we present an analysis of the bandwidth speeds for download/upload on fixed/mobile broadband relying on a study made in 2020 over the six continents by Index [2020]. Then in Subsection 1.4.2, we show with the concrete example of quantization, how bidirectional compression helps to reduce the communication cost compared to unidirectional.

1.4.1 Bandwidth speed

In a network configuration where download would be much faster than upload, bidirectional compression would present no benefit over unidirectional, as downlink communications would have a negligible cost. However, this is not the case in practice: to assess this point, we gathered broadband speeds, for both download and upload communications, for fixed broadband (cable, T1, DSL ...) or mobile (cellphones, smartphones, tablets, laptops ...) from studies carried out in 2020 over the 6 continents by *Speedtest.net* [see Index, 2020]. Results are provided in Figure 1.7, comparing download and upload speeds. The ratios (averaged by continents) between upload and download speeds stand between 1 (in Asia, for fixed broadband) and 3.5 (in Europe, for mobile broadband): there is thus no apparent reason to simply disregard the downlink communication, and bidirectional compression is unavoidable to achieve substantial speedup. More precisely, if we denote v_d and v_u the speed of download and upload (in Mbits per second), we typically have $v_d = \rho v_u$, with $1 < \rho < 3.5$. Using quantization with $s = 1$ (see Definition 1.2), for unidirectional compression, each iteration takes $O\left(\frac{Nd}{\rho v_u}\right)$ seconds, while for a bidirectional one it takes only $O\left(\frac{N\sqrt{d}\log_2(d)}{v_u}\right)$ seconds.

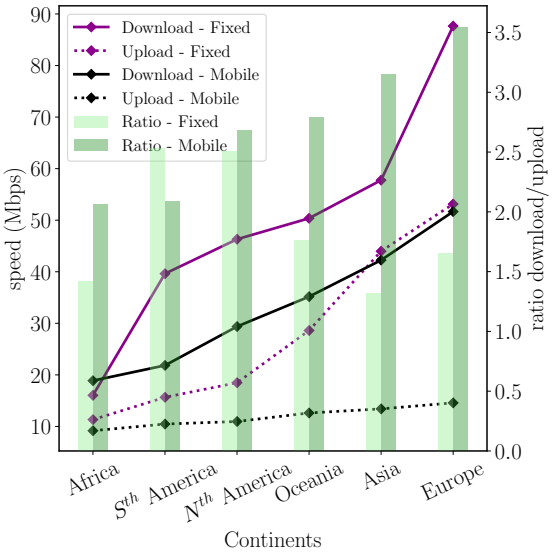


Figure 1.7: Left axis: upload and download speed for mobile and fixed broadband. Left axis: speeds (in Mbps), right axis: ratio (green bars). The dataset is gathered from *Speedtest.net*, see Index [2020].

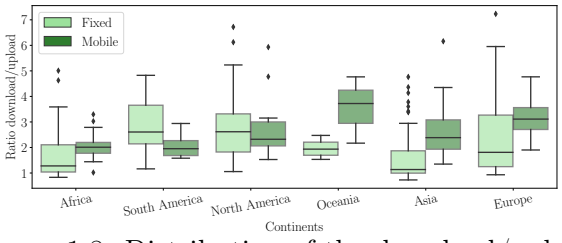


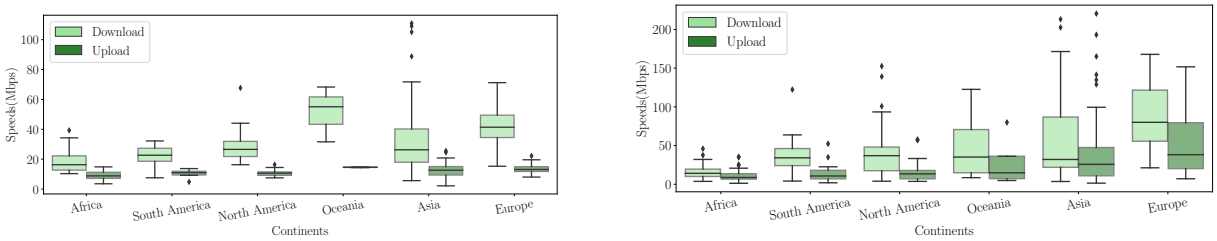
Figure 1.8: Distribution of the download/upload speeds ratio by continents.

In Figures 1.8, 1.9a and 1.9b, unlike Figure 1.7, we do not aggregate data by countries of the same continents. This allows us to analyze the speed ratio between upload and download with the *proper* value of each country. Looking at Figures 1.8, 1.9a and 1.9b, it is noticeable that in the world, the ratio between upload and download speed is between 1 and 5, and not between 1 and 3.5 as Figure 1.7 was suggesting since we were aggregating data by continents. There are only nine countries in the world having a ratio higher than 5. In Europe: Malta, Belgium, and Montenegro. In Asia: South Korea. In North America: Canada, Saint Vincent and the Grenadines, Panama, and Costa Rica. In Africa: Western Sahara. The highest ratio is 7.7 observed in Malta.

1.4.2 Communication cost: an example using the quantization scheme

In the following, we define the s-quantization operator $\mathcal{C}_s$ which we use in most of the experiments in Chapters 2 and 3. After giving its definition, we explain [based on Alistarh et al., 2017] how it helps to reduce the number of bits to broadcast at each iteration.

Definition 1.2 (s-quantization operator)**.** *Given $z \in \mathbb{R}^d$, the s-quantization operator $\mathcal{C}_s$ is defined by $\mathcal{C}_s(z) := \text{sign}(z) \times \|z\|_2 \times \frac{\chi}{s}$. $\chi \in \mathbb{R}^d$ is a random vector with j-th element defined as:*



(a) Mobile broadband. (b) Fixed broadband.

Figure 1.9: Upload/download speed (in Mbps).

$$\chi := \begin{cases} l+1 & \text{with probability } s\frac{|z_j|}{\|z\|_2} - l\,, \\ l & \text{otherwise} \end{cases} \quad \text{where the level } l \text{ is such that } \frac{s|z_j|}{\|z\|_2} \in [l, l+1[.$$

The s-quantization scheme verifies Assumption 1.7 with $\omega = \min(d/s^2, \sqrt{d}/s)$. Proof can be found in [Alistarh et al., 2017, see Appendix A.1].

Now, for any vector $v \in \mathbb{R}^d$, we are in possession of the tuple $(\|v\|^2, \phi, \chi)$, where ϕ is the vector of signs of $(v_j)_{j=1}^d$, and χ is the vector of integer values $(\chi_j)_{j=1}$. To broadcast the quantized value, we use the `Elias` encoding Elias [1975]. Using this encoding scheme, it can be shown (Theorem 3.2 of Alistarh et al. [2017]) that:

Proposition 1.1. *For any vector v, the number of bits needed to communicate $\mathcal{C}_s(v)$ is upper bounded by:*

$$\left(3 + \left(\frac{3}{2} + o(1)\right)\log_2\left(\frac{2(s^2+d)}{s(s+\sqrt{d})}\right)\right) s(s+\sqrt{d}) + 32\,.$$

With $s = 1$, it means that we will employ $O(\sqrt{d}\log_2 d)$ bits per iteration instead of $32d$, which reduces by a factor $\frac{\sqrt{d}}{\log_2 d}$ the number of bits used by iteration. Now, in a FL settings, at each iteration we have a double communication (device to the main server, main server to the device) for each of the N clients. It means that at each iteration, we need to communicate $2 \times N \times 32d$ bits if compression is not used. Obviously, unidirectional compression can at best result in a factor 2 reduction in term of total number of bits, while for bidirectional compression, we need to broadcast $O(N\sqrt{d}\log_2 d)$ bits using the `Elias` encoding [defined in Elias, 1975]. Denoting v_d and v_u the speed of download and upload (in bits per second), we typically have $v_d = \rho v_u$, $3.5 > \rho > 1$. Then for unidirectional compression, each iteration takes $O\left(\frac{Nd}{v_d} + \frac{N\sqrt{d}\log_2(d)}{v_u}\right) = O\left(\frac{Nd}{\rho v_u}\right)$ seconds, while for a bidirectional one, it takes only $O\left(\frac{N\sqrt{d}\log_2(d)}{v_u}\right)$ seconds.

In other words, unless ρ is really large (which is not the case in practice as stressed by Figure 1.7), double compression reduces by several orders of magnitude the global time complexity, and bidirectional compression is superior to unidirectional.

1.5 Summary of the contributions of this thesis

The naive way of doing bidirectional compression has been given in Equation (1.3), it consists in simply compressing the local gradients, and then compressing their average on the central server before using it to update the model: $w_{k+1} = w_k - \gamma \mathcal{C}_{\text{dwn}}\left(\frac{1}{N}\sum_{i=1}^N \mathcal{C}_{\text{up}}(g_{k+1}^i(w_k))\right)$.

In a such setting – considering that the compressors are random processes whose variance are constants $\omega_{\text{up}}, \omega_{\text{dwn}}$ – it is possible to prove that the variance of SGD iterates is increased by a factor $\omega_{\text{up}} \times \omega_{\text{dwn}}$.

In the next subsection, we summarize the key contributions of this thesis to address the challenge of compression, we include the most representative theorems of each chapter. Our results are validated by numerical experiments and the code is provided on our GitHub repositories:

- see this repository for the implementation of both `Artemis` and `MCM` used in Chapters 2 and 3,
- see this repository for the code of Chapter 4.

1.5.1 Contributions of Chapter 2

In Chapter 2, we propose `Artemis`, a framework that encompasses 6 algorithms (with or without up/down compression, with or without memory), the update being given for any k $\mathbb{N}^*$ by:

$$\begin{cases} \forall i \in [\![1, N]\!], \quad \widehat{\Delta}_k^i = \mathcal{C}_{\text{up}}\left(g_k^i - h_{k-1}^i\right), \text{ and then } h_k^i = h_{k-1} + \alpha \widehat{\Delta}_k^i, \\ \Omega_k = \mathcal{C}_{\text{dwn}}\left(\frac{1}{N}\sum_{i=1}^N (\widehat{\Delta}_k^i + h_{k-1}^i)\right) \\ w_k = w_{k-1} - \gamma \Omega_k. \end{cases}$$

Constants $\gamma, \alpha \in \mathbb{R}^* \times \mathbb{R}_+$ are learning rates for respectively the iterate sequence and the memory sequence $(h_k^i)_{k\in\mathbb{N}^*, i\in\{1,\dots,N\}}$. The consequence of introducing the memory is that at iteration k in $\mathbb{N}^*$, instead of compressing the gradient g_k^i which expectation tends to $\nabla F_i(w_*) \neq 0$ (Assumption 1.8), we compress a difference which tends now to zero in expectation like in the homogeneous scenario. We provide a fast rate of convergence – exponential convergence up to a threshold proportional to σ_*^2, the noise at the optimal point –, obtaining tighter bounds than in other works on compression.

Theorem 1.4 (Convergence of `Artemis`)**.** *Under Assumptions 1.2 and 1.5 to 1.8, for a step size γ satisfying some conditions, for a learning rate α_{up} verifying some conditions, and for any k in $\mathbb{N}$, the mean squared distance of w_k to w_* decreases at a linear rate up to a constant of the order of E:*

$$\mathbb{E}\left[\|w_k - w_*\|^2\right] \leq (1-\gamma\mu)^k \left(\|w_0 - w_*\|^2 + 2\gamma^2 C B^2\right) + \frac{2\gamma E}{\mu N},$$

for constants C and E depending on the variant (independent of k) of `Artemis`.

We explicitly tackle heterogeneity using Assumption 1.8, proving that the limit variance of `Artemis` with memory is independent from the difference between distributions (as for SGD). Indeed, we prove that memory makes the saturation threshold E independent of B^2. This is one of the first theoretical guarantee for double compression that explicitly quantifies the impact of non-i.i.d. data.

We prove *convergence in distribution* of the iterates, and subsequently provide *a lower bound* on the asymptotic variance. This sheds light on the limits of (double) compression, which results in an increase of the algorithm's variance, and can thus only accelerate the learning process for *early iterations* and up to a *"moderate" accuracy*. It also means that the upper bound on the saturation level *is tight* w.r.t. $\sigma_*^2, \omega_{\text{up}}, \omega_{\text{dwn}}, B^2, N$ and γ.

Theorem 1.5 (Convergence in distribution and lower bound on the variance)**.** *Under Assumptions 1.2 and 1.5 to 1.8, for $\gamma, \alpha_{\text{up}}, E$ satisfying some condition, when k goes to infinity, the second order moment $\mathbb{E}[\|w_k - w_*\|^2]$ converges to a limit variance* lower bounded *by $\Omega(\gamma E/\mu N)$, with E depending on the variant of `Artemis`.*

1.5.2 Contributions of Chapter 3

`Artemis` has a drawback, in order to be able to broadcast back the aggregate of the received local gradient, it compresses it *before* applying the update, resulting in a waste of valuable information. The advantage is that the model held on the central server and the one used on the local workers (to

query the gradient oracle) are identical. However, this means that the model on the central server has been artificially degraded: instead of using all the received information, it is updated with the compressed information. We propose a new algorithm `MCM` which updates the global model w_{k+1} independently of the downlink compression, hence **leading to a non-degraded update**. `MCM` is entirely defined by the following uplink and downlink equations.

"Server-to-client" equations

$$\begin{cases} \Omega_k = w_k - H_{k-1}\,, \\ \widehat{w}_k = H_{k-1} + \mathcal{C}_{\rm dwn}(\Omega_k) \\ H_k = H_{k-1} + \alpha_{\rm dwn}\mathcal{C}_{\rm dwn}(\Omega_k). \end{cases}$$

"Clients-to-server" equations

$$\begin{cases} \forall i \in [\![1, N]\!], \Delta_k^i = g_k^i(\widehat{w}_{k-1}) - h_{k-1}^i \\ w_k = w_{k-1} - \frac{\gamma}{N}\sum_{i=1}^N \mathcal{C}_{\rm up}(\Delta_k^i) + h_{k-1}^i \\ h_k^i = h_{k-1}^i + \alpha_{\rm up}\mathcal{C}_{\rm up}(\Delta_k^i). \end{cases}$$

This implies that the local models are *different* from the central model. The local gradients are thus measured on a *"perturbed model"* (or *"perturbed iterate"*). Such an approach requires a more involved analysis and the deviation between the local and global models must be carefully controlled [Mania et al., 2016].

In this thesis, for the sake of simplicity, we summarize results in a homogeneous setting [see our paper for results in heterogeneous setting, Philippenko and Dieuleveut, 2021, Appendix G]. Therefore, we set $\alpha_{\rm up}$ to zero; indeed we show in Chapter 2 that uplink memory is useful only in the heterogeneous setting. Additionally, we choose $\alpha_{\rm dwn} = (8\omega_{\rm dwn})^{-1}$ and denote $\Phi(\gamma) := (1+\omega_{\rm up})\left(1 + 64\gamma L\omega_{\rm dwn}^2\right)$. We show that `MCM` achieves the same rate of convergence as single compression in strongly-convex, convex and non-convex regimes. Note that this behavior has later been also recovered by Zou et al. [2022] for the special case of Top-K compression. We consider $\gamma_{\max}$ and $\widetilde{L}$, two constants defined in Chapter 3, then we have the following bounds of convergence.

Theorem 1.6 (Convergence of `MCM` in the homogeneous and strongly-convex case)**.** *Under Assumptions 1.2 to 1.4 and 1.7 with $\mu > 0$, for k in $\mathbb{N}$, if $\sigma^2 = 0$ (noiseless case), for $\gamma_k = \gamma_{\max}$ we recover a linear convergence rate: $\mathbb{E}[\|w_k - w_*\|^2] \leq (1 - \gamma_{\max}\mu/2)^k \|w_0 - w_*\|^2$.*

Furthermore, if $\sigma^2 > 0$, taking for all K in $\mathbb{N}$, $\gamma_K = 4/(\mu(K+1) + \widetilde{L})$, for the weighted Polyak-Ruppert average $\bar{w}_K = \sum_{k=1}^K \lambda_k w_{k-1} / \sum_{k=1}^K \lambda_k$, with $\lambda_k := (\gamma_{k-1})^{-1}$, we have:

$$\mathbb{E}\left[F(\bar{w}_K) - F(w_*)\right] \leq \frac{\square}{\mu K^2}\|w_0 - w_*\|^2 + \frac{8\sigma^2(1+\omega_{\rm up})}{\mu K N}\left(1 + \frac{\square'\omega_{\rm dwn}^2}{\mu K}\ln(\mu K + \square'')\right),$$

where $\square, \square', \square''$ are three constants given in Chapter 3.

On Figure 1.10, we observe that `MCM` meets `Diana` (unidirectional compression) while `Artemis` saturates at a higher level (scaling as $\omega_{\rm up} \times \omega_{\rm dwn}$), which illustrates the behavior stated in Theorem 1.6.

We also propose a variant, `Rand-MCM` incorporating diversity into models shared with the local clients and show that it improves convergence for quadratic functions. `Rand-MCM` simply consists in applying independent downlink compressions for each client.

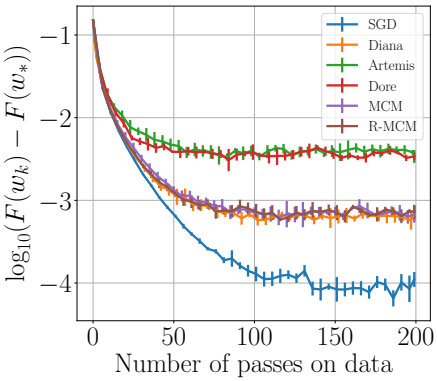


Figure 1.10: `MCM` and `Rand-MCM` on quantum, they achieves a rate of convergence identical to unidirectional compression.

Theorem 1.7 (Convergence in the quadratic case)**.** *Under Assumptions 1.2 to 1.4 and 1.7 with $\mu = 0$, if the function is quadratic, after running $K > 0$ iterations, for any $\gamma \leq \gamma_{\max}$, we have:*

$$\mathbb{E}[F(\bar{w}_K) - F_*] \leq \frac{\|w_0 - w_*\|^2}{\gamma K} + \frac{\gamma\sigma^2(1+\omega_{\rm up})}{N}\left(1 + \frac{4\gamma^2 L^2\omega_{\rm dwn}}{K}\left(\frac{1}{\mathbf{C}} + \frac{\omega_{\rm up}}{N}\right)\right),$$

with $\mathbf{C} = N$ for `Rand-MCM` and $\mathbf{C} = 1$ for `MCM`.

1.5.3 Contributions of Chapter 4

The goal of this Chapter is to provide an in-depth analysis of compression within a fundamental learning framework, namely least-squares regression (see Subsection 1.1.4), in order to highlight the differences in convergence between several unbiased compression schemes having the *same* variance increase. More precisely, we consider linear stochastic approximation recursion, to find a zero of the linear mean field ∇F.

Definition 1.3 (Linear Stochastic Approximation, LSA)**.** *Let $w_0 \in \mathbb{R}^d$ be the initialization, the linear*[3] *stochastic approximation recursion is defined as:*

$$w_k = w_{k-1} - \gamma \nabla F(w_{k-1}) + \gamma \xi_k(w_{k-1} - w_*), \quad k \in \mathbb{N}, \tag{LSA}$$

where $\gamma > 0$ is the step size and $(\xi_k)_{k\in\mathbb{N}^*}$ is a sequence of i.i.d. zero-centered random fields that characterizes the stochastic oracle on $\nabla F(\cdot)$. For any $k \in \mathbb{N}^*$, we denote by $\mathcal{F}_k = \sigma(\xi_1, \dots, \xi_k)$, such that the filtration $(\mathcal{F}_k)_{k\geq 0}$ is adapted to $(w_k)_{k\geq 0}$.

We assume that F is quadratic, we denote H_F its Hessian. For any k in $\mathbb{N}$, with $\eta_k = w_k - w_*$, we get equivalently:

$$\eta_k = (\mathrm{I} - \gamma H_F)\eta_{k-1} + \gamma \xi_k(\eta_{k-1}), \quad k \in \mathbb{N}.$$

Although there is abundant literature on the study of (LSA), the application to the case of federated least-mean-squares poses novel challenges. Especially, most analyses of LSA assume that the field ξ_k is linear (i.e. for any $z, z' \in \mathbb{R}^d$, $\xi_k(z) - \xi_k(z') = \xi_k(z - z')$). More general non-asymptotic results on stochastic approximation with a Lipschitz mean field (i.e. SGD with a smooth objective) also assume that the noise-field is Lipschitz in squared expectation i.e. for any $z, z' \in \mathbb{R}^d, \mathbb{E}[\|\xi_k(z) - \xi_k(z')\|^2] \leq C\|z - z'\|^2$. One major specificity and bottleneck in the case of compression is the fact that the resulting field **does not** satisfy such an assumption. Instead, we consider the following Hölder-type assumption on the compressor:

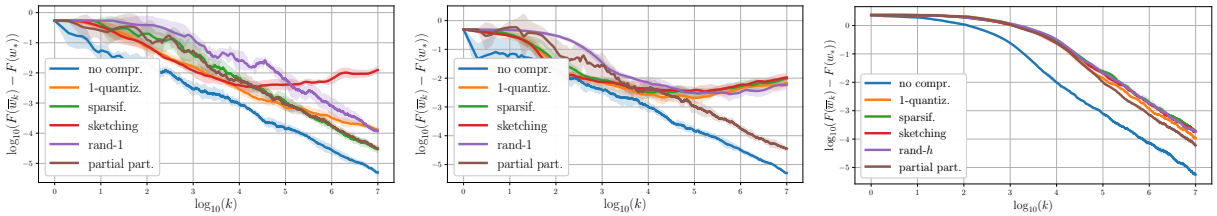
Assumption 1.9 (Compression.)**.** *We suppose that there exists two constants $\omega, \Omega \in \mathbb{R}_+^*$, such that the random operator $\mathcal{C}$ satisfies for all z in $\mathbb{R}^d$ the following property:*

$$\mathbb{E}[\|\mathcal{C}(z) - \mathcal{C}(z')\|^2] \leq \Omega \min(\|z\|, \|z'\|)\|z - z'\| + 3(\omega + 1)\|z - z'\|^2 .$$

It enables to provide a non-asymptotic analysis of (LSA) under weak regularity assumptions of the noise field $(\xi_k)_k$. We show that the asymptotically dominant term depends on the covariance matrix $\mathfrak{C}_{\text{ania}}$ of the *additive noise induced by the algorithm* (nicknamed the *ania's covariance*), as expected from the classical asymptotic literature Polyak and Juditsky [1992]. The backbone theorem of the chapter generalizes the results from Bach and Moulines [2013] obtained in the scenario of centralized LMS. It shows that the variance term scales with $\mathrm{Tr}\left(\mathfrak{C}_{\text{ania}} H_F^{-1}\right)$, which highlights the interaction between the Hessian of the optimization problem H_F, and the ania's covariance $\mathfrak{C}_{\text{ania}}$.

We then consider the simple configuration of compressed central LMS, it enables to describe the impact of the compressor choice on the dependency between the features' covariance H (which is also the Hessian H_F of the optimization problem) and the ania's covariance $\mathfrak{C}_{\text{ania}}$. Contrary to the classical scenario without compression for which the noise is said to be *structured*, i.e., the ania's covariance is proportional to the Hessian H_F, applying a random compression mechanism on the gradient breaks this structure. This phenomenon is noteworthy: for an ill-conditioned H_F, it may lead to a drastic increase in $\mathrm{Tr}\left(\mathfrak{C}_{\text{ania}} H_F^{-1}\right)$ and thus, to a degradation in convergence. By calculating the ania's covariance for various compression mechanisms, we identify differences that classical literature was unable to capture.

[3] While in LSA literature, both the mean-field ∇F and the noise-field (ξ_k) are linear, we do not here consider the noise fields to be linear.



(a) Features' covariance diagonal with high eigenvalues' decay.

(b) Features' covariance not diagonal with high eigenvalues' decay.

(c) Features' covariance not diagonal with slow eigenvalues' decay.

Figure 1.11: Logarithm excess loss of the Polyak-Ruppert iterate after $K = 10^7$ iterations for a single client ($N = 1$).

For instance, we show that Rand-h and partial participation (see definitions in Chapter 4) with probability (h/d) satisfy the same variance condition. Yet the convergence of compressed least-mean-squares for PP is more robust to ill-conditioned problems. To illustrate these findings, we run a gradient descent on a LSR problem and plot the logarithm excess loss of the Polyak-Ruppert iterate on Figure 1.11 in three scenarios: when the features' covariance is diagonal or not in the case of an ill-conditioned problem, i.e., with high eigenvalues' decay (Figures 1.11a and 1.11b, $\mu = 10^{-8}$); when the features' covariance is not diagonal and with slow eigenvalues' decay (Figure 1.11c, $\mu = 10^{-2}$).

Finally, we study the case of federated learning with heterogeneous clients. We examine two different sources of heterogeneity. First, the case of heterogeneous features' covariances $(H_i)_{i=1}^N$ (covariate-shift), second, the case of heterogeneous local optimal points $(w_*^i)_{i=1}^N$ (concept-shift). In the covariate-shift case, most insights from the centralized case remain valid and we explain how to compute the ania's covariance. On the contrary, despite that the concept-shift scenario keeps the noise structured (without compression), it hinders the limit convergence rate, suffering from the dispersion of the optimal points.

1.5.4 Key messages of this thesis

Throughout the chapters of this thesis, three key take-away messages can be identified.

1. The relationship between compression and heterogeneity is non-trivial. Our research has shown that the primary factor that affects convergence is the noise on the gradient computed on the optimal point. Based on this finding, we have developed an algorithm that performs variance reduction for compressed SGD in scenarios where clients are heterogeneous. Our algorithm is specifically designed to cancel the impact of heterogeneity and improve the accuracy of compressed SGD in these scenarios. With our algorithm, we aim to address the challenges posed by the interaction between compression and heterogeneity.
2. During downlink compression, two quantities are typically observed: x in $\mathbb{R}^d$ and $\mathcal{C}_{\text{dwn}}(x)$, however, only one of these quantities $\mathcal{C}_{\text{dwn}}(x)$ is transmitted. If $\mathcal{C}_{\text{dwn}}(x)$ is used to update the central model, as in `Artemis`-like algorithms, it results in an increase by a factor ω_{dwn} of the variance. If x is used to update the central model, it implies that the local models are different from the central model. This leads to compute the local gradients on a "perturbed model", which is more challenging. Taking advantage of this scenario, we have developed an algorithm that asymptotically cancels the impact of downlink compression.
3. Compression has a significant impact on the regularity of the optimization problem. For example, quantization is neither linear nor Lipschitz in squared expectation. Despite this challenge, we have conducted an analysis in the fundamental learning framework of LSR. Using a Hölder-type condition, we have identified differences in convergence rates between several unbiased compression operators that all satisfy the same condition on their variance, thus going beyond the classical worst-case analysis.

2

`Artemis`: tight convergence guarantees for bidirectional compression with heterogeneous clients

"C'est la nuit qu'il est beau de croire à la lumière."

Chantecler, Edmond Rostand.

In this Chapter, we focus on the intertwined effect of compression and client (statistical) heterogeneity. We introduce a framework – `Artemis` – to tackle the problem of learning in a distributed or federated setting with communication constraints. Several clients perform the optimization process using a central server to aggregate their computations. To alleviate the communication cost, `Artemis` allows to compress the information sent in *both directions* (from the clients to the server and conversely) combined with a memory mechanism. It improves on existing algorithms that only consider unidirectional compression (to the server), or use very strong assumptions on the compression operator. We provide fast rates of convergence (linear up to a threshold) under weak assumptions on the stochastic gradients (noise's variance bounded only *at optimal point*) in non-i.i.d. setting, highlight the impact of memory for unidirectional and bidirectional compression, and analyze Polyak-Ruppert averaging. We use convergence in distribution to obtain a *lower bound* of the asymptotic variance that highlights the practical limits of compression. We provide experimental results to demonstrate the validity of our analysis.

This chapter is based on our work *Bidirectional compression in heterogeneous settings for distributed or federated learning with partial participation: tight convergence guarantees* [Philippenko and Dieuleveut, 2021].

Contents

2.1 Introduction

In modern large scale machine learning applications, optimization has to be processed in a distributed fashion, using a potentially large number N in $\mathbb{N}$ of clients. In the data-parallel framework, each client only accesses a fraction of the data: new challenges have arisen, especially when communication constraints between the workers are present.

In this chapter, we focus on first-order methods, especially stochastic gradient descent [Bottou, 1999, Robbins and Monro, 1951] in a centralized framework: a central machine aggregates the computation of the N workers in a synchronized way. This applies to both the *distributed* [e.g. Li et al., 2014] and the *federated learning* [introduced in Konečný et al., 2016, McMahan et al., 2017] settings.

Formally, we consider a number of features $d \in \mathbb{N}^*$, and a convex cost function $F : \mathbb{R}^d \to \mathbb{R}$. We want to solve the following convex optimization problem:

$$\min_{w \in \mathbb{R}^d} F(w) \text{ with } F(w) = \frac{1}{N} \sum_{i=1}^{N} F_i(w)\,, \tag{2.1}$$

where $(F_i)_{i=1}^N$ is a *local* risk function for the model w on the worker i. Especially, in the classical supervised machine learning framework, we fix a loss ℓ and access, on a worker i, n_i observations $(z_k^i)_{1 \leq k \leq n_i}$ following a distribution $\mathcal{D}_i$. In this framework, F_i can be either the (weighted) local empirical risk, $w \mapsto (n_i^{-1}) \sum_{k=1}^{n_i} \ell(w, z_k^i)$ or the expected risk $w \mapsto \mathbb{E}_{z \sim \mathcal{D}_i}[\ell(w, z)]$. At each iteration of the algorithm, each client can get an *unbiased oracle* on the gradient of the function F_i (typically either by choosing uniformly an observation in its dataset or in a *streaming fashion*, getting a new observation at each step).

Our goal is to reduce the amount of information exchanged between workers, to accelerate the learning process, limit the bandwidth usage, and reduce energy consumption. Indeed, the communication cost has been identified as an important bottleneck in the distributed settings [e.g. Strom, 2015]. In their overview of the federated learning framework, Kairouz et al. [2019] also underline in Section 3.5 two possible directions to reduce this cost: (1) compressing communication from workers to the central server (uplink) (2) compressing the downlink communication.

Most of the papers considering the problem of reducing the communication cost [Alistarh et al., 2017, Agarwal et al., 2018, Wu et al., 2018, Karimireddy et al., 2019, Mishchenko et al., 2019, Horváth et al., 2022, Li et al., 2020b, Horváth and Richtárik, 2020] only focus on compressing the message sent from the workers to the central node. This direction has the highest potential to reduce the total runtime given that (i) the bandwidth for upload is generally more limited than for download, and that (ii) for some regimes with a large number of workers, the downlink communication, that

corresponds to a "one-to-N" communication, may not be the bottleneck compared to the "N-to-one" uplink.

Nevertheless, there are several reasons to also consider downlink compression. First, the difference between upload and download speeds is not significant enough at all to ignore the impact of the downlink direction (see Section 1.4 for an analysis of bandwidth). If we consider for instance a small number N of workers training a very heavy model – the size of Deep Learning models generally exceeds hundreds of MB [Dean et al., 2012, Huang et al., 2019] –, the training speed will be limited by the exchange time of the updates, thus using downlink compression is key to accelerating the process. Secondly, in a different framework in which a network of smartphones collaborate to train a large scale model in a federated framework, participants to the training would not be eager to download a hundreds of MB for each update on their phone. Here again, downlink compression appears to be necessary. To encompass all situations, our framework implements compression in either or both directions with possibly different compression levels.

Bidirectional compression (i.e. compressing both uplink and downlink) raises new challenges. In the downlink step, if we compress the *model*, the quantity compressed does *not* tend to zero. Consequently the compression error significantly hinders convergence. To circumvent this problem we compress the *gradient* that may asymptotically approach zero. Prior to this work, bidirectional compression had been considered by Tang et al. [2019], Zheng et al. [2019], Liu et al. [2020], Yu et al. [2019]. In particular, Liu et al. [2020] developed (concomitantly and independently to our work) an algorithm called `Dore`, which combines error compensation, a memory mechanism, and model compression, and assumes a uniform bound on the gradient variance. In this chapter, we provide new results on `Dore`-like algorithms, considering a framework *without error-feedback* using tighter assumptions, and quantifying precisely the impact of data heterogeneity on the convergence.

Indeed, we focus on a *heterogeneous* setting: the data distribution depends on each worker (thus non i.i.d.). We *explicitly control the differences between distributions*. In such a setting, the local gradient at the optimal point $\nabla F_i(w_*)$ may not vanish: to get a vanishing compression error, we introduce a "memory" process [Mishchenko et al., 2019].

Assumptions made on the gradient oracle directly influence the convergence rate of the algorithm: in this Chapter, we neither assume that the gradients are uniformly bounded [as in Zheng et al., 2019] nor that their variance is uniformly bounded [Assumption 1.4, as in Alistarh et al., 2017, Mishchenko et al., 2019, Liu et al., 2020, Tang et al., 2019, Horváth et al., 2022]: instead we only assume that the variance is bounded by a constant σ_*^2 *at the optimal point* w_*, and provide linear convergence rates up to a threshold *proportional to* σ_*^2 (as in [Dieuleveut et al., 2020, Gower et al., 2019] for non distributed optimization). This is a fundamental difference as the variance bound at the optimal point can be orders of magnitude smaller than the uniform bound used in previous work: this is striking when all loss functions have the same critical point, and thus the noise at the optimal point is null! This happens for example in the *interpolation regime*, which has recently gained importance in the machine learning community [Belkin et al., 2019]. As the empirical risk at the optimal point is null or very close to zero, so are all the loss functions with respect to one example. This is often the case in deep learning [e.g., Zhang et al., 2017] or in large dimension regression [Mei and Montanari, 2019].

Overall, we make the following contributions:

1. We describe a framework – `Artemis` – **that encompasses 6 algorithms** (with or without up/down compression, with or without memory). We provide and analyze in Theorem 2.1 a fast rate of convergence – exponential convergence up to a threshold proportional to σ_*^2, the noise at the optimal point –, **obtaining tighter bounds** than in [Alistarh et al., 2017, Mishchenko et al., 2019].
2. We explicitly tackle heterogeneity using Assumption 2.4, proving that the limit variance of `Artemis` with memory is independent from the difference between distributions (as for SGD).

Table 2.1: Comparison of frameworks for main algorithms handling (bidirectional) compression. By "non i.i.d.", we mean that the theoretical framework encompasses *and* explicitly quantifies the impact of data heterogeneity on convergence (Assumption 2.4), e.g., `Dore` does not assume i.i.d. workers but does not quantify differences between distributions. References: see Alistarh et al. [2017] for `QSGD`, Mishchenko et al. [2019] for `Diana`, Horváth and Richtárik [2020] for [HR20], Liu et al. [2020] for `Dore` and Tang et al. [2019] for `DoubleSqueeze`.

	`QSGD`	`Diana`	[HR20]	`Dore`	`Double Squeeze`	`Dist EF-SGD`	`Artemis` **(new)**
Data	i.i.d.	non i.i.d.	non i.i.d.	i.i.d.	i.i.d.	i.i.d.	non i.i.d.
Bounded variance	Uniformly	Uniformly	Uniformly	Uniformly	Uniformly	Uniformly	At optimal point
Compression	One-way	One-way	One-way	Two-way	Two-way	Two-way	Two-way
Error-feedback			✓	✓	✓	✓	
Memory		✓		✓			✓
Partial part.			✓				✓

This is the first theoretical guarantee for double compression that explicitly quantifies the impact of non i.i.d. data.

3. In the non-strongly-convex case, we prove the convergence using Polyak-Ruppert averaging in Theorem 2.2.
4. We prove *convergence in distribution* of the iterates, and subsequently **provide *a lower bounds* on the asymptotic variance.** This sheds light on the limits of (double) compression, which results in an increase of the algorithm's variance, and can thus only accelerate the learning process for *early iterations* and up to a *"moderate" accuracy*. Interestingly, this *"moderate" accuracy* has to be understood with respect to the *reduced* noise σ_*^2.

Furthermore, we support our analysis with various experiments illustrating the behavior of our new algorithm and we provide the code to reproduce our experiments, see this repository. In Table 2.1, we highlight the main features and assumptions of `Artemis` compared to recent algorithms using compression.

The rest of the chapter is organized as follows: in Section 2.2 we introduce the framework of `Artemis`. In Subsection 2.2.1 we describe the assumptions, and we review related work in Subsection 2.2.2. We then give the theoretical results in Section 2.3, we present experiments in Section 2.4, and finally, we conclude in Section 2.5.

2.2 Problem statement

We consider the problem described in Equation (2.1). In the convex case, we assume that there exist at least one optimal point which we denote w_*, we also denote $h_*^i = \nabla F_i(w_*)$, for i in $[\![1, N]\!]$. To solve this problem, we rely on a stochastic gradient descent (SGD) algorithm.

A stochastic gradient g_k^i is provided at iteration k in $\mathbb{N}^*$ to the client i in $[\![1, N]\!]$. This function is then evaluated at point w_{k-1}: to alleviate notation, we will use $g_k^i = g_k^i(w_{k-1})$ and $g_{k,*}^i = g_k^i(w_*)$ to denote the stochastic gradient vectors at points w_{k-1} and w_* on client i. In the classical centralized framework (without compression), SGD corresponds to:

$$w_k = w_{k-1} - \gamma \frac{1}{N} \sum_{i=1}^{N} g_k^i \tag{2.2}$$

where γ is the learning rate.

However, computing such a sequence would require the nodes to send either the gradient g_k^i or the updated local model to the central server (*uplink* communication), and the central server to broadcast back either the averaged gradient g_k or the updated global model (*downlink* communication). Here, in order to reduce communication cost, we perform a *bidirectional* compression. More precisely, we combine two main tools: (1) an *unbiased compression operator* $\mathcal{C} : \mathbb{R}^d \to \mathbb{R}^d$ that reduces the number of bits exchanged, and (2) a *memory* process that reduces the size of the signal to compress, and consequently the error [Mishchenko et al., 2019, Li et al., 2020b]. That is, instead of directly compressing the gradient, we first approximate it by the memory term and, afterwards, we compress the difference. As a consequence, the compressed term tends in expectation to zero, and the error of compression is reduced. Following Tang et al. [2019], we only broadcast gradients, or difference of gradients, and never models. To distinguish the two compression operations we denote $\mathcal{C}_{\text{up}}$ and $\mathcal{C}_{\text{dwn}}$ the compression operator for uplink and downlink. At each iteration, we thus have the following steps:

1. First, each active local node sends to the central server a compression of gradient differences: $\widehat{\Delta}_k^i = \mathcal{C}_{\text{up}}(g_k^i - h_{k-1}^i)$, and updates the *memory term* $h_k^i = h_{k-1}^i + \alpha_{\text{up}}\widehat{\Delta}_k^i$ with $\alpha_{\text{up}} \in \mathbb{R}^*$. The server recovers the approximated gradients' values by adding the received term to the memories kept on its side.
2. Then, the central server sends back the compression of the sum of compressed gradients: $\Omega_k = \mathcal{C}_{\text{dwn}}\left(\frac{1}{N}\sum_{i=1}^N \widehat{\Delta}_k^i + h_{k-1}^i\right)$. No memory mechanism needs to be used, as the sum of gradients tends to zero in the absence of regularization.

The update is thus given by:

$$\begin{cases} \forall i \in [\![1, N]\!], \quad \widehat{\Delta}_k^i = \mathcal{C}_{\text{up}}\left(g_k^i - h_{k-1}^i\right) \\ \Omega_k = \mathcal{C}_{\text{dwn}}\left(\frac{1}{N}\sum_{i=1}^N (\widehat{\Delta}_k^i + h_{k-1}^i)\right) \\ w_k = w_{k-1} - \gamma\Omega_k \,. \end{cases} \tag{2.3}$$

Constants $\gamma, \alpha_{\text{up}} \in \mathbb{R}^* \times \mathbb{R}_+$ are learning rates for respectively the iterate sequence and the memory sequence.

Partial participation. As underlined in Subsection 1.3.1, an important setting of FL is the partial participation (PP) of clients at each round: clients only participate in a fraction p of the training steps. This can be addressed theoretically by modelizing it as a compression scheme $\mathcal{C}_{\text{PP}}$, which compresses a vector z as either z/p or 0. As such, our analysis of uplink compression naturally encompasses the PP scenario. In the PP setting, the main difficulty is to keep all clients synchronized when they return to the training process. This requires sharing any updates they missed or the latest iterate, depending on which option is more efficient. This step is commonly referred to as a "catching-up" process. This approach has also been proposed by Sattler et al. [2019, see the remark preceding Equation (20) in Section VI.C] or by Tang et al. [2019, v2 on arxiv for the distributed case], who use a buffer. We present the pseudo-code of `Artemis` with the catching-up step in Algorithm 2.

As a summary, the `Artemis` framework encompasses, in particular, these four algorithms: the variant with unidirectional compression ($\omega_{\text{dwn}} = 0$) w.o. or with memory ($\alpha_{\text{up}} = 0$ or $\alpha_{\text{up}} \neq 0$) recovers `QSGD` defined by Alistarh et al. [2017] and `DIANA` proposed by Mishchenko et al. [2019]. The variant using bidirectional compression ($\omega_{\text{dwn}} \neq 0$) w.o memory ($\alpha_{\text{up}} = 0$) is called `Bi-QSGD`. The last and most effective variant combines bidirectional compression *with* memory and is the one we refer to as `Artemis` if no precision is given. It corresponds to a simplified version of `Dore` without error-feedback, but this additional mechanism did not lead to any theoretical improvement in the case of unbiased compressors [Remark 2 in Sec. 4.1., Liu et al., 2020].

Remark 2.1 (Local steps)**.** *An obvious independent direction to reduce communication is to increase the number of steps performed before communication. This is the spirit of Local-SGD [Stich, 2019].*

Algorithm 2: Pseudocode of `Artemis` – set $\alpha_{\text{up}} > 0$ to use memory.

Input: Mini-batch size b, learning rates $\alpha_{\text{up}}, \gamma > 0$, initial model $w_0 \in \mathbb{R}^d$, operators $\mathcal{C}_{\text{up}}$ and $\mathcal{C}_{\text{dwn}}$, M_1 and M_2 the sizes of the full/compressed gradients.

Initialization: Index of last participation: $k_i = 0$. Local memory: $\forall i \in \{1, \dots, N\}, h_0^i = g_1^i(w_0)$ (smart initialization). Central memory: $h_0 = \sum_{i=1}^N h_0^i / N$.

Output: Model w_K

for $k = 1, 2, \dots, K$ **do**

- Get the set of active devices $S_k \subset \{1, \dots, N\}$
- **for** *each device* $i \in S_k$ **do**
  - **Catching up.**
  - If $k - k_i > \lfloor M_1/M_2 \rfloor$, send the model w_{k-1}
  - Else receive $(\widehat{\Omega}_j)_{j=k_i+1}^k$ and update local model: $\forall j \in [\![k_i + 1, k]\!], \quad w_j = w_{j-1} - \gamma \Omega_j$
  - Update index of its last participation: $k_i = k$
  - **Local training.**
  - Compute stochastic gradient $g_k^i = g_k(w_{k-1})$ (with mini-batch)
  - Set $\Delta_k^i = g_k^i - h_{k-1}^i$, compress it $\widehat{\Delta}_k^i = \mathcal{C}_{\text{up}}(\Delta_k^i)$
  - Update memory term: $h_k^i = h_{k-1}^i + \alpha_{\text{up}} \widehat{\Delta}_k^i$
  - Send $\widehat{\Delta}_k^i$ to central server
- Compute $\widehat{g}_k = h_{k-1} + \frac{1}{N} \sum_{i=1}^N \widehat{\Delta}_k^i$
- Update central memory: $h_k = h_{k-1} + \alpha_{\text{up}} \frac{1}{N} \sum_{i=1}^N \widehat{\Delta}_k^i$
- Back compression: $\Omega_k = \mathcal{C}_{\text{dwn}}(\widehat{g}_k)$
- Broadcast Ω_k to all workers.
- Update model on central server: $w_k = w_{k-1} - \gamma \Omega_k$

It is an interesting extension to incorporate this into our framework. We do not consider it in order to focus on the compression insights.

In the following section, we present and discuss assumptions over the function F, the data distribution and the compression operator.

2.2.1 Assumptions

We make classical assumptions on $F : \mathbb{R}^d \to \mathbb{R}$.

Assumption 2.1 (Strong-convexity)**.** *F is μ-strongly-convex, that is for all vectors z, z' in $\mathbb{R}^d$: $F(z') \geq F(w) + (z' - z)^T \nabla F(z) + \frac{\mu}{2} \|z' - z\|_2^2$.*

Note that we do not need each F_i to be strongly convex, but only F. Also remark that we only use this inequality for $z' = w_*$ in the proof of Theorems 2.1 and 2.2.

Below, we introduce cocoercivity [see Zhu and Marcotte, 1996, for more details about this hypothesis]. This assumption implies that all $(F_i)_{i \in [\![1,N]\!]}$ are L-smooth.

Assumption 2.2 (Cocoercivity of stochastic gradients in quadratic mean)**.** *We suppose that for all k in $\mathbb{N}$, stochastic gradients functions $(g_k^i)_{i \in [\![1,N]\!]}$ are L-cocoercive in quadratic mean. That is, for k in $\mathbb{N}$, i in $[\![1, N]\!]$ and for all vectors z, z' in $\mathbb{R}^d$, we have:*

$$\mathbb{E}[\|g_k^i(z) - g_k^i(z')\|^2] \leq L \left\langle \nabla F_i(z) - \nabla F_i(z'), z - z' \right\rangle .$$

E.g., this is true under the much stronger assumption that stochastic gradients functions $(g_k^i)_{i\in[\![1,N]\!]}$ are *almost surely* L-cocoercive, i.e.: $\|g_k^i(z) - g_k^i(z')\|^2 \leq L\left\langle g_k^i(z) - g_k^i(z'), z - z'\right\rangle$. Next, we present the assumption on the stochastic gradient's noise. Again, we highlight that the noise is only controlled at the optimal point. To carefully control the noises process (gradient oracle, uplink, and downlink compression), we introduce three filtrations $(\mathcal{H}_k, \mathcal{G}_k, \mathcal{F}_k)_{k\geq 0}$, such that w_k is $\mathcal{H}_k$-measurable for any $k \in \mathbb{N}$. Detailed definitions are given in Section B.2.

Assumption 2.3 (Noise over stochastic gradients computation)**.** *The noise over stochastic gradients at the global optimal point, for a mini-batch of size b, is bounded: there exists a constant $\sigma_* \in \mathbb{R}$, s. t. for all k in $\mathbb{N}$, for all i in $[\![1, N]\!]$, we have a.s.:*$\mathbb{E}[\|g_{k,*}^i - \nabla F_i(w_*)\|^2 | \mathcal{H}_{k-1}] \leq \frac{\sigma_*^2}{b}$.

The constant σ_*^2 is null, for example, if we use deterministic (batch) gradients, *or* in the interpolation regime for i.i.d. observations, as discussed in the Introduction of this Chapter. As we have also incorporated here a mini-batch parameter, this reduces the variance by a factor b.

Unlike `Diana` [Mishchenko et al., 2019, Li et al., 2020b], `Dore` [Liu et al., 2020], `Dist-EF-SGD` [Zheng et al., 2019] or `Double-Squeeze` [Tang et al., 2019], we assume that the variance of the noise is bounded *only at optimal point* w_* and not *at any point* w in $\mathbb{R}^d$. It results that if the variance is null ($\sigma_*^2 = 0$) at the optimal point, we obtain a linear convergence while previous results obtain this rate solely if the variance is null *at any point* (i.e. only for deterministic GD). Also remark that Assumptions 2.2 and 2.3 both stand for the simplest Least-Square Regression (LSR) setting, while the uniform bound on the gradient's variance *does not*. Next, we give the assumption that links the distributions on the different machines.

Assumption 2.4 (Bounded gradient at w_*)**.** *There exists a constant $B \in \mathbb{R}_+$, s.t.:*

$$\frac{1}{N}\sum_{i=0}^{N} \|\nabla F_i(w_*)\|^2 = B^2 \,.$$

This assumption is used to quantify how different the distributions are on the different clients. In the streaming *i.i.d. setting* – $D_1 = \cdots = D_N$ and $F_1 = \cdots = F_N$ – the assumption is satisfied with $B = 0$. Combining Assumptions 2.3 and 2.4 results in an upper bound on the averaged squared norm of stochastic gradients at w_*: for all k in $\mathbb{N}$, we have a.s. $\frac{1}{N}\sum_{i=1}^N \mathbb{E}[\|g_{k,*}^i\|^2 | \mathcal{H}_{k-1}] \leq \frac{\sigma_*^2}{b} + B^2$. In fact, Assumption 2.4 only requires that for any $i \in [\![1, N]\!]$, $\mathbb{E}[\|g_{k,*}^i - \nabla F_i(w_*)\|^2 \mid \mathcal{H}_{k-1}] \leq \frac{\sigma_{*,i}^2}{b}$, and the results then hold for $\sigma_* = \frac{1}{N}\sum_{i=1}^N \sigma_{*,i}^2$. In other words, the bounds do not need to be uniform over workers, only the average truly matters.

Finally, compression operators can be classified in two main categories: quantization [as in Rabbat and Nowak, 2005, Alistarh et al., 2017, Seide et al., 2014, Zhou et al., 2018, Wen et al., 2017, Reisizadeh et al., 2020, Horváth et al., 2022] and random projection [as in Vempala, 2005, Rahimi and Recht, 2008, Stich et al., 2018, Alistarh et al., 2018, Khirirat et al., 2020b]. Theoretical guarantees provided in this chapter do not rely on a particular kind of compression, as we only consider the following assumption on the compression operators $\mathcal{C}_{\text{up}}$ and $\mathcal{C}_{\text{dwn}}$:

Assumption 2.5. *There exist two constants $\omega_{\text{up}}, \omega_{\text{dwn}} \in \mathbb{R}_+^*$, such that for* dir $\in \{\text{up}, \text{dwn}\}$*, the compression operators $\mathcal{C}_{\text{dir}}$ verify the two following properties for all z in $\mathbb{R}^d$:*

$$\begin{cases} \mathbb{E}[\mathcal{C}_{\text{dir}}(z)] = z \,, \\ \mathbb{E}[\|\mathcal{C}_{\text{dir}}(z) - z\|^2] \leq \omega_{\text{dir}} \|z\|^2 \,. \end{cases}$$

In other words, the compression operators are unbiased and their variances are relatively bounded. Note that Horváth and Richtárik [2020] have shown that using an unbiased operator leads to better performances. Unlike us, Tang et al. [2019] assume uniformly bounded compression error, which is

a much more restrictive assumption. We now provide additional details on related papers dealing with compression. Also note that $\omega_{\mathrm{up/dwn}}$ can be considered as *parameters* of the algorithm, as the compression levels can be chosen.

Remark 2.2 (I.i.d. compressions)**.** *Assumption 2.5 requires in fact to access a* sequence of i.i.d. compression operators $\mathcal{C}_{\mathrm{up/dwn},k}$ *for* $k \in \mathbb{N}$*; but for simplicity, we generally omit the* k *index.*

2.2.2 Related work on compression

Quantization is a common method for compression and is used in various algorithms. For instance, Seide et al. [2014] are one of the first to propose to quantize each gradient component by either -1 or 1. This approach has been extended in Karimireddy et al. [2019]. Alistarh et al. [2017] define a new algorithm – `QSGD` – which instead of sending gradients, broadcasts their quantized version, getting robust results with this approach. On top of gradient compression, Wu et al. [2018] add an error-compensation mechanism that accumulates quantization errors and corrects the gradient computation at each iteration. In the case of quadratic problems, Khirirat et al. [2020a] have further shown that contrary to the simple error-compensation mechanism, it is possible, when considering compressors with uniformly bounded variance, to remove all of the accumulated error using instead a Hessian-aided error compensation mechanism. `Diana` [introduced in Mishchenko et al., 2019] introduces a "memory" term in the place of accumulating errors. Li et al. [2020b] extend this algorithm and improve its convergence by using an accelerated gradient descent. Reisizadeh et al. [2020] combine unidirectional quantization with client sampling, leading to a framework closer to federated learning settings where clients can easily be switched off. In the same perspective, Horváth and Richtárik [2020] detail results that also consider PP. Tang et al. [2019] are the first to suggest a bidirectional compression scheme for a decentralized network. For both uplink and downlink, the method consists in sending a compression of gradients combined with an error compensation. Later, Yu et al. [2019] choose to compress models instead of compressing gradients. This approach is enhanced by Liu et al. [2020] who combine model compression with a memory mechanism and an error compensation drawing from Mishchenko et al. [2019]. Both Tang et al. [2019] and Zheng et al. [2019] compress gradients without using a memory mechanism. However, as proved in the following Section, memory is key to reducing the asymptotic variance in the heterogeneous case. Beyond compressing down communications, Grishchenko et al. [2021] proposed an algorithm that reduces the down communication by using a proximal operator which, combined with a ℓ_1-regularisation and a sparsification of ascending communications, produces sparse iterates after some steps of communication.

We now provide theoretical results about the convergence of bidirectional compression.

2.3 Theoretical results

In this Section, we present our main theoretical results on the convergence of `Artemis` and its variants. To ensure clarity, the most complete and tightest versions of theorems are given in Appendices, while offering simplified versions here.

The main linear convergence rates are given in Theorem 2.1, and in Theorem 2.2 we show that `Artemis` combined with Polyak-Ruppert averaging reaches a sub-linear convergence rate. We denote $\delta_0^2 = \|w_0 - w_*\|^2$.

Theorem 2.1 (Convergence of `Artemis`)**.** *Under Assumptions 2.1 to 2.5, for a step-size* γ *satisfying the conditions in Table 2.3, for a learning rate* α_{up} *and for any* k *in* $\mathbb{N}$*, the mean squared distance to*

Table 2.2: Details on constants C and E defined in Theorem 2.1. $C = 0$ for $\alpha_{\text{up}} = 0$, see Th. B.2 for $\alpha_{\text{up}} \neq 0$.

α_{up}	E
0	$(\omega_{\text{dwn}}+1)\left((\omega_{\text{up}}+1)\frac{\sigma_*^2}{b}+\omega_{\text{up}}B^2\right)$
$\neq 0$	$\frac{\sigma_*^2}{b}\left((2\omega_{\text{up}}+1)(\omega_{\text{dwn}}+1)+4\alpha_{\text{up}}^2 C(\omega_{\text{up}}+1)-2\alpha_{\text{up}}C\right)$

w_* decreases at a linear rate up to a constant of the order of E:

$$\mathbb{E}\left[\|w_k - w_*\|^2\right] \leq (1-\gamma\mu)^k\left(\delta_0^2 + 2C\gamma^2 B^2\right) + \frac{2\gamma E}{\mu N},$$

for constants C and E depending on the variant (independent of k) given in Table 2.2 or in the appendix. Variants with $\alpha_{\text{up}} \neq 0$ require $\alpha_{\text{up}} \in [1/2(\omega_{\text{up}}+1), \alpha_{\max}]$, the upper bound $\alpha_{\max}$ is given in Theorem B.2.

This theorem is derived from Theorems B.1 and B.2 which are respectively proved in Subsections B.4.1 and B.4.2.

We can make the following remarks:

1. **Linear convergence.** The convergence rate given in Theorem 2.1 can be decomposed into two terms: a bias term, forgotten at linear speed $(1-\gamma\mu)^k$, and a variance residual term which corresponds to the *saturation level* of the algorithm. The rate of convergence $(1-\gamma\mu)$ does not depend on the variant of the algorithm. However, the variance and initial bias do vary.
2. **Bias term.** The initial bias always depends on $\|w_0 - w_*\|^2$, and when using memory (i.e. $\alpha_{\text{up}} \neq 0$) it also depends on the difference between distributions (constant B^2).
3. **Variance term and memory.** On the other hand, the variance depends (1) on both σ_*^2/b, and the distributions' difference B^2 without memory (2) only on the gradients' variance *at the optimum* σ_*^2/b with memory. Similar theorems in related literature [Liu et al., 2020, Alistarh et al., 2017, Mishchenko et al., 2019, Yu et al., 2019, Tang et al., 2019, Zheng et al., 2019] systematically had a worse bound for the variance term depending on a *uniform bound of the noise variance* or under much stronger conditions on the compression operator. This work and [Liu et al., 2020] are also the first to give a linear convergence up to a threshold for bidirectional compression.
4. **Impact of memory.** To the best of our knowledge, this is the first work on double compression that explicitly tackles the non i.i.d. case. We prove that memory makes the saturation threshold independent of B^2 for `Artemis`.
5. **Variance term.** The variance term increases with a factor proportional to ω_{up} for the unidirectional compression, and proportional to $\omega_{\text{up}} \times \omega_{\text{dwn}}$ for bidirectional. This is the counterpart of compression, each compression resulting in a multiplicative factor on the noise. A similar increase in the variance appears in [Mishchenko et al., 2019] and [Liu et al., 2020]. The noise level is attenuated by the number of clients N, to which it is inversely proportional.
6. **Link with classical SGD.** For variant of `Artemis` with $\alpha_{\text{up}} = 0$, if $\omega_{\text{up/dwn}} = 0$ (i.e. no compression) we recover SGD results: convergence does not depend on B^2, but only on the noise's variance.

Conclusion: Overall, it appears that `Artemis` is able to efficiently accelerate the learning during first iterations, enjoying the same linear rate as SGD with lower communication complexity, but it saturates at a higher level, proportional to σ_*^2 and independent of B^2.

The range of acceptable learning rates is an important feature for first order algorithms. In Table 2.3, we summarize the upper bound $\gamma_{\max}$ on γ, to guarantee a $(1-\gamma\mu)$ convergence of `Artemis`.

Table 2.3: Upper bound on $\gamma_{\max}$ to guarantee convergence. For unidirectional compression (resp. no compr.), $\omega_{\rm dwn} = 0$ (resp. $\omega_{\rm up/dwn} = 0$, recovering classical rates for SGD).

Memory	$\alpha_{\rm up} = 0$	$\alpha_{\rm up} \neq 0$
$N \gg \omega_{\rm up}$	$\frac{1}{(\omega_{\rm dwn}+1)L}$	$\frac{1}{2(\omega_{\rm dwn}+1)L}$
$N \approx \omega_{\rm up}$	$\frac{1}{3(\omega_{\rm dwn}+1)L}$	$\frac{1}{5(\omega_{\rm dwn}+1)L}$
$\omega_{\rm up} \gg N$	$\frac{N}{2\omega_{\rm up}(\omega_{\rm dwn}+1)L}$	$\frac{N}{4\omega_{\rm up}(\omega_{\rm dwn}+1)L}$

These bounds are derived from Theorems B.1 and B.2, in three main asymptotic regimes: $N \gg \omega^{\rm up}$, $N \approx \omega^{\rm up}$ and $\omega^{\rm up} \gg N$. Using bidirectional compression impacts $\gamma_{\max}$ by a factor $\omega_{\rm dwn} + 1$ in comparison to unidirectional compression. For unidirectional compression, if the number of machines is at least of the order of $\omega_{\rm up}$, then $\gamma_{\max}$ nearly corresponds to $\gamma_{\max}$ for vanilla (serial) SGD.

We now provide a convergence guarantee for the averaged iterate without strong-convexity.

Theorem 2.2 (Convergence of `Artemis` with Polyak-Ruppert averaging)**.** *Under Assumptions 2.2 to 2.5 (convex case) with constants C and E as in Theorem 2.1 (see Table 2.2 for precision), after running K in $\mathbb{N}^*$ iterations, for a learning rate $\gamma = \min\left(\sqrt{\frac{N\delta_0^2}{2EK}}; \gamma_{\max}\right)$, with $\gamma_{\max}$ as in Table 2.3, we have a sublinear convergence rate for the Polyak-Ruppert averaged iterate $\overline{w}_{K-1} = \frac{1}{K}\sum_{k=0}^{K-1} w_k$:*

$$\mathbb{E}\left[F(\overline{w}_{K-1}) - F(w_*)\right] \leq 2\max\left(\sqrt{\frac{2\delta_0^2 E}{NK}}; \frac{\delta_0^2}{\gamma_{\max} K}\right) + \frac{2\gamma_{\max} C B^2}{K}\,.$$

This theorem is proved in Subsection B.4.3. Several comments can be made on this theorem:

1. **Importance of averaging** This is the first theorem given for averaging for double compression. In the context of convex optimization, averaging has been shown to be optimal [Rakhlin et al., 2012].

2. **Speed of convergence, if $\sigma_* = 0$, $B \neq 0$, $K \to \infty$.** For $\alpha_{\rm up} \neq 0$, $E = 0$, while for $\alpha_{\rm up} = 0$, $E \propto B^2$. Memory thus accelerates the convergence from a rate $O(K^{-1/2})$ to $O(K^{-1})$.

3. **Speed of convergence, general case.** More generally, we always get a $K^{-1/2}$ sublinear speed of convergence, and a faster rate K^{-1} when using memory and if $E \leq \delta_0^2 N/(2K\gamma_{\max}^2)$ – i.e. in the context of a low noise σ_*^2, as $E \propto \sigma_*^2$. Again, it appears that bi-compression is mostly useful in low-σ_*^2 regimes or during the first iterations: intuitively, for a fixed communication budget, while bi-compression allows to perform $\min\{\omega_{\rm up}, \omega_{\rm dwn}\}$-times more iterations, this is no longer beneficial if the convergence rate is dominated by $\sqrt{2\delta_0^2 E/NK}$, as E increases by a factor $\omega_{\rm up} \times \omega_{\rm dwn}$.

4. **Memoryless case, impact of minibatch.** In the variant of `Artemis` *without memory*, the asymptotic convergence rate is $\sqrt{2\delta_0^2 E/NK}$ with the constant $E \propto \sigma_*^2/b + B^2$: interestingly, it appears that in the case of non i.i.d. data ($B^2 > 0$), the *convergence rate saturates when the size of the mini-batch increases*: large mini-batches *do not help*. On the contrary, with memory, the variance is, as classically, reduced by a factor proportional to the size of the batch, without saturation.

The increase in the variance (in Item 3) is not an artifact of the proof: indeed we provide a corresponding (algorithm-specific) lower bound based on proving the existence of a limit distribution for the iterates of `Artemis`, and analyzing its variance, see Theorem 2.3 in next Section.

2.3.1 Convergence in distribution and lower bound

The increase in the variance (in Item 3) is not an artifact of the proof: we prove the existence of a limit distribution for the iterates of `Artemis`, and analyze its variance. More precisely, we show a linear rate of convergence for the distribution Θ_k of w_k (launched from w_0), w.r.t. the Wasserstein distance $\mathcal{W}_2$ [Villani, 2009]: this gives us a lower bound on the asymptotic variance. Here, we further assume that the compression operator $\mathcal{C}$ is *linear* (i.e. for any z, z' in $\mathbb{R}$, we have $\mathcal{C}(z) - \mathcal{C}(z') = \mathcal{C}(z - z')$, it is the case for instance for sparsification, sketching, rand-h, PP).

Theorem 2.3 (Convergence in distribution and lower bound on the variance)**.** *Under Assumptions 2.1 to 2.5, for $\gamma, \alpha_{\rm up}, E$ given in Theorem 2.1 and Table 2.3:*

1. *There exists a limit distribution $\pi_{\gamma,v}$ depending on the variant v of the algorithm, s.t. for any $k \geq 1$, $\mathcal{W}_2(\Theta_k, \pi_{\gamma,v}) \leq (1-\gamma\mu)^k C_0$, with C_0 a constant.*
2. *When k goes to infinity, the second order moment $\mathbb{E}[\|w_k - w_*\|^2]$ converges to $\mathbb{E}_{w\sim\pi_{\gamma,v}}[\|w - w_*\|^2]$, which is* lower bounded *by $\Omega(\gamma E/\mu N)$ as in Theorem 2.1 as $\gamma \to 0$, with E depending on the variant.*

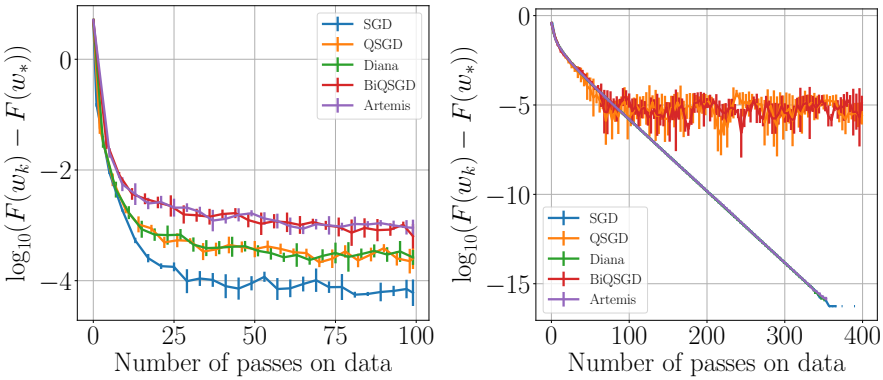
Interpretation. The second point (2.) means that the upper bound on the saturation level provided in Theorem 2.1 *is tight* w.r.t. $\sigma_*^2, \omega_{\rm up}, \omega_{\rm dwn}, B^2, N$ and γ. Especially, it proves that there is indeed a quadratic increase in the variance w.r.t. $\omega_{\rm up}$ and $\omega_{\rm dwn}$ when using bidirectional compression (which is itself rather intuitive). Altogether, these three theorems prove that bidirectional compression can become strictly worse than usual stochastic gradient descent in high precision regimes, a fact of major importance in practice and barely (if ever) even mentioned in previous literature. To the best of our knowledge, only Mayekar and Tyagi [2020] are giving a lower bound on the asymptotic variance for algorithms using compression. Their result is more general, i.e., valid for any algorithm using unidirectional compression, but weaker (worst case on the oracle does not highlight the importance of noise at the optimal point and is incompatible with linear rates).

Proof and assumptions. This theorem also naturally requires, for the second point, Assumptions 2.3 to 2.5 to be "tight": that is, e.g., $\mathrm{Var}(g_{k,*}^i) \geq \Omega(\sigma_*^2/b)$; more details and the proof are given in Subsection B.4.4. Extension to other types of compression reveals to be surprisingly non-simple, and is thus out of the scope of this chapter and a promising direction.

2.4 Experiments

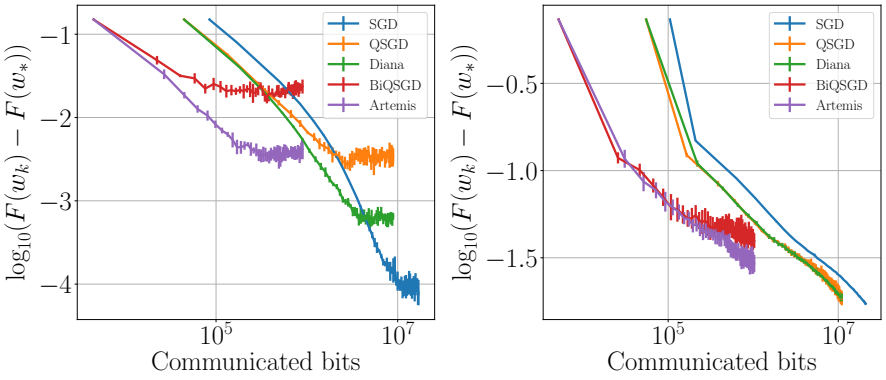
In this Section, we illustrate our theoretical guarantees on both synthetic and real datasets. The goal of this section is to confirm the theoretical findings in Theorems 2.1 to 2.3, and to underline the impact of the memory. Therefore, we focus on five of the algorithms covered by our framework: `Artemis` with bidirectional compression (simply denoted `Artemis`), `QSGD`, `Diana`, `Bi-QSGD`, and usual `SGD` without any compression. In the end of this Section, we compare `Artemis` with other existing benchmarks: `Double-Squeeze`, `Dore`, `FedSGD` and `FedPAQ` [see Reisizadeh et al., 2020]. In the Appendix, we also perform experiments with *optimized* learning rates (Figure B.13).

In all experiments, we display the logarithm excess loss $\log_{10}(F(w_{k-1}) - F(w_*))$ w.r.t. the number of iterations k or the number of communicated bits. For w in $\mathbb{R}$, in the case of linear regression, we have $F(w) = \frac{1}{N}\sum_{i=1}^N \sum_{(x,y)\in\mathcal{D}_i} x^\top w - y$, and in the case of logistic regression we have $F(w) = \frac{-1}{N}\sum_{i=1}^N \sum_{(x,y)\in\mathcal{D}_i} \log(\mathrm{Sigm}(yx^\top w))$. We use a quantization scheme (defined in Definition 1.2, see Chapter 1) with $s = 2^0$. Curves are averaged over 5 runs, we plot error bars on all figures. These error bars correspond to $\pm$ the standard deviation of the logarithm excess loss over the five runs. For each figure, the model is initialized at zero and we plot the corresponding excess loss such that all algorithms start at the same point.



(a) LSR (i.i.d.): $\sigma_*^2 \neq 0$ (b) LR (non-i.i.d.): $\sigma_* = 0$.

Figure 2.1: Left: illustration of the saturation when $\sigma_* \neq 0$ and data is i.i.d., right: illustration of the memory benefits when $\sigma_* = 0$ but with non-i.i.d. data.



(a) Quantum (b) Superconduct

Figure 2.2: **Real dataset** (non-i.i.d.): $\sigma_* \neq 0$, $N = 20$ workers, $p = 1$, $b > 1$ (150 iter.). X-axis in # bits.

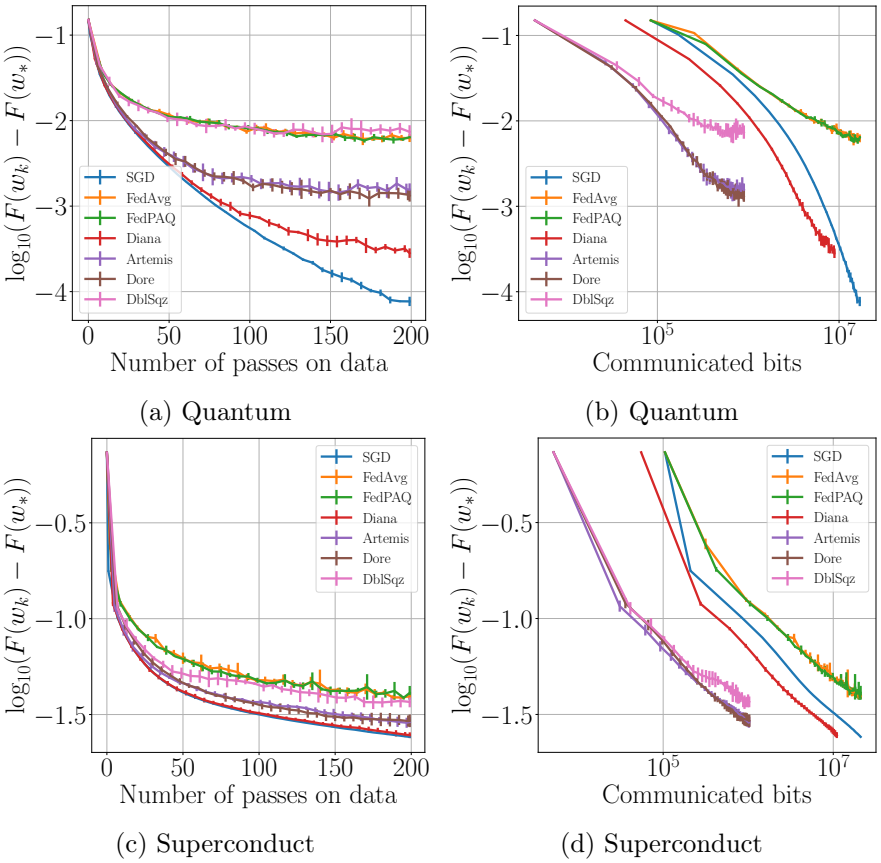
We first consider two simple synthetic datasets: one for least-squares regression (with the same distribution over each machine), and one for logistic regression (with varying distributions across clients). More details are given in Section B.1 on the way data is generated. We use $N = 20$ clients, each holding 200 points of dimension $d = 20$, and run algorithms over 100 epochs.

To illustrate theorems on real data and higher dimension, we then consider two real-world dataset: *superconduct* [see Hamidieh, 2018, with 21 263 points and 81 features] and *quantum* [see Caruana et al., 2004, with 50 000 points and 65 features] with $N = 20$ workers. To simulate non-i.i.d. *and* unbalanced workers, we split the dataset in heterogeneous groups, using a Gaussian mixture clustering on the TSNE representations (defined by Maaten and Hinton [2008]). Thus, the distribution and number of points hold by each worker largely differs between clients, see Figure B.6.

Convergence. Figure 2.1a presents the convergence of each algorithm w.r.t. the number of iterations k. During first iterations all algorithms make fast progress. However, because $\sigma_*^2 \neq 0$, all algorithms saturate; and the saturation level is higher for double compression (`Artemis`, `Bi-QSGD`), than for simple compression (`Diana`, `QSGD`), or than for SGD. This corroborates findings in Theorem 2.1 and Theorem 2.3.

Complexity. On Figure 2.2, the loss is plotted w.r.t. the theoretical number of bits exchanged after k iterations for the *quantum* and *superconduct* dataset. This confirms that double compression should be the method of choice to achieve a reasonable precision (*w.r.t.* σ_*), whereas for high precision, a simple method like SGD results *in a lower complexity.*

Linear convergence under *null variance at the optimum.* To highlight the significance



(a) Quantum (b) Quantum

(c) Superconduct (d) Superconduct

Figure 2.3: **Artemis compared to other existing algorithms.** $\gamma = 1/(2L)$, X-axis in # epoch or in # bits.

of our new condition on the noise, we compare $\sigma_*^2 \neq 0$ and $\sigma_*^2 = 0$ on Figure 2.1. Saturation is observed in Figure 2.1a, but if we consider a situation in which $\sigma_*^2 = 0$, and where the uniform bound on the gradient's variance *is not null* (as opposed to experiments in Liu et al. [2020] who consider batch gradient descent), *a linear convergence rate is observed.* This illustrates that our new condition is sufficient to reach a linear convergence. Comparing Figure 2.1a with Figure B.3a sheds light on the fact that the saturation level (before which double compression is indeed beneficial) is truly proportional to the noise variance *at optimal point* i.e. σ_*^2. And when $\sigma_*^2 = 0$, bidirectional compression is much more effective than the other methods (see Figure B.3 in Subsection B.1.1.1).

Heterogeneity and real datasets. While in Figure 2.1a, data is i.i.d. on machines, and `Artemis` is thus not expected to outperform `Bi-QSGD` (the difference between the two being the memory), in Figures 2.1b and 2.2 we use **non-i.i.d. data**. None of the previous papers on compression directly illustrated the impact of heterogeneity on simple examples, neither compared it with i.i.d. situations.

Comparing `Artemis` with other existing algorithms. On Figure 2.3 we compare `Artemis` with `FedAvg`, `FedPAQ`, `Diana`, `Dore` and `Double-Squeeze`. We take $\gamma = 1/(2L)$ because otherwise `FedAvg` and `FedPAQ` diverge. These two algorithms present worse performance because they have not been designed for non-i.i.d. datasets. We can observe that `Double-Squeeze` (which only uses error-feedback) is outperformed by `Artemis`. Besides, we observe that `Dore` (which combines EF with memory) has an identical rate of convergence than `Artemis`, which underlines that for unbiased operators of compression, in a heterogeneous setting, **the enhancement comes from the memory and not from the error-feedback**. `FedPAQ` (unidirectional compression) has a very fast convergence during first iterations, but then saturates at a level higher than for `Artemis`-like algorithms.

2.5 Conclusion

We propose `Artemis`, a framework using bidirectional compression to reduce the number of bits needed to perform distributed or federated learning. On top of compression, `Artemis` includes a memory mechanism which improves convergence over non-i.i.d. data. We provide three tight theorems giving guarantees of a fast convergence (linear up to a threshold), highlighting the impact of memory, analyzing Polyak-Ruppert averaging and obtaining lowers bound by studying convergence in distribution of our algorithm. Altogether, this improves the understanding of compression combined with a memory mechanism and sheds light on challenges ahead.

3

MCM: a preserved central model for faster bidirectional compression in distributed settings

"Where sky and water meet,
Where the waves grow sweet,
Doubt not, Reepicheep,
To find all you seek,
There is the utter East."

The voyage of the Dawn Treader, C.S. Lewis.

In this Chapter, we move the focus toward feedback loops to reduce the impact of compression. We develop a new approach to tackle communication constraints in distributed learning problems with a central server. We propose and analyze an algorithm that performs bidirectional compression and achieves asymptotically the same convergence rate as algorithms using only uplink (from the local clients to the central server) compression. This algorithm, MCM, is such that the downlink compression *only impacts local models*, while the global model is preserved. As a result, and contrary to previous works, the gradients on local servers are computed on *perturbed models*. Consequently, convergence proofs are more challenging and require a precise control of this perturbation. To ensure it, MCM additionally combines model compression with a memory mechanism. This analysis opens new doors, e.g. incorporating worker dependent randomized-models and partial participation.

This chapter is based on our work *Preserved central model for faster bidirectional compression in distributed settings* [Philippenko and Dieuleveut, 2021] published at Neurips 2021.

Contents

3.1 Introduction

Large scale distributed machine learning is widely used in many modern applications [Abadi et al., 2016, Caldas et al., 2019, Seide and Agarwal, 2016]. The training is distributed over a potentially large number N of clients that communicate either with a central server [see Konečný et al., 2016, McMahan et al., 2017, on federated learning], or using peer-to-peer communication [Colin et al., 2016, Vanhaesebrouck et al., 2017, Tang et al., 2018].

In this work, we consider a setting using a central server that aggregates updates from remote nodes. Formally, we have a number of features $d \in \mathbb{N}^*$, and a convex cost function $F : \mathbb{R}^d \to \mathbb{R}$. We want to solve the following distributed convex optimization problem using stochastic gradient algorithms [Robbins and Monro, 1951, Bottou, 2010]:

$$\min_{w \in \mathbb{R}^d} F(w) \text{ with } F(w) = \frac{1}{N} \sum_{i=1}^{N} F_i(w) \,,$$

where $(F_i)_{i=1}^N$ is a *local* risk function (empirical risk or expected risk in a streaming framework). This applies to both instances of *distributed* and *federated* learning.

An important issue of those frameworks is the high communication cost between the clients and the central server [Kairouz et al., 2019, Sec. 3.5]. This cost is a concern from several points of view. First, exchanging information can be the bottleneck in terms of speed. Second, the data consumption and the bandwidth usage of training large distributed models can be problematic; and furthermore, the energetic and environmental impact of those exchanges is a growing concern. Over the last few years, new algorithms were introduced, compressing messages in the *upload communications* (i.e., from remote devices to the central server) in order to reduce the size of those exchanges [Seide et al., 2014, Alistarh et al., 2017, Wu et al., 2018, Agarwal et al., 2018, Wangni et al., 2018, Stich et al., 2018, Stich and Karimireddy, 2020, Mishchenko et al., 2019, Li et al., 2020b]. More recently, a new trend has emerged to also compress the *downlink communication*: this is *bidirectional compression*.

The necessity for bidirectional compression can depend on the situation. For example, a single uplink compression could be sufficient in *asymmetric* regimes in which broadcasting a message to N clients ("one to N") is faster than aggregating the information coming from each node ("N to one"). However, in other regimes, e.g. with few machines, where the bottleneck is the transfer

time of a heavy model (up to several GB in modern deep learning architectures) the downlink communication cannot be disregarded, as the upload and download speed are of the same order (see an analysis of bandwidth usage in Section 1.4) Furthermore, in a situation in which participants have to systematically download an update (e.g., on their smartphones) to participate in the training, participants would prefer to receive a small size update (compressed) rather than a heavier one. To encompass all situations, we consider algorithms for which the information exchanged is compressed in both directions.

To perform downlink communication, existing bidirectional algorithms [Tang et al., 2019, Zheng et al., 2019, Sattler et al., 2019, Liu et al., 2020, Philippenko and Dieuleveut, 2020, Horváth and Richtárik, 2020, Xu et al., 2021, Gorbunov et al., 2020b] first aggregate all the information they have received, compress them and then carry out the broadcast. Both the main "global" model and the "local" ones perform the *same* update with this compressed information. Consequently, the model hold on the central server and the one used on the local clients (to query the gradient oracle) are identical. However, this means that the model on the central server has been artificially *degraded*: instead of using all the information it has received, it is updated with the compressed information.

Here, we focus on *preserving* (instead of *degrading*) the central model: the update made on its side does not depend on the downlink compression. This implies that the local models are *different* from the central model. The local gradients are thus measured on a *"perturbed model"* (or *"perturbed iterate"*): such an approach requires a more involved analysis and the algorithm must be carefully designed to control the deviation between the local and global models [Mania et al., 2016]. For example, algorithms directly compressing the model or the update would simply not converge.

We propose `MCM` - *Model Compression with Memory* - an algorithm that 1) preserves the central model, and 2) uses a memory scheme to reduce the variance of the local model. We prove that the convergence of this method is similar to the one of algorithms using only unidirectional compression.

Potential Impact. Proposing an analysis that handles perturbed iterates is the key to unlock three major challenges of distributed learning run with bidirectionally compressed gradients. First, we show that it is possible to improve the convergence rate by sending *different* ***randomized*** *models* to the different clients, this is `Rand-MCM`. Secondly, this analysis also paves the way to deal with partially participating clients: the adaptation of `Rand-MCM` to this framework is straightforward; while adapting existing algorithms to partial participation is not practical (see the "catching-up" process described in Section 2.2). Thirdly, this framework is also promising in terms of business applications, e.g., in the situation of learning with privacy guarantees and *with a trusted central server*. We detail those three possible extensions in Subsection 3.4.1.

Contributions. We make the following contributions:

1. We propose a new algorithm `MCM`, combining a memory process to the "preserved" update. To convey the key steps of the proof, we also introduce an auxiliary hypothetical algorithm, `Ghost`.
2. For those algorithms, we carefully control the variance of the local models w.r.t. the global one. We provide a *contraction equation* involving the control on the local model's variance and show that `MCM` achieves the same rate of convergence as single compression in strongly-convex, convex and non-convex regimes. We give a comparisons of `MCM`'s rates with existing algorithms in Table 3.2.
3. We propose a variant, `Rand-MCM` incorporating diversity into models shared with the local clients and show that it improves convergence for quadratic functions.

This is the first algorithm for double compression to focus on a **preserved central model**. We underline, both theoretically and in practice, that we get the same asymptotic convergence rate for simple and double compression - which is a major improvement. Our approach is one of the first to allow for client dependent model, and to naturally adapt to client dependent compression levels.

Table 3.1: Features of the main existing algorithms performing compression. e_k^i (resp. E_k) denotes the use of error-feedback at uplink (resp. downlink). h_k^i (resp. H_k) denotes the use of a memory at uplink (resp. downlink). Note that `Dist-EF-SGD` is identical to `Double-Squeeze` but has been developed simultaneously and independently.

	Compr.	e_k^i	h_k^i	E_k	H_k	Rand.	update point
`Qsgd` [AGL+17]	one-way						
`ECQ-sgd` [WHHZ18]	one-way	✓					
`Diana` [MGTR19]	one-way		✓				
`Dore` [CL11]	two-way		✓	✓			degraded
`Double-Squeeze` [TYL+19], `Dist-EF-SGD` [ZHK19]	two-way	✓		✓			degraded
`Artemis` [see Chapter 2]	two-way		✓				degraded
`MCM`	two-way		✓		✓		non-degraded
`Rand-MCM`	two-way		✓		✓	✓	non-degraded

The rest of the chapter is organized as follows: in Section 3.2 we present the problem statement and introduce `MCM` and `Rand-MCM`. Theoretical results on these algorithms are successively presented in Sections 3.3 and 3.4. Finally, we present experiments supporting the theory in Section 3.5.

3.2 Problem statement

We consider the minimization problem described in Section 3.1. In the convex case, we assume there exists an optimal parameter w_*, and denote $F_* = F(w_*)$. To solve this problem, we rely on a stochastic gradient descent (SGD) algorithm. A stochastic gradient g_{k+1}^i is provided at iteration k in $\mathbb{N}$ to the device i in $[\![1, N]\!]$. This gradient oracle can be computed on a mini-batch of size b. This function is then evaluated at point w_k. In the classical centralized framework (without compression), for a learning rate γ, SGD corresponds to:

$$w_k = w_{k-1} - \gamma \frac{1}{N} \sum_{i=1}^{N} g_k^i(w_{k-1}) \,. \tag{3.1}$$

We now describe the framework used for compression.

3.2.1 Bidirectional compression framework

Bidirectional compression consists in compressing communications in both directions between the central server and remote devices. We use two different compression operators, respectively $\mathcal{C}_{\text{up}}$ and $\mathcal{C}_{\text{dwn}}$ to compress the message in each direction. Roughly speaking, the update in Equation (3.1) becomes:

$$w_k = w_{k-1} - \gamma \mathcal{C}_{\text{dwn}} \left(\frac{1}{N} \sum_{i=1}^{N} \mathcal{C}_{\text{up}}(g_k^i(w_{k-1})) \right) .$$

However, this approach has a major drawback. The central server receives and aggregates information $\frac{1}{N}\sum_{i=1}^{N} \mathcal{C}_{\text{up}}(g_k^i(w_{k-1}))$. But in order to be able to broadcast it back, it compresses it, *before* applying the update. We refer to this strategy as the "degraded update" approach. Its major advantage is simplicity, and it was used in all previous papers performing double compression. Yet, it appears to be a waste of valuable information. In this Chapter, we update the global model w_{k+1} independently of the downlink compression:

$$\begin{cases} w_k = w_{k-1} - \gamma \frac{1}{N} \sum_{i=1}^{N} \mathcal{C}_{\text{up}} \left(g_k^i(\widehat{w}_{k-1}) \right) \\ \hat{w}_k = C_{\text{dwn}}(w_k) \,. \end{cases} \tag{3.2}$$

However, bluntly compressing w_k in Equation (3.2) hinders convergence, thus the second part of the update needs to be refined by adding a memory mechanism. **We now describe both communication stages of the real MCM, which is entirely defined by the following "clients-to-server" and "server-to-client" equations.**

"Server-to-client" equations

$$\begin{cases} \Omega_k = w_k - H_{k-1} \\ \widehat{w}_k = H_{k-1} + \mathcal{C}_{\text{dwn}}(\Omega_k) \\ H_k = H_{k-1} + \alpha_{\text{dwn}}\mathcal{C}_{\text{dwn}}(\Omega_k) \end{cases}$$

"Clients-to-server" equations

$$\begin{cases} \forall i \in [\![1, N]\!], \Delta_k^i = g_k^i(\widehat{w}_{k-1}) - h_{k-1}^i \\ w_k = w_{k-1} - \frac{\gamma}{N}\sum_{i=1}^N \mathcal{C}_{\text{up}}(\Delta_k^i) + h_{k-1}^i \\ h_k^i = h_{k-1}^i + \alpha_{\text{up}}\mathcal{C}_{\text{up}}(\Delta_k^i)\,. \end{cases} \tag{3.3}$$

Downlink Communication. We introduce a *downlink memory term* $(H_k)_k$, which is available on both clients and central server. The downlink memory term is initialized at the same point on the server and the clients and then follows the same update process, therefore it does not lead to any additional communication cost. The difference Ω_k between the model w_k and the memory H_{k-1} is compressed and exchanged, and the memory term is then used to compute the model $\hat{w}_k = \mathcal{C}_{\text{dwn}}(\Omega_k) + H_{k-1}$ held on each client. Next, the memory is updated with a learning rate α_{dwn} using the compressed term $\mathcal{C}_{\text{dwn}}(\Omega_k)$ sent from the server to the client, as given on the left part of Equation (3.3).

Introducing this memory mechanism is crucial to control the variance of the local model $\widehat{w}_k$. To the best of our knowledge, MCM is the first algorithm that uses such a memory mechanism for downlink compression. This mechanism was introduced by Mishchenko et al. [2019] for the uplink compression but with the other purpose of mitigating the impact of heterogeneity, while we use it here to avoid divergence of the local model's variance.

Uplink Communication. The motivation to introduce an uplink memory term h_k^i for each device $i \in [\![1, N]\!]$ is different, and better understood. Indeed, for the uplink direction, this mechanism is only necessary (and then crucial) to handle heterogeneous clients (i.e., with different data distributions, see Chapter 2). Here, the difference Δ_k^i between the stochastic gradient $g_k^i(\widehat{w}_{k-1})$ evaluated at the local model $\widehat{w}_{k-1}$ (as defined in Equation (3.3)) and the memory term is compressed and exchanged. The memory is then updated as defined on right part of Equation (3.3) with a rate α_{dwn}.

Remark 3.1 (Rate α_{dwn})**.** *It is necessary to use $\alpha_{\text{dwn}} < 1$. Otherwise, the compression noise tends to propagate and is amplified, because of the multiplicative nature of the compression. In Figure 3.1 we compare* ***MCM****, with 3 other strategies: compressing only the update, compressing $w_k - \widehat{w}_{k-1}$, (i.e., $\alpha_{\text{dwn}} = 1$), and compressing the model (i.e., $H_k = 0$), showing that only* ***MCM*** *converges.*

Remark 3.2 (Memory vs Error Feedback)**.** *Error feedback is another technique, introduced by Seide et al. [2014]. In the context of double compression, it has been shown to improve convergence for a restrictive class of* contracting *compression operators (which are generally biased) by Zheng et al. [2019], Tang et al. [2019]. However, we note several differences to our approach. (1) For unbiased operators - as considered in* ***Dore****, it did not lead to any theoretical improvement [Remark 2 in Sec. 4.1., Liu et al., 2020]. (2) Moreover, only a fraction (namely $(1+\omega_{\text{dwn}})^{-1}$) of the "error" $w_k - \hat{w}_k$ can be preserved in the EF term (see line 18 in algo 1 in Liu et al.). It is thus impossible to recover the central preserved model as a function of the degraded model and the EF term. (3) Zheng et al. [2019] consider a biased operator and the same compression level for uplink and downlink compression. They also rely on stronger assumptions on the gradient (uniformly bounded) and only tackle the homogeneous case.*

In Table 3.1 we summarize the main algorithms for compression in distributed training. As downlink communication can be more efficient than uplink, we consider distinct operators $\mathcal{C}_{\text{dwn}}$, $\mathcal{C}_{\text{up}}$ and allow the corresponding compressions levels to be distinct: those quantities are defined in Assumption 3.1.

Assumption 3.1. *There exist two constants* $\omega_{\mathrm{up}}, \omega_{\mathrm{dwn}} \in \mathbb{R}_+^*$*, such that for* $\mathrm{dir} \in \{\mathrm{up}, \mathrm{dwn}\}$*, the compression operators* $\mathcal{C}_{\mathrm{dir}}$ *verify the two following properties for all* z *in* $\mathbb{R}^d$*:* $\mathbb{E}[\mathcal{C}_{\mathrm{dir}}(z)] = z$*, and* $\mathbb{E}[\|\mathcal{C}_{\mathrm{dir}}(z) - z\|^2] \leq \omega_{\mathrm{dir}} \|z\|^2$*. The higher is* ω_{dir}*, the more aggressive the compression is.*

In general, compression operators can be biased or unbiased, and their effects on convergence can vary widely (a detailed analysis of the impact of compressors on convergence is given in Chapter 4). For instance, algorithms with error-feedback may diverge if the operator is not contracting. While Horváth and Richtárik [2020] have proposed a method to unbiase biased operators, and Beznosikov et al. [2020] have conducted a general study of biased operators, our focus in this chapter is solely on unbiased operators; which includes sparsification, quantization, and sketching.

The choice of the operator of compression is crucial when compressing data. Operators of compression may be classified into two mains categories: quantization [as in Rabbat and Nowak, 2005, Alistarh et al., 2017, Seide et al., 2014, Zhou et al., 2018, Wen et al., 2017, Reisizadeh et al., 2020, Horváth et al., 2022] and random projection [as in Vempala, 2005, Rahimi and Recht, 2008, Stich et al., 2018, Alistarh et al., 2018, Khirirat et al., 2020b].

Remark 3.3 (Extension to biased operator of compression)**.** *Our analysis could be extended to biased uplink operators, following similar lines of proof given in [Beznosikov et al., 2020]. However, the extension for the downlink operator seems more difficult as our analysis relies on numerous occurrences on the fact that the expectation of* $\widehat{w}_{k-1}$ *knowing* w_{k-1} *is* w_{k-1}*.*

Remark 3.4 (Related work on Perturbed iterate analysis)**.** *The theory of perturbed iterate analysis was introduced by Mania et al. [2016] to deal with asynchronous SGD. More recently, it was used by Stich and Karimireddy [2020], Gorbunov et al. [2020b] to analyze the convergence of algorithms with uplink compressions, error feedback and asynchrony. Using gradients at randomly perturbed points can also be seen as a form of randomized smoothing [Scaman et al., 2018], a point we discuss below.*

Our approach can also be related to randomized smoothing. Formally, $\nabla F(\widehat{w}_{k-1})$ can be considered as an unbiased gradient of the smoothed function F_ρ at point w_{k-1}, with $F_\rho : w \mapsto \mathbb{E}[F(w + \widehat{w}_{k-1} - w_{k-1})]$. Then $\mathbb{E}\left\langle \nabla F(\widehat{w}_{k-1}), w_{k-1} - w_* \right\rangle = \mathbb{E}\left\langle \nabla F_\rho(w_{k-1}), w_{k-1} - w_* \right\rangle$. One key aspect is that the condition number μ_ρ / L_ρ of F_ρ is always larger (better) than the one for F. However, the minimum of F_ρ is different and moving, thus the proof techniques from randomized smoothing are not adapted to a varying noise which distribution is unknown. Providing a theoretical result that quantifies the smoothing impact of MCM is an interesting open direction.

Randomized smoothing has been applied to non-smooth problems by Duchi et al. [2012]. The aim is to transform a non-smooth function into a smooth function, before computing the gradient. This is achieved by adding a Gaussian noise to the point where the gradient is computed. This mechanism has been applied by Scaman et al. [2018] to convex problems. We consider in this work a randomized version of compression: at iteration k in $\mathbb{N}$ each client i in $[\![1, N]\!]$ receives a noisy estimate $\widehat{w}_k^i$ of the global model w_k kept on central server. Thus, we compute the local gradient at a perturbed point $w_k + \delta_k^i$. Unlike the randomization process as defined by Duchi et al. [2012], the noise here is not chosen to improve the function's regularity but results from the compression.

3.2.2 The randomization mechanism, Rand-MCM

In this subsection, we describe the key feature introduced in Rand-MCM: *randomization*. It consists in performing an independent compression for each device instead of performing a single one for all of them. As a consequence, each client holds a different model centered around the global one. This introduces some supplementary randomness that stabilizes the algorithm. Formally, we will consider N mutually independent compression operators $\mathcal{C}_{\mathrm{dwn},i}$ instead of a single one $\mathcal{C}_{\mathrm{dwn}}$, and the central server will send to the device i at iteration k the compression of the difference between

Algorithm 3: Pseudocode of `Rand-MCM` - set $\alpha_{\text{up/dwn}} > 0$ to use memory

Input: Mini-batch size b, learning rates $\alpha_{\text{up}}, \alpha_{\text{dwn}}, \gamma > 0$, initial model $w_0 \in \mathbb{R}^d$ (on all devices), operators $\mathcal{C}_{\text{up}}$ and $\mathcal{C}_{\text{dwn}}$.
Initialization: $\forall i \in [\![1, N]\!]$, $h_0^i = g_1^i(w_0)$ (smart initialization) and $\widehat{\Omega}_{-1}^i = H_{-1}^i = w_0$
Output: Model w_K
for $k = 1, 2, \ldots, K$ **do**
  for *each device* $i = 1, 2, 3, \ldots, N$ **do**
    Receive $\widehat{\Omega}_{k-1}^i$, and set: $w_{k-1}^i = \widehat{\Omega}_{k-1}^i + H_{k-2}^i$
    Update down memory: $H_{k-1}^i = H_{k-2}^i + \alpha_{\text{dwn}} \widehat{\Omega}_{k-1}^i$
    Compute $g_k^i(w_{k-1}^i)$ (with mini-batch)
    Set $\Delta_k^i = g_k^i(w_{k-1}^i) - h_{k-1}^i$, compress it: $\widehat{\Delta}_k^i = \mathcal{C}_{\text{up}}(\Delta_k^i)$
    Update uplink memory: $h_k^i = h_{k-1}^i + \alpha_{\text{up}} \widehat{\Delta}_k^i$
    Send $\widehat{\Delta}_k^i$ to central server
  Receive $(\widehat{\Delta}_k^i)_{i=1}^N$ from all remote clients
  Compute $\widehat{g}_k = \frac{1}{N} \sum_{i=1}^N \widehat{\Delta}_k^i + h_{k-1}^i$
  Update uplink memory: $\forall i \in [\![1, N]\!]$, $h_k^i = h_{k-1}^i + \alpha_{\text{up}} \widehat{\Delta}_k^i$
  Non-degraded update: $w_k = w_{k-1} - \gamma \widehat{g}_k$
  Down compression: $\forall i \in [\![1, N]\!]$, $\widehat{\Omega}_k^i = \mathcal{C}_{\text{dwn},i}(w_k - H_{k-1}^i)$
  Update downlink memory: $H_k^i = H_{k-1}^i + \alpha_{\text{dwn}} \widehat{\Omega}_k^i$
  Send $(\widehat{\Omega}_k^i)_{i=1}^N$ to all remote clients

its model and the local memory on client i: $\mathcal{C}_{\text{dwn},i}(w_k - H_{k-1}^i)$. The trade-offs associated with this modification are discussed in Section 3.4.

The pseudocode of `Rand-MCM` is given in Algorithm 3. It incorporates all components described above: 1) the bidirectional compression, 2) the model update using the non-degraded point, 3) the two memories, 4) the up and down compression operators, 5) the randomization mechanism.

3.2.3 The `Ghost` algorithm

To convey the best understanding of the theorems and the spirit of the proof, we define a *ghost* algorithm (that is impossible to implement in practice). `Ghost` is introduced only to get some intuition of the theoretical insight.

Definition 3.1 (`Ghost` algorithm)**.** *The `Ghost` algorithm is defined as follows, for $k \in \mathbb{N}$, for all $i \in [\![1, N]\!]$ we have:*

$$w_k = w_{k-1} - \gamma \frac{1}{N} \sum_{i=1}^N \widehat{g}_k^i(\widehat{w}_{k-1}) \quad \text{and} \quad \widehat{w}_k = w_{k-1} - \gamma \mathcal{C}_{\text{dwn}} \left(\frac{1}{N} \sum_{i=1}^N \widehat{g}_k^i(\widehat{w}_{k-1}) \right). \tag{3.4}$$

While the global model is unchanged (1^{st} line), the local model $\widehat{w}_k$ (2^{nd} line) is updated using the global model w_{k-1} at the previous step, which is not available locally.

3.3 Assumptions and theoretical analysis

We make standard assumptions on $F : \mathbb{R}^d \to \mathbb{R}$. We first assume that the loss function F is smooth.

Assumption 3.2 (Smoothness). *F is twice continuously differentiable, and is L-smooth, that is for all vectors z, z' in $\mathbb{R}^d$: $\|\nabla F(z) - \nabla F(z')\| \leq L\|z - z'\|$.*

Results in Section 3.3 are provided in a convex, strongly-convex and non-convex setting.

Assumption 3.3 (Strong convexity). *F is μ-strongly convex (or convex if $\mu = 0$), that is for all vectors z, z' in $\mathbb{R}^d$: $F(z') \geq F(z) + (z' - z)^T \nabla F(z) + \frac{\mu}{2}\|z' - z\|_2^2$.*

Next, we present the assumption on the stochastic gradients.

Assumption 3.4 (Noise over stochastic gradients computation). *The noise over stochastic gradients for a mini-batch of size b, is uniformly bounded: there exists a constant $\sigma \in \mathbb{R}_+$, such that for all k in $\mathbb{N}$, for all i in $[\![1, N]\!]$ and for all w in $\mathbb{R}^d$ we have: $E[\|g_k^i(w) - \nabla F(w)\|^2] \leq \sigma^2/b$.*

Unlike in Chapter 2, we do not assume the noise over stochastic gradients to be bounded only at the optimal point w_*. However, our results could be extended to this setting.

To prove the convergence of MCM, we combine two results: a control of the variance of the local model, $\mathbb{E}[\|w_k - \widehat{w}_k\|^2 | w_k]$[1], and then the convergence under this result. Proposition 3.1 and Theorem 3.1 below provide those results for Ghost algorithm.

3.3.1 Theoretical results: Ghost algorithm

Proposition 3.1. *Consider the Ghost update in Equation (3.4), under Assumptions 3.1, 3.2 and 3.4, for all k in $\mathbb{N}$ with the convention $\nabla F(w_{-1}) = 0$:*

$$\mathbb{E}\left[\|w_k - \widehat{w}_k\|^2 \;\middle|\; \widehat{w}_{k-1}\right] \leq \gamma^2 \omega_{\mathrm{dwn}} \left(1 + \frac{\omega_{\mathrm{up}}}{N}\right) \|\nabla F(\widehat{w}_{k-1})\|^2 + \frac{\gamma^2 \omega_{\mathrm{dwn}}(1 + \omega_{\mathrm{up}})\sigma^2}{Nb}\,.$$

This proposition is proved in Subsection C.3.1. From Equation (3.4) and Assumption 3.1, it appears that $\|w_k - \widehat{w}_k\|^2 \leq \omega_{\mathrm{dwn}} \|\gamma \sum_{i=1}^N \widehat{g}_k^i(w_{k-1})\|^2$: the result follows from controlling the variance of $\widehat{g}_k^i(w_{k-1})$. The take-away from this Proposition is that we are able to bound the variance of the local model by an affine function of the squared norm of the *previous* stochastic gradients $\nabla F(\widehat{w}_{k-1})$. For Ghost only the previous gradient is involved, while for MCM, we obtain an additional recursive process.

To obtain the convergence, we then follow the classical approach of Mania et al. [2016], expanding $\mathbb{E}[\|w_k - w_*\|^2]$ as $\mathbb{E}[\|w_{k-1} - w_*\|^2] - 2\gamma\mathbb{E}[\langle \nabla F(\widehat{w}_{k-1}), w_{k-1} - w_* \rangle] + \gamma^2 \mathbb{E}[\|\widehat{g}_k(\widehat{w}_{k-1})\|^2]$. The critical aspect is that the inner product does not directly result in a contraction, as the support point of the gradient differs from w_{k-1}. Using the fact that $\mathbb{E}\left[\widehat{w}_{k-1} \mid w_{k-1}\right] = w_{k-1}$, we further decompose it as

$$-2\gamma\mathbb{E}\left\langle \nabla F(\widehat{w}_{k-1}), \widehat{w}_{k-1} - w_* \right\rangle + 2\gamma\mathbb{E}\left\langle \nabla F(\widehat{w}_{k-1}) - \nabla F(w_{k-1}), w_{k-1} - \widehat{w}_{k-1} \right\rangle . \tag{3.5}$$

The first part of Equation (3.5), corresponds to a "strong contraction": by (strong-)convexity, we can upper bound it by $-2\gamma(\mu \|\widehat{w}_{k-1} - w_*\|^2 + F(\widehat{w}_{k-1}) - F_*)$, which is on average larger than $-2\gamma(\mu \|w_{k-1} - w_*\|^2 + F(w_{k-1}) - F_*)$ (Jensen's inequality). Moreover, as the function is smooth and convex, it can also be upper bounded by $-2\gamma \|\nabla F(\widehat{w}_{k-1})\|^2 / L$. This is a crucial term: we "gain" something of the order of a squared norm of the gradient at $\widehat{w}_{k-1}$, which will *in fine* compensate the variance of the local model. The second part of Equation (3.5), corresponds to a positive residual term, proportional to the variance of the compressed model, that can be controlled thanks to Proposition C.1 (at w_{k-1}!). Putting things together, we get, in the convex case ($\mu = 0$):

[1]For clarity here, we use $\mathbb{E}\left[\cdot \mid w_{k-1}\right]$ to denote the conditioning w.r.t. all randomness before w_{k-1}.

Theorem 3.1 (Contraction for Ghost, convex case)**.** *Under Assumptions 3.1 to 3.4, with $\mu = 0$, if $\gamma L(1 + \omega_{\rm up}/N) \leq \frac{1}{2}$.*

$$\mathbb{E}\|w_k - w_*\|^2 \leq \mathbb{E}\|w_{k-1} - w_*\|^2 - \gamma\mathbb{E}(F(w_{k-1}) - F_*) - \frac{\gamma}{2L}\mathbb{E}\left[\|\nabla F(\widehat{w}_{k-1})\|^2\right] + 2\gamma^3\omega_{\rm dwn}L\left(1 + \frac{\omega_{\rm up}}{N}\right)\mathbb{E}\|\nabla F(\hat{w}_{k-2})\|^2 + \gamma^2\frac{(1+\omega_{\rm up})\sigma^2}{Nb}(1 + 2\gamma L\omega_{\rm dwn})\,.$$

We can make the following observations:

1. At step k, the residual can be upper bounded by a constant times squared norm of the gradient at point $\hat{w}_{k-2}$. When using recursively this upper bound, if $2\gamma^3\omega_{\rm dwn}L(1 + \omega_{\rm up}/N) \leq \gamma/(2L)$, then these terms cancel out. This is equivalent to $2\gamma L\sqrt{\omega_{\rm dwn}\,(1 + \omega_{\rm up}/N)} \leq 1$. It is natural to chose $\gamma \leq 1/(2L\max(1 + \omega_{\rm up}/N, 1 + \omega_{\rm dwn}))$.
2. The bound is in fact proved conditionally to w_{k-1}, recursive conditioning is required to propagate the inequality. We carefully handle conditioning in the proofs.

Theorem 3.2 (Convergence of Ghost, convex case)**.** *Under Assumptions 3.1 to 3.4 with $\mu = 0$ (convex case), for all k in $\mathbb{N}$, defining $V_k := \mathbb{E}\left[w_k - w_*\right] + \frac{\gamma}{2L}\mathbb{E}[\|\nabla F(\widehat{w}_{k-1})\|^2] + 2\gamma L\mathbb{E}[\|\widehat{w}_k - w_k\|^2]$, we have:*

$$V_k \leq V_{k-1} - \gamma\mathbb{E}\left[F(w_{k-1}) - F(w_*)\right] + \frac{\gamma^2\sigma^2\Phi^{\mathcal{G}}(\gamma)}{Nb}\,,$$

with $\Phi^{\mathcal{G}}(\gamma) := (1 + \omega_{\rm up})(1 + 2\gamma L\omega_{\rm dwn})$.

3.3.2 Results for MCM

We here provide guarantees of convergence for MCM which incorporates an uplink memory term, designed to handle heterogeneous clients. But to highlight our main contribution, that concerns the downlink compression, we present the results in the homogeneous setting, that is with $F_i = F_j$ and $\alpha_{\rm up} = 0$. In the heterogeneous setting, similar results (almost identical, up to constant numerical factors) are obtained [see our paper Philippenko and Dieuleveut, 2021, Appendix G]. Experiments are performed on heterogeneous clients. We provide here convergence results in the strongly-convex, convex and non-convex cases.

Notations and settings. For k in $\mathbb{N}$, we denote $\Upsilon_k = \|w_k - H_{k-1}\|^2$, and define $V_k = \mathbb{E}[\|w_k - w_*\|^2] + 32\gamma L\omega_{\rm dwn}^2\mathbb{E}[\Upsilon_k]$, which serves as Lyapunov function. V_k is composed of two terms: the first one controls the quadratic distance to the optimal model, and the second controls the variance of the local models $\hat{w}_k$. For both theorems, we choose $\alpha_{\rm dwn} = (8\omega_{\rm dwn})^{-1}$. We denote $\Phi(\gamma) := (1 + \omega_{\rm up})\left(1 + 64\gamma L\omega_{\rm dwn}^2\right)$.

Limit learning rate. There exists a maximal learning rate to ensure convergence. More specifically, we define $\gamma_{\max} := \min(\gamma_{\max}^{\rm up}, \gamma_{\max}^{\rm dwn}, \gamma_{\max}^{\Upsilon})$, where $\gamma_{\max}^{\rm up} := (2L\,(1 + \omega_{\rm up}/N))^{-1}$ corresponds to the classical constraint on the learning rate in the unidirectional regime [see Mishchenko et al., 2019, Philippenko and Dieuleveut, 2020], $\gamma_{\max}^{\rm dwn} := (8L\omega_{\rm dwn})^{-1}$ is a similar constraint coming from the downlink compression, and $\gamma_{\max}^{\Upsilon} := \left(8\sqrt{2}L\omega_{\rm dwn}\sqrt{8\omega_{\rm dwn} + \omega_{\rm up}/N}\right)^{-1}$ is a combined constraint that arises when controlling the variance term Υ.[2] Overall, this constraints are weaker than in the "degraded" framework Liu et al. [2020], Philippenko and Dieuleveut [2020], in which $\gamma_{\max}^{\rm Dore} \leq \left(8L(1+\omega_{\rm dwn})(1+\omega_{\rm up}/N)\right)^{-1}$. Especially, in the regime in which $\omega_{\rm up,dwn} \to \infty$ and $\omega_{\rm dwn} \simeq \omega_{\rm up} \simeq: \omega$, the maximal learning rate for MCM is $(L\omega^{3/2})^{-1}$, while it is $(L\omega^2)^{-1}$ in Liu et al. [2020], Philippenko and Dieuleveut [2020]. Our $\gamma_{\max}$ is thus larger by a factor $\sqrt{\omega}$, see Table 3.2. We define $\widetilde{L}$ such that $\gamma_{\max} = (2\widetilde{L})^{-1}$.

[2] The dependency in $\omega^{3/2}$ is similar to the one obtained by Horváth et al. [2022] in unidirectional compression in the non-convex case (Theorem 4).

Theorem 3.3 (Convergence of MCM in the strongly-convex case)**.** *Under Assumptions 3.1 to 3.4 with $\mu > 0$, for k in $\mathbb{N}$, for any sequence $(\gamma_k)_{k\geq 0} \leq \gamma_{\max}$ we have:*

$$V_k \leq (1 - \frac{\gamma_k \mu}{2})V_{k-1} - \gamma_k \mathbb{E}\left[F(\widehat{w}_{k-1}) - F(w_*)\right] + \frac{\gamma_k^2 \sigma^2 \Phi(\gamma_k)}{Nb}\,, \tag{3.6}$$

Consequently, (1) if $\sigma^2 = 0$ (noiseless case), for $\gamma_k \equiv \gamma_{\max}$ we recover a linear convergence rate: $\mathbb{E}[\|w_k - w_*\|^2] \leq (1 - \gamma_{\max}\mu/2)^k V_0$; (2) if $\sigma^2 > 0$, taking for all K in $\mathbb{N}$, $\gamma_K = 4/(\mu(K+1) + \widetilde{L})$, for the weighted Polyak-Ruppert average $\bar{w}_K = \sum_{k=1}^K \lambda_k w_{k-1} / \sum_{k=1}^K \lambda_k$, with $\lambda_k := (\gamma_{k-1})^{-1}$,

$$\mathbb{E}\left[F(\bar{w}_K) - F(w_*)\right] \leq \frac{(\mu + \widetilde{L})\widetilde{L}}{16\mu K^2}\|w_0 - w_*\|^2 + \frac{8\sigma^2(1+\omega_{\rm up})}{\mu K N b}\left(1 + \frac{256 L \omega_{\rm dwn}^2}{\mu K}\ln(\mu(K+1) + \widetilde{L})\right). \tag{3.7}$$

Limit Variance (Equation (3.6)). For a constant γ, the variance term (i.e., term proportional to σ^2) in Equation (3.6) is upper bounded by $\frac{\gamma^2\sigma^2}{Nb}(1+\omega_{\rm up})(1 + 64\gamma L \omega_{\rm dwn}^2)$. The impact of the downlink compression is attenuated by a factor γ. As γ decreases, this makes the limit variance similar to the one of Diana, i.e. without downlink compression [Mishchenko et al., 2019, Eq. 16 in Th. 2], and much lower than the variance for previous algorithms using double compression for which the variance scales quadratically with the compression constants as $\gamma^2\sigma^2(1+\omega_{\rm up})(1+\omega_{\rm dwn})/N$: (1) for Dore, see Corollary 1 in Liu et al. [2020] (who indicate $(1-\rho)^{-1} \geq (1+\omega_{\rm up}/N)(1+\omega_{\rm dwn})$), (2) for Artemis see Table 2.2 and point 2 of Theorem 2.3 in Chapter 2, (3) for Gorbunov et al. [2020b], see Theorem I.1. (with $\gamma D_1' \propto \gamma^2\sigma^2(1+\omega_{\rm up})(1+\omega_{\rm dwn})/N$).

Bound 3.7 has a quadratic dependence on $\omega_{\rm dwn}$, but the corresponding term is divided by an extra factor K, the number of iterations. For example in experiments, for *w8a* using quantization with $s = 2^0$, we have $\omega_{\rm dwn} \simeq 17$, and after only 50 epoch with a batch size $b = 12$, we have $K \simeq 2500$. Hence, the term ω^2/K is vanishing through iterations and we asymptotically recover a rate of convergence equivalent to algorithms using unidirectional compression.

Convergence and complexity: With a decaying sequence of steps, we obtain a convergence rate scaling as $O(K^{-1})$ in Equation (3.7), without dependency on the $\omega_{\rm dwn}$ in the dominating term, which only appears in faster decaying terms scaling as K^{-2}. The iteration complexity (i.e., number of iterations to achieve ϵ expected error) is thus at first order $O_{\epsilon\to 0}(\frac{\sigma^2(1+\omega_{\rm up})}{\mu\epsilon N b})$. Again, this matches the complexity of Diana [Horváth et al., 2022, see Theorem 1 and Corollary 1] and is smaller by a factor $1+\omega_{\rm dwn}$ than the one of Artemis, Dore, DIANAsr-DQ (see Corollary I.1. in Gorbunov et al. [2020b]). Next, we give a convergence result in the convex case.

Theorem 3.4 (Convergence of MCM, convex case)**.** *Under Assumptions 3.1 to 3.4 with $\mu = 0$. For all $k > 0$, for any $\gamma \leq \gamma_{\max}$, we have, for $\bar{w}_k = \frac{1}{k}\sum_{i=0}^{k-1} w_i$,*

$$\gamma\mathbb{E}\left[F(w_{k-1}) - F(w_*)\right] \leq V_{k-1} - V_k + \frac{\gamma^2\sigma^2\Phi(\gamma)}{Nb} \Longrightarrow \mathbb{E}[F(\bar{w}_k) - F_*] \leq \frac{V_0}{\gamma k} + \frac{\gamma\sigma^2\Phi(\gamma)}{Nb}\,. \tag{3.8}$$

Consequently, for K in $\mathbb{N}$ large enough, a step-size $\gamma = \sqrt{\frac{\|w_0 - w_*\|^2 Nb}{(1+\omega_{\rm up})\sigma^2 K}}$, we have:

$$\mathbb{E}[F(\bar{w}_K) - F_*] \leq 2\sqrt{\frac{\|w_0 - w_*\|^2 (1+\omega_{\rm up})\sigma^2}{NbK}} + O(K^{-1}). \tag{3.9}$$

Moreover if $\sigma^2 = 0$ (noiseless case), we recover a faster convergence: $\mathbb{E}[F(\bar{w}_K) - F_*] = O(K^{-1})$.

Limit Variance (Eq. (3.8)). The variance term is identical to the strongly-convex case.

Convergence and complexity (Equation (3.9)). The downlink compression constant only appears in the second-order term, scaling as $1/K$. In other words, the convergence rate is equivalent to the convergence rate of Diana, in the non-strongly-convex. As K increases, this complexity scales as $\frac{(1+\omega_{\text{up}})}{n\epsilon^2}$ independently of the downlink compression. Again, for previous algorithms with double compression the complexity is at least $O\left(\frac{(1+\omega_{\text{up}})(1+\omega_{\text{dwn}})}{n\epsilon^2}\right)$ (see Corollary I.2 in Gorbunov et al. [2020b]).

Control of the variance of the local model. We here present the backbone Lemma of MCM's proof. It allows to control the variance of the local model $\mathbb{E}[\|\hat{w}_k - w_k\|^2 \,|w_k]$ (which is upper-bounded by $\omega_{\text{dwn}}\mathbb{E}[\|\Upsilon_k\|^2 \,|w_k]$) and to build the Lyapunov function defined in Theorems 3.3 and 3.4.

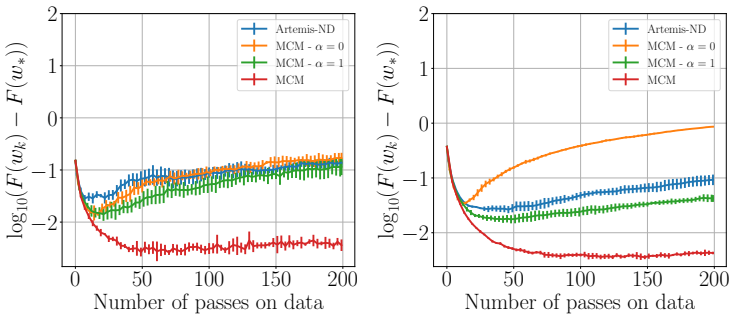


Figure 3.1: Comparing MCM on two datasets (left quantum, right a9a) with three other algorithms using a non-degraded update, $\gamma = 1/L$. Artemis-ND stands for Artemis with a non-degraded update.

This result highlights the impact of the downlink memory term. Without memory, i.e., with $\alpha_{\text{dwn}} = 0$, the variance of the local model $\|\hat{w}_k - w_k\|^2$ increases with the number of iterations. On the other hand, if α_{dwn} is too large (close to 1), this variance diverges. In fact, Theorem 3.5 states that α_{dwn} must be lower than $\alpha_{\max} = (8\omega_{\text{dwn}})^{-1}$ in order to obtain a $(1 - \alpha_{\text{dwn}}/2)$-contraction, which later allows convergence. However, it also involves an additional term $\mathbb{E}[\|\nabla F(\widehat{w}_{k-1})\|^2]$ depending on α_{dwn}^{-1} which results in a bound $\gamma_{\max}^{\Upsilon}$ on the admissible step-size. Therefore, this term must be as small as possible and we chose α_{dwn} as the maximal possible memory's learning rate. This is why, Theorems 3.3 and 3.4 are presented with $\alpha_{\text{dwn}} = (8\omega_{\text{dwn}})^{-1}$. This trade-off is illustrated on two real datasets on Figure 3.1. This phenomenon is similar to the divergence observed in frameworks involving error-feedback, when the compression operator is not contractive.

Theorem 3.5. *Consider the MCM update as in Equation (3.2). Under Assumptions 3.1, 3.2 and 3.4 with $\mu = 0$, if $\gamma \leq (8\omega_{\text{dwn}}L)^{-1}$ and $\alpha_{\text{dwn}} \leq (8\omega_{\text{dwn}})^{-1}$, then for all k in $\mathbb{N}$:*

$$\mathbb{E}\left[\Upsilon_k\right] \leq \left(1 - \frac{\alpha_{\text{dwn}}}{2}\right)\mathbb{E}\left[\Upsilon_{k-1}\right] + 2\gamma^2\left(\frac{1}{\alpha_{\text{dwn}}} + \frac{\omega_{\text{up}}}{N}\right)\mathbb{E}\left[\|\nabla F(\widehat{w}_{k-1})\|^2\right] + \frac{2\gamma^2\sigma^2(1+\omega_{\text{up}})}{Nb}\,.$$

This bound provides a recursive control on Υ_k. Beyond the $(1 - \alpha_{\text{dwn}})$ contraction, the bound comprises the squared norm of the gradient at the previous perturbed iterate, and a noise term.

Summary of rates. In Table 3.2, we summarize the rates and complexities, and maximal learning rate for Diana, Artemis, Dore and MCM. For simplicity, we ignore absolute constants, and provide asymptotic values for large ω_{up}, ω_{dwn}, and complexities for $\epsilon \to 0$.

The last important theorem of this section ensures the convergence of MCM in the non-convex settings. The demonstration is given in Subsection C.4.4 and follows a different approach than the one presented in Subsection 3.3.1.

Theorem 3.6 (Convergence of MCM in the non-convex case)**.** *Under Assumptions 3.1, 3.2 and 3.4 (non-convex case), for a learning rate $\alpha_{\text{dwn}} = \frac{1}{8\omega_{\text{dwn}}}$, for any step-size γ s.t. $\gamma \leq \gamma_{\max}$, after running K in $\mathbb{N}^*$ iterations, we have:*

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\|\nabla F(w_{k-1})\|^2\right] \leq \frac{2\left(F(w_0) - F(w_*)\right)}{\gamma K} + \frac{\gamma L\sigma^2\Phi^{\text{non-cvx}}(\gamma)}{Nb}\,,$$

Table 3.2: Summary of rates on the initial condition, limit variance, asympt. complexities and $\gamma_{\max}$.

Problem		Diana	Artemis, Dore	MCM, Rand-MCM
	$L\gamma_{\max} \propto$	$1/(1+\omega_{\rm up})$	$1/(1+\omega_{\rm up})(1+\omega_{\rm dwn})$	$1/(1+\omega_{\rm dwn})\sqrt{1+\omega_{\rm up}} \wedge 1/(1+\omega_{\rm up})$
	Lim. var. $\propto$ $\gamma^2\sigma^2/n\times$	$(1+\omega_{\rm up})$	$(1+\omega_{\rm up})(1+\omega_{\rm dwn})$	$(1+\omega_{\rm up})(1+\gamma L\omega_{\rm dwn}^2)$
Str.-convex	Rate on init. cond. (SC)	$(1-\gamma\mu)^k$	$(1-\gamma\mu)^k$	$(1-\gamma\mu)^k$
	Complexity	$(1+\omega_{\rm up})/\mu\epsilon N$	$(1+\omega_{\rm dwn})(1+\omega_{\rm up})/\mu\epsilon N$	$(1+\omega_{\rm up})/\mu\epsilon N$
Convex	Complexity	$(\omega_{\rm up}+1)/\epsilon^2$	$(1+\omega_{\rm up})(1+\omega_{\rm dwn})/\epsilon^2$	$(\omega_{\rm up}+1)/\epsilon^2$

with $\Phi^{\rm non-cvx}(\gamma) := (1+\omega_{\rm up})\left(1+32\gamma L\omega_{\rm dwn}^2\right)$. *Consequently, for* K *in* $\mathbb{N}^*$ *large enough, we have taking* $\gamma = \sqrt{\frac{2Nb(F(w_0)-F(w_*))}{\sigma^2 L(1+\omega_{\rm up})K}}$*:*

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\|\nabla F(w_{k-1})\|^2\right] \leq 2\sqrt{\frac{2L\sigma^2(1+\omega_{\rm up})\left(F(w_0)-F(w_*)\right)}{NbK}} + O(K^{-1})\,.$$

Proofs. To convey the best understanding of the theorems and the spirit of the proof, we have introduced the Ghost algorithm in Subsection 3.2.3. The sketch of the proof of Ghost and MCM are similar. Proofs of Theorems 3.1 and 3.2 is given in Section C.3. Proofs of Theorems 3.3 to 3.5 are given in Section C.4. And Theorem 3.6 is proved in Subsection C.4.4. Note that the proof for non-convex follows a different approach than the one in Theorems 3.3 and 3.4.

Proof in the heterogeneous case. To extend Theorems 3.3 to 3.5 in the heterogeneous setting for a convex objective, as in Chapter 2, we assume that there exists a constant B in $\mathbb{R}_+$, s.t.: $\frac{1}{N}\sum_{i=0}^{N}\|\nabla F_i(w_*)\|^2 = B^2$. We further define $\Xi_k = \frac{1}{N^2}\sum_{i=1}^{N}\left\|h_k^i - \nabla F_i(w_*)\right\|^2$, where for all i in $[\![1, N]\!]$. This term is recursively controlled and combined into the Lyapunov function, as in Chapter 2 For sake of the manuscript brevity, we have not included the demonstration and refer to [Philippenko and Dieuleveut, 2021, see Appendix G].

Remark 3.5 (Communication budget)**.** *How to split a given communication budget between uplink and downlink to optimize the convergence is an open question that is intrinsically related to the situation. Indeed it depends on many factors like the selected operators of compression, the upload/downlink speed or the number of participating clients at each iteration. However, our approach provides some insights on this question. Because asymptotically the impact of double compression is marginal, for a fixed budget, Theorem 3.4 suggests to strongly compress on the downlink direction (which leads to a large* ω_{dwn}*), but to perform a weaker compression in the uplink direction.*

As mentioned in the introduction, our analysis of perturbed iterate in the context of double compression opens new directions: in particular, it opens the door to handling a different model for each client. In the next section, we detail those possibilities, and provide theoretical guarantees for Rand-MCM, the variant of MCM in which instead of sending the same model to all clients, the compression noises are mutually *independent*.

3.4 Extension to Rand-MCM

3.4.1 Communication and convergence trade-offs

In Rand-MCM, we leverage the fact that the compressions used for each client need not to be identical. On the contrary, it is possible to consider *independent* compressions. By doing so, we reduce the impact of the downlink compression.

The relevance of such a modification depends on the framework: while the convergence rate will be improved, the computational time can be slightly increased. Indeed, N compressions need to be computed instead of one: however, this computational time is typically not a bottleneck w.r.t. the communication time. A more important aspect is the communication cost. While the size of each message will remain identical, a different message needs to be sent to each client. That is, we go from a "one to N" configuration to N "one to one" communications. While this is a drawback, it is not an issue when the bandwidth/transfer time are the bottlenecks, as `Rand-MCM` will result in a better convergence with almost no cost. Furthermore, we argue that handling client dependent models is essential for several major applications. `Rand-MCM` can directly be adapted to those frameworks.

1. client dependent compression. A first simple situation is the case in which clients are allowed to choose the size (or equivalently the compression level) of their updates.

2. Partial participation (PP). Similarly, having N different messages to send to each client may be unavoidable in the case of *partial participation* of the clients. This is a key feature in federated learning frameworks, see Subsection 1.3.1. In the classical distributed framework (without downlink constraints) it is easy to deal with it, as each available client just queries the global model to compute its gradient on it [see for example Horváth and Richtárik, 2020]. On the other hand, for bidirectional compression, to ensure that all the local models match the central model, the adaptation to partial participation relies on a *synchronization step*. During this step, each client that has not participated in the last S steps receives the last S corresponding messages as long as it costs less to send this sequence than a full uncompressed model, see the pseudo-code of `Artemis` given in Algorithm 2 in Chapter 2.

On the contrary, `Rand-MCM` naturally handles a different model, memory, and update per client. The adaptation to partial participation is thus straightforward, without the need to catch up on missed updates or to synchronize memory. This is because the partial participation of clients is modeled as a compression scheme $\mathcal{C}_{\mathrm{PP}}$, which compresses a vector z as either z/p or 0 (see Section 2.2). Consequently, our theoretical analysis covers the case of partial participation. For sake of clarity, we present below the equations defining the downlink stage of `Rand-MCM`.

$$\begin{cases} \Omega_{k+1} = w_{k+1} - \bar{H}_k \,, \\ \widehat{w}^i_{k+1} = H^i_k + \mathcal{C}_{\mathrm{dwn},i}(\Omega_{k+1}) \\ H^i_{k+1} = H^i_k + \alpha_{\mathrm{dwn}} \mathcal{C}_{\mathrm{dwn},i}(\Omega_{k+1}) \\ \bar{H}_{k+1} = \bar{H}_k + \frac{\alpha_{\mathrm{dwn}}}{N} \sum_{i=1}^N \mathcal{C}_{\mathrm{dwn},i}(\Omega_{k+1}). \end{cases} \tag{3.10}$$

Remark 3.6 (Protecting the global model from honest-but-curious clients)**.** *Another business advantage of `MCM` and `Rand-MCM` is that providing degraded models to the participants can be used to guarantee privacy, or to ensure the clients participate in good faith, and not only to obtain the model. This issue of detecting ill-intentioned clients (free-riders) that want to obtain the model without actually contributing has been studied by Fraboni et al. [2021b].*

3.4.2 Theoretical results

In this Section, we provide two main theoretical results for `Rand-MCM`. First, Theorem 3.7 ensures that the theoretical guarantees are at least as good for `Rand-MCM` as for `MCM`. Then, in Theorem 3.8, we provide convergence result for both `MCM` and `Rand-MCM` in the case of quadratic functions.

Theorem 3.7. *Theorems 3.3 to 3.5 are valid for `Rand-MCM`*

The improvement in `Rand-MCM` comes from the fact that we are ultimately averaging the gradients at several random points, reducing the variance coming from this aspect. The goal is obviously to reduce the impact of ω_{dwn}. Keeping in mind that the dominating term in the rate is independent of

ω_{dwn}, *we can thus only expect to reduce the second-order term.* Next, the uplink compression noise increases with the variance of the randomized model, which will not be directly reduced by Rand-MCM. As a consequence, we only expect the improvement to be visible in the part of the second-order term that does not depend on ω_{up} (that is, the effect would be the most significant if ω_{up} is small or 0).

This intuition is corroborated by the following result, in which we show that the convergence is improved when adding the randomization process for a quadratic function. Extending the proof beyond quadratic functions is possible, though it requires an assumption on third or higher order derivatives of F (e.g., using self-concordance [Bach, 2010]) to control of $\mathbb{E}\left[||\nabla F(\widehat{w}_{k-1}) - \mathbb{E}[\nabla F(\widehat{w}_{k-1})]||^2 \;\middle|\; w_{k-1}\right]$.

Theorem 3.8 (Convergence in the quadratic case)**.** *Under Assumptions 3.1 to 3.4 with $\mu = 0$, if the function is quadratic, after running $K > 0$ iterations, for any $\gamma \leq \gamma_{\max}$, and we have*

$$\mathbb{E}[F(\bar{w}_K) - F_*] \leq \frac{V_0}{\gamma K} + \frac{\gamma \sigma^2 \Phi^{\mathrm{Rd}}(\gamma)}{Nb},$$

with $\Phi^{\mathrm{Rd}}(\gamma) = (1+\omega_{\mathrm{up}})\left(1 + \frac{4\gamma^2 L^2 \omega_{\mathrm{dwn}}}{K}\left(\frac{1}{\mathbf{C}} + \frac{\omega_{\mathrm{up}}}{N}\right)\right)$ *and* $\mathbf{C} = N$ *for* Rand-MCM, $\mathbf{C} = 1$ *for* MCM.

This result is derived in Section C.5. We can make the following comments. (1) The convergence rate for quadratic functions is slightly better than for smooth functions. More specifically, the right hand term in Φ is multiplied by an additional $\gamma\left(\frac{1}{\mathbf{C}} + \frac{\omega_{\mathrm{up}}}{N}\right)$ (w.r.t. Theorem 3.4), which is decaying at the same rate as γ. Besides, the proof for Rand-MCM is substantially modified, as $\mathbb{E}[\nabla F(\widehat{w}_{k-1})]$ is an unbiased estimator of $\nabla F(w_{k-1})$. (2) Moreover, the randomization in Rand-MCM further reduces by a factor N this term. Depending on the relative sizes of ω_{up} and N, this can lead to a significant improvement up to a factor of N. In practice, the impact of Rand-MCM is noticeable, as illustrated in the following experiments.

3.5 Experiments

In this section, we illustrate the validity of the theoretical results given in the previous section on both synthetic and real datasets, on (1) least-squares regression (LSR), (2) logistic regression (LR), and (3) non-convex deep learning. We compare MCM with classical algorithms used in distributed settings: Diana, Artemis, Dore and of course the simplest setting - SGD, which is the baseline.

In these experiments, we provide results on the log of the excess loss $F(w_k) - F_*$, averaged on 5 runs (resp. 2) in convex settings (resp. deep learning), with errors bars displayed on each figure, corresponding to the standard deviation of $\log_{10}(F(w_k) - F_*)$. On Figure 3.3, the X-axis is respectively the number of iterations and the number of bits exchanged.

Each experiment has been run with $N = 20$ clients using stochastic scalar quantization [Alistarh et al., 2017], w.r.t. 2-norm. To maximize compression, we always quantize on a single level ($s = 2^0$), unless for neural network (the value of s depends on the dataset).

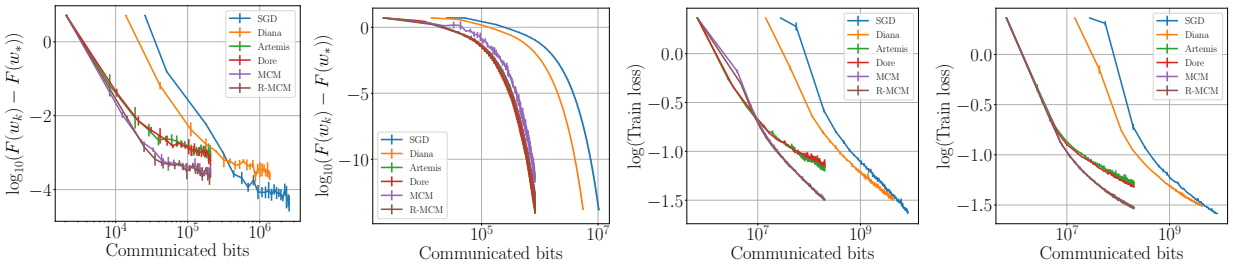
We used 9 different datasets.

- One toy dataset devoted to linear regression in an homogeneous setting. This toy dataset allows to illustrate MCM properties in a simple framework, and in particular to ilustrate that when $\sigma^2 = 0$, we recover a linear convergence[3], see Figure 3.2b.
- Five datasets commonly used in convex optimization (a9a, quantum, phishing, superconduct and w8a); see Table C.1 for more details. Experiments were conducted with heterogeneous clients obtained by clustering (using *TSNE* [Maaten and Hinton, 2008], as in Chapter 2) the input points.

[3]Even stronger, we show in experiments that we recover a linear rate if we have $\sigma_* = 0$ (the noise over stochastic gradient computation at the optimum point $w_\star$).

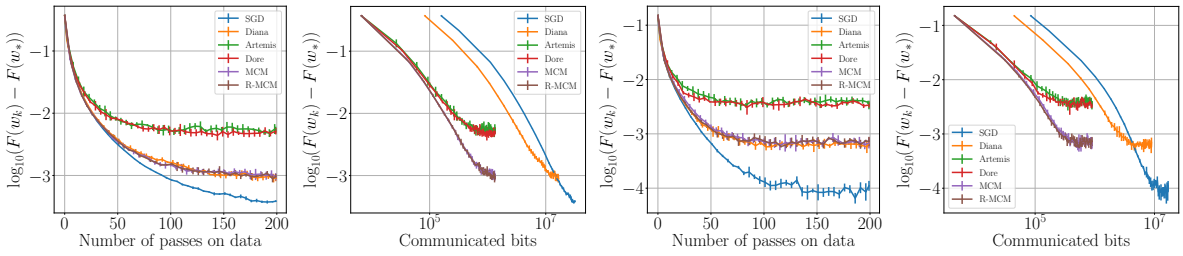
Table 3.3: Summary of experiment results for MCM - convex experiments, b is the batch size

Excess loss after 450 epochs	SGD	Diana	MCM	Dore	Ref
a9a ($b = 128$)	-3.5	-2.7	-2.7	-1.8	Chang and Lin [2011]
quantum ($b = 256$)	-3.4	-3.2	-3.2	-2.6	Caruana et al. [2004]
phishing ($b = 64$)	-3.7	-3.5	-3.4	-2.7	Chang and Lin [2011]
superconduct ($b = 64$)	-1.6	-1.6	-1.55	-1.45	Hamidieh [2018]
w8a ($b = 12$)	-3.5	-3.0	-2.5	-1.75	Chang and Lin [2011]
Compression	no	uni-dir	bi-dir	bi-dir	



(a) $\sigma^2 \neq 0$, $\gamma = (L\sqrt{k})^{-1}$ (b) $\sigma^2 = 0$, $\gamma = L^{-1}$ (c) MNIST with a CNN (d) FEMNIST with a CNN

Figure 3.2: Convergence on synthetic datasets (left) and neural networks (right).



(a) a9a in #iter. (b) a9a in #bits (c) quantum in #iter. (d) quantum in #bits

Figure 3.3: Experiments on real datasets with $\gamma = 1/L$, quantization with $s = 1$ (LR, for LSR see superconduct on Figure 2.2a).

- Four datasets in a non-convex settings (CIFAR10, Fashion-MNIST, FE-MNIST, MNIST); see Table C.2 for more details.

All experiments are performed without any tuning of the algorithms, (e.g., with the same learning rate for all algorithms and without reducing it after a certain number of epochs). Indeed, our goal is to show that our method achieves a performance close to the unidirectional-compression framework (Diana), while performing an important downlink compression. More details about experiments can be found in Section C.1.

On Figure 3.3, we display the excess loss for quantum and a9a w.r.t. the number of iteration and number of communicated bits. The plots of phising, superconduct and w8a are not provided but can be found on our github repository. We only report their excess loss after 450 iterations in Table 3.3.

Saturation level. All experiments are performed with a *constant learning rate* γ to observe the bias (initial reduction) and the variance (saturation level) independently. Stochastic gradient descent results in a fast convergence during the first iterations, and then reaches a saturation at a given level proportional to σ^2. Theorem 3.4 states that the variance of MCM is proportional to ω_{up}, this is experimentally observed on Tables 3.3 and 3.4 and Figures 3.2 and 3.3: MCM meets Diana while Artemis and Dore saturate at a higher level (scaling as $\omega_{\text{up}} \times \omega_{\text{dwn}}$). These trade-offs are preserved with optimized learning rates.

Linear convergence when $\sigma^2 = 0$. The six algorithms present a linear convergence when

$\sigma^2 = 0$. This is illustrated by Figure 3.2b: we ran experiments with a full gradient descent. Note that in these settings MCM has a slightly worse performance than other methods; however, this slow-down is compensated by Rand-MCM.

Deep learning. Table 3.4 and Figures 3.2c and 3.2d illustrate experiments with neural networks, details on dataset settings and networks architecture are given in Subsection C.1.2. Again, MCM meets Diana rates as stated by Theorem 3.6 (non-convex case).

Table 3.4: Accuracy and train loss in non-convex experiments, detailed settings can be found in Table C.2.

	Algorithm	MNIST	Fashion MNIST	FE-MNIST	CIFAR-10
Accuracy after 300 epochs	SGD:	99.0%	92.4%	99.0%	69.1%
	Diana:	98.9%	92.4%	98.9%	64.0%
	MCM:	98.8%	90.6%	98.9%	63.5%
	Artemis:	97.9%	86.7%	98.3%	54.8%
	Dore:	97.9%	87.9%	98.5%	56.3%
Train loss after 300 epochs	SGD:	0.025	0.093	0.026	0.909
	Diana:	0.034	0.141	0.031	1.047
	MCM:	0.033	0.209	0.030	1.096
	Artemis:	0.075	0.332	0.052	1.342
	Dore:	0.072	0.300	0.048	1.292

Impact of randomization. The impact of randomization is noticeable on Figures 3.2b and C.5b. Randomization helps to stabilize convergence, it reduces the variance over the runs, and when $\sigma^2 = 0$, it performs identically to SGD.

Overall, these experiments show the benefits of MCM and Rand-MCM, that reach the saturation level of Diana while exchanging at 10x to 100x fewer bits. All the code is provided on our github repository.

3.6 Conclusion

In this work, we propose a new algorithm to perform bidirectional compression while achieving the convergence rate of algorithms using compression in a single direction. One of the main application of this framework is federated learning. With MCM we stress the importance of not degrading the global model. In addition, we add the concept of randomization which allows to reduce the variance associated with the downlink compression. The analysis of MCM is challenging as the algorithm involves perturbed iterates. Proposing such an analysis is the key to unlocking numerous challenges in distributed learning, e.g., proposing practical algorithms for partial participation, incorporating privacy-preserving schemes *after* the global update is performed, dealing with local steps, etc. This approach could also be pivotal in non-smooth frameworks, as it can be considered as a weak form of randomized smoothing.

4

Convergence rates for distributed, compressed and averaged least-squares regression: application to Federated Learning

"Бог уже потому мне необходим, что это единственное существо, которое можно вечно любить ..."

Бесы, Федор Михайлович Достоевский.

In this Chapter, we go beyond the classical worst-case assumption on the variance of compressors (Assumption 2.5 in Chapter 2 or Assumption 3.1 in Chapter 3) and provide a fine-grained analysis of the impact of compression within the fundamental learning framework of least-squares regression (LSR). Within this setting, we underline differences in terms of convergence rates between several unbiased compression operators, that all satisfy the same condition on their variance, thus going beyond the classical worst-case analysis. To do so, we analyze a general stochastic approximation algorithm for minimizing quadratic functions relying on a random field. We consider weak assumptions on the random field, tailored to the analysis (specifically, expected Hölder regularity), and on the noise covariance, enabling the analysis of various randomizing mechanisms, including compression. We then extend our results to the case of federated learning.

More formally, we highlight the impact on the convergence of the covariance $\mathfrak{C}_{\text{ania}}$ of the *additive noise induced by the algorithm*. We demonstrate despite the non-regularity of the stochastic field, that the limit variance term scales with $\operatorname{Tr}\left(\mathfrak{C}_{\text{ania}} H_F^{-1}\right)/K$ (where H_F is the Hessian of the optimization problem and K the number of iterations) generalizing the rate for the vanilla LSR case where it is $\sigma^2 \operatorname{Tr}\left(H_F H_F^{-1}\right)/K = \sigma^2 d/K$ [Bach and Moulines, 2013]. Then, we analyze the dependency of $\mathfrak{C}_{\text{ania}}$ on the compression strategy and ultimately its impact on convergence, first in the centralized case, then in two heterogeneous FL frameworks.

This Chapter is based on our work *Convergence rates for distributed, compressed and averaged least-squares regression: application to federated learning* [Philippenko and Dieuleveut, 2023] submitted at JMLR.

Contents

4.1 Introduction

Large-scale optimization [Bottou and Bousquet, 2007] has become ubiquitous in today's learning problems due to the incredible growth of data collection. It becomes computationally extremely hard to process a full dataset or even, to store it on a single device [Abadi et al., 2016, Seide and Agarwal, 2016, Caldas et al., 2019]. This led practitioners to either process each observation only once in a streaming fashion and to design distributed algorithms. This Chapter is part of this line of work and considers in particular stochastic federated algorithms [Konečný et al., 2016, McMahan et al., 2017] that use a central server to orchestrate the training over a network of N in $\mathbb{N}^*$ clients.

A well-identified challenge in this framework is the communication cost of the learning process [Seide et al., 2014, Chilimbi et al., 2014, Strom, 2015] based on stochastic gradient algorithms. Indeed, iteratively exchanging gradient or model information between the local workers and the central server generates a huge computational and bandwidth bottleneck. To reduce this communication cost, two strategies have been widely implemented and analyzed: performing local updates [see e.g. McMahan et al., 2017, Karimireddy et al., 2020], or reducing the size of the exchanged messages by passing them through a compression operator, on the uplink channel [Seide et al., 2014, Alistarh et al., 2017, 2018, Mishchenko et al., 2019, Karimireddy et al., 2019, Wu et al., 2018, Horvath et al., 2022, Mishchenko et al., 2019, Khirirat et al., 2020a, Grishchenko et al., 2021, Richtarik et al., 2021], or on both uplink and downlink channels [Harrane et al., 2018, Tang et al., 2019, Liu et al., 2020, Zheng et al., 2019, Philippenko and Dieuleveut, 2020, 2021, Gorbunov et al., 2020b, Sattler et al., 2019, Fatkhullin et al., 2021]. These two strategies, although typically analyzed independently, are often combined. We focus on compression; to reduce the cost of exchanging a vector, three techniques are combined: (1) sending the message to only a few clients, (2) sending only a fraction of the coordinates, (3) sending low-precision updates.

Most analyses of the impact of compression schema rely on generic assumptions on the compression operator $\mathcal{C}$, typically either *contractive*, i.e. for any z in $\mathbb{R}^d$, $\|\mathcal{C}(z) - z\| < (1-\delta)\|z\|$ with $\delta \in]0;1[$ [almost surely or in expectation, see for instance Seide et al., 2014, Stich et al., 2018, Karimireddy et al.,

2019, Ivkin et al., 2019, Koloskova et al., 2019, Gorbunov et al., 2020b, Beznosikov et al., 2020], or unbiased with bounded variance increase, i.e., for any z in $\mathbb{R}^d$, $\mathbb{E}[\mathcal{C}(z)] = z$ and $\mathbb{E}[\|\mathcal{C}(z)-z\|^2] \leq \omega\|z\|^2$ for a parameter $\omega > 1$ [see among others Alistarh et al., 2017, Wu et al., 2018, Mishchenko et al., 2019, Chraibi et al., 2019, Gorbunov et al., 2020a, Reisizadeh et al., 2020, Horvath et al., 2022, Kovalev et al., 2021, Philippenko and Dieuleveut, 2020, 2021, Haddadpour et al., 2021, Li and Richtárik, 2021, Khirirat et al., 2018]. Unlike biased – and often deterministic – operators, unbiased operators typically benefit from a variance reduction proportional to the number of clients (e.g., Gorbunov et al., 2020b vs Horváth et al., 2019).

In parallel, a line of work has thus focused on the design of compression schemes satisfying one of these two assumptions [Bernstein et al., 2018, Dai et al., 2019, Beznosikov et al., 2020, Horvath et al., 2022, Xu et al., 2020, Leconte et al., 2021, Gandikota et al., 2021, Ramezani-Kebrya et al., 2021, Horvath et al., 2022]. Two fundamental strategies are typically combined: (1) quantization [Rabbat and Nowak, 2005, Gersho and Gray, 2012, Alistarh et al., 2018], and (2) random projection [Vempala, 2005, Rahimi and Recht, 2008, Nesterov, 2012, Nutini et al., 2015]. These methods are compared based on (1) the number of bits required for storing or exchanging a d dimensional vector and (2) the resulting variance increase ω or contractiveness constant δ. Consequently, convergence results are *worst-case* results over the class of compression operators: two compression operators satisfying the same variance assumption are regarded as producing the same convergence rate.

The goal of this Chapter is to provide an in-depth analysis of compression within a fundamental learning framework, namely least-squares regression [LSR, Legendre, 1806], in order to highlight the differences in convergence between several unbiased compression schemes having the *same* variance increase.

Especially, this analysis will highlight the impact of (1) the compression scheme's regularity (Lipschitz in squared expectation or not) and of (2) the correlation between the compression of the different coordinates. We highlight three examples of possible take-aways from our analysis, that will be detailed in Section 4.3.

Take-away 1. *Quantization-based compression schemes do not have* Lipschitz in squared expectation *regularity but satisfy a Hölder condition. Because of that, their convergence is degraded, yet they asymptotically achieve a rate comparable to projection-based compressors, in which the limit covariance is similar.*

Take-away 2. *Rand-h and partial participation with probability (h/d) satisfy the same variance condition. Yet the convergence of compressed least mean squares algorithms for partial participation is more robust to ill-conditioned problems.*

Take-away 3. *The asymptotic convergence rate is expected to be at least as good for quantization than for sparsification or randomized coordinate selection,* if *the features are standardized. On the contrary, if the features are independent and the feature vector is normalized, then quantization is worse than sparsification or randomized coordinate selection.*

We consider a random-design LSR framework and make the following assumption on the input-output pairs distribution

Model 1 (Federated case)**.** *We consider N clients. Each client i in $\{1, \dots, N\}$ accesses K in $\mathbb{N}^*$ i.i.d. observations $(x_k^i, y_k^i)_{k\in\{1,\dots,K\}} \sim \mathcal{D}_i^{\otimes K}$, such that there exists a well-defined client-dependent model w_*^i:*

$$\forall k \in \{1, \dots, K\}, \quad y_k^i = \langle x_k^i, w_*^i \rangle + \varepsilon_k^i, \qquad \text{with} \;\; \varepsilon_k^i \sim \mathcal{N}(0, \sigma^2) \;\; , \tag{4.1}$$

for an i.i.d. sequence $\left((\varepsilon_k^i)_{k\in\{1,\dots,K\},i\in\{1,\dots,N\}}\right)$ *independent from* $\left((x_k^i)_{k\in\{1,\dots,K\},i\in\{1,\dots,N\}}\right)$. *We use the generic notation* $(x^i, y^i, \varepsilon^i)$ *for such an input-output-noise triplet on client* i. *Moreover, we assume that the inputs' second moment*[1] *is bounded to define* $\mathbb{E}[x^i \otimes x^i] = H_i$ *and* $\mathbb{E}[\|x^i\|^2] = R_i^2$; *such that* $\mathbb{E}[\|x^i\|^2 x^i \otimes x^i] \preccurlyeq R_i^2 H_i$. *For any* $i \in \{1,\dots,N\}$, *we consider the expected squared loss on client* i *of a model* w *as* $F_i(w) := \frac{1}{2}\mathbb{E}_{(x^i,y^i)\sim\mathcal{D}_i}[(\langle x^i, w\rangle - y^i)^2]$.

Remark 4.1 (Almost surely bounded features). *In the case of linear compressors, we will also assume that for each client* i *in* $\{1,\dots,N\}$, *features are almost surely bounded by* R_i^2.

This model is classical in the single worker case [e.g. Hsu et al., 2012, Bach and Moulines, 2013]:

Model 2 (Centralized case). *We consider Model 1 with* $N = 1$ *client. For simplicity, we then omit the* i *superscript.*

We focus on the problem of minimizing the global expected risk $F : \mathbb{R}^d \to \mathbb{R}$, thus finding the optimal model w_* in $\mathbb{R}^d$ such that:

$$w_* = \operatorname*{arg\,min}_{w\in\mathbb{R}^d} \left\{ F(w) := \frac{1}{N}\sum_{i=1}^{N} F_i(w) \right\}. \tag{OPT}$$

The $(F_i)_{i=1}^N$ are the squared loss defined on each client defined in Model 1. Note that we assume that $\mathrm{Span}\{\mathrm{Supp}(x^i), i \in \{1,\dots,N\}\} = \mathbb{R}^d$ to ensure the existence and uniqueness of w_*.

The empirical version of the risk minimization admits an explicit formula, yet is computationally too expensive to compute for large problems. This is why, in practice, LSR is solved using iterative stochastic algorithms, for example Stochastic Gradient Descent [SGD, see Robbins and Monro, 1951]. SGD for LSR is often referred to as the *Least Mean Squares* (LMS) algorithm [Bershad, 1986, Macchi, 1995]. Analysis of LMS [Györfi and Walk, 1996, Bach and Moulines, 2013] and its variants received a lot of interest over the last decades. Indeed despite its simplicity, LSR is a model of choice for practitioners because of its efficiency to train good and interpretable models [see e.g. Molnar, 2018, chapter 5.1]. Moreover, its simplicity enables to isolate and analyze challenges faced in specific configurations, for instance, non-strong convexity [Bach and Moulines, 2013], interaction between acceleration and stochasticity [Dieuleveut et al., 2017, Jain et al., 2018a, Varre and Flammarion, 2022], non-uniform iterate averaging [Jain et al., 2018b, Neu and Rosasco, 2018, Muecke et al., 2019], infinite-dimensional frameworks [Dieuleveut and Bach, 2016], biased compression and error-compensation mechanism [Khirirat et al., 2020a], or over-parametrized regimes and double descent phenomena [Belkin et al., 2019].

Our approach follows this line of work: our goal is to analyze the impact of *compression* in FL algorithms, by providing a careful study of compressed LMS, based on a fine-grained analysis of Stochastic Approximation (SA) under weak assumptions on the random field. More precisely, we consider linear stochastic approximation recursion, to find a zero of the linear mean-field ∇F.

Definition 4.1 (Linear Stochastic Approximation, LSA). *Let* $w_0 \in \mathbb{R}^d$ *be the initialization, the linear*[2] *stochastic approximation recursion is defined as:*

$$w_k = w_{k-1} - \gamma \nabla F(w_{k-1}) + \gamma \xi_k (w_{k-1} - w_*), \quad k \in \mathbb{N}, \tag{LSA}$$

where $\gamma > 0$ *is the step-size and* $(\xi_k)_{k\in\mathbb{N}^*}$ *is a sequence of i.i.d. zero-centered random fields that characterizes the stochastic oracle on* $\nabla F(\cdot)$. *For any* $k \in \mathbb{N}^*$, *we denote* $\mathcal{F}_k = \sigma(\xi_1,\dots,\xi_k)$, *such that the filtration* $(\mathcal{F}_k)_{k\geq 0}$ *is adapted to* $(w_k)_{k\geq 0}$.

[1] In the following, we may refer to this matrix H as the *covariance* (in the case of centered features, covariance is equal to the second moment)

[2] While in LSA literature, both the mean-field ∇F and the noise-field (ξ_k) are linear, we do not here consider the noise fields to be linear.

We assume that F is quadratic, we denote H_F its Hessian, $R_F^2 := \operatorname{Tr}(H_F)$ its trace and μ its smallest eigenvalue. For any k in $\mathbb{N}$, with $\eta_k = w_k - w_*$, we get equivalently:

$$\eta_k = (\mathrm{I} - \gamma H_F)\eta_{k-1} + \gamma\xi_k(\eta_{k-1}), \quad k \in \mathbb{N}.$$

As underlined by Bach and Moulines [2013], (LSA) corresponds to a homogeneous Markov chain. A study of stochastic approximation using results and techniques from the Markov chain literature can be found for instance in Freidlin and Wentzell [1998] or more recently in Dieuleveut et al. [2020].

(LSA) encompasses three examples of interest, the first one is the classical LMS algorithm. Indeed, with the observations in Models 1 and 2, for any client $i \in \{1, \dots, N\}$, any iteration k in $\{1, \dots, K\}$, any model $w \in \mathbb{R}^d$,

$$g_k^i(w) := (\langle x_k^i, w\rangle - y_k^i)x_k^i \tag{4.2}$$

is an unbiased oracle of $\nabla F_i(w)$. This can be used to define the following three algorithms.

Algorithm 1 (LMS). *For LMS algorithm, with a single worker, we have for all $k \in \mathbb{N}$, $w_k = w_{k-1} - \gamma g_k(w_{k-1}) = w_{k-1} - \gamma(\langle x_k, w_{k-1}\rangle - y_k)x_k$, thus equivalently $\xi_k(\cdot) = (x_k x_k^\top - \mathbb{E}[x_1 x_1^\top])(\cdot) + (\langle w_*, x_k\rangle - y_k)x_k$.*

Second, the case of a single client compressed LMS algorithm.

Algorithm 2 (Centralized compressed LMS). *A single client ($N = 1$) observes at any step $k \in \{1, \dots, K\}$ an oracle $g_k(\cdot)$ on the gradient of the objective function F, and applies a random compression mechanism $\mathcal{C}_k(\cdot)$. Thus, for any step-size $\gamma > 0$ and any $k \in \mathbb{N}^*$, the resulting sequence of iterates $(w_k)_{k\in\mathbb{N}}$ satisfies: $w_k = w_{k-1} - \gamma\mathcal{C}_k(g_k(w_{k-1}))$.*

And finally, the extension to the distributed case.

Algorithm 3 (Distributed compressed LMS). *In our motivating example, each client $i \in \{1, \dots, N\}$ observes at any step $k \in \{1, \dots, K\}$ an oracle $g_k^i(\cdot)$ on the gradient of the local objective function F_i, and applies a random compression mechanism $\mathcal{C}_k^i(\cdot)$. Thus, for any step-size $\gamma > 0$ and any $k \in \mathbb{N}^*$, the resulting sequence of iterates $(w_k)_{k\in\mathbb{N}}$ satisfies: $w_k = w_{k-1} - \frac{\gamma}{N}\sum_{i=1}^N \mathcal{C}_k^i(g_k^i(w_{k-1}))$ (we consider the randomization made on clients $(\mathcal{C}_k^i(\cdot))_{i\in\{1\cdots,N\}}$ to be independent)*

Remark 4.2. *The analysis naturally covers any randomized postprocessing $\mathcal{C}_k^i(\cdot)$, beyond the compression case.*

Challenges, contributions and structure of the Chapter. Although there is abundant literature on the study of (LSA), the application to Algorithms 2 and 3 poses novel challenges. Especially, most analyses of LSA [Blum, 1954, Ljung, 1977, Ljung and Söderström, 1983] assume that the field ξ_k is linear [i.e. for any $z, z' \in \mathbb{R}^d$, $\xi_k(z) - \xi_k(z') = \xi_k(z - z')$, see Konda and Tsitsiklis, 2003, Benveniste et al., 2012, Leluc and Portier, 2022]. More general non-asymptotic results on SA with a Lipschitz mean-field (i.e. SGD with a smooth objective) also assume that the noise-field is Lipschitz-in-squared-expectation i.e. for any $z, z' \in \mathbb{R}^d, \mathbb{E}[\|\xi_k(z) - \xi_k(z')\|^2] \leq C\|z - z'\|^2$ [Moulines and Bach, 2011, Bach, 2014, Dieuleveut et al., 2020, Gadat and Panloup, 2023]. One major specificity and bottleneck in the case of compression is the fact that the resulting field **does not** satisfy such an assumption. The rest of the Chapter is thus organized as follows:

1. In Section 4.2, we provide a non-asymptotic analysis of (LSA) under weak regularity assumptions of the noise field $(\xi_k)_k$. We show that the asymptotically dominant term depends on the covariance matrix $\mathfrak{C}_{\text{ania}}$ of the *additive noise induced by the algorithm*, as expected from the classical asymptotic literature [Polyak and Juditsky, 1992]. The backbone results of our Chapter are Theorems 4.1 and 4.2 which generalize the results from Bach and Moulines [2013] for Algorithm 1. The limit convergence rate term scales with $\operatorname{Tr}\left(\mathfrak{C}_{\text{ania}}H_F^{-1}\right)/K$, which highlights the interaction between the Hessian of the optimization problem H_F, and the additive noise's covariance $\mathfrak{C}_{\text{ania}}$.

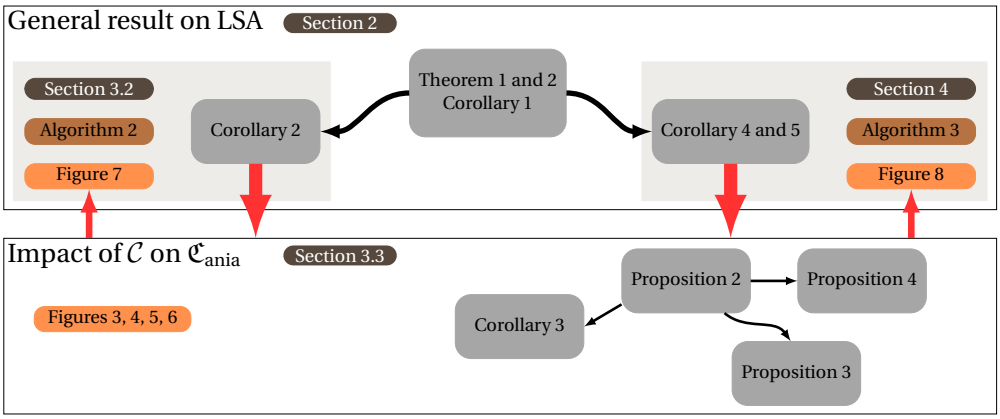


Figure 4.1: Flow chart summarizing our results.

2. In Section 4.3, we prove that assumptions made in Section 4.2 are valid for Algorithm 2 with classical compression schemes. Although this single-client case is a simple configuration, it enables to describe the impact of the compressor choice on the dependency between the features' covariance H (which is also the Hessian H_F of the optimization problem) and the additive noise's covariance $\mathfrak{C}_{\text{ania}}$. Contrary to Algorithm 1, for which the noise is said to be *structured*, i.e. the additive noise's covariance is proportional to the Hessian H_F, applying a random compression mechanism on the gradient breaks this structure. This phenomenon is noteworthy: for an ill-conditioned H_F, it may lead to a drastic increase in $\operatorname{Tr}\left(\mathfrak{C}_{\text{ania}} H_F^{-1}\right)$ and thus, to a degradation in convergence. By calculating the additive noise's covariance for various compression mechanisms, we identify differences that classical literature was unable to capture.
3. In Section 4.4, we study the distributed Algorithm 3 with heterogeneous clients. We examine two different sources of heterogeneity for which we show that Theorems 4.1 and 4.2 remain valid. First, the case of heterogeneous features' covariances $(H_i)_{i=1}^N$ in Subsection 4.4.1; second, the case of heterogeneous local optimal points $(w_*^i)_{i=1}^N$ in Subsection 4.4.2.

These results are validated by numerical experiments which help to get an intuition of the underlying mechanisms. The code is provided on our GitHub repository. We summarize the structure of the Chapter in Figure 4.1.

Notations. We denote by $\preccurlyeq$ the order between self-adjoint operators, i.e., $A \preccurlyeq B$ if and only if $B - A$ is positive semi-definite (p.s.d.) and $A \overset{\sim}{\preccurlyeq} B$ if $A \preccurlyeq B$ and $A = B + O(\frac{1}{d})$. We denote by $A^{1/2}$ the p.s.d. square root of any symmetric p.s.d. matrix A. For two vectors x, y in $\mathbb{R}^d$, the Kronecker product is defined as $x \otimes y := xy^\top$, the element-wise product is denoted as $x \odot y$, and the Euclidean norm is $\|x\|^2 := \sum_{i=1}^d x_i^2$. For any rectangular matrix A in $\mathbb{R}^{n \times m}$ s.t. $AA^\top$ is inversible, we denote $A^\dagger := A^\top (AA^\top)^{-1}$ the Moore–Penrose pseudo inverse. For x, y in $\mathbb{R}^d$, we use $x \wedge y$ for the minimum between two values, and $x \overset{\sim}{\leq} y$ if $x \leq y$ and $x = y + O(\frac{1}{d})$. For any sequence of vector $(x_k)_{k \in \{0,\dots,K\}}$ we denote $\overline{x}_{K-1} = \sum_{k=0}^{K-1} x_k / K$. We use e_i to denote the vector in $\mathbb{R}^d$ with zero everywhere except at coordinate i, and $\mathcal{O}_d(\mathbb{R})$ the group of orthogonal matrices. Finally, all random variables are defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, $\mathbb{E}$ is the expectation associated with the probability $\mathbb{P}$ and $\mathcal{A}$ is a σ-algebra. We define the set of probability distribution function $\mathcal{P}_M$ whose second moment is equal to M in $\mathbb{R}^{d \times d}$: $\mathcal{P}_M = \{\text{probability distribution } p_M \text{ over } \mathbb{R}^d \text{ s.t.}, \mathbb{E}_{\varepsilon \sim p_M}[\varepsilon^{\otimes 2}] = M\}$. Any such distribution p_M is indexed with its matrix of covariance.

4.2 Non asymptotic convergence result for (LSA)

4.2.1 Definition of the additive noise's covariance and assumptions on the random fields

For any k in $\mathbb{N}^*$, we define the additive noise ξ_k^{add} and the multiplicative noise $\xi_k^{\text{mult}}(\cdot)$.

Definition 4.2 (Additive and multiplicative noise)**.** *Under the setting of Definition 4.1, for any k in $\mathbb{N}^*$, we define:*

$$\xi_k^{\text{add}} := \xi_k(0) \qquad \text{and} \qquad \xi_k^{\text{mult}} : z \in \mathbb{R}^d \mapsto \xi_k(z) - \xi_k^{\text{add}} .$$

Remark 4.3. *Observe that $(\xi_k^{\text{add}})_{k\in\mathbb{N}^*}$ is an i.i.d. sequence of random variables and $(\xi_k^{\text{mult}})_{k\in\mathbb{N}^*}$ is an i.i.d. sequence of random field. The following assumptions, made for $k = 1$, are thus equivalently valid for any $k \geq 1$.*

Assumption 4.1 (Second moment)**.** *ξ_1^{add} admits a second order moment. We note $\mathcal{A}$ in $\mathbb{R}^d$ s.t. $\mathbb{E}[\|\xi_1^{\text{add}}\|^2] \leq \mathcal{A}$.*

Assumption 4.1 and Remark 4.3 enable us to define the covariance of the additive noise induced by the algorithm.

Definition 4.3 (Additive noise's induced by the algorithm's covariance.)**.** *Under the setting of Definition 4.1, we define the* additive noise's covariance *as the covariance of the additive noise:* $\mathfrak{C}_{\text{ania}} = \mathbb{E}[\xi_1^{\text{add}} \otimes \xi_1^{\text{add}}]$.

Secondly, we state our assumptions on the multiplicative part of the noise, especially its regularity around 0 (note that $\xi_1^{\text{mult}}(0) = 0$).

Assumption 4.2 (Second moment of the multiplicative noise)**.** *There exist two constants $\mathcal{M}_1, \mathcal{M}_2 > 0$ such that, for any η in $\mathbb{R}^d$, the following hold:*

A4.2.1: $\mathbb{E}[\|\xi_1^{\text{mult}}(\eta)\|^2] \leq 2\mathcal{M}_2 \|H_F^{1/2}\eta\|^2 + 4\mathcal{A}$.

A4.2.2: $\mathbb{E}[\|\xi_1^{\text{mult}}(\eta)\|^2] \leq \mathcal{M}_1 \|H_F^{1/2}\eta\| + 3\mathcal{M}_2 \|H_F^{1/2}\eta\|^2$.

The main originality of this section is the analysis under Assumption 4.2.2. This Hölder-type condition will appear naturally for compression in Section 4.3. Up to our knowledge, (LSA) has not been analyzed under this particular condition.

Under these assumptions, asymptotic results from Polyak and Juditsky [1992] can be applied. Especially, we establish the asymptotic normality of $(\sqrt{K}\overline{\eta}_{K-1})_{K>0}$, with an asymptotic variance equal to $H_F^{-1}\mathfrak{C}_{\text{ania}}H_F^{-1}$.

Proposition 4.1 (CLT for (LSA))**.** *Under Assumptions 4.1 and 4.2, consider a sequence $(w_k)_{k\in\mathbb{N}^*}$ produced in the setting of Definition 4.1 for a step-size $(\gamma_k)_{k\in\mathbb{N}^*}$ s.t. $\gamma_k = k^{-\alpha}$, $\alpha \in]0, 1[$. Then $(\sqrt{K}\overline{\eta}_{K-1})_{K>0}$ is asymptotically normal and converge in distribution to $\mathcal{N}(0, H_F^{-1}\mathfrak{C}_{\text{ania}}H_F^{-1})$.*

The proof of this result is almost straightforward and is recalled in Subsection D.1.3. In the following, we establish non-asymptotic results in Theorems 4.1 and 4.2, that highlight the impact of Assumption 4.2.2.

4.2.2 Convergence rates for (LSA), general case

In this section, we present non-asymptotic convergence rates for (LSA) under the assumptions above. These results build upon the work of Bach and Moulines [2013]. Our first result is the main result, under the Hölder assumption on the noise field, it is demonstrated in Section D.2.

Theorem 4.1 (Non-linear multiplicative noise)**.** *Under Assumptions 4.1 and 4.2, consider a sequence $(w_k)_{k\in\mathbb{N}^*}$ produced in the setting of Definition 4.1 for a constant step-size γ such that*

$\gamma(R_F^2 + 2\mathcal{M}_2) \leq 1/2$. *Then for any horizon* K, *we have:*

$$\mathbb{E}[F(\overline{w}_{K-1}) - F(w_*)] \leq \frac{1}{2K}\left(\frac{\|H_F^{-1/2}\eta_0\|}{\gamma\sqrt{K}} \wedge \frac{\|\eta_0\|}{\sqrt{\gamma}} + \sqrt{\mathrm{Tr}\left(\mathfrak{C}_{\mathrm{ania}}H_F^{-1}\right)} + (10\mathcal{A}\gamma)^{1/4}\sqrt{\mathcal{M}_1\mu^{-1}} + (30\mathcal{A}\gamma)^{1/2}\sqrt{\mathcal{M}_2\mu^{-1}}\right)^2.$$

The first two terms of the RHS correspond respectively to the impact of the initial condition η_0 and the impact of the additive noise. The dependency on these two terms is similar to the one established in Bach and Moulines [2013] in the case of LMS. Note that following Defossez and Bach [2015], we improve the dependency on the initial condition to $\frac{\|\eta_0\|^2}{\gamma K} \wedge \frac{\|H_F^{-1/2}\eta_0\|^2}{\gamma^2 K^2}$. Regarding the noise term, the dependency on $\frac{\mathrm{Tr}\left(\mathfrak{C}_{\mathrm{ania}}H_F^{-1}\right)}{2K}$ corresponds to the classical asymptotic noise term in CLT for Stochastic Approximation [e.g., Delyon, 1996, Duflo, 1997, Györfi and Walk, 1996]. In fact, for a sequence of step sizes γ_t decreasing to zero, we recover the variance from Proposition 4.1. Remark that in [Bach and Moulines, 2013] and several follow up works, the algorithm under consideration is LMS (Algorithm 1, which enables to ensure that $\mathfrak{C}_{\mathrm{ania}} \preccurlyeq \sigma^2 H_F$: the variance term thus scales as $\sigma^2 d/K$. On the contrary, Algorithms 2 and 3 do not always satisfy $\mathfrak{C}_{\mathrm{ania}} \preccurlyeq \sigma^2 H_F$: in such case, $\mathrm{Tr}\left(\mathfrak{C}_{\mathrm{ania}}H_F^{-1}\right)$ may scale as $1/\mu$.

The third and fourth term, that scale respectively as $\sqrt{\gamma}/K$ and γ/K, are asymptotically negligible for $\gamma = o(1)$. Those terms are proportional to the Hölder-regularity constants $\mathcal{M}_1, \mathcal{M}_2$, and also increase with μ^{-1}. The dominant term is $\frac{\mathcal{M}_1\sqrt{10\mathcal{A}\gamma}}{\mu K}$. Interestingly, when γ is constant (not decreasing with K), then the limit variance of the algorithm is affected. Moreover, contrary to [Bach and Moulines, 2013], we do not recover a convergence rate independent of μ. This dependency is un-avoidable as the multiplicative noise is only controlled around w_*: without strong-convexity, the iterates may not converge to w_*. While these additional terms in the variance may be considered as a drawback, it can be mitigated by taking a step-size γ proportional to $1/K^\alpha$ with $\alpha > 0$ small (γ is horizon dependent, but constant).

Corollary 4.1. *Under the assumptions of Theorem 4.1, with* $\gamma = 1/K^\alpha$, *and* $\alpha \in]0, 1/2[$, *we have:*

$$\mathbb{E}[F(\overline{w}_{K-1}) - F(w_*)] \leq \frac{60}{K}\left(\mathrm{Tr}\left(\mathfrak{C}_{\mathrm{ania}}H_F^{-1}\right) + \frac{\|H_F^{-1/2}\eta_0\|^2}{K^{(1-2\alpha)}} + \frac{\mathcal{M}_1\sqrt{\mathcal{A}}}{\mu K^{\alpha/2}} + \frac{\mathcal{M}_2\mathcal{A}}{\mu K^\alpha}\right).$$

The decrease of the second order terms is then optimized for $\alpha = 2/5$. To highlight the impact of the non-linearity in compression schemes, we provide for comparison the result for a linear multiplicative noise.

4.2.3 Convergence rates for (LSA), linear case

Alternatively, to cover the particular case of a linear multiplicative noise (e.g., to recover LMS or projection-based compressed LMS) we make the following stronger hypothesis:

Assumption 4.3. *The multiplicative noise is linear i.e. there exists a random matrix* Ξ_1 *in* $\mathbb{R}^{d\times d}$ *s.t. for any* η *in* $\mathbb{R}^d$, *we have a.s.* $\xi_1^{\mathrm{mult}}(\eta) = \Xi_1\eta$. *Moreover* $\mathbb{E}[\|\xi_1^{\mathrm{mult}}(\eta)\|^2] \leq \mathcal{M}_2\|H_F^{1/2}\eta\|^2$.

Remark 4.4. *Note that* Ξ_1 *is not necessarily symmetric (in Algorithms 2 and 3, this results from the compression).*

In addition to Assumption 4.3, in the case of linear multiplicative noise, we also consider the following assumption.

Assumption 4.4. *The following hold.*

A4.4.1: *There exists a constant*[3] $\text{Ш}_{\text{add}} > 0$ *s.t.* $\mathfrak{C}_{\text{ania}} \preccurlyeq \text{Ш}_{\text{add}} H_F$.

A4.4.2: *There exists a constant* $\text{Ш}_{\text{mult}} > 0$*, such that* $\mathbb{E}\left[\Xi_1 \Xi_1^\top\right] \preccurlyeq \text{Ш}_{\text{mult}} H_F$.

Remark 4.5 (Link between Assumptions 4.1, 4.2 and 4.4)**.** *Assumption 4.1 (resp. Assumption 4.2) corresponds to an assumption on the second order moment of the additive noise (resp. multiplicative), while Assumption 4.4.1 (resp. Assumption 4.4.2) is a (stronger) assumption on its covariance.*

Theorem 4.2 (Linear multiplicative noise)**.** *Under Assumptions 4.1, 4.3 and 4.4, i.e., with a linear multiplicative noise. Consider a sequence* $(w_k)_{k\in\mathbb{N}^*}$ *produced in the setting of Definition 4.1, for a constant step-size* γ *such that* $\gamma(R_F^2 + \mathcal{M}_2) \leq 1$ *and* $4\text{Ш}_{\text{mult}}\gamma \leq 1$*. Then for any horizon* K*, we have*

$$\mathbb{E}[F(\overline{w}_{K-1}) - F(w_*)] \leq \frac{1}{2K}\left(\frac{\|\eta_0\|}{\sqrt{\gamma}} + \sqrt{\operatorname{Tr}\left(\mathfrak{C}_{\text{ania}} H_F^{-1}\right)} + 2\left(\gamma d \text{Ш}_{\text{add}} \text{Ш}_{\text{mult}}\right)^{1/2}\right)^2 .$$

Theorem 4.2 generalizes Theorem 1 from Bach and Moulines [2013]. It also highlights the impact of additive noise's covariance, and the comparison between Theorem 4.1 and Theorem 4.2 shows the advantage of linear compression schemes. Indeed the variance scales as $K^{-1}(\operatorname{Tr}\left(\mathfrak{C}_{\text{ania}} H_F^{-1}\right) + 4\gamma d \text{Ш}_{\text{add}} \text{Ш}_{\text{mult}})$. As before, the first term $\operatorname{Tr}\left(\mathfrak{C}_{\text{ania}} H_F^{-1}\right)$ corresponds to the asymptotic variance given in Proposition 4.1, and the second term is negligible: (i) for all $4\text{Ш}_{\text{mult}}\gamma \leq 1$ it can be upper bounded by $d\text{Ш}_{\text{add}}$, and for LMS [see Bach and Moulines, 2013], the variance term is $\operatorname{Tr}\left(\mathfrak{C}_{\text{ania}} H_F^{-1}\right) = d\sigma^2$, which is thus at least as large, (ii) it scales with γ thus is asymptotically negligible as γ tends to 0. Overall, depending on $\mathfrak{C}_{\text{ania}}$, the algorithm may or may not suffer from the lack of strong-convexity (μ tending to 0). More precisely, in the case of linear multiplicative noise, we can obtain a $O(K^{-1})$ rate independent of μ if and only if $\mathfrak{C}_{\text{ania}} \preccurlyeq a H_F$, with a in $\mathbb{R}$. The proof of Theorem 4.2 is given in Section D.3, and follows the line of proof of Bach and Moulines [2013].

Conclusion: we established rates for (LSA) for both the Hölder-noise case and the linear noise case. In the former, convergence requires strong convexity while in the latter, we can achieve $O(K^{-1})$ for $\mathfrak{C}_{\text{ania}} \preccurlyeq a H_F$. In both cases, the dominant term for an optimal choice of γ scales as $\frac{\operatorname{Tr}\left(\mathfrak{C}_{\text{ania}} H_F^{-1}\right)}{K}$.

In the following section, we turn to the analysis of Algorithm 2: we show how the choice of the compression impacts both the linearity of the noise and the structure of $\mathfrak{C}_{\text{ania}}$.

4.3 Application to Algorithm 2: compressed LSR on a single worker

In this section, we analyze Algorithm 2, i.e. compressed LSR. In Subsection 4.3.1, we introduce the compression operators of interest and verify in Subsection 4.3.2 that Theorems 4.1 and 4.2 can be applied. Then, in Subsection 4.3.3, we provide explicit formulas of $\operatorname{Tr}\left(\mathfrak{C}_{\text{ania}} H^{-1}\right)$ for various compression schemes. Finally, in Subsection 4.3.4, we validate our findings with numerical experiments.

4.3.1 Compression operators

Our analysis applies to most unbiased compression operators.

Definition 4.4 (Compression operators)**.** *Let* $z \in \mathbb{R}^d$.

1. ***1-quantization*** *is defined as* $\mathcal{C}_{\text{q}}(z) := \|z\|\text{sign}(z) \odot \chi$ *with* $\chi \sim \otimes_{i=1}^d (\text{Bern}(|z_i|/\|z\|_2))$.

[3] This letter Ш is the Russian upper letter "sha".

2. ***Stabilized* 1*-quantization*** *is defined as* $\mathcal{C}_{\mathrm{sq}}(z) := U^\top \mathcal{C}_q(Uz)$*, with* $U \in \mathrm{Unif}(\mathcal{O}_d)$.
3. ***Rand-h*** *is defined as* $\mathcal{C}_{\mathrm{rd}h}(z) := \frac{d}{h} B(S) \odot z$ *with* $S \sim \mathrm{Unif}(\mathcal{P}_h([d]))$ *and* $B(S)_i = \mathbb{1}_{i \in S}$.
4. ***Sparsification*** *is defined as* $\mathcal{C}_{\mathrm{s}}(z) := \frac{1}{p} B \odot z \in \mathbb{R}^d$ *with* $B \sim \otimes_{i=1}^d (\mathrm{Bern}(p))$.
5. ***Partial participation*** *is defined* $\mathcal{C}_{\mathrm{PP}}(z) := \frac{b_0}{p} z$ *with* $b_0 \sim \mathrm{Bern}(p)$.
6. ***Random Projection***, *also referred to as* sketching, *is defined as* $\mathcal{C}_\Phi(z) := \frac{1}{p} \Phi^\dagger \Phi z$, *where* $h \ll d \in \mathbb{N}$, $p = h/d$ *and* $\Phi \in \mathbb{R}^{h \times d}$ *is a random projection matrix onto a lower-dimension space [Vempala, 2005, Li et al., 2006]. In the following, we consider Gaussian projection, where each element* $i, j \in [\![1, h]\!] \times [\![1, d]\!]$ *follows an independent zero-centered normal distribution.*

We refer to the introduction for related work on compression. Operators $\mathcal{C}_{\mathrm{q}}, \mathcal{C}_{\mathrm{sq}}$ are quantization-based schemes while $\mathcal{C}_{\mathrm{rd1}}, \mathcal{C}_{\mathrm{s}}, \mathcal{C}_{\mathrm{PP}}, \mathcal{C}_\Phi$ are projection-based. Indeed sparsification can be seen as a random projection (for $h \ll d$, $p = h/d$ and h randomly sampled coordinates $\mathcal{I}$ from $[\![1, d]\!]$ such that for any $i \in \mathcal{I}$, the i^{th} lines of Φ are equal to $e_i \in \mathbb{R}^d$, and equal to zero otherwise). For $\mathcal{C}_{\mathrm{PP}}$, the motivation is distributed settings, in which the intermittent availability of clients prevents them from systematically participating in the training. This can be modeled through *partial participation*: clients only participate in a fraction p of the training steps. In theoretical analyses, this can be handled as a compression scheme $\mathcal{C}_{\mathrm{PP}}$, in which the compression of a vector z is either z/p or 0. Observe that in the centralized case, this is slightly artificial as it actually means that no update is performed at most steps and that the step-size is scaled at the other steps. Finally, we denote $\mathcal{C}_{\mathrm{I}_d} : z \in \mathbb{R}^d \mapsto z$ the operator that does not carry out any compression.

Remark 4.6. *The analysis of random projection is related to Random features [Rahimi and Recht, 2008], usually used for Kernel learning in infinite dimensions. Nyström method [introduced by Kumar et al., 2009] is another similar technique of compression often used in this setting, it consists of removing a subset* $\mathcal{S} \subset \{1, \cdots, d\}$ *of lines and columns in the kernel matrix* K*. Both techniques have been extensively studied in the context of linear and non-linear kernel learning [Rudi et al., 2015, 2017, Rudi and Rosasco, 2017, Lin and Rosasco, 2017]. Recently, the combination of SGD and random features has been analyzed by Carratino et al. [2018]. However, their results cannot be directly applied to our setting for two reasons. Firstly, their analysis is for infinite dimensions, where they obtain a* $O(1/\sqrt{K})$ *rate of convergence. Secondly, the compressions used in their approach are not independent at each iteration.*

Remark 4.7. *Diffusion LMS (i.e. distributed learning without a central server) has also been studied from the perspective of low-cost training by Arablouei et al. [2015], Harrane et al. [2018], but using only clients' partial participation or sparsification. Contrary to our work they use biased compression and an adaptive correction step to compensate for the induced error. They provide results guarantying asymptotic convergence [Harrane et al., 2018, see Equations (28)-(37)].*

4.3.2 Applicability of the results on (LSA) from Section 4.2

We first show that our results from Section 4.2 can be applied for Algorithm 2 with a random compression operator $\mathcal{C}$, in the case of Model 2.

Lemma 4.1. *For any compressor* $\mathcal{C} \in \{\mathcal{C}_{\mathrm{q}}, \mathcal{C}_{\mathrm{sq}}, \mathcal{C}_{\mathrm{rd}h}, \mathcal{C}_{\mathrm{s}}, \mathcal{C}_\Phi, \mathcal{C}_{\mathrm{PP}}\}$*, there exists constants* $\omega, \Omega \in \mathbb{R}_+^*$*, such that the random operator* $\mathcal{C}$ *satisfies the following properties for all* $z, z' \in \mathbb{R}^d$.

L.1: $\mathbb{E}[\mathcal{C}(z)] = z$ *and* $\mathbb{E}[\|\mathcal{C}(z) - z\|^2] \leq \omega \|z\|^2$ *(unbiasedness and variance relatively bounded),*
L.2: $\mathbb{E}[\|\mathcal{C}(z) - \mathcal{C}(z')\|^2] \leq \Omega \min(\|z\|, \|z'\|)\|z - z'\| + 3(\omega + 1)\|z - z'\|^2$ *(Hölder-type bound),*

with $\omega = \sqrt{d}$ *and* $\Omega = 12\sqrt{d}$ *(resp.* $\omega = (1-p)/p$ *and* $\Omega = 0$*) for* $\mathcal{C}_{\mathrm{q}}$ *and* $\mathcal{C}_{\mathrm{sq}}$ *(resp.* $\mathcal{C}_{\mathrm{rd}h}, \mathcal{C}_{\mathrm{s}}$, $\mathcal{C}_\Phi, \mathcal{C}_{\mathrm{PP}}$*).*

We note $\mathbb{C}$ the set of unbiased compressors verifying Lemma 4.1. Item L.1 is frequently established in the literature and corresponds to the worst-case assumption, see the introduction for references. On the other hand, Item L.2 is the Hölder-type bound, which is not used in the literature up to our knowledge. The expected squared distance between the compression of two nearby points scales with the *non-squared* norm of the distance. Moreover, the distance is multiplied by an unavoidable coefficient scaling with z, z'. Remark that in Item L.2, we assume the compression randomness to be the same for the compression of z and z': formally, we control $\mathcal{W}_2(\mathcal{C}(z), \mathcal{C}(z'))^2$, with $\mathcal{W}_2$ the Wassertein-2 distance. This lemma is demonstrated in Subsection D.5.1.

Remark 4.8. *For a given ω, note that the* communication cost *c for quantization-based and projection-based compressors is not always equivalent. For 1-quantization we have $c \approx \frac{3}{2}\sqrt{d}\log_2 d + 32$ while for projection-based we have $c \approx 32\sqrt{d}$, for $\sqrt{d}$-quantization we have $c \approx 3d + 32$ while for projection-based we have $c = 16d$.*

Lemma 4.1 enables to show that Theorems 4.1 and 4.2 Algorithm 2 are valid in the context of Model 2.

Corollary 4.2. *Consider Algorithm 2 in the context of Model 2, with a compressor $\mathcal{C} \in \{\mathcal{C}_{\mathrm{q}}, \mathcal{C}_{\mathrm{sq}}, \mathcal{C}_{\mathrm{rd}h}, \mathcal{C}_{\mathrm{s}}, \mathcal{C}_{\Phi}, \mathcal{C}_{\mathrm{PP}}\}$. With Lemma 4.1 above, Assumptions 4.1 and 4.2 on the resulting random field $(\xi_k)_{k\in\mathbb{N}^*}$ are valid, with in particular $H_F = H$, $R_F^2 = R^2$, $\mathcal{A} = (\omega+1)R^2\sigma^2$, $\mathcal{M}_2 = (\omega+1)R^2$, $\mathcal{M}_1 = \Omega R^2\sigma$. Therefore, it follows that Theorem 4.1 holds.*

Moroever for any linear compressor $\mathcal{C} \in \{\mathcal{C}_{\mathrm{rd}h}, \mathcal{C}_{\mathrm{s}}, \mathcal{C}_{\Phi}, \mathcal{C}_{\mathrm{PP}}\}$, under Remark 4.1, we also have that Assumptions 4.3 and 4.4 are valid with $\amalg_{\mathrm{add}} = \sigma^2 \amalg_H$ and $\amalg_{\mathrm{mult}} = R^2 \amalg_H$, with $\amalg_H$ given below. Therefore, it follows that Theorem 4.2 holds.

Compressor	$\mathcal{C}_{\mathrm{rd}h}$	$\mathcal{C}_{\mathrm{s}}$	$\mathcal{C}_{\mathrm{PP}}$	$\mathcal{C}_{\Phi}$
$\amalg_H$	$\frac{h-1}{p(d-1)} + (1 - \frac{h-1}{d-1})\frac{\tau}{p}$	$1 + \frac{(1-p)\tau}{p}$	$\frac{1}{p}$	$\frac{\alpha-\beta}{p} + \frac{\beta\tau}{p}$
$\amalg_H$ *(if H diagonal)*	$\frac{1}{p}$	$\frac{1}{p}$	$\frac{1}{p}$	$\frac{\alpha-\beta}{p} + \frac{\beta\tau}{p}$

Where $p = h/d$, $\tau = \mathrm{Tr}\,(H)/\mu$, and for sketching $\alpha = \frac{h+2}{d+2}$ and $\beta = \frac{d-h}{(d-1)(d+2)}$.

This corollary is proved in Section D.4. We observe that a first difference in terms of convergence exists between quantization-based compression and projection-based: for the former, *only* Theorem 4.1 can be applied and the lower-order terms always have a *poorer dependency on μ* while for the latter, Theorem 4.2 is applicable and lower-order terms do not necessarily depends on μ. Indeed, the constants $\amalg_H$ do not depend on μ for $\mathcal{C}_{\mathrm{PP}}$, and for $\mathcal{C}_{\mathrm{rd1}}, \mathcal{C}_{\mathrm{s}}$, *when* the features' covariance H is diagonal. On the contrary, there is always a dependency on μ for $\mathcal{C}_{\Phi}$, and for $\mathcal{C}_{\mathrm{rd1}}, \mathcal{C}_{\mathrm{s}}$ when H is not diagonal. In practice, this means that, among projection-based compressors, regarding lower-order terms, the convergence is expected to be slower for random Gaussian projection.

We now turn to the analysis of the impact of the choice of the compression on the dominant asymptotic term $\mathrm{Tr}\left(H_F^{-1}\mathfrak{C}_{\mathrm{ania}}\right)$.

4.3.3 Impact of the compression on the additive noise covariance

In this section, we illustrate how distinct compressors lead to different covariances for the additive noise. This shows how $\mathrm{Tr}\left(H_F^{-1}\mathfrak{C}_{\mathrm{ania}}\right)$ is impacted by the choice of a compressor.

First recall that for Algorithm 2 in the context of Model 2, with any compressor $\mathcal{C}$, the additive noise writes for any $k \in \{1, \ldots, K\}$, as:

$$\xi_k^{\mathrm{add}} \overset{\text{def. } 4.2}{=} \xi_k(0) \overset{\text{algo } 2}{=} \nabla F(w_*) - \mathcal{C}_k(g_k(w_*)) \overset{\text{eq. } 4.2}{=} -\mathcal{C}_k((\langle x_k, w_*\rangle - y_k)x_k) \overset{\text{model } 2}{=} \mathcal{C}_k(\varepsilon_k x_k)\,.$$

Also recall that $\mathfrak{C}_{\text{ania}}$ is defined as $\mathfrak{C}_{\text{ania}} := \mathbb{E}[(\xi_k^{\text{add}})^{\otimes 2}] = \mathbb{E}[\mathcal{C}(\varepsilon_k x_k)^{\otimes 2}]$. Moreover, note that $\mathcal{C}(\varepsilon_k x_k) \overset{\text{a.s.}}{=} \varepsilon_k \mathcal{C}(x_k)$ for all operators under consideration (this is immediate for linear operators and results from the scaling for quantization-based ones). Consequently

$$\mathfrak{C}_{\text{ania}} = \mathbb{E}[\varepsilon_k^2 \mathcal{C}(x_k)^{\otimes 2}] = \sigma^2 \mathbb{E}[\mathcal{C}(x_k)^{\otimes 2}], \tag{4.3}$$

as $\mathbb{E}[\varepsilon_k^2 | x_k] = \sigma^2$. Ultimately, we have to study the covariance of $\mathcal{C}(x_k)$, for x_k a random variable with second-moment H.

We thus generically study the covariance of $\mathcal{C}(E)$, for E a random vector with distribution p_M with second moment[4] $\mathbb{E}[E^{\otimes 2}] = M$.

Definition 4.5 (Compressor' covariance on p_M)**.** *We define the following operator $\mathfrak{C}$ which returns* the covariance of a random mechanism $\mathcal{C}$ *acting on a distribution $p_M \in \mathcal{P}_M$,*

$$\mathfrak{C} : \begin{array}{ccc} \mathbb{C} \times \mathcal{P}_M & \to & \mathbb{R}^{d \times d} \\ (\mathcal{C}, \quad p_M) & \mapsto & \mathbb{E}[\mathcal{C}(E)^{\otimes 2}] \,, \end{array}$$

where $E \sim p_M$ and the expectation is over the joint randomness of $\mathcal{C}$ and E, which are considered independent, that is $\mathbb{E}[\mathcal{C}(E)^{\otimes 2}] = \int_{\mathbb{R}^d} \mathbb{E}[\mathcal{C}(e)^{\otimes 2}] \mathrm{d}p_M(e)$.

Using a compressor $\mathcal{C} \in \mathbb{C}$, we therefore have by Equation (4.3):

$$\mathfrak{C}_{\text{ania}} = \sigma^2 \mathfrak{C}(\mathcal{C}, p_H), \tag{4.4}$$

where p_H is the marginal distribution of x_k (for any k).

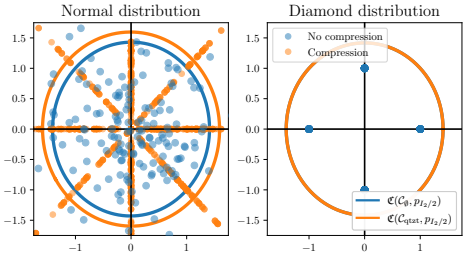


Figure 4.2: Illustration of Remark 4.9

Remark 4.9 (Dependence on p_M, not only M)**.** *Note that, for $\mathcal{C} = \mathcal{C}_{\text{q}}$, there exist two distributions p_M, p'_M with the same covariance M, such that $\mathfrak{C}(\mathcal{C}, p_M) \neq \mathfrak{C}(\mathcal{C}, p'_M)$. This is why we cannot simply denote $\mathfrak{C}(\mathcal{C}, M)$.*

Indeed, consider $d = 2$ and (1) a normal distribution $E_1 \sim \mathcal{N}(0, I_2/2)$, vs (2) a *diamond* distribution $E_2 \sim \mathbb{P}_\diamond$, such that $\mathbb{P}_\diamond\{(1,0)\} = \mathbb{P}_\diamond\{(-1,0)\} = \mathbb{P}_\diamond\{(0,1)\} = \mathbb{P}_\diamond\{(0,-1)\} = 1/4$, and thus $\mathrm{Cov}\,[E_1] = \mathrm{Cov}\,[E_2] = I_2/2$. Then $\mathrm{Cov}\,[E_1] \prec \mathrm{Cov}\,[\mathcal{C}_{\text{q}}(E_1)]$, but $\mathcal{C}_{\text{q}}(E_2) \overset{\text{a.s.}}{=} E_2$ thus $\mathrm{Cov}\,[E_2] = \mathrm{Cov}\,[\mathcal{C}_{\text{q}}(E_2)]$. We illustrate this on Figure 4.2: we represent E_i in blue and $\mathcal{C}_{\text{q}}(E_i)$ in orange for $i = 1$ (left) and $i = 2$ (right). We also represent the covariance matrices by plotting the ellipses $\mathcal{E}_{\mathrm{Cov}[E_i]}$ and $\mathcal{E}_{\mathrm{Cov}[\mathcal{C}_{\text{q}}(E_i)]}$, where $\mathcal{E}_M = \{x \in \mathbb{R}^d, x^\top M^{-1} x = 4\}$ (see Definition A.1)[5].

We now compute for the compression operators, the value or an upper bound on $\mathfrak{C}(\mathcal{C}, p_H)$.

Proposition 4.2 (Compression and covariance)**.** *The following formulas hold:*

$$\begin{aligned}
\mathfrak{C}(\mathcal{C}_{\mathrm{I}_d}, p_M) &= M \\
\mathfrak{C}(\mathcal{C}_q, p_M) &\preccurlyeq \widetilde{\mathfrak{C}}(\mathcal{C}_{\text{q}}, M) := M + \sqrt{\mathrm{Tr}\,(M)} \sqrt{\mathrm{Diag}\,(M)} - \mathrm{Diag}\,(M) \\
&\quad \textit{(with equality if } \|E\| \textit{ is a.s. constant under } p_M\textit{)} \\
\mathfrak{C}(\mathcal{C}_s, p_M) &= M + (1-p) p^{-1} \mathrm{Diag}\,(M) \\
\mathfrak{C}(\mathcal{C}_\Phi, p_M) &= p^{-1} \left((\tfrac{h+1}{d+2} + \delta_{hd}) M + \left(1 - \tfrac{h-1}{d-1}\right) \tfrac{\mathrm{Tr}(M)}{d+2} \mathrm{I}_d \right), \quad \textit{with } \delta_{hd} = \tfrac{h-1}{(d-1)(d+2)} = O(\tfrac{1}{d}) \\
\mathfrak{C}(\mathcal{C}_{\mathrm{rd}h}, p_M) &= p^{-1} \left(\tfrac{h-1}{d-1} M + \left(1 - \tfrac{h-1}{d-1}\right) \mathrm{Diag}\,(M) \right) \\
\mathfrak{C}(\mathcal{C}_{\mathrm{PP}}, p_M) &= p^{-1} M \,.
\end{aligned}$$

[4] Remark that we do not assume $\mathbb{E}[E] = 0$. Indeed, all computations only depend on the *second-order moment* M of E, not on its variance (and the convergence depends of the second-*order moment* H of x, not its variance). It is clear, that $\mathbb{E}[\mathcal{C}(E)^{\otimes 2}]$ does not depend on the fact that E is centered: indeed, for R a Rademacher $1/2$ independent of E, we have $\mathbb{E}[\mathcal{C}(E)^{\otimes 2}] = \mathbb{E}[R^2]\mathbb{E}[\mathcal{C}(E)^{\otimes 2}] \overset{+}{=} \mathbb{E}[(R\mathcal{C}(E))^{\otimes 2}] = \mathbb{E}[\mathcal{C}(RE)^{\otimes 2}]$ and RE is (1) centered (2) has the same second-moment as E. Remark that centering the covariates before learning does impact H: indeed $H = \mathbb{E}[(x)^{\otimes 2}] = \mathbb{E}[(x - \mathbb{E}[X])^{\otimes 2}] + (\mathbb{E}[X])^{\otimes 2})$. Centering subtracts $(\mathbb{E}[X])^{\otimes 2}$ to the second moment, which is a rank-1 matrix, typically does not affect the smallest eigenvalue, but it can affect the top-eigenvalue.

[5] The constant 4 is chosen so that for Gaussian distributions, the expected fraction of points within the ellipse is $86,4\% \simeq 1 - F_{\chi^2(2)}(4)$

Conclusion and interpretation. Most compression operators induce *both* a *structured* noise [Flammarion and Bach, 2015] which covariance scales with H and an *unstructured* noise, which covariance scales with $\mathrm{Diag}\,(H)$ or I_d – thus corresponding to an *isotropic* noise.

From the convergence standpoint, the asymptotic convergence rate scales with $\mathrm{Tr}\left(\mathfrak{C}_{\mathrm{ania}}H^{-1}\right) = \sigma^2\mathrm{Tr}\left(\mathfrak{C}(\mathcal{C}, p_H)H^{-1}\right)$. Therefore, the un-structured part in the noise is problematic as $\mathrm{Tr}\left(\mathfrak{C}_{\mathrm{ania}}H^{-1}\right)$ will strongly depends on the smallest eigenvalue μ. This comes from the fact that the compression induces a significant noise in directions in which the Hessian curvature is very limited (thus directions onto which the contraction towards the optimum in the algorithm is weak).

A particular case is when H is diagonal (e.g. the features are *centered* and *independent*), we get the following corollary.

Corollary 4.3 (Compression and covariance, diagonal case)**.** *If M is diagonal, then Proposition 4.2 is simplified to the following (with the same δ_{hd}):*

$$\begin{array}{lll} \mathfrak{C}(\mathcal{C}_{\mathrm{I}_d}, p_M) & = M & \mathfrak{C}(\mathcal{C}_\Phi, p_M) = p^{-1}\left((\frac{h+1}{d+2}+\delta_{hd})M + (1-\frac{h-1}{d-1})\frac{\mathrm{Tr}(M)}{d+2}\mathrm{I}_d\right) \\ \mathfrak{C}(\mathcal{C}_q, p_M) & \preccurlyeq \sqrt{\mathrm{Tr}\,(M)}\sqrt{M} & \mathfrak{C}(\mathcal{C}_{\mathrm{rd}h}, p_H) = p^{-1}M \\ \mathfrak{C}(\mathcal{C}_s, p_M) & = p^{-1}M & \mathfrak{C}(\mathcal{C}_{\mathrm{PP}}, p_M) = p^{-1}M. \end{array}$$

Remark 4.10 (Composition of compressors)**.** *For all compression schemes but $\mathcal{C}_{\mathrm{q}}$, we observe that $\mathfrak{C}(\mathcal{C}, p_M)$ is a function of M, which complements Remark 4.9. In that particular case, we can then denote $\mathfrak{C}(C, M)$. This means that the lemma can be extended to any composition of compression schemes, for example to compute $\mathfrak{C}(C_1 \circ C_2, M) = \mathfrak{C}(C_1, \mathfrak{C}(C_2, M))$.*

From Proposition 4.2 and Corollary 4.3 we can deduce certain generic comparisons between the asymptotic convergence rates, depending on the compression operator (for compression operators having the same variance bound). They are proven in Subsection D.5.3. In the following, for any $a, b \in \mathbb{R}$, we use the notation $a \overset{\sim}{\leq} b$, to denote a *systematic inequality* (i.e., $a \leq b$) with a negligible difference as $d \to \infty$ (i.e., $a = b + O(1/d)$), and similarly for any two symmetric matrices $A, B \in \mathcal{S}_d(\mathbb{R})$, $A \overset{\sim}{\preccurlyeq} B$, for $A \preccurlyeq B$ and $A = B + O(1/d)$ as $d \to \infty$.

Proposition 4.3 (Comparison between $\mathcal{C}_{\mathrm{PP}}, \mathcal{C}_{\mathrm{s}}, \mathcal{C}_{\mathrm{rd}h}, \mathcal{C}_\Phi$, $\omega = d/h - 1$)**.** *We consider $\mathcal{C} \in \{\mathcal{C}_{\mathrm{PP}}, \mathcal{C}_{\mathrm{s}}, \mathcal{C}_{\mathrm{rd}h}, \mathcal{C}_\Phi\}$ with $p = h/d$, such that $\mathcal{C}$ always satisfies Lemma 4.1 with $\omega = d/h - 1$. For any matrix $M \in \mathbb{R}^{d\times d}$:*

1. *If M is diagonal, then:*
   - $\mathfrak{C}(\mathcal{C}_{\mathrm{PP}}, p_M) = \mathfrak{C}(\mathcal{C}_{\mathrm{s}}, p_M) = \mathfrak{C}(\mathcal{C}_{\mathrm{rd}h}, p_M) = \frac{d}{h}M$,
   - $\mathrm{Tr}\big(\mathfrak{C}(\mathcal{C}_{\mathrm{PP/s/rd}h}, p_M)M^{-1}\big) \leq \mathrm{Tr}\left(\mathfrak{C}(\mathcal{C}_\Phi, p_M)M^{-1}\right)$.

   This means that the asymptotic convergence rate does not depend on the choice of the compressor between $\mathcal{C}_{\mathrm{PP}}, \mathcal{C}_{\mathrm{s}}, \mathcal{C}_{\mathrm{rd1}}$ in the diagonal case.
2. *Moreover, for any matrix M with a* constant diagonal *(e.g., we standardize[6] the data in the pre-processing step, such that $\mathrm{Diag}\,(M) = \mathrm{I}_d$), we have:*

$$\mathrm{Tr}(\mathfrak{C}(\mathcal{C}_{\mathrm{PP}}, p_M)M^{-1}) \leq \mathrm{Tr}(\mathfrak{C}(\mathcal{C}_\Phi, p_M)M^{-1}) \leq \mathrm{Tr}(\mathfrak{C}(\mathcal{C}_s, p_M)M^{-1}) \leq \mathrm{Tr}(\mathfrak{C}(\mathcal{C}_{\mathrm{rd}h}, p_M)M^{-1})\,.$$

   With strict inequalities if M is not proportional to I_d. This means that we expect the asymptotic convergence rate to be faster for PP than Sparsification, Sketching, or Rand-h (illustrated in experiments).

In the next proposition, we compare compressors $\mathcal{C}_{\mathrm{s}}, \mathcal{C}_{\mathrm{PP}}$ to $\mathcal{C}_{\mathrm{q}}$ for equal $\omega = \sqrt{d}$ (we exclude $\mathcal{C}_{\mathrm{rd}h}$ and $\mathcal{C}_\Phi$ for which h must be an integer and that are shown to be worse than $\mathcal{C}_{\mathrm{s}}$ in Proposition 4.3).

[6] That means we center and rescale to get a variance of one for each feature.

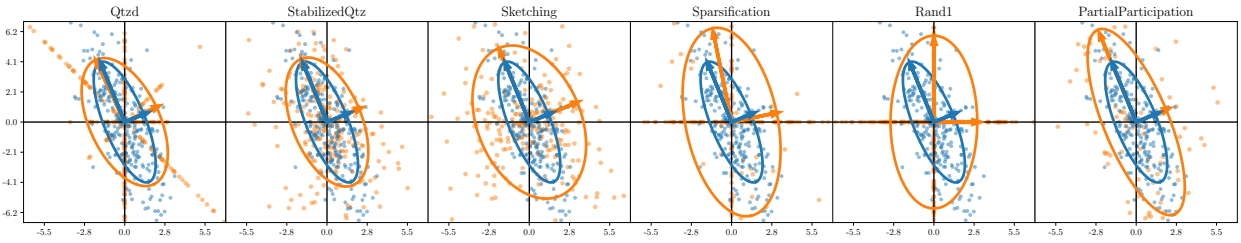


Figure 4.3: H not diagonal. Scatter plot of $(x_k)_{i=1}^K$/ $(\mathcal{C}(x_k))_{i=1}^K$ with its ellipse $\mathcal{E}_{\mathrm{Cov}[x_k]}/\mathcal{E}_{\mathrm{Cov}[\mathcal{C}(x_k)]}$.

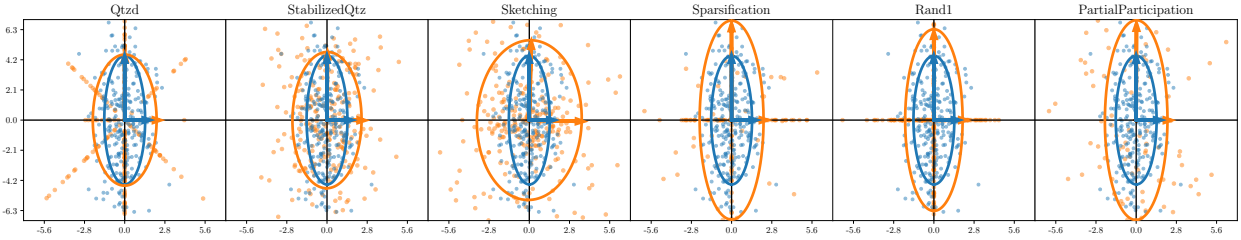


Figure 4.4: H diagonal. Scatter plot of $(x_k)_{i=1}^K/(\mathcal{C}(x_k))_{i=1}^K$ with its ellipse $\mathcal{E}_{\mathrm{Cov}[x_k]}/\mathcal{E}_{\mathrm{Cov}[\mathcal{C}(x_k)]}$.

Proposition 4.4 (Comparison between $\mathcal{C}_{\mathrm{PP}}, \mathcal{C}_{\mathrm{q}}, \mathcal{C}_{\mathrm{s}}$, $\omega = \sqrt{d}$). *We consider $\mathcal{C} \in \{\mathcal{C}_{\mathrm{PP}}, \mathcal{C}_{\mathrm{q}}, \mathcal{C}_{\mathrm{s}}\}$ with $p = (\sqrt{d}+1)^{-1}$, such that $\mathcal{C}$ always satisfies Lemma 4.1 with $\omega = \sqrt{d}$.*

1. *For any symmetric matrix M diagonal, we have:*

$$\mathrm{Tr}\left(\mathfrak{C}(\mathcal{C}_{\mathrm{PP}}, p_M)M^{-1}\right) = \mathrm{Tr}\left(\mathfrak{C}(\mathcal{C}_s, p_M)M^{-1}\right) \overset{\text{possib.}}{\leq} \ll \left(1 + \frac{1}{\sqrt{d}}\right) \mathrm{Tr}\left(\widetilde{\mathfrak{C}}(\mathcal{C}_{\mathrm{q}}, M)M^{-1}\right).$$

2. *If M is not necessarily diagonal but with a* constant diagonal *(e.g., after standardization), then*
   - $\widetilde{\mathfrak{C}}(\mathcal{C}_{\mathrm{q}}, M) \preccurlyeq \mathfrak{C}(\mathcal{C}_{\mathrm{s}}, p_M)$
   - $\mathrm{Tr}\left(\mathfrak{C}(\mathcal{C}_{\mathrm{PP}}, p_M)M^{-1}\right) \leq \left(1 + \frac{1}{\sqrt{d}}\right) \mathrm{Tr}\left(\widetilde{\mathfrak{C}}(\mathcal{C}_{\mathrm{q}}, M)M^{-1}\right)$

This means that sparsification is expected to always result in a poorer asymptotic convergence rate than quantization. Moreover, the upper bound on the covariance $\widetilde{\mathfrak{C}}(\mathcal{C}_{\mathrm{q}}, M)$ for quantization itself leads to a worst bound than for PP.[7]

We now propose a detailed illustration of the results of Proposition 4.2 and Corollary 4.3, first in a low-dimensional setting ($d = 2$) and then in higher dimension on synthetic and real datasets.

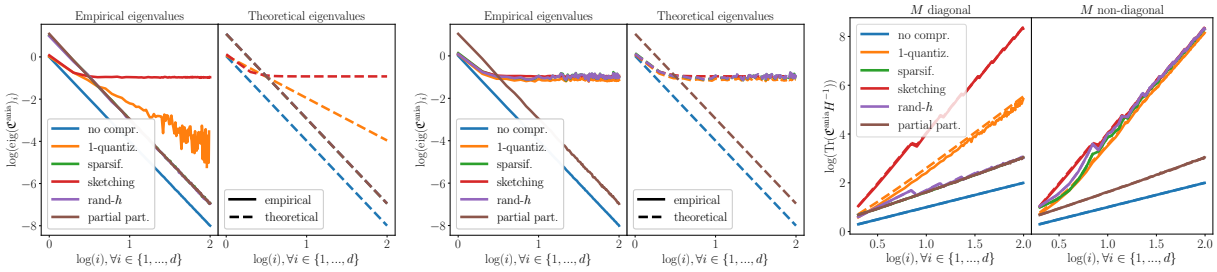
4.3.3.1 Illustration of Proposition 4.2 and Corollary 4.3 in dimension 2.

In order to build intuition, we illustrate Proposition 4.2 and Corollary 4.3 in Figures 4.3 and 4.4, showing how compression affects the additive noise covariance, in a simple 2-dimensional case, for both a non-diagonal matrix M (Figure 4.3) and a diagonal one (Figure 4.4).

More specifically, we consider features $(x_k)_{k\in\{1,\dots,K\}}$ sampled from $\mathcal{N}(0, M)$ where $M = QDQ$, $D = \mathrm{Diag}\,(1, 10)$ and Q is rotation matrix with angle $\pi/8$ (resp. 0) in Figure 4.3 (resp. 4.4). We represent the values of x_k and $\mathcal{C}(x_k)$, unit-ellipses of the corresponding covariance matrices $\mathcal{E}_{\mathrm{Cov}[x_k]}$ and $\mathcal{E}_{\mathrm{Cov}[\mathcal{C}(x_k)]}$ (see Definition A.1 – recall that $\mathcal{E}_{\mathrm{Cov}[x_k]} \subset \mathcal{E}_{\mathrm{Cov}[\mathcal{C}(x_k)]} \Leftrightarrow \mathrm{Cov}\,[x_k] \preccurlyeq \mathrm{Cov}\,[\mathcal{C}(x_k)]$), as well as their two eigenvectors; we take $p = (1 + \sqrt{d})^{-1} = 0.41$, hence for $\mathcal{C} \in \{\mathcal{C}_{\mathrm{q}}, \mathcal{C}_{\mathrm{sq}}, \mathcal{C}_{\mathrm{s}}, \mathcal{C}_{\mathrm{PP}}\}$ we have $\omega = 1.41$ but for sketching and rand-1, we have $p = 1/2$ and $\omega = (1-p)/p = 1$.

We make the following observations:

[7] Note that the behavior for quantization, apart from the upper bound $\widetilde{\mathfrak{C}}(\mathcal{C}_{\mathrm{q}}, M)$ is not quantified, it is thus possible that quantization performs even better than PP.



(a) M diagonal, $d = 100$ (b) M non-diagonal, $d = 100$ (c) $\mathrm{Tr}\left(\mathfrak{C}_{\text{ania}} M^{-1}\right)$, $d \in [\![2, 100]\!]$

Figure 4.5: Figures 4.5a & 4.5b: Eigenvalues of $\mathfrak{C}(\mathcal{C}, p_M)$. Figure 4.5c: $\mathrm{Tr}\left(\mathfrak{C}(\mathcal{C}, p_M) M^{-1}\right)$. $K = 10^4, \omega = 10$, $M = Q\mathrm{Diag}\left((1/i^4)_{i=1}^d\right) Q^T$ and $Q = \mathrm{I}_d$ (on 4.5a & 4.5c-l) or $Q \sim \mathrm{Unif}(\mathcal{O}_d)$ (on 4.5b & 4.5c-r). Plain lines: empirical values; dashed lines: theoretical formula or upper bound given by Proposition 4.2.

[Qtz] For quantization and stabilized quantization, in the non-diagonal case, the eigenvectors of $\mathcal{E}_{\mathrm{Cov}[x_k]}$ and $\mathcal{E}_{\mathrm{Cov}[\mathcal{C}(x_k)]}$ are slightly[8] different (as $\sqrt{\mathrm{Diag}\,(M)}$ and M are not jointly diagonalizable, as well as if $\mathrm{Diag}\,(M)$ is constant, although this case is not presented here, but in Figure D.5 in Subsection D.5.3). They are equal for the diagonal case (as $\sqrt{\mathrm{Diag}\,(M)}$ and M are both diagonal so the eigenvectors are aligned with the axis). In both cases, the eigenvalue decay is reduced (from $\lambda_2/\lambda_1 = 1/10$ without compression to $1/\sqrt{10}$ with compression, which visually corresponds to a "wider" ellipse).
This slower eigenvalue decay results from the *unstructured-noise*, i.e., large noise on the weak-curvature direction, which is particularly visible on Figure 4.4. This is critical as it results in a potentially much larger limit rate, as $\mathrm{Tr}\left(\mathfrak{C}(\mathcal{C}_q, p_M) M^{-1}\right) \simeq \mathrm{Tr}\left(M^{-1/2}\right)$.

[Skt] For sketching, the eigenvectors remain the same for $\mathcal{E}_{\mathrm{Cov}[x_k]}$ and $\mathcal{E}_{\mathrm{Cov}[\mathcal{C}(x_k)]}$ (as I_2 and M are jointly diagonalizable, see Corollary 4.3), both in the diagonal and non-diagonal case. However, the isotropic noise with covariance I_2 is visible (wide ellipse), also drastically impacting $\mathrm{Tr}\left(\mathfrak{C}(\mathcal{C}_{\mathrm{PP}}, p_M) M^{-1}\right) \propto \mathrm{Tr}\left(M^{-1}\right)$.

[Sp] For p-sparsification, eigenvectors are not aligned with the ones of M in the non-diagonal case, but are in the diagonal case. In this latter case, the covariance $\mathfrak{C}(\mathcal{C}_s, p_M)$ is proportional to M.

[Rd] Same remarks hold for Rand-1 than for sparsification. We see that $\mathfrak{C}(\mathcal{C}_{\mathrm{rd1}}, p_M)$ is diagonal, as expected. Again, both operators induce an unstructured-noise in the non-diagonal case.

[PP] For PP, the covariances are always proportional (with factor p^{-1}), i.e., the ellipses have the same axis and $\mathcal{E}_{\mathrm{Cov}[\mathcal{C}(x_k)]}$ is a scaled version of $\mathcal{E}_{\mathrm{Cov}[x_k]}$.

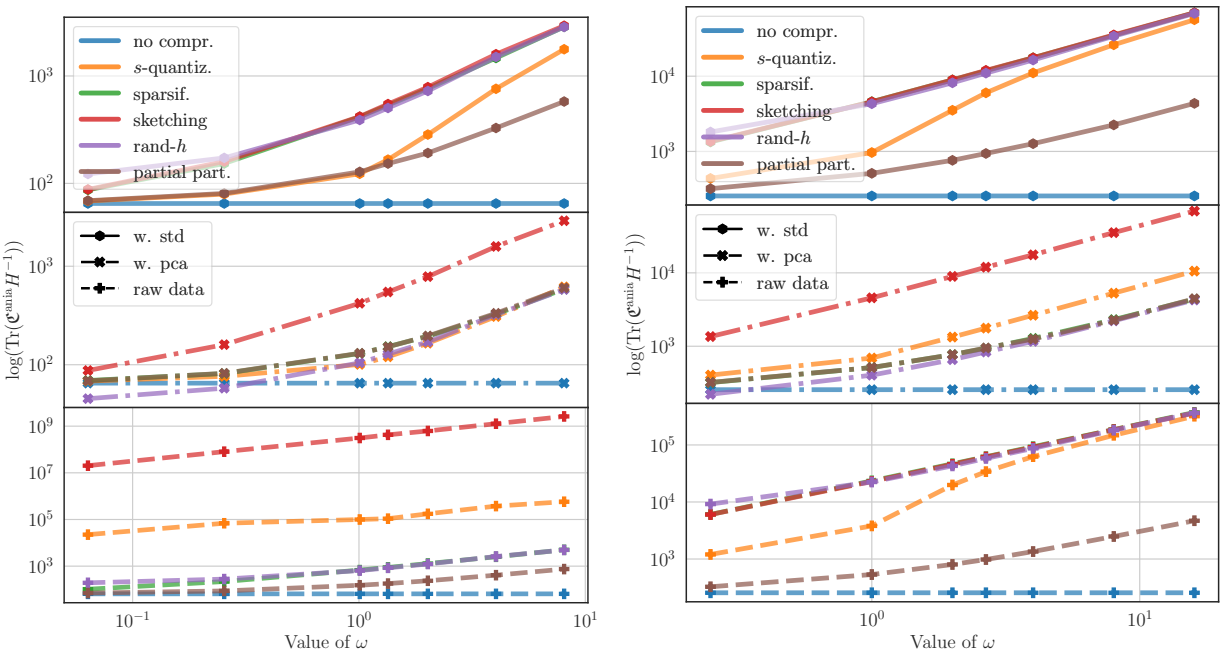
We highlight the following points regarding pairwise comparisons:

- In the diagonal case, as stated by Item 1 in Proposition 4.3, $\mathrm{Cov}\left[\mathcal{C}_{\mathrm{s}}(x_k)\right]$ and $\mathrm{Cov}\left[\mathcal{C}_{\mathrm{PP}}(x_k)\right]$ are identical. $\mathrm{Cov}\left[\mathcal{C}_{\mathrm{rd1}}(x_k)\right]$ would have been identical too if $p = 1/d$ (but here we observe $\mathfrak{C}(\mathcal{C}_{\mathrm{rd1}}, p_M) \preccurlyeq \mathfrak{C}(\mathcal{C}_{s/\mathrm{PP}}, p_M)$ because the variance of rand-1 is smaller that for sparsification/PP).
- In the non-diagonal case, from Item 2 in Proposition 4.3, we have $\mathrm{Tr}\left(\mathfrak{C}(\mathcal{C}_{\mathrm{PP}}, p_M) M^{-1}\right) \leq \mathrm{Tr}\left(\mathfrak{C}(\mathcal{C}_s, p_M) M^{-1}\right)$, however we do not have $\mathfrak{C}(\mathcal{C}_{\mathrm{PP}}, p_M) \preccurlyeq \mathfrak{C}(\mathcal{C}_s, p_M)$, hence we can not conclude anything on $\mathrm{Cov}\left[\mathcal{C}_{\mathrm{PP}}(x_k)\right]$ and $\mathrm{Cov}\left[\mathcal{C}_{\mathrm{s}}(x_k)\right]$.
- In the non-diagonal scenario, we observe on Figure 4.3, that $\mathfrak{C}(\mathcal{C}_q, p_M) \preccurlyeq \mathfrak{C}(\mathcal{C}_s, p_M)$ (as in Item 2 in Proposition 4.4).

4.3.3.2 Illustration of Proposition 4.2 and Corollary 4.3 in dimension $d > 2$

Another way of visualizing the structured and isotropic parts of the noise is by plotting the eigenvalues of $\mathfrak{C}(\mathcal{C}, p_M)$ in dimension $d = 100$. This is done in Figure 4.5, in which we plot the eigenvalues in

[8] On the figure, there are nearly aligned, but actually differ.



(a) quantum (tabular dataset)

(b) cifar10 (images)

Figure 4.6: $\mathrm{Tr}\left(\mathfrak{E}(\mathcal{C}, p_M)M^{-1}\right)$ w.r.t the level of ω for quantum and cifar10. X/Y-axis are in log scale. Note that the plots may have different magnitudes.

decreasing order for both M and $\mathfrak{E}(\mathcal{C}, p_M)$, with Gaussian $p_M = \mathcal{N}(0, M)$ and $\mathrm{Sp}(M) = \{(1/i^4)_{i=1}^d\}$. We see that in the diagonal case, in Figure 4.5a, all operators but $\mathcal{C}_\mathrm{q}, \mathcal{C}_\Phi$ have a covariance proportional to M (thus a slope -4 on a log/log scale), while $\mathcal{C}_\mathrm{q}$ is proportional to $\sqrt{M}$ (thus a slope -2) and $\mathcal{C}_\Phi$ has an isotropic component (thus eigenvalues not decreasing to 0). In Figure 4.5b we see that only $\mathcal{C}_\mathrm{PP}$ has a covariance proportional to M while all other ones have an isotropic component (thus eigenvalues not decreasing to 0). We plot both empirical values and the ones obtained in Proposition 4.2, which shows that the upper bound on quantization is reasonable in practice and acts as a safety check for other compression schemes.

We plot on Figure 4.5c the theoretical and empirical $\mathrm{Tr}\left(\mathfrak{E}(\mathcal{C}, p_M)M^{-1}\right)$ again in two cases, diagonal and non-diagonal. In the diagonal case, PP, sparsification, and rand-h have the same behavior; their traces have the smallest value among all compressors. However, in the general case of non-diagonal features' covariance, all compression operators have similar slow performance except for PP. For $d = 100$, all the compressors *have* $\omega = 10$, *but* $\mathrm{Tr}\left(\mathfrak{E}(\mathcal{C}, p_M)M^{-1}\right)$ *varies by several orders depending on the compressor*, illustrating again that compressors satisfying Lemma 4.1 with the same ω may have vastly different behaviors.

Lastly, we perform the same experiments on $\mathrm{Tr}\left(\mathfrak{E}(\mathcal{C}, p_M)M^{-1}\right)$, but on non-simulated datasets, namely `quantum` [Caruana et al., 2004] and `cifar-10` [Krizhevsky et al., 2009]: in Figure 4.6 we plot $\mathrm{Tr}\left(\mathfrak{E}(\mathcal{C}, p_M)M^{-1}\right)$ w.r.t. the worst-case-variance-level ω of the compression in three scenarios: **(top-row)** – with data standardization, thus $\mathrm{Diag}\,(M)$ is constant equal to 1; **(middle-row)** – with a PCA, thus with a diagonal covariance M (note that this is for illustration purpose: performing a PCA would be more expensive computationally than running Algorithm 2); and **(bottom-row)** – without any data transformation. As a pre-processing, we have resized images of the `cifar-10` dataset to a 16×16 dimension. We adjust the level $s \in \mathcal{C}_\mathrm{q}$, $h \in \mathcal{C}_{\mathrm{rd}h}, \mathcal{C}_\Phi$, and $p \in \mathcal{C}_\mathrm{PP}, \mathcal{C}_\mathrm{s}$ to make ω vary.

Interpretation. (Top-row): with standardization, the order predicted from Proposition 4.3.2 (large ω), and Proposition 4.4.2 (low ω) is obtained for both `quantum` and `cifar-10`: $\mathcal{C}_\mathrm{PP} \leq \mathcal{C}_\mathrm{q} \leq$

$\mathcal{C}_{\rm s} \simeq \mathcal{C}_{{\rm rd}h} \simeq \mathcal{C}_\Phi$. For quantization, we observe two regimes: 1) when ω tends to zero, quantization and PP outperform sketching, sparsification, and rand-h, that are equivalent. 2) when ω increases, quantization changes from scaling as PP to scaling as the second group. **(Middle-row):** in the diagonal regime, comments made for Figure 4.5c-l are still valid. **(Bottom-row):** We observe that for a generic matrix M (obtained from raw-data) there is no systematic order between compression schemes. This is un-avoidable as the order for a "M diagonal" and "M with constant-diagonal" is *not* the same. We observe that:

- for `quantum`, $\mathcal{C}_{\rm PP} \leq \mathcal{C}_{\rm s} \lesssim \mathcal{C}_{{\rm rd}h} \ll \mathcal{C}_{\rm q} \ll \mathcal{C}_\Phi$
- for `cifar-10`, $\mathcal{C}_{\rm PP} \ll \mathcal{C}_{\rm q} \ll \mathcal{C}_{\rm s} \simeq \mathcal{C}_{{\rm rd}h} \simeq \mathcal{C}_\Phi$.

We also observe that $\mathcal{C}_\Phi$, which is the only operator to always induce an isotropic component, may be much worse than all other compressors (e.g., on `quantum`). Ultimately, the order depends on the covariance matrix M. Here we observe that the raw-data behavior is close for `cifar-10` to the standardized version, while for `quantum` the order between compressors is the same for raw-data and diagonal (although the ratios are different). In Subsection D.5.4 (Table D.3), we provide an illustration of the covariance matrices, that supports such interpretation.

4.3.4 Numerical experiments on Algorithm 2

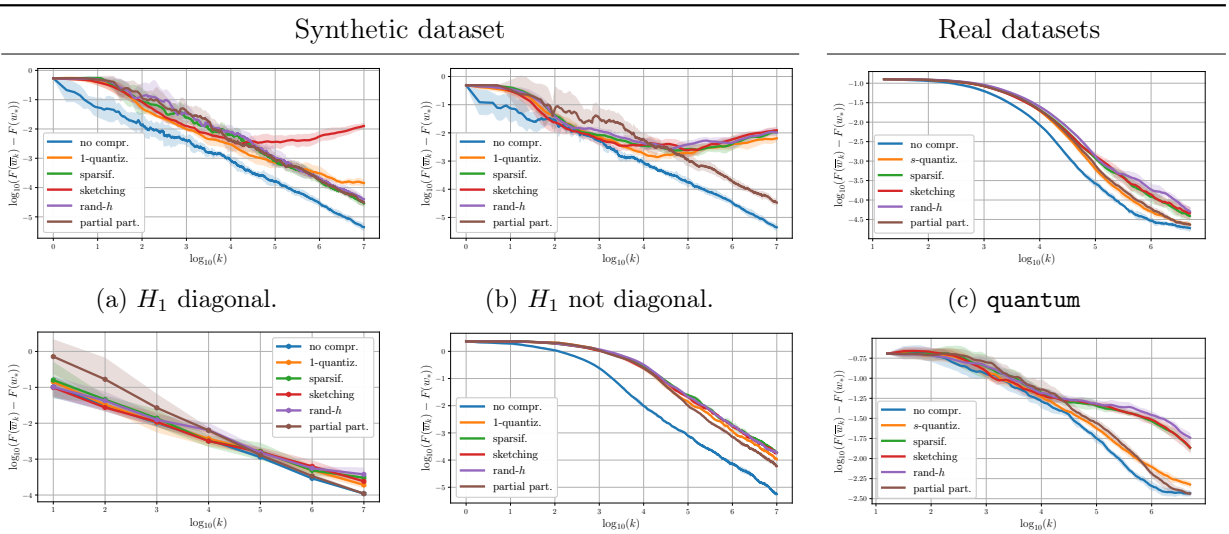
In this section, we run Algorithm 2 on both synthetic and real datasets to illustrate the combined theoretical results of Sections 4.2 and 4.3. In Figure 4.7, we compare the compression operators to the baseline of no-compression. We plot the excess loss of the Polyak-Ruppert iterate $F(\overline{w}_k) - F(w)$, versus the index in log/log scale. Each experiment is conducted 5 times, with a new dataset generated from a new seed. The standard deviation of $\log_{10}(F(\overline{w}_k) - F(w))$ is indicated by the shadow-area.

Setting: (a) *Synthetic dataset generation:* The dataset is generated using Model 2 with $K = 10^7$, $\sigma^2 = 1$, an optimal point w_* set as a constant vector of ones and a geometric eigenvalues decay of $D_1 = \mathrm{Diag}\left((1/i^4)_{i=1}^d\right)$ (resp. $D_2 = \mathrm{Diag}\left((1/i)_{i=1}^d\right)$). For $i \in \{1, 2\}$, the covariance matrix is $H_{\{i\}} = QD_{\{i\}}Q^T$, where Q is either orthogonal matrix, or $Q = \mathrm{I}_d$ in the case of a diagonal features' matrix. (b) *Real datasets processing:* We resize images of the `cifar-10` dataset to a 16×16 dimension, and then for both datasets, we apply standardization. To compute the optimal point (and so to compute the excess loss), we run SGD over 200 passes on the whole dataset and consider the last Polyak-Ruppert average as the optimal point w_*. (c) *Algorithm 2:* We take a constant step-size $\gamma = 1/(2(\omega + 1)R^2)$ with R^2 the trace of the features' covariance, and $w_0 = 0$ as initial point. We set the batch-size $b = 1$ (resp. $b = 16$) and the compressor variance $\omega = 10$ (resp. $\omega = 1$, thus a factor 4 compression for `quantum` and factor 2 for `cifar-10`) for synthetic datasets (resp. for real datasets). For `cifar-10` and `quantum`, we run Algorithm 2 for 5×10^6 iterations (it corresponds to 100 passes on the whole dataset). These settings are summarized in Tables D.1 and D.2 in Subsection D.1.1. Additionally, to illustrate Corollary 4.1, we plot on Figure 4.7d the final excess loss after running Algorithm 2 with an horizon-dependent step-size $\gamma = K^{-2/5}$, computed for seven values of $K \in \{10^i, i \in [\![1, 7]\!]\}$.

Interpretation – H diagonal (Figure 4.7a). For sparsification, rand-h, and PP (linear compressors), the rate of convergence is given by Theorem 4.2. As stated by Corollary 4.3, the covariance $\mathfrak{C}_{\rm ania}$ is proportional to H leading to a $O(1/K)$ rate. We indeed observe in Figure 4.7a that excess loss is linear in a log/log scale.

For non-linear compression operators, the rate is given by Theorem 4.1. On the one hand, 1-quantization results in a slower eigenvalues' decay, leading to a larger $\mathrm{Tr}\left(\mathfrak{C}_{\rm ania}H^{-1}\right)$, thus a slower convergence than linear compressors. On the other hand, for sketching, covariance has a purely isotropic part scaling with I_d, which causes $\mathrm{Tr}\left(\mathfrak{C}_{\rm ania}H^{-1}\right)$ to strongly depend on the strong-convexity coefficient μ resulting in an extremely large constant. Both behaviors are observed in Figure 4.7a.

Interpretation – H not diagonal (Figures 4.7b and 4.7e). In the case of the high eigenvalues'



(a) H_1 diagonal. (b) H_1 not diagonal. (c) `quantum`

(d) H_1 not diagonal, $\gamma = K^{-2/5}$. (e) H_2 not diagonal. (f) `cifar10`

Figure 4.7: Logarithm excess loss of the Polyak-Ruppert iterate for a single client ($N = 1$).

decay of H_1 ($\mu = 10^{-8}$), the only compressor that shows in Figure 4.7b a linear rate of convergence in the log/log scale is PP. All others exhibit a saturation phenomenon after a certain number of iterations. This is again due to the unstructured part of the noise for all other compressors, as given by Proposition 4.2. Besides, we also note an increase of the excess loss after some iterations that is likely caused by the accumulation of noise on axis onto which the curvature of H is weak (but the isotropic noise is not). However, taking the optimal horizon-dependent step-size given by Corollary 4.1, we recover on Figure 4.7d for all compressor $\mathcal{C}$ the sub-linear convergence rate of PP shown at Figure 4.7b, reducing by a factor 100 the excess loss w.r.t. to the scenario where $\gamma = 1/2(\omega + 1)R^2$). While using a small step-size is slightly worse for SGD, it reduces the second and third term of the variance in Theorem 4.1 that depends on μ for other compressors. And in the scenario of a slow eigenvalues' decay ($\mu = 10^{-2}$), we observe on Figure 4.7e that all compressors reach the sub-linear rate (same slope -1 on the log/log plot), but with different constants. This illustrates Theorems 4.1 and 4.2 in the case of moderate coefficient μ where we expect the second and third parts of the variance term to be negligible.

Interpretation - real datasets (Figures 4.7c and 4.7f). We observe that quantization performs competitively with PP and outperforms all other compressors. The asymptotic behavior is consistent with Figure 4.6 (top-row) for $\omega = 1$, where the order $\mathcal{C}_{\text{PP}} \simeq \mathcal{C}_{\text{q}} \ll \mathcal{C}_{\text{s}} \simeq \mathcal{C}_{\text{rd}h} \simeq \mathcal{C}_{\Phi}$ is observed. This experience is going beyond Proposition 4.2 which only applies to 1-quantization.

4.3.5 Conclusion

In this section, we investigated how the compression scheme choice impacts the convergence rate, first by showing that quantization-based and projection-based methods respectively satisfy Theorem 4.1 and Theorem 4.2, resulting in different non-asymptotic behaviors. In the asymptotic regime, in both cases, the averaged excess loss scales as $\text{Tr}\left(H^{-1}\mathfrak{C}_{\text{ania}}\right)/K$. We then analyzed the impact of the most-used schemes on this limit rate. Overall, it appears that all compression schemes typically generate an *unstructured-noise*, which covariance does not scale with H, contrarily to the classical un-compressed Algorithm 1. The one exception is PP, which corresponds (on a single worker) to performing fewer iterations. For other compression schemes, we show the impact of the covariance H: depending on the correlation between features (H diagonal or not) and on the pre-processing (e.g., standardization for which H has diagonal constant), the ordering between compression scheme varies.

In many cases, this highlights the need for an additional regularisation when running Algorithm 2: all compression schemes (but PP) result in a significant noise that accumulates along the low curvature directions. Our results can be extended to the ridge (a.k.a., Tikhonov) regularized case [see Dieuleveut et al., 2017], which creates an additional bias but changes the rate $\mathrm{Tr}\left(H^{-1}\mathfrak{C}_{\mathrm{ania}}\right)/K$ into $\mathrm{Tr}\left((H+\lambda \mathrm{I})^{-1}\mathfrak{C}_{\mathrm{ania}}\right)/K$. The theoretical optimal choice for λ depending on H and the compression scheme could be obtained from our analysis but is left as future work.

We now turn to the distributed/federated case, which motivates the study of compression schemes for practical applications.

4.4 Application to federated learning

In this section, we consider Algorithm 3 under Model 1, which corresponds to heterogeneous federated learning on a network composed of N clients. We hereafter consider two particular cases of Model 1. First, in Subsection 4.4.1, the *covariate-shift* case, i.e., Model 1 with $w_*^i = w_*$ for all i (thus the distribution of y^i conditional to x^i does not change between workers), but on the other hand, the features' marginal distributions are different, in particular, $H_i \neq H_j$. Second, in Subsection 4.4.2, the *optimal-point-shift* case, i.e., for each client $i, j \in \{1, \dots, N\}$, their optimal points are different $w_*^i \neq w_*^j$, but $H_i = H_j$. In the rest of the section, we denote $\overline{H} := \frac{1}{N}\sum_{i=1}^N H_i$, $\overline{R}^2 := \frac{1}{N}\sum_{i=1}^N R_i^2$, and we have $F(w_k) - F(w_*) = \frac{1}{2}\left\langle \eta_{k-1}, \overline{H}\eta_{k-1}\right\rangle$.

4.4.1 Heterogeneous covariance

In this section, we first show that Theorems 4.1 and 4.2 on (LSA) from Section 4.2 can be applied to the federated learning case within the scenario of covariate-shift. Corollary 4.4 is proved in Subsection D.6.1.

Corollary 4.4 (Algorithm 3 with covariate-shift)**.** *Consider Algorithm 3 under Model 1 with $w_*^i = w_*^j$ (and potentially $H_i \neq H_j$).*

1. *For a compressor $\mathcal{C} \in \{\mathcal{C}_{\mathrm{q}}, \mathcal{C}_{\mathrm{sq}}, \mathcal{C}_{\mathrm{rdh}}, \mathcal{C}_{\mathrm{s}}, \mathcal{C}_{\Phi}, \mathcal{C}_{\mathrm{PP}}\}$, Theorem 4.1 holds, with $H_F = \overline{H}$, $R_F^2 = \overline{R}^2$, $\mathcal{A} = (\omega+1)\overline{R}^2\sigma^2/N$, $\mathcal{M}_2 = (\omega+1)\max_{i\in\{1,\dots,N\}}(R_i^2)/N$, $\mathcal{M}_1 = \Omega\sigma\max_{i\in\{1,\dots,N\}}(R_i^2)/N$.*
2. *Moreover for any linear compressor $\mathcal{C} \in \{\mathcal{C}_{\mathrm{rdh}}, \mathcal{C}_{\mathrm{s}}, \mathcal{C}_{\Phi}, \mathcal{C}_{\mathrm{PP}}\}$, Theorem 4.2 holds, with the same constants and $\mathrm{III}_{\mathrm{add}} = \sigma^2\max_{i\in\{1,\dots,N\}}(\mathrm{III}_{H_i})/N$ and $\mathrm{III}_{\mathrm{mult}} = \max_{i\in\{1,\dots,N\}}(R_i^2\mathrm{III}_{H_i})/N$, with $(\mathrm{III}_{H_i})_{i=1}^N$ given in Corollary 4.2.*

The Hessian of the objective function is now $\overline{H}$, and Theorems 4.1 and 4.2 still hold. The proof consists in showing that with Lemma 4.1, Assumptions 4.1 to 4.4 on the resulting random field $(\xi_k)_{k\in\mathbb{N}^*}$ are valid, with the constants given above.

In order to understand the impact of the compressor on the limit convergence rate, we establish a formula for $\mathfrak{C}_{\mathrm{ania}}$ similar to Equation (4.4). In the setting of covariate-shift, we have for any clients $i, j \in \{1, \dots, N\}$, $w_*^i = w_*^j$, thus

$$\xi_k^{\mathrm{add}} \overset{\text{def. 4.2}}{=} \xi_k(0) \overset{\text{algo 3}}{=} \nabla F(w_*) - \frac{1}{N}\sum_{i=1}^N \mathcal{C}_k^i(g_k^i(w_*))$$
$$\overset{\text{eq. 4.2}}{=} -\frac{1}{N}\sum_{i=1}^N \mathcal{C}_k^i((\langle x_k^i, w_*\rangle - y_k^i)x_k^i) \underset{\text{with } w_*^i = w_*^j}{\overset{\text{model 1}}{=}} \frac{1}{N}\sum_{i=1}^N \mathcal{C}_k^i(\varepsilon_k^i x_k^i)\,.$$

Next for all operators under consideration we have $\mathcal{C}_k^i(\varepsilon_k^i x_k^i) \overset{\text{a.s.}}{=} \varepsilon_k^i \mathcal{C}_k^i(x_k^i)$, thus, with p_{H_i} denoting the distribution of x_k^i with covariance H_i, we have:

$$\mathfrak{C}_{\text{ania}} = \mathbb{E}\left[(\xi_k^{\text{add}})^{\otimes 2}\right] = \mathbb{E}\left[\left(\frac{1}{N}\sum_{i=1}^N \mathcal{C}_k^i(\varepsilon_k^i x_k^i)\right)^{\otimes 2}\right] \overset{\text{indep. of } (\mathcal{C}_k^i)_{i=1}^d}{=} \frac{1}{N^2}\sum_{i=1}^N \mathbb{E}\left[\mathcal{C}_k^i(\varepsilon_k^i x_k^i)^{\otimes 2}\right]$$
$$= \frac{\sigma^2}{N^2}\sum_{i=1}^N \mathbb{E}\left[\mathcal{C}_k^i(x_k^i)^{\otimes 2}\right] \overset{\text{Def. 4.5}}{=} \frac{\sigma^2}{N^2}\sum_{i=1}^N \mathfrak{C}(\mathcal{C}_k^i, p_{H_i}) \overset{\text{notation}}{=:} \frac{\sigma^2}{N}\overline{\mathfrak{C}((\mathcal{C}^i, p_{H_i})_{i=1}^N)}\,. \tag{4.5}$$

The operator $\overline{\mathfrak{C}((\mathcal{C}^i, p_{H_i})_{i=1}^N)}$ generalizes the notion of *compressor's covariance* (Definition 4.5) to the case of multiple clients, and Equation (4.5) corresponds to Equation (4.4).

Remark 4.11 (All clients use the same *linear* compressor)**.** *If for all $i \in \{1, \dots, N\}$, $\mathcal{C}^i \overset{(d)}{=} \mathcal{C}$ and $\mathcal{C} \in \{\mathcal{C}_{\text{PP}}, \mathcal{C}_{\text{s}}, \mathcal{C}_{\text{rd}h}, \mathcal{C}_\Phi\}$, leveraging Remark 4.10, we have*

$$\overline{\mathfrak{C}((\mathcal{C}^i, p_{H_i})_{i=1}^N)} = \mathfrak{C}(\mathcal{C}, \overline{H})\,.$$

The analysis of (LSA) *on a single worker made in Section 4.3 is still valid in this setting with now the Hessian of the problem being equal to the average of covariance $\overline{H}$. Corollary 4.4 and Equation* (4.5) *prove that the case of covariate-shift is identical to the centralized setting with a variance reduced by a factor N.*

Remark 4.12 (Varying compressor, or compression level, or non-linear compression)**.** *In most other cases, the computation of $\frac{\sigma^2}{N}\overline{\mathfrak{C}((\mathcal{C}^i, p_{H_i})_{i=1}^N)} = \frac{\sigma^2}{N^2}\sum_{i=1}^N \mathfrak{C}(\mathcal{C}_i^i, p_{H_i})$ is possible using the results of Subsection 4.3.3*

Overall, in the covariate-shift case, most insights from the centralized case remain valid, especially, client sampling (i.e., PP) is the safest way to limit the impact of compression. Moreover, the trade-offs and ordering between compressors remain preserved, especially regimes in which quantization outperforms other competitors.

4.4.2 Heterogeneous optimal point

Hereafter, we focus on the case of heterogeneous optimal points and consider that all clients share the same covariance matrix, i.e. for any $i, j \in \{1, \dots, N\}$, $H_i = H$, but potentially $w_*^i \neq w_*^j$. This can be seen as a case of *concept-shift* [Kairouz et al., 2019], and we also refer to the situation as *optimal-point-shift*. This setting could eventually be combined with the covariate-shift case. Similarly, Theorems 4.1 and 4.2 on (LSA) from Section 4.2 can be applied.

Corollary 4.5 (Algorithm 3 with concept-shift)**.** *Consider Algorithm 3 under Model 1 with $H_i = H_j$ (and potentially $w_*^i \neq w_*^j$).*

1. *For a compressor $\mathcal{C} \in \{\mathcal{C}_{\text{q}}, \mathcal{C}_{\text{sq}}, \mathcal{C}_{\text{rd}h}, \mathcal{C}_{\text{s}}, \mathcal{C}_\Phi, \mathcal{C}_{\text{PP}}\}$, Theorem 4.1 holds, with $H_F = H$, $R_F^2 = R^2$, $\mathcal{A} = \frac{R^2(\omega+1)}{N}(\kappa \text{Tr}\left(H\text{Cov}\left[W_*\right]\right) + \sigma^2)$ with $W_* \sim \text{Unif}(\{w_*^i, i \in \{1, \dots, N\}\})$, $\mathcal{M}_2 = (\omega+1)^2/N$, and $\mathcal{M}_1 = \Omega R^2 \sigma/N$.*
2. *Moreover for any linear compressor $\mathcal{C} \in \{\mathcal{C}_{\text{rd}h}, \mathcal{C}_{\text{s}}, \mathcal{C}_\Phi, \mathcal{C}_{\text{PP}}\}$, Theorem 4.2 holds, with the same constants and $\text{III}_{\text{add}} = \sigma^2 \text{III}_H/N$ and $\text{III}_{\text{mult}} = R^2 \text{III}_H/N$, with III_H given in Corollary 4.2.*

Corollary 4.5 can be proved reusing computation made for Corollary 4.4 and using below Proposition 4.5. We next aim at computing the additive noise covariance. We note $g_{k,*}^i = g_k^i(w_*)$ the local stochastic gradient evaluated at optimal point w_*. We have, in Model 1, for any $w \in \mathbb{R}^d$,

$F_i(w) := \mathbb{E}(\langle x_k^i, w - w_*^i\rangle - x_k^i\varepsilon_k^i)^2$, thus $\nabla F(w) = \frac{1}{N}\sum_{i=1}^N H(w - w_*^i)$, and $w_* = \sum_{i=1}^N w_*^i/N$. The setting of Definition 4.1 is verified with $H_F = H$, and for any $w \in \mathbb{R}^d$, that the random field ξ_k can be computed as:

$$\xi_k(w - w_*) \overset{\text{Def. 4.1\&Alg.3}}{=} H_F(w - w_*) - \frac{1}{N}\sum_{i=1}^N \mathcal{C}^i(g_k^i(w)), \text{ thus } \xi_k^{\text{add}} \overset{\text{Def. 4.2}}{=} -\frac{1}{N}\sum_{i=1}^N \mathcal{C}^i(g_{k,*}^i),$$

with $g_{k,*}^i = (x_k^i \otimes x_k^i)(w_* - w_*^i) + x_k^i\varepsilon_k^i$. We thus have, for any $k \in \mathbb{N}$:

$$\begin{aligned}
\mathfrak{C}_{\text{ania}} &= \mathbb{E}\left[(\xi_k^{\text{add}})^{\otimes 2}\right] \overset{\nabla F(w_*)=0}{=} \mathbb{E}\left[\left(\frac{1}{N}\sum_{i=1}^N \mathcal{C}^i(g_{k,*}^i) - \nabla F_i(w_*)\right)^{\otimes 2}\right] \\
&\overset{\substack{\forall i\neq j,\ \mathcal{C}_k^i \perp \mathcal{C}_k^j \\ \mathbb{E}\mathcal{C}_k^i(g_{k,*}^i) = \nabla F_i(w_*)}}{=} \frac{1}{N^2}\sum_{i=1}^N \mathbb{E}\left[\left(\mathcal{C}_k^i(g_{k,*}^i) - \nabla F_i(w_*)\right)^{\otimes 2}\right] \\
&= \frac{1}{N^2}\sum_{i=1}^N \left(\mathbb{E}[\mathcal{C}_k^i(g_{k,*}^i)^{\otimes 2}] - \nabla F_i(w_*)^{\otimes 2}\right) \\
&= \frac{\sigma^2}{N^2}\sum_{i=1}^N \mathfrak{C}(\mathcal{C}^i, p_{\Theta_i}) - \frac{1}{N^2}H\sum_{i=1}^N (w_* - w_*^i)^{\otimes 2}H \preccurlyeq \frac{\sigma^2}{N}\overline{\mathfrak{C}((\mathcal{C}^i, p_{\Theta_i})_{i=1}^N)},
\end{aligned}$$

where p_{Θ_i} is the distribution of $g_{k,*}^i$ (for any k). In the last inequality, we simply discarded the non-positive term $-H\sum_{i=1}^N (w_* - w_*^i)^{\otimes 2}H$. For linear compressors, by Proposition 4.2, $\mathfrak{C}_{\text{ania}}$ is a linear function of $\frac{1}{N}\sum_{i=1}^N \Theta_i$ – the averaged second-order moment of the local gradients $(g_{k,*}^i)_{i=1}^N$. In order to bound this quantity, following Dieuleveut et al. [2017], we make the following assumption.

Assumption 4.5. *The kurtosis for the projection of the covariates x_1^i (or equivalently x_k^i for any k) is bounded on any direction $z \in \mathbb{R}^d$, i.e., there exists $\kappa > 0$, such that:*

$$\forall i \in \{1, \dots, N\},\ \forall z \in \mathbb{R}^d, \quad \mathbb{E}\left[\left\langle z, x_1^i\right\rangle^4\right] \leq \kappa\langle z, Hz\rangle^2$$

For instance, it is verified for Gaussian vectors with $\kappa = 3$. By Cauchy-Schwarz inequality, it implies that $\mathbb{E}[\left\langle z, x_1^i\right\rangle^2 (x_1^i)^{\otimes 2}] \preccurlyeq \kappa\langle z, Hz\rangle H$ for all $z \in \mathbb{R}^d$. We obtain the following proposition.

Proposition 4.5 (Impact of client-heterogeneity.)**.** *Let W_* be a random variable uniformly distributed over $\{w_*^i, i \in \{1, \dots, N\}\}$, thus such that, $\operatorname{Cov}[W_*] = \frac{1}{N}\sum_{i=1}^N (w_* - w_*^i)^{\otimes 2}$, then:*

$$\frac{1}{N}\sum_{i=1}^N \Theta_i \preccurlyeq \left(\kappa \operatorname{Tr}\left(H\operatorname{Cov}[W_*]\right) + \sigma^2\right) H\,.$$

Proof We have:

$$\begin{aligned}
\Theta_i &= \mathbb{E}[((x_k^i \otimes x_k^i)(w_* - w_*^i) + x_k^i\varepsilon_k^i)^{\otimes 2}] \overset{(\varepsilon_k^i)\perp(x_k^i)}{=} \mathbb{E}[(x_k^i \otimes x_k^i)(w_* - w_*^i)^{\otimes 2}(x_k^i \otimes x_k^i)] + \sigma^2 H \\
&\overset{\text{Ass. 4.5}}{\preccurlyeq} \kappa\left\langle w_* - w_*^i, H(w_* - w_*^i)\right\rangle H + \sigma^2 H = \kappa \operatorname{Tr}\left(H(w_* - w_*^i)^{\otimes 2}\right) H + \sigma^2 H\,.
\end{aligned}$$

■

In words, we have the following two main observations.

Remark 4.13 (**Structured noise before compression.**)**.** *Before compression is possibly applied, the noise remains structured, i.e., with covariance proportional to H, in the case of concept-shift. As a consequence, the rate for un-compressed Equation (LSA) will remain independent of the smallest eigenvalue of H. This remark extends to the case where $\mathcal{C}_{\text{PP}}$ is applied.*

Remark 4.14 (**Heterogeneous vs homogeneous case**). *Compared to the homogeneous case, in which* $\Theta_i = \sigma^2 H_i$ *and* $\mathfrak{C}_{\text{ania}} = \frac{\sigma^2}{N}\overline{\mathfrak{C}((\mathcal{C}^i, p_{H_i})_{i=1}^N)}$*, the averaged second-order moment increases from* $\sigma^2 H$ *to* $(\kappa \mathrm{Tr}\left(H \mathrm{Cov}\left[W_*\right]\right) + \sigma^2)H$*, showing the impact of the dispersion of the optimal points* $(w_*^i)_{i=1}^N$*. This corresponds to the typical variance increase in the compressed heterogeneous SGD case, see Section 2.3.*

Concept-shift thus hinders the limit convergence rate. To limit this effect, in Chapter 2, we introduced a control-variate term $(h_k^i)_{k\in\mathbb{N}^*, i\in\{1,\dots,N\}}$, that is subtracted to the gradient before compression and asymptotically approximate $\nabla F_i(w_*)$ for any $i \in \{1, \dots, N\}$. We explore the impact of memory on the asymptotic convergence in Subsection D.6.2.

4.4.3 Numerical experiments

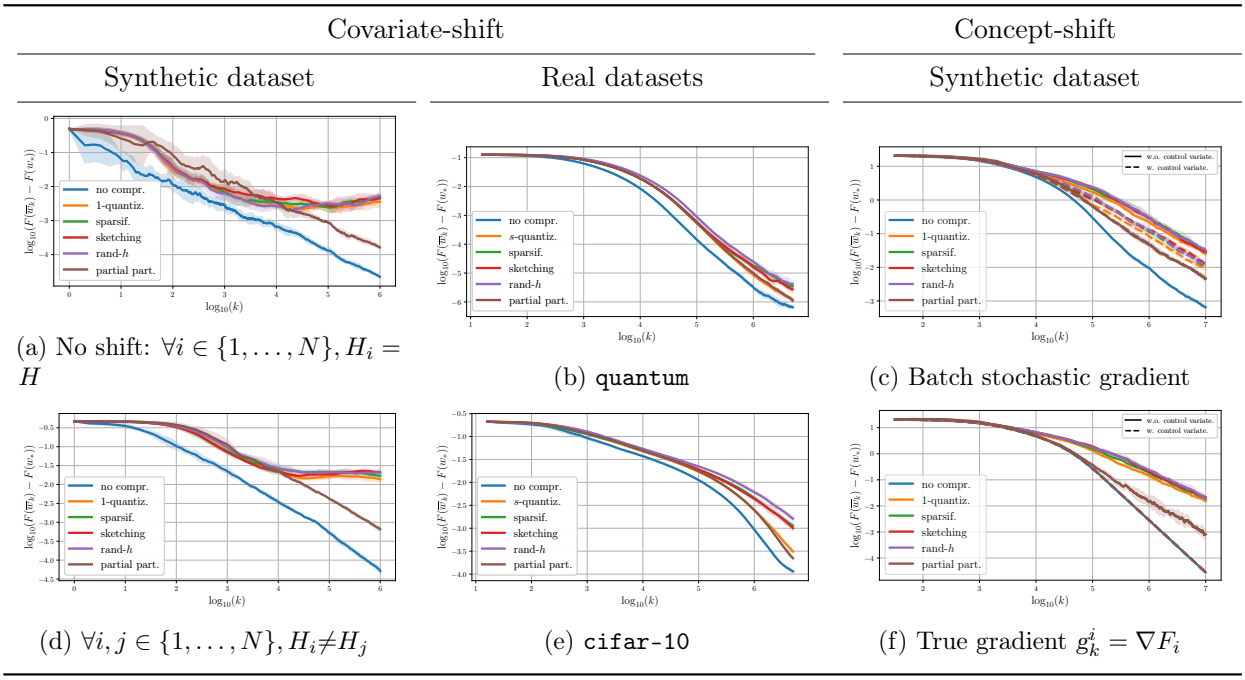
We support the theoretical results from Subsections 4.4.1 and 4.4.2 by performing experiments in the FL framework that extend the ones from Section 4.3.

On figures Figure 4.8, we present the results of the excess loss of the Polyak-Ruppert iterate $F(\overline{w}_k) - F(w_*)$ versus the number of iterations in log/log scale. The experiments were run 5 times, each time with different datasets (dispersion is shown by shaded area).

Settings. (a) *Synthetic dataset generation:* The dataset is generated using Model 1 with $N = 10$, $K = 10^6$ on each client, $\sigma^2/N = 1$. For any clients i in $\{1, \dots, N\}$, the covariance matrix is $H_i = Q_i D_i Q_i^T$, where Q_i is an orthogonal matrix. For heterogeneous clients, the dataset generation is as follows. *Covariate shift:* The rotation matrix Q_i is sampled independently for each client and the diagonal matrix D_i is $\mathrm{Diag}\left((1/j^{\beta_i})_{j=1}^d\right)$ where $\beta_i \sim \mathrm{Unif}(\{3,4,5,6\})$. *Concept-shift:* The optimal models of the clients $i \in \{1, \dots, N\}$ were drawn from a zero-centered normal distribution with a variance of $100\mathrm{I}_d$, that is, $w_*^i \sim \mathcal{N}(0, 100\mathrm{I}_d)$. We also take for all client i in $\{1, \dots, N\}$, $H_i = QDQ^T$, with $D = \mathrm{Diag}\left((1/j)\right)_{j=1}^d$. (b) *Real-dataset and covariate-shift:* To simulate non-i.i.d. clients, we split the dataset in heterogeneous groups (with equal number of points) using a K-nearest neighbors clustering on the TSNE representations [defined by Maaten and Hinton, 2008]. Thus, the marginal feature distribution significantly varies between clients, providing a covariate-shift, while keeping the same distribution for the output conditional to the features on all clients. (c) *Algorithm 3:* We take a constant step-size $\gamma = 1/(2(\omega+1)R^2)$ with $R^2 = \mathrm{Tr}\left(H\right)$ and $w_0 = 0$ as initial point. We set the batch-size $b = 1$ for synthetic datasets and $b = 16$ for real datasets, the compressor variance is $\omega = 10$. (d) *Algorithm 3* vs *Algorithm 4*: We take a bigger constant step-size $\gamma = (2R^2)^{-1}$ in order to emphasize the difference between the case w./w.o. control variate, we set $w_0 = 0$ as initial point and the compressor variance is $\omega = 10$. We set the batch-size $b = 32$ for Figure 4.8c and $b = K$ for Figure 4.8f.

Interpretation – homogeneous case and covariate-shift case (Figures 4.8a, 4.8b, 4.8d and 4.8e). These experiments extend those presented in Subsection 4.3.4 in the case of a single client. The observations made in the centralized case (Figure 4.7), especially on the impact of the compressor choice on the convergence and the ordering between limit convergence rates remain valid. This illustrates Corollary 4.4 and Remark 4.11: Theorems 4.1 and 4.2 hold in the case of homogeneous client or in the case of heterogeneous covariance and the only compressor that ensures that the noise is structured is client sampling (partial participation). On the real datasets, quantization is also competitive.

Interpretation – concept-shift case (Figures 4.8c and 4.8f). These experiments extend those presented on Figure 4.7e (slow eigenvalues' decay with $\mu = 10^{-2}$) to the scenario of concept shift. First, we observe on Figure 4.8c that for all compressors the convergence rate remains in $O(1/K)$, (though vanilla SGD converges faster during the first iterations). Second, we observe that control-variates improve convergence for compressors inducing un-structured noise ; this is



(a) No shift: $\forall i \in \{1, \dots, N\}, H_i = H$

(b) `quantum`

(c) Batch stochastic gradient

(d) $\forall i, j \in \{1, \dots, N\}, H_i \neq H_j$

(e) `cifar-10`

(f) True gradient $g_k^i = \nabla F_i$

Figure 4.8: Logarithm excess loss of the Polyak-Ruppert iterate iterations for $N = 10$ clients.

predicted by theory, see Theorem D.3. Third, on Figure 4.8f, at each iteration $k \in \{1, \dots, K\}$, we use deterministic gradients $g_k^i = \nabla F_i$ which leads to having a.s. $\xi_k^{\text{add}} = 0$, and in the absence of compression, we obtain a $O(1/K^2)$ convergence rate for $\overline{w}_K$ which corresponds in Theorem 4.1 to the case where the dependency on the initial condition is dominated by $\frac{\|H_F^{-1/2}\eta_0\|^2}{\gamma^2 K^2}$. Overall, these experiments illustrate and support our theoretical insights.

4.5 Conclusion

Conclusion. In short, we investigate the impact of the choice of compression scheme on the convergence of the Polyak-Ruppert averaged iterate. By analysing the case of compressed least-squares regression, we shed light on the interplay between the Hessian of the optimization problem H_F, the features' distribution, the additive noise's covariance $\mathfrak{C}_{\text{ania}}$, and the compression scheme. This shows fundamental differences between compression that deemed equivalent under the classical worst-case-variance assumption. We extend our analysis to the case of heterogeneous federated learning, a setting in which compression is widely used and its impact not fully understood.

More precisely, first, the analysis of the generic stochastic approximation algorithm (LSA) provides (1) the fact that projection based compressions achieve a faster convergence rate than quantization based, and that yet, their asymptotic rate is similar; (2) the analysis of quantization-based compression requires introducing a new Hölder-type regularity assumption for the analysis of the stochastic approximation scheme, and showing that such an assumption is satisfied for quantization.

Second, the computation of the additive noise's covariance $\mathfrak{C}_{\text{ania}}$ reveals the impact of the compression scheme and the data distribution on the asymptotically dominant term. We obtain that (1) partial participation (i.e., client sampling in the federated case) is the only method that systematically ensures a convergence without a dependency on the strong-convexity constant; (2) other compressors may all induce an un-structured noise, with covariance scaling with I or $\sqrt{H}$, that strongly hinders convergence by accumulating noise on low curvature directions; (3) the relative

performance or various schemes changes depending on the pre-processing applied to the data, making quantization the best method (apart from PP) when standardization is applied, but one of the worst (with random Gaussian projection) when the features are independent and the eigenvalues of the covariance decay rapidly (4) in that particular last setting, all projection based methods (but Gaussian projection) behave similarly.

Third, we discuss how these results apply to the federated case, that corresponds to the initial motivation. We show that we encompass two particular heterogeneity situations and how our analysis applies. Overall, these results are a step towards a better understanding of the impact of a widely used tool.

Open directions. This analysis can be extended to include various aspects that are beyond the scope of this work. First, one natural improvement for application in FL would be to consider also the scenario where each client runs several *local iterations* [McMahan et al., 2017, Karimireddy et al., 2020] before sending their updates, reducing further the cost of communication. Similar approach can be used, although the additive noise field would be more complicated, which potentially implies a different additive noise's covariance. Second, as mentioned in Subsection 4.3.5, our analysis could also be extended to the case of stochastic approximation with ridge regularization [e.g., following Dieuleveut et al., 2017] which in practice is helpful to mitigate the impact of the lack of strong convexity. Third, an obvious direction is to extend beyond quadratic functions and considering other objective functions, such as logistic regression or even shallow neural networks. Several results in the literature can be leveraged to tackle non quadratic but self-concordant losses Bach [2010], Gadat and Panloup [2023]. Fourth, our analysis still only relies on second moments (variance and covariance) of the stochastic field. One major drawback of partial participation is to induce a significant increase on higher order moments. Incorporating higher order bounds may also bring novel insights to the use of compression in FL. Finally, all our analysis is made in finite dimension and our asymptotic focuses on $K \to \infty$: further works should analyze the case of infinite dimension: within the reproducing kernel Hilbert space [Dieuleveut and Bach, 2016] framework or within the overparametrized setting [Belkin et al., 2019].

5

Conclusion and perspectives

‘Ἐν ἀρχῇ ἦν ὁ λόγος, καὶ ὁ λόγος ἦν πρὸς τὸν θεόν, καὶ θεὸς ἦν ὁ λόγος.’

Ἰωάννην 1:1.

5.1 Conclusion

In this thesis, we investigated several aspects of stochastic optimization for federated learning with the objective of reducing the cost of communication in a setting of heterogeneous clients.

In the opening Chapter of this thesis, we provide an overview of the general setting of statistical learning, convex optimization, stochastic approximation, and federated learning, which is the main motivation of this thesis. More particularly, we introduce the bidirectional compression setting, which is the focus of Chapters 2 and 3, and highlight its relevance for distributed algorithms.

In our first contribution, we focus on the intertwined effect of compression and client (statistical) heterogeneity. We introduce a framework – `Artemis` – to tackle the problem of learning in a distributed or federated setting with communication constraints. To alleviate the communication cost, `Artemis` allows to compress the information sent in *both directions* (from the clients to the server and conversely) combined with a memory mechanism. We provide three tight theorems giving guarantees of a fast convergence (linear up to a threshold), highlighting the impact of memory, analyzing Polyak-Ruppert averaging, and obtaining lowers bounds by studying convergence in distribution of our algorithm. Altogether, this improves the understanding of compression combined with a memory mechanism and sheds light on the challenges ahead.

In our second contribution, we move the focus toward feedback loops to reduce the impact of compression. We propose and analyze an algorithm that performs bidirectional compression and achieves asymptotically the same convergence rate as algorithms using only uplink (from the local clients to the central server) compression. This algorithm, `MCM`, is such that the downlink compression *only impacts local models*, while the global model is preserved. As a result, and contrary to previous works, the gradients on local servers are computed on *perturbed models*. Proposing such an analysis is the key to unlocking numerous challenges in distributed learning, e.g., proposing practical algorithms for partial participation, incorporating privacy-preserving schemes *after* the global update is performed, dealing with local steps, etc.

In our third contribution, we go beyond the classical worst-case assumption on the variance of compressors and provide a fine-grained analysis of the impact of compression within the fundamental learning framework of least-squares regression. More precisely, we analyze a general stochastic

algorithm for minimizing quadratic functions relying upon a random field. We consider weak assumptions on the random field, tailored to the analysis (specifically, expected Hölder regularity), and on the noise covariance, enabling the analysis of various randomizing mechanisms, including compression. It underlines differences in terms of convergence rates between several unbiased compression operators, that all satisfy the same condition on their variance. We then extend our results to two heterogeneous FL frameworks.

Overall, this thesis proposes contributions to the field of Federated Learning by addressing central challenges and proposing solutions for efficient and sustainable learning in a distributed and heterogeneous framework. This work aligns with a global effort to make the use of large-scale Federated Learning viable by minimizing its environmental impact. Although benefits are expected, at least with respect to energy concerns, cautiousness is still required, as a rebound effect could occur: having faster and less energy-consuming algorithms could lead to a sharp increase in their applications, reducing or even canceling out the gains made by progress in their design.

5.2 Perspectives

From a theoretical point of view, several questions were triggered by our results, I think they are worth exploring in order to delve deeper into the understanding of the effects of compression in a heterogeneous environment.

1. In Chapter 2, we have modeled heterogeneity by considering that local gradients evaluated at the optimal point are not zero, which explicitly assumes the existence of such a point. This raises three questions. *Firstly*, how to extend results on `Artemis` to non-convex scenarios where such an optimal point does not exist? Following Karimireddy et al. [2020], an approach is to consider bounded gradient dissimilarity which consists in assuming that there exist constants $B \geq 0$ and $G \geq 1$ such that for any w in $\mathbb{R}^d$, we have $\frac{1}{N}\sum_{i=1}^N \|\nabla F_i(w)\|^2 \leq B^2 + G^2\|\nabla F(w)\|^2$. Note that this recovers our Assumption 1.8 if F is convex and if there exists at least one optimal point w_*. *Secondly*, is this assumption on the gradient relevant to describe any statistical heterogeneity? Indeed such an assumption is closely related to the optimization process and not to the dataset itself; for instance, in Chapter 4 we consider clients with heterogeneous features' covariances, yet simultaneously, we have $\nabla F_i(w_*) = 0$ for any client i in $\{1, \dots, N\}$. In another example, in Chapter 4, we consider clients with heterogeneous optimal models, which indeed leads to having $\nabla F_i(w_*) \neq 0$, but it also has an impact on the constant σ_* that bounds the variance of stochastic gradient evaluated at optimal points (Assumption 1.5), it depends on the distance $\|w_* - w_*^i\|^2$ between the global and local optimal models. Beyond the scope of optimization, the question of statistical heterogeneity is complex as it may be of different kinds [Kairouz et al., 2019], e.g., covariate-shift, concept-shift, prior-probability-shift, or unbalancedness. Thereby, the issue of modeling all these types of heterogeneity is important in order to design algorithms adapted to each scenario. *Thirdly*, from a practical point of view, the question is next how to provide a measure that evaluates the type and degree of heterogeneity within a network of clients in order to select the algorithm accordingly? Our joint contribution to [du Terrail et al., 2022] which aims to provide such metrics, paves the way to solve this challenge.

2. In Chapter 2, we have shown the key role of memory in the setting of heterogeneous clients, and in Chapter 3 we have combined it with a model-preservation mechanism (corresponding to a feedback-loop with three sequences $(w_k, \widehat{w}_k, H_k)_{k\in\mathbb{N}}$), enabling to recover the rate of convergence of unidirectional compression for unbiased compressors. But as model-preservation mechanism is to be compared with the historical EF mechanism, a question triggered our attention: *what* is the relationship between EF and model-preservation, and in which scenario,

one will outperform the other? This question is open and emphasizes that the relationship between these two mechanisms is not yet completely understood. A first answer can be given based on Theorem 3.5 of Chapter 3 where we show that the third sequence $(H_k)_{k\in\mathbb{N}}$ involved in model-preservation (and not in EF) enables controlling the variance of the degraded model $(\widehat{w}_k)_{k\in\mathbb{N}}$. However, a unified theory regrouping EF [Seide et al., 2014], EF 21 [Richtarik et al., 2021, Fatkhullin et al., 2021], memory [Mishchenko et al., 2019] and model-preservation [Philippenko and Dieuleveut, 2021] appears to be necessary in order to fully leverage the potential of each of this mechanism and to understand the impact of *each* of the involved sequence. Such an analysis could lead to the design of better algorithms taking advantage of all the properties of these four mechanisms. This is a topic of great interest as underlined by the recent works of Gorbunov et al. [2020a] and Condat et al. [2022], which propose a unified framework recovering various algorithms (but not `MCM`), enabling their analysis under very general assumptions [e.g., Gorbunov et al., 2020a, see Assumption 2.3, 4.1 and 4.2, see Table 2].

3. In Chapter 4, we have provided a fine-grained analysis of the impact of compression within the fundamental learning framework of least-squares regression, highlighting the key role of the covariance induced by the additive noise. A relevant perspective would be to go beyond quadratic functions. A first step could be to extend our analysis to logistic regression following the line of proof given in Bach and Moulines [2013]. To extend the result of Chapter 4 to the non-convex case (a starting point could be to consider a neural network with one hidden layer), it appears necessary to investigate the distribution followed by the gradients evaluated at a local optimum. In particular, it would be relevant to highlight the relationship between the features' covariance, the structure of the network, the covariance of the stochastic gradients evaluated at a local optimum point, and the compression scheme. The goal of such an analysis would be to (1) generalize the results of Chapter 4 to neural networks, in particular, to describe the impact of the covariances of the additive noise, and (2) select at each step the optimal compressor for each layer. But then the results on asymptotic normality given in Proposition 4.1 can not be applied anymore. To overcome this difficulty, one solution could be to take inspiration from the recent work of Gadat and Gavra [2022] (Theorems 1 and 2). In this article, the authors have studied asymptotic properties of adaptive algorithms (Adagrad and Rmsprop) in the non-convex setting; their assumptions can be compared with those presented in Section 4.2.

4. Results presented in Chapter 4 can be extended to the ridge (a.k.a., Tikhonov) regularized case [see Dieuleveut et al., 2017], it creates an additional bias but changes the rate $\mathrm{Tr}\left(H^{-1}\mathfrak{C}_{\text{ania}}\right)/K$ into $\mathrm{Tr}\left((H+\lambda \mathrm{I})^{-1}\mathfrak{C}_{\text{ania}}\right)/K$. From our analysis, one could obtain the optimal theoretical choice for λ depending on H and the compression scheme. Adding regularization will help in the case of ill-conditioned problems and will make quantization-based and sparsification-based compressors compete with PP.

5. The analysis of compressors in Chapter 4 shed to light that PP is the most robust compressor to ill-conditioned problems. This is due to the fact that, unlike other compressors, the coordinates are not compressed independently. This property enables the induced noise to be structured and suggests designing compressors where the compression of coordinates is not independent. For example, for 1-quantization, a simple modification could be to choose the same seed for each coordinate to build a coordinate-dependent compressor, but experimentally, we find that this naive approach does not help in practice, leaving the door open for further refinements and improvements.

A

Technical preliminaries

In this Chapter, we provide some classical results that are used throughout this thesis. In particular, we recall some classical inequalities, or some classical results for random vectors and optimization.

Contents

A.1 Identities and inequalities

In this Subsection, we recall some very classical inequalities; for all $a, b \in \mathbb{R}^d$, $\beta > 0$ we have:

$$\langle a, b\rangle \leq \frac{\|a\|^2}{2\beta} + \frac{\beta \|b\|^2}{2}\,, \tag{A.1}$$

$$\|a+b\|^2 \leq (1+\frac{1}{\beta})\,\|a\|^2 + (1+\beta)\,\|a\|^2\,, \tag{A.2}$$

$$\|a+b\|^2 \leq 2\left(\|a\|^2 + \|b\|^2\right)\,, \tag{A.3}$$

$$|\,\langle a, b\rangle\,| \leq \|a\| \cdot \|b\| \qquad \text{(Cauchy-Schwarz's inequality)}\,, \tag{A.4}$$

$$\langle a, b\rangle = \frac{1}{2}\left(\|a\|^2 + \|b\|^2 - \|a-b\|^2\right) \qquad \text{(Polarization identity)}. \tag{A.5}$$

Inequality 1. *Let $N \in \mathbb{N}$ and $d \in \mathbb{N}$. For any sequence of vector $(a_i)_{i=1}^N \in \mathbb{R}^d$, we have the following inequalities:*

$$\left\|\sum_{i=1}^N a_i\right\|^2 \leq \left(\sum_{i=1}^N \|a_i\|\right)^2 \leq N\sum_{i=1}^N \|a_i\|^2\,.$$

The first part of the inequality corresponds to the triangular inequality, while the second part is Cauchy's inequality.

Inequality 2. *Let x in $\mathbb{R}^d$ and A in $\mathcal{M}_{d,d}(\mathbb{R})$, then we have $\|Ax\| \leq |||A|||\|x\|$.*

Lemma A.1. *Let $\alpha \in [0,1]$ and $x, y \in (\mathbb{R}^d)^2$, then:*

$$\|\alpha x + (1-\alpha)y\|^2 = \alpha\,\|x\|^2 + (1-\alpha)\,\|y\|^2 - \alpha(1-\alpha)\,\|x-y\|^2\,.$$

This is a norm's decomposition of a convex combination.

Lemma A.2. *Let X be a random vector of $\mathbb{R}^d$, then for any vector $x \in \mathbb{R}^d$:*

$$\mathbb{E}[\|X - \mathbb{E}X\|^2] = \mathbb{E}[\|X-x\|^2] - \|\mathbb{E}[X]-x\|^2\,.$$

This equality is a generalization of the well know decomposition of the variance (with $x = 0$). A consequence is the next lemma which will be used several times in the proofs.

Lemma A.3. *Let a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ with Ω a sample space, $\mathcal{A}$ a σ-algebra, $\mathbb{P}$ a probability measure, and $\mathcal{F}$ a $\sigma-$algebra. For any $a \in \mathbb{R}^d$ and for any random vector in $\mathbb{R}^d$ we have:*

$$\mathbb{E}\left[\|X - \mathbb{E}X\|^2\right] \leq \mathbb{E}\left[\|X-a\|^2\right]\,,$$

indeed $\mathbb{E}[X] = \arg\min_{a\in\mathbb{R}^d} \mathbb{E}[\|X-a\|^2]$. Similarly, for any random vector Y in $\mathbb{R}^d$ which is $\mathcal{F}$-measurable, we have:

$$\mathbb{E}\left[\|X - \mathbb{E}\left[X \mid \mathcal{F}\right]\|^2 \;\middle|\; \mathcal{F}\right] \leq \mathbb{E}\left[\|X-Y\|^2 \;\middle|\; \mathcal{F}\right]\,.$$

In Chapter 4, we use ellipses to visual quadratic functions, therefore we provide in Definition A.1 the mathematical definition.

Definition A.1 (Representing positive matrices through ellipsoids)**.** *Any symmetric positive definite matrix M in $\mathcal{S}_d^{++}(\mathbb{R})$ defines an ellipsoid $\mathcal{E}_M = \{x \in \mathbb{R}^d, x^\top M^{-1} x = 1\}$ centered around zero. The eigenvectors of M are the principal axes of the ellipsoid, and the squared root of the eigenvalues are the half-lengths of these axes. The ellipse corresponds to the sphere of radius 1 associated with the norm $N_{M^{-1}} = \sqrt{x^\top M^{-1} x}$.*

A.2 Classical results for random vectors

Below, we recall Minkowski's and Jensen's inequalities. Additionally, we recall the Cauchy-Schwarz's inequality for conditional expectations. Let a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ with Ω a sample space, $\mathcal{A}$ a σ-algebra, and $\mathbb{P}$ a probability measure.

Minkowski's inequality. Let $p > 1$ and suppose that X, Y are two random variables in $\mathcal{L}^p(\Omega, \mathcal{A}, \mathbb{P})$ (i.e. their p^{th} moment is bounded), we have the following triangular inequality:

$$\mathbb{E}[\|X + Y\|^p]^{1/p} \leq \mathbb{E}[\|X\|^p]^{1/p} + \mathbb{E}[\|Y\|^p]^{1/p} . \tag{A.6}$$

Jensen's inequality. Suppose that $X : \Omega \longrightarrow \mathbb{R}^d$ is a random variable, then for any convex function $f : \mathbb{R}^d \longrightarrow \mathbb{R}$ we have:

$$f\left(\mathbb{E}(X)\right) \leq \mathbb{E} f(X) . \tag{A.7}$$

Cauchy-Schwarz's inequality for conditional expectations. Suppose that X, Y are two random variables in $\mathcal{L}^2(\Omega, \mathcal{A}, \mathbb{P})$ (i.e. their second moment is bounded), then for any σ-algebra $\mathcal{F} \subset \mathcal{A}$ we have a.s.:

$$\mathbb{E}\left[XY \mid \mathcal{F}\right]^2 \leq \mathbb{E}\left[X^2 \mid \mathcal{F}\right] \mathbb{E}\left[Y^2 \mid \mathcal{F}\right] . \tag{A.8}$$

Convergence in L^p-norm. Suppose that $(X_n)_{n\in\mathbb{N}}$ is a sequence of random variables in $\mathcal{L}^p(\Omega, \mathcal{A}, \mathbb{P})$, and that X is a random variable in $\mathcal{L}^p(\Omega, \mathcal{A}, \mathbb{P})$. We say that $(X_n)_{n\in\mathbb{N}}$ converges in L^p-norm towards X, if the p-th absolute moments $\mathbb{E}(\|X_n\|^p)$ and $\mathbb{E}(\|X\|^p)$ of X_n and X exist, and if $\mathbb{E}(\|X_n - X\|^p) \xrightarrow[n\to+\infty]{} 0$. This type of convergence is denoted by:

$$X_n \xrightarrow[n\to+\infty]{L^p} X . \tag{A.9}$$

A.3 Classical results in optimization

In this section, we provide classical inequalities used in optimization that can be found in Nesterov [2004]. We use these inequalities in the demonstrations given in Chapters B and C.

Proposition A.1. *If $F : \mathcal{X} \subset \mathbb{R}^d \to \mathbb{R}$ is strongly convex, then the following inequality holds:*

$$\forall (x, y) \in \mathbb{R}^d, \langle \nabla F(x) - \nabla F(y), x - y \rangle \geq \mu \|x - y\|^2 .$$

This inequality is a consequence of strong convexity and can be found in [Nesterov, 2004, equation 2.1.22]. The next proposition presents two inequalities used in Chapter 3 when invoking convexity or strong-convexity.

Proposition A.2. *If a function F is convex, then it satisfies for all w in $\mathbb{R}^d$ [Nesterov, 2004, see equation 2.1.7]:*

$$\langle \nabla F(w), w - w_* \rangle \geq \frac{1}{2}(F(w) - F(w_*)) + \frac{1}{2L} \|\nabla F(w)\|^2 \,. \tag{A.10}$$

If a function F is strongly-convex, then it satisfies for all w in $\mathbb{R}^d$ [Nesterov, 2004, see equation 2.1.8 and 2.1.16]:

$$\langle \nabla F(w), w - w_* \rangle \geq \frac{1}{2}(F(w) - F(w_*)) + \frac{\mu}{4} \|w - w_*\|^2 + \frac{1}{2L} \|\nabla F(w)\|^2 \,. \tag{A.11}$$

Polyak and Juditsky [1992] show the following theorem guaranteeing the asymptotic normality of the Polyak-Ruppert iterate. This result is used in Chapter 4.

Theorem A.1. ***From Polyak and Juditsky [1992], see Theorem 1].***

For k in $\mathbb{N}^*$, we denote $\eta_k = w_k - w_*$ and we define $w_k = w_{k-1} - \gamma_k \nabla F(w_{k-1}) + \gamma_k \xi(\eta_{k-1})$. If we assume that:

- $\gamma_k \xrightarrow[k \to +\infty]{} 0$ *and* $\gamma_k^{-1}(\gamma_k - \gamma_{k+1}) = \underset{k \to +\infty}{o}(\gamma_k)$,
- *F is strongly convex and* $\left\|\nabla^2 F\right\|_\infty < \infty$,
- *the convergence in probability of the conditional covariance to a matrix Σ holds, i.e. we have a.s.:*

$$\mathbb{E}\left[\xi(\eta_{k-1})\xi(\eta_{k-1})^\top \;\middle|\; \mathcal{F}_{k-1}\right] \xrightarrow[k \to +\infty]{\mathbb{P}} \Sigma \,.$$

Then for any K in $\mathbb{N}^*$, we have the asymptotic normality of $\sqrt{K}\overline{\eta}_{K-1}$:

$$\sqrt{K}\overline{\eta}_{K-1} \xrightarrow[K \to +\infty]{\mathcal{L}} \mathcal{N}(0, \Sigma^*) \text{ with } \Sigma^\star = \left\{\nabla^2 F\left(w_*\right)\right\}^{-1} \Sigma \left\{\nabla^2 F\left(w_*\right)\right\}^{-1} .$$

Appendix to Artemis

In this Chapter, we provide additional details to our work. First, in Section B.1, we present the detailed framework of our experiments and give further illustrations to our theorems. Secondly, in Section B.2, we define the filtrations used in the following demonstrations. In Section B.3, we gather a few technical results and introduce the lemmas required in the proofs of the main results. Those proofs are given in Section B.4. More precisely, Theorem 2.1 follows from Theorems B.1 and B.2, which are proved in Subsections B.4.1 and B.4.2, while Theorems 2.2 and 2.3 are respectively proved in Subsections B.4.3 and B.4.4.

Contents

B.1 Experiments

In this section we provide additional details about our experiments. We recall that we use two kind of datasets: 1) toy-ish synthetic datasets and 2) real datasets: *superconduct* [Hamidieh, 2018, 21263 points, 81 features] and *quantum* [Caruana et al., 2004, 50,000 points, 65 features]. The aim of using synthetic datasets is mainly to underline the properties resulting from Theorems 2.1 to 2.3. We estimate in Subsection B.1.3 the carbon footprint of the experiments presented in this chapter.

We use the same 1-quantization scheme (see Definition 1.2, $s = 1$ is the most drastic compression) for both uplink and downlink, and thus, we consider that $\omega_{\text{up}} = \omega_{\text{dwn}}$. In addition, we choose $\alpha_{\text{up}} = \frac{1}{2(1+\omega^{\text{up}})}$.

For each figure, we plot the convergence w.r.t. the number of iteration k or w.r.t. the theoretical number of bits exchanged after k iterations. On the Y-axis we display $\log_{10}(F(w_{k-1}) - F(w_*))$, with k in $\mathbb{N}$. All experiments have been run 5 times and averaged before displaying the curves. We plot error bars on all figures. To compute error bars we take the standard deviation of $\log_{10}(F(w_{k-1}) - F(w_*))$, we then plot the curve $\pm$ this standard deviation.

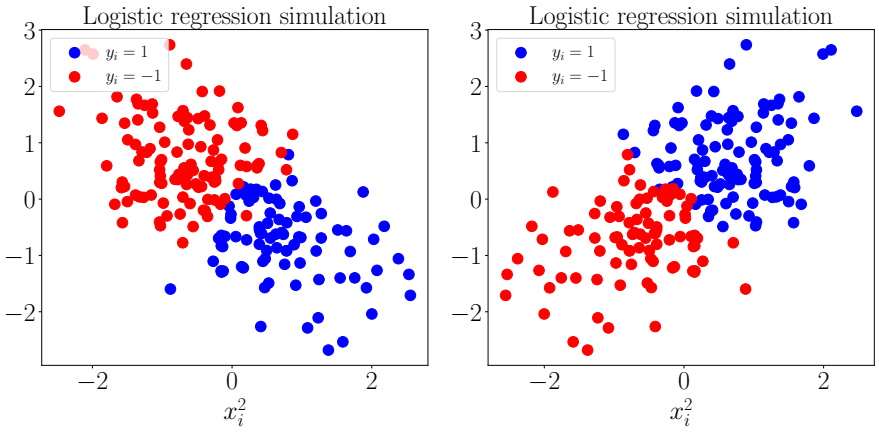
All the code is available on GitHub.

B.1.1 Synthetic dataset

We build two different synthetic dataset for i.i.d. or non-i.i.d. cases. We use linear regression to tackle the i.i.d case and logistic regression to handle the non-i.i.d. settings. As explained in Section 2.1, each worker i holds n_i observations $(z_j^i)_{1\leq j\leq n_i} = (x_j^i, y_j^i)_{1\leq j\leq n_i} = (X^i, Y^i)$ following a distribution D_i.

We use $N = 20$ devices, each holding 200 points of dimension $d = 20$ for least-square regression and $d = 2$ for logistic regression. We ran algorithms over 100 epochs.

Choice of the step-size for the synthetic datasets. For stochastic descent, we use a decreasing step-size $\gamma_k = \frac{1}{L\sqrt{k}}$ with k in $\mathbb{N}$, and for the full gradient descent we choose $\gamma = \frac{1}{L}$.



(a) Dataset 1 (b) Dataset 2

Figure B.1: Data distribution for logistic regression to simulate non-i.i.d. data. Half of the device holds the first dataset, and the other half the second one.

For i.i.d. setting, we use a linear regression model without bias. For each worker i, data points are generated from a normal distribution $(x_j^i)_{1\leq j\leq n_i} \sim \mathcal{N}(0,\Sigma)$. And then, for all j in $[\![1,n_i]\!]$, we have: $y_j^i = \left\langle w, x_j^i \right\rangle + e_i$ with $e_i \sim \mathcal{N}(0,\lambda^2)$ and w the true model.

To obtain $\sigma_* = 0$, it is enough to remove the noise e_i by setting the variance λ^2 of the dataset distribution to 0. Indeed, using a least-square regression, for all i in $[\![1,N]\!]$, the cost function evaluated at point w is $F_i(w) = \frac{1}{2}\|{X^i}^T w - Y^i\|^2$. Thus the stochastic gradient j in $[\![1,n_i]\!]$ on device i in $[\![1,N]\!]$ is $g_j^i(w) = ({X_j^i}^T w - Y_j^i)X_j^i$. On the other hand, the true gradient is $\nabla F_i(w) = \mathbb{E}X^i{X^i}^T(w-w^*)$. Computing the difference, we have for all device i in $[\![1,N]\!]$ and all j in $[\![1,n_i]\!]$:

$$g_j^i(w) - F_i(w) = \underbrace{(X_j^i{X_j^i}^T - \mathbb{E}X^i{X^i}^T)(w-w_*)}_{\text{multiplicative noise equal to 0 in } w_*} + \underbrace{({X_j^i}^T w_* - Y_j^i)}_{\sim\mathcal{N}(0,\lambda^2)} X_j^i \tag{B.1}$$

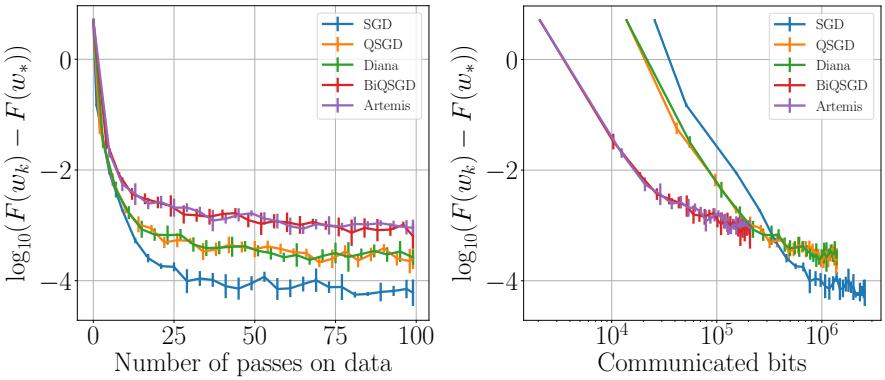
This is why, if we set $\lambda = 0$ and evaluate Equation (B.1) at w_*, we get back Assumption 2.3 with $\sigma_* = 0$, and as a consequence, the stochastic noise at the optimum is removed. Remark that it remains a stochastic gradient descent, and the uniform bound on the gradients noise **is not 0**. We set $\lambda^2 = 0 (\Leftrightarrow \sigma_*^2 = 0)$ in Figure B.3. Otherwise, we set $\lambda^2 = 0.4$.

For non-i.i.d. setting, we generate two different datasets based on a logistic model with two different parameters: $w_1 = (10,10)$ and $w_2 = (10,-10)$. Thus the model is expected to converge to $w_* = (10,0)$. We have two different data distributions $x_1 \sim \mathcal{N}(0,\Sigma_1)$ and $x_2 \sim \mathcal{N}(0,\Sigma_2)$, and for all i in $[\![1,N]\!]$, for all k in $[\![1,n_i]\!]$, $y_k^i = \mathcal{R}\left(\text{Sigm}\left(\left\langle w_{(i \mod 2)+1}, x^k_{(i \mod 2)+1}\right\rangle\right)\right) \in \{-1,+1\}$. That is, half the machines use the first distribution $\mathcal{N}(0,\Sigma_1)$ for inputs and model w_1 and the other half the second distribution for inputs and model w_2. Here, $\mathcal{R}$ is the Rademacher distribution and Sigm is the sigmoid function defined as Sigm: $x \mapsto \frac{e^x}{1+e^x}$. These two distributions are presented on Figure B.1.

B.1.1.1 Least-squares regression

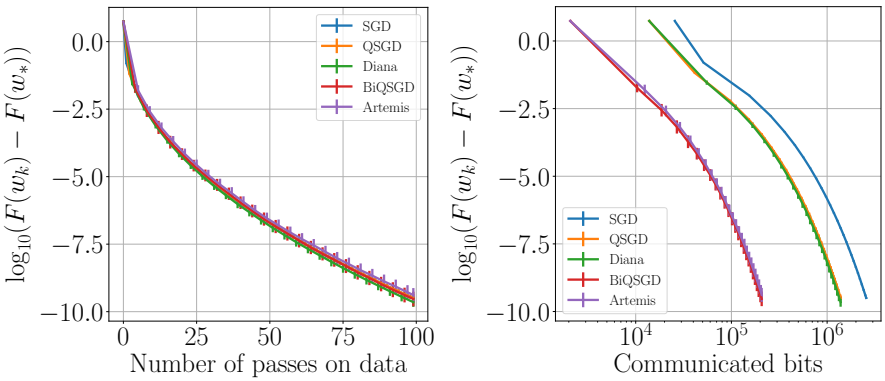
In this section, we present all figures generated using Least-squares regression. Figure B.2 corresponds to Figure 2.1a.

As explained in Chapter 2, in the case of $\sigma_* \neq 0$ (Figure B.2), algorithms using memory (i.e `Diana` and `Artemis`) are not expected to outperform those without (i.e `QSGD` and `Bi-QSGD`). On



(a) LSR: $\sigma_*^2 \neq 0$ (b) X-axis in # bits.

Figure B.2: **Synthetic dataset, Least-Square Regression with noise** ($\sigma_* \neq 0$). In a situation where data is i.i.d., the memory does not present much interest and has no impact on the convergence. Because $\sigma_*^2 \neq 0$, all algorithms saturate; and saturation level is higher for double compression (`Artemis`, `Bi-QSGD`), than for simple compression (`Diana`, `QSGD`) or than for SGD. This corroborates findings in Theorem 2.1 and Theorem 2.3.



(a) LSR: $\sigma_*^2 = 0$ (b) X-axis in # bits.

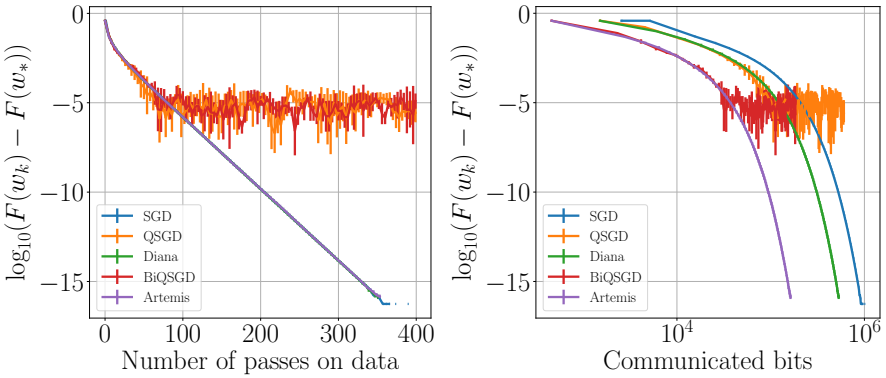
Figure B.3: **Synthetic dataset, Least-Square Regression without noise** ($\sigma_* = 0$). Without surprise, with i.i.d data and $\sigma_* = 0$, the convergence of each algorithm is linear. Thus, in i.i.d. settings, the impact of the memory is negligible, but this will not be the case in the non-i.i.d. settings as underlined by Figure B.4.

the contrary, they saturate at a higher level. However, as soon as the noise at the optimum is 0 (Figure B.3), all algorithms (regardless of memory), converge at a linear rate exactly as classical SGD.

B.1.1.2 Logistic regression

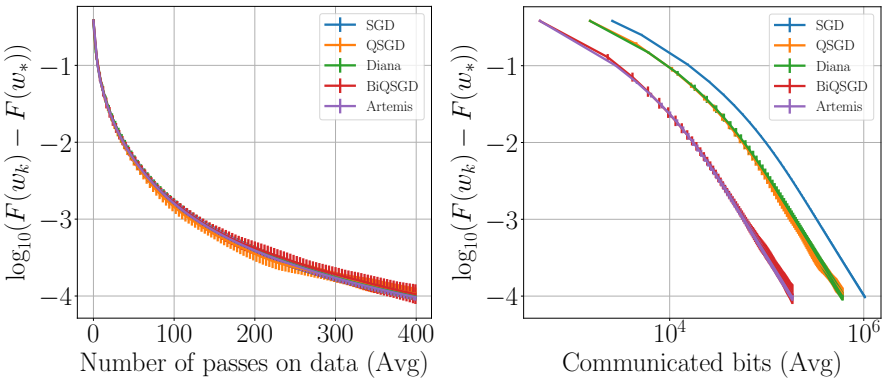
In this section, we present all figures generated using a logistic regression model. Figure B.4 corresponds to Figure 2.1b. Data is non-i.i.d. and we use a full batch gradient descent to get $\sigma_* = 0$ to shed light on the impact of memory on convergence.

Figure B.5 is using same data and configuration as Figure B.4, except that *it is combined with a Polyak-Ruppert averaging*. Note that in the absence of memory the variance increases compared to algorithms using memory. To generate these figures, we didn't take the optimal step-size. But if we took it, the trade-off between variance and bias would be worse and algorithms using memory would outperform those without.



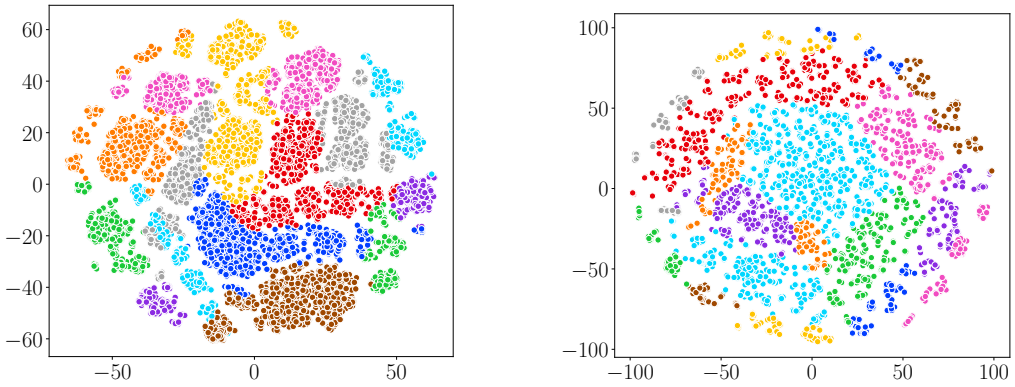
(a) LR: $\sigma_*^2 = 0$ (b) X-axis in # bits.

Figure B.4: **Synthetic dataset, Logistic Regression on non-i.i.d. data** using a full batch gradient descent (to get $\sigma_* = 0$). The benefit of memory is obvious, it makes the algorithm converge linearly, while algorithms without are saturating at a higher level. This stresses the importance of using the memory in non-i.i.d. settings.



(a) LR: $\sigma_*^2 = 0$ (b) X-axis in # bits.

Figure B.5: **Polyak-Ruppert averaging, synthetic dataset.** Logistic regression on non-i.i.d. data using a full batch gradient descent (to get $\sigma_* = 0$) and a Polyak-Ruppert averaging. The convergence is sublinear as predicted by Theorem 2.2 because $\sigma_* = 0$.



(a) Quantum dataset: 20 clusters. Each cluster has between 900 and 10500 points with a median at 2300 points.

(b) Superconduct dataset: 20 cluster. Each cluster has between 250 and 3900 points with a median at 750 points.

Figure B.6: TSNE representations.

Table B.1: Settings of experiments.

Settings	quantum	superconduct
references	Caruana et al. [2004]	Hamidieh [2018]
model	LR	LSR
dimension d	66	82
training dataset size	$50,000$	$21,200$
batch size b	256	64
compression rate s	2^0 (*i.e.* two levels)	
norm quantization	$\|\cdot\|_2$	
momentum m	no momentum	
step-size γ	$1/L$	

B.1.2 Real datasets: *Quantum* and *Superconduct*

In this Section, we present details about experiments conducted on real-life datasets: *superconduct* (from Caruana et al. [2004]) where we use least-squares regression, and *quantum* (from Hamidieh [2018]) with logistic regression. All figures can be found in the notebooks provided on our GitHub repository.

In the following, we present results on superconduct and quantum in the setting of full device participation. Next, we address in Subsection B.1.2.1 the issue of the optimal step-size.

In order to simulate non-i.i.d. data and to make the experiments closer to real-life usage, we split the dataset in heterogeneous groups using a Gaussian mixture clustering on TSNE representations (defined by Maaten and Hinton [2008]). Thus, the data are highly non-i.i.d. *and* unbalanced over devices. We plot on Figure B.6 the TSNE representation of the two real datasets.

There are $N = 20$ devices for *superconduct* and *quantum* datasets. For *superconduct*, there are between 250 and 3900 points by worker, with a median at 750 ; and for *quantum*, there are between 900 and 10500 points, with a median at 2300. On each figure, we indicate which step-size γ has been used.

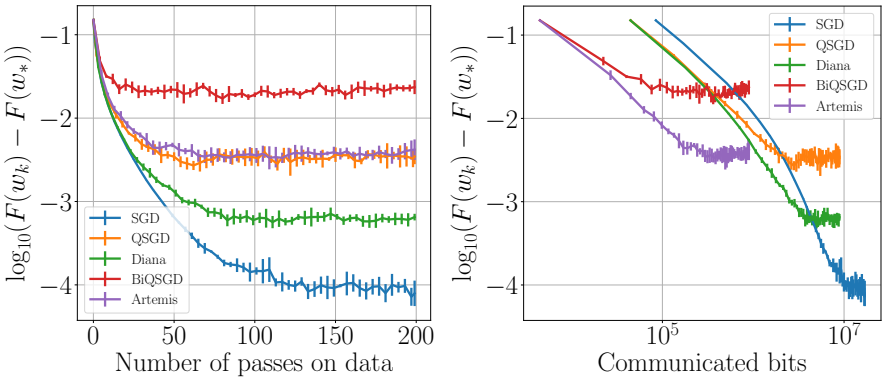
Convex settings are given in Table B.1. Experiments have been performed with 200 epochs in the stochastic regime, and 400 epochs in the full batch regime. We use quantization [defined in Alistarh et al., 2017] with $s = 2^0$ for all experiments.

Figures B.7 to B.10 underline the benefit of using memory in the stochastic and full batch regime for non-i.i.d. datasets. Figures B.7 and B.9 correspond to Figure 2.2. We observe on these figures the benefit of the memory. The level of saturation of algorithms using memory is much lower than those without memory. Additionally, Theorem 2.1 highlights that the level of saturation (see constant E of Table 2.2) is proportional to the level of compression $\omega_{\text{up/dwn}}$. This is indeed observed on Figures B.7 to B.10.

In the case of the *quantum* dataset (see Figure B.7), `Artemis` is not only better than `Bi-QSGD`, but in fact, as good as `QSGD`. That is to say, we achieve to make an algorithm doing bidirectional compression, as good as an algorithm doing unidirectional compression.

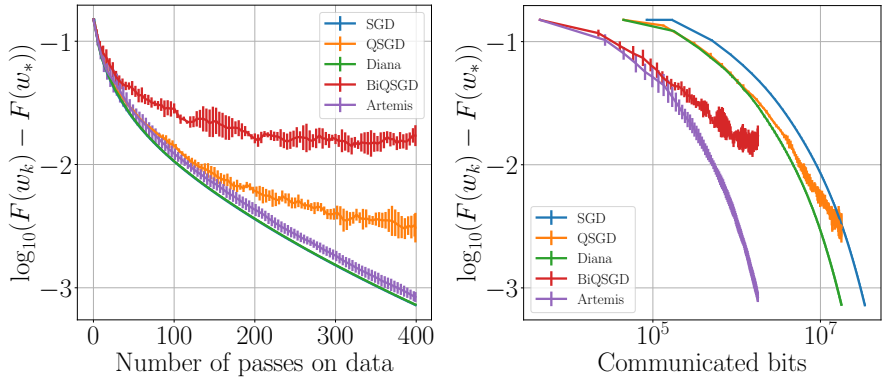
On Figures B.8 and B.10, we run the five algorithms with full gradient descent, resulting in $\sigma_* = 0$. In this case, as the dependency on B^2 is removed, Theorem 2.1 predicts that we must have a linear convergence for algorithms using memory. This is experimentally observed.

Memory trade-off: batch size, noise at the optimum, and heterogeneity. Because the variance of the algorithm (see constant E of Table 2.2) is divided by the batch size b, the choice of this hyperparameter is not without importance. Indeed, reducing the batch size will increase the impact of σ_* on the convergence's rate, while the impact of B^2 will remain constant. Thus, there is



(a) X-axis in # epoch (b) X-axis in # bit

Figure B.7: ***Quantum***. Least-squares regression, $\sigma_* \neq 0$, $\gamma = 1/L$, $b = 256$, non-i.i.d..



(a) X-axis in # epoch (b) X-axis in # bit

Figure B.8: ***Quantum***. Least-squares regression, $\sigma_* = 0$, $\gamma = 1/L$, full gradient descent, non-i.i.d..

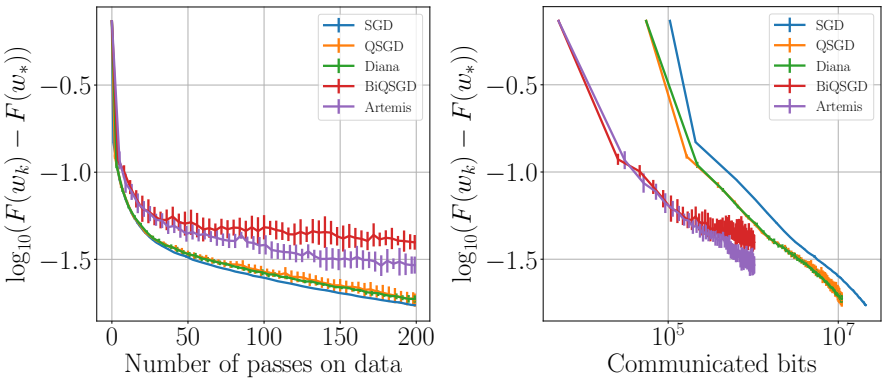
a *trade-off*: if the batch-size is too small, the quantity σ_*/b will become larger than B^2, and the impact of the memory will be hidden by the second term depending on the dataset heterogeneity. This will lead `Artemis`-like algorithms to fail: the memory term is canceled by the high heterogeneity. On the other hand, if the dataset does not present enough heterogeneity, the constant B^2, will be negligible making memory useless, or even penalizing.

B.1.2.1 Optimized step-size

In this section, we want to address the issue of the optimal step-size. On Figure B.11 we plot the minimal loss after 250 iterations for each of the 5 algorithms. We can see that algorithms with memory clearly outperform those without. Then, on Figure B.12 we present the loss of `Artemis` after 250 iterations for various step-size: $\frac{N=20}{2L}$, $\frac{5}{L}$, $\frac{2}{L}$, $\frac{1}{L}$, $\frac{1}{2L}$, $\frac{1}{4L}$, $\frac{1}{8L}$ and $\frac{1}{16L}$. This helps to understand which step-size should be taken to obtain the best accuracy after k in $[\![1, 150]\!]$ iterations. Finally, on Figure B.13, we plot the loss obtained with the optimal step-size γ_{opt} of each algorithms (found with Figure B.11) w.r.t the number of communicated bits.

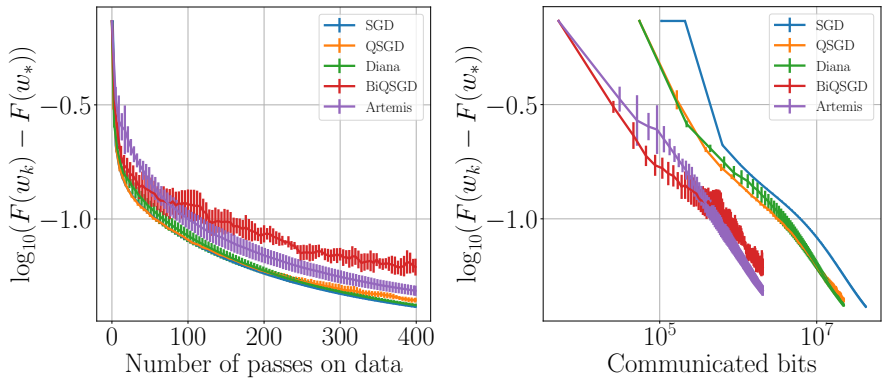
On Figure B.11, it is interesting to note that the memory allows to increase the maximal step-size. So, the optimal step-size is $\gamma_{opt} = \frac{1}{L}$ for `Artemis` , but is $\gamma_{opt} = \frac{1}{2L}$ for `BiQSGD`.

We plot the loss of `Artemis` after 250 iterations for different step-size on Figure B.12. As stressed by Figure B.11, after 250 iterations, the best accuracy for both datasets is indeed obtained with $\gamma_{opt} = \frac{1}{L}$. And we observe that (as for Vanilla `SGD`), the optimal step-size of `Artemis` decreases with the number of iterations (e.g., for *quantum*, it is $1/L$ before 50 iterations and $1/2L$ after). This is consistent with Theorem 2.1.



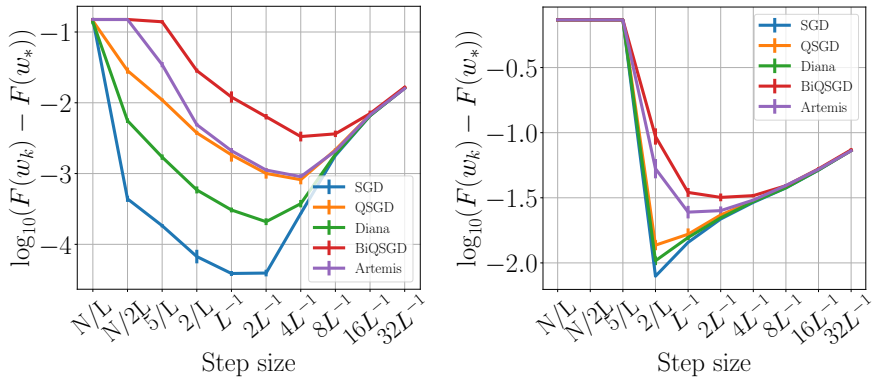
(a) X-axis in # epochs (b) X-axis in # bit

Figure B.9: ***Superconduct***. Least-squares regression, $\sigma_* \neq 0$, $\gamma = 1/L$, $b = 64$, non-i.i.d..



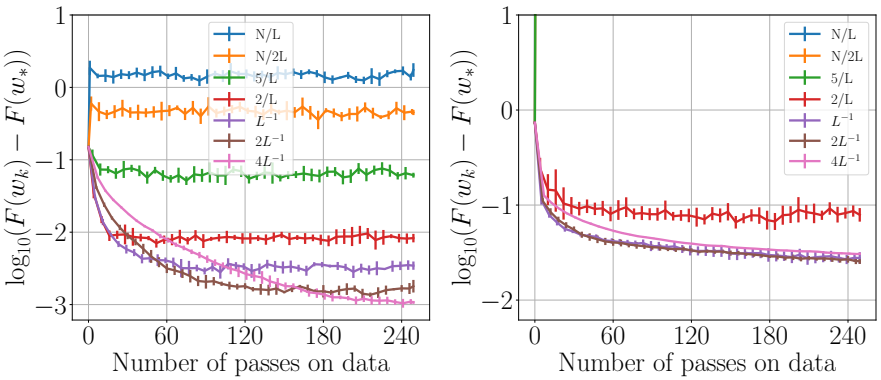
(a) X-axis in # epochs (b) X-axis in # bit

Figure B.10: ***Superconduct***. Least-squares regression, $\sigma_* = 0$, $\gamma = 1/L$, full batch gradiend descent, non-i.i.d..



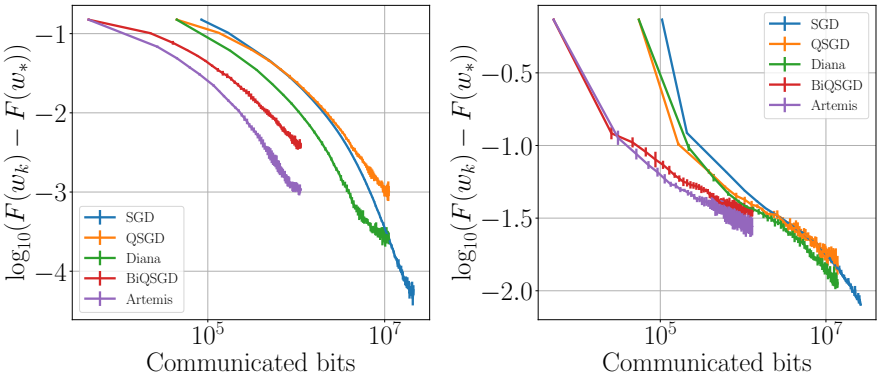
(a) Quantum (b) Superconduct

Figure B.11: **Searching for the optimal step-size γ_{opt} for each algorithm.** X-axis - value on step-size, Y-axis - minimal loss after running 250 iterations



(a) Quantum (b) Superconduct

Figure B.12: Loss w.r.t. step-size γ.



(a) Quantum (b) Superconduct

Figure B.13: **Optimal step-size for each of the algorithms.** X-axis in # bits.

Figure B.13 plots the loss of each algorithm obtained with its optimal step-size γ i.e. the step-size that attains the lowest error after 150 iterations. For instance $\gamma = \frac{1}{L}$ for `Artemis`, but $\gamma = \frac{2}{L}$ for `SGD`. For both *superconduct* and *quantum* datasets, taking the optimal step-size leads `Artemis` to superior performance than other variants w.r.t. both accuracy and number of bits.

In conclusion of this subsection, Figures B.11 to B.13 allow to conclude on the significant impact of memory in a non-i.i.d. settings, and to claim that bidirectional compression with memory is by far superior (up to a threshold) to the four other algorithm: `SGD, QSGD, Diana` and `BiQSGD`.

B.1.3 CPU usage and carbon footprint

As part as a community effort to report the amount of experiments that were performed, we estimated that overall our experiments ran for 220 to 270 hours end to end. We used an Intel(R) Xeon(R) CPU E5-2667 processor with 16 cores.

The carbon emissions caused by this work were subsequently evaluated with `Green Algorithm` built by Lannelongue et al. [2021]. It estimates our computations to generate 30 to 35 kg of CO2, requiring 100 to 125 kWh. To compare, it corresponds to about 160 to 200km by car. This is a relatively moderate impact, matching the goal to keep the experiments for an illustrative purpose.

B.2 Filtrations

In this section, we provide some explanations about filtrations - especially a rigorous definition - and how it is used in the proofs of Theorems 2.1 to 2.3. We recall that we denote by ω_{up} and ω_{dwn} the

$$w_{k-1} \xrightarrow{\xi_k^i} g_k^i \xrightarrow{\epsilon_k^i} \widehat{g}_k^i \longrightarrow \widehat{g}_k = \sum_{i=1}^N \widehat{g}_k^i \xrightarrow{\epsilon_k} \Omega_k = \mathcal{C}\left(\widehat{g}_k\right)$$

Figure B.14: The sequence of successive noises in the algorithm.

variance factors for respectively uplink and downlink compression.

Let a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ with Ω a sample space, $\mathcal{A}$ a σ-algebra, and $\mathbb{P}$ a probability measure. We recall that the σ-algebra generated by a random variable $X : \Omega \to \mathbb{R}^m$ is

$$\sigma(X) = \{X^{-1}(A) : A \in \mathcal{B}(\mathbb{R}^m)\},$$

where $\mathcal{B}(\mathbb{R}^m)$ is the Borel set of $\mathbb{R}^m$.

Furthermore, we recall that a filtration of $(\Omega, \mathcal{A}, \mathbb{P})$ is defined as an increasing sequence $(\mathcal{F}_n)_{n\in\mathbb{N}}$ of σ-algebras:

$$\mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots \subset \mathcal{F}.$$

Randomness in our algorithm comes from three sources, therefore, we define three sequences of i.i.d. zero-centered random fields $(\xi_k^i)_{k\in\mathbb{N},i\in\{1,\dots,N\}}$, $(\epsilon_k^i)_{k\in\mathbb{N},i\in\{1,\dots,N\}}$, $(\epsilon_k)_{k\in\mathbb{N}}$.

1. Stochastic gradients. It corresponds to the noise associated with the computation of the stochastic gradient on device i at epoch k. We have:

$$\forall k \in \mathbb{N}^*,\ \forall i \in [\![0, ..., N]\!], \quad g_k^i = \nabla F_i(w_{k-1}) + \xi_k^i(w_{k-1}).$$

2. Uplink compression: this noise corresponds to the uplink compression when local gradients are compressed. Let $k \in \mathbb{N}$ and $i \in [\![0, ..., N]\!]$, suppose, we want to compress $\Delta_k^i \in \mathbb{R}^d$, then:

$$\forall k \in \mathbb{N}^*,\ \forall i \in [\![0, ..., N]\!], \quad \widehat{\Delta}_k^i = \Delta_k^i + \epsilon_k^i(\Delta_k^i) \iff \widehat{g}_k^i = g_k^i + \epsilon_k^i(\Delta_k^i).$$

3. Downlink compression. This noise corresponds to the downlink compression when the global model parameter is compressed. Let $k \in \mathbb{N}$, suppose we want to compress $\widehat{g}_k \in \mathbb{R}^d$, then:

$$\forall k \in \mathbb{N}^*, \quad \Omega_k = \mathcal{C}_s(\widehat{g}_k) = \widehat{g}_k + \epsilon_k(\widehat{g}_k).$$

This "succession of noises" in the algorithm is illustrated in Figure B.14. In order to handle these three sources of randomness, we define three sequences of nested σ-algebras.

Definition B.1. *We note $(\mathcal{F}_k)_{k\in\mathbb{N}}$ the filtration associated to the stochastic gradient computation noise, $(\mathcal{G}_k)_{k\in\mathbb{N}}$ the filtration associated to the uplink compression noise and $(\mathcal{H}_k)_{k\in\mathbb{N}}$ the filtration associated to the downlink compression noise. For $k \in \mathbb{N}^*$, we define:*

$$\begin{aligned}
\mathcal{F}_k &= \sigma\left(\Gamma_{k-1}, (\xi_k^i)_{i=1}^N\right) \\
\mathcal{G}_k &= \sigma\left(\Gamma_{k-1}, (\xi_k^i)_{i=1}^N, (\epsilon_k^i)_{i=1}^N\right) \\
\mathcal{H}_k &= \sigma\left(\Gamma_{k-1}, (\xi_k^i)_{i=1}^N, (\epsilon_k^i)_{i=1}^N, \epsilon_k\right)
\end{aligned}$$

with $\Gamma_k = \{(\xi_t^i)_{i\in[\![1,N]\!]}, (\epsilon_t^i)_{i\in[\![1,N]\!]}, \epsilon_t\}_{t\in[\![1,k]\!]}$ and $\Gamma_0 = \{\varnothing\}$.

We can make the following observations for all $k \geq 1$:

- From these three definitions, it follows that our sequences are nested.

$$\mathcal{F}_1 \subset \mathcal{G}_1 \subset \mathcal{H}_1 \subset \mathcal{F}_2 \subset \cdots \subset \mathcal{H}_k.$$

- w_{k-1} is $\mathcal{H}_{k-1}$-measurable.
- g_k is $\mathcal{F}_k$-measurable.
- $\widehat{g}_k$ is $\mathcal{G}_k$-measurable.

As a consequence, we have Propositions B.1 to B.5. Below Proposition B.1 gives the expectation over stochastic gradients conditionally to σ-algebras $\mathcal{H}_{k-1}$ and $\mathcal{F}_k$.

Proposition B.1 (Stochastic Expectation). *Let $k \in \mathbb{N}^*$ and $i \in [\![1, N]\!]$. Then on each local device $i \in [\![1, N]\!]$ we have almost surely (a.s.) $\mathbb{E}[g_k^i \mid \mathcal{F}_k] = g_k^i$ and $\mathbb{E}[g_k^i \mid \mathcal{H}_{k-1}] = \nabla F_i(w_{k-1})$.*

Proposition B.2 gives expectation of uplink compression (information sent from remote devices to central server) conditionally to σ-algebras $\mathcal{F}_k$ and $\mathcal{G}_k$.

Proposition B.2 (Uplink Compression Expectation). *Let $k \in \mathbb{N}^*$ and $i \in [\![1, N]\!]$. Recall that $\widehat{g}_k^i = g_k^i + \epsilon_k^i$, then on each local device $i \in [\![1, N]\!]$, we have a.s. $\mathbb{E}[\widehat{g}_k^i \mid \mathcal{G}_k] = \widehat{g}_k^i$ and $\mathbb{E}[\widehat{g}_k^i \mid \mathcal{F}_k] = g_k^i$.*

From Assumption 2.5, it follows that variance over uplink compression can be bounded as expressed in Proposition B.3.

Proposition B.3 (Uplink Compression Variance). *Let $k \in \mathbb{N}^*$ and $i \in [\![1, N]\!]$. Recall that $\Delta_k^i = g_k^i + h_{k-1}^i$, using Assumption 2.5 following hold a.s.:*

$$\mathbb{E}\left[\|\widehat{\Delta}_k^i - \Delta_k^i\|^2 \;\middle|\; \mathcal{F}_k\right] \leq \omega_{\text{up}} \|\Delta_k^i\|^2 \tag{B.2}$$

$$(\iff \quad \mathbb{E}\left[\|\widehat{g}_k^i - g_k^i\|^2 \;\middle|\; \mathcal{F}_k\right] \leq \omega_{\text{up}} \|g_k^i\|^2 \text{ when no memory }). \tag{B.3}$$

Concerning downlink compression (information sent from central server to each node), Proposition B.4 gives its expectation w.r.t σ-algebras $\mathcal{G}_k$ and $\mathcal{H}_k$.

Proposition B.4 (Downlink Compression Expectation). *Let $k \in \mathbb{N}^*$, recall that $\Omega_k = \mathcal{C}_{\text{dwn}}(\widehat{g}_k) = \widehat{g}_k + \epsilon_k$, then a.s. $\mathbb{E}[\Omega_k \mid \mathcal{H}_k] = \Omega_k$ and $\mathbb{E}[\Omega_k \mid \mathcal{G}_k] = \widehat{g}_k$.*

The next proposition states that downlink compression can be bounded as for Proposition B.3.

Proposition B.5 (Downlink Compression Variance). *Let $k \in \mathbb{N}$, using Assumption 2.5, we have a.s. $\mathbb{E}\left[\|\Omega_k - \widehat{g}_k\|^2 \;\middle|\; \mathcal{G}_k\right] \leq \omega_{\text{dwn}} \|\widehat{g}_k\|^2$.*

B.3 Technical results

In this section, we introduce a few technical lemmas that will be used in the proofs of Theorems B.1 to B.3. We first present lemmas common to the proofs with/without memory and which are needed to prove the contraction of the Lyapunov function. Then, in respectively Subsections B.3.1 and B.3.2, we give lemmas adapted to the cases without and with memory.

The first lemma is very simple and straightforward from the definition of Δ_k^i. We remind that Δ_k^i is the difference between the computed gradient and the memory hold on device i. It corresponds to the information which will be compressed and sent from device i to the central server.

Lemma B.1 (Bounding the compressed term). *The squared norm of $(\Delta_k^i)_{k \in \mathbb{N}^*, i \in \{1,\dots,N\}}$, the term sent by each node to the central server, can be bounded as follows:*

$$\forall k \in \mathbb{N}^*, \forall i \in [\![1, N]\!], \quad \left\|\Delta_k^i\right\|^2 \leq 2\left(\left\|g_k^i - h_*^i\right\|^2 + \left\|h_{k-1}^i - h_*^i\right\|^2\right).$$

Proof

Let $k \in \mathbb{N}$ and $i \in \{1, \dots, N\}$, we have by definition:

$$\left\|\Delta_k^i\right\|^2 = \left\|g_k^i - h_{k-1}^i\right\|^2 = \left\|(g_k^i - h_*^i) + (h_*^i - h_{k-1}^i)\right\|^2 .$$

Applying Inequality 1 gives the expected result. ■

Below, we show up a recursion over the memory term h_{k-1}^i involving the stochastic gradients. This recursion will be used in Lemma B.8. This recursion has been first shed into light by Mishchenko et al. [2019].

Lemma B.2 (Expectation of memory term)**.** *The memory term h_k^i can be expressed using a recursion involving the stochastic gradient g_k^i:*

$$\forall k \in \mathbb{N}^* , \forall i \in [\![1, N]\!] , \quad \mathbb{E}\left[h_k^i \mid \mathcal{F}_k\right] = (1 - \alpha_{\text{up}})h_{k-1}^i + \alpha_{\text{up}} g_k^i .$$

Proof Let $k \in \mathbb{N}$ and $i \in \{1, \dots, N\}$. We just need to decompose h_k^i using its definition:

$$h_k^i = h_{k-1}^i + \alpha_{\text{up}}\widehat{\Delta}_k^i = h_{k-1}^i + \alpha_{\text{up}}(\widehat{g}_k^i - h_{k-1}^i) = (1 - \alpha_{\text{up}})h_{k-1}^i + \alpha_{\text{up}}\widehat{g}_k^i ,$$

and considering that $\mathbb{E}\left[\widehat{g}_k^i \mid \mathcal{F}_k\right] = g_k^i$ (Proposition B.2), the proof is completed. ■

In Lemma B.3, we rewrite $\|g_k\|^2$ and $\left\|g_k - h_*^i\right\|^2$ to make appears:

1. the noise over stochasticity,
2. $\|g_k - g_{k,*}\|^2$ which is the term on which will later be applied cocoercivity (see Assumption 2.2).

Lemma B.3 is required to correctly apply cocoercivity in Lemma B.9.

Lemma B.3 (Before using co-coercivity)**.** *Let $k \in [\![0, K]\!]$ and $i \in [\![1, N]\!]$. The noise on the stochastic gradients as defined in Assumptions 2.3 and 2.4 can be controlled as following:*

$$\frac{1}{N}\sum_{i=1}^N \mathbb{E}\left[\left\|g_k^i\right\|^2 \mid \mathcal{H}_{k-1}\right] \leq \frac{2}{N}\sum_{i=1}^N \left(\mathbb{E}\left[\left\|g_k^i - g_{k,*}^i\right\|^2 \mid \mathcal{H}_{k-1}\right] + (\frac{\sigma_*^2}{b} + B^2)\right) , \tag{B.4}$$

$$\frac{1}{N}\sum_{i=1}^N \mathbb{E}\left[\left\|g_k^i - h_*^i\right\|^2 \mid \mathcal{H}_{k-1}\right] \leq \frac{2}{N}\sum_{i=1}^N \left(\mathbb{E}\left[\left\|g_k^i - g_{k,*}^i\right\|^2 \mid \mathcal{H}_{k-1}\right] + \frac{\sigma_*^2}{b}\right) . \tag{B.5}$$

Proof Let $k \in \mathbb{N}$ and i in $\{1, \dots, N\}$. We obtain Equation (B.4) using Inequality 1:

$$\left\|g_k^i\right\|^2 = \left\|g_k^i - g_{k,*}^i + g_{k,*}^i\right\|^2 \leq 2\left(\left\|g_k^i - g_{k,*}^i\right\|^2 + \left\|g_{k,*}^i\right\|^2\right) .$$

Taking expectation with regards to filtration $\mathcal{H}_{k-1}$ and using Assumptions 2.3 and 2.4 gives the first result.

For Equation (B.5), we use again Inequality 1 and we write (by definition, $h_*^i = \nabla F_i(w_*)$):

$$\left\|g_k^i - h_*^i\right\|^2 = \left\|(g_k^i - g_{k,*}^i) + (g_{k,*}^i - \nabla F_i(w_*))\right\|^2 \leq 2(\left\|g_k^i - g_{k,*}^i\right\|^2 + \left\|g_{k,*}^i - \nabla F_i(w_*)\right\|^2) .$$

Taking expectation, we have:

$$\begin{aligned}\mathbb{E}\left[\left\|g_k^i - h_*^i\right\|^2 \mid \mathcal{H}_{k-1}\right] &\leq 2\left(\mathbb{E}\left[\left\|g_k^i - g_{k,*}^i\right\|^2 \mid \mathcal{H}_{k-1}\right] + \mathbb{E}\left[\left\|g_{k,*}^i - \nabla F_i(w_*)\right\|^2 \mid \mathcal{H}_{k-1}\right]\right) \\ &\leq 2\left(\mathbb{E}\left[\left\|g_k^i - g_{k,*}^i\right\|^2 \mid \mathcal{H}_{k-1}\right] + \frac{\sigma_*^2}{b}\right) \quad \text{using Assumption 2.3.}\end{aligned}$$

■

Demonstrating that the Lyapunov function is a contraction requires to bound $\|g_k\|^2$ which needs to control each term $(\|g_k^i\|^2)_{i=1}^N$ of the sum. This leads to invoke smoothness of F (consequence of Assumption 2.2).

Lemma B.4. *Regardless if we use memory, we have the following bound on the squared norm of the gradient, for all k in $\mathbb{N}^*$:*

$$\mathbb{E}\left[\|g_k\|^2 \;\middle|\; \mathcal{H}_{k-1}\right] \leq \frac{1}{N^2}\sum_{i=1}^N \mathbb{E}\left[\left\|g_k^i - h_*^i\right\|^2 \;\middle|\; \mathcal{H}_{k-1}\right] + L\left\langle \nabla F(w_{k-1}), w_{k-1} - w_*\right\rangle .$$

Proof

Let $k \in \mathbb{N}^*$, taking expectation w.r.t the σ-algebra $\mathcal{H}_{k-1}$:

$$\mathbb{E}\left[\|g_k\|^2 \;\middle|\; \mathcal{H}_{k-1}\right] = \mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^N g_k^i - \nabla F_i(w_{k-1}) + \frac{1}{N}\sum_{i=1}^N \nabla F_i(w_{k-1})\right\|^2 \;\middle|\; \mathcal{H}_{k-1}\right] .$$

Decomposing the squared norm:

$$\begin{aligned}\mathbb{E}\left[\|g_k\|^2 \;\middle|\; \mathcal{H}_{k-1}\right] = \mathbb{E}&\left[\left\|\frac{1}{N}\sum_{i=1}^N g_k^i - \nabla F_i(w_{k-1})\right\|^2 \;\middle|\; \mathcal{H}_{k-1}\right] \\ &+ 2\mathbb{E}\left[\left\langle \frac{1}{N}\sum_{i=1}^N g_k^i - \nabla F_i(w_{k-1}), \frac{1}{N}\sum_{i=1}^N \nabla F_j(w_{k-1})\right\rangle \;\middle|\; \mathcal{H}_{k-1}\right] \\ &+ \mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^N \nabla F_i(w_{k-1})\right\|^2 \;\middle|\; \mathcal{H}_{k-1}\right] .\end{aligned}$$

Moreover, $\forall i,j \in \{1,\dots,N\}^2, \mathbb{E}\left[\left\langle g_k^i - \nabla F_i(w_{k-1}), \nabla F_j(w_{k-1})\right\rangle \mid \mathcal{H}_{k-1}\right] = 0$ and $\nabla F(w_{k-1})$ is $\mathcal{H}_{k-1}$-measurable, hence:

$$\mathbb{E}\left[\|g_k\|^2 \;\middle|\; \mathcal{H}_{k-1}\right] \leq \mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^N g_k^i - \nabla F_i(w_{k-1})\right\|^2 \;\middle|\; \mathcal{H}_{k-1}\right] + \|\nabla F(w_{k-1})\|^2 . \tag{B.6}$$

To compute $\|\nabla F(w_{k-1})\|^2$, we apply cocoercivity (Assumption 2.2):

$$\|\nabla F(w_{k-1})\|^2 \leq L\left\langle \nabla F(w_{k-1}), w_{k-1} - w_*\right\rangle .$$

We note $\square_k = \left\|\frac{1}{N}\sum_{i=1}^N g_k^i - \nabla F_i(w_{k-1})\right\|^2$, then expending the squared norm:

$$\begin{aligned}\mathbb{E}\left[\square_k \mid \mathcal{H}_{k-1}\right] &= \frac{1}{N^2}\sum_{i=1}^N \mathbb{E}\left[\left\|g_k^i - \nabla F_i(w_{k-1})\right\|^2 \;\middle|\; \mathcal{H}_{k-1}\right] \\ &\quad + \frac{1}{N^2}\sum_{i,j\in\{1,\dots,N\}/i\neq j} \underbrace{\mathbb{E}\left[\left\langle g_k^i - \nabla F_i(w_{k-1}), g_k^j - \nabla F_j(w_{k-1})\right\rangle \;\middle|\; \mathcal{H}_{k-1}\right]}_{=0 \text{ by independence of } (g_k^i)_{i=0}^N} \\ &= \frac{1}{N^2}\sum_{i=1}^N \mathbb{E}\left[\left\|(g_k^i - \nabla F_i(w_*)) + (\nabla F_i(w_*) - \nabla F_i(w_{k-1}))\right\|^2 \;\middle|\; \mathcal{H}_{k-1}\right] .\end{aligned}$$

Developing the squared norm a second time:

$$\begin{aligned}
\mathbb{E}\left[\square_k \mid \mathcal{H}_{k-1}\right] &= \frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}\left[\left\|g_k^i - \nabla F_i(w_*)\right\|^2 \,\middle|\, \mathcal{H}_{k-1}\right] \\
&\quad + \frac{2}{N^2}\sum_{i=1}^{N}\mathbb{E}\left[\left\langle g_k^i - \nabla F_i(w_*), \nabla F_i(w_*) - \nabla F_i(w_{k-1})\right\rangle \,\middle|\, \mathcal{H}_{k-1}\right] \\
&\quad + \frac{1}{N^2}\sum_{i=1}^{N}\left\|\nabla F_i(w_{k-1}) - \nabla F_i(w_*)\right\|^2 \\
&= \frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}\left[\left\|g_k^i - \nabla F_i(w_*)\right\|^2 \,\middle|\, \mathcal{H}_{k-1}\right] - \frac{1}{N^2}\sum_{i=1}^{N}\left\|\nabla F_i(w_{k-1}) - \nabla F_i(w_*)\right\|^2 \\
&\leq \frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}\left[\left\|g_k^i - \nabla F_i(w_*)\right\|^2 \,\middle|\, \mathcal{H}_{k-1}\right] .
\end{aligned}$$

Recall that we note $h_*^i = \nabla F_i(w_*)$, returning to Equation (B.6), we have:

$$\mathbb{E}\left[\left\|g_k\right\|^2 \,\middle|\, \mathcal{H}_{k-1}\right] \leq \frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}\left[\left\|g_k^i - h_*^i\right\|^2 \,\middle|\, \mathcal{H}_{k-1}\right] + L\left\langle \nabla F(w_{k-1}), w_{k-1} - w_*\right\rangle ,$$

which allows to conclude. ■

B.3.1 Lemmas for the case without memory

In this subsection, we give lemmas that are used only to demonstrate Theorem B.1 (i.e. without memory).

Lemma B.5 is used to remove the uplink compression noise.

Lemma B.5 (Expectation of the squared norm of the compressed gradient when no memory)**.** *In the case without memory, we have the following bound on the squared norm of the compressed gradient, for all k in $\mathbb{N}^*$:*

$$\begin{aligned}
\mathbb{E}\left[\left\|\widehat{g}_k\right\|^2 \,\middle|\, \mathcal{H}_{k-1}\right] \leq \frac{\omega_{\mathrm{up}}}{N^2}\sum_{i=0}^{N}\mathbb{E}\left[\left\|g_k^i\right\|^2 \,\middle|\, \mathcal{H}_{k-1}\right] + \frac{1}{N^2}\sum_{i=0}^{N}\mathbb{E}\left[\left\|g_k^i - h_*^i\right\|^2 \,\middle|\, \mathcal{H}_{k-1}\right] \\
+ L\left\langle \nabla F(w_{k-1}), w_{k-1} - w_*\right\rangle .
\end{aligned}$$

Proof Let k in $\mathbb{N}^*$, first, we write as following:

$$\left\|\widehat{g}_k\right\|^2 = \left\|\widehat{g}_k - g_k\right\|^2 + 2\left\langle \widehat{g}_k - g_k, g_k\right\rangle + \left\|g_k\right\|^2 .$$

Taking stochastic expectation (recall that g_k is $\mathcal{F}_k$-measurable and that $\mathcal{H}_{k-1} \subset \mathcal{F}_k$):

$$\begin{aligned}
\mathbb{E}\left[\mathbb{E}\left[\left\|\widehat{g}_k\right\|^2 \,\middle|\, \mathcal{F}_k\right] \,\middle|\, \mathcal{H}_{k-1}\right] &= \mathbb{E}\left[\mathbb{E}\left[\left\|\widehat{g}_k - g_k\right\|^2 \,\middle|\, \mathcal{F}_k\right] \,\middle|\, \mathcal{H}_{k-1}\right] \\
&\quad + 2 \times \mathbb{E}\left[\mathbb{E}\left[\left\langle \widehat{g}_k - g_k, g_k\right\rangle \mid \mathcal{F}_k\right] \mid \mathcal{H}_{k-1}\right] \\
&\quad + \mathbb{E}\left[\left\|g_k\right\|^2 \,\middle|\, \mathcal{H}_{k-1}\right] .
\end{aligned} \tag{B.7}$$

We need to find a bound for each of the terms of above Equation (B.7). The second term is zero in expectation and the last term is handled in Lemma B.4. It follows that we just need to bound $\|\widehat{g}_k - g_k\|^2$:

$$\begin{aligned}\mathbb{E}\left[\|\widehat{g}_k - g_k\|^2 \;\middle|\; \mathcal{F}_k\right] &= \mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^N \widehat{g}_k^i - g_k^i\right\|^2 \;\middle|\; \mathcal{F}_k\right] \\ &= \frac{1}{N^2}\sum_{i=0}^N \mathbb{E}\left[\left\|\widehat{g}_k^i - g_k^i\right\|^2 \;\middle|\; \mathcal{F}_k\right] + \underbrace{\frac{1}{N}\sum_{i\neq j}\mathbb{E}\left[\left\langle \widehat{g}_k^i - g_k^i, \widehat{g}_k^j - g_k^j\right\rangle \;\middle|\; \mathcal{F}_k\right]}_{=0 \text{ because } (\widehat{g}_k^i)_{i=1}^N \text{ are independents}} \\ &= \frac{1}{N^2}\sum_{i=1}^N \mathbb{E}\left[\left\|\widehat{g}_k^i - g_k^i\right\|^2 \;\middle|\; \mathcal{F}_k\right] .\end{aligned}$$

Combining with Proposition B.3, we hold that $\mathbb{E}\left[\|\widehat{g}_k - g_k\|^2 \;\middle|\; \mathcal{F}_k\right] \leq \frac{\omega_{\mathrm{up}}}{N^2}\sum_{i=1}^N \left\|g_k^i\right\|^2$. Furthermore, we have that:

- $\mathbb{E}\left[\langle \widehat{g}_k - g_k, g_k\rangle \mid \mathcal{F}_k\right] = 0$ (Proposition B.2)
- $\mathbb{E}\left[\|g_k\|^2 \;\middle|\; \mathcal{H}_{k-1}\right] \leq \frac{1}{N^2}\sum_{i=0}^N \mathbb{E}\left[\left\|g_k^i - h_*^i\right\|^2 \;\middle|\; \mathcal{H}_{k-1}\right] + L\left\langle \nabla F(w_{k-1}), w_{k-1} - w_*\right\rangle$ (Lemma B.4).

Thus, we obtain from Equation (B.7):

$$\begin{aligned}\mathbb{E}\left[\|\widehat{g}_k\|^2 \;\middle|\; \mathcal{H}_{k-1}\right] \leq \frac{\omega_{\mathrm{up}}}{N^2}\sum_{i=1}^N \mathbb{E}\left[\left\|g_k^i\right\|^2 \;\middle|\; \mathcal{H}_{k-1}\right] + \frac{1}{N^2}\sum_{i=1}^N \mathbb{E}\left[\left\|g_k^i - h_*^i\right\|^2 \;\middle|\; \mathcal{H}_{k-1}\right] \\ + L\left\langle \nabla F(w_{k-1}), w_{k-1} - w_*\right\rangle .\end{aligned}$$

■

Lemma B.6. *In the case without memory, we have the following bound on the squared norm of the local compressed gradient, for all k in $\mathbb{N}^*$, for all i in $[\![1, N]\!]$: $\mathbb{E}[\|\widehat{g}_k^i\|^2 \mid \mathcal{F}_k] \leq (\omega_{\mathrm{up}} + 1)\|g_k^i\|^2$*

Proof Let k in $\mathbb{N}^*$ and i in $[\![1, N]\!]$:

$$\begin{aligned}\mathbb{E}\left[\left\|\widehat{g}_k^i\right\|^2 \;\middle|\; \mathcal{F}_k\right] &= \mathbb{E}\left[\left\|\widehat{g}_k^i - g_k^i + g_k^i\right\|^2 \;\middle|\; \mathcal{F}_k\right] \\ &= \mathbb{E}\left[\left\|\widehat{g}_k^i - g_k^i\right\|^2 \;\middle|\; \mathcal{F}_k\right] + 2\underbrace{\mathbb{E}\left[\left\langle \widehat{g}_k^i - g_k^i, g_k^i\right\rangle \mid \mathcal{F}_k\right]}_{=0} + \mathbb{E}\left[\left\|g_k^i\right\|^2 \;\middle|\; \mathcal{F}_k\right]\end{aligned}$$

We obtain the result because $\left\|g_k^i\right\|^2$ is $\mathcal{F}_{k+1}$-measurable and using Proposition B.5. ■

B.3.2 Lemmas for the case with memory

In this Subsection, we give lemmas that are used only to demonstrate Theorems B.2 and B.3 (i.e. with memory).

In order to derive an upper bound on the squared norm of $\|w_k - w_*\|^2$, for k in $\mathbb{N}^*$, we need to control $\|\widehat{g}_k\|^2$. This term is decomposed as a sum of three terms depending on:

1. the recursion over the memory term (h_{k-1}^i)

2. the difference between the stochastic gradient at the current point and at the optimal point (later controlled by co-coercivity)
3. the noise over stochasticity.

Lemma B.7. *In the case with memory, we have the following upper bound on the squared norm of the compressed gradient, for all k in $\mathbb{N}^*$:*

$$\mathbb{E}\left[\|\widehat{g}_k\|^2 \,\middle|\, \mathcal{H}_{k-1}\right] \leq \frac{2(2\omega_{\rm up}+1)}{N^2}\sum_{i=1}^N \mathbb{E}\left[\left\|g_k^i - g_{k,*}^i\right\|^2 \,\middle|\, \mathcal{H}_{k-1}\right] + \frac{2\omega_{\rm up}}{N^2}\sum_{i=1}^N \mathbb{E}\left[\left\|h_{k-1}^i - h_*^i\right\|^2 \,\middle|\, \mathcal{H}_{k-1}\right] + L\left\langle \nabla F(w_{k-1}), w_{k-1} - w_*\right\rangle + \frac{2(2\omega_{\rm up}+1)\sigma_*}{Nb}\,.$$

Proof

Let k in $\mathbb{N}^*$. We take the expectation w.r.t. the σ-algebra $\mathcal{H}_{k-1}$, with a bias-variance decomposition and we obtain $\mathbb{E}[\|\widehat{g}_k\|^2 \mid \mathcal{H}_{k-1}] = \mathbb{E}[\|g_k\|^2 \mid \mathcal{H}_{k-1}] + \mathbb{E}[\|\widehat{g}_k - g_k\|^2 \mid \mathcal{H}_{k-1}]$. The first term is handled with Lemma B.4:

$$\mathbb{E}\left[\|g_k\|^2 \,\middle|\, \mathcal{H}_{k-1}\right] \leq \frac{1}{N^2}\sum_{i=1}^N \mathbb{E}\left[\left\|g_k^i - h_*^i\right\|^2 \,\middle|\, \mathcal{H}_{k-1}\right] + L\left\langle \nabla F(w_{k-1}), w_{k-1} - w_*\right\rangle\,.$$

Furthermore, by the independence of the "N" compressions:

$$\mathbb{E}\left[\|\widehat{g}_k - g_k\|^2 \,\middle|\, \mathcal{H}_{k-1}\right] = \frac{1}{N^2}\sum_{i=1}^N \mathbb{E}\left[\left\|\widehat{\Delta}_k^i - \Delta_k^i\right\|^2 \,\middle|\, \mathcal{H}_{k-1}\right]\,,$$

because $\mathcal{H}_{k-1} \subset \mathcal{F}_k$, we can use Proposition B.3 to obtain $\mathbb{E}\left[\|\widehat{g}_k - g_k\|^2 \,\middle|\, \mathcal{H}_{k-1}\right] \leq \frac{\omega_{\rm up}}{N^2}\sum_{i=1}^N \left\|\Delta_k^i\right\|^2$ and next with Lemma B.1, we have:

$$\mathbb{E}\left[\|\widehat{g}_k - g_k\|^2 \,\middle|\, \mathcal{H}_{k-1}\right] \leq \frac{2\omega_{\rm up}}{N^2}\sum_{i=1}^N \mathbb{E}\left[\left\|g_k^i - h_*^i\right\|^2 \,\middle|\, \mathcal{H}_{k-1}\right] + \mathbb{E}\left[\left\|h_{k-1}^i - h_*^i\right\|^2 \,\middle|\, \mathcal{H}_{k-1}\right]\,.$$

At the end:

$$\mathbb{E}\left[\|\widehat{g}_k\|^2 \,\middle|\, \mathcal{H}_{k-1}\right] = \frac{2\omega_{\rm up}+1}{N^2}\sum_{i=1}^N \mathbb{E}\left[\left\|g_k^i - h_*^i\right\|^2 \,\middle|\, \mathcal{H}_{k-1}\right] + \frac{2\omega_{\rm up}}{N^2}\sum_{i=1}^N \mathbb{E}\left[\left\|h_{k-1}^i - h_*^i\right\|^2 \,\middle|\, \mathcal{H}_{k-1}\right] + L\left\langle \nabla F(w_{k-1}), w_{k-1} - w_*\right\rangle\,.$$

We can now apply Lemma B.3 to conclude the proof:

$$\mathbb{E}\left[\|\widehat{g}_k\|^2 \,\middle|\, \mathcal{H}_{k-1}\right] \leq \frac{2(2\omega_{\rm up}+1)}{N^2}\sum_{i=1}^N \mathbb{E}\left[\left\|g_k^i - g_{k,*}^i\right\|^2 \,\middle|\, \mathcal{H}_{k-1}\right] + \frac{2\omega_{\rm up}}{N^2}\sum_{i=1}^N \mathbb{E}\left[\left\|h_{k-1}^i - h_*^i\right\|^2 \,\middle|\, \mathcal{H}_{k-1}\right] + L\left\langle \nabla F(w_{k-1}), w_{k-1} - w_*\right\rangle + \frac{2(2\omega_{\rm up}+1)\sigma_*}{Nb}\,.$$

■

To show that the Lyapunov function is a contraction, we need to find a bound for each terms. Bounding $\|w_k - w_*\|^2$, for k in $\mathbb{N}$, flows from update schema (see Equation (2.3)) decomposition. However the memory term $\left\|h_k^i - h_*^i\right\|^2$ involved in the Lyapunov function doesn't show up naturally.

The aim of Lemma B.8 is precisely to provide a recursive bound over the memory term to highlight the contraction. Like Lemma B.2, the following lemma comes from Mishchenko et al. [2019].

Lemma B.8 (Recursive inequalities over memory term)**.** *Let $k \in \mathbb{N}^*$ and let $i \in [\![1, N]\!]$. The memory term used in the uplink broadcasting can be bounded using a recursion:*

$$\begin{aligned}\mathbb{E}\left[\left\|h_k^i - h_*^i\right\|^2 \,\middle|\, \mathcal{H}_{k-1}\right] \leq \left(1 + 2\alpha_{\text{up}}^2\omega_{\text{up}} + 2\alpha_{\text{up}}^2 - 3\alpha_{\text{up}}\right)\left\|h_{k-1}^i - h_*^i\right\|^2 \\ + 2(2\alpha_{\text{up}}^2\omega_{\text{up}} + 2\alpha_{\text{up}}^2 - \alpha_{\text{up}})\mathbb{E}\left[\left\|g_k - g_{k,*}\right\|^2 \,\middle|\, \mathcal{H}_{k-1}\right] \\ + \frac{2\sigma_*^2}{b}\left(2\alpha_{\text{up}}^2(\omega_{\text{up}} + 1) - \alpha_{\text{up}}\right) .\end{aligned}$$

Proof

Let $k \in \mathbb{N}^*$ and let $i \in [\![1, N]\!]$, using Lemma A.2 we have:

$$\mathbb{E}\left[\left\|h_k^i - h_*^i\right\|^2 \,\middle|\, \mathcal{F}_k\right] = \left\|\mathbb{E}\left[h_k^i \mid \mathcal{F}_k\right] - h_*^i\right\|^2 + \mathbb{E}\left[\left\|h_k^i - \mathbb{E}\left[h_k^i \mid \mathcal{F}_k\right]\right\|^2 \,\middle|\, \mathcal{F}_k\right] ,$$

and now with Lemma B.2:

$$\mathbb{E}\left[\left\|h_k^i - h_*^i\right\|^2 \,\middle|\, \mathcal{F}_k\right] = \left\|(1 - \alpha_{\text{up}})h_{k-1}^i + \alpha_{\text{up}}g_k^i - h_*^i\right\|^2 + \mathbb{E}\left[\left\|h_k^i - \mathbb{E}\left[h_k^i \mid \mathcal{F}_k\right]\right\|^2 \,\middle|\, \mathcal{F}_k\right] .$$

Now recall that $h_k^i = h_{k-1}^i + \alpha_{\text{up}}\widehat{\Delta}_k^i$, with $\mathbb{E}[\widehat{\Delta}_k^i \mid \mathcal{F}_k] = \Delta_k^i$ and h_{k-1}^i being $\mathcal{F}_k$-measurable:

$$\mathbb{E}\left[\left\|h_k^i - h_*^i\right\|^2 \,\middle|\, \mathcal{F}_k\right] = \left\|(1 - \alpha_{\text{up}})(h_{k-1}^i - h_*^i) + \alpha_{\text{up}}(g_k^i - h_*^i)\right\|^2 + \alpha_{\text{up}}^2\mathbb{E}\left[\left\|\widehat{\Delta}_k^i - \Delta_k^i\right\|^2 \,\middle|\, \mathcal{F}_k\right] .$$

Using Lemma A.1 of Section A.1 and Proposition B.3:

$$\begin{aligned}\mathbb{E}\left[\left\|h_k^i - h_*^i\right\|^2 \,\middle|\, \mathcal{F}_k\right] \leq (1 - \alpha_{\text{up}})\left\|h_{k-1}^i - h_*^i\right\|^2 + \alpha_{\text{up}}\left\|g_k^i - h_*^i\right\|^2 \\ - \alpha_{\text{up}}(1 - \alpha_{\text{up}})\left\|h_{k-1}^i - g_k^i\right\|^2 + \alpha_{\text{up}}^2\omega_{\text{up}}\left\|\Delta_k^i\right\|^2 .\end{aligned}$$

Because $h_{k-1}^i - g_k^i = \Delta_k^i$:

$$\mathbb{E}\left[\left\|h_k^i - h_*^i\right\|^2 \,\middle|\, \mathcal{F}_k\right] \leq (1 - \alpha_{\text{up}})\left\|h_{k-1}^i - h_*^i\right\|^2 + \alpha_{\text{up}}\left\|g_k^i - h_*^i\right\|^2 + \alpha_{\text{up}}\left(\alpha_{\text{up}}(\omega_{\text{up}} + 1) - 1\right)\left\|\Delta_k^i\right\|^2 ,$$

and using Lemma B.1:

$$\begin{aligned}\mathbb{E}\left[\left\|h_k^i - h_*^i\right\|^2 \,\middle|\, \mathcal{F}_k\right] &\leq (1 - \alpha_{\text{up}})\left\|h_{k-1}^i - h_*^i\right\|^2 + \alpha_{\text{up}}\left\|g_k^i - h_*^i\right\|^2 \\ &\quad + 2\alpha_{\text{up}}\left(\alpha_{\text{up}}(\omega_{\text{up}} + 1) - 1\right)\left(\left\|h_{k-1}^i - h_*^i\right\|^2 + \left\|g_k - h_*^i\right\|^2\right) \\ &\leq \left(1 + 2\alpha_{\text{up}}^2\omega_{\text{up}} + 2\alpha_{\text{up}}^2 - 3\alpha_{\text{up}}\right)\left\|h_{k-1}^i - h_*^i\right\|^2 \\ &\quad + \alpha_{\text{up}}(2\alpha_{\text{up}}\omega_{\text{up}} + 2\alpha_{\text{up}} - 1)\left\|g_k - h_*^i\right\|^2 .\end{aligned}$$

Finally taking expectation w.r.t. the σ-algebra $\mathcal{H}_{k-1}$ ($\mathcal{H}_{k-1} \subset \mathcal{F}_k$) and using Equation (B.5) of Lemma B.3, we have:

$$\begin{aligned}\mathbb{E}\left[\left\|h_k^i - h_*^i\right\|^2 \;\middle|\; \mathcal{H}_{k-1}\right] \leq (1 + 2\alpha_{\text{up}}^2\omega_{\text{up}} + 2\alpha_{\text{up}}^2 - 3\alpha_{\text{up}})\left\|h_{k-1}^i - h_*^i\right\|^2 \\ + 2(2\alpha_{\text{up}}^2\omega_{\text{up}} + 2\alpha_{\text{up}}^2 - \alpha_{\text{up}})\mathbb{E}\left[\left\|g_k - g_{k,*}\right\|^2 \;\middle|\; \mathcal{H}_{k-1}\right] \\ + \frac{2\sigma_*^2}{b}\left(2\alpha_{\text{up}}^2(\omega_{\text{up}} + 1) - \alpha_{\text{up}}\right) ,\end{aligned}$$

which concludes the proof.

■

After successfully invoking all previous lemmas, we will finally be able to use co-coercivity. Lemma B.9 shows how Assumption 2.2 is used to do it. After this stage, proof will be continued by applying strong-convexity of F.

Lemma B.9 (Applying co-coercivity)**.** *This lemma shows how to apply co-coercivity on stochastic gradients. For all k in $\mathbb{N}^*$, we have* $\frac{1}{N}\sum_{i=1}^N \mathbb{E}\left[\left\|g_k^i - g_{k,*}\right\|^2 \;\middle|\; \mathcal{H}_{k-1}\right] \leq L\left\langle \nabla F(w_{k-1}), w_{k-1} - w_*\right\rangle$.

Proof Let $k \in \mathbb{N}^*$, using Assumption 2.2, we have:

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^N \mathbb{E}\left[\left\|g_k^i - g_{k,*}^i\right\|^2 \;\middle|\; \mathcal{H}_{k-1}\right] &\leq \frac{1}{N}\sum_{i=1}^N L\left\langle \mathbb{E}\left[g_k^i - g_{k,*}^i \;\middle|\; \mathcal{H}_{k-1}\right], w_{k-1} - w_*\right\rangle \\ &\leq L\left\langle \frac{1}{N}\sum_{i=1}^N \nabla F_i(w_{k-1}) - \nabla F_i(w_*), w_{k-1} - w_*\right\rangle .\end{aligned}$$

■

B.4 Proofs of Theorems

In this Section, we give demonstrations of all our theorems, that is to say, first the proofs of Theorems B.1 and B.2 from which flow Theorem 2.1. Their demonstration sketch is drawn from Mishchenko et al. [2019]. And in a second time, we give a complete demonstration of theorems stated in Chapter 2: Theorems 2.2 and 2.3.

For the sake of demonstration, we define a Lyapunov function V_k [as in Mishchenko et al., 2019, Liu et al., 2020], for k in $\mathbb{N}$:

$$V_k = \|w_k - w_*\|^2 + 2\gamma^2 C\frac{1}{N}\sum_{i=1}^N \left\|h_k^i - h_*^i\right\|^2 ,$$

with C in $\mathbb{R}_+^*$. The Lyapunov function is defined by combining two terms.

1. The distance from parameter w_k to optimal parameter w_*.
2. The memory term, the distance between the next element prediction h_k^i and the true gradient $h_*^i = \nabla F_i(w_*)$.

The aim is to proof that this function is a $(1 - \gamma\mu)$ contraction for each variant of `Artemis`. To show that it's a contraction, we need three stages:

1. we develop the update schema defined in Equation (2.3) to get a first bound on $\|w_k - w_*\|^2$,
2. we find a recurrence over the memory term $\left\|h_k^i - h_*^i\right\|^2$,
3. and finally we combines the two equations to obtain the expected contraction using co-coercivity and strong-convexity.

B.4.1 Proof of main Theorem for Artemis - variant without memory

Theorem B.1 (Unidirectional or bidirectional compression without memory)**.** *Considering that Assumptions 2.1 to 2.5 hold. Taking γ such that*

$$\gamma \leq \frac{N}{L(\omega_{\mathrm{dwn}}+1)\left(N+2(\omega_{\mathrm{up}}+1)\right)},$$

then running **Artemis** *with $\alpha_{\mathrm{up}}=0$ (i.e without memory), we have for all k in $\mathbb{N}^*$:*

$$\mathbb{E}\left\|w_k-w_*\right\|^2 \leq (1-\gamma\mu)^k\left\|w_0-w_*\right\|^2+2\gamma\frac{E}{\mu N},$$

with $E=(\omega_{\mathrm{dwn}}+1)\left(\frac{(\omega_{\mathrm{up}}+1)\sigma_*^2}{b}+\omega_{\mathrm{up}}B^2\right)$. In the case of unidirectional compression (resp. no compression), we have $\omega_{\mathrm{dwn}}=0$ (resp. $\omega_{\mathrm{up/dwn}}=0$).

Proof

In the case of the variant of Artemis with $\alpha_{\mathrm{up}}=0$, we don't have any memory term, thus $C=0$ and we don't need to use the Lyapunov function.

Let k in $\mathbb{N}^*$, we start by writing that by definition of Equation (2.3):

$$\begin{aligned}\left\|w_k-w_*\right\|^2 &= \left\|w_{k-1}-\gamma\Omega_k-w_*\right\|^2\\ &= \left\|w_{k-1}-w_*\right\|^2-2\gamma\left\langle\Omega_k,w_{k-1}-w_*\right\rangle+\gamma^2\left\|\Omega_k\right\|^2,\end{aligned}$$

with $\Omega_k=\mathcal{C}_{\mathrm{dwn}}(\widehat{g}_k)$ and $\widehat{g}_k=\frac{1}{N}\sum_{i=1}^N\widehat{g}_k^i$. First, we have $\mathbb{E}\left[\Omega_k\mid\mathcal{G}_{k-1}\right]=\widehat{g}_k$ (Proposition B.4) secondly considering that $\mathbb{E}[\left\|\Omega_k\right\|^2\mid\mathcal{G}_{k-1}]=\mathbb{V}(\Omega_k)+\left\|\mathbb{E}\left[\Omega_k\mid\mathcal{G}_{k-1}\right]\right\|^2=(\omega_{\mathrm{dwn}}+1)\left\|\widehat{g}_k\right\|^2$ leads to:

$$\mathbb{E}\left[\left\|w_k-w_*\right\|^2\,\middle|\,\mathcal{G}_{k-1}\right]=\mathbb{E}\left[\left\|w_{k-1}-w_*\right\|^2\,\middle|\,\mathcal{G}_{k-1}\right]-2\gamma\left\langle\widehat{g}_k,w_{k-1}-w_*\right\rangle+\gamma^2(\omega_{\mathrm{dwn}}+1)\left\|\widehat{g}_k\right\|^2.$$

Now, we take expectation w.r.t σ-algebra $\mathcal{H}_{k-1}\subset\mathcal{G}_{k-1}$, (with use of Propositions B.1 and B.2, we obtain :

$$\begin{aligned}\mathbb{E}\left[\left\|w_k-w_*\right\|^2\,\middle|\,\mathcal{H}_{k-1}\right]=\mathbb{E}\left[\left\|w_{k-1}-w_*\right\|^2\,\middle|\,\mathcal{H}_{k-1}\right]&-2\gamma\left\langle\nabla F(w_{k-1}),w_{k-1}-w_*\right\rangle\\ &+\gamma^2(\omega_{\mathrm{dwn}}+1)\mathbb{E}\left[\left\|\widehat{g}_k\right\|^2\,\middle|\,\mathcal{H}_{k-1}\right].\end{aligned}\tag{B.8}$$

Lemma B.5 gives:

$$\begin{aligned}\mathbb{E}[\left\|\widehat{g}_k\right\|^2\mid\mathcal{H}_{k-1}]\leq\frac{\omega_{\mathrm{up}}}{N^2}\sum_{i=0}^N\mathbb{E}[\left\|g_k^i\right\|^2\mid\mathcal{H}_{k-1}]+\frac{1}{N}\sum_{i=0}^N\mathbb{E}[\left\|g_k^i-h_*^i\right\|^2\mid\mathcal{H}_{k-1}]\\ +L\left\langle\nabla F(w_{k-1}),w_{k-1}-w_*\right\rangle.\end{aligned}$$

Lets introducing the noise at optimal point w_* with the two equations of Lemma B.3:

$$\begin{aligned}\mathbb{E}\left[\left\|\widehat{g}_k\right\|^2\,\middle|\,\mathcal{H}_{k-1}\right]\leq\frac{\omega_{\mathrm{up}}}{N^2}\sum_{i=1}^N 2\left(\mathbb{E}\left[\left\|g_k^i-g_{k,*}^i\right\|^2\,\middle|\,\mathcal{H}_{k-1}\right]+(\frac{\sigma_*^2}{b}+B^2)\right)\\ +\frac{1}{N^2}\sum_{i=1}^N 2\left(\mathbb{E}\left[\left\|g_k^i-g_{k,*}^i\right\|^2\,\middle|\,\mathcal{H}_{k-1}\right]+\frac{\sigma_*^2}{b}\right)\\ +L\left\langle\nabla F(w_{k-1}),w_{k-1}-w_*\right\rangle.\end{aligned}$$

Invoking cocoercivity (Assumption 2.2):

$$\begin{aligned}\mathbb{E}\left[\|\widehat{g}_k\|^2 \,\middle|\, \mathcal{H}_{k-1}\right] &\leq \frac{2(\omega_{\rm up}+1)}{N^2}\sum_{i=1}^{N}\mathbb{E}\left[L\left\langle g_k^i - g_{k,*}^i, w_{k-1}-w_*\right\rangle \,\middle|\, \mathcal{H}_{k-1}\right] \\ &\quad + L\left\langle \nabla F(w_{k-1}), w_{k-1}-w_*\right\rangle + \frac{2}{N}\left(\frac{(\omega_{\rm up}+1)\sigma_*^2}{b} + \omega_{\rm up}B^2\right) \\ &\leq \frac{2(\omega_{\rm up}+1)L}{N}\left\langle \nabla F(w_{k-1}), w_{k-1}-w_*\right\rangle \\ &\quad + L\left\langle \nabla F(w_{k-1}), w_{k-1}-w_*\right\rangle + \frac{2}{N}\left(\frac{(\omega_{\rm up}+1)\sigma_*^2}{b} + \omega_{\rm up}B^2\right). \end{aligned} \tag{B.9}$$

Finally, we can inject Equation (B.9) in Equation (B.8) to obtain:

$$\begin{aligned}\mathbb{E}\left[\|w_k - w_*\|^2 \,\middle|\, \mathcal{H}_{k-1}\right] &\leq \|w_{k-1}-w_*\|^2 \\ &\quad - 2\gamma\left(1 - \frac{\gamma L(\omega_{\rm dwn}+1)(\omega_{\rm up}+1)}{N} - \frac{\gamma L(\omega_{\rm dwn}+1)}{2}\right)\left\langle \nabla F(w_{k-1}), w_{k-1}-w_*\right\rangle \\ &\quad + \frac{2\gamma^2(\omega_{\rm dwn}+1)\left(\frac{(\omega_{\rm up}+1)\sigma_*^2}{b} + \omega_{\rm up}B^2\right)}{N}. \end{aligned} \tag{B.10}$$

We note:

1. $\square = 1 - \frac{\gamma L(\omega_{\rm dwn}+1)(\omega_{\rm up}+1)}{N} - \frac{\gamma L(\omega_{\rm dwn}+1)}{2}$
2. $E = (\omega_{\rm dwn}+1)\left(\frac{(\omega_{\rm up}+1)\sigma_*^2}{b} + \omega_{\rm up}B^2\right)$.

We need $\square \geq 0$ in order to further apply strong-convexity. However, in order to later obtain a convergence in $(1-\gamma\mu)$, we will use a stronger condition and, instead, state that we need $\square \geq 1/2$, which is equivalent to:

$$\frac{1}{2} \geq \frac{\gamma L(\omega_{\rm dwn}+1)(\omega_{\rm up}+1)}{N} + \frac{\gamma L(\omega_{\rm dwn}+1)}{2} \iff \quad \gamma \leq \frac{N}{L(\omega_{\rm dwn}+1)\left(N + 2(\omega_{\rm up}+1)\right)},$$

Using strong-convexity of F (Assumption 2.1), we rewrite Equation (B.10) as follows:

$$\mathbb{E}\left[\|w_k - w_*\|^2 \,\middle|\, \mathcal{H}_{k-1}\right] \leq \|w_{k-1}-w_*\|^2 - 2\gamma\mu\square\|w_{k-1}-w_*\|^2 + 2\gamma^2\frac{E}{N}, \text{ equivalent to:}$$

$$\leq (1 - 2\gamma\mu\square)\|w_{k-1}-w_*\|^2 + 2\gamma^2\frac{E}{N}.$$

To guarantee a $(1-\gamma\mu)$ convergence, we need $\square \geq 1/2$, which is already verified, hence taking full expectation, we are allowed to write:

$$\begin{aligned} & \mathbb{E}[\|w_k - w_*\|^2] \leq (1-\gamma\mu)\mathbb{E}[\|w_{k-1}-w_*\|^2] + 2\gamma^2\frac{E}{N} \\ \iff \quad & \mathbb{E}[\|w_k - w_*\|^2] \leq (1-\gamma\mu)^k\|w_0 - w_*\|^2 + 2\gamma^2\frac{E}{N} \times \frac{1-(1-\gamma\mu)^k}{\gamma\mu} \\ \iff \quad & \mathbb{E}[\|w_k - w_*\|^2] \leq (1-\gamma\mu)^k\|w_0 - w_*\|^2 + 2\gamma\frac{E}{\mu N}, \end{aligned}$$

and the proof is complete.

■

B.4.2 Proof of main Theorem for Artemis - variant with memory

Theorem B.2 (Unidirectional or bidirectional compression with memory)**.** *Considering that Assumptions 2.1 to 2.5 hold. We use w_* to indicate the optimal parameter such that $\nabla F(w_*) = 0$, and we note $h_*^i = \nabla F_i(w_*)$. We define the Lyapunov function for any k in $\mathbb{N}$:*

$$V_k = \|w_k - w_*\|^2 + 2\gamma^2 C \frac{1}{N} \sum_{i=1}^{N} \left\|h_k^i - h_*^i\right\|^2 .$$

We defined $C \in \mathbb{R}_+^*$, such that:

$$\frac{\omega_{\text{up}}(\omega_{\text{dwn}}+1)}{\alpha_{\text{up}}\left(3 - 2\alpha_{\text{up}}(\omega_{\text{up}}+1)\right)} \leq C \leq \frac{N - \gamma L(\omega_{\text{dwn}}+1)\left(N + 2(2\omega_{\text{up}}+1)\right)}{4\gamma L \alpha_{\text{up}}\left(2\alpha_{\text{up}}(\omega_{\text{up}}+1) - 1\right)} . \tag{B.11}$$

Then, using **Artemis** *with a memory mechanism ($\alpha_{\text{up}} \neq 0$), the convergence of the algorithm is guaranteed if:*

1. $\dfrac{1}{2(\omega_{\text{up}}+1)} \leq \alpha_{\text{up}} < \min\left(\dfrac{3}{2(\omega_{\text{up}}+1)}, \quad \dfrac{3N - \gamma L(\omega_{\text{dwn}}+1)\left(3N + 8\omega_{\text{up}} + 6\right)}{2(\omega_{\text{up}}+1)(N - \gamma L(\omega_{\text{dwn}}+1)(N+2))}\right)$
2. $$\gamma < \min\left\{ \frac{1}{(\omega_{\text{dwn}}+1)\left(1 + \dfrac{2}{N}\right)L}, \quad \frac{3}{(\omega_{\text{dwn}}+1)\left(3 + \dfrac{8\omega_{\text{up}}+6}{N}\right)L}, \quad \frac{N}{(\omega_{\text{dwn}}+1)\left(N + 2(2\omega_{\text{up}}+1)\right)L} \right\} .$$

And we have a bound for the Lyapunov function:

$$\mathbb{E}V_k \leq (1-\gamma\mu)^k \left(\|w_0 - w_*\|^2 + 2C\gamma^2 B^2\right) + 2\gamma \frac{E}{\mu N} ,$$

with $E = \dfrac{\sigma_*^2}{b}\left((2\omega_{\text{up}}+1)(\omega_{\text{dwn}}+1) + 2C\left(2\alpha_{\text{up}}^2(\omega_{\text{up}}+1) - \alpha_{\text{up}}\right)\right)$. In the case of unidirectional compression (resp. no compression), we have $\omega_{\text{dwn}} = 0$ (resp. $\omega_{\text{up/dwn}} = 0$).

Proof Let $k \in \mathbb{N}^*$, by definition of the update schema (Algorithm 2), we have: $w_k = w_{k-1} - \gamma\Omega_k$, with $\Omega_k = \mathcal{C}_{\text{dwn}}(\widehat{g}_k)$ and $\widehat{g}_k = h_{k-1} + \frac{1}{N}\sum_{i=1}^N \widehat{\Delta}_k^i$, thus $\|w_k - w_*\|^2 = \|w_{k-1} - w_* + \gamma\Omega_k\|^2 = \|w_{k-1} - w_*\| - 2\gamma\left\langle \Omega_k, w_{k-1} - w_*\right\rangle + \gamma^2\left\|\Omega_k\right\|^2$. Taking expectation w.r.t. the σ-algebra $\mathcal{G}_{k-1}$:

$$\mathbb{E}\left[\|w_k - w_*\|^2 \;\middle|\; \mathcal{G}_{k-1}\right] = \mathbb{E}\left[\|w_{k-1} - w_*\|^2 \;\middle|\; \mathcal{G}_{k-1}\right] - 2\gamma\left\langle \widehat{g}_k, w_{k-1} - w_*\right\rangle + \gamma^2(\omega_{\text{dwn}}+1)\left\|\widehat{g}_k\right\|^2 .$$

We take expectation w.r.t σ-algebra $\mathcal{H}_{k-1} \subset \mathcal{G}_{k-1}$ and invoke Lemma B.7:

$$\begin{aligned}
\mathbb{E}\left[\|w_k - w_*\|^2 \;\middle|\; \mathcal{H}_{k-1}\right] \leq \|w_{k-1} - w_*\|^2 &- 2\gamma\mathbb{E}\left[\left\langle \widehat{g}_k, w_{k-1} - w_*\right\rangle \mid \mathcal{H}_{k-1}\right] \\
&+ \frac{2(2\omega_{\text{up}}+1)(\omega_{\text{dwn}}+1)\gamma^2}{N^2} \sum_{i=1}^N \mathbb{E}\left[\left\|g_k^i - g_{k,*}^i\right\|^2 \;\middle|\; \mathcal{H}_{k-1}\right] \\
&+ \frac{2\omega_{\text{up}}(\omega_{\text{dwn}}+1)\gamma^2}{N^2} \sum_{i=1}^N \mathbb{E}\left[\left\|h_{k-1}^i - h_*^i\right\|^2 \;\middle|\; \mathcal{H}_{k-1}\right] \\
&+ \gamma^2(\omega_{\text{dwn}}+1)L\left\langle \nabla F(w_{k-1}), w_{k-1} - w_*\right\rangle \\
&+ \frac{2(2\omega_{\text{up}}+1)(\omega_{\text{dwn}}+1)\gamma^2\sigma_*}{Nb} .
\end{aligned} \tag{B.12}$$

Note that in the case of unidirectional compression, we have $\Omega_k = \widehat{g}_k$, and the steps above are more straightforward. Recall that according to Lemma B.8 (and taking the sum), we have:

$$\begin{aligned}
&\frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}\left[\left\|h_k^i - h_*^i\right\|^2 \;\middle|\; \mathcal{H}_{k-1}\right] \\
&\qquad \leq \left(1 + 2\alpha_{\text{up}}^2\omega_{\text{up}} + 2\alpha_{\text{up}}^2 - 3\alpha_{\text{up}}\right)\frac{1}{N^2}\sum_{i=1}^{N}\left\|h_{k-1}^i - h_*^i\right\|^2 \\
&\qquad + 2(2\alpha_{\text{up}}^2\omega_{\text{up}} + 2\alpha_{\text{up}}^2 - \alpha_{\text{up}})\frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}\left[\left\|g_k^i - g_{k,*}^i\right\|^2 \;\middle|\; \mathcal{H}_{k-1}\right] \\
&\qquad + \frac{2\sigma_*^2}{Nb}\left(2\alpha_{\text{up}}^2(\omega_{\text{up}}+1) - \alpha_{\text{up}}\right) .
\end{aligned} \tag{B.13}$$

With a linear combination (B.12) + $2\gamma^2 C$ (B.13):

$$\begin{aligned}
&\mathbb{E}\left[\left\|w_k - w_*\right\|^2 \;\middle|\; \mathcal{H}_{k-1}\right] + 2\gamma^2 C\frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}\left[\left\|h_k^i - h_*^i\right\|^2 \;\middle|\; \mathcal{H}_{k-1}\right] \\
&\quad \leq \left\|w_{k-1} - w_*\right\|^2 - 2\gamma\mathbb{E}\left[\langle \widehat{g}_k, w_{k-1} - w_*\rangle \mid \mathcal{H}_{k-1}\right] \\
&\qquad + 2\gamma^2\frac{(2\omega_{\text{up}}+1)(\omega_{\text{dwn}}+1) + 2C(2\alpha_{\text{up}}^2\omega_{\text{up}} + 2\alpha_{\text{up}}^2 - \alpha_{\text{up}})}{N^2}\sum_{i=1}^{N}\mathbb{E}\left[\left\|g_k^i - g_{k,*}^i\right\| \mid \mathcal{H}_{k-1}\right] \\
&\qquad + 2\gamma^2 C\left(\frac{\omega_{\text{up}}(\omega_{\text{dwn}}+1)}{C} + 1 + 2\alpha_{\text{up}}^2\omega_{\text{up}} + 2\alpha_{\text{up}}^2 - 3\alpha_{\text{up}}\right) \times \frac{1}{N^2}\sum_{i=1}^{N}\left\|h_{k-1}^i - h_*^i\right\|^2 \\
&\qquad + \gamma^2(\omega_{\text{dwn}}+1)L\left\langle \nabla F(w_{k-1}), w_{k-1} - w_*\right\rangle \\
&\qquad + \frac{2\gamma^2}{N}\left(\frac{\sigma_*^2}{b}\left((2\omega_{\text{up}}+1)(\omega_{\text{dwn}}+1) + 2C\left(2\alpha_{\text{up}}^2(\omega_{\text{up}}+1) - \alpha_{\text{up}}\right)\right)\right) .
\end{aligned}$$

We transform $\|g_k^i - g_{k,*}^i\|^2$ applying co-coercivity (Lemma B.9) and note:

- $\square = 1 - \gamma L(\omega_{\text{dwn}}+1)/2 - \gamma L\left((2\omega_{\text{up}}+1)(\omega_{\text{dwn}}+1) + 2C(2\alpha_{\text{up}}^2\omega_{\text{up}} + 2\alpha_{\text{up}}^2 - \alpha_{\text{up}})\right)/N$
- $\diamondsuit = \frac{\omega_{\text{up}}(\omega_{\text{dwn}}+1)}{C} + 1 + 2\alpha_{\text{up}}^2\omega_{\text{up}} + 2\alpha_{\text{up}}^2 - 3\alpha_{\text{up}}$
- $E = \frac{\sigma_*^2}{b}\left((2\omega_{\text{up}}+1)(\omega_{\text{dwn}}+1) + 2C\left(2\alpha_{\text{up}}^2(\omega_{\text{up}}+1) - \alpha_{\text{up}}\right)\right)$.

Now, because $\mathbb{E}\left[\widehat{g}_k \mid \mathcal{H}_{k-1}\right] = \mathbb{E}\left[h_{k-1} + \frac{1}{N}\sum_{i=1}^{N}\widehat{\Delta}_k^i \;\middle|\; \mathcal{H}_{k-1}\right] = \nabla F(w_{k-1})$, we have:

$$\begin{aligned}
\mathbb{E}\left[V_k \mid \mathcal{H}_{k-1}\right] \leq \left\|w_{k-1} - w_*\right\|^2 - 2\gamma\square\left\langle \nabla F(w_{k-1}), w_{k-1} - w_*\right\rangle \\
+ \frac{2\gamma^2\diamondsuit}{N^2}\sum_{i=1}^{N}\left\|h_{k-1}^i - h_*^i\right\|^2 + \frac{2\gamma^2 E}{N} .
\end{aligned} \tag{B.14}$$

Now, the goal is to apply strong-convexity of F (Assumption 2.1) using the inequality presented in Proposition A.1. But then we must have $\square \geq 0$. However, in order to later obtain a convergence in $(1-\gamma\mu)$, we will use a stronger condition and, instead, state that we need $\square \geq 1/2$, which is

equivalent to:

$$\begin{aligned}
&\frac{\omega_{\text{dwn}}+1}{2}+\frac{1}{N}\left((2\omega_{\text{up}}+1)(\omega_{\text{dwn}}+1)+2C(2\alpha_{\text{up}}^2\omega_{\text{up}}+2\alpha_{\text{up}}^2-\alpha_{\text{up}})\right)\leq\frac{1}{2\gamma L}\\
\iff\quad&(2\omega_{\text{up}}+1)(\omega_{\text{dwn}}+1)+2C(2\alpha_{\text{up}}^2\omega_{\text{up}}+2\alpha_{\text{up}}^2-\alpha_{\text{up}})\leq\frac{(1-\gamma L(\omega_{\text{dwn}}+1))N}{2\gamma L}\\
\iff\quad&C\leq\frac{N-\gamma L(\omega_{\text{dwn}}+1)\left(N+2(2\omega_{\text{up}}+1)\right)}{4\gamma L\alpha_{\text{up}}\left(2\alpha_{\text{up}}(\omega_{\text{up}}+1)-1\right)}\,.
\end{aligned}$$

This holds only if the numerator and the denominator are positive:

$$\begin{cases}
N-\gamma L(\omega_{\text{dwn}}+1)\left(N+2(2\omega_{\text{up}}+1)\right)>0\iff\gamma<\dfrac{N}{(\omega_{\text{dwn}}+1)\left(N+2(2\omega_{\text{up}}+1)\right)L}\\
2\alpha_{\text{up}}(\omega_{\text{up}}+1)-1\leq 0\iff\alpha_{\text{up}}\geq\dfrac{1}{2(\omega_{\text{up}}+1)}\,.
\end{cases}$$

strong-convexity is applied, and we obtain:

$$\mathbb{E}\left[V_k\mid\mathcal{H}_{k-1}\right]\leq(1-2\gamma\mu\square)\left\|w_{k-1}-w_*\right\|^2+\frac{2\gamma^2C\diamondsuit}{N}\sum_{i=1}^{N}\left\|h_{k-1}^i-h_*^i\right\|^2+\frac{2\gamma^2E}{N}\,. \tag{B.15}$$

To guarantee a $(1-\gamma\mu)$ convergence, constants must verify: (1) $\square\geq 1/2$ and (2) $\diamondsuit\leq 1-\gamma\mu$. The first condition is already verified, and the second one leads to:

$$\begin{aligned}
\diamondsuit\leq 1-\gamma\mu&\iff\frac{\omega_{\text{dwn}}+1}{C}\omega_{\text{up}}\leq 3\alpha_{\text{up}}-2\alpha_{\text{up}}^2\omega_{\text{up}}-2\alpha_{\text{up}}-\gamma\mu\\
&\iff C\geq\frac{\omega_{\text{up}}(\omega_{\text{dwn}}+1)}{\alpha_{\text{up}}\left(3-2\alpha_{\text{up}}(\omega_{\text{up}}+1)\right)-\gamma\mu}\,.
\end{aligned}$$

In the following we will consider that $\frac{\gamma\mu}{\alpha_{\text{up}}}=\underset{\mu\to 0}{o}(1)$ which is possible because α_{up} is independent of μ (it depends only of ω_{up} and ω_{dwn}) and it result to:

$$\alpha_{\text{up}}\left(3-2\alpha_{\text{up}}(\omega_{\text{up}}+1)\right)-\gamma\mu\underset{\mu\to 0}{\sim}\alpha_{\text{up}}\left(3-2\alpha_{\text{up}}(\omega_{\text{up}}+1)\right)$$

Thus, the condition on C becomes $\dfrac{\omega_{\text{up}}(\omega_{\text{dwn}}+1)}{\alpha_{\text{up}}\left(3-2\alpha_{\text{up}}(\omega_{\text{up}}+1)\right)}\leq C$, which is correct only if $\alpha_{\text{up}}\leq\dfrac{3}{2(\omega_{\text{up}}+1)}$. And we obtain the following conditions on C:

$$\frac{\omega_{\text{up}}(\omega_{\text{dwn}}+1)}{\alpha_{\text{up}}\left(3-2\alpha_{\text{up}}(\omega_{\text{up}}+1)\right)}\leq C\leq\frac{N-\gamma L(\omega_{\text{dwn}}+1)\left(N+2(2\omega_{\text{up}}+1)\right)}{4\gamma L\alpha_{\text{up}}\left(2\alpha_{\text{up}}(\omega_{\text{up}}+1)-1\right)}\,.$$

It follows, that the above interval is not empty if:

$$\frac{\omega_{\text{up}}(\omega_{\text{dwn}}+1)}{\alpha_{\text{up}}\left(3-2\alpha_{\text{up}}(\omega_{\text{up}}+1)\right)}\leq\frac{N-\gamma L(\omega_{\text{dwn}}+1)\left(N+2(2\omega_{\text{up}}+1)\right)}{4\gamma L\alpha_{\text{up}}\left(2\alpha_{\text{up}}(\omega_{\text{up}}+1)-1\right)}\,.$$

For sake of clarity we denote momentarily $\widetilde{\gamma}=(\omega_{\text{dwn}}+1)\gamma L$, hence the above condition becomes:

$$\begin{aligned}
&8\alpha_{\text{up}}\omega_{\text{up}}(\omega_{\text{up}}+1)\widetilde{\gamma}-4\omega_{\text{up}}\widetilde{\gamma}\leq 3N-3\widetilde{\gamma}\left(N+2+2(2\omega_{\text{up}}+1)\right))\\
&\qquad\qquad-2\alpha_{\text{up}}(\omega_{\text{up}}+1)N+2\alpha_{\text{up}}\widetilde{\gamma}(\omega_{\text{up}}+1)\left(N+2(2\omega_{\text{up}}+1)\right)\\
\iff\quad&2\alpha_{\text{up}}(\omega_{\text{up}}+1)(N-\widetilde{\gamma}(N+2))\leq 3N-\widetilde{\gamma}\left(3N+8\omega_{\text{up}}+6\right)\,.
\end{aligned}$$

And at the end, we obtain:

$$\alpha_{\text{up}} \leq \frac{3N - \gamma L(\omega_{\text{dwn}}+1)\left(3N + 8(\omega_{\text{up}}+6)\right)}{2(\omega_{\text{up}}+1)(N - \gamma L(\omega_{\text{dwn}}+1)(N+2))}\,.$$

Again, this implies two conditions on γ:

$$\begin{cases} 3N - \gamma L(\omega_{\text{dwn}}+1)\left(3N + 8\omega_{\text{up}} + 6\right) > 0 \iff \gamma < \dfrac{3}{(\omega_{\text{dwn}}+1)\left(3 + \dfrac{8\omega_{\text{up}}+6}{N}\right)L} \\ N - \gamma L(\omega_{\text{dwn}}+1)(N+2) > 0 \iff \gamma < \dfrac{1}{(\omega_{\text{dwn}}+1)\left(1+\dfrac{2}{N}\right)L}\,. \end{cases}$$

The constant C exists, and from Equation (B.15), taking full expectation, we are allowed to write $\mathbb{E}[V_k] \leq (1-\gamma\mu)\mathbb{E}[V_{k-1}] + 2\gamma^2\frac{E}{N}$. Unrolling the inequality we obtain:

$$\begin{aligned} & \mathbb{E}[V_k] \leq (1-\gamma\mu)^k \mathbb{E}V_0 + 2\gamma^2\frac{E}{N} \times \frac{1-(1-\gamma\mu)^k}{\gamma\mu} \\ \implies \quad & \mathbb{E}[V_k] \leq (1-\gamma\mu)^k V_0 + 2\gamma\frac{E}{\mu N}\,. \end{aligned}$$

Because $V_0 = \mathbb{E}\left\|w_0 - w_*\right\|^2 + 2\gamma^2 C\frac{1}{N}\sum_{i=0}^N \left\|h_*^i\right\|^2 \leq \left\|w_0 - w_*\right\|^2 + 2C\gamma^2B^2$ (Assumption 2.4), we can write:

$$\mathbb{E}[V_k] = (1-\gamma\mu)^k\left(\left\|w_0 - w_*\right\|^2 + 2C\gamma^2B^2\right) + 2\gamma\frac{E}{\mu N}\,.$$

Thus, we highlighted that the Lyapunov function V_k is a $(1-\gamma\mu)$ contraction if C is taken in a given interval, with γ and α_{up} satisfying some conditions. This guarantees the convergence of the Artemis using version 1 or 2 with $\alpha_{\text{up}} \neq 0$ (algorithm with uni-compression or bi-compression combined with a memory mechanism).

■

B.4.3 Proof of Theorem 2.2 - Polyak-Ruppert averaging

Theorem B.3 (Unidirectional or bidirectional compression using memory and averaging)**.** *Considering now that F is convex, thus $\mu = 0$ and considering that Assumptions 2.2 to 2.5 hold. We use w_* to indicate the optimal parameter such that $\nabla F(w_*) = 0$, and we note $h_*^i = \nabla F_i(w_*)$. A Lyapunov function is defined for any k in $\mathbb{N}$:*

$$V_k = \left\|w_k - w_*\right\|^2 + 2\gamma^2 C\frac{1}{N}\sum_{i=1}^N \left\|h_k^i - h_*^i\right\|^2\,.$$

We defined $C \in \mathbb{R}_+^*$, such that:

$$\frac{\omega_{\text{up}}(\omega_{\text{dwn}}+1)}{\alpha_{\text{up}}\left(3 - 2\alpha_{\text{up}}(\omega_{\text{up}}+1)\right)} \leq C \leq \frac{N - \gamma L(\omega_{\text{dwn}}+1)\left(N + 2(2\omega_{\text{up}}+1)\right)}{4\gamma L\alpha_{\text{up}}\left(2\alpha_{\text{up}}(\omega_{\text{up}}+1) - 1\right)}\,.$$

Then running the variant of Artemis *with $\alpha_{\text{up}} \neq 0$, hence with a memory mechanism, and using Polyak-Ruppert averaging, the convergence of the algorithm is guaranteed if:*

1. $\dfrac{1}{2(\omega_{\text{up}}+1)} \leq \alpha_{\text{up}} < \min\left(\dfrac{3}{2(\omega_{\text{up}}+1)}, \quad \dfrac{3N-\gamma L(\omega_{\text{dwn}}+1)\,(3N+8\omega_{\text{up}}+6)}{2(\omega_{\text{up}}+1)(N-\gamma L(\omega_{\text{dwn}}+1)(N+2))}\right)$

2. $$\gamma < \min\left\{ \dfrac{1}{(\omega_{\text{dwn}}+1)\left(1+\dfrac{2}{N}\right)L}, \quad \dfrac{3}{(\omega_{\text{dwn}}+1)\left(3+\dfrac{8\omega_{\text{up}}+6}{N}\right)L}, \quad \dfrac{N}{(\omega_{\text{dwn}}+1)\,(N+2(2\omega_{\text{up}}+1))\,L} \right\}. \tag{B.16}$$

And we have the following bound for the Polyak-Ruppert averaged iterate $\overline{w}_{K-1} = \frac{1}{K}\sum_{k=0}^{K-1} w_k$*:*

$$\mathbb{E}\left[F(\overline{w}_{K-1}) - F(w_*)\right] \leq \frac{\|w_0 - w_*\|^2 + 2C\gamma^2 B^2}{\gamma K} + 2\gamma\frac{E}{N}\,, \tag{B.17}$$

with $E = \dfrac{\sigma_*^2}{b}\left((2\omega_{\text{up}}+1)(\omega_{\text{dwn}}+1) + 2C\left(2\alpha_{\text{up}}^2(\omega_{\text{up}}+1) - \alpha_{\text{up}}\right)\right)$*. Equation (B.17) can be written as in Theorem 2.2 if we take* $\gamma = \min\left(\sqrt{\frac{N\delta_0^2}{2EK}}; \gamma_{\max}\right)$*, where* $\gamma_{\max}$ *is the maximal possible value of* γ *as precised by Equation (B.16):*

$$\mathbb{E}\left[F(\overline{w}_{K-1}) - F(w_*)\right] \leq 2\max\left(\sqrt{\frac{2\delta_0^2 E}{NK}}; \frac{\delta_0^2}{\gamma_{\max}K}\right) + \frac{2\gamma_{\max}CB^2}{K}$$

Proof

Let k in $\mathbb{N}^*$, starting from Equation (B.14) from the proof of Theorem B.2, we have:

$$\mathbb{E}\left[V_k \mid \mathcal{H}_{k-1}\right] \leq \|w_{k-1} - w_*\|^2 - 2\gamma\square\left\langle \nabla F(w_{k-1}), w_{k-1} - w_*\right\rangle + \frac{2\gamma^2\diamondsuit}{N^2}\sum_{i=1}^{N}\left\|h_{k-1}^i - h_*^i\right\|^2 + \frac{2\gamma^2 E}{N}\,.$$

But this time, instead of applying strong-convexity of F, we apply convexity (Assumption 2.1 but with $\mu = 0$):

$$\mathbb{E}V_k \leq \|w_{k-1} - w_*\|^2 - 2\gamma\square\left(F(w_{k-1}) - F(w_*)\right) + \frac{2\gamma^2 C\diamondsuit}{N^2}\sum_{i=1}^{N}\left\|h_{k-1}^i - h_*^i\right\|^2 + \frac{2\gamma^2}{N}E \tag{B.18}$$

As in Theorem B.2, we want $\square \geq 1/2$, which is equivalent to:

$$\begin{aligned} &\frac{\omega_{\text{dwn}}+1}{2} + \frac{1}{N}\left((2\omega_{\text{up}}+1)(\omega_{\text{dwn}}+1) + 2C(2\alpha_{\text{up}}^2\omega_{\text{up}} + 2\alpha_{\text{up}}^2 - \alpha_{\text{up}})\right) \leq \frac{1}{2\gamma L} \\ \iff \quad & C \leq \frac{N - \gamma L(\omega_{\text{dwn}}+1)\,(N+8\omega_{\text{up}}+6)}{4\gamma L\alpha_{\text{up}}\,(2\alpha_{\text{up}}(\omega_{\text{up}}+1) - 1)}\,. \end{aligned} \tag{B.19}$$

It holds only if the numerator and the denominator are positive:

$$\begin{cases} N - \gamma L(\omega_{\text{dwn}}+1)\,(N+8\omega_{\text{up}}+6) > 0 \iff \gamma < \dfrac{N}{(\omega_{\text{dwn}}+1)\,(N+8\omega_{\text{up}}+6)\,L} \\ 2\alpha_{\text{up}}(\omega_{\text{up}}+1) - 1 \leq 0 \iff \alpha_{\text{up}} \geq \dfrac{1}{2(\omega_{\text{up}}+1)}\,. \end{cases}$$

Returning to Equation (B.18), taking benefit of Equation (B.19) and passing $F(w_{k-1}) - F(w_*)$ on the left side gives:

$$\gamma(F(w_{k-1}) - F(w_*)) \leq \|w_{k-1} - w_*\|^2 + \frac{2\gamma^2 C\diamondsuit}{N^2}\sum_{i=1}^{N}\left\|h_{k-1}^i - h_*^i\right\|^2 - \mathbb{E}V_k + \frac{2\gamma^2}{N}E\,.$$

If $\diamondsuit \leq 1$, we have $\gamma \mathbb{E}\left[F(w_{k-1}) - F(w_*)\right] \leq \mathbb{E}V_{k-1} - \mathbb{E}V_k + 2\gamma^2 E/N$, and summing over all K in $\mathbb{N}^*$ iterations gives:

$$\begin{aligned}\gamma\left(\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[F(w_{k-1}) - F(w_*)\right]\right) &\leq \frac{1}{K}\sum_{k=1}^{K}\left(\mathbb{E}V_{k-1} - \mathbb{E}V_k + 2\gamma^2\frac{E}{N}\right)\\ &\leq \frac{\mathbb{E}V_0 - \mathbb{E}V_K}{K} + 2\gamma^2\frac{E}{N} \quad \text{because } E \text{ is independent of } K.\end{aligned}$$

Thus, by convexity:

$$\mathbb{E}\left[F\left(\frac{1}{K}\sum_{k=1}^{K}w_{k-1}\right) - F(w_*)\right] \leq \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[F(w_{k-1}) - F(w_*)\right] \leq \frac{V_0}{\gamma K} + 2\gamma\frac{E}{N}\,.$$

Last step is to extract conditions over γ and α_{up} from requirement $\diamondsuit \leq 1$:

$$\diamondsuit < 1 \iff \frac{2\omega_{\text{up}}(\omega_{\text{dwn}}+1)}{2C} < 3\alpha_{\text{up}} - 2\alpha_{\text{up}}^2\omega_{\text{up}} - 2\alpha_{\text{up}} \iff C > \frac{\omega_{\text{up}}(\omega_{\text{dwn}}+1)}{\alpha_{\text{up}}\left(3 - 2\alpha_{\text{up}}(\omega_{\text{up}}+1)\right)}\,,$$

and the second inequality is correct only if $\alpha_{\text{up}} \leq \frac{3}{2(\omega_{\text{up}}+1)}$. From this development follows the following conditions on C, which are equivalent to those obtain in Theorem B.2

$$\frac{\omega_{\text{up}}(\omega_{\text{dwn}}+1)}{\alpha_{\text{up}}\left(3 - 2\alpha_{\text{up}}(\omega_{\text{up}}+1)\right)} \leq C \leq \frac{N - \gamma L(\omega_{\text{dwn}}+1)\left(N + 2(2\omega_{\text{up}}+1)\right)}{4\gamma L\alpha_{\text{up}}\left(2\alpha_{\text{up}}(\omega_{\text{up}}+1) - 1\right)}\,.$$

This interval is not empty:

$$\begin{aligned}&\frac{\omega_{\text{up}}(\omega_{\text{dwn}}+1)}{\alpha_{\text{up}}\left(3 - 2\alpha_{\text{up}}(\omega_{\text{up}}+1)\right)} \leq \frac{N - \gamma L(\omega_{\text{dwn}}+1)\left(N + 2(2\omega_{\text{up}}+1)\right)}{4\gamma L\alpha_{\text{up}}\left(2\alpha_{\text{up}}(\omega_{\text{up}}+1) - 1\right)}\\ \iff \quad &\alpha_{\text{up}} \leq \frac{3N - \gamma L(\omega_{\text{dwn}}+1)\left(3N + 8\omega_{\text{up}} + 6\right)}{2(\omega_{\text{up}}+1)(N - \gamma L(\omega_{\text{dwn}}+1)(N+2))}\,.\end{aligned}$$

Again, this implies two conditions on γ:

$$\begin{cases} 3N - \gamma L(\omega_{\text{dwn}}+1)\left(3N + 8\omega_{\text{up}} + 6\right) > 0 \iff \gamma < \dfrac{3}{(\omega_{\text{dwn}}+1)\left(3 + \dfrac{8\omega_{\text{up}}+6}{N}\right)L} \\ N - \gamma L(\omega_{\text{dwn}}+1)(N+2) > 0 \iff \gamma < \dfrac{1}{(\omega_{\text{dwn}}+1)\left(1 + \dfrac{2}{N}\right)L}\,. \end{cases}$$

which guarantees the existence of C and thus the validity of the above development. In conclusion:

$$\begin{aligned}\mathbb{E}\left[F(\overline{w}_{K-1}) - F(w_*)\right] \leq \frac{V_0}{\gamma K} + 2\gamma\frac{E}{N} &\leq \frac{\|w_0 - w_*\|^2 + 2C\gamma^2B^2}{\gamma K} + 2\gamma\frac{E}{N}\\ &\leq \frac{\|w_0 - w_*\|^2}{\gamma K} + 2\gamma\left(\frac{E}{N} + \frac{CB^2}{K}\right)\,.\end{aligned}$$

Next, our goal is to define the optimal step-size γ_{opt}. With this aim, we bound $2\gamma\frac{CB^2}{K}$ by $2\gamma_{\max}\frac{CB^2}{K}$. This leads to ignore this term when optimizing the step-size and thus to obtain a simpler expression of γ_{opt}. This approximation is relevant, because B^2/K is "small". And we obtain:

$$\mathbb{E}\left[F(\overline{w}_{K-1}) - F(w_*)\right] \leq \frac{\|w_0 - w_*\|^2}{\gamma K} + 2\gamma\frac{E}{N} + 2\gamma_{\max}\frac{CB^2}{K}.$$

This is valid for all variants of Artemis, with step-size in Table 2.3 and E in Theorem 2.1. Subsequently, the "optimal" step-size (at least the one minimizing the upper bound) is

$$\gamma_{\text{opt}} = \sqrt{\frac{\|w_0 - w_*\|^2 N}{2EK}},$$

resulting in a convergence rate as $2\sqrt{\frac{2\|w_0-w_*\|^2 E}{NK}} + \frac{2\gamma_{\max}CB^2}{K}$, if this step-size is allowed. If $\sqrt{\frac{\|w_0-w_*\|^2 N}{2EK}} \geq \gamma_{\max}$ $\left(\Longrightarrow \frac{2\gamma_{\max}E}{N} \leq \frac{\|w_0-w_*\|^2}{\gamma_{\max}K}\right)$, then the bias term dominates and the upper bound is $2\frac{\|w_0-w_*\|^2}{\gamma_{\max}K} + \frac{2\gamma_{\max}CB^2}{K}$. Overall, the convergence rate is given by:

$$\mathbb{E}\left[F(\overline{w}_{K-1}) - F(w_*)\right] \leq 2\max\left(\sqrt{\frac{2\|w_0 - w_*\|^2 E}{NK}}; \frac{\|w_0 - w_*\|^2}{\gamma_{\max}K}\right) + \frac{2\gamma_{\max}CB^2}{K}.$$

■

B.4.4 Proof of Theorem 2.3 - convergence in distribution

In this Section, we give the proof of Theorem 2.3. The theorem is decomposed into two main points, that are respectively derived from Propositions B.6 and B.7, given in Subsections B.4.4.2 and B.4.4.3. Throughout this Section, we consider a *linear* compression operator $\mathcal{C}$, for instance, sparsification (Definition 1.1), then for any $z, z' \in \mathbb{R}^d$, we have that $\mathcal{C}(z) - \mathcal{C}(z') = \mathcal{C}(z - z')$. We first introduce a few notations in Subsection B.4.4.1.

B.4.4.1 Background on distributions and Markov chains

We consider Artemis iterates $(w_{k-1}, (h^i_{k-1})_{i\in[\![1,N]\!]})_{k\in\mathbb{N}} \in \mathbb{R}^{d(1+N)}$ with the following update equation:

$$\left\{\begin{array}{lll} & w_k &= w_{k-1} - \gamma\mathcal{C}_{\text{dwn}}\left(\frac{1}{N}\sum_{i=1}^N \mathcal{C}_{\text{up}}\left(g^i_k - h^i_{k-1}\right) + h^i_{k-1}\right) \\ \forall i \in [\![1,N]\!], & h^i_k &= h^i_{k-1} + \alpha_{\text{up}}\mathcal{C}_{\text{up}}\left(g^i_k - h^i_{k-1}\right) \end{array}\right. \tag{B.20}$$

We see the iterates, for a constant step-size γ, as a homogeneous Markov chain, and denote $R_{\gamma,v}$ the *Markov kernel*, which is the equivalent for continuous spaces of the *transition matrix* in finite state spaces. Let $R_{\gamma,v}$ be the Markov kernel on $(\mathbb{R}^{d(1+N)}, \mathcal{B}(\mathbb{R}^{d(1+N)}))$ associated with the SGD iterates $(w_{k-1}, \tau(h^i_{k-1})_{i\in[\![1,N]\!]})_{k\geq 0}$ for a variant v of Artemis, as defined in Algorithm 2 and with τ a constant specified afterwards, where $\mathcal{B}(\mathbb{R}^{d(1+N)})$ is the Borel σ-field of $\mathbb{R}^{d(1+N)}$. Meyn and Tweedie [2009] provide an introduction to Markov chain theory. For readability, we now denote $(h^i_{k-1})_i$ for $(h^i_{k-1})_{i\in[\![1,N]\!]}$.

Definition B.2. *For any initial distribution ν_0 on $\mathcal{B}(\mathbb{R}^{d(1+N)})$ and $k \in \mathbb{N}^*$, $\nu_0 R^k_{\gamma,v}$ denotes the distribution of $(w_{k-1}, \tau(h^i_{k-1})_i)$ starting at $(w_0, \tau(h^i_0)_i)$ distributed according to ν_0.*

We can make the following comments:

1. **Initial distribution.** We consider deterministic initial points, i.e., $(w_0, \tau(h^i_0)_i)$ follows a Dirac at point $(w_0, \tau(h^i_0)_i)$. We denote this Dirac $\delta_{w_0} \otimes \otimes_{i=1}^N \delta_{\tau h^i_0} \overset{\text{not.}}{=} \delta_{w_0} \otimes \delta_{\tau h^1_0} \otimes \cdots \otimes \delta_{\tau h^N_0}$.

2. **Notation in the main text:** In the main text, for simplicity, we used Θ_k to denote the distribution of w_{k-1} when launched from $(w_0, \tau(h_0^i)_i)$. Thus Θ_k corresponds to the distribution of the projection on first d coordinates of $((\delta_{w_0} \otimes \otimes_{i=1}^N \delta_{\tau h_0^i}) R_\gamma^k)$.

3. **Case without memory:** In the memory-less case, we have $(h_{k-1}^i)_{k\in\mathbb{N}} \equiv 0$, and could restrict ourselves to a Markov kernel on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$.

For any variant v of Artemis, we prove that $(w_{k-1}, (h_{k-1}^i)_i)_{k\geq 0}$ admits a limit stationary distribution

$$\Pi_{\gamma,v} = \pi_{\gamma,v,w} \otimes \pi_{\gamma,v,(h)} \tag{B.21}$$

and quantify the convergence of $((\delta_{w_0} \otimes \otimes_{i=1}^N \delta_{\tau h_0^i}) R_\gamma^k)_{k\geq 0}$ to $\Pi_{\gamma,v}$, in terms of Wasserstein metric $\mathcal{W}_2$.

Definition B.3. *For all probability measures ν and λ on $\mathcal{B}(\mathbb{R}^d)$, such that $\int_{\mathbb{R}^d} \|w\|^2 \,\mathrm{d}\nu(w) < +\infty$ and $\int_{\mathbb{R}^d} \|w\|^2 \,\mathrm{d}\lambda(w) \leq +\infty$, define the squared Wasserstein distance of order 2 between λ and ν by*

$$\mathcal{W}_2^2(\lambda,\nu) := \inf_{\zeta\in\Gamma(\lambda,\nu)} \int \|x-y\|^2 \zeta(dx,dy), \tag{B.22}$$

where $\Gamma(\lambda,\nu)$ is the set of probability measures ζ on $\mathcal{B}(\mathbb{R}^d \times \mathbb{R}^d)$ satisfying for all $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$, $\zeta(\mathsf{A}\times\mathbb{R}^d) = \nu(\mathsf{A})$, $\zeta(\mathbb{R}^d \times \mathsf{A}) = \lambda(\mathsf{A})$.

B.4.4.2 Proof of the first point in Theorem 2.3

We prove the following proposition:

Proposition B.6. *Under Assumptions 2.1 to 2.5, for any linear compression operator $\mathcal{C}$, for any variant v of the algorithm, there exists a limit distribution $\Pi_{\gamma,v}$, which is stationary, such that for any k in $\mathbb{N}$, for any γ satisfying conditions given in Theorems B.1 and B.2:*

$$\mathcal{W}_2^2((\delta_{w_0} \otimes \otimes_{i=1}^N \delta_{\tau h_0^i}) R_\gamma^k, \Pi_{\gamma,v}) \leq (1-\gamma\mu)^k \int_{(w',h')\in\mathbb{R}^{d(1+N)}} \left\|(w_0, \tau(h_0^i)_i) - (w', \tau(h^i)_i')\right\|^2 \mathrm{d}\Pi_{\gamma,v}(w', (h^i)_i').$$

Point 1 in Theorem 2.3 is derived from the proposition above using $\pi_{\gamma,v} = \pi_{\gamma,v,w}$, with $\pi_{\gamma,v,w}$ as in Equation (B.21), the limit distribution of the main iterates $(w_{k-1})_{k\in\mathbb{N}}$ and the observation that:

$$\begin{aligned}
\mathcal{W}_2^2(\Theta_k, \pi_{\gamma,v}) &\leq \mathcal{W}_2^2((\delta_{w_0} \otimes \otimes_{i=1}^N \delta_{\tau h_0^i}) R_{\gamma,v}^k, \Pi_{\gamma,v}) \\
&\leq (1-\gamma\mu)^k \int_{(w',h')\in\mathbb{R}^{d(1+N)}} \left\|(w_0, \tau(h_0^i)_i) - (w', \tau(h^i)_i')\right\|^2 \mathrm{d}\Pi_{\gamma,v}(w', (h^i)_i') \\
&= (1-\gamma\mu)^k C_0.
\end{aligned}$$

The sketch of the proof is simple:

- We introduce a *coupling of random variables* following respectively $\nu_0^a R_{\gamma,v}^k$ and $\nu_0^b R_{\gamma,v}^k$, and show that under the assumptions given in the proposition:

  $$\mathcal{W}_2^2(\nu_0^a R_{\gamma,v}^k, \nu_0^b R_{\gamma,v}^k) \leq (1-\gamma\mu)\mathcal{W}_2^2(\nu_0^a R_{\gamma,v}^{k-1}, \nu_0^b R_{\gamma,v}^{k-1}).$$

  This proof follows the same line as the proof of Theorems B.1 and B.2.

- We deduce that $((\delta_{w_0} \otimes \otimes_{i=1}^N \delta_{\tau h_0^i})) R_{\gamma,v}^k)$ is a Cauchy sequence in a Polish space, thus the existence and stability of the limit, we show that this limit is independent from $(\delta_{w_0} \otimes \otimes_{i=1}^N \delta_{\tau h_0^i}))$ and conclude.

Proof We consider two initial distributions ν_0^a and ν_0^b for $(w_0, \tau(h_0^i)_i)$ with finite second moment and $\gamma > 0$. Let $(w_0^a, \tau(h_0^{i,a})_i)$ and $(w_0^b, \tau(h_0^{i,b})_i)$ be respectively distributed according to ν_0^a and ν_0^b. Let $(w_k^a, \tau(h_k^{i,a})_i)_{k\geq 0}$ and $(w_k^b, \tau(h_k^{i,b})_i)_{k\geq 0}$ the `Artemis` iterates, respectively starting from $(w_0^a, \tau(h_0^{i,a})_i)$ and $(w_0^b, \tau(h_0^{i,b})_i)$, and *sharing the same sequence of noises*, i.e.,

- built with the same gradient oracles $g_k^{i,a} = g_k^{i,b}$ for all $k \in \mathbb{N}, i \in [\![1, N]\!]$.
- the compression operator used for both recursions is almost surely the same, for any iteration k, and both uplink and downlink compression. We denote these operators $\mathcal{C}_{\text{dwn},k}$ and $\mathcal{C}_{\text{up},k}$ the compression operators at iteration k for respectively the uplink compression and downlink compression.

We thus have the following updates, for any $u \in \{a, b\}$:

$$\left\{ \begin{array}{lll} & w_k^u & = w_{k-1}^u - \gamma \mathcal{C}_{\text{dwn},k}\left(\frac{1}{N}\sum_{i=1}^N \mathcal{C}_{\text{up},k}\left(g_k^i - h_{k-1}^{i,u}\right) + h_{k-1}^{i,u}\right) \\ \forall i \in [\![1;n]\!] & h_k^{i,u} & = h_{k-1}^{i,u} + \alpha_{\text{up}} \mathcal{C}_{\text{up},k}\left(g_k^i - h_{k-1}^{i,u}\right) . \end{array} \right. \tag{B.23}$$

The proof is obtained by induction. For a k in $\mathbb{N}$, let $\left((w_k^a, \tau(h_k^{i,a})_i), (w_k^b, \tau(h_k^{i,b})_i)\right)$ be a coupling of random variable in $\Gamma(\nu_0^a R_{\gamma,v}^k, \nu_0^b R_{\gamma,v}^k)$ – as in Definition B.3 –, that achieve the equality in the definition, i.e.,

$$\mathcal{W}_2^2(\nu_0^a R_{\gamma,v}^k, \nu_0^b R_{\gamma,v}^k) = \mathbb{E}\left[\left\|(w_k^a, \tau(h_k^{i,a})_i) - (w_k^b, \tau(h_k^{i,b})_i)\right\|^2\right]. \tag{B.24}$$

Existence of such a couple is given by [Villani, 2009, theorem 4.1]. Then $\left((w_k^a, \tau(h_k^{i,a})_i), (w_k^b, \tau(h_k^{i,b})_i)\right)$ obtained after one update from Equation (B.23) belongs to $\Gamma(\nu_0^a R_{\gamma,v}^k, \nu_0^b R_{\gamma,v}^k)$, and as a consequence:

$$\begin{aligned} \mathcal{W}_2^2(\nu_0^a R_{\gamma,v}^k, \nu_0^b R_{\gamma,v}^k) &\leq \mathbb{E}\left[\left\|(w_k^a, \tau(h_k^{i,a})_i) - (w_k^b, \tau(h_k^{i,b})_i))\right\|^2\right] \\ &= \mathbb{E}\left[\left\|w_k^a - w_k^b\right\|^2\right] + \tau^2 \sum_{i=1}^N \mathbb{E}\left[\left\|h_k^{i,a} - h_k^{i,b}\right\|^2\right] \\ &= \mathbb{E}\left[\left\|w_k^a - w_k^b\right\|^2\right] + 2\gamma^2 \frac{C}{N} \sum_{i=1}^N \mathbb{E}\left[\left\|h_k^{i,a} - h_k^{i,b}\right\|^2\right], \end{aligned}$$

with $\tau^2 = 2\gamma^2 \frac{C}{N}$, where C depends on the variant as in Theorem 2.1. We now follow the proof of the previous theorems to control respectively $\mathbb{E}[\|w_k^a - w_k^b\|^2]$ and $\mathbb{E}[\|h_k^{i,a} - h_k^{i,b}\|^2]$. First, following the proof of Equation (B.12), we get, using the fact that the compression operator is linear, thus that $\mathcal{C}(x) - \mathcal{C}(y) = \mathcal{C}(x - y)$:

$$\begin{aligned} \mathbb{E}\left[\left\|w_k^a - w_k^b\right\|^2 | \mathcal{H}_{k-1}\right] \leq \left\|w_k^a - w_k^b\right\|^2 &- 2\gamma\left\langle \nabla F(w_{k-1}^a) - \nabla F(w_{k-1}^b), w_k^a - w_k^b \right\rangle \\ &+ \frac{2(2\omega_{\text{up}} + 1)(\omega_{\text{dwn}} + 1)\gamma^2}{N^2} \sum_{i=1}^N \mathbb{E}\left[\left\|g_k^i(w_{k-1}^a) - g_k^i(w_{k-1}^b)\right\|^2 \,\middle|\, \mathcal{H}_{k-1}\right] \\ &+ \frac{2\omega_{\text{up}}(\omega_{\text{dwn}} + 1)\gamma^2}{N^2} \sum_{i=1}^N \mathbb{E}\left[\left\|h_{k-1}^{i,a} - h_{k-1}^{i,b}\right\|^2 \,\middle|\, \mathcal{H}_{k-1}\right] \\ &+ \gamma^2(\omega_{\text{dwn}} + 1) L\left\langle \nabla F(w_{k-1}^a) - \nabla F(w_{k-1}^b), w_k^a - w_k^b \right\rangle. \end{aligned}$$

This expression is nearly the same as in Equation (B.12), apart from the constant term depending on σ_*^2 that disappears. Note that with a more general compression operator, for example for quantization, it is not possible to derive such a result. Similarly, we control $\mathbb{E}[\|h_k^{i,a} - h_k^{i,b}\|^2]$ using the same line of proof as for Equation (B.13), resulting in:

$$\frac{1}{N^2}\sum_{i=0}^{N}\mathbb{E}\left[\left\|h_k^{a,i} - h_k^{b,i}\right\|^2 \,\middle|\, \mathcal{H}_{k-1}\right] \leq \left(1 + p\left(2\alpha_{\text{up}}^2\omega_{\text{up}} + 2\alpha_{\text{up}}^2 - 3\alpha_{\text{up}}\right)\right)\frac{1}{N^2}\sum_{i=0}^{N}\mathbb{E}\left[\left\|h_k^{a,i} - h_k^{b,i}\right\|^2 \,\middle|\, \mathcal{H}_{k-1}\right]$$
$$+ 2(2\alpha_{\text{up}}^2\omega_{\text{up}} + 2\alpha_{\text{up}}^2 - \alpha_{\text{up}})\frac{1}{N^2}\sum_{i=0}^{N}\mathbb{E}\left[\left\|g_k^i(w_{k-1}^a) - g_k^i(w_{k-1}^b)\right\|^2 \,\middle|\, \mathcal{H}_{k-1}\right].$$

Combining both equations, and using Assumptions 2.1 and 2.2 and Equation (B.24) we get, under conditions on the learning rates $\alpha_{\text{up}}, \gamma$ similar to the ones in Theorems B.1 and B.2, that

$$\mathcal{W}_2^2(\nu_0^a R_{\gamma,v}^k, \nu_0^b R_{\gamma,v}^k) \leq (1-\gamma\mu)\mathcal{W}_2^2(\nu_0^a R_{\gamma,v}^{k-1}, \nu_0^b R_{\gamma,v}^{k-1}).$$

And by induction:

$$\mathcal{W}_2^2(\nu_0^a R_{\gamma,v}^k, \nu_0^b R_{\gamma,v}^k) \leq (1-\gamma\mu)^k\mathcal{W}_2^2(\nu_0^a, \nu_0^b).$$

■

From the contraction above, it is easy to derive the existence of a unique stationnary limit distribution: we use Picard fixed point theorem, as in Dieuleveut et al. [2020]. This concludes the proof of Proposition B.6.

B.4.4.3 Proof of the second point of Theorem 2.3

To prove the second point, we first detail the complementary assumptions mentioned in the text, then show the convergence to the mean squared distance under the limit distribution, and finally give a lower bound on this quantity.

Complementary assumptions.
To prove the lower bound given by the second point, we need to assume that the constants given in the assumptions are tight, in other words, that corresponding lower bounds exist in Assumptions 2.3 to 2.5.

Assumption B.1 (Lower bound on noise over stochastic gradients computation)**.** *The noise over stochastic gradients at optimal global point for a mini-batch of size b is* lower bounded*. In other words, there exists a constant $\sigma_* \in \mathbb{R}$, such that for all k in $\mathbb{N}$, for all i in $[\![1, N]\!]$, we have a.s:*

$$\mathbb{E}\left[\|g_{k,*}^i - \nabla F_i(w_*)\|^2 \mid \mathcal{H}_{k-1}\right] \geq \frac{\sigma_*^2}{b}.$$

Assumption B.2 (Lower bound on local gradient at w_*)**.** *There exists a constant $B \in \mathbb{R}$, s.t.:*

$$\frac{1}{N}\sum_{i=1}^{N}\|\nabla F_i(w_*)\|^2 \geq B^2.$$

Assumption B.3 (Lower bound on the compression operator's variance)**.** *There exists a constant $\omega \in \mathbb{R}^*$ such that the compression operators $\mathcal{C}_{\text{up}}$ and $\mathcal{C}_{\text{dwn}}$ verify the following property:*

$$\forall \Delta \in \mathbb{R}^d, \mathbb{E}[\left\|\mathcal{C}_{\text{up/dwn}}(\Delta) - \Delta\right\|^2] = \omega_{\text{up/dwn}}\|\Delta\|^2 .$$

This last assumption is valid for sparsification, sketching, rand-h, PP.

Moreover, we also assume some extra regularity on the function. This restricts the regularity of the function beyond Assumption 2.2 and is a purely technical assumption in order to conduct the detailed asymptotic analysis. It is valid in practice for least-squares or logistic regression.

Assumption B.4 (Regularity of the functions)**.** *The function F is also times continuously differentiable with second to fifth uniformly bounded derivatives: for all $k \in \{2, \dots, 5\}$, $\sup_{w \in \mathbb{R}^d} \|F^{(k)}(w)\| < \infty$.*

Convergence of moments.
We first prove that $\mathbb{E}[\|w_{k-1} - w_*\|^2]$ converges to $\mathbb{E}_{w \sim \pi_{\gamma,v}}[\|w - w_*\|^2]$ as k increases to ∞.

We have that the difference satisfies, for random variables w_{k-1} and w following distributions $\delta_{w_0} R_{\gamma,v}^k$ and $\pi_{\gamma,v}$, and coupled such that they achieve the equality in Equation (B.22):

$$
\begin{aligned}
\Delta_{\mathbb{E},k-1} :&= \mathbb{E}[\|w_{k-1} - w_*\|^2] - \mathbb{E}_{w \sim \pi_{\gamma,v}}[\|w - w_*\|^2] \\
&= \mathbb{E}_{w_{k-1}, w \sim \pi_{\gamma,v}} \left[\|w_{k-1} - w_*\|^2 - \|w - w_*\|^2\right] \\
&= \mathbb{E}_{w_{k-1}, w \sim \pi_{\gamma,v}} \left[(\|w_{k-1} - w_*\| - \|w - w_*\|)(\|w_{k-1} - w_*\| + \|w - w_*\|)\right] \\
&\overset{\text{C.S}}{\leq} \left(\mathbb{E}_{w_{k-1}, w \sim \pi_{\gamma,v}} \left[(\|w_{k-1} - w_*\| - \|w - w_*\|)^2\right] \mathbb{E}_{w_{k-1}, w} \left[(\|w_{k-1} - w_*\| + \|w - w_*\|)^2\right]\right)^{1/2} \\
&\overset{\text{T.I.}}{\leq} \left(\mathbb{E}_{w_{k-1}, w \sim \pi_{\gamma,v}} \left[(\|w_{k-1} - w\|)^2\right] \mathbb{E}_{w_{k-1}, w \sim \pi_{\gamma,v}} \left[(\|w_{k-1} - w_*\| + \|w - w_*\|)^2\right]\right)^{1/2} \\
&\overset{\text{(i)}}{\leq} \left(\mathbb{E}_{w_{k-1}, w \sim \pi_{\gamma,v}} \left[(\|w_{k-1} - w\|)^2\right] 2L\right)^{1/2} \\
&\overset{\text{(ii)}}{\leq} \left(\mathcal{W}_2^2(\delta_{w_0} R_{\gamma,v}^{k-1}, \pi_{\gamma,v}) 2L\right)^{1/2} \\
&\overset{\text{(iii)}}{\to} 0.
\end{aligned}
$$

Where we have used Cauchy-Schwarz inequality at line C.S., triangular inequality at line T.I., the fact that the moments are bounded by a constant L at line (i), the fact that the distributions are coupled such that they achieve the equality in Equation (B.22) at line (ii), and finally Proposition B.6 for the conclusion at line (iii).

Overall, this shows that the mean squared distance (i.e., saturation level) converges to the mean squared distance under the limit distribution.

Evaluation of $\mathbb{E}_{w \sim \pi_{\gamma,v}}[\|w - w_*\|^2]$.
In this section, we denote $\xi(w_{k-1}, h_{k-1})$ the *global* noise, defined by

$$
\xi(w_{k-1}, h_{k-1}) = \nabla F(w_{k-1}) - \mathcal{C}_{\text{dwn}} \left(\frac{1}{N} \sum_{i=1}^N \mathcal{C}_{\text{up}}(g_k^i(w_{k-1}) - h_{k-1}^i) + h_{k-1}^i \right),
$$

such that $w_k = w_{k-1} - \gamma \nabla F(w_{k-1}) + \gamma \xi(w_{k-1}, h_{k-1})$. In fact, $(\xi)_{k \in \mathbb{N}*}$ is a zero-centered random field characterizing the stochastic oracle on $\nabla F(\cdot)$, same notation is used in Chapter 4.

In the following, we denote $a^{\otimes 2} := aa^T$ the second order moment of a. We define Tr the trace operator and Cov the covariance operator such that $\text{Cov}(\xi(w,h)) = \mathbb{E}\left[(\xi(w,h))^{\otimes 2}\right]$, where the expectation is taken on the randomness of both compressions and the gradient oracle. We make a final technical assumption on the regularity of the covariance matrix.

Assumption B.5. *We assume that:*

1. $\text{Cov}(\xi(w,h))$ *is continuously differentiable, and there exists constants C and C' such that for all $w, h \in \mathbb{R}^{d(1+N)}$, $\max_{o=1,2,3} \text{Cov}^{(o)}(w,h) \leq C + C' ||(w,h) - (w_*, h_*)||^2$.*

2. $(\xi(w_*, h_*))$ *has finite order moments up to order 8.*

Remark: with the *linear* operators, this assumption can directly be translated into an assumption on the moments and regularity of g_k^i, this is done in Assumptions 4.1 and 4.2 in the setting of LSR. Note that Point 2 in Assumption B.5 is an extension of Assumption 2.3 to higher order moments, but **still at the optimal point**. Under this assumption, we have the following lemma:

Lemma B.10. *Under Assumptions 2.1 to B.5, we have that*

$$\mathbb{E}_{\pi_{\gamma,v}}\left[\|w - w_*\|^2\right] \underset{\gamma\to 0}{=} \gamma \mathrm{Tr}\left(A\ \mathrm{Cov}(\xi(w_*, h_*))\right) + O(\gamma^2), \tag{B.25}$$

with $A := (F''(w_*) \otimes I + I \otimes F''(w_*))^{-1}$.

The intuition of the proof is natural: using the stability of the limit distribution, we have that if we start from the stationary distribution, i.e., $(w_0, h_0) \sim \Pi_{\gamma,v}$, then $(w_1, h_1) \sim \Pi_{\gamma,v}$.

We can thus write:

$$\begin{aligned}\mathbb{E}_{\pi_{\gamma,v}}\left[(w - w_*)^{\otimes 2}\right] &= \mathbb{E}\left[(w_1 - w_*)^{\otimes 2}\right] \\ &= \mathbb{E}\left[(w_0 - w_* - \gamma\nabla F(w_0) + \gamma\xi(w_0, h_0))^{\otimes 2}\right].\end{aligned}$$

Then, expanding the right hand side and using the fact that $\mathbb{E}[\xi(w_0, h_0)|\mathcal{H}_0] = 0$, then the fact that $\mathbb{E}\left[(w_1 - w_*)^{\otimes 2}\right] = \mathbb{E}\left[(w_0 - w_*)^{\otimes 2}\right]$, and expanding the derivative of F around w_* (this is where we require the regularity assumption Assumption B.4), we get that:

$$\gamma\left(F''(w_*) \otimes I + I \otimes F''(w_*) + O(\gamma)\right)\mathbb{E}_{\pi_{\gamma,v}}\left[(w - w_*)^{\otimes 2}\right] \underset{\gamma\to 0}{=} \gamma^2 \mathbb{E}_{(w,h)\sim\Pi_{\gamma,v}}\left[\xi(w, h)^{\otimes 2}\right].$$

Thus:

$$\begin{aligned}\mathbb{E}_{\pi_{\gamma,v}}\left[(w - w_*)^{\otimes 2}\right] &\underset{\gamma\to 0}{=} \gamma A \mathbb{E}_{(w,h)\sim\Pi_{\gamma,v}}\left[\xi(w, h)^{\otimes 2}\right] + O(\gamma^2). \\ \Rightarrow \mathbb{E}_{\pi_{\gamma,v}}\left[\|(w - w_*)\|^2\right] &\underset{\gamma\to 0}{=} \gamma \mathrm{Tr}\left(A\mathbb{E}_{(w,h)\sim\Pi_{\gamma,v}}\left[\xi(w, h)^{\otimes 2}\right]\right) + O(\gamma^2).\end{aligned}$$

Finally, we use that $\mathbb{E}_{(w,h)\sim\Pi_{\gamma,v}}\left[\mathrm{Cov}(\xi(w, h))\right] \underset{\gamma\to 0}{=} \mathrm{Cov}(\xi(w_*, h_*)) + O(\gamma)$ (which is derived from Assumption B.5) to get Lemma B.10. More formally, we can rely on Theorem 4 in Dieuleveut et al. [2020]: under Assumptions 2.1 to 2.5 and Assumptions B.4 and B.5, all assumptions required for the application of the theorem are verified and the result follows.

To conclude the proof, it only remains to control $\mathrm{Cov}(\xi(w_*, h_*))$. We have the following Lemma:

Lemma B.11. *Under Assumptions B.1 to B.3, we have that, for any variant v of the algorithm, with the constant E given in Theorem 2.1 depending on the variant:*

$$\mathrm{Tr}\left(\mathrm{Cov}(\xi(w_*, h_*))\right) = \Omega\left(\frac{\gamma E}{\mu N}\right). \tag{B.26}$$

Combining Lemmas B.10 and B.11 and using the observation that A is lower bounded by $\frac{1}{2L}$ independently of γ, N, σ_*, B, we have proved the following proposition:

Proposition B.7. *Under Assumptions B.1 to B.5, we have that*

$$\mathbb{E}[\|w_{k-1} - w_*\|^2] \underset{k\to\infty}{\to} \mathbb{E}_{\pi_{\gamma,v}}\left[\|w - w_*\|^2\right] \underset{\gamma\to 0}{=} \Omega\left(\frac{\gamma E}{\mu N}\right) + O(\gamma^2), \tag{B.27}$$

where the constant in the Ω is independent of N, σ_*, γ, B (it depends only on the regularity of the operator A).

Before giving the proof, we make a couple of observations:

1. This shows that the upper bound on the limit mean squared error given in Theorem 2.1 **is tight** with respect to N, σ_*, γ, B. This underlines that the conditions on the problem that we have used are the correct ones to understand convergence.

2. The upper bound is possibly not tight with respect to μ, as is clear from the proof: the tight bound is actually $\operatorname{Tr}\left(A\operatorname{Cov}(\xi(w_*, h_*))\right)$. Getting a tight upper bound involving the eigenvalue decomposition of A instead of only μ is an open direction.

3. In the memory-less case, $h \equiv 0$ and all the proof can be carried out analyzing only the distribution of the iterates $(w_{k-1})_k$ and not necessarily the couple $(w_{k-1}, (h^i_{k-1})_i)_k$.

We now give the proof of Lemma B.11.

Proof With memory, we have the following:

$$\begin{aligned}
\operatorname{Tr}\left(\operatorname{Cov}(\xi(w_*, h_*))\right) &= \mathbb{E}\left[\left\|\mathcal{C}_{\text{dwn}}\left(\frac{1}{N}\sum_{i=1}^N \mathcal{C}_{\text{up}}(g_1^i(w_*) - h_*^i) + h_*^i\right)\right\|^2\right] \\
&\overset{\text{(i)}}{=} (1+\omega_{\text{dwn}})\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^N \mathcal{C}_{\text{up}}(g_1^i(w_*) - h_*^i) + h_*^i\right\|^2\right] \\
&\overset{\text{(ii)}}{=} \frac{(1+\omega_{\text{dwn}})}{N^2}\sum_{i=1}^N \mathbb{E}\left[\left\|\mathcal{C}_{\text{up}}(g_1^i(w_*) - h_*^i)\right\|^2\right] \\
&\overset{\text{(iii)}}{=} \frac{(1+\omega_{\text{dwn}})}{N^2}\sum_{i=1}^N (1+\omega_{\text{up}})\mathbb{E}\left[\left\|g_1^i(w_*) - h_*^i\right\|^2\right] \\
&\overset{\text{(iv)}}{\geq} \frac{(1+\omega_{\text{dwn}})}{N}(1+\omega_{\text{up}})\frac{\sigma_*^2}{b}.
\end{aligned}$$

At line (i) we use Assumption B.3 for the downlink compression operator with constant ω_{dwn}. At line (ii) we use the fact that $\sum_{i=1}^N h_*^i = \nabla F(w_*) = 0$, the independence of the random variables $\mathcal{C}_{\text{up}}(g_1^i(w_*) - h_*^i), \mathcal{C}_{\text{up}}(g_1^j(w_*) - h_*^j)$ for $i \neq j$ and the fact that they have 0 mean. We use Assumption B.3 for the uplink compression operator with constant ω_{up} in line (iii); and finally Assumption B.1 at line (iv) to lower bound the variance of the gradients at the optimum. This proof applies to both simple and double compression with $\omega_{\text{dwn}} = 0$ or not.

Remark that for the variant 2 of `Artemis`, the constant E given in Theorem 2.1 has a factor $\alpha_{\text{up}}^2 C(\omega+1)$: combining with the value of C, this term is indeed of the order of $(1+\omega_{\text{dwn}})(1+\omega_{\text{up}})$.

Without memory, we have the following computation:

$$\begin{aligned}
\operatorname{Tr}\left(\operatorname{Cov}(\xi(w_*, 0))\right) &= \mathbb{E}\left[\left\|\mathcal{C}_{\text{dwn}}\left(\frac{1}{N}\sum_{i=1}^N \mathcal{C}_{\text{up}}(g_1^i(w_*))\right)\right\|^2\right] \\
&\overset{\text{(i)}}{=} (1+\omega_{\text{dwn}})\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^N \mathcal{C}_{\text{up}}(g_1^i(w_*)) - h_*^i\right\|^2\right] \\
&\overset{\text{(ii)}}{=} \frac{(1+\omega_{\text{dwn}})}{N^2}\sum_{i=1}^N \mathbb{E}\left[\left\|\mathcal{C}_{\text{up}}(g_1^i(w_*)) - h_*^i\right\|^2\right] \\
&\overset{\text{(iii)}}{=} \frac{(1+\omega_{\text{dwn}})}{N^2}\sum_{i=1}^N \mathbb{E}\left[\left\|\mathcal{C}_{\text{up}}(g_1^i(w_*)) - g_1^i(w_*)\right\|^2 + \left\|g_1^i(w_*) - h_*^i\right\|^2\right]
\end{aligned}$$

At line (i) we use Assumption B.3 for the downlink compression operator with constant ω_{dwn} and the fact that $\sum_{i=1}^N h_*^i = \nabla F(w_*) = 0$, then at line (ii) the independence of the random variables $\mathcal{C}_{\text{up}}(g_1^i(w_*)) - h_*^i$ with mean 0, then a Bias Variance decomposition at line (iii).

$$\begin{aligned}
\text{Tr}\left(\text{Cov}(\xi(w_*, 0))\right) &\overset{\text{(iv)}}{=} \frac{(1+\omega_{\text{dwn}})}{N^2} \sum_{i=1}^N \mathbb{E}\left[\omega_{\text{up}} \left\|(g_1^i(w_*))\right\|^2 + \left\|g_1^i(w_*) - h_*^i\right\|^2\right] \\
&\overset{\text{(v)}}{=} \frac{(1+\omega_{\text{dwn}})}{N^2} \sum_{i=1}^N \mathbb{E}\left[\omega_{\text{up}} \left(\left\|g_1^i(w_*) - h_i^*\right\|^2 + \|h_i^*\|^2\right) + \left\|g_1^i(w_*) - h_*^i\right\|^2\right] \\
&\overset{\text{(vi)}}{=} \frac{(1+\omega_{\text{dwn}})}{N} \left((\omega_{\text{up}}+1)\frac{\sigma_*^2}{b} + \omega_{\text{up}} B^2\right).
\end{aligned}$$

Next we use Assumption B.3 for the uplink compression operator with constant ω_{up} at line (iv). Line (v) is another Bias-Variance decomposition and we finally conclude by using Assumptions B.1 and B.2 at line (vi) and reorganizing terms.

We have showed the lower bound both with or without memory, which concludes the proof. ■

C

Appendix to MCM

In this Chapter, we provide additional details about our work. First, in Section C.1, we enlarge figures provided in Section 3.5 and complete them with a comparison between MCM and other algorithms using non-degraded updates. The next sections are all devoted to theoretical results. In Section C.2, we detail some technical results required to demonstrate Theorems 3.3 to 3.5, 3.7 and 3.8, in Section C.3, we highlight the key stages of the demonstration in the easier case of Ghost, in Section C.4, we completely prove the given guarantees of convergence in three regimes: convex, strongly-convex and non-convex. In Section C.5, we show the benefit of Rand-MCM compared to MCM in the context of quadratic functions.

Contents

C.1 Experiments

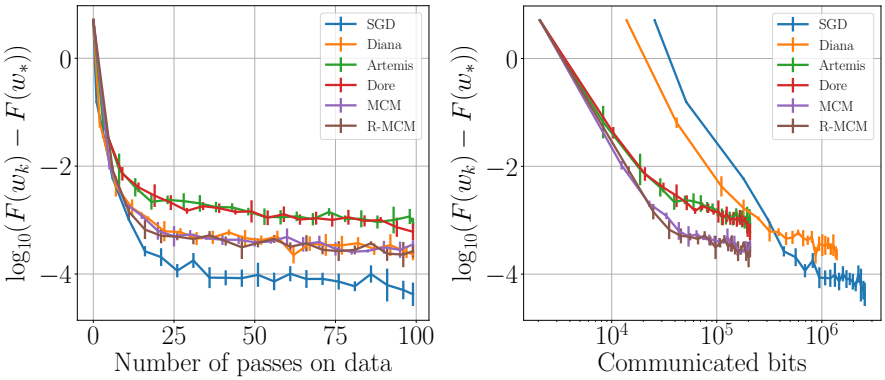
In this Section, we provide additional details about our experiments. We first give the settings of our experiments in Tables C.1 and C.2. Next, we describe the numerical results obtained on our 9 datasets. Finally, we provide an estimation of the carbon footprint required by this Chapter.

We use the same operator of compression for uplink and downlink, thus we consider that $\omega_{\rm up} = \omega_{\rm dwn}$. In addition, we choose $\alpha_{\rm up} = \alpha_{\rm dwn} = \dfrac{1}{2(1+\omega_{\rm up/dwn})}$.

Convex settings are given in Table C.1. We obtain non-i.i.d. data distributions by computing a TSNE representation [defined in Maaten and Hinton, 2008] followed by a clustering. Experiments have been performed with 200 epochs, we use quantization [defined in Alistarh et al., 2017] with $s = 2^0$.

Table C.1: Settings of experiment in the convex mode.

Settings	a9a	quantum	phishing	superconduct	w8a
references	[CL11]	[CTL04]	[CL11]	[Ham18]	[CL11]
model	LR	LR	LR	LSR	LR
dimension d	124	66	69	82	301
training dataset size	32,561	50,000	11,055	21,200	49,749
batch size b	218	256	64	64	12
compression rate s	2^0 (*i.e.* two levels)				
norm quantization	$\|\cdot\|_2$				
momentum m	no momentum				
step-size γ	$1/L$				



(a) X axis in # iterations. (b) X axis in # bits.

Figure C.1: Least-squares regression, toy dataset: $\gamma = (L\sqrt{k})^{-1}$, $\sigma \neq 0$.

Deep-learning settings are provided in Table C.2. All experiments have been performed with 300 epochs

Table C.2: Settings of experiments in the non-convex mode.

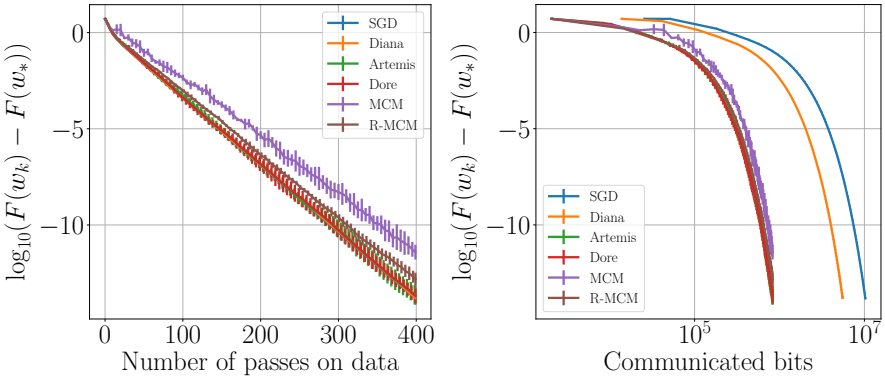
Settings	MNIST	Fashion-MNIST	FE-MNIST	CIFAR10
references	[LBBH98]	[XRV17]	[CJB+19]	[Kri09]
model	CNN	Fashion CNN	CNN	LeNet
trainable parameters d	20×10^3	400×10^3	20×10^3	62×10^3
training dataset size	$60,000$	$60,000$	$805,263$	$60,000$
compression rate s	2^2	2^2	2^2	2^4
momentum m	0	0	0	0.9
norm quantization	$\|\cdot\|_2$			
batch size b	128			
step-size γ	0.1			
loss	Cross Entropy			

C.1.1 Convex settings

In this section, we provide the plot of excess loss for the toy dataset, for quantum and for a9a datasets. For results on superconduct, phishing and w8a, see our github repository. For these last three datasets, we give only the excess loss w.r.t. number of iteration in the basic settings of full participation on Figure C.5. At the left side (resp. right side) we display the result w.r.t. the number of iterations (resp. number of communicated bits).

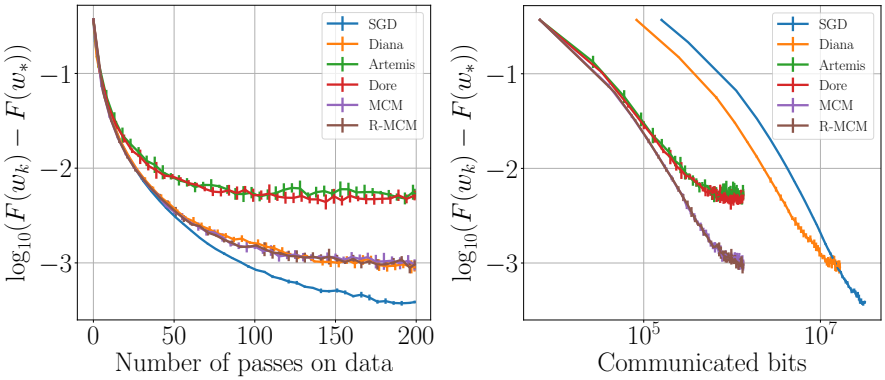
We provide results on the log of the excess loss $F(w_k) - F_*$, with error bars displayed on each figure, corresponding to the standard deviation of $\log_{10}(F(w_k) - F_*)$. Figures C.1b, C.2b, C.3 and C.4 correspond to Figures 3.2a, 3.2b and 3.3 given in Section 3.5. Additionally, we provide results for the synthetic dataset (Figures 3.2a and 3.2b) w.r.t to the number of iterations in Figure C.1 (stochastic gradient) and Figure C.2 (full batch gradient). As predicted by Theorem 3.4, when $\sigma = 0$, we observe a linear convergence.

On Figure C.6, we present a9a, quantum and phishing with a different operator of compression than in all other experiments. We use sparsification: each coordinate has a likelihood $p = 0.1$ to be selected. Unlike experiences with quantization for which $\omega = \sqrt{d}$, hence depending on the datasets' dimensionality, we have $\omega = (1-p)/p = 9$ for all datasets. A deeper analysis of the impact of each compressor is given in Chapter 4.



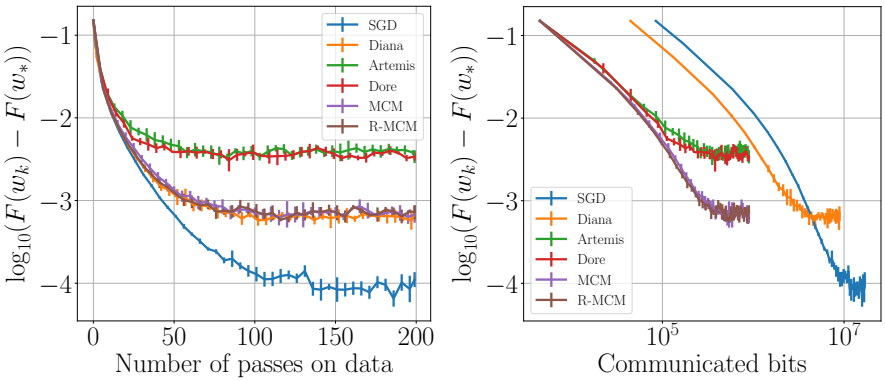
(a) X axis in # iterations. (b) X axis in # bits.

Figure C.2: Least-squares regression, toy dataset: $\gamma = 1/L$, $\sigma_*^2 = 0$.



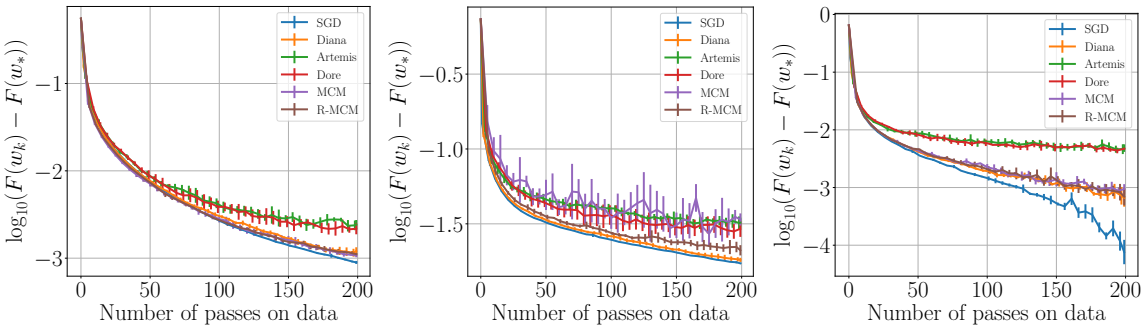
(a) X axis in # iterations. (b) X axis in # bits.

Figure C.3: a9a with $b = 128$, $\gamma = 1/L$.



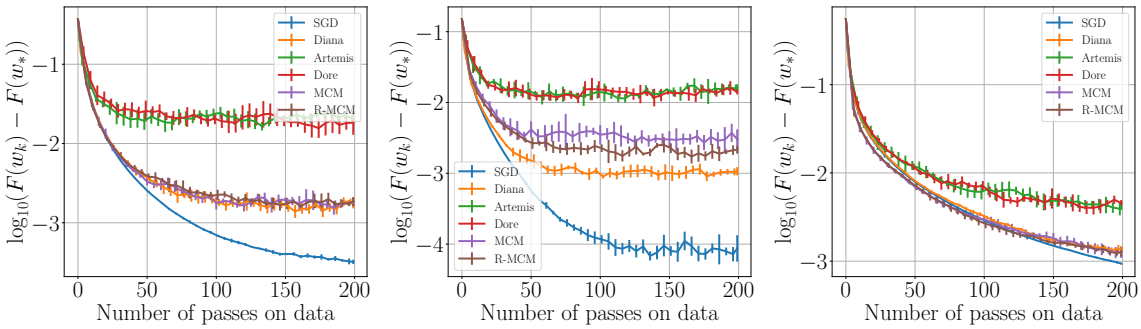
(a) X axis in # iterations. (b) X axis in # bits.

Figure C.4: quantum with $b = 256$, $\gamma = 1/L$.



(a) phishing – $b = 64$. (b) superconduct – $b = 64$. (c) w8a – $b = 12$.

Figure C.5: X axis in # iterations.



(a) a9a - $b = 128$. (b) quantum – $b = 256$. (c) phishing – $b = 64$.

Figure C.6: X axis in # iterations using random sparsification with $p = 0.1$.

C.1.1.1 Comparing MCM with other algorithm using non-degraded update

The aim of this section is to show the importance to set $\alpha < 1$, for this purpose we compare MCM with three other algorithms:

1. Artemis with a non-degraded update i.e. unlike the version proposed in Chapter 2, we do not update the global model with the compression sent to all remote nodes. *It means that we compress only the update that has already been performed on the global server.* It corresponds to:

$$\begin{cases} \forall i \in [\![1, N]\!], \Delta_k^i = g_{k+1}^i(\widehat{w}_k) - h_k^i \\ w_{k+1} = w_k - \frac{\gamma}{N} \sum_{i=1}^N \mathcal{C}_{\text{up}}(\Delta_k^i) + h_k^i \\ \widehat{w}_{k+1} = \widehat{w}_k - \gamma \mathcal{C}_{\text{dwn}} \left(\frac{1}{N} \sum_{i=1}^N \mathcal{C}_{\text{up}}(\Delta_k^i) + h_k^i \right) \\ h_{k+1}^i = h_k^i + \alpha_{\text{up}} \mathcal{C}_{\text{up}}(\Delta_k^i). \end{cases}$$

2. MCM with $\alpha = 0$, *thus without memory.*

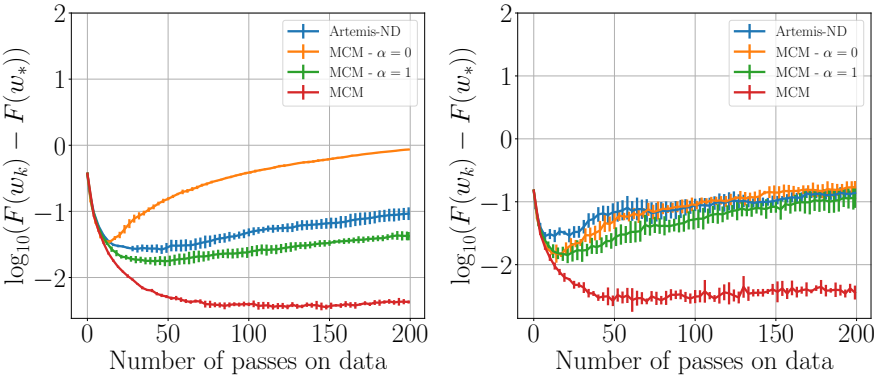
3. MCM with $\alpha = 1$, in other words, for k in $\mathbb{N}^*$ it corresponds to the case $H_{k+1} = \widehat{w}_{k+1}$. Indeed by definition we have $H_{k+1} = H_k + \alpha \widehat{\Omega}_{k+1}$, and furthermore, when we rebuild the compressed model on remote device, we have: $\widehat{w}_{k+1} = \widehat{\Omega}_{k+1} + H_k$. *In this case, we use the compressed model as memory.*

Figures C.7a and C.7b clearly show the superiority of MCM over the three other variants. Some conclusions can be drawn from the observation of these figures.

- MCM without downlink memory (orange curve, $\alpha = 0$) does not converge. As stressed in Subsection 3.2.1, this mechanism is crucial to control the variance of the local model w_{k+1}, for k in $\mathbb{N}$.

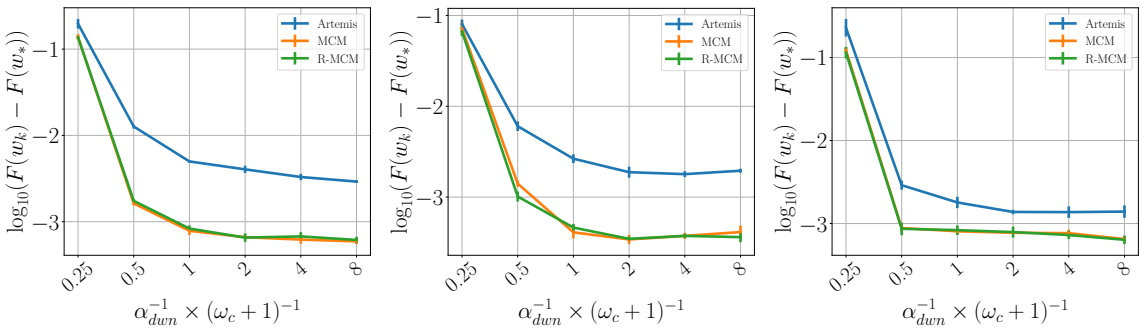
- Intuitively, while it appears reasonable to consider as memory the model that has been compressed at the previous step, experiments (green curves) show that this is not the case in practice and that α must be small enough to ensure convergence.

- Compressing only the update (blue curve) gives results similar to MCM with $\alpha = 1$. It means that combining model preservation and Artemis-like algorithms is not enough to guarantee convergence.



(a) a9a - $b = 128$. (b) quantum - $b = 256$.

Figure C.7: Comparing MCM with three other algorithms using a non-degraded update, $\gamma = 1/L$. Artemis-ND stands for Artemis with a non-degraded update.



(a) a9a – $b = 256$. (b) quantum – $b = 256$. (c) phishing – $b = 64$

Figure C.8: On X axis is displayed different values of $\frac{1}{\alpha(\omega_{\text{dwn}}+1)}$. On Y axis is given the excess loss after 250 epochs. In all other experiments, we choose $\alpha_{\text{dwn}} = \frac{1}{2(\omega_{\text{dwn}}+1)} (= \alpha_{\text{up}})$.

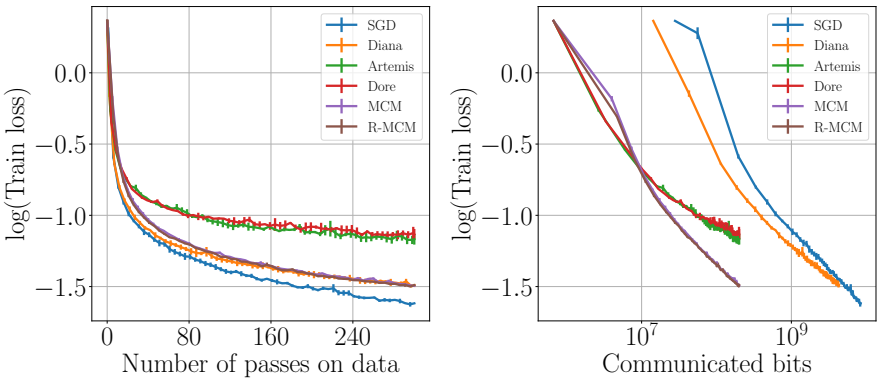
C.1.1.2 Impact of the learning rate α

On Figure C.8, we plot the value of the excess loss obtained after 250 epochs w.r.t. to the value of $\frac{1}{2(1+\omega_{\text{up/dwn}})}$. We observe that if α is too big, MCM converges slowly; but after reaching a threshold, the value of α does not impact anymore the rate of convergence. This confirms theory that suggests to use the largest possible α_{dwn} but smaller than a given value. The condition $\alpha_{\text{dwn}} \leq \frac{1}{4(\omega_{\text{dwn}}+1)}$ results from the proofs of Theorem C.3. But because the constant 4 is partially an artifact of the proof, in experiments we used $\alpha_{\text{dwn}} = \frac{1}{2(\omega_{\text{dwn}}+1)}$ as in Chapter 2 (Theorem B.2), and this choice is confirmed by Figure C.8.

C.1.2 Experiments in deep learning

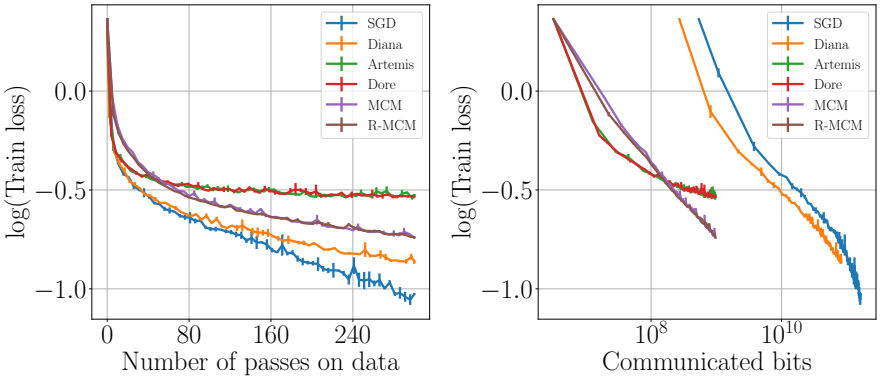
In this section, we show the robustness of MCM in high dimension using more complex data and applying the algorithm to non-convex problems (see Theorem C.6 for a guarantee of convergence in this scenario). We carried out experiments on MNIST/FE-MNIST/Fashion-MNIST using a CNN (Figures C.9 to C.11), and on CIFAR using the LeNet model (Figure C.12). We plot the logarithm of the train loss w.r.t the number of iterations and the number of communicated bits. The accuracy has been given in Section 3.5, see Table 3.4. Settings of the experiments can be found in Table C.2, all experiments are averaged over 2 runs.

As for experiments in convex case, MCM presents identical rates of convergence than Diana but with a small shift that makes Artemis better during the first iterations.



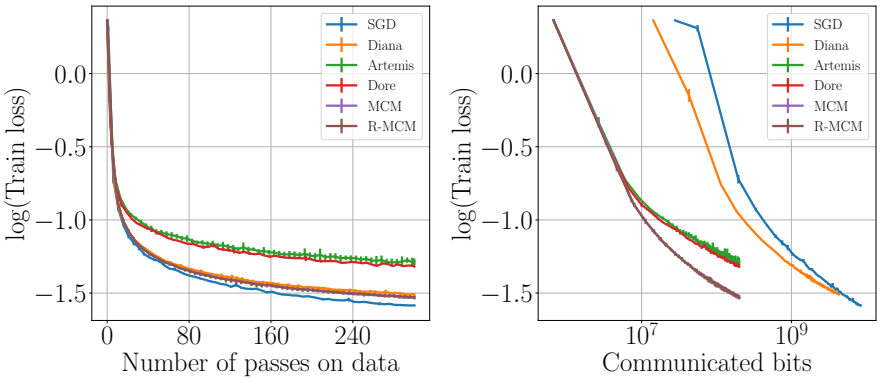
(a) X axis in # iterations. (b) X axis in # bits.

Figure C.9: Convergence on MNIST using a CNN.



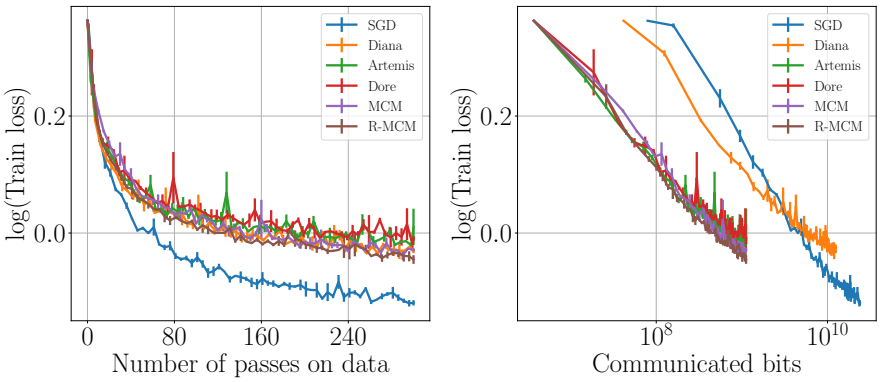
(a) X axis in # iterations. (b) X axis in # bits.

Figure C.10: Convergence on Fashion-MNIST.



(a) X axis in # iterations. (b) X axis in # bits.

Figure C.11: Convergence on FE-MNIST.



(a) X axis in # iterations. (b) X axis in # bits.

Figure C.12: Convergence on CIFAR10.

C.1.3 Hardware and Carbon footprint

As part as a community effort to report the carbon footprint of experiments, we describe in this subsection the hardware used and the total computation time.

We have two kind of experiments : for deep learning models we ran experiments on a GPU, and for linear/logistic regression on a CPU. We used an Intel(R) Xeon(R) CPU E5-2667 processor with 16 cores; and we used an Nvidia Tesla V100 GPU with 4 nodes.

To generate all figures in this chapter and in Chapter 3, our code ran (if run in a sequential mode) for 150 hours on a CPU. In overall, we consider that the whole chapter writing process required (code development, debugging, exploring settings ...) at least 600 hours end to end on the CPU. The carbon emissions caused by this work were subsequently evaluated with the `Green Algorithm`, built by Lannelongue et al. [2021]. It estimates our computations to generate around 100kg of CO2, requiring 2.5MWh. To compare, this corresponds to about 570km by car.

Overall, we consider that the full chapter writing process required at least 280 hours end to end on the GPU. The `Green Algorithm` estimates our computations to generate 220kg of CO2, requiring 5.7MWh. To compare, this corresponds to about $1,270$km by car.

C.2 Two lemmas

In this subsection, we give two lemmas required to prove the convergences of `Ghost`[1], `MCM` and `Rand-MCM`.

In Sections C.3 and C.4, for ease of notation we denote, for k in $\mathbb{N}^*$, $\widetilde{g}_k = \frac{1}{N}\sum_{i=1}^N \widehat{g}_k^i(\widehat{w}_{k-1})$. Furthermore we use the convention $\nabla F(w_{-1}) = 0$.

The first lemma will be used to show that `MCM` indeed satisfies Theorem 3.5. The proof is straightforward from the definition of w_k and H_{k-1}.

Lemma C.1 (Expectation of $w_k - H_{k-1}$)**.** *For any k in $\mathbb{N}^*$, the expectation of $(w_k - H_{k-1})$ conditionally to w_{k-1} can be decomposed as follows:*

$$\mathbb{E}\left[w_k - H_{k-1} \mid w_{k-1}\right] = (1-\alpha_{\mathrm{dwn}})(w_{k-1} - H_{k-2}) - \gamma \mathbb{E}\left[\nabla F(\widehat{w}_{k-1}) \mid w_{k-1}\right] .$$

Proof Let k in $\mathbb{N}^*$, by definition and with Assumption 3.1:

$$\begin{aligned}\mathbb{E}\left[w_k - H_{k-1} \mid w_{k-1}\right] &= \mathbb{E}\left[w_{k-1} - \gamma \widehat{g}_k(\widehat{w}_{k-1}) - (H_{k-2} + \alpha_{\mathrm{dwn}}\mathcal{C}(w_{k-1} - H_{k-2})) \mid w_{k-1}\right] \\ &= (w_{k-1} - H_{k-2}) - \alpha_{\mathrm{dwn}}\mathbb{E}\left[\mathcal{C}\left(w_{k-1} - H_{k-2}\right) \mid w_{k-1}\right] - \gamma \mathbb{E}\left[\widetilde{g}_k \mid w_{k-1}\right] ,\end{aligned}$$

from which the result follows.

■

The following lemma provides a control of the impact of the uplink compression. It decomposes the squared-norm of stochastic gradients into two terms: 1) the true gradient 2) the variance of the stochastic gradient σ^2.

Lemma C.2 (Squared-norm of stochastic gradients)**.** *For any k in $\mathbb{N}^*$, the second moment and*

[1] `Ghost` is defined in Subsection 3.2.3.

variance of the compressed gradients can be bounded a.s.:

$$\mathbb{E}\left[\|\widetilde{g}_k\|^2 \;\middle|\; \widehat{w}_{k-1}\right] \leq \left(1+\frac{\omega_{\mathrm{up}}}{N}\right)\|\nabla F(\widehat{w}_{k-1})\|^2 + \frac{\sigma^2(1+\omega_{\mathrm{up}})}{Nb}\,,$$
$$\mathbb{E}\left[\|\widetilde{g}_k - \nabla F(\widehat{w}_{k-1})\|^2 \;\middle|\; \widehat{w}_{k-1}\right] \leq \frac{\omega_{\mathrm{up}}}{N}\|\nabla F(\widehat{w}_{k-1})\|^2 + \frac{\sigma^2(1+\omega_{\mathrm{up}})}{Nb}\,.$$

Interpretation:

- *If $\omega_{\mathrm{up}} = 0$ (i.e. no up compression), the variance corresponds to a mini-batch.*
- *If $\sigma = 0$ and $N = 1$ (i.e. full batch descent with a single device), it becomes:*

$$\mathbb{E}\left[\|\mathcal{C}(\nabla F(w_{k-1})) - \nabla F(w_{k-1})\|^2\right] \leq \omega_{\mathrm{up}}\|\nabla F(w_{k-1})\|^2\,,$$

  which is consistent with Assumption 3.1.

Proof Let k in $\mathbb{N}^*$, then $\mathbb{E}\left[\|\widetilde{g}_k\|^2 \;\middle|\; \widehat{w}_{k-1}\right] = \|\nabla F(\widehat{w}_{k-1})\|^2 + \mathbb{E}\left[\|\widetilde{g}_k - \nabla F(\widehat{w}_{k-1})\|^2 \;\middle|\; \widehat{w}_{k-1}\right]$.

Secondly:

$$\begin{aligned}
&\mathbb{E}\left[\|\widetilde{g}_k - \nabla F(\widehat{w}_{k-1})\|^2 \;\middle|\; \widehat{w}_{k-1}\right] \\
&\quad= \mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^N \left(\widehat{g}_k^i(\widehat{w}_{k-1}) - g_k^i(\widehat{w}_{k-1}) + g_k^i(\widehat{w}_{k-1}) - \nabla F(\widehat{w}_{k-1})\right)\right\|^2 \;\middle|\; \widehat{w}_{k-1}\right] \\
&\quad= \mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^N \left(\widehat{g}_k^i(\widehat{w}_{k-1}) - g_k^i(\widehat{w}_{k-1})\right)\right\|^2 \;\middle|\; \widehat{w}_{k-1}\right] \\
&\qquad+ \mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^N \left(g_k^i(\widehat{w}_{k-1}) - \nabla F(\widehat{w}_{k-1})\right)\right\|^2 \;\middle|\; \widehat{w}_{k-1}\right]\,,
\end{aligned}$$

the inner product being null.

Next expanding the squared norm again, and because the two sums of inner products are null as the stochastic oracle and uplink compressions are independent:

$$\begin{aligned}
\mathbb{E}\left[\|\widetilde{g}_k - \nabla F(\widehat{w}_{k-1})\|^2 \;\middle|\; \widehat{w}_{k-1}\right] &= \frac{1}{N^2}\sum_{i=1}^N \mathbb{E}\left[\left\|\widehat{g}_k^i(\widehat{w}_{k-1}) - g_k^i(\widehat{w}_{k-1})\right\|^2 \;\middle|\; \widehat{w}_{k-1}\right] \\
&\quad+ \frac{1}{N^2}\sum_{i=1}^N \mathbb{E}\left[\left\|g_k^i(\widehat{w}_{k-1}) - \nabla F(\widehat{w}_{k-1})\right\|^2 \;\middle|\; \widehat{w}_{k-1}\right]\,.
\end{aligned}$$

Then using Assumption 3.1 we have:

$$\begin{aligned}
\mathbb{E}\left[\|\widetilde{g}_k - \nabla F(\widehat{w}_{k-1})\|^2 \;\middle|\; \widehat{w}_{k-1}\right] &= \frac{\omega_{\mathrm{up}}}{N^2}\sum_{i=1}^N \mathbb{E}\left[\left\|g_k^i(\widehat{w}_{k-1})\right\|^2 \;\middle|\; \widehat{w}_{k-1}\right] \\
&\quad+ \frac{1}{N^2}\sum_{i=1}^N \mathbb{E}\left[\left\|g_k^i(\widehat{w}_{k-1}) - \nabla F(\widehat{w}_{k-1})\right\|^2 \;\middle|\; \widehat{w}_{k-1}\right]\,.
\end{aligned}$$

Furthermore $\mathbb{E}\left[\left\|g_k^i(\widehat{w}_{k-1})\right\|^2 \;\middle|\; \widehat{w}_{k-1}\right] = \mathbb{E}\left[\left\|g_k^i(\widehat{w}_{k-1}) - \nabla F(\widehat{w}_{k-1})\right\|^2 \;\middle|\; \widehat{w}_{k-1}\right] + \|\nabla F(\widehat{w}_{k-1})\|^2$, and using Assumption 3.4:

$$\mathbb{E}\left[\|\widetilde{g}_k - \nabla F(\widehat{w}_{k-1})\|^2 \;\middle|\; \widehat{w}_{k-1}\right] = \frac{\omega_{\mathrm{up}}}{N}\|\nabla F(\widehat{w}_{k-1})\|^2 + \frac{\sigma^2(1+\omega_{\mathrm{up}})}{Nb}\,,$$

from which we derive the two inequalities of the lemma.

■

C.3 Proof for Ghost

C.3.1 Control of the Variance of the local model for Ghost (Proposition 3.1)

The proof of Proposition 3.1 is straightforward using Definition 3.1 that defines the Ghost algorithm.

Proposition C.1. *Consider the Ghost update in Equation (3.4), under Assumptions 3.1, 3.2 and 3.4, for all k in $\mathbb{N}$ with the convention $\nabla F(w_{-1}) = 0$:*

$$\mathbb{E}\left[\left\|w_k - \widehat{w}_k\right\|^2 \,\middle|\, \widehat{w}_{k-1}\right] \leq \gamma^2 \omega_{\mathrm{dwn}} \left(1 + \frac{\omega_{\mathrm{up}}}{N}\right) \left\|\nabla F(\widehat{w}_{k-1})\right\|^2 + \frac{\gamma^2 \omega_{\mathrm{dwn}}(1+\omega_{\mathrm{up}})\sigma^2}{Nb}\,.$$

Proof The proof of Proposition C.1 is straightforward using Definition 3.1. Let k in $\mathbb{N}$, by Definition 3.1 we have:

$$\begin{aligned}\left\|w_k - \widehat{w}_k\right\|^2 &= \left\| \left(w_{k-1} - \gamma \mathcal{C}_{\mathrm{dwn}} \left(\frac{1}{N} \sum_{i=1}^N \widehat{g}_k^i(\widehat{w}_{k-1}) \right) \right) - \left(w_{k-1} - \gamma \frac{1}{N} \sum_{i=1}^N \widehat{g}_k^i(\widehat{w}_{k-1}) \right) \right\|^2 \\ &= \gamma^2 \left\| \mathcal{C}_{\mathrm{dwn}} \left(\frac{1}{N} \sum_{i=1}^N \widehat{g}_k^i(\widehat{w}_{k-1}) \right) - \frac{1}{N} \sum_{i=1}^N \widehat{g}_k^i(\widehat{w}_{k-1}) \right\|^2 .\end{aligned}$$

Taking expectation w.r.t. down compression, as $\frac{1}{N}\sum_{i=1}^N \widehat{g}_k^i(\widehat{w}_{k-1})$ is w_k-measurable:

$$\mathbb{E}\left[\left\|w_k - \widehat{w}_k\right\|^2 \,\middle|\, w_k\right] = \gamma^2 \omega_{\mathrm{dwn}} \mathbb{E}\left[\left\| \frac{1}{N} \sum_{i=1}^N \widehat{g}_k^i(\widehat{w}_{k-1}) \right\|^2 \,\middle|\, w_k \right] = \gamma^2 \omega_{\mathrm{dwn}} \left\|\widetilde{g}_k\right\|^2 ,$$

and Lemma C.2 gives the upper bound $\mathbb{E}[\|\widetilde{g}_k\|^2 \mid \widehat{w}_{k-1}]$. ■

Theorem C.1 (Contraction for Ghost, convex case)**.** *Under Assumptions 3.1 to 3.4, with $\mu = 0$, if $\gamma L(1 + \omega_{\mathrm{up}}/N) \leq \frac{1}{2}$.*

$$\begin{aligned}\mathbb{E}[\|w_k - w_*\|^2] \leq \mathbb{E}[\|w_{k-1} - w_*\|^2] - \gamma \mathbb{E}[(F(w_{k-1}) - F_*)] - \frac{\gamma}{2L} \mathbb{E}\left[\left\|\nabla F(\widehat{w}_{k-1})\right\|^2\right] \\ + 2\gamma^3 \omega_{\mathrm{dwn}} L \left(1 + \frac{\omega_{\mathrm{up}}}{N}\right) \mathbb{E}[\|\nabla F(\hat{w}_{k-2})\|^2] + \gamma^2 \frac{(1+\omega_{\mathrm{up}})\sigma^2}{Nb} \left(1 + 2\gamma L \omega_{\mathrm{dwn}}\right).\end{aligned}$$

We can make the following observations:

1. At step k, the residual can be upper bounded by a constant times squared norm of the gradient at point $\hat{w}_{k-2}$. When using recursively this upper bound, if $2\gamma^3 \omega_{\mathrm{dwn}} L(1 + \omega_{\mathrm{up}}/N) \leq \gamma/(2L)$, then these terms cancel out. This is equivalent to $2\gamma L \sqrt{\omega_{\mathrm{dwn}}\,(1 + \omega_{\mathrm{up}}/N)} \leq 1$. It is natural to chose $\gamma \leq 1/(2L \max(1 + \omega_{\mathrm{up}}/N, 1 + \omega_{\mathrm{dwn}}))$.
2. The bound is in fact proved conditionally to w_{k-1}, recursive conditioning is required to propagate the inequality. We carefully handle conditioning in the proofs.

C.3.2 Convergence of Ghost, complete proof (Theorem 3.2)

In this Subsection, we provide the complete proof of convergence for Ghost. Thus, in the following demonstration, we give the key concepts required to later prove the convergence of MCM.

Theorem C.2 (Convergence of Ghost, convex case)**.** *Under Assumptions 3.1 to 3.4 with $\mu = 0$ (convex case), for all k in $\mathbb{N}$, defining $V_k := \mathbb{E}[w_k - w_*] + \frac{\gamma}{2L}\mathbb{E}[\|\nabla F(\widehat{w}_{k-1})\|^2] + 2\gamma L\mathbb{E}[\|\widehat{w}_k - w_k\|^2]$, we have:*

$$V_k \leq V_{k-1} - \gamma\mathbb{E}\left[F(w_{k-1}) - F(w_*)\right] + \frac{\gamma^2\sigma^2\Phi^{\mathcal{G}}(\gamma)}{Nb},$$

with $\Phi^{\mathcal{G}}(\gamma) := (1+\omega_{\text{up}})(1+2\gamma L\omega_{\text{dwn}})$.

Remark C.1. *This result is similar to Equation (3.8) but with a different function $\Phi^{\mathcal{G}}$ that has a weaker dependency on ω_{dwn}.*

Proof Let k in $\mathbb{N}^*$, by definition $\|w_k - w_*\|^2 \leq \|w_{k-1} - w_*\|^2 - 2\gamma\left\langle \widetilde{g}_k, w_{k-1} - w_*\right\rangle + \gamma^2\left\|\widetilde{g}_k\right\|^2$. Next, we expend the inner product as following:

$$\|w_k - w_*\|^2 \leq \|w_{k-1} - w_*\|^2 - 2\gamma\left\langle \widetilde{g}_k, \widehat{w}_{k-1} - w_*\right\rangle - 2\gamma\left\langle \widetilde{g}_k, w_{k-1} - \widehat{w}_{k-1}\right\rangle + \gamma^2\left\|\widetilde{g}_k\right\|^2.$$

Taking expectation conditionally to w_{k-1}, and using $\mathbb{E}\left[\widetilde{g}_k \mid w_{k-1}\right] = \mathbb{E}\left[\mathbb{E}\left[\widetilde{g}_k \mid \widehat{w}_{k-1}\right] \mid w_{k-1}\right] = \mathbb{E}\left[\nabla F(\widehat{w}_{k-1}) \mid w_{k-1}\right]$, we obtain:

$$\begin{aligned}\mathbb{E}\left[\|w_k - w_*\|^2 \;\middle|\; w_{k-1}\right] \leq \|w_{k-1} - w_*\|^2 &- \mathbb{E}\left[2\gamma\left\langle \nabla F(\widehat{w}_{k-1}), \widehat{w}_{k-1} - w_*\right\rangle \mid w_{k-1}\right] \\ &- 2\gamma\mathbb{E}\left[\left\langle \nabla F(\widehat{w}_{k-1}), w_{k-1} - \widehat{w}_{k-1}\right\rangle \mid w_{k-1}\right] \\ &+ \gamma^2\mathbb{E}\left[\left\|\widetilde{g}_k\right\|^2 \;\middle|\; w_{k-1}\right].\end{aligned}$$

Then, invoking Lemma C.2 to upper bound the squared norm of the stochastic gradients, and noticing that $\mathbb{E}\left[\left\langle \nabla F(w_{k-1}), \widehat{w}_{k-1} - w_{k-1}\right\rangle \mid w_{k-1}\right] = 0$ leads to:

$$\begin{aligned}\mathbb{E}\left[\|w_k - w_*\|^2 \;\middle|\; w_{k-1}\right] \leq \|w_{k-1} - w_*\|^2 &- 2\gamma\mathbb{E}\left[\left\langle \nabla F(\widehat{w}_{k-1}), \widehat{w}_{k-1} - w_*\right\rangle \mid w_{k-1}\right] \\ &- 2\gamma\mathbb{E}\left[\left\langle \nabla F(\widehat{w}_{k-1}) - \nabla F(w_{k-1}), w_{k-1} - \widehat{w}_{k-1}\right\rangle \mid w_{k-1}\right] \\ &+ \gamma^2\left(\left(1 + \frac{\omega_{\text{up}}}{Nb}\right)\mathbb{E}\left[\|\nabla F(\widehat{w}_{k-1})\|^2 \;\middle|\; w_{k-1}\right] + \frac{\sigma^2\left(1+\omega_{\text{up}}\right)}{Nb}\right).\end{aligned} \tag{C.1}$$

In the upper inequality:

1. the term $\mathbb{E}\left[\left\langle \nabla F(\widehat{w}_{k-1}), \widehat{w}_{k-1} - w_*\right\rangle \mid w_{k-1}\right]$ allows the "strong contraction"
2. the terms $\mathbb{E}\left[\left\langle \nabla F(\widehat{w}_{k-1}) - \nabla F(w_{k-1}), w_{k-1} - \widehat{w}_{k-1}\right\rangle \mid w_{k-1}\right]$ and $\mathbb{E}[\|\nabla F(\widehat{w}_{k-1})\|^2 \mid w_{k-1}]$ are two positives terms that we treat as residuals.
3. the last term $\sigma^2\left(1+\omega_{\text{up}}\right)/(Nb)$ is due to the stochastic noise.

Now using Cauchy-Schwarz's inequality (Equation (A.4)) and smoothness:

$$\begin{aligned}&-\mathbb{E}\left[2\gamma\left\langle \nabla F(\widehat{w}_{k-1}) - \nabla F(w_{k-1}), w_{k-1} - \widehat{w}_{k-1}\right\rangle \mid w_{k-1}\right] \\ &\quad= 2\gamma\mathbb{E}\left[\left\langle \nabla F(\widehat{w}_{k-1}) - \nabla F(w_{k-1}), \widehat{w}_{k-1} - w_{k-1}\right\rangle \mid w_{k-1}\right] \\ &\quad\leq 2\gamma L\mathbb{E}\left[\|\widehat{w}_{k-1} - w_{k-1}\|^2 \;\middle|\; w_{k-1}\right],\end{aligned}$$

and thus:

$$\begin{aligned}\mathbb{E}\left[\|w_k - w_*\|^2 \;\middle|\; w_{k-1}\right] \leq \|w_{k-1} - w_*\|^2 &- 2\gamma\mathbb{E}\left[\langle \nabla F(\widehat{w}_{k-1}), \widehat{w}_{k-1} - w_*\rangle \mid w_{k-1}\right] \\ &+ 2\gamma L\mathbb{E}\left[\|\widehat{w}_{k-1} - w_{k-1}\|^2 \;\middle|\; w_{k-1}\right] \\ &+ \gamma^2\left(1 + \frac{\omega_{\text{up}}}{N}\right)\mathbb{E}\left[\|\nabla F(\widehat{w}_{k-1})\|^2 \;\middle|\; w_{k-1}\right] + \frac{\gamma^2\sigma^2(1+\omega_{\text{up}})}{Nb}\,.\end{aligned} \tag{C.2}$$

Now, using convexity with Proposition A.2:

$$\begin{aligned}\mathbb{E}\left[\|w_k - w_*\|^2 \;\middle|\; w_{k-1}\right] \leq \|w_{k-1} - w_*\|^2 &\\ - \gamma\mathbb{E}\left[\left(F(\widehat{w}_{k-1}) - F(w_*) + \frac{1}{L}\|\nabla F(\widehat{w}_{k-1})\|^2\right) \;\middle|\; w_{k-1}\right] &\\ + 2\gamma L\mathbb{E}\left[\|\widehat{w}_{k-1} - w_{k-1}\|^2 \;\middle|\; w_{k-1}\right] &\\ + \gamma^2\left(1 + \frac{\omega_{\text{up}}}{N}\right)\mathbb{E}\left[\|\nabla F(\widehat{w}_{k-1})\|^2 \;\middle|\; w_{k-1}\right] + \frac{\gamma^2\sigma^2(1+\omega_{\text{up}})}{Nb}\,.&\end{aligned}$$

Taking the full expectation (without conditioning over any random vectors), and because invoking Jensen's inequality (A.7) leads to $\mathbb{E}\left[F(\widehat{w}_{k-1})\right] \geq \mathbb{E}\left[F(w_{k-1})\right]$, we finally obtain this intermediate result:

$$\begin{aligned}\mathbb{E}\left[\|w_k - w_*\|^2\right] \leq \mathbb{E}\left[\|w_{k-1} - w_*\|^2\right] &- \gamma\left(\mathbb{E}\left[F(w_{k-1})\right] - F(w_*)\right) \\ &- \frac{\gamma}{2L}\mathbb{E}\left[\|\nabla F(\widehat{w}_{k-1})\|^2\right] \\ &+ 2\gamma L\mathbb{E}\left[\|\widehat{w}_{k-1} - w_{k-1}\|^2\right] + \frac{\gamma^2\sigma^2(1+\omega_{\text{up}})}{Nb}\,,\end{aligned} \tag{C.3}$$

where we considered that $\gamma L(1 + \omega_{\text{up}}/N) \leq 1/2$, which implies that $\gamma\left(1 - \gamma L\left(1 + \frac{\omega_{\text{up}}}{N}\right)\right) \geq \frac{\gamma}{2}$.

Remark that Equation (C.3) is valid for both `Ghost` and `MCM`, and that the proof of `MCM` will follow the same initial line. Now, because $\mathbb{E}\left[\|w_{k-1} - \widehat{w}_{k-1}\|^2\right] = \mathbb{E}\left[\mathbb{E}\left[\|w_{k-1} - \widehat{w}_{k-1}\|^2 \;\middle|\; \widehat{w}_{k-2}\right]\right]$, we can use Proposition 3.1 which is specific to `Ghost` and we recover Theorem 3.1:

$$\begin{aligned}\mathbb{E}\|w_k - w_*\|^2 \leq \mathbb{E}\,\|w_{k-1} - w_*\|^2 - \gamma\mathbb{E}(F(w_{k-1}) - F_*) - \frac{\gamma}{2L}\mathbb{E}\left[\|\nabla F(\widehat{w}_{k-1})\|^2\right] \\ + 2\gamma^3\omega_{\text{dwn}}L\left(1 + \frac{\omega_{\text{up}}}{N}\right)\mathbb{E}\,\|\nabla F(\hat{w}_{k-2})\|^2 + \gamma^2\frac{(1+\omega_{\text{up}})\sigma^2}{Nb}\left(1 + 2\gamma L\omega_{\text{dwn}}\right).\end{aligned}$$

As a reminder, Proposition 3.1 gives the following contraction; we use it now to define a Lyapunov function:

$$\mathbb{E}\left[\|w_k - \widehat{w}_k\|^2 \;\middle|\; \widehat{w}_{k-1}\right] \leq \gamma^2\omega_{\text{dwn}}\left(1 + \frac{\omega_{\text{up}}}{N}\right)\|\nabla F(\widehat{w}_{k-1})\|^2 + \frac{\gamma^2\omega_{\text{dwn}}(1+\omega_{\text{up}})\sigma^2}{Nb}\,. \tag{C.4}$$

Defining $V_k := \mathbb{E}\left[w_k - w_*\right] + \frac{\gamma}{2L}\mathbb{E}\left[\|\nabla F(\widehat{w}_{k-1})\|^2\right] + C\mathbb{E}[\|\widehat{w}_k - w_k\|^2]$ with $C = 2\gamma L$, and combining this two equations as following (C.3) + C(C.4) leads to:

$$\begin{aligned}\mathbb{E}\left[\|w_k - w_*\|^2\right] + C\mathbb{E}\left[\|\widehat{w}_{k-1} - w_{k-1}\|^2\right] + \frac{\gamma}{2L}\mathbb{E}\left[\|\nabla F(\widehat{w}_{k-1})\|^2\right] &\\ \leq \mathbb{E}\left[\|w_{k-1} - w_*\|^2\right] - \gamma\left(\mathbb{E}\left[F(w_{k-1})\right] - F(w_*)\right) &\\ + 2\gamma L\mathbb{E}\left[\|\widehat{w}_{k-1} - w_{k-1}\|^2\right] + \frac{\gamma^2\sigma^2(1+\omega_{\text{up}})}{Nb} &\\ + 2\gamma L \times \gamma^2\omega_{\text{dwn}}\left(1 + \frac{\omega_{\text{up}}}{N}\right)\|\nabla F(\widehat{w}_{k-1})\|^2 + 2\gamma L \times \frac{\gamma^2\omega_{\text{dwn}}(1+\omega_{\text{up}})\sigma^2}{Nb}\,.&\end{aligned}$$

To ensure a contraction of the Lyapunov function we require:

$$\gamma^2\omega_{\text{dwn}}\left(1+\frac{\omega_{\text{up}}}{N}\right) \leq \frac{\gamma}{2L} \iff \gamma L \leq \frac{1}{2\sqrt{\omega_{\text{dwn}}\left(1+\frac{\omega_{\text{up}}}{N}\right)}}$$

Under this condition, we obtain $V_k \leq V_{k-1} - \gamma\mathbb{E}\left[F(w_{k-1}) - F(w_*)\right] + \frac{\gamma^2\sigma^2\Phi^{\mathcal{G}}(\gamma)}{Nb}$ with $\Phi^{\mathcal{G}}(\gamma) := (1+\omega_{\text{up}})(1+2\gamma L\omega_{\text{dwn}})$. By recurrence and for $k = K$:

$$V_K \leq V_0 - \sum_{k=1}^{K}\gamma\mathbb{E}\left[F(w_{k-1}) - F(w_*)\right] + \sum_{k=1}^{K}\frac{\gamma^2\sigma^2\Phi^{\mathcal{G}}(\gamma)}{Nb}\,,$$

which leads to: $\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[F(w_{k-1}) - F(w_*)\right] \leq \frac{V_0 - V_k}{\gamma K} + \frac{\gamma\sigma^2\Phi^{\mathcal{G}}(\gamma)}{Nb}$. Finally, for any K in $\mathbb{N}^*$, with $\gamma L \leq \min\left\{\frac{1}{2\left(1+\frac{\omega_{\text{up}}}{N}\right)}, \frac{1}{2\sqrt{\omega_{\text{dwn}}\left(1+\frac{\omega_{\text{up}}}{N}\right)}}\right\}$ we have:

$$\frac{\gamma}{K}\sum_{t=1}^{K}\mathbb{E}\left[F(w_t) - F(w_*)\right] \leq \frac{\|w_0 - w_*\|^2}{K} + \frac{\gamma\sigma^2\Phi^{\mathcal{G}}(\gamma)}{Nb}\,.$$

Note that the bound of γL encompasses the case $\omega_{\text{dwn}} = 0$ (i.e. no downlink compression), but in the general case of bidirectional compression, we nearly always have $\omega_{\text{dwn}} > 1$, and thus the dominant term is, in fact, $\frac{1}{2\sqrt{\omega_{\text{dwn}}\left(1+\frac{\omega_{\text{up}}}{N}\right)}}$. And by Jensen, it implies that:

$$\mathbb{E}\left[F(\bar{w}_K) - F(w_*)\right] \leq \frac{\|w_0 - w_*\|^2}{\gamma K} + \frac{\gamma\sigma^2\Phi(\gamma)}{Nb} \quad \text{with } \Phi^{\mathcal{G}}(\gamma) := (1+\omega_{\text{up}})(1+2\omega_{\text{dwn}}\gamma L)\,.$$

■

C.4 Proofs for MCM (and Rand-MCM)

In this Section, we provide the proofs for MCM in the convex, strongly-convex, and non-convex cases in respectively Theorems C.4 to C.6. The proofs for Rand-MCM (see Theorem 3.7) are identical and only require to adapt notations as explained in Subsection C.4.5.

We denote for γ in $\mathbb{R}$, $\Phi(\gamma) := (1+\omega_{\text{up}})\left(1+\frac{8\gamma L\omega_{\text{dwn}}}{\alpha_{\text{dwn}}}\right)$, for k in $\mathbb{N}$, $\Upsilon_k = \|w_k - H_{k-1}\|^2$ and we define $\gamma_{\max}$ such that:

$$\gamma_{\max}L \leq \min\left\{\frac{1}{8\omega_{\text{dwn}}}, \frac{1}{2\left(1+\frac{\omega_{\text{up}}}{N}\right)}, \frac{1}{4\sqrt{\frac{\omega_{\text{dwn}}}{\alpha_{\text{dwn}}}\left(\frac{1}{\alpha_{\text{dwn}}}+\frac{\omega_{\text{up}}}{N}\right)}}\right\}.$$

Note that this is equivalent to notations given in Section 3.3 if we take $\alpha_{\text{dwn}} = 1/8\omega_{\text{dwn}}$.

C.4.1 Control of the Variance of the local model for MCM (Theorem 3.5)

In this Subsection, we provide a control of the variance of the local model for MCM, as done previously in Proposition C.1 for Ghost: this corresponds to Theorem 3.5. The demonstration is more complex than for Ghost and it highlights the trade-offs for the learning rate α_{dwn}. The demonstration builds a bias-variance decomposition of $\|\Omega_k\|^2 = \|w_k - H_k\|^2$. The variance is then decomposed in three terms, as a result we will need to compute four terms:

$$\|w_k - H_{k-1}\|^2 = \text{Bias}^2 + 2\gamma^2(\text{Var}_{11} + \text{Var}_{12}) + 2\alpha_{\text{dwn}}^2 \text{Var}_2 \,. \tag{C.5}$$

Theorem C.3. *Consider the MCM update as in Equation (3.2). Under Assumptions 3.1, 3.2 and 3.4 with $\mu = 0$, if $\gamma \leq (8\omega_{\text{dwn}} L)^{-1}$ and $\alpha_{\text{dwn}} \leq (8\omega_{\text{dwn}})^{-1}$, then for all k in $\mathbb{N}$:*

$$\mathbb{E}\left[\Upsilon_k\right] \leq \left(1 - \frac{\alpha_{\text{dwn}}}{2}\right) \mathbb{E}\left[\Upsilon_{k-1}\right] + 2\gamma^2 \left(\frac{1}{\alpha_{\text{dwn}}} + \frac{\omega_{\text{up}}}{N}\right) \mathbb{E}\left[\|\nabla F(\widehat{w}_{k-1})\|^2\right] + \frac{2\gamma^2\sigma^2(1+\omega_{\text{up}})}{Nb} \,.$$

Proof Let k in $\mathbb{N}$, we recall that by definition:

$$\begin{cases} \Omega_k = w_k - H_{k-1} \\ \widehat{\Omega}_k = \mathcal{C}_{\text{dwn}}(\Omega_k) \\ \widehat{w}_k = \widehat{\Omega}_k + H_{k-1} \,. \end{cases}$$

We start the proof by introducing $\|\Omega_k\|^2$: $\mathbb{E}[\|w_k - \widehat{w}_k\|^2 \mid w_k] = \mathbb{E}[\left\|\widehat{\Omega}_k - \Omega_k\right\|^2 \mid w_k] \leq \omega_{\text{dwn}} \|\Omega_k\|^2$. Next, we perform a bias-variance decomposition:

$$\begin{aligned} \|\Omega_k\|^2 = \|w_k - H_{k-1}\|^2 = &\|w_k - H_{k-1} - \mathbb{E}\left[w_k - H_{k-1} \mid w_{k-1}\right]\|^2 \\ &+ \|\mathbb{E}\left[w_k - H_{k-1} \mid w_{k-1}\right]\|^2 \\ &+ 2\left\langle w_k - H_{k-1} - \mathbb{E}\left[w_k - H_{k-1} \mid w_{k-1}\right], \mathbb{E}\left[w_k - H_{k-1} \mid w_{k-1}\right]\right\rangle \,, \end{aligned}$$

taking expectation w.r.t. w_{k-1}:

$$\mathbb{E}\left[\Upsilon_k \mid w_{k-1}\right] = \underbrace{\mathbb{E}\left[\|w_k - H_{k-1} - \mathbb{E}\left[w_k - H_{k-1} \mid w_{k-1}\right]\|^2 \,\middle|\, w_{k-1}\right]}_{\text{Var}} + \underbrace{\|\mathbb{E}\left[w_k - H_{k-1} \mid w_{k-1}\right]\|^2}_{\text{Bias}^2} \,.$$

The first term is the variance Var, and the second term corresponds to the squared bias Bias^2.

Let's handle first the variance, by definition:

$$\begin{aligned} \text{Var} &= \mathbb{E}\left[\|w_k - H_{k-1} - \mathbb{E}\left[w_k - H_{k-1} \mid w_{k-1}\right]\|^2 \,\middle|\, w_{k-1}\right] \\ &= \mathbb{E}\left[\|w_{k-1} - \gamma\widetilde{g}_k - H_{k-2} - \alpha_{\text{dwn}}\mathcal{C}(w_{k-1} - H_{k-2}) \right. \\ &\qquad \left. - w_{k-1} - \gamma\mathbb{E}\left[\widetilde{g}_k \mid w_{k-1}\right] - H_{k-2} - \alpha_{\text{dwn}}\mathbb{E}\left[\mathcal{C}(w_{k-1} - H_{k-2} \mid w_{k-1})\right)\|^2 \middle| w_{k-1}\right] \,. \end{aligned}$$

After simplification and using Equation (A.3):

$$\begin{aligned} \text{Var} &= \mathbb{E}\left[\| - \gamma\left(\widetilde{g}_k + \mathbb{E}\left[\nabla F(\widehat{w}_{k-1}) \mid w_{k-1}\right]\right) + \alpha_{\text{dwn}}\left(\mathcal{C}(w_{k-1} - H_{k-2})\right) \right. \\ &\qquad \left. - (w_{k-1} - H_{k-2})\|^2 \middle| w_{k-1}\right] \\ &\leq 2\gamma^2 \mathbb{E}\left[\|\widetilde{g}_k - \mathbb{E}\left[\nabla F(\widehat{w}_{k-1}) \mid w_{k-1}\right]\|^2 \,\middle|\, w_{k-1}\right] \\ &\qquad + 2\alpha_{\text{dwn}}^2 \mathbb{E}\left[\|\mathcal{C}(w_{k-1} - H_{k-2}) - (w_{k-1} - H_{k-2})\|^2 \,\middle|\, w_{k-1}\right] \\ &\leq 2\gamma^2 \underbrace{\mathbb{E}\left[\|\widetilde{g}_k - \mathbb{E}\left[\nabla F(\widehat{w}_{k-1}) \mid w_{k-1}\right]\|^2 \,\middle|\, w_{k-1}\right]}_{\text{Var}_1} + 2\alpha_{\text{dwn}}^2 \underbrace{\omega_{\text{dwn}} \|w_{k-1} - H_{k-2}\|^2}_{\text{Var}_2} \\ &\leq 2\gamma^2 \text{Var}_1 + 2\alpha_{\text{dwn}}^2 \text{Var}_2 \,. \end{aligned}$$

An interpretation of the above decomposition is that:

- Var_1 is the part of the downlink compression caused by the increment $\widetilde{g}_k$, it is similar to `Ghost`.
- Var_2 is the impact of the propagation of the previous noise.

We compute the first term by introducing $\nabla F(\widehat{w}_{k-1})$, the second being kept as it is:

$$\begin{aligned}
\mathrm{Var}_1 &= \mathbb{E}\left[\left\|\widetilde{g}_k - \mathbb{E}\left[\nabla F(\widehat{w}_{k-1}) \mid w_{k-1}\right]\right\|^2 \,\middle|\, w_{k-1}\right] \\
&= \mathbb{E}\left[\left\|\widetilde{g}_k - \nabla F(\widehat{w}_{k-1}) + \nabla F(\widehat{w}_{k-1}) - \mathbb{E}\left[\nabla F(\widehat{w}_{k-1}) \mid w_{k-1}\right]\right\|^2 \,\middle|\, w_{k-1}\right] \\
&= \underbrace{\mathbb{E}\left[\left\|\widetilde{g}_k - \nabla F(\widehat{w}_{k-1})\right\|^2 \,\middle|\, w_{k-1}\right]}_{\mathrm{Var}_{11}} + \underbrace{\mathbb{E}\left[\left\|\nabla F(\widehat{w}_{k-1}) - \mathbb{E}\left[\nabla F(\widehat{w}_{k-1}) \mid w_{k-1}\right]\right\|^2 \,\middle|\, w_{k-1}\right]}_{\mathrm{Var}_{12}} \\
&= \mathrm{Var}_{11} + \mathrm{Var}_{12}\,,
\end{aligned}$$

the inner product is null given that $\mathbb{E}\left[\nabla F(\widehat{w}_{k-1}) - \mathbb{E}\left[\nabla F(\widehat{w}_{k-1}) \mid w_{k-1}\right] \mid w_{k-1}\right] = 0$. Moreover:

$$\mathrm{Var}_{11} = \mathbb{E}\left[\left\|\widetilde{g}_k - \nabla F(\widehat{w}_{k-1})\right\|^2 \,\middle|\, w_{k-1}\right] = \mathbb{E}\left[\mathbb{E}\left[\left\|\widetilde{g}_k - \nabla F(\widehat{w}_{k-1})\right\|^2 \,\middle|\, \widehat{w}_{k-1}\right] \,\middle|\, w_{k-1}\right]\,,$$

so, we can use Lemma C.2: $\mathrm{Var}_{11} = \mathbb{E}\left[\frac{\sigma^2}{Nb}(1+\omega_{\mathrm{up}}) + \frac{\omega_{\mathrm{up}}}{N}\left\|\nabla F(\widehat{w}_{k-1})\right\|^2 \,\middle|\, w_{k-1}\right]$. And now we use smoothness for the second term:

$$\begin{aligned}
\mathrm{Var}_{12} &= \mathbb{E}\left[\left\|\nabla F(\widehat{w}_{k-1}) - \mathbb{E}\left[\nabla F(\widehat{w}_{k-1}) \mid w_{k-1}\right]\right\|^2 \,\middle|\, w_{k-1}\right] \\
&\leq \mathbb{E}\left[\left\|\nabla F(\widehat{w}_{k-1}) - \nabla F(w_{k-1})\right\|^2 \,\middle|\, w_{k-1}\right] \quad \text{by Lemma A.3,} \\
&\leq L^2\mathbb{E}\left[\left\|\widehat{w}_{k-1} - w_{k-1}\right\|^2 \,\middle|\, w_{k-1}\right] \quad \text{using smoothness,} \\
&\leq L^2\omega_{\mathrm{dwn}}\Upsilon_{k-1} \quad \text{with Assumption 3.1}\,.
\end{aligned}$$

At the end:

$$\begin{aligned}
\mathrm{Var} \leq 2\gamma^2 &\left(\frac{\sigma^2(1+\omega_{\mathrm{up}})}{Nb} + \frac{\omega_{\mathrm{up}}}{N}\mathbb{E}\left[\left\|\nabla F(\widehat{w}_{k-1})\right\|^2 \,\middle|\, w_{k-1}\right] + L^2\omega_{\mathrm{dwn}}\Upsilon_{k-1}\right) \\
&+ 2\alpha_{\mathrm{dwn}}^2\omega_{\mathrm{dwn}}\Upsilon_{k-1}\,.
\end{aligned} \tag{C.6}$$

Now we focus on the squared bias Bias^2, with Lemma C.1:

$$\begin{aligned}
\mathrm{Bias}^2 &= \left\|\mathbb{E}\left[w_k - H_{k-1} \mid w_{k-1}\right]\right\|^2 \\
&= \left\|(1-\alpha_{\mathrm{dwn}})(w_{k-1} - H_{k-2}) - \gamma\mathbb{E}\left[\nabla F(\widehat{w}_{k-1}) \mid w_{k-1}\right]\right\|^2\,, \quad \text{and with Equation (A.2)}\,, \\
&\leq (1-\alpha_{\mathrm{dwn}})^2\left(1+\alpha_{\mathrm{dwn}}\right)\Upsilon_{k-1} + \gamma^2\left(1+\frac{1}{\alpha_{\mathrm{dwn}}}\right)\left\|\mathbb{E}\left[\nabla F(\widehat{w}_{k-1}) \mid w_{k-1}\right]\right\|^2\,.
\end{aligned}$$

And because $(1-\alpha_{\mathrm{dwn}})(1+\alpha_{\mathrm{dwn}}) < 1$, we finally get that:

$$\mathrm{Bias}^2 \leq (1-\alpha_{\mathrm{dwn}})\Upsilon_{k-1} + \gamma^2\left(1+\frac{1}{\alpha_{\mathrm{dwn}}}\right)\left\|\mathbb{E}\left[\nabla F(\widehat{w}_{k-1}) \mid w_{k-1}\right]\right\|^2\,. \tag{C.7}$$

Combining all Equations (C.6) and (C.7) into Equation (C.5):

$$\begin{aligned}
\mathbb{E}\left[\Upsilon_k \mid w_{k-1}\right] \leq\ & (1-\alpha_{\mathrm{dwn}})\Upsilon_{k-1} + \gamma^2\left(1+\frac{1}{\alpha_{\mathrm{dwn}}}\right)\left\|\mathbb{E}\left[\nabla F(\widehat{w}_{k-1}) \mid w_{k-1}\right]\right\|^2 \\
&+ 2\gamma^2\left(\frac{\sigma^2(1+\omega_{\mathrm{up}})}{Nb} + \frac{\omega_{\mathrm{up}}}{N}\mathbb{E}\left[\left\|\nabla F(\widehat{w}_{k-1})\right\|^2 \,\middle|\, w_{k-1}\right]\right) \\
&+ 2\gamma^2\left(L^2\omega_{\mathrm{dwn}}\Upsilon_{k-1}\right) + 2\alpha_{\mathrm{dwn}}^2\omega_{\mathrm{dwn}}\Upsilon_{k-1}\,,
\end{aligned}$$

that is:

$$\begin{aligned}\mathbb{E}\left[\Upsilon_k \mid w_{k-1}\right] \leq & \left(1-\alpha_{\mathrm{dwn}}+2\gamma^2 L^2 \omega_{\mathrm{dwn}}+2\alpha_{\mathrm{dwn}}^2 \omega_{\mathrm{dwn}}\right)\left\|w_{k-1}-H_{k-2}\right\|^2 \\ & +\gamma^2\left(1+\frac{1}{\alpha_{\mathrm{dwn}}}\right)\left\|\mathbb{E}\left[\nabla F(\widehat{w}_{k-1}) \mid w_{k-1}\right]\right\|^2 \\ & +\frac{2\gamma^2\omega_{\mathrm{up}}}{N}\mathbb{E}\left[\left\|\nabla F(\widehat{w}_{k-1})\right\|^2 \,\middle|\, w_{k-1}\right]+\frac{2\gamma^2\sigma^2(1+\omega_{\mathrm{up}})}{Nb}\,.\end{aligned}$$

Next, we require:

$$\begin{cases} 2\alpha_{\mathrm{dwn}}^2\omega_{\mathrm{dwn}} \leq \frac{1}{4}\alpha_{\mathrm{dwn}} \Longleftrightarrow \alpha_{\mathrm{dwn}} \leq \frac{1}{8\omega_{\mathrm{dwn}}}\,, \\ 2\gamma^2L^2\omega_{\mathrm{dwn}} \leq \frac{1}{4}\alpha_{\mathrm{dwn}} = \frac{1}{32\omega_{\mathrm{dwn}}}\,, \text{ by taking } \alpha_{\mathrm{dwn}} = \frac{1}{8\omega_{\mathrm{dwn}}} \\ 1+\frac{1}{\alpha_{\mathrm{dwn}}} \leq \frac{2}{\alpha_{\mathrm{dwn}}} \text{ which is not restrictive if } \omega_{\mathrm{dwn}} \geq 1. \end{cases} \Longleftrightarrow \gamma \leq \frac{1}{8\omega_{\mathrm{dwn}}L}\,,$$

Thus, it leads to:

$$\begin{aligned}\mathbb{E}\left[\Upsilon_k \mid w_{k-1}\right] \leq & \left(1-\frac{\alpha_{\mathrm{dwn}}}{2}\right)\Upsilon_{k-1}+\frac{2\gamma^2}{\alpha_{\mathrm{dwn}}}\left\|\mathbb{E}\left[\nabla F(\widehat{w}_{k-1}) \mid w_{k-1}\right]\right\|^2 \\ & +\frac{2\gamma^2\omega_{\mathrm{up}}}{N}\mathbb{E}[\left\|\nabla F(\widehat{w}_{k-1})\right\|^2 \mid w_{k-1}]+\frac{2\gamma^2\sigma^2(1+\omega_{\mathrm{up}})}{Nb}\,.\end{aligned}$$

Next, we bound $\left\|\mathbb{E}\left[\nabla F(\widehat{w}_{k-1}) \mid w_{k-1}\right]\right\|^2$ with $\mathbb{E}[\left\|\nabla F(\widehat{w}_{k-1})\right\|^2 \mid w_{k-1}]$, and we obtain:

$$\begin{aligned}\mathbb{E}\left[\Upsilon_k \mid w_{k-1}\right] \leq & \left(1-\frac{\alpha_{\mathrm{dwn}}}{2}\right)\Upsilon_{k-1}+2\gamma^2\left(\frac{1}{\alpha_{\mathrm{dwn}}}+\frac{\omega_{\mathrm{up}}}{N}\right)\mathbb{E}[\left\|\nabla F(\widehat{w}_{k-1})\right\|^2 \mid w_{k-1}] \\ & +\frac{2\gamma^2\sigma^2(1+\omega_{\mathrm{up}})}{Nb}\,.\end{aligned}$$

Taking the unconditional expectation gives the result.

■

C.4.2 Convex case (Theorem 3.4)

In this section, we give the demonstration of MCM in the convex case (Theorem 3.4).

Theorem C.4 (Convergence of MCM in the convex case)**.** *Under Assumptions 3.1 to 3.4 with $\mu = 0$, for a learning rate $\alpha_{\mathrm{dwn}} \leq \frac{1}{8\omega_{\mathrm{dwn}}}$, for all $k > 0$, for any $\gamma \leq \gamma_{\max}$, defining $V_k := \mathbb{E}[\left\|w_k - w_*\right\|^2] + 32\gamma L\omega_{\mathrm{dwn}}^2\mathbb{E}[\Upsilon_k]$, for $\bar{w}_k = \frac{1}{k}\sum_{i=0}^{k-1} w_i$, we have:*

$$\gamma\mathbb{E}\left[F(w_{k-1}) - F(w_*)\right] \leq V_{k-1} - V_k + \frac{\gamma^2\sigma^2\Phi(\gamma)}{Nb} \Longrightarrow \mathbb{E}[F(\bar{w}_k) - F_*] \leq \frac{V_0}{\gamma k} + \frac{\gamma\sigma^2\Phi(\gamma)}{Nb}\,.$$

Consequently, for K in $\mathbb{N}$ large enough, a step-size $\gamma = \sqrt{\frac{\left\|w_0-w_*\right\|^2 Nb}{(1+\omega_{\mathrm{up}})\sigma^2 K}}$ and a learning rate $\alpha_{\mathrm{dwn}} = \frac{1}{8\omega_{\mathrm{dwn}}}$, we have:

$$\mathbb{E}[F(\bar{w}_K) - F_*] \leq 2\sqrt{\frac{\left\|w_0 - w_*\right\|^2(1+\omega_{\mathrm{up}})\sigma^2}{NbK}} + O(K^{-1}).$$

Moreover if $\sigma^2 = 0$ (noiseless case), we recover a faster convergence: $\mathbb{E}[F(\bar{w}_K) - F_*] = O(K^{-1})$.

Proof Let k in $\mathbb{N}^*$, the proof follows the one for Ghost, and we start from Equation (C.3):

$$\mathbb{E}\left[\|w_k - w_*\|^2\right] \leq \mathbb{E}\left[\|w_{k-1} - w_*\|^2\right] - \gamma\left(\mathbb{E}\left[F(w_{k-1})\right] - F(w_*)\right) - \frac{\gamma}{2L}\mathbb{E}\left[\|\nabla F(\widehat{w}_{k-1})\|^2\right] + 2\gamma L\mathbb{E}\left[\|\widehat{w}_{k-1} - w_{k-1}\|^2\right] + \frac{\gamma^2\sigma^2(1+\omega_{\mathrm{up}})}{Nb},$$

with Assumption 3.1, it easily becomes:

$$\mathbb{E}\left[\|w_k - w_*\|^2\right] \leq \mathbb{E}\left[\|w_{k-1} - w_*\|^2\right] - \gamma\left(\mathbb{E}\left[F(w_{k-1})\right] - F(w_*)\right) - \frac{\gamma}{2L}\mathbb{E}\left[\|\nabla F(\widehat{w}_{k-1})\|^2\right] + 2\gamma L\omega_{\mathrm{dwn}}\mathbb{E}\left[\Upsilon_{k-1}\right] + \frac{\gamma^2\sigma^2(1+\omega_{\mathrm{up}})}{Nb}.$$

Theorem 3.5 which is specific to MCM gives:

$$\mathbb{E}\left[\Upsilon_k\right] \leq \left(1 - \frac{\alpha_{\mathrm{dwn}}}{2}\right)\mathbb{E}\left[\Upsilon_{k-1}\right] + 2\gamma^2\left(\frac{1}{\alpha_{\mathrm{dwn}}} + \frac{\omega_{\mathrm{up}}}{N}\right)\mathbb{E}\left[\|\nabla F(\widehat{w}_{k-1})\|^2\right] + \frac{2\gamma^2\sigma^2(1+\omega_{\mathrm{up}})}{Nb}.$$

Defining: $V_k := \mathbb{E}\left[\|w_k - w_*\|^2\right] + \gamma LC\mathbb{E}\left[\Upsilon_k\right]$ with $C = \frac{4\omega_{\mathrm{dwn}}}{\alpha_{\mathrm{dwn}}}$, and, combining the two last equations:

$$\begin{aligned}\mathbb{E}\left[\|w_k - w_*\|^2\right] + \gamma LC\mathbb{E}\left[\Upsilon_k\right] \leq\ & \mathbb{E}\left[\|w_{k-1} - w_*\|^2\right] - \gamma\mathbb{E}\left[F(w_{k-1}) - F(w_*)\right] \\ & + 2\gamma L\omega_{\mathrm{dwn}}\mathbb{E}\left[\Upsilon_{k-1}\right] \\ & - \frac{\gamma}{2L}\mathbb{E}\left[\|\nabla F(\widehat{w}_{k-1})\|^2\right] + \frac{\gamma^2\sigma^2(1+\omega_{\mathrm{up}})}{Nb} \\ & + \left(1 - \frac{\alpha_{\mathrm{dwn}}}{2}\right)\gamma LC\mathbb{E}\left[\Upsilon_{k-1}\right] \\ & + 2\gamma^3 LC\left(\frac{1}{\alpha_{\mathrm{dwn}}} + \frac{\omega_{\mathrm{up}}}{N}\right)\mathbb{E}\left[\|\nabla F(\widehat{w}_{k-1})\|^2\right] \\ & + \frac{2\gamma^3 L\sigma^2(1+\omega_{\mathrm{up}})C}{N},\end{aligned}$$

and reordering the terms gives:

$$\begin{aligned}V_k \leq\ & \mathbb{E}\left[\|w_{k-1} - w_*\|^2\right] + \left(2\gamma L\omega_{\mathrm{dwn}} + \left(1 - \frac{\alpha_{\mathrm{dwn}}}{2}\right)\gamma LC\right)\mathbb{E}\left[\|w_{k-1} - H_{k-1}\|^2\right] \\ & + \left(2\gamma^3 LC\left(\frac{1}{\alpha_{\mathrm{dwn}}} + \frac{\omega_{\mathrm{up}}}{N}\right) - \frac{\gamma}{2L}\right)\mathbb{E}\left[\|\nabla F(\widehat{w}_{k-1})\|^2\right] \\ & - \gamma\mathbb{E}\left[F(w_{k-1}) - F(w_*)\right] \\ & + (2\gamma LC + 1)\frac{\gamma^2\sigma^2(1+\omega_{\mathrm{up}})}{Nb}.\end{aligned}$$

We observe that:

$$2\gamma L\omega_{\mathrm{dwn}} + \left(1 - \frac{\alpha_{\mathrm{dwn}}}{2}\right)\gamma LC \leq \gamma LC \iff C \geq \frac{4\omega_{\mathrm{dwn}}}{\alpha_{\mathrm{dwn}}} \quad \text{which is true by definition of } C.$$

Secondly, to get the contraction requires

$$\begin{aligned}2\gamma^3 LC\left(\frac{1}{\alpha_{\mathrm{dwn}}} + \frac{\omega_{\mathrm{up}}}{N}\right) - \frac{\gamma}{2L} \leq 0 &\iff \gamma^2 L \leq \frac{1}{4LC\left(\frac{1}{\alpha_{\mathrm{dwn}}} + \frac{\omega_{\mathrm{up}}}{N}\right)} \\ &\iff \gamma L \leq \frac{1}{4\sqrt{\frac{\omega_{\mathrm{dwn}}}{\alpha_{\mathrm{dwn}}}\left(\frac{1}{\alpha_{\mathrm{dwn}}} + \frac{\omega_{\mathrm{up}}}{N}\right)}},\end{aligned}$$

because $C = 4\omega_{\text{dwn}}/\alpha$. Thus, we have that $V_k \leq V_{k-1} - \gamma \mathbb{E}\left[F(w_{k-1}) - F(w_*)\right] + \frac{\gamma^2\sigma^2\Phi(\gamma)}{Nb}$, denoting $\Phi(\gamma) := (1+\omega_{\text{up}})\left(1+\frac{8\gamma L\omega_{\text{dwn}}}{\alpha_{\text{dwn}}}\right)$, and then for $k = K \in \mathbb{N}^*$, by recurrence: $V_K \leq V_0 - \gamma\sum_{k=1}^K \mathbb{E}\left[F(w_{k-1}) - F(w_*)\right] + \frac{\gamma^2\sigma^2\Phi(\gamma)}{Nb}$, which implies:

$$\frac{1}{K}\sum_{k=1}^K \mathbb{E}\left[F(w_{k-1}) - F(w_*)\right] \leq \frac{V_0 - V_K}{\gamma K} + \frac{\gamma^2\sigma^2\Phi(\gamma)}{Nb}\,,$$

Finally, by Jensen, for any K in $\mathbb{N}^*$ such that $\gamma L \leq \min\left\{\frac{1}{8\omega_{\text{dwn}}}, \frac{1}{2\left(1+\frac{\omega_{\text{up}}}{N}\right)}, \frac{1}{4\sqrt{\frac{\omega_{\text{dwn}}}{\alpha_{\text{dwn}}}\left(\frac{1}{\alpha_{\text{dwn}}}+\frac{\omega_{\text{up}}}{N}\right)}}\right\}$, we have $\mathbb{E}\left[F(\bar{w}_K) - F(w_*)\right] \leq \frac{V_0}{\gamma K} + \frac{\gamma\sigma^2\Phi(\gamma)}{Nb}$, which concludes the proof.

■

C.4.3 Strongly-convex case (Theorem 3.3)

In this section, we give the demonstration for MCM in the strongly-convex case (Theorem 3.3).

Theorem C.5 (Convergence of MCM in the strongly-convex case)**.** *Under Assumptions 3.1 to 3.4 with $\mu > 0$, for k in $\mathbb{N}$, for a learning rate $\alpha_{\text{dwn}} \leq \frac{1}{8\omega_{\text{dwn}}}$, for any sequence $(\gamma_k)_{k\geq 0} \leq \gamma_{\max}$, defining $V_k := \mathbb{E}[\|w_k - w_*\|^2] + 32\gamma L\omega_{\text{dwn}}^2\mathbb{E}[\Upsilon_k]$, we have:*

$$V_k \leq (1-\gamma_k\mu/2)V_{k-1} - \gamma_k\mathbb{E}\left[F(\widehat{w}_{k-1}) - F(w_*)\right] + \frac{\gamma_k^2\sigma^2\Phi(\gamma_k)}{Nb}\,,$$

Consequently,

1. *if $\sigma^2 = 0$ (noiseless case), for $\gamma_k \equiv \gamma_{\max}$ we recover a linear convergence rate: $\mathbb{E}[\|w_K) - w_*\|^2] \leq (1-\gamma_{\max}\mu)^k V_0$;*
2. *if $\sigma^2 > 0$, defining $\widetilde{L}$ such that $\gamma_{\max} = (2\widetilde{L})^{-1}$, taking for all k in $\mathbb{N}$, $\gamma_k = 4/(\mu(k+1)+\widetilde{L})$, for the weighted Polyak-Ruppert average $\bar{w}_K = \sum_{k=1}^K \lambda_k w_{k-1}/\sum_{k=1}^K \lambda_k$, with $\lambda_k := \gamma_{k-1}^{-1}$, we have:*

$$\mathbb{E}\left[F(\bar{w}_K) - F(w_*)\right] \leq \frac{(\mu+\widetilde{L})\widetilde{L}}{16\mu K^2}\|w_0 - w_*\|^2 + \frac{8\sigma^2(1+\omega_{\text{up}})}{\mu KNb}\left(1 + \frac{256L\omega_{\text{dwn}}^2}{\mu K}\ln(\mu(K+1)+\widetilde{L})\right).$$

Proof Let k in $\mathbb{N}^*$, the proof starts like the one for Ghost, and we start from Equation (C.2) but we consider a variable step-size $\gamma_k = 2/(\mu(k+1)+\widetilde{L})$ that depends of the iteration k in $\mathbb{N}$.

$$\begin{aligned}\mathbb{E}\left[\|w_k - w_*\|^2\right] \leq{}& \mathbb{E}\left[\|w_{k-1} - w_*\|^2\right] - 2\gamma_k\mathbb{E}\left[\langle\nabla F(\widehat{w}_{k-1}), \widehat{w}_{k-1} - w_*\rangle\right] \\ &+ 2\gamma_k L\mathbb{E}\left[\|\widehat{w}_{k-1} - w_{k-1}\|^2\right] + \gamma_k^2\left(1+\frac{\omega_{\text{up}}}{N}\right)\mathbb{E}\left[\|\nabla F(\widehat{w}_{k-1})\|^2\right] + \frac{\gamma_k^2\sigma^2(1+\omega_{\text{up}})}{Nb}\,.\end{aligned}$$

Now we apply strong-convexity (Equation (A.11) of Proposition A.2):

$$\begin{aligned}\mathbb{E}\left[\|w_k - w_*\|^2\right] \leq{}& \mathbb{E}\left[\|w_{k-1} - w_*\|^2\right] + 2\gamma_k L\mathbb{E}\left[\|\widehat{w}_{k-1} - w_{k-1}\|^2\right] \\ &- \gamma_k\mathbb{E}\left[F(\widehat{w}_{k-1}) - F(w_*)\right] - \gamma_k\left(\frac{\mu}{2}\|\widehat{w}_{k-1} - w_*\|^2 + \frac{1}{L}\|\nabla F(\widehat{w}_{k-1})\|^2\right) \\ &+ \gamma_k^2\left(1+\frac{\omega_{\text{up}}}{N}\right)\mathbb{E}\left[\|\nabla F(\widehat{w}_{k-1})\|^2\right] + \frac{\gamma_k^2\sigma^2(1+\omega_{\text{up}})}{Nb}\,.\end{aligned}$$

As $\gamma_k \leq \frac{2}{\widetilde{L}} \leq \frac{1}{2L\left(1+\frac{\omega_{\text{up}}}{N}\right)}$, and thus $\left(1-\gamma_k L\left(1+\frac{\omega_{\text{up}}}{N}\right)\right) \geq 1/2$; this allows to simplify the coefficient of $\mathbb{E}[\|\nabla F(\widehat{w}_{k-1})\|^2]$:

$$\begin{aligned}\mathbb{E}\left[\|w_k - w_*\|^2\right] \leq (1-\frac{\gamma_k\mu}{2})\|w_{k-1}-w_*\|^2 - \gamma_k\mathbb{E}\left[F(\widehat{w}_{k-1}) - F(w_*)\right] \\ - \frac{\gamma_k}{2L}\mathbb{E}\left[\|\nabla F(\widehat{w}_{k-1})\|^2\right] + 2\gamma_k L\mathbb{E}\left[\|\widehat{w}_{k-1} - w_{k-1}\|^2\right] + \frac{\gamma_k^2\sigma^2(1+\omega_{\text{up}})}{Nb}\end{aligned}$$

equivalent to:

$$\begin{aligned}\mathbb{E}\left[\|w_k - w_*\|^2\right] \leq (1-\frac{\gamma_k\mu}{2})\|w_{k-1}-w_*\|^2 - \gamma_k\mathbb{E}\left[F(\widehat{w}_{k-1}) - F(w_*)\right] \\ - \frac{\gamma_k}{2L}\mathbb{E}\left[\|\nabla F(\widehat{w}_{k-1})\|^2\right] + 2\gamma_k L\omega_{\text{dwn}}\mathbb{E}\left[\|w_{k-1} - H_{k-1}\|^2\right] \\ + \frac{\gamma_k^2\sigma^2(1+\omega_{\text{up}})}{Nb}.\end{aligned} \tag{C.8}$$

Theorem 3.5 adapted to the case of decaying steps gives:

$$\mathbb{E}\left[\Upsilon_k\right] \leq \left(1-\frac{\alpha_{\text{dwn}}}{2}\right)\mathbb{E}\left[\Upsilon_{k-1}\right] + 2\gamma_k^2\left(\frac{1}{\alpha_{\text{dwn}}} + \frac{\omega_{\text{up}}}{N}\right)\mathbb{E}\left[\|\nabla F(\widehat{w}_{k-1})\|^2\right] + \frac{2\gamma_k^2\sigma^2(1+\omega_{\text{up}})}{Nb}. \tag{C.9}$$

Defining $V_k := \mathbb{E}\left[\|w_k - w_*\|^2\right] + \gamma_k LC\mathbb{E}\left[\Upsilon_k\right]$ with $C = 4\omega_{\text{dwn}}/\alpha$, combining the two later equations (A.11) + $\gamma_k LC$ (C.9):

$$\begin{aligned}\mathbb{E}\left[\|w_k - w_*\|^2\right] + \gamma_k LC\mathbb{E}\left[\Upsilon_k\right] \leq (1-\frac{\gamma_k\mu}{2})\|w_{k-1}-w_*\|^2 - \gamma_k\mathbb{E}\left[F(\widehat{w}_{k-1}) - F(w_*)\right] \\ - \frac{\gamma_k}{2L}\mathbb{E}\left[\|\nabla F(\widehat{w}_{k-1})\|^2\right] + 2\gamma_k L\omega_{\text{dwn}}\mathbb{E}\left[\|w_{k-1} - H_{k-1}\|^2\right] + \frac{\gamma_k^2\sigma^2(1+\omega_{\text{up}})}{Nb} \\ + \left(1-\frac{\alpha_{\text{dwn}}}{2}\right)\gamma_k LC\mathbb{E}\left[\Upsilon_{k-1}\right] + 2\gamma_k^3 LC\left(\frac{1}{\alpha_{\text{dwn}}} + \frac{\omega_{\text{up}}}{N}\right)\mathbb{E}\left[\|\nabla F(\widehat{w}_{k-1})\|^2\right] \\ + \frac{2\gamma_k^3 L\sigma^2(1+\omega_{\text{up}})C}{Nb},\end{aligned}$$

and reordering the terms gives:

$$\begin{aligned}V_k \leq (1-\frac{\gamma_k\mu}{2})\|w_{k-1}-w_*\|^2 - \gamma_k\mathbb{E}\left[F(\widehat{w}_{k-1}) - F(w_*)\right] \\ + \left(1-\frac{\alpha_{\text{dwn}}}{2}+\frac{2\omega_{\text{dwn}}}{C}\right)\gamma_k LC\mathbb{E}\left[\|w_{k-1} - H_{k-1}\|^2\right] \\ + \left(2\gamma_k^3 LC\left(\frac{1}{\alpha_{\text{dwn}}} + \frac{\omega_{\text{up}}}{N}\right) - \frac{\gamma_k}{2L}\right)\mathbb{E}\left[\|\nabla F(\widehat{w}_{k-1})\|^2\right] \\ + (2\gamma_k LC + 1)\frac{\gamma_k^2\sigma^2(1+\omega_{\text{up}})}{Nb},\end{aligned}$$

To reach a $(1-\frac{\gamma_k\mu}{2})$-convergence we first need $\left(1-\frac{\alpha_{\text{dwn}}}{2}+\frac{2\omega_{\text{dwn}}}{C}\right)\gamma_k LC \leq (1-\frac{\gamma_k\mu}{2})\gamma_{k-1}LC$ i.e $1-\frac{\alpha_{\text{dwn}}}{2}+\frac{2\omega_{\text{dwn}}}{C} \leq \frac{(1-\gamma_k\mu/2)\gamma_{k-1}}{\gamma_k}$.

We need that for all $k \in \mathbb{N}$, $\frac{1-\gamma_k\mu/2}{\gamma_k} \leq \frac{1}{\gamma_{k-1}}$ i.e., $1-\frac{\gamma_k\mu}{2} \leq \frac{\gamma_k}{\gamma_{k-1}}$, but:

$$\frac{\gamma_k}{\gamma_{k-1}} = \frac{\mu k - \mu + \widetilde{L}}{\mu k + \widetilde{L}} = 1 - \frac{\mu}{\mu k + \widetilde{L}} \quad \text{and} \quad 1 - \frac{\gamma_k\mu}{2} = 1 - \frac{\mu}{\mu k + \widetilde{L}},$$

and so, the inequality is always true. Thus we must have $2\omega_{\text{dwn}}/C \leq \alpha_{\text{dwn}}/2$ which is true by definition of C. Secondly, it requires:

$$2\gamma_k^2 C\left(\frac{1}{\alpha_{\text{dwn}}}+\frac{\omega_{\text{up}}}{N}\right)-\frac{\gamma_k}{2L}\leq 0 \iff \gamma_k L \leq \frac{1}{4C\left(\frac{1}{\alpha_{\text{dwn}}}+\frac{\omega_{\text{up}}}{N}\right)}$$
$$\iff \gamma_k L \leq \frac{1}{4\sqrt{\frac{\omega_{\text{dwn}}}{\alpha_{\text{dwn}}}\left(\frac{1}{\alpha_{\text{dwn}}}+\frac{\omega_{\text{up}}}{N}\right)}},$$

by definition of C. And it follows that the first part of the theorem is proved:

$$V_k \leq (1-\frac{\gamma_k\mu}{2})V_{k-1}-\gamma_k\mathbb{E}\left[F(\widehat{w}_{k-1})-F(w_*)\right]+\frac{\gamma_k^2\sigma^2\Phi(\gamma_k)}{Nb},$$

where $\Phi(\gamma_k) := (1+\omega_{\text{up}})\left(1+\frac{8\gamma_k L\omega_{\text{dwn}}}{\alpha_{\text{dwn}}}\right)$. We now prove the second part, which requires carefully handling the term of noise. By definition $\gamma_k = \frac{2}{\mu(k+1)+L}$, we denote $\lambda_k = \gamma_{k-1}^{-1}$ and we sum the above equation weighted with the sequence of $(\lambda_k)_{k=1}^K$:

$$\frac{1}{\sum_{k=1}^K\lambda_k}\sum_{k=1}^K\lambda_k\mathbb{E}\left[F(\widehat{w}_{k-1})-F(w_*)\right] \leq \frac{1}{\sum_{k=1}^K\lambda_k}\sum_{k=1}^K\frac{(1-\gamma_k\mu/2)\lambda_k}{\gamma_k}V_{k-1}-\frac{\lambda_k}{\gamma_k}V_k$$
$$+\frac{1}{\sum_{k=1}^K\lambda_k}\sum_{k=1}^K\lambda_k\frac{\gamma_k\sigma^2\Phi(\gamma_k)}{Nb}.$$

The weights are chosen to ensure that the sum of $(V_k)_{k=1}^K$ is telescopic. Because $\frac{1-\gamma_k\mu/2}{\gamma_k}=\gamma_{k-2}^{-1}$, we have:

$$\frac{1}{\sum_{k=1}^K\lambda_k}\sum_{k=1}^K\lambda_k\mathbb{E}\left[F(\widehat{w}_{k-1})-F(w_*)\right] \leq \frac{1}{\sum_{k=1}^K\lambda_k}\sum_{k=1}^K\frac{1}{\gamma_{k-2}\gamma_{k-1}}V_{k-1}-\frac{1}{\gamma_k\gamma_{k-1}}V_k$$
$$+\frac{1}{\sum_{k=1}^K\lambda_k}\sum_{k=1}^K\lambda_k\frac{\gamma_k\sigma^2\Phi(\gamma_k)}{Nb},$$

and because for $K\in\mathbb{N}^*$ big enough $\frac{1}{\sum_{k=1}^K\lambda_k}=\frac{1}{\mu(K+1)K/8+(\widetilde{L}K)/4}\leq\frac{8}{\mu K^2}$, it results that:

$$\frac{1}{\sum_{k=1}^K\lambda_k}\sum_{k=1}^K\lambda_k\mathbb{E}\left[F(\widehat{w}_{k-1})-F(w_*)\right] \leq \frac{V_0}{\gamma_0\gamma_{-1}\mu K^2}+\frac{8}{\mu K^2}\sum_{k=1}^K\lambda_k\frac{\gamma_k\sigma^2\Phi(\gamma_k)}{Nb}. \tag{C.10}$$

At the end, using the Jensen's inequality - $F(w_{k-1}) = F(\mathbb{E}\left[\widehat{w}_{k-1}\mid w_{k-1}\right]) \leq \mathbb{E}\left[F(\widehat{w}_{k-1})\mid w_{k-1}\right]$, see Equation (A.7) - we have for all K in $\mathbb{N}$:

$$\frac{1}{\sum_{k=1}^K\lambda_k}\sum_{k=1}^K\lambda_k\mathbb{E}\left[F(w_{k-1})-F(w_*)\right]$$
$$\leq \frac{V_0}{\gamma_0\gamma_{-1}\mu K^2}+\frac{8}{\mu K^2}\sum_{k=1}^K\frac{1}{\gamma_{k-1}}\left(1+\frac{8\gamma_k L\omega_{\text{dwn}}}{\alpha_{\text{dwn}}}\right)\frac{\gamma_k\sigma^2(1+\omega_{\text{up}})}{Nb}$$
$$\leq \frac{V_0}{\gamma_0\gamma_{-1}\mu K^2}+\frac{8}{\mu K^2}\sum_{k=1}^K\left(1+\frac{8\gamma_k L\omega_{\text{dwn}}}{\alpha_{\text{dwn}}}\right)\frac{\sigma^2(1+\omega_{\text{up}})}{Nb},$$

because for all k in N^*, $\gamma_k \leq \gamma_{k-1}$. We need to compute the following classical sum:

$$\sum_{k=1}^{K} \frac{1}{\mu(k+1)+\widetilde{L}} \leq \int_{x=0}^{K} \frac{1}{\mu(x+1)+\widetilde{L}} \mathrm{dx} \leq \frac{1}{\mu} \ln\left(\mu(\mathrm{K}+1)+\tilde{\mathrm{L}}\right) .$$

At the end, using again the Jensen inequality, defining $\widetilde{L} = \max\left\{4L\sqrt{\frac{\omega_{\mathrm{dwn}}}{\alpha_{\mathrm{dwn}}}\left(\frac{1}{\alpha_{\mathrm{dwn}}}+\frac{\omega_{\mathrm{up}}}{N}\right)},\ 4L\left(1+\frac{\omega_{\mathrm{up}}}{N}\right)\right\}$, taking for all k in $\mathbb{N}$, $\gamma_k = 4(\mu(k+1)+\widetilde{L}^{-1}$, for all k in N^*, $\lambda_k = \gamma_{k-1}^{-1}$ and denoting $\bar{w}_K = \frac{\sum_{k=1}^{K}\lambda_k w_{k-1}}{\sum_{k=1}^{K}\lambda_k}$, then for any K in $\mathbb{N}^*$, we have:

$$\mathbb{E}\left[F(\bar{w}_K) - F(w_*)\right] \leq \frac{(\mu+\widetilde{L})\widetilde{L}}{16\mu K^2}\left\|w_0 - w_*\right\|^2 + \left(1+\frac{256L\omega_{\mathrm{dwn}}^2}{\mu K}\ln\left(\mu(K+1)+\widetilde{L}\right)\right)\cdot\frac{8\sigma^2(1+\omega_{\mathrm{up}})}{\mu K N b} .$$

■

C.4.4 Non-convex case (Theorem 3.6)

In this section, we detail the convergence guarantee given for MCM in the non-convex case. In this scenario, the theorem hold on the average of gradients after K in $\mathbb{N}^*$ iterations. The structure of the proof is different from the one used for Ghost and MCM in convex and strongly-convex case. Instead, the demonstration starts from the equation resulting from smoothness and use the polarization identity to handle the inner product of gradients taken at two different points.

Theorem C.6 (Convergence of MCM in the non-convex case)**.** *Under Assumptions 3.1, 3.2 and 3.4 (non-convex case), for a learning rate* $\alpha_{\mathrm{dwn}} = \frac{1}{8\omega_{\mathrm{dwn}}}$*, for any step-size* γ *s.t.*

$$\gamma L \leq \min\left\{\frac{1}{8\omega_{\mathrm{dwn}}}, \frac{1}{2\left(1+\frac{\omega_{\mathrm{up}}}{N}\right)}, \frac{1}{8\sqrt{\omega_{\mathrm{dwn}}^2\left(8\omega_{\mathrm{dwn}}+\frac{\omega_{\mathrm{up}}}{N}\right)}}\right\},$$

after running K *in* $\mathbb{N}^*$ *iterations, we have:*

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\nabla F(w_{k-1})\right\|^2\right] \leq \frac{2\left(F(w_0)-F(w_*)\right)}{\gamma K} + \frac{\gamma L\sigma^2\Phi^{\mathrm{non-cvx}}(\gamma)}{Nb},$$

with $\Phi^{\mathrm{non-cvx}}(\gamma) := (1+\omega_{\mathrm{up}})\left(1+32\gamma L\omega_{\mathrm{dwn}}^2\right)$*. Thus, for* K *in* $\mathbb{N}^*$ *large enough, taking* $\gamma = \sqrt{\frac{2Nb\left(F(w_0)-F(w_*)\right)}{\sigma^2 L(1+\omega_{\mathrm{up}})K}}$*:*

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\nabla F(w_{k-1})\right\|^2\right] \leq 2\sqrt{\frac{2L\sigma^2(1+\omega_{\mathrm{up}})\left(F(w_0)-F(w_*)\right)}{NbK}} + O(K^{-1}) .$$

Proof Let k in $\mathbb{N}^*$, then smoothness (see Assumption 3.2) implies:

$$\begin{aligned} & F(w_k) \leq F(w_{k-1}) + \langle\nabla F(w_{k-1}), w_k - w_{k-1}\rangle + \frac{L}{2}\left\|w_k - w_{k-1}\right\|^2 \\ \iff\quad & F(w_k) \leq F(w_{k-1}) - \gamma\left\langle\nabla F(w_{k-1}), \widetilde{g}_k\right\rangle + \frac{\gamma^2 L}{2}\left\|\widetilde{g}_k\right\|^2 . \end{aligned}$$

The inner product is not easy to handle because it implies two gradients computed at two different points: w_{k-1} and $\widehat{w}_{k-1}$. To turn around this difficulty, we use the polarization identity, and so we have:

$$\begin{aligned}-\mathbb{E}\left[\langle \nabla F(w_{k-1}), \widetilde{g}_k\rangle \mid w_{k-1}\right] &= -\left\langle \nabla F(w_{k-1}), \mathbb{E}\left[\nabla F(\widehat{w}_{k-1}) \mid w_{k-1}\right]\right\rangle \\ &= \frac{1}{2}\Big(-\left\|\nabla F(w_{k-1})\right\|^2 - \mathbb{E}\left[\left\|\nabla F(\widehat{w}_{k-1})\right\|^2 \;\middle|\; w_{k-1}\right] \\ &\quad + \mathbb{E}\left[\left\|\nabla F(w_{k-1}) - \nabla F(\widehat{w}_{k-1})\right\|^2 \;\middle|\; w_{k-1}\right]\Big)\,.\end{aligned}$$

where we used the Polarization identity (Equation (A.5)), and next with smoothness:

$$\begin{aligned}-\mathbb{E}\left[\langle \nabla F(w_{k-1}), \widetilde{g}_k\rangle \mid w_{k-1}\right] \leq \frac{1}{2}\Big(&-\left\|\nabla F(w_{k-1})\right\|^2 - \mathbb{E}\left[\left\|\nabla F(\widehat{w}_{k-1})\right\|^2 \;\middle|\; w_{k-1}\right] \\ &+L^2\mathbb{E}\left[\left\|w_{k-1} - \widehat{w}_{k-1}\right\|^2 \;\middle|\; w_{k-1}\right]\Big)\,.\end{aligned}$$

Combining with Lemma C.2, we obtain:

$$\begin{aligned}F(w_k) \leq F(w_{k-1}) &- \frac{\gamma}{2}\left\|\nabla F(w_{k-1})\right\|^2 - \frac{\gamma}{2}\mathbb{E}\left[\left\|\nabla F(\widehat{w}_{k-1})\right\|^2 \;\middle|\; w_{k-1}\right] \\ &+ \frac{\gamma L^2}{2}\mathbb{E}\left[\left\|w_{k-1} - \widehat{w}_{k-1}\right\|^2 \;\middle|\; w_{k-1}\right] \\ &+ \frac{\gamma^2 L}{2}\left(\left(1 + \frac{\omega_{\mathrm{up}}}{N}\right)\left\|\nabla F(\widehat{w}_{k-1})\right\|^2 + \frac{\sigma^2(1+\omega_{\mathrm{up}})}{Nb}\right)\,.\end{aligned}$$

Taking the full expectation and re-ordering the terms gives:

$$\begin{aligned}\mathbb{E}\left[F(w_k)\right] \leq \mathbb{E}\left[F(w_{k-1})\right] &- \frac{\gamma}{2}\mathbb{E}\left[\left\|\nabla F(w_{k-1})\right\|^2\right] - \frac{\gamma}{2}\left(1 - \gamma L\left(1 + \frac{\omega_{\mathrm{up}}}{N}\right)\right)\mathbb{E}\left[\left\|\nabla F(\widehat{w}_{k-1})\right\|^2\right] \\ &+ \frac{\gamma L^2}{2}\mathbb{E}\left[\left\|w_{k-1} - \widehat{w}_{k-1}\right\|^2\right] + \frac{\gamma^2 L}{2} \times \frac{\sigma^2(1+\omega_{\mathrm{up}})}{Nb}\,.\end{aligned}$$

Exactly like the convex case, we consider that $\gamma L(1+\omega_{\mathrm{up}}/N) \leq 1/2$ and because $\mathbb{E}[\|w_{k-1} - \widehat{w}_{k-1}\|^2] = \mathbb{E}[\mathbb{E}[\|w_{k-1} - \widehat{w}_{k-1}\|^2 \mid \widehat{w}_{k-2}]]$ we can use Assumption 3.1:

$$\begin{aligned}\mathbb{E}\left[F(w_k)\right] \leq \mathbb{E}\left[F(w_{k-1})\right] &- \frac{\gamma}{2}\mathbb{E}\left[\left\|\nabla F(w_{k-1})\right\|^2\right] - \frac{\gamma}{4}\mathbb{E}\left[\left\|\nabla F(\widehat{w}_{k-1})\right\|^2\right] \\ &+ \frac{\omega_{\mathrm{dwn}}\gamma L^2}{2}\mathbb{E}\left[\Upsilon_k\right] + \frac{\gamma^2 L}{2} \times \frac{\sigma^2(1+\omega_{\mathrm{up}})}{Nb}\,.\end{aligned} \tag{C.11}$$

Next, Theorem 3.5 gives:

$$\mathbb{E}\left[\Upsilon_k\right] \leq \left(1 - \frac{\alpha_{\mathrm{dwn}}}{2}\right)\mathbb{E}\left[\Upsilon_{k-1}\right] + 2\gamma^2\left(\frac{1}{\alpha_{\mathrm{dwn}}} + \frac{\omega_{\mathrm{up}}}{N}\right)\mathbb{E}\left[\left\|\nabla F(\widehat{w}_{k-1})\right\|^2\right] + \frac{2\gamma^2\sigma^2(1+\omega_{\mathrm{up}})}{Nb}\,.$$

We iterate over k and compute the resulting geometric sum, it gives:

$$\begin{aligned}\mathbb{E}\left[\Upsilon_k\right] \leq \left(1 - \frac{\alpha_{\mathrm{dwn}}}{2}\right)^k\left\|\Upsilon_0\right\|^2 &+ 2\gamma^2\left(\frac{1}{\alpha_{\mathrm{dwn}}} + \frac{\omega_{\mathrm{up}}}{N}\right)\sum_{t=1}^{k}\left(1 - \frac{\alpha}{2}\right)^{k-t}\mathbb{E}\left[\left\|\nabla F(\widehat{w}_{t-1})\right\|^2\right] \\ &+ \frac{4\gamma^2\sigma^2(1+\omega_{\mathrm{up}})}{\alpha_{\mathrm{dwn}}Nb}\,,\end{aligned}$$

where we considered for the last term of the above equation that $\sum_{t=1}^{k}\left(1-\frac{\alpha_{\text{dwn}}}{2}\right)^{k} \leq \frac{2}{\alpha_{\text{dwn}}}$. This is equivalent to:

$$\mathbb{E}\left[\Upsilon_k\right] \leq 2\gamma^2\left(\frac{1}{\alpha_{\text{dwn}}}+\frac{\omega_{\text{up}}}{N}\right)\sum_{t=1}^{k}\left(1-\frac{\alpha_{\text{dwn}}}{2}\right)^{k-t}\mathbb{E}\left[\left\|\nabla F(\widehat{w}_{t-1})\right\|^2\right]+\frac{4\gamma^2\sigma^2(1+\omega_{\text{up}})}{\alpha_{\text{dwn}}Nb}\,.$$

We apply this last result to Equation (C.11):

$$\begin{aligned}
\frac{\gamma}{2}\mathbb{E}\left[\left\|\nabla F(w_{k-1})\right\|^2\right] &\leq \mathbb{E}\left[F(w_{k-1})-F(w_k)\right]-\frac{\gamma}{4}\mathbb{E}\left[\left\|\nabla F(\widehat{w}_{k-1})\right\|^2\right]\\
&\quad+\frac{\gamma L^2}{2}\left(\frac{4\omega_{\text{dwn}}\gamma^2\sigma^2(1+\omega_{\text{up}})}{Nb\alpha_{\text{dwn}}}\right.\\
&\qquad\left.+2\omega_{\text{dwn}}\gamma^2\left(\frac{1}{\alpha_{\text{dwn}}}+\frac{\omega_{\text{up}}}{N}\right)\sum_{t=1}^{k}\left(1-\frac{\alpha_{\text{dwn}}}{2}\right)^{k-t}\mathbb{E}\left[\left\|\nabla F(\widehat{w}_{t-1})\right\|^2\right]\right)\\
&\quad+\frac{\gamma^2 L}{2}\times\frac{\sigma^2(1+\omega_{\text{up}})}{Nb}\\
&\leq \mathbb{E}\left[F(w_{k-1})-F(w_k)\right]-\frac{\gamma}{4}\mathbb{E}\left[\left\|\nabla F(\widehat{w}_{k-1})\right\|^2\right]\\
&\quad+\gamma^3L^2\omega_{\text{dwn}}\left(\frac{1}{\alpha_{\text{dwn}}}+\frac{\omega_{\text{up}}}{N}\right)\sum_{t=1}^{k}\left(1-\frac{\alpha_{\text{dwn}}}{2}\right)^{k-t}\mathbb{E}\left[\left\|\nabla F(\widehat{w}_{t-1})\right\|^2\right]\\
&\quad+\frac{\gamma^2\sigma^2L(1+\omega_{\text{up}})}{2Nb}\left(1+\frac{4\gamma L\omega_{\text{dwn}}}{\alpha_{\text{dwn}}}\right)\,.
\end{aligned}$$

Summing this equation, for k in range 1 to K:

$$\begin{aligned}
\frac{\gamma}{2}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\nabla F(w_{k-1})\right\|^2\right] &\leq \mathbb{E}\left[F(w_0)-F(w_k)\right]-\frac{\gamma}{4}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\nabla F(\widehat{w}_{k-1})\right\|^2\right]\\
&\quad+\gamma^3L^2\omega_{\text{dwn}}\left(\frac{1}{\alpha_{\text{dwn}}}+\frac{\omega_{\text{up}}}{N}\right)\sum_{k=1}^{K}\sum_{t=1}^{k}\left(1-\frac{\alpha_{\text{dwn}}}{2}\right)^{k-t}\mathbb{E}\left[\left\|\nabla F(\widehat{w}_{t-1})\right\|^2\right]\\
&\quad+\frac{\gamma^2\sigma^2L(1+\omega_{\text{up}})}{2Nb}\left(1+\frac{4\gamma L\omega_{\text{dwn}}}{\alpha_{\text{dwn}}}\right)K\,.
\end{aligned}$$

We need to invert the double-sum and we obtain:

$$\begin{aligned}
\frac{\gamma}{2}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\nabla F(w_{k-1})\right\|^2\right] &\leq \gamma F(w_0)-F(w_k)-\frac{\gamma}{4}\sum_{i=1}^{K}\mathbb{E}\left[\left\|\nabla F(\widehat{w}_{k-1})\right\|^2\right]\\
&\quad+\frac{2}{\alpha_{\text{dwn}}}\times\gamma^3L^2\omega_{\text{dwn}}\left(\frac{1}{\alpha_{\text{dwn}}}+\frac{\omega_{\text{up}}}{N}\right)\sum_{k=1}^{K}\mathbb{E}\left[\left\|\nabla F(\widehat{w}_{k-1})\right\|^2\right]\\
&\quad+\frac{\gamma^2\sigma^2L(1+\omega_{\text{up}})}{2Nb}\left(1+\frac{4\gamma L\omega_{\text{dwn}}}{\alpha_{\text{dwn}}}\right)K\\
&\leq \mathbb{E}\left[F(w_0)-F(w_k)\right]\\
&\quad+\left(2\gamma^3L^2\frac{\omega_{\text{dwn}}}{\alpha_{\text{dwn}}}\left(\frac{1}{\alpha_{\text{dwn}}}+\frac{\omega_{\text{up}}}{N}\right)-\frac{\gamma}{4}\right)\sum_{k=1}^{K}\mathbb{E}\left[\left\|\nabla F(\widehat{w}_{k-1})\right\|^2\right]\\
&\quad+\frac{\gamma^2\sigma^2L(1+\omega_{\text{up}})}{2Nb}\left(1+\frac{4\gamma L\omega_{\text{dwn}}}{\alpha_{\text{dwn}}}\right)K\,.
\end{aligned}$$

Now we consider that $2\gamma^3 L^2 \frac{\omega_{\rm dwn}}{\alpha_{\rm dwn}} \left(\frac{1}{\alpha_{\rm dwn}} + \frac{\omega_{\rm up}}{N}\right) \leq \gamma/4$, and because for all k in $\mathbb{N}$, $F(w_0) - F(w_k) \leq F(w_0) - F(w_*)$:

$$\frac{1}{K}\sum_{k=1}^{K} \mathbb{E}\left[\|\nabla F(w_{k-1})\|^2\right] \leq \frac{2\left(F(w_0) - F(w_*)\right)}{\gamma K} + \frac{\gamma\sigma^2 L(1+\omega_{\rm up})}{Nb}\left(1 + \frac{4\gamma L\omega_{\rm dwn}}{\alpha_{\rm dwn}}\right).$$

Finally, for any K in $\mathbb{N}^*$, such that $\gamma L \leq \min\left\{\frac{1}{8\omega_{\rm dwn}}, \frac{1}{2\left(1+\frac{\omega_{\rm up}}{N}\right)}, \frac{1}{2\sqrt{2\frac{\omega_{\rm dwn}}{\alpha_{\rm dwn}}\left(\frac{1}{\alpha_{\rm dwn}} + \frac{\omega_{\rm up}}{N}\right)}}\right\}$ and $\alpha_{\rm dwn} \leq \frac{1}{8\omega_{\rm dwn}}$, we have:

$$\frac{1}{K}\sum_{k=1}^{K} \mathbb{E}\left[\|\nabla F(w_{k-1})\|^2\right] \leq \frac{2\left(F(w_0) - F(w_*)\right)}{\gamma K} + \frac{\gamma L\sigma^2 \Phi^{\rm non-cvx}(\gamma)}{Nb},$$

denoting $\Phi^{\rm non-cvx}(\gamma) := (1+\omega_{\rm up})\left(1 + \frac{4\gamma L\omega_{\rm dwn}}{\alpha_{\rm dwn}}\right)$.

Thus, for K in $\mathbb{N}^*$ large enough, taking $\gamma = \sqrt{\frac{2Nb\left(F(w_0) - F(w_*)\right)}{\sigma^2 L(1+\omega_{\rm up})K}}$ and $\alpha_{\rm dwn} = 1/(8\omega_{\rm dwn})$:

$$\frac{1}{K}\sum_{k=1}^{K} \mathbb{E}\left[\|\nabla F(w_{k-1})\|^2\right] \leq 2\sqrt{\frac{2L\sigma^2(1+\omega_{\rm up})\left(F(w_0) - F(w_*)\right)}{NbK}} + O(K^{-1}).$$

■

C.4.5 Proof for Rand-MCM (Theorem 3.7)

The proof for Rand-MCM is almost identical to the MCM-scenario. It only requires to modify some notations because each device i in $[\![1, N]\!]$ holds a unique model $\widehat{w}_{k-1}^i$.

For k in $\mathbb{N}$:

1. $\widetilde{g}_k$ is now defined as $\widetilde{g}_k = \frac{1}{N}\sum_{i=1}^N \widehat{g}_k^i(\widehat{w}_{k-1}^i)$,
2. for all i in $[\![1, N]\!]$, $\widehat{g}_k^i(\widehat{w}_{k-1})$ and $\nabla F(\widehat{w}_{k-1})$ must be replaced by $\widehat{g}_k^i(\widehat{w}_{k-1}^i)$ and $\nabla F(\widehat{w}_{k-1}^i)$,
3. instead of having a unique memory H_k, there is N memories $(H_k^i)_{i=1}^N$ that keep track of the updates done on each worker,
4. furthermore the notation $w_{k-1} - H_{k-2}$ is no more correct as we have N different memories. Thus, it must be replaced by $\frac{1}{N}\sum_{i=1}^N w_{k-1} - H_{k-2}^i$.

C.5 Proofs in the quadratic case for MCM and Rand-MCM

In this section, for ease of notation we denote for k in $\mathbb{N}^*$, $\widetilde{g}_k = \frac{1}{N}\sum_{i=1}^N \widehat{g}_k^i(\widehat{w}_{k-1}^i)$.

MCM has a unique memory H_k, and Rand-MCM has N different memories $(H_k^i)_{i=1}^N$. But for the sake of factorization, we will consider that both algorithm have N memories, thus we will always consider the

quantity $\frac{1}{N}\sum_{i=1}^{N}\left\|w_{k-1}-H_{k-2}^{i}\right\|^{2}$, while we should consider the quantity $\frac{1}{N}\sum_{i=1}^{N}\|w_{k-1}-H_{k-2}\|^{2}$ for MCM. However this notation is correct considering that for MCM, for all i in $[\![1,N]\!]$, $H_k^i = H_k$.

Unlike the previous sections where the proofs for MCM and Rand-MCM do not require any distinction, here in the quadratic case, we will on the contrary stress on the difference between the two. The difference appears in Lemma C.3 and comes from the way we handle the expectation of $\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla F(\widehat{w}_{k-1}^{i})-\nabla F(w_{k-1})\right\|^{2}$ for k in $\mathbb{N}^*$. For this purpose we define a constant $\mathbf{C}$ such that $\mathbf{C}=1$ in the MCM-case and $\mathbf{C}=N$ in the Rand-MCM-case.

The proofs for quadratic functions relies on the fact that for any k in $\mathbb{N}^*$, $\mathbb{E}\left[\nabla F(\widehat{w}_{k-1}) \mid w_{k-1}\right] = \nabla F(w_{k-1})$.

Definition C.1 (Quadratic function)**.** *A function $f:\mathbb{R}^d\mapsto\mathbb{R}$ is said to be quadratic if there exists a symmetric matrix A in $\mathcal{M}_{d,d}(\mathbb{R})$ such that for all x in $\mathbb{R}^d$: $f(x)-f(x_*)=\frac{1}{2}(x-x_*)^TA(x-x_*)$. And then its gradient is defined for all x in $\mathbb{R}^d$ as: $\nabla f(x)=A(x-x_*)$.*

C.5.1 Two other lemmas

In this section, we detail two lemmas required to prove the convergence of MCM and Rand-MCM in the case of quadratic functions.

The first lemma allows to factorize all the results obtained for both MCM and Rand-MCM algorithms. For k in $\mathbb{N}^*$ and i in $[\![1,N]\!]$, the difference between the MCM-case and the Rand-MCM-case results from the tigher control of $\|\sum_{i=1}^{N}\nabla F(\widehat{w}_{k-1}^{i})-\nabla F(w_{k-1})\|^2$.

Lemma C.3. *We define $\mathbf{C}$ such that $\mathbf{C}=1$ in the MCM-case and $\mathbf{C}=N$ in the Rand-MCM-case. Then for any k in $\mathbb{N}^*$, we have:*

$$\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla F(\widehat{w}_{k-1}^{i})-\nabla F(w_{k-1})\right\|^{2}\,\middle|\, w_{k-1}\right]\leq\frac{L^{2}\omega_{\mathrm{dwn}}}{\mathbf{C}}\frac{1}{N}\sum_{i=1}^{N}\left\|w_{k-1}-H_{k-2}^{i}\right\|^{2}.$$

Proof Let k in $\mathbb{N}^*$, we apply smoothness (Assumption 3.2), and then we upper bound the variance of the quantization operator with Assumption 3.1. But we must distinguish MCM and Rand-MCM because in the first case we have $\widehat{w}_{k-1}^{i}$ equal to $\widehat{w}_{k-1}$ for all i in $[\![1,N]\!]$.

In the MCM-case:

$$\begin{aligned}\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla F(\widehat{w}_{k-1}^{i})-\nabla F(w_{k-1})\right\|^{2}\,\middle|\, w_{k-1}\right] &= \mathbb{E}\left[\|\nabla F(\widehat{w}_{k-1})-F(w_{k-1})\|^{2}\mid w_{k-1}\right]\\ &\leq L^{2}\mathbb{E}\left[\|\widehat{w}_{k-1}-w_{k-1}\|^{2}\,\middle|\, w_{k-1}\right]\\ &\leq L^{2}\omega_{\mathrm{dwn}}\|\Omega_{k-1}\|^{2}\\ &\leq L^{2}\omega_{\mathrm{dwn}}\frac{1}{N}\sum_{i=1}^{N}\left\|w_{k-1}-H_{k-2}^{i}\right\|^{2},\end{aligned}$$

because we consider that $\|\Omega_{k-1}\|^2=\|w_{k-1}-H_{k-2}\|^2=\frac{1}{N}\sum_{i=1}^{N}\left\|w_{k-1}-H_{k-2}^{i}\right\|^{2}$.

In the Rand-MCM-case, by independence of the compressions on the downlink direction:

$$\begin{aligned}\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla F(\widehat{w}_{k-1}^i)-\nabla F(w_{k-1})\right\|^2\,\middle|\,w_{k-1}\right] &= \frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}\left[\left\|\nabla F(\widehat{w}_{k-1}^i)-\nabla F(w_{k-1})\right\|^2\,\middle|\,w_{k-1}\right]\\ &\leq \frac{L^2}{N^2}\sum_{i=1}^{N}\left\|\widehat{w}_{k-1}^i-w_{k-1}\right\|^2\\ &\leq \frac{L^2\omega_{\mathrm{dwn}}}{N}\times\frac{1}{N}\sum_{i=1}^{N}\left\|w_{k-1}-H_{k-2}^i\right\|^2\\ &\leq \frac{L^2\omega_{\mathrm{dwn}}}{N}\frac{1}{N}\sum_{i=1}^{N}\left\|w_{k-1}-H_{k-2}^i\right\|^2.\end{aligned}$$

We factorize the two results and define $\mathbf{C}$ such that $\mathbf{C}=1$ in the MCM-case and $\mathbf{C}=N$ in the Rand-MCM-case, and the result follows.

$$\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla F(\widehat{w}_{k-1}^i)-\nabla F(w_{k-1})\right\|^2\,\middle|\,w_{k-1}\right]\leq\frac{L^2\omega_{\mathrm{dwn}}}{\mathbf{C}}\frac{1}{N}\sum_{i=1}^{N}\left\|w_{k-1}-H_{k-2}^i\right\|^2.$$

■

The next lemma replaces Lemma C.2 in the context of randomization and quadratic functions. Note that the conditioning in Lemma C.2 is w.r.t. to $\widehat{w}_{k-1}$ while here we take the expectation w.r.t. w_{k-1}. This is because we remove $\widehat{w}_{k-1}$ from the gradient and give a result which depends of $\|\nabla F(w_{k-1})\|^2$ instead of $\|\nabla F(\widehat{w}_{k-1})\|^2$. This is made possible by the fact that for all k in $\mathbb{N}$, for quadratic functions, we have $\mathbb{E}\left[\nabla F(\widehat{w}_{k-1})\right]=\nabla F(w_{k-1})$.

Lemma C.4 (Squared-norm of stochastic gradients)**.** *For any k in $\mathbb{N}^*$, the squared-norm of gradients can be bounded a.s.:*

$$\begin{aligned}\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\widehat{g}_k^i(\widehat{w}_{k-1}^i)-\nabla F(\widehat{w}_{k-1}^i)\right\|^2\,\middle|\,w_{k-1}\right] &\leq \frac{\omega_{\mathrm{up}}}{N}\left\|\nabla F(w_{k-1})\right\|^2+\frac{\sigma^2(1+\omega_{\mathrm{up}})}{Nb}\\ &\quad+\frac{\omega_{\mathrm{up}}\omega_{\mathrm{dwn}}L^2}{N}\frac{1}{N}\sum_{i=1}^{N}\left\|w_{k-1}-H_{k-2}^i\right\|^2,\end{aligned}\tag{C.12}$$

$$\begin{aligned}\mathbb{E}\left[\left\|\widetilde{g}_k\right\|^2\,\middle|\,w_{k-1}\right] &\leq \left(1+\frac{\omega_{\mathrm{up}}}{N}\right)\left\|\nabla F(w_{k-1})\right\|^2+\frac{\sigma^2(1+\omega_{\mathrm{up}})}{Nb}\\ &\quad+L^2\omega_{\mathrm{dwn}}\left(\frac{1}{\mathbf{C}}+\frac{\omega_{\mathrm{up}}}{N}\right)\frac{1}{N}\sum_{i=1}^{N}\left\|w_{k-1}-H_{k-2}^i\right\|^2.\end{aligned}\tag{C.13}$$

The demonstration will be in two stages. We first show Equation (C.12), and in a second time, we show Equation (C.13).

Proof Let k in $\mathbb{N}^*$.

First part (Equation (C.12)). We can decompose the squared-norm in two terms:

$$\begin{aligned}
&\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\left(\widehat{g}_k^i(\widehat{w}_{k-1}^i)-\nabla F(\widehat{w}_{k-1}^i)\right)\right\|^2 \;\middle|\; w_{k-1}\right] \\
&\quad=\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\left(\widehat{g}_k^i(\widehat{w}_{k-1}^i)-g_k^i(\widehat{w}_{k-1}^i)\right)\right\|^2 \;\middle|\; w_{k-1}\right] \\
&\quad+\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\left(g_k^i(\widehat{w}_{k-1}^i)-\nabla F(\widehat{w}_{k-1}^i)\right)\right\|^2 \;\middle|\; w_{k-1}\right],
\end{aligned}$$

the first term is bounded by Assumption 3.1 and the last term by Assumption 3.4:

$$\begin{aligned}
&\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\left(\widehat{g}_k^i(\widehat{w}_{k-1}^i)-\nabla F(\widehat{w}_{k-1}^i)\right)\right\|^2 \;\middle|\; w_{k-1}\right] \\
&\quad\le\frac{\omega_{\mathrm{up}}}{N^2}\sum_{i=1}^{N}\mathbb{E}\left[\left\|g_k^i(\widehat{w}_{k-1}^i)\right\|^2 \;\middle|\; w_{k-1}\right]+\frac{\sigma^2}{Nb} \\
&\quad\le\frac{\omega_{\mathrm{up}}}{N^2}\sum_{i=1}^{N}\mathbb{E}\left[\left\|g_k^i(\widehat{w}_{k-1}^i)-\nabla F(\widehat{w}_{k-1}^i)\right\|^2 \;\middle|\; w_{k-1}\right] \\
&\qquad+\frac{\omega_{\mathrm{up}}}{N^2}\sum_{i=1}^{N}\mathbb{E}\left[\left\|\nabla F(\widehat{w}_{k-1}^i)\right\|^2 \;\middle|\; w_{k-1}\right]+\frac{\sigma^2}{Nb}.
\end{aligned}$$

And again applying Assumption 3.4 on $\mathbb{E}\left[\left\|g_k^i(\widehat{w}_{k-1}^i)-\nabla F(\widehat{w}_{k-1}^i)\right\|^2 \;\middle|\; w_{k-1}\right]$ for i in $\{1,\cdots N\}$:

$$\begin{aligned}
\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\left(\widehat{g}_k^i(\widehat{w}_{k-1}^i)-\nabla F(\widehat{w}_{k-1}^i)\right)\right\|^2 \;\middle|\; w_{k-1}\right] &= \frac{\omega_{\mathrm{up}}}{N^2}\sum_{i=1}^{N}\mathbb{E}\left[\left\|\nabla F(\widehat{w}_{k-1}^i)\right\|^2 \;\middle|\; w_{k-1}\right] \\
&\quad+\frac{\sigma^2(1+\omega_{\mathrm{up}})}{Nb}.
\end{aligned}$$

Now, we have:

$$\begin{aligned}
\frac{\omega_{\mathrm{up}}}{N^2}\sum_{i=1}^{N}\mathbb{E}\left[\left\|\nabla F(\widehat{w}_{k-1}^i)\right\|^2 \;\middle|\; w_{k-1}\right] &= \frac{\omega_{\mathrm{up}}}{N^2}\sum_{i=1}^{N}\mathbb{E}\left[\left\|\nabla F(\widehat{w}_{k-1}^i)-\nabla F(w_{k-1})\right\|^2 \;\middle|\; w_{k-1}\right] \\
&\quad+\frac{\omega_{\mathrm{up}}}{N^2}\sum_{i=1}^{N}\mathbb{E}\left[\left\|\nabla F(w_{k-1})\right\|^2 \;\middle|\; w_{k-1}\right],
\end{aligned}$$

using smoothness (Assumption 3.2) gives:

$$\frac{\omega_{\mathrm{up}}}{N^2}\sum_{i=1}^{N}\mathbb{E}\left[\left\|\nabla F(\widehat{w}_{k-1}^i)\right\|^2 \;\middle|\; w_{k-1}\right]=\frac{\omega_{\mathrm{up}}\omega_{\mathrm{dwn}}L^2}{N}\frac{1}{N}\sum_{i=1}^{N}\left\|w_{k-1}-H_{k-2}^i\right\|^2+\frac{\omega_{\mathrm{up}}}{N}\left\|\nabla F(w_{k-1})\right\|^2,$$

and putting everythings together allows to conclude for Equation (C.12).

Second part (Equation (C.13)). We start by introducing $\|\nabla F(w_{k-1})\|^2$:

$$\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\widehat{g}_k^i(\widehat{w}_{k-1}^i)\right\|^2 \,\middle|\, w_{k-1}\right] = \mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\widehat{g}_k^i(\widehat{w}_{k-1}^i) - \nabla F(w_{k-1})\right\|^2 \,\middle|\, w_{k-1}\right] + \|\nabla F(w_{k-1})\|^2$$
$$= \mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\widehat{g}_k^i(\widehat{w}_{k-1}^i) - \nabla F(\widehat{w}_{k-1}^i)\right\|^2 \,\middle|\, w_{k-1}\right]$$
$$+ \mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla F(\widehat{w}_{k-1}^i) - \nabla F(w_{k-1})\right\|^2 \,\middle|\, w_{k-1}\right]$$
$$+ \|\nabla F(w_{k-1})\|^2 \,.$$

The second term of the previous line is controlled by Lemma C.3 which distinguish the MCM and Rand-MCM-cases by defining a constant $\mathbf{C}$ such that $\mathbf{C} = 1$ for MCM and $\mathbf{C} = N$ for Rand-MCM:

$$\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla F(\widehat{w}_{k-1}^i) - \nabla F(w_{k-1})\right\|^2 \,\middle|\, w_{k-1}\right] \leq \frac{L^2\omega_{\mathrm{dwn}}}{\mathbf{C}}\frac{1}{N}\sum_{i=1}^{N}\left\|w_{k-1} - H_{k-2}^i\right\|^2 \,.$$

Thus, we have:

$$\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\widehat{g}_k^i(\widehat{w}_{k-1}^i)\right\|^2 \,\middle|\, w_{k-1}\right] = \mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\widehat{g}_k^i(\widehat{w}_{k-1}^i) - \nabla F(\widehat{w}_{k-1}^i)\right\|^2 \,\middle|\, w_{k-1}\right]$$
$$+ \frac{\omega_{\mathrm{dwn}}L^2}{\mathbf{C}}\frac{1}{N}\sum_{i=1}^{N}\left\|w_{k-1} - H_{k-2}^i\right\|^2 + \|\nabla F(w_{k-1})\|^2 \,,$$

and Equation (C.12) allows to conclude. ∎

C.5.2 Control of the Variance of the local model for quadratic function (both MCM and Rand-MCM)

The next theorem replaces the Theorem 3.5 in the case of quadratic functions. The results are almost identical except that in these settings we control the variance using non-degraded points $(w_t)_{t\in\mathbb{N}}$. This is necessary because, for quadratic functions, the analysis is slightly different. Previously, we upper-bounded the inner product in the decomposition (Equation (C.1)) by a "strong contraction" that was allowing to subtract $\|\nabla F(\widehat{w}_{k-1})\|^2$ and an extra residual term. Here we instead directly get a smaller contraction proportional to $\|\nabla F(w_{k-1})\|^2$ (but without any residual!). Indeed for all k in $\mathbb{N}$, we have $\mathbb{E}\left[\nabla F(\widehat{w}_{k-1})\right] = \nabla F(w_{k-1})$. This difference will appear in Subsection C.5.3.

As a consequence, we need to also control the variance of the local iterates that will appear when expanding the expected squared gradient $\mathbb{E}\|\tilde{g}_k\|^2$ by an affine function of the squared norms of the gradients **at the non perturbed points**. This is what Theorem C.7 provides.

Theorem C.7. *Consider the* MCM *update as in Equation (3.2) or the* Rand-MCM *update as described in Subsection 3.2.2. Under Assumptions 3.1 to 3.4 with $\mu = 0$, if $\gamma \leq \frac{1}{8L\omega_{\mathrm{dwn}}\sqrt{(1/\mathbf{C} + \omega_{\mathrm{up}}/N)}}$ and $\alpha_{\mathrm{dwn}} \leq 1/(8\omega_{\mathrm{dwn}})$, then for all k in $\mathbb{N}$:*

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\left\|w_k - H_{k-1}^i\right\|^2 \;\middle|\; w_{k-1}\right] \leq 2\gamma^2\left(\frac{1}{\alpha_{\mathrm{dwn}}} + \frac{\omega_{\mathrm{up}}}{N}\right)\sum_{t=1}^{k}(1-\frac{\alpha_{\mathrm{dwn}}}{2})^{k-t}\mathbb{E}\left[\left\|\nabla F(w_{t-1})\right\|^2 \;\middle|\; w_{t-1}\right] + \frac{4\gamma^2\sigma^2(1+\omega_{\mathrm{up}})}{\alpha_{\mathrm{dwn}}Nb}\,.$$

Proof Let k in $\mathbb{N}^*$ and i in $\{1,\dots N\}$, from Theorem C.3 we have:

$$\mathbb{E}[\|w_k - H_{k-1}^i\|^2 \mid w_{k-1}] = \mathrm{Var} + \mathrm{Bias}^2 = 2\gamma^2\mathrm{Var}_1 + 2\alpha_{\mathrm{dwn}}^2\mathrm{Var}_2 + \mathrm{Bias}^2\,,$$

with

$$\begin{cases} \mathrm{Var}_1 & = \mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^N \widehat{g}_k^i(\widehat{w}_{k-1}^i) + \mathbb{E}\left[\nabla F(\widehat{w}_{k-1}^i) \mid w_{k-1}\right]\right\|^2 \;\middle|\; w_{k-1}\right] \\ \mathrm{Var}_2 & = \omega_{\mathrm{dwn}}\frac{1}{N}\sum_{i=1}^N \left\|w_{k-1} - H_{k-2}^i\right\|^2 \\ \mathrm{Bias}^2 & = \left\|\mathbb{E}\left[w_k - H_{k-1} \mid w_{k-1}\right]\right\|^2\,. \end{cases}$$

Recall that in the case of quadratic functions, we have for all i in $[\![1, N]\!]$: $\mathbb{E}\left[\nabla F(\widehat{w}_{k-1}^i) \mid w_{k-1}\right] = \nabla F(w_{k-1})$. And so, for the first term of variance, we can decompose as follows:

$$\begin{aligned} \mathrm{Var}_1 &= \mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^N \widehat{g}_k^i(\widehat{w}_{k-1}^i) - \nabla F(w_{k-1})\right\|^2 \;\middle|\; w_{k-1}\right] \\ &= \mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^N \widehat{g}_k^i(\widehat{w}_{k-1}^i) - \nabla F(\widehat{w}_{k-1}^i)\right\|^2 \;\middle|\; w_{k-1}\right] \\ &\quad + \mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^N \nabla F(\widehat{w}_{k-1}^i) - \nabla F(w_{k-1})\right\|^2 \;\middle|\; w_{k-1}\right]\,. \end{aligned}$$

The first part is handled by Equation (C.12) of Lemma C.4:

$$\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^N \widehat{g}_k^i(\widehat{w}_{k-1}^i) - \nabla F(\widehat{w}_{k-1}^i)\right\|^2 \;\middle|\; w_{k-1}\right] = \frac{\omega_{\mathrm{up}}\omega_{\mathrm{dwn}}L^2}{N}\frac{1}{N}\sum_{i=1}^N \left\|w_{k-1} - H_{k-2}^i\right\|^2 + \frac{\omega_{\mathrm{up}}}{N}\left\|\nabla F(w_{k-1})\right\|^2 + \frac{\sigma^2(1+\omega_{\mathrm{up}})}{Nb}\,,$$

and the second part is tackled by Lemma C.3 where is defined a constant $\mathbf{C}$ such that $\mathbf{C} = 1$ in the MCM-case, and $\mathbf{C} = N$ in the Rand-MCM-case: $\mathbb{E}[\|\frac{1}{N}\sum_{i=1}^N \nabla F(\widehat{w}_{k-1}^i) - \nabla F(w_{k-1})\|^2 \mid w_{k-1}] \leq \frac{L^2\omega_{\mathrm{dwn}}}{\mathbf{C}}\frac{1}{N}\sum_{i=1}^N \left\|w_{k-1} - H_{k-2}^i\right\|^2$.

Finally, given that $\mathrm{Var} = 2\gamma^2\mathrm{Var}_1 + 2\alpha_{\mathrm{dwn}}^2\mathrm{Var}_2$ we have:

$$\mathrm{Var} \leq 2\gamma^2L^2\omega_{\mathrm{dwn}}\left(\frac{1}{\mathbf{C}} + \frac{\omega_{\mathrm{up}}}{N}\right)\frac{1}{N}\sum_{i=1}^N \left\|w_{k-1} - H_{k-2}^i\right\|^2 + 2\alpha_{\mathrm{dwn}}^2\omega_{\mathrm{dwn}}\left\|w_{k-1} - H_{k-2}^i\right\|^2 + \frac{2\gamma^2\omega_{\mathrm{up}}}{N}\left\|\nabla F(w_{k-1})\right\|^2 + \frac{2\gamma^2\sigma^2(1+\omega_{\mathrm{up}})}{Nb}\,.$$

Now we focus on the squared bias Bias^2 exactly like in Theorem C.3 and we obtain:

$$\text{Bias}^2 \leq (1-\alpha_{\text{dwn}}) \left\|w_{k-1} - H_{k-2}^i\right\|^2 + \gamma^2(1+\frac{1}{\alpha_{\text{dwn}}}) \left\|\nabla F(w_{k-1})\right\|^2 .$$

In the end:

$$\begin{aligned}\mathbb{E}\left[\left\|w_k - H_{k-1}^i\right\|^2 \;\middle|\; w_{k-1}\right] \leq\; & 2\gamma^2 L^2 \omega_{\text{dwn}} \left(\frac{1}{\mathbf{C}} + \frac{\omega_{\text{up}}}{N}\right) \frac{1}{N}\sum_{i=1}^N \left\|w_{k-1} - H_{k-2}^i\right\|^2 \\ & + \gamma^2(1 + \frac{1}{\alpha_{\text{dwn}}} + \frac{2\omega_{\text{up}}}{N}) \left\|\nabla F(w_{k-1})\right\|^2 \\ & + \left((1-\alpha_{\text{dwn}}) + 2\alpha_{\text{dwn}}^2\omega_{\text{dwn}}\right) \left\|w_{k-1} - H_{k-2}^i\right\|^2 + \frac{2\gamma^2\sigma^2(1+\omega_{\text{up}})}{Nb} .\end{aligned}$$

Summing this last equation over the N devices gives:

$$\begin{aligned}&\frac{1}{N}\sum_{i=1}^N \mathbb{E}\left[\left\|w_k - H_{k-1}^i\right\|^2 \;\middle|\; w_{k-1}\right] \\ &\quad\leq \left(1 - \alpha_{\text{dwn}} + 2\alpha_{\text{dwn}}^2\omega_{\text{dwn}} + \gamma^2 L^2 \omega_{\text{dwn}} \left(\frac{1}{\mathbf{C}} + \frac{\omega_{\text{up}}}{N}\right)\right) \frac{1}{N}\sum_{i=1}^N \left\|w_{k-1} - H_{k-2}^i\right\|^2 \\ &\qquad + \gamma^2(1 + \frac{1}{\alpha_{\text{dwn}}} + \frac{2\omega_{\text{up}}}{N}) \left\|\nabla F(w_{k-1})\right\|^2 + \frac{2\gamma^2\sigma^2(1+\omega_{\text{up}})}{Nb} .\end{aligned}$$

Exactly like in Theorem C.3, we need and by taking $\alpha_{\text{dwn}} = 1/(8\omega_{\text{dwn}})$:

$$\begin{cases} 2\alpha_{\text{dwn}}^2\omega_{\text{dwn}} \leq \frac{1}{4}\alpha_{\text{dwn}} \iff \alpha_{\text{dwn}} \leq \frac{1}{8\omega_{\text{dwn}}} , \\ 2\gamma^2 L^2 \omega_{\text{dwn}} \left(\frac{1}{\mathbf{C}} + \frac{\omega_{\text{up}}}{N}\right) \leq \frac{1}{4}\alpha_{\text{dwn}} = \frac{1}{32\omega_{\text{dwn}}} \iff \gamma \leq \dfrac{1}{8L\omega_{\text{dwn}}\sqrt{(1/\mathbf{C} + \omega_{\text{up}}/N)}} , \\ 1 + \frac{1}{\alpha_{\text{dwn}}} \leq \frac{2}{\alpha_{\text{dwn}}} \text{ which is not restrictive.} \end{cases}$$

Thus, we can write:

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^N \mathbb{E}\left[\left\|w_k - H_{k-1}^i\right\|^2 \;\middle|\; w_{k-1}\right] \leq\; & \left(1 - \frac{\alpha_{\text{dwn}}}{2}\right) \frac{1}{N}\sum_{i=1}^N \left\|w_{k-1} - H_{k-2}^i\right\|^2 \\ & + 2\gamma^2(\frac{1}{\alpha_{\text{dwn}}} + \frac{\omega_{\text{up}}}{N}) \left\|\nabla F(w_{k-1})\right\|^2 + \frac{2\gamma^2\sigma^2(1+\omega_{\text{up}})}{Nb} .\end{aligned}$$

Finally, we take the full expectation without any conditioning, we iterate over k and compute the geometric sums:

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^N \mathbb{E}\left[\left\|w_k - H_{k-1}^i\right\|^2\right] \leq\; & (1 - \frac{\alpha_{\text{dwn}}}{2})^k \left\|w_0 - H_{-1}\right\|^2 + \frac{4\gamma^2\sigma^2(1+\omega_{\text{up}})}{\alpha_{\text{dwn}} Nb} \\ & + 2\gamma^2(\frac{1}{\alpha_{\text{dwn}}} + \frac{\omega_{\text{up}}}{N}) \sum_{t=1}^k (1 - \frac{\alpha_{\text{dwn}}}{2})^{k-t} \mathbb{E}\left[\left\|\nabla F(w_{t-1})\right\|^2\right] .\end{aligned}$$

and the result follows.

■

C.5.3 Proof for quadratic function (Theorem 3.8)

Theorem C.8. *Under Assumptions 3.1 to 3.4 with $\mu = 0$, if the function is quadratic, for $\gamma = 1/(L\sqrt{K})$ and a given learning rate $\alpha_{\text{dwn}} = 1/(8\omega_{\text{dwn}})$, after running K iterations:*

$$\mathbb{E}\left[F(\bar{w}_K) - F_*\right] \leq \frac{\|w_0 - w_*\|^2 L}{\sqrt{K}} + \frac{\sigma^2 \Phi(\gamma)}{NbL\sqrt{K}}\,.$$

with $\Phi = (1+\omega_{\text{up}})\left(1 + 32\frac{\omega_{\text{dwn}}^2}{\sqrt{K}} \times \frac{1}{\sqrt{K}}\left(\frac{1}{\mathbf{C}} + \frac{\omega_{\text{up}}}{N}\right)\right)$ *and* $\mathbf{C} = N$ *for* `Rand-MCM`*, and 1 for* `MCM`*.*

The structure of the proof is different from the one used in Sections C.3 and C.4.

Proof Let k in $\mathbb{N}$, by definition: $\|w_k - w_*\|^2 \leq \|w_{k-1} - w_*\|^2 - 2\gamma\left\langle \widetilde{g}_k, w_{k-1} - w_*\right\rangle + \gamma^2 \|\widetilde{g}_k\|^2$. Because F is quadratic, we have $\mathbb{E}\left[\nabla F(\widehat{w}_{k-1}) \mid w_{k-1}\right] = \nabla F(w_{k-1})$, thus taking expectation gives:

$$\mathbb{E}[\|w_k - w_*\|^2 \mid w_{k-1}] \leq \|w_{k-1} - w_*\|^2 - 2\gamma\left\langle \nabla F(w_{k-1}), w_{k-1} - w_*\right\rangle + \gamma^2 \mathbb{E}[\|\widetilde{g}_k\|^2 \mid w_{k-1}]\,.$$

We can directly apply convexity with Equation (A.10) from Proposition A.2:

$$\begin{aligned}\mathbb{E}\left[\|w_k - w_*\|^2 \,\middle|\, w_{k-1}\right] \leq \|w_{k-1} - w_*\|^2 - \gamma\left(F(w_{k-1}) - F(w_*) + \frac{1}{L}\|\nabla F(w_{k-1})\|^2\right)\\ + \gamma^2 \mathbb{E}\left[\|\widetilde{g}_k\|^2 \,\middle|\, w_{k-1}\right]\,.\end{aligned}$$

Now, with Equation (C.13) of Lemma C.4:

$$\begin{aligned}\mathbb{E}\left[\|w_k - w_*\|^2 \,\middle|\, w_{k-1}\right] \leq \|w_{k-1} - w_*\|^2 - \gamma(F(w_{k-1}) - F(w_*)) - \frac{\gamma}{L}\|\nabla F(w_{k-1})\|^2\\ + \gamma^2\left(\left(1 + \frac{\omega_{\text{up}}}{N}\right)\|\nabla F(w_{k-1})\|^2\right.\\ \left. + L^2\omega_{\text{dwn}}\left(\frac{1}{\mathbf{C}} + \frac{\omega_{\text{up}}}{N}\right)\frac{1}{N}\sum_{i=1}^N \left\|w_{k-1} - H_{k-2}^i\right\|^2 + \frac{\sigma^2(1+\omega_{\text{up}})}{Nb}\right)\,,\end{aligned}$$

which gives:

$$\begin{aligned}\mathbb{E}\left[\|w_k - w_*\|^2 \,\middle|\, w_{k-1}\right] \leq \|w_{k-1} - w_*\|^2 - \gamma(F(w_{k-1}) - F(w_*)) - \frac{\gamma}{L}\|\nabla F(w_{k-1})\|^2\\ + \gamma^2\left(1 + \frac{\omega_{\text{up}}}{N}\right)\|\nabla F(w_{k-1})\|^2\\ + \gamma^2 L^2\omega_{\text{dwn}}\left(\frac{1}{\mathbf{C}} + \frac{\omega_{\text{up}}}{N}\right)\frac{1}{N}\sum_{i=1}^N \left\|w_{k-1} - H_{k-2}^i\right\|^2 + \frac{\sigma^2\gamma^2(1+\omega_{\text{up}})}{Nb}\,.\end{aligned}$$

Taking full expectation, we use the inequality controlling $\frac{1}{N}\sum_{i=1}^N \left\|w_{k-1} - H_{k-2}^i\right\|^2$ (Theorem C.7):

$$\begin{aligned}\mathbb{E}\left[\|w_k - w_*\|^2\right] \leq \mathbb{E}\left[\|w_{k-1} - w_*\|^2\right] - \gamma\mathbb{E}\left[F(w_{k-1}) - F(w_*)\right]\\ - \frac{\gamma}{L}\left(1 - \gamma L\left(1 + \frac{\omega_{\text{up}}}{N}\right)\right)\mathbb{E}\left[\|\nabla F(w_{k-1})\|^2\right]\\ + \gamma^2 L^2\omega_{\text{dwn}}\left(\frac{1}{\mathbf{C}} + \frac{\omega_{\text{up}}}{N}\right) \times 2\gamma^2\left(\frac{1}{\alpha_{\text{dwn}}} + \frac{\omega_{\text{up}}}{N}\right)\sum_{t=1}^k (1 - \frac{\alpha_{\text{dwn}}}{2})^{k-t}\mathbb{E}\left[\|\nabla F(w_{t-1})\|^2\right]\\ + \frac{\sigma^2\gamma^2(1+\omega_{\text{up}})}{Nb} + \gamma^2 L^2\omega_{\text{dwn}}\left(\frac{1}{\mathbf{C}} + \frac{\omega_{\text{up}}}{N}\right) \times \frac{4\sigma^2\gamma^2(1+\omega_{\text{up}})}{\alpha_{\text{dwn}}Nb}\,.\end{aligned}$$

Next, we consider – as in previous proofs – that $\gamma L(1+\omega_{\rm up}/N) \leq 1/2$, and thus $\frac{\gamma}{L}\left(1-\gamma L\left(1+\frac{\omega_{\rm up}}{N}\right)\right) \geq \frac{\gamma}{2}$. Next we carry out the "top-down recurrence":

$$\begin{aligned}\mathbb{E}\left[\|w_k - w_*\|^2\right] &\leq \|w_0 - w_*\|^2 - \gamma\sum_{j=1}^{k}\mathbb{E}\left[F(w_{k-j}) - F(w_*)\right] - \frac{\gamma}{2L}\sum_{j=1}^{k}\mathbb{E}\left[\|\nabla F(w_{k-j-1})\|^2\right] \\ &+ \sum_{j=1}^{k} 2\gamma^4 L^2 \omega_{\rm dwn}\left(\frac{1}{\mathbf{C}}+\frac{\omega_{\rm up}}{N}\right)\left(\frac{1}{\alpha_{\rm dwn}}+\frac{\omega_{\rm up}}{N}\right)\sum_{t=1}^{k-j}\left(1-\frac{\alpha_{\rm dwn}}{2}\right)^{k-j-t}\mathbb{E}\left[\|\nabla F(w_{t-1})\|^2\right] \\ &+ \sum_{j=1}^{k}\frac{\gamma^2\sigma^2(1+\omega_{\rm up})}{Nb}\left(1+\frac{4\gamma^2L^2\omega_{\rm dwn}}{\alpha_{\rm dwn}}\left(\frac{1}{\mathbf{C}}+\frac{\omega_{\rm up}}{N}\right)\right).\end{aligned}$$

We invert the double-sum, it leads to:

$$\begin{aligned}\mathbb{E}\left[\|w_k - w_*\|^2\right] &\leq \|w_0 - w_*\|^2 - \gamma\sum_{j=1}^{k}\mathbb{E}\left[F(w_{j-1}) - F(w_*)\right] - \frac{\gamma}{2L}\mathbb{E}\left[\|\nabla F(w_{k-1})\|^2\right] \\ &+ \frac{2}{\alpha_{\rm dwn}}\times 2\gamma^4 L^2\omega_{\rm dwn}\left(\frac{1}{\mathbf{C}}+\frac{\omega_{\rm up}}{N}\right)\left(\frac{1}{\alpha_{\rm dwn}}+\frac{\omega_{\rm up}}{N}\right)\mathbb{E}\left[\|\nabla F(w_{-1})\|^2\right] \\ &+ \sum_{j=1}^{k-1}\left(\frac{2}{\alpha_{\rm dwn}}\times 2\gamma^4 L^2\omega_{\rm dwn}\left(\frac{1}{\mathbf{C}}+\frac{\omega_{\rm up}}{N}\right)\left(\frac{1}{\alpha_{\rm dwn}}+\frac{\omega_{\rm up}}{N}\right)-\frac{\gamma}{2L}\right)\mathbb{E}\left[\|\nabla F(w_{j-1})\|^2\right] \\ &+ \frac{\gamma^2\sigma^2(1+\omega_{\rm up})}{Nb}\left(1+\frac{4\gamma^2L^2\omega_{\rm dwn}}{\alpha_{\rm dwn}}\left(\frac{1}{\mathbf{C}}+\frac{\omega_{\rm up}}{N}\right)\right)\times k\,.\end{aligned}$$

Now, we consider that $\frac{4\omega_{\rm dwn}\gamma^4L^2}{\alpha_{\rm dwn}}\left(\frac{1}{\mathbf{C}}+\frac{\omega_{\rm up}}{N}\right)\left(\frac{1}{\alpha_{\rm dwn}}+\frac{\omega_{\rm up}}{N}\right) < \frac{\gamma}{2L}$, thus we have:

$$\frac{\gamma}{k}\sum_{t=1}^{k}\mathbb{E}\left[F(w_{t-1}) - F(w_*)\right] \leq \frac{\|w_0 - w_*\|^2}{k} + \frac{\gamma^2\sigma^2(1+\omega_{\rm up})}{Nb}\left(1+\frac{4\gamma^2L^2\omega_{\rm dwn}}{\alpha_{\rm dwn}}\left(\frac{1}{\mathbf{C}}+\frac{\omega_{\rm up}}{N}\right)\right).$$

Finally, by Jensen, for any K in $\mathbb{N}^*$, taking γ such that:

$$\gamma L \leq \min\left\{\frac{1}{8\omega_{\rm dwn}\sqrt{\frac{1}{\mathbf{C}}+\frac{\omega_{\rm up}}{N}}}, \frac{1}{2\left(1+\frac{\omega_{\rm up}}{N}\right)}, \frac{1}{\sqrt[3]{\frac{8\omega_{\rm dwn}}{\alpha_{\rm dwn}}\left(\frac{1}{\mathbf{C}}+\frac{\omega_{\rm up}}{N}\right)\left(\frac{1}{\alpha_{\rm dwn}}+\frac{\omega_{\rm up}}{N}\right)}}\right\}$$

and with $\alpha_{\rm dwn} \leq \frac{1}{8\omega_{\rm dwn}}$, we recover Theorem 3.8 $\mathbb{E}\left[F(\bar{w}_K) - F(w_*)\right] \leq \frac{\|w_0 - w_*\|^2}{\gamma K} + \frac{\gamma\sigma^2\Phi^{\rm Rd}(\gamma)}{Nb}$ denoting $\Phi^{\rm Rd}(\gamma) = (1+\omega_{\rm up})\left(1+\frac{4\gamma^2L^2\omega_{\rm dwn}}{K}\left(\frac{1}{\mathbf{C}}+\frac{\omega_{\rm up}}{N}\right)\right)$.

■

Appendix to Distributed, compressed and averaged LSR

In this Chapter, we provide additional information to supplement our work. We begin by detailing technical results, by introducing an auxiliary lemma in Section D.1 and by proving Proposition 4.1 which gives a CLT for (LSA). Secondly, in respectively Section D.2 and Section D.3, we give the proof of Theorems 4.1 and 4.2. Thirdly, in Section D.4, we verify that the setting of single-client compressed LSR fulfills the setting presented in Section 4.2. In Section D.5 we prove that Lemma 4.1 hold and compute the compressors' covariance to establish Proposition 4.2 and Corollary 4.3. Finally, in Section D.6, we provide demonstrations for the federated learning case, including verifying assumptions (covariate-shift scenario) on random fields in Subsection D.6.1, and proving a Central Limit Theorem D.3 in Subsection D.6.2 for the concept-shift scenario.

Contents

D.1 Technical results

Additional notations. We use the Frobenius norm $\|A\|^2 := \mathrm{Tr}\left(A^\top A\right)$, which is the same notation as the vector euclidean norm (no ambiguity in general), J_r to denote the $d \times d$ diagonal matrix whose r first diagonal elements are equal to one and all the other matrix's coefficients equal to zero, $\mathcal{S}_d^{++}(\mathbb{R})$ the cone of positive definite symmetric matrices, and $\mathcal{L}^p(\Omega, \mathcal{A}, \mathbb{P})$ the set of random vectors defined on the probability space $(\Omega, \mathcal{A}, \mathbb{P})$ such that $\mathbb{E}[\|X\|^p] < \infty$. We define also the operator norm $|||A||| := \sqrt{\max \mathrm{eig}(A^\top A)}$.

D.1.1 Settings of experiments

In Tables D.1 and D.2, we summarize the settings of experiments presented in Subsection 4.3.4.

Table D.1: Settings of experiments for a single client ($N = 1$) on synthetic data (Figures 4.7a and 4.7b).

Parameter	K	d	$\mathrm{eig}(H)_i$	w_*	σ^2	ω	γ^{-1}	w_0	$\#runs$
Values	10^7	100	$1/i^4$	$(1)_{i=1}^d$	1	10	$2(\omega+1)R^2$	0	5

Table D.2: Settings of experiments for a single client ($N = 1$) on real data (Figures 4.7c and 4.7f).

Dataset	d	standardization	b	ω	γ^{-1}	w_0	$\#runs$	reference
`quantum`	65	✓	16	1	$2(\omega+1)R^2$	0	5	[CTL04]
`cifar-10`	256							[Kri09]

D.1.2 An auxiliary inequality

In this Section, we provide an auxiliary lemma that is specific to the framework considered in Section 4.2. It will be used in the proof of Theorem 4.1 and corresponds to an adaptation of Lemma 1 from Bach and Moulines [2013].

Lemma D.1 (Auxiliary inequality on $\sum_{k=1}^{K}\mathbb{E}[\|H_F^{1/2}\eta_k\|^2]/K$)**.** *Under Assumptions 4.2.1 and 4.1, for any K in $\mathbb{N}^*$ and any step-size $\gamma \in \mathbb{R}^+$ s.t. $\gamma(R_F^2 + 2\mathcal{M}_2) \leq 1$, the sequence $(w_k)_{k\in\mathbb{N}^*}$ produced by a setting such as in Definition 4.1, verifies the following bound:*

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|H_F^{1/2}(w_k - w_*)\|^2] \leq \frac{\|\eta_0\|^2}{2\gamma K(1-\gamma(R_F^2+2\mathcal{M}_2))} + \frac{5\mathcal{A}\gamma}{1-\gamma(R_F^2+2\mathcal{M}_2)}\,.$$

Proof Let k in $\mathbb{N}^*$, we start writing that by Definition 4.1, we have $w_k = w_{k-1} - \gamma\nabla F(w_{k-1}) + \gamma\xi_k(\eta_{k-1})$. Thus taking the squared norm and developing it, gives:

$$\|\eta_k\|^2 = \|\eta_{k-1}\|^2 - 2\gamma\left\langle\eta_{k-1}, \nabla F(w_{k-1}) - \xi_k(\eta_{k-1})\right\rangle + \gamma^2\left\|\nabla F(w_{k-1}) - \xi_k(\eta_{k-1})\right\|^2\,. \tag{D.1}$$

We need to bound the last term. By Definition 4.2, we have $\xi_k(\eta_{k-1}) = \xi_k^{\text{mult}}(\eta_{k-1}) + \xi_k^{\text{add}}$, hence using Inequality 1, we have:

$$\|\nabla F(w_{k-1}) - \xi_k(\eta_{k-1})\|^2 \leq 2\left\|\nabla F(w_{k-1}) - \xi_k^{\text{mult}}(\eta_{k-1})\right\|^2 + 2\left\|\xi_k^{\text{add}}\right\|^2\,,$$

taking expectation w.r.t the σ-algebra $\mathcal{F}_{k-1}$, developping $\left\|\nabla F(w_{k-1}) - \xi_k^{\text{mult}}(\eta_{k-1})\right\|^2$ and because $\mathbb{E}\left[\xi_k^{\text{mult}}(\eta_{k-1}) \,\middle|\, \mathcal{F}_{k-1}\right] = 0$ (the random fields $(\xi_k)_{k\in\mathbb{N}^*}$ are zero-centered, see Definition 4.1), we have:

$$\begin{aligned}
&\mathbb{E}\left[\|\nabla F(w_{k-1}) - \xi_k(\eta_{k-1})\|^2 \,\middle|\, \mathcal{F}_{k-1}\right] \\
&\quad\leq 2\mathbb{E}\left[\|\nabla F(w_{k-1})\|^2 \,\middle|\, \mathcal{F}_{k-1}\right] + 2\mathbb{E}\left[\left\|\xi_k^{\text{mult}}(\eta_{k-1})\right\|^2 \,\middle|\, \mathcal{F}_{k-1}\right] + 2\mathbb{E}\left[\left\|\xi_k^{\text{add}}\right\|^2 \,\middle|\, \mathcal{F}_{k-1}\right]\,.
\end{aligned}$$

Now, we use Definition 4.1 and Assumptions 4.2.1 and 4.1: it leads to:

$$\begin{aligned}
\mathbb{E}\left[\|\nabla F(w_{k-1}) - \xi_k(\eta_{k-1})\|^2 \,\middle|\, \mathcal{F}_{k-1}\right] &\leq 2R_F^2\left\|H_F^{1/2}\eta_{k-1}\right\|^2 + 4\mathcal{M}_2\left\|H_F^{1/2}\eta_{k-1}\right\|^2 + 8\mathcal{A} + 2\mathcal{A} \\
&\leq 2(R_F^2 + 2\mathcal{M}_2)\left\|H_F^{1/2}\eta_{k-1}\right\|^2 + 10\mathcal{A}
\end{aligned}$$

Because the sequence of random field $(\xi_k)_{k\in\mathbb{N}^*}$ is zero-centered (Definition 4.1), we have:

$$\mathbb{E}\left[\left\langle\eta_{k-1}, \nabla F(w_{k-1}) - \xi_k(\eta_{k-1})\right\rangle \mid \mathcal{F}_{k-1}\right] = \left\langle\eta_{k-1}, H_F\eta_{k-1}\right\rangle = \left\|H_F^{1/2}\eta_{k-1}\right\|^2\,,$$

hence back to Equation (D.1), we obtain:

$$\mathbb{E}\left[\|\eta_k\|^2 \,\middle|\, \mathcal{F}_{k-1}\right] \leq \|\eta_{k-1}\|^2 - 2\gamma(1-\gamma(R_F^2+2\mathcal{M}_2))\left\|H_F^{1/2}\eta_{k-1}\right\|^2 + 10\mathcal{A}\gamma^2\,. \tag{D.2}$$

It follows that if $\gamma(R_F^2 + 2\mathcal{M}_2) \leq 1$, summing the previous bound and taking full expectation gives:

$$\frac{1}{K}\sum_{k=1}^{K} \mathbb{E}\left\|H_F^{1/2}\eta_{k-1}\right\|^2 \leq \frac{\|\eta_0\|^2 - \mathbb{E}\|\eta_K\|^2}{2\gamma K(1-\gamma(R_F^2+2\mathcal{M}_2))} + \frac{5\mathcal{A}\gamma}{1-\gamma(R_F^2+2\mathcal{M}_2)},$$

which allows concluding.

■

D.1.3 Asymptotic results: central limit theorem for (LSA)

The demonstration of Proposition 4.1 uses the following theorem from guaranteeing the asymptotic normality of the Polyak-Ruppert iterate.

Below we present our CLT that gives the asymptotic normality of $(\sqrt{K}\eta_{K-1})_{K\in\mathbb{N}^*}$ in the case of strongly-convex case and decreasing step size, it is based on Theorem A.1 [Polyak and Juditsky, 1992].

Proposition D.1 (CLT for (LSA) in the strongly convex-case and deacreasing step-size)**.** *Under Assumptions 4.1 and 4.2, consider a sequence $(w_k)_{k\in\mathbb{N}^*}$ produced in the setting of Definition 4.1 using a step-size $(\gamma_k)_{k\in\mathbb{N}^*}$ s.t. $\gamma_k = k^{-\alpha}$, $\alpha \in (0,1)$. Then $(\eta_K)_{K\geq 0}$ converges in L^2-norm to 0, i.e. $\eta_K \xrightarrow[K\to+\infty]{L^2} 0$.*
Furthermore, $(\sqrt{K}\overline{\eta}_{K-1})_{K\geq 0}$ is asymptotically normal with mean zero and covariance such that:

$$\sqrt{K}\overline{\eta}_{K-1} \xrightarrow[K\to+\infty]{\mathcal{L}} \mathcal{N}(0, H_F^{-1}\mathfrak{C}_{\text{ania}}H_F^{-1}).$$

Proof

First, we have that in the case of decreasing step size s.t. for any k in $\mathbb{N}$, $\gamma_k = k^{-\alpha}$, we have: $\eta_K \xrightarrow[K\to+\infty]{L^2} 0$. This is a classical computation for SGD with bounded variance (Assumptions 4.2.1 and 4.1.). Detailed computations are for instance given in lectures notes of Bach [2022, pages 164-167 and 182], and Kushner and Yin [2003].

To apply Theorem 1 from Polyak and Juditsky [1992, recalled in Theorem A.1], which gives the desired result, it suffices to prove the convergence in probability of the covariance of the noise $\xi_k(\eta_{k-1})$ towards $\mathfrak{C}_{\text{ania}}$, as $k \to \infty$.

In the following, we show that $\lim_{k\to+\infty} \mathbb{E}\left[\xi_k(\eta_{k-1})\xi_k(\eta_{k-1})^\top \mid \mathcal{F}_{k-1}\right] \overset{\mathbb{P}}{=} \mathfrak{C}_{\text{ania}}$. We start writing:

$$\begin{aligned}\xi_k(\eta_{k-1})\xi_k(\eta_{k-1})^\top &= (\xi_k^{\text{add}} - \xi_k^{\text{mult}}(\eta_{k-1}))(\xi_k^{\text{add}} - \xi_k^{\text{mult}}(\eta_{k-1}))^\top \\ &= (\xi_k^{\text{add}})^{\otimes 2} - \xi_k^{\text{add}}\xi_k^{\text{mult}}(\eta_{k-1})^\top - \xi_k^{\text{mult}}(\eta_{k-1})(\xi_k^{\text{add}})^\top + \xi_k^{\text{mult}}(\eta_{k-1})^{\otimes 2}.\end{aligned}$$

(i) First, from Definition 4.2, it flows that $\mathbb{E}\left[\xi_k^{\text{add}} \otimes \xi_k^{\text{add}} \mid \mathcal{F}_{k-1}\right] = \mathfrak{C}_{\text{ania}}$.

(ii) Second, we show that $\mathbb{E}[\xi_k^{\text{mult}}(\eta_{k-1})^{\otimes 2} \mid \mathcal{F}_{k-1}]$ converges to 0 in probability: it is sufficient to show that: $\mathbb{E}[\|\xi_k^{\text{mult}}(\eta_{k-1})^{\otimes 2}\|_F \mid \mathcal{F}_{k-1}] \xrightarrow[k\to+\infty]{} 0$. To do so, we use the fact that $\|\xi_k^{\text{mult}}(\eta_{k-1})^{\otimes 2}\|_F = \|\xi_k^{\text{mult}}(\eta_{k-1})\|_2^2$, then with Assumption 4.2.2: $\mathbb{E}[\|\xi_k^{\text{mult}}(w - w_*)\|^2 \mid \mathcal{F}_{k-1}] \leq \mathcal{M}_1\|H^{1/2}\eta_{k-1}\| + \mathcal{M}_2\|H^{1/2}\eta_{k-1}\|^2$. And we have the result as we showed that $\eta_{k-1} \xrightarrow[k\to+\infty]{L^2} 0$.

(iii) Third, it remains to show that $\mathbb{E}[\xi_k^{\text{mult}}(\eta_{k-1})(\xi_k^{\text{add}})^\top \mid \mathcal{F}_{k-1}] \xrightarrow[k\to+\infty]{L^1} 0$. We use the Cauchy-

Schwarz inequality's A.8 for conditional expectation:

$$\mathbb{E}\left[\|\xi_k^{\text{mult}}(\eta_{k-1})(\xi_k^{\text{add}})^\top\|_F \;\middle|\; \mathcal{F}_{k-1}\right]^2 = \mathbb{E}\left[\|\xi_k^{\text{mult}}(\eta_{k-1})\|_2\|(\xi_k^{\text{add}})^\top\|_2 \;\middle|\; \mathcal{F}_{k-1}\right]^2$$
$$\leq \mathbb{E}\left[\|\xi_k^{\text{mult}}(\eta_{k-1})\|_2^2 \;\middle|\; \mathcal{F}_{k-1}\right]\mathbb{E}\left[\|\xi_k^{\text{add}}\|_2 \;\middle|\; \mathcal{F}_{k-1}\right].$$

The sequence of random vectors $(\xi_k^{\text{add}})_{k\in\mathbb{N}^*}$ is i.i.d., and moreover we have shown previously that $\mathbb{E}[\|\xi_k^{\text{mult}}(\eta_{k-1})\|^2 \mid \mathcal{F}_{k-1}]$ tends to 0, hence $\mathbb{E}[\xi_k^{\text{mult}}(\eta_{k-1})(\xi_k^{\text{add}})^\top \mid \mathcal{F}_{k-1}]$ converges to 0 in probability. Consequently, we can state that $\mathbb{E}[\xi_k(\eta_{k-1})^{\otimes 2} \mid \mathcal{F}_{k-1}] \xrightarrow[k\to+\infty]{\mathbb{P}} \mathfrak{C}_{\text{ania}}$.

■

D.2 Generalization of Bach and Moulines (2013).

In this section, we give the demonstration of Theorem 4.1 which extends Theorem 1 from Bach and Moulines [2013]; the demonstration is close to the original one.

D.2.1 Proof principle

For k in $\mathbb{N}^*$, the proof relies (1) on decomposing $\mathbb{E}[\|H_F^{1/2}\overline{\eta}_{K-1}\|^2]$ in two terms using the Minkowski inequality A.6 to make appear a recursion $(\eta_k^0)_{k\in\mathbb{N}^*}$ without multiplicative noise, and another $(\alpha_k)_{k\in\mathbb{N}^*}$ without additive noise, (2) on an expansion of η_k^0 and $\overline{\eta}_k^0$ as polynomials in γ, and (3) on using the Hölder-type Assumption 4.2.2 to bound α_k. We define the sequence $(\eta_k^0)_{k\in\mathbb{N}^*}$ such that it involves only an additive noise:

$$\eta_k^0 = (\mathrm{I}_d - \gamma H_F)\eta_{k-1}^0 + \gamma\xi_k^{\text{add}}. \tag{D.3}$$

Then, we decompose $\mathbb{E}[\|H_F^{1/2}\overline{\eta}_{K-1}\|^2]$ in the following way using Minkowski inequality A.6:

$$\mathbb{E}\left[\|H_F^{1/2}\overline{\eta}_{K-1}\|^2\right] \leq \left(\mathbb{E}\left[\|H_F^{1/2}\overline{\eta}_{K-1}^0\|^2\right]^{1/2} + \mathbb{E}\left[\|H_F^{1/2}(\overline{\eta}_{K-1} - \overline{\eta}_{K-1}^0)\|^2\right]^{1/2}\right)^2. \tag{D.4}$$

The goal is then to establish a bound for the two above quantities.

1. Bounding $\mathbb{E}[\|H_F^{1/2}\overline{\eta}_{K-1}^0\|^2]$.

The bound on $\mathbb{E}[\|H_F^{1/2}\overline{\eta}_{K-1}^0\|^2]$ is given in Lemma D.2. For k in $\mathbb{N}^*$, the proof relies on an expansion of η_k^0 and $\overline{\eta}_k^0$ as polynomials in γ. The recursion defining the sequence $(\eta_k^0)_{k\in\mathbb{N}^*}$ is $\eta_k^0 = (\mathrm{I}_d - \gamma H_F)\eta_{k-1}^0 + \gamma\xi_k^{\text{add}}$. If we denote $M_i^k = (\mathrm{I}_d - \gamma H_F)^{k-i}$ and $M_i^{i-1} = \mathrm{I}_d$, we have:

$$\eta_k^0 = M_1^k\eta_0^0 + \gamma\sum_{i=1}^{k} M_{i+1}^k\xi_k^{\text{add}}.$$

For K in $\mathbb{N}^*$, it leads to $\overline{\eta}_{K-1}^0 = \frac{1}{K}\sum_{k=0}^{K-1} M_1^k\eta_0^0 + \frac{\gamma}{K}\sum_{k=1}^{K-1}\left(\sum_{i=k}^{K} M_{k+1}^i\right)\xi_k^{\text{add}}$, and with Minkowski inequality A.6 to:

$$\mathbb{E}\left[\|H_F^{1/2}\overline{\eta}_{K-1}^0\|^2\right]^{1/2} \leq \mathbb{E}\left[\left\|\frac{H_F^{1/2}}{K}\sum_{k=0}^{K-1} M_1^k\eta_0^0\right\|^2\right]^{1/2} + \mathbb{E}\left[\left\|\frac{\gamma H_F^{1/2}}{K}\sum_{k=1}^{K-1}\sum_{i=k}^{K} M_{k+1}^i\xi_k^{\text{add}}\right\|^2\right]^{1/2}. \tag{D.5}$$

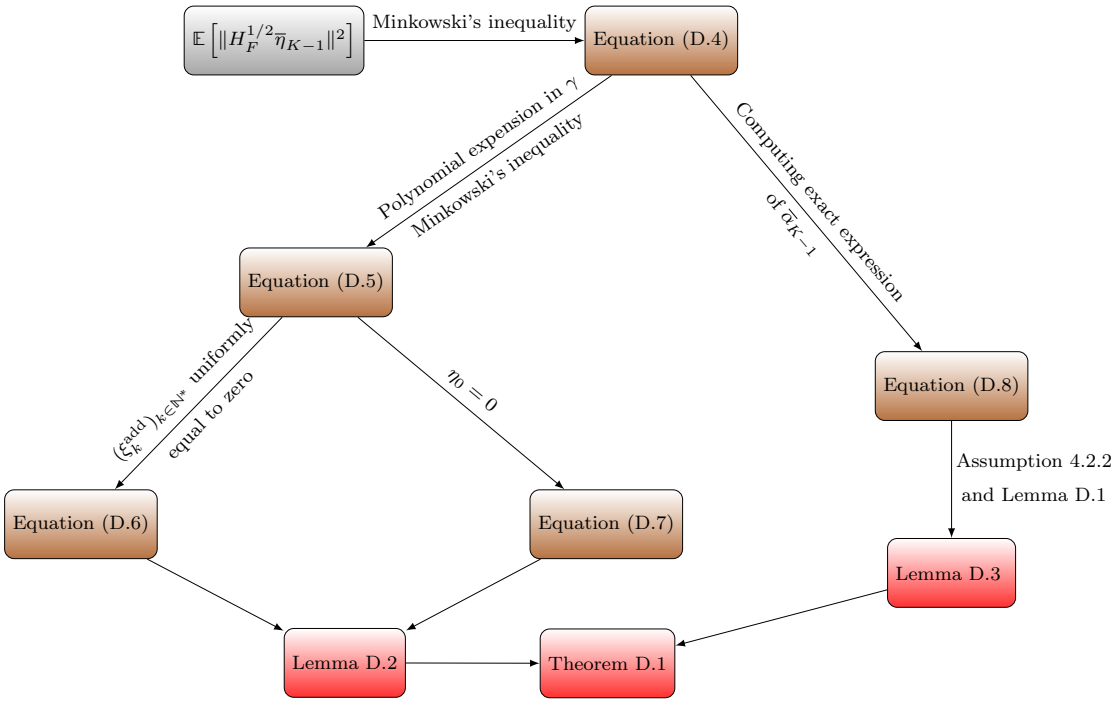


Figure D.1: Proof principle of Theorem D.1

The left term depends only on initial conditions η_0^0 $(= \eta_0)$ and the right term depends only on the additive noise. This is why, in the proof, we expend η_{k-1}^0 and $\overline{\eta}_{k-1}^0$ separately for the noise process (i.e., when assuming $\eta_0 = 0$) and for the noise-free process that depends only on the initial conditions (i.e. when assuming that the additive noise $(\xi_k^{\text{add}})_{k\in\mathbb{N}^*}$ is uniformly equal to zero). In the end, the two bounds computed separately may be added.

2. Bounding $\mathbb{E}[\|H_F^{1/2}(\overline{\eta}_{K-1} - \overline{\eta}_{K-1}^0)\|^2]$.

The bound on $\mathbb{E}[\|H_F^{1/2}(\overline{\eta}_{K-1} - \overline{\eta}_{K-1}^0)\|^2]$ is given in Lemma D.3. For k in $\mathbb{N}^*$, the demonstration is based on an exact expression of $\alpha_k = \eta_k - \eta_k^0$ and $\overline{\alpha}_k$ computed by unrolling the recursion from α_k to α_0. Because $\alpha_0 = 0$ and because there is no additive noise involved in α_k, we obtain for K in $\mathbb{N}^*$, an expression of $\overline{\alpha}_{K-1}$ that depends only on the multiplicative noise at iteration k in $\{1, \cdots, K\}$:

$$\overline{\alpha}_{K-1} = \frac{\gamma}{K} \sum_{k=1}^{K-1} (\mathrm{I}_d - (\mathrm{I}_d - \gamma H_F)^{K-k})(\gamma H_F)^{-1} \xi_k^{\text{mult}}(\eta_{k-1}) \,.$$

We then show (Equation (D.8)) that bounding $\mathbb{E}[\|H_F^{1/2}(\overline{\eta}_{K-1} - \overline{\eta}_{K-1}^0)\|^2]$ leads to bound the following sum $\frac{1}{K^2}\sum_{k=1}^{K-1} \mathbb{E}[\|H_F^{-1/2}\xi_k^{\text{mult}}(\eta_{k-1})\|^2 \mid \mathcal{F}_{k-1}]$, and this bound is established using the Hölder-type Assumption 4.2.2; which concludes this part of the proof.

D.2.2 Two bounds

In this subsection, we give two lemmas that provide a bound on $\mathbb{E}[\|H_F^{1/2}\overline{\eta}_{K-1}^0\|^2]$ and $\mathbb{E}[\|H_F^{1/2}(\overline{\eta}_{K-1} - \overline{\eta}_{K-1}^0)\|^2]$. These bounds are required due to the decomposition of $\mathbb{E}[\|H_F^{1/2}\overline{\eta}_{K-1}\|^2]$ done in Equation (D.4).

- The bound on $\mathbb{E}[\|H_F^{1/2}\overline{\eta}_k^0\|^2]$ is given in Lemma D.2. It is established by decomposing the noise process and the noise-free process. The bound on the noise process comes from Lemma 2 [Bach and Moulines, 2013] and involves the additive noise's covariance $\mathfrak{C}_{\text{ania}}$.
- The bound on $\mathbb{E}[\|H_F^{1/2}(\overline{\eta}_K - \overline{\eta}_K^0)\|^2]$ is established in Lemma D.3.

Note that in order to demonstrate Lemma D.3, we need to bound $\sum_{k=1}^{K} \|H_F^{1/2}\eta_k\|^2/K$. This is done in Lemma D.1 which is an adaptation of Lemma 1 from Bach and Moulines [2013] to random mechanisms. This auxiliary lemma holds for any kind of multiplicative noise – linear or non-linear.

Below lemma provides a bound on $\mathbb{E}[\|H_F^{1/2}\overline{\eta}_k^0\|^2]$.

Lemma D.2 (Bound on $\mathbb{E}[\|H_F^{1/2}\overline{\eta}_k^0\|^2]$)**.** *Under the setting considered in Definition 4.1, under Assumption 4.1, for any K in $\mathbb{N}^*$ and any step-size $\gamma \in \mathbb{R}^+$ s.t. $\gamma R_F^2 \leq 1$, the sequence $(\eta_k^0)_{k\in\mathbb{N}^*}$ defined in Equation (D.3) verifies the following bound:*

$$\mathbb{E}\left[\|H_F^{1/2}\overline{\eta}_{K-1}^0\|^2\right]^{1/2} \leq \frac{1}{\sqrt{K}}\left(\frac{\|H_F^{-1/2}\eta_0\|}{\gamma\sqrt{K}} \wedge \frac{\|\eta_0\|}{\sqrt{\gamma}} + \sqrt{\operatorname{Tr}\left(\mathfrak{C}_{\text{ania}}H_F^{-1}\right)}\right).$$

Proof

The proof relies on the proof presented by Bach and Moulines [2013] and is done separately for the noise process and for the noise-free process that depends only on the initial condition. The bounds may then be added (see the discussion in Subsection D.2.1).

Noise-free process.

As in section A.3 from Bach and Moulines [2013], we assume in this section that the random fields $(\xi_k^{\text{add}})_{k\in\mathbb{N}^*}$ is uniformly equal to zero and that $\gamma R_F^2 \leq 1$. We thus have for any k in $\mathbb{N}^*$ that $\eta_k^0 = (\mathrm{I}_d - \gamma H_F)\eta_{k-1}^0$.

First inequality. By recursion, we have $\eta_k^0 = (\mathrm{I}_d - \gamma H_F)^k\eta_0^0$, averaging over K in $\mathbb{N}^*$ and computing the resulting geometric sum, we have:

$$\overline{\eta}_{K-1}^0 = \frac{1}{K}\sum_{k=0}^{K-1}(\mathrm{I}_d - \gamma H_F)^k\eta_0^0 = \frac{1}{K}(\mathrm{I}_d - (\mathrm{I}_d - \gamma H_F)^{K-1})(\gamma H_F)^{-1}\eta_0^0 \preccurlyeq \frac{1}{\gamma K}H_F^{-1}\eta_0^0.$$

And because $\eta_0^0 = \eta_0$, it gives $\mathbb{E}\left[\left\langle\overline{\eta}_{K-1}^0, H_F\overline{\eta}_{K-1}^0\right\rangle\right] \leq \frac{\|H_F^{1/2}\eta_0\|^2}{\gamma^2K^2}$.

Second inequality. From the expression of η_k^0 flows:

$$\mathbb{E}[\|\eta_k^0\|^2] = \mathbb{E}[\|\eta_{k-1}^0\|^2] - 2\gamma\left\langle\eta_{k-1}^0, H_F\eta_{k-1}^0\right\rangle + \gamma^2\left\langle\eta_{k-1}^0, H_F^2\eta_{k-1}^0\right\rangle.$$

Considering that $H_F \preccurlyeq \operatorname{Tr}(H_F)\,\mathrm{I}_d \preccurlyeq R_F^2\mathrm{I}_d$ (Definition 4.1) and that $\gamma R_F^2 \leq 1$, because $\eta_0^0 = \eta_0$, by convexity we have: $\mathbb{E}[\|H_F^{1/2}\overline{\eta}_{K-1}^0\|^2] \leq \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|H_F^{1/2}\eta_{k-1}^0\|^2] \leq \frac{\|\eta_0\|^2}{\gamma K}$.

Putting things together.

In the end, we take the minimum of the two above bounds and obtain that:

$$\mathbb{E}[\|H_F^{1/2}\overline{\eta}_{K-1}^0\|^2] \leq \frac{\|H_F^{-1/2}\eta_0\|^2}{\gamma^2K^2} \wedge \frac{\|\eta_0\|^2}{\gamma K}. \tag{D.6}$$

Noise process.

We assume in this part that $\eta_0^0 = \eta_0 = 0$. We apply Lemma 2 from Bach and Moulines [2013] to η_{k-1}^0. This sequence of iterates has an i.i.d. noise process $(\xi_k^{\text{add}})_{k\in\mathbb{N}^*}$ which is such that $\mathbb{E}\left[\xi_k^{\text{add}} \otimes \xi_k^{\text{add}}\right] = \mathfrak{C}_{\text{ania}}$ (existence guaranteed by Assumption 4.1). Therefore we have:

$$\mathbb{E}[\|H_F^{1/2}\overline{\eta}_{K-1}^0\|^2] \leq \frac{\operatorname{Tr}\left(\mathfrak{C}_{\text{ania}}H_F^{-1}\right)}{K}. \tag{D.7}$$

Putting things together. We now take results derived from the part without noise and the part with noise, and we get from Minkowski inequality:

$$\mathbb{E}\left[\|H_F^{1/2}\overline{\eta}_{K-1}^0\|^2\right]^{1/2} \leq \frac{1}{\sqrt{K}}\left(\frac{\|H_F^{-1/2}\eta_0\|}{\gamma\sqrt{K}} \wedge \frac{\|\eta_0\|}{\sqrt{\gamma}} + \sqrt{\mathrm{Tr}\left(\mathfrak{C}_{\mathrm{ania}}H_F^{-1}\right)}\right).$$

■

Below lemma provides a bound on $\mathbb{E}[\|H_F^{1/2}(\overline{\eta}_K - \overline{\eta}_K^0)\|^2]$.

Lemma D.3 (Bound on $\mathbb{E}[\|H_F^{1/2}(\overline{\eta}_K - \overline{\eta}_K^0)\|^2]$)**.** *Under the setting considered in Definition 4.1 with $\mu > 0$, under Assumption 4.1 , under Assumptions 4.2.1 and 4.2.2, for any K in $\mathbb{N}^*$ and any step-size $\gamma \in \mathbb{R}^+$ s.t. $\gamma(R_F^2 + 2\mathcal{M}_2) < 1$, the sequence $(\overline{\eta}_k - \overline{\eta}_k^0)_{k\in\mathbb{N}^*}$ verifies the following bound:*

$$\mathbb{E}\left[\|H_F^{1/2}(\overline{\eta}_K - \overline{\eta}_K^0)\|^2\right]^{1/2}] \leq \frac{1}{\sqrt{K}}\left(\sqrt{\mathcal{M}_1\mu^{-1}}\left(\frac{5\mathcal{A}\gamma}{1-\gamma(R_F^2+2\mathcal{M}_2)}\right)^{1/4} + \sqrt{\mathcal{M}_2\mu^{-1}}\left(\frac{15\mathcal{A}\gamma}{1-\gamma(R_F^2+2\mathcal{M}_2)}\right)^{1/2}\right).$$

Remark D.1. *To demonstrate Lemma D.3, we use the Hölder-type Assumption 4.2.2. This is why we obtain a term with a square root in the bound.*

Proof

Let k in $\mathbb{N}^*$, we denote $\alpha_k = \eta_k - \eta_k^0$, with $\eta_k = (\mathrm{I}_d - \gamma H_F)\eta_{k-1} + \gamma\xi_k(\eta_{k-1})$ and $\eta_k^0 = (\mathrm{I}_d - \gamma H_F)\eta_{k-1}^0 + \gamma\xi_k^{\mathrm{add}}$. First, we write the exact expression of α_{k-1}:

$$\begin{aligned}\alpha_k &= (\mathrm{I}_d - \gamma H_F)\alpha_{k-1} + \gamma(\xi_k(\eta_{k-1}) - \xi_k^{\mathrm{add}}) \\ &= (\mathrm{I}_d - \gamma H_F)^k\alpha_0 + \gamma\sum_{i=1}^{k}(\mathrm{I}_d - \gamma H_F)^{k-i}(\xi_i(\eta_{i-1}) - \xi_i^{\mathrm{add}}),\end{aligned}$$

and because $\eta_0^0 = \eta_0$, it follows that $\alpha_0 = \eta_0 - \eta_0^0 = 0$. Averaging over K in $\mathbb{N}^*$, we have the exact expression of $\overline{\alpha}_{K-1}$:

$$\begin{aligned}\overline{\alpha}_{K-1} &= \frac{\gamma}{K}\sum_{k=0}^{K-1}\sum_{i=1}^{k}(\mathrm{I}_d - \gamma H_F)^{k-i}(\xi_i(\eta_{i-1}) - \xi_i^{\mathrm{add}})) \\ &= \frac{\gamma}{K}\sum_{i=1}^{K-1}\left(\sum_{k=i}^{K-1}(\mathrm{I}_d - \gamma H_F)^{k-i}\right)(\xi_i(\eta_{i-1}) - \xi_i^{\mathrm{add}})).\end{aligned}$$

Computing the geometric sum results in:

$$\overline{\alpha}_{K-1} = \frac{\gamma}{K}\sum_{k=1}^{K-1}(\mathrm{I}_d - (\mathrm{I}_d - \gamma H_F)^{K-k})(\gamma H_F)^{-1}(\xi_k(\eta_{k-1}) - \xi_k^{\mathrm{add}}).$$

And because for any k in $\mathbb{N}$, $0 \preccurlyeq (\mathrm{I}_d - \gamma H_F)^k \preccurlyeq \mathrm{I}_d$, we obtain:

$$\overline{\alpha}_{K-1} \preccurlyeq \frac{1}{K}\sum_{k=1}^{K-1}H_F^{-1}(\xi_k(\eta_{k-1}) - \xi_k^{\mathrm{add}}),$$

hence $\|H_F^{1/2}\overline{\alpha}_{K-1}\|^2 = \|\frac{1}{K}\sum_{k=1}^{K-1} H_F^{-1/2}(\xi_k(\eta_{k-1}) - \xi_k^{\text{add}})\|^2$. We take full expectation, because for any k in $\mathbb{N}^*$, by Definitions 4.1 and 4.2, $\xi_k^{\text{mult}}(\eta_{k-1}) = \xi_k(\eta_{k-1}) - \xi_k^{\text{add}}$ is $\mathcal{F}_k$-measurable and $\mathbb{E}\left[\xi_k^{\text{mult}}(\eta_{k-1}) \,\middle|\, \mathcal{F}_{k-1}\right] = 0$, we can unroll the sum and we have in the end that the variance of the sum is the sum of variances:

$$\mathbb{E}\left[\left\|H_F^{1/2}\overline{\alpha}_{K-1}\right\|^2\right] \leq \frac{1}{K^2}\sum_{k=1}^{K-1}\mathbb{E}\left[\left\|H_F^{-1/2}\xi_k^{\text{mult}}(\eta_{k-1})\right\|^2 \,\middle|\, \mathcal{F}_{k-1}\right]. \tag{D.8}$$

Computing $\mathbb{E}[\|H_F^{-1/2}\xi_k^{\text{mult}}(\eta_{k-1})\|^2 \mid \mathcal{F}_{k-1}]$ for k in $\mathbb{N}$, we first have:

$$\|H_F^{-1/2}\xi_k^{\text{mult}}(\eta_{k-1})\|^2 \leq \left|\left|\left|H_F^{-1/2}\right|\right|\right|^2 \|\xi_k^{\text{mult}}(\eta_{k-1})\|^2,$$

where we used Inequality 2. Because H_F is a symmetric semi-positive matrix, we have $\left|\left|\left|H_F^{-1/2}\right|\right|\right|^2 = 1/\mu$, hence: $\|H_F^{-1/2}\xi_k^{\text{mult}}(\eta_{k-1})\|^2 \leq \mu^{-1}\|\xi_k^{\text{mult}}(\eta_{k-1})\|^2$. Taking expectation conditionally to the σ-algebra $\mathcal{F}_{k-1}$ and invoking Assumption 4.2.2 gives:

$$\mathbb{E}[\|H_F^{-1/2}\xi_k^{\text{mult}}(\eta_{k-1})\|^2 \mid \mathcal{F}_{k-1}] \leq \mu^{-1}(\mathcal{M}_1\|H_F^{1/2}\eta_{k-1}\| + 3\mathcal{M}_2\|H_F^{1/2}\eta_{k-1}\|^2). \tag{D.9}$$

Combining equations D.8 and D.9, we obtain:

$$\mathbb{E}[\|H_F^{1/2}\overline{\alpha}_{K-1}\|^2] \leq \frac{\mathcal{M}_1}{\mu K^2}\sum_{k=1}^{K-1}\mathbb{E}[\|H_F^{1/2}\eta_{k-1}\|] + \frac{3\mathcal{M}_2}{\mu K^2}\sum_{k=1}^{K-1}\mathbb{E}[\|H_F^{1/2}\eta_{k-1}\|^2].$$

Now using Jensen's inequality for concave function allows us to write:

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|H_F^{1/2}(w-w_*)\|] \leq \frac{1}{K}\sum_{k=1}^{K}\sqrt{\mathbb{E}[\|H_F^{1/2}(w-w_*)\|^2]} \leq \sqrt{\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|H_F^{1/2}(w-w_*)\|^2]},$$

thus we have:

$$\mathbb{E}[\|H_F^{1/2}\overline{\alpha}_{K-1}\|^2] \leq \frac{\mathcal{M}_1}{\mu K}\sqrt{\frac{1}{K}\sum_{k=1}^{K-1}\mathbb{E}[\|H_F^{1/2}\eta_{k-1}\|^2]} + \frac{3\mathcal{M}_2}{\mu K^2}\sum_{k=1}^{K-1}\mathbb{E}[\|H_F^{1/2}\eta_{k-1}\|^2].$$

Using Lemma D.1 (with $\eta_0 = 0$), we finally obtain:

$$\mathbb{E}[\|H_F^{1/2}\overline{\alpha}_{K-1}\|^2] \leq \frac{1}{K}\left(\mathcal{M}_1\mu^{-1}\sqrt{\frac{5\mathcal{A}\gamma}{1-\gamma(R_F^2+2\mathcal{M}_2)}} + \frac{15\mathcal{A}\gamma\mathcal{M}_2\mu^{-1}}{1-\gamma(R_F^2+2\mathcal{M}_2)}\right).$$

In the end, we take the square root (and use that for any a, b in $\mathbb{R}_+$, $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$) which allows concluding:

$$\mathbb{E}\left[\|H_F^{1/2}(\overline{\eta}_K - \overline{\eta}_K^0)\|^2\right]^{1/2} \leq \frac{1}{\sqrt{K}}\left(\sqrt{\mathcal{M}_1\mu^{-1}}\left(\frac{5\mathcal{A}\gamma}{1-\gamma(R_F^2+2\mathcal{M}_2)}\right)^{1/4} + \sqrt{\mathcal{M}_2\mu^{-1}}\left(\frac{15\mathcal{A}\gamma}{1-\gamma(R_F^2+2\mathcal{M}_2)}\right)^{1/2}\right).$$

■

D.2.3 Final theorem

In this section, we gather the pieces of proof required to demonstrate Theorem 4.1.

Theorem D.1 (Non-linear multiplicative noise)**.** *Under Assumptions 4.1 and 4.2, considering any constant step-size γ such that $\gamma(R_F^2 + 2\mathcal{M}_2) \leq 1/2$, then for any K in $\mathbb{N}^*$, the sequence $(w_k)_{k\in\mathbb{N}^*}$ produced by a setting such as in Definition 4.1 verifies the following bound:*

$$\mathbb{E}[F(\overline{w}_{K-1}) - F(w_*)] \leq \frac{1}{2K}\left(\frac{\|H_F^{-1/2}\eta_0\|}{\gamma\sqrt{K}} \wedge \frac{\|\eta_0\|}{\sqrt{\gamma}} + \sqrt{\mathrm{Tr}\left(\mathfrak{C}_{\mathrm{ania}}H_F^{-1}\right)} + (10\mathcal{A}\gamma)^{1/4}\sqrt{\mathcal{M}_1\mu^{-1}} + (30\mathcal{A}\gamma)^{1/2}\sqrt{\mathcal{M}_2\mu^{-1}}\right)^2.$$

Proof

As explained in the discussion in Subsection D.2.1 (Equation (D.4)), we define the sequence $(\eta_k^0)_{k\in\mathbb{N}^*}$ which involves only an additive noise $\eta_k^0 = (\mathrm{I}_d - \gamma H_F)\eta_{k-1}^0 + \gamma\xi_k^{\mathrm{add}}$. Then, we decompose $\mathbb{E}[\|H_F^{1/2}\overline{\eta}_{K-1}\|]$ using Minkowski's inequality A.6:

$$\mathbb{E}\left[\|H_F^{1/2}\overline{\eta}_{K-1}\|^2\right] \leq \left(\mathbb{E}\left[\|H_F^{1/2}\overline{\eta}_{K-1}^0\|^2\right]^{1/2} + \mathbb{E}\left[\|H_F^{1/2}(\overline{\eta}_{K-1} - \overline{\eta}_{K-1}^0)\|^2\right]^{1/2}\right)^2. \tag{D.10}$$

First term.

To bound the first term, we use Lemma D.2 which gives:

$$\mathbb{E}\left[\|H_F^{1/2}\overline{\eta}_{K-1}^0\|^2\right]^{1/2} \leq \frac{1}{\sqrt{K}}\left(\frac{\|H_F^{-1/2}\eta_0\|}{\gamma\sqrt{K}} \wedge \frac{\|\eta_0\|}{\sqrt{\gamma}} + \sqrt{\mathrm{Tr}\left(\mathfrak{C}_{\mathrm{ania}}H_F^{-1}\right)}\right).$$

Second term.

From Lemma D.3, we have:

$$\mathbb{E}\left[\|H_F^{1/2}(\overline{\eta}_K - \overline{\eta}_K^0)\|^2\right]^{1/2} \leq \frac{1}{\sqrt{K}}\left(\sqrt{\mathcal{M}_1\mu^{-1}}\left(\frac{5\mathcal{A}\gamma}{1-\gamma(R_F^2+2\mathcal{M}_2)}\right)^{1/4} + \sqrt{\mathcal{M}_2\mu^{-1}}\left(\frac{15\mathcal{A}\gamma}{1-\gamma(R_F^2+2\mathcal{M}_2)}\right)^{1/2}\right).$$

Final bound. Hence, back to Equation (D.10), we get:

$$\mathbb{E}\left[\|H_F^{1/2}\overline{\eta}_{K-1}\|^2\right]^{1/2} \leq \frac{1}{\sqrt{K}}\left(\frac{\|H_F^{-1/2}\eta_0\|}{\gamma\sqrt{K}} \wedge \frac{\|\eta_0\|}{\sqrt{\gamma}} + \sqrt{\mathrm{Tr}\left(\mathfrak{C}_{\mathrm{ania}}H_F^{-1}\right)} + \sqrt{\mathcal{M}_1\mu^{-1}}\left(\frac{5\mathcal{A}\gamma}{1-\gamma(R_F^2+2\mathcal{M}_2)}\right)^{1/4} + \sqrt{\mathcal{M}_2\mu^{-1}}\left(\frac{15\mathcal{A}\gamma}{1-\gamma(R_F^2+2\mathcal{M}_2)}\right)^{1/2}\right),$$

and considering $\gamma(R_F^2 + 2\mathcal{M}_2) \leq 1/2$, it concludes the proof because $\mathbb{E}[F(\overline{w}_{K-1}) - F(w_*)] =$

$\mathbb{E}[\|H_F^{1/2}\overline{\eta}_{K-1}\|^2]/2$:

$$\mathbb{E}[F(\overline{w}_{K-1}) - F(w_*)] \leq \frac{1}{2K}\left(\frac{\|H_F^{-1/2}\eta_0\|}{\gamma\sqrt{K}} \wedge \frac{\|\eta_0\|}{\sqrt{\gamma}} + \sqrt{\mathrm{Tr}\left(\mathfrak{C}_{\mathrm{ania}}H_F^{-1}\right)} + (10\mathcal{A}\gamma)^{1/4}\sqrt{\mathcal{M}_1\mu^{-1}} + (30\mathcal{A}\gamma)^{1/2}\sqrt{\mathcal{M}_2\mu^{-1}}\right)^2 .$$

■

D.3 Generalisation of Bach and Moulines (2013) for linear multiplicative noise.

In this Section, we give the demonstration of Theorem 4.2 which extends Theorem 1 from Bach and Moulines [2013] to the case of linear multiplicative noise. The demonstration follows the same steps as the one given by Bach and Moulines [2013]. The minor differences lie in the generality of the form of the multiplicative noise in our approach. Bach and Moulines [2013] only analyse LMS algorithm, while we here consider (LSA) with assumptions on the linear multiplicative noise process. Moreover, our theorem decomposes into 3 terms instead of 2.

D.3.1 Proof principle

For k in $\mathbb{N}^*$, the proof relies on an expansion of η_k and $\overline{\eta}_k$ as polynomials in γ. Because we consider a linear multiplicative noise, there exists a matrix Ξ_k in $\mathbb{R}^{d\times d}$ s.t. for any z in $\mathbb{R}^d$, $\xi_k^{\mathrm{mult}}(z) = \Xi_k z$ (Assumption 4.3); hence the recursion defined in Definition 4.1 can be rewritten as:

$$\eta_k = \eta_{k-1} - \gamma\nabla F(\eta_{k-1}) + \gamma\xi_k^{\mathrm{mult}}(\eta_{k-1}) + \gamma\xi_k^{\mathrm{add}} = (\mathrm{I}_d - \gamma H_F + \gamma\Xi_k)\eta_{k-1} + \gamma\xi_k^{\mathrm{add}} .$$

We denote $M_i^k = (\mathrm{I}_d - \gamma H_F + \gamma\Xi_k)\cdots(\mathrm{I}_d - \gamma H_F + \gamma\Xi_i)$ and $M_i^{i-1} = \mathrm{I}_d$, then we have that $\eta_k = M_1^k\eta_0 + \gamma\sum_{i=1}^k M_{i+1}^k\xi_k^{\mathrm{add}}$.

For K in $\mathbb{N}^*$, it leads to $\overline{\eta}_{K-1} = \frac{1}{K}\sum_{k=0}^{K-1} M_1^k\eta_0 + \frac{\gamma}{K}\sum_{k=1}^{K-1}\left(\sum_{i=k}^K M_{k+1}^i\right)\xi_k^{\mathrm{add}}$, and with Minkowski's inequality A.6 to:

$$\sqrt{\mathbb{E}\left[\left\|H_F^{1/2}\overline{\eta}_{K-1}\right\|^2\right]} \leq \mathbb{E}\left[\left\|\frac{H_F^{1/2}}{K}\sum_{k=0}^{K-1} M_1^k\eta_0\right\|^2\right]^{1/2} + \mathbb{E}\left[\left\|\frac{\gamma H_F^{1/2}}{K}\sum_{k=1}^{K-1}\sum_{i=k}^{K} M_{k+1}^i\xi_k^{\mathrm{add}}\right\|^2\right]^{1/2} . \quad \text{(D.11)}$$

The left term depends only on initial conditions and the right term depends only on the noise process.

This is why, in the proof, we expend η_{k-1} and $\overline{\eta}_{k-1}$ separately for the noise process (i.e., when assuming $\eta_0 = 0$) and for the noise-free process that depends only on the initial conditions (i.e. when assuming that the additive noise $(\xi_k^{\mathrm{add}})_{k\in\mathbb{N}^*}$ is uniformly equal to zero). In the end, the two bounds computed separately may be added.

To study the noise process, inspiring from Bach and Moulines [2013], we define the following sequence:

$$\begin{cases} \eta_k^0 = (\mathrm{I}_d - \gamma H_F)\eta_{k-1}^0 + \gamma\xi_k^{\mathrm{add}} \\ \eta_k^r = (\mathrm{I}_d - \gamma H_F)\eta_{k-1}^r + \gamma\xi_k^{\mathrm{mult}}\left(\eta_{k-1}^{r-1}\right) \qquad \text{with} \qquad \forall r \geq 0\,, \eta_0^r = 0\,. \end{cases} \quad \text{(D.12)}$$

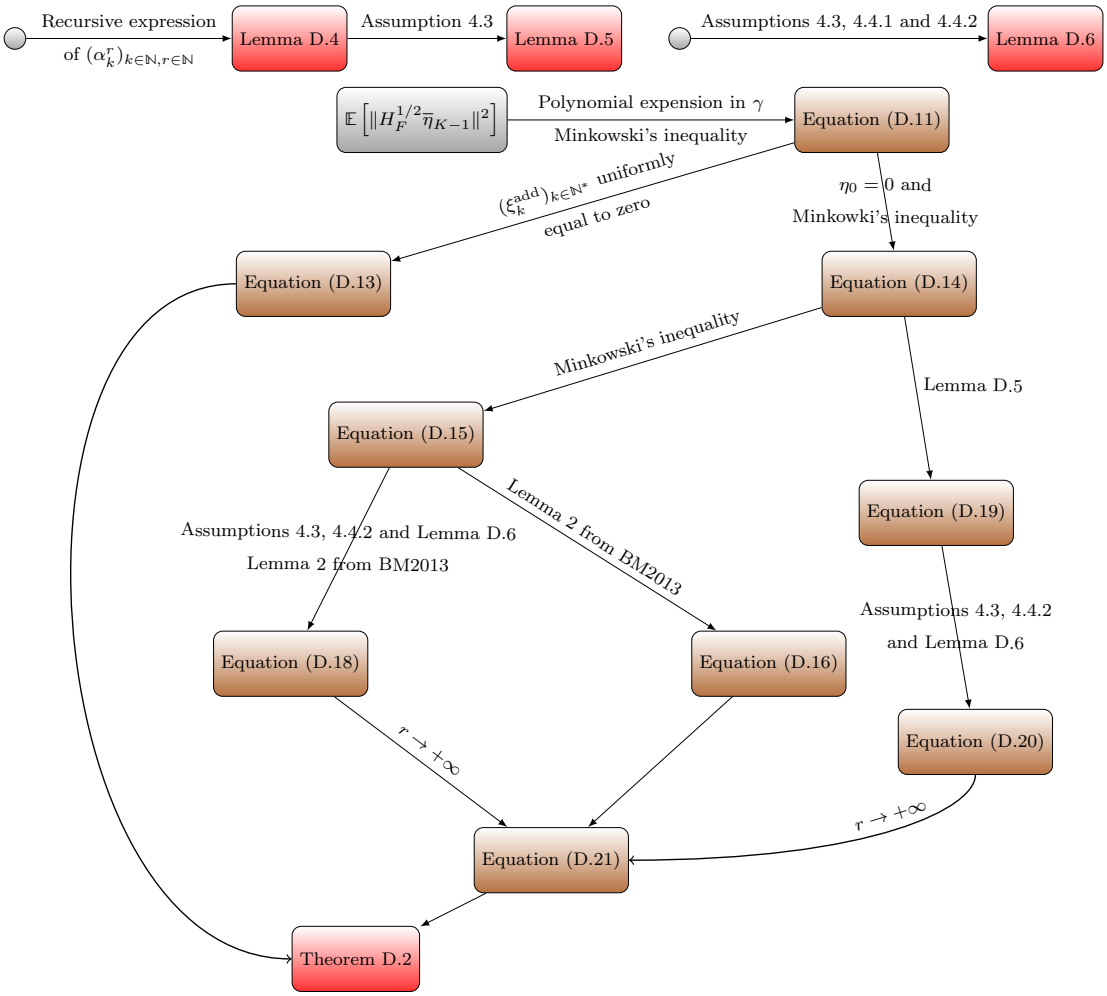


Figure D.2: Proof principle of Theorem D.2.

Then, we decompose $\mathbb{E}[\|H_F^{1/2}\overline{\eta}_{K-1}\|^2]$ in the following way using Minkowski's inequality A.6:

$$\sqrt{\mathbb{E}\left[\|H_F^{1/2}\overline{\eta}_{K-1}\|^2\right]} \leq \mathbb{E}[\|H_F^{1/2}\sum_{i=0}^{r}\overline{\eta}_{K-1}^i\|^2]^{1/2} + \mathbb{E}[\|H_F^{1/2}(\overline{\eta}_{K-1} - \sum_{i=0}^{r}\overline{\eta}_{K-1}^i)\|^2]^{1/2}.$$

The goal is then to establish a bound for the two above quantities.

D.3.2 Lemmas for the noise process

In this Subsection, we provide lemmas for the noise process, and thus we suppose that $\eta_0 = 0$. The noise-free process is later considered in Subsection D.3.3 and puts together with the results of the coming Subsection. The sketch of the proof relies on establishing two bounds.

- For r, k in $\mathbb{N} \times \mathbb{N}^*$, noting $\alpha_k^r = \eta_k - \sum_{i=0}^r \eta_k^i$, the first one is a bound on $\mathbb{E}[\|H_F^{1/2}\overline{\alpha}_{K-1}^r\|^2]$ that tends to zero when r tends to $+\infty$.
- The second one is on $\sum_{i=0}^r \mathbb{E}[\|H_F^{1/2}\overline{\eta}_{K-1}^i\|^2]$ and is established using Lemma 2 from [Bach and Moulines, 2013]. It will correspond to the final variance term and it involves the additive noise's covariance $\mathfrak{C}_{\text{ania}}$.

In the following, we provide Lemmas D.4 to D.6. Let r, k in $\mathbb{N} \times \mathbb{N}^*$.

- Lemma D.4 builds a recursive expression of $\alpha_k^r = \eta_k - \sum_{i=0}^r \eta_k^i$.
- Lemma D.5 provides a bound on $\mathbb{E}[\|H_F^{1/2}\overline{\alpha}_{K-1}^r\|^2]$ which involves $\mathbb{E}\|\xi_k^{\text{mult}}\left(\eta_{k-1}^r\right)\|^2$.

- Lemma D.6 bounds the covariance of η_{k-1}^r, this result will be necessary when computing the expectation of $\xi_k^{\text{mult}}\left(\eta_{k-1}^r\right)^{\otimes 2}$.

Below, we provide the lemma that builds a recursive expression of $\eta_k - \sum_{i=0}^r \eta_k^i$, with k, r in $\mathbb{N}^*$.

Lemma D.4 (A recursion on $\eta_k - \sum_{i=0}^r \eta_k^i$). *Under the setting given in Definition 4.1, considering that $\xi_k^{\text{mult}}(\cdot)$ is linear (Assumption 4.3), for any k in $\mathbb{N}^*$ and any step-size $\gamma > 0$, considering $(\eta_k^r)_{r\in\mathbb{N}}$ as given by Equation (D.12), denoting for r in $\mathbb{N}$, $\alpha_k^r = \eta_k - \sum_{i=0}^r \eta_k^i$, we have the following recursive expression for the sequence of iterate $(\alpha_k^r)_{r\in\mathbb{N}}$:*

$$\forall r \geq 0, \alpha_k^r = (\mathrm{I}_d - \gamma H_F)\alpha_{k-1}^r + \xi_k^{\text{mult}}\left(\alpha_{k-1}^r\right) + \gamma\xi_k^{\text{mult}}\left(\eta_{k-1}^r\right) .$$

Proof Let k in $\mathbb{N}^*$, the proof is done by recursion. For $r = 0$, by Definitions 4.1 and 4.2, we have $\eta_k = \eta_{k-1} - \gamma\nabla F(w_{k-1}) + \gamma\xi_k(\eta_{k-1}) = (\mathrm{I}_d - \gamma H_F)\eta_{k-1} + \gamma\xi_k^{\text{add}} + \gamma\xi_k^{\text{mult}}(\eta_{k-1})$, which gives:

$$\begin{aligned}
\alpha_k^0 = \eta_k - \eta_k^0 &= \left\{(\mathrm{I}_d - \gamma H_F)\eta_{k-1} + \gamma\xi_k^{\text{add}} + \gamma\xi_k^{\text{mult}}(\eta_{k-1})\right\} - \left\{(\mathrm{I}_d - \gamma H_F)\eta_{k-1}^0 + \gamma\xi_k^{\text{add}}\right\} \\
&= (\mathrm{I}_d - \gamma H_F)(\eta_{k-1} - \eta_{k-1}^0) + \gamma\xi_k^{\text{mult}}(\eta_{k-1}) \\
&= (\mathrm{I}_d - \gamma H_F)(\eta_{k-1} - \eta_{k-1}^0) + \gamma\xi_k^{\text{mult}}\left(\eta_{k-1} - \eta_{k-1}^0\right) + \gamma\xi_k^{\text{mult}}\left(\eta_{k-1}^0\right) ,
\end{aligned}$$

which is possible because ξ_k^{mult} is linear (Assumption 4.3). To go from r to $r+1$, we have $\alpha_k^{r+1} = \eta_k - \sum_{i=0}^{r+1}\eta_k^i = \eta_k - \sum_{i=0}^r \eta_k^i - \eta_k^{r+1}$. Then by definition of η_k^{r+1} and using the hypothesis:

$$\begin{aligned}
\alpha_k^{r+1} &= (\mathrm{I}_d - \gamma H_F)\left(\eta_{k-1} - \sum_{i=0}^r \eta_{k-1}^i\right) + \xi_k^{\text{mult}}\left(\eta_{k-1} - \sum_{i=0}^r \eta_{k-1}^i\right) + \gamma\xi_k^{\text{mult}}\left(\eta_{k-1}^r\right) \\
&\quad - (\mathrm{I}_d - \gamma H_F)\eta_{k-1}^{r+1} - \gamma\xi_k^{\text{mult}}\left(\eta_{k-1}^r\right) \\
&= (\mathrm{I}_d - \gamma H_F)\left(\eta_{k-1} - \sum_{i=0}^{r+1} \eta_{k-1}^i\right) + \xi_k^{\text{mult}}\left(\eta_{k-1} - \sum_{i=0}^{r+1} \eta_{k-1}^i\right) + \gamma\xi_k^{\text{mult}}\left(\eta_{k-1}^{r+1}\right) ,
\end{aligned}$$

again by linearity. This concludes the proof. ∎

The next lemma is the adaptation to our settings of Lemma 1 from Bach and Moulines [2013]. We give a bound on $\mathbb{E}[\|H_F^{1/2}\overline{\alpha}_{K-1}^r\|^2]$ with a quantity that tends to 0. This result will be used in the final demonstration of Theorem D.2.

Lemma D.5 (Bound on $\eta_K - \sum_{i=0}^r \eta_K^i$). *Under the setting given in Definition 4.1, considering that ξ_k^{mult} is linear (Assumption 4.3), for any r, K in $\mathbb{N}\times\mathbb{N}^*$ and any step-size γ s.t. $\gamma(R_F^2 + \mathcal{M}_2) \leq 1$, the recursion $\alpha_K^r = \eta_K - \sum_{i=0}^r \eta_K^i$ verifies the following bound:*

$$\forall r \geq 0, (1 - \gamma(R_F^2 + \mathcal{M}_2))\mathbb{E}\left\langle \overline{\alpha}_{K-1}^r, H_F\overline{\alpha}_{K-1}^r\right\rangle \leq \frac{\gamma}{K}\sum_{k=1}^K \mathbb{E}\|\xi_k^{\text{mult}}\left(\eta_{k-1}^r\right)\|^2 .$$

Proof Let r, k in $\mathbb{N}\times\mathbb{N}^*$, we denote $\alpha_k^r = \eta_k - \sum_{i=0}^r \eta_k^i$, then we have shown in Lemma D.4 that:

$$\alpha_k^r = (\mathrm{I}_d - \gamma H_F)\alpha_{k-1}^r + \xi_k^{\text{mult}}\left(\alpha_{k-1}^r\right) + \gamma\xi_k^{\text{mult}}\left(\eta_{k-1}^r\right) .$$

Taking the squared norm and developing it:

$$\begin{aligned}
\|\alpha_k^r\|^2 = \left\|\alpha_{k-1}^r\right\|^2 &+ 2\gamma\left\langle \alpha_{k-1}^r, \xi_k^{\text{mult}}\left(\alpha_{k-1}^r\right) + \xi_k^{\text{mult}}\left(\eta_{k-1}^r\right) - H_F\alpha_{k-1}^r\right\rangle \\
&+ \gamma^2\|\xi_k^{\text{mult}}\left(\alpha_{k-1}^r\right) + \xi_k^{\text{mult}}\left(\eta_{k-1}^r\right) - H_F\alpha_{k-1}^r\|^2 ,
\end{aligned}$$

and developing the last term with Inequality 1 leads to:

$$\|\alpha_k^r\|^2 \leq \left\|\alpha_{k-1}^r\right\|^2 + 2\gamma \left\langle \alpha_{k-1}^r, \xi_k^{\text{mult}}\left(\alpha_{k-1}^r\right) + \xi_k^{\text{mult}}\left(\eta_{k-1}^r\right) - H_F \alpha_{k-1}^r \right\rangle + 2\gamma^2 \left\{ \|\xi_k^{\text{mult}}\left(\eta_{k-1}^r\right)\|^2 + \|H_F \alpha_{k-1}^r - \xi_k^{\text{mult}}\left(\alpha_{k-1}^r\right)\|^2 \right\}.$$

Because α_{k-1}^r is $\mathcal{F}_{k-1}$-measurable and $\mathbb{E}[\xi_k^{\text{mult}}\left(\alpha_{k-1}^r\right) \mid \mathcal{F}_{k-1}] = 0$ (expectation of $\xi_k^{\text{mult}}(\cdot)$ is zero, see Definitions 4.1 and 4.2), taking expectation w.r.t. the σ-algebra $\mathcal{F}_{k-1}$, using Assumption 4.3 and again Definition 4.1 gives:

$$\begin{aligned}\mathbb{E}[\|H_F \alpha_{k-1}^r - \xi_k^{\text{mult}}\left(\alpha_{k-1}^r\right)\|^2 \mid \mathcal{F}_{k-1}] &= \mathbb{E}[\|H_F \alpha_{k-1}^r\|^2 \mid \mathcal{F}_{k-1}] \\ &\quad + \mathbb{E}[\|\xi_k^{\text{mult}}\left(\alpha_{k-1}^r\right)\|^2 \mid \mathcal{F}_{k-1}] \\ &\leq (R_F^2 + \mathcal{M}_2)\|H_F^{1/2} \alpha_{k-1}^r\|^2 .\end{aligned}$$

Hence:

$$\begin{aligned}\mathbb{E}[\|\alpha_k^r\|^2 \mid \mathcal{F}_{k-1}] \leq \|\alpha_{k-1}^r\|^2 &- 2\gamma(1 - \gamma(R_F^2 + \mathcal{M}_2)) \left\langle \alpha_{k-1}^r, H_F \alpha_{k-1}^r \right\rangle \\ &+ 2\gamma^2 \mathbb{E}[\|\xi_k^{\text{mult}}\left(\eta_{k-1}^r\right)\|^2 \mid \mathcal{F}_{k-1}] ,\end{aligned}$$

which gives when taking full expectation and averaging over K in $\mathbb{N}^*$:

$$\begin{aligned}(1 - \gamma(R_F^2 + \mathcal{M}_2)) \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}\left\langle \alpha_{k-1}^r, H_F \alpha_{k-1}^r \right\rangle &\leq \frac{1}{2\gamma}(\|\alpha_0^r\|^2 - \left\|\alpha_{k-1}^r\right\|^2) \\ &\quad + \frac{\gamma}{K} \sum_{k=1}^{K} \mathbb{E}[\|\xi_k^{\text{mult}}\left(\eta_{k-1}^r\right)\|^2] ,\end{aligned}$$

and by convexity $\left\langle \overline{\alpha}_{K-1}^r, H \overline{\alpha}_{K-1}^r \right\rangle \leqslant \frac{1}{K} \sum_{k=1}^{K} \left\langle \alpha_{k-1}^r, H_F \alpha_{k-1}^r \right\rangle$, which allows to conclude as $\alpha_0^r = 0$. ■

In below lemma, we bound $\mathbb{E}\left[\eta_{k-1}^r \otimes \eta_{k-1}^r\right]$ for r, k in $\mathbb{N} \times \mathbb{N}^*$. It is required because we will use Lemma 2 from Bach and Moulines [2013] and apply it to the sequence $(\eta_{k-1}^r)_{k \in \mathbb{N}^*, r \in \mathbb{N}}$. The noise process of this sequence is equal to $\xi_k^{\text{mult}}\left(\eta_{k-1}^{r-1}\right)$; and computing the expectation of its covariance involves knowing $\mathbb{E}\left[\eta_{k-1}^r \otimes \eta_{k-1}^r\right]$.

Lemma D.6 (Bounding the covariance of η_{k-1}^r)**.** *Under the setting in Definition 4.1, under Assumptions 4.1, 4.3 and 4.4, i.e. considering that $\xi_k^{\text{mult}}(\cdot)$ is linear, for any K in $\mathbb{N}^*$, any step-size $\gamma > 0$, and for any $r \geq 0$, we have the following bound on the covariance of η_{k-1}^r:*

$$\mathbb{E}\left[\eta_{k-1}^r \otimes \eta_{k-1}^r\right] \preccurlyeq \gamma^{r+1} \text{Ш}_{\text{add}} \text{Ш}_{\text{mult}}^r \mathrm{I}_d .$$

Proof

Let $r > 0$, we first prove by recursion that we have:

$$\forall k > 0 \,, \eta_k^{r+1} = \gamma \sum_{i=1}^{k} (\mathrm{I}_d - \gamma H_F)^{k-i} \xi_i^{\text{mult}}(\eta_{i-1}^r) .$$

For $k = 0$, we indeed have $\eta_0^{r+1} = 0$. To go from k to $k+1$:

$$\begin{aligned}\eta_{k+1}^{r+1} &= (\mathrm{I}_d - \gamma H_F)\eta_k^{r+1} + \gamma \xi_{k+1}^{\text{mult}}(\eta_k^r) \qquad \text{by definition,} \\ &= \gamma \sum_{i=1}^{k} (\mathrm{I}_d - \gamma H_F)^{k-i} \xi_i^{\text{mult}}(\eta_{i-1}^r) + \gamma (\mathrm{I}_d - \gamma H_F)^{(k+1)-(k+1)} \xi_{k+1}^{\text{mult}}(\eta_k^r) ,\end{aligned}$$

by hypothesis, which allows concluding.

We now prove by recursion the main result of the lemma.

Initialization. For $r = 0$, by definition, we have $\eta_k^0 = (\mathrm{I}_d - \gamma H_F)\eta_{k-1}^0 + \gamma \xi_k^{\mathrm{add}}$, unrolling the sum gives $\eta_k^0 = (\mathrm{I}_d - \gamma H_F)^k \eta_0^0 + \gamma \sum_{i=1}^k (\mathrm{I}_d - \gamma H_F)^{k-i} \xi_i^{\mathrm{add}}$. Because we consider $\eta_0^0 = 0$ and given that the sequence of noise $(\xi_i^{\mathrm{add}})_{i \in [\![1,k]\!]}$ is independent at each iterations, we have:

$$\mathbb{E}\left[\eta_k^0 \otimes \eta_k^0\right] = \gamma^2 \sum_{i=1}^k (\mathrm{I}_d - \gamma H_F)^{k-i} \mathbb{E}\left[\xi_i^{\mathrm{add}} \otimes \xi_i^{\mathrm{add}}\right] (\mathrm{I}_d - \gamma H_F)^{k-i} .$$

Because the sequence of additive noise $(\xi_i^{\mathrm{add}})_{i \in \mathbb{N}^*}$is i.i.d., for any i in $\{1, \cdots, k\}$, we have that $\mathbb{E}\left[\xi_i^{\mathrm{add}} \otimes \xi_i^{\mathrm{add}}\right] = \mathfrak{C}_{\mathrm{ania}} \preccurlyeq \text{Ш}_{\mathrm{add}} H_F$ (Assumption 4.4.1), hence:

$$\mathbb{E}\left[\eta_k^0 \otimes \eta_k^0\right] \preccurlyeq \gamma^2 \sum_{i=1}^k (\mathrm{I}_d - \gamma H_F)^{k-i} \text{Ш}_{\mathrm{add}} H_F (\mathrm{I}_d - \gamma H_F)^{k-i} .$$

These matrices commute:

$$\begin{aligned}
\mathbb{E}\left[\eta_k^0 \otimes \eta_k^0\right] &\preccurlyeq \gamma^2 \text{Ш}_{\mathrm{add}} \sum_{i=1}^k (\mathrm{I}_d - \gamma H_F)^{2k-2i} H_F \,, \text{ and because it is a geometric sum:} \\
&\preccurlyeq \gamma^2 \text{Ш}_{\mathrm{add}} \left(\mathrm{I}_d - (\mathrm{I}_d - \gamma H_F)^{2k-2}\right) \left(\mathrm{I}_d - (\mathrm{I}_d - \gamma H_F)^2\right)^{-1} H_F \\
&\preccurlyeq \gamma^2 \text{Ш}_{\mathrm{add}} \left(\mathrm{I}_d - (\mathrm{I}_d - \gamma H_F)^{2k-2}\right) \left(2\gamma H_F - \gamma^2 H_F^2\right)^{-1} H_F \\
&\preccurlyeq \gamma \text{Ш}_{\mathrm{add}} H_F^{-1} H_F \qquad \text{because } \gamma H_F \preccurlyeq \mathrm{I}_d, \\
&\preccurlyeq \gamma \text{Ш}_{\mathrm{add}} \mathrm{I}_d \,.
\end{aligned}$$

Recursion. Let $r \geq 0$, to go from r to $r+1$, we start writing:

$$\eta_k^{r+1} \otimes \eta_k^{r+1} = \gamma^2 \sum_{i=1}^k (\mathrm{I}_d - \gamma H_F)^{k-1-i} \xi_i^{\mathrm{mult}}(\eta_{i-1}^r) \otimes \xi_i^{\mathrm{mult}}(\eta_{i-1}^r) (\mathrm{I}_d - \gamma H_F)^{k-1-i} .$$

Now we use linearity of the multiplicative noise (Assumption 4.3), thus there exists a matrix Ξ_k in $\mathbb{R}^{d \times d}$ s.t. for any z in $\mathbb{R}^d$, we have $\xi_k^{\mathrm{mult}}(z) = \Xi_k z$, and it leads to:

$$\eta_k^{r+1} \otimes \eta_k^{r+1} = \gamma^2 \sum_{i=1}^k (\mathrm{I}_d - \gamma H_F)^{k-i} \Xi_i (\eta_{i-1}^r \otimes \eta_{i-1}^r) \Xi_i^\top (\mathrm{I}_d - \gamma H_F)^{k-i} .$$

Taking full expectation, we have:

$$\begin{aligned}
\mathbb{E}\left[\eta_k^{r+1} \otimes \eta_k^{r+1}\right] &= \gamma^2 \sum_{i=1}^k (\mathrm{I}_d - \gamma H_F)^{k-i} \mathbb{E}\left[\mathbb{E}\left[\Xi_i (\eta_{i-1}^r \otimes \eta_{i-1}^r) \Xi_i^\top \;\middle|\; \sigma(\Xi_i)\right]\right] (\mathrm{I}_d - \gamma H_F)^{k-i} \\
&= \gamma^2 \sum_{i=1}^k (\mathrm{I}_d - \gamma H_F)^{k-i} \mathbb{E}\left[\Xi_i \mathbb{E}[\eta_{i-1}^r \otimes \eta_{i-1}^r \mid \sigma(\Xi_i)] \Xi_i^\top\right] (\mathrm{I}_d - \gamma H_F)^{k-i} ,
\end{aligned}$$

and because for any i in $\{1, \cdots, k\}$, η_{i-1}^r is independent of Ξ_i, we have $\mathbb{E}\left[\eta_{i-1}^r \otimes \eta_{i-1}^r \;\middle|\; \sigma(\Xi_i)\right] = \mathbb{E}\left[\eta_{i-1}^r \otimes \eta_{i-1}^r\right] \preccurlyeq \gamma^{r+1} \text{Ш}_{\mathrm{add}} \text{Ш}_{\mathrm{mult}}^r \mathrm{I}_d$, where we use the hypothesis for r. We have in the end:

$$\mathbb{E}\left[\eta_k^{r+1} \otimes \eta_k^{r+1}\right] \preccurlyeq \gamma^{r+3} \text{Ш}_{\mathrm{add}} \text{Ш}_{\mathrm{mult}}^r \sum_{i=1}^k (\mathrm{I}_d - \gamma H_F)^{k-i} \mathbb{E}\left[\Xi_i \Xi_i^\top\right] (\mathrm{I}_d - \gamma H_F)^{k-i} .$$

Furthermore, by Assumption 4.4.2 we have $\mathbb{E}\left[\Xi_i \Xi_i^\top\right] \preccurlyeq \text{Ш}_{\text{mult}} H_F$, thus:

$$\begin{aligned}\mathbb{E}\left[\eta_k^{r+1} \otimes \eta_k^{r+1}\right] &\preccurlyeq \gamma^{r+3} \text{Ш}_{\text{add}} \text{Ш}_{\text{mult}}^{r+1} \sum_{i=1}^{k} (\mathrm{I}_d - \gamma H_F)^{2k-2-2i} H_F \\ &\preccurlyeq \gamma^{r+3} \text{Ш}_{\text{add}} \text{Ш}_{\text{mult}}^{r+1} \gamma^{-1} H_F^{-1} H_F \,,\end{aligned}$$

because $\sum_{i=1}^{k} (\mathrm{I}_d - \gamma H_F)^{2k-2-2i} = \left(\mathrm{I}_d - (\mathrm{I}_d - \gamma H_F)^{2k}\right) \left(2\gamma H_F - \gamma^2 H_F^2\right)^{-1} \preccurlyeq \gamma^{-1} H_F^{-1}$. In the end, we have $\mathbb{E}[\eta_k^{r+1} \otimes \eta_k^{r+1}] \preccurlyeq \gamma^{r+2} \text{Ш}_{\text{add}} \text{Ш}_{\text{mult}}^{r+1} \mathrm{I}_d$, which concludes the proof.

■

D.3.3 Final theorem

In this section, we gather the pieces of proof required to demonstrate Theorem 4.2. As done in Section D.2, we consider separately the noise process and the noise-free process, then put them together to obtain the final result.

Theorem D.2 (Linear multiplicative noise, convex case)**.** *Under Assumption 4.1, under Assumptions 4.3 and 4.4 i.e. with a linear multiplicative noise, considering any constant step-size γ such that $\gamma(R_F^2 + \mathcal{M}_2) \leq 1$ and $4\gamma \text{Ш}_{\text{mult}} R_F^2 \leq 1$, then for any K in $\mathbb{N}^*$, the sequence $(w_k)_{k \in \mathbb{N}^*}$ produced by a setting such as in Definition 4.1, verifies the following bound:*

$$\mathbb{E}[F(\overline{w}_{K-1}) - F(w_*)] \leq \frac{1}{2K} \left(\frac{\|\eta_0\|}{\sqrt{\gamma}} + \sqrt{\operatorname{Tr}\left(\mathfrak{C}_{\text{ania}} H_F^{-1}\right)} + \frac{(\gamma d \text{Ш}_{\text{add}} \text{Ш}_{\text{mult}})^{1/2}}{1 - \sqrt{\gamma \text{Ш}_{\text{mult}}}} \right)^2 .$$

Proof Let K in $\mathbb{N}^*$, the proof relies on the proof presented by Bach and Moulines [2013] and is done separately for the noise process and for the noise-free process that depends only on the initial condition. The bounds may then be added (see the discussion in Subsection D.3.1).

Noise-free process. As in section A.3 from Bach and Moulines [2013], we assume here that the additive noise $(\xi_k^{\text{add}})_{k \in \mathbb{N}^*}$ is uniformly equal to zero and that $\gamma(R_F^2 + \mathcal{M}_2) \leq 1$. Using Definitions 4.1 and 4.2, we thus have for any k in $\mathbb{N}^*$ that $\eta_k = \eta_{k-1} - \gamma H_F \eta_{k-1} + \gamma \xi_k^{\text{mult}}(\eta_{k-1})$, it flows:

$$\begin{aligned}\mathbb{E}[\|\eta_k\|^2] &= \mathbb{E}[\|\eta_{k-1}\|^2] - 2\gamma \mathbb{E}[\langle \eta_{k-1}, H_F \eta_{k-1} \rangle] + \gamma^2 \mathbb{E}[\|H_F \eta_{k-1} - \xi_k^{\text{mult}}(\eta_{k-1})\|^2] \\ &= \mathbb{E}[\|\eta_{k-1}\|^2] - 2\gamma \mathbb{E}[\langle \eta_{k-1}, H_F \eta_{k-1} \rangle] + \gamma^2 \mathbb{E}[\|H_F \eta_{k-1}\|^2] + \gamma^2 \mathbb{E}[\|\xi_k^{\text{mult}}(\eta_{k-1})\|^2] \,.\end{aligned}$$

Considering that $H_F \preccurlyeq \operatorname{Tr}(H_F) \mathrm{I}_d \preccurlyeq R_F^2 \mathrm{I}_d$ and using Assumption 4.3, we obtain:

$$\mathbb{E}[\|\eta_k\|^2] \leq \mathbb{E}[\|\eta_{k-1}\|^2] - 2\gamma \mathbb{E}[\|H_F^{1/2} \eta_{k-1}\|^2] + \gamma^2 (R_F^2 + \mathcal{M}_2) \mathbb{E}[\|H_F^{1/2} \eta_{k-1}\|^2] \,.$$

Because the step-size γ is s.t. $\gamma(R_F^2 + \mathcal{M}_2) \leq 1$, we recover that in the absence of noise, we have:

$$\mathbb{E}[\|H_F^{1/2} \overline{\eta}_{K-1}\|^2] \leq \frac{\|\eta_0\|^2}{\gamma K} \,. \tag{D.13}$$

Noise process. Now, all the following results comes from Subsection D.3.2 where we assume that $\eta_0 = w_0 - w_* = 0$, we start using Minkowski's inequality A.6:

$$\mathbb{E}\left[\|H_F^{1/2} \overline{\eta}_{K-1}\|^2\right]^{1/2} \leq \mathbb{E}\left[\|H_F^{1/2} \sum_{i=0}^{r} \overline{\eta}_{K-1}^i\|^2\right]^{1/2} + \mathbb{E}\left[\|H_F^{1/2} (\overline{\eta}_{K-1} - \sum_{i=0}^{r} \overline{\eta}_{K-1}^i)\|^2\right]^{1/2} . \tag{D.14}$$

First term.

Let $r \in \mathbb{N}$, again using Minkowski's inequality A.6, we have

$$\mathbb{E}[\|H_F^{1/2}\sum_{i=0}^{r}\overline{\eta}_{K-1}^{i}\|^2]^{1/2} \leq \sum_{i=0}^{r}\mathbb{E}[\|H_F^{1/2}\overline{\eta}_{K-1}^{i}\|^2]^{1/2}$$
$$= \mathbb{E}[\|H_F^{1/2}\overline{\eta}_{K-1}^{0}\|^2]^{1/2} + \sum_{i=1}^{r}\mathbb{E}[\|H_F^{1/2}\overline{\eta}_{K-1}^{i}\|^2]^{1/2}\,. \tag{D.15}$$

By Equation (D.12), we have $\eta_k^0 = (\mathrm{I}_d - \gamma H_F)\eta_{k-1}^0 + \gamma\xi_k^{\mathrm{add}}$, hence to bound the first term, we have to apply Lemma 2 from Bach and Moulines [2013] to the sequence $(\eta_{k-1}^0)_{k\in\mathbb{N}^*}$ and we obtain

$$\mathbb{E}[\|H_F^{1/2}\overline{\eta}_{K-1}^{0}\|^2] \leq \mathrm{Tr}\left(\mathfrak{C}_{\mathrm{ania}}H_F^{-1}\right)/K\,. \tag{D.16}$$

Let i in $\{1,\cdots,r\}$, to bound the second term, we have to apply Lemma 2 from Bach and Moulines [2013] to the sequence $(\eta_{k-1}^i)_{k\in\mathbb{N}^*}$. To do so, we bound the covariance of the noise which is here equal to $\xi_k^{\mathrm{mult}}\left(\eta_{k-1}^{i-1}\right)$ (by definition of η_{k-1}^i, see Equation (D.12)).

Because the multiplicative noise is linear, using Assumption 4.3, there exists a matrix Ξ_k in $\mathbb{R}^{d\times d}$ s.t. $\xi_k^{\mathrm{mult}}\left(\eta_{k-1}^{i-1}\right) = \Xi_k\eta_{k-1}^{i-1}$. It follows that taking the expectation w.r.t to the σ-algebra $\sigma(\Xi_k)$, and because η_{k-1}^{i-1} is independent of it, using Lemma D.6, we have:

$$\mathbb{E}\left[\eta_{k-1}^{i-1}\otimes\eta_{k-1}^{i-1} \mid \sigma(\Xi_k)\right] = \mathbb{E}\left[\eta_{k-1}^{i-1}\otimes\eta_{k-1}^{i-1}\right] \preccurlyeq \gamma^i\mathrm{III}_{\mathrm{add}}\mathrm{III}_{\mathrm{mult}}^{i-1}\mathrm{I}_d\,.$$

Thus, the noise $\xi_k^{\mathrm{mult}}\left(\eta_{k-1}^{i-1}\right)$ is such that:

$$\mathbb{E}[\xi_k^{\mathrm{mult}}\left(\eta_{k-1}^{i-1}\right)\otimes\xi_k^{\mathrm{mult}}\left(\eta_{k-1}^{i-1}\right) \mid \sigma(\Xi_k)] = \Xi_k\mathbb{E}\left[\eta_{k-1}^{i-1}\otimes\eta_{k-1}^{i-1}\right]\Xi_k^\top \preccurlyeq \gamma^i\mathrm{III}_{\mathrm{add}}\mathrm{III}_{\mathrm{mult}}^{i-1}\Xi_k\Xi_k^\top\,.$$

Taking full expectation, we furthermore consider Assumption 4.4.2 which gives that: $\mathbb{E}\left[\Xi_i\Xi_i^\top\right] \preccurlyeq \mathrm{III}_{\mathrm{mult}}H_F$, hence:

$$\mathbb{E}\left[\xi_k^{\mathrm{mult}}\left(\eta_{k-1}^{i-1}\right)\otimes\xi_k^{\mathrm{mult}}\left(\eta_{k-1}^{i-1}\right)\right] \leq \gamma^i\mathrm{III}_{\mathrm{add}}\mathrm{III}_{\mathrm{mult}}^{i}H_F\,. \tag{D.17}$$

Using Lemma 2 from Bach and Moulines [2013] results to:

$$\sum_{i=1}^{r}\mathbb{E}[\|H_F^{1/2}\overline{\eta}_{K-1}^{i}\|^2]^{1/2} \leq \sum_{i=1}^{r}\gamma^i\mathrm{III}_{\mathrm{add}}\mathrm{III}_{\mathrm{mult}}^{i}\mathrm{Tr}\left(H_FH_F^{-1}\right)/K\,. \tag{D.18}$$

In the end, we obtain from Equation (D.15):

$$\mathbb{E}[\|H_F^{1/2}\sum_{i=0}^{r}\overline{\eta}_{K-1}^{i}\|^2]^{1/2} \leq \frac{\sqrt{\mathrm{Tr}\left(\mathfrak{C}_{\mathrm{ania}}H_F^{-1}\right)}}{\sqrt{K}} + \frac{\sqrt{d\mathrm{III}_{\mathrm{add}}}}{\sqrt{K}}\sum_{i=1}^{r}\gamma^{i/2}\mathrm{III}_{\mathrm{mult}}^{i/2}$$
$$\leq \frac{\sqrt{\mathrm{Tr}\left(\mathfrak{C}_{\mathrm{ania}}H_F^{-1}\right)}}{\sqrt{K}} + \frac{\sqrt{\gamma d\mathrm{III}_{\mathrm{add}}\mathrm{III}_{\mathrm{mult}}}\left(1-(\gamma\mathrm{III}_{\mathrm{mult}})^{r/2}\right)}{\sqrt{K}\left(1-\sqrt{\gamma\mathrm{III}_{\mathrm{mult}}}\right)}\,.$$

Second term.

If $\gamma(R_F^2+\mathcal{M}_2) \leq 1$, Lemma D.5 gives:

$$\mathbb{E}\left\langle\overline{\eta}_{K-1}-\sum_{i=0}^{r}\overline{\eta}_{K-1}^{i}, H(\overline{\eta}_{K-1}-\sum_{i=0}^{r}\overline{\eta}_{K-1}^{i})\right\rangle \leq \frac{\gamma}{(1-\gamma(R_F^2+\mathcal{M}_2))K}\sum_{k=1}^{K}\mathbb{E}\left[\|\xi_k^{\mathrm{mult}}\left(\eta_{k-1}^{r}\right)\|^2\right]\,. \tag{D.19}$$

Furthermore, $\left\|\xi_k^{\text{mult}}\left(\eta_{k-1}^r\right)\right\|^2 = \operatorname{Tr}\left(\xi_k^{\text{mult}}\left(\eta_{k-1}^r\right)^{\otimes 2}\right)$, by reusing what has been written in the previous paragraph (Equation (D.17)), we obtain:

$$\begin{aligned}\|\xi_k^{\text{mult}}\left(\eta_{k-1}^r\right)\|^2 &\leq \gamma^{r+1}\text{III}_{\text{add}}\text{III}_{\text{mult}}^{r+1}\operatorname{Tr}\left(H_F\right)\\ &\leq \gamma^{r+1}\text{III}_{\text{add}}\text{III}_{\text{mult}}^{r+1}R_F^2 \qquad \text{(Definition 4.1)}.\end{aligned}$$

It follows that we have:

$$\mathbb{E}\left\langle \overline{\eta}_{K-1} - \sum_{i=0}^{r}\overline{\eta}_{K-1}^i, H(\overline{\eta}_{K-1} - \sum_{i=0}^{r}\overline{\eta}_{K-1}^i)\right\rangle \leq \frac{\gamma^{r+2}\text{III}_{\text{add}}\text{III}_{\text{mult}}^{r+1}R_F^2}{(1-\gamma(R_F^2+\mathcal{M}_2))}\,. \tag{D.20}$$

Putting things together. In the end, from the Minkowski decomposition done in Equation (D.14), we combine the two terms and it leads to:

$$\begin{aligned}\mathbb{E}\left[\left\langle \overline{\eta}_{K-1}, H_F\overline{\eta}_{K-1}\right\rangle\right]^{1/2} \leq &\left(\frac{\gamma^{r+2}\text{III}_{\text{add}}\text{III}_{\text{mult}}^{r+1}R_F^2}{(1-\gamma(R_F^2+\mathcal{M}_2))}\right)^{1/2} + \frac{\sqrt{\operatorname{Tr}\left(\mathfrak{C}_{\text{ania}}H_F^{-1}\right)}}{\sqrt{K}}\\ &+ \frac{\sqrt{\gamma d\text{III}_{\text{add}}\text{III}_{\text{mult}}}\left(1-(\gamma\text{III}_{\text{mult}})^{r/2}\right)}{\sqrt{K}\left(1-\sqrt{\gamma\text{III}_{\text{mult}}}\right)}\,.\end{aligned}$$

This implies that for any $\gamma\text{III}_{\text{mult}} \leq 1$, we obtain, by letting r tend to $+\infty$:

$$\mathbb{E}\left[\left\langle \overline{\eta}_{K-1}, H_F\overline{\eta}_{K-1}\right\rangle\right]^{1/2} \leq \frac{1}{\sqrt{K}}\left(\sqrt{\operatorname{Tr}\left(\mathfrak{C}_{\text{ania}}H_F^{-1}\right)} + \frac{(\gamma d\text{III}_{\text{add}}\text{III}_{\text{mult}})^{1/2}}{1-\sqrt{\gamma\text{III}_{\text{mult}}}}\right)\,. \tag{D.21}$$

Final bound. We now take results derived from the part without noise, and the part with noise, to get:

$$\mathbb{E}[\left\langle \overline{\eta}_{K-1}, H_F\overline{\eta}_{K-1}\right\rangle]^{1/2} \leq \frac{1}{\sqrt{K}}\left(\frac{\|\eta_0\|}{\sqrt{\gamma}} + \sqrt{\operatorname{Tr}\left(\mathfrak{C}_{\text{ania}}H_F^{-1}\right)} + \frac{(\gamma d\text{III}_{\text{add}}\text{III}_{\text{mult}})^{1/2}}{1-\sqrt{\gamma\text{III}_{\text{mult}}}}\right)\,,$$

which leads to the desired result considering that $4\gamma\text{III}_{\text{mult}} \leq 1$. ■

D.4 Validity of the assumptions made on the random fields

In this section, we verify that all the assumptions on the random fields done in Subsection 4.2.1 are fulfilled in the setting of compressed least-squares regression analyzed in Section 4.3. To do so, we first need to define the filtrations considered in this section.

For k in $\mathbb{N}^*$, we note u_k the noise that controls the randomization $\mathcal{C}_k(\cdot)$ at round k. In Section 4.2, we have denoted by $\mathcal{F}_k$ the σ-algebra generated by $(x_1, \varepsilon_1, u_1, \cdots, x_k, \varepsilon_k)$. In particular, w_k and $\overline{w}_k$ are $\mathcal{F}_k$-measurable. We also consider the following filtrations.

Definition D.1. *We note $(\mathcal{G}_k)_{k\in\mathbb{N}}$ the filtration associated with the features noise, $(\mathcal{H}_k)_{k\in\mathbb{N}}$ the filtration associated with the output noise, and $(\mathcal{I}_k)_{k\in\mathbb{N}}$ the filtration associated with the stochastic gradient noise, which is the union of the two previous filtrations. Thus, we define $\mathcal{F}_0 = \{\varnothing\}$ and for $k \in \mathbb{N}^*$:*

$$\begin{aligned}\mathcal{G}_k &= \sigma\left(\mathcal{F}_{k-1} \cup \{x_k\}\right)\\ \mathcal{H}_k &= \sigma\left(\mathcal{F}_{k-1} \cup \{\varepsilon_k\}\right)\\ \mathcal{I}_k &= \sigma\left(\mathcal{F}_{k-1} \cup \{x_k, \varepsilon_k\}\right)\\ \mathcal{F}_k &= \sigma\left(\mathcal{F}_{k-1} \cup \{x_k, \varepsilon_k, u_k\}\right)\,.\end{aligned}$$

Note that there are two filtrations $\mathcal{G}$ and $\mathcal{H}$ for the two independent noises that are both involved to compute the stochastic gradient. This will help us to compute the scalar product of these two quantities.

We start by providing a bound on the distance between two compressions, this lemma will be used to prove Property D.3.

Lemma D.7. *For any compressor $\mathcal{C}$ in $\mathbb{C}$ verifying Lemma 4.1, for all x, y in $\mathbb{R}^d$, we have:*

$$\mathbb{E}[\|\mathcal{C}(x) - \mathcal{C}(y)\|^2] \leq 2(\omega+1)\|x\|^2 + 2(\omega+1)\|y\|^2 .$$

Proof Let a compressor $\mathcal{C}$ in $\mathbb{C}$ and x, y in $\mathbb{R}^d$, using Inequality 1, we have that:

$$\|\mathcal{C}(x) - \mathcal{C}(y)\|^2 \leq 2\|\mathcal{C}(x)\|^2 + 2\|\mathcal{C}(y)\|^2 .$$

Taking full expectation and using Lemma 4.1 allows to conclude:

$$\mathbb{E}\left[\|\mathcal{C}(x) - \mathcal{C}(y)\|^2\right] \leq 2(\omega+1)\|x\|^2 + 2(\omega+1)\|y\|^2 .$$

■

Now we prove that all the assumptions done in Section 4.2 are correct.

Property D.1 (Validity of the setting presented in Definition 4.1)**.** *Consider Algorithm 2 in the context of Model 2, we have that the setting presented in Definition 4.1 is verified.*

Proof From Algorithm 2, we have for any k in $\mathbb{N}^*$ and any w in $\mathbb{R}^d$ $\xi_k(w - w_*) = \nabla F(w) - \mathcal{C}_k(g_k(w))$. Because $(g_k)_{k\in\mathbb{N}^*}$ and $(\mathcal{C}_k)_{k\in\mathbb{N}^*}$ are by definition two sequences of i.i.d. random fields (Algorithm 2), it follows that their composition is also i.i.d., *hence $(\xi_k)_{k\in\mathbb{N}^*}$ is a sequence of i.i.d. random fields.*

Taking expectation w.r.t. the σ-algebra $\mathcal{I}_k$, we have $\mathbb{E}\left[\mathcal{C}_k(g_k(w)) \mid \mathcal{I}_k\right] = g_k(w)$ (Lemma 4.1), next with the σ-algebra $\mathcal{F}_{k-1}$, we have $\mathbb{E}\left[g_k(w) \mid \mathcal{F}_{k-1}\right] = \nabla F(w)$ (Equation 4.2). *Hence, the random fields are zero-centered.*

From Model 2, we have for any k in $\mathbb{N}^*$ and any w in $\mathbb{R}^d$ that:

$$\begin{aligned} F(w) &= \frac{1}{2}\mathbb{E}\left[(\langle x_k, w\rangle - y_k)^2\right] = \frac{1}{2}\mathbb{E}\left[(w - w_*)^\top (x_k \otimes x_k)(w - w_*) - 2\varepsilon_k \langle x_k, w - w_*\rangle + \varepsilon_k^2\right] \\ &= \frac{1}{2}((w - w_*)^\top H (w - w_*) + \sigma^2) , \end{aligned}$$

hence F is quadratic with Hessian equal to H whose trace is equal to R^2. ■

Property D.2 (Validity of Assumption 4.1)**.** *Considering Algorithm 2 under the setting of Model 2 with Lemma 4.1, for any iteration k in $\mathbb{N}^*$, the second moment of the additive noise ξ_k^{add} can be bounded by $(\omega+1)R^2\sigma^2$, i.e., Assumption 4.1 is verified.*

Proof Let k in $\mathbb{N}^*$. Because we consider Algorithm 2, with Definitions 4.1 and 4.2, we first have $\xi_k^{\text{add}} = -\mathcal{C}_k(g_{k,*})$, then with Lemma 4.1 we obtain $\mathbb{E}[\|\mathcal{C}_k(g_{k,*})\|^2 \mid \mathcal{I}_k] \leq (\omega+1)\|g_{k,*}\|^2$. Next, we first have from Model 2 and Equation (4.2) that $g_{k,*} = \varepsilon_k x_k$, secondly because $\left((\varepsilon_k)_{k\in\{1,\dots,K\}}\right)$ is independent from $\left((x_k)_{k\in\{1,\dots,K\}}\right)$ (Model 2), we have that $\mathbb{E}[\|\varepsilon_k x_k\|^2] \leq \sigma^2 R^2$, hence it results to:

$$\mathbb{E}[\|\xi_k^{\text{add}}\|^2 \mid \mathcal{F}_{k-1}] = \mathbb{E}[\|\xi_k^{\text{add}}\|^2] \leq (\omega+1)\sigma^2 R^2 .$$

■

Property D.3 (Validity of Assumption 4.2.1)**.** *Considering Algorithm 2, under the setting of Model 2 with Lemma 4.1, for any iteration k in $\mathbb{N}^*$, the second moment of the multiplicative noise $\xi_k^{\text{mult}}(w)$ can be bounded for any w in $\mathbb{R}^d$ by $2(\omega+1)R^2\left\|H^{1/2}(w-w_*)\right\|^2 + 4(\omega+1)\sigma^2R^2$, i.e., Assumption 4.2.1 is verified.*

Proof Let k in $\mathbb{N}^*$, we note $\eta = w - w_*$. First, because we consider Algorithm 2, with Definitions 4.1 and 4.2, we have $\xi_k(\eta) = \nabla F(w) - \mathcal{C}_k(g_k(w))$ and $\xi_k^{\text{add}} = -\mathcal{C}_k(g_{k,*})$, hence:

$$\xi_k^{\text{mult}}(\eta) = \xi_k(\eta) - \xi_k^{\text{add}} = \nabla F(w) - \mathcal{C}_k(g_k(w)) + \mathcal{C}_k(g_{k,*})\,,$$

thus developing the squared-norm of $\xi_k^{\text{mult}}(\eta)$ gives:

$$\|\xi_k^{\text{mult}}(\eta)\|^2 = \|\nabla F(w)\|^2 + 2\left\langle \nabla F(w), \mathcal{C}_k(g_{k,*}) - \mathcal{C}_k(g_k(w))\right\rangle + \|\mathcal{C}_k(g_{k,*}) - \mathcal{C}_k(g_k(w))\|^2\,.$$

On the first side we have $\mathbb{E}\left[\mathbb{E}\left[\mathcal{C}_k(g_{k,*}) - \mathcal{C}_k(g_k(w)) \mid \mathcal{I}_k\right] \mid \mathcal{F}_{k-1}\right] = -\nabla F(w_{k-1})$. On the second side, we use Lemma D.7; this allows us to write:

$$\mathbb{E}\left[\|\mathcal{C}_k(g_{k,*}) - \mathcal{C}_k(g_k(w))\|^2 \;\middle|\; \mathcal{I}_k\right] \leq 2(\omega+1)\|g_k(w)\|^2 + 2(\omega+1)\|g_{k,*}\|^2\,.$$

Note that this bound is far from being optimal when $g_k(w) = g_{k,*}$ or if $\mathcal{C}$ is the identity. Next, we decompose as follows:

$$\begin{aligned}\mathbb{E}\left[\|\mathcal{C}_k(g_{k,*}) - \mathcal{C}_k(g_k(w))\|^2 \;\middle|\; \mathcal{I}_k\right] &\leq 2(\omega+1)\|g_k(w) - g_{k,*}\|^2 \\ &\quad + 4(\omega+1)\left\langle g_k(w) - g_{k,*}, g_{k,*}\right\rangle + 4(\omega+1)\|g_{k,*}\|^2\,.\end{aligned}$$

Taking expectation w.r.t. the σ-algebra $\mathcal{G}_k$, recalling that $g_k(w) - g_{k,*}$ is $\mathcal{G}_k$-measurable (Definition D.1) and considering Model 2 allows to write:

$$\begin{aligned}\mathbb{E}\left[\|\mathcal{C}_k(g_{k,*}) - \mathcal{C}_k(g_k(w)\|^2 \;\middle|\; \mathcal{G}_k\right] &\leq 2(\omega+1)\|g_{k,*} - g_k(w)\|^2 + 4(\omega+1)\sigma^2R^2 \\ &\leq 2(\omega+1)\|(x_k\otimes x_k)\eta_{k-1}\|^2 + 4(\omega+1)\sigma^2R^2\,,\end{aligned}$$

and now taking expectation w.r.t the σ-algebra $\mathcal{F}_{k-1}$ leads to conclude the proof:

$$\mathbb{E}[\|\mathcal{C}_k(g_{k,*}) - \mathcal{C}_k(g_k(w))\|^2 \mid \mathcal{F}_{k-1}] \leq 2(\omega+1)R^2\|H^{1/2}(w_k - w_*)\|^2 + 4(\omega+1)\sigma^2R^2\,.$$

■

Property D.4 (Validity of Assumption 4.2.2)**.** *Considering Algorithm 2, under the setting of Model 2 with Lemma 4.1, for any iteration k in $\mathbb{N}^*$, the second moment of the multiplicative noise $\xi_k^{\text{mult}}(w)$ can be bounded for any w in $\mathbb{R}^d$ by $\Omega R^2\sigma\|H^{1/2}(w-w_*)\| + 3(\omega+1)R^2\|H^{1/2}(w-w_*)\|^2$, i.e. Assumption 4.2.2 is verified.*

Proof Let k in $\mathbb{N}^*$, we note $\eta = w - w_*$. Because we consider Algorithm 2, with Definitions 4.1 and 4.2, we have the following decomposition:

$$\xi_k^{\text{mult}}(\eta) = \|\nabla F(w)\|^2 + 2\left\langle \nabla F(w), \mathcal{C}_k(g_{k,*}) - \mathcal{C}_k(g_k(w))\right\rangle + \|\mathcal{C}_k(g_{k,*}) - \mathcal{C}_k(g_k(w))\|^2\,.$$

We take expectation w.r.t. the σ-algebra $\mathcal{I}_k$ and use Item L.2 of Lemma 4.1:

$$\begin{aligned}\mathbb{E}\left[\xi_k^{\text{mult}}(\eta) \;\middle|\; \mathcal{I}_k\right] &\leq \|\nabla F(w)\|^2 + 2\left\langle \nabla F(w), g_{k,*} - g_k(w)\right\rangle \\ &\quad + \Omega\min(\|g_{k,*}\|, \|g_k(w)\|)\|g_{k,*} - g_k(w)\| + 3(\omega+1)\|g_{k,*} - g_k(w)\|^2\,.\end{aligned}$$

Then, we have $\min(\|g_{k,*}\|, \|g_k(w)\|)\|g_{k,*} - g_k(w)\| \leq \|g_{k,*}\|\|g_{k,*} - g_k(w)\|$, taking expectation conditionally to the σ-algebra $\mathcal{F}_{k-1}$, applying the Cauchy-Schwarz's Equation (A.8) and considering Model 2, we have:

$$\begin{aligned}\mathbb{E}[\|g_{k,*}\|\|g_{k,*} - g_k(w)\| \mid \mathcal{F}_{k-1}]^2 &\leq \mathbb{E}[\|g_{k,*}\|^2 \mid \mathcal{F}_{k-1}]\mathbb{E}[\|g_{k,*} - g_k(w)\|^2 \mid \mathcal{F}_{k-1}] \\ &\leq \sigma^2 R^4 \|H^{1/2}(w - w_*)\|^2 .\end{aligned}$$

Therefore,

$$\mathbb{E}\left[\xi_k^{\text{mult}}(\eta) \;\middle|\; \mathcal{F}_{k-1}\right] \leq -\|\nabla F(w)\|^2 + \sigma R^2 \Omega \|H^{1/2}(w - w_*)\| + 3(\omega + 1)R^2 \|H^{1/2}(w - w_*)\|^2 ,$$

which allows concluding. ■

Property D.5 (Validity of Assumption 4.3)**.** *Considering Algorithm 2, under the setting of Model 2 with Lemma 4.1, if the compressor $\mathcal{C}$ is linear, then for any iteration k in $\mathbb{N}^*$, the multiplicative noise ξ_k^{mult} is linear, thus there exist a matrix Ξ_k in $\mathbb{R}^{d\times d}$ such that for any w in $\mathbb{R}^d$, $\xi_k^{\text{mult}}(w) = \Xi_k w$. Furthermore, the second moment of the multiplicative noise can be bounded for any w in $\mathbb{R}^d$ by $(\omega + 1)R^2 \left\|H^{1/2}(w - w_*)\right\|^2$, hence Assumption 4.3 is verified.*

Proof Let k in $\mathbb{N}^*$, we note $\eta = w - w_*$. First, because we consider Algorithm 2, with Definitions 4.1 and 4.2, we have $\xi_k(\eta) = \nabla F(w) - \mathcal{C}_k(g_k(w))$ and $\xi_k^{\text{add}} = -\mathcal{C}_k(g_{k,*})$, hence:

$$\xi_k^{\text{mult}}(\eta) = \xi_k(\eta) - \xi_k^{\text{add}} = \nabla F(w) - \mathcal{C}_k(g_k(w)) + \mathcal{C}_k(g_{k,*}) .$$

Because the random mechanism $\mathcal{C}_k$ is linear, there exists a random matrix Π_k in $\mathbb{R}^{d\times d}$ such that for any z in $\mathbb{R}^d$, we have $\mathcal{C}_k(z) = \Pi_k z$, it follows that:

$$\xi_k^{\text{mult}}(\eta) = \nabla F(w) + \mathcal{C}_k(g_{k,*} - g_k(w)) = (H - \Pi_k(x_k \otimes x_k))\eta .$$

Hence, the first part of Assumption 4.3 is verified with $\Xi_k = H - \Pi_k(x_k \otimes x_k)$. Now, we compute the second moment of the multiplicative noise. We start by developing its squared norm:

$$\|\xi_k^{\text{mult}}(\eta)\|^2 = \|\nabla F(w)\|^2 + 2\langle \nabla F(w), \mathcal{C}_k(g_{k,*} - g_k(w))\rangle + \|\mathcal{C}_k(g_{k,*} - g_k(w))\|^2 .$$

Taking expectation conditionally to the σ-algebra $\mathcal{I}_k$, and using Lemma 4.1 gives:

$$\mathbb{E}\left[\|\xi_k^{\text{mult}}(\eta))\|^2 \;\middle|\; \mathcal{I}_k\right] = \|\nabla F(w)\|^2 + 2\langle \nabla F(w), g_{k,*} - g_k(w)\rangle + (\omega + 1)\|g_{k,*} - g_k(w)\|^2 .$$

Finally, with σ-algebra $\mathcal{F}_{k-1}$ and considering Model 2 we have:

$$\mathbb{E}\left[\|\xi_k^{\text{mult}}(\eta)\|^2 \;\middle|\; \mathcal{F}_{k-1}\right] = -\|\nabla F(w)\|^2 + (\omega + 1)R^2 \|H^{1/2}(w - w_*)\|^2 ,$$

which allows to conclude. ■

Property D.6 (Validity of Assumption 4.4)**.** *Considering Algorithm 2 under the setting of Model 2 with Remark 4.1 and Lemma 4.1, if the compressor $\mathcal{C}$ is linear, then for any k in $\mathbb{N}^*$, there exists a constant $\text{Ш}_H > 0$ s.t. $\mathfrak{C}_{\text{ania}} \preccurlyeq \sigma^2 \text{Ш}_H H_F$ and $\mathbb{E}\left[\Xi_k \Xi_k^\top\right] \preccurlyeq R^2 \text{Ш}_H H$; Assumption 4.4 is thus verified.*

Proof

Let k in $\mathbb{N}^*$, we note $\eta = w - w_*$. We first need to compute $\amalg_H$ in $\mathbb{R}^d$ for each compressor $\mathcal{C}$ in $\{\mathcal{C}_{\mathrm{q}}, \mathcal{C}_{\mathrm{sq}}, \mathcal{C}_{\mathrm{rd1}}, \mathcal{C}_{\mathrm{s}}, \mathcal{C}_{\Phi}, \mathcal{C}_{\mathrm{PP}}\}$, it comes from Proposition 4.2 which results having a constant $\amalg_H$ s.t.:

$$\mathfrak{C}(\mathcal{C}, p_H) = \mathbb{E}_{E\sim p_H}[\mathcal{C}(E)^{\otimes 2}] \preccurlyeq \amalg_H H \,. \tag{D.22}$$

Indeed, $\mathrm{Diag}\,(H)$ car be bounded by $\mathrm{Tr}\,(H)\,\mathrm{I}_d$, and then I_d by $\mu^{-1}H$. This constant $\amalg_H$ can be computed from Proposition 4.2 for any compressor:

Compressor	$\mathcal{C}_{\mathrm{rd}h}$	$\mathcal{C}_{\mathrm{s}}$	$\mathcal{C}_{\mathrm{PP}}$	$\mathcal{C}_{\Phi}$
$\amalg_H$	$\frac{h-1}{p(d-1)} + (1 - \frac{h-1}{d-1})\frac{\tau}{p}$	$1 + \frac{(1-p)\tau}{p}$	$\frac{1}{p}$	$\frac{\alpha-\beta}{p} + \frac{\beta\tau}{p}$
$\amalg_H$ (if H diagonal)	$\frac{1}{p}$	$\frac{1}{p}$	$\frac{1}{p}$	$\frac{\alpha-\beta}{p} + \frac{\beta\tau}{p}$

Where $p = h/d$, $\tau = \mathrm{Tr}\,(H)\,/\mu$, and for sketching $\alpha = \frac{h+2}{d+2}$ and $\beta = \frac{d-h}{(d-1)(d+2)}$.

We now show that the two inequalities given in Assumption 4.4 are valid.

First inequality.

By Definition 4.3, we have $\mathfrak{C}_{\mathrm{ania}} = \mathbb{E}\left[\xi_k^{\mathrm{add}} \otimes \xi_k^{\mathrm{add}}\right] = \mathbb{E}\left[\mathcal{C}_k(\varepsilon_k x_k)^{\otimes 2}\right]$, because $\left((\varepsilon_k)_{k\in\{1,\dots,K\}}\right)$ is independent from $\left((x_k)_{k\in\{1,\dots,K\}}\right)$ (Model 2) and using compressor linearity and Equation (D.22), it gives: $\mathfrak{C}_{\mathrm{ania}} = \sigma^2\mathbb{E}\left[\mathcal{C}_k(x_k)^{\otimes 2}\right] = \sigma^2\mathfrak{C}(\mathcal{C}, p_H) \preccurlyeq \sigma^2\amalg_H H$.

Second inequality.

Using Property D.6, because the compressor $\mathcal{C}$ is linear, there exists two matrices Π_k, Ξ_k in $\mathbb{R}^{d\times d}$ s.t. for any z in $\mathbb{R}^d$, we have $\mathcal{C}_k(z) = \Pi_k z$ and $\xi_k^{\mathrm{mult}}(z) = \Xi_k z$, which gives that $\Xi_k = H - \Pi_k(x_k \otimes x_k)$. It follows that:

$$\Xi_k\Xi_k^\top = HH^\top - H\Pi_k(x_k\otimes x_k) - \Pi_k(x_k\otimes x_k)H + \Pi_k(x_k\otimes x_k)(x_k\otimes x_k)\Pi_k^\top \,.$$

Given that the compression is unbiased (Lemma 4.1) we have $\mathbb{E}\left[\Pi_k \mid \mathcal{I}_k\right] = \mathrm{I}_d$, hence:

$$\mathbb{E}\left[\Xi_k\Xi_k^\top \;\middle|\; \mathcal{I}_k\right] = HH^\top - H(x_k\otimes x_k) - (x_k\otimes x_k)H + \mathbb{E}\left[\Pi_k(x_k\otimes x_k)(x_k\otimes x_k)\Pi_k^\top \;\middle|\; \mathcal{I}_k\right] \,,$$

and now taking expectation w.r.t the σ-algebra $\mathcal{F}_{k-1}$:

$$\mathbb{E}\left[\Xi_k\Xi_k^\top \;\middle|\; \mathcal{F}_{k-1}\right] = -HH^\top + \mathbb{E}\left[\Pi_k(x_k\otimes x_k)(x_k\otimes x_k)\Pi_k^\top \;\middle|\; \mathcal{F}_{k-1}\right] \,.$$

In the end, we have that $\mathbb{E}\left[\Xi_k\Xi_k^\top \;\middle|\; \mathcal{F}_{k-1}\right] \preccurlyeq \mathbb{E}\left[\Pi_k(x_k\otimes x_k)(x_k\otimes x_k)\Pi_k^\top \;\middle|\; \mathcal{F}_{k-1}\right]$, and if we consider that the second moment of the features $(x_k)_{k\in\mathbb{N}^*}$ is almost surely bounded (Remark 4.1), we obtain:

$$\mathbb{E}\left[\Xi_k\Xi_k^\top \;\middle|\; \mathcal{F}_{k-1}\right] \preccurlyeq R^2\mathbb{E}\left[\Pi_k(x_k\otimes x_k)\Pi_k^\top \;\middle|\; \mathcal{F}_{k-1}\right] \preccurlyeq R^2\mathbb{E}\left[\mathcal{C}_k(x_k)^{\otimes 2} \;\middle|\; \mathcal{F}_{k-1}\right] \,. \tag{D.23}$$

Thus, using Equation (D.22), we can state that $\mathbb{E}\left[\Xi_k\Xi_k^\top \;\middle|\; \mathcal{F}_{k-1}\right] \preccurlyeq R^2\amalg_H H$, which concludes the second part of the verification of Assumption 4.4.

■

D.5 Compression operators

In this Section, we provide additional details about compression operators. First, we prove in Subsection D.5.1 that Lemma 4.1 hold and compute the compressor's covariance given in Proposition 4.2.

The specific computations for sketching are given separately in Subsection D.5.2 because they are more complex. Third, it allows to prove Propositions 4.3 and 4.4 in Subsection D.5.3. And finally, in Subsection D.5.4, we plot the covariance matrix induced by quantization and sparsification for `quantum` and `cifar-10`.

D.5.1 Computation of the variance and covariance of the compression operators

In this Subsection, we first prove Lemma 4.1. Item L.1 is frequently established in the literature and corresponds to the worst-case assumption, see the introduction for references. On the other hand, Item L.2 is the Hölder-type bound, which is not used in the literature up to our knowledge. Next, we compute the compressors' covariances that have been given in Proposition 4.2.

Lemma D.8. *For any compressor $\mathcal{C} \in \{\mathcal{C}_{\rm q}, \mathcal{C}_{\rm sq}, \mathcal{C}_{\rm rdh}, \mathcal{C}_{\rm s}, \mathcal{C}_{\Phi}, \mathcal{C}_{\rm PP}\}$, there exists constants $\omega, \Omega \in \mathbb{R}_+^*$, such that the random operator $\mathcal{C}$ satisfies the following properties for all $z, z' \in \mathbb{R}^d$.*

L.1: *$\mathbb{E}[\mathcal{C}(z)] = z$ and $\mathbb{E}[\|\mathcal{C}(z) - z\|^2] \leq \omega \|z\|^2$ (unbiasedness and variance relatively bounded),*
L.2: *$\mathbb{E}[\|\mathcal{C}(z) - \mathcal{C}(z')\|^2] \leq \Omega \min(\|z\|, \|z'\|)\|z - z'\| + 3(\omega + 1)\|z - z'\|^2$ (Hölder-type bound),*

with $\omega = \sqrt{d}$ and $\Omega = 12\sqrt{d}$ (resp. $\omega = (1-p)/p$ and $\Omega = 0$) for $\mathcal{C}_{\rm q}$ and $\mathcal{C}_{\rm sq}$ (resp. $\mathcal{C}_{\rm rdh}, \mathcal{C}_{\rm s}$, $\mathcal{C}_{\Phi}, \mathcal{C}_{\rm PP}$).

Proof

Value of ω (Item L.1 of Lemma 4.1). For projection-based compressors, the proof is straightforward, for quantization-based, the proof can be found in Alistarh et al. [2017] and gives $\omega = \sqrt{d}$.

Value of Ω (Item L.2 of Lemma 4.1). For linear compressors, it is straightforward to obtain $\Omega = 0$.

For quantization, we take x, y in $\mathbb{R}^d$, we note $(u_i)_{i=1}^d$ the vector controlling the randomness of compression, and we write $\mathcal{C}_{\rm q}(x) - \mathcal{C}_{\rm q}(y) = A + B + C$, with:

1. $A := \|x\|\text{sign}(x)\text{Bern}(\frac{|x|}{\|x\|}) - \|x\|\text{sign}(x)\text{Bern}(\frac{|x|}{\|y\|})$
2. $B := \|x\|\text{sign}(x)\text{Bern}(\frac{|x|}{\|y\|}) - \|x\|\text{sign}(y)\text{Bern}(\frac{|y|}{\|y\|})$
3. $C := \|x\|\text{sign}(y)\text{Bern}(\frac{|y|}{\|y\|}) - \|y\|\text{sign}(y)\text{Bern}(\frac{|y|}{\|y\|})$.

We note $\|\cdot\|$ the 2-norm and $\|\cdot\|_1$ the 1-norm. By symmetry, we suppose that $\|y\|^2 \geq \|x\|^2$.

First term. We have $\|A\|^2 = \|x\|^2 \sum_{i=1}^d (\mathbb{1}_{u_i \leq \frac{|x_i|}{\|x\|}} - \mathbb{1}_{u_i \leq \frac{|x_i|}{\|y\|}})^2 = \|x\|^2 \sum_{i=1}^d \mathbb{1}^2_{\frac{|x_i|}{\|y\|} \leq u_i \leq \frac{|x_i|}{\|x\|}}$ because $\|y\|^2 \geq \|x\|^2$. Taking expectation, it gives $\mathbb{E}[\|A\|^2] = \|x\|^2 \sum_{i=1}^d \frac{|x_i|}{\|x\|} - \frac{|x_i|}{\|y\|} = \|x\|^2 \|x\|_1 \frac{\|y\| - \|x\|}{\|y\|\|x\|}$. Now with triangular inequality, we have:

$$\mathbb{E}[\|A\|^2] \leq \frac{\|x\|}{\|y\|}\|x\|_1\|y - x\| \leq \|x\|_1\|y - x\| \leq \sqrt{d}\|x\|\|y - x\|\,,$$

and by symmetry $\mathbb{E}[\|A\|^2] \leq \sqrt{d}\min(\|x\|, \|y\|)\|y - x\|$.

Second term.

We have $\|B\|^2 = \|x\|^2 \sum_{i=1}^d (\text{sign}(x_i)\mathbb{1}_{u_i \leq \frac{|x_i|}{\|y\|}} - \text{sign}(y_i)\mathbb{1}_{u_i \leq \frac{|y_i|}{\|y\|}})^2$. Let i in $[d]$, if $\text{sign}(x_i) = \text{sign}(y_i)$, then:

$$\mathbb{E}\left[\|B\|^2\right] = \|x\|^2 \sum_{i=1}^d \mathbb{E}\left[\mathbb{1}^2_{\frac{\min(|x_i|,|y_i|)}{\|y\|} \leq u_i \leq \frac{\max(|x_i|,|y_i|)}{\|y\|}}\right] = \frac{\|x\|^2}{\|y\|} \sum_{i=1}^d |y_i - x_i| \leq \|x\|\|x - y\|_1\,.$$

If $\text{sign}(x_i) \neq \text{sign}(y_i)$, developping $(\text{sign}(x_i)\mathbb{1}_{u_i \leq \frac{|x_i|}{\|y\|}} - \text{sign}(y_i)\mathbb{1}_{u_i \leq \frac{|y_i|}{\|y\|}})^2$, we have:

$$\begin{aligned}\mathbb{E}\left[\|B\|^2\right] &= \|x\|^2 \sum_{i=1}^{d} \frac{|x_i|}{\|y\|} + \frac{|y_i|}{\|y\|} - 2\text{sign}(x_i)\text{sign}(y_i)\frac{\min(|x_i|,|y_i|)}{\|y\|} \\ &= \frac{\|x\|^2}{\|y\|} \sum_{i=1}^{d} \max(|x_i|,|y_i|) + 3\min(|x_i|,|y_i|)\,.\end{aligned}$$

Next, we have $\max(|x_i|,|y_i|) + \min(|x_i|,|y_i|) = |x_i| + |y_i| \overset{\text{sign}(x_i)\neq\text{sign}(y_i)}{=} |x_i - y_i|$, which results to $\mathbb{E}\left[\|B\|^2\right] \leq 3\frac{\|x\|^2}{\|y\|}\sum_{i=1}^{d}|y_i - x_i| \leq 3\|x\|\|x-y\|_1 \leq 3\sqrt{d}\|x\|\|x-y\|$.

Third term. We have $\|C\|^2 = (\|x\| - \|y\|)^2 \sum_{i=1}^{d} \mathbb{1}^2_{u_i \leq \frac{|y_i|}{\|y\|}}$, taking expectation, it gives:

$$\mathbb{E}[\|C\|^2] = (\|x\| - \|y\|)^2 \sum_{i=1}^{d} \frac{|y_i|}{\|y\|} \leq \|x-y\|^2 \frac{\|y\|_1}{\|y\|} \leq \sqrt{d}\|x-y\|^2\,.$$

Overall, using Inequality 1, we have:

$$\mathbb{E}[\|\mathcal{C}_{\text{q}}(x) - \mathcal{C}_{\text{q}}(y)\|^2] \leq 12\sqrt{d}\min(\|x\|,\|y\|)\|x-y\| + 3\sqrt{d}\|x-y\|^2\,,$$

which allows to conclude that $\Omega = 12\sqrt{d}$, and the Hölder-type bound is verified as for 1-quantization, we have $\omega = \sqrt{d}$. ■

We now compute the compressors' covariance given in Proposition 4.2 and Corollary 4.3. However, sketching requires more involved computations, they are provided in Subsection D.5.2.

Proposition D.2 (Structure of the compressor's covariance)**.** *The following formulas of compressors' covariance hold:*

- $\mathfrak{C}(\mathcal{C}_\emptyset, p_M) = M$
- $\mathfrak{C}(\mathcal{C}_q, p_M) \preccurlyeq M + \sqrt{\text{Tr}\,(M)}\sqrt{\text{Diag}\,(M)} - \text{Diag}\,(M)$
- $\mathfrak{C}(\mathcal{C}_s, p_M) = M + \frac{1-p}{p}\text{Diag}\,(M)$
- $\mathfrak{C}(\mathcal{C}_\Phi, p_M) = \frac{1}{p}\left((\alpha-\beta)M + \beta\text{Tr}\,(M)\,\text{I}_d\right)$ *with* $\alpha = \frac{h+2}{d+2}$ *and* $\beta = \frac{d-h}{(d-1)(d+2)}$
- $\mathfrak{C}(\mathcal{C}_{\text{rd}h}, p_M) = \frac{d(h-1)}{h(d-1)}M + \left(\frac{d}{h} - \frac{d(h-1)}{h(d-1)}\right)\text{Diag}\,(M)$
- $\mathfrak{C}(\mathcal{C}_{\text{PP}}, p_M) = \frac{1}{p}M\,.$

Proof

In this proof, we denote $\mathcal{F}$ the σ-field generated by the random sampling of $E \sim p_M \in \mathcal{P}_M$, and $\mathcal{G}$ the σ-field generated by the noise from the compression process. Let $E \sim p_M \in \mathcal{P}_M$.

Quantization. By definition, we have $\mathcal{C}_q(E) = \|E\|_2 \text{sign}(E) \odot \chi$, with $\chi = \left(\text{Bern}(\frac{|E_i|}{\|E\|_2})\right)_{i=1}^{d}$. It follows that $\mathcal{C}_q(E)^{\otimes 2} = \|E\|_2^2 \text{sign}(E)^{\otimes 2} \odot \chi^{\otimes 2}$.

Because:

$$\mathbb{E}\left[\chi^{\otimes 2} \mid \mathcal{F}\right] = \begin{cases} \frac{|E_i|}{\|E\|_2} & \text{if} \quad i = j \\ \frac{|E_i|\,|E_j|}{\|E\|_2^2} & \text{else,} \end{cases}$$

and considering that $\operatorname{sign}(E)^{\otimes 2} = \begin{pmatrix} 1 & & \operatorname{sign}(E_i)\operatorname{sign}(E_j) \\ & \ddots & \\ \operatorname{sign}(E_i)\operatorname{sign}(E_j) & & 1 \end{pmatrix}$, we have:

$$\mathbb{E}\left[\mathcal{C}_q(E)^{\otimes 2} \mid \mathcal{F}\right] = \begin{cases} \|E\|_2 \ |E_i| & \text{if} \quad i = j\,, \\ E_i E_j & \text{else.} \end{cases}$$

Taking the complete expectation gives:

$$\mathbb{E}\left[\mathcal{C}_q(E)^{\otimes 2}\right] = \begin{cases} \mathbb{E}\left[\|E\|_2 \ |E_i|\right] & \text{if} \quad i = j \\ M_{ij} & \text{else.} \end{cases}$$

Changing the diagonal to make appear M, we obtain:

$$\mathbb{E}\left[\mathcal{C}_q(E)^{\otimes 2}\right] = M + \mathbb{E}\left[\|E\|_2 \operatorname{Diag}\left(|E_i|\right)_{i=1}^d\right] - \mathbb{E}\left[\operatorname{Diag}\left(E_i^2\right)_{i=1}^d\right] .$$

Furthermore, we first have that $\mathbb{E}\left[\operatorname{Diag}\left(E_i^2\right)_{i=1}^d\right] = \operatorname{Diag}\left(M\right)$ and secondly, by Cauchy-Schwarz Equation (A.8) that:

$$\mathbb{E}\left[\|E\|_2 \operatorname{Diag}\left(|E_i|\right)_{i=1}^d\right]^2 \preccurlyeq \mathbb{E}\left[\|E\|_2^2\right] \mathbb{E}\left[\operatorname{Diag}\left(E_i^2\right)_{i=1}^d\right] = \operatorname{Tr}\left(M\right) \operatorname{Diag}\left(M\right) ,$$

which finally gives $\mathbb{E}\left[\mathcal{C}_q(E)^{\otimes 2}\right] \preccurlyeq M + \sqrt{\operatorname{Tr}\left(M\right)}\sqrt{\operatorname{Diag}\left(M\right)} - \operatorname{Diag}\left(M\right)$.

Sparsification. By definition, we have $\mathcal{C}_s(E) = \frac{1}{p} B \odot E \in \mathbb{R}^d$, with $B \sim (\operatorname{Bern}(p))_{i=1}^d$, thus $\mathcal{C}_s(E)^{\otimes 2} = \frac{1}{p^2} B^{\otimes 2} \odot E^{\otimes 2}$. Taking the expectation w.r.t. to the σ-filtration $\mathcal{F}$, we have $\mathbb{E}\left[\mathcal{C}_s(E)^{\otimes 2} \mid \mathcal{F}\right] = \frac{1}{p^2} P \odot E^{\otimes 2}$ with $P = \begin{pmatrix} p & & p^2 \\ & \ddots & \\ p^2 & & p \end{pmatrix}$, because for all i, j in $[\![1, d]\!]$, we have $\mathbb{E}\left[B_i^2 \mid \mathcal{F}\right] = p$ and $\mathbb{E}\left[B_i B_j \mid \mathcal{F}\right] = p^2$. This naturally gives: $\mathbb{E}\left[\mathcal{C}_s(E)^{\otimes 2}\right] = \frac{1}{p^2} P \odot M$.

Sketching. The proof is more complex and therefore is given separately, in Subsection D.5.2.3.

Rand-h. By definition, we have $\mathcal{C}_{\mathrm{rd}h}(E) := \frac{d}{h} B(S) \odot E$ with $S \sim \operatorname{Unif}(\mathcal{P}_h([d]))$ and $B(S)_i = \mathbb{1}_{i \in S}$, thus $\mathcal{C}_{\mathrm{rd}h}(E)^{\otimes 2} = \frac{1}{p^2} B^{\otimes 2} \odot E^{\otimes 2}$ $(p = h/d)$. We have that for any i, j in $\{1, \ldots, d\}$, B_i and B_j are *not* independent and that $B_i \sim (\operatorname{Bern}(p))$, therefore we have that $\mathbb{E}[B_i^2] = p$ and that: $h^2 = \left(\sum_{i=1}^d B_i\right)^2 = \sum_{i=1}^d B_i^2 + \sum_{i \neq j} B_i B_j$. Taking expectation, it gives $h^2 = h + d(d-1)\mathbb{E}[B_i B_j]$ i.e. $\mathbb{E}[B_i B_j] = \frac{h(h-1)}{d(d-1)}$. Taking the expectation w.r.t. to the σ-filtration $\mathcal{F}$, we have :

$$\mathbb{E}\left[\mathcal{C}_{\mathrm{rd}h}(E)^{\otimes 2} \mid \mathcal{F}\right] = \frac{d(h-1)}{h(d-1)} E^{\otimes 2} + \left(\frac{d}{h} - \frac{d(h-1)}{h(d-1)}\right) \operatorname{Diag}\left(E^{\otimes 2}\right) .$$

And taking full expectation allows conclusion.

Partial Participation. This result is straightforward.

■

D.5.2 Variance and covariance of sketching

In this Subsection, we compute the expectation, the variance, and the covariance of sketching. In Subsection D.5.2.1, we give the proof principle of our computation, in Subsection D.5.2.2, we compute the expectation and the variance, and in Subsection D.5.2.3, we compute the covariance.

We thank Baptiste Goujaud (École polytechnique, CMAP) who greatly helped to prove the following.

D.5.2.1 Proof principle

Let y in $\mathbb{R}^d$ with $\|y\|^2 = 1$, and x in $\mathbb{R}^d$. By Definition 4.4, for Φ in $\mathbb{R}^{h\times d}$, we have $\mathcal{C}_\Phi(x) = \frac{1}{p}\Phi^\dagger\Phi x$ with $\Phi^\dagger = \Phi^\top(\Phi\Phi^T)^{-1}$ and $p = h/d$.

To compute the expectation, the variance, and the covariance of $\mathcal{C}_\Phi(x)$, the idea is to compute $\mathbb{E}[y^\top C_\Phi(x)]$ and $\mathbb{E}[(y^\top C_\Phi(x))^2]$ by establishing Equation (D.24) which allows controlling the randomness of sketching by using Equation (D.25). To establish Equation (D.24), first observe that $pC_\Phi(\cdots)$ is a projector into a subspace of dimension h, indeed we have $(pC_\Phi \odot pC_\Phi)(x) = pC_\Phi(x)$. Then there exists a random matrix P in $\mathcal{O}_d$ s.t. $pC_\Phi(x) = P^\top J_h P x$. It leads to:

$$y^\top C_\Phi(x) = \frac{1}{p} y^\top P^\top J_h P x = \frac{1}{p}(Py)^\top J_h (Px)\,.$$

Now we note $X = Px/\|x\|$ and $Y = Py$, hence $y^\top C_\Phi(x) = \frac{\|x\|}{p} Y^\top J_h X$, and because P is in $\mathcal{O}_d$, we have:

$$\begin{cases} \|X\|^2 = 1 \\ \|Y\|^2 = \|y\|^2 = 1 \\ \langle X, Y\rangle = \langle x, y\rangle / \|x\|\,. \end{cases}$$

Furthermore, P is a random projector, it follows that X and Y are sampled uniformly from the zero-center sphere of radius 1; i.e. $X \sim \mathrm{Unif}(\mathcal{S}_d(0,1))$ and $Y \sim \mathrm{Unif}(\mathcal{S}_d(0,1))$. However, X and Y are not independent, this is why, we consider that $X \sim \mathrm{Unif}(\mathcal{S}_d(0,1))$ and write Y s.t. $Y = aX + bu$ with u a random vector in $\mathbb{R}^d$ of norm 1 orthogonal to X, that is to say, $u|X$ is uniformly sampled on a zero-centered hyper-sphere of radius 1 orthogonal to the vector X (see illustration on Figure D.3). It comes that:

$$y^\top C_\Phi(x) = \frac{\|x\|}{p} Y^\top J_h X = \frac{\|x\|}{p}(aX^\top + bu^T) J_h X = \frac{\|x\|}{p}(aX^\top J_h X + bu^\top J_h X)\,. \tag{D.24}$$

Observe that for any i, j in $\{1, \cdots, d\}$, X_i, X_j (resp. u_i, u_j) have the same law, it results to:

$$\forall (i,j) \in \{1,\cdots,d\}^2,\ \forall k \in \mathbb{N}, \qquad \mathbb{E}[X_i^k] = \mathbb{E}[X_j^k] \quad \text{and} \quad \mathbb{E}[u_i^k] = \mathbb{E}[u_j^k]\,. \tag{D.25}$$

This property is the key to compute the expectation, the variance, and the covariance of sketching.

We now compute a and b. First, by definition, we have:

$$\frac{\langle x, y\rangle}{\|x\|} = \langle X, Y\rangle = a\,\|X\|^2 = a\,,$$

then we write that:

$$1 = \|Y\|^2 = \frac{\langle x, y\rangle^2}{\|x\|^4}\|X\|^2 + b^2\|u\|^2 = \frac{\langle x, y\rangle^2}{\|x\|^2} + b^2\,,$$

which gives $b = \sqrt{1 - \frac{\langle x,y\rangle^2}{\|x\|^2}}$.

At the end, we have: $Y = aX + bu = \frac{\langle x,y\rangle}{\|x\|}X + \sqrt{1 - \frac{\langle x,y\rangle^2}{\|x\|^2}}u$.

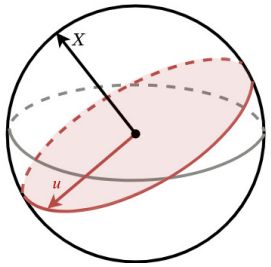


Figure D.3: Sphere zero-center with radius 1: X and u are orthogonal.

D.5.2.2 Expectation and variance of sketching

In this Subsection, we prove that sketching verifies Item L.1 in Lemma 4.1; for this purpose, we show that it is unbiased, then we compute its variance.

Proposition D.3. *Sketching is unbiased and its variance is relatively bounded, i.e., it verifies Item L.1 in Lemma 4.1 with* $\omega = (1-p)/p$ *where* $p = h/d$.

Proof Starting from Equation (D.24), we have $y^\top C_\Phi(x) = \frac{\|x\|}{p}(aX^\top J_h X + bu^\top J_h X)$. We first compute the expectation w.r.t. the σ-algebra $\sigma(\{X\})$ generated by the noise involved in the random vector X, it gives:

$$\mathbb{E}[y^\top C_\Phi(x) \mid \sigma(\{X\})] = \frac{\|x\|}{p}\sum_{i=1}^{h} aX_i^2 + bX_i\mathbb{E}\left[u_i \mid \sigma(\{X\})\right] .$$

Because u is sampled uniformly from the zero-center sphere of radius 1 s.t. it is orthogonal to X, for any i in $\{1, \cdots, d\}$, we have $\mathbb{E}[u_i \mid \sigma(\{X\})] = 0$, hence taking full expectation, we obtain:

$$\mathbb{E}[y^\top C_\Phi(x)] = \frac{\|x\|}{p}\sum_{i=1}^{h} a\mathbb{E}[X_i^2] .$$

Using Equation (D.25), we have $\mathbb{E}[X_i^2] = \frac{1}{d}\sum_{j=1}^{d}\mathbb{E}[X_j^2]$, next recalling that $p = h/d$ and $\|X\|^2 = 1$, it leads to $\mathbb{E}[y^\top C_\Phi(x)] = a\|x\|\mathbb{E}[\sum_{j=1}^{d} X_j^2] = a\|x\|\mathbb{E}[\|X\|^2] = a\|x\|$. And because $a = \langle x, y\rangle / \|x\|$, we have at the end that $\mathbb{E}[C_\Phi(x)] = x$. Now we compute the variance:

$$\mathbb{E}[C_\Phi(x)^\top C_\Phi(x)] = \frac{1}{p^2}\mathbb{E}[x^\top P^\top J_h P P^\top J_h P x] = \frac{1}{p^2}\mathbb{E}[x^\top P^\top J_h P x] = \frac{\|x\|^2}{p^2}\mathbb{E}[X^\top J_h X] .$$

$\mathbb{E}[X^\top J_h X]$ has been computed above and is equal to p, it results that $\mathbb{E}[C_\Phi(x)^\top C_\Phi(x)] = \|x\|^2/p$. In the end, sketching verifies Lemma 4.1 with $\omega = (1-p)/p$. ■

D.5.2.3 Covariance of sketching.

In this Subsection, we compute the covariance of sketching. For the sake of demonstration, we need to compute the 4^{th}-moment of X_1 and the 2^{nd}-moment of u_1. For any i in $[d]$ and any vector v in $\mathbb{R}^d$, we note $v_{-i} = (v_j)_{j\in[d], j\neq i}$ in $\mathbb{R}^{d-1}$.

Computing the 4^{th}*-moment of* X_1.

The marginal density of X_1 is $f_{X_1} : x \mapsto B(\frac{d-1}{2}, \frac{1}{2})^{-1}(1-x^2)^{(d-3)/2}$ where B is the beta function defined as $B : x, y \mapsto \int_0^1 t^{x-1}(1-t)^{y-1} = 2\int_0^{\pi/2}\sin^{2x-1}(t)\cos^{2y-1}(t)\mathrm{d}t$. This result can be obtained either by an application of the formula for the surface area of a sphere [Li, 2010, Sidiropoulos, 2014], either by writing that $X_1 = \frac{Z_1}{\|Z\|}$ with Z a Gaussian vector with d components. Therefore we have that:

$$\mathbb{E}[X_1^4] = \frac{\int_{-1}^{1} x^4(1-x^2)^{(d-3)/2}\mathrm{d}x}{2\int_0^{\pi/2}\sin^{d-2}(t)\mathrm{d}t} \overset{\text{(i)}}{=} \frac{2\int_0^{\pi/2}\cos^4(t)\sin^{d-2}(t)\mathrm{d}t}{2\int_0^{\pi/2}\sin^{d-2}(t)\mathrm{d}t} \overset{\text{(ii)}}{=} \frac{W_{d-2} - 2W_d + W_{d+2}}{W_{d-2}} ,$$

where at (i) we set $x = \cos(t)$ and at (ii) we make appears the Wallis' integrals defined for any n in $\mathbb{N}$ as $W_n = \int_0^{\pi/2}\sin^n(t)\mathrm{d}t$. Furthermore, we have the following recursion using integration by parts:

$W_{d+2} = \frac{d+1}{d+2} W_d$, therefore, we have:

$$\mathbb{E}[X_1^4] = \left(1 - \frac{2(d-1)}{d} + \frac{(d-1)(d+1)}{d(d+2)}\right) = \frac{3}{d(d+2)}\,. \tag{D.26}$$

Computing the 2^{nd}*-moment of* u_1 *w.r.t the* σ*-algebra* $\sigma(X)$*.*

We define three $(d-2)$–dimensional manifolds, two parallel hyperplanes P, P' and a sphere S, as follows:

$$\begin{cases} P = \{\tilde{u} \in \mathbb{R}^{d-1} \mid \langle \tilde{u}, X_{-i} \rangle = -X_i u_i\} \\ P' = \{\tilde{u} \in \mathbb{R}^{d-1} \mid \langle \tilde{u}, X_{-i} \rangle = 0\} \\ S = S_{d-1}(0, \sqrt{1-u_1^2}) \end{cases}$$

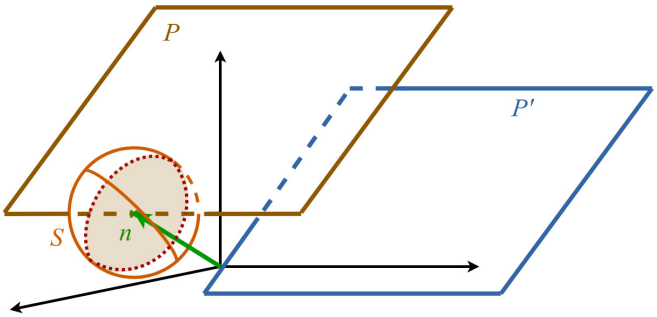


Figure D.4: Parallel hyperplanes P and P' with the sphera S.

Obviously u_{-i} is in $P \cap S$; then we decompose u_{-i} in two terms $n + v$, with $v \sim \text{Unif}(P')$ orthogonal to X and independent of u_i: n is the center of the sphere $S \cap P$ and v is its radius, n corresponds also to the normal vector of both P, P' with norm equal to the distance between the two hyperplanes, hence $n = \frac{\langle u_{-i}, X_{-i} \rangle}{\|X_{-i}\|^2} X_{-i} = -\frac{u_i X_i}{\|X_{-i}\|^2} X_{-i}$.

First, because $u_{-1} \in S$, we have $\|n+v\|^2 = 1 - u_1^2$, next by Pythagorean theorem this is equivalent to $\|v\|^2 = 1 - u_1^2 - \|n\|^2 = 1 - \frac{u_1^2}{\|X_{-1}\|^2}$. Second, because $u_{-1} \in P$, we have $u_1 = \frac{-\langle u_{-1}, X_{-1} \rangle}{X_1}$, that is to say the probability density function of $u_1 \mid X$ is proportional to the number of possible values for u_{-1}, which corresponds to the surface area of the hypersphere $P \cap S$. This surface is proportional to the radius $\|v\|^{d-4} = (1 - \frac{u_1^2}{\|X_{-1}\|^2})^{(d-4)/2}$ given that $P \cap S$ is a $(d-3)$–dimensional manifold, therefore:

$$\begin{aligned} \mathbb{E}[u_1^2 \mid \sigma(\{X\})] = \frac{\int_{-\|X_{-1}\|}^{\|X_{-1}\|} x^2 \left(1 - \frac{x^2}{\|X_{-i}\|^2}\right)^{(d-4)/2} \mathrm{d}x}{\int_{-\|X_{-1}\|}^{\|X_{-1}\|} \left(1 - \frac{x^2}{\|X_{-i}\|^2}\right)^{(d-4)/2} \mathrm{d}x} &\overset{\text{(i)}}{=} \frac{\|X_{-1}\|^2 \int_{-1}^{1} y^2 \left(1-y^2\right)^{(d-4)/2} \mathrm{d}y}{\int_{-1}^{1} \left(1-y^2\right)^{(d-4)/2} \mathrm{d}y} \\ &\overset{\text{(ii)}}{=} \|X_{-1}\|^2 \frac{W_{d-3} - W_{d-1}}{W_{d-3}}\,, \end{aligned}$$

where at (i) we set $y = \frac{x}{\|X_{-1}\|}$ and at (ii) we reuse the previous computations to make appear the Wallis' integral. In the end, we obtain:

$$\mathbb{E}[u_1^2 \mid \sigma(\{X\})] = (1 - \frac{d-2}{d-1})\|X_{-1}\|^2 = \frac{\|X_{-1}\|^2}{d-1}\,. \tag{D.27}$$

Note that this result is consistent with the fact that $\sum_{i=1}^d \mathbb{E}[u_i^2 \mid \sigma(\{X\})] = \frac{d - \sum_{i=1}^d X_i^2}{d-1} = 1$. Now we can compute the covariance of the sketching operator.

Proposition D.4. *Let* x *in* p_M*, the covariance of sketching is equal to:*

$$\mathbb{E}[\mathcal{C}_\Phi(x)^{\otimes 2}] = \frac{1}{p}((\alpha - \beta)M + \beta \mathrm{Tr}\,(M)\,\mathrm{I}_d)\,,$$

with $\alpha = \frac{h+2}{d+2}$ *and* $\beta = \frac{d-h}{(d-1)(d+2)}$*.*

Proof

Let x in $\mathbb{R}^d$ and y in $\mathbb{R}^d$ with $\|y\|^2 = 1$, starting from Equation (D.24), we have:

$$(y^\top C_\Phi(x))^2 = \frac{\|x\|^2}{p^2}(aX^\top J_h X + bu^\top J_h X)^2$$
$$= \frac{\|x\|^2}{p^2}\left(a^2(X^\top J_h X)^2 + 2ab(X^\top J_h X u^\top J_h X) + b^2(u^\top J_h X)^2\right).$$

First term. Taking expectation, we have $\mathbb{E}[(X^\top J_h X)^2] = \sum_{i=1}^h \left(\mathbb{E}[X_i^4] + \sum_{j=1,j\neq i}^h \mathbb{E}[X_i^2 X_j^2]\right)$. However:

$$\sum_{j=1,j\neq i}^h \mathbb{E}[X_i^2 X_j^2] = \mathbb{E}\left[X_i^2 \sum_{j=1,j\neq i}^h X_j^2\right] \overset{\text{(i)}}{=} \mathbb{E}\left[X_i^2 \sum_{j=1,j\neq i}^h \frac{1}{d-1}\sum_{k=1,k\neq i}^d X_k^2\right]$$
$$\overset{\text{(ii)}}{=} \frac{h-1}{d-1}\mathbb{E}\left[X_i^2(1-X_i^2)\right],$$

where we use at line (i) Equation (D.25) and at line (ii) $\sum_{i=1}^d X_i^2 = 1$. It follows that:

$$\mathbb{E}[(X^\top J_h X)^2] = \sum_{i=1}^h \left(\frac{d-h}{d-1}\mathbb{E}[X_i^4] + \frac{h-1}{d-1}\mathbb{E}\left[X_i^2\right]\right)$$
$$\overset{\text{(i)}}{=} \frac{h(d-h)}{d-1}\mathbb{E}[X_1^4] + \frac{h-1}{d-1}\sum_{i=1}^h \mathbb{E}\left[X_i^2\right]$$
$$\overset{\text{(iii)}}{=} \frac{h(d-h)}{d-1}\mathbb{E}[X_1^4] + \frac{h(h-1)}{d(d-1)}$$
$$\overset{\text{eq. D.26}}{=} \frac{3h(d-h)}{d(d-1)(d+2)} + \frac{h(h-1)}{d(d-1)} = \frac{h(h+2)}{d(d+2)} := \alpha'.$$

Where we considered at line (i) that for any i in $\{1,\cdots,h\}$, $\mathbb{E}[X_i^4] = \mathbb{E}[X_1^4]$, and at line (ii) that $\sum_{i=1}^h \mathbb{E}\left[X_i^2\right] = \frac{h}{d}\mathbb{E}[\|X\|^2] = h/d$.

Second term. We compute the expectation w.r.t. the σ-algebra $\sigma(\{X\})$ generated by the noise involved in the random vector X. It gives $\mathbb{E}\left[X^\top J_h X u^\top J_h X \mid \sigma(\{X\})\right] = 0$, because $u|X$ is uniformly sampled on a zero-centered hyper-sphere, and thus for any i in $\{1,\cdots,d\}$, we have $\mathbb{E}[u_i \mid \sigma(\{X\})] = 0$.

Third term. We have $(u^\top J_h X)^2 = \sum_{i=1}^h u_i^2 X_i^2 + \sum_{j=1,j\neq i}^h u_i u_j X_i X_j$. On one side, we compute the expectation w.r.t. the σ-algebra $\sigma(\{X\})$ generated by the noise involved in the random vector X:

$$\sum_{i=1}^h \mathbb{E}\left[u_i^2 X_i^2 \mid \sigma(\{X\})\right] = \sum_{i=1}^h X_i^2 \mathbb{E}\left[u_i^2 \mid \sigma(\{X\})\right] \overset{\text{eq. D.27}}{=} \frac{1}{d-1}\sum_{i=1}^h X_i^2 \|X_{-i}\|^2.$$

Taking full expectation, we have $\sum_{i=1}^h \mathbb{E}[u_i^2 X_i^2] = \frac{1}{d-1}\sum_{i=1}^h \mathbb{E}[X_i^2(1-X_i^2)] = \frac{h}{d-1}(\frac{1}{d} - \mathbb{E}[X_1^4])$, because for any i in $\{1,\ldots,h\}$, $\mathbb{E}[X_i^4] = \mathbb{E}[X_1^4]$ and $\sum_{i=1}^h \mathbb{E}\left[X_i^2\right] = \frac{h}{d}\mathbb{E}[\|X\|^2] = h/d$.

Let i in $[d]$, on the other side, we compute the expectation w.r.t. the σ-algebra $\sigma(\{X, u_i\})$ generated by the noise involved in the random vector X and the random variable u_i, hence we requires to compute $\mathbb{E}\left[u_j \mid \sigma(\{X, u_i\})\right]$. To do so, as before, we decompose u_{-i} in two terms $n + v$ (see Figure D.4), with $v \sim \text{Unif}(P')$ orthogonal to X and independent of u_i, hence $\mathbb{E}\left[v \mid \sigma(\{X, u_i\})\right] = 0$. It gives that $\mathbb{E}\left[u_{-i} \mid \sigma(\{X, u_i\})\right] = -\frac{u_i X_i}{\|X_{-i}\|^2} X_{-i}$. Thereby, replacing for any coordinate $j \neq i$ in $[d]$

the value of u_{-i} and taking expectation w.r.t. the σ-algebra $\sigma(\{X\})$, we obtain:

$$\sum_{i=1}^{h}\sum_{j=1,j\neq i}^{h} X_iX_j\mathbb{E}\left[u_iu_j \mid \sigma(\{X\})\right] = -\sum_{i=1}^{h}\sum_{j=1,j\neq i}^{h}\frac{1}{\|X_{-i}\|^2}X_i^2X_j^2\mathbb{E}\left[u_i^2 \mid \sigma(\{X\})\right]$$
$$\overset{\text{eq. D.27}}{=} -\frac{1}{d-1}\sum_{i=1}^{h}\sum_{j=1,j\neq i}^{h} X_i^2X_j^2$$
$$= -\frac{1}{d-1}\sum_{i=1}^{h}\sum_{j=1,j\neq i}^{h} X_i^2\frac{1-X_i^2}{d-1}\,.$$

Finally, we have: $\sum_{i=1}^{h}\sum_{j=1,j\neq i}^{h}\mathbb{E}[X_iX_ju_iu_j] = -\frac{h(h-1)}{d(d-1)^2}(1-\sum_{i=1}^{d}\mathbb{E}[X_i^4])$. Putting together the two terms, we have that:

$$\mathbb{E}\left[(u^\top J_hX)^2\right] = \frac{h}{d-1}(\frac{1}{d}-\mathbb{E}[X_i^4]) - \frac{h(h-1)}{d(d-1)^2}(1-d\mathbb{E}[X_1^4]) \overset{\text{eq. D.26}}{=} \frac{h(d-h)}{d(d-1)(d+2)} := \beta'\,.$$

In the end, we have $\mathbb{E}[(y^\top C_\Phi(x))^2] = \frac{\|x\|^2}{p^2}(a^2\alpha' + b^2\beta')$. And because $\|y\|^2 = 1$, $a = \langle x,y\rangle/\|x\|$ and $b = \sqrt{1-\langle x,y\rangle^2/\|x\|^2}$, replacing them by their values gives:

$$y^\top\mathbb{E}[C_\Phi(x))^{\otimes 2}]y = \frac{\|x\|^2}{p^2}\left(\alpha'\frac{\langle x,y\rangle^2}{\|x\|^2} + \beta'\left(y^\top y - \frac{\langle x,y\rangle^2}{\|x\|^2}\right)\right)\,,$$

hence $\mathbb{E}[C_\Phi(x))^{\otimes 2}] = \frac{1}{p^2}\left((\alpha'-\beta')xx^\top + \beta'\|x\|^2\,\mathrm{I}_d\right)$. To conclude, we consider that x is a random variable sampled from a distribution p_M, then taking expectation on this random variable we have: $\mathbb{E}C_\Phi(x)^{\otimes 2} = \frac{1}{p}\left((\alpha-\beta)M + \beta\mathrm{Tr}\,(M)\,\mathrm{I}_d\right)$, with $\alpha = \frac{\alpha'}{p} = \frac{h+2}{d+2}$ and $\beta = \frac{\beta'}{p} = \frac{d-h}{(d-1)(d+2)}$. ■

D.5.3 Proof of Propositions 4.3 and 4.4

In this Subsection, we give the proof of Propositions 4.3 and 4.4 which provides generic comparisons between the asymptotic convergence rate of compressors. We first give a lemma resulting from the Cauchy-Schwarz's inequality necessary to establish these proofs.

Lemma D.9 (Cauchy-Schwarz's inequality on matrices' traces)**.** *For any matrix M in $\mathbb{R}^{d\times d}$, we have $\mathrm{Tr}\,(M)\,\mathrm{Tr}\left(M^{-1}\right) \geq d^2$, with strict inequalities if M is not proportional to I_d. And if M is with constant diagonal equal to c in $\mathbb{R}$, we have $c\mathrm{Tr}\left(M^{-1}\right) \geq d$.*

Proof Let M in $\mathbb{R}^{d\times d}$, using the Cauchy-Schwarz inequality, we have:

$$d^2 = \mathrm{Tr}\,(\mathrm{I}_d)^2 = \mathrm{Tr}\left(M^{1/2}M^{-1/2}\right)^2 \overset{\text{C.S}}{\leq} \mathrm{Tr}\,(M)\,\mathrm{Tr}\left(M^{-1}\right)\,,$$

and we have equality if M is proportional to I_d. ■

Now we give the demonstration of Propositions 4.3 and 4.4. On Figure D.5, we complete the numerical illustration provided in Subsection 4.3.3.1 by illustrating the scenario of standardized features, i.e., when the diagonal of M is the identity.

Proposition D.5 (Comparison between $\mathcal{C}_{\mathrm{PP}}, \mathcal{C}_{\mathrm{s}}, \mathcal{C}_{\mathrm{rdh}}, \mathcal{C}_\Phi$, $\omega = d/h-1$)**.** *We consider $\mathcal{C} \in \{\mathcal{C}_{\mathrm{PP}}, \mathcal{C}_{\mathrm{s}}, \mathcal{C}_{\mathrm{rdh}}, \mathcal{C}_\Phi\}$ with $p = h/d$, such that $\mathcal{C}$ always satisfies Lemma 4.1 with $\omega = d/h-1$. For any matrix $M \in \mathbb{R}^{d\times d}$:*

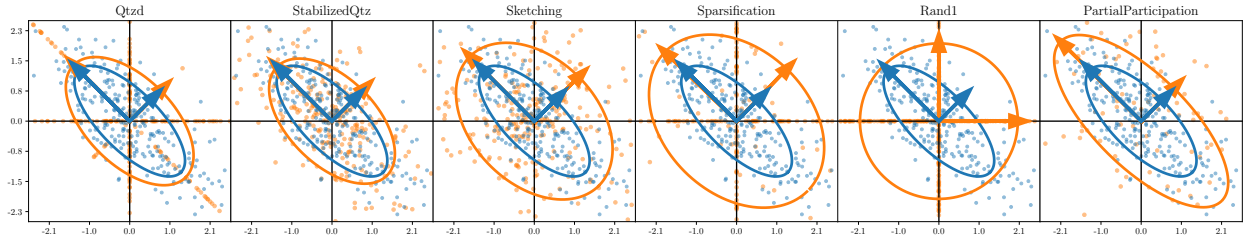


Figure D.5: H not diagonal, scenario using features standardization. Scatter plot of $(x_k)_{i=1}^K$/ $(\mathcal{C}(x_k))_{i=1}^K$ with its ellipse $\mathcal{E}_{\mathrm{Cov}[x_k]}$/$\mathcal{E}_{\mathrm{Cov}[\mathcal{C}(x_k)]}$.

1. *If M is diagonal, then:*
   - $\mathfrak{C}(\mathcal{C}_{\mathrm{PP}}, p_M) = \mathfrak{C}(\mathcal{C}_{\mathrm{s}}, p_M) = \mathfrak{C}(\mathcal{C}_{\mathrm{rd}h}, p_M) = \frac{d}{h} M$,
   - $\mathrm{Tr}\left(\mathfrak{C}(\mathcal{C}_{\mathrm{PP/s/rd}h}, p_M) M^{-1}\right) \leq \mathrm{Tr}\left(\mathfrak{C}(\mathcal{C}_{\Phi}, p_M) M^{-1}\right)$.
2. *Moreover, for any matrix M with a* constant diagonal *(e.g., after standardization), we have:*

$$\mathrm{Tr}(\mathfrak{C}(\mathcal{C}_{\mathrm{PP}}, p_M) M^{-1}) \leq \mathrm{Tr}(\mathfrak{C}(\mathcal{C}_{\Phi}, p_M) M^{-1}) \leq \mathrm{Tr}(\mathfrak{C}(\mathcal{C}_{s}, p_M) M^{-1}) \leq \mathrm{Tr}(\mathfrak{C}(\mathcal{C}_{\mathrm{rd}h}, p_M) M^{-1}) .$$

With strict inequalities if M is not proportional to I_d.

Proof

Let M in $\mathbb{R}^{d\times d}$ and take $p = h/d$.

Proof of Item 1 in Proposition 4.3. In the diagonal case, the first equalities are straightforward as we have $\mathfrak{C}(\mathcal{C}_{\mathrm{PP}}, p_M) = \mathfrak{C}(\mathcal{C}_{\mathrm{s}}, p_M) = \mathfrak{C}(\mathcal{C}_{\mathrm{rd}1}, p_M) = \frac{d}{h} M$. Next, we have (regardless if M is diagonal or not):

$$\begin{aligned}
\mathrm{Tr}\left(\left(\mathfrak{C}(\mathcal{C}_{\Phi}, p_M) - \mathfrak{C}(\mathcal{C}_{\mathrm{PP}}, p_M)\right) M^{-1}\right) &= \left(\frac{h+1}{d+2} + \delta_{hd} - 1\right) \frac{\mathrm{Tr}\left(\mathrm{I}_d\right)}{p} + \left(1 - \frac{h-1}{d-1}\right) \frac{\mathrm{Tr}\left(M\right) \mathrm{Tr}\left(M^{-1}\right)}{p(d+2)} \\
&\overset{\text{Lemma D.9}}{\geq} \frac{d}{p}\left(\frac{h+1}{d+2} + \delta_{hd} - 1 + \frac{d}{d+2}\left(1 - \frac{h-1}{d-1}\right)\right) \\
&= 0 .
\end{aligned}$$

Proof of Item 2 in Proposition 4.3. Suppose now that $\mathrm{Diag}\left(M\right) = c\mathrm{I}_d$, then we have $\mathfrak{C}(\mathcal{C}_{\mathrm{PP}}, p_M) = \frac{d}{h} M$, $\mathfrak{C}(\mathcal{C}_{\mathrm{s}}, p_M) = M + (\frac{d}{h} - 1) c\mathrm{I}_d$, $\mathfrak{C}(\mathcal{C}_{\mathrm{rd}h}, p_M) = \frac{d(h-1)}{h(d-1)} M + \frac{d}{h}(1 - \frac{h-1}{d-1}) c\mathrm{I}_d$ and $\mathfrak{C}(\mathcal{C}_{\Phi}, p_M) = \frac{d}{h}\left((\frac{h+1}{d+2} - \delta_{hd}) M + \left(1 - \frac{h-1}{d-1}\right) \frac{\mathrm{Tr}(M)}{d+2} \mathrm{I}_d\right)$. Firstly, from previous item, we have

$$\mathrm{Tr}\left(\mathfrak{C}(\mathcal{C}_{\mathrm{PP}}, p_M) M^{-1}\right) \leq \mathrm{Tr}\left(\left(\mathfrak{C}(\mathcal{C}_{\Phi}, p_M)\right) M^{-1} .\right.$$

Secondly, we write:

$$\begin{aligned}
\mathrm{Tr}\left(\left(\mathfrak{C}(\mathcal{C}_{\Phi}, p_M) - \mathfrak{C}(\mathcal{C}_{\mathrm{s}}, p_M)\right) M^{-1}\right) &= \frac{d}{p}\left(\frac{h+1}{d+2} + \delta_{hd} - \frac{h}{d}\right) \\
&\quad + \frac{c\mathrm{Tr}\left(M^{-1}\right)}{p}\left(\frac{d}{d+2}\left(1 - \frac{h-1}{d-1}\right) - \left(1 - \frac{h}{d}\right)\right) \\
&= \frac{d}{p}\left(\frac{h+1}{d+2} + \delta_{hd} - \frac{h}{d}\right) - \frac{c\mathrm{Tr}\left(M^{-1}\right)}{p} \cdot \frac{(d-2)(d-h)}{d(d-1)(d+2)} \\
&\overset{\text{Lemma D.9}}{\leq} \frac{d}{p}\left(\frac{h+1}{d+2} + \delta_{hd} - \frac{h}{d} - \frac{(d-2)(d-h)}{d(d-1)(d+2)}\right) = 0 .
\end{aligned}$$

Thirdly, we have:

$$\begin{aligned}\mathrm{Tr}\left(\left(\mathfrak{C}(\mathcal{C}_{\mathrm{rd}h}, p_M) - \mathfrak{C}(\mathcal{C}_{\mathrm{s}}, p_M)\right) M^{-1}\right) &= \frac{h-d}{h(d-1)}\mathrm{Tr}\left(\mathrm{I}_d\right) + \frac{d-h}{h(d-1)} c\mathrm{Tr}\left(M^{-1}\right) \\ &\overset{\text{Lemma D.9}}{\geq} \frac{d}{h}\left(\frac{h-d}{d-1} + \frac{d-h}{d-1}\right) = 0\,.\end{aligned}$$

■

Proposition D.6 (Comparison between $\mathcal{C}_{\mathrm{PP}}, \mathcal{C}_{\mathrm{q}}, \mathcal{C}_{\mathrm{s}}$, $\omega = \sqrt{d}$)**.** *We consider $\mathcal{C} \in \{\mathcal{C}_{\mathrm{PP}}, \mathcal{C}_{\mathrm{q}}, \mathcal{C}_{\mathrm{s}}\}$ with $p = (\sqrt{d}+1)^{-1}$, such that $\mathcal{C}$ always satisfies Lemma 4.1 with $\omega = \sqrt{d}$.*

1. *For any symmetric matrix M diagonal, we have:*

$$\mathrm{Tr}\left(\mathfrak{C}(\mathcal{C}_{\mathrm{PP}}, p_M) M^{-1}\right) = \mathrm{Tr}\left(\mathfrak{C}(\mathcal{C}_s, p_M) M^{-1}\right) \overset{possib.}{\leq\!\ll} \left(1 + \frac{1}{\sqrt{d}}\right) \mathrm{Tr}\left(\widetilde{\mathfrak{C}}(\mathcal{C}_{\mathrm{q}}, M) M^{-1}\right)\,.$$

2. *If M is not necessarily diagonal but with a* constant diagonal *(e.g., after standardization), then*
   - $\widetilde{\mathfrak{C}}(\mathcal{C}_{\mathrm{q}}, M) \preccurlyeq \mathfrak{C}(\mathcal{C}_{\mathrm{s}}, p_M)$
   - $\mathrm{Tr}\left(\mathfrak{C}(\mathcal{C}_{\mathrm{PP}}, p_M) M^{-1}\right) \leq \left(1 + \frac{1}{\sqrt{d}}\right) \mathrm{Tr}\left(\widetilde{\mathfrak{C}}(\mathcal{C}_{\mathrm{q}}, M) M^{-1}\right)$.

Proof

Let M in $\mathbb{R}^{d\times d}$ and take $p = \frac{1}{1+\sqrt{d}}$.

Proof of Item 1 in Proposition 4.4. In the diagonal case with $p = \frac{1}{1+\sqrt{d}}$, we have $\widetilde{\mathfrak{C}}(\mathcal{C}_{\mathrm{q}}, M) = \sqrt{\mathrm{Tr}\left(M\right)}\sqrt{M}$ and $\mathfrak{C}(\mathcal{C}_{\mathrm{PP}}, p_M) = (1+\sqrt{d})M$, hence $\mathrm{Tr}\left(\widetilde{\mathfrak{C}}(\mathcal{C}_{\mathrm{q}}, M) M^{-1}\right) = \sqrt{\mathrm{Tr}\left(M\right)}\mathrm{Tr}\left(\sqrt{M^{-1}}\right)$ and $\mathrm{Tr}\left(\mathfrak{C}(\mathcal{C}_{\mathrm{PP}}, p_M) M^{-1}\right) = (1+\sqrt{d})d$. Noting $(\lambda_i)_{i\in[d]}$ the eigenvalues of M, and using the Cauchy-Schwarz inequality's, we have:

$$\begin{aligned}d^2 = \left(\sum_{i=1}^d 1\right)^2 = \left(\sum_{i=1}^d \lambda_i^{1/4}\lambda_i^{-1/4}\right)^2 &\overset{\text{C.S}}{\leq} \left(\sum_{i=1}^d \lambda_i^{1/2}\right)\left(\sum_{i=1}^d \lambda_i^{-1/2}\right) \\ &\overset{\text{C.S}}{\leq} \sqrt{\sum_{i=1}^d \lambda_i}\sqrt{\sum_{i=1}^d 1}\left(\sum_{i=1}^d \lambda_i^{-1/2}\right) = \sqrt{d\mathrm{Tr}\left(M\right)}\mathrm{Tr}\left(M^{-1/2}\right) = \sqrt{d}\mathrm{Tr}\left(\widetilde{\mathfrak{C}}(\mathcal{C}_{\mathrm{q}}, M) M^{-1}\right)\,.\end{aligned}$$

Which follows that $\mathrm{Tr}\left(\widetilde{\mathfrak{C}}(\mathcal{C}_{\mathrm{q}}, M) M^{-1}\right) \geq d^{3/2} = \sqrt{d}(1+\sqrt{d})^{-1}\mathrm{Tr}\left(\mathfrak{C}(\mathcal{C}_{\mathrm{PP}}, p_M) M^{-1}\right)$ and it allows to conclude.

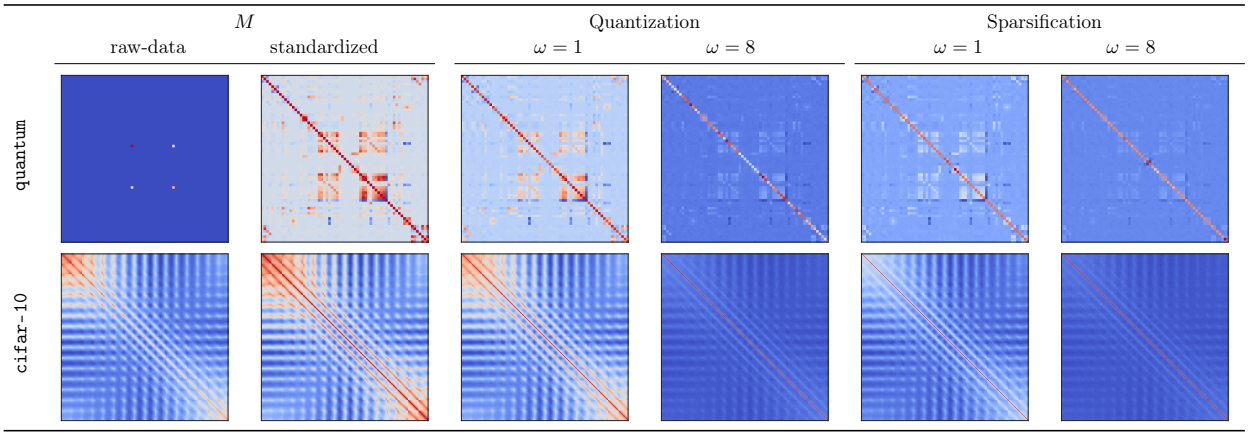
Proof of Item 2 in Proposition 4.4. Suppose now that $\mathrm{Diag}\left(M\right) = c\mathrm{I}_d$, then we have $\mathfrak{C}(\mathcal{C}_{\mathrm{PP}}, p_M) = (\sqrt{d}+1)M$, $\widetilde{\mathfrak{C}}(\mathcal{C}_{\mathrm{q}}, M) = M + (\sqrt{d}-1)c\mathrm{I}_d$, and $\mathfrak{C}(\mathcal{C}_{\mathrm{s}}, p_M) = M + c\sqrt{d}\mathrm{I}_d$. Firstly, it follows that:

$$\mathfrak{C}(\mathcal{C}_{\mathrm{s}}, p_M) - \widetilde{\mathfrak{C}}(\mathcal{C}_{\mathrm{q}}, M) = \left(M + \sqrt{d}c\mathrm{I}_d\right) - \left(M + (\sqrt{d}-1)c\mathrm{I}_d\right) = c\mathrm{I}_d \succcurlyeq 0\,,$$

Secondly, we have $(1+\frac{1}{\sqrt{d}})\widetilde{\mathfrak{C}}(\mathcal{C}_{\mathrm{q}}, M) - \mathfrak{C}(\mathcal{C}_{\mathrm{PP}}, p_M) = -(1-\frac{1}{\sqrt{d}})M + (\sqrt{d}-\frac{1}{\sqrt{d}})c\mathrm{I}_d$, which gives:

$$\begin{aligned}\mathrm{Tr}\left(\left((\sqrt{d}-\frac{1}{\sqrt{d}})\widetilde{\mathfrak{C}}(\mathcal{C}_{\mathrm{q}}, M) - \mathfrak{C}(\mathcal{C}_{\mathrm{PP}}, p_M)\right) M^{-1}\right) &= (\sqrt{d}-\frac{1}{\sqrt{d}})c\mathrm{Tr}\left(M^{-1}\right) - (1-\frac{1}{\sqrt{d}})\mathrm{Tr}\left(\mathrm{I}_d\right) \\ &\geq (\sqrt{d}-\frac{1}{\sqrt{d}})d - (1-\frac{1}{\sqrt{d}})d \quad \text{(Lemma D.9)} \\ &\geq d(\sqrt{d}-1) \geq 0\,.\end{aligned}$$

Table D.3: (1) Data covariances for `quantum` and `cifar-10`. (2) Covariance $\mathfrak{C}(\mathcal{C}_M, p_H)$ w./w.o. standardization for quantization and sparsification; see Figure 4.6 to have the corresponding trace of $\mathfrak{C}(\mathcal{C}_M, p_H)M^{-1}$.



And the proof is concluded.

■

D.5.4 Empirical covariances computed on quantum and cifar10

On Table D.3, for both `quantum` and `cifar-10`, we first plot the covariance matrix (1) without any processing and (2) with standardization. In this latter case, we then plot the covariances induced by quantization and sparsification for $\omega = 1$ and 8. For `quantum`, without standardization, only four points are visible; it is caused by some rows having extremely large values at features 27 and 43, resulting in a feature mean 100 times greater than the others.

Looking at the covariance induced by the compressors, we observe that for small ω, quantization better preserves the matrix structure compared to sparsification. This fact is consistent with Figure 4.6 where is given the trace of $\mathfrak{C}(\mathcal{C}_M, p_H)M^{-1}$ for these eight covariances: the traces for quantization are indeed smaller than for sparsification. This is also consistent with Figures 4.7c and 4.7f where $\omega = 1$ and where quantization outperforms sparsification.

D.6 Technical results on federated learning.

D.6.1 Validity of the assumptions made on the random fields in the case of covariate-shift

In this Subsection, we examine the setting of federated and compressed LSR under the scenario of covariate-shift (Subsection 4.4.1). Specifically, we consider the case where for any i, j in $[\![1, N]\!]$, we have heterogeneous covariances, i.e., $H_i \neq H_j$, but a unique optimal model i.e. $w_*^i = w_*$. We verify that all the assumptions on the random fields done in Subsection 4.2.1 are fulfilled in the setting. For this purpose, we redefine the filtration given in Section D.4 to align them with the FL setting. For k in $\mathbb{N}^*$ and for i in $\{1, \dots, N\}$, we note u_k^i the noise that controls the compression $\mathcal{C}_k^i(\cdot)$ at round k.

Definition D.2. *We note $(\mathcal{G}_k)_{k\in\mathbb{N}}$ the filtration associated with the features noise, $(\mathcal{H}_k)_{k\in\mathbb{N}}$ the filtration associated with the label noise, and $(\mathcal{I}_k)_{k\in\mathbb{N}}$ the filtration associated to the stochastic gradient*

noise, which is the union of the two previous filtrations. For $k \in \mathbb{N}^*$, we define $\mathcal{F}_0 = \{\varnothing\}$ and

$$\begin{aligned}
\mathcal{G}_k &= \sigma\left(\mathcal{F}_{k-1} \cup \{(x_k^i)_{i=1}^N\}\right) \\
\mathcal{H}_k &= \sigma\left(\mathcal{F}_{k-1} \cup \{(\varepsilon_k^i)_{i=1}^N\}\right) \\
\mathcal{I}_k &= \sigma\left(\mathcal{F}_{k-1} \cup \{(x_k^i, \varepsilon_k^i)_{i=1}^N\}\right) \\
\mathcal{F}_k &= \sigma\left(\mathcal{F}_{k-1} \cup \{(x_k, \varepsilon_k^i, u_k^i)_{i=1}^N\}\right) .
\end{aligned}$$

Now we prove that all assumptions done in Section 4.2 are correct in this setting.

Property D.7 (Validity of the setting presented in Definition 4.1)**.** *Consider Algorithm 3 in the context of Model 1, we have that the setting presented in Definition 4.1 is verified.*

Proof From Algorithm 3, we have for any k in $\mathbb{N}^*$ and any w in $\mathbb{R}^d$, $\xi_k(w - w_*) = \nabla F(w) - \frac{1}{N}\sum_{i=1}^N \mathcal{C}_k^i(g_k^i(w))$. Because $(g_k^i)_{k\in\mathbb{N}^*, i\in[\![1,N]\!]}$ and $(\mathcal{C}_k^i)_{k\in\mathbb{N}^*, i\in[\![1,N]\!]}$ are by definition two sequences of i.i.d. random fields (Algorithm 3), it follows that their composition is also i.i.d., *hence $(\xi_k)_{k\in\mathbb{N}^*}$ is a sequence of i.i.d. random fields.*

Taking expectation w.r.t. the σ-algebra $\mathcal{I}_k$ we have $\mathbb{E}\left[\mathcal{C}_k^i(g_k^i(w)) \,\middle|\, \mathcal{I}_k\right] = g_k^i(w)$ (Lemma 4.1), next with the σ-algebra $\mathcal{F}_{k-1}$, we have $\mathbb{E}\left[g_k^i(w) \,\middle|\, \mathcal{F}_{k-1}\right] = \nabla F_i(w)$ (Equation (4.2)). And because $\frac{1}{N}\sum_{i=1}^N \nabla F_i(w) = \nabla F(w)$, *we obtain that the random fields are zero-centered.*

From Model 1, we have for any k in $\mathbb{N}^*$ and any w in $\mathbb{R}^d$ that:

$$\begin{aligned}
F(w) &= \frac{1}{2N}\sum_{i=1}^N \mathbb{E}\left[(\langle x_k^i, w\rangle - y_k^i)^2\right] \\
&= \frac{1}{2N}\sum_{i=1}^N \mathbb{E}\left[(w - w_*)^\top (x_k^i \otimes x_k^i)(w - w_*) - 2\varepsilon_k^i \langle x_k^i, w - w_*\rangle + (\varepsilon_k^i)^2\right] \\
&= \frac{1}{2N}\sum_{i=1}^N (w - w_*)^\top H_i (w - w_*) + \sigma^2 = \frac{1}{2}((w - w_*)^\top \overline{H}(w - w_*) + \sigma^2) .
\end{aligned}$$

And we have from Model 1: $\operatorname{Tr}\left(\overline{H}\right) = \frac{1}{N}\sum_{i=1}^N \operatorname{Tr}\left(H_i\right) = \frac{1}{N}\sum_{i=1}^N R_i^2 =: \overline{R}^2$, which concludes the verification. ■

Property D.8 (Validity of Assumption 4.1)**.** *Consider Algorithm 3 in the context of Model 1 with Lemma 4.1, for any iteration k in $\mathbb{N}^*$, the second moment of the additive noise ξ_k^{add} can be bounded by $(\omega+1)\overline{R}^2\sigma^2/N$ i.e. Assumption 4.1 is verified.*

Proof Let k in $\mathbb{N}^*$. Because we consider Algorithm 3, with Definitions 4.1 and 4.2, we first have $\xi_k^{\text{add}} = -\frac{1}{N}\sum_{i=1}^N \mathcal{C}_k^i(g_{k,*}^i)$, hence taking expectation w.r.t the σ-algebra $\mathcal{I}_k$ and because the N compressions are independent (Algorithm 3), using Lemma 4.1, we have that:

$$\begin{aligned}
\mathbb{E}\left[\|\xi_k^{\text{add}}\|^2 \,\middle|\, \mathcal{I}_k\right] &= \frac{1}{N^2}\sum_{i=1}^N \mathbb{E}\left[\left\|\mathcal{C}_k^i(g_{k,*}^i)\right\|^2 \,\middle|\, \mathcal{I}_k\right] + \frac{1}{N^2}\sum_{i\neq j}\left\langle g_{k,*}^i, g_{k,*}^j\right\rangle \\
&\leq \frac{\omega+1}{N^2}\sum_{i=1}^N \left\|g_{k,*}^i\right\|^2 + \frac{1}{N^2}\sum_{i\neq j}\left\langle g_{k,*}^i, g_{k,*}^j\right\rangle .
\end{aligned}$$

Next, we first have from Model 1 and Equation (4.2) that for any i in $\{1,\dots,N\}$, $g_{k,*}^i = -\varepsilon_k^i x_k^i$, secondly because $\left((\varepsilon_k^i)_{k\in\{1,\dots,K\}, i\in\{1,\dots,N\}}\right)$ are independent from $\left((x_k^i)_{k\in\{1,\dots,K\}, i\in\{1,\dots,N\}}\right)$ (Model 1), we

have that $\mathbb{E}[\|\varepsilon_k^i x_k^i\|^2] \leq \sigma^2 R_i^2$, hence it results to $\mathbb{E}\left[\|\xi_k^{\text{add}}\|^2 \mid \mathcal{F}_{k-1}\right] = \mathbb{E}[\|\xi_k^{\text{add}}\|^2] = \frac{\omega+1}{N^2}\sum_{i=1}^N \sigma^2 R_i^2$. ■

Property D.9 (Validity of Assumption 4.2.1)**.** *Consider Algorithm 3 in the context of Model 1 with Lemma 4.1, for any iteration k in $\mathbb{N}^*$, the second moment of the multiplicative noise $\xi_k^{\text{mult}}(w)$ can be bounded for any w in $\mathbb{R}^d$ by $2(\omega+1)\max_{i\in\{1,\dots,N\}}(R_i^2)\left\|\overline{H}^{1/2}(w-w_*)\right\|^2/N + 4(\omega+1)\overline{R}^2\sigma^2/N$ i.e. Assumption 4.2.1 is verified.*

Proof Let k in $\mathbb{N}^*$, we note $\eta = w - w_*$. Because we consider Algorithm 3, with Definitions 4.1 and 4.2, we write $\xi_k^{\text{mult}}(\eta) = \frac{1}{N}\sum_{i=1}^N \xi_k^{i,\text{mult}}(\eta)$, where $\xi_k^{i,\text{mult}}(\eta) = H_i\eta - \mathcal{C}(g_k^i(w)) + \mathcal{C}(g_{k,*}^i)$ is the multiplicative noise on client i in $\{1,\dots,N\}$, hence developing the squared norm gives:

$$\left\|\xi_k^{\text{mult}}(\eta)\right\|^2 = \left\|\frac{1}{N}\sum_{i=1}^N \xi_k^{i,\text{mult}}(\eta)\right\|^2 = \frac{1}{N^2}\sum_{i=1}^N \left\|\xi_k^{i,\text{mult}}(\eta)\right\|^2 + \frac{1}{N^2}\sum_{i\neq j}\left\langle \xi_k^{i,\text{mult}}(\eta), \xi_k^{j,\text{mult}}(\eta)\right\rangle .$$

Taking expectation w.r.t. the σ-algebra $\mathcal{F}_{k-1}$, using that the N compressions are independent (Algorithm 3) and that for any i in $\{1,\dots,N\}$, $\mathbb{E}[\xi_k^{i,\text{mult}}(\eta) \mid \mathcal{F}_{k-1}] = 0$ (Lemma 4.1) results to have:

$$\mathbb{E}[\|\xi_k^{\text{mult}}(\eta)\|^2 \mid \mathcal{F}_{k-1}] = \frac{1}{N^2}\sum_{i=1}^N \mathbb{E}[\|\xi_k^{i,\text{mult}}(\eta)\|^2 \mid \mathcal{F}_{k-1}] .$$

Next, we use the result of Property D.3 for each client i in $\{1,\dots,N\}$ and we obtain:

$$\begin{aligned}\mathbb{E}\left[\|\xi_k^{\text{mult}}(\eta)\|^2 \,\middle|\, \mathcal{F}_{k-1}\right] &\leq \frac{1}{N^2}\sum_{i=1}^N \left(2(\omega+1)R_i^2\|H_i^{1/2}(w-w_*)\|^2 + 4(\omega+1)R_i^2\sigma^2\right) \\ &\leq \frac{2(\omega+1)\max_{i\in\{1,\dots,N\}}(R_i^2)}{N}\|\overline{H}^{1/2}(w-w_*)\|^2 + \frac{4(\omega+1)\overline{R}^2\sigma^2}{N} ,\end{aligned}$$

which allows concluding.

■

Property D.10 (Validity of Assumption 4.2.2)**.** *Consider Algorithm 3 in the context of Model 1 with Lemma 4.1, for any iteration k in $\mathbb{N}^*$, the second moment of the multiplicative noise $\xi_k^{\text{mult}}(w)$ can be bounded for any w in $\mathbb{R}^d$ by $(\Omega\sigma\max_{i\in\{1,\dots,N\}}(R_i^2)\|\overline{H}^{1/2}(w-w_*)\| + (\omega+1)\max_{i\in\{1,\dots,N\}}(R_i^2)\|\overline{H}^{1/2}(w-w_*)\|^2)/N$ i.e. Assumption 4.2.2 is verified.*

Proof Let k in $\mathbb{N}^*$, we note $\eta = w - w_*$. From Property D.9, taking expectation w.r.t. the σ-algebra $\mathcal{F}_{k-1}$, decomposing the multiplicative noise results to have:

$$\mathbb{E}[\|\xi_k^{\text{mult}}(\eta)\|^2 \mid \mathcal{F}_{k-1}] = \frac{1}{N^2}\sum_{i=1}^N \mathbb{E}[\|\xi_k^{i,\text{mult}}(\eta)\|^2 \mid \mathcal{F}_{k-1}] .$$

Next we use the result of Property D.4 for each client i in $\{1,\dots,N\}$ and we obtain:

$$\mathbb{E}[\|\xi_k^{\text{mult}}(\eta)\|^2 \mid \mathcal{F}_{k-1}] \leq \frac{1}{N^2}\sum_{i=1}^N \Omega R_i^2\sigma\sqrt{\|H_i^{1/2}(w-w_*)\|^2} + (\omega+1)R_i^2\|H_i^{1/2}(w-w_*)\|^2 .$$

With Jensen's inequality A.7 used for concave function:

$$\begin{aligned}\mathbb{E}\left[\left\|\xi_k^{\text{mult}}(\eta)\right\|^2 \,\middle|\, \mathcal{F}_{k-1}\right] &\leq \frac{\Omega\sigma \max_{i\in\{1,\dots,N\}}(R_i^2)}{N}\sqrt{\frac{1}{N}\sum_{i=1}^N \|H_i^{1/2}(w-w_*)\|^2} \\ &\quad + \frac{(\omega+1)\max_{i\in\{1,\dots,N\}}(R_i^2)}{N^2}\sum_{i=1}^N \|H_i^{1/2}(w-w_*)\|^2 \\ &\leq \frac{\Omega\sigma \max_{i\in\{1,\dots,N\}}(R_i^2)}{N}\sqrt{\|\overline{H}^{1/2}(w-w_*)\|^2} \\ &\quad + \frac{1}{N}(\omega+1)\max_{i\in\{1,\dots,N\}}(R_i^2)\|\overline{H}^{1/2}(w-w_*)\|^2\,,\end{aligned}$$

which allows concluding.

■

Property D.11 (Validity of Assumption 4.3)**.** *Consider Algorithm 3 in the context of Model 1 with Lemma 4.1, if the compressor $\mathcal{C}$ is linear, then for any iteration k in $\mathbb{N}^*$, the multiplicative noise ξ_k^{mult} is linear, thus there exist a matrix Ξ_k in $\mathbb{R}^{d\times d}$ such that for any w in $\mathbb{R}^d$, $\xi_k^{\text{mult}}(w) = \Xi_k w$. Furthermore the second moment of the multiplicative noise can be bounded for any w in $\mathbb{R}^d$ by $(\omega+1)\max_{i\in\{1,\dots,N\}}(R_i^2)\left\|\overline{H}^{1/2}(w-w_*)\right\|^2/N$, hence Assumption 4.3 is verified.*

Proof Let k in $\mathbb{N}^*$, we note $\eta = w - w_*$. Because we consider Algorithm 3, with Definitions 4.1 and 4.2, we write $\xi_k^{\text{mult}}(\eta) = \frac{1}{N}\sum_{i=1}^N \xi_k^{i,\text{mult}}(\eta)$, where $\xi_k^{i,\text{mult}}(\eta) = H_i\eta - \mathcal{C}(g_k^i(w)) + \mathcal{C}(g_{k,*}^i)$ is the multiplicative noise on client i in $\{1,\dots,N\}$. And because for any clients i in $\{1,\cdots N\}$ the random mechanism $\mathcal{C}_k^i$ is linear, there exists a random matrix Π_k^i in $\mathbb{R}^{d\times d}$ s.t. for any z in $\mathbb{R}^d$, we have $\mathcal{C}_k^i(z) = \Pi_k^i z$, it follows that:

$$\xi_k^{\text{mult}}(\eta) = \nabla F(w) - \frac{1}{N}\sum_{i=1}^N \mathcal{C}_k^i(g_k^i(w)) + \mathcal{C}_k^i(g_{k,*}^i) = \left(\overline{H} - \frac{1}{N}\sum_{i=1}^N \Pi_k^i(x_k^i\otimes x_k^i)\right)\eta\,.$$

Hence, the first part of Assumption 4.2.2 is verified with $\Xi_k = \frac{1}{N}\sum_{i=1}^N H_i - \Pi_k^i(x_k^i\otimes x_k^i)$. From Property D.9, taking expectation w.r.t. the σ-algebra $\mathcal{F}_{k-1}$, decomposing the multiplicative noise results to have:

$$\mathbb{E}\left[\left\|\xi_k^{\text{mult}}(\eta)\right\|^2 \,\middle|\, \mathcal{F}_{k-1}\right] = \frac{1}{N^2}\sum_{i=1}^N \mathbb{E}\left[\left\|\xi_k^{i,\text{mult}}(\eta)\right\|^2 \,\middle|\, \mathcal{F}_{k-1}\right]\,.$$

Next we use the result of Property D.5 for each client i in $\{1,\dots,N\}$ and we obtain:

$$\begin{aligned}\mathbb{E}\left[\left\|\xi_k^{\text{mult}}(\eta)\right\|^2 \,\middle|\, \mathcal{F}_{k-1}\right] &\leq \frac{1}{N}\sum_{i=1}^N (\omega+1)R_i^2\left\|H_i^{1/2}(w-w_*)\right\|^2 \\ &\leq \frac{(\omega+1)\max_{i\in\{1,\dots,N\}}(R_i^2)}{N^2}\left\|\frac{1}{N}\sum_{i=1}^N H_i^{1/2}(w-w_*)\right\|^2\,,\end{aligned}$$

which allows concluding.

■

Property D.12 (Validity of Assumption 4.4)**.** *Considering Algorithm 3 under the setting of Model 2 with Remark 4.1 and Lemma 4.1, if the compressor $\mathcal{C}$ is linear, then for any k in $\mathbb{N}^*$, we have $\mathfrak{C}_{\text{ania}} \preccurlyeq \sigma^2 \max_{i\in\{1,\dots,N\}}(\text{Ш}_{H_i})\overline{H}/N$ and $\mathbb{E}\left[\Xi_k\Xi_k^\top\right] \preccurlyeq \max_{i\in\{1,\dots,N\}}(R_i^2\text{Ш}_{H_i})\overline{H}/N$, with $(\text{Ш}_{H_i})_{i\in\{1,\dots,N\}}$ given in Corollary 4.2. Overall, Assumption 4.4 is thus verified.*

Proof

First inequality.

By Definition 4.3, we have $\mathfrak{C}_{\text{ania}} = \mathbb{E}[\xi_k^{\text{add}} \otimes \xi_k^{\text{add}} \mid \mathcal{F}_{k-1}] = \frac{1}{N^2}\sum_{i=1}^N \mathbb{E}[\mathcal{C}_k^i(g_{k,*}^i)^{\otimes 2} \mid \mathcal{F}_{k-1}]$, because for any client i in $\{1,\dots,N\}$ $\left((\varepsilon_k^i)_{k\in\{1,\dots,K\}}\right)$ is independent from $\left((x_k^i)_{k\in\{1,\dots,K\}}\right)$ (Model 1) and using compressor linearity and Equation (D.22), it gives:

$$\mathfrak{C}_{\text{ania}} = \sigma^2\frac{1}{N^2}\sum_{i=1}^N \mathbb{E}\left[\mathcal{C}_k^i(x_k^i)^{\otimes 2}\right] = \frac{\sigma^2}{N^2}\sum_{i=1}^N \mathfrak{C}(\mathcal{C}^i, p_{H_i}) \preccurlyeq \frac{\sigma^2}{N^2}\sum_{i=1}^N \text{Ш}_{H_i} H \preccurlyeq \frac{\sigma^2 \max_{i\in\{1,\dots,N\}}(\text{Ш}_{H_i})}{N}\overline{H}\,.$$

Second inequality.

Using Property D.11, because the random mechanism $\mathcal{C}^i$ is linear, there exists two matrices Π_k^i, Ξ_k^i in $\mathbb{R}^{d\times d}$ s.t. for any z in $\mathbb{R}^d$, we have $\mathcal{C}_k^i(z) = \Pi_k^i z$ and $\xi_k^{\text{mult,i}}(z) = \Xi_k^i z = (H_i - \Pi_k^i(x_k^i \otimes x_k^i))z$, which gives that $\Xi_k = \frac{1}{N}\sum_{i=1}^N H_i - \Pi_k^i(x_k^i \otimes x_k^i)$. It follows that:

$$\Xi_k\Xi_k^\top = \frac{1}{N^2}\sum_{i=1}^N (\Xi_k^i)(\Xi_k^i)^\top + \frac{1}{N^2}\sum_{i\neq j}(\Xi_k^i)(\Xi_k^j)^\top\,.$$

Taking the σ-algebra $\mathcal{F}_{k-1}$, using that the N compressions are independent (Algorithm 3) and that for any i in $\{1,\dots,N\}$, $\mathbb{E}\left[\xi_k^{i,\text{mult}} \;\middle|\; \mathcal{F}_{k-1}\right] = 0$ (Lemma 4.1) results to have $\mathbb{E}[\Xi_k\Xi_k^\top \mid \mathcal{F}_{k-1}] = \frac{1}{N^2}\sum_{i=1}^N \mathbb{E}\left[(\Xi_k^i)(\Xi_k^i)^\top \mid \mathcal{F}_{k-1}\right]$. Now, we can reuse the computations given in Property D.6 to obtain $\mathbb{E}\left[(\Xi_k^i)(\Xi_k^i)^\top \mid \mathcal{F}_{k-1}\right] \preccurlyeq R_i^2\text{Ш}_{H_i}H_i$. Therefore, we can state that $\mathbb{E}\left[\Xi_k\Xi_k^\top \mid \mathcal{F}_{k-1}\right] \preccurlyeq \max_{i\in\{1,\dots,N\}}(R_i^2\text{Ш}_{H_i})\overline{H}/N$, which concludes the second part of the verification of Assumption 4.4.

■

D.6.2 Heterogeneous optimal point

In this Section, we explore further the scenario of concept-shift by adding a memory mechanism, as in Section 2.3. Indeed, Theorem 2.1 shows that this mechanism improves the convergence in the case of heterogeneous clients. We give below the updates equation defining the algorithm of distributed compressed LSR with memory.

Algorithm 4 (Distributed compressed LMS with control variates)**.** *Each client $i \in \{1,\dots,N\}$ maintains a sequence $(h_k^i)_{i\in\{1,\dots,N\}}$ in $\mathbb{R}^d$, observes at any step $k \in \{1,\dots,K\}$ an oracle $g_k^i(\cdot)$ on the gradient of the local objective function F_i and applies an independent random compression mechanism $\mathcal{C}_k^i(\cdot)$ to the difference $g_k^i - h_k^i$. And for any step-size $\gamma > 0$, any $k \in \mathbb{N}^*$, the sequence of iterates $(w_k)_{k\in\mathbb{N}}$ satisfies:*

$$\begin{cases} w_k = w_{k-1} - \dfrac{\gamma}{N}\displaystyle\sum_{i=1}^N \mathcal{C}_k^i(g_k^i(w_{k-1}) - h_{k-1}^i) + h_{k-1}^i \\ h_k^i = h_{k-1}^i + \alpha\mathcal{C}_k^i(g_k^i(w_{k-1}) - h_{k-1}^i)\,, \end{cases} \tag{D.28}$$

with $\alpha = 1/2(\omega+1)$.

The counterpart of adding memory is that the random fields are no more identically distributed, thus Definition 4.1 is not fulfilled, and results from Section 4.2 cannot be applied, especially because $\mathbb{E}\left[\xi_k^{\text{add}} \otimes \xi_k^{\text{add}}\right]$ changes along iterations. To remedy this problem, we define here the *limit* of the covariance of the additive noise i.e. $\mathfrak{C}_{\text{ania}}^{\infty} = \lim_{k \to +\infty} \mathbb{E}\left[\xi_k^{\text{add}} \otimes \xi_k^{\text{add}}\right]$. In the following result, we establish an asymptotic result on the convergence, similar to Theorem 4.1.

Theorem D.3 (CLT for concept-shift heterogeneity)**.** *Consider Algorithm 4 under Model 1 with $\mu > 0$ and Lemma 4.1, for any step-size $(\gamma_k)_{k \in \mathbb{N}^*}$ s.t. $\gamma_k = 1/\sqrt{k}$. Then*

1. $(\sqrt{K}\overline{\eta}_{K-1})_{K>0} \xrightarrow[K \to +\infty]{\mathcal{L}} \mathcal{N}(0, H_F^{-1} \mathfrak{C}_{\text{ania}}^{\infty} H_F^{-1})$,
2. $\mathfrak{C}_{\text{ania}}^{\infty} = \overline{\mathfrak{C}((\mathcal{C}^i, p_{\Theta_i'})_{i=1}^N)}$, *where $p_{\Theta_i'}$ is the distribution of $g_{k,*}^i - \nabla F_i(w_*)$.*

Theorem D.3 shows that when using memory, in the settings of heterogeneous optimal points $(w_*^i)_{i=1}^N$, convergence is still impacted by heterogeneity but with a smaller additive noise's covariance as $\Theta_i' \prec \Theta_i$. In particular, in the case of deterministic gradients (batch case), we case $\Theta_i' \equiv 0$. From a technical standpoint, it shows that we recover asymptotically the results stated by Theorems 4.1 and 4.2 in the general setting of i.i.d. random fields $(\xi_k(\eta_{k-1}))_{k \in \mathbb{N}^*}$. To prove this theorem, we show that the assumptions required by Theorem A.1 are fulfilled by this framework.

Proof

For the sake of demonstration, we define a Lyapunov function V_k as in Section B.4, with k in $[\![1, K]\!]$:

$$V_k = \|\eta_k\|^2 + 2\gamma_k^2 C \frac{1}{N} \sum_{i=1}^N \left\|h_{k-1}^i - \nabla F_i(w_*)\right\|^2 ,$$

with C in $\mathbb{R}^*$ being a Lyapunov constant that verifies some conditions given in Theorem B.2. For any k in $\mathbb{N}$, the Lyapunov function is defined combining two terms: (1) the distance from parameter w_k to the optimal parameter w_*, (2) for any client i in $\{1, \dots, N\}$, the distance between the memory term h_{k-1}^i and the true gradient $\nabla F_i(w_*)$.

First, we have that in the case of decreasing step size s.t. for any k in $\mathbb{N}$, $\gamma_k = k^{-\alpha}$, we have $\eta_K \xrightarrow[K \to +\infty]{L^2} 0$ and $h_K^i \xrightarrow[K \to +\infty]{L^2} \nabla F_i(w_*)$.

Let k in $\mathbb{N}^*$, from the demonstration of the `Artemis` algorithm with memory, we have from Theorem B.2 that (1) combining Equation (B.12) and Equation (B.13) in the form (B.12)+$2\gamma_k^2 C$(B.13), (2) and applying strong-convexity, allows to obtain Equation (B.15) but adapted to decreasing step-size:

$$\mathbb{E}\left[V_k \mid \mathcal{F}_{k-1}\right] \leq (1 - 2\gamma_k \mu \square_k) \|w_{k-1} - w_*\|^2 + \frac{2\gamma_k^2 C \diamondsuit}{N} \sum_{i=1}^N \left\|h_{k-1}^i - \nabla F_i(w_*)\right\|^2 + \frac{2\gamma_k^2 \sigma \triangle}{N} ,$$

with $\square_k, \diamondsuit, \triangle$ being three constants in $\mathbb{R}$ whose exact expression is given in the proof of Theorem B.2. Furthermore, in the same article, they verify that to obtain a $(1 - \gamma_k \mu)$ convergence, the following condition on $\square_k, \diamondsuit, \triangle$ are fulfilled for any k in $\mathbb{N}^*$: $\square_k \leq 1/2$ and $\diamondsuit \leq 1 - \gamma_k \mu$.

These properties are valid under some conditions on the Lyapunov constant C, the step-size γ_k, and the learning rate α; these conditions are provided in the statement of Theorem B.2 and we don't recall them here. Hence, we can write that we have:

$$\mathbb{E}\left[V_k \mid \mathcal{F}_{k-1}\right] \leq (1 - \gamma_k \mu) \left(\|w_{k-1} - w_*\|^2 + \frac{2\gamma_k^2 C}{N} \sum_{i=1}^N \left\|h_{k-1}^i - \nabla F_i(w_*)\right\|^2 \right) + \frac{2\gamma_k^2 \sigma^2 \triangle}{N} ,$$

and because for any k in N, the step-size is decreasing, we have $\gamma_k \leq \gamma_{k-1}$, which makes to recover the Lyapunov function V_{k-1} at step $k-1$: $\mathbb{E}\left[V_k \mid \mathcal{F}_{k-1}\right] \leq (1-\gamma_k\mu)V_{k-1} + \frac{2\gamma_k^2\sigma^2\triangle}{N}$. Taking full expectation and unrolling the sequence $(V_k)_{k\in\mathbb{N}}$, we obtain:

$$\mathbb{E}V_k \leq \prod_{i=1}^{k}(1-\gamma_i\mu)V_0 + \frac{2\sigma^2\triangle}{N}\sum_{j=1}^{k}\gamma_j^2\prod_{i=j+1}^{k}(1-\gamma_i\mu)\,. \tag{D.29}$$

To show that each part of the bound given in Equation (D.29) tends to zero when k grows to infinity is classical computations and can be find for instance in lectures notes of Bach [2022, pages 164-167 and 182], and Kushner and Yin [2003].

To apply Theorem 1 from Polyak and Juditsky [1992, recalled in Theorem A.1], which gives the desired result, it suffices to prove the convergence in probability of the covariance of the noise $\xi_k(\eta_{k-1})$ towards $\mathfrak{C}_{\text{ania}}$, as $k \to \infty$.

In the following, we show that $\lim_{k\to+\infty}\mathbb{E}\left[\xi_k(\eta_{k-1})\xi_k(\eta_{k-1})^\top \mid \mathcal{F}_{k-1}\right] \overset{\mathbb{P}}{=} \mathfrak{C}_{\text{ania}}^{\infty}$**.** Let k in $\mathbb{N}^*$, for this purpose, we consider the following additive/multiplicative noise decomposition:

$$\begin{cases} \xi_{k,*}^{\mathrm{A}} = -\dfrac{1}{N}\displaystyle\sum_{i=1}^{N}\mathcal{C}_k^i(g_{k,*}^i - \nabla F_i(w_*)) \\ \xi_k^{\mathrm{M}}(\eta_k) = H_F\eta_k - \dfrac{1}{N}\displaystyle\sum_{i=1}^{N}\mathcal{C}_k^i(g_k^i(w_{k-1}) - h_{k-1}^i) + \dfrac{1}{N}\displaystyle\sum_{i=1}^{N}\mathcal{C}_k^i(g_{k,*} - \nabla F_i(w_*)) + h_{k-1}^i\,. \end{cases} \tag{D.30}$$

Furthermore, we have that $\xi_k^{\text{add}} \xrightarrow[k\to+\infty]{L^2} \xi_{k,*}^{\mathrm{A}}$ because of the Hölder-inequality (Lemma 4.1) and because we shown that $h_K^i \xrightarrow[K\to+\infty]{L^2} \nabla F_i(w_*)$; thereby $\mathbb{E}[\xi_k^{\text{add}} \otimes \xi_k^{\text{add}}] \xrightarrow[k\to+\infty]{L^1} \mathfrak{C}_{\text{ania}}^{\infty}$. Next, from Equation (D.30), we write:

$$\begin{aligned} \xi_k(\eta_{k-1})\xi_k(\eta_{k-1})^\top &= (\xi_{k,*}^{\mathrm{A}} - \xi_k^{\mathrm{M}}(\eta_{k-1}))(\xi_{k,*}^{\mathrm{A}} - \xi_k^{\mathrm{M}}(\eta_{k-1}))^\top \\ &= \xi_{k,*}^{\mathrm{A}} \otimes \xi_{k,*}^{\mathrm{A}} - \xi_{k,*}^{\mathrm{A}}\xi_k^{\mathrm{M}}(\eta_{k-1})^\top - \xi_k^{\mathrm{M}}(\eta_{k-1})(\xi_{k,*}^{\mathrm{A}})^\top + \xi_k^{\mathrm{M}}(\eta_{k-1}) \otimes \xi_k^{\mathrm{M}}(\eta_{k-1})\,. \end{aligned}$$

(i) Developing the covariance of the additive noise and considering Model 1 and Algorithm 3 results to $\mathbb{E}[\xi_{k,*}^{\mathrm{A}} \otimes \xi_{k,*}^{\mathrm{A}} \mid \mathcal{F}_{k-1}] = \frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}[\mathcal{C}_k^i(g_{k,*}^i - \nabla F_i(w_*))^{\otimes 2} \mid \mathcal{F}_{k-1}]$. For any iteration k in $\mathbb{N}^*$ and any client i in $\{1,\dots,N\}$, we note Θ_i' the covariance of $g_{k,*}^i - \nabla F_i(w_*)$, then $g_{k,*}^i - \nabla F_i(w_*)$ is an i.i.d. zero-centered random vectors draw from a distribution $p_{\Theta_i'}$, hence we have for any iteration k in $\mathbb{N}^*$, $\mathfrak{C}_{\text{ania}}^{\infty} = \mathbb{E}[\xi_{k,*}^{\mathrm{A}} \otimes \xi_{k,*}^{\mathrm{A}} \mid \mathcal{F}_{k-1}] = \overline{\mathfrak{C}(\mathcal{C}^i, (p_{\Theta_i'})_{i=1}^N)}$.

(ii) Second, we show that $\mathbb{E}\left[\xi_k^{\mathrm{M}}(\eta_{k-1})^{\otimes 2} \mid \mathcal{F}_{k-1}\right]$ converge to 0 in probability: it is sufficient to show that $\|\xi_k^{\mathrm{M}}(\eta_{k-1})^{\otimes 2}\|_F$ tends to 0. To do so, we use the fact that $\|\xi_k^{\mathrm{M}}(\eta_{k-1})^{\otimes 2}\|_F = \|\xi_k^{\mathrm{M}}(\eta_{k-1})\|_2^2$, it results to the following decomposition:

$$\begin{aligned} \|\xi_k^{\mathrm{M}}(\eta_{k-1})^{\otimes 2}\| \leq 3\,\|H\eta_{k-1}\|^2 &+ 3\left\|\frac{1}{N}\sum_{i=1}^{N}\mathcal{C}_k^i(g_k^i(w_{k-1}) - h_{k-1}^i) - \mathcal{C}_k^i(g_{k,*}^i - \nabla F_i(w_*))\right\|^2 \\ &+ 3\left\|\frac{1}{N}\sum_{i=1}^{N}h_{k-1}^i - \nabla F_i(w_*)\right\|^2\,. \end{aligned}$$

Considering the Hölder inequality given in Item L.2 from Lemma 4.1, because $\eta_k \xrightarrow[k\to+\infty]{L^2} 0$ and $h_k^i \xrightarrow[k\to+\infty]{L^2} \nabla F_i(w_*)$, we deduce that $\mathbb{E}\left[\xi_k^{\mathrm{M}}(\eta_{k-1})^{\otimes 2} \mid \mathcal{F}_{k-1}\right]$ tends to 0 in L^1-norm.

(iii) Third, it remains to show that $\mathbb{E}[\xi_k^{\mathrm{M}}(\eta_{k-1})(\xi_{k,*}^{\mathrm{A}})^\top \mid \mathcal{F}_{k-1}] \xrightarrow[k\to+\infty]{\mathbb{P}} 0$. We use the Cauchy-Schwarz inequality's A.8 for conditional expectation:

$$\begin{aligned}\mathbb{E}\left[\xi_k^{\mathrm{M}}(\eta_{k-1})(\xi_{k,*}^{\mathrm{A}})^\top\|_F \;\middle|\; \mathcal{F}_{k-1}\right]^2 &= \mathbb{E}\left[\xi_k^{\mathrm{M}}(\eta_{k-1})\|_2\|(\xi_{k,*}^{\mathrm{A}})^\top\|_2 \;\middle|\; \mathcal{F}_{k-1}\right]^2 \\ &\leq \|\mathbb{E}\left[\xi_k^{\mathrm{M}}(\eta_{k-1})\|_2^2 \;\middle|\; \mathcal{F}_{k-1}\right] \mathbb{E}\left[\|(\xi_{k,*}^{\mathrm{A}})^\top\|_2^2 \;\middle|\; \mathcal{F}_{k-1}\right].\end{aligned}$$

The sequence of random vectors $(\xi_{k,*}^{\mathrm{A}})_{k\in\mathbb{N}}$ is i.i.d., and moreover we have shown previously that $\xi_k^{\mathrm{M}}(\eta_{k-1})^{\otimes 2}$ tends to 0, hence $\mathbb{E}[\xi_k^{\mathrm{M}}(\eta_{k-1})(\xi_{k,*}^{\mathrm{A}})^\top \mid \mathcal{F}_{k-1}]$ converges to 0 in distribution. Consequently, noting $\Theta_i' = \mathbb{E}[g_{k,*}^i - \nabla F_i(w_*))^{\otimes 2}]$ we can state that:

$$\mathbb{E}\left[\xi_k(\eta_{k-1})^{\otimes 2} \;\middle|\; \mathcal{F}_{k-1}\right] \xrightarrow[k\to+\infty]{\mathbb{P}} \mathfrak{C}_{\mathrm{ania}}^{\infty} = \overline{\mathfrak{C}(\mathcal{C}^i, (p_{\Theta_i'})_{i=1}^N)}.$$

■

Bibliography

M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: a system for large-scale machine learning. In *Proceedings of the 12th USENIX conference on Operating Systems Design and Implementation*, OSDI'16, pages 265–283, USA, Nov. 2016. USENIX Association. ISBN 978-1-931971-33-1.

N. Agarwal, A. T. Suresh, F. X. X. Yu, S. Kumar, and B. McMahan. cpSGD: Communication-efficient and differentially-private distributed SGD. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 7564–7575. Curran Associates, Inc., 2018.

V. Albino, U. Berardi, and R. M. Dangelico. Smart cities: Definitions, dimensions, performance, and initiatives. *Journal of urban technology*, 22(1):3–21, 2015.

D. Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic. QSGD: Communication-Efficient SGD via Gradient Quantization and Encoding. *Advances in Neural Information Processing Systems*, 30: 1709–1720, 2017.

D. Alistarh, T. Hoefler, M. Johansson, N. Konstantinov, S. Khirirat, and C. Renggli. The Convergence of Sparsified Gradient Methods. *Advances in Neural Information Processing Systems*, 31:5973–5983, 2018.

L. F. W. Anthony, B. Kanding, and R. Selvan. Carbontracker: Tracking and predicting the carbon footprint of training deep learning models. *arXiv preprint arXiv:2007.03051*, 2020.

R. Arablouei, S. Werner, K. Doğançay, and Y.-F. Huang. Analysis of a reduced-communication diffusion lms algorithm. *Signal Processing*, 117:355–361, 2015.

M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

F. Bach. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4(none): 384–414, Jan. 2010. ISSN 1935-7524, 1935-7524. doi: 10.1214/09-EJS521. Publisher: Institute of Mathematical Statistics and Bernoulli Society.

F. Bach. Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *The Journal of Machine Learning Research*, 15(1):595–627, 2014.

F. Bach. Lecture notes on statistical machine learning and convex optimization, 2022.

F. Bach and E. Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate O(1/n). In *Neural Information Processing Systems (NIPS)*, pages –, United States, Dec. 2013.

C. Badue, R. Guidolini, R. V. Carneiro, P. Azevedo, V. B. Cardoso, A. Forechi, L. Jesus, R. Berriel, T. M. Paixao, F. Mutz, et al. Self-driving cars: A survey. *Expert Systems with Applications*, 165: 113816, 2021.

M. Beaussart, F. Grimberg, M.-A. Hartley, and M. Jaggi. WAFFLE: Weighted Averaging for Personalized Federated Learning, Dec. 2021. arXiv:2110.06978 [cs].

M. Belkin, D. Hsu, S. Ma, and S. Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32): 15849–15854, 2019. Publisher: National Acad Sciences.

R. M. Bell, Y. Koren, and C. Volinsky. The bellkor solution to the netflix prize. *KorBell Team's Report to Netflix*, 2007.

A. Benveniste, M. Métivier, and P. Priouret. *Adaptive algorithms and stochastic approximations*, volume 22. Springer Science & Business Media, 2012.

J. Bernstein, Y.-X. Wang, K. Azizzadenesheli, and A. Anandkumar. signsgd: Compressed optimisation for non-convex problems. In *International Conference on Machine Learning*, pages 560–569. PMLR, 2018.

N. Bershad. Analysis of the normalized lms algorithm with gaussian inputs. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34(4):793–806, 1986.

A. Beznosikov, S. Horváth, P. Richtárik, and M. Safaryan. On Biased Compression for Distributed Learning. *arXiv:2002.12410 [cs, math, stat]*, Feb. 2020. arXiv: 2002.12410.

S. Bitam, A. Mellouk, and S. Zeadally. Vanet-cloud: a generic cloud computing model for vehicular ad hoc networks. *IEEE Wireless Communications*, 22(1):96–102, 2015.

J. R. Blum. Multidimensional stochastic approximation methods. *The Annals of Mathematical Statistics*, pages 737–744, 1954.

L. Bottou. Online learning and stochastic approximations. 1999. doi: 10.1017/CBO9780511569920.003.

L. Bottou. Large-Scale Machine Learning with Stochastic Gradient Descent. In Y. Lechevallier and G. Saporta, editors, *Proceedings of COMPSTAT'2010*, pages 177–186, Heidelberg, 2010. Physica-Verlag HD. ISBN 978-3-7908-2604-3. doi: 10.1007/978-3-7908-2604-3_16.

L. Bottou and O. Bousquet. The Tradeoffs of Large Scale Learning. *Advances in Neural Information Processing Systems*, 20:161–168, 2007.

S. Boyd, S. P. Boyd, and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

S. Brandi, M. S. Piscitelli, M. Martellacci, and A. Capozzoli. Deep reinforcement learning to optimise indoor temperature control and heating energy consumption in buildings. *Energy and Buildings*, 224:110225, 2020.

L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

S. Bubeck. Convex Optimization: Algorithms and Complexity. *arXiv:1405.4980 [cs, math, stat]*, Nov. 2015. arXiv: 1405.4980.

S. Caldas, S. M. K. Duddu, P. Wu, T. Li, J. Konečný, H. B. McMahan, V. Smith, and A. Talwalkar. LEAF: A Benchmark for Federated Settings. *arXiv:1812.01097 [cs, stat]*, Dec. 2019. arXiv: 1812.01097.

M. Campbell, M. Egerstedt, J. P. How, and R. M. Murray. Autonomous driving in urban environments: approaches, lessons and challenges. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 368(1928):4649–4672, 2010.

L. Carratino, A. Rudi, and L. Rosasco. Learning with SGD and Random Features. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

R. Caruana, T. Joachims, and L. Backstrom. KDD-Cup 2004: results and analysis. *ACM SIGKDD Explorations Newsletter*, 6(2):95–108, Dec. 2004. ISSN 1931-0145. doi: 10.1145/1046456.1046470.

C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27:1–27:27, May 2011. ISSN 2157-6904. doi: 10.1145/1961189.1961199.

J. H. Chen and S. M. Asch. Machine learning and prediction in medicine—beyond the peak of inflated expectations. *The New England journal of medicine*, 376(26):2507, 2017.

W. Chen, S. Horvath, and P. Richtarik. Optimal Client Sampling for Federated Learning, Oct. 2020. arXiv:2010.13723 [cs] version: 1.

H.-T. Cheng, L. Koc, J. Harmsen, T. Shaked, T. Chandra, H. Aradhye, G. Anderson, G. Corrado, W. Chai, M. Ispir, et al. Wide & deep learning for recommender systems. In *Proceedings of the 1st workshop on deep learning for recommender systems*, pages 7–10, 2016.

T. Chilimbi, Y. Suzue, J. Apacible, and K. Kalyanaraman. Project adam: Building an efficient and scalable deep learning training system. In *11th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 14)*, pages 571–582, 2014.

E. Choi, M. T. Bahadori, A. Schuetz, W. F. Stewart, and J. Sun. Doctor ai: Predicting clinical events via recurrent neural networks. In *Machine learning for healthcare conference*, pages 301–318. PMLR, 2016.

S. Chraibi, A. Khaled, D. Kovalev, P. Richtárik, A. Salim, and M. Takáč. Distributed fixed point methods with compressed iterates. *arXiv preprint arXiv:1912.09925*, 2019.

I. Colin, A. Bellet, J. Salmon, and S. Clémençon. Gossip Dual Averaging for Decentralized Optimization of Pairwise Functions. In *International Conference on Machine Learning*, pages 1388–1396. PMLR, June 2016. ISSN: 1938-7228.

L. Collins, H. Hassani, A. Mokhtari, and S. Shakkottai. Exploiting shared representations for personalized federated learning. In *International Conference on Machine Learning*, pages 2089–2099. PMLR, 2021.

L. Condat and P. Richtarik. Murana: A generic framework for stochastic variance-reduced optimization. In B. Dong, Q. Li, L. Wang, and Z.-Q. J. Xu, editors, *Proceedings of Mathematical and Scientific Machine Learning*, volume 190 of *Proceedings of Machine Learning Research*, pages 81–96. PMLR, 15–17 Aug 2022.

L. Condat, K. Yi, and P. Richtárik. EF-BV: A unified theory of error feedback and variance reduction mechanisms for biased and unbiased compression in distributed optimization. In A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=PeJO709WUup.

J. Contreras-Castillo, S. Zeadally, and J. A. Guerrero Ibáñez. A seven-layered model architecture for internet of vehicles. *Journal of Information and Telecommunication*, 1(1):4–22, 2017.

C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

P. Covington, J. Adams, and E. Sargin. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM conference on recommender systems*, pages 191–198, 2016.

D. Csiba and P. Richtárik. Importance sampling for minibatches. *The Journal of Machine Learning Research*, 19(1):962–982, 2018.

X. Dai, X. Yan, K. Zhou, H. Yang, K. K. Ng, J. Cheng, and Y. Fan. Hyper-sphere quantization: Communication-efficient sgd for federated learning. *arXiv preprint arXiv:1911.04655*, 2019.

Dall-E. An old castle on a cloud in a miyazaki style. https://labs.openai.com/, 2023. Accessed: 2023-04-01.

I. Dayan, H. R. Roth, A. Zhong, A. Harouni, A. Gentili, A. Z. Abidin, A. Liu, A. B. Costa, B. J. Wood, C.-S. Tsai, et al. Federated learning for predicting clinical outcomes in patients with covid-19. *Nature medicine*, 27(10):1735–1743, 2021.

J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, M. Mao, M. Ranzato, A. Senior, P. Tucker, K. Yang, Q. Le, and A. Ng. Large Scale Distributed Deep Networks. *Advances in Neural Information Processing Systems*, 25, 2012.

A. Defazio, F. Bach, and S. Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. *Advances in neural information processing systems*, 27, 2014.

A. Defossez and F. Bach. Averaged Least-Mean-Squares: Bias-Variance Trade-offs and Optimal Sampling Distributions. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, pages 205–213. PMLR, Feb. 2015. ISSN: 1938-7228.

B. Delyon. General results on the convergence of stochastic algorithms. *IEEE Transactions on Automatic Control*, 41(9):1245–1255, 1996.

Y. Deng, M. M. Kamani, and M. Mahdavi. Adaptive Personalized Federated Learning, Nov. 2020. arXiv:2003.13461 [cs, stat].

A. Dieuleveut and F. Bach. Nonparametric stochastic approximation with large step-sizes. *Ann. Statist.*, 44(4):1363–1399, 2016. ISSN 0090-5364. doi: 10.1214/15-AOS1391.

A. Dieuleveut, N. Flammarion, and F. Bach. Harder, better, faster, stronger convergence rates for least-squares regression. *The Journal of Machine Learning Research*, 18(1):3520–3570, 2017. Publisher: JMLR. org.

A. Dieuleveut, A. Durmus, and F. Bach. Bridging the gap between constant step size stochastic gradient descent and markov chains. *Ann. Statist.*, 48(3):1348–1382, 06 2020. doi: 10.1214/19-AOS1850. URL https://doi.org/10.1214/19-AOS1850.

A. Dieuleveut, G. Fort, E. Moulines, and G. Robin. Federated-em with heterogeneity mitigation and variance reduction. *Advances in Neural Information Processing Systems*, 34:29553–29566, 2021.

J. O. du Terrail, A. Leopold, C. Joly, C. Beguier, M. Andreux, C. Maussion, B. Schmauch, E. W. Tramel, E. Bendjebbar, M. Zaslavskiy, et al. Collaborative federated learning behind hospitals' firewalls for predicting histological response to neoadjuvant chemotherapy in triple-negative breast cancer. *medRxiv*, pages 2021–10, 2021.

J. O. du Terrail, S.-S. Ayed, E. Cyffers, F. Grimberg, C. He, R. Loeb, P. Mangold, T. Marchand, O. Marfoq, E. Mushtaq, B. Muzellec, C. Philippenko, S. Silva, M. Teleńczuk, S. Albarqouni, S. Avestimehr, A. Bellet, A. Dieuleveut, M. Jaggi, S. P. Karimireddy, M. Lorenzi, G. Neglia, M. Tommasi, and M. Andreux. FLamby: Datasets and benchmarks for cross-silo federated learning in realistic healthcare settings. In *Thirty-sixth conference on neural information processing systems datasets and benchmarks track*, 2022.

J. C. Duchi, P. L. Bartlett, and M. J. Wainwright. Randomized Smoothing for Stochastic Optimization. *SIAM Journal on Optimization*, 22(2):674–701, Jan. 2012. ISSN 1052-6234. doi: 10.1137/110831659. Publisher: Society for Industrial and Applied Mathematics.

R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern classification and scene analysis*, volume 3. Wiley New York, 1973.

M. Duflo. *Random iterative models*, volume 34. Springer Science & Business Media, 1997.

H. Eichner, T. Koren, H. B. McMahan, N. Srebro, and K. Talwar. Semi-Cyclic Stochastic Gradient Descent. Apr. 2019.

P. Elias. Universal codeword sets and representations of the integers, Sept. 1975.

A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun, and J. Dean. A guide to deep learning in healthcare. *Nature medicine*, 25(1):24–29, 2019.

M. Even, L. Massoulié, and K. Scaman. Sample Optimality and All-for-all Strategies in Personalized Federated and Collaborative Learning, Feb. 2022. arXiv:2201.13097 [math].

A. Fallah, A. Mokhtari, and A. Ozdaglar. Personalized Federated Learning: A Meta-Learning Approach, Oct. 2020. arXiv:2002.07948 [cs, math, stat].

I. Fatkhullin, I. Sokolov, E. Gorbunov, Z. Li, and P. Richtárik. EF21 with Bells & Whistles: Practical Algorithmic Extensions of Modern Error Feedback, Oct. 2021. arXiv:2110.03294 [cs, math].

P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–1645, 2009.

N. Flammarion and F. Bach. From averaging to acceleration, there is only a step-size. In P. Grünwald, E. Hazan, and S. Kale, editors, *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 658–695, Paris, France, 03–06 Jul 2015. PMLR. URL https://proceedings.mlr.press/v40/Flammarion15.html.

Y. Fraboni, R. Vidal, L. Kameni, and M. Lorenzi. Clustered Sampling: Low-Variance and Improved Representativity for Clients Selection in Federated Learning. In *Proceedings of the 38th International Conference on Machine Learning*, pages 3407–3416. PMLR, July 2021a. ISSN: 2640-3498.

Y. Fraboni, R. Vidal, and M. Lorenzi. Free-rider attacks on model aggregation in federated learning. In *International Conference on Artificial Intelligence and Statistics*, pages 1846–1854. PMLR, 2021b.

Y. Fraboni, R. Vidal, L. Kameni, and M. Lorenzi. A General Theory for Client Sampling in Federated Learning. In *International Workshop on Trustworthy Federated Learning in Conjunction with IJCAI 2022 (FL-IJCAI'22)*, Vienna, Austria, July 2022.

M. I. Freidlin and A. D. Wentzell. Random perturbations. In *Random perturbations of dynamical systems*, pages 15–43. Springer, 1998.

Y. Freund and R. E. Schapire. Experiments with a new boosting algorithm. In *Machine learning: Proceedings of the thirteenth international conference*, pages 148–156. Morgan Kaufmann, 1996.

S. Gadat and I. Gavra. Asymptotic study of stochastic adaptive algorithms in non-convex landscape. *The Journal of Machine Learning Research*, 23(1):10357–10410, 2022.

S. Gadat and F. Panloup. Optimal non-asymptotic analysis of the Ruppert–Polyak averaging stochastic algorithm. *Stochastic Processes and their Applications*, 156:312–348, Feb. 2023. ISSN 0304-4149. doi: 10.1016/j.spa.2022.11.012.

V. Gandikota, D. Kane, R. K. Maity, and A. Mazumdar. vqsgd: Vector quantized stochastic gradient descent. In *International Conference on Artificial Intelligence and Statistics*, pages 2197–2205. PMLR, 2021.

H. Gao, A. Xu, and H. Huang. On the convergence of communication-efficient local sgd for federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 7510–7518, 2021.

C. F. Gauss. *Theoria motus corporum coelestium in sectionibus conicis solem ambientium*, volume 7. FA Perthes, 1809.

A. Gersho and R. M. Gray. *Vector quantization and signal compression*, volume 159. Springer Science & Business Media, 2012.

H. S. Ghadikolaei, S. Stich, and M. Jaggi. Lena: Communication-efficient distributed learning with self-triggered gradient uploads. In *International Conference on Artificial Intelligence and Statistics*, pages 3943–3951. PMLR, 2021.

H. Ghayvat, S. Mukhopadhyay, X. Gui, and N. Suryadevara. Wsn-and iot-based smart homes and their extension to smart buildings. *Sensors*, 15(5):10350–10379, 2015.

R. Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

E. Gorbunov, F. Hanzely, and P. Richtárik. A unified theory of sgd: Variance reduction, sampling, quantization and coordinate descent. In *International Conference on Artificial Intelligence and Statistics*, pages 680–690. PMLR, 2020a.

E. Gorbunov, D. Kovalev, D. Makarenko, and P. Richtarik. Linearly Converging Error Compensated SGD. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 20889–20900. Curran Associates, Inc., 2020b.

E. Gorbunov, K. P. Burlachenko, Z. Li, and P. Richtárik. Marina: Faster non-convex distributed learning with compression. In *International Conference on Machine Learning*, pages 3788–3798. PMLR, 2021.

R. M. Gower, N. Loizou, X. Qian, A. Sailanbayev, E. Shulgin, and P. Richtárik. SGD: General Analysis and Improved Rates. In *International Conference on Machine Learning*, pages 5200–5209. PMLR, May 2019. ISSN: 2640-3498.

F. Grimberg, M.-A. Hartley, M. Jaggi, and S. P. Karimireddy. Weight erosion: An update aggregation scheme for personalized collaborative machine learning. In *Domain adaptation and representation transfer, and distributed and collaborative learning*, pages 160–169. Springer, 2020.

D. Grishchenko, F. Iutzeler, J. Malick, and M.-R. Amini. Distributed learning with sparse communications by identification. *SIAM Journal on Mathematics of Data Science*, 3(2):715–735, 2021.

K. Gruntkowska, A. Tyurin, and P. Richtárik. EF21-P and Friends: Improved Theoretical Communication Complexity for Distributed Optimization with Bidirectional Compression, Sept. 2022. arXiv:2209.15218 [cs, math].

V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama*, 316(22): 2402–2410, 2016.

L. Györfi and H. Walk. On the averaged stochastic approximation for linear regression. *SIAM Journal on Control and Optimization*, 34(1):31–61, 1996.

F. Haddadpour, M. M. Kamani, A. Mokhtari, and M. Mahdavi. Federated learning with compression: Unified analysis and sharp guarantees. In A. Banerjee and K. Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 2350–2358. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/haddadpour21a.html.

K. Hamidieh. A data-driven statistical model for predicting the critical temperature of a superconductor. *Computational Materials Science*, 154:346–354, Nov. 2018. ISSN 0927-0256. doi: 10.1016/j.commatsci.2018.07.052.

I. E. K. Harrane, R. Flamary, and C. Richard. On reducing the communication cost of the diffusion lms algorithm. *IEEE Transactions on Signal and Information Processing over Networks*, 5(1): 100–112, 2018.

T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

P. Henderson, J. Hu, J. Romoff, E. Brunskill, D. Jurafsky, and J. Pineau. Towards the systematic reporting of the energy and carbon footprints of machine learning. *Journal of Machine Learning Research*, 21(248):1–43, 2020.

S. Horvath, C.-Y. Ho, L. Horvath, A. N. Sahu, M. Canini, and P. Richtárik. Natural compression for distributed deep learning. In *Mathematical and Scientific Machine Learning*, pages 129–141. PMLR, 2022.

S. Horváth, D. Kovalev, K. Mishchenko, P. Richtárik, and S. Stich. Stochastic distributed learning with gradient quantization and double-variance reduction. *Optimization Methods and Software*, pages 1–16, 2022.

S. Horváth and P. Richtárik. A Better Alternative to Error Feedback for Communication-Efficient Distributed Learning. *arXiv:2006.11077 [cs, stat]*, June 2020. arXiv: 2006.11077.

S. Horváth, D. Kovalev, K. Mishchenko, S. Stich, and P. Richtárik. Stochastic Distributed Learning with Gradient Quantization and Variance Reduction. *arXiv:1904.05115 [math]*, Apr. 2019. arXiv: 1904.05115.

D. Hsu, S. M. Kakade, and T. Zhang. Random Design Analysis of Ridge Regression. In *Proceedings of the 25th Annual Conference on Learning Theory*, pages 9.1–9.24. JMLR Workshop and Conference Proceedings, June 2012. ISSN: 1938-7228.

T.-M. H. Hsu, H. Qi, and M. Brown. Measuring the Effects of Non-Identical Data Distribution for Federated Visual Classification, Sept. 2019. arXiv:1909.06335 [cs, stat].

Y. Huang, Y. Cheng, A. Bapna, O. Firat, D. Chen, M. Chen, H. Lee, J. Ngiam, Q. V. Le, Y. Wu, and z. Chen. GPipe: Efficient Training of Giant Neural Networks using Pipeline Parallelism. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d. Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Z. Huang, H. Chen, and D. Zeng. Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering. *ACM Transactions on Information Systems (TOIS)*, 22(1): 116–142, 2004.

E. B. Hunt, M. Marin, and P. J. Stone. *Experiments in Induction*. Academic Press, 1966.

R. Hussain and S. Zeadally. Autonomous cars: Research results, issues, and future challenges. *IEEE Communications Surveys & Tutorials*, 21(2):1275–1313, 2018.

S. G. Index. Speedtest Global Index – Monthly comparisons of internet speeds from around the world, 2020.

N. Ivkin, D. Rothchild, E. Ullah, V. Braverman, I. Stoica, and R. Arora. Communication-efficient Distributed SGD with Sketching. *Advances in Neural Information Processing Systems*, 32:13144–13154, 2019.

P. Jain, S. M. Kakade, R. Kidambi, P. Netrapalli, and A. Sidford. Accelerating Stochastic Gradient Descent for Least Squares Regression. In *Proceedings of the 31st Conference On Learning Theory*, pages 545–604. PMLR, July 2018a. ISSN: 2640-3498.

P. Jain, S. M. Kakade, R. Kidambi, P. Netrapalli, and A. Sidford. Parallelizing Stochastic Gradient Descent for Least Squares Regression: Mini-batching, Averaging, and Model Misspecification. *Journal of Machine Learning Research*, 18(223):1–42, 2018b. ISSN 1533-7928.

D. Jhunjhunwala, P. Sharma, A. Nagarkatti, and G. Joshi. Fedvarp: Tackling the variance due to partial client participation in federated learning. In *Uncertainty in Artificial Intelligence*, pages 906–916. PMLR, 2022.

R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. *Advances in neural information processing systems*, 26, 2013.

P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, R. G. L. D'Oliveira, S. E. Rouayheb, D. Evans, J. Gardner, Z. Garrett, A. Gascón, B. Ghazi, P. B. Gibbons, M. Gruteser, Z. Harchaoui, C. He, L. He, Z. Huo, B. Hutchinson, J. Hsu, M. Jaggi, T. Javidi, G. Joshi, M. Khodak, J. Konečný, A. Korolova, F. Koushanfar, S. Koyejo, T. Lepoint, Y. Liu, P. Mittal, M. Mohri, R. Nock, A. Özgür, R. Pagh, M. Raykova, H. Qi, D. Ramage, R. Raskar, D. Song, W. Song, S. U. Stich, Z. Sun, A. T. Suresh, F. Tramèr, P. Vepakomma, J. Wang, L. Xiong, Z. Xu, Q. Yang, F. X. Yu, H. Yu, and S. Zhao. Advances and Open Problems in Federated Learning. *arXiv:1912.04977 [cs, stat]*, Dec. 2019. arXiv: 1912.04977.

T. Kanade, C. Thorpe, and W. Whittaker. Autonomous land vehicle project at cmu. In *Proceedings of the 1986 ACM Fourteenth Annual Conference on Computer Science*, CSC '86, page 71–80, New York, NY, USA, 1986. Association for Computing Machinery. ISBN 0897911776. doi: 10.1145/324634.325197. URL https://doi.org/10.1145/324634.325197.

S. P. Karimireddy, Q. Rebjock, S. Stich, and M. Jaggi. Error Feedback Fixes SignSGD and other Gradient Compression Schemes. In *International Conference on Machine Learning*, pages 3252–3261. PMLR, May 2019. ISSN: 2640-3498.

S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, pages 5132–5143. PMLR, 2020.

S. P. Karimireddy, M. Jaggi, S. Kale, M. Mohri, S. Reddi, S. U. Stich, and A. T. Suresh. Breaking the centralized barrier for cross-device federated learning. *Advances in Neural Information Processing Systems*, 34:28663–28676, 2021.

T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

S. Khirirat, H. R. Feyzmahdavian, and M. Johansson. Distributed learning with compressed gradients. *arXiv preprint arXiv:1806.06573*, 2018.

S. Khirirat, S. Magnússon, and M. Johansson. Compressed gradient methods with hessian-aided error compensation. *IEEE Transactions on Signal Processing*, 69:998–1011, 2020a.

S. Khirirat, S. Magnússon, A. Aytekin, and M. Johansson. Communication Efficient Sparsification for Large Scale Machine Learning. *arXiv:2003.06377 [math, stat]*, Mar. 2020b. arXiv: 2003.06377.

A. Koloskova, S. Stich, and M. Jaggi. Decentralized Stochastic Optimization and Gossip Algorithms with Compressed Communication. In *International Conference on Machine Learning*, pages 3478–3487. PMLR, May 2019. ISSN: 2640-3498.

A. Koloskova, N. Loizou, S. Boreiri, M. Jaggi, and S. Stich. A unified theory of decentralized sgd with changing topology and local updates. In *International Conference on Machine Learning*, pages 5381–5393. PMLR, 2020.

V. R. Konda and J. N. Tsitsiklis. Linear stochastic approximation driven by slowly varying markov chains. *Systems & control letters*, 50(2):95–102, 2003.

J. Konečný, H. B. McMahan, F. X. Yu, P. Richtarik, A. T. Suresh, and D. Bacon. Federated Learning: Strategies for Improving Communication Efficiency. In *NIPS Workshop on Private Multi-Party Machine Learning*, 2016.

Y. Koren. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 426–434, 2008.

Y. Koren. Collaborative filtering with temporal dynamics. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 447–456, 2009.

D. Kovalev, E. Gasanov, P. Richtárik, and A. Gasnikov. Lower Bounds and Optimal Algorithms for Smooth and Strongly Convex Decentralized Optimization Over Time-Varying Networks. *arXiv:2106.04469 [cs, math]*, June 2021. arXiv: 2106.04469.

A. Krizhevsky, G. Hinton, and others. Learning multiple layers of features from tiny images. 2009. Publisher: Citeseer.

S. Kumar, M. Mohri, and A. Talwalkar. Ensemble Nystrom Method. In *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc., 2009.

H. J. Kushner and G. Yin. Stochastic approximation and recursive algorithms and applications. 2003.

Y. Laguel, K. Pillutla, J. Malick, and Z. Harchaoui. Device Heterogeneity in Federated Learning: A Superquantile Approach. *arXiv:2002.11223 [cs, math, stat]*, Feb. 2020. arXiv: 2002.11223.

Y. Laguel, K. Pillutla, J. Malick, and Z. Harchaoui. A superquantile approach to federated learning with heterogeneous devices. In *2021 55th Annual Conference on Information Sciences and Systems (CISS)*, pages 1–6. IEEE, 2021.

L. Lannelongue, J. Grealey, and M. Inouye. Green algorithms: Quantifying the carbon footprint of computation. *Advanced Science*, page 2100707, 2021. Publisher: Wiley Online Library.

P. S. Laplace. *Théorie analytique des probabilités*, volume 7. Courcier, 1820.

L. Leconte, A. Dieuleveut, E. Oyallon, E. Moulines, and G. Pages. DoStoVoQ: Doubly Stochastic Voronoi Vector Quantization SGD for Federated Learning. May 2021.

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, Nov. 1998. ISSN 1558-2256. doi: 10.1109/5.726791. Conference Name: Proceedings of the IEEE.

Y. LeCun, P. Haffner, L. Bottou, and Y. Bengio. Object Recognition with Gradient-Based Learning. In D. A. Forsyth, J. L. Mundy, V. di Gesú, and R. Cipolla, editors, *Shape, Contour and Grouping in Computer Vision*, Lecture Notes in Computer Science, pages 319–345. Springer, Berlin, Heidelberg, 1999. ISBN 978-3-540-46805-9. doi: 10.1007/3-540-46805-6_19.

Y. LeCun, C. Cortes, and C. Burges. MNIST handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

A. M. Legendre. *Nouvelles méthodes pour la détermination des orbites des comètes*. Firmin Didot, Libraire pour les mathematiques, 1806.

R. Leluc and F. Portier. Sgd with coordinate sampling: Theory and practice. *Journal of Machine Learning Research*, 23(342):1–47, 2022.

M. Li, D. G. Andersen, J. W. Park, A. J. Smola, A. Ahmed, V. Josifovski, J. Long, E. J. Shekita, and B.-Y. Su. Scaling distributed machine learning with the parameter server. In *Proceedings of the 11th USENIX conference on Operating Systems Design and Implementation*, OSDI'14, pages 583–598, USA, Oct. 2014. USENIX Association. ISBN 978-1-931971-16-4.

P. Li, T. J. Hastie, and K. W. Church. Very sparse random projections. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 287–296, 2006.

Q. Li, Y. Diao, Q. Chen, and B. He. Federated Learning on Non-IID Data Silos: An Experimental Study. In *2022 IEEE 38th International Conference on Data Engineering (ICDE)*, pages 965–978, May 2022a. doi: 10.1109/ICDE53745.2022.00077. ISSN: 2375-026X.

S. Li. Concise formulas for the area and volume of a hyperspherical cap. *Asian Journal of Mathematics & Statistics*, 4(1):66–70, 2010.

T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith. Federated Optimization in Heterogeneous Networks. *arXiv:1812.06127 [cs, stat]*, Sept. 2019a. arXiv: 1812.06127.

T. Li, A. K. Sahu, A. Talwalkar, and V. Smith. Federated learning: Challenges, methods, and future directions. *IEEE signal processing magazine*, 37(3):50–60, 2020a.

T. Li, S. Hu, A. Beirami, and V. Smith. Ditto: Fair and robust federated learning through personalization. In *International conference on machine learning*, pages 6357–6368, 2021. tex.organization: PMLR.

X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang. On the Convergence of FedAvg on Non-IID Data. Oct. 2019b.

Z. Li and P. Richtárik. CANITA: Faster Rates for Distributed Convex Optimization with Communication Compression. *arXiv:2107.09461 [cs, math]*, July 2021. arXiv: 2107.09461.

Z. Li, D. Kovalev, X. Qian, and P. Richtarik. Acceleration for Compressed Gradient Descent in Distributed and Federated Optimization. In *International Conference on Machine Learning*, pages 5895–5904. PMLR, Nov. 2020b. ISSN: 2640-3498.

Z. Li, H. Zhao, B. Li, and Y. Chi. Soteriafl: A unified framework for private federated learning with communication compression. *arXiv preprint arXiv:2206.09888*, 2022b.

J. Lin and L. Rosasco. Optimal Rates for Learning with Nyström Stochastic Gradient Methods. *arXiv preprint arXiv:1710.07797*, 2017.

T. Lin, S. U. Stich, K. K. Patel, and M. Jaggi. Don't use large mini-batches, use local sgd. *arXiv preprint arXiv:1808.07217*, 2018.

X. Liu, Y. Li, J. Tang, and M. Yan. A Double Residual Compression Algorithm for Efficient Distributed Learning. In *International Conference on Artificial Intelligence and Statistics*, pages 133–143, June 2020. ISSN: 1938-7228 Section: Machine Learning.

L. Ljung. Analysis of recursive stochastic algorithms. *IEEE transactions on automatic control*, 22 (4):551–575, 1977.

L. Ljung and T. Söderström. *Theory and practice of recursive identification*. MIT press, 1983.

L. Lü, M. Medo, C. H. Yeung, Y.-C. Zhang, Z.-K. Zhang, and T. Zhou. Recommender systems. *Physics reports*, 519(1):1–49, 2012.

B. Luo, W. Xiao, S. Wang, J. Huang, and L. Tassiulas. Tackling system and statistical heterogeneity for federated learning with adaptive client sampling. In *IEEE INFOCOM 2022-IEEE conference on computer communications*, pages 1739–1748. IEEE, 2022.

L. v. d. Maaten and G. Hinton. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008. ISSN ISSN 1533-7928.

O. Macchi. *Adaptative processing: the least mean squares approach with applications in transmission*, volume 71. New York: John Wiley & Sons, Ltd, 1995.

G. Malinovskiy, D. Kovalev, E. Gasanov, L. Condat, and P. Richtarik. From local sgd to local fixed-point methods for federated learning. In *International Conference on Machine Learning*, pages 6692–6701. PMLR, 2020.

H. Mania, X. Pan, D. Papailiopoulos, B. Recht, K. Ramchandran, and M. I. Jordan. Perturbed Iterate Analysis for Asynchronous Stochastic Optimization. *arXiv:1507.06970 [cs, math, stat]*, Mar. 2016. arXiv: 1507.06970.

Y. Mansour, M. Mohri, J. Ro, and A. T. Suresh. Three approaches for personalization with applications to federated learning. *arXiv preprint arXiv:2002.10619*, 2020.

O. Marfoq, G. Neglia, R. Vidal, and L. Kameni. Personalized federated learning through local memorization. In *International Conference on Machine Learning*, pages 15070–15092. PMLR, 2022.

P. Mayekar and H. Tyagi. RATQ: A Universal Fixed-Length Quantizer for Stochastic Optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 1399–1409. PMLR, June 2020. ISSN: 2640-3498.

W. S. McCulloch and W. Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5:115–133, 1943.

B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y. Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Artificial Intelligence and Statistics*, pages 1273–1282. PMLR, Apr. 2017. ISSN: 2640-3498.

S. Mei and A. Montanari. The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv:1908.05355 [math, stat]*, Oct. 2019. arXiv: 1908.05355.

S. Meyn and R. Tweedie. *Markov chains and stochastic stability*. Cambridge University Press, New York, NY, USA, 2 edition, 2009. ISBN 0-521-73182-8 978-0-521-73182-9.

R. Miotto, L. Li, B. A. Kidd, and J. T. Dudley. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Scientific reports*, 6(1):1–10, 2016.

M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

K. Mishchenko, F. Iutzeler, J. Malick, and M.-R. Amini. A delay-tolerant proximal-gradient algorithm for distributed learning. In *International Conference on Machine Learning*, pages 3587–3595. PMLR, 2018.

K. Mishchenko, E. Gorbunov, M. Takáč, and P. Richtárik. Distributed Learning with Compressed Gradient Differences. *arXiv:1901.09269 [cs, math, stat]*, June 2019. arXiv: 1901.09269.

T. Mitchell. *Machine Learning*. McGraw Hill, 1997.

A. Mitra, R. Jaafar, G. J. Pappas, and H. Hassani. Linear convergence in federated learning: Tackling client heterogeneity and sparse gradients. *Advances in Neural Information Processing Systems*, 34: 14606–14619, 2021.

C. Molnar. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. 2018.

B. Morvaj, L. Lugaric, and S. Krajcar. Demonstrating smart buildings and smart grid features in a smart energy city. In *Proceedings of the 2011 3rd international youth conference on energetics (IYCE)*, pages 1–8. IEEE, 2011.

E. Moulines and F. Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. *Advances in neural information processing systems*, 24, 2011.

N. Muecke, G. Neu, and L. Rosasco. Beating SGD Saturation with Tail-Averaging and Minibatching. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Applied Optimization. Springer US, 2004. ISBN 978-1-4020-7553-7. doi: 10.1007/978-1-4419-8853-9.

Y. Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.

G. Neu and L. Rosasco. Iterate averaging as regularization for stochastic gradient descent. In *Conference On Learning Theory*, pages 3222–3242. PMLR, 2018.

J. Nutini, M. Schmidt, I. Laradji, M. Friedlander, and H. Koepke. Coordinate descent converges faster with the gauss-southwell rule than random selection. In *International Conference on Machine Learning*, pages 1632–1641. PMLR, 2015.

S. Pati, U. Baid, M. Zenk, B. Edwards, M. Sheller, G. A. Reina, P. Foley, A. Gruzdev, J. Martin, S. Albarqouni, et al. The federated tumor segmentation (fets) challenge. *arXiv preprint arXiv:2105.05874*, 2021.

C. Philippenko and A. Dieuleveut. Artemis: tight convergence guarantees for bidirectional compression in Federated Learning. *arXiv:2006.14591 [cs, stat]*, Nov. 2020. arXiv: 2006.14591.

C. Philippenko and A. Dieuleveut. Preserved central model for faster bidirectional compression in distributed settings. *Advances in Neural Information Processing Systems*, 34, 2021.

C. Philippenko and A. Dieuleveut. Convergence rates for distributed, compressed and averaged least-squares regression: application to federated learning. *arXiv[cs, stat]*, 2023.

K. Pillutla, Y. Laguel, J. Malick, and Z. Harchaoui. Tackling distribution shifts in federated learning with superquantile aggregation. In *NeurIPS 2022 Workshop on Distribution Shifts (DistShift)*, 2022a.

K. Pillutla, K. Malik, A.-R. Mohamed, M. Rabbat, M. Sanjabi, and L. Xiao. Federated learning with partial model personalization. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 17716–17758. PMLR, 17–23 Jul 2022b. URL https://proceedings.mlr.press/v162/pillutla22a.html.

A. P. Plageras, K. E. Psannis, C. Stergiou, H. Wang, and B. B. Gupta. Efficient iot-based sensor big data collection–processing and analysis in smart buildings. *Future Generation Computer Systems*, 82:349–357, 2018.

B. Polyak and A. Juditsky. Acceleration of Stochastic Approximation by Averaging. *SIAM Journal on Control and Optimization*, 30:838–855, July 1992. doi: 10.1137/0330046.

X. Qian, P. Richtárik, and T. Zhang. Error compensated distributed sgd can be accelerated. *Advances in Neural Information Processing Systems*, 34:30401–30413, 2021.

M. G. Rabbat and R. D. Nowak. Quantized incremental algorithms for distributed optimization. *IEEE Journal on Selected Areas in Communications*, 23(4):798–808, 2005.

A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

A. Rahimi and B. Recht. Random Features for Large-Scale Kernel Machines. In *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2008.

A. Rajkomar, J. Dean, and I. Kohane. Machine learning in medicine. *New England Journal of Medicine*, 380(14):1347–1358, 2019.

A. Rakhlin, O. Shamir, and K. Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. *ICML*, 2012.

A. Ramezani-Kebrya, F. Faghri, I. Markov, V. Aksenov, D. Alistarh, and D. M. Roy. Nuqsgd: Provably communication-efficient data-parallel sgd via nonuniform quantization. *The Journal of Machine Learning Research*, 22(1):5074–5116, 2021.

J. Redmon and A. Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017.

J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

A. Reisizadeh, A. Mokhtari, H. Hassani, A. Jadbabaie, and R. Pedarsani. FedPAQ: A Communication-Efficient Federated Learning Method with Periodic Averaging and Quantization. In *International Conference on Artificial Intelligence and Statistics*, pages 2021–2031. PMLR, June 2020. ISSN: 2640-3498.

S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

P. Resnick and H. R. Varian. Recommender systems. *Communications of the ACM*, 40(3):56–58, 1997.

M. T. Ribeiro, S. Singh, and C. Guestrin. "why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

P. Richtarik, I. Sokolov, and I. Fatkhullin. EF21: A New, Simpler, Theoretically Better, and Practically Faster Error Feedback. In *Advances in Neural Information Processing Systems*, volume 34, pages 4384–4396. Curran Associates, Inc., 2021.

P. Richtárik, I. Sokolov, E. Gasanov, I. Fatkhullin, Z. Li, and E. Gorbunov. 3pc: Three point compressors for communication-efficient distributed training and a better theory for lazy aggregation. In *International Conference on Machine Learning*, pages 18596–18648. PMLR, 2022.

N. Rieke, J. Hancox, W. Li, F. Milletari, H. R. Roth, S. Albarqouni, S. Bakas, M. N. Galtier, B. A. Landman, K. Maier-Hein, et al. The future of digital health with federated learning. *NPJ digital medicine*, 3(1):119, 2020.

H. Robbins and S. Monro. A Stochastic Approximation Method. *Annals of Mathematical Statistics*, 22(3):400–407, Sept. 1951. ISSN 0003-4851, 2168-8990. doi: 10.1214/aoms/1177729586. Number: 3 Publisher: Institute of Mathematical Statistics.

A. Rodio, F. Faticanti, O. Marfoq, G. Neglia, and E. Leonardi. Federated learning under heterogeneous and correlated client availability. *arXiv preprint arXiv:2301.04632*, 2023.

F. Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

A. Rudi and L. Rosasco. Generalization Properties of Learning with Random Features. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

A. Rudi, R. Camoriano, and L. Rosasco. Less is More: Nyström Computational Regularization. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

A. Rudi, L. Carratino, and L. Rosasco. FALKON: An Optimal Large Scale Kernel Method. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

M. Safaryan, R. Islamov, X. Qian, and P. Richtárik. Fednl: Making newton-type methods applicable to federated learning. *arXiv preprint arXiv:2106.02969*, 2021.

T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

F. Sattler, S. Wiedemann, K.-R. Müller, and W. Samek. Robust and Communication-Efficient Federated Learning From Non-i.i.d. Data. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–14, 2019. ISSN 2162-2388. doi: 10.1109/TNNLS.2019.2944481. Conference Name: IEEE Transactions on Neural Networks and Learning Systems.

K. Scaman, F. Bach, S. Bubeck, L. Massoulié, and Y. T. Lee. Optimal Algorithms for Non-Smooth Distributed Optimization in Networks. *Advances in Neural Information Processing Systems*, 31: 2740–2749, 2018.

M. Schmidt and N. L. Roux. Fast Convergence of Stochastic Gradient Descent under a Strong Growth Condition. *arXiv:1308.6370 [math]*, Aug. 2013. arXiv: 1308.6370.

M. Schmidt, N. Le Roux, and F. Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1):83–112, Mar. 2017. ISSN 1436-4646. doi: 10.1007/s10107-016-1030-6.

F. Seide and A. Agarwal. CNTK: Microsoft's Open-Source Deep-Learning Toolkit. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 2135, New York, NY, USA, Aug. 2016. Association for Computing Machinery. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2945397.

F. Seide, H. Fu, J. Droppo, G. Li, and D. Yu. 1-bit stochastic gradient descent and its application to data-parallel distributed training of speech dnns. In *Fifteenth Annual Conference of the International Speech Communication Association*. Citeseer, 2014.

M. J. Sheller, B. Edwards, G. A. Reina, J. Martin, S. Pati, A. Kotrotsou, M. Milchenko, W. Xu, D. Marcus, R. R. Colen, et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Scientific reports*, 10(1):1–12, 2020.

B. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi. Deep ehr: a survey of recent advances in deep learning techniques for electronic health record (ehr) analysis. *IEEE journal of biomedical and health informatics*, 22(5):1589–1604, 2017.

P. Sidiropoulos. N-sphere chord length distribution. *arXiv preprint arXiv:1411.5639*, 2014.

K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition, Apr. 2015. arXiv:1409.1556 [cs].

V. Smith, C.-K. Chiang, M. Sanjabi, and A. S. Talwalkar. Federated multi-task learning. *Advances in neural information processing systems*, 30, 2017.

S. U. Stich. Local SGD Converges Fast and Communicates Little. *arXiv:1805.09767 [cs, math]*, May 2019. arXiv: 1805.09767.

S. U. Stich and S. P. Karimireddy. The error-feedback framework: Better rates for SGD with delayed gradients and compressed updates. *Journal of Machine Learning Research*, 21:1–36, 2020.

S. U. Stich, J.-B. Cordonnier, and M. Jaggi. Sparsified SGD with Memory. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 4447–4458. Curran Associates, Inc., 2018.

N. Strom. Scalable distributed DNN training using commodity GPU cloud computing. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2018.

C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

H. Tang, X. Lian, M. Yan, C. Zhang, and J. Liu. D^2: Decentralized Training over Decentralized Data. In *International Conference on Machine Learning*, pages 4848–4856. PMLR, July 2018. ISSN: 2640-3498.

H. Tang, C. Yu, X. Lian, T. Zhang, and J. Liu. DoubleSqueeze: Parallel Stochastic Gradient Descent with Double-pass Error-Compensated Compression. In *International Conference on Machine Learning*, pages 6155–6165. PMLR, May 2019. ISSN: 2640-3498.

D. S. W. Ting, C. Y.-L. Cheung, G. Lim, G. S. W. Tan, N. D. Quang, A. Gan, H. Hamzah, R. Garcia-Franco, I. Y. San Yeo, S. Y. Lee, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *Jama*, 318(22):2211–2223, 2017.

P. Vanhaesebrouck, A. Bellet, and M. Tommasi. Decentralized Collaborative Learning of Personalized Models over Networks. In *Artificial Intelligence and Statistics*, pages 509–517. PMLR, Apr. 2017. ISSN: 2640-3498.

V. Vapnik. Estimation of dependences based on empirical data: Springer series in statistics (springer series in statistics), 1982.

V. Vapnik. *The nature of statistical learning theory*. Springer science & business media, 1999.

A. Varre and N. Flammarion. Accelerated SGD for Non-Strongly-Convex Least Squares. In *Proceedings of Thirty Fifth Conference on Learning Theory*, pages 2062–2126. PMLR, June 2022. ISSN: 2640-3498.

S. S. Vempala. *The random projection method*, volume 65. American Mathematical Soc., 2005.

C. Villani. *Optimal transport : old and new*. Grundlehren der mathematischen wissenschaften. Springer, Berlin, 2009. ISBN 978-3-540-71049-3.

M. Vono, V. Plassier, A. Durmus, A. Dieuleveut, and E. Moulines. Qlsd: Quantised langevin stochastic dynamics for bayesian federated learning. In *International Conference on Artificial Intelligence and Statistics*, pages 6459–6500. PMLR, 2022.

H. Wang and J. Xu. Friends to help: Saving federated learning from client dropout. *arXiv preprint arXiv:2205.13222*, 2022.

H. Wang, S. Marella, and J. Anderson. Fedadmm: A federated primal-dual algorithm allowing partial participation. In *2022 IEEE 61st Conference on Decision and Control (CDC)*, pages 287–294. IEEE, 2022.

J. Wangni, J. Wang, J. Liu, and T. Zhang. Gradient Sparsification for Communication-Efficient Distributed Optimization. *Advances in Neural Information Processing Systems*, 31:1299–1309, 2018.

W. Wen, C. Xu, F. Yan, C. Wu, Y. Wang, Y. Chen, and H. Li. TernGrad: Ternary Gradients to Reduce Communication in Distributed Deep Learning. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 1509–1519. Curran Associates, Inc., 2017.

J. Wu, W. Huang, J. Huang, and T. Zhang. Error Compensated Quantized SGD and its Applications to Large-scale Distributed Optimization. In *International Conference on Machine Learning*, pages 5325–5333. PMLR, July 2018. ISSN: 2640-3498.

H. Xiao, K. Rasul, and R. Vollgraf. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. *arXiv:1708.07747 [cs, stat]*, Sept. 2017. arXiv: 1708.07747.

A. Xu, Z. Huo, and H. Huang. Optimal gradient quantization condition for communication-efficient distributed training. *arXiv preprint arXiv:2002.11082*, 2020.

A. Xu, Z. Huo, and H. Huang. Step-ahead error feedback for distributed training with compressed gradient. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10478–10486, 2021.

H. Yang, M. Fang, and J. Liu. Achieving linear speedup with partial worker participation in non-iid federated learning. *arXiv preprint arXiv:2101.11203*, 2021.

YoloV2. Object detection on *James Bond* (Skyfall). https://youtu.be/VOC3huqHrss, 2023. Accessed: 2023-04-01.

Y. Yu, J. Wu, and L. Huang. Double Quantization for Communication-Efficient Distributed Optimization. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d. Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 4438–4449. Curran Associates, Inc., 2019.

C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. In *5th international conference on learning representations, ICLR 2017, toulon, france, april 24-26, 2017, conference track proceedings*. OpenReview.net, 2017.

M. Zhang, K. Sapra, S. Fidler, S. Yeung, and J. M. Alvarez. Personalized federated learning with first order model optimization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=ehJqJQk9cw.

S. Zhang, L. Yao, A. Sun, and Y. Tay. Deep learning based recommender system: A survey and new perspectives. *ACM computing surveys (CSUR)*, 52(1):1–38, 2019.

H. Zhao, K. Burlachenko, Z. Li, and P. Richtárik. Faster rates for compressed federated learning with client-variance reduction. *arXiv preprint arXiv:2112.13097*, 2021.

P. Zhao and T. Zhang. Stochastic optimization with importance sampling for regularized loss minimization. In *international conference on machine learning*, pages 1–9. PMLR, 2015.

S. Zheng, Z. Huang, and J. Kwok. Communication-Efficient Distributed Blockwise Momentum SGD with Error-Feedback. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

S. Zhou, Y. Wu, Z. Ni, X. Zhou, H. Wen, and Y. Zou. DoReFa-Net: Training Low Bitwidth Convolutional Neural Networks with Low Bitwidth Gradients. *arXiv:1606.06160 [cs]*, Feb. 2018. arXiv: 1606.06160.

T. Zhou, Z. Kuscsik, J.-G. Liu, M. Medo, J. R. Wakeling, and Y.-C. Zhang. Solving the apparent diversity-accuracy dilemma of recommender systems. *Proceedings of the National Academy of Sciences*, 107(10):4511–4515, 2010.

D. L. Zhu and P. Marcotte. Co-Coercivity and Its Role In the Convergence of Iterative Schemes For Solving Variational Inequalities, Mar. 1996.

C.-N. Ziegler, S. M. McNee, J. A. Konstan, and G. Lausen. Improving recommendation lists through topic diversification. In *Proceedings of the 14th international conference on World Wide Web*, pages 22–32, 2005.

W. Zou, H. De Sterck, and J. Liu. Downlink Compression Improves TopK Sparsification, Sept. 2022. arXiv:2209.15203 [cs].

ECOLE
DOCTORALE

Titre: Compression bidirectionnelle pour l'apprentissage fédéré hétérogène

Mots clés: Apprentissage fédéré, optimisation, compression bidirectionnelle, hétérogénéité.

Résumé: Les deux dernières décennies ont été marquées par une augmentation sans précédent de la puissance de calcul et du volume de données disponibles. En conséquence, les algorithmes d'apprentissage automatique ont évolué pour s'adapter à cette nouvelle situation. En particulier, beaucoup d'applications modernes utilisent désormais des réseaux de clients pour stocker les données et calculer les modèles : un apprentissage efficace dans ce cadre est plus difficile, en particulier en raison des contraintes de communication. C'est pourquoi, une nouvelle approche, l'apprentissage fédéré, a été développée au cours de ces dernières années : les données sont conservées sur leur serveur d'origine et un serveur central orchestre l'entraînement. Cette thèse vise à aborder deux aspects fondamentaux de l'apprentissage fédéré. Le premier objectif est d'analyser les compromis de l'apprentissage distribué sous contraintes de communication ; le but étant de réduire le coût énergétique et l'empreinte environnementale. Le second objectif est d'aborder les problèmes résultant de l'hétérogénéité des clients qui complexifie la convergence de l'algorithme vers une solution optimale. Cette thèse se concentre sur la compression bidirectionnelle et résume mes contributions à ce domaine de recherche.
Dans notre première contribution, nous nous concentrons sur l'effet entremêlé de la compression et de l'hétérogénéité (statistique) des clients. Nous introduisons un framework d'algorithmes, appelé `Artemis`, qui s'attaque au problème des coûts de communication de l'apprentissage fédéré. Dans notre deuxième contribution, nous mettons l'accent sur les boucles de rétroaction afin de réduire l'impact de la compression. Nous introduisons un algorithme, `MCM`, qui s'appuie sur `Artemis` et propose un nouveau paradigme qui préserve le modèle central lors de la compression descendante. Ce mécanisme permet d'effectuer une compression bidirectionnelle tout en atteignant asymptotiquement des taux de convergence identiques à ceux de la compression unidirectionnelle. Dans notre troisième contribution, nous allons au-delà de l'hypothèse classique du pire cas sur la variance et fournissons une analyse fine de l'impact de la compression dans le cadre de la régression des moindres carrés. Dans cette configuration, nous mettons en évidence les différences de convergence entre plusieurs schémas de compression sans biais ayant pourtant la même variance.

Title: Bidirectional compression for federated learning in heterogeneous setting

Keywords: Federated learning, optimization, bidirectionnel compression, heterogeneity.

Abstract: The last two decades have witnessed an unprecedented increase in computational power, leading to a vast surge in the volume of available data. As a consequence, machine learning algorithms have evolved to adapt to this new situation. Especially, many modern applications now use a network of clients to store the data and compute the models: efficient learning in this framework is harder, especially under communication constraints. This is why, a new approach, federated learning, has been developed in recent years: the data is kept on the original server and a central server orchestrates the training. This thesis aims to address two fundamental aspects of federated learning. The first goal is to analyze the trade-offs of distributed learning with communication constraints, with the objective of reducing its energy cost and environmental footprint. The second goal is to tackle problems resulting from heterogeneity among clients. This thesis focuses on bidirectional compression and summarizes my contributions to this field of research.
In our first contribution, we focus on the intertwined effect of compression and client (statistical) heterogeneity. We introduce a framework of algorithms, named `Artemis`, that tackles the problem of learning in a federated setting with communication constraints. In our second contribution, we move the focus toward feedback loops to reduce the impact of compression. We introduce an algorithm, coined `MCM`; it builds upon `Artemis` and introduces a new paradigm that preserves the central model from down compression. This mechanism allows to carry out bidirectional compression while asymptotically achieving the rates of convergence of unidirectional compression. In our third contribution, we go beyond the classical worst-case assumption on the variance of compressors and provide a fine-grained analysis of the impact of compression within the fundamental learning framework of least-squares regression. Within this setting, we highlight differences in convergence between several unbiased compression schemes having the same variance increase.

Institut Polytechnique de Paris
91120 Palaiseau, France