Near Optimal Decentralized Optimization with Compression and Momentum Tracking

Rustem Islamov*
University of Basel
rustem.islamov@unibas.ch

Yuan Gao*
CISPA†
yuan.gao@cispa.de

Sebastian U. Stich
CISPA†
stich@cispa.de

Abstract

Communication efficiency has garnered significant attention as it is considered the main bottleneck for large-scale decentralized Machine Learning applications in distributed and federated settings. In this regime, clients are restricted to transmitting small amounts of quantized information to their neighbors over a communication graph. Numerous endeavors have been made to address this challenging problem by developing algorithms with compressed communication for decentralized non-convex optimization problems. Despite considerable efforts, the current results suffer from various issues such as non-scalability with the number of clients, requirements for large batches, or bounded gradient assumption. In this paper, we introduce MoTEF, a novel approach that integrates communication compression with **Mo**mentum **T**racking and **E**rror **F**eedback. Our analysis demonstrates that MoTEF achieves most of the desired properties, and significantly outperforms existing methods under arbitrary data heterogeneity. We provide numerical experiments to validate our theoretical findings and confirm the practical superiority of MoTEF.

1 Introduction

Decentralized machine learning approaches are increasingly popular in numerous applications such as the internet-of-things (IoT) and networked autonomous systems [Marvasti et al., 2014, Savazzi et al., 2020], primarily due to their scalability to larger datasets and systems, as well as their respect for data locality and privacy concerns. In this work, we focus on decentralized optimization techniques that operate without a central coordinator, relying solely on on-device computation and local communication with neighboring devices. This encompasses traditional scenarios like training Machine Learning models in large data centers, as well as emerging applications where computations occur directly on devices. Such a setting is preferred over centralized topology which often poses a significant bottleneck on the central node in terms of communication latency, bandwidth, and fault tolerance.

Considering the enormous size of modern Machine Learning models, classic single-node training is often impossible. Moreover, the training of large models requires a huge amount of data that does not fit the memory of a single machine. Therefore, modern training techniques heavily rely on distributed computations over a set of computation nodes/clients [Shoeybi et al., 2019, Wang et al., 2020, Ramesh et al., 2021, 2022]. One of the instances of distributed training is Federated Learning (FL) [Konecnỳ et al., 2016, Kairouz et al., 2021] which has recently gathered a lot of attention. In this setting, clients, such as hospitals or owners of edge devices, collaboratively train a model on their devices while retaining their data locally.

*Equal contribution.
†CISPA Helmholtz Center for Information Security

One of the key issues in distributed optimization is the communication bottleneck [Seide et al., 2014, Ström, 2015] that limits the scaling properties of distributed deep learning training [Seide et al., 2014, Alistarh et al., 2017, Stich, 2018, Stich et al., 2021]. One of the remedies to decrease communication expenses involves communication compression, where only quantized messages (with fewer bits) are exchanged between clients using compression operators. When used appropriately, contractive compressors (see Definition 1), such as Top-K, are often empirically preferable. However, the naive application of contractive compression operators might lead to divergence [Beznosikov et al., 2023]. To make compression suitable for distributed training, the Error Feedback (EF) mechanism [Seide et al., 2014, Stich et al., 2018] is widely used in practice. It plays a crucial role in achieving high compression ratios.

However, most of the works analyzed EF mechanism in the centralized setting [Stich and Karimireddy, 2019, Gorbunov et al., 2020, Stich, 2020]. Recent research achievements [Gao et al., 2024a, Fatkhullin et al., 2024] demonstrate that in this regime properly constructed EF mechanism can handle both client drift [Mishchenko et al., 2019, Karimireddy et al., 2020b] and stochastic noise from the gradients, and can achieve near-optimal convergence rates. In the more challenging decentralized setting, a series of studies [Zhao et al., 2022, Yan et al., 2023] introduced algorithms capable of effectively managing the drift but fail to achieve a linear acceleration in parallel training, i.e. increasing the number of devices used for training does not lead to a decrease in the training time. Yau and Wai [2022] partially solved this issue under stronger assumptions, achieving linear speed-up using variance reduction, but with worse dependency on the variance of the noise

Designing a method that addresses client drift while preserving linear acceleration in decentralized training has been challenging due to the complex interplay between client drift, Error Feedback mechanism and the communication topology. In our study, we introduce MoTEF, a novel method that tackles these challenges concurrently. Our primary contributions can be outlined as follows.

- We propose a novel method MoTEF that incorporates momentum tracking with compression and Error Feedback, and provably works under standard assumptions (*i*) without imposing any data heterogeneity bounds, (*ii*) without any impractical assumptions such as large batches, (*iii*) with arbitrary contractive compressor, and (*iv*) achieves linear speed-up with the number of clients n. We provide convergence guarantees for the general class of non-convex functions, and for the structured class of non-convex functions satisfying the Polyak-Łojasiewicz (PŁ) condition.

- We propose MoTEF-VR, a momentum-based STORM-type [Cutkosky and Orabona, 2019] variance-reduced variant of our base method that improves further the asymptotic rate of convergence.

- Finally, we provide an extensive numerical study of MoTEF demonstrating the superiority of the proposed method in practice and supporting theoretical claims.

1.1 Related works

Decentralized optimization and gradient tracking. First works in the field studied gossip averaging procedures that are typically used to reach consensus [Kempe et al., 2003, Xiao and Boyd, 2004]. Nevertheless, direct use of gossip averaging might be sub-optimal as it often results in slow convergence [Nedic and Ozdaglar, 2009, Koloskova et al., 2020b]. Gradient tracking [Qu and Li, 2017, Nedic et al., 2017, Koloskova et al., 2021] is one the most popular remedies to this issue. It has been widely applied to obtain faster decentralized algorithms [Sun et al., 2020, Xin et al., 2022, 2021, Li et al., 2022, Zhao et al., 2022, Liu et al., 2024]. In this work, we follow a similar approach but perform a tracking step on momentum term instead of gradients.

Momentum in distributed training. Lately, the utilization of momentum [Polyak, 1964] has attracted attention in distributed optimization. Several works empirically showed that momentum can improve performance in distributed setting [Wang et al., 2019a, Karimireddy et al., 2020a, Das et al., 2022]. Besides, it has recently been shown that the use of momentum improves convergence guarantees [Yau and Wai, 2022, Fatkhullin et al., 2024, Cheng et al., 2024, Huang et al., 2024, Gao et al., 2024b] fully removing dependencies on data heterogeneity bounds. In this work, we follow this approach and apply the momentum technique to the more challenging decentralized setting.

Short history of Error Feedback. Initially, the Error Feedback mechanism was introduced as a heuristic [Seide et al., 2014] and was subsequently analyzed within a simple single-node framework [Stich et al., 2018, Karimireddy et al., 2019]. The first findings in the distributed context were achieved under strong assumptions such as IID data distributions [Karimireddy et al., 2019] or bounded gradients [Cordonnier, 2018, Alistarh et al., 2018, Koloskova et al., 2019, 2020a]. EF21 [Richtárik et al., 2021] was proven to operate with any contractive compressors and under arbitrary heterogeneity, albeit failing to converge when clients are limited to using only stochastic gradients [Fatkhullin et al., 2024]. Subsequently, EF21 was extended to diverse practical scenarios [Fatkhullin et al., 2021] and decentralized training [Zhao et al., 2022] improving the dependencies on some problem parameters. Recent advancements [Gao et al., 2024a, Fatkhullin et al., 2024] have demonstrated that a carefully designed EF mechanism (through the control of feedback signal strength or the use of momentum) results in nearly optimal convergence guarantees in a centralized setting.

Issues of Error Feedback in decentralized setting. Despite having been studied in the centralized setting extensively, EF-based algorithms in the decentralized regime still fail to achieve desirable properties.

- **Strong assumptions.** Many earlier theoretical results for EF require strong assumptions, such as either the bounded gradient assumption [Koloskova et al., 2019, 2020a] or global heterogeneity bound [Lian et al., 2017, Tang et al., 2019, Lu and De Sa, 2021, Singh et al., 2021].
- **Mega batches.** Convergence of the BEER algorithm [Zhao et al., 2022] requires large batches that can be costly or even infeasible in some applications. For example, in medical applications [Rieke et al., 2020] or Reinforcement Learning [Khodadadian et al., 2022, Jin et al., 2022, Mitra et al., 2023] sampling large batches is often intractable. Moreover, it has been shown that training with small batch sizes improves generalization and convergence [Wilson and Martinez, 2003, Keskar et al., 2016, Sekhari et al., 2021].
- **Suboptimal rates.** The stochastic term of several algorithms does not improve with n the number of clients [Zhao et al., 2022, Yan et al., 2023], while the opposite is often desirable, and can be achieved in the centralized training setting [Fatkhullin et al., 2024, Gao et al., 2024a]. Other work achieves speed-up with n, but requires stronger smoothness assumptions and has a worse dependency on the noise variance [Yau and Wai, 2022]. Moreover, [Koloskova et al., 2019, 2020a] do not achieve the standard $\mathcal{O}(1/\varepsilon^2)$ convergence rate in the noiseless regime.
- **Necessity of unbiased compression.** Finally, early works analyzed decentralized algorithms only for a more restricted class of *unbiased* compressors [Tang et al., 2018a, Kovalev et al., 2021]. Huang and Pu [2023] modify any contractive compressor using an additional unbiased compressor following the results of [Horváth and Richtárik, 2020]. This approach enables the creation of a better sequence of gradient estimators, albeit with twice the per-iteration communication cost.

In Table 1, we provide a summary of known theoretical results in decentralized training with compression. We highlight the main issues of existing algorithms.

Table 1: Summary of convergence guarantees for decentralized methods supporting contractive compressors. **nCVX** = supports non-convex functions; **PŁ** = supports functions satisfying PŁ condition. We present the convergence in terms of $\mathbb{E}\left[\|\nabla f(\mathbf{x}_{\text{out}})\|^2\right] \le \varepsilon^2$ and $\mathbb{E}\left[f(\mathbf{x}_{\text{out}}) - f^\star\right] \le \varepsilon$ in PŁ regimes for specifically chosen $\mathbf{x}_{\text{out}}$. Here $F^0 := \mathbb{E}\left[f(\mathbf{x}^0) - f^*\right]$, L and ℓ are smoothness constants, ρ is a spectral gap, and σ^2 is stochastic variance bound.

Method	Asymptotic Complexity nCVX	PŁ	Large Batches?	Extra Assumptions?
Choco-SGD [Koloskova et al., 2019]	$\frac{LF^0\sigma^2}{n\varepsilon^4}$	✗	✗	Bounded Gradients $\mathbb{E}\left[\|\nabla f_i(\mathbf{x},\xi)\|^2\right] \le G^2$
BEER [Zhao et al., 2022]	$\frac{LF^0\sigma^2}{\alpha^2\rho^3\varepsilon^4}$	$\frac{LF^0}{\mu^2\alpha^2\rho^3\varepsilon}$	Batch size of order $\frac{\sigma^2}{\alpha\varepsilon^2}$	✗
CEDAS [Huang and Pu, 2023]	$\frac{LF^0\sigma^2}{n\varepsilon^4}$	✗	✗	Additional Unbiased Compressor
DeepSqueeze [Tang et al., 2019]	$\frac{LF^0\sigma^2}{n\varepsilon^4}$	✗	✗	Bounded Heterogeneity $n^{-1}\sum_i \|\nabla f_i(\mathbf{x}) - \nabla f(\mathbf{x})\|^2 \le \zeta^2$
DoCoM [Yau and Wai, 2022]	$\frac{\ell F^0\sigma^3}{n\varepsilon^3}$	$\frac{\ell F^0\sigma^3}{\mu^2 n\varepsilon}$	✗	✗
CDProxSGT [Yan et al., 2023]	$\frac{LF^0\sigma^2}{\alpha^2\rho^2\varepsilon^4}$	✗	✗	✗
MoTEF **[This work]**	$\frac{LF^0\sigma^2}{n\varepsilon^4}$	$\frac{LF^0\sigma^2}{\mu^2 n\varepsilon}$	✗	✗
MoTEF-VR **[This work]**	$\frac{\ell F^0\sigma^2}{n\varepsilon^3}$	✗	✗	✗

2 Problem setup

Formally, we consider the following optimization problem

$$\min_{\mathbf{x}\in\mathbb{R}^d} \left\{ f(\mathbf{x}) := \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}) \right\}, \tag{1}$$

where n is the number of clients participating in the training, $\mathbf{x}$ are the parameters of a model, $f(\mathbf{x})$ is the global objective, and $f_i(\mathbf{x}) := \mathbb{E}_{\xi_i\sim\mathcal{D}_i}[f_i(\mathbf{x},\xi_i)]$ is the local objective over local dataset $\mathcal{D}_i$. Throughout this work, we assume that the global function f is bounded below by $f^\star > -\infty$.

In the setting of decentralized communication, the clients are restricted to communicating with their neighbors only over a certain undirected communication graph $\mathcal{G}([n], E)$. Each vertex in $[n]$ represents a client, and each edge in E represents a communication link between clients. Besides, we assign a positive weight to w_{ij} if there is an edge $(i,j)\in E$, and $w_{ij}=0$ if $(i,j)\notin E$. Weights w_{ij} form a mixing matrix $\mathbf{W}\in\mathbb{R}^{n\times n}$ (sometimes also called gossip or interaction matrix). The mixing matrix $\mathbf{W}$ should satisfy the following standard assumption.

Assumption 1. *We assume that* $\mathbf{W}$ *is symmetric* ($\mathbf{W} = \mathbf{W}^\top$) *and doubly stochastic* ($\mathbf{W}\mathbf{1} = \mathbf{1}, \mathbf{1}^\top\mathbf{W} = \mathbf{1}^\top$) *matrix with eigenvalues* $1 = |\lambda_1(\mathbf{W})| > |\lambda_2(\mathbf{W})| \ge \dots \ge |\lambda_n(\mathbf{W})|$. *We denote the spectral gap of* $\mathbf{W}$ *as*

$$\rho := 1 - |\lambda_2(\mathbf{W})| \in (0,1]. \tag{2}$$

The spectral gap is typically used to measure the influence of network topology in the training [Aldous and Fill, 2002, Nedić et al., 2018].

In our work, we consider algorithms combined with compressed communication. Formally, we analyze methods utilizing practically useful contractive compression operators.

Definition 1. *We say that a (possibly randomized) mapping $\mathcal{C}: \mathbb{R}^d \to \mathbb{R}^d$ is a contractive compression operator if for some constant $0 < \alpha \leq 1$ it holds*

$$\mathbb{E}\left[\|\mathcal{C}(\mathbf{x}) - \mathbf{x}\|^2\right] \leq (1-\alpha)\|\mathbf{x}\|^2. \tag{3}$$

One of the classic examples of compressors satisfying (3) is Top-K [Stich et al., 2018]. It acts on the input by preserving K largest by magnitude entries while zeroing the rest. The class of contractive compressors includes well-known sparsification [Alistarh et al., 2018, Stich et al., 2018] and quantization [Wen et al., 2017, Bernstein et al., 2018, Horváth et al., 2022] operators. We refer to [Beznosikov et al., 2023, Safaryan et al., 2022, Islamov et al., 2023] for more examples of contractive compressors.

In decentralized training, typically, each client receives the messages from its neighbors and transfers back to them the aggregated information. We highlight that, contrary to many prior works, our analysis supports an arbitrarily heterogeneous setting, i.e. it does not require any assumptions on the heterogeneity level, which means that local data distributions might be distant from each other. Next, we provide standard assumptions on the function class and noise model.

Assumption 2. *We assume that each local function f_i is L-smooth, i.e. for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, and $i \in [n]$ it holds*

$$\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|. \tag{4}$$

Next, we assume that each client has access to an unbiased gradient estimator with bounded variance.

Assumption 3. *We assume that we have access to a gradient oracle $\mathbf{g}^i(\mathbf{x}): \mathbb{R}^d \to \mathbb{R}^d$ for each local function f_i such that for all $\mathbf{x} \in \mathbb{R}^d$ and $i \in [n]$ it holds*

$$\mathbb{E}\left[\mathbf{g}^i(\mathbf{x})\right] = \nabla f_i(\mathbf{x}), \quad \mathbb{E}\left[\|\mathbf{g}^i(\mathbf{x}) - \nabla f_i(\mathbf{x})\|^2\right] \leq \sigma^2. \tag{5}$$

It is important to mention that mini-batches are allowed as well, effectively reducing the variance by the local batch size. Nevertheless, there is no requirement for any specific (minimal) batch size, and for simplicity, we consistently assume a batch size of one.

Finally, we consider the structural class of non-convex functions satisfying Polyak-Łojasiewicz condition [Polyak, 1963]. This assumption is one of the weakest conditions under which vanilla Gradient Descent converges linearly [Karimi et al., 2016].

Assumption 4. *We assume that the global function f is μ-PŁ for some $\mu > 0$, i.e. for all $\mathbf{x} \in \mathbb{R}^d$ it holds*

$$\|\nabla f(\mathbf{x})\|^2 \geq 2\mu(f(x) - f^\star). \tag{6}$$

Note that the PŁ condition is a relaxation of strong convexity, i.e. if strong convexity with parameter μ implies μ-PŁ condition.

3 Theoretical analysis

In this section, we list the main results of our work. Our base method MoTEF is summarized in Algorithm 1, and our variance-reduced variant is summarized in Algorithm 2. We defer the proofs of all theoretical claims to the appendix; see Appendices A and B.

Algorithm 1 MoTEF

Input: $\mathbf{X}^0 = \mathbf{x}^0\mathbf{1}^\top, \mathbf{G}^0, \mathbf{H}^0, \mathbf{V}^0, \gamma, \eta, \lambda,$
$\mathcal{C}_\alpha$
for $t = 0, 1, 2, \ldots$ **do**
  $\mathbf{X}^{t+1} = \mathbf{X}^t + \gamma\mathbf{H}^t(\mathbf{W} - \mathbf{I}) - \eta\mathbf{V}^t$
  $\mathbf{Q}_h^{t+1} = \mathcal{C}_\alpha(\mathbf{X}^{t+1} - \mathbf{H}^t)$
  $\mathbf{H}^{t+1} = \mathbf{H}^t + \mathbf{Q}_h^{t+1}$
  $\mathbf{M}^{t+1} = (1-\lambda)\mathbf{M}^t + \lambda\widetilde{\nabla}F(\mathbf{X}^{t+1})$
  $\mathbf{V}^{t+1} = \mathbf{V}^t + \gamma\mathbf{G}^t(\mathbf{W} - \mathbf{I}) + \mathbf{M}^{t+1} - \mathbf{M}^t$
  $\mathbf{Q}_g^{t+1} = \mathcal{C}_\alpha(\mathbf{V}^{t+1} - \mathbf{G}^t)$
  $\mathbf{G}^{t+1} = \mathbf{G}^t + \mathbf{Q}_g^{t+1}$

Algorithm 2 MoTEF-VR

Input: $\mathbf{X}^0 = \mathbf{x}^0\mathbf{1}^\top, \mathbf{G}^0, \mathbf{H}^0, \mathbf{V}^0, \gamma, \eta, \lambda,$
$\mathcal{C}_\alpha$
for $t = 0, 1, 2, \ldots$ **do**
  $\mathbf{X}^{t+1} = \mathbf{X}^t + \gamma\mathbf{H}^t(\mathbf{W} - \mathbf{I}) - \eta\mathbf{V}^t$
  $\mathbf{Q}_h^{t+1} = \mathcal{C}_\alpha(\mathbf{X}^{t+1} - \mathbf{H}^t)$
  $\mathbf{H}^{t+1} = \mathbf{H}^t + \mathbf{Q}_h^{t+1}$
  $\mathbf{M}^{t+1} = \widetilde{\nabla}F(\mathbf{X}^{t+1}, \Xi^{t+1})$
  $\quad +(1-\lambda)(\mathbf{M}^t - \widetilde{\nabla}F(\mathbf{X}^t, \Xi^{t+1}))$
  $\mathbf{V}^{t+1} = \mathbf{V}^t + \gamma\mathbf{G}^t(\mathbf{W} - \mathbf{I}) + \mathbf{M}^{t+1} - \mathbf{M}^t$
  $\mathbf{Q}_g^{t+1} = \mathcal{C}_\alpha(\mathbf{V}^{t+1} - \mathbf{G}^t)$
  $\mathbf{G}^{t+1} = \mathbf{G}^t + \mathbf{Q}_g^{t+1}$

3.1 Notation

Before going into details, we introduce a notation that we use throughout the paper. We stack the local parameters $\mathbf{x}_i^t$ stored at each clients into a matrix $\mathbf{X}^t \coloneqq [\mathbf{x}_1^t, \ldots, \mathbf{x}_n^t] \in \mathbb{R}^{n\times d}$, and denote the average model $\bar{\mathbf{x}}^t \coloneqq \frac{1}{n}\mathbf{X}^t\mathbf{1}$, where $\mathbf{1}$ is a vector of ones. Other quantities are defined similarly. To track local gradients, we define $\nabla F(\mathbf{X}^t) \coloneqq [\nabla f_1(\mathbf{x}_1^t), \ldots, \nabla f_n(\mathbf{x}_n^t)] \in \mathbb{R}^{d\times n}$. Similarly we write $\widetilde{\nabla}F(\mathbf{X}^t)$ as the collection of local stochastic gradients. Finally, $\mathcal{C}_\alpha(\mathbf{X})$ denotes the contractive compression operator $\mathcal{C}_\alpha$ applied column-wise on a matrix $\mathbf{X}$, i.e. $\mathcal{C}_\alpha(\mathbf{X}) \coloneqq [\mathcal{C}(\mathbf{x}_1), \ldots, \mathcal{C}(\mathbf{x}_n)] \in \mathbb{R}^{d\times n}$.

3.2 Convergence of MoTEF

Now we are ready to present convergence guarantees for MoTEF. Below we summarize the convergence guarantees for Algorithm 1 in general non-convex and PŁ settings. Our analysis relies on the Lyapunov function of the form

$$\Phi^t \coloneqq F^t + \frac{c_1}{n^2L}\hat{G}^t + \frac{c_2\tau}{nL}\widetilde{G}^t + \frac{c_3L}{\rho^3n\tau}\Omega_1^t + \frac{c_4\tau}{\rho nL}\Omega_2^t + \frac{c_5L}{\rho^3n\tau}\Omega_3^t + \frac{c_6\tau}{\rho nL}\Omega_4^t, \tag{7}$$

where $\{c_k\}_{k=1}^6$ are absolute constants defined in (34)[1], $F^t \coloneqq \mathbb{E}\left[f(\bar{\mathbf{x}}^t) - f^\star\right]$ represents the sub-optimality function gap, and the error terms are defined as follows

$$\begin{gathered}\hat{G}^t \coloneqq \mathbb{E}\left[\|\nabla F(\mathbf{X}^t)\mathbf{1} - \mathbf{M}^t\mathbf{1}\|_{\mathrm{F}}^2\right], \quad \widetilde{G}^t \coloneqq \mathbb{E}\left[\left\|\nabla F(\mathbf{X}^t) - \mathbf{M}^t\right\|_{\mathrm{F}}^2\right], \quad \Omega_1^t \coloneqq \mathbb{E}\left[\left\|\mathbf{H}^t - \mathbf{X}^t\right\|_{\mathrm{F}}^2\right] \\ \Omega_2^t \coloneqq \mathbb{E}\left[\left\|\mathbf{G}^t - \mathbf{V}^t\right\|_{\mathrm{F}}^2\right], \quad \Omega_3^t \coloneqq \mathbb{E}\left[\left\|\mathbf{X}^t - \bar{\mathbf{x}}^t\mathbf{1}^T\right\|_{\mathrm{F}}^2\right] \\ \Omega_4^t \coloneqq \mathbb{E}\left[\left\|\mathbf{V}^t - \bar{\mathbf{v}}^t\mathbf{1}^T\right\|_{\mathrm{F}}^2\right], \quad \Omega_5^t \coloneqq \mathbb{E}\left[\left\|\bar{\mathbf{v}}^t\right\|^2\right].\end{gathered} \tag{8}$$

Our theory relies on the descent of the Lyapunov function Φ^t introduced above.

Lemma 1 (Descent of the Lyapunov function)**.** *Let Assumptions 1 and 3 hold. Then there exist absolute constants $c_\gamma, c_\lambda, c_\eta$, and $\tau \le 1$ such that if we set stepsizes $\gamma = c_\gamma\alpha\rho, \lambda = c_\lambda\alpha\rho^3\tau, \eta = c_\eta L^{-1}\alpha\rho^3\tau$ such*

[1] To find a suitable choice of constants we use Symbolic Math Toolbox in MATLAB [Inc., 2023]. Our code can be found at `https://github.com/mlolab/MoTEF.git`.

that the Lyapunov function Φ^t decreases as

$$\Phi^{t+1} \leq \Phi^t - \frac{c_\eta \alpha \rho^3 \tau}{2L} \mathbb{E}\left[\|\nabla f(\bar{\mathbf{x}}^t)\|^2\right] + \frac{c_\lambda^2 c_1 \alpha^2 \rho^6}{nL} \cdot \tau^2 \sigma^2 + \left(\frac{6c_\lambda^2 c_4 \alpha \rho^5}{L} + \frac{2c_\lambda^2 c_2 \alpha^2 \rho^6}{L} + \frac{6c_6 c_\lambda^2 \alpha \rho^4}{c_\gamma L}\right) \tau^3 \sigma^2. \tag{9}$$

Using the above descent of the Lyapunov function, we demonstrate the convergence guarantees for MoTEF.

Theorem 1 (Convergence of MoTEF)**.** *Let Assumptions 1 to 3 hold. Then there exist absolute constants $c_\gamma, c_\lambda, c_\eta$, and some $\tau \leq 1$ such that if we set stepsizes $\gamma = c_\gamma \alpha \rho, \lambda = c_\lambda \alpha \rho^3 \tau, \eta = c_\eta L^{-1} \alpha \rho^3 \tau$, and choosing the initial batch size $B_{\text{init}} \geq \lceil \frac{LF^0}{\sigma^2} \rceil$, then after at most*

$$T = \mathcal{O}\left(\frac{\sigma^2}{n\varepsilon^4} + \frac{\sigma}{\alpha\rho^2\varepsilon^3} + \frac{\sigma}{\alpha^{1/2}\rho^{3/2}\varepsilon^3} + \frac{\sigma}{\alpha\rho^{5/2}\varepsilon^3} + \frac{1}{\alpha\rho^3\varepsilon^2}\right) LF^0 \tag{10}$$

iterations of Algorithm 1 it holds $\mathbb{E}\left[\|\nabla f(\mathbf{x}_{\text{out}})\|^2\right] \leq \varepsilon^2$, where $\mathbf{x}_{\text{out}}$ is chosen uniformly at random from $\{\bar{\mathbf{x}}_0, \ldots, \bar{\mathbf{x}}_{T-1}\}$, and $\mathcal{O}$ suppresses absolute constants.

Remark 2. *Note that using a large initial batch size B_{int} is not required for convergence of* MoTEF*. If we set $B_{\text{init}} = 1$, the above theorem still holds by replacing F^0 by Φ^0.*

We observe that the use of momentum in MoTEF allows us to improve convergence guarantees over BEER. Indeed, Algorithm 1 achieves optimal asymptotic complexity[2] with a desirable linear speed-up with the number of clients n. Moreover, MoTEF provably converges for any batch size in contrast to BEER. To the best of our knowledge, MoTEF is the first decentralized algorithm supporting contractive compressors that achieves optimal asymptotic complexity under Assumptions 2 and 3 without data heterogeneity restrictions.

Now we derive convergence guarantees of MoTEF for the class of functions satisfying Assumption 4.

Theorem 2 (Convergence of MoTEF)**.** *Let Assumptions 1 to 4 hold. Then there exist absolute constants $c_\gamma, c_\lambda, c_\eta$, and some $\tau \leq 1$ such that if we set stepsizes $\gamma = c_\gamma \alpha \rho, \lambda = c_\lambda \alpha \rho^3 \tau, \eta = c_\eta L^{-1} \alpha \rho^3 \tau$, and choosing the initial batch size $B_{\text{init}} \geq \lceil \frac{LF^0}{\sigma^2} \rceil$, then after at most*

$$T = \widetilde{\mathcal{O}}\left(\frac{L\sigma^2}{\mu^2 n\varepsilon} + \frac{L\sigma}{\alpha\rho^2\mu^{3/2}\varepsilon^{1/2}} + \frac{L\sigma}{\alpha\rho^{5/2}\mu^{3/2}\varepsilon^{1/2}} + \frac{L\sigma}{\alpha\rho^2\mu^{3/2}\varepsilon^{1/2}} + \frac{L}{\mu\alpha\rho^3}\right) \tag{11}$$

iterations of Algorithm 1 it holds $\mathbb{E}\left[f(\mathbf{x}^T) - f^*\right] \leq \varepsilon$, and $\widetilde{\mathcal{O}}$ suppresses absolute constants and poly-logarithmic factors.

Remark 3. *Note that using a large initial batch size B_{int} is not required for convergence of* MoTEF*. If we set $B_{\text{init}} = 1$, the above theorem still holds by replacing F^0 by Φ^0, which is hidden in the logarithmic terms.*

Contrary to BEER, we demonstrate that the asymptotic rate of MoTEF in the PŁ setting improves with n and does not require large batches. To the best of our knowledge, MoTEF is the first decentralized algorithm that supports contractive compressors and achieves optimal asymptotic complexity under Assumptions 2 to 4. Moreover, we highlight that in the noiseless regime MoTEF converges linearly as expected. Another momentum-based algorithm DoCom was analyzed under more restricted Assumption 5 only. Therefore, its applicability in this setting is not known. Moreover, DoCoM achieves linear speed-up with n, but with sub-optimal dependency on the noise variance σ^2.

[2]This means the regime when $\varepsilon \to 0$.

3.3 Convergence of MoTEF-VR

Though MoTEF achieves optimal asymptotic complexity under Assumptions 2 and 3, we consider strengthening of Assumption 2, the mean-squared-smoothness assumption, under which further acceleration on the stochastic term might be achieved via variance reduction. In this section, we introduce MoTEF-VR, our variance-reduced algorithm for decentralized optimization with compression. First, we introduce the mean-squared-smoothness assumption.

Assumption 5. *We assume that each local function f_i is ℓ-mean-squared-smooth, i.e. for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d, i \in [n]$, it holds*

$$\mathbb{E}_\xi \left[\|\nabla f_i(\mathbf{x}, \xi) - \nabla f_i(\mathbf{y}, \xi)\|^2 \right] \leq \ell^2 \|\mathbf{x} - \mathbf{y}\|. \tag{12}$$

Note that this also implicitly assumes that the clients can take the same randomness for different points $\mathbf{x}$ and $\mathbf{y}$. Assumption 5 is the standard assumption made for variance reduction techniques, and is the key assumption for circumventing existing lower bounds on stochastic methods [Fang et al., 2018, Cutkosky and Orabona, 2019, Tran-Dinh et al., 2022, Wang et al., 2019b, Xu and Xu, 2022].

In MoTEF-VR, instead of a simple momentum term, each client now maintains a momentum-based variance reduction term, similar to the STORM estimator [Cutkosky and Orabona, 2019]. The algorithm also maintains a momentum parameter λ, and it turns out that the additional variance reduction terms and Assumption 5 allow us to set the momentum parameter more aggressively, leading to an improved convergence rate. For the analysis of MoTEF-VR, we introduce the following Lyapunov function of the form

$$\Psi^t := F^t + \frac{d_1}{\alpha\rho^3 n\tau\ell}\hat{G}^t + \frac{d_2}{n\ell}\widetilde{G}^t + \frac{d_3\ell}{\rho^3 n\tau}\Omega_1^t + \frac{d_4}{\rho n\ell}\Omega_2^t + \frac{d_5\ell}{\rho^3 n\tau}\Omega_3^t + \frac{d_6}{\rho n\ell}\Omega_4^t, \tag{13}$$

where $\{d_k\}_{k=1}^6$ are absolute constants defined in (49), and other error terms are defined similarly as before. Again, we present the descent lemma on the Lyapunov function Ψ^t.

Lemma 4 (Descent of the Lyapunov function). *Let Assumptions 1, 3 and 5 hold. Then there exists absolute constants $c_\gamma, c_\lambda, c_\eta$ and $\tau < 1$ such that if we set stepsizes $\gamma = c_\gamma\alpha\rho, \lambda = c_\lambda n^{-1}\alpha^2\rho^6\tau^2, \eta = c_\eta\ell^{-1}\alpha\rho^3\tau$ then the Lyapunov function Ψ^t decreases as*

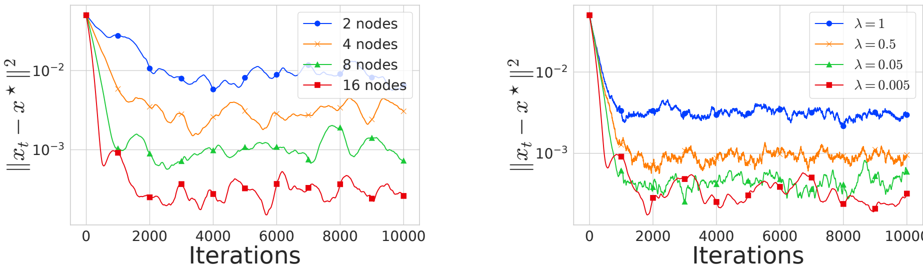
$$\begin{aligned}
\Psi^{t+1} \quad &\leq \quad \Psi^t - \frac{c_\eta\alpha\rho^3}{2\ell}\tau\mathbb{E}\left[\|\nabla f(\bar{\mathbf{x}}^t)\|^2\right] + \frac{2c_1c_\lambda^2}{n^2\ell}\alpha^3\rho^9\tau^3\sigma^2 \\
&\quad + \left(\frac{2c_2c_\lambda^2\alpha^4\rho^{12}}{n^2\ell} + \frac{12c_4c_\lambda^2\alpha^3\rho^{11}}{n^3\ell} + \frac{6c_6c_\lambda^2\alpha^3\rho^{10}}{n^3\ell}\right)\tau^4\sigma^2.
\end{aligned} \tag{14}$$

Remark 5. *Compared to Lemma 1, in Lemma 4, the leading stochastic term has a cubic dependence on τ, whereas in Lemma 1 the dependence is quadratic. The improved dependence on τ is the key ingredient to the speed-up for variance reduction type methods.*

Theorem 3 (Convergence of MoTEF-VR). *Let Assumptions 1, 3 and 5 hold. Then there exists absolute constants $c_\gamma, c_\lambda, c_\eta$ and some $\tau < 1$ such that if we stepsizes $\gamma = c_\gamma\alpha\rho, \lambda = c_\lambda n^{-1}\alpha^2\rho^6\tau^2, \eta = c_\eta\ell^{-1}\alpha\rho^3\tau$, and initial batch size $B_{\text{init}} \geq \lceil \frac{\sigma^2}{LF^0\alpha\rho^3} \rceil$, then after at most*

$$T = \mathcal{O}\left(\frac{\sigma}{n\varepsilon^3} + \frac{\sigma^{2/3}}{n^{2/3}\varepsilon^{8/3}} + \frac{\sigma^{2/3}}{n\alpha^{1/3}\rho^{1/3}\varepsilon^{8/3}} + \frac{\sigma^{2/3}}{n\alpha^{1/3}\rho^{2/3}\varepsilon^{8/3}} + \frac{1}{\alpha\rho^3\varepsilon^2}\right)\ell F^0 \tag{15}$$

iterations of Algorithm 2 it holds $\mathbb{E}\left[\|\nabla f(\mathbf{x}_{\text{out}})\|^2\right] \leq \varepsilon^2$, where $\mathbf{x}_{\text{out}}$ is chosen uniformly at random from $\{\bar{\mathbf{x}}_0, \cdots, \bar{\mathbf{x}}_{T-1}\}$, and $\mathcal{O}$ suppresses absolute constants and poly-logarithmic factors.



(a) (b)

Figure 1: Performance of MoTEF (a) with different number of clients n; (b) varying momentum parameter λ. In all cases, we set $d = 20, \zeta = 10, \sigma = 10$, and apply Top-K compressor with $\alpha = {}^{K}/_{d} = 0.1$. We fix the parameters $\gamma = 0.1, \eta = 0.0005, \lambda = 0.005$, and $n = 16$, if the opposite is not stated.

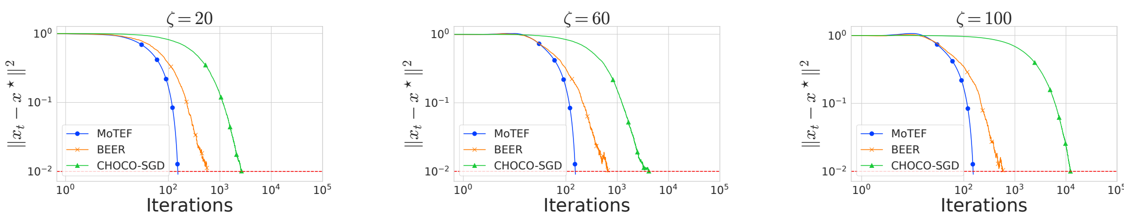


Figure 2: Performance of MoTEF, BEER and CHOCO-SGD with varying data heterogeneity ζ and fixed noise level $\sigma = 5$. We set $d = 20, n = 4$ and apply Top-K compressor with $\alpha = {}^{K}/_{d} = 0.1$. We set the target error to be 0.01.

Remark 6. *Note that using a large initial batch size B_{init} is not required for convergence of* MoTEF-VR*. If we set $B_{\text{init}} = 1$, the above theorem still holds replacing F^0 by Ψ^0.*

Compared to MoTEF, MoTEF-VR achieves an improved asymptotic rate. Moreover, all stochastic terms (the ones with σ) have a speed-up with n in contrast to the convergence of DoCoM, where only asymptotic term improves with n.

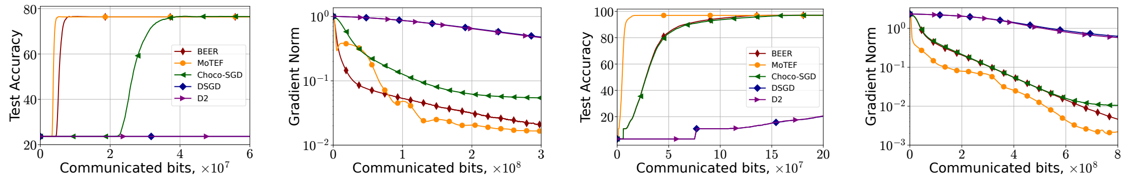
We point out that MoTEF-VR applies the STORM mechanism locally to achieve the variance reduction effect. STORM is specifically designed for non-convex optimization problems, and its convergence rate in the more structured class of functions satisfying Assumption 4 is still unclear in the literature [Cutkosky and Orabona, 2019, Xu and Xu, 2022] even for the simplest centralized SGD setting. In this work, we also do not consider the rate of MoTEF-VR under the PŁ condition.

4 Numerical experiments

In this section, we complement the theoretical results on the convergence of Algorithm 1 with numerical evaluations. We run our experiments on AMD EPYC 9554 64-Core Processor.

4.1 Synthetic least squares problem

We first consider a simple synthetic least squares problem to demonstrate some of the important theoretical properties of Algorithm 1. This problem is designed by Koloskova et al. [2020b] and studied in [Gao et al., 2024a]. For each client i, $f_i(\mathbf{x}) \coloneqq \frac{1}{2}\|\mathbf{A}_i\mathbf{x} - \mathbf{b}_i\|^2$, where $\mathbf{A}_i^2 \coloneqq {}^{i^2}/_{n} \cdot \mathbf{I}_d$ and each $\mathbf{b}_i$



(a) a9a (b) a9a (c) w8a (d) w8a

Figure 3: Comparison of MoTEF, BEER, Choco-SGD, DSGD, D2 in terms of communication complexity on logistic regression with non-convex regularization on ring topology with batch size 5 and gsgd$_b$ compressor.

is sampled from $\mathcal{N}(0, \zeta^2/i^2 \mathbf{I}_d)$ for some parameter ζ which controls the gradient dissimilarity of the problem [Koloskova et al., 2020b]. It is easy to see that when $\zeta = 0, \nabla f_i(\mathbf{x}^\star) = 0, \forall i$. We add Gaussian noise to the gradients to control the stochastic level σ^2 of the gradient. We use the ring network topology for the synthetic experiment. The codes to reproduce our synthetic experiment can be accessed here.

Increasing the number of nodes. In Figure 1-(a) we study the effect of increasing the number of nodes on the convergence of Algorithm 1. A crucial property of Algorithm 1 is that its convergence rate provably improves linearly with the number of nodes, which BEER does not possess. Here we fix a small stepsize and investigate the error that Algorithm 1 achieves with an increasing number of nodes. We observe that the error decreases linearly with the number of nodes, which is consistent with the theoretical results.

Effect of the momentum parameter. In Figure 1-(b) we investigate the effect of the momentum parameter λ. In particular, how it affects the convergence in the noisy regime. Our theoretical analysis suggests that the momentum parameter $\lambda \propto \eta$ is crucial for the convergence of MoTEF. We observe that the error increases as the momentum parameter increases. Note that when $\lambda = 1$, we recover BEER which is known to not converge with the presence of noise in the local gradients, which our experiment confirms.

Effect of changing heterogeneity. In Figure 2 we investigate the effect of changing data heterogeneity ζ on the performance of MoTEF, BEER, and Choco-SGD. The hyperparameters were tuned; the detailed description is given in Appendix C.1. We observe that MoTEF outperforms other algorithms and is not affected by the changing ζ. BEER is also not affected by the changing ζ, while CHOCO-SGD's performance degrades as ζ increases. This is consistent with the theoretical results.

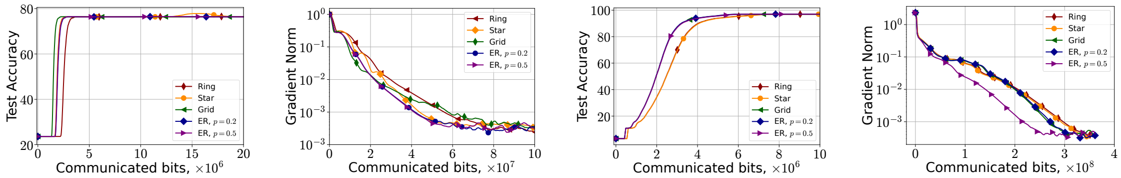
4.2 Non-convex logistic regression

Following [Khirirat et al., 2023, Makarenko et al., 2023, Islamov et al., 2024] we compare algorithms on logistic regression problem with non-convex regularization[3]

$$\min_{\mathbf{x}\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}) + \lambda \sum_{j=1}^{d} \frac{x_j^2}{1+x_j^2}, \quad f_i(\mathbf{x}) \coloneqq \frac{1}{m}\sum_{j=1}^{m} \log(1+\exp(-b_{ij}\mathbf{a}_{ij}^\top \mathbf{x})), \tag{16}$$

where $\{b_{ij}, \mathbf{a}_{ij}\}_{j=1}^m$ is a local dataset. We set $\lambda = 0.05, n = 100$ and use LibSVM datasets [Chang and Lin, 2011]. We do not shuffle datasets to have a more heterogeneous setting. Besides, each dataset is equally distributed among all clients. In all experiments on logistic regression, we use gsgd$_b$ compressor [Alistarh et al., 2017] with $b = 5$. More details of this experiments are given in Appendix C.

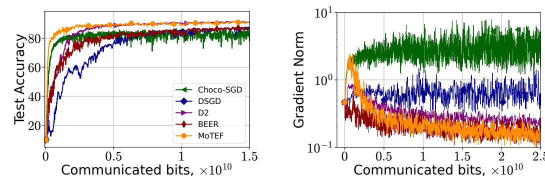
[3] Our implementation is based on open-source code from [Zhao et al., 2022] https://github.com/liboyue/beer. Our code can be accessed at https://github.com/mlolab/MoTEF.git.



(a) a9a (b) a9a (c) w8a (d) w8a

Figure 5: Performance of MoTEF changing of network topology tested on logistic regression with non-convex regularization. We set $n = 40, \lambda = 0.05$, and batch size 100.

Comparison against other methods. We compare BEER [Zhao et al., 2022], Choco-SGD [Koloskova et al., 2019], DSGD [Alistarh et al., 2017], and D2 [Tang et al., 2018b] algorithms with MoTEF on ring topology. Detailed description is given in Appendix C.3. For each algorithm, we fine-tune all stepsizes to achieve better convergence. According to the results in Figure 3, we observe that MoTEF outperforms other algorithms in terms of communication complexity in both cases, when the convergence is measured by training gradient norm and test accuracy.



(a) MLP (b) MLP

Figure 4: Comparison of MoTEF, BEER, Choco-SGD, DSGD, D2 in terms of communication complexity on training MLP with 1 hidden layer.

Robustness to communication topology. Next, we study the effect of the network topology on the convergence of MoTEF. We run experiments for ring, star, grid, Erdös-Rènyi ($p = 0.2$ and $p = 0.5$) topologies. Note the spectral gaps of these networks $0.012, 0.049, 0.063, 0.467, 0.755$ correspondingly. The hyperparameters of algorithms are given in Appendix C.2. Despite the convergence of Algorithm 1 is affected by the spectral gap ρ, in Figure 5 we demonstrate that convergence MoTEF is not affected much by the change in spectral gap. These results demonstrate the robustness of MoTEF to the change of network topology.

Training of MLP. Finally, we consider training MLP with 1 hidden layer of size 32. We present the results in Figure 4. We observe that MLP trained with MoTEF and BEER achieve similar gradient norm, but MoTEF is much faster in accuracy metric showing the advantage from using momentum tracking.

5 Conclusion, limitations, and future work

In this work, we proposed new efficient algorithms, MoTEF and its variance-reduced version MoTEF-VR, for decentralized training with compressed communication that incorporates momentum tracking and Error Feedback. We provide theoretical convergence of our algorithms in general non-convex regimes. Besides, we extend the convergence of MoTEF to the class of functions satisfying the PŁ condition. In the non-convex regime, we achieve an optimal asymptotic complexity without imposing any assumption on batch sizes, bounded gradients or data heterogeneity. We support our theoretical findings with an extensive experimental study.

We believe that it might be possible to improve the dependency on the spectral gap via a more careful choice of the Lyapunov function. However, this might require computer assistance with the search of the parameters. In our study, we focus only on compressed communication while there are many approaches such as performing several local steps [Stich, 2018, Mishchenko et al., 2022b, Gorbunov et al., 2021, Jiang et al., 2024] or asynchronous communication [Islamov et al., 2024, Mishchenko et al., 2022a, Ghadikolaei et al., 2021] that might be useful. We also note that some recent works attempt to improve the dependencies on the smoothness parameters for variants of Error Feedback algorithms [Richtárik et al., 2024], where each local objective is assumed to be L_i-smooth, and a more careful analysis of the method gives a dependency on the average-smoothness $\bar{L} = n^{-1}\sum_{i=1}^{n} L_i$ instead of the maximum smoothness $L = \max_{i\in[n]} L_i$. Therefore, combining the aforementioned research directions with our proof techniques might lead to more improved results. We defer the exploration of these possible extensions to future research endeavors.

References

David Aldous and James Allen Fill. Reversible markov chains and random walks on graphs. *Unfinished monograph, recompiled 2014*, 2002. URL http://www.stat.berkeley.edu/$\sim$aldous/RWG/book.html. (Cited on page 4)

Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. Qsgd: Communication-efficient sgd via gradient quantization and encoding. *Advances in neural information processing systems*, 30, 2017. (Cited on pages 2, 10, and 11)

Dan Alistarh, Torsten Hoefler, Mikael Johansson, Nikola Konstantinov, Sarit Khirirat, and Cédric Renggli. The convergence of sparsified gradient methods. *Advances in Neural Information Processing Systems*, 31, 2018. (Cited on pages 3 and 5)

Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. signsgd: Compressed optimisation for non-convex problems. In *International Conference on Machine Learning*, 2018. (Cited on page 5)

Aleksandr Beznosikov, Samuel Horváth, Peter Richtárik, and Mher Safaryan. On biased compression for distributed learning. *Journal of Machine Learning Research*, 2023. (Cited on pages 2 and 5)

Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2011. (Cited on page 10)

Ziheng Cheng, Xinmeng Huang, Pengfei Wu, and Kun Yuan. Momentum benefits non-iid federated learning simply and provably. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=TdhkAcXkRi. (Cited on page 3)

Jean-Baptiste Cordonnier. Convex optimization using sparsified stochastic gradient descent with memory. 2018. (Cited on page 3)

Ashok Cutkosky and Francesco Orabona. Momentum-based variance reduction in non-convex sgd. *Advances in neural information processing systems*, 32, 2019. (Cited on pages 2, 8, and 9)

Rudrajit Das, Anish Acharya, Abolfazl Hashemi, Sujay Sanghavi, Inderjit S Dhillon, and Ufuk Topcu. Faster non-convex federated learning via global and local momentum. In *Uncertainty in Artificial Intelligence*, 2022. (Cited on page 3)

Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. *Advances in neural information processing systems*, 31, 2018. (Cited on page 8)

Ilyas Fatkhullin, Igor Sokolov, Eduard Gorbunov, Zhize Li, and Peter Richtárik. Ef21 with bells & whistles: Practical algorithmic extensions of modern error feedback, 2021. (Cited on page 3)

Ilyas Fatkhullin, Alexander Tyurin, and Peter Richtárik. Momentum provably improves error feedback! *Advances in Neural Information Processing Systems*, 2024. (Cited on pages 2 and 3)

Yuan Gao, Rustem Islamov, and Sebastian U Stich. EControl: Fast distributed optimization with compression and error control. In *The Twelfth International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=lsvlvWB9vz. (Cited on pages 2, 3, and 9)

Yuan Gao, Anton Rodomanov, and Sebastian U Stich. Non-convex stochastic composite optimization with polyak momentum. *arXiv preprint arXiv:2403.02967*, 2024b. (Cited on page 3)

Hossein Shokri Ghadikolaei, Sebastian Stich, and Martin Jaggi. Lena: Communication-efficient distributed learning with self-triggered gradient uploads. In *International Conference on Artificial Intelligence and Statistics*, pages 3943–3951. PMLR, 2021. (Cited on page 12)

Eduard Gorbunov, Dmitry Kovalev, Dmitry Makarenko, and Peter Richtárik. Linearly converging error compensated sgd. *Advances in Neural Information Processing Systems*, 2020. (Cited on page 2)

Eduard Gorbunov, Filip Hanzely, and Peter Richtárik. Local sgd: Unified theory and new efficient methods. In *International Conference on Artificial Intelligence and Statistics*, 2021. (Cited on page 11)

Samuel Horváth and Peter Richtárik. A better alternative to error feedback for communication-efficient distributed learning. *arXiv preprint arXiv:2006.11077*, 2020. (Cited on page 3)

Samuel Horváth, Chen-Yu Ho, Ludovit Horvath, Atal Narayan Sahu, Marco Canini, and Peter Richtárik. Natural compression for distributed deep learning. In *Mathematical and Scientific Machine Learning*, 2022. (Cited on page 5)

Kun Huang and Shi Pu. Cedas: A compressed decentralized stochastic gradient method with improved convergence. *arXiv preprint arXiv:2301.05872*, 2023. (Cited on pages 3 and 4)

Xinmeng Huang, Ping Li, and Xiaoyun Li. Stochastic controlled averaging for federated learning with communication compression. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/pdf?id=jj5ZjZsWJe. (Cited on page 3)

The MathWorks Inc. Symbolic math toolbox version: 23.2 (r2023b), 2023. URL https://www.mathworks.com. (Cited on page 6)

Rustem Islamov, Xun Qian, Slavomír Hanzely, Mher Safaryan, and Peter Richtárik. Distributed newton-type methods with communication compression and bernoulli aggregation. *Transactions on Machine Learning Research*, 2023. (Cited on page 5)

Rustem Islamov, Mher Safaryan, and Dan Alistarh. AsGrad: A sharp unified analysis of asynchronous-SGD algorithms. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, 2024. (Cited on pages 10 and 12)

Xiaowen Jiang, Anton Rodomanov, and Sebastian U Stich. Federated optimization with doubly regularized drift correction. *arXiv preprint arXiv:2404.08447*, 2024. (Cited on page 12)

Hao Jin, Yang Peng, Wenhao Yang, Shusen Wang, and Zhihua Zhang. Federated reinforcement learning with environment heterogeneity. In *International Conference on Artificial Intelligence and Statistics*, 2022. (Cited on page 3)

Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and trends® in machine learning*, 2021. (Cited on page 1)

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19-23, 2016, Proceedings, Part I 16*, 2016. (Cited on page 5)

Sai Praneeth Karimireddy, Quentin Rebjock, Sebastian Stich, and Martin Jaggi. Error feedback fixes signsgd and other gradient compression schemes. In *International Conference on Machine Learning*, 2019. (Cited on page 3)

Sai Praneeth Karimireddy, Martin Jaggi, Satyen Kale, Mehryar Mohri, Sashank J Reddi, Sebastian U Stich, and Ananda Theertha Suresh. Mime: Mimicking centralized stochastic algorithms in federated learning. *arXiv preprint arXiv:2008.03606*, 2020a. (Cited on page 3)

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International conference on machine learning*. PMLR, 2020b. (Cited on page 2)

David Kempe, Alin Dobra, and Johannes Gehrke. Gossip-based computation of aggregate information. In *44th Annual IEEE Symposium on Foundations of Computer Science, 2003. Proceedings.*, 2003. (Cited on page 2)

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016. (Cited on page 3)

Sarit Khirirat, Eduard Gorbunov, Samuel Horváth, Rustem Islamov, Fakhri Karray, and Peter Richtárik. Clip21: Error feedback for gradient clipping. *arXiv preprint: arXiv 2305.18929*, 2023. (Cited on page 10)

Sajad Khodadadian, Pranay Sharma, Gauri Joshi, and Siva Theja Maguluri. Federated reinforcement learning: Linear speedup under markovian sampling. In *International Conference on Machine Learning*, 2022. (Cited on page 3)

Anastasia Koloskova, Sebastian Stich, and Martin Jaggi. Decentralized stochastic optimization and gossip algorithms with compressed communication. In *International Conference on Machine Learning*, 2019. (Cited on pages 3, 4, and 11)

Anastasia Koloskova, Tao Lin, Sebastian U Stich, and Martin Jaggi. Decentralized deep learning with arbitrary communication compression. In *International Conference on Learning Representations*, 2020a. URL `https://openreview.net/forum?id=SkgGCkrKvH`. (Cited on page 3)

Anastasia Koloskova, Nicolas Loizou, Sadra Boreiri, Martin Jaggi, and Sebastian Stich. A unified theory of decentralized sgd with changing topology and local updates. In *Proceedings of International Conference on Machine Learning*, pages 5381–5393. PMLR, 2020b. (Cited on pages 2, 9, and 10)

Anastasiia Koloskova, Tao Lin, and Sebastian U Stich. An improved analysis of gradient tracking for decentralized machine learning. *Advances in Neural Information Processing Systems*, 2021. (Cited on page 2)

Jakub Konecnỳ, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016. (Cited on page 1)

Dmitry Kovalev, Anastasia Koloskova, Martin Jaggi, Peter Richtarik, and Sebastian Stich. A linearly convergent algorithm for decentralized optimization: Sending less bits for free! In *International Conference on Artificial Intelligence and Statistics*, 2021. (Cited on page 3)

Boyue Li, Zhize Li, and Yuejie Chi. Destress: Computation-optimal and communication-efficient decentralized nonconvex finite-sum optimization. *SIAM Journal on Mathematics of Data Science*, 2022. (Cited on page 2)

Xiangru Lian, Ce Zhang, Huan Zhang, Cho-Jui Hsieh, Wei Zhang, and Ji Liu. Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. *Advances in neural information processing systems*, 2017. (Cited on page 3)

Yue Liu, Tao Lin, Anastasia Koloskova, and Sebastian U Stich. Decentralized gradient tracking with local steps. *Optimization Methods and Software*, pages 1–28, 2024. (Cited on page 2)

Yucheng Lu and Christopher De Sa. Optimal complexity in decentralized training. In *International Conference on Machine Learning*, 2021. (Cited on page 3)

Maksim Makarenko, Elnur Gasanov, Abdurakhmon Sadiev, Rustem Islamov, and Peter Richtárik. Adaptive compression for communication-efficient distributed training. *Transactions on Machine Learning Research*, 2023. (Cited on page 10)

Amin Kargarian Marvasti, Yong Fu, Saber DorMohammadi, and Masoud Rais-Rohani. Optimal operation of active distribution grids: A system of systems framework. *IEEE Transactions on Smart Grid*, 2014. (Cited on page 1)

Konstantin Mishchenko, Eduard Gorbunov, Martin Takáč, and Peter Richtárik. Distributed learning with compressed gradient differences. *arXiv preprint arXiv:1901.09269*, 2019. (Cited on page 2)

Konstantin Mishchenko, Ahmed Khaled, and Peter Richtárik. Random reshuffling: Simple analysis with vast improvements. *Advances in Neural Information Processing Systems*, 2020. (Cited on page 28)

Konstantin Mishchenko, Francis Bach, Mathieu Even, and Blake E Woodworth. Asynchronous sgd beats minibatch sgd under arbitrary delays. *Advances in Neural Information Processing Systems*, 2022a. (Cited on page 12)

Konstantin Mishchenko, Grigory Malinovsky, Sebastian Stich, and Peter Richtárik. Proxskip: Yes! local gradient steps provably lead to communication acceleration! finally! In *International Conference on Machine Learning*, 2022b. (Cited on page 11)

Aritra Mitra, George J Pappas, and Hamed Hassani. Temporal difference learning with compressed updates: Error-feedback meets reinforcement learning. *arXiv preprint arXiv:2301.00944*, 2023. (Cited on page 3)

Angelia Nedic and Asuman Ozdaglar. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 2009. (Cited on page 2)

Angelia Nedic, Alex Olshevsky, and Wei Shi. Achieving geometric convergence for distributed optimization over time-varying graphs. *SIAM Journal on Optimization*, 2017. (Cited on page 2)

Angelia Nedić, Alex Olshevsky, and Michael G Rabbat. Network topology and communication-computation tradeoffs in decentralized optimization. *Proceedings of the IEEE*, 2018. (Cited on page 4)

Boris T Polyak. Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics*, 1963. (Cited on page 5)

Boris T Polyak. Some methods of speeding up the convergence of iteration methods. *Ussr computational mathematics and mathematical physics*, 1964. (Cited on page 2)

Guannan Qu and Na Li. Harnessing smoothness to accelerate distributed optimization. *IEEE Transactions on Control of Network Systems*, 2017. (Cited on page 2)

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, 2021. (Cited on page 1)

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. (Cited on page 1)

Peter Richtárik, Igor Sokolov, and Ilyas Fatkhullin. Ef21: A new, simpler, theoretically better, and practically faster error feedback. *Advances in Neural Information Processing Systems*, 2021. (Cited on page 3)

Peter Richtárik, Elnur Gasanov, and Konstantin Pavlovich Burlachenko. Error feedback reloaded: From quadratic to arithmetic mean of smoothness constants. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=Ch7WqGcGmb`. (Cited on page 12)

Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletari, Holger R Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N Galtier, Bennett A Landman, Klaus Maier-Hein, et al. The future of digital health with federated learning. *NPJ digital medicine*, 2020. (Cited on page 3)

Mher Safaryan, Rustem Islamov, Xun Qian, and Peter Richtarik. FedNL: Making Newton-type methods applicable to federated learning. In *Proceedings of the 39th International Conference on Machine Learning*, 2022. (Cited on page 5)

Stefano Savazzi, Monica Nicoli, and Vittorio Rampa. Federated learning with cooperating devices: A consensus approach for massive iot networks. *IEEE Internet of Things Journal*, 2020. (Cited on page 1)

Frank Seide, Hao Fu, Jasha Droppo, Gang Li, and Dong 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*, 2014. (Cited on pages 2 and 3)

Ayush Sekhari, Karthik Sridharan, and Satyen Kale. Sgd: The role of implicit regularization, batch-size and multiple-epochs. *Advances In Neural Information Processing Systems*, 2021. (Cited on page 3)

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*, 2019. (Cited on page 1)

Navjot Singh, Deepesh Data, Jemin George, and Suhas Diggavi. Squarm-sgd: Communication-efficient momentum sgd for decentralized optimization. *IEEE Journal on Selected Areas in Information Theory*, 2021. (Cited on page 3)

Sebastian Stich, Amirkeivan Mohtashami, and Martin Jaggi. Critical parameters for scalable distributed learning with large batches and asynchronous updates. In *International Conference on Artificial Intelligence and Statistics*, pages 4042–4050. PMLR, 2021. (Cited on page 2)

Sebastian U Stich. Local sgd converges fast and communicates little. *arXiv preprint arXiv:1805.09767*, 2018. (Cited on pages 2 and 11)

Sebastian U Stich. On communication compression for distributed optimization on heterogeneous data. *arXiv preprint arXiv:2009.02388*, 2020. (Cited on page 2)

Sebastian U Stich and Sai Praneeth Karimireddy. The error-feedback framework: Better rates for sgd with delayed gradients and compressed communication. *arXiv preprint arXiv:1909.05350*, 2019. (Cited on page 2)

Sebastian U Stich, Jean-Baptiste Cordonnier, and Martin Jaggi. Sparsified sgd with memory. *Advances in neural information processing systems*, 31, 2018. (Cited on pages 2, 3, and 5)

Nikko Ström. Scalable distributed dnn training using commodity gpu cloud computing. 2015. (Cited on page 2)

Haoran Sun, Songtao Lu, and Mingyi Hong. Improving the sample and communication complexity for decentralized non-convex optimization: Joint gradient estimation and tracking. In *International conference on machine learning*, 2020. (Cited on page 2)

Hanlin Tang, Shaoduo Gan, Ce Zhang, Tong Zhang, and Ji Liu. Communication compression for decentralized training. *Advances in Neural Information Processing Systems*, 2018a. (Cited on page 3)

Hanlin Tang, Xiangru Lian, Ming Yan, Ce Zhang, and Ji Liu. D2: Decentralized training over decentralized data. In *International Conference on Machine Learning*, 2018b. (Cited on page 11)

Hanlin Tang, Xiangru Lian, Shuang Qiu, Lei Yuan, Ce Zhang, Tong Zhang, and Ji Liu. Deepsqueeze: Decentralization meets error-compensated compression. *arXiv preprint arXiv:1907.07346*, 2019. (Cited on pages 3 and 4)

Quoc Tran-Dinh, Nhan H Pham, Dzung T Phan, and Lam M Nguyen. A hybrid stochastic optimization framework for composite nonconvex optimization. *Mathematical Programming*, 191(2):1005–1071, 2022. (Cited on page 8)

Jianyu Wang, Vinayak Tantia, Nicolas Ballas, and Michael Rabbat. Slowmo: Improving communication-efficient distributed sgd with slow momentum. *arXiv preprint arXiv:1910.00643*, 2019a. (Cited on page 3)

Meng Wang, Weijie Fu, Xiangnan He, Shijie Hao, and Xindong Wu. A survey on large-scale machine learning. *IEEE Transactions on Knowledge and Data Engineering*, 2020. (Cited on page 1)

Zhe Wang, Kaiyi Ji, Yi Zhou, Yingbin Liang, and Vahid Tarokh. Spiderboost and momentum: Faster variance reduction algorithms. *Advances in Neural Information Processing Systems*, 32, 2019b. (Cited on page 8)

Wei Wen, Cong Xu, Feng Yan, Chunpeng Wu, Yandan Wang, Yiran Chen, and Hai Li. Terngrad: Ternary gradients to reduce communication in distributed deep learning. *Advances in neural information processing systems*, 2017. (Cited on page 5)

D Randall Wilson and Tony R Martinez. The general inefficiency of batch training for gradient descent learning. *Neural networks*, 2003. (Cited on page 3)

Lin Xiao and Stephen Boyd. Fast linear iterations for distributed averaging. *Systems & Control Letters*, 2004. (Cited on page 2)

Ran Xin, Usman A Khan, and Soummya Kar. A fast randomized incremental gradient method for decentralized nonconvex optimization. *IEEE Transactions on Automatic Control*, 2021. (Cited on page 2)

Ran Xin, Usman A Khan, and Soummya Kar. Fast decentralized nonconvex finite-sum optimization with recursive variance reduction. *SIAM Journal on Optimization*, 2022. (Cited on page 2)

Yangyang Xu and Yibo Xu. Momentum-based variance-reduced proximal stochastic gradient method for composite nonconvex stochastic optimization, 2022. (Cited on pages 8 and 9)

Yonggui Yan, Jie Chen, Pin-Yu Chen, Xiaodong Cui, Songtao Lu, and Yangyang Xu. Compressed decentralized proximal stochastic gradient method for nonconvex composite problems with heterogeneous data. In *International Conference on Machine Learning*, 2023. (Cited on pages 2, 3, and 4)

Chung-Yiu Yau and Hoi-To Wai. Docom: Compressed decentralized optimization with near-optimal sample complexity. *arXiv preprint arXiv:2202.00255*, 2022. (Cited on pages 2, 3, and 4)

Haoyu Zhao, Boyue Li, Zhize Li, Peter Richtárik, and Yuejie Chi. Beer: Fast $o(1/t)$ rate for decentralized nonconvex optimization with communication compression. *Advances in Neural Information Processing Systems*, 2022. (Cited on pages 2, 3, 4, 10, 11, 19, and 22)

A Missing proof for MoTEF

We recall the notation we use to prove convergence of MoTEF:

$$
\begin{aligned}
&\hat{G}^t \coloneqq \mathbb{E}\left[\left\|\nabla F(\mathbf{X}^t)\mathbf{1} - \mathbf{M}^t\mathbf{1}\right\|^2\right] \\
&\widetilde{G}^t \coloneqq \sum_{i=1}^n \mathbb{E}\left[\left\|\nabla f_i(\mathbf{x}_i^t) - \mathbf{m}_i^t\right\|^2\right] = \mathbb{E}\left[\left\|\nabla F(\mathbf{X}^t) - \mathbf{M}^t\right\|_F^2\right] \\
&\Omega_1^t \coloneqq \mathbb{E}\left[\left\|\mathbf{H}^t - \mathbf{X}^t\right\|_{\mathrm{F}}^2\right] \\
&\Omega_2^t \coloneqq \mathbb{E}\left[\left\|\mathbf{G}^t - \mathbf{V}^t\right\|_{\mathrm{F}}^2\right] \\
&\Omega_3^t \coloneqq \mathbb{E}\left[\left\|\mathbf{X}^t - \bar{\mathbf{x}}^t\mathbf{1}^T\right\|_{\mathrm{F}}^2\right] \\
&\Omega_4^t \coloneqq \mathbb{E}\left[\left\|\mathbf{V}^t - \bar{\mathbf{v}}^t\mathbf{1}^T\right\|_{\mathrm{F}}^2\right] \\
&\Omega_5^t \coloneqq \mathbb{E}\left[\left\|\bar{\mathbf{v}}^t\right\|^2\right].
\end{aligned}
$$

Moreover, $F^t \coloneqq \mathbb{E}\left[f(\bar{\mathbf{x}}^t)\right] - f^\star$. Let us define $\boldsymbol{\Omega}^t \coloneqq [\hat{G}^t, \widetilde{G}^t, \Omega_1^t, \Omega_2^t, \Omega_3^t, \Omega_4^t]^\top$. In addition, we denote $\widetilde{\nabla} F(\mathbf{X}^t) \coloneqq [\mathbf{g}^1(\mathbf{x}_i^t), \ldots, \mathbf{g}^n(\mathbf{x}_n^t)] \in \mathbb{R}^{d\times n}$ a matrix that contains local stochastic gradients. We denote $C \coloneqq \sigma_{\max}^2(\mathbf{W} - \mathbf{I}) \leq 4$.

Lemma 7 (Lemma B.2 from [Zhao et al., 2022])**.** *Let $\mathbf{W}$ be a mixing matrix with a spectral gap ρ. Then for any matrix $\mathbf{X} \in \mathbb{R}^{d\times n}$ and $\bar{\mathbf{x}} = \frac{1}{n}\mathbf{X}\mathbf{1}$ we have*

$$
\|\mathbf{X}\mathbf{W} - \bar{\mathbf{x}}\mathbf{1}^\top\|_{\mathrm{F}}^2 \leq (1-\rho)\|\mathbf{X} - \bar{\mathbf{x}}\mathbf{1}^\top\|_{\mathrm{F}}^2. \tag{17}
$$

Moreover, for any $\gamma \in (0,1]$ the matrix $\widetilde{\mathbf{W}} = \mathbf{I} + \gamma(\mathbf{W} - \mathbf{I})$ has a spectral gap at least $\gamma\rho$.

Lemma 8. *The iterates of Algorithm 1 satisfy*

$$
\bar{\mathbf{v}}^{t+1} = \frac{1}{n}\mathbf{M}^{t+1}\mathbf{1}, \tag{18}
$$

and

$$
\bar{\mathbf{x}}^{t+1} = \bar{\mathbf{x}}^t - \frac{\eta}{n}\mathbf{M}^t\mathbf{1}. \tag{19}
$$

Proof. By induction, we can show that $\bar{\mathbf{v}}^t = \frac{1}{n}\mathbf{M}^t\mathbf{1}$, if we initialize $\mathbf{V}^0 = \mathbf{M}^0$. Indeed, we have

$$
\begin{aligned}
\bar{\mathbf{v}}^{t+1} &= \frac{1}{n}\mathbf{V}^{t+1}\mathbf{1} \\
&= \frac{1}{n}\mathbf{V}^t\mathbf{1} + \frac{1}{n}\gamma\mathbf{G}^t(\mathbf{W} - \mathbf{I})\mathbf{1} + \frac{1}{n}(\mathbf{M}^{t+1} - \mathbf{M}^t)\mathbf{1} \\
&= \frac{1}{n}\mathbf{V}^t\mathbf{1} + \frac{1}{n}(\mathbf{M}^{t+1} - \mathbf{V}^t)\mathbf{1} \\
&= \frac{1}{n}\mathbf{M}^{t+1}\mathbf{1}.
\end{aligned}
$$

Therefore, we have

$$
\begin{aligned}
\bar{\mathbf{x}}^{t+1} &= \bar{\mathbf{x}}^t + \frac{\gamma}{n}\mathbf{H}^t(\mathbf{W} - \mathbf{I})\mathbf{1} - \frac{\eta}{n}\mathbf{V}^t\mathbf{1} \\
&= \bar{\mathbf{x}}^t - \eta\bar{\mathbf{v}}^t = \bar{\mathbf{x}}^t - \frac{\eta}{n}\mathbf{M}^t\mathbf{1}.
\end{aligned}
$$

□

A.1 General non-convex setting.

Lemma 9. *Assume that Assumption 2 holds. Then we have the following descent on F^t*

$$F_{t+1} \quad \le F_t - \tfrac{\eta}{2}\mathbb{E}\left[\|\nabla f(\bar{\mathbf{x}}^t)\|^2\right] + \tfrac{\eta}{n^2}\hat{G}^t + \tfrac{\eta L^2}{n}\Omega_3^t - (-\eta/2 - \eta^2 L/2)\Omega_5^t. \tag{20}$$

Proof. Using smoothness we get

$$\begin{aligned}
F_{t+1} &\le F_t - \eta\mathbb{E}\left[\langle\nabla f(\bar{\mathbf{x}}^t), \bar{\mathbf{v}}^t\rangle\right] + \frac{\eta^2 L}{2}\mathbb{E}\left[\|\mathbf{v}^t\|^2\right] \\
&= F_t - \frac{\eta}{2}\mathbb{E}\left[\|\nabla f(\bar{\mathbf{x}}^t)\|^2\right] + \frac{\eta}{2}\mathbb{E}\left[\|\nabla f(\bar{\mathbf{x}}^t) - \bar{\mathbf{v}}^t\|^2\right] - (-\eta/2 - \eta^2 L/2)\mathbb{E}\left[\|\bar{\mathbf{v}}^t\|^2\right] \\
&= F_t - \frac{\eta}{2}\mathbb{E}\left[\|\nabla f(\bar{\mathbf{x}}^t)\|^2\right] + \frac{\eta}{2}\mathbb{E}\left[\left\|\frac{1}{n}\nabla F(\bar{\mathbf{x}}^t)\mathbf{1} - \frac{1}{n}\mathbf{M}^t\mathbf{1}\right\|^2\right] - (-\eta/2 - \eta^2 L/2)\mathbb{E}\left[\|\bar{\mathbf{v}}^t\|^2\right] \\
&\le F_t - \frac{\eta}{2}\mathbb{E}\left[\|\nabla f(\bar{\mathbf{x}}^t)\|^2\right] + \eta\mathbb{E}\left[\left\|\frac{1}{n}\nabla F(\mathbf{X}^t)\mathbf{1} - \frac{1}{n}\mathbf{M}^t\mathbf{1}\right\|^2\right] \\
&\quad + \eta\mathbb{E}\left[\left\|\frac{1}{n}\nabla F(\bar{\mathbf{x}}^t)\mathbf{1} - \frac{1}{n}\nabla F(\mathbf{X}^t)\mathbf{1}\right\|^2\right] - (-\eta/2 - \eta^2 L/2)\mathbb{E}\left[\|\bar{\mathbf{v}}^t\|^2\right] \\
&\le F_t - \frac{\eta}{2}\mathbb{E}\left[\|\nabla f(\bar{\mathbf{x}}^t)\|^2\right] + \frac{\eta}{n^2}\hat{G}^t + \frac{\eta L^2}{n}\mathbb{E}\left[\left\|\mathbf{X}^t - \bar{\mathbf{x}}^t\mathbf{1}^\top\right\|^2\right] - (-\eta/2 - \eta^2 L/2)\Omega_5^t \\
&= F_t - \frac{\eta}{2}\mathbb{E}\left[\|\nabla f(\bar{\mathbf{x}}^t)\|^2\right] + \frac{\eta}{n^2}\hat{G}^t + \frac{\eta L^2}{n}\Omega_3^t - (-\eta/2 - \eta^2 L/2)\Omega_5^t.
\end{aligned}$$

□

Lemma 10. *Assume that Assumptions 2 and 3 hold. Then we have the following descent on $\hat{G}^t$*

$$\hat{G}^{t+1} \le (1-\lambda)\mathbb{E}\left[\left\|\nabla F(\mathbf{X}^t)\mathbf{1} - \mathbf{M}^t\mathbf{1}\right\|^2\right] + \frac{(1-\lambda)^2 nL^2}{\lambda}\mathbb{E}\left[\left\|\mathbf{X}^t - \mathbf{X}^{t+1}\right\|_{\mathrm{F}}^2\right] + \lambda^2 n\sigma^2. \tag{21}$$

Proof. Using the update rules of Algorithm 1 we get

$$\begin{aligned}
\hat{G}^{t+1} \quad &= \quad \mathbb{E}\left[\left\|\nabla F(\mathbf{X}^{t+1})\mathbf{1} - \mathbf{M}^{t+1}\mathbf{1}\right\|^2\right] \\
&= \quad \mathbb{E}\left[\left\|\nabla F(\mathbf{X}^{t+1})\mathbf{1} - (1-\lambda)\mathbf{M}^t\mathbf{1} - \lambda\widetilde{\nabla}F(\mathbf{X}^{t+1})\mathbf{1}\right\|^2\right] \\
&= \quad \mathbb{E}\left[\left\|(1-\lambda)(\nabla F(\mathbf{X}^t) - \mathbf{M}^t)\mathbf{1} + \lambda(\nabla F(\mathbf{X}^{t+1}) - \widetilde{\nabla}F(\mathbf{X}^{t+1}))\mathbf{1} + (1-\lambda)(\nabla F(\mathbf{X}^{t+1}) - \nabla F(\mathbf{X}^t))\mathbf{1}\right\|^2\right] \\
&\le \quad (1-\lambda)^2\mathbb{E}\left[\left\|(\nabla F(\mathbf{X}^t) - \mathbf{M}^t)\mathbf{1} + (\nabla F(\mathbf{X}^{t+1}) - \nabla F(\mathbf{X}^t))\mathbf{1}\right\|^2\right] + \lambda^2 n\sigma^2 \\
&\le \quad (1-\lambda)\mathbb{E}\left[\left\|\nabla F(\mathbf{X}^t)\mathbf{1} - \mathbf{M}^t\mathbf{1}\right\|^2\right] + \frac{(1-\lambda)^2 nL^2}{\lambda}\mathbb{E}\left[\left\|\mathbf{X}^t - \mathbf{X}^{t+1}\right\|_{\mathrm{F}}^2\right] + \lambda^2 n\sigma^2,
\end{aligned}$$

where in the first inequality we use the fact that $\mathbb{E}\left[\widetilde{\nabla}F(\mathbf{X}^{t+1})\right] = \nabla F(\mathbf{X}^{t+1})$ and Assumption 3, and in the second inequality we use $\|\mathbf{a}+\mathbf{b}\|^2 \le (1+\beta)\|\mathbf{a}\|^2 + (1+\beta^{-1})\|\mathbf{b}\|^2$ for any vectors $\mathbf{a}, \mathbf{b}$ and constant $\mathbf{b}$. □

Lemma 11. *Assume that Assumptions 2 and 3 hold. Then we have the following descent on $\widetilde{G}^t$*

$$\widetilde{G}^{t+1} \le \lambda^2\sigma^2 n + \frac{(1-\lambda)^2 L^2}{\lambda}\mathbb{E}\left[\left\|\mathbf{X}^{t+1} - \mathbf{X}^t\right\|_{\mathrm{F}}^2\right] + (1-\lambda)\widetilde{G}^t. \tag{22}$$

Proof.

$$\begin{aligned}
\widetilde{G}^{t+1} &= \mathbb{E}\left[\left\|\nabla F(\mathbf{X}^{t+1}) - \mathbf{M}^{t+1}\right\|_{\mathrm{F}}^2\right] \\
&= \mathbb{E}\left[\left\|\nabla F(\mathbf{X}^{t+1}) - (1-\lambda)\mathbf{M}^t - \lambda\widetilde{\nabla} F(\mathbf{X}^{t+1})\right\|_{\mathrm{F}}^2\right] \\
&\leq \lambda^2\sigma^2 n + (1-\lambda)^2\mathbb{E}\left[\left\|\nabla F(\mathbf{X}^{t+1}) - \mathbf{M}^t\right\|_{\mathrm{F}}^2\right] \\
&\leq \lambda^2\sigma^2 n + (1-\lambda)^2(1+\beta_1^{-1})\mathbb{E}\left[\left\|\nabla F(\mathbf{X}^{t+1}) - \nabla F(\mathbf{X}^t)\right\|_{\mathrm{F}}^2\right] \\
&\quad + (1-\lambda)^2(1+\beta_1)\mathbb{E}\left[\left\|\mathbf{M}^t - \nabla F(\mathbf{X}^t)\right\|_{\mathrm{F}}^2\right] \\
&\leq \lambda^2\sigma^2 n + \frac{(1-\lambda)^2}{\lambda}\mathbb{E}\left[\left\|\nabla F(\mathbf{X}^{t+1}) - \nabla F(\mathbf{X}^t)\right\|_{\mathrm{F}}^2\right] + (1-\lambda)\mathbb{E}\left[\left\|\mathbf{M}^t - \nabla F(\mathbf{X}^t)\right\|_{\mathrm{F}}^2\right] \\
&\leq \lambda^2\sigma^2 n + \frac{(1-\lambda)^2 L^2}{\lambda}\mathbb{E}\left[\left\|\mathbf{X}^{t+1} - \mathbf{X}^t\right\|_{\mathrm{F}}^2\right] + (1-\lambda)\widetilde{G}^t.
\end{aligned}$$

where we choose $\beta_1 = \frac{\lambda}{(1-\lambda)}$. □

Lemma 12. *Let $\mathcal{C}_\alpha$ be any contractive compressor with parameter α. Then we have the following descent on Ω_1^t*

$$\Omega_1^{t+1} \leq (1 - \alpha/2)\mathbb{E}\left[\left\|\mathbf{H}^t - \mathbf{X}^t\right\|_{\mathrm{F}}^2\right] + \frac{2}{\alpha}\mathbb{E}\left[\left\|\mathbf{X}^t - \mathbf{X}^{t+1}\right\|_{\mathrm{F}}^2\right]. \tag{23}$$

Proof. We have

$$\begin{aligned}
\Omega_1^{t+1} &= \mathbb{E}\left[\left\|\mathbf{H}^{t+1} - \mathbf{X}^{t+1}\right\|_{\mathrm{F}}^2\right] \\
&= \mathbb{E}\left[\left\|\mathbf{H}^t + \mathcal{C}_\alpha(\mathbf{X}^{t+1} - \mathbf{H}^t) - \mathbf{X}^{t+1}\right\|_{\mathrm{F}}^2\right] \\
&\leq (1-\alpha)\mathbb{E}\left[\left\|\mathbf{H}^t - \mathbf{X}^{t+1}\right\|_{\mathrm{F}}^2\right] \\
&\leq (1 - \alpha/2)\mathbb{E}\left[\left\|\mathbf{H}^t - \mathbf{X}^t\right\|_{\mathrm{F}}^2\right] + \frac{2}{\alpha}\mathbb{E}\left[\left\|\mathbf{X}^t - \mathbf{X}^{t+1}\right\|_{\mathrm{F}}^2\right].
\end{aligned}$$

□

Lemma 13. *Let $\mathcal{C}_\alpha$ be any contractive compressor with parameter α. Then we have the following descent on Ω_2^t*

Proof. The proof is similar to the one of Lemma 12

$$\begin{aligned}
\Omega_2^{t+1} &= \mathbb{E}\left[\left\|\mathbf{G}^{t+1} - \mathbf{V}^{t+1}\right\|_{\mathrm{F}}^2\right] \\
&\leq (1 - \alpha/2)\mathbb{E}\left[\left\|\mathbf{G}^t - \mathbf{V}^t\right\|_{\mathrm{F}}^2\right] + \frac{2}{\alpha}\mathbb{E}\left[\left\|\mathbf{V}^t - \mathbf{V}^{t+1}\right\|_{\mathrm{F}}^2\right].
\end{aligned}$$

□

Lemma 14. *We have the following descent on Ω_3^t*

$$\Omega_3^{t+1} \quad \leq (1 - \gamma\rho/2)\Omega_3^t + (1 + 2/\gamma\rho)2\gamma^2 C\Omega_1^t + (1 + 2/\gamma\rho)2\eta^2\Omega_4^t. \tag{24}$$

Proof.

$$
\begin{aligned}
\Omega_3^{t+1} &= \mathbb{E}\left[\left\|\mathbf{X}^{t+1} - \bar{\mathbf{x}}^{t+1}\mathbf{1}^\top\right\|_{\mathrm{F}}^2\right] \\
&= \mathbb{E}\left[\left\|\mathbf{X}^t + \gamma\mathbf{H}^t(\mathbf{W}-\mathbf{I}) - \eta\mathbf{V}^t - \bar{\mathbf{x}}^t\mathbf{1}^T + \eta\bar{\mathbf{v}}^t\mathbf{1}^T\right\|_{\mathrm{F}}^2\right] \\
&= \mathbb{E}\left[\left\|\mathbf{X}^t\widetilde{\mathbf{W}} - \bar{\mathbf{x}}^t\mathbf{1}^T + \gamma(\mathbf{H}^t - \mathbf{X}^t)(\mathbf{W}-\mathbf{I}) - \eta\mathbf{V}^t + \eta\bar{\mathbf{v}}^t\mathbf{1}^\top\right\|_{\mathrm{F}}^2\right] \\
&\leq (1+\beta)(1-\gamma\rho)\mathbb{E}\left[\left\|\mathbf{X}^t - \bar{\mathbf{x}}^t\mathbf{1}^\top\right\|_{\mathrm{F}}^2\right] + (1+\beta^{-1})(2\gamma^2\mathbb{E}\left[\left\|(\mathbf{H}^t - \mathbf{X}^t)(\mathbf{W}-\mathbf{I})\right\|_{\mathrm{F}}^2\right] \\
&\quad + 2\eta^2\mathbb{E}\left[\left\|\mathbf{V}^t - \bar{\mathbf{v}}^t\mathbf{1}^\top\right\|_{\mathrm{F}}^2\right]) \\
&\leq (1-\gamma\rho/2)\mathbb{E}\left[\left\|\mathbf{X}^t - \bar{\mathbf{x}}^t\mathbf{1}^\top\right\|_{\mathrm{F}}^2\right] + (1+2/\gamma\rho)(2\gamma^2 C\mathbb{E}\left[\left\|\mathbf{H}^t - \mathbf{X}^t\right\|_{\mathrm{F}}^2\right] \\
&\quad + 2\eta^2\mathbb{E}\left[\left\|\mathbf{V}^t - \bar{\mathbf{v}}^t\mathbf{1}^\top\right\|_{\mathrm{F}}^2\right]) \\
&= (1-\gamma\rho/2)\Omega_3^t + (1+2/\gamma\rho)2\gamma^2 C\Omega_1^t + (1+2/\gamma\rho)2\eta^2\Omega_4^t.
\end{aligned}
$$

where $\beta = \frac{\gamma\rho/2}{1-\gamma\rho}$ and we define $\widetilde{\mathbf{W}} \coloneqq \mathbf{I} + \gamma(\mathbf{W}-\mathbf{I})$ which has a spectral gap at least $\gamma\rho$ by Lemma 7. □

Lemma 15. *We have the following descent on* Ω_4^t

$$
\begin{aligned}
\Omega_4^{t+1} \quad \leq \quad & (1-\gamma\rho/2)\mathbb{E}\left[\left\|\mathbf{V}^t - \bar{\mathbf{v}}^t\mathbf{1}^T\right\|_{\mathrm{F}}^2\right] \\
& + (1+2/\gamma\rho)\left(2\gamma^2 C\mathbb{E}\left[\left\|\mathbf{G}^t - \mathbf{V}^t\right\|_{\mathrm{F}}^2\right] + 2\mathbb{E}\left[\left\|\mathbf{M}^{t+1} - \mathbf{M}^t\right\|_{\mathrm{F}}^2\right]\right).
\end{aligned} \tag{25}
$$

Proof.

$$
\begin{aligned}
\Omega_4^{t+1} &= \mathbb{E}\left[\left\|\mathbf{V}^{t+1} - \bar{\mathbf{v}}^t\mathbf{1}^T + \bar{\mathbf{v}}^t\mathbf{1}^T - \bar{\mathbf{v}}^{t+1}\mathbf{1}^T\right\|_{\mathrm{F}}^2\right] \\
&= \mathbb{E}\left[\left\|\mathbf{V}^{t+1} - \bar{\mathbf{v}}^t\mathbf{1}^T\right\|_{\mathrm{F}}^2\right] - n\mathbb{E}\left[\left\|\bar{\mathbf{v}}^t - \bar{\mathbf{v}}^{t+1}\right\|^2\right] \\
&\leq \mathbb{E}\left[\left\|\mathbf{V}^{t+1} - \bar{\mathbf{v}}^t\mathbf{1}^T\right\|_{\mathrm{F}}^2\right] \\
&= \mathbb{E}\left[\left\|\mathbf{V}^t + \gamma\mathbf{G}^t(\mathbf{W}-\mathbf{I}) + \mathbf{M}^{t+1} - \mathbf{M}^t - \bar{\mathbf{v}}^t\mathbf{1}^T\right\|_{\mathrm{F}}^2\right] \\
&= \mathbb{E}\left[\left\|\mathbf{V}^t\widetilde{\mathbf{W}} - \bar{\mathbf{v}}^t\mathbf{1}^T + \gamma(\mathbf{G}^t - \mathbf{V}^t)(\mathbf{W}-\mathbf{I}) + \mathbf{M}^{t+1} - \mathbf{M}^t\right\|_{\mathrm{F}}^2\right] \\
&\leq (1-\gamma\rho/2)\mathbb{E}\left[\left\|\mathbf{V}^t - \bar{\mathbf{v}}^t\mathbf{1}^T\right\|_{\mathrm{F}}^2\right] + (1+2/\gamma\rho)(2\gamma^2 C\mathbb{E}\left[\left\|\mathbf{G}^t - \mathbf{V}^t\right\|_{\mathrm{F}}^2\right] \\
&\quad + 2\mathbb{E}\left[\left\|\mathbf{M}^{t+1} - \mathbf{M}^t\right\|_{\mathrm{F}}^2\right]).
\end{aligned}
$$

□

Lemma 16 (Lemma B.4, Eq. (18) from [Zhao et al., 2022])**.** *We have the following control of the iterates at iterations t and $t+1$*

$$
\mathbb{E}\left[\left\|\mathbf{X}^{t+1} - \mathbf{X}^t\right\|_{\mathrm{F}}^2\right] \leq 3\gamma^2 C\Omega_1^t + 3\gamma^2 C\Omega_3^t + 3\eta^2\Omega_4^t + 3\eta^2 n\Omega_5^t. \tag{26}
$$

Lemma 17. *Assume Assumptions 2 and 3 hold. Then we have the following control of the momentum at iterations t and $t+1$*

$$
\mathbb{E}\left[\left\|\mathbf{M}^{t+1} - \mathbf{M}^t\right\|_{\mathrm{F}}^2\right] \leq \lambda^2 n\sigma^2 + 2\lambda^2\mathbb{E}\left[\left\|\nabla F(\mathbf{X}^t) - \mathbf{M}^t\right\|_{\mathrm{F}}^2\right] + 2\lambda^2 L^2\mathbb{E}\left[\left\|\mathbf{X}^t - \mathbf{X}^{t+1}\right\|_{\mathrm{F}}^2\right]. \tag{27}
$$

Proof.

$$\begin{aligned}
\mathbb{E}\left[\left\|\mathbf{M}^{t+1}-\mathbf{M}^t\right\|_{\mathrm{F}}^2\right] &= \lambda^2 \mathbb{E}\left[\left\|\widetilde{\nabla} F(\mathbf{X}^{t+1})-\mathbf{M}^t\right\|_{\mathrm{F}}^2\right] \\
&= \lambda^2 \mathbb{E}\left[\left\|\widetilde{\nabla} F(\mathbf{X}^{t+1})-\nabla F(\mathbf{X}^{t+1})+\nabla F(\mathbf{X}^{t+1})-\mathbf{M}^t\right\|_{\mathrm{F}}^2\right] \\
&\leq \lambda^2 n\sigma^2+\lambda^2 \mathbb{E}\left[\left\|\nabla F(\mathbf{X}^{t+1})-\mathbf{M}^t\right\|_{\mathrm{F}}^2\right] \\
&\leq \lambda^2 n\sigma^2+2\lambda^2 \mathbb{E}\left[\left\|\nabla F(\mathbf{X}^{t})-\mathbf{M}^t\right\|_{\mathrm{F}}^2\right]+2\lambda^2 L^2 \mathbb{E}\left[\left\|\mathbf{X}^t-\mathbf{X}^{t+1}\right\|_{\mathrm{F}}^2\right].
\end{aligned}$$

□

Lemma 18. *We have the following control of the gradient estimator* $\mathbf{V}^t$ *at iterations* t *and* $t+1$

$$\mathbb{E}\left[\left\|\mathbf{V}^{t+1}-\mathbf{V}^t\right\|_{\mathrm{F}}^2\right] \leq 3\gamma^2 C\Omega_2^t+3\gamma^2 C\Omega_4^t+3\mathbb{E}\left[\left\|\mathbf{M}^{t+1}-\mathbf{M}^t\right\|_{\mathrm{F}}^2\right]. \tag{28}$$

Proof.

$$\begin{aligned}
\mathbb{E}\left[\left\|\mathbf{V}^{t+1}-\mathbf{V}^t\right\|_{\mathrm{F}}^2\right] &= \mathbb{E}\left[\left\|\gamma\mathbf{G}^t(\mathbf{W}-\mathbf{I})+\mathbf{M}^{t+1}-\mathbf{M}^t\right\|_{\mathrm{F}}^2\right] \\
&= \mathbb{E}\left[\left\|\gamma(\mathbf{G}^t-\mathbf{V}^t)(\mathbf{W}-\mathbf{I})+\gamma(\mathbf{V}^t-\bar{\mathbf{v}}^t\mathbf{1}^T)(\mathbf{W}-\mathbf{I})+\mathbf{M}^{t+1}-\mathbf{M}^t\right\|_{\mathrm{F}}^2\right] \\
&\leq 3\gamma^2 C\mathbb{E}\left[\left\|\mathbf{G}^t-\mathbf{V}^t\right\|_{\mathrm{F}}^2\right]+3\gamma^2 C\mathbb{E}\left[\left\|\mathbf{V}^t-\bar{\mathbf{v}}^t\mathbf{1}^T\right\|_{\mathrm{F}}^2\right] \\
&\quad + 3\mathbb{E}\left[\left\|\mathbf{M}^{t+1}-\mathbf{M}^t\right\|_{\mathrm{F}}^2\right] \\
&= 3\gamma^2 C\Omega_2^t+3\gamma^2 C\Omega_4^t+3\mathbb{E}\left[\left\|\mathbf{M}^{t+1}-\mathbf{M}^t\right\|_{\mathrm{F}}^2\right].
\end{aligned}$$

□

Theorem 1 (Convergence of MoTEF)**.** *Let Assumptions 1 to 3 hold. Then there exist absolute constants* $c_\gamma, c_\lambda, c_\eta$*, and some* $\tau \leq 1$ *such that if we set stepsizes* $\gamma = c_\gamma\alpha\rho, \lambda = c_\lambda\alpha\rho^3\tau, \eta = c_\eta L^{-1}\alpha\rho^3\tau$*, and choosing the initial batch size* $B_{\mathrm{init}} \geq \lceil \frac{LF^0}{\sigma^2} \rceil$*, then after at most*

$$T = \mathcal{O}\left(\frac{\sigma^2}{n\varepsilon^4}+\frac{\sigma}{\alpha\rho^2\varepsilon^3}+\frac{\sigma}{\alpha^{1/2}\rho^{3/2}\varepsilon^3}+\frac{\sigma}{\alpha\rho^{5/2}\varepsilon^3}+\frac{1}{\alpha\rho^3\varepsilon^2}\right)LF^0 \tag{10}$$

iterations of Algorithm 1 it holds $\mathbb{E}\left[\|\nabla f(\mathbf{x}_{\mathrm{out}})\|^2\right] \leq \varepsilon^2$*, where* $\mathbf{x}_{\mathrm{out}}$ *is chosen uniformly at random from* $\{\bar{\mathbf{x}}_0, \ldots, \bar{\mathbf{x}}_{T-1}\}$*, and* $\mathcal{O}$ *suppresses absolute constants.*

Proof. From Lemma 17 and Lemma 16 we get

$$\begin{aligned}
\mathbb{E}\left[\left\|\mathbf{M}^{t+1}-\mathbf{M}^t\right\|_{\mathrm{F}}^2\right] &\leq \lambda^2 n\sigma^2+2\lambda^2\widetilde{G}^t+6\lambda^2\gamma^2L^2C\Omega_1^t+6\lambda^2\gamma^2L^2C\Omega_3^t+6\lambda^2\eta^2L^2\Omega_4^t \\
&\quad +6\lambda^2\eta^2L^2n\Omega_5^t.
\end{aligned} \tag{29}$$

Using the above and Lemma 18 we get

$$\begin{aligned}
\mathbb{E}\left[\left\|\mathbf{V}^{t+1}-\mathbf{V}^t\right\|_{\mathrm{F}}^2\right] &\leq 3\lambda^2 n\sigma^2+6\lambda^2\widetilde{G}^t+18\lambda^2\gamma^2L^2C\Omega_1^t+3\gamma^2C\Omega_2^t+18\lambda^2\gamma^2L^2C\Omega_3^t \\
&\quad +(3\gamma^2C+18\lambda^2\eta^2L^2)\Omega_4^t+18\lambda^2\eta^2L^2n\Omega_5^t.
\end{aligned} \tag{30}$$

Using (29), (30), and Lemma 10 we get the following descent on $\hat{G}^t$

$$\hat{G}^{t+1} \leq (1-\lambda)\hat{G}^t + \frac{3L^2 n\gamma^2 C}{\lambda}\Omega_1^t + \frac{3L^2 n\gamma^2 C}{\lambda}\Omega_3^t + \frac{3L^2 n\eta^2}{\lambda}\Omega_4^t + \frac{3L^2 n^2\eta^2}{\lambda}\Omega_5^t + \lambda^2 n\sigma^2.$$

Using (29), (30), and Lemma 11 we get the following descent on $\widetilde{G}^t$

$$\widetilde{G}^{t+1} \leq (1-\lambda)\widetilde{G}^t + \frac{3L^2\gamma^2 C}{\lambda}\Omega_1 + \frac{3L^2\gamma^2 C}{\lambda}\Omega_3^t + \frac{3L^2\eta^2}{\lambda}\Omega_4^t + \frac{3L^2 n\eta^2}{\lambda}\Omega_5^t + \lambda^2 n\sigma^2.$$

Using (29), (30), and Lemma 12 we get the following descent on Ω_1^t

$$\Omega_1^{t+1} \leq \left(1 - \frac{\alpha}{2} + \frac{6\gamma^2 C}{\alpha}\right)\Omega_1^t + \frac{6\gamma^2 C}{\alpha}\Omega_3^t + \frac{6\eta^2}{\alpha}\Omega_4^t + \frac{6\eta^2 n}{\alpha}\Omega_5^t.$$

Using (29), (30), and Lemma 13 we get the following descent on Ω_2^t

$$\begin{aligned}\Omega_2^{t+1} \leq& \left(1 - \frac{\alpha}{2} + \frac{6\gamma^2 C}{\alpha}\right)\Omega_2^t + \frac{6\lambda^2}{\alpha}\widetilde{G}^t + \frac{36\lambda^2\gamma^2 L^2 C}{\alpha}\Omega_1^t + \frac{36\lambda^2\gamma^2 L^2 C}{\alpha}\Omega_3^t \\ &+ \left(\frac{6\gamma^2 C}{\alpha} + \frac{36\eta^2\lambda^2 L^2}{\alpha}\right)\Omega_4^t + \frac{36\eta^2\lambda^2 L^2 n}{\alpha}\Omega_5^t + \frac{6\lambda^2 n}{\alpha}\sigma^2.\end{aligned}$$

Using (29), (30), and Lemma 14 we get the following descent on Ω_3^t

$$\Omega_3^{t+1} \leq (1 - \frac{\gamma\rho}{2})\Omega_3^t + \frac{6\gamma C}{\rho}\Omega_1^t + \frac{6\eta^2}{\gamma\rho}\Omega_4^t. \tag{31}$$

Finally, using (29), (30), and Lemma 15 we get the following descent on Ω_4^t:

$$\begin{aligned}\Omega_4^{t+1} \leq& (1 - \frac{\gamma\rho}{2} + \frac{36\eta^2\lambda^2 L^2}{\gamma\rho})\Omega_4^t + \frac{12\lambda^2}{\gamma\rho}\widetilde{G}^t + \frac{36\gamma\lambda^2 L^2 C}{\rho}\Omega_1^t + \frac{6\gamma C}{\rho}\Omega_2^t + \frac{36\gamma\lambda^2 L^2 C}{\rho}\Omega_3^t \\ &+ \frac{36\eta^2\gamma L^2 n}{\rho}\Omega_5^t + \frac{6\lambda^2 n}{\gamma\rho}\sigma^2.\end{aligned}$$

Now we can gather all together

$$\begin{aligned}\mathbf{\Omega}^{t+1} \leq& \underbrace{\begin{pmatrix} 1-\lambda & 0 & \frac{3L^2 n\gamma^2 C}{\lambda} & 0 & \frac{3L^2 n\gamma^2 C}{\lambda} & \frac{3L^2 n\eta^2}{\lambda} \\ 0 & 1-\lambda & \frac{3L^2\gamma^2 C}{\lambda} & 0 & \frac{3L^2\gamma^2 C}{\lambda} & \frac{3L^2\eta^2}{\lambda} \\ 0 & 0 & 1-\frac{\alpha}{2}+\frac{6\gamma^2 C}{\alpha} & 0 & \frac{6\gamma^2 C}{\alpha} & \frac{6\eta^2}{\alpha} \\ 0 & \frac{6\lambda^2}{\alpha} & \frac{36\lambda^2\gamma^2 L^2 C}{\alpha} & 1-\frac{\alpha}{2}+\frac{6\gamma^2 C}{\alpha} & \frac{36\lambda^2\gamma^2 L^2 C}{\alpha} & \frac{6\gamma^2 C}{\alpha}+\frac{36\lambda^2\eta^2 L^2}{\alpha} \\ 0 & 0 & \frac{6\gamma C}{\rho} & 0 & 1-\frac{\gamma\rho}{2} & \frac{6\eta^2}{\gamma\rho} \\ 0 & \frac{12\lambda^2}{\gamma\rho} & \frac{36\gamma\lambda^2 L^2 C}{\rho} & \frac{6\gamma C}{\rho} & \frac{36\gamma\lambda^2 L^2 C}{\rho} & 1-\frac{\gamma\rho}{2}+\frac{36\eta^2\lambda^2 L^2}{\gamma\rho} \end{pmatrix}}_{:=\mathbf{A}} \mathbf{\Omega}^t \\ &+ \underbrace{\begin{pmatrix} \frac{3L^2 n^2\eta^2}{\lambda} \\ \frac{3L^2 n\eta^2}{\lambda} \\ \frac{6\eta^2 n}{\alpha} \\ \frac{36\eta^2\lambda^2 L^2 n}{\alpha} \\ 0 \\ \frac{36\eta^2\gamma L^2 n}{\rho} \end{pmatrix}}_{:=\mathbf{b}_1} \Omega_5^t + \underbrace{\begin{pmatrix} n \\ 2n \\ 0 \\ \frac{6n}{\alpha} \\ 0 \\ \frac{6n}{\gamma\rho} \end{pmatrix}}_{:=\mathbf{b}_2} \lambda^2\sigma^2.\end{aligned} \tag{32}$$

We remind that the Lyapunov function Φ^t has the following form

$$\Phi^t \coloneqq F^t + \frac{c_1}{n^2 L}\hat{G}^t + \frac{c_2\tau}{nL}\widetilde{G}^t + \frac{c_3 L}{\rho^3 n\tau}\Omega_1^t + \frac{c_4\tau}{\rho n L}\Omega_2^t + \frac{c_5 L}{\rho^3 n\tau}\Omega_3^t + \frac{c_6\tau}{\rho n L}\Omega_4^t = F^t + \mathbf{c}^\top \mathbf{\Omega}^t,$$

where $\{c_k\}_{k=1}^6$ are absolute constants. Let

$$\mathbf{c} \coloneqq \left(\frac{c_1}{n^2 L}, \frac{c_2\tau}{nL}, \frac{c_3 L}{\rho^3 n\tau}, \frac{c_4\tau}{\rho n L}, \frac{c_5 L}{\rho^3 n\tau}, \frac{c_6\tau}{\rho n L}\right)^\top.$$

Therefore, the descent on Φ^t for is the following

$$\begin{aligned}
\Phi^{t+1} &= F^{t+1} + \mathbf{c}^\top \mathbf{\Omega}^t \\
&\leq F_t - \frac{\eta}{2}\mathbb{E}\left[\left\|\nabla f(\bar{\mathbf{x}}^t)\right\|^2\right] + \frac{\eta}{n^2}\hat{G}^t + \frac{\eta L^2}{n}\Omega_3^t - (\eta/2 - \eta^2 L/2)\Omega_5^t \\
&+ \ \mathbf{c}^\top(\mathbf{A}\mathbf{\Omega}^t + \Omega_5^t \mathbf{b}_1 + \lambda^2\sigma^2\mathbf{b}_2) \\
&= F^t - \frac{\eta}{2}\mathbb{E}\left[\left\|\nabla f(\bar{\mathbf{x}}^t)\right\|^2\right] + \mathbf{c}^\top\mathbf{\Omega}^t + (\mathbf{q}^\top + \mathbf{c}^\top\mathbf{A} - \mathbf{c}^\top)\mathbf{\Omega}^t - (\eta/2 - \eta^2 L/2 - \mathbf{c}^\top\mathbf{b}_1)\Omega_5^t \\
&+ \ \mathbf{c}^\top\mathbf{b}_2\lambda^2\sigma^2 \\
&= \Phi^t - \frac{\eta}{2}\mathbb{E}\left[\left\|\nabla f(\bar{\mathbf{x}}^t)\right\|^2\right] + (\mathbf{q}^\top + \mathbf{c}^\top\mathbf{A} - \mathbf{c}^\top)\mathbf{\Omega}^t - (\eta/2 - \eta^2 L/2 - \mathbf{c}^\top\mathbf{b}_1)\Omega_5^t \\
&+ \ \mathbf{c}^\top\mathbf{b}_2\lambda^2\sigma^2,
\end{aligned}$$

where $\mathbf{q} \coloneqq (\eta/n^2, 0, 0, 0, \eta L^2/n, 0)^\top$. We need coefficients next to $\mathbf{\Omega}^t$ and Ω_5^t to be negative. This is equivalent to finding $\mathbf{c}$ such that

$$\begin{bmatrix} \mathbf{I} - \mathbf{A}^\top \\ -\mathbf{b}_1^\top \end{bmatrix} \mathbf{c} \geq \begin{bmatrix} \mathbf{q} \\ \frac{\eta^2 L}{2} - \frac{\eta}{2} \end{bmatrix}. \tag{33}$$

We make the following choice of stepsizes

$$\lambda \coloneqq c_\lambda \alpha\rho^3\tau, \quad \gamma \coloneqq c_\gamma\alpha\rho, \quad \eta \coloneqq \frac{c_\eta\alpha\rho^3\tau}{L}.$$

with the following choice of constants:

$$\begin{gathered}
c_\lambda = \frac{1}{200}, c_\gamma = \frac{1}{200}, c_\eta = \frac{1}{100000}, \quad \text{and,} \\
c_1 = \frac{1}{500}, c_2 = \frac{13}{200000}, c_3 = \frac{1}{20}, c_4 = \frac{1}{400000}, c_5 = \frac{9}{100}, c_6 = \frac{1}{200000}.
\end{gathered} \tag{34}$$

The system of inequalities (33) are satisfied when $\tau \leq 1$.

Given the complexity of the inequalities and the choices of the parameters, we do not attempt to write down a proof for the correctness of the choices manually, instead, we verify these choices using the Symbolic Math Toolbox in MATLAB. We also perform such verification for our parameters and constants choices for MoTEF in PŁ case and MoTEF-VR. The code performing all the verification can be found at here. We also note that, when c_λ, c_γ and c_η are fixed, we can search for a feasible $\{c_i\}_{i\in[6]}$ efficiently using the Linear Program solver with MATLAB as well. But searching for a feasible set of choices for c_λ, c_γ and c_η is very much a trial-and-error process.

Note that this choice gives us both λ and γ smaller than 1. This choice of constants gives the following result

$$\begin{aligned}
\Phi^{t+1} &\leq \Phi^t - \frac{c_\eta \alpha\rho^3\tau}{2L}\mathbb{E}\left[\|\nabla f(\bar{\mathbf{x}}^t)\|^2\right] + \frac{c_1}{n^2 L}\cdot n c_\lambda^2 \alpha^2\rho^6\tau^2\sigma^2 \\
&\quad + \frac{c_2\tau}{nL}\cdot 2n c_\lambda^2\alpha^2\rho^6\tau^2\sigma^2 \\
&\quad + \frac{c_4\tau}{\rho n L}\cdot\frac{6n}{\alpha}c_\lambda^2\alpha^2\rho^6\tau^2\sigma^2 \\
&\quad + \frac{c_6\tau}{\rho n L}\cdot\frac{6n}{c_\gamma\alpha\rho}c_\lambda^2\alpha^2\rho^6\tau^2\sigma^2 \\
&= \Phi^t - \frac{c_\eta \alpha\rho^3\tau}{2L}\mathbb{E}\left[\|\nabla f(\bar{\mathbf{x}}^t)\|^2\right] + \frac{c_\lambda^2 c_1\alpha^2\rho^6}{nL}\cdot\tau^2\sigma^2 \\
&\quad + \left(\frac{6c_\lambda^2 c_4\alpha\rho^5}{L} + \frac{2c_\lambda^2 c_2\alpha^2\rho^6}{L} + \frac{6c_6 c_\lambda^2\alpha\rho^4}{c_\gamma L}\right)\tau^3\sigma^2. \qquad (35)
\end{aligned}$$

By this, we proved Lemma 1. Let us define constants

$$\begin{aligned}
B &\coloneqq \frac{c_\eta\alpha\rho^3}{2L}, \\
C &\coloneqq \frac{c_\lambda^2 c_1\alpha^2\rho^6}{nL}, \\
D &\coloneqq \left(\frac{6c_\lambda^2 c_4\alpha\rho^5}{L} + \frac{2c_\lambda^2 c_2\alpha^2\rho^6}{L} + \frac{6c_6 c_\lambda^2\alpha\rho^4}{c_\gamma L}\right), \\
E &\coloneqq 1.
\end{aligned}$$

Using $\tau \leq E$ and unrolling (35) for T iterations we get

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla f(\bar{\mathbf{x}}^t)\|^2\right] \leq \frac{\Phi^0}{\tau B T} + \frac{C}{B}\tau\sigma^2 + \frac{D}{B}\tau^2\sigma^2.$$

So we need to choose $\tau = \min\left\{\frac{1}{E}, \left(\frac{\Phi^0}{CT\sigma^2}\right)^{1/2}, \left(\frac{\Phi^0}{DT\sigma^2}\right)^{1/3}\right\}$ and we get the following rate

$$\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla f(\bar{\mathbf{x}}^t)\|^2\right] &\leq \mathcal{O}\left(\frac{\Phi^0 E}{BT} + \left(\frac{C\Phi^0\sigma^2}{B^2 T}\right)^{1/2} + \left(\frac{\sqrt{D}\Phi^0\sigma}{B^{3/2}T}\right)^{2/3}\right) \\
&= \mathcal{O}\left(\frac{L\Phi^0}{\alpha\rho^3 T} + \left(\frac{L\Phi^0\sigma^2}{nT}\right)^{1/2}\right. \\
&\qquad \left. + \left(\frac{\sqrt{\alpha\rho^5 + \alpha^2\rho^6 + \alpha\rho^4}L\Phi^0\sigma}{\alpha^{3/2}\rho^{9/2}T}\right)^{2/3}\right) \\
&= \mathcal{O}\left(\frac{L\Phi^0}{T} + \left(\frac{L\Phi^0\sigma^2}{nT}\right)^{1/2} + \left(\frac{(\rho^{1/2} + \alpha^{1/2}\rho + 1)L\Phi^0\sigma}{\alpha\rho^{7/2}T}\right)^{2/3}\right),
\end{aligned}$$

that translates to the rate in terms of ε to

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla f(\bar{\mathbf{x}}^t)\|^2\right] \leq \varepsilon^2 \Rightarrow T = \mathcal{O}\left(\frac{L\Phi^0}{\alpha\rho^3\varepsilon^2} + \frac{L\Phi^0\sigma^2}{n\varepsilon^4} + \frac{L\Phi^0\sigma}{\alpha\rho^2\varepsilon^3} + \frac{L\Phi^0\sigma}{\alpha^{1/2}\rho^{3/2}\varepsilon^3} + \frac{L\Phi^0\sigma}{\alpha\rho^{5/2}\varepsilon^3}\right).$$

Note that with the choice $\mathbf{V}^0 = \mathbf{G}^0 = \mathbf{M}^0 = \widetilde{\nabla} F(\mathbf{X}^0), \mathbf{H}^0 = \mathbf{X}^0 = \mathbf{x}^0 \mathbf{1}^\top$, we get

$$\hat{G}^0 \leq \sigma^2 n, \quad \widetilde{G}^0 \leq \sigma^2 n, \quad \Omega_1^0 = \Omega_2^0 = \Omega_3^0 = \Omega_4^0 = 0.$$

$$\Phi^0 \leq F^0 + \frac{c_1}{n^2 L}\sigma^2 n + \frac{c_2 \tau}{nL}\sigma^2 n. \tag{36}$$

If we choose the initial batch size $B_{\text{init}} \geq \lceil \frac{\sigma^2}{LF^0} \rceil$, we get

$$\Phi^0 \leq F^0 + \frac{1}{nL}\frac{\sigma^2}{B_{\text{init}}} + \frac{1}{L}\frac{\sigma^2}{B_{\text{init}}} \leq 3F^0. \tag{37}$$

□

A.2 PŁ setting.

Theorem 4 (Convergence of MoTEF)**.** *Let Assumptions 1 to 4 hold. Then there exist absolute constants $c_\gamma, c_\lambda, c_\eta$, and some $\tau \leq 1$ such that if we set stepsizes $\gamma = c_\gamma \alpha\rho, \lambda = c_\lambda \alpha \rho^3 \tau, \eta = c_\eta L^{-1}\alpha\rho^3\tau$, and choosing the initial batch size $B_{\text{init}} \geq \lceil \frac{LF^0}{\sigma^2} \rceil$, then after at most*

$$T = \widetilde{\mathcal{O}}\left(\frac{L\sigma^2}{\mu^2 n\varepsilon} + \frac{L\sigma}{\alpha\rho^2\mu^{3/2}\varepsilon^{1/2}} + \frac{L\sigma}{\alpha\rho^{5/2}\mu^{3/2}\varepsilon^{1/2}} + \frac{L\sigma}{\alpha\rho^2\mu^{3/2}\varepsilon^{1/2}} + \frac{L}{\mu\alpha\rho^3}\right) \tag{38}$$

iterations of Algorithm 1 it holds $\mathbb{E}\left[f(\mathbf{x}^T) - f^*\right] \leq \varepsilon$, and $\widetilde{\mathcal{O}}$ suppresses absolute constants and poly-logarithmic factors.

Proof. The only change in the proof is the descent of the Lyapunov function. In PŁ case, the descent on Φ^t becomes

$$\begin{aligned}
\Phi^{t+1} &= F^{t+1} + \mathbf{b}^\top \boldsymbol{\Omega}^t \\
&\leq F_t - \frac{\eta}{2}\mathbb{E}\left[\left\|\nabla f(\bar{\mathbf{x}}^t)\right\|^2\right] + \frac{\eta}{n^2}\hat{G}^t + \frac{\eta L^2}{n}\Omega_3^t - ({\eta}/{2} - {\eta^2 L}/{2})\Omega_5^t \\
&\quad + \mathbf{b}^\top(\mathbf{A}\boldsymbol{\Omega}^t + \Omega_5^t \mathbf{b}_1 + \lambda^2\sigma^2\mathbf{b}_2) \\
&\leq (1-\eta\mu)F^t + (1-\eta\mu)\mathbf{b}^\top\boldsymbol{\Omega}^t + (\mathbf{q}^\top + \mathbf{b}^\top\mathbf{A} - (1-\eta\mu)\mathbf{b}^\top)\boldsymbol{\Omega}^t \\
&\quad - ({\eta}/{2} - {\eta^2 L}/{2} - \mathbf{b}^\top\mathbf{b}_1)\Omega_5^t + \mathbf{b}^\top\mathbf{b}_2\lambda^2\sigma^2 \\
&= (1-\eta\mu)\Phi^t + (\mathbf{q}^\top + \mathbf{b}^\top\mathbf{A} - (1-\eta\mu)\mathbf{b}^\top)\boldsymbol{\Omega}^t - ({\eta}/{2} - {\eta^2 L}/{2} - \mathbf{b}^\top\mathbf{b}_1)\Omega_5^t + \mathbf{b}^\top\mathbf{b}_2\lambda^2\sigma^2,
\end{aligned}$$

where in the second inequality we use PŁ condition. Similar to the proof of Theorem 1, we need to satisfy

$$\begin{bmatrix} (1-\mu\eta)\mathbf{I} - \mathbf{A}^\top \\ -\mathbf{b}_1^\top \end{bmatrix} \mathbf{b} \geq \begin{bmatrix} \mathbf{q} \\ \frac{\eta^2 L}{2} - \frac{\eta}{2} \end{bmatrix}$$

for some coefficients $\mathbf{b}$. We set the stepsizes such that

$$\lambda \coloneqq c_\lambda \alpha\rho^3\tau, \quad \gamma \coloneqq c_\gamma\alpha\rho, \quad \eta \coloneqq \frac{c_\eta\alpha\rho^3\tau}{L},$$

and

$$\mathbf{b} \coloneqq \left(\frac{b_1}{n^2 L}, \frac{b_2\tau}{nL}, \frac{b_3 L}{\rho^3 n\tau}, \frac{b_4\tau}{\rho nL}, \frac{b_5 L}{\rho^3 n\tau}, \frac{b_6\tau}{\rho nL}\right)^\top$$

with the choice

$$c_\lambda = \frac{1}{200000}, c_\gamma = \frac{1}{200000}, c_\eta = \frac{1}{100000000},$$

and

$$b_1 = \frac{1}{250}, b_2 = \frac{13}{200000}, b_3 = \frac{1}{20}, b_4 = \frac{1}{400000}, b_5 = 2, b_6 = \frac{1}{200000},$$

gives the following descent on Φ^t (note that both γ and λ are smaller than 1 with this choice of constants)

$$\begin{aligned}\Phi^{t+1} &\leq \left(1 - \frac{c_\eta \alpha \rho^3 \tau \mu}{L}\right)\Phi^t + \frac{c_\lambda^2 b_1 \alpha^2 \rho^6}{nL} \cdot \tau^2 \sigma^2 \\ &\quad + \left(\frac{6c_\lambda^2 b_4 \alpha \rho^5}{L} + \frac{2c_\lambda^2 b_2 \alpha^2 \rho^6}{L} + \frac{6b_6 c_\lambda^2 \alpha \rho^4}{c_\gamma L}\right)\tau^3 \sigma^2. \qquad (39)\end{aligned}$$

Let us define constants

$$\begin{aligned} B &:= \frac{c_\eta \alpha \rho^3 \mu}{2L}, \\ C &:= \frac{c_\lambda^2 c_1 \alpha^2 \rho^6}{nL}, \\ D &:= \left(\frac{6c_\lambda^2 c_4 \alpha \rho^5}{L} + \frac{2c_\lambda^2 c_2 \alpha^2 \rho^6}{L} + \frac{6c_6 c_\lambda^2 \alpha \rho^4}{c_\gamma L}\right), \\ E &:= 1. \end{aligned}$$

Unrolling (39) for T iterations we get

$$\Phi^T \leq (1 - B\tau)^T \Phi^0 + \frac{C}{B\tau}\tau^2\sigma^2 + \frac{D}{B\tau}\tau^3\sigma^2 = (1 - B\tau)^T \Phi^0 + \frac{C}{B}\sigma^2 + \frac{D}{B\tau}\tau^3\sigma^2$$

where we use the fact that

$$\sum_{l=0}^{m-1} (1 - B\tau)^l = \frac{1 - (1 - B\tau)^m}{1 - (1 - B\tau)} \leq \frac{1}{B\tau}.$$

Choosing $\tau = \min\left\{\frac{1}{E}, \frac{1}{BT}\log\left(\frac{\Phi^0 B^2 T}{C\sigma^2}\right), \frac{1}{BT}\log\left(\frac{\Phi^0 B^3 T^2}{D\sigma^2}\right)\right\}$ leads to the following rate

$$\Phi^T \leq \widetilde{\mathcal{O}}\left(\exp\left(-\frac{B}{E}T\right)\Phi^0 + \frac{C\sigma^2}{B^2 T} + \frac{D\sigma^2}{B^3 T^2}\right).$$

We refer to [Mishchenko et al., 2020] for a more detailed derivation (proof of Corollary 1, page 20). To achieve $F^T \leq \varepsilon$, we need to perform

$$\begin{aligned} T &= \widetilde{\mathcal{O}}\left(\frac{E}{B} + \frac{C\sigma^2}{B^2\varepsilon} + \frac{\sqrt{D}\sigma}{B^{3/2}\varepsilon^{1/2}}\right) \\ &= \widetilde{\mathcal{O}}\left(\frac{L}{\mu\alpha\rho^3} + \frac{L\sigma^2}{\mu^2 n\varepsilon} + \frac{L\sigma}{\alpha\rho^2\mu^{3/2}\varepsilon^{1/2}} + \frac{L\sigma}{\alpha\rho^{5/2}\mu^{3/2}\varepsilon^{1/2}} + \frac{L\sigma}{\alpha\rho^2\mu^{3/2}\varepsilon^{1/2}}\right) \end{aligned}$$

iterations.

□

B Missing proofs for MoTEF-VR

In this section, we provide the proof of convergence of Algorithm 2. Note that in this case Lemma 9 remains unchanged.

Lemma 19. *Let Assumptions 3 and 5 hold. Then we have the following descent on $\hat{G}^t$*

$$\hat{G}^{t+1} \quad \leq \quad (1-\lambda)\hat{G}^t + 2\lambda^2\sigma^2 n + \ell^2 \mathbb{E}\left[\|\mathbf{X}^{t+1} - \mathbf{X}^t\|_{\mathrm{F}}^2\right]. \tag{40}$$

Proof. We have

$$\begin{aligned}
\hat{G}^{t+1} &= \mathbb{E}\left[\left\|\mathbf{M}^{t+1}\mathbf{1} - \nabla F(\mathbf{X}^{t+1})\mathbf{1}\right\|^2\right] \\
&= \mathbb{E}\left[\left\|[\widetilde{\nabla} F(\mathbf{X}^{t+1}, \Xi^{t+1}) + (1-\lambda)(\mathbf{M}^t - \widetilde{\nabla} F(\mathbf{X}^t, \Xi^{t+1}) - \nabla F(\mathbf{X}^{t+1})]\mathbf{1}\right\|^2\right] \\
&= \mathbb{E}\Big[\Big\|\Big(\lambda(\widetilde{\nabla} F(\mathbf{X}^{t+1}, \Xi^{t+1}) - \nabla F(\mathbf{X}^{t+1})) \\
&\qquad + (1-\lambda)(\widetilde{\nabla} F(\mathbf{X}^{t+1}, \Xi^{t+1}) - \nabla F(\mathbf{X}^{t+1}) + \nabla F(\mathbf{X}^t) - \widetilde{\nabla} F(\mathbf{X}^t, \Xi^{t+1})) \\
&\qquad + (1-\lambda)(\mathbf{M}^t - \nabla F(\mathbf{X}^t))\Big)\mathbf{1}\Big\|^2\Big] \\
&\leq (1-\lambda)^2\,\mathbb{E}\,\|(\mathbf{M}^t - \nabla F(\mathbf{X}^t))\mathbf{1}\|^2 \\
&\quad + 2\lambda^2\,\mathbb{E}\,\|(\widetilde{\nabla} F(\mathbf{X}^{t+1}, \Xi^{t+1}) - \nabla F(\mathbf{X}^{t+1}))\mathbf{1}\|^2 \\
&\quad + 2(1-\lambda)^2\,\mathbb{E}\,\|(\widetilde{\nabla} F(\mathbf{X}^{t+1}, \Xi^{t+1} - \nabla F(\mathbf{X}^{t+1}) + \nabla F(\mathbf{X}^t) - \widetilde{\nabla} F(\mathbf{X}^t, \Xi^{t+1}))\mathbf{1}\|^2 \\
&\leq (1-\lambda)\hat{G}^t + 2\lambda^2\sigma^2 n \\
&\quad + 2\,\mathbb{E}\,\|(\widetilde{\nabla} F(\mathbf{X}^{t+1}, \Xi^{t+1}) - \nabla F(\mathbf{X}^{t+1}) + \nabla F(\mathbf{X}^t) - \widetilde{\nabla} F(\mathbf{X}^t, \Xi^{t+1}))\mathbf{1}\|^2.
\end{aligned} \tag{41}$$

For the last term above we continue as follows

$$\begin{aligned}
&\mathbb{E}\left[\|(\widetilde{\nabla} F(\mathbf{X}^{t+1}, \Xi^{t+1}) - \nabla F(\mathbf{X}^{t+1}) + \nabla F(\mathbf{X}^t) - \widetilde{\nabla} F(\mathbf{X}^t, \Xi^{t+1}))\mathbf{1}\|^2\right] \\
&= \mathbb{E}\left[\left\|\sum_{i=1}^n \nabla f_i(\mathbf{x}_i^{t+1}, \xi_i^{t+1}) - \nabla f_i(\mathbf{x}_i^{t+1}) + \nabla f_i(\mathbf{x}_i^t) - \nabla f_i(\mathbf{x}_i^t, \xi_i^{t+1})\right\|^2\right] \\
&= \sum_{i=1}^n \mathbb{E}\left[\left\|\nabla f_i(\mathbf{x}_i^{t+1}, \xi_i^{t+1}) - \nabla f_i(\mathbf{x}_i^{t+1}) + \nabla f_i(\mathbf{x}_i^t) - \nabla f_i(\mathbf{x}_i^t, \xi_i^{t+1})\right\|^2\right] \\
&\leq \sum_{i=1}^n \mathbb{E}\left[\left\|\nabla f_i(\mathbf{x}_i^{t+1}, \xi_i^{t+1}) - \nabla f_i(\mathbf{x}_i^t, \xi_i^{t+1})\right\|^2\right] \\
&\leq \ell^2 \mathbb{E}\left[\|\mathbf{X}^{t+1} - \mathbf{X}^t\|_{\mathrm{F}}^2\right].
\end{aligned} \tag{42}$$

Therefore, from (41) we get

$$\hat{G}^{t+1} \quad \leq \quad (1-\lambda)\hat{G}^t + 2\lambda^2\sigma^2 n + \ell^2 \mathbb{E}\left[\|\mathbf{X}^{t+1} - \mathbf{X}^t\|_{\mathrm{F}}^2\right]. \tag{43}$$

□

Lemma 20. *Assume Assumptions 3 and 5 hold. Then we have the following descent on $\hat{G}^t$*

$$\widetilde{G}^{t+1} \quad \leq \quad (1-\lambda)\widetilde{G}^t + 2\lambda^2\sigma^2 n + \ell^2 \mathbb{E}\left[\|\mathbf{X}^{t+1} - \mathbf{X}^t\|_{\mathrm{F}}^2\right]. \tag{44}$$

Proof. The proof is similar to the one of Lemma 19. □

Note that Lemmas 12 to 16 and 18 do not change in this setting, thus, we do not repeat them.

Lemma 21. *Assume Assumptions 3 and 5 hold. Then we have the following control of momentum at iterations t and $t+1$*

$$\mathbb{E}\left[\|\mathbf{M}^{t+1}-\mathbf{M}^t\|_{\mathrm{F}}^2\right] \le \lambda^2\widetilde{G}^t + 2\lambda^2 n\sigma^2 + 2\ell^2\mathbb{E}\left[\|\mathbf{X}^{t+1}-\mathbf{X}^t\|_{\mathrm{F}}^2\right]. \tag{45}$$

Proof. Using the update of $\mathbf{M}^t$ we have

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{M}^{t+1}-\mathbf{M}^t\|_{\mathrm{F}}^2\right] &= \mathbb{E}\left[\|\widetilde{\nabla}F(\mathbf{X}^{t+1},\Xi^{t+1}) + (1-\lambda)(\mathbf{M}^t - \widetilde{\nabla}F(\mathbf{X}^t,\Xi^{t+1})) - \mathbf{M}^t\|_{\mathrm{F}}^2\right] \\
&= \mathbb{E}\left[\|\widetilde{\nabla}F(\mathbf{X}^{t+1},\Xi^{t+1}) - \lambda\mathbf{M}^t - (1-\lambda)\widetilde{\nabla}F(\mathbf{X}^t,\Xi^{t+1})\|_{\mathrm{F}}^2\right] \\
&= \mathbb{E}\left[\left\|\lambda(\nabla F(\mathbf{X}^t)-\mathbf{M}^t) + \lambda(\widetilde{\nabla}F(\mathbf{X}^t,\Xi^{t+1}) - \nabla F(\mathbf{X}^t))\right.\right. \\
&\qquad \left.\left. + (\widetilde{\nabla}F(\mathbf{X}^{t+1},\Xi^{t+1}) - \widetilde{\nabla}F(\mathbf{X}^t,\Xi^{t+1}))\right\|_{\mathrm{F}}^2\right] \\
&= \lambda^2\widetilde{G}^t + \mathbb{E}\left[\left\|\lambda(\widetilde{\nabla}F(\mathbf{X}^t,\Xi^{t+1}) - \nabla F(\mathbf{X}^t))\right.\right. \\
&\qquad \left.\left. + \ (\widetilde{\nabla}F(\mathbf{X}^{t+1},\Xi^{t+1}) - \widetilde{\nabla}F(\mathbf{X}^t,\Xi^{t+1}))\right\|_{\mathrm{F}}^2\right] \\
&\le \lambda^2\widetilde{G}^t + 2\lambda^2 n\sigma^2 + 2\ell^2\mathbb{E}\left[\|\mathbf{X}^{t+1}-\mathbf{X}^t\|_{\mathrm{F}}^2\right].
\end{aligned}$$

□

Theorem 3 (Convergence of MoTEF-VR)**.** *Let Assumptions 1, 3 and 5 hold. Then there exists absolute constants $c_\gamma, c_\lambda, c_\eta$ and some $\tau < 1$ such that if we stepsizes $\gamma = c_\gamma\alpha\rho, \lambda = c_\lambda n^{-1}\alpha^2\rho^6\tau^2, \eta = c_\eta\ell^{-1}\alpha\rho^3\tau$, and initial batch size $B_{\mathrm{init}} \ge \lceil\frac{\sigma^2}{LF^0\alpha\rho^3}\rceil$, then after at most*

$$T = \mathcal{O}\left(\frac{\sigma}{n\varepsilon^3} + \frac{\sigma^{2/3}}{n^{2/3}\varepsilon^{8/3}} + \frac{\sigma^{2/3}}{n\alpha^{1/3}\rho^{1/3}\varepsilon^{8/3}} + \frac{\sigma^{2/3}}{n\alpha^{1/3}\rho^{2/3}\varepsilon^{8/3}} + \frac{1}{\alpha\rho^3\varepsilon^2}\right)\ell F^0 \tag{15}$$

iterations of Algorithm 2 it holds $\mathbb{E}\left[\|\nabla f(\mathbf{x}_{\mathrm{out}})\|^2\right] \le \varepsilon^2$, where $\mathbf{x}_{\mathrm{out}}$ is chosen uniformly at random from $\{\bar{\mathbf{x}}_0,\cdots,\bar{\mathbf{x}}_{T-1}\}$, and $\mathcal{O}$ suppresses absolute constants and poly-logarithmic factors.

Proof. From Lemmas 16 and 21 we get

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{M}^{t+1}-\mathbf{M}^t\|_{\mathrm{F}}^2\right] &\le \lambda^2\widetilde{G}^t + 2\lambda^2 n\sigma^2 + 2\ell^2(3\gamma^2C\Omega_1^t + 3\gamma^2C\Omega_3^t + 3\eta^2\Omega_4^t + 3\eta^2 n\Omega_5^t) \\
&= 2\lambda^2 n\sigma^2 + \lambda^2\widetilde{G}^t + 6C\gamma^2\ell^2\Omega_1^t + 6C\gamma^2\ell^2\Omega_3^t + 6\eta^2\ell^2\Omega_4^t + 6n\eta^2\ell^2\Omega_5^t.
\end{aligned} \tag{46}$$

From the above inequality (46) and Lemma 18 we get

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{V}^{t+1}-\mathbf{V}^t\|_{\mathrm{F}}^2\right] &\le 3\gamma^2C\Omega_2^t + 3\gamma^2C\Omega_4^t \\
&\quad + \ 3\left(2\lambda^2 n\sigma^2 + \lambda^2\widetilde{G}^t + 6C\gamma^2\ell^2\Omega_1^t + 6C\gamma^2\ell^2\Omega_3^t + 6\eta^2\ell^2\Omega_4^t + 6n\eta^2\Omega_5^t\right) \\
&= 6\lambda^2 n\sigma^2 + 3\lambda^2\widetilde{G}^t + 18C\gamma^2\ell^2\Omega_1^t + 3C\gamma^2\Omega_2^t + 18C\gamma^2\ell^2\Omega_3^t \\
&\quad + \ (3C\gamma^2 + 18\eta^2\ell^2)\Omega_4^t + 18n\eta^2\ell^2\Omega_5^t.
\end{aligned} \tag{47}$$

From Lemmas 16 and 19 we get the following descent on $\hat{G}^t$

$$\begin{aligned}
\hat{G}^{t+1} &\le (1-\lambda)\hat{G}^t + 2\lambda^2\sigma^2 n + \ell^2(3\gamma^2C\Omega_1^t + 3\gamma^2C\Omega_3^t + 3\eta^2\Omega_4^t + 3\eta^2 n\Omega_5^t) \\
&= 2\lambda^2\sigma^2 n + (1-\lambda)\hat{G}^t + 3C\gamma^2\ell^2\Omega_1^t + 3C\gamma^2\ell^2\Omega_3^t + 3\eta^2\ell^2\Omega_4^t + 3n\eta^2\ell^2\Omega_5^t.
\end{aligned}$$

Similarly, from Lemmas 16 and 20 we get the following descent on $\widetilde{G}^t$

$$\widetilde{G}^{t+1} \leq 2\lambda^2\sigma^2 n + (1-\lambda)\widetilde{G}^t + 3C\gamma^2\ell^2\Omega_1^t + 3C\gamma^2\ell^2\Omega_3^t + 3\eta^2\ell^2\Omega_4^t + 3n\eta^2\ell^2\Omega_5^t.$$

From Lemmas 12 and 16 we get the following descent on Ω_1^t

$$\begin{aligned}
\Omega_1^{t+1} &\leq (1-\alpha/2)\mathbb{E}\left[\|\mathbf{H}^t - \mathbf{X}^t\|_{\mathrm{F}}^2\right] + \frac{2}{\alpha}(3\gamma^2 C\Omega_1^t + 3\gamma^2 C\Omega_3^t + 3\eta^2\Omega_4^t + 3\eta^2 n\Omega_5^t) \\
&= (1-\alpha/2 + 6C\gamma^2/\alpha)\Omega_1^t + \frac{6C\gamma^2}{\alpha}\Omega_3^t + \frac{6\eta^2}{\alpha}\Omega_4^t + \frac{6n\eta^2}{\alpha}\Omega_5^t.
\end{aligned}$$

From Lemma 13 and (47) we get the following descent on Ω_2^t

$$\begin{aligned}
\Omega_2^{t+1} \leq\ & (1-\alpha/2)\Omega_2^t + \frac{2}{\alpha}\Big(6\lambda^2 n\sigma^2 + 3\lambda^2\widetilde{G}^t + 18C\gamma^2\ell^2\Omega_1^t + 3C\gamma^2\Omega_2^t + 18C\gamma^2\ell^2\Omega_3^t \\
&+ (3C\gamma^2 + 18C\eta^2\ell^2)\Omega_4^t + 18n\eta^2\ell^2\Omega_5^t.\Big) \\
=\ & \frac{12n\lambda^2}{\alpha}\sigma^2 + \frac{6\lambda^2}{\alpha}\widetilde{G}^t + \frac{36C\gamma^2\ell^2}{\alpha}\Omega_1^t + (1-\alpha/2 + 6C\gamma^2/\alpha)\Omega_2^t + \frac{36C\gamma^2\ell^2}{\alpha}\Omega_3^t \\
&+ \frac{2}{\alpha}(3C\gamma^2 + 18\eta^2\ell^2)\Omega_4^t + \frac{36n\eta^2\ell^2}{\alpha}\Omega_5^t.
\end{aligned}$$

The descent on Ω_3^t (31) from the proof of MoTEF remains unchanged

$$\Omega_3^{t+1} \leq (1-\frac{\gamma\rho}{2})\Omega_3^t + \frac{6\gamma C}{\rho}\Omega_1^t + \frac{6\eta^2}{\gamma\rho}\Omega_4^t.$$

From Lemma 15 and (46) we get the following descent on Ω_4^t

$$\begin{aligned}
\Omega_4^{t+1} \leq\ & (1-\gamma\rho/2)\Omega_4^t + 2\gamma^2 C(1+2/\gamma\rho)\Omega_2^t \\
&+ 2(1+2/\gamma\rho)(2\lambda^2 n\sigma^2 + \lambda^2\widetilde{G}^t + 6C\gamma^2\ell^2\Omega_1^t + 6C\gamma^2\ell^2\Omega_3^t + 6\eta^2\ell^2\Omega_4^t + 6n\eta^2\Omega_5^t) \\
\leq\ & \frac{6n\lambda^2}{\gamma\rho}\sigma^2 + \frac{3\lambda^2}{\gamma\rho}\widetilde{G}^t + \frac{18C\gamma\ell^2}{\rho}\Omega_1^t + \frac{6C\gamma}{\rho}\Omega_2^t + \frac{18C\gamma\ell^2}{\rho}\Omega_3^t + (1-\gamma\rho/2 + 18\eta^2\ell^2/\gamma\rho)\Omega_4^t \\
&+ \frac{18n\eta^2\ell^2}{\gamma\rho}\Omega_5^t.
\end{aligned}$$

We remind that $\mathbf{\Omega} = (\hat{G}^t, \widetilde{G}^t, \Omega_1^t \Omega_2^t, \Omega_3^t, \Omega_4^t)^\top$. Now we can gather all inequalities together

$$\mathbf{\Omega}^{t+1} \leq \underbrace{\begin{pmatrix} 1-\lambda & 0 & 3C\gamma^2\ell^2 & 0 & 3C\gamma^2\ell^2 & 3\eta^2\ell^2 \\ 0 & 1-\lambda & 3C\gamma^2\ell^2 & 0 & 3C\gamma^2\ell^2 & 3\eta^2\ell^2 \\ 0 & 0 & 1-\frac{\alpha}{2}+\frac{6C\gamma^2}{\alpha} & 0 & \frac{6C\gamma^2}{\alpha} & \frac{6\eta^2}{\alpha} \\ 0 & \frac{6\lambda^2}{\alpha} & \frac{36C\gamma^2\ell^2}{\alpha} & 1-\frac{\alpha}{2}+\frac{6C\gamma^2}{\alpha} & \frac{36C\gamma^2\ell^2}{\alpha} & \frac{6C\gamma^2}{\alpha}+\frac{36\eta^2\ell^2}{\alpha} \\ 0 & 0 & \frac{6\gamma C}{\rho} & 0 & 1-\frac{\gamma\rho}{2} & \frac{6\eta^2}{\gamma\rho} \\ 0 & \frac{3\lambda^2}{\gamma\rho} & \frac{18C\gamma\ell^2}{\rho} & \frac{6C\gamma}{\rho} & \frac{18C\gamma\ell^2}{\rho} & 1-\frac{\gamma\rho}{2}+\frac{18\eta^2\ell^2}{\gamma\rho} \end{pmatrix}}_{:=\mathbf{A}} \mathbf{\Omega}^t$$

$$+ \underbrace{\begin{pmatrix} 3n\eta^2\ell^2 \\ 3n\eta^2\ell^2 \\ \frac{6n\eta^2}{\alpha} \\ \frac{36n\eta^2\ell^2}{\alpha} \\ 0 \\ \frac{18n\eta^2\ell^2}{\gamma\rho} \end{pmatrix}}_{:=\mathbf{b}_1} \Omega_5^t + \underbrace{\begin{pmatrix} 2n \\ 2n \\ 0 \\ \frac{12n}{\alpha} \\ 0 \\ \frac{6n}{\gamma\rho} \end{pmatrix}}_{:=\mathbf{b}_2} \lambda^2\sigma^2. \tag{48}$$

Now we consider the following choice of stepsizes

$$\lambda := \frac{c_\lambda \alpha^2\rho^6\tau^2}{n}, \quad \gamma := c_\gamma \alpha\rho, \quad \eta := \frac{c_\eta \alpha\rho^3\tau}{\ell},$$

and constants

$$\mathbf{d} := \left(\frac{d_1}{\alpha\rho^3 n\tau\ell}, \frac{d_2}{n\ell}, \frac{d_3\ell}{\rho^3 n\tau}, \frac{d_4}{\rho n\ell}, \frac{d_5\ell}{\rho^3 n\tau}, \frac{d_6}{\rho n\ell}\right)^\top,$$

where

$$c_\lambda = \frac{1}{200}, c_\gamma = \frac{1}{200}, c_\eta = \frac{1}{100000},$$
$$c_1 = 0.0020, c_2 = 0.000065, c_3 = 0.005, c_4 = 0.0000025, c_5 = 0.01, c_6 = 0.000005 \tag{49}$$

Note that choosing $\tau \leq 1$ makes the system of inequalities (48) hold. Using this choice, we get the following descent on $\Psi^t = F^t + \mathbf{d}^\top \mathbf{\Omega}^t$

$$\begin{aligned} \Psi^{t+1} \quad \leq \quad & \Psi^t - \frac{c_\eta \alpha\rho^3\tau}{2\ell}\mathbb{E}\left[\|\nabla f(\bar{\mathbf{x}}^t)\|^2\right] + \frac{c_1}{\alpha\rho^3 n\tau\ell} \cdot 2nc_\lambda^2\alpha^4\rho^{12}n^{-2}\tau^4\sigma^2 \\ & + \frac{c_2\tau}{n\ell} \cdot 2nc_\lambda^2\alpha^4\rho^{12}n^{-2}\tau^4\sigma^2 \\ & + \frac{c_4}{\rho n\ell} \cdot \frac{12n}{\alpha} c_\lambda^2\alpha^4\rho^{12}n^{-2}\tau^4\sigma^2 \\ & + \frac{c_6}{\rho n\ell} \cdot \frac{6n}{c_\gamma\alpha\rho} c_\lambda^2\alpha^4\rho^{12}\tau^4 n^{-2}\sigma^2 \\ = \quad & \Psi^t - \frac{c_\eta\alpha\rho^3}{2\ell}\tau\mathbb{E}\left[\|\nabla f(\bar{\mathbf{x}}^t)\|^2\right] + \frac{2c_1c_\lambda^2}{n^2\ell}\alpha^3\rho^9\tau^3\sigma^2 \\ & + \left(\frac{2c_2c_\lambda^2\alpha^4\rho^{12}}{n^2\ell} + \frac{12c_4c_\lambda^2\alpha^3\rho^{11}}{n^3\ell} + \frac{6c_6c_\lambda^2\alpha^3\rho^{10}}{n^3\ell}\right)\tau^4\sigma^2. \end{aligned} \tag{50}$$

By this, we proved Lemma 4. Let us define constants

$$\begin{aligned}
B &:= \frac{c_\eta \alpha \rho^3}{2\ell}, \\
C &:= \frac{2c_1 c_\lambda^2}{n^2\ell}\alpha^3\rho^9, \\
D &:= \left(\frac{2c_2c_\lambda^2\alpha^4\rho^{12}}{n^2\ell} + \frac{12c_4c_\lambda^2\alpha^3\rho^{11}}{n^3\ell} + \frac{6c_6c_\lambda^2\alpha^3\rho^{10}}{n^3\ell}\right), \\
E &:= 1.
\end{aligned}$$

Unrolling (50) for T iterations we get

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla f(\bar{\mathbf{x}}^t)\|^2\right] \le \frac{\Phi^0}{\tau BT} + \frac{C}{B}\tau^2\sigma^2 + \frac{D}{B}\tau^3\sigma^2.$$

Choosing $\tau = \min\left\{\frac{1}{E}, \left(\frac{\Psi^0}{C\sigma^2 T}\right)^{1/3}, \left(\frac{\Psi^0}{D\sigma^2 T}\right)^{1/4}\right\}$ gives the following rate

$$\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla f(\bar{\mathbf{x}}^t)\|^2\right] &\le \frac{E\Psi^0}{BT} + \left(\frac{\sqrt{C}\Psi^0\sigma}{B^{3/2}T}\right)^{2/3} + \left(\frac{D^{1/3}\Psi^0\sigma^{2/3}}{B^{4/3}T}\right)^{3/4} \\
&= \mathcal{O}\left(\frac{\ell\Psi^0}{\alpha\rho^3 T} + \left(\frac{\ell\Psi^0\sigma}{nT}\right)^{2/3}\right. \\
&\quad \left. + \left(\frac{(n^{-2/3} + \alpha^{-1/3}\rho^{-1/3}n^{-1} + \alpha^{-1/3}\rho^{-2/3}n^{-1})\ell\Psi^0\sigma^{2/3}}{T}\right)^{3/4}\right),
\end{aligned}$$

that translates into the rate in terms of ε to

$$\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla f(\bar{\mathbf{x}}^t)\|^2\right] \le \varepsilon^2 \Rightarrow \quad = \quad & \mathcal{O}\left(\frac{\ell\Psi^0}{\alpha\rho^3\varepsilon^2} + \frac{\ell\Psi^0\sigma}{n\varepsilon^3} + \frac{\ell\Psi^0\sigma^{2/3}}{n^{2/3}\varepsilon^{8/3}} + \frac{\ell\Psi^0\sigma^{2/3}}{\alpha^{1/3}\rho^{1/3}n\varepsilon^{8/3}}\right. \\
& \left. + \frac{\ell\Psi^0\sigma^{2/3}}{\alpha^{1/3}\rho^{2/3}n\varepsilon^{8/3}}\right).
\end{aligned}$$

Note that with the choice $\mathbf{V}^0 = \mathbf{G}^0 = \mathbf{M}^0 = \widetilde{\nabla}F(\mathbf{X}^0), \mathbf{H}^0 = \mathbf{X}^0 = \mathbf{x}^0\mathbf{1}^\top$, we get

$$\hat{G}^0 \le \sigma^2 n, \quad \widetilde{G}^0 \le \sigma^2 n, \quad \Omega_1^0 = \Omega_2^0 = \Omega_3^0 = \Omega_4^0 = 0.$$

$$\Psi^0 \le F^0 + \frac{d_1}{\alpha\rho^3 n\tau\ell}\sigma^2 n + \frac{d_2}{n\ell}\sigma^2 n. \tag{51}$$

If we choose the initial batch size $B_{\text{init}} \ge \lceil \frac{\sigma^2}{LF^0\alpha\rho^3} \rceil$, we get

$$\Psi^0 \le F^0 + \frac{1}{\alpha\rho^3\ell}\frac{\sigma^2}{B_{\text{init}}} + \frac{1}{\ell}\frac{\sigma^2}{B_{\text{init}}} \le 3F^0. \tag{52}$$

□

C Experiment details

C.1 Effect of changing heterogeneity

We perform a grid search for the parameters γ from $\{0.1, 0.01, 0.001\}$, η from the log space from 10^{-4} to 10^{-1} and the log space from 5×10^{-4} to 5×10^{-1}. For MoTEF we search the momentum parameter λ from the same log space as η as well.

C.2 Robustness to communication topology.

Next, we study the effect of the network topology on the convergence of MoTEF. We set $n = 40, \lambda = 0.05$, choose batch size 100, and run experiments for ring, star, grid, Erdös-Rènyi ($p = 0.2$ and $p = 0.5$) topologies. For all topologies, we use $\eta = 0.05, \gamma = 0.5, \lambda = 0.01$ for a9a dataset and $\eta = 0.05, \gamma = 0.5, \lambda = 0.01$ for w8a. Note that the spectral gaps of these networks $0.012, 0.049, 0.063, 0.467, 0.755$ correspondingly.

C.3 Hyperparameters for section 4.2

For MoTEF we tune stepsize as follows $\eta \in \{0.001, 0.01, 0.05\}, \gamma \in \{0.1, 0.2, 0.5, 0.9\}, \lambda \in \{0.005, 0.01, 0.05, 0.1\}$. For BEER we tune the stepsizes in the range $\eta \in \{0.001, 0.01, 0.05\}, \gamma \in \{0.1, 0.2, 0.5, 0.9\}$. For Choco-SGD we tune the stepsizes in the range $\eta \in \{0.01, 0.05\}, \gamma \in \{0.1, 0.5, 0.9\}$. Finally, for DSGD and D2 we choose the stepsize $\eta = 0.01$.