# A Tight Theory of Error Feedback Algorithms in Distributed Optimization

**Daniel Berg Thomsen** [1 2] **Adrien Taylor** [1] **Aymeric Dieuleveut** [2]

## Abstract

Communication costs are a major bottleneck in distributed learning and first-order optimization. A common approach to alleviate this issue is to compress the gradient information exchanged between agents. However, such compression typically degrades the convergence guarantees of gradient-based methods. *Error feedback* mechanisms provide a simple and computationally cheap remedy for this issue, but numerous variants have been proposed, and their relative performance remains poorly understood. This paper provides *tight* convergence analyses for two of the main error-feedback algorithms from the literature, namely error feedback and EF$^{21}$, by identifying optimal step-size choices, and by constructing a set of optimal Lyapunov functions tailored to each method. The results hold independently of the number of agents and recover the known best guarantees possible in the single-agent regime.

## 1. Introduction

The trend toward larger model usage in machine learning has made training in distributed environments a practical necessity. In many applications, including federated learning, data are partitioned across $n$ agents and a central server coordinates the process (McMahan et al., 2017; Kairouz et al., 2019). Consider the finite-sum problem

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

where only agent $i$ has access to first-order information about $f^{(i)}$. Communication follows a star topology: at each round, agents send messages to the server, the server aggregates them into a new iterate and broadcasts this iterate to all agents. Because many agents communicate concurrently, the agents-to-server traffic is often the dominant bottleneck (Seide et al., 2014; Chilimbi et al., 2014; Strom, 2015), especially for large models seen in deep learning.

A standard approach to mitigate communication costs is to reduce the frequency of communication (McMahan et al., 2017; Karimireddy et al., 2020; Mishchenko et al., 2022) and/or to transmit *compressed* messages. Compression may be applied on the uplink (agents→server) (Seide et al., 2014; Alistarh et al., 2017; Richtárik et al., 2021) and/or downlink (server→agents) (Harrane et al., 2018; Philippenko & Dieuleveut, 2020; Gorbunov et al., 2020); the focus here is on uplink compression. Common compressors include low-precision quantization (Alistarh et al., 2017), sparsification (e.g., Top-$K$, Alistarh et al. 2018), and using random projections (Vempala, 2005) like in SKETCHED-SGD (Ivkin et al., 2019).

To analyze algorithms independently of a particular compressor, one typically assumes general properties of a (possibly random) compression operator $\mathcal{C}$. Compression is modeled as a family of deterministic maps indexed by a seed $\omega$, written $\mathcal{C}(\cdot;\omega):\mathbb{R}^d\to\mathbb{R}^d$; each invocation uses an independent seed (possibly shared across agents). When needed, $\omega_k^{(i)}$ denotes the seed used by agent $i$ at iteration $k$, and $\omega_k \coloneqq (\omega_k^{(i)})_{i=1}^n$; expectations are taken with respect to the relevant seeds. A classical assumption is unbiasedness, i.e., $\mathbb{E}_\omega[\mathcal{C}(x;\omega)] = x$ for all $x\in\mathcal{X}$. Another widely used assumption is *contractiveness*:

**Assumption 1.1** (Contractive compressor)**.** The compression operator $\mathcal{C}(\cdot;\omega)$ is such that, for some $\epsilon\in[0,1)$,

$$\text{for all}\quad x\in\mathbb{R}^d,\quad \mathbb{E}_\omega\left[\|x-\mathcal{C}(x;\omega)\|^2\right]\le\epsilon\|x\|^2.$$

A natural baseline is *Compressed Gradient Descent* (CGD):

$$x_{k+1} \coloneqq x_k - \frac{\eta}{n}\sum_{i=1}^{n}\mathcal{C}(\nabla f^{(i)}(x_k);\omega_k^{(i)}),$$

where $\eta>0$ is the step size. However, CGD generally fails to converge in the multi-agent setting, as shown for instance when $\mathcal{C}$ is biased in Beznosikov et al. (2023).

To mitigate the effects of compression, each agent can use *Error Feedback* (EF). The classic mechanism (Seide et al., 2014; Karimireddy et al., 2019) keeps track of past compression errors and "reinjects" them into later messages, as described in Algorithm 1.

[1]INRIA, D.I. École Normale Supérieure, PSL Research University, 75005 Paris, France [2]CMAP, CNRS, École polytechnique, Institut Polytechnique de Paris, 91120 Palaiseau, France. Correspondence to: Daniel Berg Thomsen <daniel.berg-thomsen@inria.fr>.

*Preprint. February 26, 2026.*

Another challenge in distributed environments is heterogeneity across agents: the local objectives $f^{(i)}$ can differ substantially, for example due to mismatched data distributions or uneven scaling induced by non-uniform data partitioning. Such heterogeneity can obstruct convergence. In particular, in the strongly convex setting, Beznosikov et al. (2023) prove linear convergence of distributed EF with biased compression, but only under the additional assumption that all local objectives share the same minimizer—i.e., the interpolation regime (Ma et al., 2018; Vaswani et al., 2019).

Motivated by the need to directly handle heterogeneity, Richtárik et al. (2021) proposed *Error Feedback 21* ($\mathrm{EF}^{21}$), given in Algorithm 2. In $\mathrm{EF}^{21}$, each agent maintains an estimator of the gradient $d_k^{(i)}$ and communicates compressed differences $\mathcal{C}(\nabla f^{(i)}(x_{k+1}) - d_k^{(i)}; \omega_k^{(i)})$. The server then performs a gradient step using this sum, which is argued to be more stable around the global optimum of the sum total objective (1).

By now, a substantial literature has been devoted to the analysis of both algorithms under different assumptions on the communication model (Koloskova et al., 2019; Philippenko & Dieuleveut, 2021), the compression operator (Alistarh et al., 2017; Stich et al., 2018; Beznosikov et al., 2023), and the objective functions (Karimireddy et al., 2019; Stich & Karimireddy, 2020; Richtárik et al., 2021). The vast literature on the subject is further complicated by the many variants of these methods that have been proposed, see e.g., Zheng et al. (2019); Li & Li (2022); Tang et al. (2021); Fatkhullin et al. (2023); Tian et al. (2026); Condat et al. (2022); Fatkhullin et al. (2025); Gruntkowska et al. (2025); Egger et al.; Redie et al. (2026), among others.

Many of the above papers provide *theoretical guarantees*, most often in the form of convergence rates that upper-bound the value of a hand-crafted metric. Yet these bounds can be loose, the analysis pessimistic, or the metric itself poorly matched to the phenomenon of interest, which complicates meaningful comparisons between algorithms. Building on recent advances in computer-aided proofs for optimization and automated construction of Lyapunov functions, Berg Thomsen et al. (2025) derived *tight* convergence rates for CGD, EF, and $\mathrm{EF}^{21}$. Their rates are optimistic compared to existing results and give a *definitive* characterization of the worst-case convergence of these methods on smooth and strongly convex functions; notably, they also show that EF and $\mathrm{EF}^{21}$ attain the same optimized worst-case rate. However, this analysis is restricted to the single-agent setting $n = 1$, which remains a significant theoretical and practical limitation.

### 1.1. Contributions.

In this paper, we provide *tight analyses* of both EF and $\mathrm{EF}^{21}$ in the multi-agent setting, $n > 1$, under a common set of assumptions: contractive compression (Assumption 1.1) together with strong convexity and smoothness of the individual objective functions. We show that the behavior in the distributed setting can differ from the single agent case, in particular due to *heterogeneity* between agents. We distinguish between two types of heterogeneity: statistical heterogeneity (the local minimizer can differ across agents) and heterogeneity of the regularity parameters (the smoothness and strong-convexity constants being different for each agent).

These rates are *tight* in that they correspond to the *best tuning*, the *best Lyapunov function* (within the considered class), and the *best possible convergence rate*. To obtain them, we combine advanced proof techniques with numerical evaluations that corroborate the theoretical predictions. Altogether, our results provide a *complete picture* of the worst-case behavior of error-feedback methods in the multi-agent setting.

**Rigor: Theorems, Certificates, and Empirical Laws.** Most of our results are stated either as a theorem with a complete proof or as a counterexample. To facilitate verification, we also attach **proof certificates** to each formal statement, which serve as either analytical or numerical validation of the stated guarantee. Specifically, certificates of correctness are provided either

(i) by a *Computer Algebra System* (CAS) via a Wolfram Language script, or

(ii) by numerically solving the associated *Performance Estimation Problems* (PEP).

CAS certificates verify algebraic identities, whereas PEP certificates provide numerical confirmation of complete statements. In the paper, these certificates are indicated by the CAS and PEP markers, which link directly to the corresponding Wolfram Language script or Jupyter notebook in the public GitHub repository.[1]

For cases in which a formal proof could not be identified, we state certain results as **empirical laws** instead. These laws are supported by extensive numerical evidence indicating that a proposed closed-form expression matches, up to numerical precision, the corresponding analytical quantity, which may itself lack a closed-form representation.

[1]These certificates do not replace the mathematical proofs in the paper; rather, they provide an additional layer of transparency and error checking, analogous to unit tests in software development. They offer a reproducible, independently verifiable record supporting the theoretical claims and help reduce the risk of oversights in complex derivations.

| Agents | Heterogeneity | EF | EF$^{21}$ | EControl |
|---|---|---|---|---|
| $n = 1$ | — | Berg Thomsen et al. (2025) | | Empirical Law 4.4 PEP |
| $n > 1$ | None | Theorem 3.4 PEP CAS | Theorem 3.1 PEP CAS | |
| | Statistical† | Cycles (Proposition 2.1) | | |
| | Regularity‡ | Corollary 3.2 (Linear)* PEP | | |
| $n = 2$ | Both | Empirical Law 4.3 PEP | | |

*Table 1.* Summary of convergence results for EF$^{21}$ and EF. † Statistical heterogeneity means that the local minimizers $x_\star^{(i)}$ are not identical; ‡ Regularity heterogeneity means that the smoothness and strong-convexity constants are not identical; $*$ the linear corollary additionally assumes (beyond contractive compression) that $\mathcal{C}$ is deterministic, additive, and positively homogeneous ($\mathcal{C}(x+y) = \mathcal{C}(x) + \mathcal{C}(y), \mathcal{C}(\alpha x) = \alpha\mathcal{C}(x)$ for $\alpha \geq 0$). Heterogeneous guarantees for EF$^{21}$ and EF are currently limited to linear compressors and the $n = 2$ empirical law. Badges link to certificates.

**Detailed contributions and outline.** More precisely, we make the following contributions, summarized in Table 1.

(i) In Section 2, we exhibit simple quadratic counterexamples for contractive compressors, both CGD and classic EF exhibit cycles in heterogeneous multi-agent settings and therefore do not converge in general.

(ii) For EF$^{21}$ under contractive compression and statistical heterogeneity, we derive in Subsection 3.1 a tight Lyapunov function, optimal step size, and worst-case contraction factor, and the optimal rate is shown to be independent of the number of agents.

(iii) For deterministic, additive, positively homogeneous compressors under regularity heterogeneity, we also prove a sharp linear convergence guarantee for both EF$^{21}$ and EF with averaged parameters $(\bar{L}, \bar{\mu})$.

(iv) In Subsection 3.2, we construct a tight Lyapunov function for classic EF under statistical homogeneity, and show that its optimal step size and rate coincide with those of EF$^{21}$, recovering the single-agent results for any $n$.

(v) Finally, in Section 4, we state *empirical laws* that characterize the optimal step size for EF and EF$^{21}$ under general heterogeneity, an optimal Lyapunov function for EF$^{21}$, an explicit polynomial rate formula for the $n = 2$ case, and an optimal two-parameter tuning rule for EControl (Gao et al., 2023), supported by extensive numerical verification via numerical solution of the corresponding performance estimation problems.

In the next section, we introduce the algorithms, assumptions and counterexamples.

## 2. Background

This section motivates our multi-agent analysis by showing that classic EF can fail under heterogeneity unless additional structure is imposed. Specifically, Subsection 2.1 provides counterexamples, and Subsection 2.2 reviews existing theoretical results on EF and EF$^{21}$. The definitions and notation needed to state the main results of this paper are provided in Subsections 2.3 and 2.4.

### 2.1. Non-Convergence under Statistical Heterogeneity

The following counterexamples demonstrate that neither compressed gradient descent nor EF can achieve arbitrary accuracy in the presence of statistical heterogeneity.

**Compressed Gradient Descent.** Consider the case where there are $n = 2$ agents, each having access to first-order oracles querying the one-dimensional quadratic functions

$$f^{(1)}(x) := \frac{\mu}{2}x^2 - x, \qquad f^{(2)}(x) := \frac{\mu}{2}x^2 + x,$$

where $\mu > 0$ is a constant. It follows immediately that these functions belong to the class $\mathcal{F}_{\mu,L}$, for any $L > \mu$. Set

$$x_0 := \eta\frac{\sqrt{\epsilon}}{2 - \eta\mu}.$$

By definition, for CGD

$$x_1 = x_0 - \frac{\eta}{2}\left[\mathcal{C}((f^{(1)})'(x_0); \omega_0^{(1)}) + \mathcal{C}((f^{(2)})'(x_0); \omega_0^{(2)})\right].$$

Under Assumption 1.1, the compression oracle may respond

$$\mathcal{C}((f^{(1)})'(x_0); \omega_0^{(1)}) = (1 - \sqrt{\epsilon})(f^{(1)})'(x_0),$$
$$\mathcal{C}((f^{(2)})'(x_0); \omega_0^{(2)}) = (1 + \sqrt{\epsilon})(f^{(2)})'(x_0).$$

Plugging this, the derivatives of $f^{(1)}$ and $f^{(2)}$, and the definition of $x_0$ into subsection 2.1,

$$x_1 = x_0 - \frac{\eta}{2}\left[2\mu x_0 + 2\sqrt{\epsilon}\right] = x_0(1 - \eta\mu) - \eta\sqrt{\epsilon}$$
$$= \frac{\eta\sqrt{\epsilon}}{2 - \eta\mu}\left[(1 - \eta\mu) - (2 - \eta\mu)\right] = -x_0.$$

**Algorithm 1** Classic error feedback — EF

**initialization:** $x_0 \in \mathbb{R}^d, \eta > 0, e_0^{(i)} = 0$ for $i \in [n]$; seeds $(\omega_k^{(i)})_{k\geq 0}$ given
**for** $k = 0, 1, 2, \ldots, K$ **do**
  Agent $i \in [n]$ compresses $e_k^{(i)} + \eta \nabla f^{(i)}(x_k)$ and communicates $m_k^{(i)} := \mathcal{C}(e_k^{(i)} + \eta \nabla f^{(i)}(x_k); \omega_k^{(i)})$
  Agent $i \in [n]$ updates $e_k^{(i)} \leftarrow e_k^{(i)} + \eta \nabla f^{(i)}(x_k) - \mathcal{C}(e_k^{(i)} + \eta \nabla f^{(i)}(x_k); \omega_k^{(i)})$
  Server updates $x_{k+1} \leftarrow x_k - \frac{1}{n} \sum_{i=1}^n m_k^{(i)}$
**end for**

**Algorithm 2** Error Feedback 21 — EF$^{21}$

**initialization:** $x_0 \in \mathbb{R}^d$; step size $\eta > 0$; seeds $(\omega_k^{(i)})_{k\geq 0}$ given; $d_0^{(i)} = \mathcal{C}(\nabla f^{(i)}(x_0); \omega_0^{(i)})$ for $i \in [n]$;
**for** $k = 0, 1, 2, \ldots, K$ **do**
  Server updates $x_{k+1} \leftarrow x_k - \eta \cdot \frac{1}{n} \sum_{i=1}^n d_k^{(i)}$
  Agent $i \in [n]$ compresses $\nabla f^{(i)}(x_{k+1}) - d_k^{(i)}$ and communicates $m_k^{(i)} := \mathcal{C}(\nabla f^{(i)}(x_{k+1}) - d_k^{(i)}; \omega_k^{(i)})$
  Agent $i \in [n]$ updates $d_{k+1}^{(i)} \leftarrow d_k^{(i)} + m_k^{(i)}$
**end for**

By symmetry, if the compression oracle next responds with

$$\mathcal{C}((f^{(1)})'(x_1); \omega_1^{(1)}) = (1 + \sqrt{\epsilon})(f^{(1)})'(x_1),$$
$$\mathcal{C}((f^{(2)})'(x_1); \omega_1^{(2)}) = (1 - \sqrt{\epsilon})(f^{(2)})'(x_1).$$

then the same computation gives $x_2 = x_0$, yielding a 2-step cycle.

**Error Feedback.** Consider now exactly the same functions defined in subsection 2.1. The following Proposition shows that the same behavior can be observed in EF.

**Proposition 2.1.** *Let there be $n = 2$ agents, each having access to first-order oracles querying the one-dimensional quadratic functions defined in subsection 2.1. Then there are 2-step cycles for the following scenarios where the step size $\eta > 0$ is fixed:*

1. $\eta < \frac{2}{\mu}$, *with* $x_0 = \eta \frac{\sqrt{\epsilon}}{2 - \eta\mu}$.
2. $\eta > \frac{2}{\mu}$, *with* $x_0 = -\eta \frac{\sqrt{\epsilon}}{2 - \eta\mu}$.
3. $\eta = \frac{2}{\mu}$, *with any* $x_0$.

A simple proof of this is provided in Appendix A.1.

### 2.2. Related Work

Error feedback has a long history in signal processing, where it is used to compensate for quantization in communication (Cutler, 1952; Inose & Yasuda, 2005). In distributed optimization, it dates back to the classical work of Seide et al. (2014). Following its introduction, subsequent work has studied the behavior of error feedback under various assumptions, including analyses for specific compression operators such as deterministic Top-$K$ (Alistarh et al., 2018) and stochastic compressors (Wu et al., 2018).

A complementary line of work leverages the contractive properties of compression operators to derive general convergence guarantees for a broad class of operators. In this setting, it is common to distinguish between *biased* and *unbiased* compressors (Beznosikov et al., 2023). Convergence rates for error feedback with contractive compressors have been established for strongly convex functions (Stich et al., 2018), quasi-convex and nonconvex functions (Karimireddy et al., 2019), and using stochastic gradients (Stich & Karimireddy, 2020). Tight rates have also been established under the same assumptions as this work, for the single-agent regime (Berg Thomsen et al., 2025).

Error Feedback 21 (EF$^{21}$) (Richtárik et al., 2021) is a variant of error feedback that was designed specifically to handle the heterogeneity of the multi-agent setting. EF$^{21}$ and its variants have been studied in many different scenarios, some of which include using stochastic gradients, momentum (Fatkhullin et al., 2023) and practical extensions such as bidirectional compression, variance reduction, and proximal setups (Fatkhullin et al., 2021). EControl (Gao et al., 2023) introduces a controllable error-compensation mechanism that combines the notions of error feedback from both EF and EF$^{21}$ by adding a second hyperparameter, and is shown to converge in strongly convex, convex and nonconvex settings.

### 2.3. Assumptions and Notations

The following definition is used throughout this work.

**Definition 2.2** (Class $\mathcal{F}_{\mu,L}$)**.** Denote by $\mathcal{F}_{\mu,L}$ the set of functions $h : \mathbb{R}^d \to \mathbb{R}$ that are $L$-smooth and $\mu$-strongly convex. That is, for all $x, y \in \mathbb{R}^d$,

$$h(y) \leq h(x) + \langle \nabla h(x), y - x \rangle + \frac{L}{2}\|y - x\|^2,$$

and

$$h(y) \geq h(x) + \langle \nabla h(x), y - x \rangle + \frac{\mu}{2}\|y - x\|^2.$$

Throughout this work, each local function $f^{(i)}$ is assumed to belong to $\mathcal{F}_{\mu^{(i)},L^{(i)}}$.

The symbol $\mathbb{S}^\ell$ denotes the set of symmetric matrices, and $\mathbb{S}^\ell_+$ denotes the set of positive semidefinite matrices. For

any two matrices $A \in \mathbb{S}^\ell$ and $B \in \mathbb{S}^d$, the Kronecker product is denoted by $A \otimes B$. The condition number is denoted by $\kappa := \frac{L}{\mu}$. For any objective function $f \in \mathcal{F}_{\mu,L}$, the minimizer is denoted by $x_\star := \arg\min_{x\in\mathbb{R}^d} f$, and its minimum value is denoted by $f_\star := \min_{x\in\mathbb{R}^d} f(x)$. For each local function $f^{(i)}$, its unique minimizer is denoted by $x_\star^{(i)}$ and its minimum value by $f_\star^{(i)}$.

### 2.4. Methodology

The analysis contained in Section 3 relied on the systematic identification of Lyapunov functions that provide a tight convergence rate for each method. These Lyapunov functions are *optimal* with respect to a large class of Lyapunov functions, defined in this section.

**Lyapunov functions.** Let $\mathcal{M} : \mathbb{R}^{\ell\times d} \times \mathbb{R}^d \times \mathcal{F} \to \mathbb{R}^{\ell\times d} \times \mathbb{R}^d$ denote a first-order method acting on a set of functions $\mathcal{F}$ of dimension $d$, for an integer $\ell \in \mathbb{N}$ different *state variables*. Such a method, given a function $f \in \mathcal{F}$, is applied to an initial *state* $\xi_0 \in \mathbb{R}^{\ell\times d}$ and iterate $x_0 \in \mathbb{R}^d$, and generates a following state $\xi_1$, and iteration $x_1$. The *states* represent information summarizing the current point in the optimization trajectory that the algorithms may depend on beyond the current iterate—for example, error-related quantities in error feedback algorithms.

**Definition 2.3** (Candidate Lyapunov function)**.** A function $\mathcal{V} : \mathbb{R}^{\ell\times d}\times\mathbb{R}^d \to \mathbb{R}$ is called a candidate Lyapunov function for $f$ if it satisfies the following conditions:

1. (Non-negativity) $\mathcal{V}(\xi, x; f) \geq 0$, for any $\xi \in \mathbb{R}^{\ell\times d}$, $x \in \mathbb{R}^d$,
2. (Zero at fixed-point) $\mathcal{V}(\xi, x; f) = 0$ if and only if $x = x_\star$ and $\xi = \xi_\star$ for a unique $\xi_\star \in \mathbb{R}^{\ell\times d}$.
3. (Meaningful lower bound) there exists a positive semidefinite matrix $P \in \mathbb{S}^\ell_+$ and a scalar $p \geq 0$ such that $\mathcal{V}(\xi, x; f) \geq (\xi-\xi_\star)^\top(P\otimes I_d)(\xi-\xi_\star)+p(f(x)-f_\star)$, $\forall \xi \in \mathbb{R}^{\ell\times d}$, $\forall x \in \mathbb{R}^d$, and $\mathrm{Tr}(P) + p = 1$.

The objective is to find candidate Lyapunov functions $\mathcal{V}$ satisfying the recurrence

$$\mathcal{V}(\xi_1, x_1; f) \leq \rho \cdot \mathcal{V}(\xi_0, x_0; f),$$

for some constant $\rho < 1$, uniformly over $\mathcal{F}$. Finding the *optimal* Lyapunov function within a parameterized class for a method $\mathcal{M}$ amounts to solving the following problem:

$$\begin{aligned}\rho_\star(\mathcal{M}) := \min_{\mathcal{V}} \quad & \max_{\substack{f\in\mathcal{F}_{\mu,L}\\ \xi_0,x_0}} \frac{\mathcal{V}(\xi_1, x_1; f)}{\mathcal{V}(\xi_0, x_0; f)} \\ & \text{s.t. } (\xi_1, x_1) = \mathcal{M}(\xi_0, x_0; f)\,.\end{aligned} \tag{2}$$

The goal of this work is to identify optimal Lyapunov functions for EF and EF$^{21}$ and formally prove that they achieve the convergence rate defined in (2). It has been shown that optimal candidate Lyapunov functions can be identified by solving semidefinite programs (SDPs), yielding numerical convergence guarantees (Taylor et al., 2018).

We used numerical tools both to obtain guarantees (Taylor et al., 2017a; Goujaud et al., 2024) and to search for Lyapunov functions (Taylor et al., 2018; Upadhyaya et al., 2025). Such numerical evidence is not, on its own, a theoretical proof, but it can suggest symbolic expressions; we leverage symbolic regression (Cranmer, 2023) to infer them. In the multi-agent setting, the optimal Lyapunov coefficients need not be unique, so several heuristics were used in order to infer closed-form formulas.

**Algorithm 3** Error Control — EControl

**initialization:** $x_0 \in \mathbb{R}^d$; step sizes $\eta > 0, \gamma > 0$; $e_0^{(i)} = 0, d_0^{(i)} = 0$ for $i \in [n]$; seeds $(\omega_k^{(i)})_{k\geq 0}$ given
**for** $k = 0, 1, 2, \ldots, K$ **do**
  Agent $i \in [n]$ compresses $\eta e_k^{(i)} + \nabla f^{(i)}(x_k) - d_k^{(i)}$ and communicates $m_k^{(i)} := \mathcal{C}(\eta e_k^{(i)} + \nabla f^{(i)}(x_k) - d_k^{(i)}; \omega_k^{(i)})$
  Agent $i \in [n]$ updates $d_{k+1}^{(i)} \leftarrow d_k^{(i)} + m_k^{(i)}$
  Server updates $x_{k+1} \leftarrow x_k - \gamma \cdot \frac{1}{n}\sum_{i=1}^n d_{k+1}^{(i)}$
  Agent $i$ updates $e_{k+1}^{(i)} \leftarrow e_k^{(i)} + \nabla f^{(i)}(x_k) - d_{k+1}^{(i)}$
**end for**

## 3. Main Results

This section states tight convergence guarantees for EF$^{21}$ under contractive compression, a linear-compressor corollary when the regularity parameters are heterogenous, and corresponding guarantees for EF under statistical homogeneity. In light of the counterexample in Proposition 2.1, the result for EF requires the additional assumption that the local objectives $f^{(i)}$ share the same minimizer.

### 3.1. Error Feedback 21

For EF$^{21}$, the state is defined as

$$\xi_k^{\mathrm{EF}^{21}} := \begin{bmatrix} [x_k - x_\star^{(1)}, \ldots, x_k - x_\star^{(n)}]^\top \\ [\nabla f^{(1)}(x_k) - \nabla f^{(1)}(x_\star), \ldots]^\top \\ [d_k^{(1)}, \ldots, d_k^{(n)}]^\top \end{bmatrix}. \tag{3}$$

This state contains the per-agent optimality residuals, current gradients, and error-feedback terms. Candidate Lyapunov functions may thus include many terms, but the Lyapunov function given in Theorem 3.1 is relatively simple, and is worst-case optimal under the optimal step size tuning.

**Theorem 3.1.** CAS PEP
*Let $\epsilon \in [0, 1)$ and assume that the compression operator $\mathcal{C}$*

*satisfies Assumption 1.1. Let $f^{(i)} \in \mathcal{F}_{\mu,L}$ for each $i \in [n]$. Let the step size be given by*

$$\eta^\star = \left(\frac{2}{L+\mu}\right) \cdot \left(\frac{1-\sqrt{\epsilon}}{1+\sqrt{\epsilon}}\right). \tag{4}$$

*Then, the Lyapunov function*

$$\begin{aligned}\mathcal{V}(\xi^{\mathrm{EF21}}, x; f) :=& \frac{\sqrt{\epsilon}}{n}\left\|\sum_{i=1}^n \nabla f^{(i)}(x_k)\right\|^2 \\ &+ \sum_{i=1}^n \|\nabla f^{(i)}(x_k) - d_k^{(i)}\|^2\end{aligned} \tag{5}$$

*is optimal (i.e., solves (2)) and satisfies*

$$\mathbb{E}_\omega\left[\mathcal{V}\left(\xi_{k+1}^{\mathrm{EF21}}, x_{k+1}; f\right)\right] \le \rho_\star \cdot \mathbb{E}_\omega\left[\mathcal{V}\left(\xi_k^{\mathrm{EF21}}, x_k; f\right)\right]$$

*where the rate is given by*

$$\rho_\star := \sqrt{\epsilon} + \left(\frac{1-\sqrt{\epsilon}}{2}\right)\left(\frac{\kappa-1}{\kappa+1}\right)^2 \Psi(\kappa,\epsilon) \tag{6}$$

*and*

$$\Psi(\kappa,\epsilon) := 1 - \sqrt{\epsilon} + \sqrt{(1+\sqrt{\epsilon})^2 + \sqrt{\epsilon}\,16\frac{\kappa}{(\kappa-1)^2}}.$$

*Finally, the step size in (4) is worst-case optimal for $\mathrm{EF}^{21}$: within the candidate Lyapunov class based on $\xi^{\mathrm{EF}^{21}}$, it achieves the minimal worst-case one-step contraction factor, $\rho_\star$ in (6). This bound is also multi-step tight, meaning that after $k$ iterations the worst-case contraction equals $\rho_\star^k$.*

A proof appears in Appendix A.2. The numerical and symbolic certificates can be accessed by clicking the PEP and CAS icons in the theorem header.

The symmetry across agents implies that the worst-case rate in Theorem 3.1 matches the single-agent rate: the worst-case instance for $n = 1$ can be replicated across agents. Though the single-agent Lyapunov function is recovered when setting $n = 1$, the multi-agent Lyapunov function is not a trivial extension of the single-agent case. One could have expected a weighted sum of the single-agent Lyapunov functions defined in (Berg Thomsen et al., 2025) to be sufficient, but it is clear that the weighting placed on the individual terms of (5) do not correspond to that.

The next corollary extends Theorem 3.1 to regularity heterogeneity (heterogeneous smoothness and strong convexity parameters) under deterministic, additive, positively homogeneous compression, a setup in which Richtárik et al. (2021) showed an equivalence between EF and $\mathrm{EF}^{21}$, under a reparametrization.

**Corollary 3.2.**
*Let $\epsilon \in [0,1)$ and assume that the compression operator $\mathcal{C}$ satisfies Assumption 1.1 and is deterministic, additive, and positively homogeneous (so $\mathcal{C}(x+y) = \mathcal{C}(x) + \mathcal{C}(y)$ and $\mathcal{C}(\alpha x) = \alpha\mathcal{C}(x)$ for all $\alpha \ge 0$). Let $f^{(i)} \in \mathcal{F}_{\mu^{(i)},L^{(i)}}$ for each $i \in [n]$, and define*

$$\bar{L} := \frac{1}{n}\sum_{i=1}^n L^{(i)}, \qquad \bar{\mu} := \frac{1}{n}\sum_{i=1}^n \mu^{(i)}.$$

$$\kappa_\Sigma := \frac{\bar{L}}{\bar{\mu}} = \frac{\sum_{i=1}^n L^{(i)}}{\sum_{i=1}^n \mu^{(i)}}.$$

*Let the step size be given by*

$$\eta^\star = \left(\frac{2}{\bar{L}+\bar{\mu}}\right) \cdot \left(\frac{1-\sqrt{\epsilon}}{1+\sqrt{\epsilon}}\right).$$

*Then the Lyapunov function*

$$\mathcal{V}_{\mathrm{lin}}\left(\xi^{\mathrm{EF21}}, x; f\right) := \sqrt{\epsilon}\,\|\bar{g}_k\|^2 + \|\bar{g}_k - \bar{d}_k\|^2$$

*satisfies*

$$\mathbb{E}_\omega\left[\mathcal{V}_{\mathrm{lin}}\left(\xi_{k+1}^{\mathrm{EF21}}, x_{k+1}; f\right)\right] \le \rho_\star \cdot \mathbb{E}_\omega\left[\mathcal{V}_{\mathrm{lin}}\left(\xi_k^{\mathrm{EF21}}, x_k; f\right)\right],$$

*where $\bar{g}_k := \frac{1}{n}\sum_{i=1}^n \nabla f^{(i)}(x_k)$, $\bar{d}_k := \frac{1}{n}\sum_{i=1}^n d_k^{(i)}$, and the rate $\rho_\star$ is given by (6) with $\kappa = \kappa_\Sigma$.*

A proof appears in Appendix A.3. The same step size and rate apply to EF, matching Theorems 1–2 in Berg Thomsen et al. (2025) evaluated at $\kappa_\Sigma$. The result follows by observing that the barred variables evolve exactly as a single-agent $\mathrm{EF}^{21}$ instance with parameters $(\bar{L}, \bar{\mu})$, and then invoking the EF–$\mathrm{EF}^{21}$ equivalence. Extending tight guarantees beyond this deterministic linear setting appears technically challenging, motivating the next section.

### 3.2. Classic Error Feedback under Statistical Homogeneity

Due to the counterexample given in Proposition 2.1, an additional assumption is required to prove convergence for EF. This is the same assumption as that required for the linear rate of convergence given in (Beznosikov et al., 2023)—namely, that the minimizers of the local functions are all the same, a condition commonly known as the interpolation regime (Ma et al., 2018; Vaswani et al., 2019).

**Assumption 3.3** (Statistical homogeneity)**.** Each function $f^{(i)}$ is *statistically homogeneous*, i.e.,

$$x_\star^{(i)} = x_\star^{(j)} \quad \text{for all} \quad i, j \in [n]^2.$$

For EF, the state is defined as

$$\xi_k^{\mathrm{EF}} := \begin{bmatrix} x_k - x_\star \\ [\nabla f^{(1)}(x_k), \ldots, \nabla f^{(n)}(x_k)]^\top \\ [m_k^{(1)}, \ldots, m_k^{(n)}]^\top \\ [e_k^{(1)}, \ldots, e_k^{(n)}]^\top \end{bmatrix}.$$

Similar to Theorem 3.1, Theorem 3.4 provides a relatively simple Lyapunov function which is worst-case optimal under the optimal step size tuning.

**Theorem 3.4.** CAS PEP
*Let $\epsilon \in [0,1)$ and assume that the compression operator $\mathcal{C}$ satisfies Assumption 1.1. Let $f^{(i)} \in \mathcal{F}_{\mu,L}$ for each $i \in [n]$, and assume Assumption 3.3.*

*Let the step size be given by $\eta^\star$ as defined in (4). Then, the Lyapunov function*

$$\mathcal{V}\big(\xi^{\mathrm{EF}}, x; f\big) := \frac{1}{n\sqrt{\epsilon}} \left\| \sum_{i=1}^n e_k^{(i)} \right\|^2 + \sum_{i=1}^n \|x_k - x_\star - e_k^{(i)}\|^2.$$

*is optimal and satisfies*

$$\rho_\star(\mathrm{EF}) = \rho_\star,$$

*where $\rho_\star$ is defined in (6). Finally, the step size in (4) is worst-case optimal for* EF*: within the candidate Lyapunov class based on $\xi^{\mathrm{EF}}$, it achieves the minimal worst-case one-step contraction factor, $\rho_\star$ in (6). This bound is also multi-step tight, meaning that after $k$ iterations the worst-case contraction equals $\rho_\star^k$.*

A proof appears in Appendix A.4. Like in the case of $\mathrm{EF}^{21}$, the single-agent results from (Berg Thomsen et al., 2025) are recovered for any $n \geq 1$, including the optimal step size and the Lyapunov function (by setting $n = 1$ in Theorem 3.4).

The results in this section establish that, when each $f^{(i)} \in \mathcal{F}_{\mu,L}$, the worst-case contraction of EF and $\mathrm{EF}^{21}$ does not depend on the number of agents $n$. In particular, the class-optimal step sizes and contraction factors coincide with their single-agent counterparts, implying that the single-agent setting suffices to characterize the worst-case behavior of these distributed methods.

## 4. Empirical Laws under Regularity Heterogeneity

The results of the previous sections either assume regularity homogeneity (shared parameters $(L,\mu)$) or rely on linear deterministic compressors to handle regularity heterogeneity. When this is relaxed to allow general regularity heterogeneity, with agent-specific parameters $f^{(i)} \in \mathcal{F}_{\mu^{(i)},L^{(i)}}$, numerical evidence indicates that the worst-case behavior depends on the local smoothness and strong convexity parameters. Based on extensive numerical experiments, this section states empirical laws for (i) the optimal step size for both EF and $\mathrm{EF}^{21}$, (ii) a corresponding Lyapunov function for $\mathrm{EF}^{21}$, (iii) the optimal convergence rate when $n = 2$, and (iv) a two-parameter tuning rule for EControl (Gao et al., 2023) in the general case.

The empirical laws below are formulated based on numerical validation using the Performance Estimation Problem (PEP) framework (Drori, 2014; Taylor et al., 2017b). A detailed account of the verification methodology, including exact grid searches and parameter sweeps, is provided in Section C. The empirical formulas were tested across a wide range of problem parameters.

**Empirical Law 4.1** (General optimal step size)**.** PEP
*Consider the setting with heterogeneous regularity parameters, i.e., let $f^{(i)} \in \mathcal{F}_{\mu^{(i)},L^{(i)}}$. The optimal step size for both* EF *and* $\mathrm{EF}^{21}$ *is given by*

$$\eta^\star = \left(\frac{2n}{\sum_{i=1}^n L^{(i)} + \mu^{(i)}}\right) \cdot \left(\frac{1-\sqrt{\epsilon}}{1+\sqrt{\epsilon}}\right).$$

This step size reduces to (4) when all agents share the same parameters.

Numerical evidence further suggests that the following Lyapunov function is optimal for $\mathrm{EF}^{21}$ among Lyapunov functions built from the class defined by (3).

**Empirical Law 4.2** (General Lyapunov function)**.** PEP
*Let $S := \sum_{i=1}^n L^{(i)} + \mu^{(i)}$ and $w_i := \frac{S}{L^{(i)}+\mu^{(i)}}$. Let each agent have heterogeneous regularity parameters, i.e., $f^{(i)} \in \mathcal{F}_{\mu^{(i)},L^{(i)}}$. The optimal Lyapunov function for* $\mathrm{EF}^{21}$ *is given by*

$$\mathcal{V}\big(\xi^{\mathrm{EF21}}, x; f\big) := \frac{1}{n}\sum_{i=1}^n w_i \|\nabla f^{(i)}(x_k) - d_k^{(i)}\|^2 + \frac{\sqrt{\epsilon}}{n}\left\|\sum_{i=1}^n \nabla f^{(i)}(x_k)\right\|^2$$

This family of Lyapunov functions reduces to (5) when all agents share the same parameters, and motivates a weighting based on relative condition numbers.

Finally, numerical evidence suggests an explicit characterization of the agent-sensitive convergence rate when $n = 2$:

**Empirical Law 4.3** (General rate for $n = 2$)**.** PEP
*Let $n = 2$ and let the agents have heterogeneous regularity parameters, i.e., $f^{(i)} \in \mathcal{F}_{\mu^{(i)},L^{(i)}}$. The optimal convergence rate of* $\mathrm{EF}^{21}$ *is given by the largest root of the polynomial*

$$Q(\rho) = \rho^3 - \left[s(2+s) + r(s)(sK_1 + K_2)\right]\rho^2 + s^2\left[1 + 2s + r(s)(K_1 + sK_2)\right]\rho - s^4$$

*where $s := \sqrt{\epsilon}$, $r(s) := \frac{(1-s)^2}{1+s}$, and the constants $K_1, K_2$ are defined as*

$$K_1 := \frac{\Delta_2^2\Sigma_1 + \Delta_1^2\Sigma_2}{\Sigma_1\Sigma_2(\Sigma_1+\Sigma_2)}, \qquad K_2 := \frac{(\Delta_1+\Delta_2)^2}{(\Sigma_1+\Sigma_2)^2},$$

*with $\Sigma_i := L^{(i)} + \mu^{(i)}$ and $\Delta_i := L^{(i)} - \mu^{(i)}$ for $i \in \{1,2\}$.*

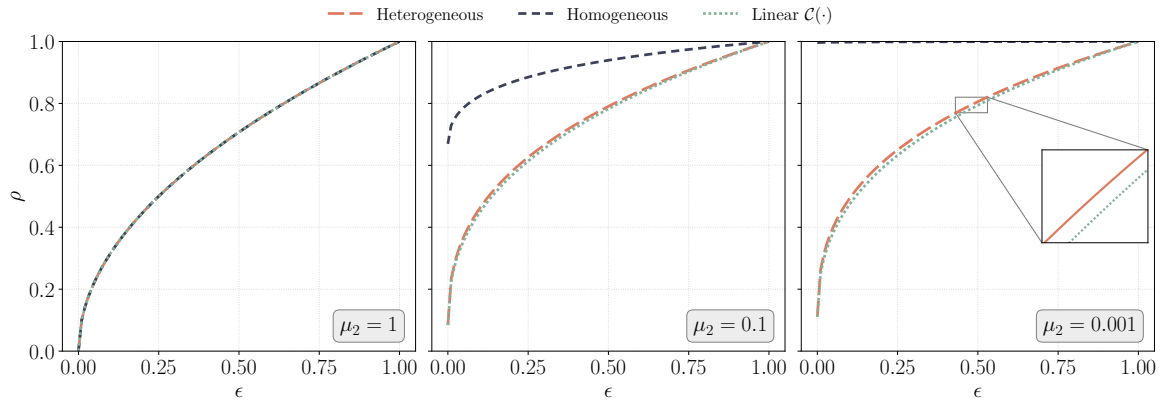


*Figure 1.* Comparison of the rates predicted by Empirical Law 4.3, Theorem 3.1 with worst-case parameters $\max_i L^{(i)}$ and $\min_i \mu^{(i)}$, and Corollary 3.2 with averaged parameters. The inset in the rightmost panel magnifies the gap between the empirical-law rate and the linear-compressor rate. Here $L^{(1)} = L^{(2)} = 1$, $\mu^{(1)} = 1$, and $\mu^{(2)} \in \{1, 0.1, 10^{-3}\}$ from left to right.

The roots of the cubic polynomial in Empirical Law 4.3 are visualized in Figure 2 (Appendix). A comparison to the homogeneous worst-case and linear-compressor rates is shown in Figure 1.

Numerical evidence further suggests a simple tuning rule for the two-parameter EControl method (Gao et al., 2023), where $\gamma$ denotes the model step size and $\eta$ denotes the error-feedback gain.

**Empirical Law 4.4** (EControl tuning)**.** PEP
*Consider the case of heterogeneous regularity parameters, i.e., $f^{(i)} \in \mathcal{F}_{\mu^{(i)},L^{(i)}}$. The empirically optimal tuning for* EControl *satisfies*

$$\eta^\star_{\text{EControl}} = 0,$$
$$\gamma^\star_{\text{EControl}} = \left(\frac{2n}{\sum_{i=1}^n L^{(i)} + \mu^{(i)}}\right) \cdot \left(\frac{1-\sqrt{\epsilon}}{1+\sqrt{\epsilon}}\right),$$

*which coincides with the optimal step size in Empirical Law 4.1.*

When $\eta = 0$ in EControl, the method can easily be shown to be equivalent to EF$^{21}$ under a reordering of the corresponding parameter $d_k^{(i)}$, and by instead initializing $d_0^{(i)} = 0$. In particular, this means the algorithm inherits the same worst-case convergence rates as EF$^{21}$.

## 5. Conclusion

This paper provides tight worst-case analyses for distributed error-feedback algorithms in the multi-agent setting. Under homogenous regularity parameters (the same strong convexity and smoothness constants across agents), the optimal step sizes and contraction factors for EF and EF$^{21}$ are independent of the number of workers $n$ and coincide with the corresponding single-agent guarantees.

For EF$^{21}$ under contractive compression, the analysis holds under statistical heterogeneity, and yields an optimal Lyapunov function together with a worst-case optimal step size and contraction factor. In contrast, a simple counterexample demonstrates that classic EF fails without additional assumptions when local minimizers differ. Under statistical homogeneity (shared minimizer), we provide a tight analysis for EF and recover the same optimal rate as EF$^{21}$; for deterministic linear compressors we further obtain sharp guarantees under regularity heterogeneity via an averaged-parameter reduction.

The Lyapunov functions proposed in this analysis could provide insights into the behavior of these methods under different sets of assumptions and, given the close relationship between EF and EF$^{21}$, could inspire the construction of Lyapunov functions for other error-compensated algorithms. Overall, these results demonstrate that the single-agent setting captures the core worst-case behavior and optimal parameter settings observed in larger systems under homogeneity of the regularization parameters, thereby unifying existing theory.

Finally, a number of empirical laws have been formulated about the behavior of EF$^{21}$ under general regularity heterogeneity, with strong numerical evidence supporting their validity. Establishing these laws rigorously is left for future work; caution is warranted, as the dependence of the optimal rate on heterogeneous parameters already appears technically intricate, even for $n = 2$.

## Acknowledgments

D. Berg Thomsen and A. Taylor are supported by the European Union (ERC grant CASPER 101162889). The work of A. Dieuleveut is partly supported by ANR-19-CHIA-0002-01/chaire SCAI, and Hi!Paris FLAG project, PEPR Redeem. The French government also partly funded this work under the management of Agence Nationale de la Recherche as part of the "France 2030" program, references ANR-23-IACL-0008 "PR[AI]RIE-PSAI", ANR-23-PEIA-005 (REDEEM project) and ANR-23-IACL-0005.

## Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of this work, none of which the authors feel must be specifically highlighted here.

## References

Alistarh, D., Grubic, D., Li, J., Tomioka, R., and Vojnovic, M. QSGD: Communication-Efficient SGD via Gradient Quantization and Encoding. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017.

Alistarh, D., Hoefler, T., Johansson, M., Konstantinov, N., Khirirat, S., and Renggli, C. The Convergence of Sparsified Gradient Methods. *Advances in Neural Information Processing Systems (NeurIPS)*, 31, 2018.

ApS, M. *MOSEK Optimizer API for Python 11.0.21*, 2025. URL https://docs.mosek.com/11.0/pythonapi/index.html.

Berg Thomsen, D., Taylor, A., and Dieuleveut, A. Tight analyses of first-order methods with error feedback. *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Beznosikov, A., Horváth, S., Richtárik, P., and Safaryan, M. On Biased Compression for Distributed Learning. *Journal of Machine Learning Research*, 24(276):1–50, 2023.

Chilimbi, T., Suzue, Y., Apacible, J., and Kalyanaraman, K. Project adam: Building an efficient and scalable deep learning training system. In *USENIX Symposium on Operating Systems Design and Implementation (OSDI 14)*, 2014.

Condat, L., Yi, K., and Richtárik, P. Ef-bv: A unified theory of error feedback and variance reduction mechanisms for biased and unbiased compression in distributed optimization. *Advances in Neural Information Processing Systems*, 35:17501–17514, 2022.

Cranmer, M. Interpretable Machine Learning for Science with PySR and SymbolicRegression.jl. *arXiv:2305.01582 [physics]*, May 2023.

Cutler, C. C. Differential quantization of communication signals, July 29 1952. US Patent 2,605,361.

Drori, Y. *Contributions to the Complexity Analysis of Optimization Algorithms*. PhD thesis, Tel-Aviv University, 2014.

Egger, M., Bitar, R., Wachter-Zeh, A., Weinberger, N., and Gunduz, D. Bicompfl: Bi-directional compression for stochastic federated learning. In *ICML 2025 Workshop on Machine Learning for Wireless Communication and Networks (ML4Wireless)*.

Fatkhullin, I., Sokolov, I., Gorbunov, E., Li, Z., and Richtárik, P. EF21 with Bells & Whistles: Practical Algorithmic Extensions of Modern Error Feedback, October 2021. arXiv:2110.03294 [cs, math].

Fatkhullin, I., Tyurin, A., and Richtárik, P. Momentum provably improves error feedback! *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Fatkhullin, I., Sokolov, I., Gorbunov, E., Li, Z., and Richtárik, P. Ef21 with bells & whistles: Six algorithmic extensions of modern error feedback. *Journal of Machine Learning Research*, 26(189):1–50, 2025.

Gao, Y., Islamov, R., and Stich, S. Econtrol: Fast distributed optimization with compression and error control. *arXiv preprint arXiv:2311.05645*, 2023.

Gorbunov, E., Kovalev, D., Makarenko, D., and Richtarik, P. Linearly Converging Error Compensated SGD. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Goujaud, B., Moucer, C., Glineur, F., Hendrickx, J. M., Taylor, A. B., and Dieuleveut, A. PEPit: computer-assisted worst-case analyses of first-order optimization methods in Python. *Mathematical Programming Computation*, 16 (3):337–367, 2024.

Gruntkowska, K., Gaponov, A., Tovmasyan, Z., and Richtárik, P. Error feedback for muon and friends. *arXiv preprint arXiv:2510.00643*, 2025.

Harrane, I. E. K., Flamary, R., and Richard, C. On reducing the communication cost of the diffusion lms algorithm. *IEEE Transactions on Signal and Information Processing over Networks*, 5(1):100–112, 2018.

Inose, H. and Yasuda, Y. A unity bit coding method by negative feedback. *Proceedings of the IEEE*, 51(11): 1524–1535, 2005.

Ivkin, N., Rothchild, D., Ullah, E., Braverman, V., Stoica, I., and Arora, R. Communication-efficient Distributed SGD with Sketching. *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., D'Oliveira, R. G. L., Rouayheb, S. E., Evans, D., Gardner, J., Garrett, Z., Gascón, A., Ghazi, B., Gibbons, P. B., Gruteser, M., Harchaoui, Z., He, C., He, L., Huo, Z., Hutchinson, B., Hsu, J., Jaggi, M., Javidi, T., Joshi, G., Khodak, M., Konečný, J., Korolova, A., Koushanfar, F., Koyejo, S., Lepoint, T., Liu, Y., Mittal, P., Mohri, M., Nock, R., Özgür, A., Pagh, R., Raykova, M., Qi, H., Ramage, D., Raskar, R., Song, D., Song, W., Stich, S. U., Sun, Z., Suresh, A. T., Tramèr, F., Vepakomma, P., Wang, J., Xiong, L., Xu, Z., Yang, Q., Yu, F. X., Yu, H., and Zhao, S. Advances and Open Problems in Federated Learning. *arXiv:1912.04977 [cs, stat]*, December 2019.

Karimireddy, S. P., Rebjock, Q., Stich, S., and Jaggi, M. Error Feedback Fixes SignSGD and other Gradient Compression Schemes. In *International Conference on Machine Learning (ICML)*, 2019.

Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S., Stich, S., and Suresh, A. T. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning (ICML)*, 2020.

Koloskova, A., Stich, S., and Jaggi, M. Decentralized Stochastic Optimization and Gossip Algorithms with Compressed Communication. In *International Conference on Machine Learning (ICML)*, 2019.

Li, X. and Li, P. Analysis of error feedback in federated non-convex optimization with biased compression. *arXiv preprint arXiv:2211.14292*, 2022.

Ma, S., Bassily, R., and Belkin, M. The power of interpolation: Understanding the effectiveness of sgd in modern over-parametrized learning. In *International Conference on Machine Learning*, pp. 3325–3334. PMLR, 2018.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and Arcas, B. A. y. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, April 2017.

Mishchenko, K., Malinovsky, G., Stich, S., and Richtárik, P. Proxskip: Yes! local gradient steps provably lead to communication acceleration! finally! In *International Conference on Machine Learning*, pp. 15750–15769. PMLR, 2022.

Philippenko, C. and Dieuleveut, A. Bidirectional compression in heterogeneous settings for distributed or federated learning with partial participation: tight convergence guarantees. *arXiv:2006.14591 [cs, stat]*, 2020.

Philippenko, C. and Dieuleveut, A. Preserved central model for faster bidirectional compression in distributed settings. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Redie, D. K., Arablouei, R., and Werner, S. Sa-pef: Step-ahead partial error feedback for efficient federated learning, 2026. URL https://arxiv.org/abs/2601.20738.

Richtárik, P., Sokolov, I., and Fatkhullin, I. EF21: A New, Simpler, Theoretically Better, and Practically Faster Error Feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Seide, F., Fu, H., Droppo, J., Li, G., and Yu, D. 1-Bit Stochastic Gradient Descent and its Application to Data-Parallel Distributed Training of Speech DNNs. In *Annual Conference of the International Speech Communication Association*, 2014.

Stich, S. U. and Karimireddy, S. P. The error-feedback framework: Better rates for SGD with delayed gradients and compressed updates. *Journal of Machine Learning Research*, 21:1–36, 2020.

Stich, S. U., Cordonnier, J.-B., and Jaggi, M. Sparsified SGD with Memory. In *Advances in Neural Information Processing Systems (NeurIPS)*. 2018.

Strom, N. Scalable distributed DNN training using commodity GPU cloud computing. In *Annual Conference of the International Speech Communication Association*, 2015.

Tang, H., Li, Y., Liu, J., and Yan, M. Errorcompensatedx: error compensation for variance reduced algorithms. *Advances in Neural Information Processing Systems*, 34: 18102–18113, 2021.

Taylor, A., Van Scoy, B., and Lessard, L. Lyapunov Functions for First-Order Methods: Tight Automated Convergence Guarantees. In *International Conference on Machine Learning (ICML)*, 2018.

Taylor, A. B., Hendrickx, J. M., and Glineur, F. Performance estimation toolbox (PESTO): automated worst-case analysis of first-order optimization methods. In *Conference on Decision and Control (CDC)*, 2017a.

Taylor, A. B., Hendrickx, J. M., and Glineur, F. Smooth strongly convex interpolation and exact worst-case performance of first-order methods. *Mathematical Programming*, 161(1-2):307–345, 2017b.

Tian, H., Li, X., Liu, S., and Shi, Y. Ef21-rr: Fast o (1/t) rate for non-convex federated optimization with error feedback. *Automatica*, 183:112655, 2026.

Upadhyaya, M., Banert, S., Taylor, A. B., and Giselsson, P. Automated tight Lyapunov analysis for first-order methods. *Mathematical Programming*, 209(1):133–170, 2025.

Vaswani, S., Bach, F., and Schmidt, M. Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. In *The 22nd international conference on artificial intelligence and statistics*, pp. 1195–1204. PMLR, 2019.

Vempala, S. S. *The random projection method*, volume 65. American Mathematical Soc., 2005.

Wu, J., Huang, W., Huang, J., and Zhang, T. Error Compensated Quantized SGD and its Applications to Large-scale Distributed Optimization. In *International Conference on Machine Learning (ICML)*, 2018.

Zheng, S., Huang, Z., and Kwok, J. Communication-Efficient Distributed Blockwise Momentum SGD with Error-Feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

## Organization of the Appendix



This appendix provides additional content and details complementing the paper. In particular, Section A provides the complete missing proofs for the main results of the paper. Section B presents additional numerical results. Finally, Section C details the verification methodology used to validate the empirical laws.

# A. Proofs

## A.1. Proof of Proposition 2.1

**Proposition 2.1.** *Let there be $n = 2$ agents, each having access to first-order oracles querying the one-dimensional quadratic functions defined in subsection 2.1. Then there are 2-step cycles for the following scenarios where the step size $\eta > 0$ is fixed:*

1. $\eta < \frac{2}{\mu}$, *with* $x_0 = \eta\frac{\sqrt{\epsilon}}{2-\eta\mu}$.
2. $\eta > \frac{2}{\mu}$, *with* $x_0 = -\eta\frac{\sqrt{\epsilon}}{2-\eta\mu}$.
3. $\eta = \frac{2}{\mu}$, *with any* $x_0$.

*Proof.* Begin by dealing with the first case, where $\eta < \frac{2}{\mu}$. Use the same initialization as in subsection 2.1. Set the compression oracle responses to also be identical to the compression oracle responses in subsection 2.1. Since the initial error terms are all zero, the first step will be the same as in subsection 2.1. After the first round of communication, the error terms are given by

$$\begin{aligned} e_1^{(1)} &= e_0^{(1)} + \eta(f^{(1)})'(x_0) - \mathcal{C}(e_0^{(1)} + \eta(f^{(1)})'(x_0); \omega_0^{(1)}) = \sqrt{\epsilon}\eta(f^{(1)})'(x_0), \\ e_1^{(2)} &= e_0^{(2)} + \eta(f^{(2)})'(x_0) - \mathcal{C}(e_0^{(2)} + \eta(f^{(2)})'(x_0); \omega_0^{(2)}) = -\sqrt{\epsilon}\eta(f^{(2)})'(x_0). \end{aligned}$$

Now, set the compression oracles to give the gradients at $x_1$ without any compression, i.e., $m_1^{(1)} = (f^{(1)})'(x_1)$ and $m_1^{(2)} = (f^{(2)})'(x_1)$. This will result in the following updates to the error terms:

$$\begin{aligned} e_2^{(1)} &= e_1^{(1)} + \eta(f^{(1)})'(x_1) - \mathcal{C}\left(e_1^{(1)} + \eta(f^{(1)})'(x_1); \omega_1^{(1)}\right) = 0, \\ e_2^{(2)} &= e_1^{(2)} + \eta(f^{(2)})'(x_1) - \mathcal{C}\left(e_1^{(2)} + \eta(f^{(2)})'(x_1); \omega_1^{(2)}\right) = 0, \end{aligned}$$

meaning that the error terms are zero after the second step. This also results in the update

$$\begin{aligned} x_2 &= x_1 - \frac{1}{2}\left[e_1^{(1)} + \eta(f^{(1)})'(x_1) + e_1^{(2)} + \eta(f^{(2)})'(x_1)\right] \\ &= x_1 - \frac{\eta}{2}\left[\sqrt{\epsilon}(\mu x_0 - 1) + (\mu x_1 - 1) - \sqrt{\epsilon}(\mu x_0 + 1) + (\mu x_1 + 1)\right] \\ &= x_1 - \eta\left(\mu x_1 - \sqrt{\epsilon}\right) = x_0, \end{aligned}$$

where the last step follows from subsection 2.1 and some basic algebra. Since the iterate is back at the starting point, with the error terms being zero, the cycle is complete.

The second case where $\eta > \frac{2}{\mu}$ is given by the converse argument, with the starting point $x_0 = -\frac{\eta\sqrt{\epsilon}}{2-\eta\mu}$ and compression oracles that on the first step give the responses

$$\mathcal{C}((f^{(1)})'(x_0); \omega_0^{(1)}) = (1 + \sqrt{\epsilon})(f^{(1)})'(x_0), \qquad \mathcal{C}((f^{(2)})'(x_0); \omega_0^{(2)}) = (1 - \sqrt{\epsilon})(f^{(2)})'(x_0).$$

The third case is given by simply considering what would happen if all the compression oracles just gave the true gradient updates during all communication rounds. Then there would be no compression nor error feedback, making the iteration equivalent to distributed gradient descent. The cycle follows from the following computations:

$$\begin{aligned} x_1 &= x_0 - \frac{\eta}{2}\left[(f^{(1)})'(x_0) + (f^{(2)})'(x_0)\right] \\ &= x_0 - 2\eta\mu x_0 \\ &= -x_0, \end{aligned}$$

and,

$$\begin{aligned} x_2 &= x_1 - \frac{\eta}{2}\left[(f^{(1)})'(x_1) + (f^{(2)})'(x_1)\right] \\ &= x_1 - 2\eta\mu x_1 \\ &= -x_1. \end{aligned}$$

□

### A.2. Proof of Theorem 3.1

**Theorem 3.1.** *Let $\epsilon \in [0,1)$ and assume that the compression operator $\mathcal{C}$ satisfies Assumption 1.1. Let $f^{(i)} \in \mathcal{F}_{\mu,L}$ for each $i \in [n]$. Let the step size be given by*

$$\eta^\star = \left(\frac{2}{L+\mu}\right)\cdot\left(\frac{1-\sqrt{\epsilon}}{1+\sqrt{\epsilon}}\right). \tag{4}$$

*Then, the Lyapunov function*

$$\begin{aligned} \mathcal{V}\big(\xi^{\mathrm{EF21}}, x; f\big) \coloneqq & \frac{\sqrt{\epsilon}}{n}\left\|\sum_{i=1}^n \nabla f^{(i)}(x_k)\right\|^2 \\ & + \sum_{i=1}^n \|\nabla f^{(i)}(x_k) - d_k^{(i)}\|^2 \end{aligned} \tag{5}$$

*is optimal (i.e., solves (2)) and satisfies*

$$\mathbb{E}_\omega\left[\mathcal{V}\big(\xi_{k+1}^{\mathrm{EF21}}, x_{k+1}; f\big)\right] \le \rho_\star \cdot \mathbb{E}_\omega\left[\mathcal{V}\big(\xi_k^{\mathrm{EF21}}, x_k; f\big)\right]$$

*where the rate is given by*

$$\rho_\star \coloneqq \sqrt{\epsilon} + \left(\frac{1-\sqrt{\epsilon}}{2}\right)\left(\frac{\kappa-1}{\kappa+1}\right)^2 \Psi(\kappa,\epsilon) \tag{6}$$

*and*

$$\Psi(\kappa,\epsilon) \coloneqq 1 - \sqrt{\epsilon} + \sqrt{(1+\sqrt{\epsilon})^2 + \sqrt{\epsilon}\,16\frac{\kappa}{(\kappa-1)^2}}.$$

*Finally, the step size in (4) is worst-case optimal for $\mathrm{EF}^{21}$: within the candidate Lyapunov class based on $\xi^{\mathrm{EF}^{21}}$, it achieves the minimal worst-case one-step contraction factor, $\rho_\star$ in (6). This bound is also multi-step tight, meaning that after $k$ iterations the worst-case contraction equals $\rho_\star^k$.*

*Proof.* The proof of the announced convergence rate follows. Denote $g_k^{(i)} \coloneqq \nabla f^{(i)}(x_k)$, so that Algorithm 2 writes as

$$x_{k+1} = x_k - \eta\cdot\frac{1}{n}\sum_{i=1}^n d_k^{(i)}, \qquad d_{k+1}^{(i)} = d_k^{(i)} + \mathcal{C}(g_{k+1}^{(i)} - d_k^{(i)}; \omega_k^{(i)}).$$

Consider the following inequalities, and associate with each of them the assigned multiplier:

$$\begin{aligned} I_{\mathcal{F}_{\mu,L}}^{(i,1)} \coloneqq & f^{(i)}(x_k) - f^{(i)}(x_{k+1}) + \frac{\|g_{k+1}^{(i)} - g_k^{(i)}\|^2}{2L} + \langle g_k^{(i)}, x_{k+1} - x_k\rangle \\ & + \frac{\mu}{2(1-\mu/L)}\left\|x_k - x_{k+1} - \frac{1}{L}\left(g_k^{(i)} - g_{k+1}^{(i)}\right)\right\|^2 \le 0, \end{aligned} \quad : \lambda_{\mathrm{EF}^{21}}$$

$$\begin{aligned} I_{\mathcal{F}_{\mu,L}}^{(i,2)} \coloneqq & f^{(i)}(x_{k+1}) - f^{(i)}(x_k) + \frac{\|g_k^{(i)} - g_{k+1}^{(i)}\|^2}{2L} + \langle g_{k+1}^{(i)}, x_k - x_{k+1}\rangle \\ & + \frac{\mu}{2(1-\mu/L)}\left\|x_{k+1} - x_k - \frac{1}{L}\left(g_{k+1}^{(i)} - g_k^{(i)}\right)\right\|^2 \le 0, \end{aligned} \quad : \lambda_{\mathrm{EF}^{21}}$$

$$I_{\mathcal{C}}^{(i)} \coloneqq \mathbb{E}_{\omega_k^{(i)}}\left[\|g_{k+1}^{(i)} - d_k^{(i)} - \mathcal{C}(g_{k+1}^{(i)} - d_k^{(i)}; \omega_k^{(i)})\|^2\right] - \epsilon\,\mathbb{E}_{\omega_k^{(i)}}\left[\|g_{k+1}^{(i)} - d_k^{(i)}\|^2\right] \le 0, \quad : \nu_{\mathrm{EF}^{21}},$$

where $i \in [n]$, $\nu_{\mathrm{EF}^{21}} \coloneqq 1$, and where $\lambda_{\mathrm{EF}^{21}}$ is defined as

$$\lambda_{\mathrm{EF}^{21}} \coloneqq \frac{\sqrt{\epsilon}}{\eta^\star(L+\mu)}\left[(1-\sqrt{\epsilon})(L-\mu)+(1+\sqrt{\epsilon})\sqrt{(L-\mu)^2+\frac{16L\mu\sqrt{\epsilon}}{(1+\sqrt{\epsilon})^2}}\right].$$

Summing the inequalities with their multipliers, the following algebraic identity holds:

$$\lambda_{\mathrm{EF}^{21}}\sum_{i=1}^{n}(I^{(i,1)}_{\mathcal{F}_{\mu,L}}+I^{(i,2)}_{\mathcal{F}_{\mu,L}})+\nu_{\mathrm{EF}^{21}}\sum_{i=1}^{n}I^{(i)}_{\mathcal{C}}=\mathbb{E}_{\omega_k}\left[\mathcal{V}(\xi_{k+1})\right]-\rho\mathcal{V}(\xi_k)+S, \tag{7}$$

where the residual $S$ is a sum of squares defined as

$$S \coloneqq naS_{\text{mean}}+cS_{\text{var}}+(\rho-\epsilon)S_{\text{mix}},$$

with the components

$$\begin{aligned}
S_{\text{mean}} &\coloneqq \left\|\bar{d}_k+\frac{1}{a}\left[(\epsilon+b)\bar{g}_{k+1}-(\rho+b)\bar{g}_k\right]\right\|^2,\\
S_{\text{var}} &\coloneqq \sum_{i=1}^{n}\left\|(g^{(i)}_{k+1}-\bar{g}_{k+1})-(g^{(i)}_k-\bar{g}_k)\right\|^2,\\
S_{\text{mix}} &\coloneqq \sum_{i=1}^{n}\left\|(d^{(i)}_k-\bar{d}_k)+\frac{\epsilon}{\rho-\epsilon}(g^{(i)}_{k+1}-\bar{g}_{k+1})-\frac{\rho}{\rho-\epsilon}(g^{(i)}_k-\bar{g}_k)\right\|^2,
\end{aligned}$$

where $\bar{g} \coloneqq \frac{1}{n}\sum_{i=1}^{n} g^{(i)}$ and $\bar{d} \coloneqq \frac{1}{n}\sum_{i=1}^{n} d^{(i)}$ denote the averages, and the coefficients are given by

$$a \coloneqq \rho-\epsilon+\lambda_{\mathrm{EF}^{21}}\eta^2\frac{L\mu}{L-\mu},\quad b \coloneqq \frac{\eta\lambda_{\mathrm{EF}^{21}}}{2}\cdot\frac{L+\mu}{L-\mu},\quad c \coloneqq \frac{\lambda_{\mathrm{EF}^{21}}}{L-\mu}-\frac{\epsilon\rho}{\rho-\epsilon}.$$

Note that the positivity of $a$ is guaranteed for $\eta^\star$, as noted in the single-agent proof in (Berg Thomsen et al., 2025). To see that $c>0$, set $t \coloneqq \sqrt{(1+\sqrt{\epsilon})^2+\frac{16L\mu\sqrt{\epsilon}}{(L-\mu)^2}}$ and $u \coloneqq 1-\sqrt{\epsilon}+t$ (so $u \geq 2$). Then

$$(L-\mu)(\rho-\epsilon)c = a_2(u-2)^2+a_1(u-2)+\frac{\sqrt{\epsilon}}{(L+\mu)^2}\left[(1+\sqrt{\epsilon}+\epsilon)(L-\mu)^2+4\sqrt{\epsilon}\,L\mu\right],$$

with $a_2, a_1 > 0$, hence $c>0$.

Since $\lambda_{\mathrm{EF}^{21}} \geq 0$, the weighted sum of inequalities (LHS of (7)) is nonpositive. The statement now follows by plugging in $\eta=\eta^\star$ and $\rho=\rho_\star$ and checking that all coefficients in (7) are nonnegative.

To show tightness of the bound, consider the case where all agents have the same objective function $f^{(i)}=f$. Initialize $d^{(i)}_0 = d_0 = \mathcal{C}(\nabla f(x_0);\omega_0)$ and assume identical compression realizations. Then $d^{(i)}_k = d_k$ for all $k$, and the algorithm updates match the single-agent $\mathrm{EF}^{21}$. The single-agent tight lower bound thus limits the multi-agent performance. Furthermore, since this scenario is a worst-case instance, the optimal step size for the single-agent setting is also worst-case optimal for the multi-agent setting. In particular, since the dynamics of the identical-functions case are invariant to the number of agents $n$, the worst-case optimal step size must also be independent of $n$.

The optimality and multi-step tightness of the Lyapunov function follow from the theory of linear dynamical systems. For the methods and function classes considered, the worst-case contraction is governed by the spectral radius of the iteration matrix on quadratic instances. The class of candidate Lyapunov functions matches this spectral behavior, ensuring that the single-step analysis is optimal and remains tight over multiple iterations (i.e., $\mathcal{V}_k \leq \rho^k\mathcal{V}_0$ is achievable). □

### A.3. Proof of Corollary 3.2

**Corollary 3.2.** *Let $\epsilon \in [0,1)$ and assume that the compression operator $\mathcal{C}$ satisfies Assumption 1.1 and is deterministic, additive, and positively homogeneous (so $\mathcal{C}(x+y)=\mathcal{C}(x)+\mathcal{C}(y)$ and $\mathcal{C}(\alpha x)=\alpha\mathcal{C}(x)$ for all $\alpha \geq 0$). Let $f^{(i)} \in \mathcal{F}_{\mu^{(i)},L^{(i)}}$*

*for each $i \in [n]$, and define*

$$\bar{L} \coloneqq \frac{1}{n}\sum_{i=1}^{n} L^{(i)}, \qquad \bar{\mu} \coloneqq \frac{1}{n}\sum_{i=1}^{n} \mu^{(i)}.$$

$$\kappa_\Sigma \coloneqq \frac{\bar{L}}{\bar{\mu}} = \frac{\sum_{i=1}^{n} L^{(i)}}{\sum_{i=1}^{n} \mu^{(i)}}.$$

*Let the step size be given by*

$$\eta^\star = \left(\frac{2}{\bar{L}+\bar{\mu}}\right)\cdot\left(\frac{1-\sqrt{\epsilon}}{1+\sqrt{\epsilon}}\right).$$

*Then the Lyapunov function*

$$\mathcal{V}_{\mathrm{lin}}\left(\xi^{\mathrm{EF21}}, x; f\right) \coloneqq \sqrt{\epsilon}\,\|\bar{g}_k\|^2 + \|\bar{g}_k - \bar{d}_k\|^2$$

*satisfies*

$$\mathbb{E}_\omega\left[\mathcal{V}_{\mathrm{lin}}\left(\xi^{\mathrm{EF21}}_{k+1}, x_{k+1}; f\right)\right] \le \rho_\star \cdot \mathbb{E}_\omega\left[\mathcal{V}_{\mathrm{lin}}\left(\xi^{\mathrm{EF21}}_{k}, x_{k}; f\right)\right],$$

*where $\bar{g}_k \coloneqq \frac{1}{n}\sum_{i=1}^{n} \nabla f^{(i)}(x_k)$, $\bar{d}_k \coloneqq \frac{1}{n}\sum_{i=1}^{n} d_k^{(i)}$, and the rate $\rho_\star$ is given by (6) with $\kappa = \kappa_\Sigma$.*

*Proof.* Recall $f(x) \coloneqq \frac{1}{n}\sum_{i=1}^{n} f^{(i)}(x)$, and denote $\bar{g}_k \coloneqq \nabla f(x_k) = \frac{1}{n}\sum_{i=1}^{n} \nabla f^{(i)}(x_k)$. Since each $f^{(i)}$ is $L^{(i)}$-smooth and $\mu^{(i)}$-strongly convex, $f$ is $\bar{L}$-smooth and $\bar{\mu}$-strongly convex with $\bar{L} = \frac{1}{n}\sum_i L^{(i)}$ and $\bar{\mu} = \frac{1}{n}\sum_i \mu^{(i)}$. Consequently, the standard interpolation (cocoercivity) inequalities for $\mathcal{F}_{\bar{\mu},\bar{L}}$ apply directly to $\bar{g}_k$ and $\bar{g}_{k+1}$.

The averaged dynamics satisfy

$$x_{k+1} = x_k - \eta\,\bar{d}_k,$$

and, due to additivity and positive homogeneity of $\mathcal{C}$,

$$\begin{aligned}\bar{d}_{k+1} &= \frac{1}{n}\sum_{i=1}^{n} d_{k+1}^{(i)} = \frac{1}{n}\sum_{i=1}^{n}\left(d_k^{(i)} + \mathcal{C}\left(g_{k+1}^{(i)} - d_k^{(i)}\right)\right)\\ &= \bar{d}_k + \mathcal{C}\left(\frac{1}{n}\sum_{i=1}^{n}\left(g_{k+1}^{(i)} - d_k^{(i)}\right)\right) = \bar{d}_k + \mathcal{C}(\bar{g}_{k+1} - \bar{d}_k).\end{aligned}$$

Therefore, the barred variables evolve as a single-agent $\mathrm{EF}^{21}$ instance on $f$ with parameters $(\bar{\mu}, \bar{L})$. The conclusion follows by applying Theorems 1–2 in Berg Thomsen et al. (2025) to this averaged single-agent instance. The optimal step size and rate are obtained by substituting $\kappa_\Sigma = \bar{L}/\bar{\mu} = \sum_i L^{(i)}/\sum_i \mu^{(i)}$ into (6). Theorem 3 in Richtárik et al. (2021) establishes that EF and $\mathrm{EF}^{21}$ are equivalent under deterministic, additive, positively homogeneous compressors, so the same rate applies to EF. □

### A.4. Proof of Theorem 3.4

**Theorem 3.4.** *Let $\epsilon \in [0,1)$ and assume that the compression operator $\mathcal{C}$ satisfies Assumption 1.1. Let $f^{(i)} \in \mathcal{F}_{\mu,L}$ for each $i \in [n]$, and assume Assumption 3.3.*

*Let the step size be given by $\eta^\star$ as defined in (4). Then, the Lyapunov function*

$$\mathcal{V}\left(\xi^{\mathrm{EF}}, x; f\right) \coloneqq \frac{1}{n\sqrt{\epsilon}}\left\|\sum_{i=1}^{n} e_k^{(i)}\right\|^2 + \sum_{i=1}^{n}\|x_k - x_\star - e_k^{(i)}\|^2.$$

*is optimal and satisfies*

$$\rho_\star(\mathrm{EF}) = \rho_\star,$$

*where $\rho_\star$ is defined in (6). Finally, the step size in (4) is worst-case optimal for* EF*: within the candidate Lyapunov class based on $\xi^{\mathrm{EF}}$, it achieves the minimal worst-case one-step contraction factor, $\rho_\star$ in (6). This bound is also multi-step tight, meaning that after $k$ iterations the worst-case contraction equals $\rho_\star^k$.*

*Proof.* The proof of the announced convergence rate follows. Algorithm 1 can be rewritten as

$$x_{k+1} = x_k - \frac{1}{n}\sum_{i=1}^{n} m_k^{(i)}, \qquad e_{k+1}^{(i)} = e_k^{(i)} + \eta\nabla f^{(i)}(x_k) - m_k^{(i)},$$

where $m_k^{(i)} = \mathcal{C}(e_k^{(i)} + \eta\nabla f^{(i)}(x_k); \omega_k^{(i)})$. Consider the following inequalities:

$$\begin{aligned} I_{\mathcal{F}_{\mu,L}}^{(i,1)} &:= f^{(i)}(x_k) - f^{(i)}(x^\star) - \langle \nabla f^{(i)}(x_k), x_k - x^\star\rangle + \frac{1}{2L}\|\nabla f^{(i)}(x_k)\|^2 \\ &\quad + \frac{\mu}{2(1-\mu/L)}\left\|x_k - x^\star - \frac{1}{L}\nabla f^{(i)}(x_k)\right\|^2 \le 0, && : \lambda_{\mathrm{EF}} \\ I_{\mathcal{F}_{\mu,L}}^{(i,2)} &:= f^{(i)}(x^\star) - f^{(i)}(x_k) + \frac{1}{2L}\|\nabla f^{(i)}(x_k)\|^2 \\ &\quad + \frac{\mu}{2(1-\mu/L)}\left\|x_k - x^\star - \frac{1}{L}\nabla f^{(i)}(x_k)\right\|^2 \le 0, && : \lambda_{\mathrm{EF}} \\ I_{\mathcal{C}}^{(i)} &:= \mathbb{E}_{\omega_k^{(i)}}\left[\|e_{k+1}^{(i)}\|^2\right] - (1-\epsilon)\mathbb{E}_{\omega_k^{(i)}}\left[\|e_k^{(i)} + \eta g_k^{(i)}\|^2\right] - \mathbb{E}_{\omega_k^{(i)}}\left[\|\mathcal{C}(e_k^{(i)} + \eta g_k^{(i)}; \omega_k^{(i)})\|^2\right] \le 0, && : \nu_{\mathrm{EF}}, \end{aligned}$$

where $i \in [n]$, $\nu_{\mathrm{EF}} := 1/\sqrt{\epsilon}$, and $\lambda_{\mathrm{EF}}$ is defined as

$$\lambda_{\mathrm{EF}} := \frac{\eta^\star}{L+\mu}\left[(1-\sqrt{\epsilon})(L-\mu) + (1+\sqrt{\epsilon})\sqrt{(L-\mu)^2 + \frac{16L\mu\sqrt{\epsilon}}{(1+\sqrt{\epsilon})^2}}\right].$$

Summing these inequalities with their multipliers yields the algebraic identity:

$$\lambda_{\mathrm{EF}}\sum_{i=1}^{n}(I_{\mathcal{F}_{\mu,L}}^{(i,1)} + I_{\mathcal{F}_{\mu,L}}^{(i,2)}) + \nu_{\mathrm{EF}}\sum_{i=1}^{n} I_{\mathcal{C}}^{(i)} = \mathbb{E}_{\omega_k}\left[\mathcal{V}(\xi_{k+1})\right] - \rho\mathcal{V}(\xi_k) + S,$$

where the residual $S$ is given by

$$S := naS_{\text{mean}} + cS_{\text{var}} + (\rho - \sqrt{\epsilon})S_{\text{mix}} + (\nu_{\mathrm{EF}} - 1)S_{\text{mix-comp}},$$

with the components

$$\begin{aligned} S_{\text{mean}} &:= \left\|\bar{e}_k - \frac{\rho-1}{a}(x_k - x^\star) + \frac{2(\sqrt{\epsilon}-1)}{a(L+\mu)}\bar{g}_k\right\|^2, \\ S_{\text{var}} &:= \sum_{i=1}^{n}\left\|g_k^{(i)} - \bar{g}_k\right\|^2, \\ S_{\text{mix}} &:= \sum_{i=1}^{n}\left\|(e_k^{(i)} - \bar{e}_k) - \frac{\sqrt{\epsilon}\,\eta}{\rho-\sqrt{\epsilon}}(g_k^{(i)} - \bar{g}_k)\right\|^2, \\ S_{\text{mix-comp}} &:= \sum_{i=1}^{n}\mathbb{E}_{\omega_k}\left[\|\mathcal{C}(e_k^{(i)} + \eta g_k^{(i)}; \omega_k^{(i)}) - \frac{1}{n}\sum_{j=1}^{n}\mathcal{C}(e_k^{(j)} + \eta g_k^{(j)}; \omega_k^{(j)})\|^2\right]. \end{aligned}$$

Here $\bar{e}_k, \bar{g}_k$ denote averages. As in the single-agent proof, the positivity of $\rho - \sqrt{\epsilon}$ (hence $a$) for $\eta^\star$ follows directly from the single-agent analysis. Let $s := \sqrt{\epsilon}$ and $t := \sqrt{(L-\mu)^2 + \frac{16L\mu s}{(1+s)^2}}$. This also gives $\nu_{\mathrm{EF}} - 1 = \frac{1-s}{s} > 0$, and

$$c = \frac{2(1-s)\big((1-s)(L-\mu) + (1+s)t\big)}{(1+s)^2(L-\mu)(L+\mu)^2} > 0.$$

To show tightness of the bound, consider the case where all agents have the same objective function $f^{(i)} = f$. With initialization $e_0^{(i)} = e_0 = 0$ and identical compression realizations, $e_k^{(i)} = e_k$ for all $k$, and the algorithm updates match

the single-agent EF. The single-agent tight lower bound thus limits the multi-agent performance. Furthermore, since this scenario is a worst-case instance, the optimal step size for the single-agent setting is also particular, since the dynamics of the identical-functions case are invariant to the number of agents $n$, the worst-case optimal step size must also be independent of $n$.

The optimality of the Lyapunov function and tightness over multiple iterations follow from the same argument as in the proof of Theorem 3.1. □

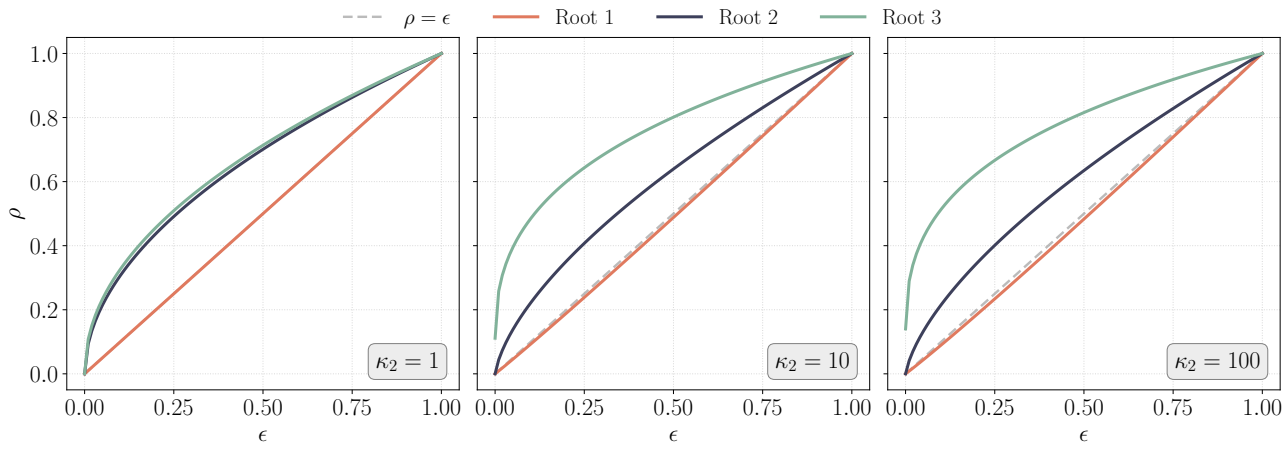


*Figure 2.* Bifurcation plot of the three roots of the cubic polynomial in Empirical Law 4.3 as $\epsilon$ varies. Parameters are $L^{(1)} = L^{(2)} = 1$, $\mu^{(1)} = 0.9$, and $\mu^{(2)} \in \{1, 0.1, 0.01\}$ from left to right.

## B. Additional Experiments

Figure 2 contains plots of the roots of the cubic polynomial defined in Empirical Law 4.3. None of the roots appears to have a directly identifiable closed form.

## C. Verification Details

Validation of the empirical laws was performed using the Performance Estimation Problem (PEP) framework (Drori, 2014; Taylor et al., 2017b). The verifications follow the SDP-based Lyapunov identification techniques of Taylor et al. (2018); Upadhyaya et al. (2025), implemented in Python using the `PEPit` library (Goujaud et al., 2024), with LMIs solved by MOSEK (ApS, 2025). Unless stated otherwise, the parameter grid uses 10 linearly spaced values for $\epsilon \in [0.05, 0.95]$, $L^{(i)} \in \{1, 10, 100\}$, and $\kappa^{(i)} \in \{2, 10, 100\}$, with $n \in \{2, 3\}$. To test tightness or optimality, we generally check feasibility at an improved target rate $\rho_{\text{imp}} = 0.99\rho$; any feasible point yielding a strict improvement is recorded as a failure. No such failures were found in any of the experiments below.

### C.1. Verification of Optimal Step Size (Empirical Law 4.1)

For each configuration, the theoretical step size $\eta^\star$ is computed and the corresponding rate $\rho_{\text{th}}$ is obtained by bisection on the Lyapunov LMI (full $\text{EF}^{21}$ class). This rate is also checked for feasibility in the EF LMI. If $\eta^\star < 10^{-2}$, the instance is skipped to avoid numerical issues. To test optimality, a grid of 20 step sizes over $[\eta_{\min}, 2n/\sum_i (L^{(i)} + \mu^{(i)})]$ with $\eta_{\min} = 10^{-2}$ is scanned. For each candidate $\eta$, feasibility is checked at $\rho_{\text{imp}} = 0.99\,\rho_{\text{th}}$ for both EF and $\text{EF}^{21}$. Due to the large number of configurations, redundant combinations of parameters $(\mu^{(i)}, L^{(i)}, \kappa^{(i)})$ are skipped, which is without loss-of-generality due to agent symmetry.

### C.2. Verification of Lyapunov Function Structure (Empirical Law 4.2)

At $\eta^\star$, PEPit is used to compute the worst-case rate $\rho_{\text{simp}}$ restricted to the simplified Lyapunov structure in Empirical Law 4.2. An improved target $\rho_{\text{imp}} = 0.99\,\rho_{\text{simp}}$ is then tested against the full Lyapunov class using the LMI feasibility oracle.

### C.3. Verification of $n = 2$ Rate (Empirical Law 4.3)

For $n = 2$ only, the candidate rate $\rho_\star$ is taken as the largest real root of the cubic in Empirical Law 4.3 and the step size is set to $\eta^\star$. Feasibility at $(\rho_\star, \eta^\star)$ is checked for $\text{EF}^{21}$ and for EF using the Lyapunov LMI. If the $\text{EF}^{21}$ LMI fails, a fallback check compares $\rho_\star$ to the simplified Lyapunov rate computed by PEPit and accepts if it is within $10^{-4}$. To test tightness, an improved target $\rho_{\text{imp}} = 0.99\,\rho_\star$ is attempted.

### C.4. Verification of EControl Tuning (Empirical Law 4.4)

For each configuration, the tuning $(\eta, \gamma) = (0, \gamma^\star_{\text{EControl}})$ is evaluated by computing the best feasible rate $\rho_\star$ via bisection on the EControl Lyapunov LMI. An improved target $\rho_{\text{imp}} = 0.99\,\rho_\star$ is then tested over a $10 \times 10$ grid with $\eta \in [0, 2n/\sum_i (L^{(i)} + \mu^{(i)})]$ and $\gamma \in [10^{-2}, 2n/\sum_i (L^{(i)} + \mu^{(i)})]$. Only $n = 2$ was considered due to the computational cost of the experiment.