ERROR FEEDBACK FOR ACCELERATED DISTRIBUTED OPTIMIZATION WITH SINGLE COMPRESSED UPLINK

Anonymous authors
Paper under double-blind review

ABSTRACT

Distributed optimization across multiple workers can incur substantial communication costs due to frequent exchange with a central server. Compressing high-dimensional gradients alleviates this bottleneck, but biased compression introduces persistent errors that complicate acceleration. Error feedback can correct compression error, yet combining it with Nesterov acceleration is subtle because acceleration accumulates gradient-estimation errors with time-varying weights. We introduce **EC-NAG**, a Nesterov-aligned error-compensation method for smooth distributed convex optimization. Each worker compresses a gradient-difference message corrected using its two most recent controlled-error states, with coefficients chosen so that the Nesterov-weighted history of gradient-estimation errors telescopes to the current averaged error state. Consequently, EC-NAG requires only one compressed uplink vector per worker per iteration. We develop a non-asymptotic analysis and show that EC-NAG preserves the deterministic $\mathcal{O}(1/T^2)$ accelerated rate, while in the stochastic setting it achieves the standard $\mathcal{O}(1/\sqrt{T})$ decay. Thus, we match the stochastic lower-bound exponent in ε. Experiments on distributed least-squares and classification problems demonstrate favorable convergence as a function of communicated uplink bits.

1 INTRODUCTION

Large-scale learning increasingly distributes data and computation across clusters, institutions, and edge devices. This architecture underlies data-parallel training of large models and federated learning across devices or organizations that cannot centralize their data (Shoeybi et al., 2019; McMahan et al., 2017; Kairouz et al., 2021; Jiang et al., 2024). In the parameter-server model, worker i stores a local objective $f_i\colon \mathbb{R}^d \to \mathbb{R}$, and the workers jointly solve

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

Although computation is parallelized across workers, every optimization round requires communication with the server. For contemporary high-dimensional models, repeatedly transmitting dense gradient- or model-sized vectors can dominate latency, energy use, and total training time. Communication compression therefore reduces not the local arithmetic, but the quantity that is often most expensive in a distributed system: the information exchanged between workers and the server.

Quantization, sparsification, and low-rank approximation reduce this communication burden (Alistarh et al., 2017; Stich et al., 2018; Vogels et al., 2019). A particularly useful class consists of contractive compressors, which includes important biased operators such as sparsifiers and fixed-rate quantizers. Biased compression is attractive in practice, but simply replacing an exact gradient by a compressed one can leave a persistent bias and can even destroy convergence (Beznosikov et al., 2023). Error feedback (EF) addresses this problem by retaining the information discarded by the compressor and feeding it into future messages (Seide et al., 2014; Alistarh et al., 2018; Karimireddy et al., 2019). This simple principle has become one of the standard mechanisms for using biased compression without permanently losing the discarded update.

The distributed setting introduces a second obstruction. Even at a minimizer $\mathbf{x}^*$ of the global objective, the local gradients $\nabla f_i(\mathbf{x}^*)$ need not vanish. Hence, under arbitrary data heterogeneity,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

repeatedly compressing raw local gradients may continue to inject a non-vanishing local signal after the global optimization error has become small. Gradient-difference (gradient-tracking) mechanisms avoid this difficulty by communicating changes in local gradients and maintaining a local estimate of the missing information. This viewpoint appears in DIANA, EF21, and the unified EF-BV framework (Mishchenko et al., 2025; Richtárik et al., 2021; Condat et al., 2022). These methods established that suitable memory mechanisms can remove restrictive data-similarity assumptions and make biased or unbiased compression compatible with heterogeneous distributed optimization.

Stochastic gradients make this tracking problem more delicate. A full-gradient tracking error can contract as the iterates approach one another, whereas a fixed-size stochastic mini-batch continuously injects fresh oracle noise into the memory state. This distinction is important for EF21: the basic tracking mechanism is naturally suited to the full-gradient regime, while stochastic variants require large batch size (Richtárik et al., 2021). EF21-SGDM uses Polyak momentum to overcome the large-batch difficulty of stochastic error feedback, while EControl introduces a tunable feedback strength and obtains guarantees for stochastic strongly convex, general convex, and non-convex problems without requiring bounded gradient dissimilarity (Fatkhullin et al., 2023; Gao et al., 2024). These developments resolve important stochastic-stability issues, but neither result by itself supplies Nesterov's smooth general-convex acceleration: Polyak momentum and Nesterov acceleration have different dynamics and different proof invariants.

For a smooth convex objective, gradient descent has the worst-case function-value rate $\mathcal{O}(1/T)$, whereas Nesterov acceleration achieves the first-order-optimal $\mathcal{O}(1/T^2)$ deterministic rate (Nesterov, 1983; 2018; Bubeck, 2015). This improvement is driven by a time-varying weighted accumulation of gradients. Under compression, the same weights also accumulate gradient-estimation errors. Consequently, an error-feedback rule that cancels an ordinary, unweighted history need not cancel the error history seen by the accelerated dynamics. In fact, this incompatibility has motivated acceleration-specific feedback mechanisms (Tyurin & Richtárik, 2023).

Other accelerated compressed methods occupy still different regimes. ADIANA obtains acceleration under unbiased-compressor models (Li et al., 2020). ECLK combines error compensation with a Katyusha-type variance-reduction construction in a strongly convex finite-sum setting (Qian et al., 2021). More recently, EF-OLGA combines acceleration and biased-compression error feedback under data-similarity assumptions, with its theory built around a strongly convex finite-sum structure (Bylinkin & Beznosikov, 2025). NEOLITHIC, on the other hand, treats broad stochastic regimes and approaches the corresponding communication lower bounds up to logarithmic factors, but it controls compression error by repeated communication between server updates (Huang et al., 2022; He et al., 2023). Thus, acceleration under communication compression is possible, but the available guarantees differ substantially in compressor assumptions, stochastic-oracle model, convexity regime, and the number of messages used to realize one optimization step.

The closest prior work to ours is ADEF (Gao et al., 2026). ADEF established an accelerated rate for stochastic distributed optimization with contractive compression in the smooth general-convex regime. Importantly, however, its accelerated estimator is built from two separately compressed worker messages: one compressed gradient-difference message maintains a control variate, while a second compressed message carries the error-feedback correction. Each worker therefore performs two compression operations and sends two compressed high-dimensional vectors to the server in every iteration. ADEF establishes that acceleration is possible in this difficult stochastic setting, but it leaves open whether these two compressed uplinks are structurally necessary.

This is the question addressed in the present work. We are first to show that the two roles played by the compressed messages can be coupled into a *single* worker-to-server message without destroying the accelerated error-cancellation structure. We introduce EC-NAG, in which worker i compresses one gradient-difference message corrected by two consecutive controlled-error states. The coefficients of these two residual terms are not generic momentum coefficients: they are determined by the Nesterov weights a_t/a_{t+1} together with a feedback gain γ. This construction gives a weighted identity, which collapses the complete Nesterov-weighted history of gradient-estimation errors to the current averaged controlled state.

The single-uplink construction also determines the analysis. We derive a self-contained non-asymptotic inexact-NAG Lyapunov recursion in which the compression error enters through the weighted controlled state rather than through an ever-growing history. The proof retains the nega-

tive Bregman-divergence terms generated by Nesterov's method and uses them to absorb the perturbations produced by gradient-difference compression. We then analyze the coupled compressor-input and controlled-state recursions, derive the admissible feedback range, and obtain a smooth general-convex convergence guarantee. In the deterministic regime, EC−NAG retains the accelerated $\mathcal{O}(1/T^2)$ function-value rate. With stochastic gradients, the error decays as $\mathcal{O}(1/\sqrt{T})$ and the dependence on target accuracy has the standard ε^{-2} stochastic order.

We compare EC−NAG with full-precision NAG, classical error feedback EF, EControl, and the closest accelerated baseline ADEF, and report progress against communicated uplink bits. The choice of direct experimental baselines follows the oracle and compressor model of the theorem: EF, EControl, and ADEF can be run under the same biased contractive-compression and stochastic-oracle protocol. EC−NAG reaches the reported accuracy levels using fewer uplink bits in the tested configurations.

The remainder of the paper develops this argument in that order. Section 2 derives the one-uplink acceleration-aligned feedback rule and its weighted telescoping identity. Section 3 gives the convergence theorem and quantizer specialization, Section 4 presents the communication-normalized experiments, and Section 5 concludes. The appendices give the design justification, complete convergence proof, extra experiments, and the implementation specifications.

2 Proposed Error Feedback for NAG

This section presents the error-feedback mechanism underlying EC−NAG. Its construction is best understood by following the progression from standard gradient descent to acceleration, and finally extending to compressed communication.

In the uncompressed centralized setting, gradient descent updates

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t \, \mathbf{g}_t, \tag{2}$$

where $\eta_t > 0$ is the stepsize and $\mathbf{g}_t$ denotes the exact or stochastic gradient available at $\mathbf{x}_t$. In a distributed environment, each worker i transmits a difference-of-gradients (gradient tracking) message:

$$\mathbf{u}_t^i = \mathbf{g}_t^i - \mathbf{g}_{t-1}^i, \tag{3}$$

where $\mathbf{g}_t^i$ is a (possibly stochastic) gradient $\nabla f_i(\mathbf{x}_t)$. The server recursively updates its gradient estimator via:

$$\widehat{\mathbf{g}}_t = \widehat{\mathbf{g}}_{t-1} + \frac{1}{n}\sum_{i=1}^n \mathbf{u}_t^i = \frac{1}{n}\sum_{i=1}^n \mathbf{g}_t^i = \mathbf{g}_t. \tag{4}$$

The gradient tracking-based representation is crucial for compressed methods (elaborated in Appendix B.1), and in the absence of compression, it reconstructs the true average gradient exactly.

Nesterov's Accelerated Gradient Descent (NAG) algorithm Nesterov (1983) improves the GD dynamics by coupling a "proximal" sequence $\{\mathbf{y}_t\}$ with an extrapolating "look-ahead" sequence $\{\mathbf{x}_t\}$ (Nesterov, 2018; Su et al., 2016). At each iteration, the algorithm computes a gradient step and then projects forward:

$$\mathbf{y}_{t+1} = \mathbf{x}_t - \eta_t \, \widehat{\mathbf{g}}_t, \tag{5}$$

$$\mathbf{x}_{t+1} = \mathbf{y}_{t+1} + \beta_t \left(\mathbf{y}_{t+1} - \mathbf{y}_t\right), \tag{6}$$

where $\beta_t > 0$ is the momentum parameter controlling the extrapolation weight (Beck & Teboulle, 2009; Bubeck, 2015). In the distributed uncompressed setting, the exact gradient in (4) can be obtained through the same gradient-tracking message (3). See Appendix A for a self-contained discussion of this classical scheme.

For high-dimensional distributed problems, compression is imperative and indispensable. However, because workers are heterogeneous and compress independently, the server's update is perturbed by the mixture of worker compression states $\mathbf{e}_t := \frac{1}{n}\sum_{i=1}^n \mathbf{e}_t^i$. Without feedback, compression errors corrupt the gradient estimator which is particularly harmful to NAG since it naturally propagates perturbations through its momentum step.

Algorithm 1: EC-NAG

Input: starting points $\mathbf{x}_0, \mathbf{y}_0 \in \mathbb{R}^d$, weight $A_0 \geq 0$, positive $\{a_t\}_{t\geq 1}$, $A_{t+1} = A_t + a_{t+1}$, feedback gain $\gamma \in (0, 1]$;

Initialize: $\mathbf{e}^i_{-1} = \mathbf{e}^i_{-2} = \mathbf{g}^i_{-1} = \mathbf{0}$ for all i and $\widehat{\mathbf{g}}_{-1} = \mathbf{0}$;

for $t = 0, 1, 2, \ldots, T-1$ **do**

  Worker (parallel for each i)

    Compute (stochastic) gradient $\mathbf{g}^i_t$ at $\mathbf{x}_t$;

    $\mathbf{u}^i_t = \left(\mathbf{g}^i_t - \mathbf{g}^i_{t-1}\right) + \frac{a_{t-1}}{a_t}\mathbf{e}^i_{t-2} - \left(\gamma + \frac{a_t}{a_{t+1}}\right)\mathbf{e}^i_{t-1}$;

    $\mathbf{q}^i_t = \mathcal{C}(\mathbf{u}^i_t)$;

    $\mathbf{e}^i_t = (1-\gamma)\mathbf{e}^i_{t-1} + \left(\mathbf{q}^i_t - \mathbf{u}^i_t\right)$;

    Send $\mathbf{q}^i_t$ to the server ;

  Server

    $\widehat{\mathbf{g}}_t = \widehat{\mathbf{g}}_{t-1} + \frac{1}{n}\sum_{i=1}^n \mathbf{q}^i_t$;

    $\mathbf{y}_{t+1} = \mathbf{x}_t - \eta_t \widehat{\mathbf{g}}_t$;

    $\mathbf{x}_{t+1} = \mathbf{y}_{t+1} + \beta_t\left(\mathbf{y}_{t+1} - \mathbf{y}_t\right)$;

    Broadcast $\mathbf{x}_{t+1}$ to all workers ;

Our proposed EC-NAG (Algorithm 1) modifies the gradient-tracking message by correcting it with two past compression states. Given $A_0 > 0$, positive weights $\{a_t\}_{t\geq 1}$, and a feedback gain $\gamma \in (0, 1]$, define $A_{t+1} = A_t + a_{t+1}$ and set $a_0 = 0$. The first message is $\mathbf{u}^i_0 = \mathbf{g}^i_0$. For $t \geq 1$,

$$\mathbf{u}^i_t = \mathbf{g}^i_t - \mathbf{g}^i_{t-1} + \frac{a_{t-1}}{a_t}\mathbf{e}^i_{t-2} - \left(\gamma + \frac{a_t}{a_{t+1}}\right)\mathbf{e}^i_{t-1}. \tag{7}$$

Here $\mathbf{e}^i_t$ is a controlled error state, not the usual compressor residual. With $\mathbf{e}^i_{-1} = \mathbf{0}$, it satisfies

$$\mathbf{e}^i_t - (1-\gamma)\mathbf{e}^i_{t-1} = \mathbf{q}^i_t - \mathbf{u}^i_t, \tag{8}$$

where $\mathbf{q}^i_t = \mathcal{C}(\mathbf{u}^i_t)$. Thus, $\gamma = 1$ recovers the ordinary compression residual, whereas $\gamma < 1$ retains a fraction $1 - \gamma$ of the previous state.

Substituting (8) and (7), the worker-local gradient tracker becomes:

$$\begin{aligned}
\widehat{\mathbf{g}}^i_t &\coloneqq \widehat{\mathbf{g}}^i_{t-1} + \mathbf{q}^i_t && (9)\\
&= \widehat{\mathbf{g}}^i_{t-1} + \mathbf{e}^i_t - (1-\gamma)\mathbf{e}^i_{t-1} + \mathbf{u}^i_t && (10)\\
&= \mathbf{g}^i_t + \mathbf{e}^i_t - \frac{a_t}{a_{t+1}}\mathbf{e}^i_{t-1} && (11)
\end{aligned}$$

after simplifying the recursion. Consequently, the server-side tracker (surrogate gradient) becomes:

$$\widehat{\mathbf{g}}_t \coloneqq \widehat{\mathbf{g}}_{t-1} + \frac{1}{n}\sum_{i=1}^n \mathbf{q}^i_t = \frac{1}{n}\sum_{i=1}^n \widehat{\mathbf{g}}^i_t = \mathbf{g}_t + \mathbf{e}_t - \frac{a_t}{a_{t+1}}\mathbf{e}_{t-1}. \tag{12}$$

Equation (7) represents the central structural property of EC-NAG. Crucially, the errors telescope and do not accumulate over time:

$$\frac{1}{a_{t+1}}\sum_{\tau=0}^t a_{\tau+1}(\widehat{\mathbf{g}}_\tau - \mathbf{g}_\tau) = \frac{1}{n}\sum_{i=1}^n \mathbf{e}^i_t = \mathbf{e}_t. \tag{13}$$

The identity explains why two state lags appear. Nesterov acceleration does not sum gradient errors uniformly: iteration τ is weighted by $a_{\tau+1}$. A standard one-residual correction cancels an unweighted history but generally leaves the changing coefficients behind. The gain γ changes the internal state dynamics without changing this identity: the term $-(1-\gamma)\mathbf{e}^i_{t-1}$ in (8) combines with $-(\gamma + a_t/a_{t+1})\mathbf{e}^i_{t-1}$ in (7), leaving a simple acceleration-aligned coefficient. Refer Appendix B.2 for detailed explanation.

Additionally, the stepsize and momentum parameter are related to $\{a_t, A_t\}$ as

$$\eta_t = \frac{a_{t+1}^2}{A_{t+1}} \quad \text{and} \quad \beta_t = \frac{a_{t+2}A_t}{a_{t+1}A_{t+2}}. \tag{14}$$

When the compressor is contractive and the feedback-stability condition holds, the controlled state remains bounded and EC−NAG behaves as NAG driven by a structured inexact gradient.

Overall, EC−NAG strikes a principled balance between communication efficiency and optimization speed, requiring only a single compressed message per worker per iteration.

3 Theoretical Convergence Results

Recall the distributed objective

$$f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}). \tag{15}$$

Throughout, $\|\cdot\|$ denotes the Euclidean norm. Below we specify the function and gradient assumptions, together with the compressor, used in the paper.

(A1) *Objective class.* The objective $f\colon \mathbb{R}^d \to \mathbb{R}$ is convex, closed, proper, and continuously differentiable. Additionally, f is L-smooth, i.e., $\frac{L}{2}\|\mathbf{x}\|^2 - f(\mathbf{x})$ is convex, equivalently, ∇f is L-Lipschitz:

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \le L\|\mathbf{x} - \mathbf{y}\|, \quad \forall\, \mathbf{x}, \mathbf{y} \in \mathbb{R}^d. \tag{16}$$

Assume there exists a minimizer, i.e., $\|\mathbf{x}^* - \mathbf{x}_0\| < \infty$, and that $f(\mathbf{x}_0) - f^* < \infty$, where $f^* \coloneqq f(\mathbf{x}^*)$.

(A2) *Average Smoothness condition.* We do not impose a uniform bound on gradient dissimilarity across workers. Instead, we assume that there exists $\ell > 0$ such that, for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$,

$$\frac{1}{n}\sum_{i=1}^{n} \|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\|^2 \le 2\ell\, \beta_f(\mathbf{x}, \mathbf{y}), \tag{17}$$

where $\beta_f(\mathbf{x}, \mathbf{y})$ denotes the Bregman divergence associated with f, defined as

$$\beta_f(\mathbf{x}, \mathbf{y}) \coloneqq f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle. \tag{18}$$

For example, if every f_i is convex and L_i-smooth, then (17) holds with $\ell \le L_{\max}$. This assumption controls changes in local gradients without requiring the local gradients themselves to be uniformly bounded or similar at the optimum.

(A3) *Stochastic gradients.* For each worker's local function f_i, the stochastic gradient oracle $g_i(\mathbf{x}, \xi_i)\colon \mathbb{R}^d \to \mathbb{R}^d$ is unbiased and has bounded variance, that is, for all $\mathbf{x} \in \mathbb{R}^d$,

$$\mathbb{E}_{\xi_i}[g_i(\mathbf{x}, \xi_i)] = \nabla f_i(\mathbf{x}). \tag{19}$$

$$\mathbb{E}_{\xi_i}\left[\|g_i(\mathbf{x}, \xi_i) - \nabla f_i(\mathbf{x})\|^2\right] \le \frac{\sigma^2}{b}. \tag{20}$$

Consequently, under independence across workers, the aggregate oracle $\mathbf{g}_t = \frac{1}{n}\sum_{i=1}^n \mathbf{g}_t^i$ is unbiased and has bounded variance of $\frac{\sigma^2}{nb}$.

The algorithm itself is written in terms of $\mathbf{g}_t^i$, so the deterministic method is obtained as a special case. See Appendix C for the probability space and filtration.

Here b denotes the mini-batch size, so the per-worker variance bound is $\frac{\sigma^2}{b}$. Communication-round comparisons use $b = 1$, the local-oracle-efficient setting in which each participating worker makes one stochastic-gradient call per round.

(A4) *Compressor model.* We consider the class of δ-contractive compressors. A (possibly) randomized compressor $\mathcal{C}\colon \mathbb{R}^d \to \mathbb{R}^d$ is called δ-contractive if, for every $\mathbf{u} \in \mathbb{R}^d$,

$$\mathbb{E}_{\mathcal{C}}\left[\|\mathcal{C}(\mathbf{u}) - \mathbf{u}\|^2\right] \le (1-\delta)\|\mathbf{u}\|^2, \qquad \delta \in (0, 1]. \tag{21}$$

A canonical example is Top-K, which retains the K coordinates with the largest absolute values and sets all remaining entries to zero (Stich et al., 2018). Fixed-rate quantizer (Lin et al., 2022) also satisfies the contraction inequality directly.

The main idea is to view the compressed, error-feedback gradient used by EC−NAG as an inexact stochastic gradient. The accumulated inexactness is controlled by a scalar error envelope, which is then substituted into the inexact-NAG recursion. This separates the proof into two parts: first, controlling the error-feedback/compression dynamics; second, choosing the accelerated parameters so that the compression-induced terms are absorbed by the Bregman terms in the Lyapunov inequality. Therefore, the proof of EC−NAG in Appendix D.2 is a specialization of the Lyapunov recursion for inexact NAG developed in Appendix D.1. We measure convergence in the convex stochastic setting by the expected function suboptimality $\mathbb{E}[f(\mathbf{y}_T)] - f^*$. The appendix proves the result for a general initial compression-error envelope. To state the main theorem in its cleanest form, one generally sets

$$E_0^2 \coloneqq \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\|\mathbf{q}_0^i - \mathbf{u}_0^i\|^2\right] = 0. \tag{22}$$

Thus, the theorem below is for an exactly initialized gradient tracker. The general $E_0 \neq 0$ estimate is (151) in Appendix D.2.

Theorem 1 (Iteration complexity of EC−NAG)**.** *Suppose* (A1)–(A4) *hold, let* $D > 0$ *be initial distance from the optimizer, and use exact tracker initialization. Choose* $0 < \gamma \leq \frac{\delta}{8}$, $c = \frac{10}{\gamma}$, *and for a prescribed horizon* $T \geq 1$, *set*

$$a_t = \frac{t+c}{M}, \qquad A_0 = \frac{c^2}{2M}, \tag{23}$$

$$M = \max\left\{32L + 731\frac{(1-\delta)\ell}{\gamma^2\delta^2}, \frac{\sigma T^{3/2}}{D\sqrt{b}}\sqrt{\frac{5}{n} + 357\frac{1-\delta}{\gamma^2\delta^2}}\right\}. \tag{24}$$

The,

$$T = \mathcal{O}\left(\frac{D}{\gamma\sqrt{\varepsilon}}\sqrt{L + \frac{(1-\delta)\ell}{\delta^2}} + \frac{\sigma^2 D^2}{b\varepsilon^2}\left(\frac{1}{n} + \frac{1-\delta}{\gamma^2\delta^2}\right)\right). \tag{25}$$

is a sufficient number of iterations to guarantee $\mathbb{E}\left[f(\mathbf{y}_T)\right] - f^* \leq \varepsilon$.

The theorem follows directly from the finite-horizon estimate (150) in the appendix after setting $E_0 = 0$. The two terms in (25) have different origins. The first is the accelerated deterministic contribution and decays as T^{-2}; the second contains stochastic-oracle noise and decays as $T^{-1/2}$.

When $\sigma = 0$, (25) reduces to

$$T = \mathcal{O}\left(\frac{D}{\gamma\sqrt{\varepsilon}}\sqrt{L + \frac{(1-\delta)\ell}{\delta^2}}\right), \tag{26}$$

equivalently $\mathbb{E}[f(\mathbf{y}_T)] - f^* = \mathcal{O}(T^{-2})$. Thus, the deterministic part retains the characteristic Nesterov $\varepsilon^{-1/2}$ dependence.

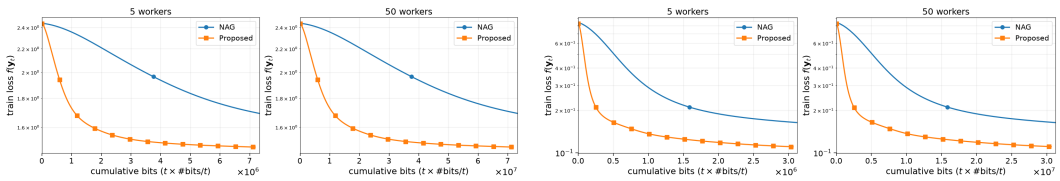
When $\sigma > 0$, the leading small-ε contribution is

$$\mathcal{O}\left(\frac{\sigma^2 D^2}{b\varepsilon^2}\left(\frac{1}{n} + \frac{1-\delta}{\gamma^2\delta^2}\right)\right). \tag{27}$$

The first term inside the parentheses is the ordinary $1/n$ reduction of independent stochastic-gradient noise. The second is generated by the stochastic controlled-error recursion and does not carry a $1/n$ factor. Accordingly, the theorem has the standard stochastic ε^{-2} exponent, but it does not establish optimal dependence on the number of workers or on the compression parameter. This is similar to ADEF.

The sufficient interval $0 < \gamma \leq \delta/8$ is used for simplification. The appendix derives the larger exact finite-c stability region before making this simplification. Within the explicit family used in Theorem 1, both γ-dependent contributions in (25) decrease as γ increases. Hence the largest permitted value, $\gamma = \frac{\delta}{8}$, gives the strongest specialization of this particular sufficient bound:

$$T = \mathcal{O}\left(\frac{D}{\delta\sqrt{\varepsilon}}\sqrt{L + \frac{(1-\delta)\ell}{\delta^2}} + \frac{\sigma^2 D^2}{b\varepsilon^2}\left(\frac{1}{n} + \frac{1-\delta}{\delta^4}\right)\right), \tag{28}$$



(a) `ILLc1850`: exact local gradients.

(b) `w8a`: exact local gradients.

Figure 1: Deterministic communication–loss comparison of EC−NAG and full-precision NAG. Panels correspond to $n = 5$ and $n = 50$ workers. Train loss $f(\mathbf{y}_t)$ versus cumulative uplink bits; quantizer with $z = d$ at R_c.

where universal numerical constants are absorbed into $\mathcal{O}(\cdot)$. We keep γ explicit in the main theorem because it is an algorithmic parameter and because the exact feasible region in the appendix is strictly larger than the convenient interval used for the uniform coefficient bounds.

The quantity M in (24) is a finite-horizon, analysis-optimal tuning parameter. It depends on the horizon and on problem quantities such as D and σ. The theorem should therefore be interpreted as an achievability statement rather than as an oracle-free parameter-selection rule. If only valid upper bounds on these quantities are known, they may be substituted in (24); increasing M decreases the effective accelerated stepsizes and preserves the sufficient stability inequalities established in the appendix.

We next instantiate the δ-contractive compressor using the classical fixed-rate quantizer from information theory. Let $\mathsf{q}\colon [-1,1]^d \to \mathbb{R}^d$ denote a fixed-rate quantizer of rate $R > 0$ (bits/dim) such that $|\operatorname{Im}(\mathsf{q})| = 2^{dR}$. At iteration t, worker i applies the scaled quantizer (Lin et al., 2022, Sec. II.A):

$$\mathcal{C}(\mathbf{u}_t^i) := r_t^i \, \mathsf{q}\left(\frac{\mathbf{u}_t^i}{r_t^i}\right), \tag{29}$$

where $r_t^i > 0$ is a positive *shrinkage factor* such that $\mathbf{u}_t^i / r_t^i \in \mathcal{B}(1)$. With the operational choice $r_t^i = \|\mathbf{u}_t^i\|$, the worker's quantizer is contractive with

$$\delta := 1 - \rho_d^2 2^{-2R}, \tag{30}$$

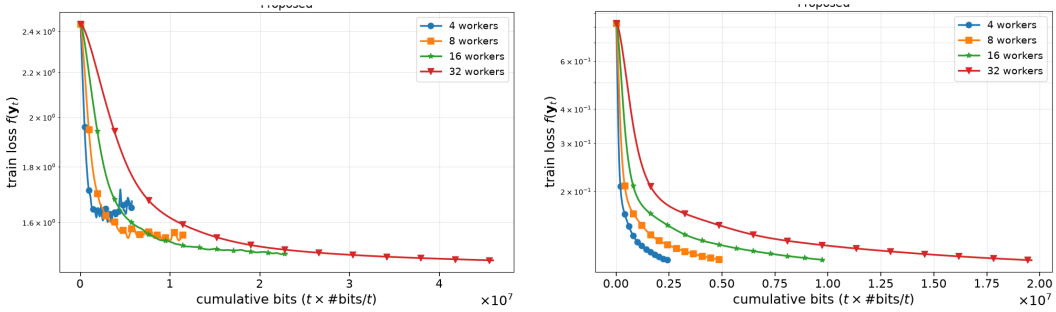
where ρ_d is the covering efficiency of the d-dimensional codebook. Smaller ρ_d corresponds to a more efficient covering of the unit ball. For example, standard coordinate-wise uniform quantization has $\rho_d = \sqrt{d}$. Then, the smallest integer rate R_c giving a strictly positive contraction parameter is $R_c = \left\lfloor \log_2(\sqrt{d}) \right\rfloor + 1$.

4 Experimental Results and Discussion

We evaluate EC−NAG on two convex problems: least squares on the `ILLc1850` matrix from the SuiteSparse collection, and unregularized binary logistic regression on `w8a` from LIBSVM. Both objectives are used without a regularizer. Feature rows are always normalized to unit ℓ_2 norm before partitioning. The global objective is the uniform average $f = \frac{1}{n}\sum_{i=1}^n f_i$, with shard losses scaled proportionally to their sample counts so that this average equals the global empirical risk. SuiteSparse shards are drawn at random; `w8a` shards are label-based and therefore heterogeneous.

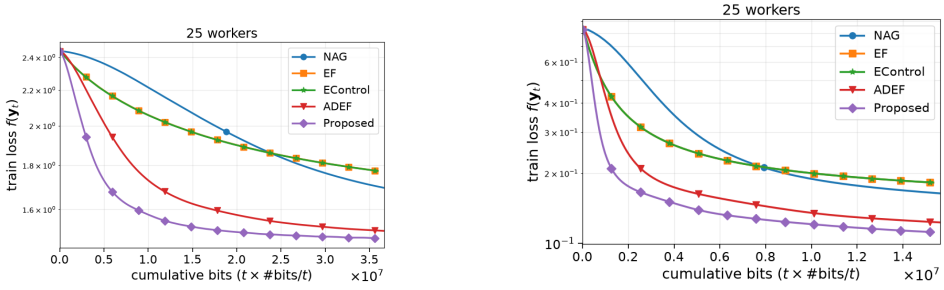
The plotted metric is the exact full-batch train loss $f(\mathbf{y}_t)$ at the gradient-step iterate. The horizontal axis is cumulative client→server uplink bits (summed over workers), excluding broadcast. With the fixed-rate uniform quantizer, a message of dimension d, and rate R costs $B = dR + 32$ bits and has contraction $\delta = 1 - d\,2^{-2R}$. All linear experiments run at the automatic critical rate $R = R_c = \lfloor \log_2 \sqrt{d} \rfloor + 1$. Full-precision NAG is charged $32d$ bits per worker per round; EF, EControl, and EC−NAG are charged one compressed message; ADEF is charged two. Step sizes and feedback gains are selected by nested tune-select (absolute powers of two on η and $\eta_\infty = 2/M$; for EControl/EC−NAG, γ is searched downward from δ). Complete details are in Section F.

Figure 1 isolates the cost of compression under exact local gradients. On both datasets, EC−NAG reaches a given train-loss level with substantially fewer uplink bits than full-precision NAG, while



(a) `ILLc1850`, batch size $b = 64$. (b) `w8a`, batch size $b = 256$.

Figure 2: Worker-count study for stochastic `EC-NAG` with $n \in \{4, 8, 16, 32\}$. Curves are single runs with seed 0; step sizes are re-tuned at each worker count.



(a) `ILLc1850`, $n = 25$, batch size $b = 64$. (b) `w8a`, $n = 25$, batch size $b = 256$.

Figure 3: Comparison of `NAG`, `EF`, `EControl`, `ADEF`, and `EC-NAG`. Train loss $f(\mathbf{y}_t)$ versus cumulative uplink bits; `ADEF` is charged two compressed uplinks per worker per round.

remaining competitive on a per-round basis. The bit advantage is expected: a message at R_c costs a fraction of $32d$ bits. These observations support an empirical communication advantage for the tested configurations; they should not be read as evidence that compression improves the iteration complexity of NAG.

Figure 2 varies the number of workers under stochastic gradients with a fixed per-worker batch size (effective batch $\min\{b, m_i\}$ on small shards). Increasing n generally improves late-stage behaviour, consistent with averaging additional worker gradients, though the ordering need not be monotone for every finite single-seed run. These experiments support practical benefits from worker averaging but do not establish uniform linear speedup in n, in line with the qualification following Theorem 1.

Figure 3 places all methods on the same bit axis at $n = 25$, including full-precision `NAG`. The unaccelerated compressed methods separate first: `EF` and `EControl` descend without the accelerated slope. The accelerated compressed methods are the closest comparison. Per round, `ADEF` and `EC-NAG` are of comparable quality; on the bit axis, `EC-NAG` benefits from transmitting only one compressed vector per round. Full-precision `NAG` is the least bit-efficient reference in both panels. Small late crossings in single-seed runs should not be over-interpreted; the trends we rely on are the ordering over the bulk of the budget. The same protocol on two further datasets appears in Section E.

5 Conclusion and Future Directions

We introduced `EC-NAG`, an acceleration-aligned error-feedback method for smooth distributed convex optimization with contractive worker-to-server compression. The central construction is a two-lag controlled-error correction whose coefficients are matched to the Nesterov weights. This makes the complete Nesterov-weighted history of gradient-estimation errors telescope to the current averaged error state, allowing the accelerated estimator to be maintained using a single compressed

worker-to-server message per iteration rather than the two or more compressed messages used by the closest same-setting accelerated method.

Our analysis shows that this communication reduction is compatible with Nesterov acceleration. Under smooth convexity, average smoothness, independent unbiased stochastic gradients, and contractive compression, EC−NAG retains the deterministic $\mathcal{O}(1/T^2)$ function-value rate. In the stochastic setting, the bound has the standard $\mathcal{O}(1/\sqrt{T})$ decay, and hence the ε^{-2} stochastic dependence on target accuracy. Moreover, we substantially narrow the gap between the conservative analytical parameter choices, for instance, $\gamma = \frac{\delta}{8}$ suggested here versus $\gamma = \frac{\delta}{400}$ in (Gao et al., 2024).

Experiments on distributed least-squares and logistic-regression problems support the communication motivation of the method. Across the tested configurations, EC−NAG reaches comparable train-loss levels using fewer transmitted uplink bits than full-precision NAG, and its single-uplink structure provides a direct communication advantage over ADEF when the two methods have comparable per-round behavior.

Several questions remain open. Of particular interest are sharper control of the stochastic feedback residual so that compression noise also benefits from worker averaging, eliminating or compressing the exact tracker initialization used in the cleanest theorem statement, and further reducing the gap between analytical parameter choices and the empirically effective ones. Extensions to partial participation and adaptive communication rates are also natural directions for future work.

REFERENCES

Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. QSGD: Communication-Efficient SGD via Gradient Quantization and Encoding. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., December 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/6c340f25839e6acdc73414517203f5f0-Paper.pdf.

Dan Alistarh, Torsten Hoefler, Mikael Johansson, Sarit Khirirat, Nikola Konstantinov, and Cédric Renggli. The convergence of sparsified gradient methods. In *Advances in Neural Information Processing Systems*, NIPS'18, pp. 5977–5987, Red Hook, NY, USA, December 2018. Curran Associates Inc. URL https://proceedings.neurips.cc/paper/2018/hash/314450613369e0ee72d0da7f6fee773c-Abstract.html.

Amir Beck and Marc Teboulle. A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, March 2009. URL https://doi.org/10.1137/080716542.

Dimitri P. Bertsekas and John N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, 1st edition, January 1996. ISBN 1886529108. URL http://www.athenasc.com/ndpbook.html.

Aleksandr Beznosikov, Samuel Horváth, Peter Richtárik, and Mher Safaryan. On biased compression for distributed learning. *Journal of Machine Learning Research*, 24(276):1–50, January 2023. ISSN 1532-4435. URL https://www.jmlr.org/papers/v24/21-1548.html.

Sébastien Bubeck. Convex Optimization: Algorithms and Complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–357, November 2015. ISSN 1935-8237. doi: 10.1561/2200000050. URL https://doi.org/10.1561/2200000050.

Dmitry Bylinkin and Aleksandr Beznosikov. Accelerated Methods with Compressed Communications for Distributed Optimization Problems Under Data Similarity. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 15632–15640, April 2025. doi: 10.1609/aaai.v39i15.33716. URL https://ojs.aaai.org/index.php/AAAI/article/view/33716.

Laurent Condat, Kai Yi, and Peter Richtárik. EF-BV: A unified theory of error feedback and variance reduction mechanisms for biased and unbiased compression in distributed optimization. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi: 10.52202/068431-1272.

Ilyas Fatkhullin, Alexander Tyurin, and Peter Richtárik. Momentum Provably Improves Error Feedback!, May 2023. URL https://arxiv.org/abs/2305.15155.

Michael P. Friedlander and Mark Schmidt. Hybrid Deterministic-Stochastic Methods for Data Fitting. *SIAM Journal on Scientific Computing*, 34(3):A1380–A1405, May 2012. doi: 10.1137/110830629. URL https://doi.org/10.1137/110830629.

Yuan Gao, Rustem Islamov, and Sebastian U. Stich. EControl: Fast Distributed Optimization with Compression and Error Control. In *International Conference on Learning Representations*, May 2024. URL https://openreview.net/forum?id=lsvlvWB9vz.

Yuan Gao, Anton Rodomanov, Jeremy Rack, and Sebastian U Stich. Accelerated Distributed Optimization with Compression and Error Feedback. In *The 29th International Conference on Artificial Intelligence and Statistics*, May 2026. URL https://openreview.net/forum?id=JdX5bJOJLC.

Eduard Gorbunov, Konstantin P. Burlachenko, Zhize Li, and Peter Richtarik. MARINA: Faster Non-Convex Distributed Learning with Compression. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 3788–3798. PMLR, July 2021. URL https://proceedings.mlr.press/v139/gorbunov21a.html.

Yutong He, Xinmeng Huang, Yiming Chen, Wotao Yin, and Kun Yuan. Lower Bounds and Accelerated Algorithms in Distributed Stochastic Optimization with Communication Compression, May 2023. URL https://arxiv.org/abs/2305.07612.

Xinmeng Huang, Yiming Chen, Wotao Yin, and Kun Yuan. Lower Bounds and Nearly Optimal Algorithms in Distributed Learning with Communication Compression. In *Advances in Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, November 2022. Curran Associates Inc. ISBN 9781713871088. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/77f2d0c271e508278ea13e24cd8773d5-Abstract-Conference.html.

Ziheng Jiang, Haibin Lin, Yinmin Zhong, Qi Huang, Yangrui Chen, Zhi Zhang, Yanghua Peng, Xiang Li, Cong Xie, Shibiao Nong, Yulu Jia, Sun He, Hongmin Chen, Zhihao Bai, Qi Hou, Shipeng Yan, Ding Zhou, Yiyao Sheng, Zhuo Jiang, Haohan Xu, Haoran Wei, Zhang Zhang, Pengfei Nie, Leqi Zou, Sida Zhao, Liang Xiang, Zherui Liu, Zhe Li, Xiaoying Jia, Jianxi Ye, Xin Jin, and Xin Liu. MegaScale: Scaling Large Language Model Training to more than 10,000 GPUs. In *21st USENIX Symposium on Networked Systems Design and Implementation*, NSDI'24, pp. 745–760, USA, April 2024. USENIX Association. ISBN 978-1-939133-39-7. URL https://www.usenix.org/conference/nsdi24/presentation/jiang-ziheng.

Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista A. Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaïd Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konečný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Mariana Raykova, Hang Qi, Daniel Ramage, Ramesh Raskar, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. Advances and Open Problems in Federated Learning. *Foundations and Trends in Machine Learning*, 14 (1-2):1–210, June 2021. doi: 10.1561/2200000083. URL https://doi.org/10.1561/2200000083.

Sai Praneeth Karimireddy, Quentin Rebjock, Sebastian Stich, and Martin Jaggi. Error Feedback Fixes SignSGD and other Gradient Compression Schemes. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3252–3261. PMLR, June 2019. URL https://proceedings.mlr.press/v97/karimireddy19a.html.

Sarit Khirirat, Hamid Reza Feyzmahdavian, and Mikael Johansson. Distributed learning with compressed gradients. arXiv:1806.06573, November 2018. URL https://arxiv.org/abs/1806.06573.

Zhize Li and Peter Richtárik. A Unified Analysis of Stochastic Gradient Methods for Nonconvex Federated Optimization. arxiv:2006.07013, June 2020. URL https://arxiv.org/abs/2006.07013.

Zhize Li, Dmitry Kovalev, Xun Qian, and Peter Richtarik. Acceleration for Compressed Gradient Descent in Distributed and Federated Optimization. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *ICML'20*, pp. 5895–5904. PMLR, July 2020. URL https://proceedings.mlr.press/v119/li20g.html.

Chung-Yi Lin, Victoria Kostina, and Babak Hassibi. Differentially Quantized Gradient Methods. *IEEE Transactions on Information Theory*, 68(9):6078–6097, September 2022. doi: 10.1109/TIT.2022.3171173. URL https://doi.org/10.1109/TIT.2022.3171173.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In Aarti Singh and Jerry Zhu (eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1273–1282. PMLR, April 2017. URL https://proceedings.mlr.press/v54/mcmahan17a.html.

K. Mishchenko, E. Gorbunov, M. Takáč, and P. Richtárik. Distributed learning with compressed gradient differences. *Optimization Methods and Software*, 40(5):1181–1196, October 2025. doi: 10.1080/10556788.2024.2358790. URL https://doi.org/10.1080/10556788.2024.2358790. (arXiv preprint circulated in 2019 and 2023).

Y. Nesterov. A method for solving the convex programming problem with convergence rate $O(1/k^2)$. *Doklady Akademii Nauk SSSR*, 269:543–547, January 1983. URL https://www.mathnet.ru/eng/dan46009.

Yurii Nesterov. *Lectures on Convex Optimization*, volume 137 of *Springer Optimization and Its Applications*. Springer, November 2018. doi: 10.1007/978-3-319-91578-4. URL https://doi.org/10.1007/978-3-319-91578-4.

Boris T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, September 1964. doi: 10.1016/0041-5553(64)90137-5. URL https://doi.org/10.1016/0041-5553(64)90137-5.

Xun Qian, Peter Richtarik, and Tong Zhang. Error Compensated Distributed SGD Can Be Accelerated. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 30401–30413. Curran Associates, Inc., December 2021. URL https://proceedings.neurips.cc/paper/2021/hash/ff1ced3097ccf17c1e67506cdad9ac95-Abstract.html.

Peter Richtárik, Igor Sokolov, and Ilyas Fatkhullin. EF21: A New, Simpler, Theoretically Better, and Practically Faster Error Feedback. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34 of *NIPS '21*, pp. 4384–4396, Red Hook, NY, USA, December 2021. Curran Associates Inc. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/231141b34c82aa95e48810a9d1b33a79-Abstract.html.

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 *Interspeech*, pp. 1058–1062, September 2014. doi: 10.21437/Interspeech.2014-274. URL https://doi.org/10.21437/Interspeech.2014-274.

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:1909.08053, September 2019. URL https://arxiv.org/abs/1909.08053.

Sebastian U. Stich, Jean-Baptiste Cordonnier, and Martin Jaggi. Sparsified SGD with memory. In *Advances in Neural Information Processing Systems*, NIPS'18, pp. 4452–4463, Red Hook, NY, USA, December 2018. Curran Associates Inc. URL https://dl.acm.org/doi/10.5555/3327345.3327357.

Weijie Su, Stephen Boyd, and Emmanuel J. Candès. A Differential Equation for Modeling Nesterov's Accelerated Gradient Method: Theory and Insights. *Journal of Machine Learning Research*, 17(153):1–43, September 2016. URL https://jmlr.org/papers/v17/15-084.html.

Alexander Tyurin and Peter Richtárik. 2Direction: Theoretically faster distributed training with bidirectional communication compression. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-0516.

Thijs Vogels, Sai Praneeth Karimireddy, and Martin Jaggi. PowerSGD: Practical Low-Rank Gradient Compression for Distributed Optimization. In *Advances in Neural Information Processing Systems*, pp. 10159–10170, December 2019. URL https://proceedings.neurips.cc/paper/2019/hash/d9fbed9da256e344c1fa46bb46c34c5f-Abstract.html.

A Technical Summary of Classical NAG

Nesterov's accelerated gradient method (NAG) (Nesterov, 1983) is a momentum-based refinement of gradient descent that achieves faster convergence rates by combining a gradient step with an extrapolation step. Unlike classical heavy-ball momentum (Polyak, 1964), Nesterov-type momentum is often interpreted as a *look-ahead* mechanism: the method forms an extrapolated point and then incorporates gradient information in a way that anticipates the direction in which the iterates are moving. This structure improves stability and helps reduce overshooting near the optimum.

The most common is a two-sequence formulation of NAG, maintaining iterates $(\mathbf{x}_t, \mathbf{y}_t)$. The update (31) performs a gradient step from the extrapolated point $\mathbf{x}_t$ to produce the proximal point $\mathbf{y}_{t+1}$, while (32) applies momentum by extrapolating along the most recent displacement $\mathbf{y}_{t+1} - \mathbf{y}_t$. This two-sequence representation is algebraically equivalent, after a change of variables, to the three-point formulation as shown in Table 1; the corresponding main-text accelerated updates are (5)–(6).

Table 1: Formulations of Nesterov's accelerated gradient method (NAG). The left panel shows the two-point look-ahead form, while the right panel shows the three-point coupled-sequence form. The two schemes are algebraically equivalent for the choices $\eta_t = a_{t+1}^2/A_{t+1}$ and $\beta_t = a_{t+2}A_t/(a_{t+1}A_{t+2})$.

Two-point form (look-ahead)

Given $\mathbf{x}_0, \mathbf{y}_0 \in \mathbb{R}^d$, step sizes $\eta_t > 0$, and momentum parameters $\beta_t \geq 0$.
For $t = 0, 1, 2, \ldots$:

$$\mathbf{y}_{t+1} = \mathbf{x}_t - \eta_t \nabla f(\mathbf{x}_t), \tag{31}$$

$$\mathbf{x}_{t+1} = \mathbf{y}_{t+1} + \beta_t\big(\mathbf{y}_{t+1} - \mathbf{y}_t\big). \tag{32}$$

The point $\mathbf{y}_{t+1}$ is obtained by a descent step from $\mathbf{x}_t$, while $\mathbf{x}_{t+1}$ extrapolates along the latest displacement, producing the accelerated dynamics.

Three-point form (coupled sequences)

Given $\mathbf{v}_0, \mathbf{y}_0 \in \mathbb{R}^d$, positive weights $\{a_t\}_{t\geq 1}$, and $A_0 > 0$, define $A_{t+1} = A_t + a_{t+1}$.
Initialize $\mathbf{x}_0 = \frac{A_0}{A_1}\mathbf{y}_0 + \frac{a_1}{A_1}\mathbf{v}_0$.
For $t = 0, 1, 2, \ldots$:

$$\mathbf{v}_{t+1} = \mathbf{v}_t - a_{t+1}\nabla f(\mathbf{x}_t), \tag{33}$$

$$\mathbf{y}_{t+1} = \frac{A_t}{A_{t+1}}\mathbf{y}_t + \frac{a_{t+1}}{A_{t+1}}\mathbf{v}_{t+1}, \tag{34}$$

$$\mathbf{x}_{t+1} = \frac{A_{t+1}}{A_{t+2}}\mathbf{y}_{t+1} + \frac{a_{t+2}}{A_{t+2}}\mathbf{v}_{t+1}. \tag{35}$$

The sequence $\mathbf{v}_{t+1}$ accumulates past gradients, while $\mathbf{y}_{t+1}$ and $\mathbf{x}_{t+1}$ are specific convex combinations.

For an L-smooth and μ-strongly convex function f, with condition number $\kappa = L/\mu$, the classical constant-parameter choice is

$$\eta_t \equiv \frac{1}{L}, \qquad \beta_t \equiv \beta_{\mathrm{NAG}} = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}. \tag{36}$$

With this choice, NAG achieves the accelerated linear rate

$$f(\mathbf{y}_T) - f^\star = \mathcal{O}\left(\left(1 - \frac{1}{\sqrt{\kappa}}\right)^T\right), \tag{37}$$

In ill-conditioned problems,

$$\sigma_{\mathrm{NAG}} = \sqrt{1 - \frac{1}{\sqrt{\kappa}}} = 1 - \frac{1}{2\sqrt{\kappa}} + \mathcal{O}\left(\frac{1}{\kappa}\right) \tag{38}$$

improves substantially over the GD contraction factor

$$\sigma_{\mathrm{GD}} = \frac{\kappa - 1}{\kappa + 1} = 1 - \frac{1}{\kappa} + \mathcal{O}\left(\frac{1}{\kappa^2}\right). \tag{39}$$

The improvement is optimal in its dependence on κ: the optimal first-order contraction factor has the expansion

$$\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} = 1 - \frac{2}{\sqrt{\kappa}} + \mathcal{O}\left(\frac{1}{\kappa}\right). \tag{40}$$

See, for example, (Nesterov, 2018, Th. 2.1.13).

For an L-smooth convex function f, one typically uses the Nesterov/FISTA sequence

$$\theta_0 = 1, \qquad \theta_{t+1} = \frac{1+\sqrt{1+4\theta_t^2}}{2}, \qquad \beta_t = \frac{\theta_t - 1}{\theta_{t+1}}, \qquad \eta_t \equiv \frac{1}{L}. \tag{41}$$

Then NAG satisfies the optimal accelerated convex rate (Nesterov, 1983; Beck & Teboulle, 2009; Bubeck, 2015):

$$f(\mathbf{y}_T) - f(\mathbf{x}^*) \leq \frac{2L\|\mathbf{x}_0 - \mathbf{x}^*\|^2}{(T+1)^2}, \tag{42}$$

Equivalently, in the three-point coupled-sequence formulation, choosing the weights so that $a_t^2 = A_t/L$ is used, then

$$A_{T-1} \geq \frac{(T+1)^2}{4L}, \tag{43}$$

and (42) follows.

The preceding discussion on centralized NAG is algebraically identical to the distributed setting without compression. Each worker sends the gradient-tracking increment in (3), and the server averages them according to (4) to recover the exact global gradient direction for the objective in (1). Therefore, the classical convergence guarantees stated above carry over directly to the distributed, uncompressed implementation.

B Essential Appendix: Traits of a "Good" Algorithm

B.1 Why the compressor input contains gradient tracking?

To see the issue with using raw local gradients, suppose that the global iterate reaches the global minimizer $\mathbf{x}^*$. Even then, a worker that feeds $\nabla f_i(\mathbf{x}^*)$ into the quantizer will generally continue to produce a nonzero signal, because $\nabla f_i(\mathbf{x}^*)$ need not vanish even though $\nabla f(\mathbf{x}^*) = 0$. In the presence of local error memory, the worker-side quantizer input can be expressed in the form

$$\mathbf{u}_t^i \in \nabla f_i(\mathbf{x}^*) + \operatorname{span}\{\mathbf{e}_0^i, \ldots, \mathbf{e}_{t-1}^i\},$$

so unless the scheme can coordinate across workers to cancel the constant bias $\nabla f_i(\mathbf{x}^*)$, the local inputs need not decay to zero. In practice, this manifests as persistent worker-side activity and a tendency for each worker to pull toward its own local minimizer, even when the global objective has already been optimized.

This obstruction appears more vividly in theory. Consider the multi-worker NQGD scheme (naively quantized GD) as stated in (Lin et al., 2022, Algo. 4). Linear convergence to $\mathbf{x}^*$ in the strongly convex setting is tied to driving both the iterate error and the quantizer inputs to zero; in particular, one requires

$$\|\mathbf{x}_t - \mathbf{x}^*\| \to 0 \qquad \text{and} \qquad \|\mathbf{u}_t^i\| \leq \|\nabla f_i(\mathbf{x}_t)\| + \|\mathbf{e}_t^i\| \to 0 \quad \forall\, i.$$

Since the second condition forces $\|\nabla f_i(\mathbf{x}_t)\| \to 0$ for every i, and each f_i is strongly convex with a unique minimizer $\mathbf{x}_{f_i}^*$, we obtain $\mathbf{x}_t \to \mathbf{x}_{f_i}^*$ for each i. Compatibility with $\mathbf{x}_t \to \mathbf{x}^*$ is therefore possible only if all local minimizers coincide:

$$\mathbf{x}^* = \mathbf{x}_{f_i}^* \quad \forall\, i = 1, \ldots, n. \tag{44}$$

This is precisely the *interpolation* (shared-minimizer) assumption imposed in the multi-worker case (Lin et al., 2022; Khirirat et al., 2018). The same logic applies to any method whose quantizer input is based solely on raw local gradients.

Under *strong convexity* plus the *interpolation setting*, the analysis is tractable and linear convergence holds. However, these assumptions are extremely restrictive in practice as

- strong convexity for every f_i is uncommon for modern models and datasets, and
- since all f_i share exactly the same minimizer then in principle the server could optimize a *single* function and ignore the others without changing the solution, rendering all but one worker essentially redundant.

However, alleviating assumption (44), even in the strongly convex case, the best one can prove is convergence to a ball around the minimizer (Khirirat et al., 2018, Th. 5.2), (Lin et al., 2022, Th. D.2); convergence all the way to $\mathbf{x}^*$ is no longer guaranteed. Thus, for multi-worker problems with heterogeneous objectives, any scheme that quantizes the local gradient is effectively forced into some version of the interpolation assumption to prove linear convergence.

In non-convex settings and SGD, it is standard to impose a gradient-dissimilarity bound of the form (see, for instance, (Alistarh et al., 2017, Def. 2.1), (Gorbunov et al., 2021, Eq. (10)), (Mishchenko et al., 2025, Eq. (18))):

$$\frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(\mathbf{x})-\nabla f(\mathbf{x})\|^2 \le B^2 \tag{45}$$

or a bounded-gradient condition ($\|\nabla f_i(\mathbf{x})\| \le B_1$) or more general forms (see, for instance, (Bertsekas & Tsitsiklis, 1996, Sec. 4.2), (Friedlander & Schmidt, 2012; Li & Richtárik, 2020)) such as:

$$\|\nabla f_i(x)\|^2 \le B_1 + B_2\|\nabla f(x)\|^2 \quad \forall\, i, \tag{46}$$

for some constants $B, B_1, B_2 \ge 0$. In classical GD, with the goal of *geometric decay* of the convergence error, a fixed additive constant (like $B > 0$) in the state recursion is fatal: if the recursion for the state (or for $\|\mathbf{u}_t^i\|$) contains a nonvanishing additive term, the best one can hope for is convergence to a neighborhood of $\mathbf{x}^*$ whose size depends on B, rather than true convergence to the minimizer. In other words, if $\|\mathbf{u}_t^i\|$ does not vanish, then neither can the convergence error.

This motivates the upgrade from raw gradient to tracking-based messages: by compressing difference-of-gradients evaluated at consecutive iterates, we can ensure that the global descent direction converges to zero without requiring each local gradient $\nabla f_i(\cdot)$ to vanish individually. This decouples the local stationary points $\mathbf{x}_{f_i}^*$ from the global minimizer $\mathbf{x}^*$, which allows us to handle heterogeneous workers without resorting to extraneous assumptions.

B.2 Development and Justification of the Proposed Scheme

We now explain why the particular error-feedback mechanism used in EC−NAG is needed for accelerated methods. The main issue is that, in Nesterov's accelerated gradient method, gradient errors are not merely used in the current gradient step; they are also accumulated in a momentum-like sequence with time-varying weights. Consequently, an error-feedback rule that is sufficient for gradient descent need not be adequate for acceleration.

We begin with a simple uncompensated gradient-tracking scheme based on the main-text tracker (3). Suppose each worker i transmits a compressed tracking message of the local gradient difference, $\mathbf{q}_t^i = \mathcal{C}(\mathbf{g}_t^i - \mathbf{g}_{t-1}^i)$, without any error compensation. Defining the worker's local compression error at iteration t as $\mathbf{e}_t^i \coloneqq \mathbf{q}_t^i - (\mathbf{g}_t^i - \mathbf{g}_{t-1}^i)$, the server's aggregated gradient estimator unrolls to:

$$\widehat{\mathbf{g}}_t = \widehat{\mathbf{g}}_{t-1} + \frac{1}{n}\sum_{i=1}^{n}\mathbf{q}_t^i = \mathbf{g}_t + \sum_{\tau=0}^{t}\mathbf{e}_\tau, \tag{47}$$

where $\mathbf{e}_t \coloneqq \frac{1}{n}\sum_{i=1}^{n}\mathbf{e}_t^i$ is the average compression error. Under this uncompensated formulation, the server-side estimator is corrupted by the *cumulative historical sum* of all past compression errors. This accumulated error can be highly unstable in optimization and can lead to divergence.

The EF21 mechanism (Richtárik et al., 2021) eliminates this pathology for GD by subtracting the previous local error from the tracking message:

$$\mathbf{u}_t^i = (\mathbf{g}_t^i - \mathbf{g}_{t-1}^i) - \mathbf{e}_{t-1}^i, \tag{48}$$

which yields the clean server-side update:

$$\widehat{\mathbf{g}}_t = \mathbf{g}_t + \mathbf{e}_t. \tag{49}$$

Hence, the gradient step $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t\widehat{\mathbf{g}}_t$ is perturbed only by the immediate error $\mathbf{e}_t$, rather than by the accumulated history.

However, for Nesterov's accelerated gradient method, the standard EF21 correction is no longer adequate. In the three-point form, the gradient estimator is scaled by a time-varying weight and

accumulated into a momentum variable. In particular, the accelerated update contains a recursion of the form:

$$\mathbf{v}_{t+1} = \mathbf{v}_t - a_{t+1}\widehat{\mathbf{g}}_t, \tag{50}$$

where $a_{t+1} > 0$ is the acceleration weight. If one naively combines this accelerated recursion with the standard EF21 relation (49), then

$$\mathbf{v}_{t+1} = \mathbf{v}_0 - \sum_{\tau=0}^{t} a_{\tau+1}\widehat{\mathbf{g}}_\tau = \mathbf{v}_0 - \sum_{\tau=0}^{t} a_{\tau+1}\mathbf{g}_\tau - \sum_{\tau=0}^{t} a_{\tau+1}\mathbf{e}_\tau. \tag{51}$$

The last term is a weighted accumulation of all past compression errors. Since the acceleration weights a_t typically grow over time, for example linearly in the classical accelerated regime, this weighted error history can grow rapidly. This uncompensated growth of weighted historical error acts as a persistent drift in the momentum trajectory, destroying the acceleration property and causing optimization to fail.

Therefore, to preserve acceleration under compression, the error-feedback mechanism must be aligned with the weighted momentum accumulation. The goal is to force the weighted cumulative discrepancy between the surrogate gradients and the true gradients to collapse to a stable residual depending only on the current compression state. Specifically, matching the main-text identity (13), we design the estimator so that

$$\frac{1}{a_{t+1}} \sum_{\tau=0}^{t} a_{\tau+1}\big(\widehat{\mathbf{g}}_\tau - \mathbf{g}_\tau\big) = \mathbf{e}_t. \tag{52}$$

This is satisfied by enforcing the surrogate-gradient relation

$$\widehat{\mathbf{g}}_t - \mathbf{g}_t = \mathbf{e}_t - \frac{a_t}{a_{t+1}}\mathbf{e}_{t-1}. \tag{53}$$

To realize (53) through local compressed messages, each worker constructs the two-term error-corrected tracking-based input $\mathbf{u}_t^i$ in (7). Because the controlled state satisfies (8), the compressed message can be written as

$$\begin{aligned}\mathbf{q}_t^i &= \mathbf{u}_t^i + \mathbf{e}_t^i - (1-\gamma)\mathbf{e}_{t-1}^i \\ &= \left(\mathbf{g}_t^i + \mathbf{e}_t^i - \frac{a_t}{a_{t+1}}\mathbf{e}_{t-1}^i\right) - \left(\mathbf{g}_{t-1}^i + \mathbf{e}_{t-1}^i - \frac{a_{t-1}}{a_t}\mathbf{e}_{t-2}^i\right).\end{aligned} \tag{54}$$

The first equality shows exactly why the coefficient γ also appears in (7): the two contributions involving $\mathbf{e}_{t-1}^i$ combine as $-(1-\gamma) - (\gamma + a_t/a_{t+1}) = -(1 + a_t/a_{t+1})$. Therefore, when the server averages these messages and updates the global estimator, the intermediate terms perfectly telescope. Thus, the weighted history of gradient-estimation errors is represented exactly by the current averaged state $\mathbf{e}_t$, as formalized in (53) and in the main-text identities (12) and (13). Since the remaining correction coefficients are matched to the momentum ratio a_t/a_{t+1}, the feedback mechanism is precisely aligned with the NAG dynamics.

The gain γ is needed to separate two roles that coincide in ordinary error feedback: preserving the exact accelerated telescoping identity and controlling the internal error state. At $\gamma = 1$, $\mathbf{e}_t^i = \mathbf{q}_t^i - \mathbf{u}_t^i$ and the method reduces to the undamped two-residual rule. For $0 < \gamma < 1$, the update retains the fraction $1-\gamma$ of the previous state, while the coefficient multiplying $\mathbf{e}_{t-1}^i$ in the compressor input decreases by the same amount. Consequently, the message identity (54) and hence the NAG trajectory seen by the server keep exactly the required telescoping form, while the sequence presented to the compressor changes. This is a genuine stability–responsiveness tradeoff, not a free improvement.

This control-gain viewpoint is related to EControl (Gao et al., 2024), which also regulates the strength of an error-compensation signal instead of always feeding back the full stored error. The mechanisms are nevertheless different. EC-NAG places γ inside the two-lag correction dictated by the Nesterov's coefficients. The defining property here is the exact weighted identity (52); neither the two-lag coefficients nor this acceleration-weighted telescoping relation follows from EControl. Thus, γ just imports the useful idea, while the surrounding algebra is specific to NAG.

The space of possible local error-feedback corrections is broad. One could, for example, construct schemes with longer historical windows or use different linear combinations of previous compression residuals. Our two-step correction in EC-NAG is most appropriate given it satisfies the following guiding principles.

- **Minimal memory footprint.** The correction uses only the two most recent local controlled states, $\mathbf{e}_{t-1}^i$ and $\mathbf{e}_{t-2}^i$. Thus, keeping the local memory overhead low.
- **Low effective state dimension.** The induced error-propagation dynamics are low-dimensional. The parameter γ controls the one-step persistence of this state and makes the required stability margin explicit. This keeps the convergence analysis tractable and yields an explicit non-asymptotic convergence bound.
- **Exact alignment with acceleration.** The time-varying step size η_t and momentum parameter β_t are matched to the sequences $\{a_t, A_t\}_{t\geq 1}$. Under contractive compression, EC−NAG behaves like Nesterov's accelerated gradient method driven by a controlled and stable perturbation of the true gradient, rather than by an accumulating weighted history of compression errors.

C Probability Space, Filtration, and Measurability of the Algorithmic State

This section rigorously specifies the probability space, filtration, and all sources of randomness (the per-worker stochastic gradient oracles and the per-worker heterogeneous contractive compressors) underlying the distributed algorithm and its analysis.

The **sample spaces for each randomness source** is given below:

- *Stochastic gradient oracle randomness.* For each worker $i \in \{1, \dots, n\}$, let $(\Xi_i, \mathcal{A}_i, \mu_i)$ be a measurable space equipped with a probability measure μ_i, representing the randomness of a single stochastic-gradient query at worker i (e.g., a minibatch index or data sample). Define the per-iteration, all-workers gradient sample space
$$(\Xi, \mathcal{A}, \mu) \coloneqq \left(\prod_{i=1}^n \Xi_i,\ \bigotimes_{i=1}^n \mathcal{A}_i,\ \bigotimes_{i=1}^n \mu_i \right). \tag{55}$$
An element $\xi = (\xi^1, \dots, \xi^n) \in \Xi$ collects, for one iteration, the independent gradient-sampling randomness across all n workers.
- *Compressor randomness.* For each worker $i \in \{1, \dots, n\}$, the compressor $\mathcal{C}_i$ may itself be randomized (e.g., random sparsification) with its own contraction parameter $\delta_i \in (0, 1]$. Let $(\mathrm{H}_i, \mathcal{B}_i, \nu_i)$ be the measurable space carrying the internal randomness of $\mathcal{C}_i$, independent across i and independent of all gradient-sampling randomness. Define the per-iteration, all-workers compressor sample space
$$(\mathrm{H}, \mathcal{B}, \nu) \coloneqq \left(\prod_{i=1}^n \mathrm{H}_i,\ \bigotimes_{i=1}^n \mathcal{B}_i,\ \bigotimes_{i=1}^n \nu_i \right). \tag{56}$$
An element $h = (h^1, \dots, h^n) \in \mathrm{H}$ collects the compressor randomization used by all workers at one iteration; we write $\mathcal{C}_i(\mathbf{u}; h^i)$ for the realized (measurable) ompression map applied to $\mathbf{u} \in \mathbb{R}^d$ using randomness h^i.

Consequently, the combined per-iteration randomness is defined as
$$(\Theta, \mathcal{L}, \theta) \coloneqq (\Xi \times \mathrm{H},\ \mathcal{A} \otimes \mathcal{B},\ \mu \otimes \nu), \tag{57}$$
so that a generic element $\theta = (\xi, h) \in \Theta$ represents *all* randomness, gradient sampling and compression across all n workers, consumed during a single iteration.

Then, the **global probability space** is defined as the infinite product over all iterations $t = 0, 1, 2, \dots$:
$$(\Omega, \mathcal{F}, \mathbb{P}) \coloneqq \left(\Theta^{\mathbb{N}_0},\ \mathcal{L}^{\otimes \mathbb{N}_0},\ \theta^{\otimes \mathbb{N}_0}\right). \tag{58}$$
An element $\omega \in \Omega$ is written
$$\omega = \big(\theta_0(\omega), \theta_1(\omega), \dots\big), \quad \theta_t(\omega) = \big(\xi_t(\omega), h_t(\omega)\big) = \Big(\big(\xi_t^1(\omega), \dots, \xi_t^n(\omega)\big),\ \big(h_t^1(\omega), \dots, h_t^n(\omega)\big)\Big), \tag{59}$$
where the coordinate maps $\theta_t \colon \Omega \to \Theta$, $t \geq 0$, are i.i.d. with law $\theta = \mu \otimes \nu$. By construction of the product measure:

- $\xi_t^1, \dots, \xi_t^n$ are mutually independent for each fixed t (independence *across workers*);
- $h_t^1, \dots, h_t^n$ are mutually independent for each fixed t, and independent of $(\xi_t^1, \dots, \xi_t^n)$ (independence of compressor noise from gradient noise);
- $\theta_0, \theta_1, \dots$ are mutually independent across iterations (independence *across time*).

This brings us to the **filtration**. For $t \geq 0$, define

$$\mathcal{F}_t \coloneqq \sigma(\theta_0, \dots, \theta_t) = \sigma\big(\xi_0^1, \dots, \xi_0^n, h_0^1, \dots, h_0^n, \dots, \xi_t^1, \dots, \xi_t^n, h_t^1, \dots, h_t^n\big), \tag{60}$$

and set $\mathcal{F}_{-1} \coloneqq \{\varnothing, \Omega\}$. Then $(\mathcal{F}_t)_{t \geq -1}$ is an increasing sequence of sub-σ-algebras of $\mathcal{F}$ (a filtration), and by independence of the θ_t's,

$$\theta_t = (\xi_t, h_t) \text{ is independent of } \mathcal{F}_{t-1}, \qquad \forall\, t \geq 0. \tag{61}$$

For an integrable random vector $\mathbf{Z} : \Omega \to \mathbb{R}^m$ and a sub-σ-algebra $\mathcal{G} \subseteq \mathcal{F}$, the conditional expectation $\mathbb{E}[\mathbf{Z} \mid \mathcal{G}]$ is defined, as usual, as the (a.s. unique) $\mathcal{G}$-measurable random vector satisfying

$$\int_A \mathbb{E}[\mathbf{Z} \mid \mathcal{G}]\, d\mathbb{P} = \int_A \mathbf{Z}\, d\mathbb{P}, \qquad \forall\, A \in \mathcal{G}. \tag{62}$$

All expectations $\mathbb{E}[\cdot]$ appearing below are with respect to $\mathbb{P}$, i.e., taken jointly over *all* gradient-sampling and compressor randomness of *all* workers and *all* iterations.

With everything defined above, the per-worker gradient oracle and compressor are given below.

- *Stochastic gradient oracles (per worker).*
  Let $g_i : \mathbb{R}^d \times \Xi_i \to \mathbb{R}^d$, $i = 1, \dots, n$, be the local stochastic first-order oracles. Assume the main-text oracle conditions (19)–(20) for every $\mathbf{x} \in \mathbb{R}^d$ and every i,

  $$\mathbb{E}_{\xi^i \sim \mu^i}[g_i(\mathbf{x}, \xi^i)] = \nabla f_i(\mathbf{x}), \qquad \mathbb{E}_{\xi^i \sim \mu^i}\big[\|g_i(\mathbf{x}, \xi^i) - \nabla f_i(\mathbf{x})\|^2\big] \leq \frac{\sigma_i^2}{b}. \tag{63}$$

  For simplicity, one usually considers a uniform bound across all workers $\sigma = \max_i \sigma_i$, as also done in this paper.
  At iteration t, define

  $$\mathbf{g}_t^i \coloneqq g_i(\mathbf{x}_t, \xi_t^i), \qquad \mathbf{g}_t \coloneqq \frac{1}{n} \sum_{i=1}^n \mathbf{g}_t^i. \tag{64}$$

  Since $\mathbf{x}_t$ is $\mathcal{F}_{t-1}$-measurable and $\xi_t^1, \dots, \xi_t^n$ are independent of $\mathcal{F}_{t-1}$ and mutually independent,

  $$\mathbb{E}[\mathbf{g}_t^i \mid \mathcal{F}_{t-1}] = \nabla f_i(\mathbf{x}_t) \quad \Longrightarrow \quad \mathbb{E}[\mathbf{g}_t \mid \mathcal{F}_{t-1}] = \nabla f(\mathbf{x}_t), \tag{65}$$

  and, using conditional independence across i (so that cross terms vanish under conditional expectation),

  $$\mathbb{E}\big[\|\mathbf{g}_t - \nabla f(\mathbf{x}_t)\|^2 \mid \mathcal{F}_{t-1}\big] = \frac{1}{n^2} \sum_{i=1}^n \mathbb{E}\big[\|\mathbf{g}_t^i - \nabla f_i(\mathbf{x}_t)\|^2 \mid \mathcal{F}_{t-1}\big] \leq \frac{\sigma^2}{nb} =: \sigma_g^2. \tag{66}$$

- *Compressors (per worker, heterogeneous contraction rates).*
  For each worker i, let $\mathcal{C}_i(\cdot\,; h^i) : \mathbb{R}^d \to \mathbb{R}^d$ be a (possibly randomized) compression operator satisfying the main-text contractivity condition (21) with parameter $\delta_i \in (0, 1]$:

  $$\mathbb{E}_{h^i \sim \nu^i}\big[\|\mathcal{C}_i(\mathbf{u}; h^i) - \mathbf{u}\|^2\big] \leq (1 - \delta_i)\|\mathbf{u}\|^2, \qquad \forall \mathbf{u} \in \mathbb{R}^d, \tag{67}$$

  where the expectation is taken over the compressor's *own* internal randomness h^i, holding $\mathbf{u}$ fixed. In the algorithm, the input $\mathbf{u}_t^i$ to $\mathcal{C}_i$ at iteration t is a deterministic function of $\mathcal{F}_{t-1}$ (or of $\mathcal{F}_{t-1}$ together with the freshly drawn ξ_t^i), while h_t^i is independent of $\mathcal{F}_{t-1}$ and of $\mathbf{u}_t^i$, and is independent across i. Hence,

  $$\mathbb{E}\big[\|\mathcal{C}_i(\mathbf{u}_t^i; h_t^i) - \mathbf{u}_t^i\|^2 \mid \mathcal{F}_{t-1}, \mathbf{u}_t^i\big] \leq (1 - \delta_i)\|\mathbf{u}_t^i\|^2 \quad \text{a.s.} \tag{68}$$

  Taking total expectation via the tower property yields the unconditional bound

  $$\mathbb{E}\big[\|\mathcal{C}_i(\mathbf{u}_t^i; h_t^i) - \mathbf{u}_t^i\|^2\big] \leq (1 - \delta_i)\, \mathbb{E}\big[\|\mathbf{u}_t^i\|^2\big]. \tag{69}$$

  For EC-NAG, the raw discrepancy on the left is $\mathbf{e}_t^i - (1 - \gamma)\mathbf{e}_{t-1}^i$ by (8); it is therefore this difference, not $\mathbf{e}_t^i$ alone, to which compressor contractivity applies. For simplicity, one usually considers a uniform contraction rate across all workers $\delta_1 = \cdots = \delta_n = \delta$, as also done in this paper.

Let us close the discussion with **measurability of the algorithmic state**. The compressed/inexact local gradient $\widehat{\mathbf{g}}_t^i$ at worker i, iteration t, is a deterministic measurable function of $(\theta_0, \ldots, \theta_t)$; equivalently, of the algorithmic state up to time t (which is $\mathcal{F}_{t-1}$-measurable) together with the freshly drawn ξ_t^i and h_t^i. Consequently, $\widehat{\mathbf{g}}_t^i$, the aggregate $\widehat{\mathbf{g}}_t := \frac{1}{n}\sum_{i=1}^n \widehat{\mathbf{g}}_t^i$, and the updated iterates $\mathbf{v}_{t+1}$, $\mathbf{y}_{t+1}$, $\mathbf{x}_{t+1}$ are all $\mathcal{F}_t$-measurable. The gradient-evaluation point $\mathbf{x}_t$ and the objective iterate $\mathbf{y}_t$ are $\mathcal{F}_{t-1}$-measurable, consistent with the standing measurability assumptions used throughout the convergence analysis.

D Convergence analysis for EC-NAG

We study an accelerated method with inexact gradient information and aim to establish a unified Lyapunov recursion that captures the effect of stochasticity and inexactness. This recursion will serve as the main analytical tool for deriving convergence guarantees, and will later be specialized to compression-induced error.

D.1 Algorithm and Analysis of the General inexact-NAG

Algorithm 2 describes `inexact-NAG` in its three-point formulation. At each iteration t, the algorithm queries a stochastic gradient $\mathbf{g}_t := g(\mathbf{x}_t, \xi_t)$ at the point $\mathbf{x}_t$, where $g(\cdot,\cdot)$ is a stochastic first-order oracle, $\mathbf{x}_t$ is $\mathcal{F}_{t-1}$-measurable, and ξ_t is a freshly-drawn stochastic-oracle randomness independent of $\mathcal{F}_{t-1}$. The updates follow the standard accelerated scheme, the only modification being that the momentum step is driven by an inexact estimate $\widehat{\mathbf{g}}_t$. We do not impose any structural assumptions on the inexactness mechanism just like Gao et al. (2026).

Algorithm 2: `inexact-NAG`: NAG updates with inexact gradient

Given $\mathbf{v}_0, \mathbf{y}_0 \in \mathbb{R}^d$, $A_0 \geq 0$, a positive sequence $\{a_t\}_{t=1}^{\infty}$;
Set $A_1 = A_0 + a_1$ and $\mathbf{x}_0 = \frac{A_0}{A_1}\mathbf{y}_0 + \frac{a_1}{A_1}\mathbf{v}_0$;
for $t = 0, 1, 2, \ldots, T-1$ **do**
  Compute inexact gradient $\widehat{\mathbf{g}}_t \approx \mathbf{g}_t$;
  $A_{t+2} = A_{t+1} + a_{t+2}$;
  $\mathbf{v}_{t+1} = \mathbf{v}_t - a_{t+1}\widehat{\mathbf{g}}_t$; `// Perform the momentum update`
  $\mathbf{y}_{t+1} = \frac{A_t}{A_{t+1}}\mathbf{y}_t + \frac{a_{t+1}}{A_{t+1}}\mathbf{v}_{t+1}$; `// Update the objective iterate`
  $\mathbf{x}_{t+1} = \frac{A_{t+1}}{A_{t+2}}\mathbf{y}_{t+1} + \frac{a_{t+2}}{A_{t+2}}\mathbf{v}_{t+1}$; `// Update the gradient-evaluation point`
  Broadcast the gradient-access point $\mathbf{x}_{t+1}$ to all workers;

In order to analyze `inexact-NAG`, we introduce a virtual sequence that tracks the evolution of momentum under exact stochastic gradient:

$$\widetilde{\mathbf{v}}_{t+1} := \widetilde{\mathbf{v}}_t - a_{t+1}\mathbf{g}_t, \qquad \widetilde{\mathbf{v}}_0 := \mathbf{v}_0. \tag{70}$$

By comparing this sequence with the actual iterates, we can explicitly isolate the accumulated error due to inexact gradients. In particular,

$$\mathbf{v}_t = \mathbf{v}_0 - \sum_{\tau=0}^{t-1} a_{\tau+1}\widehat{\mathbf{g}}_\tau, \qquad \widetilde{\mathbf{v}}_t = \mathbf{v}_0 - \sum_{\tau=0}^{t-1} a_{\tau+1}\mathbf{g}_\tau, \tag{71}$$

which implies

$$\mathbf{v}_t = \widetilde{\mathbf{v}}_t - a_t\mathbf{e}_{t-1}, \tag{72}$$

where the error sequence $\{\mathbf{e}_t\}$ is defined as

$$\mathbf{e}_t := \frac{1}{a_{t+1}}\sum_{\tau=0}^{t} a_{\tau+1}\left(\widehat{\mathbf{g}}_\tau - \mathbf{g}_\tau\right), \qquad \mathbf{e}_{-1} = \mathbf{0}. \tag{73}$$

Consequently, the increment in the momentum admits the decomposition

$$\mathbf{v}_{t+1} - \mathbf{v}_t = (\widetilde{\mathbf{v}}_{t+1} - \widetilde{\mathbf{v}}_t) - (a_{t+1}\mathbf{e}_t - a_t\mathbf{e}_{t-1}), \tag{74}$$

which, together with the algorithmic identities (Line 6–8 in inexact-NAG), yields

$$A_{t+1}(\mathbf{y}_{t+1} - \mathbf{x}_t) = a_{t+1}(\mathbf{v}_{t+1} - \mathbf{v}_t). \tag{75}$$

The probability space and filtration $\mathcal{F}_t$ are described in C. Recall that we assume that stochastic first-order oracle satisfies

$$\mathbb{E}[\mathbf{g}_t \mid \mathcal{F}_{t-1}] = \nabla f(\mathbf{x}_t), \tag{76a}$$

$$\mathbb{E}\left[\|\mathbf{g}_t - \nabla f(\mathbf{x}_t)\|^2 \mid \mathcal{F}_{t-1}\right] \le \sigma_g^2. \tag{76b}$$

Our analysis is based on a Lyapunov potential that combines function suboptimality and a momentum error term. Specifically, we define

$$\Phi_t \coloneqq A_t\big(f(\mathbf{y}_t) - f(\mathbf{x}^*)\big) + \frac{1}{2}\|\widetilde{\mathbf{v}}_t - \mathbf{x}^*\|^2, \tag{77}$$

and denote

$$F_t \coloneqq \mathbb{E}[f(\mathbf{y}_t)] - f(\mathbf{x}^*). \tag{78}$$

The proof proceeds by expanding the Lyapunov potential around the gradient evaluation point, invoking a proximal interpretation of the virtual update, and carefully tracking the deviation between exact and inexact gradients through the error sequence $\{\mathbf{e}_t\}$. The argument combines smoothness-based descent inequalities, conditional unbiasedness of stochastic gradients, and repeated applications of Young's inequality to control stochastic and inexactness-induced cross terms. This leads to a one-step recursion for $\mathbb{E}[\Phi_t]$, which is then telescoped over iterations to obtain the desired bound. Unlike Gao et al. (2026), we will also retain the free parameters for now.

Lemma 1 (Recursion for inexact-NAG)**.** *Let f be convex and gradients satisfy* (76)*. For every $T \ge 1$, $\{(\mathbf{v}_t, \mathbf{y}_t, \mathbf{x}_t)\}_{t\ge 0}$ generated by Algorithm 2, satisfy*

$$\begin{aligned} A_T F_T \le A_0 F_0 + \frac{D^2}{2} + \frac{\varepsilon_1 + \varepsilon_3^{-1}}{2}\sigma_g^2 \sum_{t=0}^{T-1} a_{t+1}^2 + \sum_{t=0}^{T-1} w_{t+1} a_{t+1}^2 \mathbb{E}\left[\|\mathbf{e}_t\|^2\right] \\ - \sum_{t=0}^{T-1}\left(\frac{1 - \varepsilon_1^{-1} - \varepsilon_2^{-1}}{2} - \frac{2La_{t+1}^2}{A_{t+1}}\right) a_{t+1}^2 \mathbb{E}\left[\|\mathbf{g}_t\|^2\right] \\ - \sum_{t=1}^{T-1} A_t \mathbb{E}\left[\beta_f(\mathbf{x}_t, \mathbf{y}_t) + \beta_f(\mathbf{y}_t, \mathbf{x}_{t-1})\right], \end{aligned} \tag{79}$$

where $\varepsilon_1, \varepsilon_2, \varepsilon_3 > 0$ are arbitrary, $D^2 \coloneqq \|\mathbf{v}_0 - \mathbf{x}^*\|^2$ and

$$w_t \coloneqq \frac{\varepsilon_2 + \varepsilon_3}{2} + \frac{4La_t^2}{A_t} + \frac{4La_{t+1}^2}{A_{t+1}}. \tag{80}$$

Proof. For clarity, the proof is carried out in a couple of steps.

Step 1: Expand the accelerated potential around the gradient point. By the definition of β_f,

$$f(\mathbf{y}_t) = f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), \mathbf{y}_t - \mathbf{x}_t\rangle + \beta_f(\mathbf{x}_t, \mathbf{y}_t), \tag{81}$$

$$f(\mathbf{x}^*) = f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), \mathbf{x}^* - \mathbf{x}_t\rangle + \beta_f(\mathbf{x}_t, \mathbf{x}^*). \tag{82}$$

Substituting (81) and (82) in (77) gives

$$\begin{aligned} \Phi_t &= A_t f(\mathbf{y}_t) + a_{t+1} f(\mathbf{x}^*) - A_{t+1} f(\mathbf{x}^*) + \frac{1}{2}\|\widetilde{\mathbf{v}}_t - \mathbf{x}^*\|^2 \\ &= A_{t+1}\big(f(\mathbf{x}_t) - f(\mathbf{x}^*)\big) + A_t\left(\langle \nabla f(\mathbf{x}_t), \mathbf{y}_t - \mathbf{x}_t\rangle + \beta_f(\mathbf{x}_t, \mathbf{y}_t)\right) \\ &\quad + a_{t+1}\left(\langle \nabla f(\mathbf{x}_t), \mathbf{x}^* - \mathbf{x}_t\rangle + \beta_f(\mathbf{x}_t, \mathbf{x}^*)\right) + \frac{1}{2}\|\widetilde{\mathbf{v}}_t - \mathbf{x}^*\|^2. \end{aligned} \tag{83}$$

Since $\mathbf{x}_t$, $\mathbf{y}_t$, and $\widetilde{\mathbf{v}}_t$ are $\mathcal{F}_{t-1}$-measurable and $\mathbb{E}[\mathbf{g}_t \mid \mathcal{F}_{t-1}] = \nabla f(\mathbf{x}_t)$, we have $\mathbb{E}\left[\langle \nabla f(\mathbf{x}_t), \mathbf{x}^* - \mathbf{x}_t\rangle\right] = \mathbb{E}\left[\langle \mathbf{g}_t, \mathbf{x}^* - \mathbf{x}_t\rangle\right]$. Then, taking expectations in (83) gives

$$\begin{aligned} \mathbb{E}[\Phi_t] = \mathbb{E}\left[A_{t+1}\big(f(\mathbf{x}_t) - f(\mathbf{x}^*)\big) + A_t\left(\langle \nabla f(\mathbf{x}_t), \mathbf{y}_t - \mathbf{x}_t\rangle + \beta_f(\mathbf{x}_t, \mathbf{y}_t)\right)\right] \\ + \mathbb{E}\left[a_{t+1}\big(\langle \mathbf{g}_t, \mathbf{x}^* - \mathbf{x}_t\rangle + \beta_f(\mathbf{x}_t, \mathbf{x}^*)\big) + \frac{1}{2}\|\widetilde{\mathbf{v}}_t - \mathbf{x}^*\|^2\right]. \end{aligned} \tag{84}$$

Step 2: Euclidean proximal identity. The virtual update (70) is equivalent to

$$\widetilde{\mathbf{v}}_{t+1} = \arg\min_{\mathbf{v}\in\mathbb{R}^d} \left\{ a_{t+1}\langle \mathbf{g}_t, \mathbf{v}\rangle + \frac{1}{2}\|\mathbf{v} - \widetilde{\mathbf{v}}_t\|^2 \right\}. \tag{85}$$

By 1-strong convexity of the minimized objective,

$$a_{t+1}\langle \mathbf{g}_t, \mathbf{x}^*\rangle + \frac{1}{2}\|\mathbf{x}^* - \widetilde{\mathbf{v}}_t\|^2 \geq a_{t+1}\langle \mathbf{g}_t, \widetilde{\mathbf{v}}_{t+1}\rangle + \frac{1}{2}\|\widetilde{\mathbf{v}}_{t+1} - \widetilde{\mathbf{v}}_t\|^2 + \frac{1}{2}\|\mathbf{x}^* - \widetilde{\mathbf{v}}_{t+1}\|^2. \tag{86}$$

Substituting (86) in (84) gives

$$\begin{aligned}\mathbb{E}[\Phi_t] \geq\ & \mathbb{E}\left[A_{t+1}\big(f(\mathbf{x}_t) - f(\mathbf{x}^*)\big) + A_t\langle \nabla f(\mathbf{x}_t), \mathbf{y}_t - \mathbf{x}_t\rangle + A_t\beta_f(\mathbf{x}_t, \mathbf{y}_t)\right] \\ & + \mathbb{E}\left[a_{t+1}\big(\langle \mathbf{g}_t, \widetilde{\mathbf{v}}_{t+1} - \mathbf{x}_t\rangle + \beta_f(\mathbf{x}_t, \mathbf{x}^*)\big)\right] \\ & + \mathbb{E}\left[\frac{1}{2}\|\widetilde{\mathbf{v}}_t - \widetilde{\mathbf{v}}_{t+1}\|^2 + \frac{1}{2}\|\widetilde{\mathbf{v}}_{t+1} - \mathbf{x}^*\|^2\right]. \end{aligned} \tag{87}$$

Step 3: Identify Φ_{t+1} on the right-hand side. First, since $\widetilde{\mathbf{v}}_t - \mathbf{x}_t$ is $\mathcal{F}_{t-1}$-measurable and $\mathbb{E}[\mathbf{g}_t - \nabla f(\mathbf{x}_t) \mid \mathcal{F}_{t-1}] = \mathbf{0}$ (76a), we have $\mathbb{E}\left[\langle \mathbf{g}_t - \nabla f(\mathbf{x}_t), \widetilde{\mathbf{v}}_t - \mathbf{x}_t\rangle\right] = 0$. Combining this with $\widetilde{\mathbf{v}}_{t+1} - \mathbf{x}_t = (\widetilde{\mathbf{v}}_{t+1} - \widetilde{\mathbf{v}}_t) + (\widetilde{\mathbf{v}}_t - \mathbf{x}_t)$ gives

$$\mathbb{E}\left[\langle \mathbf{g}_t - \nabla f(\mathbf{x}_t), \widetilde{\mathbf{v}}_{t+1} - \mathbf{x}_t\rangle\right] = \mathbb{E}\left[\langle \mathbf{g}_t - \nabla f(\mathbf{x}_t), \widetilde{\mathbf{v}}_{t+1} - \widetilde{\mathbf{v}}_t\rangle\right]. \tag{88}$$

Second, using (72) and Lines 5,7 in Algorithm 2,

$$A_t(\mathbf{y}_t - \mathbf{x}_t) + a_{t+1}(\widetilde{\mathbf{v}}_{t+1} - \mathbf{x}_t) = A_{t+1}(\mathbf{y}_{t+1} - \mathbf{x}_t) + a_{t+1}^2\mathbf{e}_t; \tag{89}$$

therefore,

$$\begin{aligned} & A_t\langle \nabla f(\mathbf{x}_t), \mathbf{y}_t - \mathbf{x}_t\rangle + a_{t+1}\langle \nabla f(\mathbf{x}_t), \widetilde{\mathbf{v}}_{t+1} - \mathbf{x}_t\rangle \\ & = A_{t+1}\langle \nabla f(\mathbf{x}_t), \mathbf{y}_{t+1} - \mathbf{x}_t\rangle + a_{t+1}\langle \nabla f(\mathbf{x}_t), a_{t+1}\mathbf{e}_t\rangle. \end{aligned} \tag{90}$$

Substituting (88) and (90) in (87), and identifying Φ_{t+1} (77) on the right-hand side, gives

$$\begin{aligned}\mathbb{E}[\Phi_t] \geq\ & \mathbb{E}[\Phi_{t+1}] + A_{t+1}\mathbb{E}\left[f(\mathbf{x}_t) - f(\mathbf{y}_{t+1}) + \langle \nabla f(\mathbf{x}_t), \mathbf{y}_{t+1} - \mathbf{x}_t\rangle\right] \\ & + a_{t+1}\mathbb{E}\left[\langle \mathbf{g}_t - \nabla f(\mathbf{x}_t), \widetilde{\mathbf{v}}_{t+1} - \widetilde{\mathbf{v}}_t\rangle\right] + a_{t+1}\mathbb{E}\left[\langle \nabla f(\mathbf{x}_t), a_{t+1}\mathbf{e}_t\rangle\right] \\ & + \mathbb{E}\left[A_t\beta_f(\mathbf{x}_t, \mathbf{y}_t) + a_{t+1}\beta_f(\mathbf{x}_t, \mathbf{x}^*)\right] + \frac{1}{2}\mathbb{E}\left[\|\widetilde{\mathbf{v}}_t - \widetilde{\mathbf{v}}_{t+1}\|^2\right]. \end{aligned} \tag{91}$$

Step 4: Use smoothness to lower bound the objective term. By the reverse Bregman identity,

$$\begin{aligned} f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), \mathbf{y}_{t+1} - \mathbf{x}_t\rangle = & f(\mathbf{y}_{t+1}) + \beta_f(\mathbf{y}_{t+1}, \mathbf{x}_t) \\ & - \langle \nabla f(\mathbf{y}_{t+1}) - \nabla f(\mathbf{x}_t), \mathbf{y}_{t+1} - \mathbf{x}_t\rangle. \end{aligned} \tag{92}$$

By L-smoothness,

$$|\langle \nabla f(\mathbf{y}_{t+1}) - \nabla f(\mathbf{x}_t), \mathbf{y}_{t+1} - \mathbf{x}_t\rangle| \leq L\|\mathbf{y}_{t+1} - \mathbf{x}_t\|^2. \tag{93}$$

Combining (92) and (93),

$$f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), \mathbf{y}_{t+1} - \mathbf{x}_t\rangle \geq f(\mathbf{y}_{t+1}) + \beta_f(\mathbf{y}_{t+1}, \mathbf{x}_t) - L\|\mathbf{y}_{t+1} - \mathbf{x}_t\|^2. \tag{94}$$

Use (74), (75), and Young's inequality to bound $\|\mathbf{y}_{t+1} - \mathbf{x}_t\|^2$, then plugging into (94) gives

$$\begin{aligned} & A_{t+1}\big(f(\mathbf{x}_t) - f(\mathbf{y}_{t+1}) + \langle \nabla f(\mathbf{x}_t), \mathbf{y}_{t+1} - \mathbf{x}_t\rangle\big) \\ & \geq A_{t+1}\beta_f(\mathbf{y}_{t+1}, \mathbf{x}_t) - \frac{2La_{t+1}^2}{A_{t+1}}\|\widetilde{\mathbf{v}}_{t+1} - \widetilde{\mathbf{v}}_t\|^2 - \frac{2La_{t+1}^2}{A_{t+1}}\|a_{t+1}\mathbf{e}_t - a_t\mathbf{e}_{t-1}\|^2. \end{aligned} \tag{95}$$

Substituting (95) into (91) gives

$$\begin{aligned}\mathbb{E}[\Phi_t] \geq\ & \mathbb{E}[\Phi_{t+1}] + a_{t+1}\mathbb{E}\left[\langle \mathbf{g}_t - \nabla f(\mathbf{x}_t), \widetilde{\mathbf{v}}_{t+1} - \widetilde{\mathbf{v}}_t\rangle\right] + a_{t+1}\mathbb{E}\left[\langle \nabla f(\mathbf{x}_t), a_{t+1}\mathbf{e}_t\rangle\right] \\ & + \left(\frac{1}{2} - \frac{2La_{t+1}^2}{A_{t+1}}\right)\mathbb{E}\left[\|\widetilde{\mathbf{v}}_t - \widetilde{\mathbf{v}}_{t+1}\|^2\right] - \frac{2La_{t+1}^2}{A_{t+1}}\mathbb{E}\left[\|a_{t+1}\mathbf{e}_t - a_t\mathbf{e}_{t-1}\|^2\right] \\ & + \mathbb{E}\left[A_{t+1}\beta_f(\mathbf{y}_{t+1}, \mathbf{x}_t) + A_t\beta_f(\mathbf{x}_t, \mathbf{y}_t) + a_{t+1}\beta_f(\mathbf{x}_t, \mathbf{x}^*)\right]. \end{aligned} \tag{96}$$

Step 5: Bounding the inner products. Applying Young's inequality[1] to the second term in (96), then using (76b), gives

$$a_{t+1}\mathbb{E}\left[\langle \mathbf{g}_t - \nabla f(\mathbf{x}_t), \widetilde{\mathbf{v}}_{t+1} - \widetilde{\mathbf{v}}_t\rangle\right] \geq -\frac{\varepsilon_1 a_{t+1}^2}{2}\mathbb{E}\left[\|\mathbf{g}_t - \nabla f(\mathbf{x}_t)\|^2\right] - \frac{1}{2\varepsilon_1}\mathbb{E}\left[\|\widetilde{\mathbf{v}}_{t+1} - \widetilde{\mathbf{v}}_t\|^2\right]$$
$$\geq -\frac{\varepsilon_1 a_{t+1}^2}{2}\sigma_g^2 - \frac{1}{2\varepsilon_1}\mathbb{E}\left[\|\widetilde{\mathbf{v}}_{t+1} - \widetilde{\mathbf{v}}_t\|^2\right]. \tag{97}$$

Next, we upper bound the third term in (96). Using $\widetilde{\mathbf{v}}_t - \widetilde{\mathbf{v}}_{t+1} = a_{t+1}\mathbf{g}_t$ (70), then applying Young's inequality and using (76b), gives

$$\begin{aligned}
-a_{t+1}\mathbb{E}\left[\langle \nabla f(\mathbf{x}_t), a_{t+1}\mathbf{e}_t\rangle\right] =& -a_{t+1}\mathbb{E}\left[\langle \mathbf{g}_t, a_{t+1}\mathbf{e}_t\rangle\right] - a_{t+1}\mathbb{E}\left[\langle \nabla f(\mathbf{x}_t) - \mathbf{g}_t, a_{t+1}\mathbf{e}_t\rangle\right] \\
=& -\mathbb{E}\left[\langle \widetilde{\mathbf{v}}_t - \widetilde{\mathbf{v}}_{t+1}, a_{t+1}\mathbf{e}_t\rangle\right] - a_{t+1}\mathbb{E}\left[\langle \nabla f(\mathbf{x}_t) - \mathbf{g}_t, a_{t+1}\mathbf{e}_t\rangle\right] \\
\leq& \left(\frac{1}{2\varepsilon_2}\mathbb{E}\left[\|\widetilde{\mathbf{v}}_t - \widetilde{\mathbf{v}}_{t+1}\|^2\right] + \frac{\varepsilon_2}{2}a_{t+1}^2\mathbb{E}\left[\|\mathbf{e}_t\|^2\right]\right) \\
&+ \left(\frac{a_{t+1}^2}{2\varepsilon_3}\mathbb{E}\left[\|\nabla f(\mathbf{x}_t) - \mathbf{g}_t\|^2\right] + \frac{\varepsilon_3}{2}a_{t+1}^2\mathbb{E}\left[\|\mathbf{e}_t\|^2\right]\right) \\
\leq& \frac{1}{2\varepsilon_2}\mathbb{E}\left[\|\widetilde{\mathbf{v}}_t - \widetilde{\mathbf{v}}_{t+1}\|^2\right] + \frac{a_{t+1}^2}{2\varepsilon_3}\sigma_g^2 + \frac{\varepsilon_2+\varepsilon_3}{2}a_{t+1}^2\mathbb{E}\left[\|\mathbf{e}_t\|^2\right].
\end{aligned} \tag{98}$$

Plugging (97) and (98) in (96) and rearranging gives

$$\begin{aligned}
\mathbb{E}[\Phi_{t+1}] \leq \mathbb{E}[\Phi_t] &+ \frac{\varepsilon_1 + \varepsilon_3^{-1}}{2}a_{t+1}^2\sigma_g^2 - \left(\frac{1 - \varepsilon_1^{-1} - \varepsilon_2^{-1}}{2} - \frac{2La_{t+1}^2}{A_{t+1}}\right)\mathbb{E}\left[\|\widetilde{\mathbf{v}}_t - \widetilde{\mathbf{v}}_{t+1}\|^2\right] \\
&+ \frac{\varepsilon_2+\varepsilon_3}{2}a_{t+1}^2\mathbb{E}\left[\|\mathbf{e}_t\|^2\right] + \frac{2La_{t+1}^2}{A_{t+1}}\mathbb{E}\left[\|a_{t+1}\mathbf{e}_t - a_t\mathbf{e}_{t-1}\|^2\right] \\
&- a_{t+1}\mathbb{E}\left[\beta_f(\mathbf{x}_t, \mathbf{x}^*)\right] - A_t\mathbb{E}\left[\beta_f(\mathbf{x}_t, \mathbf{y}_t)\right] - A_{t+1}\mathbb{E}\left[\beta_f(\mathbf{y}_{t+1}, \mathbf{x}_t)\right].
\end{aligned} \tag{99}$$

Step 6: One-step recursion. Applying Young's inequality gives

$$\|a_{t+1}\mathbf{e}_t - a_t\mathbf{e}_{t-1}\|^2 \leq 2\|a_{t+1}\mathbf{e}_t\|^2 + 2\|a_t\mathbf{e}_{t-1}\|^2. \tag{100}$$

Substituting (100) into (99) yields the one-step recursion:

$$\begin{aligned}
\mathbb{E}[\Phi_{t+1}] \leq \mathbb{E}[\Phi_t] &+ \frac{\varepsilon_1 + \varepsilon_3^{-1}}{2}a_{t+1}^2\sigma_g^2 - \left(\frac{1 - \varepsilon_1^{-1} - \varepsilon_2^{-1}}{2} - \frac{2La_{t+1}^2}{A_{t+1}}\right)a_{t+1}^2\mathbb{E}\left[\|\mathbf{g}_t\|^2\right] \\
&+ \left(\frac{\varepsilon_2+\varepsilon_3}{2} + \frac{4La_{t+1}^2}{A_{t+1}}\right)a_{t+1}^2\mathbb{E}\left[\|\mathbf{e}_t\|^2\right] + \frac{4La_{t+1}^2}{A_{t+1}}a_t^2\mathbb{E}\left[\|\mathbf{e}_{t-1}\|^2\right] \\
&- a_{t+1}\mathbb{E}\left[\beta_f(\mathbf{x}_t, \mathbf{x}^*)\right] - A_t\mathbb{E}\left[\beta_f(\mathbf{x}_t, \mathbf{y}_t)\right] - A_{t+1}\mathbb{E}\left[\beta_f(\mathbf{y}_{t+1}, \mathbf{x}_t)\right].
\end{aligned} \tag{101}$$

Recall weights w_t defined in (80). Summing (101) over $t = 0, \ldots, T-1$ gives

$$\begin{aligned}
\mathbb{E}[\Phi_T] \leq \mathbb{E}[\Phi_0] &+ \frac{\varepsilon_1 + \varepsilon_3^{-1}}{2}\sigma_g^2\sum_{t=0}^{T-1}a_{t+1}^2 + \sum_{t=0}^{T-1}w_{t+1}a_{t+1}^2\mathbb{E}\left[\|\mathbf{e}_t\|^2\right] \\
&- \sum_{t=0}^{T-1}\left(\frac{1 - \varepsilon_1^{-1} - \varepsilon_2^{-1}}{2} - \frac{2La_{t+1}^2}{A_{t+1}}\right)a_{t+1}^2\mathbb{E}\left[\|\mathbf{g}_t\|^2\right] \\
&- \sum_{t=0}^{T-1}a_{t+1}\mathbb{E}\left[\beta_f(\mathbf{x}_t, \mathbf{x}^*)\right] \\
&- \sum_{t=0}^{T-1}A_t\mathbb{E}\left[\beta_f(\mathbf{x}_t, \mathbf{y}_t)\right] - \sum_{t=1}^{T}A_t\mathbb{E}\left[\beta_f(\mathbf{y}_t, \mathbf{x}_{t-1})\right],
\end{aligned} \tag{102}$$

[1] For $\varepsilon > 0$, we have $\|\mathbf{a} - \mathbf{b}\|^2 \leq (1 + \varepsilon^{-1})\|\mathbf{a}\|^2 + (1 + \varepsilon)\|\mathbf{b}\|^2$.

where the terms involving $\{E_t\}_{t=0}^{T}$ (recall $E_{-1}=0$) were simplified as follows:

$$\begin{aligned}
&\sum_{t=0}^{T-1}\left[\left(\frac{\varepsilon_2+\varepsilon_3}{2}+\frac{4La_{t+1}^2}{A_{t+1}}\right)a_{t+1}^2\mathbb{E}\left[\|\mathbf{e}_t\|^2\right]+\frac{4La_{t+1}^2}{A_{t+1}}a_t^2\mathbb{E}\left[\|\mathbf{e}_{t-1}\|^2\right]\right]\\
&=\sum_{t=0}^{T-1}\left(\frac{\varepsilon_2+\varepsilon_3}{2}+\frac{4La_{t+1}^2}{A_{t+1}}\right)a_{t+1}^2\mathbb{E}\left[\|\mathbf{e}_t\|^2\right]+\sum_{\tau=0}^{T-2}\frac{4La_{\tau+2}^2}{A_{\tau+2}}a_{\tau+1}^2\mathbb{E}\left[\|\mathbf{e}_\tau\|^2\right] \qquad (103a)\\
&\le\sum_{t=0}^{T-1}\left(\frac{\varepsilon_2+\varepsilon_3}{2}+\frac{4La_{t+1}^2}{A_{t+1}}+\frac{4La_{t+2}^2}{A_{t+2}}\right)a_{t+1}^2\,\mathbb{E}\left[\|\mathbf{e}_t\|^2\right] \qquad (103b)\\
&=\sum_{t=0}^{T-1}w_{t+1}a_{t+1}^2\mathbb{E}\left[\|\mathbf{e}_t\|^2\right] \qquad (103c)
\end{aligned}$$

by the inequality in (103b) due to the addition of the nonnegative terminal contribution $\frac{4La_{T+1}^2}{A_{T+1}}a_T^2\mathbb{E}\left[\|\mathbf{e}_{T-1}\|^2\right]$.

Substituting the expressions for Φ_T and Φ_0 as per (77) in (102) and dropping the nonnegative term $\frac{1}{2}\|\widetilde{\mathbf{v}}_t-\mathbf{x}^*\|^2$ from the left-hand side as well as the nonnegative terms $A_0\mathbb{E}\left[\beta_f(\mathbf{x}_0,\mathbf{y}_0)\right]$, $A_T\mathbb{E}\left[\beta_f(\mathbf{y}_T,\mathbf{x}_{T-1})\right]$, and $a_{t+1}\mathbb{E}\left[\beta_f(\mathbf{x}_t,\mathbf{x}^*)\right]$ from the right-hand side, yield the desired result. □

D.2 Analysis and Proof of EC-NAG

Recall the stochastic-gradient assumptions:

$$\mathbb{E}\left[\mathbf{g}_t^i-\nabla f_i(\mathbf{x}_t)\mid\mathcal{F}_{t-1}\right]=\mathbf{0}, \qquad (104a)$$

$$\mathbb{E}\left[\left\|\mathbf{g}_t^i-\nabla f_i(\mathbf{x}_t)\right\|^2\mid\mathcal{F}_{t-1}\right]\le\frac{\sigma^2}{b}. \qquad (104b)$$

Consequently,

$$\mathbb{E}[\mathbf{g}_t\mid\mathcal{F}_{t-1}]=\nabla f(\mathbf{x}_t),\qquad \mathbb{E}\left[\|\mathbf{g}_t-\nabla f(\mathbf{x}_t)\|^2\mid\mathcal{F}_{t-1}\right]\le\frac{\sigma^2}{nb}.$$

Thus, in the one-step recursion, $\sigma_g^2=\sigma^2/(nb)$ because the stochastic-gradient noises are independent across workers.

Define

$$E_t^2\coloneqq\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\|\mathbf{e}_t^i\|^2\right]\ge\mathbb{E}\left[\|\mathbf{e}_t\|^2\right],\qquad \mathbf{e}_t\coloneqq\frac{1}{n}\sum_{i=1}^{n}\mathbf{e}_t^i.$$

Proof of Theorem 1. The proof proceeds through a coupled envelope argument for the compressor inputs and the controlled error states.

Step 1: Recursion for the error envelope. Substituting the controlled-state update (8) at round $t-1$ into the compressor input (7) gives, for every $t\ge1$,

$$\begin{aligned}
\mathbf{u}_t^i=\mathbf{g}_t^i-\mathbf{g}_{t-1}^i&-\left(\gamma+\frac{a_t}{a_{t+1}}\right)\left(\mathbf{q}_{t-1}^i-\mathbf{u}_{t-1}^i\right)\\
&+\left(\frac{a_{t-1}}{a_t}-\left(\gamma+\frac{a_t}{a_{t+1}}\right)(1-\gamma)\right)\mathbf{e}_{t-2}^i. \qquad (105)
\end{aligned}$$

The two feedback terms must be combined before applying Young's inequality; otherwise, the cancellation generated by γ would be lost.

For arbitrary $\varepsilon_4, \varepsilon_5 > 0$, two applications of Young's inequality to (105), followed by compressor contractivity, give

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\|\mathbf{u}_t^i\|^2\right] \le (1+\varepsilon_4)\left(\gamma+\frac{a_t}{a_{t+1}}\right)^2\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\|\mathbf{q}_{t-1}^i-\mathbf{u}_{t-1}^i\|^2\right] + (1+\varepsilon_4^{-1})(1+\varepsilon_5)\left(\frac{a_{t-1}}{a_t}-\left(\gamma+\frac{a_t}{a_{t+1}}\right)(1-\gamma)\right)^2 E_{t-2}^2 + (1+\varepsilon_4^{-1})(1+\varepsilon_5^{-1})\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\|\mathbf{g}_t^i-\mathbf{g}_{t-1}^i\|^2\right]. \tag{106}$$

Using

$$\mathbf{g}_t^i-\mathbf{g}_{t-1}^i = \nabla f_i(\mathbf{x}_t)-\nabla f_i(\mathbf{x}_{t-1})+\left(g_i(\mathbf{x}_t,\xi_t^i)-\nabla f_i(\mathbf{x}_t)\right) - \left(g_i(\mathbf{x}_{t-1},\xi_{t-1}^i)-\nabla f_i(\mathbf{x}_{t-1})\right), \tag{107}$$

conditional unbiasedness gives

$$\mathbb{E}\left[\|\mathbf{g}_t^i-\mathbf{g}_{t-1}^i\|^2 \mid \mathcal{F}_{t-1}\right] = \left\|\nabla f_i(\mathbf{x}_t)-\nabla f_i(\mathbf{x}_{t-1})-\left(g_i(\mathbf{x}_{t-1},\xi_{t-1}^i)-\nabla f_i(\mathbf{x}_{t-1})\right)\right\|^2 + \mathbb{E}\left[\left\|g_i(\mathbf{x}_t,\xi_t^i)-\nabla f_i(\mathbf{x}_t)\right\|^2 \mid \mathcal{F}_{t-1}\right]. \tag{108}$$

Applying Young's inequality and (104) yields

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\|\mathbf{g}_t^i-\mathbf{g}_{t-1}^i\|^2\right] \le \frac{2}{n}\sum_{i=1}^{n}\mathbb{E}\left[\|\nabla f_i(\mathbf{x}_t)-\nabla f_i(\mathbf{x}_{t-1})\|^2\right]+3\frac{\sigma^2}{b}. \tag{109}$$

The coefficient 3 accounts for the stochastic-gradient noise at both iterations, noting that the previous stochastic gradient is already $\mathcal{F}_{t-1}$-measurable.

Inserting $\mathbf{y}_t$, applying Young's inequality, and then applying the average-smoothness assumption (17) gives

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\|\mathbf{g}_t^i-\mathbf{g}_{t-1}^i\|^2\right] \le 8\ell\,\mathbb{E}\left[\beta_f(\mathbf{x}_t,\mathbf{y}_t)+\beta_f(\mathbf{y}_t,\mathbf{x}_{t-1})\right]+3\frac{\sigma^2}{b}. \tag{110}$$

The controlled-state update and

$$\|(1-\gamma)\mathbf{a}+\mathbf{b}\|^2 \le (1-\gamma)\|\mathbf{a}\|^2+\frac{1}{\gamma}\|\mathbf{b}\|^2, \qquad 0<\gamma\le 1,$$

give

$$E_t^2 \le (1-\gamma)E_{t-1}^2+\frac{1}{\gamma n}\sum_{i=1}^{n}\mathbb{E}\left[\|\mathbf{q}_t^i-\mathbf{u}_t^i\|^2\right] \le (1-\gamma)E_{t-1}^2+\frac{1-\delta}{\gamma n}\sum_{i=1}^{n}\mathbb{E}\left[\|\mathbf{u}_t^i\|^2\right] \tag{111}$$

At initialization, $\mathbf{u}_0^i=\mathbf{g}_0^i$ and $\mathbf{e}_{-1}^i=\mathbf{0}$, so

$$E_0^2 \coloneqq \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\|\mathbf{q}_0^i-\mathbf{u}_0^i\|^2\right] \le \frac{1-\delta}{n}\sum_{i=1}^{n}\mathbb{E}\left[\|\mathbf{g}_0^i\|^2\right].$$

Step 2: Bound on the weighted cumulative error. Consider the standard accelerated schedule

$$a_t \coloneqq \frac{t+c}{M}, \qquad t\ge 1,$$
$$A_t = A_0+\sum_{\tau=1}^{t}a_\tau = \frac{(t+c)^2+t}{2M}, \qquad A_0 \coloneqq \frac{c^2}{2M}. \tag{112}$$

Here $c, M > 0$ will be chosen below. It remains to specify $\varepsilon_1, \varepsilon_2, \varepsilon_3, c, M$, and the admissible range of the feedback gain γ.

To avoid confusing the schedule ratio with the feedback gain γ, define

$$\vartheta_t \coloneqq \frac{a_{t+1}}{a_t} = \frac{t+c+1}{t+c} \le \frac{A_{t+1}}{A_t}. \tag{113}$$

The inequality follows because

$$\begin{aligned}&(t+c)\big((t+c+1)^2+t+1\big) - (t+c+1)\big((t+c)^2+t\big)\\ &= (t+c)^2 + t + 2c > 0.\end{aligned}$$

Also, since $A_t \ge (t+c)^2/(2M)$,

$$\eta_{t-1} \coloneqq \frac{a_t^2}{A_t} = \frac{2(t+c)^2}{M\big((t+c)^2+t\big)} \le \frac{2}{M} = \eta^\bullet. \tag{114}$$

Consequently, using (113),

$$\frac{\eta_t}{\eta_{t-1}} = \vartheta_t^2 \frac{A_t}{A_{t+1}} \le \vartheta_t. \tag{115}$$

Using $1 \le \vartheta_t$, the monotonicity $\vartheta_{t+1} \le \vartheta_t$, and (115), the weights in (80) satisfy

$$\begin{aligned}w_{t+1} &= \frac{\varepsilon_2+\varepsilon_3}{2} + 4L\eta_t + 4L\eta_{t+1}\\ &\le \frac{\varepsilon_2+\varepsilon_3}{2}\vartheta_t + 4L\vartheta_t\eta_{t-1} + 4L\vartheta_{t+1}\eta_t\\ &\le \vartheta_t w_t.\end{aligned} \tag{116}$$

Therefore, for $t \ge 1$,

$$\frac{w_{t+1}a_{t+1}^2}{w_t a_t^2} \le \vartheta_t^3 \le \vartheta_1^3 = \left(\frac{c+2}{c+1}\right)^3. \tag{117}$$

Assume temporarily that, for some B to be determined later,

$$\left|\frac{a_{t-1}}{a_t} - \left(\gamma + \frac{a_t}{a_{t+1}}\right)(1-\gamma)\right| \le B, \qquad t \ge 2. \tag{118}$$

Now that we have the prerequisites, multiply (111) by $w_{t+1}a_{t+1}^2$, sum over $t = 1, \dots, T-1$, and use (117); after adding the initial and terminal nonnegative terms, we obtain

$$\begin{aligned}&\big(1-(1-\gamma)\vartheta_1^3\big)\sum_{t=0}^{T-1} w_{t+1}a_{t+1}^2 E_t^2\\ &\le w_1 a_1^2 E_0^2 + \frac{1-\delta}{\gamma}\sum_{t=1}^{T-1} w_{t+1}a_{t+1}^2 \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\|\mathbf{u}_t^i\|^2\right].\end{aligned} \tag{119}$$

Clearly, we need a bound on the weighted cumulative sum of $\|\mathbf{u}_t^i\|$. So, multiply (106) by $w_{t+1}a_{t+1}^2$ and sum over $t = 1, \dots, T-1$ to get

$$\begin{aligned}&\sum_{t=1}^{T-1} w_{t+1}a_{t+1}^2 \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\|\mathbf{u}_t^i\|^2\right]\\ &\le (1+\varepsilon_4)(1-\delta)\sum_{\tau=1}^{T-2} w_{\tau+2}a_{\tau+2}^2\left(\gamma + \frac{a_{\tau+1}}{a_{\tau+2}}\right)^2 \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\|\mathbf{u}_\tau^i\|^2\right]\\ &\quad + (1+\varepsilon_4)w_2 a_2^2\left(\gamma + \frac{a_1}{a_2}\right)^2 E_0^2\\ &\quad + (1+\varepsilon_4^{-1})(1+\varepsilon_5)\sum_{t=1}^{T-1} w_{t+1}a_{t+1}^2\left(\frac{a_{t-1}}{a_t} - \left(\gamma + \frac{a_t}{a_{t+1}}\right)(1-\gamma)\right)^2 E_{t-2}^2\\ &\quad + (1+\varepsilon_4^{-1})(1+\varepsilon_5^{-1})\sum_{t=1}^{T-1} w_{t+1}a_{t+1}^2 \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\|\mathbf{g}_t^i - \mathbf{g}_{t-1}^i\|^2\right].\end{aligned} \tag{120}$$

This is the required weighted compressor-input recursion. We retain its coefficients when coupling it to the controlled-state recursion. For the shifted compressor-input term, (117) gives

$$\begin{aligned}\frac{w_{s+2}a_{s+2}^2}{w_{s+1}a_{s+1}^2}\left(\gamma+\frac{a_{s+1}}{a_{s+2}}\right)^2 &\le \vartheta_{s+1}^3\left(\gamma+\frac{1}{\vartheta_{s+1}}\right)^2\\ &= \vartheta_{s+1}\left(1+\gamma\vartheta_{s+1}\right)^2\\ &\le \vartheta_1\left(1+\gamma\vartheta_1\right)^2.\end{aligned}$$

For $t \ge 2$,

$$\frac{w_{t+1}a_{t+1}^2}{w_{t-1}a_{t-1}^2} \le \vartheta_{t-1}^3\vartheta_t^3 \le \vartheta_1^3\vartheta_2^3.$$

Using these bounds and (118) in (120) gives

$$\begin{aligned}&\left(1-(1+\varepsilon_4)(1-\delta)\vartheta_1\left(1+\gamma\vartheta_1\right)^2\right)\sum_{t=1}^{T-1}w_{t+1}a_{t+1}^2\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\|\mathbf{u}_t^i\|^2\right]\\ &\le (1+\varepsilon_4)w_1a_1^2\vartheta_1(1+\gamma\vartheta_1)^2E_0^2+(1+\varepsilon_4^{-1})(1+\varepsilon_5)B^2\vartheta_1^3\vartheta_2^3\sum_{t=0}^{T-1}w_{t+1}a_{t+1}^2E_t^2\\ &\quad+(1+\varepsilon_4^{-1})(1+\varepsilon_5^{-1})\sum_{t=1}^{T-1}w_{t+1}a_{t+1}^2\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\|\mathbf{g}_t^i-\mathbf{g}_{t-1}^i\|^2\right].\end{aligned} \tag{121}$$

When $\delta = 1$, contractivity gives $\mathbf{q}_t^i = \mathbf{u}_t^i$ almost surely, so $E_t = 0$. For the remainder of Step 2, assume $0 < \delta < 1$; the final bound extends to $\delta = 1$ by the preceding observation. After dividing (121) by its left-hand coefficient, the two resulting coefficients containing $1+\varepsilon_4^{-1}$ are minimized by

$$\varepsilon_4 = \frac{1}{\sqrt{(1-\delta)\vartheta_1\left(1+\gamma\vartheta_1\right)^2}} - 1, \tag{122}$$

provided $(1-\delta)\vartheta_1(1+\gamma\vartheta_1)^2 < 1$. Substituting (122) into (121) gives the next display because, for every $0 < r < 1$,

$$\min_{\substack{\varepsilon>0:\\(1+\varepsilon)r<1}}\frac{1+\varepsilon^{-1}}{1-(1+\varepsilon)r} = \frac{1}{(1-\sqrt{r})^2}, \qquad \varepsilon_\star = \frac{1}{\sqrt{r}}-1,$$

and $(1+\varepsilon_\star)\left(1-(1+\varepsilon_\star)r\right)^{-1} = \left(\sqrt{r}(1-\sqrt{r})\right)^{-1}$. Thus, (122) gives

$$\begin{aligned}&\sum_{t=1}^{T-1}w_{t+1}a_{t+1}^2\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\|\mathbf{u}_t^i\|^2\right]\\ &\le \frac{w_1a_1^2\vartheta_1\left(1+\gamma\vartheta_1\right)^2}{\sqrt{(1-\delta)\vartheta_1\left(1+\gamma\vartheta_1\right)^2}\left(1-\sqrt{(1-\delta)\vartheta_1\left(1+\gamma\vartheta_1\right)^2}\right)}E_0^2\\ &\quad+\frac{(1+\varepsilon_5)B^2\vartheta_1^3\vartheta_2^3}{\left(1-\sqrt{(1-\delta)\vartheta_1\left(1+\gamma\vartheta_1\right)^2}\right)^2}\sum_{t=0}^{T-1}w_{t+1}a_{t+1}^2E_t^2\\ &\quad+\frac{1+\varepsilon_5^{-1}}{\left(1-\sqrt{(1-\delta)\vartheta_1\left(1+\gamma\vartheta_1\right)^2}\right)^2}\sum_{t=1}^{T-1}w_{t+1}a_{t+1}^2\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\|\mathbf{g}_t^i-\mathbf{g}_{t-1}^i\|^2\right].\end{aligned} \tag{123}$$

Next, substituting (123) into (119) gives

$$\left(1-(1-\gamma)\vartheta_1^3-\frac{(1-\delta)(1+\varepsilon_5)B^2\vartheta_1^3\vartheta_2^3}{\gamma\left(1-\sqrt{(1-\delta)\vartheta_1\left(1+\gamma\vartheta_1\right)^2}\right)^2}\right)\sum_{t=0}^{T-1}w_{t+1}a_{t+1}^2E_t^2$$

$$\leq w_1a_1^2E_0^2+\frac{w_1a_1^2\sqrt{(1-\delta)\vartheta_1\left(1+\gamma\vartheta_1\right)^2}}{\gamma\left(1-\sqrt{(1-\delta)\vartheta_1\left(1+\gamma\vartheta_1\right)^2}\right)}E_0^2$$

$$+\frac{(1-\delta)(1+\varepsilon_5^{-1})}{\gamma\left(1-\sqrt{(1-\delta)\vartheta_1\left(1+\gamma\vartheta_1\right)^2}\right)^2}\sum_{t=1}^{T-1}w_{t+1}a_{t+1}^2\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\|\mathbf{g}_t^i-\mathbf{g}_{t-1}^i\|^2\right]. \quad (124)$$

After absorbing the error-sum term on the left and dividing by its positive coefficient, the gradient-difference coefficient is minimized by

$$\varepsilon_5=\sqrt{\frac{\left(1-(1-\gamma)\vartheta_1^3\right)\left(1-\sqrt{(1-\delta)\vartheta_1\left(1+\gamma\vartheta_1\right)^2}\right)^2}{\frac{1-\delta}{\gamma}B^2\vartheta_1^3\vartheta_2^3}}-1. \quad (125)$$

This choice is positive precisely when the second inequality in (127) below holds. Substituting (125) into (124) yields the optimized bound. Here we use the scalar identity, valid for $A_1>A_2>0$,

$$\min_{\substack{\varepsilon>0:\\ A_2(1+\varepsilon)<A_1}}\frac{1+\varepsilon^{-1}}{A_1-A_2(1+\varepsilon)}=\frac{1}{(\sqrt{A_1}-\sqrt{A_2})^2},\qquad \varepsilon_\star=\sqrt{\frac{A_1}{A_2}}-1,$$

while $A_1-A_2(1+\varepsilon_\star)=\sqrt{A_1}(\sqrt{A_1}-\sqrt{A_2})$ for the initial terms. Therefore,

$$\sum_{t=0}^{T-1}w_{t+1}a_{t+1}^2E_t^2$$

$$\leq\frac{w_1a_1^2E_0^2+\frac{w_1a_1^2\sqrt{(1-\delta)\vartheta_1\left(1+\gamma\vartheta_1\right)^2}}{\gamma\left(1-\sqrt{(1-\delta)\vartheta_1\left(1+\gamma\vartheta_1\right)^2}\right)}E_0^2}{\sqrt{1-(1-\gamma)\vartheta_1^3}\left(\sqrt{1-(1-\gamma)\vartheta_1^3}-\frac{\sqrt{\frac{1-\delta}{\gamma}}B(\vartheta_1\vartheta_2)^{3/2}}{1-\sqrt{(1-\delta)\vartheta_1\left(1+\gamma\vartheta_1\right)^2}}\right)}$$

$$+\frac{(1-\delta)\left(1-\sqrt{(1-\delta)\vartheta_1\left(1+\gamma\vartheta_1\right)^2}\right)^{-2}}{\gamma\left(\sqrt{1-(1-\gamma)\vartheta_1^3}-\frac{\sqrt{\frac{1-\delta}{\gamma}}B(\vartheta_1\vartheta_2)^{3/2}}{1-\sqrt{(1-\delta)\vartheta_1\left(1+\gamma\vartheta_1\right)^2}}\right)^2}$$

$$\times\sum_{t=1}^{T-1}w_{t+1}a_{t+1}^2\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\|\mathbf{g}_t^i-\mathbf{g}_{t-1}^i\|^2\right]. \quad (126)$$

The positivity conditions for (122) and (125) are

$$(1-\delta)\vartheta_1\left(1+\gamma\vartheta_1\right)^2<1, \quad (127a)$$

$$\left(1-(1-\gamma)\vartheta_1^3\right)\left(1-\sqrt{(1-\delta)\vartheta_1\left(1+\gamma\vartheta_1\right)^2}\right)^2>\frac{1-\delta}{\gamma}B^2\vartheta_1^3\vartheta_2^3. \quad (127b)$$

The two inequalities in (127) should be solved jointly. The first (127a) makes the second factor on the left of (127b) positive, while the second inequality (127b) necessarily makes

$$1-(1-\gamma)\vartheta_1^3>0. \tag{128}$$

Based on this, we determine $B=\gamma^2$ as shown below. For $t\geq 2$, direct substitution of (112) gives

$$\frac{a_{t-1}}{a_t}-\left(\gamma+\frac{a_t}{a_{t+1}}\right)(1-\gamma)=\gamma^2-\frac{\gamma}{t+c+1}-\frac{1}{(t+c)(t+c+1)}. \tag{129}$$

The corresponding term at $t=1$ multiplies $\mathbf{e}_{-1}^i=\mathbf{0}$. For $t\geq 2$, the right-hand side of (129) is increasing in t and is always at most γ^2. Moreover, the state-stability condition (128) implies that its value at $t=2$ is nonnegative. Indeed, let $x=1/(c+2)\in(0,1/2)$. Since $\vartheta_1=1/(1-x)$, the condition $1-(1-\gamma)\vartheta_1^3>0$ gives

$$\gamma>1-(1-x)^3=x(3-3x+x^2).$$

Since $3-3x+x^2\geq 7/4$, we have

$$\begin{aligned}&(1+x)(3-3x+x^2)^2-(3-3x+x^2)-1\\&=(3-3x+x^2)\big((1+x)(3-3x+x^2)-1\big)-1\\&\geq\frac{5}{16}>0.\end{aligned}$$

Because $(1+x)\gamma^2-x\gamma-x^2$ is increasing for $\gamma\geq x(3-3x+x^2)$, it follows that

$$(1+x)\gamma^2-x\gamma-x^2>x^2\left((1+x)(3-3x+x^2)^2-(3-3x+x^2)-1\right)>0.$$

This is precisely the nonnegativity of the right-hand side of (129) at $t=2$. Consequently,

$$1-(1-\gamma)\vartheta_1^3>0\quad\Longrightarrow\quad 0\leq\frac{a_{t-1}}{a_t}-\left(\gamma+\frac{a_t}{a_{t+1}}\right)(1-\gamma)\leq\gamma^2,\quad t\geq 2. \tag{130}$$

Thus, $B=\gamma^2$ in (118)–(127), and no separate lower bound on c is needed for the cancellation term.

Substituting $B=\gamma^2$ into (127), taking positive square roots, and rearranging shows that the joint conditions are equivalent to

$$1-(1-\gamma)\vartheta_1^3>0, \tag{131a}$$

$$\sqrt{1-\delta}\left(\sqrt{\vartheta_1}\,(1+\gamma\vartheta_1)+\frac{\gamma^{3/2}(\vartheta_1\vartheta_2)^{3/2}}{\sqrt{1-(1-\gamma)\vartheta_1^3}}\right)<1. \tag{131b}$$

The second inequality (131b) also implies (127a), so no positivity condition has been lost.

Step 3: Applying Lemma 1 for the Lyapunov bound. Both $\vartheta_1=(c+2)/(c+1)$ and $\vartheta_2=(c+3)/(c+2)$ strictly decrease with c. On the domain (131a), the left-hand side of (131b) is therefore strictly decreasing in c. Its limiting value as $c\to\infty$ is $\sqrt{1-\delta}(1+2\gamma)$. This limit is used only to determine whether a finite admissible c exists; the finite-c condition itself is retained. Thus, for $0<\delta<1$, a finite admissible c exists exactly when

$$0<\gamma\leq 1,\qquad \gamma<\frac{1-\sqrt{1-\delta}}{2\sqrt{1-\delta}}. \tag{132}$$

This condition on γ in (132) is for every $\delta\in(0,1)$; hence imposes no lower bound on δ.

For any γ satisfying (132), choose the finite $c>0$ so that

$$c>\frac{1}{(1-\gamma)^{-1/3}-1}-1,\qquad 0<\gamma<1, \tag{133a}$$

$$\sqrt{1-\delta}\left(\sqrt{\frac{c+2}{c+1}}\left(1+\gamma\frac{c+2}{c+1}\right)+\frac{\gamma^{3/2}\left(\frac{c+3}{c+1}\right)^{3/2}}{\sqrt{1-(1-\gamma)\left(\frac{c+2}{c+1}\right)^3}}\right)<1. \tag{133b}$$

The first inequality (133a) is exactly (131a) solved for c; for $\gamma = 1$ it is vacuous. On the part of $c > 0$ where (133a) holds, the left-hand side of (133b) decreases strictly to the limiting value used in (132). If the stability boundary lies in $c \geq 0$, the left-hand side diverges there; otherwise, it is already finite at $c = 0$. In either case, (132) guarantees that (133b) holds for all sufficiently large finite c. Accordingly, (133) is the exact finite-c requirement. For a statement uniform over all $c \geq c_0$, it is enough to verify (133) at $c = c_0$, where ϑ_1 and ϑ_2 are largest. Notice also that (133a) is precisely the stability condition used in (130); hence no independent cancellation constraint is required.

To avoid repeatedly writing the two long coefficients already present in (126), denote the exact optimized coefficients by

$$C_{\mathrm{grad}} \coloneqq \frac{(1-\delta)\left(1-\sqrt{(1-\delta)\vartheta_1\left(1+\gamma\vartheta_1\right)^2}\right)^{-2}}{\gamma\left(\sqrt{1-(1-\gamma)\vartheta_1^3}-\dfrac{\sqrt{1-\delta}\,\gamma^{3/2}(\vartheta_1\vartheta_2)^{3/2}}{1-\sqrt{(1-\delta)\vartheta_1\left(1+\gamma\vartheta_1\right)^2}}\right)^2}, \tag{134a}$$

$$C_{\mathrm{init}} \coloneqq \frac{\left(1+\dfrac{\sqrt{(1-\delta)\vartheta_1\left(1+\gamma\vartheta_1\right)^2}}{\gamma\left(1-\sqrt{(1-\delta)\vartheta_1\left(1+\gamma\vartheta_1\right)^2}\right)}\right)}{\sqrt{1-(1-\gamma)\vartheta_1^3}\left(\sqrt{1-(1-\gamma)\vartheta_1^3}-\dfrac{\sqrt{1-\delta}\,\gamma^{3/2}(\vartheta_1\vartheta_2)^{3/2}}{1-\sqrt{(1-\delta)\vartheta_1\left(1+\gamma\vartheta_1\right)^2}}\right)}. \tag{134b}$$

These are the coefficients obtained after the two exact scalar minimizations in (122) and (125); no additional relaxation has yet been made.

Using (110) in (126) yields

$$\sum_{t=0}^{T-1} w_{t+1}a_{t+1}^2E_t^2 \leq 8\ell C_{\mathrm{grad}}\sum_{t=1}^{T-1} w_{t+1}a_{t+1}^2\mathbb{E}\left[\beta_f(\mathbf{x}_t,\mathbf{y}_t)+\beta_f(\mathbf{y}_t,\mathbf{x}_{t-1})\right] + 3C_{\mathrm{grad}}\frac{\sigma^2}{b}\sum_{t=1}^{T-1} w_{t+1}a_{t+1}^2 + C_{\mathrm{init}}w_1a_1^2E_0^2. \tag{135}$$

Substituting (135) into (79) from Lemma 1 gives

$$A_TF_T \leq A_0F_0 + \frac{D^2}{2} + C_{\mathrm{init}}w_1a_1^2E_0^2 + \frac{\varepsilon_1+\varepsilon_3^{-1}}{2}\frac{\sigma^2}{nb}\sum_{t=0}^{T-1} a_{t+1}^2 + 3C_{\mathrm{grad}}\frac{\sigma^2}{b}\sum_{t=1}^{T-1} w_{t+1}a_{t+1}^2, \tag{136}$$

provided

$$\frac{1-\varepsilon_1^{-1}-\varepsilon_2^{-1}}{2} - \frac{2La_{t+1}^2}{A_{t+1}} \geq 0, \quad t = 0,\dots,T-1, \tag{137a}$$

$$8\ell C_{\mathrm{grad}}w_{t+1}a_{t+1}^2 \leq A_t, \qquad t = 1,\dots,T-1. \tag{137b}$$

Condition (137a) preserves the nonnegative coefficient of the gradient term in the inexact-NAG recursion, whereas (137b) absorbs the two Bregman-divergence terms generated by the error envelope (135). We now choose the remaining parameters so that both conditions hold simultaneously.

Set $\varepsilon_1 = 4$, $\varepsilon_2 = 2$, and $\varepsilon_3 = 1$. Then (137a) becomes

$$\frac{1-4^{-1}-2^{-1}}{2} - \frac{4L(t+c)^2}{M\big((t+c)^2+t\big)} \geq 0.$$

The fraction $(t+c)^2/((t+c)^2+t)$ is at most one, so the condition holds whenever

$$M \geq 32L. \tag{138}$$

Under (138), the weights satisfy

$$w_t = \frac{3}{2} + \frac{4La_t^2}{A_t} + \frac{4La_{t+1}^2}{A_{t+1}} \leq \frac{3}{2} + \frac{16L}{M} \leq 2. \tag{139}$$

This uniform bound allows us to verify the remaining absorption condition without changing the weighted-envelope argument. Indeed,

$$\frac{a_{t+1}^2}{A_t} \leq \frac{2}{M}\left(\frac{t+c+1}{t+c}\right)^2,$$

and hence (137b) follows from

$$\frac{24\ell C_{\mathrm{grad}}\vartheta_1^2}{M} + \frac{256L\ell C_{\mathrm{grad}}\vartheta_1^2}{M^2} \leq 1.$$

Solving this quadratic inequality exactly shows that both conditions in (137) hold whenever

$$M \geq \max\left\{32L,\, 12\ell C_{\mathrm{grad}}\vartheta_1^2 + \sqrt{144\ell^2 C_{\mathrm{grad}}^2\vartheta_1^4 + 256L\ell C_{\mathrm{grad}}\vartheta_1^2}\right\}. \tag{140}$$

This is the exact condition on M produced by (137). We next derive a transparent sufficient version after controlling C_{grad}.

Step 4: Parameter choices and final bound. The exact conditions (132)–(133) are the least restrictive conditions on (γ, c). To turn the exact coefficients (134) into explicit uniform bounds, we retain a quantitative margin from the boundary by requiring

$$0 < \gamma \leq \frac{\delta}{8}, \qquad c \geq \frac{10}{\gamma}. \tag{141}$$

This is a sufficient subset of the exact feasible region, rather than a replacement for (132)–(133).

We first bound the three denominator terms appearing in (134). Since $c \geq 10/\gamma$,

$$\vartheta_1 = 1 + \frac{1}{c+1} \leq 1 + \frac{\gamma}{10+\gamma} = \frac{10+2\gamma}{10+\gamma},$$
$$\vartheta_1\vartheta_2 = 1 + \frac{2}{c+1} \leq 1 + \frac{2\gamma}{10+\gamma} \leq 1 + \frac{\delta}{40}.$$

For the state-stability denominator,

$$\left(1 - \frac{7}{10}\gamma\right)(10+\gamma)^3 - (1-\gamma)(10+2\gamma)^3 = 300\gamma^2 + 92\gamma^3 + \frac{73}{10}\gamma^4 \geq 0.$$

Thus,

$$1 - (1-\gamma)\vartheta_1^3 \geq \frac{7}{10}\gamma.$$

Next, $\sqrt{\vartheta_1}(1+\gamma\vartheta_1)$ is increasing in both γ and ϑ_1. Its worst case under (141) is therefore attained at $\gamma = \delta/8$ and $c = 10/\gamma$. Squaring the desired inequality and clearing the positive denominator gives

$$\left(1 - \frac{59}{160}\delta\right)^2 - (1-\delta)\frac{80+2\delta}{80+\delta}\left(1 + \frac{\delta}{8}\frac{80+2\delta}{80+\delta}\right)^2$$
$$= \frac{\delta^2}{64(80+\delta)^3}\left(12345600 + 1156656\delta + \frac{248927}{5}\delta^2 + \frac{409881}{400}\delta^3 + 8\delta^4\right) \geq 0.$$

Consequently,

$$1 - \sqrt{(1-\delta)\vartheta_1\left(1+\gamma\vartheta_1\right)^2} \geq \frac{59}{160}\delta.$$

For the remaining coupling term, observe that

$$1-(1-\delta)\left(1+\frac{\delta}{40}\right)^3=\frac{37}{40}\delta+\frac{117}{1600}\delta^2+\frac{119}{64000}\delta^3+\frac{1}{64000}\delta^4\geq 0.$$

It follows that

$$\sqrt{1-\delta}(\vartheta_1\vartheta_2)^{3/2}\leq 1.$$

Combining this inequality with the preceding denominator bound and $\gamma\leq\delta/8$ yields

$$\frac{\sqrt{1-\delta}\,\gamma^{3/2}(\vartheta_1\vartheta_2)^{3/2}}{1-\sqrt{(1-\delta)\vartheta_1\left(1+\gamma\vartheta_1\right)^2}}\leq\frac{20}{59}\sqrt{\gamma}.$$

Altogether,

$$1-(1-\gamma)\vartheta_1^3\geq\frac{7}{10}\gamma, \tag{142a}$$

$$1-\sqrt{(1-\delta)\vartheta_1\left(1+\gamma\vartheta_1\right)^2}\geq\frac{59}{160}\delta, \tag{142b}$$

$$\sqrt{1-(1-\gamma)\vartheta_1^3}-\frac{\sqrt{1-\delta}\,\gamma^{3/2}(\vartheta_1\vartheta_2)^{3/2}}{1-\sqrt{(1-\delta)\vartheta_1\left(1+\gamma\vartheta_1\right)^2}}\geq\left(\sqrt{\frac{7}{10}}-\frac{20}{59}\right)\sqrt{\gamma}. \tag{142c}$$

The last coefficient is strictly positive. Thus, (142a) and (142b) give the two individual denominator margins, while (142c) gives the remaining combined margin. In particular, (142) verifies directly that (141) lies inside the exact feasible region.

Substituting (142a)–(142c) into (134) gives

$$C_{\mathrm{grad}}\leq\left(\frac{160}{59\left(\sqrt{7/10}-20/59\right)}\right)^2\frac{1-\delta}{\gamma^2\delta^2}<29.693\frac{1-\delta}{\gamma^2\delta^2}, \tag{143a}$$

$$C_{\mathrm{init}}\leq\frac{160}{59\sqrt{7/10}\left(\sqrt{7/10}-20/59\right)}\frac{1}{\gamma^2\delta}<6.513\frac{1}{\gamma^2\delta}. \tag{143b}$$

For (143b), we retain the useful negative term in the identity

$$\begin{aligned}&\gamma^2\delta\left(1+\frac{\sqrt{(1-\delta)\vartheta_1(1+\gamma\vartheta_1)^2}}{\gamma\left(1-\sqrt{(1-\delta)\vartheta_1(1+\gamma\vartheta_1)^2}\right)}\right)\\&=\frac{\gamma\delta}{1-\sqrt{(1-\delta)\vartheta_1(1+\gamma\vartheta_1)^2}}-\gamma\delta(1-\gamma)\\&\leq\frac{\gamma\delta}{1-\sqrt{(1-\delta)\vartheta_1(1+\gamma\vartheta_1)^2}}.\end{aligned}$$

This is why the coefficient in (143b) is smaller than the one obtained by bounding the two positive numerator terms separately.

The constants in (143) are sharp over (141). Indeed, along

$$\delta\downarrow 0,\qquad\gamma=\frac{\delta}{8},\qquad c=\frac{10}{\gamma},$$

the three normalized quantities in (142) converge, respectively, to $7/10$, $59/160$, and $20/59$. Hence the normalized coefficients in (143) converge to the displayed constants, and those constants cannot be reduced uniformly.

We now fix only the schedule offset. Since the explicit bounds in (143) already hold uniformly for every $c\geq 10/\gamma$, taking a larger c is unnecessary and would enlarge the initial-condition terms. We therefore set

$$c=\frac{10}{\gamma}. \tag{144}$$

while keeping γ free in the full interval $0 < \gamma \le \delta/8$.

We next simplify the sufficient condition on M. Under (144),

$$\vartheta_1 = \frac{10 + 2\gamma}{10 + \gamma} \le \frac{82}{81},$$

where the last inequality uses $\gamma \le \delta/8 \le 1/8$. Therefore, (143a) implies

$$\ell C_{\mathrm{grad}} \vartheta_1^2 < 29.693 \left(\frac{82}{81}\right)^2 \frac{(1-\delta)\ell}{\gamma^2 \delta^2} < 30.431 \frac{(1-\delta)\ell}{\gamma^2 \delta^2}.$$

Moreover, directly,

$$\sqrt{144 \ell^2 C_{\mathrm{grad}}^2 \vartheta_1^4 + 256 L \ell C_{\mathrm{grad}} \vartheta_1^2} \le 12 \ell C_{\mathrm{grad}} \vartheta_1^2 + 32 L,$$

because the square of the right-hand side exceeds the expression under the square root by

$$512 L \ell C_{\mathrm{grad}} \vartheta_1^2 + 1024 L^2 \ge 0.$$

Consequently, the exact condition (140) is implied by the simple sufficient condition

$$M \ge 32L + 731 \frac{(1-\delta)\ell}{\gamma^2 \delta^2}. \tag{145}$$

Indeed, the coefficient required by the preceding calculation is strictly smaller than $24(30.431) = 730.344$.

We now substitute the coefficient bounds and the choice of c before writing the first estimate for F_T. Using (139) and the schedule (112), we have

$$\sum_{t=0}^{T-1} a_{t+1}^2 \le \frac{T(T + 10/\gamma)^2}{M^2}, \quad \sum_{t=1}^{T-1} w_{t+1} a_{t+1}^2 \le \frac{2T(T + 10/\gamma)^2}{M^2}, \quad A_T \ge \frac{(T + 10/\gamma)^2}{2M}. \tag{146}$$

Substituting these estimates and (146) directly into (136), then using $T + \frac{10}{\gamma} \ge T$, yields

$$F_T \le \frac{100 F_0}{\gamma^2 T^2} + \frac{D^2 M}{T^2} + \frac{1}{M} \left(27 \frac{(1 + 10/\gamma)^2}{\gamma^2 \delta T^2} E_0^2 + T \left(\frac{5\sigma^2}{nb} + 357 \frac{1-\delta}{\gamma^2 \delta^2} \frac{\sigma^2}{b} \right) \right). \tag{147}$$

Thus, the dependence on γ and δ is explicit before the balancing step, while γ remains a free algorithmic parameter.

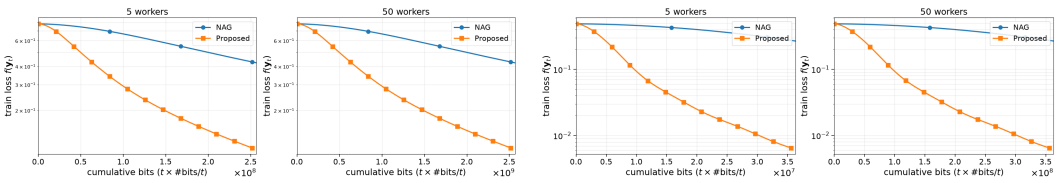
For fixed T and $D > 0$, if the bracketed term in (147) is positive, the last two terms are balanced by

$$M_\star = \frac{T}{D} \sqrt{27 \frac{(1 + 10/\gamma)^2}{\gamma^2 \delta T^2} E_0^2 + T \left(\frac{5\sigma^2}{nb} + 357 \frac{1-\delta}{\gamma^2 \delta^2} \frac{\sigma^2}{b} \right)}. \tag{148}$$

If the bracketed term vanishes, set $M_\star = 0$; then the entire M^{-1} term in (147) is zero, so the same final estimate below holds. If the actual initial distance is zero, one may use any positive upper bound for D in this tuning rule and then let that upper bound tend to zero in the resulting inequality.

Combining the sufficient absorption condition (145) with the balancing value (148), choose the simple explicit value

$$M := \max\left\{ 32L + 731 \frac{(1-\delta)\ell}{\gamma^2 \delta^2}, M_\star \right\}. \tag{149}$$



(a) `real-sim`. (b) `bcsstk15`.

Figure 4: `EC-NAG` against full-precision `NAG` with exact local gradients, $n \in \{5, 50\}$, on the additional convex datasets.

For the choice (149), $M_\star \le M$ and

$$M \le 32L + 731\frac{(1-\delta)\ell}{\gamma^2\delta^2} + M_\star,$$

substitution into (147) gives

$$F_T \le \frac{\frac{100F_0}{\gamma^2} + D^2\left(32L + 731\frac{(1-\delta)\ell}{\gamma^2\delta^2}\right)}{T^2} + \frac{2D}{T}\left(27\frac{(1+10/\gamma)^2}{\gamma^2\delta T^2}E_0^2 + T\left(\frac{5\sigma^2}{nb} + 357\frac{1-\delta}{\gamma^2\delta^2}\frac{\sigma^2}{b}\right)\right)^{1/2}. \tag{150}$$

This is the explicit finite-horizon bound obtained without fixing γ.

Finally, applying $\gamma \le \delta/8 \le 1/8$ and $1 + \frac{10}{\gamma} \le \frac{81}{8\gamma}$ to (150) and using $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ gives

$$F_T = \mathcal{O}\left(\frac{F_0/\gamma^2 + D^2\left(L + \frac{(1-\delta)\ell}{\gamma^2\delta^2}\right) + \frac{DE_0}{\gamma^2\sqrt{\delta}}}{T^2} + \frac{\sigma D}{\sqrt{bT}}\sqrt{\frac{1}{n} + \frac{1-\delta}{\gamma^2\delta^2}}\right). \tag{151}$$

The estimate (151) is the claimed simplified bound. This completes the proof. □

E Extra Experiments

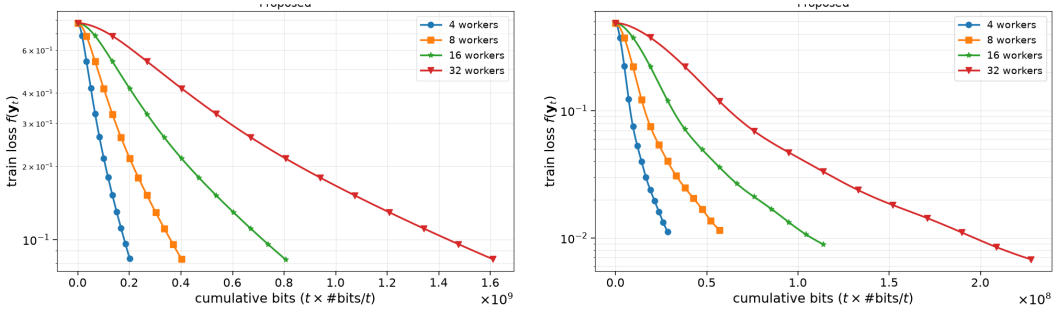
This appendix repeats the three experiment families of Section 4 on `real-sim` (binary logistic regression) and `bcsstk15` (least squares). The protocol is unchanged: row normalization, nested tune-select, train loss $f(\mathbf{y}_t)$ versus cumulative uplink bits, $R = R_c$, single runs. Large LIBSVM matrices are subsampled under a fixed seed when densifying would exceed a memory budget (affecting `real-sim`).

The qualitative trends of Section 4 carry over. In Figure 4, `EC-NAG` stays competitive with `NAG` per round and is ahead of it per uplink bit. In Figure 5 the stochastic curves again improve with the number of workers; as in the main text we read this only as a benign dependence on n and make no speedup claim. In Figure 6, the unaccelerated compressed methods descend without the accelerated slope, while `EC-NAG` remains competitive with `ADEF` on the bit axis because it sends one compressed vector per round instead of two.

Methods such as `DIANA`, `EF21`/`EF-BV`, `ADIANA`, `2Direction`, and `EF-OLGA` remain important conceptually, but their published guarantees change at least one of the compressor, stochastic-oracle, convexity, or communication-direction assumptions, so their theoretical rates are not directly interchangeable with the same-setting comparison considered (Mishchenko et al., 2025; Richtárik et al., 2021; Condat et al., 2022; Li et al., 2020; Tyurin & Richtárik, 2023; Bylinkin & Beznosikov, 2025).

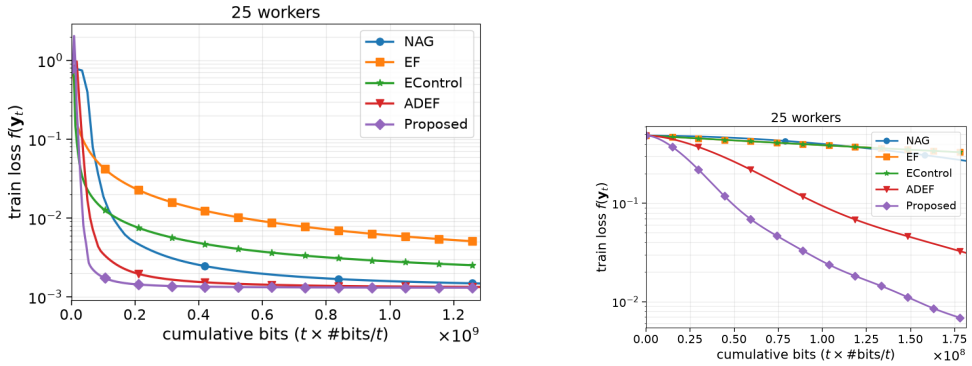
F Implementation Details

Software and seeds. The experiments were executed in 64-bit floating-point arithmetic on Microsoft Windows 10, build 10.0.26200, using Python 3.9.13, NumPy 2.0.2, SciPy 1.13.1, PyTorch



(a) `real-sim`, $b = 256$. (b) `bcsstk15`, $b = 64$.

Figure 5: Worker-count study for stochastic `EC-NAG`, $n \in \{4, 8, 16, 32\}$.



(a) `real-sim`, $n = 25$, $b = 256$. (b) `bcsstk15`, $n = 25$, $b = 64$.

Figure 6: Baselines at $n = 25$ with stochastic gradients; `ADEF` charged two compressed uplinks per round.

2.8.0 + `cpu`, scikit-learn 1.6.1, and Matplotlib 3.9.4. Workers are simulated in one process with a parameter-server pattern. The host has an AMD Ryzen 9 8940HX CPU and 15.21 GiB of RAM. The machine contains an NVIDIA GeForce RTX 5070 Ti Laptop GPU. The source code and run manifests will be released upon publication under a research-compatible license.

Data and normalization. Row normalization is always enabled: least-squares rows and responses are divided by the row ℓ_2 norms; classification features are row-normalized. `ILLc1850` and `bcsstk15` come from SuiteSparse without a right-hand side; responses are drawn i.i.d. standard normal under the data seed and then normalized with the rows. Large LIBSVM matrices are subsampled under a fixed seed when densifying would exceed a memory budget (`real-sim`).

Objectives. All reported experiments use the convex nature: least squares without ridge and logistic regression without a regularizer. The server aggregates with weights $1/n$. Each f_i is the shard loss multiplied by a scale $s_i \propto m_i$ so that $\frac{1}{n}\sum_i f_i$ equals the global sample average under unequal shards.

Partitioning. SuiteSparse least-squares problems use a random row partition. Classification problems use label-based shards (non-i.i.d.).

Compression and bit accounting. The compressor is the fixed-rate group-wise uniform quantizer with $\delta = 1 - z\,2^{-2R}$ and automatic critical rate $R = R_c = \lfloor \log_2 \sqrt{z} \rfloor + 1$. Linear experiments use $z = d$ ($G = 1$). One compressed uplink costs $B = dR + 32G$ bits. Accounting is uplink-only: `NAG` pays $32d$ per worker per round; `EF`, `EControl`, and `EC-NAG` pay B; `ADEF` pays $2B$. Exact

tracker warm-starts are local computation and are not charged as uplink. The plotted bit axis is the ledger total over all workers.

Hyperparameter selection. Free knobs are selected independently for each (algorithm, dataset, worker count) by nested tune-select. Steps η and η_∞ are searched over $\{2^{-12}, \dots, 2^{0}\}$. For `EControl` and `EC-NAG`, the search is nested at $\gamma = \delta$, then $\gamma \in \{\delta, \delta/2, \dots\}$, then a re-tuned step; $c = 1/\delta$ is fixed for `ADEF`/`EC-NAG`. Each trial is scored by the late-tail mean of $\|\nabla f(\mathbf{x}_t)\|^2$ (rejecting non-finite / diverging runs).

Horizons and batches. Main linear runs use $T = 400$ rounds with 60 tuning rounds per trial. Appendix convex runs use $T = 300$ with 50 tuning rounds. Stochastic least-squares experiments use per-worker batch size $b = 64$; stochastic logistic experiments use $b = 256$. The effective batch on worker i is $\min\{b, m_i\}$.

Evaluation. Every round records the exact full-batch train loss $f(\mathbf{y}_t)$ at the gradient-step iterate (the plotted metric) and $\|\nabla f(\mathbf{x}_t)\|^2$ at the query iterate (used only for tuning). The horizontal axis is cumulative uplink cost as above, not the round index.