Generalization Without Compression Penalty: A Stability Analysis of Error Feedback

Yifei Liang[1,2], Peng Wang[1], Yan Sun[3], Yingqi Liu[4]
Xiaoyan Wang[5], Xiaochun Cao[1], Li Shen[1,2]*
[1]Sun Yat-sen University, Shenzhen Campus, Shenzhen, China
[2]Shenzhen Loop Area Institute, Shenzhen, China
[3]Generative AI Lab, Nanyang Technological University, Singapore
[4]Wuhan University, Wuhan, China
[5]Information Technology Service Center of People's Court, Beijing, China
liangyf65@mail2.sysu.edu.cn, wangp389@mail.sysu.edu.cn
sun.yan.95@outlook.com, liuyingqi1199@gmail.com
428163395@139.com, caoxiaochun@mail.sysu.edu.cn, shenli6@mail.sysu.edu.cn

Abstract

Error Feedback (EF) compensates for aggressive compression in distributed training by carrying residuals into later updates. Prior work shows that this correction can preserve convergence rates. Its effect on generalization under biased compression and local updates is less understood. We analyze single-node EF-SGD and DiLoCo-EF with SGD at both levels through on-average model stability, without bounded-gradient assumptions. The virtual update cancels the explicit buffer difference, while gradients still depend on the actual model. We couple virtual iterates and buffers across neighboring runs. Gradient co-coercivity absorbs gradient–buffer cross terms, and the optimization bounds control the gradient energy along the trajectory. For convex losses under the stated stepsize schedules, the explicit buffer term decays with the horizon, while compression still affects the optimization constants. Without EF, residuals can accumulate across rounds. Under interpolation, the explicit buffer contribution decays faster than the $T^{-1/2}$ optimization term for fixed problem parameters and admissible stepsize constants. In DiLoCo, worker averaging attenuates sample-replacement perturbations, and a local stepsize $\eta_l = \Theta(1/(H\sqrt{T}))$ controls local drift. Under interpolation, we obtain $O((MHT)^{-1/2})$ excess risk when the local sample size permits the required compression-dependent horizon. At matched nominal value payloads, MNIST generalization gaps remain close across Top-k, Random-k, and quantization over the tested horizon, for both the single-node EF-SGD and distributed DiLoCo-EF settings.

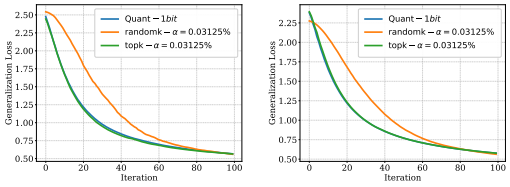
1 Introduction

The growing scale of modern machine learning models has made memory use and communication bandwidth central constraints. Model compression reduces storage [23, 25, 14, 53], while gradient compression cuts the volume of data that workers exchange during distributed training [1, 57]. Error Feedback (EF), first used in 1-bit quantization [42], compensates for compression distortion by carrying each residual into the next update. DiLoCo makes exchanges less frequent by letting workers take several local steps between synchronization rounds [16]. Variants with EF also compress the resulting model displacements, reducing the amount of data sent per round [41].

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Prior analyses show that EF can match the convergence rate of standard SGD and correct compression bias [44, 52, 28]. Subsequent work established linear convergence and acceleration for EF variants [21, 39, 40] and studied EF in distributed training [50, 55, 34]. These guarantees concern the training objective, leaving open how past residuals affect the returned model's sensitivity to its training data. Detached EF [54] uses uniform stability to study a modified feedback rule for smooth non-convex loss functions.



(a) Single-node EF-SGD (b) DiLoCo-EF

Figure 1: MNIST generalization gaps at matched nominal value payloads: 3.125% retention for Top-k and Random-k, and 1-bit QSGD.

Stability depends on how the compressor responds to small changes in the update. Top-k, for instance, can select different coordinates for nearby updates. The approximation guarantee bounds each compression residual, but does not bound the amplification of a change in the input. Adding the residual back changes the model at which the next gradient is evaluated, coupling the model and buffer even when neighboring runs use the same sample. In DiLoCo, worker trajectories can also drift over the H local steps before the server averages their displacements. The effect of feedback on stability thus depends on both the local steps and the server average.

What is the statistical effect of Error Feedback in distributed compressed training?

We study single-node EF-SGD and DiLoCo-EF with SGD at both levels. Subtracting the buffer correction from the model gives a virtual iterate whose update cancels the explicit buffer difference. Since gradients are still evaluated at the actual model, we track both virtual iterates and buffers in two runs that differ by one example. On shared samples, gradient co-coercivity absorbs the gradient–buffer cross term. In DiLoCo, auxiliary local paths that skip steps using the replaced example isolate the shared update. Each local step on the affected worker draws this example with probability $1/N$; its update enters the server average with weight $1/M$ at synchronization.

We use on-average model stability [33, 32, 47] and retain the dependence on gradient energy along the trajectory. Contractivity bounds the buffer energy by the energy of the uncompressed updates. A Lyapunov estimate combines the virtual iterate's squared distance to an empirical minimizer with weighted buffer energy to bound the average empirical risk. For nonnegative smooth losses, this risk controls gradient energy without a bounded-gradient assumption. For DiLoCo, we also account for local drift from the server model. Worker averaging reduces independent, centered sampling noise, while compression residuals can align across workers. For the time-averaged output $\bar{\boldsymbol{x}}_T$, the generalization gap and empirical optimization error together bound excess population risk.

Table 1 compares these guarantees with those for uncompressed training and Detached EF. It reports each rate with its chosen stepsize and distinguishes excess risk from the generalization gap.

- **Error feedback yields a transient buffer term** (Section 4.2). The stability bounds for single-node EF-SGD and DiLoCo-EF (Theorems 4.9 and 4.11) retain the buffer contribution to output sensitivity under replacement of a training example. The optimization bounds turn it into a decaying term in the generalization bound at fixed compression and admissible stepsize constants. Under interpolation, it decays faster than the $T^{-1/2}$ optimization term. Their relative size at a finite horizon still depends on compression through the coefficients and admissible stepsizes. The no-EF comparison explains how uncompensated residuals accumulate across rounds and how conditional centering and worker independence sharpen the bound (Remark 4.8).
- **Worker averaging and local drift in DiLoCo-EF.** The distributed bound separates the attenuation of sample-replacement perturbations from the drift caused by local updates. With $\eta_l = \Theta(1/(H\sqrt{T}))$ and outer stepsize $\eta = \sqrt{MH}$, Remark 4.16 gives $O(1/\sqrt{MHT})$ excess risk under interpolation when the local sample size permits the compression-dependent horizon. The stepsizes balance initialization and sampling noise, while the horizon restrictions keep drift, buffer, and sample-replacement terms at or below the stated excess-risk rate.
- **Experiments examine finite-horizon compression effects** (Section 5). The MNIST experiments vary stepsize and compression, along with worker count and local steps in DiLoCo. Figure 1 compares compressors within each setting. The DiLoCo sweeps on DCLM-baseline 1.0 extend the comparison to a non-convex task and show similar compression trends (Appendix A.3).

Table 1: Selected expected risk bounds and stepsizes for L-smooth losses.

Method	Step size	Quantity	Bound
Single-node			
SGD [24]*	$\eta = c/\sqrt{T}$	Excess risk	$O(T^{-1/2} + \sqrt{T}/N)$
SGD [33]	$\eta = c/\sqrt{T}$	Excess risk	$O(N^{-1/2})$
EF-SGD (Thm. 4.13)	$\eta = \frac{1}{4L(1+\sqrt{b_\alpha})\sqrt{T}}$	Excess risk	$O\left(\frac{1+\sqrt{b_\alpha}}{\sqrt{T}} + \frac{1}{N}\right)^\dagger$
Distributed			
D-SGD (avg.) [59]	$\eta = c/\sqrt{T}$	Excess risk	$O\left(\frac{1}{\sqrt{MN}} + \frac{1}{MN(1-\lambda)^2}\right)$
FedAvg [47]*	$\eta_l = c/H$	Gen. gap	$O\left(\frac{T}{MN}\left(D + \nu + \frac{1}{\sqrt[4]{MHT}} + \frac{1}{\sqrt{T}}\right)\right)$
Detached EF [54]	$\eta_t = c/(t+1)$	Gen. gap	$O\left(\frac{T}{N}\left[\frac{1}{T} + \frac{N}{T}\sqrt{\frac{(1-\alpha)(2-\alpha)}{\alpha}}\right]^{N/[N+(N-1)Lc]}\right)$
DiLoCo-EF (Thm. 4.15)	$\eta = \sqrt{MH}$	Excess risk	$O(1/\sqrt{MHT})^\dagger$

T: total iterations/rounds; t: current index; M, N, H: workers, samples/worker and local steps. α: contraction; $b_\alpha = (1-\alpha)/(1-\sqrt{1-\alpha})^2$. Blue: compression dependence. λ, D, ν: Appendix F. * bounded gradients; $^\dagger F(x^\star) = 0$. SGD [33]: $T \asymp N$; EF-SGD: $T \le N^2$. D-SGD: time-averaged output, i.i.d. data, symmetric mixing, $T \asymp MN/\sqrt{L}, \eta \le (1-\lambda)/(8L)$. Detached EF: nonconvex, $M, N \ge 2, T \gtrsim 1 + N\sqrt{(1-\alpha)(2-\alpha)/\alpha}$. DiLoCo-EF: $\eta_l = c_l/(H\sqrt{T})$, $MH(1+b_\alpha) + b_\alpha^2(MH)^3 \lesssim T \lesssim MN^2/H$.

2 Related work

Compression and error feedback. Quantization [1] and low-rank compression [49] reduce the size of messages exchanged in distributed training [22]. With biased compressors such as Top-k, the distortion can disrupt convergence [4, 43], motivating error feedback [42]. EF-SGD admits convex and non-convex convergence guarantees [44, 52, 45], recovers convergence for sign-based updates [28], and extends to momentum optimizers [60]. EF21 uses Markov compression to obtain an $\mathcal{O}(1/T)$ stationarity bound for smooth non-convex objectives without bounded gradients [40]. Other EF analyses establish linear convergence [21], acceleration [39], guarantees for adaptive optimizers [9, 10], and convergence in federated settings with biased compression and partial participation [36]. For LLMs, CocktailSGD combines compression techniques for fine-tuning [50]. DiLoCo uses local AdamW and infrequent outer Nesterov updates on model displacements [16], and its scaling behavior has been studied under a fixed compute budget [8]. SparseLoCo adds sparsified momentum and EF to compress the exchanges during pre-training [41]. When the compressor acts on a pseudo-gradient, the same buffer identity applies. We study SGD at both levels to separate residual compensation, worker averaging, and local drift. AdamW and outer momentum would add optimizer states to the coupling. Appendix B discusses other DiLoCo variants.

Stability and generalization. Algorithmic stability connects generalization to how a change in the training data affects the returned model [6, 18]. Hardt et al. [24] established uniform-stability bounds for SGD under global bounded-gradient assumptions. Lei and Ying [33] used on-average model stability to derive expected generalization bounds without those assumptions. Subsequent work sharpens high-probability bounds from uniform stability [7] and develops on-average bounds for minibatch and local SGD [32]. In distributed settings, Sun et al. [47] and Ye et al. [56] study how data heterogeneity affects generalization. The Libra framework relates stability to optimization for FedAvg [58]. The closest trajectory-dependent bounds concern uncompressed decentralized learning: Ye et al. [56] assume strong convexity, while Zeng and Lei [59] cover general convex objectives. Related analyses track the effect of stale gradients in asynchronous SGD [13] and relate decentralized SGD stability to the communication graph and the distribution of data across nodes [31]. Detached EF [54] modifies the gradient-evaluation point and studies uniform stability. Standard EF evaluates gradients at the actual model, so its virtual trajectory remains coupled to the buffer. We control this model–buffer coupling through gradient co-coercivity and cumulative buffer energy to isolate the compression contribution in stability and excess-risk bounds for convex losses.

3 Preliminaries

We first formulate the learning problem in the single-node and distributed settings, then describe the corresponding updates with error feedback. We introduce the risk measures and on-average model stability used in the analysis. Appendix F summarizes the notation.

3.1 Problem formulation

Let $f(\boldsymbol{x};\xi)$ be the loss of model $\boldsymbol{x} \in \mathbb{R}^d$ on sample ξ. Worker $m \in [M]$ holds $\mathcal{S}_m = (\xi_{1,m},\dots,\xi_{N,m})$, drawn i.i.d. from $\mathcal{D}_m$, independently across workers. Its empirical risk is $F_{\mathcal{S}_m}(\boldsymbol{x}) := N^{-1}\sum_{n=1}^{N} f(\boldsymbol{x};\xi_{n,m})$. For the full dataset $\mathcal{S} = (\mathcal{S}_1,\dots,\mathcal{S}_M)$, the population and empirical risks average the corresponding worker risks:

$$F(\boldsymbol{x}) := \frac{1}{M}\sum_{m=1}^{M} \mathbb{E}_{\xi\sim\mathcal{D}_m}\big[f(\boldsymbol{x};\xi)\big], \quad F_{\mathcal{S}}(\boldsymbol{x}) := \frac{1}{M}\sum_{m=1}^{M} F_{\mathcal{S}_m}(\boldsymbol{x}).$$

For a single node, $M = 1$ and we drop the worker index, writing $\mathcal{D} = \mathcal{D}_1$ and $\mathcal{S} = (\xi_1,\dots,\xi_N)$. Where needed, we assume a finite-risk population minimizer $\boldsymbol{x}^\star$ and a measurable empirical minimizer $\boldsymbol{x}^\star_{\mathcal{S}}$ that exists almost surely. Initial-distance terms assume $\mathbb{E}\|\boldsymbol{x}_0 - \boldsymbol{x}^\star_{\mathcal{S}}\|^2 < \infty$.

3.2 Algorithm

Both algorithms optimize the empirical risk using a possibly randomized compressor $\mathcal{C} : \mathbb{R}^d \to \mathbb{R}^d$. Its input is an SGD step or the displacement after several local steps, with the buffer correction added in either case. We measure distortion by the squared residual relative to the squared input norm.

Definition 3.1 (α-approximate compressor)**.** Fix $\alpha \in (0,1]$. A possibly randomized map $\mathcal{C} : \mathbb{R}^d \to \mathbb{R}^d$ is an α-approximate compressor if every $\boldsymbol{x} \in \mathbb{R}^d$ satisfies $\mathbb{E}_{\mathcal{C}}\|\boldsymbol{x} - \mathcal{C}(\boldsymbol{x})\|_2^2 \le (1-\alpha)\|\boldsymbol{x}\|_2^2$.

A larger α gives a smaller bound on the expected squared residual, with $\alpha = 1$ recovering the uncompressed case. The condition permits biased as well as unbiased compressors. The examples below relate α to coordinate retention, quantization error, and rank (see also Appendix D.3).

Algorithm 1: Single-node EF-SGD

Require: $\boldsymbol{x}_0$, stepsize η, T steps, compressor $\mathcal{C}$
Initialize $\boldsymbol{e}_0 = \mathbf{0}$
for $t = 0, 1, \dots, T-1$ **do**
  Sample $i_t \sim \mathrm{Unif}([N])$
  $\boldsymbol{g}_t \leftarrow \nabla f(\boldsymbol{x}_t;\xi_{i_t})$
  $\Delta_t \leftarrow \eta \boldsymbol{g}_t$
  $\boldsymbol{u}_t \leftarrow \mathcal{C}(\Delta_t + \boldsymbol{e}_t)$
  $\boldsymbol{e}_{t+1} \leftarrow \Delta_t + \boldsymbol{e}_t - \boldsymbol{u}_t$
  $\boldsymbol{x}_{t+1} \leftarrow \boldsymbol{x}_t - \boldsymbol{u}_t$
end for
Output: $\bar{\boldsymbol{x}}_T = T^{-1}\sum_{t=0}^{T-1} \boldsymbol{x}_t$

Sparsification. Top-k keeps the k largest coordinates in magnitude, while unscaled Random-k keeps a uniform subset of size k without rescaling [2, 44, 4]. The retained fraction of the squared norm is at least k/d for Top-k and equals k/d in expectation for Random-k, giving $\alpha = k/d$ for both.

Quantization. Quantizers such as QSGD [1] represent entries using a finite set of levels. A squared distortion bound $\mathbb{E}_{\mathcal{C}}\big\|\mathcal{C}(\boldsymbol{x}) - \boldsymbol{x}\big\|_2^2 \le \omega\|\boldsymbol{x}\|_2^2$ with $0 \le \omega < 1$ gives $\alpha = 1 - \omega$ in Definition 3.1. For an unbiased quantizer with a larger distortion bound, Appendix D.3 shows how rescaling makes it contractive.

Low-rank approximation. For a $p \times q$ matrix update, exact rank-r truncation keeps the r largest singular components, giving $\alpha = r/\min(p,q)$ in Frobenius norm (Appendix D.3). Approximate methods such as PowerSGD [49] need a separate residual bound.

3.2.1 Error feedback

Error feedback (EF) [42, 28] compensates for distortion by adding the stored residual to the next compression input. Algorithm 1 compresses the SGD step $\Delta_t = \eta \boldsymbol{g}_t$ plus the previous residual $\boldsymbol{e}_t$. With $\boldsymbol{e}_{t+1}$ as the new residual, the equivalent update is

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta \boldsymbol{g}_t - (\boldsymbol{e}_t - \boldsymbol{e}_{t+1}). \tag{1}$$

In distributed training, DiLoCo [16] reduces synchronization frequency. Each worker takes H local steps from $\boldsymbol{x}_t$ with stepsize η_l and treats the resulting displacement $\Delta_t^{(m)}$ as a pseudo-gradient. To further reduce communication, each worker adds its buffer before compressing this displacement and keeps the new residual for the next round [41]. In Algorithm 2, the server averages the compressed messages and takes an outer step with stepsize η.

With SGD at both levels, the H local steps give $\Delta_t^{(m)} = \eta_l \sum_{h=0}^{H-1} \boldsymbol{g}_{t,h}^{(m)}$. Each compressed message is this displacement plus the previous buffer minus the new one. Write $\overline{\boldsymbol{e}}_t = M^{-1} \sum_{m=1}^{M} \boldsymbol{e}_t^{(m)}$ for the mean buffer. The server's SGD step uses the average of these messages, giving the update

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \frac{\eta \eta_l}{M} \sum_{m=1}^{M} \sum_{h=0}^{H-1} \boldsymbol{g}_{t,h}^{(m)} - \eta\big(\overline{\boldsymbol{e}}_t - \overline{\boldsymbol{e}}_{t+1}\big). \tag{2}$$

Algorithm 2 DiLoCo with Error Feedback

Require: Initial model $\boldsymbol{x}_0$, workers M, rounds T, local steps H, stepsizes η_l, η, compressor $\mathcal{C}$, optimizers $\mathcal{O}_{\text{in}}, \mathcal{O}_{\text{out}}$

Initialize $\boldsymbol{e}_0^{(m)} = \mathbf{0}$, $\boldsymbol{s}_{\text{in},0}^{(m)}$ for all $m \in [M]$, and $\boldsymbol{s}_{\text{out},0}$
for $t = 0, 1, \ldots, T-1$ **do**
  Server: broadcast $\boldsymbol{x}_t$ to all workers
  for each worker $m \in [M]$ **in parallel do**
    $\boldsymbol{x}_{t,0}^{(m)} \leftarrow \boldsymbol{x}_t$
    for $h = 0, 1, \ldots, H-1$ **do**
      Sample $i_{t,h}^{(m)} \sim \text{Unif}([N])$
      $\boldsymbol{g}_{t,h}^{(m)} \leftarrow \nabla f(\boldsymbol{x}_{t,h}^{(m)}; \xi_{i_{t,h}^{(m)},m})$
      $(\boldsymbol{x}_{t,h+1}^{(m)}, \boldsymbol{s}_{\text{in},tH+h+1}^{(m)}) \leftarrow \mathcal{O}_{\text{in}}(\boldsymbol{x}_{t,h}^{(m)}, \boldsymbol{g}_{t,h}^{(m)}, \boldsymbol{s}_{\text{in},tH+h}^{(m)}; \eta_l)$
    end for
    $\Delta_t^{(m)} \leftarrow \boldsymbol{x}_t - \boldsymbol{x}_{t,H}^{(m)}$
    $\boldsymbol{u}_t^{(m)} \leftarrow \mathcal{C}(\Delta_t^{(m)} + \boldsymbol{e}_t^{(m)})$
    $\boldsymbol{e}_{t+1}^{(m)} \leftarrow \Delta_t^{(m)} + \boldsymbol{e}_t^{(m)} - \boldsymbol{u}_t^{(m)}$
    Send $\boldsymbol{u}_t^{(m)}$ to the server
  end for
  $\bar{\boldsymbol{u}}_t \leftarrow M^{-1} \sum_{m=1}^{M} \boldsymbol{u}_t^{(m)}$
  $(\boldsymbol{x}_{t+1}, \boldsymbol{s}_{\text{out},t+1}) \leftarrow \mathcal{O}_{\text{out}}(\boldsymbol{x}_t, \bar{\boldsymbol{u}}_t, \boldsymbol{s}_{\text{out},t}; \eta)$
end for
Output: $\bar{\boldsymbol{x}}_T = T^{-1} \sum_{t=0}^{T-1} \boldsymbol{x}_t$

For $M = H = 1$ and outer stepsize $\eta = 1$, (2) gives the single-node update (1) with stepsize η_l. In both settings, compression enters through the difference of successive buffers.

3.3 Stability and generalization

We study how these algorithms generalize to unseen data. For the time-averaged output $\bar{\boldsymbol{x}}_T$, the generalization gap and empirical optimization error bound expected excess population risk.

Definition 3.2. Let $\mathcal{A}(\mathcal{S}) \in \mathbb{R}^d$ be the output of a randomized algorithm trained on $\mathcal{S}$.

(i) The *generalization gap* compares population and empirical risk at the same output:

$$\epsilon_{\text{gen}} := \mathbb{E}_{\mathcal{S},\mathcal{A}}[F(\mathcal{A}(\mathcal{S})) - F_{\mathcal{S}}(\mathcal{A}(\mathcal{S}))].$$

(ii) The *optimization error* compares the output's empirical risk with its minimum:

$$\epsilon_{\text{opt}} := \mathbb{E}_{\mathcal{S},\mathcal{A}}[F_{\mathcal{S}}(\mathcal{A}(\mathcal{S})) - F_{\mathcal{S}}(\boldsymbol{x}_{\mathcal{S}}^{\star})].$$

To bound the generalization gap, we use algorithmic stability, which measures how much the output changes when one training example is replaced [6]. For nonnegative smooth losses, we use its on-average form to obtain expected generalization bounds without a global bound on gradient norms.

Definition 3.3 (ℓ_2 on-average model stability [33, 32])**.** Let $\tilde{\mathcal{S}}$ be an independent copy of $\mathcal{S}$. Write $\mathcal{S}^{(nm)}$ for $\mathcal{S}$ with only $\xi_{n,m}$ replaced by $\tilde{\xi}_{n,m}$, drawn from the same distribution $\mathcal{D}_m$. For $M = 1$, write $\mathcal{S}^{(n)}$ and average over the N replacements. The two runs use the same random seed, drawn independently of the data. The algorithm $\mathcal{A}$ has ℓ_2 on-average model stability at level ϵ if the mean squared distance between the paired outputs, averaged over all sample positions, satisfies

$$\frac{1}{MN} \sum_{m=1}^{M} \sum_{n=1}^{N} \mathbb{E}_{\mathcal{S},\tilde{\mathcal{S}},\mathcal{A}} \big\| \mathcal{A}(\mathcal{S}) - \mathcal{A}(\mathcal{S}^{(nm)}) \big\|_2^2 \le \epsilon^2.$$

Lemma 3.4 (Generalization from on-average stability [33, Theorem 2(b)])**.** *Suppose each loss $f(\cdot;\xi)$ is nonnegative and L-smooth. For any possibly randomized algorithm $\mathcal{A}$, use the replacements and shared seed in Definition 3.3. Then for every $\gamma > 0$,*

$$\epsilon_{\text{gen}} \le \frac{L}{\gamma} \mathbb{E} F_{\mathcal{S}}(\mathcal{A}(\mathcal{S})) + \frac{L+\gamma}{2MN} \sum_{m,n} \mathbb{E} \|\mathcal{A}(\mathcal{S}) - \mathcal{A}(\mathcal{S}^{(nm)})\|_2^2.$$

Remark 3.5 (From stability to excess risk). The two terms in Lemma 3.4 measure the empirical risk at the returned model and the mean squared change under replacement. Increasing γ lowers the weight on the empirical risk and raises the weight on stability. The first term is controlled by the optimization error and the optimal population risk, since $\mathbb{E}F_{\mathcal{S}}(\mathcal{A}(\mathcal{S})) = \epsilon_{\text{opt}} + \mathbb{E}F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}}) \leq \epsilon_{\text{opt}} + F(\boldsymbol{x}^\star)$. Once the gap is bounded, adding it to the optimization error controls the excess population risk: $\mathbb{E}F(\mathcal{A}(\mathcal{S})) - F(\boldsymbol{x}^\star) \leq \epsilon_{\text{opt}} + \epsilon_{\text{gen}}$. The proof uses sample replacements from each worker's own data distribution and covers the single-node case by setting $M = 1$ (Appendix E.1).

4 Main results

We first state the assumptions in Section 4.1 and analyze stability in Section 4.2, beginning with training without EF to examine how residuals accumulate across rounds. We then combine these stability results with the optimization bounds in Appendix C to derive the generalization guarantees in Section 4.3. Supporting lemmas and detailed proofs are given in Appendices D and E.

4.1 Assumptions

The analysis relies on the following assumptions on the sample loss, stochastic gradient noise, and differences between local and global empirical gradients evaluated at the same model.

Assumption 4.1 (Convexity)**.** For every sample ξ and all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, $f(\boldsymbol{x};\xi) \geq f(\boldsymbol{y};\xi) + \langle \nabla f(\boldsymbol{y};\xi),\, \boldsymbol{x} - \boldsymbol{y}\rangle$.

Assumption 4.2 (L-Smoothness)**.** For every sample ξ, $f(\cdot;\xi)$ is differentiable, and for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, $\|\nabla f(\boldsymbol{y};\xi) - \nabla f(\boldsymbol{x};\xi)\|_2 \leq L\|\boldsymbol{y} - \boldsymbol{x}\|_2$.

Assumption 4.3 (Nonnegative Loss)**.** For every sample ξ and every $\boldsymbol{x} \in \mathbb{R}^d$, $f(\boldsymbol{x};\xi) \geq 0$.

Assumption 4.4 (Bounded Variance)**.** Given the data and past draws, a fresh uniform draw from $\mathcal{S}_m$ gives an unbiased estimate of $\nabla F_{\mathcal{S}_m}(\boldsymbol{x})$. Almost surely in $\mathcal{S}$, for every worker m and $\boldsymbol{x} \in \mathbb{R}^d$, its conditional variance satisfies $N^{-1}\sum_{n=1}^{N} \|\nabla f(\boldsymbol{x};\xi_{n,m}) - \nabla F_{\mathcal{S}_m}(\boldsymbol{x})\|_2^2 \leq \sigma^2$.

Assumption 4.5 (Bounded Heterogeneity)**.** Almost surely in $\mathcal{S}$, for every $m \in [M]$ and $\boldsymbol{x} \in \mathbb{R}^d$, the local and global empirical gradients satisfy $\|\nabla F_{\mathcal{S}_m}(\boldsymbol{x}) - \nabla F_{\mathcal{S}}(\boldsymbol{x})\|_2^2 \leq \zeta^2$.

Remark 4.6 (Scope of the assumptions). Convexity and smoothness imply co-coercivity, which controls the difference between neighboring updates on shared samples. Both hold for ℓ_2-regularized logistic regression with bounded features. Contractivity controls buffer energy without convexity, but smoothness alone does not ensure nonexpansive updates on shared samples. The stability bounds retain gradient energy along the trajectory; nonnegativity then gives $\|\nabla f(\boldsymbol{x};\xi)\|^2 \leq 2Lf(\boldsymbol{x};\xi)$, allowing the optimization bounds to control this energy when deriving generalization. The variance bound controls within-worker sampling noise; the heterogeneity bound controls differences between local and global empirical gradients at the same model [35, 56]. These gradients can differ even when workers share a data distribution. These are bounds on gradient differences, not on gradient norms. For this model, a common quadratic regularizer cancels from both differences, although its gradient can grow with $\|\boldsymbol{x}\|$. When $M = 1$, the local and global empirical risks coincide, so the heterogeneity term vanishes. Sampling noise and compression remain in the single-node analysis.

4.2 Stability analysis

We now quantify how error feedback affects the sensitivity of the averaged model $\bar{\boldsymbol{x}}_T$ to a single-sample replacement. The bounds retain gradient energy along the training trajectory, which we later control through the optimization analysis. Stepsizes are positive and fixed within a run and may depend on its horizon. The symbol $\lesssim$ hides only universal numerical constants. Throughout the analysis, neighboring runs share a data-independent initial point, start from zero error buffers, and use the same sample indices and compression randomness. Empirical indices are drawn freshly and uniformly, independently across workers and local steps. Randomized compressors use fresh randomness at each compression call, conditional on their current input. Unless indicated otherwise, sums range over $m \in [M]$, $n \in [N]$, $0 \leq t < T$, and $0 \leq h < H$.

To isolate compression from local optimization, we compare no-EF and full-precision runs on the same dataset, with the same sample indices, local-update rules, and server averaging.

Proposition 4.7 (No-EF compression residual)**.** *Suppose Assumptions 4.1–4.2 and Definition 3.1 hold. Use SGD for both optimizers with $0 < \eta_l \le 1/(8LH)$ and $0 < \eta \le 1/(32L\eta_l H)$. Without EF, the server updates $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - (\eta/M)\sum_m \mathcal{C}_t^{(m)}(\Delta_t^{(m)})$, where $\Delta_t^{(m)} = \eta_l \sum_{h=0}^{H-1} \boldsymbol{g}_{t,h}^{(m)}$ is evaluated along this no-EF run. Let $\bar{\boldsymbol{x}}_T$ and $\bar{\boldsymbol{x}}_T^{\mathrm{FP}}$ be the pre-update averages of the no-EF and full-precision runs, with the same dataset, initialization, and sample indices. Define $\mathcal{R}_{\mathrm{noEF}}(T) := 4\mathbb{E}\|\bar{\boldsymbol{x}}_T - \bar{\boldsymbol{x}}_T^{\mathrm{FP}}\|_2^2$. Under the coupling specified above,*

$$\sqrt{\frac{1}{MN}\sum_{m,n}\mathbb{E}\|\bar{\boldsymbol{x}}_T(\mathcal{S}) - \bar{\boldsymbol{x}}_T(\mathcal{S}^{(nm)})\|_2^2} \le \sqrt{\frac{1}{MN}\sum_{m,n}\mathbb{E}\|\bar{\boldsymbol{x}}_T^{\mathrm{FP}}(\mathcal{S}) - \bar{\boldsymbol{x}}_T^{\mathrm{FP}}(\mathcal{S}^{(nm)})\|_2^2} + \sqrt{\mathcal{R}_{\mathrm{noEF}}(T)},$$

$$\mathcal{R}_{\mathrm{noEF}}(T) \le \frac{4(1-\alpha)\eta^2}{M}\left[\sum_t \frac{T-1-t}{T}\sqrt{\sum_m \mathbb{E}\|\Delta_t^{(m)}\|_2^2}\right]^2.$$

Remark 4.8 (Residual accumulation). Without EF, a round-t residual can affect the $T-1-t$ later iterates in the output, giving the weight $(T-1-t)/T$ in the bound. For admissible $\eta = \Theta(\sqrt{MH})$ and $\eta_l = c_l/(H\sqrt{T})$, contractivity gives a gradient-energy coefficient of $O((1-\alpha)MHT)$. Conditional centering reduces this to $O((1-\alpha)MH)$, since residual second moments add across rounds. Additional worker independence gives $O((1-\alpha)H)$ in Proposition C.8. The comparison gives an upper bound relative to full precision. For example, $\mathcal{C}(v) = qv$ with $0 < q \le 1$ simply rescales the outer stepsize to $q\eta$ without EF. With EF, co-coercivity absorbs the gradient–buffer cross term on shared samples, leaving the optimization analysis to control buffer energy along the trajectory.

Single-node EF-SGD isolates this feedback mechanism; DiLoCo-EF adds worker averaging and local drift. Both stability bounds use the compressor-induced buffer coefficient $b_\alpha := \frac{1-\alpha}{\left(1-\sqrt{1-\alpha}\right)^2}$, so that $b_1 = 0$ in the uncompressed case and $b_\alpha = \Theta(1/\alpha^2)$ under aggressive compression.

Theorem 4.9 (On-average Stability of Single-node EF)**.** *Suppose Assumptions 4.1–4.2 and Definition 3.1 hold. Run single-node EF-SGD with stepsize $0 < \eta \le 1/L$ under the sampling and coupling above. Its averaged output satisfies*

$$\frac{1}{N}\sum_n \mathbb{E}\left\|\bar{\boldsymbol{x}}_T(\mathcal{S}) - \bar{\boldsymbol{x}}_T(\mathcal{S}^{(n)})\right\|_2^2 \lesssim \left(1 + \frac{1}{\sqrt{L\eta T}}\right)\frac{\eta^2}{N^3}\sum_n\left(\sum_t \sqrt{\mathbb{E}\|\nabla f(\boldsymbol{x}_t;\xi_n)\|_2^2}\right)^2$$
$$+ \left[\underbrace{\left(1 + \frac{1}{\sqrt{L\eta T}}\right)\frac{\eta^2}{N}}_{\textit{SGD}} + \underbrace{b_\alpha\left(\sqrt{L\eta^3} + \frac{\eta}{\sqrt{T}}\right)^2}_{\textit{compression}}\right]\frac{1}{N}\sum_{n,t}\mathbb{E}\|\nabla f(\boldsymbol{x}_t;\xi_n)\|_2^2.$$

Remark 4.10 (From averaged SGD to EF-SGD). The first term accumulates the effect of drawing the replaced example, and the sampling term accounts for the squared one-step perturbations. The buffer term accounts for the coupled virtual steps and the return to the actual averaged output (Appendix E.5). At $\alpha = 1$, it vanishes and we recover the averaged-SGD structure of [33]. For smaller α, the larger b_α raises the buffer cost and tightens the optimization stepsize cap. Zero buffer initialization aligns the virtual and actual starting points and removes the initial buffer energy (Remark C.6). Theorem 4.13 controls the gradient energy using the empirical risk on the same trajectory.

Theorem 4.11 (On-average Stability of DiLoCo-EF)**.** *Suppose Assumptions 4.1–4.2 and Definition 3.1 hold. Run DiLoCo-EF under the sampling and coupling above, with SGD for both optimizers and stepsizes $0 < \eta_l \le 1/(8LH)$ and $0 < \eta \le 1/(32L\eta_l H)$. Its averaged output satisfies*

$$\frac{1}{MN}\sum_{m,n}\mathbb{E}\left\|\bar{\boldsymbol{x}}_T(\mathcal{S}) - \bar{\boldsymbol{x}}_T(\mathcal{S}^{(nm)})\right\|_2^2 \lesssim \frac{\eta^2\eta_l^2}{M^3N^3}\sum_{m,n}\left(\sum_{t,h}\sqrt{\mathbb{E}\|\nabla f(\boldsymbol{x}_{t,h}^{(m)};\xi_{n,m})\|_2^2}\right)^2$$
$$+ \left[\underbrace{\frac{\eta^2\eta_l^2}{M^2N}}_{\textit{DiLoCo}} + \underbrace{b_\alpha\,\eta^2\eta_l^2 H\left(L\eta\eta_l H + \frac{1}{T}\right)}_{\textit{compression}}\right]\frac{1}{MN}\sum_{m,n,t,h}\mathbb{E}\|\nabla f(\boldsymbol{x}_{t,h}^{(m)};\xi_{n,m})\|_2^2.$$

Remark 4.12 (Worker averaging and local drift). The coupling exposes a shared local update by omitting draws of the replaced example in the auxiliary paths (Appendix E.6). Each local step on the affected worker draws the replaced example with probability $1/N$, and this worker has weight $1/M$ at the server. These factors control the two replacement terms. Local drift enters the optimization bound through the $H-1$ term, controlled by $\eta_l = \Theta(1/(H\sqrt{T}))$. At $\alpha = 1$, the method reduces to full-precision local SGD with the same outer stepsize; $\eta = 1$ gives standard FedAvg. With $\alpha = 1$ and $H = 1$, it becomes mini-batch SGD with stepsize $\eta\eta_l$, averaging one gradient per worker.

4.3 Generalization analysis

We now combine stability with optimization to derive the generalization bounds. Theorems C.1 and C.3 bound the average empirical risk along the trajectory, which controls the risk of $\bar{\boldsymbol{x}}_T$ by convexity and the gradient energy by the self-bounding property. For DiLoCo, we first transfer the server risk bound to local iterates. With the risk prefactors defined below, we obtain

$$\frac{1}{T}\sum_{t<T}\mathbb{E}\|\boldsymbol{g}_t\|_2^2 \le 2L\mathcal{R}_{\mathrm{opt}}, \qquad \frac{1}{MHT}\sum_{m,t,h}\mathbb{E}\|\boldsymbol{g}_{t,h}^{(m)}\|_2^2 \le \frac{3L\widetilde{\mathcal{R}}_{\mathrm{opt}}}{\sqrt{T}}.$$

Substituting these estimates into the stability bounds, using (31) for DiLoCo, controls the second term in Lemma 3.4. The empirical risk of $\bar{\boldsymbol{x}}_T$ controls the first. Minimizing over γ balances the empirical-risk and sample-replacement contributions, giving the generalization bounds below.

Theorem 4.13 (Generalization of Single-node EF)**.** *Suppose Assumptions 4.1–4.4 and Definition 3.1 hold. For $1 \le T \le N^2$, run single-node EF-SGD with $\eta = c/\sqrt{T}$ and $0 < c \le [2L(1+\sqrt{1+2b_\alpha})]^{-1}$. The risk prefactor and generalization bound are*

$$\mathcal{R}_{\mathrm{opt}} := F(\boldsymbol{x}^\star) + \frac{2\mathbb{E}\|\boldsymbol{x}_0 - \boldsymbol{x}_S^\star\|_2^2}{c\sqrt{T}} + \frac{2c\sigma^2}{\sqrt{T}} + \frac{4Lb_\alpha c^2\sigma^2}{T},$$

$$\epsilon_{\mathrm{gen}} \le 32Lc\sqrt{T}\,\mathcal{R}_{\mathrm{opt}}\left[\underbrace{\frac{1}{N^2}+\frac{1}{NT}}_{\textit{full-precision}} + \underbrace{b_\alpha\left(\frac{Lc}{T^{3/2}}+\frac{1}{T^2}\right)}_{\textit{compression}}\right]^{1/2}.$$

Remark 4.14 (Finite-horizon compression). With fixed problem parameters and admissible c, the explicit buffer contribution is $O(\mathcal{R}_{\mathrm{opt}}T^{-1/4})$. Under interpolation, $F(\boldsymbol{x}^\star) = 0$ gives $\mathcal{R}_{\mathrm{opt}} = O(T^{-1/2})$, making this contribution $O(T^{-3/4})$; for positive $F(\boldsymbol{x}^\star)$ it is $O(T^{-1/4})$. Inside the square root, the buffer block is at most of the order of the sampling block once $T \gtrsim \max\{(b_\alpha LcN^2)^{2/3}, N\sqrt{b_\alpha}\}$ within $T \le N^2$ (Appendix E.7). Before this crossover, the buffer term can dominate while still decaying. Compression still affects the stepsize and risk prefactor.

Theorem 4.15 (Generalization of DiLoCo-EF)**.** *Suppose Assumptions 4.1–4.5 and Definition 3.1 hold. Run DiLoCo-EF with SGD for both optimizers for $T \ge 1$ with $HT \le (MN)^2$. Use constant stepsizes $\eta > 0$ and $\eta_l = c_l/(H\sqrt{T})$, $c_l > 0$, satisfying the conditions of Theorem C.3. The risk prefactor and generalization bound are*

$$\widetilde{\mathcal{R}}_{\mathrm{opt}} := \sqrt{T}\,F(\boldsymbol{x}^\star) + \frac{2\mathbb{E}\|\boldsymbol{x}_0 - \boldsymbol{x}_S^\star\|_2^2}{\eta c_l} + \frac{6\eta c_l\sigma^2}{MH} + \frac{9Lc_l^2}{H\sqrt{T}}(\sigma^2 + H\zeta^2)\big[2b_\alpha\eta^2 + (H-1)/H\big],$$

$$\epsilon_{\mathrm{gen}} \lesssim \widetilde{\mathcal{R}}_{\mathrm{opt}}\Bigg[\underbrace{\frac{L\eta c_l}{MN} + \frac{L\eta c_l}{M\sqrt{NHT}} + \frac{L^2\eta^2c_l^2\sqrt{T}}{(MN)^2} + \frac{L^2\eta^2c_l^2}{M^2NH\sqrt{T}}}_{\textit{full-precision}} + \underbrace{\frac{L\eta c_l\sqrt{b_\alpha}}{T^{3/4}}\sqrt{L\eta c_l + \frac{1}{\sqrt{T}}} + \frac{L^2\eta^2c_l^2b_\alpha}{T}\left(L\eta c_l + \frac{1}{\sqrt{T}}\right)}_{\textit{compression}}\Bigg].$$

Remark 4.16 (Distributed excess-risk rate). At fixed η, c_l, T, the schedule $\eta_l = c_l/(H\sqrt{T})$ keeps the cumulative step $\eta\eta_l HT = \eta c_l\sqrt{T}$ independent of H. Under interpolation, $\eta = \sqrt{MH}$ balances initialization and sampling in $\widetilde{\mathcal{R}}_{\mathrm{opt}}$. With fixed $c_l > 0$ and bounded problem quantities, the excess risk is $O((MHT)^{-1/2})$ for $\max\{MH(1+b_\alpha), b_\alpha^2(MH)^3\} \lesssim T \lesssim MN^2/H$, with compatible constants. The lower limit enforces the stepsize restrictions and controls the buffer terms, while the upper limit controls the sample-replacement terms (Appendix E.8). When $b_\alpha^2(MH)^3$ dominates, feasibility requires $N \gtrsim b_\alpha MH^2$. For $b_\alpha = O(1/M)$, $MH^3 \lesssim T \lesssim MN^2/H$ suffices.

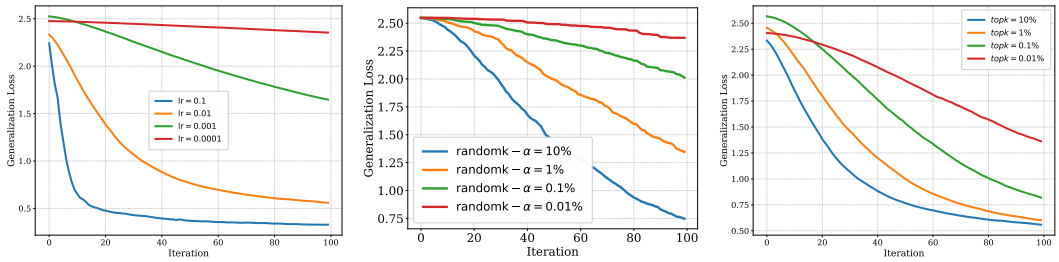
Remark 4.17 (Compression and stepsize). The $1/M$ variance reduction comes from averaging centered sampling noise. Per-worker contractivity allows buffer residuals to align, so the buffer term in $\widetilde{\mathcal{R}}_{\mathrm{opt}}$ need not share this gain. The optimization cap on $\eta\eta_l H$ is of order α/L for small α (Remark C.5). At fixed problem parameters and admissible stepsize constants, the explicit buffer contribution is $O(T^{-3/4})$ under interpolation and $O(T^{-1/4})$ otherwise. At fixed η and η_l, the optimization bound can retain nonvanishing buffer and drift terms. The bound extends distributed on-average analyses [59, 56] to nonlinear compression without assuming globally Lipschitz losses [24].

5 Experimental results

The bounds describe how optimization progress and compression enter generalization over a finite horizon. We examine these effects by varying stepsize and compression; DiLoCo additionally allows us to vary worker count and local steps (Figures 2 and 3).

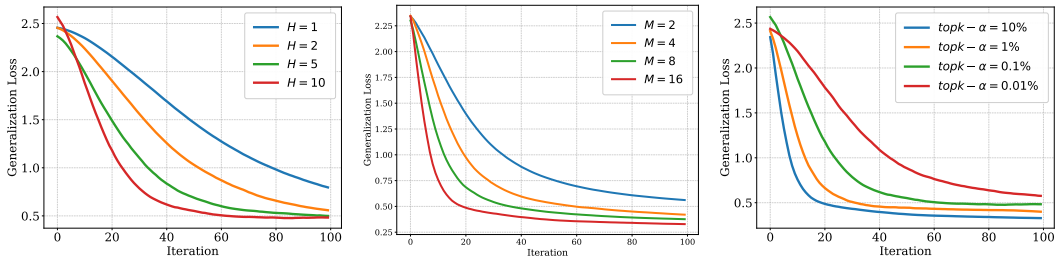
5.1 Empirical setup

We fit logistic regression [26] on MNIST [12], measuring the generalization gap as test loss minus training loss for both single-node EF-SGD and DiLoCo-EF (Definition 3.2). Figure 1 compares compressors at matched nominal value payloads. The compression sweeps vary strength within each method, including 1–16-bit quantization against a 32-bit baseline (Appendix A.2). Appendix A.1 lists the settings. We also examine whether the qualitative trends persist in non-convex DiLoCo-EF on DCLM-baseline 1.0 (Appendix A.3).[2]



(a) Stepsize η (b) Random-k (c) Top-k

Figure 2: Single-node EF-SGD on MNIST, varying stepsize and the retained fraction.



(a) Local steps H (b) Workers M (c) Top-k

Figure 3: DiLoCo-EF on MNIST, varying local steps, worker count, and the retained fraction.

[2] DCLM Team, DCLM-baseline 1.0 dataset release.

5.2 Finite-horizon behavior

At the same 100-iteration horizon, single-node EF-SGD reaches a gap of about 0.33 at $\eta = 0.1$, whereas the gap stays near 2.3 at $\eta = 0.0001$ (Figure 2a). The $1/(\eta T)$ initialization term in Theorem 4.13 represents the optimization cost associated with very small steps over a finite training horizon, while larger steps increase the sampling and buffer factors in the generalization bound.

Although the curves in Figure 1 stay close at the matched nominal payload, the lower-retention sweeps reveal differences between coordinate-selection rules. At 0.01%, the final gap is about 2.35 for Random-k and 1.36 for Top-k, while the gap trends downward for both (Figures 2b and 2c). The two compressors share a worst-case contraction factor α. At fixed input and retained count, Top-k minimizes discarded energy; Random-k introduces conditional variance (Appendix A.2).

The DiLoCo sweeps vary worker count and local computation separately. Increasing M from 2 to 16 lowers the final gap from about 0.56 to 0.33 (Figure 3b). In the bound, worker averaging attenuates sample-replacement perturbations and centered sampling noise, while compression residuals can align across workers (Remark 4.17). Increasing H from 1 to 5 also lowers the final gap, with little further change at $H = 10$ despite the extra local work per round (Figure 3a). Remark 4.16 explains how the analysis controls drift by decreasing η_l as H grows.

In DiLoCo, keeping fewer coordinates again raises the final gap (Figure 3c), while the quantization curves remain close across the tested precisions (Appendix A.2). The same contrast appears in the non-convex DCLM sweeps, even though the stepsize ordering changes (Appendix A.3). There too, Top-k reaches a smaller final gap than Random-k at the same retained count $k = 128$ (Figure 6f).

6 Conclusion

Our optimization-driven stability analysis shows how error feedback changes the compression terms in the generalization bound for convex losses. Under the analyzed stepsize schedules, the explicit buffer contribution decays, while compression still affects the optimization constants. The comparison without EF shows how uncompensated residuals can accumulate across rounds. In DiLoCo, the bounds distinguish the benefit of worker averaging from the drift due to local steps. The DiLoCo experiments on MNIST and DCLM show larger generalization gaps at lower coordinate retention, while quantization curves remain close across the tested precisions. The DCLM observations motivate extensions to non-convex losses, EF21 [40], and momentum-based distributed training.

Acknowledgments and Disclosure of Funding

This work was supported by the National Key R&D Projects (No. 2024YFC3307100), the National Natural Science Foundation of China (Nos. 62576364 and 62625602), the Guangdong Basic and Applied Basic Research Foundation (No. 2026B1515020071), and the Shenzhen Science and Technology Program (Nos. JCYJ20241202124430041, SYSRD20250529113401002, and ZDCY20250901113005006). The authors declare no competing interests.

References

[1] Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. QSGD: Communication-efficient SGD via gradient quantization and encoding. *Advances in neural information processing systems*, 30, 2017.

[2] Dan Alistarh, Torsten Hoefler, Mikael Johansson, Nikola Konstantinov, Sarit Khirirat, and Cédric Renggli. The convergence of sparsified gradient methods. *Advances in Neural Information Processing Systems*, 31, 2018.

[3] Peter L Bartlett and Shahar Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

[4] Aleksandr Beznosikov, Samuel Horváth, Peter Richtárik, and Mher Safaryan. On biased compression for distributed learning. *Journal of Machine Learning Research*, 24(276):1–50, 2023.

[5] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM (JACM)*, 36(4):929–965, 1989.

[6] Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of machine learning research*, 2(Mar):499–526, 2002.

[7] Olivier Bousquet, Yegor Klochkov, and Nikita Zhivotovskiy. Sharper bounds for uniformly stable algorithms. In *Conference on Learning Theory*, pages 610–626. PMLR, 2020.

[8] Zachary Charles, Gabriel Teston, Lucio Dery, Keith Rush, Nova Fallen, Zachary Garrett, Arthur Szlam, and Arthur Douillard. Communication-efficient language model training scales reliably and robustly: Scaling laws for DiLoCo. *arXiv preprint arXiv:2503.09799*, 2025.

[9] Congliang Chen, Li Shen, Haozhi Huang, and Wei Liu. Quantized Adam with error feedback. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 12(5):1–26, 2021.

[10] Congliang Chen, Li Shen, Wei Liu, and Zhi-Quan Luo. Efficient-Adam: Communication-efficient distributed Adam. *IEEE Transactions on Signal Processing*, 71:3257–3266, 2023.

[11] Aaron Defazio, Konstantin Mishchenko, Parameswaran Raman, Hao-Jun Michael Shi, and Lin Xiao. Smoothing DiLoCo with primal averaging for faster training of LLMs. *arXiv preprint arXiv:2512.17131*, 2025.

[12] Li Deng. The MNIST database of handwritten digit images for machine learning research [best of the web]. *IEEE signal processing magazine*, 29(6):141–142, 2012.

[13] Xiaoge Deng, Tao Sun, Shengwei Li, Dongsheng Li, and Xicheng Lu. Stability and generalization of asynchronous SGD: Sharper bounds beyond lipschitz and smoothness. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

[14] Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. GPT3.int8(): 8-bit matrix multiplication for transformers at scale. *Advances in neural information processing systems*, 35:30318–30332, 2022.

[15] Arthur Douillard, Yanislav Donchev, Keith Rush, Satyen Kale, Zachary Charles, Zachary Garrett, Gabriel Teston, Dave Lacey, Ross McIlroy, Jiajun Shen, et al. Streaming diloco with overlapping communication: Towards a distributed free lunch. *arXiv preprint arXiv:2501.18512*, 2025.

[16] Arthur Douillard, Qixuan Feng, Andrei A Rusu, Rachita Chhaparia, Yani Donchev, Adhiguna Kuncoro, Marc'Aurelio Ranzato, Arthur Szlam, and Jiajun Shen. DiLoCo: Distributed low-communication training of language models. *arXiv preprint arXiv:2311.08105*, 2023.

[17] Gintare Karolina Dziugaite and Daniel M Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv preprint arXiv:1703.11008*, 2017.

[18] Andre Elisseeff, Theodoros Evgeniou, and Massimiliano Pontil. Stability of randomized learning algorithms. *Journal of Machine Learning Research*, 6(3):55–79, 2005.

[19] Vitaly Feldman and Jan Vondrak. Generalization bounds for uniformly stable algorithms. *Advances in Neural Information Processing Systems*, 31, 2018.

[20] Vitaly Feldman and Jan Vondrak. High probability generalization bounds for uniformly stable algorithms with nearly optimal rate. In *Conference on learning theory*, pages 1270–1279. PMLR, 2019.

[21] Eduard Gorbunov, Dmitry Kovalev, Dmitry Makarenko, and Peter Richtárik. Linearly converging error compensated SGD. *Advances in Neural Information Processing Systems*, 33:20889–20900, 2020.

[22] Farzin Haddadpour, Mohammad Mahdi Kamani, Aryan Mokhtari, and Mehrdad Mahdavi. Federated learning with compression: Unified analysis and sharp guarantees. In *International Conference on Artificial Intelligence and Statistics*, pages 2350–2358. PMLR, 2021.

[23] Song Han, Huizi Mao, and William J Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding. *arXiv preprint arXiv:1510.00149*, 2015.

[24] Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International conference on machine learning*, pages 1225–1234. PMLR, 2016.

[25] Torsten Hoefler, Dan Alistarh, Tal Ben-Nun, Nikoli Dryden, and Alexandra Peste. Sparsity in deep learning: Pruning and growth for efficient inference and training in neural networks. *Journal of Machine Learning Research*, 22(241):1–124, 2021.

[26] David W Hosmer, Jr., Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*. John Wiley & Sons, 3 edition, 2013.

[27] Satyen Kale, Arthur Douillard, and Yanislav Donchev. Eager updates for overlapped communication and computation in DiLoCo. *arXiv preprint arXiv:2502.12996*, 2025.

[28] Sai Praneeth Karimireddy, Quentin Rebjock, Sebastian Stich, and Martin Jaggi. Error feedback fixes SignSGD and other gradient compression schemes. In *International conference on machine learning*, pages 3252–3261. PMLR, 2019.

[29] Jari Kolehmainen, Nikolay Blagoev, John Donaghy, Oğuzhan Ersoy, and Christopher Nies. NoLoCo: No-all-reduce low communication training method for large models. *arXiv preprint arXiv:2506.10911v1*, 2025.

[30] Nikolay Kutuzov, Makar Baderko, Stepan Kulibaba, Artem Dzhalilov, Daniel Bobrov, Maxim Mashtaler, and Alexander Gasnikov. AdLoCo: adaptive batching significantly improves communications efficiency and convergence for large language models. *arXiv preprint arXiv:2508.18182*, 2025.

[31] Batiste Le Bars, Aurélien Bellet, Marc Tommasi, Kevin Scaman, and Giovanni Neglia. Improved stability and generalization guarantees of the decentralized SGD algorithm. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 26215–26240. PMLR, 2024.

[32] Yunwen Lei, Tao Sun, and Mingrui Liu. Stability and generalization for minibatch SGD and local SGD. *arXiv preprint arXiv:2310.01139v1*, 2023.

[33] Yunwen Lei and Yiming Ying. Fine-grained analysis of stability and generalization for stochastic gradient descent. In *International Conference on Machine Learning*, pages 5809–5819. PMLR, 2020.

[34] Bingcong Li, Shuai Zheng, Parameswaran Raman, Anshumali Shrivastava, and Georgios B Giannakis. Contractive error feedback for gradient compression. *arXiv preprint arXiv:2312.08538*, 2023.

[35] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020.

[36] Xiaoyun Li and Ping Li. Analysis of error feedback in federated non-convex optimization with biased compression: Fast convergence and partial participation. In *International Conference on Machine Learning*, pages 19638–19688. PMLR, 2023.

[37] Yingqi Liu, Qinglun Li, Jie Tan, Yifan Shi, Li Shen, and Xiaochun Cao. Understanding the stability-based generalization of personalized federated learning. In *The Thirteenth International Conference on Learning Representations*, 2025.

[38] Ji Qi, WenPeng Zhu, Li Li, Ming Wu, YingJun Wu, Wu He, Xun Gao, Jason Zeng, and Michael Heinrich. DiLoCoX: A low-communication large-scale training framework for decentralized cluster. *arXiv preprint arXiv:2506.21263*, 2025.

[39] Xun Qian, Peter Richtárik, and Tong Zhang. Error compensated distributed SGD can be accelerated. *Advances in Neural Information Processing Systems*, 34:30401–30413, 2021.

[40] Peter Richtárik, Igor Sokolov, and Ilyas Fatkhullin. EF21: A new, simpler, theoretically better, and practically faster error feedback. *Advances in Neural Information Processing Systems*, 34:4384–4396, 2021.

[41] Amir Sarfi, Benjamin Thérien, Joel Lidin, and Eugene Belilovsky. Communication efficient LLM pre-training with SparseLoCo. *arXiv preprint arXiv:2508.15706v1*, 2025.

[42] 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 2014*, pages 1058–1062, 2014.

[43] Shaohuai Shi, Xiaowen Chu, Ka Chun Cheung, and Simon See. Understanding top-k sparsification in distributed deep learning. *arXiv preprint arXiv:1911.08772*, 2019.

[44] Sebastian U Stich, Jean-Baptiste Cordonnier, and Martin Jaggi. Sparsified SGD with memory. *Advances in neural information processing systems*, 31, 2018.

[45] Sebastian U Stich and Sai Praneeth Karimireddy. The error-feedback framework: Better rates for SGD with delayed gradients and compressed communication. *arXiv preprint arXiv:1909.05350*, 2019.

[46] Tao Sun, Dongsheng Li, and Bao Wang. Stability and generalization of decentralized stochastic gradient descent. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(11):9756–9764, 2021.

[47] Zhenyu Sun, Xiaochun Niu, and Ermin Wei. Understanding generalization of federated learning via stability: Heterogeneity matters. In *International conference on artificial intelligence and statistics*, pages 676–684. PMLR, 2024.

[48] Benjamin Thérien, Xiaolong Huang, Irina Rish, and Eugene Belilovsky. MuLoCo: Muon is a practical inner optimizer for DiLoCo. *arXiv preprint arXiv:2505.23725v1*, 2025.

[49] Thijs Vogels, Sai Praneeth Karimireddy, and Martin Jaggi. PowerSGD: Practical low-rank gradient compression for distributed optimization. *Advances in Neural Information Processing Systems*, 32, 2019.

[50] Jue Wang, Yucheng Lu, Binhang Yuan, Beidi Chen, Percy Liang, Christopher De Sa, Christopher Re, and Ce Zhang. CocktailSGD: Fine-tuning foundation models over 500Mbps networks. In *International Conference on Machine Learning*, pages 36058–36076. PMLR, 2023.

[51] Peng Wang, Li Shen, Zerui Tao, Shuaida He, and Dacheng Tao. Generalization analysis of stochastic weight averaging with general sampling. In *Forty-first International Conference on Machine Learning*, volume 235, pages 51442–51464. PMLR, 2024.

[52] Jiaxiang Wu, Weidong Huang, Junzhou Huang, and Tong Zhang. Error compensated quantized SGD and its applications to large-scale distributed optimization. In *International conference on machine learning*, pages 5325–5333. PMLR, 2018.

[53] Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. SmoothQuant: Accurate and efficient post-training quantization for large language models. In *International conference on machine learning*, pages 38087–38099. PMLR, 2023.

[54] An Xu and Heng Huang. Detached error feedback for distributed SGD with random sparsification. In *International conference on machine learning*, pages 24550–24575. PMLR, 2022.

[55] An Xu, Zhouyuan Huo, and Heng Huang. Step-ahead error feedback for distributed training with compressed gradient. *Proceedings of the AAAI conference on artificial intelligence*, 35(12):10478–10486, 2021.

[56] Haoxiang Ye, Tao Sun, and Qing Ling. Generalization error analysis for attack-free and byzantine-resilient decentralized learning with data heterogeneity. *arXiv preprint arXiv:2506.09438*, 2025.

[57] Mingchao Yu, Zhifeng Lin, Krishna Narra, Songze Li, Youjie Li, Nam Sung Kim, Alexander Schwing, Murali Annavaram, and Salman Avestimehr. GradiVeQ: Vector quantization for bandwidth-efficient gradient aggregation in distributed CNN training. *Advances in Neural Information Processing Systems*, 31, 2018.

[58] Dun Zeng, Zheshun Wu, Shiyu Liu, Yu Pan, Xiaoying Tang, and Zenglin Xu. Understanding generalization of federated learning: The trade-off between model stability and optimization. *arXiv preprint arXiv:2411.16303*, 2024.

[59] Shuang Zeng and Yunwen Lei. Stability and generalization analysis of decentralized SGD: Sharper bounds beyond lipschitzness and smoothness. In *Forty-second International Conference on Machine Learning*, volume 267, pages 74098–74132. PMLR, 2025.

[60] Shuai Zheng, Ziyue Huang, and James Kwok. Communication-efficient distributed blockwise momentum SGD with error-feedback. *Advances in Neural Information Processing Systems*, 32, 2019.

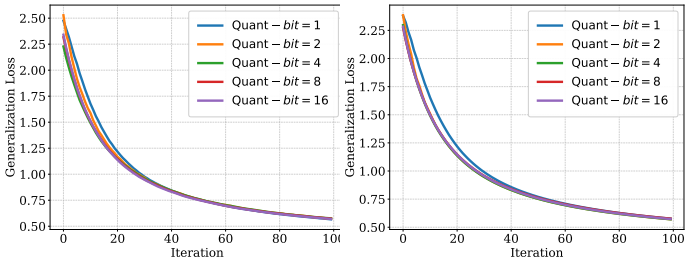
A Experimental details and additional results

A.1 Experimental settings

The MNIST sweeps span 100 iterations and use stepsizes $\{0.1, 0.01, 0.001, 0.0001\}$. Top-$k$ and Random-k retain $\{10\%, 1\%, 0.1\%, 0.01\%\}$ of coordinates, and the quantization sweeps use $\{16, 8, 4, 2, 1\}$ bits, with 32-bit floating point as the uncompressed baseline. In Figure 1, both sparsifiers keep 3.125% of coordinates at 32-bit precision, matching the 1-bit quantizer's nominal value payload before sparse indices and quantization metadata are counted.

On MNIST, DiLoCo uses worker counts $\{2, 4, 8, 16\}$ and local-step settings $\{1, 2, 5, 10\}$. The DCLM-baseline 1.0 sweeps in Figure 6 use stepsizes $\{0.01, 0.005, 0.001, 0.0005\}$, local intervals $\{15, 25, 35, 50\}$ (labeled "epoch"), and retained counts $k \in \{32, 64, 128, 256\}$.

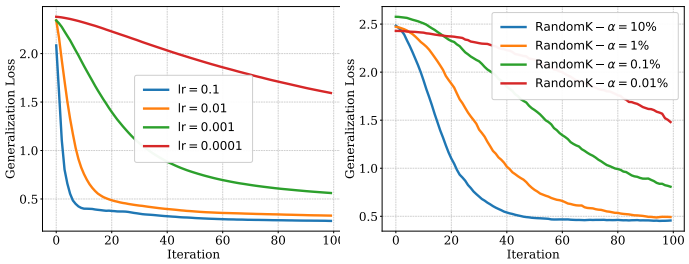
A.2 Convex experiments on MNIST



(a) Single-node EF-SGD (b) DiLoCo-EF

Figure 4: Generalization gaps on MNIST under different quantization bit-widths.

The 1-bit curve approaches the higher-precision curves in both MNIST settings (Figure 4). EF carries quantization residuals into later updates [42, 28]. The buffer bounds in Lemmas D.1 and D.7 depend on the quantizer through b_α; relating this quantity to bit-width requires the contraction factor α of the chosen quantization rule.

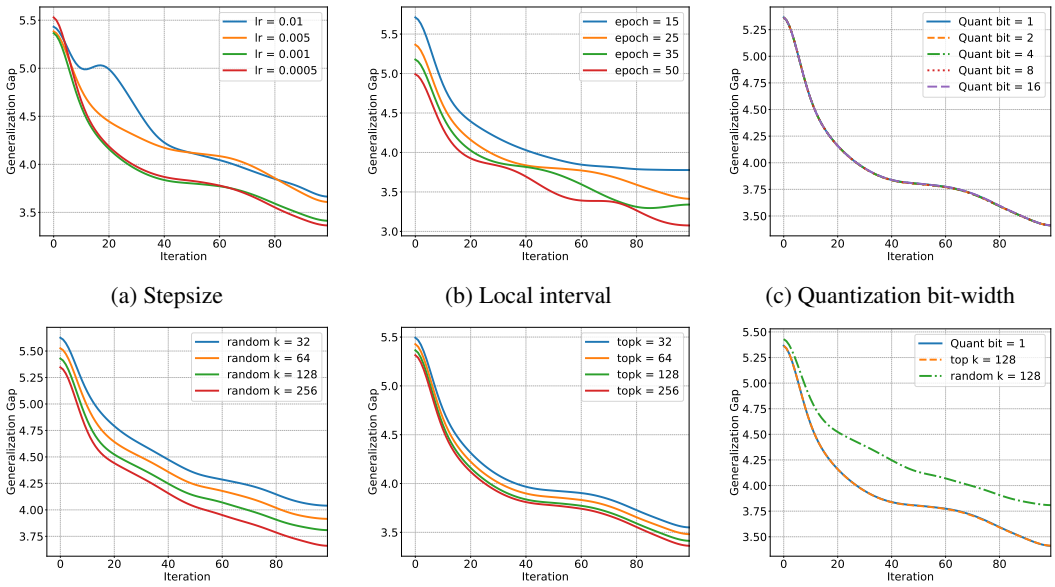


(a) Stepsize (b) Random-k

Figure 5: DiLoCo-EF on MNIST, varying stepsize and the Random-k retained fraction.

The DiLoCo stepsize sweep follows the same ordering as single-node EF-SGD on MNIST. In Figure 5a, the curves for 0.1 and 0.01 fall faster and finish lower than those for the smaller stepsizes. In the optimization bound, increasing either stepsize within the admissible range, with the other fixed, lowers the initialization term and raises the sampling term (Theorem C.3). Retaining fewer coordinates in the Random-k sweep raises the final gap, as in the Top-k sweep (Figures 5b and 3c).

The difference between the sparsifiers is also visible in their conditional variance. For a fixed input v, unscaled Random-k keeps each coordinate with probability $\alpha = k/d$, giving mean update αv. Summing the coordinate variances gives $\mathbb{E}_{\mathcal{C}}\|\mathcal{C}(v) - \alpha v\|^2 = \alpha(1-\alpha)\|v\|^2$. Top-$k$ with deterministic tie-breaking has zero conditional compressor variance and minimizes the discarded energy among masks of size k. The same worst-case contraction factor α can therefore correspond to different distortion and conditional variance at a given input.



(a) Stepsize (b) Local interval (c) Quantization bit-width

(d) Random-k (e) Top-k (f) Compressor comparison

Figure 6: DiLoCo-EF on DCLM-baseline 1.0. Sparsification varies the retained count k.

A.3 Non-convex experiments on DCLM

On DCLM-baseline 1.0, the final generalization gaps at stepsizes 0.0005 and 0.001 are smaller than those at 0.005 and 0.01 (Figure 6a), reversing the MNIST ordering. Longer local intervals generally lower the curves while changing the work per iteration (Figure 6b).

The contrast between quantization and sparsification recurs on DCLM. The quantization curves stay close down to 1 bit (Figure 6c), while retaining fewer coordinates raises the final gap for both sparsifiers (Figures 6d and 6e). At $k = 128$, Top-k nearly tracks 1-bit quantization and Random-k remains higher (Figure 6f).

B Further related work

Several DiLoCo variants address different communication bottlenecks. Streaming DiLoCo and Eager updates overlap synchronization with computation [15, 27]. MuLoCo couples Muon updates with compression and EF [48], while DiLoCoX uses adaptive compression alongside pipeline parallelism and delayed synchronization [38]. NoLoCo replaces all-reduce with pairwise partial averaging [29]; AdLoCo adapts batching and gradient accumulation to heterogeneous resources [30]. Generalized Primal Averaging recasts single-worker DiLoCo-style averaging as a single-loop method [11].

Beyond classical VC-dimension and Rademacher-complexity bounds [5, 3], PAC-Bayes bounds apply to over-parameterized networks [17]. For bounded losses, Feldman and Vondrák [19, 20] sharpen high-probability guarantees from uniform stability. Stability analyses also cover decentralized SGD [46], stochastic weight averaging [51], and personalized federated objectives [37].

C Supplementary theoretical results

The optimization results below bound the empirical risk averaged over the training trajectory, which controls the gradient energy in the stability bounds for nonnegative smooth losses. Appendix C.3 separates bias from centered compression noise without EF. Proofs are in Appendix E.

C.1 Optimization of single-node EF

Theorem C.1 (Optimization Error of Single-node EF)**.** *Suppose Assumptions 4.1–4.4 and Definition 3.1 hold, and the compressor uses fresh randomness at each iteration when it is randomized. Run single-node EF-SGD with $e_0 = 0$ and constant stepsize $0 < \eta \le [2L(1 + \sqrt{1 + 2b_\alpha})]^{-1}$. For every $T \ge 1$,*

$$\begin{aligned}\mathbb{E}[F_{\mathcal{S}}(\bar{x}_T) - F_{\mathcal{S}}(x^\star_{\mathcal{S}})] &\le \frac{1}{T}\sum_{t<T}\mathbb{E}[F_{\mathcal{S}}(x_t) - F_{\mathcal{S}}(x^\star_{\mathcal{S}})] \\ &\le \frac{2\mathbb{E}\|x_0 - x^\star_{\mathcal{S}}\|^2}{\eta T} + 2\eta\sigma^2 + 4Lb_\alpha\eta^2\sigma^2.\end{aligned}$$

Remark C.2 (Connection to classical compressed SGD analysis). Theorem C.1 retains the usual initialization and sampling-noise terms; compression contributes $4Lb_\alpha\eta^2\sigma^2$, with an extra power of η relative to the sampling term. The proof uses the virtual-iterate approach of classical EF analyses [45, 28] and weights the buffer energy by the stepsize to absorb its interaction with descent. For fixed $\alpha > 0$ and admissible $\eta = c/\sqrt{T}$ with fixed c, the compression term is $O(T^{-1})$, while initialization and sampling noise contribute $O(T^{-1/2})$. Compression still affects the allowable stepsize and finite-horizon bound; at $\alpha = 1$, the buffer vanishes and the classical averaged-SGD rate is recovered. In unbiased QSGD, compression noise enters through variance inflation [1].

C.2 Optimization of DiLoCo-EF

Theorem C.3 (Optimization Error of DiLoCo-EF)**.** *Assume that both optimizers are SGD, Assumptions 4.1–4.5 and Definition 3.1 hold, local sample indices are drawn freshly and independently across workers and inner steps, and randomized compressors use fresh randomness at every communication round. Start with $e_0^{(m)} = 0$ for every worker and hold both stepsizes constant within the run, satisfying $0 < \eta_l \le 1/(8LH)$ and $0 < \eta\eta_l H \le [32L(1 + \sqrt{b_\alpha})]^{-1}$. Then, for every $T \ge 1$,*

$$\begin{aligned}\mathbb{E}[F_{\mathcal{S}}(\bar{x}_T) - F_{\mathcal{S}}(x^\star_{\mathcal{S}})] &\le \frac{1}{T}\sum_{t<T}\mathbb{E}[F_{\mathcal{S}}(x_t) - F_{\mathcal{S}}(x^\star_{\mathcal{S}})] \\ &\le \frac{2\mathbb{E}\|x_0 - x^\star_{\mathcal{S}}\|^2}{\eta\eta_l HT} + \frac{6\eta\eta_l\sigma^2}{M} \\ &\quad + 18Lb_\alpha\eta^2\eta_l^2 H(\sigma^2 + H\zeta^2) + 9L\eta_l^2(H-1)(\sigma^2 + H\zeta^2).\end{aligned}$$

Remark C.4 (Distributed gain, compression memory, and heterogeneity). Worker averaging reduces the centered sampling variance by $1/M$. For the buffers, contractivity controls the energy at each worker and allows their residuals to align. The compression term vanishes at $\alpha = 1$, whereas the local-drift term vanishes at $H = 1$. Even at $H = 1$, heterogeneous worker gradients can still contribute to compression memory through ζ^2. For fixed M, H, α, η, c_l and admissible $\eta_l = c_l/(H\sqrt{T})$, both memory and drift decay as T^{-1}, compared with the $T^{-1/2}$ initialization and sampling terms. Thus heterogeneity leaves no stepsize-independent ζ^2/L floor in this bound.

Remark C.5 (Joint choice of compression and stepsizes). The stepsize conditions in Theorem C.3 imply those of Theorem 4.11. At fixed L, feasibility is preserved if neither $H\eta_l$ nor $\eta\eta_l H(1 + \sqrt{b_\alpha})$ increases. One can offset a larger H by decreasing η_l in inverse proportion. The upper limit on $\eta\eta_l H$ is $(1 - \sqrt{1-\alpha})/(32L) = \Theta(\alpha/L)$ as $\alpha \downarrow 0$, where α measures contraction as in Definition 3.1. Within this range, for fixed η_l, H, T, increasing η reduces the initialization term but raises sampling and compression terms linearly and quadratically, respectively. At fixed η, scaling η_l inversely with H preserves the effective stepsize but does not eliminate local drift or heterogeneity.

Remark C.6 (Initial buffers and the virtual starting point). Zero buffer initialization removes both the initial buffer energy and the shift of the virtual starting point. For square-integrable initial buffers, the proofs in Appendices E.2 and E.3 instead retain these two contributions. The initialization terms in Theorems C.1 and C.3 become, respectively,

$$\frac{2}{\eta T}\left[\mathbb{E}\|x_0 - e_0 - x^\star_{\mathcal{S}}\|^2 + \frac{2L\eta}{1 - \sqrt{1-\alpha}}\mathbb{E}\|e_0\|^2\right],$$

$$\frac{2}{\eta\eta_l HT}\left[\mathbb{E}\|x_0 - \eta\bar{e}_0 - x^\star_{\mathcal{S}}\|^2 + \frac{3L\eta^3\eta_l H}{M(1 - \sqrt{1-\alpha})}\sum_m \mathbb{E}\|e_0^{(m)}\|^2\right].$$

The other optimization terms and stepsize restrictions are unchanged. The shifted distances retain the cross term between the model and buffer initialization. The stability theorems use zero buffers. With common data-independent initial states and nonzero buffers, the coupled discrepancy still starts at zero, while the cumulative buffer bound retains the initial buffer energy.

C.3 Further analysis of the no-EF comparison

Proposition 4.7 bounds the accumulated compression residual along the actual no-EF trajectory. Under the same conditions, its weighted bound implies

$$\mathcal{R}_{\mathrm{noEF}}(T) \le \frac{2}{3}(1-\alpha)\eta^2\eta_l^2 H^2(T-1)(2T-1)\cdot\left(\frac{1}{MHT}\sum_{m,t,h}\mathbb{E}\|\boldsymbol{g}_{t,h}^{(m)}\|_2^2\right). \tag{3}$$

For admissible $\eta=\sqrt{MH}$ and $\eta_l = c_l/(H\sqrt{T})$, held fixed within the run, (3) becomes

$$\mathcal{R}_{\mathrm{noEF}}(T) \le \frac{2}{3}(1-\alpha)MHc_l^2\frac{(T-1)(2T-1)}{T}\left(\frac{1}{MHT}\sum_{m,t,h}\mathbb{E}\|\boldsymbol{g}_{t,h}^{(m)}\|_2^2\right). \tag{4}$$

Remark C.7 (EF changes how compression errors accumulate). Under admissible $\eta=\Theta(\sqrt{MH})$ and $\eta_l = c_l/(H\sqrt{T})$, (4) has coefficient $O((1-\alpha)MHT)$ multiplying the actual no-EF trajectory energy. The weighted bound can be sharper, but contractivity alone does not force this contribution to decay. EF changes this accumulation mechanism: the residual is retained in the buffer, whose cumulative energy is controlled by optimization and yields the explicit transient contribution in Theorem 4.15.

Proposition C.8 (No-EF with conditionally unbiased compression)**.** *Under the conditions of Proposition 4.7, suppose that, conditional on all information before the current compression, each worker's compressor is unbiased and the compression randomness is independent across workers. Then*

$$\mathcal{R}_{\mathrm{noEF}}(T) \le \frac{4(1-\alpha)\eta^2\eta_l^2H^2(T-1)^2}{MT}\left[\frac{1}{MHT}\sum_{m,t,h}\mathbb{E}\|g_{t,h}^{(m)}\|^2\right].$$

In particular, for admissible $\eta=\sqrt{MH}$ and $\eta_l = c_l/(H\sqrt{T})$,

$$\mathcal{R}_{\mathrm{noEF}}(T) \le 4(1-\alpha)Hc_l^2\left(\frac{T-1}{T}\right)^2\left[\frac{1}{MHT}\sum_{m,t,h}\mathbb{E}\|g_{t,h}^{(m)}\|^2\right]. \tag{5}$$

Conditional centering removes the extra T factor in (4), and worker independence removes the M factor. The coefficient is $O((1-\alpha)H)$, with further decay governed by the no-EF gradient-energy average. At $T=1$, both bounds vanish because the output is the common initial model.

Replacing $1-\alpha$ in Proposition C.8 by a finite uniform conditional relative-variance bound extends the proof to unbiased compressors outside Definition 3.1, provided the full-precision update remains nonexpansive: $0<\eta_l\le 2/L$ and $0<\eta\le[2/(L\eta_l)+H-1]/H$. Appendix E.4 gives the proofs and the transfer from trajectory discrepancy to empirical risk.

Bias alone also does not imply a persistent error floor. For $\mathcal{C}(v)=qv$, $0<q\le 1$, the contraction factor is $\alpha=1-(1-q)^2$, and the no-EF update is full-precision DiLoCo with outer stepsize $q\eta$.

D Technical lemmas

D.1 Single-node EF

Lemma D.1 (Bounded Accumulation of Error Buffer)**.** *Consider the EF recursion*

$$\boldsymbol{u}_t = \mathcal{C}(\Delta_t + \boldsymbol{e}_t), \qquad \boldsymbol{e}_{t+1} = \Delta_t + \boldsymbol{e}_t - \boldsymbol{u}_t, \qquad \boldsymbol{e}_0 = \boldsymbol{0},$$

where Δ_t is formed before the current compressor draw and b_α is defined in Sec. 4.2. Suppose Definition 3.1 holds and the compressor uses fresh randomness at each iteration when it is randomized. Then, for any $T \geq 1$,

$$\sum_{t=0}^{T-1} \mathbb{E}\|\boldsymbol{e}_t\|_2^2 \leq b_\alpha \sum_{t=0}^{T-1} \mathbb{E}\|\Delta_t\|_2^2.$$

In particular, $\Delta_t = \eta \boldsymbol{g}_t$ gives $\sum_{t<T} \mathbb{E}\|\boldsymbol{e}_t\|_2^2 \leq b_\alpha \eta^2 \sum_{t<T} \mathbb{E}\|\boldsymbol{g}_t\|_2^2$.

Proof. For $\alpha = 1$, compression is exact almost surely, so $e_t = 0$ for every t. For $0 < \alpha < 1$, condition on the information $\mathcal{F}_t^{\mathrm{comp}}$ available just before the current compression. Contractivity and Young's inequality give

$$\mathbb{E}[\|e_{t+1}\|^2 \mid \mathcal{F}_t^{\mathrm{comp}}] \leq (1-\alpha)\|e_t + \Delta_t\|^2 \leq \sqrt{1-\alpha}\,\|e_t\|^2 + \frac{1-\alpha}{1-\sqrt{1-\alpha}}\|\Delta_t\|^2.$$

Taking total expectation and summing, with $e_0 = 0$, gives

$$\begin{aligned}
\sum_{t<T} \mathbb{E}\|e_t\|^2 + \mathbb{E}\|e_T\|^2 &= \sum_{t<T} \mathbb{E}\|e_{t+1}\|^2 \\
&\leq \sqrt{1-\alpha} \sum_{t<T} \mathbb{E}\|e_t\|^2 + \frac{1-\alpha}{1-\sqrt{1-\alpha}} \sum_{t<T} \mathbb{E}\|\Delta_t\|^2, \\
(1-\sqrt{1-\alpha}) \sum_{t<T} \mathbb{E}\|e_t\|^2 &\leq \frac{1-\alpha}{1-\sqrt{1-\alpha}} \sum_{t<T} \mathbb{E}\|\Delta_t\|^2 - \mathbb{E}\|e_T\|^2 \\
&\leq \frac{1-\alpha}{1-\sqrt{1-\alpha}} \sum_{t<T} \mathbb{E}\|\Delta_t\|^2.
\end{aligned}$$

Dividing by $1-\sqrt{1-\alpha}$ yields b_α, and $\Delta_t = \eta g_t$ gives the SGD specialization. □

For any square-integrable initial buffer, the same one-step inequality gives

$$(1-\sqrt{1-\alpha}) \sum_{t<T} \mathbb{E}\|e_t\|^2 \leq \mathbb{E}\|e_0\|^2 - \mathbb{E}\|e_T\|^2 + \frac{1-\alpha}{1-\sqrt{1-\alpha}} \sum_{t<T} \mathbb{E}\|\Delta_t\|^2,$$

$$\sum_{t<T} \mathbb{E}\|e_t\|^2 \leq \frac{\mathbb{E}\|e_0\|^2}{1-\sqrt{1-\alpha}} + b_\alpha \sum_{t<T} \mathbb{E}\|\Delta_t\|^2.$$

At $\alpha = 1$, all buffers after the first compression vanish, and the same bound holds directly. Iterating the recursion bounds the direct initial contribution by $(\sqrt{1-\alpha})^t \mathbb{E}\|e_0\|^2$. Subsequent updates can also depend on initialization. The corresponding optimization bounds are given in Remark C.6.

Lemma D.2 (Virtual-iterate Recursion of Single-node EF)**.** *Consider single-node EF with SGD base update $\Delta_t = \eta \boldsymbol{g}_t$. Define $\widehat{\boldsymbol{x}}_t := \boldsymbol{x}_t - \boldsymbol{e}_t$. Then, for every $t \geq 0$, $\widehat{\boldsymbol{x}}_{t+1} = \widehat{\boldsymbol{x}}_t - \eta \boldsymbol{g}_t$.*

Proof. Using $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \boldsymbol{u}_t$ and $\boldsymbol{e}_{t+1} = \eta \boldsymbol{g}_t + \boldsymbol{e}_t - \boldsymbol{u}_t$, we obtain

$$\widehat{\boldsymbol{x}}_{t+1} = \boldsymbol{x}_{t+1} - \boldsymbol{e}_{t+1} = (\boldsymbol{x}_t - \boldsymbol{u}_t) - (\eta \boldsymbol{g}_t + \boldsymbol{e}_t - \boldsymbol{u}_t) = \widehat{\boldsymbol{x}}_t - \eta \boldsymbol{g}_t.$$

□

The virtual recursion uses the gradient at the actual iterate $\boldsymbol{x}_t$, so the analysis also requires control of the buffer $\boldsymbol{x}_t - \widehat{\boldsymbol{x}}_t = \boldsymbol{e}_t$.

D.2 DiLoCo-EF

Local updates introduce a gap between the points where gradients are evaluated and the server model. We control this drift before extending the buffer and virtual-iterate arguments to worker averages.

Lemma D.3 (Bounded Local Drift)**.** *Consider the local SGD inner loop in Algorithm 2, initialized from $\boldsymbol{x}_{t,0}^{(m)} = \boldsymbol{x}_t$ at each communication round t. Then the following pathwise bound holds:*

$$\frac{1}{M}\sum_{m=1}^{M}\sum_{h=0}^{H-1}\|\boldsymbol{x}_{t,h}^{(m)} - \boldsymbol{x}_t\|_2^2 \le \frac{\eta_l^2 H(H-1)}{2M}\sum_{m=1}^{M}\sum_{h=0}^{H-1}\|\boldsymbol{g}_{t,h}^{(m)}\|_2^2.$$

Proof. Fix t and m. Unrolling the local SGD recursion gives $\boldsymbol{x}_{t,h}^{(m)} - \boldsymbol{x}_t = -\eta_l \sum_{r<h} \boldsymbol{g}_{t,r}^{(m)}$ for $0 \le h < H$, where the empty sum is zero. Hence, by Cauchy–Schwarz,

$$\begin{aligned}
\sum_{h=0}^{H-1}\|\boldsymbol{x}_{t,h}^{(m)} - \boldsymbol{x}_t\|_2^2 &= \eta_l^2 \sum_{h=0}^{H-1}\left\|\sum_{r=0}^{h-1}\boldsymbol{g}_{t,r}^{(m)}\right\|_2^2 \\
&\le \eta_l^2 \sum_{h=0}^{H-1} h \sum_{r=0}^{h-1}\|\boldsymbol{g}_{t,r}^{(m)}\|_2^2 \\
&= \eta_l^2 \sum_{r=0}^{H-1}\left(\sum_{h=r+1}^{H-1} h\right)\|\boldsymbol{g}_{t,r}^{(m)}\|_2^2 \\
&\le \frac{\eta_l^2 H(H-1)}{2}\sum_{h=0}^{H-1}\|\boldsymbol{g}_{t,h}^{(m)}\|_2^2.
\end{aligned}$$

Averaging over m proves the claim. □

Lemma D.4 (Bounded Global Drift)**.** *Consider the local SGD inner loop in Algorithm 2, initialized from $\boldsymbol{x}_{t,0}^{(m)} = \boldsymbol{x}_t$ at each communication round t. Suppose Assumptions 4.2, 4.4, and 4.5 hold, and the local sample indices are drawn freshly and uniformly, independently across inner steps. If $0 < \eta_l \le 1/(8LH)$, then, for every communication round t,*

$$\frac{1}{MH}\sum_{m=1}^{M}\sum_{h=0}^{H-1}\mathbb{E}\|\boldsymbol{x}_{t,h}^{(m)} - \boldsymbol{x}_t\|_2^2 \le 2\eta_l^2(H-1)(\sigma^2 + H\zeta^2) + 2\eta_l^2 H(H-1)\mathbb{E}\|\nabla F_{\mathcal{S}}(\boldsymbol{x}_t)\|_2^2.$$

In particular, the bound is zero when $H = 1$.

Proof. Fix a round t. Conditional on the dataset and the past, write $g_{t,h}^{(m)} = \nabla F_{\mathcal{S}_m}(x_{t,h}^{(m)}) + \varepsilon_{t,h}^{(m)}$, where the noise is centered and its conditional second moment is at most σ^2. We separate the noise, the gradient at the server model, and the change in the local gradient. For $r < s$, conditional centering gives $\mathbb{E}\langle \varepsilon_{t,r}^{(m)}, \varepsilon_{t,s}^{(m)}\rangle = 0$. Unrolling the steps and using $\|a+b+c\|^2 \le 3\|a\|^2 + 3\|b\|^2 + 3\|c\|^2$ therefore gives

$$\begin{aligned}
\mathbb{E}\|x_{t,h}^{(m)} - x_t\|^2 &= \eta_l^2\mathbb{E}\left\|\sum_{r<h}\varepsilon_{t,r}^{(m)} + h\nabla F_{\mathcal{S}_m}(x_t) + \sum_{r<h}[\nabla F_{\mathcal{S}_m}(x_{t,r}^{(m)}) - \nabla F_{\mathcal{S}_m}(x_t)]\right\|^2 \\
&\le 3\eta_l^2 h\sigma^2 + 3\eta_l^2 h^2\mathbb{E}\|\nabla F_{\mathcal{S}_m}(x_t)\|^2 + 3\eta_l^2 L^2 h\sum_{r<h}\mathbb{E}\|x_{t,r}^{(m)} - x_t\|^2.
\end{aligned}$$

At x_t, decompose the worker gradient energy around the global gradient.

$$\begin{aligned}
\frac{1}{M}\sum_m \|\nabla F_{\mathcal{S}_m}(x_t)\|^2 &= \|\nabla F_{\mathcal{S}}(x_t)\|^2 + \frac{1}{M}\sum_m \|\nabla F_{\mathcal{S}_m}(x_t) - \nabla F_{\mathcal{S}}(x_t)\|^2 \\
&\le \|\nabla F_{\mathcal{S}}(x_t)\|^2 + \zeta^2.
\end{aligned}$$

Averaging the drift estimate over workers and local steps, we can absorb its last term into the left-hand side. Using $\sum_{h<H} h = H(H-1)/2$, $\sum_{h<H} h^2 = H(H-1)(2H-1)/6$, and $\sum_{h=r+1}^{H-1} h \le H(H-1)/2$, we obtain

$$\begin{aligned}\frac{1}{MH}\sum_{m,h}\mathbb{E}\|x_{t,h}^{(m)} - x_t\|^2 &\le \frac{3}{2}\eta_l^2(H-1)\sigma^2 + \frac{\eta_l^2(H-1)(2H-1)}{2}\left(\mathbb{E}\|\nabla F_{\mathcal{S}}(x_t)\|^2 + \zeta^2\right)\\ &\quad + \frac{3}{2}L^2\eta_l^2 H(H-1)\frac{1}{MH}\sum_{m,h}\mathbb{E}\|x_{t,h}^{(m)} - x_t\|^2\\ &\le \frac{3}{2}\eta_l^2(H-1)\sigma^2 + \eta_l^2 H(H-1)\left(\mathbb{E}\|\nabla F_{\mathcal{S}}(x_t)\|^2 + \zeta^2\right)\\ &\quad + \frac{3}{128}\frac{1}{MH}\sum_{m,h}\mathbb{E}\|x_{t,h}^{(m)} - x_t\|^2.\end{aligned}$$

The last coefficient follows from $L\eta_l H \le 1/8$. Since $128/125 < 4/3$, moving this term to the left yields

$$\begin{aligned}\frac{1}{MH}\sum_{m,h}\mathbb{E}\|x_{t,h}^{(m)} - x_t\|^2 &\le 2\eta_l^2(H-1)(\sigma^2 + H\zeta^2)\\ &\quad + 2\eta_l^2 H(H-1)\mathbb{E}\|\nabla F_{\mathcal{S}}(x_t)\|^2.\end{aligned}$$

□

Lemma D.5 (Local Empirical-Risk Transfer)**.** *Consider local SGD with fresh uniform samples, initialized at $\boldsymbol{x}_{t,0}^{(m)} = \boldsymbol{x}_t$ in each round. Under Assumptions 4.2 and 4.4, if $0 < \eta_l \le 1/L$, then, for every $T \ge 1$,*

$$\frac{1}{MHT}\sum_{m=1}^{M}\sum_{t=0}^{T-1}\sum_{h=0}^{H-1}\mathbb{E}F_{\mathcal{S}_m}(\boldsymbol{x}_{t,h}^{(m)}) \le \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}F_{\mathcal{S}}(\boldsymbol{x}_t) + \frac{L\eta_l^2(H-1)\sigma^2}{4}.$$

Proof. Condition on the dataset and the information $\mathcal{F}_{t,h}^{(m)}$ before the current local sample is drawn. Smoothness, the conditional variance bound, and $\eta_l \le 1/L$ give

$$\begin{aligned}\mathbb{E}[F_{\mathcal{S}_m}(\boldsymbol{x}_{t,h+1}^{(m)}) \mid \mathcal{F}_{t,h}^{(m)}] &\le F_{\mathcal{S}_m}(\boldsymbol{x}_{t,h}^{(m)}) - \eta_l\|\nabla F_{\mathcal{S}_m}(\boldsymbol{x}_{t,h}^{(m)})\|_2^2\\ &\quad + \frac{L\eta_l^2}{2}\left[\|\nabla F_{\mathcal{S}_m}(\boldsymbol{x}_{t,h}^{(m)})\|_2^2 + \sigma^2\right]\\ &\le F_{\mathcal{S}_m}(\boldsymbol{x}_{t,h}^{(m)}) + \frac{L\eta_l^2\sigma^2}{2},\\ \mathbb{E}F_{\mathcal{S}_m}(\boldsymbol{x}_{t,h}^{(m)}) &\le \mathbb{E}F_{\mathcal{S}_m}(\boldsymbol{x}_t) + \frac{L\eta_l^2 h\sigma^2}{2}.\end{aligned}$$

Averaging over workers, rounds, and local steps gives

$$\begin{aligned}\frac{1}{MHT}\sum_{m,t,h}\mathbb{E}F_{\mathcal{S}_m}(\boldsymbol{x}_{t,h}^{(m)}) &\le \frac{1}{MT}\sum_{m,t}\mathbb{E}F_{\mathcal{S}_m}(\boldsymbol{x}_t) + \frac{L\eta_l^2\sigma^2}{2H}\sum_{h=0}^{H-1} h\\ &= \frac{1}{T}\sum_{t<T}\mathbb{E}F_{\mathcal{S}}(\boldsymbol{x}_t) + \frac{L\eta_l^2(H-1)\sigma^2}{4}.\end{aligned}$$

□

Lemma D.6 (Local Gradient-Energy Bound)**.** *Under the conditions of Lemma D.5 and Assumption 4.3, for every $T \geq 1$,*

$$\frac{1}{MNHT} \sum_{m,n,t,h} \mathbb{E}\|\nabla f(\boldsymbol{x}_{t,h}^{(m)}; \xi_{n,m})\|_2^2 \leq \frac{2L}{T} \sum_{t<T} \mathbb{E} F_{\mathcal{S}}(\boldsymbol{x}_t) + \frac{L^2 \eta_l^2 (H-1)\sigma^2}{2}.$$

Proof. For nonnegative smooth losses, self-bounding controls each squared gradient norm by $2L$ times its loss. Averaging over the samples at each worker gives the local empirical risk, to which we apply Lemma D.5.

$$\begin{aligned}
\frac{1}{MNHT} \sum_{m,n,t,h} \mathbb{E}\|\nabla f(\boldsymbol{x}_{t,h}^{(m)}; \xi_{n,m})\|_2^2 &\leq \frac{2L}{MNHT} \sum_{m,n,t,h} \mathbb{E} f(\boldsymbol{x}_{t,h}^{(m)}; \xi_{n,m}) \\
&= \frac{2L}{MHT} \sum_{m,t,h} \mathbb{E} F_{\mathcal{S}_m}(\boldsymbol{x}_{t,h}^{(m)}) \\
&\leq \frac{2L}{T} \sum_{t<T} \mathbb{E} F_{\mathcal{S}}(\boldsymbol{x}_t) + \frac{L^2 \eta_l^2 (H-1)\sigma^2}{2}.
\end{aligned}$$

□

Lemma D.7 (Bounded Averaged Error Buffer)**.** *Under DiLoCo-EF in Algorithm 2, suppose each worker uses a compressor satisfying Definition 3.1, with fresh randomness at every communication round when it is randomized, and $\boldsymbol{e}_0^{(m)} = \mathbf{0}$ for all $m \in [M]$. Then, for any $T \geq 1$,*

$$\sum_{t=0}^{T-1} \mathbb{E}\|\overline{\boldsymbol{e}}_t\|_2^2 \leq \frac{b_\alpha}{M} \sum_{m=1}^{M} \sum_{t=0}^{T-1} \mathbb{E}\|\Delta_t^{(m)}\|_2^2.$$

For inner SGD, this gives

$$\sum_{t=0}^{T-1} \mathbb{E}\|\overline{\boldsymbol{e}}_t\|_2^2 \leq \frac{b_\alpha \eta_l^2 H}{M} \sum_{m=1}^{M} \sum_{t=0}^{T-1} \sum_{h=0}^{H-1} \mathbb{E}\|\boldsymbol{g}_{t,h}^{(m)}\|_2^2.$$

Proof. Jensen's inequality and the worker-wise buffer bound control the average buffer by local update energy. For SGD, expanding each local update and applying Cauchy–Schwarz gives

$$\begin{aligned}
\sum_{t<T} \mathbb{E}\|\overline{\boldsymbol{e}}_t\|_2^2 &= \sum_{t<T} \mathbb{E} \left\| \frac{1}{M} \sum_m \boldsymbol{e}_t^{(m)} \right\|_2^2 \\
&\leq \frac{1}{M} \sum_{m,t} \mathbb{E}\|\boldsymbol{e}_t^{(m)}\|_2^2 \leq \frac{b_\alpha}{M} \sum_{m,t} \mathbb{E}\|\Delta_t^{(m)}\|_2^2 \\
&= \frac{b_\alpha \eta_l^2}{M} \sum_{m,t} \mathbb{E} \left\| \sum_{h<H} \boldsymbol{g}_{t,h}^{(m)} \right\|_2^2 \\
&\leq \frac{b_\alpha \eta_l^2 H}{M} \sum_{m,t,h} \mathbb{E}\|\boldsymbol{g}_{t,h}^{(m)}\|_2^2.
\end{aligned}$$

□

For square-integrable worker buffers at initialization, the same argument gives

$$\sum_{t<T} \mathbb{E}\|\overline{e}_t\|^2 \leq \frac{\sum_m \mathbb{E}\|e_0^{(m)}\|^2}{M(1-\sqrt{1-\alpha})} + \frac{b_\alpha}{M} \sum_{m,t<T} \mathbb{E}\|\Delta_t^{(m)}\|^2.$$

Lemma D.8 (Virtual-iterate Recursion of DiLoCo-EF)**.** *For DiLoCo-EF with outer SGD and constant outer stepsize η, define $\overline{\Delta}_t = M^{-1}\sum_m \Delta_t^{(m)}$ and $\widehat{\boldsymbol{x}}_t = \boldsymbol{x}_t - \eta\overline{\boldsymbol{e}}_t$. Then, for every $t \geq 0$, $\widehat{\boldsymbol{x}}_{t+1} = \widehat{\boldsymbol{x}}_t - \eta\overline{\Delta}_t$. In particular, if the inner optimizer is SGD, then*

$$\widehat{\boldsymbol{x}}_{t+1} = \widehat{\boldsymbol{x}}_t - \frac{\eta\eta_l}{M}\sum_{m=1}^{M}\sum_{h=0}^{H-1}\boldsymbol{g}_{t,h}^{(m)}.$$

Proof. The worker identity $\boldsymbol{u}_t^{(m)} = \Delta_t^{(m)} + \boldsymbol{e}_t^{(m)} - \boldsymbol{e}_{t+1}^{(m)}$ and the outer update give the following chain after worker averaging:

$$\begin{aligned}\widehat{\boldsymbol{x}}_{t+1} &= \boldsymbol{x}_{t+1} - \eta\overline{\boldsymbol{e}}_{t+1} = \boldsymbol{x}_t - \eta\overline{\Delta}_t - \eta(\overline{\boldsymbol{e}}_t - \overline{\boldsymbol{e}}_{t+1}) - \eta\overline{\boldsymbol{e}}_{t+1}\\ &= \widehat{\boldsymbol{x}}_t - \eta\overline{\Delta}_t.\end{aligned}$$

For inner SGD, substitute $\overline{\Delta}_t = (\eta_l/M)\sum_{m,h}\boldsymbol{g}_{t,h}^{(m)}$. □

Lemma D.9 (Co-coercivity of Accumulated Local Updates)**.** *Fix a sequence of $H \geq 1$ gradient-step maps for convex, L-smooth losses, with a common stepsize $0 < \eta_l \leq 2/L$. Any selected maps may be replaced by the identity. Let P denote their composition and $\Delta(x) = x - P(x)$. Then, for all $x, y \in \mathbb{R}^d$,*

$$\langle x - y, \Delta(x) - \Delta(y)\rangle \geq \frac{2/(L\eta_l) + H - 1}{2H}\|\Delta(x) - \Delta(y)\|^2.$$

The same inequality holds for the arithmetic average of such update maps.

Proof. Starting from $z_0 = x$ and $z_0' = y$, apply the same sequence of maps to obtain z_h, z_h'. For an identity step, both increments below are zero. Co-coercivity at each gradient step gives

$$\begin{aligned}\|z_{h+1} - z_{h+1}'\|^2 &= \|z_h - z_h'\|^2 - 2\langle z_h - z_h', (z_h - z_{h+1}) - (z_h' - z_{h+1}')\rangle\\ &\quad + \|(z_h - z_{h+1}) - (z_h' - z_{h+1}')\|^2\\ &\leq \|z_h - z_h'\|^2 - \left(\frac{2}{L\eta_l} - 1\right)\|(z_h - z_{h+1}) - (z_h' - z_{h+1}')\|^2.\end{aligned}$$

Summing over the local steps and applying Cauchy–Schwarz gives

$$\begin{aligned}\|P(x) - P(y)\|^2 &\leq \|x - y\|^2 - \left(\frac{2}{L\eta_l} - 1\right)\sum_{h=0}^{H-1}\|(z_h - z_{h+1}) - (z_h' - z_{h+1}')\|^2\\ &\leq \|x - y\|^2 - \frac{2/(L\eta_l) - 1}{H}\left\|\sum_{h=0}^{H-1}[(z_h - z_{h+1}) - (z_h' - z_{h+1}')]\right\|^2\\ &= \|x - y\|^2 - \frac{2/(L\eta_l) - 1}{H}\|\Delta(x) - \Delta(y)\|^2.\end{aligned}$$

Expanding $P(x) - P(y) = (x - y) - (\Delta(x) - \Delta(y))$ gives

$$\begin{aligned}&2\langle x - y, \Delta(x) - \Delta(y)\rangle\\ &\quad= \|x - y\|^2 + \|\Delta(x) - \Delta(y)\|^2 - \|P(x) - P(y)\|^2\\ &\quad\geq \frac{2/(L\eta_l) + H - 1}{H}\|\Delta(x) - \Delta(y)\|^2.\end{aligned}$$

Dividing by two proves the first claim. For worker maps $\Delta^{(m)}$, linearity of the inner product and Jensen's inequality give

$$\begin{aligned}\left\langle x - y, \frac{1}{M}\sum_m[\Delta^{(m)}(x) - \Delta^{(m)}(y)]\right\rangle &\geq \frac{2/(L\eta_l) + H - 1}{2HM}\sum_m\|\Delta^{(m)}(x) - \Delta^{(m)}(y)\|^2\\ &\geq \frac{2/(L\eta_l) + H - 1}{2H}\left\|\frac{1}{M}\sum_m[\Delta^{(m)}(x) - \Delta^{(m)}(y)]\right\|^2.\end{aligned}$$ □

D.3 Further contractive compressor examples

For blockwise Top-k, let block b have d_b coordinates and retain $1 \le k_b \le d_b$. The retained coordinates contain at least a k_b/d_b fraction of block b's squared norm, hence

$$\|x - \mathcal{C}(x)\|^2 = \sum_b \|x^{(b)} - \mathrm{Top}_{k_b}(x^{(b)})\|^2 \le \sum_b \left(1 - \frac{k_b}{d_b}\right) \|x^{(b)}\|^2 \le \left(1 - \min_b \frac{k_b}{d_b}\right) \|x\|^2.$$

Thus $\alpha = \min_b k_b/d_b$ is valid, without requiring equal blocks. For an unbiased compressor Q satisfying $\mathbb{E}\|Q(x) - x\|^2 \le \omega\|x\|^2$, $\omega \ge 0$, the scaled implementation $\mathcal{C}(x) = Q(x)/(1+\omega)$ obeys

$$\mathbb{E}\|Q(x)\|^2 = \|x\|^2 + \mathbb{E}\|Q(x) - x\|^2 \le (1+\omega)\|x\|^2,$$

$$\mathbb{E}\|x - \mathcal{C}(x)\|^2 = \|x\|^2 - \frac{2\langle x, \mathbb{E}Q(x)\rangle}{1+\omega} + \frac{\mathbb{E}\|Q(x)\|^2}{(1+\omega)^2} \le \frac{\omega}{1+\omega}\|x\|^2.$$

The scaled compressor thus satisfies Definition 3.1 with $\alpha = 1/(1+\omega)$ and has mean $x/(1+\omega)$.

For a $p \times q$ matrix X, its exact best rank-r approximation X_r, with $1 \le r \le \min(p,q)$, satisfies

$$\|X - X_r\|_F^2 = \sum_{j=r+1}^{\min(p,q)} \sigma_j(X)^2 \le \left(1 - \frac{r}{\min(p,q)}\right) \sum_{j=1}^{\min(p,q)} \sigma_j(X)^2 = \left(1 - \frac{r}{\min(p,q)}\right) \|X\|_F^2,$$

where the singular values are in decreasing order. This gives $\alpha = r/\min(p,q)$ for exact truncation. An approximate factorization must meet its own Frobenius residual bound.

E Proofs of the main and supplementary results

E.1 Proof of Lemma 3.4

Proof. Replacing a sample within worker m preserves the dataset's law and makes $\mathcal{A}(\mathcal{S}^{(nm)})$ independent of the removed sample $\xi_{n,m} \sim \mathcal{D}_m$. This gives the first equality below without requiring identical worker distributions. Smoothness, Young's inequality, and the self-bound $\|\nabla f(x;\xi)\|^2 \le 2Lf(x;\xi)$ then give

$$\begin{aligned}
&\mathbb{E}[F(\mathcal{A}(\mathcal{S})) - F_{\mathcal{S}}(\mathcal{A}(\mathcal{S}))] \\
&= \frac{1}{MN} \sum_{m,n} \mathbb{E}\big[f(\mathcal{A}(\mathcal{S}^{(nm)}); \xi_{n,m}) - f(\mathcal{A}(\mathcal{S}); \xi_{n,m})\big] \\
&\le \frac{1}{MN} \sum_{m,n} \mathbb{E}\Big[\big\langle \nabla f(\mathcal{A}(\mathcal{S}); \xi_{n,m}), \mathcal{A}(\mathcal{S}^{(nm)}) - \mathcal{A}(\mathcal{S})\big\rangle + \frac{L}{2}\|\mathcal{A}(\mathcal{S}^{(nm)}) - \mathcal{A}(\mathcal{S})\|^2\Big] \\
&\le \frac{1}{2\gamma MN} \sum_{m,n} \mathbb{E}\|\nabla f(\mathcal{A}(\mathcal{S}); \xi_{n,m})\|^2 + \frac{L+\gamma}{2MN} \sum_{m,n} \mathbb{E}\|\mathcal{A}(\mathcal{S}^{(nm)}) - \mathcal{A}(\mathcal{S})\|^2 \\
&\le \frac{L}{\gamma}\mathbb{E}F_{\mathcal{S}}(\mathcal{A}(\mathcal{S})) + \frac{L+\gamma}{2MN} \sum_{m,n} \mathbb{E}\|\mathcal{A}(\mathcal{S}^{(nm)}) - \mathcal{A}(\mathcal{S})\|^2. \qquad (6)
\end{aligned}$$

Setting $M = 1$ gives the single-node bound. □

E.2 Proof of Theorem C.1

Proof. By Lemma D.2, the virtual iterate takes an SGD step using the stochastic gradient at the actual model x_t. Substituting $\widehat{x}_t = x_t - e_t$ in the conditional descent term separates the empirical-risk decrease from the gradient–buffer interaction. With $\mathcal{F}_t$ the data and history before sampling, convexity and Young's inequality give

$$\begin{aligned}
-2\eta\langle \nabla F_{\mathcal{S}}(x_t), \widehat{x}_t - x_{\mathcal{S}}^\star\rangle &= -2\eta\langle \nabla F_{\mathcal{S}}(x_t), x_t - x_{\mathcal{S}}^\star\rangle + 2\eta\langle \nabla F_{\mathcal{S}}(x_t), e_t\rangle \\
&\le -2\eta[F_{\mathcal{S}}(x_t) - F_{\mathcal{S}}(x_{\mathcal{S}}^\star)] + 2\eta\langle \nabla F_{\mathcal{S}}(x_t), e_t\rangle \\
&\le -2\eta[F_{\mathcal{S}}(x_t) - F_{\mathcal{S}}(x_{\mathcal{S}}^\star)] + \frac{\eta}{2L}\|\nabla F_{\mathcal{S}}(x_t)\|^2 + 2L\eta\|e_t\|^2.
\end{aligned}$$

To cancel $2L\eta\|e_t\|^2$, we weight the buffer energy by $2L\eta$ divided by its contraction margin $1 - \sqrt{1-\alpha}$. This gives the Lyapunov function

$$\Phi_t := \|\widehat{x}_t - x_{\mathcal{S}}^\star\|^2 + \frac{2L\eta}{1-\sqrt{1-\alpha}}\|e_t\|^2. \tag{7}$$

Conditioning the one-step estimate from the proof of Lemma D.1 on $\mathcal{F}_t$ gives

$$\begin{aligned}
&\frac{2L\eta}{1-\sqrt{1-\alpha}}\big(\mathbb{E}[\|e_{t+1}\|^2 \mid \mathcal{F}_t] - \|e_t\|^2\big) \\
&\quad\le -2L\eta\|e_t\|^2 + 2Lb_\alpha\eta^3\mathbb{E}[\|g_t\|^2 \mid \mathcal{F}_t].
\end{aligned}$$

Adding this inequality to the squared virtual-distance recursion cancels the buffer term. Using the conditional variance bound and $\|\nabla F_{\mathcal{S}}(x_t)\|^2 \le 2L[F_{\mathcal{S}}(x_t) - F_{\mathcal{S}}(x_{\mathcal{S}}^\star)]$, we obtain

$$\begin{aligned}
\mathbb{E}[\Phi_{t+1} \mid \mathcal{F}_t] - \Phi_t &= -2\eta\langle \nabla F_{\mathcal{S}}(x_t), \widehat{x}_t - x_{\mathcal{S}}^\star\rangle + \eta^2\mathbb{E}[\|g_t\|^2 \mid \mathcal{F}_t] \\
&\quad + \frac{2L\eta}{1-\sqrt{1-\alpha}}\big(\mathbb{E}[\|e_{t+1}\|^2 \mid \mathcal{F}_t] - \|e_t\|^2\big) \\
&\le -2\eta[F_{\mathcal{S}}(x_t) - F_{\mathcal{S}}(x_{\mathcal{S}}^\star)] + \frac{\eta}{2L}\|\nabla F_{\mathcal{S}}(x_t)\|^2 \\
&\quad + (\eta^2 + 2Lb_\alpha\eta^3)\big[\|\nabla F_{\mathcal{S}}(x_t)\|^2 + \sigma^2\big] \\
&\le -\eta(1 - 2L\eta - 4L^2b_\alpha\eta^2)[F_{\mathcal{S}}(x_t) - F_{\mathcal{S}}(x_{\mathcal{S}}^\star)] \\
&\quad + (\eta^2 + 2Lb_\alpha\eta^3)\sigma^2.
\end{aligned} \tag{8}$$

The stepsize restriction leaves at least $\eta/2$ of descent, since

$$2L\eta + 4L^2b_\alpha\eta^2 \le \frac{1}{1+\sqrt{1+2b_\alpha}} + \frac{b_\alpha}{(1+\sqrt{1+2b_\alpha})^2} = \frac{1}{2}.$$

Taking expectation and telescoping the Lyapunov differences over $t < T$, with $e_0 = 0$ and $\Phi_T \ge 0$, gives

$$\begin{aligned}
\frac{\eta}{2}\sum_{t<T}\mathbb{E}[F_{\mathcal{S}}(x_t) - F_{\mathcal{S}}(x_{\mathcal{S}}^\star)] &\le \mathbb{E}\Phi_0 - \mathbb{E}\Phi_T + (\eta^2 + 2Lb_\alpha\eta^3)\sigma^2T \\
&\le \mathbb{E}\|x_0 - x_{\mathcal{S}}^\star\|^2 + (\eta^2 + 2Lb_\alpha\eta^3)\sigma^2T.
\end{aligned} \tag{9}$$

Divide by $\eta T/2$ and use convexity to bound the risk of the averaged output.

$$\begin{aligned}
\mathbb{E}[F_{\mathcal{S}}(\bar{x}_T) - F_{\mathcal{S}}(x_{\mathcal{S}}^\star)] &\le \frac{1}{T}\sum_{t<T}\mathbb{E}[F_{\mathcal{S}}(x_t) - F_{\mathcal{S}}(x_{\mathcal{S}}^\star)] \\
&\le \frac{2\mathbb{E}\|x_0 - x_{\mathcal{S}}^\star\|^2}{\eta T} + 2\eta\sigma^2 + 4Lb_\alpha\eta^2\sigma^2.
\end{aligned} \tag{10}$$

□

E.3 Proof of Theorem C.3

Proof. Lemma D.8 gives $\widehat{x}_{t+1} = \widehat{x}_t - \eta\overline{\Delta}_t$ and $\widehat{x}_t = x_t - \eta\bar{e}_t$. Expanding the squared distance, we separate descent at the server model from its interaction with the buffer:

$$\begin{aligned}\|\widehat{x}_{t+1} - x_{\mathcal{S}}^\star\|^2 = \|\widehat{x}_t - x_{\mathcal{S}}^\star\|^2 - 2\eta\langle\overline{\Delta}_t, x_t - x_{\mathcal{S}}^\star\rangle \\ + 2\eta^2\langle\overline{\Delta}_t, \bar{e}_t\rangle + \eta^2\|\overline{\Delta}_t\|^2.\end{aligned}$$

For each local step, convexity and smoothness imply

$$\begin{aligned}\langle\nabla F_{\mathcal{S}_m}(x_{t,h}^{(m)}), x_t - x_{\mathcal{S}}^\star\rangle &= \langle\nabla F_{\mathcal{S}_m}(x_{t,h}^{(m)}), x_{t,h}^{(m)} - x_{\mathcal{S}}^\star\rangle \\ &\quad + \langle\nabla F_{\mathcal{S}_m}(x_{t,h}^{(m)}), x_t - x_{t,h}^{(m)}\rangle \\ &\geq F_{\mathcal{S}_m}(x_{t,h}^{(m)}) - F_{\mathcal{S}_m}(x_{\mathcal{S}}^\star) + \langle\nabla F_{\mathcal{S}_m}(x_{t,h}^{(m)}), x_t - x_{t,h}^{(m)}\rangle \\ &\geq F_{\mathcal{S}_m}(x_t) - F_{\mathcal{S}_m}(x_{\mathcal{S}}^\star) - \frac{L}{2}\|x_{t,h}^{(m)} - x_t\|^2.\end{aligned}$$

Fresh sampling and worker averaging therefore give

$$\mathbb{E}\langle\overline{\Delta}_t, x_t - x_{\mathcal{S}}^\star\rangle \geq \eta_l H\mathbb{E}[F_{\mathcal{S}}(x_t) - F_{\mathcal{S}}(x_{\mathcal{S}}^\star)] - \frac{L\eta_l H}{2MH}\sum_{m,h}\mathbb{E}\|x_{t,h}^{(m)} - x_t\|^2.$$

To control the two remaining terms in the distance recursion, decompose the local update into the round-start gradient, local drift, and sampling noise:

$$\overline{\Delta}_t = \eta_l H\nabla F_{\mathcal{S}}(x_t) + \frac{\eta_l}{M}\sum_{m,h}[\nabla F_{\mathcal{S}_m}(x_{t,h}^{(m)}) - \nabla F_{\mathcal{S}_m}(x_t)] + \frac{\eta_l}{M}\sum_{m,h}\varepsilon_{t,h}^{(m)}.$$

Here $\varepsilon_{t,h}^{(m)}$ is the centered noise from Lemma D.4. Conditioning before each fresh draw eliminates the cross terms in its second moment, giving

$$\mathbb{E}\left\|\frac{\eta_l}{M}\sum_{m,h}\varepsilon_{t,h}^{(m)}\right\|^2 = \frac{\eta_l^2}{M^2}\sum_{m,h}\mathbb{E}\|\varepsilon_{t,h}^{(m)}\|^2 \leq \frac{\eta_l^2 H\sigma^2}{M}.$$

The same centering makes its expected inner product with $\bar{e}_t$ zero, since the buffer is fixed before the local draws. Young's inequality and smoothness therefore bound the buffer interaction as follows:

$$\begin{aligned}2\eta^2\mathbb{E}\langle\overline{\Delta}_t, \bar{e}_t\rangle &\leq \frac{\eta\eta_l H}{2L}\mathbb{E}\|\nabla F_{\mathcal{S}}(x_t)\|^2 + 2L\eta^3\eta_l H\mathbb{E}\|\bar{e}_t\|^2 \\ &\quad + \frac{\eta}{L\eta_l H}\mathbb{E}\left\|\frac{\eta_l}{M}\sum_{m,h}[\nabla F_{\mathcal{S}_m}(x_{t,h}^{(m)}) - \nabla F_{\mathcal{S}_m}(x_t)]\right\|^2 + L\eta^3\eta_l H\mathbb{E}\|\bar{e}_t\|^2 \\ &\leq \frac{\eta\eta_l H}{2L}\mathbb{E}\|\nabla F_{\mathcal{S}}(x_t)\|^2 + \frac{L\eta\eta_l H}{MH}\sum_{m,h}\mathbb{E}\|x_{t,h}^{(m)} - x_t\|^2 + 3L\eta^3\eta_l H\mathbb{E}\|\bar{e}_t\|^2.\end{aligned}$$

The same decomposition, with $\|a+b+c\|^2 \leq 3\|a\|^2 + 3\|b\|^2 + 3\|c\|^2$, gives

$$\mathbb{E}\|\overline{\Delta}_t\|^2 \leq 3\eta_l^2 H^2\mathbb{E}\|\nabla F_{\mathcal{S}}(x_t)\|^2 + \frac{3\eta_l^2 H\sigma^2}{M} + \frac{3L^2\eta_l^2 H^2}{MH}\sum_{m,h}\mathbb{E}\|x_{t,h}^{(m)} - x_t\|^2,$$

$$\begin{aligned}\frac{1}{M}\sum_m\mathbb{E}\|\Delta_t^{(m)}\|^2 &\leq \frac{3\eta_l^2 H^2}{M}\sum_m\mathbb{E}\|\nabla F_{\mathcal{S}_m}(x_t)\|^2 + 3\eta_l^2 H\sigma^2 + \frac{3L^2\eta_l^2 H^2}{MH}\sum_{m,h}\mathbb{E}\|x_{t,h}^{(m)} - x_t\|^2 \\ &\leq 3\eta_l^2 H^2\mathbb{E}\|\nabla F_{\mathcal{S}}(x_t)\|^2 + 3\eta_l^2 H(\sigma^2 + H\zeta^2) + \frac{3L^2\eta_l^2 H^2}{MH}\sum_{m,h}\mathbb{E}\|x_{t,h}^{(m)} - x_t\|^2.\end{aligned}$$

The last line uses the empirical-gradient heterogeneity bound. We now choose the buffer weight to cancel the term $3L\eta^3\eta_l H\mathbb{E}\|\bar{e}_t\|^2$ in the cross-term estimate:

$$\Phi_t := \|\widehat{x}_t - x_{\mathcal{S}}^\star\|^2 + \frac{3L\eta^3\eta_l H}{1-\sqrt{1-\alpha}}\frac{1}{M}\sum_m \|e_t^{(m)}\|^2. \tag{11}$$

Indeed, $\mathbb{E}\|\bar{e}_t\|^2 \le M^{-1}\sum_m \mathbb{E}\|e_t^{(m)}\|^2$, and the one-step buffer bound gives

$$\begin{aligned}
&\frac{3L\eta^3\eta_l H}{M(1-\sqrt{1-\alpha})}\sum_m \mathbb{E}[\|e_{t+1}^{(m)}\|^2 - \|e_t^{(m)}\|^2] \\
&\quad\le -\frac{3L\eta^3\eta_l H}{M}\sum_m \mathbb{E}\|e_t^{(m)}\|^2 + \frac{3Lb_\alpha\eta^3\eta_l H}{M}\sum_m \mathbb{E}\|\Delta_t^{(m)}\|^2.
\end{aligned}$$

The decrease in worker buffer energy absorbs the term in $\mathbb{E}\|\bar{e}_t\|^2$. Adding the two energy differences and substituting the update second moments gives

$$\begin{aligned}
\mathbb{E}\Phi_{t+1} - \mathbb{E}\Phi_t \le\ & -2\eta\eta_l H\mathbb{E}[F_{\mathcal{S}}(x_t) - F_{\mathcal{S}}(x_{\mathcal{S}}^\star)] \\
&+ \eta\eta_l H\left[\frac{1}{2L} + 3\eta\eta_l H + 9Lb_\alpha\eta^2\eta_l^2 H^2\right]\mathbb{E}\|\nabla F_{\mathcal{S}}(x_t)\|^2 \\
&+ \eta\eta_l H\left[2L + 3L^2\eta\eta_l H + 9L^3 b_\alpha\eta^2\eta_l^2 H^2\right]\frac{1}{MH}\sum_{m,h}\mathbb{E}\|x_{t,h}^{(m)} - x_t\|^2 \\
&+ \frac{3\eta^2\eta_l^2 H\sigma^2}{M} + 9Lb_\alpha\eta^3\eta_l^3 H^2(\sigma^2 + H\zeta^2).
\end{aligned} \tag{12}$$

Using $\|\nabla F_{\mathcal{S}}(x_t)\|^2 \le 2L[F_{\mathcal{S}}(x_t) - F_{\mathcal{S}}(x_{\mathcal{S}}^\star)]$, the $1/(2L)$ term reduces the descent coefficient to $\eta\eta_l H$. After applying Lemma D.4, the remaining positive coefficients of the risk gap sum to less than half that amount.

$$\begin{aligned}
&6L\eta\eta_l H + 18b_\alpha(L\eta\eta_l H)^2 \\
&\quad + 4L^2\eta_l^2 H(H-1)[2 + 3L\eta\eta_l H + 9b_\alpha(L\eta\eta_l H)^2] \\
&\quad \le \frac{6}{32} + \frac{18}{32^2} + \frac{1}{16}\left(2 + \frac{3}{32} + \frac{9}{32^2}\right) < \frac{1}{2}.
\end{aligned}$$

The same stepsizes give $4[2 + 3L\eta\eta_l H + 9b_\alpha(L\eta\eta_l H)^2] < 9$ for the noise and heterogeneity part of the local drift. Hence the energy estimate reduces to

$$\begin{aligned}
\mathbb{E}\Phi_{t+1} - \mathbb{E}\Phi_t \le\ & -\frac{\eta\eta_l H}{2}\mathbb{E}[F_{\mathcal{S}}(x_t) - F_{\mathcal{S}}(x_{\mathcal{S}}^\star)] + \frac{3\eta^2\eta_l^2 H\sigma^2}{M} \\
&+ 9Lb_\alpha\eta^3\eta_l^3 H^2(\sigma^2 + H\zeta^2) + \frac{9}{2}L\eta\eta_l^3 H(H-1)(\sigma^2 + H\zeta^2).
\end{aligned} \tag{13}$$

Summing this inequality over $t < T$ telescopes the energy differences. The zero initial buffers give $\Phi_0 = \|x_0 - x_{\mathcal{S}}^\star\|^2$, while $\Phi_T \ge 0$, so

$$\begin{aligned}
\frac{1}{T}\sum_{t<T}\mathbb{E}[F_{\mathcal{S}}(x_t) - F_{\mathcal{S}}(x_{\mathcal{S}}^\star)] \le\ & \frac{2\mathbb{E}\|x_0 - x_{\mathcal{S}}^\star\|^2}{\eta\eta_l HT} + \frac{6\eta\eta_l\sigma^2}{M} \\
&+ 18Lb_\alpha\eta^2\eta_l^2 H(\sigma^2 + H\zeta^2) + 9L\eta_l^2(H-1)(\sigma^2 + H\zeta^2).
\end{aligned} \tag{14}$$

Convexity of the empirical risk gives $\mathbb{E}[F_{\mathcal{S}}(\bar{x}_T) - F_{\mathcal{S}}(x_{\mathcal{S}}^\star)] \le T^{-1}\sum_{t<T}\mathbb{E}[F_{\mathcal{S}}(x_t) - F_{\mathcal{S}}(x_{\mathcal{S}}^\star)]$. □

E.4 Proof of Proposition 4.7 and the refined no-EF bounds

Proof of Proposition 4.7. Fix a dataset and the local index sequences for round t. Write $\Delta_t^{(m)}(x)$ for the uncompressed local update from starting point x, and $\overline{\Delta}_t(x) = M^{-1}\sum_m \Delta_t^{(m)}(x)$. Lemma D.9 gives

$$\begin{aligned}
&\|(x - \eta\overline{\Delta}_t(x)) - (y - \eta\overline{\Delta}_t(y))\|^2 \\
&\quad = \|x-y\|^2 - 2\eta\langle x-y, \overline{\Delta}_t(x) - \overline{\Delta}_t(y)\rangle + \eta^2\|\overline{\Delta}_t(x) - \overline{\Delta}_t(y)\|^2 \\
&\quad \le \|x-y\|^2 - \eta\left(\frac{2/(L\eta_l) + H - 1}{H} - \eta\right)\|\overline{\Delta}_t(x) - \overline{\Delta}_t(y)\|^2 \\
&\quad \le \|x-y\|^2.
\end{aligned}$$

The stated stepsizes make the coefficient nonnegative, so the full-precision round is nonexpansive even when the outer stepsize exceeds one.

Use the same local index sequences in the two runs, with $x_0^{\mathrm{FP}} = x_0$ and $x_{t+1}^{\mathrm{FP}} = x_t^{\mathrm{FP}} - \eta\overline{\Delta}_t(x_t^{\mathrm{FP}})$. The no-EF compression residual is $\bar{r}_t = M^{-1}\sum_m[\Delta_t^{(m)} - \mathcal{C}_t^{(m)}(\Delta_t^{(m)})]$, evaluated on the actual run. Using the nonexpansive full-precision update and the common initialization,

$$\begin{aligned}
\|x_{t+1} - x_{t+1}^{\mathrm{FP}}\| &= \|(x_t - \eta\overline{\Delta}_t(x_t)) - (x_t^{\mathrm{FP}} - \eta\overline{\Delta}_t(x_t^{\mathrm{FP}})) + \eta\bar{r}_t\| \\
&\le \|x_t - x_t^{\mathrm{FP}}\| + \eta\|\bar{r}_t\| \\
&\le \eta\sum_{s=0}^{t}\|\bar{r}_s\|.
\end{aligned}$$

Averaging the pre-update iterates $x_0, \ldots, x_{T-1}$ gives

$$\begin{aligned}
\|\bar{x}_T - \bar{x}_T^{\mathrm{FP}}\| \le \frac{1}{T}\sum_{t=0}^{T-1}\|x_t - x_t^{\mathrm{FP}}\| &\le \frac{\eta}{T}\sum_{t=0}^{T-1}\sum_{s=0}^{t-1}\|\bar{r}_s\| \\
&= \eta\sum_{s=0}^{T-1}\frac{T-1-s}{T}\|\bar{r}_s\|.
\end{aligned} \tag{15}$$

The final residual has weight zero because it first affects x_T. For the neighboring runs, apply Minkowski over replacements and random draws.

$$\begin{aligned}
&\left[\frac{1}{MN}\sum_{m,n}\mathbb{E}\|\bar{x}_T(\mathcal{S}) - \bar{x}_T(\mathcal{S}^{(nm)})\|^2\right]^{1/2} \\
&\quad \le \left[\frac{1}{MN}\sum_{m,n}\mathbb{E}\|\bar{x}_T^{\mathrm{FP}}(\mathcal{S}) - \bar{x}_T^{\mathrm{FP}}(\mathcal{S}^{(nm)})\|^2\right]^{1/2} \\
&\qquad + \left[\frac{1}{MN}\sum_{m,n}\mathbb{E}\|\bar{x}_T(\mathcal{S}) - \bar{x}_T^{\mathrm{FP}}(\mathcal{S})\|^2\right]^{1/2} + \left[\frac{1}{MN}\sum_{m,n}\mathbb{E}\|\bar{x}_T(\mathcal{S}^{(nm)}) - \bar{x}_T^{\mathrm{FP}}(\mathcal{S}^{(nm)})\|^2\right]^{1/2} \\
&\quad = \left[\frac{1}{MN}\sum_{m,n}\mathbb{E}\|\bar{x}_T^{\mathrm{FP}}(\mathcal{S}) - \bar{x}_T^{\mathrm{FP}}(\mathcal{S}^{(nm)})\|^2\right]^{1/2} + \sqrt{\mathcal{R}_{\mathrm{noEF}}(T)}.
\end{aligned}$$

The two distances to full precision have the same second moment because the original and replaced datasets have the same law. To bound this distance, apply Minkowski to (15), then use Jensen across

workers and contractivity conditional on the local updates.

$$\begin{aligned}\mathcal{R}_{\mathrm{noEF}}(T) = 4\mathbb{E}\|\bar{x}_T - \bar{x}_T^{\mathrm{FP}}\|^2 &\le 4\eta^2 \left[\sum_{t<T} \frac{T-1-t}{T}\sqrt{\mathbb{E}\|\bar{r}_t\|^2}\right]^2 \\ &\le \frac{4\eta^2}{M}\left[\sum_{t<T} \frac{T-1-t}{T}\sqrt{\sum_m \mathbb{E}\|\Delta_t^{(m)} - \mathcal{C}_t^{(m)}(\Delta_t^{(m)})\|^2}\right]^2 \\ &\le \frac{4(1-\alpha)\eta^2}{M}\left[\sum_{t<T} \frac{T-1-t}{T}\sqrt{\sum_m \mathbb{E}\|\Delta_t^{(m)}\|^2}\right]^2 .\end{aligned} \tag{16}$$

This proves the weighted bound in Proposition 4.7. To express it through gradient energy as in (3), we use Cauchy–Schwarz while retaining the exact sum of squared time weights:

$$\begin{aligned}\mathcal{R}_{\mathrm{noEF}}(T) &\le \frac{4(1-\alpha)\eta^2}{M}\left[\sum_{t<T}\left(\frac{T-1-t}{T}\right)^2\right]\sum_{m,t}\mathbb{E}\|\Delta_t^{(m)}\|^2 \\ &= \frac{2(1-\alpha)\eta^2(T-1)(2T-1)}{3MT}\sum_{m,t}\mathbb{E}\left\|\eta_l\sum_{h<H} g_{t,h}^{(m)}\right\|^2 \\ &\le \frac{2(1-\alpha)\eta^2\eta_l^2 H(T-1)(2T-1)}{3MT}\sum_{m,t,h}\mathbb{E}\|g_{t,h}^{(m)}\|^2 \\ &= \frac{2}{3}(1-\alpha)\eta^2\eta_l^2 H^2(T-1)(2T-1)\left[\frac{1}{MHT}\sum_{m,t,h}\mathbb{E}\|g_{t,h}^{(m)}\|^2\right].\end{aligned}$$

Here $\sum_{t<T}(T-1-t)^2/T^2 = (T-1)(2T-1)/(6T)$. Substituting $\eta = \sqrt{MH}$ and $\eta_l = c_l/(H\sqrt{T})$ gives (4). The calculation allows the residuals to accumulate without cancellation across rounds.

□

We bound the conditional mean and the centered residual separately. Let $\mu_t = \mathbb{E}[\bar{r}_t \mid \mathcal{F}_t^{\mathrm{comp}}]$, where $\mathcal{F}_t^{\mathrm{comp}}$ contains the local sample draws and both trajectories' states before the current compression. The centered part has zero conditional inner product with the other terms, and nonexpansivity gives

$$\begin{aligned}&\mathbb{E}\|x_{t+1} - x_{t+1}^{\mathrm{FP}}\|^2 \\ &\quad= \mathbb{E}\|x_t - \eta\overline{\Delta}_t(x_t) - x_t^{\mathrm{FP}} + \eta\overline{\Delta}_t(x_t^{\mathrm{FP}}) + \eta\mu_t\|^2 + \eta^2\mathbb{E}\|\bar{r}_t - \mu_t\|^2 \\ &\quad\le \left(\sqrt{\mathbb{E}\|x_t - x_t^{\mathrm{FP}}\|^2} + \eta\sqrt{\mathbb{E}\|\mu_t\|^2}\right)^2 + \eta^2\mathbb{E}\|\bar{r}_t - \mu_t\|^2.\end{aligned}$$

Starting from zero discrepancy, induction and the triangle inequality in $\mathbb{R}^2$ give

$$\begin{aligned}\sqrt{\mathbb{E}\|x_{t+1} - x_{t+1}^{\mathrm{FP}}\|^2} &\le \eta\left[\left(\sum_{s\le t}\sqrt{\mathbb{E}\|\mu_s\|^2} + \sqrt{\sum_{s<t}\mathbb{E}\|\bar{r}_s - \mu_s\|^2}\right)^2 + \mathbb{E}\|\bar{r}_t - \mu_t\|^2\right]^{1/2} \\ &\le \eta\sum_{s\le t}\sqrt{\mathbb{E}\|\mu_s\|^2} + \eta\sqrt{\sum_{s\le t}\mathbb{E}\|\bar{r}_s - \mu_s\|^2}.\end{aligned}$$

Averaging the pre-update iterates and applying concavity of the square root to the centered-noise term yields

$$\begin{aligned}\sqrt{\mathcal{R}_{\mathrm{noEF}}(T)} &\le \frac{2}{T}\sum_{t<T}\sqrt{\mathbb{E}\|x_t - x_t^{\mathrm{FP}}\|^2} \\ &\le \frac{2\eta}{T}\sum_{t<T}\sum_{s<t}\sqrt{\mathbb{E}\|\mu_s\|^2} + \frac{2\eta}{T}\sum_{t<T}\sqrt{\sum_{s<t}\mathbb{E}\|\bar{r}_s - \mu_s\|^2} \\ &\le 2\eta\sum_{s<T}\frac{T-1-s}{T}\sqrt{\mathbb{E}\|\mu_s\|^2} + 2\eta\sqrt{\sum_{s<T}\frac{T-1-s}{T}\mathbb{E}\|\bar{r}_s - \mu_s\|^2}.\end{aligned} \tag{17}$$

The minimum of (16) and the square of (17) bounds $\mathcal{R}_{\mathrm{noEF}}(T)$. For conditionally centered residuals, $\mu_t = 0$ and the remaining time weights are at most one. With $\eta = \sqrt{MH}$ and $\eta_l = c_l/(H\sqrt{T})$, the coefficient of the no-EF gradient-energy average is then $O((1-\alpha)MH)$.

Proof of Proposition C.8. Conditional unbiasedness gives $\mathbb{E}[\bar{r}_t \mid \mathcal{F}_t^{\mathrm{comp}}] = 0$. Independence across workers then yields

$$\begin{aligned}\mathbb{E}[\|\bar{r}_t\|^2 \mid \mathcal{F}_t^{\mathrm{comp}}] &= \frac{1}{M^2}\sum_m \mathbb{E}\Big[\|\Delta_t^{(m)} - \mathcal{C}_t^{(m)}(\Delta_t^{(m)})\|^2 \mid \mathcal{F}_t^{\mathrm{comp}}\Big] \\ &\le \frac{1-\alpha}{M^2}\sum_m \|\Delta_t^{(m)}\|^2.\end{aligned}$$

The update gap is fixed before compression, so conditional centering removes the cross term.

$$\begin{aligned}\mathbb{E}\|x_{t+1} - x_{t+1}^{\mathrm{FP}}\|^2 &= \mathbb{E}\|x_t - \eta\overline{\Delta}_t(x_t) - x_t^{\mathrm{FP}} + \eta\overline{\Delta}_t(x_t^{\mathrm{FP}})\|^2 + \eta^2\mathbb{E}\|\bar{r}_t\|^2 \\ &\le \mathbb{E}\|x_t - x_t^{\mathrm{FP}}\|^2 + \frac{(1-\alpha)\eta^2}{M^2}\sum_m \mathbb{E}\|\Delta_t^{(m)}\|^2 \\ &\le \frac{(1-\alpha)\eta^2}{M^2}\sum_{s=0}^{t}\sum_m \mathbb{E}\|\Delta_s^{(m)}\|^2.\end{aligned} \tag{18}$$

The last line uses the common initialization. Since the output averages the pre-update iterates, only $T-1$ of its discrepancies can be nonzero. Minkowski and Cauchy–Schwarz over these terms yield

$$\begin{aligned}\mathcal{R}_{\mathrm{noEF}}(T) &\le \frac{4}{T^2}\left[\sum_{t=1}^{T-1}\sqrt{\mathbb{E}\|x_t - x_t^{\mathrm{FP}}\|^2}\right]^2 \\ &\le \frac{4(1-\alpha)\eta^2}{M^2T^2}\left[\sum_{t=1}^{T-1}\sqrt{\sum_{s<t}\sum_m \mathbb{E}\|\Delta_s^{(m)}\|^2}\right]^2 \\ &\le \frac{4(1-\alpha)\eta^2(T-1)}{M^2T^2}\sum_{s<T}(T-1-s)\sum_m \mathbb{E}\|\Delta_s^{(m)}\|^2 \\ &\le \frac{4(1-\alpha)\eta^2\eta_l^2H(T-1)^2}{M^2T^2}\sum_{m,t,h}\mathbb{E}\|g_{t,h}^{(m)}\|^2.\end{aligned} \tag{19}$$

Writing the last sum as MHT times the gradient-energy average gives the first bound in Proposition C.8. Substituting $\eta = \sqrt{MH}$ and $\eta_l = c_l/(H\sqrt{T})$ then gives (5). □

To convert the trajectory comparison to a loss bound, assume additionally nonnegative losses. Smoothness and the self-bound on the gradient give

$$\begin{aligned}f(x;\xi) &\le f(y;\xi) + \langle\nabla f(y;\xi), x-y\rangle + \frac{L}{2}\|x-y\|^2 \\ &\le f(y;\xi) + \sqrt{2Lf(y;\xi)}\|x-y\| + \frac{L}{2}\|x-y\|^2 = \left(\sqrt{f(y;\xi)} + \sqrt{L/2}\|x-y\|\right)^2.\end{aligned}$$

Taking square roots and applying Minkowski over the empirical samples and randomness gives

$$\begin{aligned}\sqrt{\mathbb{E}F_{\mathcal{S}}(\bar{x}_T)} &\le \sqrt{\mathbb{E}F_{\mathcal{S}}(\bar{x}_T^{\mathrm{FP}})} + \sqrt{L/2}\sqrt{\mathbb{E}\|\bar{x}_T - \bar{x}_T^{\mathrm{FP}}\|^2} \\ &= \sqrt{\mathbb{E}F_{\mathcal{S}}(\bar{x}_T^{\mathrm{FP}})} + \sqrt{L\mathcal{R}_{\mathrm{noEF}}(T)/8}.\end{aligned} \tag{20}$$

Combining Proposition 4.7, (20), and Lemma 3.4 gives the corresponding loss-stability conversion. Applying Theorem C.3 to the full-precision risk requires its variance, heterogeneity and stepsize conditions, with $\alpha = 1$.

E.5 Proof of Theorem 4.9

Throughout the stability proofs, the two runs differ in one training sample and share the initialization, sampled indices, and compression randomness specified in Definition 3.3. We use primes for the neighboring-data run and write $\delta_t := \widehat{x}_t - \widehat{x}'_t$. We first bound this virtual-iterate difference, then add back the buffers to control the actual averaged outputs.

Proof. Fix the replaced sample n. Lemma D.2 gives $\delta_{t+1} = \delta_t - \eta(g_t - g'_t)$, with $x_t - x'_t = \delta_t + e_t - e'_t$. When the current index is not n, co-coercivity and Young's inequality give

$$\begin{aligned}\|\delta_{t+1}\|^2 &= \|\delta_t\|^2 - 2\eta\langle x_t - x'_t, g_t - g'_t\rangle + 2\eta\langle e_t - e'_t, g_t - g'_t\rangle + \eta^2\|g_t - g'_t\|^2 \\ &\le \|\delta_t\|^2 - \frac{\eta(2 - L\eta)}{L}\|g_t - g'_t\|^2 + 2\eta\langle e_t - e'_t, g_t - g'_t\rangle \\ &\le \|\delta_t\|^2 + \frac{L\eta}{2 - L\eta}\|e_t - e'_t\|^2.\end{aligned}$$

The last step absorbs the cross term using $2 - L\eta > 0$. If the sampled index is n, the two gradients come from different losses, and we instead bound their contribution directly:

$$\begin{aligned}\|\delta_{t+1}\|^2 &= \|\delta_t\|^2 - 2\eta\langle \delta_t, g_t - g'_t\rangle + \eta^2\|g_t - g'_t\|^2 \\ &\le \|\delta_t\|^2 + 2\eta\|\delta_t\|\|g_t - g'_t\| + \eta^2\|g_t - g'_t\|^2.\end{aligned}$$

The fresh index hits n with probability $1/N$, conditional on the datasets and past draws. Replacement symmetry gives $(x'_t, \widetilde{\xi}_n)$ and (x_t, ξ_n) the same marginal law. Taking expectations and applying Cauchy–Schwarz gives

$$\begin{aligned}\mathbb{E}\|\delta_{t+1}\|^2 &\le \mathbb{E}\|\delta_t\|^2 + \frac{L\eta}{2 - L\eta}\mathbb{E}\|e_t - e'_t\|^2 \\ &\quad + \frac{2\eta}{N}\sqrt{\mathbb{E}\|\delta_t\|^2}\sqrt{\mathbb{E}\|\nabla f(x_t;\xi_n) - \nabla f(x'_t;\widetilde{\xi}_n)\|^2} \\ &\quad + \frac{\eta^2}{N}\mathbb{E}\|\nabla f(x_t;\xi_n) - \nabla f(x'_t;\widetilde{\xi}_n)\|^2 \\ &\le \mathbb{E}\|\delta_t\|^2 + \frac{L\eta}{2 - L\eta}\mathbb{E}\|e_t - e'_t\|^2 \\ &\quad + \frac{4\eta}{N}\sqrt{\mathbb{E}\|\delta_t\|^2}\sqrt{\mathbb{E}\|\nabla f(x_t;\xi_n)\|^2} + \frac{4\eta^2}{N}\mathbb{E}\|\nabla f(x_t;\xi_n)\|^2.\end{aligned} \tag{21}$$

Since $\delta_0 = 0$, summing this recursion leaves a term linear in the largest root-mean-square virtual gap. Young's inequality absorbs half of the corresponding squared gap:

$$\begin{aligned}\max_{s\le T}\mathbb{E}\|\delta_s\|^2 &\le \frac{4\eta}{N}\sqrt{\max_{s\le T}\mathbb{E}\|\delta_s\|^2}\sum_{t<T}\sqrt{\mathbb{E}\|\nabla f(x_t;\xi_n)\|^2} \\ &\quad + \frac{4\eta^2}{N}\sum_{t<T}\mathbb{E}\|\nabla f(x_t;\xi_n)\|^2 + \frac{L\eta}{2 - L\eta}\sum_{t<T}\mathbb{E}\|e_t - e'_t\|^2\end{aligned}$$

$$\leq \frac{1}{2}\max_{s\leq T}\mathbb{E}\|\delta_s\|^2 + \frac{8\eta^2}{N^2}\left(\sum_{t<T}\sqrt{\mathbb{E}\|\nabla f(x_t;\xi_n)\|^2}\right)^2$$
$$+ \frac{4\eta^2}{N}\sum_{t<T}\mathbb{E}\|\nabla f(x_t;\xi_n)\|^2 + \frac{L\eta}{2-L\eta}\sum_{t<T}\mathbb{E}\|e_t - e_t'\|^2$$

and hence

$$\max_{s\leq T}\mathbb{E}\|\delta_s\|^2 \leq \frac{16\eta^2}{N^2}\left(\sum_{t<T}\sqrt{\mathbb{E}\|\nabla f(x_t;\xi_n)\|^2}\right)^2$$
$$+ \frac{8\eta^2}{N}\sum_{t<T}\mathbb{E}\|\nabla f(x_t;\xi_n)\|^2 + \frac{2L\eta}{2-L\eta}\sum_{t<T}\mathbb{E}\|e_t - e_t'\|^2. \tag{22}$$

To return to the actual output, we use $x_t - x_t' = \delta_t + e_t - e_t'$. Lemma D.1 and replacement symmetry control the accumulated buffer difference, while Young's and Jensen's inequalities give, for any $\theta > 0$,

$$\sum_{t<T}\mathbb{E}\|e_t - e_t'\|^2 \leq 2\sum_{t<T}\mathbb{E}(\|e_t\|^2 + \|e_t'\|^2)$$
$$\leq 2b_\alpha\eta^2\sum_{t<T}\mathbb{E}(\|g_t\|^2 + \|g_t'\|^2) = 4b_\alpha\eta^2\sum_{t<T}\mathbb{E}\|g_t\|^2,$$
$$\mathbb{E}\|\bar{x}_T - \bar{x}_T'\|^2 = \mathbb{E}\left\|\frac{1}{T}\sum_{t<T}(\delta_t + e_t - e_t')\right\|^2$$
$$\leq (1+\theta)\max_{s\leq T}\mathbb{E}\|\delta_s\|^2 + \left(1+\frac{1}{\theta}\right)\frac{1}{T}\sum_{t<T}\mathbb{E}\|e_t - e_t'\|^2.$$

Average over n and take $\theta = \sqrt{(2-L\eta)/(L\eta T)}$. Up to the constant in (22), the buffer coefficients combine as

$$(1+\theta)\frac{L\eta^3}{2-L\eta} + \left(1+\frac{1}{\theta}\right)\frac{\eta^2}{T} = \left(\sqrt{\frac{L\eta^3}{2-L\eta}} + \frac{\eta}{\sqrt{T}}\right)^2.$$

For the generalization proof, the simpler choice $\theta = 1$ gives the explicit estimate, when $2 - L\eta \geq 1$,

$$\frac{1}{N}\sum_n \mathbb{E}\|\bar{x}_T(\mathcal{S}) - \bar{x}_T(\mathcal{S}^{(n)})\|^2$$
$$\leq \frac{32\eta^2}{N^3}\sum_n\left(\sum_{t<T}\sqrt{\mathbb{E}\|\nabla f(x_t;\xi_n)\|^2}\right)^2 + \frac{16\eta^2}{N^2}\sum_{n,t}\mathbb{E}\|\nabla f(x_t;\xi_n)\|^2$$
$$+ \left(\frac{16Lb_\alpha\eta^3}{2-L\eta} + \frac{8b_\alpha\eta^2}{T}\right)\sum_{t<T}\mathbb{E}\|g_t\|^2 \tag{23}$$
$$\leq \left[\frac{32\eta^2T^2}{N^2} + \frac{16\eta^2T}{N} + 16Lb_\alpha\eta^3T + 8b_\alpha\eta^2\right]\frac{1}{T}\sum_{t<T}\mathbb{E}\|g_t\|^2.$$

The last line uses Cauchy–Schwarz over time. By fresh uniform sampling, $\mathbb{E}\|g_t\|^2 = N^{-1}\sum_n \mathbb{E}\|\nabla f(x_t;\xi_n)\|^2$. □

E.6 Proof of Theorem 4.11

Proof. Fix the replaced sample n at worker m. A local round may draw this sample several times, so we first separate its contribution from the updates shared by the two runs. For each realization of the local indices, replace those steps by the identity map and keep all other steps. The resulting auxiliary map is shared by the two runs. Write its worker average as $\Delta_t^\circ(x)$, with evaluations Δ_t° and $\Delta_t^{\circ\prime}$ at x_t and x_t'. By Lemma D.9,

$$\langle x_t - x_t', \Delta_t^\circ - \Delta_t^{\circ\prime}\rangle \geq \frac{2/(L\eta_l) + H - 1}{2H}\|\Delta_t^\circ - \Delta_t^{\circ\prime}\|^2. \tag{24}$$

Each actual path and its auxiliary path start at the same server model. A shared step is nonexpansive, whereas a skipped step adds at most the local stepsize times the actual gradient norm to their discrepancy. Iterating this comparison bounds the difference between the actual and common updates. With $w_t = (\Delta_t^\circ - \Delta_t^{\circ\prime}) - (\overline{\Delta}_t - \overline{\Delta}_t')$, we obtain

$$\|w_t\| \leq \frac{\eta_l}{M}\sum_{h=0}^{H-1}\mathbf{1}\{i_{t,h}^{(m)} = n\}\Big(\|\nabla f(x_{t,h}^{(m)}; \xi_{n,m})\| + \|\nabla f(x_{t,h}^{\prime(m)}; \widetilde{\xi}_{n,m})\|\Big). \tag{25}$$

All gradients on the right are evaluated along actual trajectories. Using Lemma D.8, we have $x_t - x_t' = \delta_t + \eta(\bar{e}_t - \bar{e}_t')$ and the recursion $\delta_{t+1} = \delta_t - \eta(\Delta_t^\circ - \Delta_t^{\circ\prime}) + \eta w_t$. Expanding its common-update part, using (24), and completing the square gives

$$\begin{aligned}
\|\delta_t - \eta(\Delta_t^\circ - \Delta_t^{\circ\prime})\|^2 &= \|\delta_t\|^2 - 2\eta\langle x_t - x_t', \Delta_t^\circ - \Delta_t^{\circ\prime}\rangle \\
&\quad + 2\eta^2\langle \bar{e}_t - \bar{e}_t', \Delta_t^\circ - \Delta_t^{\circ\prime}\rangle + \eta^2\|\Delta_t^\circ - \Delta_t^{\circ\prime}\|^2 \\
&\leq \|\delta_t\|^2 - \eta\left(\frac{2/(L\eta_l) + H - 1}{H} - \eta\right)\|\Delta_t^\circ - \Delta_t^{\circ\prime}\|^2 \\
&\quad + 2\eta^2\langle \bar{e}_t - \bar{e}_t', \Delta_t^\circ - \Delta_t^{\circ\prime}\rangle \\
&\leq \|\delta_t\|^2 + \frac{\eta^3}{[2/(L\eta_l) + H - 1]/H - \eta}\|\bar{e}_t - \bar{e}_t'\|^2 \\
&\leq \|\delta_t\|^2 + L\eta^3\eta_l H\|\bar{e}_t - \bar{e}_t'\|^2.
\end{aligned}$$

The last line uses $[2/(L\eta_l) + H - 1]/H - \eta \geq 1/(L\eta_l H)$, which follows from $L\eta\eta_l H \leq 1/32$. Adding the hit perturbation and applying Young's inequality gives

$$\begin{aligned}
\|\delta_{t+1}\|^2 &\leq \left(\sqrt{\|\delta_t\|^2 + L\eta^3\eta_l H\|\bar{e}_t - \bar{e}_t'\|^2} + \eta\|w_t\|\right)^2 \\
&\leq \|\delta_t\|^2 + L\eta^3\eta_l H\|\bar{e}_t - \bar{e}_t'\|^2 + 2\eta\|\delta_t\|\|w_t\| \\
&\quad + 2\eta\sqrt{L\eta^3\eta_l H}\|\bar{e}_t - \bar{e}_t'\|\|w_t\| + \eta^2\|w_t\|^2 \\
&\leq \|\delta_t\|^2 + 2L\eta^3\eta_l H\|\bar{e}_t - \bar{e}_t'\|^2 + 2\eta\|\delta_t\|\|w_t\| + 2\eta^2\|w_t\|^2.
\end{aligned} \tag{26}$$

To bound w_t, condition before each local draw. Both current local iterates are fixed at that point, so the centered hit indicator times each gradient norm is a martingale difference. In the squared sum, conditioning before the later draw removes the cross terms.

$$\begin{aligned}
&\mathbb{E}\left[\left(\sum_h\left(\mathbf{1}\{i_{t,h}^{(m)} = n\} - \frac{1}{N}\right)\|\nabla f(x_{t,h}^{(m)}; \xi_{n,m})\|\right)^2\right] \\
&\quad = \frac{1}{N}\left(1 - \frac{1}{N}\right)\sum_h \mathbb{E}\|\nabla f(x_{t,h}^{(m)}; \xi_{n,m})\|^2.
\end{aligned}$$

Splitting off the conditional mean and using $(a+b)^2 \le 2a^2 + 2b^2$ gives

$$\begin{aligned}
&\mathbb{E}\left[\left(\sum_h \mathbf{1}\{i_{t,h}^{(m)} = n\}\|\nabla f(x_{t,h}^{(m)}; \xi_{n,m})\|\right)^2\right] \\
&\quad \le \frac{2}{N^2}\mathbb{E}\left[\left(\sum_h \|\nabla f(x_{t,h}^{(m)}; \xi_{n,m})\|\right)^2\right] \\
&\qquad + \frac{2}{N}\left(1 - \frac{1}{N}\right)\sum_h \mathbb{E}\|\nabla f(x_{t,h}^{(m)}; \xi_{n,m})\|^2 \\
&\quad \le 2\left(\frac{H}{N^2} + \frac{1}{N}\right)\sum_h \mathbb{E}\|\nabla f(x_{t,h}^{(m)}; \xi_{n,m})\|^2.
\end{aligned} \tag{27}$$

The two hit-weighted sums in (25) have the same law under sample replacement. For the mixed moment, conditioning before each draw replaces the hit indicator by $1/N$, since δ_t and the local gradient norms are already fixed. We then apply Cauchy–Schwarz to each run separately.

$$\begin{aligned}
\mathbb{E}\|w_t\|^2 &\le \frac{4\eta_l^2}{M^2}\mathbb{E}\left[\left(\sum_h \mathbf{1}\{i_{t,h}^{(m)} = n\}\|\nabla f(x_{t,h}^{(m)}; \xi_{n,m})\|\right)^2\right] \\
&\le \frac{8\eta_l^2}{M^2}\left(\frac{H}{N^2} + \frac{1}{N}\right)\sum_h \mathbb{E}\|\nabla f(x_{t,h}^{(m)}; \xi_{n,m})\|^2, \\
\mathbb{E}[\|\delta_t\|\|w_t\|] &\le \frac{\eta_l}{MN}\sum_h \mathbb{E}\Big[\|\delta_t\|\big(\|\nabla f(x_{t,h}^{(m)}; \xi_{n,m})\| + \|\nabla f(x_{t,h}^{\prime(m)}; \widetilde{\xi}_{n,m})\|\big)\Big] \\
&\le \frac{2\eta_l}{MN}\sqrt{\mathbb{E}\|\delta_t\|^2}\sum_h \sqrt{\mathbb{E}\|\nabla f(x_{t,h}^{(m)}; \xi_{n,m})\|^2}.
\end{aligned} \tag{28}$$

Substituting (28) into the recursion (26) and summing from $\delta_0 = 0$ gives the same maximum-gap argument as in the single-node case:

$$\begin{aligned}
\max_{s\le T}\mathbb{E}\|\delta_s\|^2 &\le \frac{4\eta\eta_l}{MN}\sqrt{\max_{s\le T}\mathbb{E}\|\delta_s\|^2}\sum_{t,h}\sqrt{\mathbb{E}\|\nabla f(x_{t,h}^{(m)}; \xi_{n,m})\|^2} \\
&\quad + \frac{16\eta^2\eta_l^2}{M^2}\left(\frac{H}{N^2} + \frac{1}{N}\right)\sum_{t,h}\mathbb{E}\|\nabla f(x_{t,h}^{(m)}; \xi_{n,m})\|^2 \\
&\quad + 2L\eta^3\eta_l H\sum_t \mathbb{E}\|\bar{e}_t - \bar{e}_t'\|^2 \\
&\le \frac{1}{2}\max_{s\le T}\mathbb{E}\|\delta_s\|^2 + \frac{8\eta^2\eta_l^2}{M^2N^2}\left(\sum_{t,h}\sqrt{\mathbb{E}\|\nabla f(x_{t,h}^{(m)}; \xi_{n,m})\|^2}\right)^2 \\
&\quad + \frac{16\eta^2\eta_l^2}{M^2}\left(\frac{H}{N^2} + \frac{1}{N}\right)\sum_{t,h}\mathbb{E}\|\nabla f(x_{t,h}^{(m)}; \xi_{n,m})\|^2 \\
&\quad + 2L\eta^3\eta_l H\sum_t \mathbb{E}\|\bar{e}_t - \bar{e}_t'\|^2
\end{aligned}$$

and hence

$$\max_{s\le T}\mathbb{E}\|\delta_s\|^2 \le \frac{16\eta^2\eta_l^2}{M^2N^2}\left(\sum_{t,h}\sqrt{\mathbb{E}\|\nabla f(x_{t,h}^{(m)};\xi_{n,m})\|^2}\right)^2 + \frac{32\eta^2\eta_l^2}{M^2}\left(\frac{H}{N^2}+\frac{1}{N}\right)\sum_{t,h}\mathbb{E}\|\nabla f(x_{t,h}^{(m)};\xi_{n,m})\|^2 + 4L\eta^3\eta_l H\sum_t \mathbb{E}\|\bar{e}_t-\bar{e}_t'\|^2. \tag{29}$$

For the averaged output, Jensen's inequality gives

$$\begin{aligned}\mathbb{E}\|\bar{x}_T-\bar{x}_T'\|^2 &= \mathbb{E}\left\|\frac{1}{T}\sum_{t<T}(\delta_t+\eta(\bar{e}_t-\bar{e}_t'))\right\|^2\\ &\le \frac{2}{T}\sum_{t<T}\mathbb{E}\|\delta_t\|^2+\frac{2\eta^2}{T}\sum_{t<T}\mathbb{E}\|\bar{e}_t-\bar{e}_t'\|^2\\ &\le 2\max_{s\le T}\mathbb{E}\|\delta_s\|^2+\frac{2\eta^2}{T}\sum_{t<T}\mathbb{E}\|\bar{e}_t-\bar{e}_t'\|^2.\end{aligned}$$

The remaining buffer term is controlled by Lemma D.7. Averaging over replacements equates the two runs' marginal buffer moments; fresh empirical sampling then expresses the update energy through the sample gradients:

$$\begin{aligned}\frac{1}{MN}\sum_{n,m}\sum_{t<T}\mathbb{E}\|\bar{e}_t(\mathcal{S})-\bar{e}_t(\mathcal{S}^{(nm)})\|^2 &\le 4\sum_{t<T}\mathbb{E}\|\bar{e}_t\|^2\\ &\le \frac{4b_\alpha\eta_l^2 H}{M}\sum_{m,t,h}\mathbb{E}\|g_{t,h}^{(m)}\|^2\\ &= \frac{4b_\alpha\eta_l^2 H}{MN}\sum_{m,n,t,h}\mathbb{E}\|\nabla f(x_{t,h}^{(m)};\xi_{n,m})\|^2.\end{aligned} \tag{30}$$

To recover the main-text form, retain the square-root sums before applying Cauchy–Schwarz in (27). Minkowski's inequality gives

$$\begin{aligned}&\sum_t\mathbb{E}\left[\left(\sum_h\|\nabla f(x_{t,h}^{(m)};\xi_{n,m})\|\right)^2\right]\\ &\le\sum_t\left[\sum_h\sqrt{\mathbb{E}\|\nabla f(x_{t,h}^{(m)};\xi_{n,m})\|^2}\right]^2\\ &\le\left[\sum_{t,h}\sqrt{\mathbb{E}\|\nabla f(x_{t,h}^{(m)};\xi_{n,m})\|^2}\right]^2.\end{aligned}$$

Using this bound in the same absorption that gives (29) replaces its first coefficient by $(16+32)\eta^2\eta_l^2/(M^2N^2)$ and leaves $32\eta^2\eta_l^2/(M^2N)$ multiplying the gradient-energy sum. The averaged-output and buffer estimates then yield

$$\begin{aligned}&\frac{1}{MN}\sum_{m,n}\mathbb{E}\|\bar{x}_T(\mathcal{S})-\bar{x}_T(\mathcal{S}^{(nm)})\|^2\\ &\le\frac{96\eta^2\eta_l^2}{M^3N^3}\sum_{m,n}\left[\sum_{t,h}\sqrt{\mathbb{E}\|\nabla f(x_{t,h}^{(m)};\xi_{n,m})\|^2}\right]^2\\ &\quad+\left[\frac{64\eta^2\eta_l^2}{M^2N}+32b_\alpha L\eta^3\eta_l^3H^2+\frac{8b_\alpha\eta^2\eta_l^2H}{T}\right]\frac{1}{MN}\sum_{m,n,t,h}\mathbb{E}\|\nabla f(x_{t,h}^{(m)};\xi_{n,m})\|^2.\end{aligned}$$

This proves the stated bound up to its universal constant. For the generalization argument, it is useful to express the whole estimate in terms of the average gradient energy. Returning to (29), Cauchy–Schwarz gives

$$\frac{1}{MN}\sum_{n,m}\left(\sum_{t,h}\sqrt{\mathbb{E}\|\nabla f(x_{t,h}^{(m)};\xi_{n,m})\|^2}\right)^2 \le \frac{HT}{MN}\sum_{m,n,t,h}\mathbb{E}\|\nabla f(x_{t,h}^{(m)};\xi_{n,m})\|^2.$$

Substituting these estimates into the averaged-output bound yields

$$\begin{aligned}
&\frac{1}{MN}\sum_{m,n}\mathbb{E}\|\bar{x}_T(\mathcal{S})-\bar{x}_T(\mathcal{S}^{(nm)})\|^2 \\
&\le \eta^2\eta_l^2H^2\left[\frac{32T^2+64T}{(MN)^2}+\frac{64T}{M^2NH}+32b_\alpha L\eta\eta_l HT+8b_\alpha\right] \\
&\quad\cdot\frac{1}{MNHT}\sum_{m,n,t,h}\mathbb{E}\|\nabla f(x_{t,h}^{(m)};\xi_{n,m})\|^2 \\
&\le 96\eta^2\eta_l^2H^2\left[\frac{T^2}{(MN)^2}+\frac{T}{M^2NH}+b_\alpha(L\eta\eta_l HT+1)\right] \\
&\quad\cdot\frac{1}{MNHT}\sum_{m,n,t,h}\mathbb{E}\|\nabla f(x_{t,h}^{(m)};\xi_{n,m})\|^2.
\end{aligned} \tag{31}$$

The final inequality uses $32T^2+64T\le 96T^2$ for $T\ge 1$.

□

E.7 Proof of Theorem 4.13

Proof. The optimization bound controls the empirical risk along the trajectory, and nonnegative smooth losses turn that risk into the gradient energy needed for stability. Specifically, the proof of Theorem C.1, together with $\mathbb{E}F_{\mathcal{S}}(\boldsymbol{x}_{\mathcal{S}}^\star)\le F(\boldsymbol{x}^\star)$, gives the first line below. Fresh uniform sampling and self-bounding then give the second:

$$\mathbb{E}F_{\mathcal{S}}(\bar{\boldsymbol{x}}_T)\le\frac{1}{T}\sum_{t<T}\mathbb{E}F_{\mathcal{S}}(\boldsymbol{x}_t)\le\mathcal{R}_{\mathrm{opt}}, \tag{32}$$

$$\frac{1}{T}\sum_{t<T}\mathbb{E}\|\boldsymbol{g}_t\|_2^2=\frac{1}{NT}\sum_{n=1}^{N}\sum_{t<T}\mathbb{E}\|\nabla f(\boldsymbol{x}_t;\xi_n)\|_2^2\le 2L\mathcal{R}_{\mathrm{opt}}. \tag{33}$$

We can therefore substitute this energy bound directly into (23). With $\eta=c/\sqrt{T}$ and $2-L\eta\ge 1$,

$$\begin{aligned}
&\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\|\bar{\boldsymbol{x}}_T(\mathcal{S})-\bar{\boldsymbol{x}}_T(\mathcal{S}^{(n)})\|_2^2 \\
&\le\left[\frac{32\eta^2T^2}{N^2}+\frac{16\eta^2T}{N}+16Lb_\alpha\eta^3T+8b_\alpha\eta^2\right]\frac{1}{T}\sum_{t<T}\mathbb{E}\|\boldsymbol{g}_t\|_2^2 \\
&\le 64Lc^2\mathcal{R}_{\mathrm{opt}}\left[\frac{T}{N^2}+\frac{1}{N}+b_\alpha\left(\frac{Lc}{\sqrt{T}}+\frac{1}{T}\right)\right].
\end{aligned} \tag{34}$$

Apply Lemma 3.4 with $M = 1$ and minimize over γ to combine the output risk and model stability.

$$\begin{aligned}\epsilon_{\text{gen}} &\le \inf_{\gamma>0}\left\{\frac{L}{\gamma}\mathbb{E}F_{\mathcal{S}}(\bar{x}_T) + \frac{L+\gamma}{2N}\sum_n \mathbb{E}\|\bar{x}_T(\mathcal{S}) - \bar{x}_T(\mathcal{S}^{(n)})\|^2\right\} \\ &= \sqrt{2L\mathbb{E}F_{\mathcal{S}}(\bar{x}_T)}\left[\frac{1}{N}\sum_n \mathbb{E}\|\bar{x}_T(\mathcal{S}) - \bar{x}_T(\mathcal{S}^{(n)})\|^2\right]^{1/2} \\ &\quad + \frac{L}{2N}\sum_n \mathbb{E}\|\bar{x}_T(\mathcal{S}) - \bar{x}_T(\mathcal{S}^{(n)})\|^2. \end{aligned} \tag{35}$$

The equality also holds by continuity when either factor is zero. Substitution of (32)–(34) yields

$$\begin{aligned}\epsilon_{\text{gen}} \le Lc\mathcal{R}_{\text{opt}}\Bigg\{&8\sqrt{2}\left[\frac{T}{N^2} + \frac{1}{N} + b_\alpha\left(\frac{Lc}{\sqrt{T}} + \frac{1}{T}\right)\right]^{1/2} \\ &+ 32Lc\left[\frac{T}{N^2} + \frac{1}{N} + b_\alpha\left(\frac{Lc}{\sqrt{T}} + \frac{1}{T}\right)\right]\Bigg\}.\end{aligned} \tag{36}$$

Using $T \le N^2$ and the stepsize restriction,

$$\begin{aligned}(Lc)^2\left[\frac{T}{N^2} + \frac{1}{N} + b_\alpha\left(\frac{Lc}{\sqrt{T}} + \frac{1}{T}\right)\right] &\le 2(Lc)^2 + b_\alpha(Lc)^2(1+Lc) \\ &\le \frac{1}{8} + \frac{5}{32} = \frac{9}{32}.\end{aligned}$$

Here $Lc \le [2(1+\sqrt{1+2b_\alpha})]^{-1} \le 1/4$ and $b_\alpha(Lc)^2 \le 1/8$. Since $8\sqrt{2}+32\sqrt{9/32} = 20\sqrt{2} < 32$, the quadratic term in (36) is absorbed into the square-root term, giving the stated bound in either of its equivalent forms:

$$\begin{aligned}\epsilon_{\text{gen}} &\le 32Lc\mathcal{R}_{\text{opt}}\left[\frac{T}{N^2} + \frac{1}{N} + b_\alpha\left(\frac{Lc}{\sqrt{T}} + \frac{1}{T}\right)\right]^{1/2} \\ &= 32Lc\sqrt{T}\,\mathcal{R}_{\text{opt}}\left[\frac{1}{N^2} + \frac{1}{NT} + b_\alpha\left(\frac{Lc}{T^{3/2}} + \frac{1}{T^2}\right)\right]^{1/2}.\end{aligned} \tag{37}$$

□

The explicit buffer contribution is bounded by $32Lc\mathcal{R}_{\text{opt}}\sqrt{b_\alpha}\big(\sqrt{Lc}/T^{1/4} + 1/T^{1/2}\big)$, by subadditivity of the square root. Substituting the definition of $\mathcal{R}_{\text{opt}}$ gives orders $T^{-3/4}$ and T^{-1} under interpolation, with faster remainders from the T^{-1} term in $\mathcal{R}_{\text{opt}}$. When $F(x^\star) > 0$, the additional orders are $T^{-1/4}$ and $T^{-1/2}$.

The buffer and full-precision blocks in (37) satisfy

$$\begin{aligned}T \ge (b_\alpha LcN^2)^{2/3} &\implies \frac{b_\alpha Lc}{\sqrt{T}} \le \frac{T}{N^2}, \\ T \ge N\sqrt{b_\alpha} &\implies \frac{b_\alpha}{T} \le \frac{T}{N^2}, \\ T \ge \max\{(b_\alpha LcN^2)^{2/3}, N\sqrt{b_\alpha}\} &\implies b_\alpha\left(\frac{Lc}{\sqrt{T}} + \frac{1}{T}\right) \le \frac{2T}{N^2} \le 2\left(\frac{T}{N^2} + \frac{1}{N}\right).\end{aligned}$$

The bound $b_\alpha Lc \le b_\alpha/[2(1+\sqrt{1+2b_\alpha})] \le \sqrt{b_\alpha}/(2\sqrt{2})$ follows from the stepsize restriction. Thus $T \gtrsim \max\{b_\alpha^{1/3}N^{4/3}, N\sqrt{b_\alpha}\}$ also suffices within the admissible horizon. Reducing c lowers the explicit buffer factor but increases the initialization term in $\mathcal{R}_{\text{opt}}$.

For Table 1, take $F(x^\star) = 0$ and $c = [4L(1 + \sqrt{b_\alpha})]^{-1}$, which satisfies the theorem's restriction. With fixed $L > 0$, variance and initial-distance bounds, constants below are uniform in α, N, T. Theorem C.1 and (37) give

$$\begin{aligned}
\epsilon_{\text{opt}} \le \mathcal{R}_{\text{opt}} &\lesssim \frac{1+\sqrt{b_\alpha}}{\sqrt{T}}, \qquad c\sqrt{T}\,\mathcal{R}_{\text{opt}} \lesssim 1, \\
\epsilon_{\text{gen}} &\lesssim \frac{1}{N} + \frac{1}{\sqrt{NT}} + \frac{\sqrt{b_\alpha Lc}}{T^{3/4}} + \frac{\sqrt{b_\alpha}}{T} \\
&\lesssim \frac{1}{N} + \frac{1}{\sqrt{NT}} + \frac{b_\alpha^{1/4}}{T^{3/4}} + \frac{\sqrt{b_\alpha}}{T}, \\
\mathbb{E}F(\bar{x}_T) - F(x^\star) \le \epsilon_{\text{opt}} + \epsilon_{\text{gen}} &\lesssim \frac{1+\sqrt{b_\alpha}}{\sqrt{T}} + \frac{1}{N}.
\end{aligned} \tag{38}$$

Here $b_\alpha Lc \le \sqrt{b_\alpha}/4$, $b_\alpha^{1/4} \le 1 + \sqrt{b_\alpha}$, and $1 \le T \le N^2$. The explicit $1 + \sqrt{b_\alpha}$ factor records the cost of this admissible stepsize; for fixed compression the rate is $O(T^{-1/2} + N^{-1})$.

E.8 Proof of Theorem 4.15

Proof. The DiLoCo stability estimate involves gradients at local iterates, so we must connect their risk to the global optimization trajectory. The trajectory-risk estimate in the proof of Theorem C.3, with $\eta_l = c_l/(H\sqrt{T})$, gives

$$\begin{aligned}
\frac{1}{T}\sum_{t<T} \mathbb{E}F_{\mathcal{S}}(x_t) \le F(x^\star) &+ \frac{2\mathbb{E}\|x_0 - x_{\mathcal{S}}^\star\|^2}{\eta c_l \sqrt{T}} + \frac{6\eta c_l \sigma^2}{MH\sqrt{T}} \\
&+ \frac{18Lb_\alpha \eta^2 c_l^2(\sigma^2 + H\zeta^2)}{HT} + \frac{9Lc_l^2(H-1)(\sigma^2+H\zeta^2)}{H^2T}.
\end{aligned}$$

We used $\mathbb{E}F_{\mathcal{S}}(x_{\mathcal{S}}^\star) \le F(x^\star)$. To pass from the server trajectory to the local gradients, Lemma D.6 gives

$$\begin{aligned}
\frac{1}{MNHT}\sum_{m,n,t,h} \mathbb{E}\|\nabla f(x_{t,h}^{(m)}; \xi_{n,m})\|^2 &\le \frac{2L}{MHT}\sum_{m,t,h} \mathbb{E}F_{\mathcal{S}_m}(x_{t,h}^{(m)}) \\
&\le \frac{2L}{T}\sum_{t<T} \mathbb{E}F_{\mathcal{S}}(x_t) + \frac{L^2\eta_l^2(H-1)\sigma^2}{2} \\
&= \frac{2L}{T}\sum_{t<T} \mathbb{E}F_{\mathcal{S}}(x_t) + \frac{L^2 c_l^2(H-1)\sigma^2}{2H^2T}.
\end{aligned}$$

The last remainder is at most $L/18$ times the local-drift term in the server risk bound. Convexity bounds the averaged output by the server risk average. Using $\widetilde{\mathcal{R}}_{\text{opt}}$ from Theorem 4.15 and $2+1/18 \le 3$ gives

$$\begin{aligned}
\mathbb{E}F_{\mathcal{S}}(\bar{x}_T) \le \frac{1}{T}\sum_{t<T} \mathbb{E}F_{\mathcal{S}}(x_t) &\le \frac{\widetilde{\mathcal{R}}_{\text{opt}}}{\sqrt{T}}, \\
\frac{1}{MNHT}\sum_{m,n,t,h} \mathbb{E}\|\nabla f(x_{t,h}^{(m)}; \xi_{n,m})\|^2 &\le \frac{3L\widetilde{\mathcal{R}}_{\text{opt}}}{\sqrt{T}}.
\end{aligned} \tag{39}$$

Apply Lemma 3.4 and minimize over γ.

$$\begin{aligned}\epsilon_{\text{gen}} &\le \inf_{\gamma>0}\left\{\frac{L}{\gamma}\mathbb{E}F_{\mathcal{S}}(\bar{x}_T)+\frac{L+\gamma}{2MN}\sum_{m,n}\mathbb{E}\|\bar{x}_T(\mathcal{S})-\bar{x}_T(\mathcal{S}^{(nm)})\|^2\right\}\\ &= \sqrt{2L\mathbb{E}F_{\mathcal{S}}(\bar{x}_T)}\left[\frac{1}{MN}\sum_{m,n}\mathbb{E}\|\bar{x}_T(\mathcal{S})-\bar{x}_T(\mathcal{S}^{(nm)})\|^2\right]^{1/2}\\ &\quad+\frac{L}{2MN}\sum_{m,n}\mathbb{E}\|\bar{x}_T(\mathcal{S})-\bar{x}_T(\mathcal{S}^{(nm)})\|^2. \end{aligned} \tag{40}$$

As in the single-node proof, zero cases follow by continuity. Inserting (31) and (39), with $\eta_l = c_l/(H\sqrt{T})$, gives

$$\begin{aligned}\epsilon_{\text{gen}} \lesssim \widetilde{\mathcal{R}}_{\text{opt}}\Bigg\{&L\eta c_l\left[\frac{1}{(MN)^2}+\frac{1}{M^2NHT}+b_\alpha\left(\frac{L\eta c_l}{T^{3/2}}+\frac{1}{T^2}\right)\right]^{1/2}\\ &+L^2\eta^2c_l^2\left[\frac{\sqrt{T}}{(MN)^2}+\frac{1}{M^2NH\sqrt{T}}+b_\alpha\left(\frac{L\eta c_l}{T}+\frac{1}{T^{3/2}}\right)\right]\Bigg\}.\end{aligned} \tag{41}$$

To display the sampling and compression contributions separately as in the theorem, we split the square root into three terms:

$$\begin{aligned}&\left[\frac{1}{(MN)^2}+\frac{1}{M^2NHT}+b_\alpha\left(\frac{L\eta c_l}{T^{3/2}}+\frac{1}{T^2}\right)\right]^{1/2}\\ &\quad\le \frac{1}{MN}+\frac{1}{M\sqrt{NHT}}+\frac{\sqrt{b_\alpha}}{T^{3/4}}\sqrt{L\eta c_l+\frac{1}{\sqrt{T}}}.\end{aligned}$$

This split costs at most a factor of $\sqrt{3}$. The three quadratic terms are already explicit in (41).

□

For the worker dependence in Table 1, take $\eta=\sqrt{MH}$ and suppose $F(x^\star)=0$, with fixed positive L, c_l and bounded $\sigma^2, \zeta^2, \mathbb{E}\|x_0 - x^\star_{\mathcal{S}}\|^2$. Constants in this rate calculation may depend on these fixed quantities but are uniform in M, N, H, T, α. A sufficient compatible horizon is

$$T \gtrsim \max\{MH(1+b_\alpha), b_\alpha^2(MH)^3\}, \qquad T \lesssim MN^2/H,$$

with constants chosen to satisfy the explicit stepsize restrictions of Theorem C.3. Indeed, the lower bound controls $L\eta_l H = Lc_l/\sqrt{T}$ and $L\eta\eta_l H(1+\sqrt{b_\alpha}) = Lc_l\sqrt{MH}(1+\sqrt{b_\alpha})/\sqrt{T}$. Since $(\sigma^2+H\zeta^2)/H = O(1)$, the optimization prefactor obeys

$$\begin{aligned}\widetilde{\mathcal{R}}_{\text{opt}} &\lesssim \frac{1}{\eta}+\frac{\eta}{MH}+\frac{b_\alpha\eta^2(\sigma^2+H\zeta^2)}{H\sqrt{T}}+\frac{(H-1)(\sigma^2+H\zeta^2)}{H^2\sqrt{T}}\\ &\lesssim \frac{1}{\eta}+\frac{b_\alpha\eta^2}{\sqrt{T}}+\frac{1}{\sqrt{T}} \lesssim \frac{1}{\eta}=\frac{1}{\sqrt{MH}}.\end{aligned}$$

The last step uses $T \gtrsim \eta^2$ and $\sqrt{T} \gtrsim b_\alpha\eta^3$. For the compression terms in Theorem 4.15,

$$\begin{aligned}\frac{L\eta c_l\sqrt{b_\alpha}}{T^{3/4}}\sqrt{L\eta c_l+\frac{1}{\sqrt{T}}} &\lesssim \frac{\eta^{3/2}\sqrt{b_\alpha}}{T^{3/4}}+\frac{\eta\sqrt{b_\alpha}}{T} \lesssim \frac{1}{\sqrt{T}},\\ \frac{L^2\eta^2c_l^2b_\alpha}{T}\left(L\eta c_l+\frac{1}{\sqrt{T}}\right) &\lesssim \frac{\eta^3b_\alpha}{T}+\frac{\eta^2b_\alpha}{T^{3/2}} \lesssim \frac{1}{\sqrt{T}}.\end{aligned}$$

The lower bound on T gives $T \gtrsim \eta^6 b_\alpha^2$ and $T \gtrsim \eta^2 b_\alpha$ for these estimates. Choosing the constant in the upper bound on T at most one ensures $HT \le (MN)^2$ and gives $\eta\sqrt{T} \lesssim MN$ for the sampling

terms. Using $\eta^2 = MH$, the full-precision terms satisfy

$$\frac{L\eta c_l}{MN} + \frac{L\eta c_l}{M\sqrt{NHT}} + \frac{L^2\eta^2 c_l^2\sqrt{T}}{(MN)^2} + \frac{L^2\eta^2 c_l^2}{M^2NH\sqrt{T}}$$
$$\lesssim \frac{\eta}{MN} + \frac{\eta}{M\sqrt{NHT}} + \frac{\eta^2\sqrt{T}}{M^2N^2} + \frac{\eta^2}{M^2NH\sqrt{T}} \lesssim \frac{1}{\sqrt{T}}.$$

Combining optimization and generalization yields

$$\mathbb{E}F(\bar{x}_T) - F(x^\star) \le \epsilon_{\text{opt}} + \epsilon_{\text{gen}} \lesssim \frac{\widetilde{\mathcal{R}}_{\text{opt}}}{\sqrt{T}} \lesssim \frac{1}{\sqrt{MHT}}. \tag{42}$$

For $b_\alpha = O(1/M)$, $MH^3 \lesssim T \lesssim MN^2/H$ is sufficient. Stronger compression is also covered when the longer required horizon is compatible with the upper bound; in particular, a dominating $b_\alpha^2(MH)^3$ lower term requires $N \gtrsim b_\alpha MH^2$. The leading worker gain comes from the choice $\eta = \sqrt{MH}$ balancing initialization and sampling in the optimization bound. The horizon condition keeps the buffer contribution within that order.

F Notation

Table 2: Main notation.

Symbol	Meaning
Data and risk	
d	Model dimension.
M, N	Workers and samples per worker; total sample size MN.
m, n	Worker index in $[M]$ and sample index in $[N]$.
$\xi_{n,m}, \mathcal{D}_m$	Sample n at worker m, drawn from $\mathcal{D}_m$.
$\mathcal{S}_m, \mathcal{S}$	Local dataset and distributed dataset $(\mathcal{S}_1, \dots, \mathcal{S}_M)$.
$\tilde{\mathcal{S}}$	Independent copy of $\mathcal{S}$.
$\mathcal{S}^{(nm)}$	Dataset with sample n at worker m replaced; $\mathcal{S}^{(n)}$ for one node.
$\mathcal{A}$	Randomized learning algorithm.
$f(x;\xi)$	Per-sample loss.
$F(x)$	Population risk, $M^{-1}\sum_m \mathbb{E}_{\xi\sim\mathcal{D}_m} f(x;\xi)$.
$F_{\mathcal{S}_m}(x)$	Local empirical risk, $N^{-1}\sum_n f(x;\xi_{n,m})$.
$F_{\mathcal{S}}(x)$	Global empirical risk, $M^{-1}\sum_m F_{\mathcal{S}_m}(x)$.
$x^\star, x^\star_{\mathcal{S}}$	Population-risk and empirical-risk minimizers.
ϵ_{gen}	Expected generalization gap (Definition 3.2).
ϵ_{opt}	Expected empirical optimization error (Definition 3.2).
Updates and averaging	
T	Number of single-node iterations or communication rounds.
H	Local SGD steps per communication round.
t, h	Iteration/round index $0 \le t < T$ and local-step index $0 \le h < H$.
x_t	Single-node iterate or global model at round t.
$x_{t,h}^{(m)}$	Local model at worker m, with $x_{t,0}^{(m)} = x_t$.
$\bar{x}_T$	Output model, $T^{-1}\sum_{t=0}^{T-1} x_t$.
$i_t, i_{t,h}^{(m)}$	Fresh uniform sample indices in single-node and local SGD.
$g_t, g_{t,h}^{(m)}$	Stochastic gradients at the corresponding actual models.
Δ_t	Single-node SGD update, ηg_t.
$\Delta_t^{(m)}$	Local update, $x_t - x_{t,H}^{(m)}$.
$e_t, e_t^{(m)}$	Single-node and worker EF buffers.
$u_t, u_t^{(m)}$	Compressed compensated updates, e.g. $u_t = \mathcal{C}(\Delta_t + e_t)$.

Table 2: Main notation (continued).

Symbol	Meaning
$\overline{\Delta}_t, \bar{e}_t, \bar{u}_t$	Arithmetic worker averages of the corresponding quantities.
$\widehat{x}_t$	Virtual iterate: $x_t - e_t$ for one node; $x_t - \eta\bar{e}_t$ for DiLoCo.
η, η_l	Single-node/outer stepsize and local stepsize.
c, c_l	Stepsize constants subject to the conditions of each result.
$\mathcal{O}_{\mathrm{in}}, \mathcal{O}_{\mathrm{out}}$	Inner and outer optimizers (Algorithm 2).
Smoothness, noise, and compression	
L	Per-sample gradient-Lipschitz constant.
σ^2	Conditional stochastic-gradient variance bound.
ζ^2	Squared local-to-global empirical gradient gap bound, per worker.
$\mathcal{C}, \alpha$	Compressor and contraction parameter $0 < \alpha \leq 1$.
b_α	Buffer-energy coefficient, $(1-\alpha)/(1-\sqrt{1-\alpha})^2$.
k	Number of retained coordinates in sparsification.
ω	Relative mean-square compression-error bound.
Stability and proof notation	
ϵ	Root-mean-square model-stability bound (Definition 3.3).
$\mathcal{R}_{\mathrm{opt}}$	Single-node empirical-risk bound for the output and for the average over iterates.
$\widetilde{\mathcal{R}}_{\mathrm{opt}}$	Prefactor in the DiLoCo empirical-risk bound $\widetilde{\mathcal{R}}_{\mathrm{opt}}/\sqrt{T}$.
x'_t, e'_t	Quantities from the neighboring-data run with matched randomness.
δ_t	Virtual-iterate difference, $\widehat{x}_t - \widehat{x}'_t$.
Φ_t	Lyapunov function used in each optimization proof.
Δ_t°	Worker-average update with sample-hit steps skipped (Appendix E.6).
w_t	Difference between the skipped-step and actual update gaps (Appendix E.6).
$\varepsilon_{t,h}^{(m)}$	Sampling noise, $g_{t,h}^{(m)} - \nabla F_{\mathcal{S}_m}(x_{t,h}^{(m)})$.
$\mathcal{F}_t, \mathcal{F}_{t,h}^{(m)}$	Data and history before the indicated fresh sample draw.
$\mathcal{F}_t^{\mathrm{comp}}$	History up to the current compression inputs, before the compressor draws.
$x_t^{\mathrm{FP}}, \bar{x}_T^{\mathrm{FP}}$	Full-precision reference iterate and its time average.
$\bar{r}_t$	Worker-average compression residual without EF (Appendix E.4).
μ_t	Conditional residual mean, $\mathbb{E}[\bar{r}_t \mid \mathcal{F}_t^{\mathrm{comp}}]$.
$\mathcal{R}_{\mathrm{noEF}}(T)$	No-EF comparison term, $4\mathbb{E}\|\bar{x}_T - \bar{x}_T^{\mathrm{FP}}\|^2$.
γ, θ	Positive parameters in the generalization and Young inequalities.
$\mathbb{E}$	Expectation over data and algorithmic randomness.
$\|\cdot\|, \|\cdot\|_2$	Euclidean norm.
$\lesssim$	Inequality up to a universal constant, with stated exceptions.
Comparison-specific quantities in Table 1	
λ	$\max\{\|\lambda_2(W)\|, \|\lambda_M(W)\|\}^2$, for the D-SGD mixing matrix W.
D	FedAvg distribution discrepancy, $\max_m d_{\mathrm{TV}}(\mathcal{D}_m, M^{-1}\sum_j \mathcal{D}_j)$.
ν^2	FedAvg local population-gradient variance bound; distinct from σ^2.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Section 4 and Appendices C–E establish the stability and generalization claims under the stated convex-SGD assumptions. Section 5 and Appendix A distinguish the convex experiments from the qualitative non-convex evidence.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Remark 4.6 and Remarks 4.14–4.17 discuss the assumptions and remaining compression dependence. Section 2 explains the restriction to SGD optimizers, and Appendix A describes the empirical scope.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The formal results state or reference their assumptions. Appendices C–E give the supporting lemmas, optimization analysis, and proofs; Appendix F collects the notation.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [No]

   Justification: Appendix A.1 reports the datasets and parameter sweeps. The full non-convex architecture, optimizer configuration, data splits, and evaluation implementation needed for independent reproduction are not specified.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The paper identifies MNIST and DCLM-baseline 1.0, but provides no public training/plotting-code link or complete reproduction instructions.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [No]

   Justification: The reported sweeps are specified in Appendix A.1; the full training and evaluation configuration, including the non-convex setup, is not reported.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The figures illustrate qualitative trends. They do not report error bars, confidence intervals or significance tests, and do not establish statistically significant differences between compressors.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: Device specifications, memory requirements and per-run runtimes are not reported in the current manuscript.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work studies optimization and generalization using MNIST and DCLM-baseline 1.0 and does not involve participant recruitment or human-subject studies. To the best of the authors' knowledge, it conforms to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The contribution is a theoretical analysis of general-purpose compressed optimization, without an application-specific deployment. We do not identify a direct application-specific societal harm from these theoretical results.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release high-risk models, scraped datasets, pre-trained language models, image generators, or other assets that would require special safeguards for responsible release.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [No]

    Justification: The manuscript identifies the experimental datasets, but does not explicitly document the licenses and terms for all datasets, models and software assets used.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not introduce a new dataset, benchmark, pre-trained model, or other standalone asset requiring separate documentation.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing, user studies, or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing, user studies, or research with human subjects, so IRB approval is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: During manuscript revision, LLM assistance was used to check derivations, refine bounds, and edit the exposition. LLMs are not a component of the EF-SGD or DiLoCo-EF training algorithms.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.