ADAPTIVE LIPSCHITZ-FREE PROXIMAL GRADIENT METHODS FOR STOCHASTIC COMPOSITE NONCONVEX OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose ALFPG, an Adaptive Lipschitz-Free Proximal Gradient framework for stochastic composite nonconvex minimization. The framework adapts its stepsizes using information collected along the optimization trajectory, without requiring prior knowledge of smoothness constants or performing line search. For finite-sum problems, ALFPG-FS combines coordinate-wise stepsizes, Nesterov's extrapolation, and the SPIDER estimator, with ALFPG-D as its full-batch specialization. Under component-wise smoothness and a bounded effective domain, ALFPG-FS achieves ϵ-stationarity in expectation using $\mathcal{O}(N + \sqrt{N}\epsilon^{-2})$ component-gradient evaluations. For expectation problems, we develop ALFPG-MVR1 and ALFPG-MVR2, where MVR stands for momentum-based variance reduction. Both use one fresh sample per iteration: ALFPG-MVR1 uses one stochastic-gradient evaluation, whereas ALFPG-MVR2 uses at most two evaluations at different points with the same sample. ALFPG-MVR1 requires smoothness of the expected loss; ALFPG-MVR2 additionally requires uniformly smooth sample losses. When the regularizer is convex or Lipschitz continuous on its effective domain, and stochastic gradients are almost surely bounded, the two methods achieve ϵ-stationarity in expectation using $\widetilde{\mathcal{O}}(\sigma^2\epsilon^{-4} + \epsilon^{-2})$ and $\widetilde{\mathcal{O}}(\sigma\epsilon^{-3} + \epsilon^{-2})$ stochastic-gradient evaluations, respectively. Here, N is the number of component functions, σ^2 bounds the gradient noise variance, and $\widetilde{\mathcal{O}}$ hides logarithmic factors. Neither expectation variant requires prior knowledge of σ, and both recover $\widetilde{\mathcal{O}}(\epsilon^{-2})$ complexity in the zero-noise regime.

1 INTRODUCTION

We consider composite nonconvex optimization of the form

$$\min_{\mathbf{x}\in\mathbb{R}^n} F(\mathbf{x}) := f(\mathbf{x}) + h(\mathbf{x}), \qquad f(\mathbf{x}) = \begin{cases} \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} f_i(\mathbf{x}), & \text{finite-sum setting,} \\ \mathbb{E}_{\xi\sim\mathcal{D}}[f(\mathbf{x};\xi)], & \text{expectation setting.} \end{cases} \tag{1}$$

The smooth loss $f(\cdot)$ may be nonconvex, and the regularizer $h(\cdot)$ is proper and lower semicontinuous. We assume that the required proximal subproblems admit exactly and efficiently computable global minimizers (Assumption 3.1). For finite-sum settings, $h(\cdot)$ can be nonconvex and non-Lipschitz under the stated domain condition, while for expectation settings, $h(\cdot)$ is required to be convex or Lipschitz continuous on its effective domain (Assumption 5.4).

Composite Optimization and Structural Regularization. The composite formulation separates data fitting from structural regularization. The term $h(\cdot)$ can promote sparsity or encode constraints through an extended-valued indicator function, allowing the optimization model to incorporate prior information about the desired solution. Applications include sparse phase retrieval Cai et al. (2024); Shechtman et al. (2014), optimization with orthogonality constraints Wen & Yin (2013), and quantized model training Bai et al. (2019). These examples motivate methods that accommodate nonconvex losses and structural regularizers while keeping gradient evaluations and parameter tuning manageable.

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

From Adaptivity to Lipschitz-Free Updates. Proximal gradient methods exploit this structure by combining gradient evaluations of $f(\cdot)$ with proximal updates involving $h(\cdot)$. Their stepsizes are commonly chosen using a smoothness constant or a backtracking procedure. Adaptive gradient and proximal methods use information collected during optimization to adjust their updates Duchi et al. (2011); Ward et al. (2020); Yun et al. (2021). However, an adaptive preconditioner may still be accompanied by smoothness-dependent parameter restrictions. We call an update *Lipschitz-free* when the parameter choices supporting its convergence guarantee require neither a smoothness constant nor line search. This property concerns algorithmic tuning; smoothness assumptions remain part of the analysis.

Challenges in Adaptive Proximal Updates. Adaptation for composite objectives should account for the interaction between the loss and the regularizer. At a stationary point $\mathbf{x}_\star$, the inclusion $0 \in \nabla f(\mathbf{x}_\star) + \partial h(\mathbf{x}_\star)$ may hold even when $\nabla f(\mathbf{x}_\star) \neq 0$. Consequently, the magnitude of the smooth gradient alone need not indicate proximity to composite stationarity. Proximal displacements incorporate the effect of h, but relating them to stationarity requires controlling the evolving metric. Extrapolation and stochastic gradient estimation introduce additional dependencies between the adaptive parameters, iterate motion, and estimation errors. Our central question is how to couple noise-sensitive weights and proximal-displacement normalization without knowing either smoothness or noise constants. The resulting random weights depend on the same trajectory as the tracking errors, so a convergence argument must control their joint moments rather than treat them as independent.

Summary of Main Contributions. We propose Adaptive Lipschitz-Free Proximal Gradient (ALFPG), a framework that uses proximal displacements as a common principle for stepsize adaptation. We measure stationarity by $r(\mathbf{x}) := \operatorname{dist}(0, \partial F(\mathbf{x}))$, where ∂ denotes the limiting subdifferential. Our main contributions are as follows.

1. **Adaptive Updates Using Proximal Displacements.** For finite-sum problems, ALFPG-FS adapts a diagonal metric using metric-scaled iterate displacements and combines this metric with Nesterov extrapolation and the SPIDER estimator. Its extrapolation coefficient is coupled to the evolving metric to support the descent analysis. The full-batch specialization is denoted by ALFPG-D. For expectation problems, ALFPG-MVR1 and ALFPG-MVR2 (MVR: Momentum-based Variance Reduction) combine normalized proximal displacements with gradient-discrepancy statistics to adapt their scalar stepsizes and momentum weights. Each iteration uses one fresh sample, with one stochastic-gradient evaluation for MVR1 and at most two evaluations for MVR2.
2. **Finite-Sum Guarantees with General Regularizers.** Under the stated finite-sum assumptions, including the bounded-domain condition, ALFPG-D and ALFPG-FS attain gradient oracle complexities $\mathcal{O}(N\epsilon^{-2})$ and $\mathcal{O}(N+\sqrt{N}\epsilon^{-2})$, respectively, for the prescribed stationarity criterion. These results establish the corresponding guarantees for Lipschitz-free diagonal adaptation combined with extrapolation and potentially nonconvex regularization.
3. **Noise-Adaptive Guarantees for Expectation Problems.** ALFPG-MVR1 requires smoothness of the expected loss, whereas ALFPG-MVR2 additionally requires smoothness of the sample losses for its same-sample correction. Under the respective smoothness assumptions, Assumption 5.4, and the bounded stochastic-gradient oracle assumption, the methods require $\widetilde{\mathcal{O}}(\epsilon^{-2} + \sigma^2\epsilon^{-4})$ and $\widetilde{\mathcal{O}}(\epsilon^{-2} + \sigma\epsilon^{-3})$ stochastic-gradient evaluations, respectively, to obtain a uniformly sampled proximal iterate $\widehat{\mathbf{x}}$ satisfying $\mathbb{E}[r(\widehat{\mathbf{x}})] \leq \epsilon$. Here, σ^2 bounds the gradient noise variance, and $\widetilde{\mathcal{O}}$ hides logarithmic factors. The algorithms do not require this variance bound as an input, and both complexities reduce to $\widetilde{\mathcal{O}}(\epsilon^{-2})$ when $\sigma = 0$.

Tables 1 and 2 compare representative methods in the two settings.

2 Related Work

Proximal Methods with Nonconvex Regularization. Proximal methods exploit the separation between a smooth loss and a regularizer or constraint. For nonconvex losses with convex regularization, RSPG establishes stochastic proximal complexity guarantees Ghadimi et al. (2016). Xu et al. (2019) study potentially nonconvex and nonsmooth regularizers. Their fixed-batch MB-SPG method attains $\mathcal{O}(\epsilon^{-4})$ stochastic-gradient complexity with a batch size of order ϵ^{-2} at a fixed noise level.

Table 1: Comparison among existing methods for composite nonconvex finite-sum minimization.

	Lipschitz-free stepsize	Regularizer class[c]	Nesterov extrapolation	Adaptive diagonal scaling	Gradient oracle complexity
APG Li & Lin (2015)	✘	NC	✔	✘	$\mathcal{O}(N\epsilon^{-2})$
SVRG-APG Li et al. (2017)	✘	NC	✔	✘	unknown[a]
ProxSVRG J. Reddi et al. (2016)	✘	CVX	✘	✘	$\mathcal{O}(N+N^{2/3}\epsilon^{-2})$
SPIDER Fang et al. (2018)	✘	$h\equiv 0$	✘	✘	$\mathcal{O}(N+\sqrt{N}\epsilon^{-2})$
SpiderBoost-M Wang et al. (2019)	✘	CVX	✔	✘	$\mathcal{O}(N+\sqrt{N}\epsilon^{-2})$
ProxSARAH Pham et al. (2020)	✘[b]	CVX	✘	✘	$\mathcal{O}(N+\sqrt{N}\epsilon^{-2})$
SPGR Xu et al. (2019)	✘	NC	✘	✘	$\mathcal{O}(N+\sqrt{N}\epsilon^{-2})$
AdaGrad-Norm Ward et al. (2020)	✔	$h\equiv 0$	✘	✘	$\mathcal{O}(N\epsilon^{-2})$
AGD Kavis et al. (2022a)	✔	$h\equiv 0$	✔	✘	$\mathcal{O}(N\epsilon^{-2})$
ADA-SPIDER Kavis et al. (2022b)	✔	$h\equiv 0$	✘	✘	$\tilde{\mathcal{O}}(N+\sqrt{N}\epsilon^{-2})$
ALFPG-D [ours]	✔	NC	✔	✔	$\mathcal{O}(N\epsilon^{-2})$
ALFPG-FS [ours]	✔	NC	✔	✔	$\mathcal{O}(N+\sqrt{N}\epsilon^{-2})$

[a] This work established linear convergence under a KL condition, but no general stationarity oracle complexity is provided.
[b] This algorithm does not qualify as using adaptive stepsizes in our strict sense (as it relies on the Lipschitz constant).
[c] NC: possibly nonconvex; CVX: convex; $h\equiv 0$: no regularizer.

Table 2: Comparison among existing methods for composite nonconvex expectation minimization.

	Lipschitz-free stepsize	Regularizer class[a]	Function $f(\cdot)$[b]	Batch size	Gradient oracle complexity
SPIDER Fang et al. (2018)	✘	$h\equiv 0$	$\mathbb{L}_+$	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-3})$
SpiderBoost Wang et al. (2019)	✘	CVX	$\mathbb{L}_+$	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-3})$
RSPG Ghadimi et al. (2016)	✘	CVX	$\mathbb{L}$	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-4})$
ProxSARAH Pham et al. (2020)	✘	CVX	$\mathbb{MS}$	$\mathcal{O}(1)$	$\mathcal{O}(\epsilon^{-3})$
SMM Davis & Drusvyatskiy (2019)	✘	CVX	$\mathbb{L}_-$	$\mathcal{O}(1)$	$\mathcal{O}(\epsilon^{-4})$
ProxHSGD Tran-Dinh et al. (2022)	✘	CVX	$\mathbb{MS}$	$\mathcal{O}(1)$	$\mathcal{O}(\epsilon^{-3})$
MB-SPG Xu et al. (2019)	✘	NC	$\mathbb{L}$	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-4})$
STORM Cutkosky & Orabona (2019)	✘	$h\equiv 0$	$\mathbb{L}_+$	$\mathcal{O}(1)$	$\mathcal{O}(\epsilon^{-3})$
PStorm Xu & Xu (2023)	✘	CVX	$\mathbb{MS}$	$\mathcal{O}(1)$	$\tilde{\mathcal{O}}(\epsilon^{-3})$
AdaGrad-Norm Ward et al. (2020)	✔	$h\equiv 0$	$\mathbb{L}$	$\mathcal{O}(1)$	$\tilde{\mathcal{O}}(\epsilon^{-4})$
ProxGen Yun et al. (2021)	✘	NC	$\mathbb{L}$	$\mathcal{O}(\epsilon^{-1})$	$\mathcal{O}(\epsilon^{-4})$
SuperADAM-$\mathbb{L}$ Huang et al. (2021)	✘	IND	$\mathbb{L}$	$\mathcal{O}(1)$	$\tilde{\mathcal{O}}(\epsilon^{-4})$
SuperADAM-$\mathbb{L}_+$ Huang et al. (2021)	✘	IND	$\mathbb{L}_+$	$\mathcal{O}(1)$	$\tilde{\mathcal{O}}(\epsilon^{-3})$
ALFPG-MVR1 [ours]	✔	CVX/LC	$\mathbb{L}$	$\mathcal{O}(1)$	$\tilde{\mathcal{O}}(\epsilon^{-2}+\sigma^2\epsilon^{-4})$
ALFPG-MVR2 [ours]	✔	CVX/LC	$\mathbb{L}_+$	$\mathcal{O}(1)$	$\tilde{\mathcal{O}}(\epsilon^{-2}+\sigma\epsilon^{-3})$

[a] NC: $h(\mathbf{x})$ is possibly nonconvex. CVX: $h(\mathbf{x})$ is convex. IND: $h(\mathbf{x})$ is the indicator function of a convex set. CVX/LC: $h(\mathbf{x})$ is convex or Lipschitz continuous on the effective domain.
[b] $\mathbb{L}$: $f(\mathbf{x})$ is L-smooth. $\mathbb{L}_+$: $f(\mathbf{x};\xi)$ is component-wise L-smooth. $\mathbb{L}_-$: $f(\mathbf{x})$ is weakly convex. $\mathbb{MS}$: the mean-squared smoothness condition displayed in the related-work discussion.

The same work develops the recursive estimator method SPGR, obtaining $\mathcal{O}(\epsilon^{-3})$ complexity under additional sample-wise smoothness. These results provide direct precedents for stochastic proximal optimization with nonconvex regularization. ALFPG studies Lipschitz-free adaptation within this class of composite problems.

Extrapolation in Nonconvex Proximal Optimization. Accelerated proximal gradient methods achieve improved objective-value convergence rates in convex optimization Beck & Teboulle (2009). Methods such as iPiano, APG, and iPALM incorporate inertial or extrapolation steps into nonconvex proximal optimization under their respective assumptions Ochs et al. (2014); Li & Lin (2015); Pock & Sabach (2016). Li et al. (2017) analyze APGnc and SVRG-APGnc, establishing linear convergence of expected function values for the latter under a global Kurdyka–Lojasiewicz condition with exponent $1/2$. Such guarantees rely on additional objective geometry. Our finite-sum analysis couples extrapolation to an adaptive diagonal metric and establishes first-order stationarity complexity under the stated finite-sum assumptions.

Variance Reduction for Finite Sums. SVRG and SAGA use finite-sum structure to reduce gradient-estimation variance Johnson & Zhang (2013); Defazio et al. (2014). Their nonconvex proximal extensions establish complexity guarantees for nonsmooth composite objectives J. Reddi et al. (2016). Recursive estimators, including SARAH, SPIDER, and PAGE, further improve gradient oracle complexity Nguyen et al. (2017); Fang et al. (2018); Li et al. (2021). For convex regularizers, Prox-SpiderBoost, Prox-SpiderBoost-M, and ProxSARAH achieve $\mathcal{O}(N + \sqrt{N}\epsilon^{-2})$ complexity Wang et al. (2019); Pham et al. (2020). SPGR attains the same dependence while permitting nonconvex regularization and directly controlling a subdifferential residual Xu et al. (2019). Thus, this complexity order and support for nonconvex regularization already coexist in prior work. The finite-sum contribution of ALFPG concerns their combination with Lipschitz-free diagonal adaptation and metric-dependent extrapolation. Although ProxSARAH includes adaptive stepsize schemes, its relevant parameter choices still involve smoothness information.

Adaptive Gradients and Proximal Preconditioning. AdaGrad and Adam adapt their updates using historical gradient information Duchi et al. (2011); Kingma & Ba (2015). Reddi et al. (2018) identify convergence failures of Adam and develop AMSGrad. Nonconvex convergence guarantees are established for AdaGrad-Norm and classes of Adam-type methods Ward et al. (2020); Chen et al. (2019). AdaSpider combines Lipschitz-free adaptation with recursive variance reduction for smooth nonconvex finite sums Kavis et al. (2022b). For composite objectives, ProxGen supports adaptive preconditioning, momentum, and potentially nonconvex regularizers Yun et al. (2021), while SUPER-ADAM studies adaptive matrices in constrained nonconvex optimization Huang et al. (2021). Their cited guarantees include smoothness-dependent parameter restrictions. Accordingly, our comparison distinguishes adaptive preconditioning from Lipschitz-free parameter selection.

Recursive Momentum for Expectation Problems. STORM uses a same-sample recursive correction to reduce variance with a constant number of gradient evaluations per iteration Cutkosky & Orabona (2019). STORM+ develops parameter-free adaptation for smooth nonconvex optimization Levy et al. (2021). More recently, Ada-STORM removes bounded-gradient and bounded-function-value requirements from its smooth analysis and obtains a rate without an additional logarithmic factor Jiang et al. (2024). Proximal developments include hybrid estimators and PStorm Tran-Dinh et al. (2022); Xu & Xu (2023). Sampling costs depend on the particular construction: ProxSARAH uses snapshot batches, while the hybrid scheme supporting its cubic complexity uses a larger initialization batch. Our expectation variants use one fresh sample at every iteration, including initialization. MVR1 uses an exponential moving average, and MVR2 adds a same-sample correction. Their composite guarantees retain a bounded stochastic-gradient oracle assumption.

Smoothness Assumptions and Noise Adaptivity. Smoothness of the expected loss, mean-squared smoothness, and uniform sample-wise smoothness impose different requirements on the stochastic oracle. The intermediate mean-squared condition is $\mathbb{E}_\xi \left[\|\nabla f(\mathbf{x};\xi) - \nabla f(\mathbf{y};\xi)\|^2\right] \leq L^2\|\mathbf{x}-\mathbf{y}\|^2$. It is used in the analyses of ProxSARAH, ProxHSGD, and PStorm Pham et al. (2020); Tran-Dinh et al. (2022); Xu & Xu (2023). Our MVR1 result assumes smoothness of the expected loss, whereas the MVR2 result uses uniform sample-wise smoothness. Variance-sensitive guarantees and recovery of deterministic rates have also been studied in smooth adaptive optimization Ward et al. (2020); Cutkosky & Orabona (2019). Our results establish these properties for the proposed Lipschitz-free proximal methods under their respective regularizer and oracle assumptions.

Stationarity Measures and Oracle Complexity. Nonconvex optimization guarantees use gradient norms, proximal-gradient mappings, and subdifferential residuals. Stochastic model-based methods for weakly convex objectives also analyze the gradient of a Moreau envelope Davis & Drusvyatskiy (2019). Our guarantees control $\mathrm{dist}(0, \partial F(\mathbf{x}))$. A proximal-gradient mapping bound can yield a residual bound at a corresponding proximal point, with suitable control of stochastic estimation errors when necessary. These distinctions matter when interpreting complexity comparisons.

Throughout, gradient oracle complexity counts every individual gradient evaluation, including initialization, periodic refreshes, and multiple evaluations using the same sample. It counts component-gradient evaluations in the finite-sum setting and stochastic-gradient evaluations in the expectation setting, and differs from the number of proximal iterations.

Notations. For a nonempty set S, $\|S\|$ denotes $\operatorname{dist}(\mathbf{0}, S) = \inf_{u\in S}\|u\|$. Thus all stationarity bounds below use the minimal subgradient norm. Vector operations are performed element-wise. Specifically, for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, the operations $(\mathbf{x} + \mathbf{y})$, $(\mathbf{x} - \mathbf{y})$, $(\mathbf{x} \odot \mathbf{y})$, and $(\mathbf{x} \div \mathbf{y})$ represent element-wise addition, subtraction, multiplication, and division, respectively. We use $\|\mathbf{x}\|_{\mathbf{v}}$ to denote the weighted vector seminorm, defined as $\|\mathbf{x}\|_{\mathbf{v}} = \sqrt{\langle \mathbf{v}, \mathbf{x} \odot \mathbf{x}\rangle}$. The notations, technical preliminaries, and relevant lemmas are provided in Appendix Section A. The notation $\tilde{\mathcal{O}}(\cdot)$ hides polylogarithmic factors, while $\mathcal{O}(\cdot)$ hides constants.

3 The Proposed Algorithms

This section presents the Adaptive Lipschitz-Free Proximal Gradient (ALFPG) framework for Problem (1). ALFPG-FS addresses finite-sum problems, with ALFPG-D as its full-batch specialization. ALFPG-MVR1 and ALFPG-MVR2 address expectation problems.

The algorithms require the following exact-proximal-computation assumption.

Assumption 3.1. *(Exact proximal computations) For $\gamma > 0$, $\mathbf{a} \in \mathbb{R}^n$, and $\mathbf{v} \in \mathbb{R}^n_{++}$, define $\operatorname{Prox}_{\gamma h}(\mathbf{a}; \mathbf{v}) := \arg\min_{\mathbf{x}\in\mathbb{R}^n} \left\{ h(\mathbf{x}) + \frac{1}{2\gamma}\|\mathbf{x} - \mathbf{a}\|^2_{\mathbf{v}} \right\}$. For every $\mathbf{a} \in \mathbb{R}^n$ and $\gamma > 0$, we assume that a global minimizer of this subproblem can be computed exactly and efficiently: (i) for every $\mathbf{v} \in \mathbb{R}^n_{++}$ in ALFPG-FS and its full-batch specialization ALFPG-D; (ii) for $\mathbf{v} = \mathbf{1}$ in ALFPG-MVR1 and ALFPG-MVR2. A minimizer is selected whenever the solution is nonunique.*

Remark 3.2. ❶ *Assumption 3.1 is commonly employed in nonconvex proximal gradient methods.* ❷ *When $\mathbf{v} = \mathbf{1}$, the diagonal preconditioner reduces to the identity preconditioner. Assumption 3.1 holds for certain regularizers $h(\mathbf{x})$. Common examples include capped-ℓ_1 penalty Zhang (2010b), log-sum penalty Candes et al. (2008), minimax concave penalty Zhang (2010a), Geman penalty Geman & Yang (1995), ℓ_p regularization with $p \in \{0, \frac{1}{2}, \frac{2}{3}, 1\}$, and indicator functions for cardinality constraints, orthogonality constraints in matrices, and rank constraints in matrices.* ❸ *When $\mathbf{v}$ is a general positive vector, the variable-metric operator can still be evaluated for certain coordinate-wise separable regularizers $h(\mathbf{x})$. Examples include ℓ_p regularization with $p \in \{0, \frac{1}{2}, \frac{2}{3}, 1\}$ (with or without bound constraints) Yun et al. (2021), and the W-shaped regularizer Bai et al. (2019).*

Algorithm 1 Proposed ALFPG-FS for Finite-Sum Settings.

Input: $\mathbf{x}_0 \in \operatorname{dom}(F)$. Let $\underline{v} > 0, \alpha > 0, \beta \geq 0$, and $\theta \in [0, 1)$. Choose integers $q, b \geq 1$ (for finite-sum problems, take $q = b = \lceil\sqrt{N}\rceil$).
Initialize: Let $\mathbf{x}_{-1} = \mathbf{y}_0 = \mathbf{x}_0$. Set $\mathbf{v}_{-1} = \mathbf{v}_0 = \underline{v}\mathbf{1}$, $\sigma_{-1} = \theta$.
For $t = 0, \ldots, T$:

1. If $\operatorname{mod}(t, q) \neq 0$, draw a fresh mini-batch $\mathcal{I}_t$ of b indices uniformly with replacement from $\{1, \ldots, N\}$. Compute:
$$\mathbf{g}_t = \begin{cases} \nabla f(\mathbf{y}_t), & \operatorname{mod}(t, q) = 0; \\ \mathbf{g}_{t-1} + \nabla f(\mathbf{y}_t; \mathcal{I}_t) - \nabla f(\mathbf{y}_{t-1}; \mathcal{I}_t), & \text{else.} \end{cases} \tag{2}$$
When $q = 1$, every iteration uses the full gradient, giving ALFPG-D.
2. Let $\mathbf{x}_{t+1} \in \operatorname{Prox}_h(\mathbf{y}_t - \mathbf{g}_t \div \mathbf{v}_t; \mathbf{v}_t)$, $\mathbf{d}_t := \mathbf{x}_{t+1} - \mathbf{x}_t$.
3. Let $\mathbf{s}_t := \alpha\|\mathbf{r}_t\|_2^2 \cdot \mathbf{1} + \beta\mathbf{r}_t \odot \mathbf{r}_t$, where $\mathbf{r}_t := \mathbf{v}_t \odot \mathbf{d}_t$.
4. Set $\mathbf{v}_{t+1} = \sqrt{\mathbf{v}_0 \odot \mathbf{v}_0 + \sum_{i=0}^{t} \mathbf{s}_i}$.
5. Let $\sigma_t := \theta(1 - \sigma_{t-1}) \cdot \min(\mathbf{v}_t \div \mathbf{v}_{t+1})$.
6. Set $\mathbf{y}_{t+1} = \mathbf{x}_{t+1} + \sigma_t \mathbf{d}_t$.

3.1 ALFPG-FS for Finite-Sum Settings

ALFPG-FS combines a displacement-based diagonal metric, metric-coupled Nesterov-type extrapolation, and the SPIDER estimator for finite-sum composite optimization. When $q = 1$, it specializes to the full-batch method ALFPG-D. Both methods are specified in Algorithm 1.

Algorithm Design and Motivation. At the extrapolated point $\mathbf{y}_t$, the algorithm computes the full gradient when $\operatorname{mod}(t, q) = 0$. Otherwise, it applies the SPIDER recursion (2) Fang et al. (2018),

with $\nabla f(\mathbf{y}; \mathcal{I}_t) := b^{-1} \sum_{i \in \mathcal{I}_t} \nabla f_i(\mathbf{y})$. The same mini-batch is evaluated at $\mathbf{y}_t$ and $\mathbf{y}_{t-1}$. Following the proximal step, the metric-scaled iterate displacement $\mathbf{r}_t = \mathbf{v}_t \odot (\mathbf{x}_{t+1} - \mathbf{x}_t)$ determines the next metric. For every coordinate j, $(\mathbf{v}_{t+1})_j^2 = \underline{v}^2 + \alpha \sum_{i=0}^{t} \|\mathbf{r}_i\|_2^2 + \beta \sum_{i=0}^{t} (\mathbf{r}_i)_j^2$. The global term weighted by α and the coordinate-wise term weighted by β jointly determine the inverse stepsizes. The extrapolation coefficient σ_t is coupled to the change in the metric. Our stationarity analysis applies to these coupled updates; it makes no claim of an accelerated worst-case rate.

Relationship to AdaGrad-Norm and Main Novelty. Consider $h \equiv 0$, $\beta = \theta = 0$, and $q = 1$. Then $\mathbf{y}_t = \mathbf{x}_t$, $\mathbf{g}_t = \nabla f(\mathbf{x}_t)$, and $\mathbf{v}_t = c_t \mathbf{1}$ for some $c_t > 0$. The proximal step implies $\mathbf{r}_t = -\mathbf{g}_t$, so the updates reduce to $c_{t+1}^2 = c_t^2 + \alpha \|\mathbf{g}_t\|_2^2$, $\mathbf{x}_{t+1} = \mathbf{x}_t - \frac{\mathbf{g}_t}{c_t}$. After matching the accumulator and step-scale parameters, AdaGrad-Norm uses the same scalar accumulator but takes the corresponding step with the updated denominator c_{t+1} Ward et al. (2020). Thus, the relevant difference in this smooth special case is the order in which the current gradient enters the denominator and the iterate update. For general h, $\mathbf{r}_t$ reflects the proximal step and need not equal $-\mathbf{g}_t$. The proposed framework combines this displacement-based metric with metric-coupled extrapolation and, for $q > 1$, SPIDER estimation.

Additional Remarks on Algorithm 1. ❶ The metric update is equivalently written $\mathbf{v}_{t+1} = \sqrt{\mathbf{v}_t \odot \mathbf{v}_t + \mathbf{s}_t}$. The choice $\beta = 0$ gives a scalar metric, while $\beta > 0$ permits different coordinate-wise stepsizes. ❷ The generalized proximal step permits nonsmooth and nonconvex h under Assumption 3.1, which requires an exactly computable global minimizer of the relevant diagonal-metric subproblem. ❸ The assumption $\alpha > 0$ gives the uniform metric-ratio bound $\max(\mathbf{v}_t)/\min(\mathbf{v}_t) \le \kappa := 1 + \sqrt{\beta/\alpha}$ (Lemma B.1(b)). Taking α very small relative to β weakens this bound. ***(iv)*** The expected stationarity and $\mathcal{O}(N + \sqrt{N}\,\epsilon^{-2})$ component-gradient complexity in Theorem 4.6 assume exact proximal computation, a bounded effective domain, component-wise smoothness, and $q = b = \lceil \sqrt{N} \rceil$.

Algorithm 2 Proposed ALFPG-MVR1 and ALFPG-MVR2 for Expectation Settings.

Input: $\theta > 0$, $\mathbf{x}_1 \in \mathrm{dom}(h)$.

Initialize: $\mathbf{x}_0 = \mathbf{x}_1$, $\mathbf{m}_0 = \mathbf{0}$, $\rho_{-1} = \rho_0 = 1$, and $(p, \lambda) = \begin{cases} (1/2, 1/2), & \text{MVR1}, \\ (2/3, 1/4), & \text{MVR2}. \end{cases}$

For $t = 1, \ldots, T$:

1. Draw a fresh sample $\xi_t \sim \mathcal{D}$. At $t = 1$, set $\xi_0 := \xi_1$.
2. Compute $\mathbf{g}_t := \nabla f(\mathbf{x}_t; \xi_t)$. For MVR2, also compute $\mathbf{g}_t^- := \nabla f(\mathbf{x}_{t-1}; \xi_t)$. At $t = 1$, set $\mathbf{g}_0 := \mathbf{g}_1$; MVR2 reuses $\mathbf{g}_1^- = \mathbf{g}_1$ because $\mathbf{x}_0 = \mathbf{x}_1$. Thus $\chi_1 = 0$ for both variants.
3. Set $\rho_t := \min_{1 \le k \le t} \left(\frac{1 + \max_{1 \le i \le k} \chi_i^2}{1 + \sum_{i=1}^{k} \chi_i^2} \right)^p$, where $\chi_t := \begin{cases} \|\mathbf{g}_{t-1} - \mathbf{g}_t\|, & \text{MVR1}, \\ \|\mathbf{g}_{t-1} - \mathbf{g}_t^-\|, & \text{MVR2}. \end{cases}$
4. Let $\alpha_t := \rho_{t-1}$, and $\mathbf{m}_t := \begin{cases} (1 - \alpha_t)\mathbf{m}_{t-1} + \alpha_t \mathbf{g}_t, & \text{MVR1}, \\ (1 - \alpha_t)(\mathbf{m}_{t-1} - \mathbf{g}_t^-) + \mathbf{g}_t, & \text{MVR2}. \end{cases}$
5. Set $\beta_t := \left(1 + \sum_{i=1}^{t-1} \|\mathbf{r}_i\|^2\right)^{-1/2}$, and $\gamma_t = \theta \beta_t \alpha_t^{\lambda}$.
6. Let $\mathbf{x}_{t+1} \in \mathrm{Prox}_{\gamma_t h}\big(\mathbf{x}_t - \gamma_t \mathbf{m}_t\big)$, and $\mathbf{r}_t := \frac{\mathbf{x}_t - \mathbf{x}_{t+1}}{\gamma_t}$.

3.2 ALFPG-MVR1 AND ALFPG-MVR2 FOR EXPECTATION SETTINGS

ALFPG-MVR1 uses a one-gradient exponential moving average (EMA) estimator Kingma & Ba (2015), whereas ALFPG-MVR2 uses a STORM-type same-sample recursive estimator with at most two gradient evaluations Cutkosky & Orabona (2019). The former assumes smoothness of the expected loss; the latter additionally assumes smoothness of each sample loss. Both methods are specified in Algorithm 2.

Algorithm Design and Motivation. Algorithm 2 combines a stochastic gradient estimator, a proximal step, and two distinct adaptive weights. For ALFPG-MVR1, χ_t is the difference between stochastic gradients at successive iterates; for ALFPG-MVR2, it compares gradients obtained from

two samples at the same point $\mathbf{x}_{t-1}$. The weight ρ_t normalizes the accumulated squared differences by their running maximum and takes the running minimum of the resulting ratios. When the differences are persistently dominated by noise of a comparable scale, its intended behavior is heuristically $\rho_t \approx (1 + t\sigma^2)^{-p}$; this is not a pathwise identity or an assumption in the analysis. The maximum reduces the effect of an isolated gradient spike, while the outer minimum makes ρ_t nonincreasing. More precisely, Lemma C.1 bounds the increments of $1/\rho_t$ and the changes in ρ_t, and provides a contraction inequality used in the gradient-tracking analysis. We use the lagged weight $\alpha_t = \rho_{t-1}$ so that the averaging coefficient and stepsize are fixed before sampling ξ_t.

Relationship to Existing Methods and Main Novelty. AdaGrad-Norm accumulates squared stochastic-gradient norms $\|\mathbf{g}_i\|^2$ in a scalar stepsize denominator Ward et al. (2020). Here, the noise-sensitive weight ρ_t instead uses the squared differences χ_i^2, while the separate factor β_t accumulates squared proximal residuals $\|\mathbf{r}_i\|^2$. These quantities serve different purposes: gradient differences respond to changes in the stochastic oracle, whereas the proximal residual incorporates the regularizer into stepsize adaptation. The ALFPG-MVR1 estimator is a one-gradient exponential moving average. ALFPG-MVR2 uses a recursive correction evaluated at $\mathbf{x}_t$ and $\mathbf{x}_{t-1}$ with the same fresh sample, as in momentum-based variance reduction methods Cutkosky & Orabona (2019); Tran-Dinh et al. (2022). Our design couples these estimators to the discrepancy-based averaging weight and the residual-based proximal stepsize. In particular, MVR1 needs smoothness of the expected loss, whereas MVR2 uses individual smoothness to control its same-sample correction.

Additional Remarks on Algorithm 2. ❶ The parameter choices are $(p, \lambda) = (1/2, 1/2)$ for ALFPG-MVR1 and $(2/3, 1/4)$ for ALFPG-MVR2. Since $\alpha_t = \rho_{t-1}$ and β_t depends only on past proximal residuals, $\gamma_t = \theta\beta_t\alpha_t^{\lambda}$ is determined before the fresh sample is drawn. ❷ ALFPG-MVR1 uses one stochastic-gradient evaluation per iteration; ALFPG-MVR2 uses at most two, both with one fresh sample. Neither variant requires L or the noise level σ as an algorithmic input. ❸ The discrepancy statistic reuses gradients already computed by each variant. For ALFPG-MVR1, evaluating $\nabla f(\mathbf{x}_{t-1}; \xi_t)$ solely to form χ_t would require a second gradient evaluation. We instead use $\chi_t = \|\mathbf{g}_{t-1} - \mathbf{g}_t\|$, which reflects both sampling noise and changes in the gradient along the iterates. For ALFPG-MVR2, the same-sample correction already computes $\mathbf{g}_t^- = \nabla f(\mathbf{x}_{t-1}; \xi_t)$ together with $\mathbf{g}_t = \nabla f(\mathbf{x}_t; \xi_t)$. Thus $\chi_t = \|\mathbf{g}_{t-1} - \mathbf{g}_t^-\|$ compares gradients evaluated at the same point $\mathbf{x}_{t-1}$ using two successive samples, without an additional gradient evaluation. When $\sigma = 0$, we have $\chi_t = 0$ and $\rho_t = 1$ for ALFPG-MVR2; for ALFPG-MVR1, iterate motion can still make $\chi_t > 0$. The analysis controls this additional gradient-change term, and both variants retain the $\widetilde{\mathcal{O}}(T^{-1/2})$ expected stationarity rate under their respective assumptions.

4 Oracle Complexity for ALFPG-D and ALFPG-FS

This section analyzes the oracle complexity of ALFPG-D and ALFPG-FS. In ALFPG-FS, the iterate $\mathbf{x}_t$ depends on the mini-batches $\varsigma_{t-1} := \{\mathcal{I}_0, \mathcal{I}_1, \ldots, \mathcal{I}_{t-1}\}$. We write $\mathbb{E}_{\varsigma_{t-1}}[\cdot] = \mathbb{E}[\cdot]$ for expectation over the algorithmic randomness.

In this section, we make the following assumptions.

Assumption 4.1. *There exists a universal positive constant $\bar{\mathrm{x}}$ such that $\|\mathbf{x}\| \leq \bar{\mathrm{x}}$ for all $\mathbf{x} \in \mathrm{dom}(F)$. Furthermore, we assume $\min_{\mathbf{x}} F(\mathbf{x}) > -\infty$.*

Assumption 4.2. *Each $f_j(\cdot)$ is L-smooth, meaning that $\|\nabla f_j(\mathbf{x}) - \nabla f_j(\tilde{\mathbf{x}})\| \leq L\|\mathbf{x} - \tilde{\mathbf{x}}\|$ for all $j \in [N]$. This property extends to $f(\mathbf{x})$, which is also L-smooth.*

Remark 4.3. ❶ *Assumption 4.1 holds by setting $h(\mathbf{x}) = \iota_\Omega(\mathbf{x})$, where $\iota_\Omega(\cdot)$ is the indicator function of a compact set Ω. For example, it can be enforced by adding simple bound constraints. Since any $\mathbf{x} \in \mathrm{dom}(F) := \{\mathbf{x} : F(\mathbf{x}) < +\infty\}$ is feasible, it follows that $\|\mathbf{x}\| \leq \bar{\mathrm{x}}$ for some $\bar{\mathrm{x}} > 0$.* ❷ *The bounded-domain assumption provides the uniform control of the adaptive metric used in our finite-sum proof; it also appears in studies of composite, fractional, and minimax optimization Liu et al. (2022); Jaggi (2013); Yun et al. (2021); Yuan (2025); Xu et al. (2023).* ❸ *Assumption 4.2 is a standard requirement in the convergence analysis of nonconvex algorithms.*

Notation. We define $\mathcal{Z}_t = \mathbb{E}[\widehat{\mathcal{Z}}_t]$, where $\widehat{\mathcal{Z}}_t := F(\mathbf{x}_t) - F(\bar{\mathbf{x}}) + \frac{1}{2}\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2_{\sigma_{t-1}(\mathbf{v}_t + L)}$, where $\bar{\mathbf{x}} \in \arg\min_{\mathbf{x}} F(\mathbf{x})$. We define $\mathcal{V}_{t+1} := \sqrt{\underline{\mathrm{v}}^2 + (\alpha + \beta)M_t}$, where $M_t := \sum_{i=0}^{t} \|\mathbf{r}_i\|_2^2$.

4.1 Analysis for ALFPG-D

This subsection provides the convergence analysis of ALFPG-D.

The following theorem states the resulting iteration complexity.

Theorem 4.4. *(Proof in Section B.2.1,* Iteration Complexity). Under Assumptions 3.1, 4.1, and 4.2, let $\{\mathbf{x}_t\}_{t=0}^{T+1}$ be generated by ALFPG-D with updates indexed by $t = 0, \dots, T$. We have $\frac{1}{T+1}\sum_{t=0}^{T} \|\nabla f(\mathbf{x}_{t+1}) + \partial h(\mathbf{x}_{t+1})\| \le \frac{C}{\sqrt{T+1}}$, where $C > 0$ is a constant. In other words, there exists $\bar{t} \in [T+1]$ such that $\|\nabla f(\mathbf{x}_{\bar{t}}) + \partial h(\mathbf{x}_{\bar{t}})\| \le \epsilon$, provided $T + 1 \ge C^2\epsilon^{-2} = \mathcal{O}(\epsilon^{-2})$.

Remark 4.5. *Theorem 4.4 gives $\mathcal{O}(\epsilon^{-2})$ proximal iterations and $\mathcal{O}(N\epsilon^{-2})$ component-gradient evaluations for deterministic composite minimization, without using the smoothness constant.*

4.2 Analysis for ALFPG-FS

This subsection provides the convergence analysis of ALFPG-FS.

The following theorem gives the expected stationarity and gradient oracle complexity guarantees.

Theorem 4.6. *(Proof in Appendix B.3.1,* Iteration Complexity) Under Assumptions 3.1, 4.1, and 4.2, let $\{\mathbf{x}_t\}_{t=0}^{T+1}$ be generated by Algorithm 1 with the SPIDER estimator. Define $\mathcal{R}_t := \operatorname{dist}\big(\mathbf{0}, \nabla f(\mathbf{x}_t) + \partial h(\mathbf{x}_t)\big)$.

- **(a)** There exists a constant $C > 0$, independent of T, such that $\frac{1}{T+1}\sum_{t=0}^{T} \mathbb{E}[\mathcal{R}_{t+1}] \le \frac{C}{\sqrt{T+1}}$. Consequently, there exists $\bar{t} \in \{1, \dots, T+1\}$ such that $\mathbb{E}[\mathcal{R}_{\bar{t}}] \le \epsilon$, provided $T + 1 \ge C^2\epsilon^{-2}$. The same expected bound holds for $\mathbf{x}_J$, where J is sampled uniformly from $\{1, \dots, T+1\}$.
- **(b)** With $b = q = \lceil\sqrt{N}\rceil$, the gradient oracle complexity for producing $\mathbf{x}_J$ satisfying $\mathbb{E}[\mathcal{R}_J] \le \epsilon$ is $\mathcal{O}(N + \sqrt{N}\epsilon^{-2})$, with the other problem and algorithm parameters fixed.

Remark 4.7. ❶ *ADA-SPIDER Kavis et al. (2022b) attains $\widetilde{\mathcal{O}}(N + \sqrt{N}\epsilon^{-2})$ gradient oracle complexity for smooth finite-sum problems. Under Assumptions 3.1, 4.1, and 4.2,* ALFPG-FS *attains $\mathcal{O}(N + \sqrt{N}\epsilon^{-2})$ without logarithmic factors in the accuracy or finite-sum size, while permitting a nonconvex regularizer.* ❷ *The deterministic proof uses the recursive bound in* (18)*; the stochastic proof instead uses the first-moment closure in* (38)*. The latter is needed because the adaptive metric is random.*

5 Oracle Complexity for ALFPG-MVR1 and ALFPG-MVR2

We analyze ALFPG-MVR1 and ALFPG-MVR2 under the following assumptions.

Assumption 5.1. *(Unbiased stochastic gradients with bounded variance) Let $\{\mathcal{F}_t\}_{t\ge0}$ be the natural filtration generated by the i.i.d. samples $\xi_1, \xi_2, \dots$. At iteration t, $\mathbf{x}_t$, $\mathbf{x}_{t-1}$, α_t, and all quantities constructed before sampling are $\mathcal{F}_{t-1}$-measurable, while ξ_t is independent of $\mathcal{F}_{t-1}$ and distributed according to $\mathcal{D}$. For every deterministic $\mathbf{x}$, $\mathbb{E}[\nabla f(\mathbf{x};\xi)] = \nabla f(\mathbf{x})$, $\mathbb{E}\|\nabla f(\mathbf{x};\xi) - \nabla f(\mathbf{x})\|^2 \le \sigma^2$.*

Assumption 5.2. *(Bounded stochastic gradients) There exists $\widehat{G} > 0$ such that $\|\nabla f(\mathbf{x};\xi)\| \le \widehat{G}$ almost surely for every $\mathbf{x} \in \mathbb{R}^n$. Consequently, Jensens inequality gives $\|\nabla f(\mathbf{x})\| \le \widehat{G}$ for all $\mathbf{x}$.*

Assumption 5.3. *(Smoothness of the expected loss and lower boundedness) The expected loss f is differentiable and L-smooth: $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \le L\|\mathbf{x} - \mathbf{y}\|$, $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. Moreover, $F_\star := \inf_{\mathbf{x}\in\mathbb{R}^n} F(\mathbf{x}) > -\infty$, $\Delta := F(\mathbf{x}_1) - F_\star < \infty$.*

Assumption 5.4. *(Regularizer: convexity or Lipschitz continuity) The function $h : \mathbb{R}^n \to (-\infty, +\infty]$ is proper, lower semicontinuous, and proximable. At least one of the following conditions holds: (i) h is convex. (ii) There exists $\hat{H} \ge 0$ such that $|h(\mathbf{x}) - h(\mathbf{y})| \le \hat{H}\|\mathbf{x} - \mathbf{y}\|$, $\forall \mathbf{x}, \mathbf{y} \in \operatorname{dom}(h)$.*

Remark 5.5. *Convex regularizers may be extended-valued and need not be Lipschitz. The second alternative permits nonconvex regularizers, including indicator functions of nonempty closed sets. Lipschitz continuity on $\operatorname{dom}(h)$ need not bound ambient subgradients.*

Notation and constants. Define $\mathbf{s}_t := \mathbf{m}_t - \nabla f(\mathbf{x}_t)$ and $\mathbf{r}_t := (\mathbf{x}_t - \mathbf{x}_{t+1})/\gamma_t$. Set $\gamma_0 = 0$, $\mathbf{r}_0 = \mathbf{0}$, $\alpha_0 = 1$, and $\mathbf{s}_0 = -\nabla f(\mathbf{x}_1)$, and write $\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot \mid \mathcal{F}_{t-1}]$. We measure composite

stationarity by $\mathcal{R}_t := \operatorname{dist}(0, \nabla f(\mathbf{x}_t) + \partial h(\mathbf{x}_t))$. Constants may depend on L, θ, $\widehat{G}$, K_h, and Δ, but are independent of T and σ. Here K_h is specified in Lemmas C.2 and C.8.

5.1 Analysis for ALFPG-MVR1

For ALFPG-MVR1, the tracking estimator is an exponential moving average. The resulting bound depends on the noise variance and requires only smoothness of the expected loss.

The following theorem establishes a noise-adaptive convergence rate for ALFPG-MVR1.

Theorem 5.6. *(Noise-adaptive composite stationarity of* ALFPG-MVR1*, proof in Theorem C.7) Suppose Assumptions 3.1, 5.1, 5.2, 5.3, and 5.4 hold, and run* ALFPG-MVR1*. Write $\ell_T := \ln(e + T)$. If $R \sim \mathrm{Unif}\{1, \dots, T\}$ is independent of the algorithmic randomness, then there exists a constant $C_R > 0$, independent of T and σ, such that*

$$\mathbb{E}\left[\operatorname{dist}\left(0, \nabla f(\mathbf{x}_{R+1}) + \partial h(\mathbf{x}_{R+1})\right)\right] \le C_R \left(\frac{\ell_T^{5/4}}{T^{1/2}} + \frac{\sigma^{1/2} \ell_T}{T^{1/4}} \right). \tag{3}$$

Thus ALFPG-MVR1 *attains $\widetilde{\mathcal{O}}(T^{-1/2} + \sigma^{1/2} T^{-1/4})$ using one stochastic-gradient evaluation per iteration. Hence, the stochastic-gradient oracle complexity for obtaining $\mathbb{E}[\mathcal{R}_{R+1}] \le \epsilon$ is $\widetilde{\mathcal{O}}(\epsilon^{-2} + \sigma^2 \epsilon^{-4})$.*

5.2 Analysis for ALFPG-MVR2

For ALFPG-MVR2, we additionally assume individual smoothness.

Assumption 5.7. *(Individual smoothness for* ALFPG-MVR2*) For almost every ξ, the sample loss $f(\cdot; \xi)$ is differentiable and L-smooth: $\|\nabla f(\mathbf{x}; \xi) - \nabla f(\mathbf{y}; \xi)\| \le L \|\mathbf{x} - \mathbf{y}\|$, $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.*

The following theorem establishes a noise-adaptive convergence rate for ALFPG-MVR2.

Theorem 5.8. *(Noise-adaptive composite stationarity of* ALFPG-MVR2*, proof in Theorem C.13) Suppose Assumptions 3.1, 5.1, 5.2, 5.3, 5.4, and 5.7 hold, and run* ALFPG-MVR2*. Write $\ell_T := \ln(e + T)$. If $R \sim \mathrm{Unif}\{1, \dots, T\}$ is independent of the algorithmic randomness, then there exists a constant $C_R > 0$, independent of T and σ, such that*

$$\mathbb{E}\left[\operatorname{dist}(0, \nabla f(\mathbf{x}_{R+1}) + \partial h(\mathbf{x}_{R+1}))\right] \le C_R \left(\frac{\ell_T}{T^{1/2}} + \frac{\sigma^{1/3} \ell_T}{T^{1/3}} \right). \tag{4}$$

Thus ALFPG-MVR2 *attains $\widetilde{O}(T^{-1/2} + \sigma^{1/3} T^{-1/3})$ using at most two stochastic-gradient evaluations per iteration. Hence, the stochastic-gradient oracle complexity for obtaining $\mathbb{E}[\mathcal{R}_{R+1}] \le \epsilon$ is $\widetilde{\mathcal{O}}(\epsilon^{-2} + \sigma \epsilon^{-3})$.*

Remark 5.9. ❶ *Theorems 5.6 and 5.8 give the gradient oracle complexities $\widetilde{\mathcal{O}}(\sigma^2 \epsilon^{-4} + \epsilon^{-2})$ and $\widetilde{\mathcal{O}}(\sigma \epsilon^{-3} + \epsilon^{-2})$, respectively. Each iteration uses one fresh sample;* ALFPG-MVR1 *uses one gradient evaluation and* ALFPG-MVR2 *uses at most two.* ❷ *The logarithmic factors arise from the adaptive-weight and motion bounds. Both complexities reduce to $\widetilde{\mathcal{O}}(\epsilon^{-2})$ for $\sigma = 0$. These results use a bounded stochastic-gradient oracle and either alternative in Assumption 5.4.*

6 Conclusions

This paper develops ALFPG, a family of Lipschitz-free proximal methods for composite nonconvex optimization. ALFPG-FS combines a displacement-based diagonal metric, metric-coupled extrapolation, and the SPIDER estimator. The analysis establishes stationarity guarantees for this combination under exact proximal computation and a bounded effective domain. With a full gradient at every iteration, ALFPG-FS reduces to ALFPG-D. Under the stated finite-sum assumptions, ALFPG-D uses $\mathcal{O}(N/\epsilon^2)$ component-gradient evaluations to reach ϵ-stationarity, while ALFPG-FS uses $\mathcal{O}(N + \sqrt{N}/\epsilon^2)$ evaluations to reach the same stationarity criterion in expectation. For expectation problems, ALFPG-MVR1 and ALFPG-MVR2 attain $\widetilde{\mathcal{O}}(\sigma^2 \epsilon^{-4} + \epsilon^{-2})$ and $\widetilde{\mathcal{O}}(\sigma \epsilon^{-3} + \epsilon^{-2})$ gradient oracle complexities under their respective assumptions. Their bounds adapt to the noise level and recover $\widetilde{\mathcal{O}}(\epsilon^{-2})$ when $\sigma = 0$.

LLM USAGE

A large language model (LLM) was used to assist with language editing, improving the presentation of proofs, and reviewing mathematical derivations, with the authors retaining full responsibility for the content and correctness of the paper.

REFERENCES

Yu Bai, Yu-Xiang Wang, and Edo Liberty. Proxquant: Quantized neural networks via proximal operators. In *International Conference on Learning Representations (ICLR)*, 2019.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

Dimitri Bertsekas. *Convex optimization algorithms*. Athena Scientific, 2015.

Jian-Feng Cai, Yu Long, Ruixue Wen, and Jiaxi Ying. A fast and provable algorithm for sparse phase retrieval. In *International Conference on Learning Representations (ICLR)*, 2024.

Emmanuel J Candes, Michael B Wakin, and Stephen P Boyd. Enhancing sparsity by reweighted ℓ_1 minimization. *Journal of Fourier analysis and applications*, 14:877–905, 2008.

Xiangyi Chen, Sijia Liu, Ruoyu Sun, and Mingyi Hong. On the convergence of a class of adam-type algorithms for non-convex optimization. In *7th International Conference on Learning Representations, ICLR 2019*, 2019.

Ashok Cutkosky and Francesco Orabona. Momentum-based variance reduction in non-convex sgd. *Advances in neural information processing systems*, 32, 2019.

Damek Davis and Dmitriy Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019.

Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. *Advances in Neural Information Processing Systems (NeurIPS)*, 27, 2014.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(7), 2011.

Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. *Advances in Neural Information Processing Systems (NeurIPS)*, 31, 2018.

Donald Geman and Chengda Yang. Nonlinear image recovery with half-quadratic regularization. *IEEE transactions on Image Processing*, 4(7):932–946, 1995.

Saeed Ghadimi, Guanghui Lan, and Hongchao Zhang. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming*, 155(1-2): 267–305, 2016.

Feihu Huang, Junyi Li, and Heng Huang. Super-adam: faster and universal framework of adaptive gradients. *Advances in Neural Information Processing Systems*, 34:9074–9085, 2021.

Sashank J. Reddi, Suvrit Sra, Barnabas Poczos, and Alexander J Smola. Proximal stochastic methods for nonsmooth nonconvex finite-sum optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 29, 2016.

Martin Jaggi. Revisiting frank-wolfe: Projection-free sparse convex optimization. In *International conference on machine learning*, pp. 427–435. PMLR, 2013.

Wei Jiang, Sifan Yang, Yibo Wang, and Lijun Zhang. Adaptive variance reduction for stochastic optimization under weaker assumptions. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

Ali Kavis, Kfir Yehuda Levy, and Volkan Cevher. High probability bounds for a class of nonconvex algorithms with adagrad stepsize. In *International Conference on Learning Representations (ICLR)*, 2022a.

Ali Kavis, Stratis Skoulakis, Kimon Antonakopoulos, Leello Tadesse Dadi, and Volkan Cevher. Adaptive stochastic variance reduction for non-convex finite-sum minimization. *Advances in Neural Information Processing Systems*, 35:23524–23538, 2022b.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

Kfir Y. Levy, Ali Kavis, and Volkan Cevher. STORM+: Fully adaptive SGD with momentum for nonconvex optimization, 2021.

Huan Li and Zhouchen Lin. Accelerated proximal gradient methods for nonconvex programming. *Advances in Neural Information Processing Systems (NeurlPS)*, 28, 2015.

Qunwei Li, Yi Zhou, Yingbin Liang, and Pramod K Varshney. Convergence analysis of proximal gradient with momentum for nonconvex optimization. In *International Conference on Machine Learning (ICML)*, pp. 2111–2119. PMLR, 2017.

Zhize Li, Hongyan Bao, Xiangliang Zhang, and Peter Richtarik. Page: A simple and optimal probabilistic gradient estimator for nonconvex optimization. In *International Conference on Machine Learning (ICML)*, pp. 6286–6295, 2021.

Zijian Liu, Ta Duy Nguyen, Alina Ene, and Huy Nguyen. Adaptive accelerated (extra-) gradient methods with variance reduction. In *International Conference on Machine Learning*, pp. 13947–13994. PMLR, 2022.

H. Brendan McMahan and Matthew J. Streeter. Adaptive bound optimization for online convex optimization. In *Conference on Learning Theory (COLT)*, pp. 244–256, 2010.

Boris S. Mordukhovich. Variational analysis and generalized differentiation i: Basic theory. *Berlin Springer*, 330, 2006.

Lam M Nguyen, Jie Liu, Katya Scheinberg, and Martin Takac. Sarah: A novel method for machine learning problems using stochastic recursive gradient. In *International Conference on Machine Learning (ICML)*, pp. 2613–2621, 2017.

Peter Ochs, Yunjin Chen, Thomas Brox, and Thomas Pock. ipiano: Inertial proximal algorithm for nonconvex optimization. *SIAM Journal on Imaging Sciences*, 7(2):1388–1419, 2014.

Nhan H Pham, Lam M Nguyen, Dzung T Phan, and Quoc Tran-Dinh. Proxsarah: An efficient algorithmic framework for stochastic composite nonconvex optimization. *Journal of Machine Learning Research*, 21(110):1–48, 2020.

Thomas Pock and Shoham Sabach. Inertial proximal alternating linearized minimization (ipalm) for nonconvex and nonsmooth problems. *SIAM journal on imaging sciences*, 9(4):1756–1787, 2016.

Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. In *International Conference on Learning Representations*, 2018.

R. Tyrrell Rockafellar and Roger J-B. Wets. Variational analysis. *Springer Science & Business Media*, 317, 2009.

Yoav Shechtman, Amir Beck, and Yonina C Eldar. Gespar: Efficient phase retrieval of sparse signals. *IEEE Transactions on Signal Processing*, 62(4):928–938, 2014.

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

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

Rachel Ward, Xiaoxia Wu, and Leon Bottou. Adagrad stepsizes: Sharp convergence over nonconvex landscapes. *Journal of Machine Learning Research*, 21(219):1–30, 2020.

Zaiwen Wen and Wotao Yin. A feasible method for optimization with orthogonality constraints. *Mathematical Programming*, 142(1):397–434, 2013.

Yangyang Xu and Yibo Xu. Momentum-based variance-reduced proximal stochastic gradient method for composite nonconvex stochastic optimization. *Journal of Optimization Theory and Applications*, 196(1):266–297, 2023.

Yi Xu, Rong Jin, and Tianbao Yang. Non-asymptotic analysis of stochastic methods for non-smooth non-convex regularized problems. *Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019.

Zi Xu, Huiling Zhang, Yang Xu, and Guanghui Lan. A unified single-loop alternating gradient projection algorithm for nonconvex–concave and convex–nonconcave minimax problems. *Mathematical Programming*, 201(1):635–706, 2023.

Ganzhao Yuan. Admm for structured fractional minimization. In *International Conference on Learning Representations (ICLR)*, 2025.

Jihun Yun, Aurélie C Lozano, and Eunho Yang. Adaptive proximal gradient methods for structured neural networks. *Advances in Neural Information Processing Systems (NeurIPS), Supplement:* `https://proceedings.neurips.cc/paper_files/paper/2021/file/cc3f5463bc4d26bc38eadc8bcffbc654-Supplemental.pdf`, 34:24365–24378, 2021.

Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, pp. 894–942, 2010a.

Tong Zhang. Analysis of multi-stage convex relaxation for sparse regularization. *Journal of Machine Learning Research*, 11(3), 2010b.

Appendix

The organization of the appendix is as follows:

Appendix A provides notations, technical preliminaries, and relevant lemmas.

Appendix B offers the proofs of the results in Section 4.

Appendix C contains the proofs of the results in Section 5.

A Notations, Technical Preliminaries, and Relevant Lemmas

A.1 Notations

In this paper, bold lowercase letters represent vectors, and uppercase letters denote real-valued matrices. The following notations are used throughout this paper.

- $[n]$: The set $\{1, 2, ..., n\}$.
- $\|\mathbf{x}\|$: Euclidean norm, defined as $\|\mathbf{x}\| = \|\mathbf{x}\|_2 = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$.
- $\langle \mathbf{a}, \mathbf{b} \rangle$: The Euclidean inner product of $\mathbf{a}$ and $\mathbf{b}$.
- $\mathbf{x} \pm \mathbf{y}$: Element-wise addition or subtraction of $\mathbf{x}$ and $\mathbf{y}$.
- $\mathbf{x} \odot \mathbf{y}$: Element-wise multiplication of $\mathbf{x}$ and $\mathbf{y}$.
- $\mathbf{x} \div \mathbf{y}$: Element-wise division of $\mathbf{x}$ by $\mathbf{y}$, provided that all components of $\mathbf{y}$ are nonzero.
- $\langle \mathbf{a}, \mathbf{b} \rangle_{\mathbf{v}}$: The weighted Euclidean inner product, defined by $\langle \mathbf{a}, \mathbf{b} \rangle_{\mathbf{v}} = \langle \mathbf{a} \odot \mathbf{b}, \mathbf{v} \rangle$.
- $\|\mathbf{x}\|_{\mathbf{v}}$: The weighted Euclidean seminorm, defined by $\|\mathbf{x}\|_{\mathbf{v}} = \sqrt{\langle \mathbf{x} \odot \mathbf{x}, \mathbf{v} \rangle}$, where $\mathbf{v} \geq \mathbf{0}$.
- $\mathbf{a} \leq \alpha$: For $\mathbf{a} \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$, this means that every component of $\mathbf{a}$ is at most α.
- $\iota_\Omega(\mathbf{x})$: Indicator function of a set Ω with $\iota_\Omega(\mathbf{x}) = 0$ if $\mathbf{x} \in \Omega$ and otherwise $+\infty$.
- $\mathbb{E}[v]$: Expected value of the random variable v.
- $\{A_i\}_{i=0}^{\infty}$, $\{B_i\}_{i=0}^{\infty}$: sequences indexed by the integers $i = 0, 1, 2, 3, \ldots$.
- $\operatorname{dist}(\mathbf{0}, \partial h(\mathbf{x}))$: distance from the origin to $\partial h(\mathbf{x})$ with $\operatorname{dist}(\mathbf{0}, \partial h(\mathbf{x})) = \inf_{\mathbf{y} \in \partial h(\mathbf{x})} \|\mathbf{y}\|$.
- $\mathbf{A}^\mathsf{T}$: the transpose of the matrix $\mathbf{A}$.

A.2 Technical Preliminaries

We introduce key concepts from nonsmooth analysis, focusing on the Fréchet subdifferential and the limiting (Fréchet) subdifferential Mordukhovich (2006); Rockafellar & Wets (2009); Bertsekas (2015). Let $F : \mathbb{R}^n \to (-\infty, +\infty]$ be an extended real-valued, not necessarily convex function. The domain of $F(\cdot)$ is defined as $\operatorname{dom}(F) := \{\mathbf{x} \in \mathbb{R}^n : |F(\mathbf{x})| < +\infty\}$. The Fréchet subdifferential of F at $\mathbf{x} \in \operatorname{dom}(F)$, denoted as $\hat{\partial} F(\mathbf{x})$, is given by

$$\hat{\partial} F(\mathbf{x}) := \{\mathbf{v} \in \mathbb{R}^n : \lim_{\mathbf{z} \to \mathbf{x}} \inf_{\mathbf{z} \neq \mathbf{x}} \frac{F(\mathbf{z}) - F(\mathbf{x}) - \langle \mathbf{v}, \mathbf{z} - \mathbf{x} \rangle}{\|\mathbf{z} - \mathbf{x}\|} \geq 0\}.$$

The limiting subdifferential of $F(\cdot)$ at $\mathbf{x} \in \operatorname{dom}(F)$, denoted $\partial F(\mathbf{x})$, is defined as:

$$\partial F(\mathbf{x}) := \{\mathbf{v} \in \mathbb{R}^n : \exists \mathbf{x}_k \to \mathbf{x}, F(\mathbf{x}_k) \to F(\mathbf{x}), \mathbf{v}^k \in \hat{\partial} F(\mathbf{x}_k) \to \mathbf{v}, \forall k\}.$$

It is important to note that $\hat{\partial} F(\mathbf{x}) \subseteq \partial F(\mathbf{x})$. If $F(\cdot)$ is continuously differentiable at $\mathbf{x}$, then $\hat{\partial} F(\mathbf{x}) = \partial F(\mathbf{x}) = \{\nabla F(\mathbf{x})\}$, where $\nabla F(\mathbf{x})$ represents the gradient of $F(\cdot)$ at $\mathbf{x}$. For convex function $F(\cdot)$, both $\hat{\partial} F(\mathbf{x})$ and $\partial F(\mathbf{x})$ reduce to the classical subdifferential for convex functions: $\hat{\partial} F(\mathbf{x}) = \partial F(\mathbf{x}) = \{\mathbf{v} \in \mathbb{R}^n : F(\mathbf{z}) - F(\mathbf{x}) - \langle \mathbf{v}, \mathbf{z} - \mathbf{x} \rangle \geq 0, \forall \mathbf{z} \in \mathbb{R}^n\}$.

A.3 Relevant Lemmas

We collect elementary inequalities used in the subsequent convergence proofs.

Lemma A.1. *(*Pythagoras Relation) For any vectors $\mathbf{a}, \mathbf{v}, \mathbf{x}, \mathbf{x}_+ \in \mathbb{R}^n$ with $\mathbf{v} \geq \mathbf{0}$, we have:

$$\tfrac{1}{2}\|\mathbf{x}-\mathbf{a}\|_{\mathbf{v}}^2 - \tfrac{1}{2}\|\mathbf{x}_+-\mathbf{a}\|_{\mathbf{v}}^2 = \tfrac{1}{2}\|\mathbf{x}-\mathbf{x}_+\|_{\mathbf{v}}^2 + \langle \mathbf{a}-\mathbf{x}_+, \mathbf{x}_+-\mathbf{x}\rangle_{\mathbf{v}}.$$

Lemma A.2. *For all $a, b \geq 0$ and $c, d > 0$, we have: $\frac{a+b}{c+d} \leq \max(\frac{a}{c}, \frac{b}{d})$.*

Proof. We consider two cases: (***i***) $\frac{a}{c} \geq \frac{b}{d}$. We derive: $b \leq \frac{ad}{c}$, leading to $\frac{a+b}{c+d} \leq \frac{a+\frac{ad}{c}}{c+d} = \frac{a}{c}\cdot\frac{c+d}{c+d} = \frac{a}{c}$. (***ii***) $\frac{a}{c} < \frac{b}{d}$. We have: $a \leq \frac{bc}{d}$, resulting in $\frac{a+b}{c+d} \leq \frac{\frac{bc}{d}+b}{c+d} = \frac{b}{d}\cdot\frac{c+d}{c+d} = \frac{b}{d}$.

□

Lemma A.3. *Assume $ax^2 \leq bx+c$, where $b, c, x \geq 0$ and $a > 0$. Then, we have: $x \leq \sqrt{c/a}+b/a$.*

Proof. Given the quadratic inequality $ax^2 \leq bx + c$, we have $\frac{b-\sqrt{b^2+4ac}}{2a} \leq x \leq \frac{b+\sqrt{b^2+4ac}}{2a}$. Since $x \geq 0$, we have $0 \leq x \leq \frac{b+\sqrt{b^2+4ac}}{2a} \leq \frac{b+b+2\sqrt{ac}}{2a} = b/a + \sqrt{c/a}$, where the last inequality uses $\sqrt{a+d} \leq \sqrt{a}+\sqrt{d}$ for all $a, d \geq 0$.

□

Lemma A.4. *Assume that $\{A_i\}_{i=0}^n$ and $\{B_i\}_{i=0}^{n+1}$ are two non-negative sequences with $A_0 \leq A_1 \leq \ldots \leq A_n$. Then, we have:*

$$\textstyle\sum_{t=0}^n A_t(B_t - B_{t+1}) \leq [\max_{i=0}^n A_i]\cdot[\max_{j=0}^n B_j].$$

Proof. We have: $\sum_{t=0}^n A_t(B_t-B_{t+1}) = [\sum_{t=1}^n (A_t-A_{t-1})B_t] + A_0B_0 - A_nB_{n+1} \overset{①}{\leq} [\sum_{t=1}^n (A_t-A_{t-1})B_t] + A_0B_0 \overset{②}{\leq} [\sum_{t=1}^n (A_t-A_{t-1})]\cdot[\max_{j=0}^n B_j] + A_0[\max_{j=0}^n B_j] = A_n[\max_{j=0}^n B_j]$, where step ① uses $A_n, B_{n+1} \geq 0$; step ② uses $\{A_t\}_{t=0}^n$ is non-decreasing.

□

Lemma A.5. *Let $\{A_t\}_{t=0}^\infty$ be a sequence of nonnegative real numbers, and let $c > 0$. Then, for every $t \geq 0$,*

$$\textstyle\sum_{i=0}^t \frac{A_i}{\sqrt{c+\sum_{j=0}^i A_j}} \leq 2\sqrt{c+\sum_{i=0}^t A_i} \tag{5}$$

Proof. This lemma extends the result of Lemma 5 in McMahan & Streeter (2010).

We define $S_t := c + \sum_{i=0}^t A_i$.

Initially, we define $h(x) := \frac{x}{\sqrt{y}} + 2\sqrt{y-x} - 2\sqrt{y}$, where $y > x \geq 0$. We have $h'(x) = y^{-1/2} - (y-x)^{-1/2} \leq 0$. Therefore, $h(x)$ is non-increasing for all $x \geq 0$. Given $h(0) = 0$, it holds that

$$h(x) := \frac{x}{\sqrt{y}} + 2\sqrt{y-x} - 2\sqrt{y} \leq 0. \tag{6}$$

We complete the proof of the lemma using mathematical induction.

Part (a). The lemma holds for $t = 0$, since $\frac{A_0}{\sqrt{c+A_0}} \leq 2\sqrt{c+A_0}$.

Part (b). Now, fix some t and assume that the lemma holds for $t-1$. We proceed as follows:

$$\begin{aligned}\textstyle\sum_{i=0}^t A_i/\sqrt{S_i} &= A_t/\sqrt{S_t} + \textstyle\sum_{i=0}^{t-1} A_i/\sqrt{S_i} \overset{①}{\leq} A_t/\sqrt{S_t} + 2\sqrt{S_{t-1}} \\ &\overset{②}{=} A_t/\sqrt{S_t} + 2\sqrt{S_t - A_t} \overset{③}{\leq} 2\sqrt{S_t},\end{aligned}$$

where step ① uses the inductive hypothesis that the conclusion of this lemma holds for $t-1$; step ② uses $S_t := c + \sum_{i=0}^t A_i$; step ③ uses Inequality (6) with $x = A_t \geq 0$ and $y = S_t > 0$.

□

Lemma A.6. *Let $\{A_t\}_{t=0}^{\infty}$ be a sequence of nonnegative real numbers, and let $c > 0$ and $p \in (0, 1]$. Then, for every $t \geq 0$,*

$$\sum_{i=0}^{t} \frac{A_i}{c+\sum_{j=0}^{i} A_j} \leq -1 + \frac{1}{p}\left(1 + \frac{1}{c}\sum_{i=0}^{t} A_i\right)^p.$$

Proof. We define $S_t := \sum_{i=0}^{t} A_i$.

We define $g(x) = \frac{x}{1+x} - \frac{1}{p}(1+x)^p + 1$ and $f(x) = x^p$, where $x \geq 0$ and $p \in (0, 1]$.

First, we have $g(0) = 1 - 1/p \leq 0$ and $g'(x) = (1+x)^{-2} - (1+x)^{p-1} = \frac{1-(1+x)^{p+1}}{(1+x)^2} \leq 0$. We derive:

$$g(x) \leq 0. \tag{7}$$

Second, since $f(x)$ is concave, we have $\forall x, y > 0,\ f(x) \leq f(y) + \langle x - y, f'(y)\rangle$, leading to

$$\forall x, y > 0,\ x^p \leq y^p - p(y - x) \cdot y^{p-1}. \tag{8}$$

We finish the proof of the lemma using mathematical induction.

Part (a). We first consider $t = 0$. We have

$$\begin{aligned} A_0/(c+S_0) - \tfrac{1}{p}(1+S_0/c)^p + 1 &\overset{①}{=} A_0/(c+A_0) - \tfrac{1}{p}(1+A_0/c)^p + 1 \\ &= \tfrac{A_0/c}{1+A_0/c} - \tfrac{1}{p}(1+A_0/c)^p + 1 \overset{②}{\leq} 0, \end{aligned}$$

where step ① uses $A_0 = S_0$; step ② uses Inequality (7) with $x = A_0/c$. Thus, the claim holds for $t = 0$.

Part (b). Now, fix some t and assume that the lemma holds for $t - 1$. We derive:

$$\begin{aligned} \textstyle\sum_{i=0}^{t} A_i/(c+S_i) &= A_t/(c+S_t) + \textstyle\sum_{i=0}^{t-1} A_i/(c+S_i) \\ &\overset{①}{\leq} -1 + A_t/(c+S_t) + \tfrac{1}{p}c^{-p}(c+S_{t-1})^p \\ &\overset{②}{=} -1 + A_t/(c+S_t) + \tfrac{1}{p}c^{-p}(c+S_t-A_t)^p \\ &\overset{③}{\leq} -1 + A_t/(c+S_t) + \tfrac{1}{p}c^{-p}\{(c+S_t)^p - pA_t(c+S_t)^{p-1}\} \\ &= -1 + \tfrac{1}{p}c^{-p}(c+S_t)^p + A_t(c+S_t)^{p-1}\cdot\{(c+S_t)^{-p} - c^{-p}\} \\ &\overset{④}{\leq} -1 + \tfrac{1}{p}c^{-p}(c+S_t)^p + 0, \end{aligned}$$

where step ① uses the inductive hypothesis that the conclusion of this lemma holds for $t - 1$; step ② uses $S_t := \sum_{i=0}^{t} A_i$; step ③ uses Inequality (8) with $x = c + S_t - A_t > 0$ and $y = c + S_t > 0$; step ④ uses $(c + S_t)^{-p} \leq c^{-p}$ for all $p \in (0, 1]$.

□

Lemma A.7. *Let $\{Z_t\}_{t=0}^{\infty}$ be a non-negative sequence satisfying $Z_{t+1} \leq a + b\sqrt{\max_{i=0}^{t} Z_i}$ for all $t \geq 0$, where $a, b \geq 0$. It follows that: $Z_t \leq \overline{\mathrm{Z}} := \max\left(Z_0, \left(\frac{1}{2}b + \frac{1}{2}\sqrt{b^2+4a}\right)^2\right)$ for all $t \geq 0$. Furthermore, an alternative valid upper bound for Z_t is given by $\overline{\mathrm{Z}}_+ := \max(Z_0, 2b^2 + 2a)$.*

Proof. We define $M_t := \max_{0 \leq i \leq t} Z_i$ for all $t \geq 0$.

Part (a). For all $t \geq 0$, we have:

$$M_{t+1} \overset{①}{=} \max(M_t, Z_{t+1}) \overset{②}{\leq} \max(M_t, a + b\sqrt{M_t}),$$

where step ① uses the definition of M_t; step ② uses the assumed recursion $Z_{t+1} \leq a + b\sqrt{M_t}$.

Part (b). We establish a fixed-point upper bound that the sequence M_t cannot exceed such that $M_t \le \overline{\mathrm{M}}$. For $\overline{\mathrm{M}}$ to be a valid upper bound, it should satisfy the recurrence relation: $\overline{\mathrm{M}} = a + b\sqrt{\overline{\mathrm{M}}}$. This is because if $M_t \le \overline{\mathrm{M}}$, then we have: $M_{t+1} \le \max\{M_t, a + b\sqrt{M_t}\} \le \overline{\mathrm{M}}$, which would imply by induction that $M_t \le \overline{\mathrm{M}}$ for all $t \ge 0$. Solving the quadratic equation $\overline{\mathrm{M}} = a + b\sqrt{\overline{\mathrm{M}}}$ yields a positive root $(\frac{1}{2}(b + \sqrt{b^2 + 4a}))^2 := c$. Taking into account the case M_0, for all $t \ge 0$, we have $M_t \le \max(M_0, c) = \max(Z_0, c) := \overline{\mathrm{Z}}$.

Part (c). Let $\overline{\mathrm{Z}} = \max(Z_0, c)$. Since $c = a + b\sqrt{c}$, we have $a + b\sqrt{\overline{\mathrm{Z}}} \le \overline{\mathrm{Z}}$. If $M_t \le \overline{\mathrm{Z}}$, then

$$M_{t+1} \le \max\{M_t, a + b\sqrt{M_t}\} \le \overline{\mathrm{Z}}.$$

As $M_0 = Z_0 \le \overline{\mathrm{Z}}$, induction proves the bound for every t.

Part (d). Finally, we have: $c := \left(\frac{1}{2}b + \frac{1}{2}\sqrt{b^2 + 4a}\right)^2 \le (\frac{1}{2}(b + b + 2\sqrt{a}))^2 = (b + \sqrt{a})^2 \le 2b^2 + 2a$. Hence, $\overline{\mathrm{Z}}_+ := \max(Z_0, 2b^2 + 2a)$ is also a valid upper bound for Z_t. □

Lemma A.8. *Let $b_1, \dots, b_T \ge 0$. Then:*

(a) For every $t = 1, \dots, T$, $\frac{b_t}{1+\sum_{i=1}^t b_i} \le \ln\left(1 + \sum_{i=1}^t b_i\right) - \ln\left(1 + \sum_{i=1}^{t-1} b_i\right)$.

(b) We have $\sum_{t=1}^T \frac{b_t}{1+\sum_{i=1}^t b_i} \le \ln\left(1 + \sum_{i=1}^T b_i\right)$.

Proof. Define $B_0 := 1$ and $B_t := 1 + \sum_{i=1}^t b_i$. Then $B_t - B_{t-1} = b_t \ge 0$.

Part (a). Since $u \mapsto 1/u$ is decreasing on $(0, \infty)$, we have: $\ln B_t - \ln B_{t-1} = \int_{B_{t-1}}^{B_t} \frac{1}{u}\, du \ge \frac{B_t - B_{t-1}}{B_t} = \frac{b_t}{B_t}$.

Part (b). Summing (a) over $t = 1, \dots, T$ gives $\sum_{t=1}^T \frac{b_t}{B_t} \le \sum_{t=1}^T (\ln B_t - \ln B_{t-1}) = \ln B_T - \ln B_0 = \ln B_T$.

□

B Proof of Section 4

The proof proceeds in three stages. We first establish metric, extrapolation, and descent estimates shared by both finite-sum variants. We then close the deterministic bounds for ALFPG-D, before incorporating the SPIDER error and taking expectations for ALFPG-FS.

Overview of the proof strategy for ALFPG-D. The deterministic analysis couples the objective gap with an extrapolation energy in $\mathcal{Z}_t$. Proximal optimality, smoothness, and the metric-dependent extrapolation rule yield a descent inequality with two adaptive sums (Lemma B.3). The accumulator estimates in Lemma B.4 control these sums in terms of $\mathcal{V}_{T+1}$. Applying summation by parts to the metric-weighted descent inequality and summing its unweighted counterpart then give a closed recursion for the largest Lyapunov value. This bounds both $\mathcal{Z}_t$ and $\mathcal{V}_t$ uniformly (Lemma B.5), without assuming that the Lyapunov sequence decreases at every iteration. The bounded accumulator implies square-summable iterate displacements; proximal optimality then converts this motion bound into the averaged $\mathcal{O}((T+1)^{-1/2})$ stationarity rate.

Overview of the proof strategy for ALFPG-FS. The finite-sum analysis starts from the same pathwise descent inequality, but must also control the SPIDER estimation error. Within each refresh epoch, its second moment is bounded by the accumulated motion of the extrapolated iterates. Summing over epochs and applying the adaptive-motion estimates bounds the total estimation error by a constant multiple of $\mathbb{E}[\sqrt{\mathcal{V}_{T+1}}]$. Taking expectations only after summing the pathwise descent inequalities yields the first-moment closure in (38), which bounds $\mathbb{E}[\mathcal{V}_t]$ and $\mathcal{Z}_t$ (Lemma B.10). This avoids treating the random metric as deterministic or requiring a uniform second-moment bound for

it. Finally, proximal optimality relates stationarity to iterate motion, estimation error, and metric-scaled displacements. Their accumulated bounds give the expected rate; counting refresh and paired mini-batch evaluations gives the oracle complexity.

B.1 Initial Theoretical Analysis

We begin by relating the diagonal metric to the accumulated scaled displacements. The coordinate-ratio and one-step growth bounds below will be used to control extrapolation and compare successive adaptive denominators.

Lemma B.1. *(*Properties of $\mathbf{v}_t$) We define $M_t := \sum_{i=0}^{t} \|\mathbf{r}_i\|_2^2 \in \mathbb{R}_+$. For all $t \geq 0$, we have:

(a) $\sqrt{\underline{\mathrm{v}}^2 + \alpha M_t} \leq \mathbf{v}_{t+1} \leq \sqrt{\underline{\mathrm{v}}^2 + (\alpha+\beta)M_t} := \mathcal{V}_{t+1}$.

(b) $\frac{\max(\mathbf{v}_t)}{\min(\mathbf{v}_t)} \leq \kappa := 1 + \sqrt{\beta/\alpha}$.

(c) $\frac{\min(\mathbf{v}_{t+1})}{\min(\mathbf{v}_t)} \leq \dot{\kappa} := 1 + 2\kappa\overline{\mathrm{x}}\sqrt{\alpha+\beta}$.

Proof. We define $M_t := \sum_{i=0}^{t} \|\mathbf{r}_i\|_2^2 \in \mathbb{R}_+$, where $\mathbf{r}_t := \mathbf{v}_t \odot \mathbf{d}_t$, and $\mathbf{d}_t := \mathbf{x}_{t+1} - \mathbf{x}_t$.

We define $\mathbf{s}_t := \alpha\|\mathbf{r}_t\|_2^2 \cdot \mathbf{1} + \beta \mathbf{r}_t \odot \mathbf{r}_t \in \mathbb{R}^n$.

Part (a). Given $\mathbf{v}_{t+1} = \sqrt{\mathbf{v}_0 \odot \mathbf{v}_0 + \sum_{i=0}^{t} \mathbf{s}^i}$, we derive:

$$\mathbf{v}_{t+1} \odot \mathbf{v}_{t+1} = (\mathbf{v}_0) \odot (\mathbf{v}_0) + \alpha \underbrace{\textstyle\sum_{i=0}^{t} \|\mathbf{r}_i\|_2^2}_{= M_t} \cdot \mathbf{1} + \beta \underbrace{\textstyle\sum_{i=0}^{t} \mathbf{r}_i \odot \mathbf{r}_i}_{\leq M_t \cdot \mathbf{1}}.$$

This results in the following lower and upper bounds for $\mathbf{v}_{t+1}$ for all $t \geq 0$:

$$\sqrt{\underline{\mathrm{v}}^2 + \alpha M_t} \leq \mathbf{v}_{t+1} \leq \sqrt{\underline{\mathrm{v}}^2 + (\alpha+\beta)M_t}.$$

Part (b). For all $t \geq 1$, we derive:

$$\frac{\max(\mathbf{v}_t)}{\min(\mathbf{v}_t)} \overset{①}{\leq} \sqrt{\frac{\underline{\mathrm{v}}^2 + (\alpha+\beta)M_{t-1}}{\underline{\mathrm{v}}^2 + \alpha M_{t-1}}} \leq \sqrt{\frac{\alpha+\beta}{\alpha}} \leq 1 + \sqrt{\beta/\alpha}, \tag{9}$$

where step ① uses **Part (a)**. Note that Inequality (9) also holds for $t = 0$ since $\mathbf{v}_0 = \underline{\mathrm{v}}\mathbf{1}$.

Part (c). For all $t \geq 0$, we derive the following results:

$$\begin{aligned}
\frac{\min(\mathbf{v}_{t+1})}{\min(\mathbf{v}_t)} &\overset{①}{=} \sqrt{\frac{\min(\mathbf{v}_t \odot \mathbf{v}_t + \alpha\|\mathbf{r}_t\|_2^2 \cdot \mathbf{1} + \beta \mathbf{r}_t \odot \mathbf{r}_t)}{\min(\mathbf{v}_t \odot \mathbf{v}_t)}} \overset{②}{\leq} \sqrt{\frac{\min(\mathbf{v}_t \odot \mathbf{v}_t) + (\alpha+\beta)\|\mathbf{r}_t\|_2^2}{\min(\mathbf{v}_t \odot \mathbf{v}_t)}} \\
&\overset{③}{\leq} \sqrt{\frac{\min(\mathbf{v}_t \odot \mathbf{v}_t) + (\alpha+\beta)\max(\mathbf{v}_t)^2\|\mathbf{x}_{t+1} - \mathbf{x}_t\|_2^2}{\min(\mathbf{v}_t \odot \mathbf{v}_t)}} \overset{④}{\leq} \sqrt{\frac{\min(\mathbf{v}_t \odot \mathbf{v}_t) + 4(\alpha+\beta)\max(\mathbf{v}_t)^2\overline{\mathrm{x}}^2}{\min(\mathbf{v}_t \odot \mathbf{v}_t)}} \\
&\overset{⑤}{\leq} \sqrt{1 + 4(\alpha+\beta)\kappa^2\overline{\mathrm{x}}^2} \leq 1 + 2\kappa\overline{\mathrm{x}}\sqrt{\alpha+\beta} := \dot{\kappa},
\end{aligned}$$

where step ① uses the update rule for $\mathbf{v}_{t+1}$ that $\mathbf{v}_{t+1} = \sqrt{\mathbf{v}_t \odot \mathbf{v}_t + \mathbf{s}_t}$; step ② uses the fact that $\mathbf{r}_t \odot \mathbf{r}_t \leq \|\mathbf{r}_t\|_2^2$; step ③ uses $\mathbf{r}_t := \mathbf{v}_t \odot \mathbf{d}_t$ with $\mathbf{d}_t := \mathbf{x}_{t+1} - \mathbf{x}_t$; step ④ uses $\|\mathbf{x}_{t+1} - \mathbf{x}_t\| \leq \|\mathbf{x}_{t+1}\| + \|\mathbf{x}_t\| \leq 2\overline{\mathrm{x}}$; step ⑤ uses $\max(\mathbf{v}_t)/\min(\mathbf{v}_t) \leq \kappa$ for all t.

□

With the metric bounds in place, we examine the extrapolation coefficients. The following inequalities quantify how their coupling to the metric preserves the negative quadratic term in the descent estimate.

Lemma B.2. *(*Properties of σ_t) For all $t \geq 0$, we have the following results.

(a) $\theta(1-\theta)/(\kappa\dot{\kappa}) \leq \sigma_t \leq \theta$.

(b) $(\sigma_{t-1} - 1)\mathbf{v}_t + \sigma_t \mathbf{v}_{t+1} \leq -(1-\theta)^2 \mathbf{v}_t$.

Proof. We define $\sigma_t := \theta(1 - \sigma_{t-1}) \cdot \min(\mathbf{v}_t \div \mathbf{v}_{t+1})$, where $t \geq 0$.

Part (a). We now prove that $\sigma_t \in [0, \theta]$. We complete the proof using mathematical induction. First, we consider $t = 0$, we have:

$$\sigma_0 \;=\; \min(\mathbf{v}_0 \div \mathbf{v}_1)\theta(1-\sigma_{-1}) \;\overset{①}{=}\; \min(\mathbf{v}_0 \div \mathbf{v}_1)\theta(1-\theta) \;\overset{②}{\leq}\; \theta(1-\theta) \;\overset{③}{\leq}\; \theta,$$

where step ① uses $\sigma_{-1} = \theta$; step ② uses $\min(\mathbf{v}_t \div \mathbf{v}_{t+1}) \in (0, 1]$; step ③ uses $\theta \in [0, 1)$. Second, we fix some t and assume that $\sigma_{t-1} \in [0, \theta]$. We analyze the following term for all $t \geq 1$:

$$\sigma_t := \theta(1-\sigma_{t-1}) \cdot \min(\mathbf{v}_t \div \mathbf{v}_{t+1}).$$

Given $\sigma_{t-1} \in [0, \theta]$, $\min(\mathbf{v}_t \div \mathbf{v}_{t+1}) \in (0, 1]$, and $\theta \in [0, 1)$, we conclude that $\sigma_t \in [0, \theta]$.

We now establish the lower bound for σ_t. For all $t \geq 0$, we have:

$$\begin{aligned}\sigma_t \;&:=\; \theta(1-\sigma_{t-1}) \cdot \min(\mathbf{v}_t \div \mathbf{v}_{t+1}) \;\overset{①}{\geq}\; \theta(1-\theta) \cdot \min(\mathbf{v}_t \div \mathbf{v}_{t+1}) \\ &\overset{②}{\geq}\; \theta(1-\theta) \cdot \tfrac{\min(\mathbf{v}_t)}{\max(\mathbf{v}_{t+1})} \;=\; \theta(1-\theta) \cdot \tfrac{\min(\mathbf{v}_{t+1})}{\max(\mathbf{v}_{t+1})} \cdot \tfrac{\min(\mathbf{v}_t)}{\min(\mathbf{v}_{t+1})} \;\overset{③}{\geq}\; \theta(1-\theta) \cdot \tfrac{1}{\kappa} \cdot \tfrac{1}{\tilde{\kappa}},\end{aligned}$$

where step ① uses $\sigma_{t-1} \leq \theta$; step ② uses $\min(\mathbf{a} \div \mathbf{b}) \geq \frac{\min(\mathbf{a})}{\max(\mathbf{b})}$ for all $\mathbf{a} \geq \mathbf{0}$ and $\mathbf{b} > \mathbf{0}$; step ③ uses $\frac{\min(\mathbf{v}_{t+1})}{\max(\mathbf{v}_{t+1})} \geq \frac{1}{\kappa}$ and $\frac{\min(\mathbf{v}_t)}{\min(\mathbf{v}_{t+1})} \geq \frac{1}{\tilde{\kappa}}$ for all t, as shown in Lemma B.1**(b,c)**.

Part (b). For all $t \geq 0$, we derive the following results:

$$\begin{aligned}(\sigma_{t-1}-1)\mathbf{v}_t + \sigma_t\mathbf{v}_{t+1} \;&\overset{①}{=}\; (\sigma_{t-1}-1)\mathbf{v}_t + \theta(1-\sigma_{t-1}) \cdot \min(\mathbf{v}_t \div \mathbf{v}_{t+1})\mathbf{v}_{t+1} \\ &\overset{②}{\leq}\; (\sigma_{t-1}-1)\mathbf{v}_t + \theta(1-\sigma_{t-1})\mathbf{v}_t \;=\; -(1-\theta)(1-\sigma_{t-1})\mathbf{v}_t \\ &\overset{③}{\leq}\; -(1-\theta)(1-\theta)\mathbf{v}_t \;=\; -(1-\theta)^2\mathbf{v}_t,\end{aligned}$$

where step ① uses the choice for σ_t for all $t \geq 0$; step ② uses $\min(\mathbf{a} \div \mathbf{v})\mathbf{v} \leq \mathbf{a}$ for all $\mathbf{a}, \mathbf{v} \in \mathbb{R}^n$ with $\mathbf{v} > \mathbf{0}$; step ③ uses $\sigma_{t-1} \leq \theta < 1$ for all $t \geq 0$.

□

We let $\bar{\mathbf{x}} \in \arg\min_{\mathbf{x}} F(\mathbf{x})$, where $F(\mathbf{x}) := f(\mathbf{x}) + h(\mathbf{x})$. We define the following sequence:

$$\mathcal{Z}_t := \mathbb{E}[\widehat{\mathcal{Z}}_t],\; \widehat{\mathcal{Z}}_t := F(\mathbf{x}_t) - F(\bar{\mathbf{x}}) + \tfrac{1}{2}\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2_{\sigma_{t-1}(\mathbf{v}_t+L)}.$$

We now combine proximal optimality, smoothness, and the extrapolation inequality to obtain a pathwise Lyapunov estimate. Keeping the gradient-estimation error explicit allows the same estimate to serve both the deterministic and stochastic analyses.

Lemma B.3. *(*Properties of $\mathcal{Z}_t$) We define $c_1 := \frac{1}{2}(\frac{1-\theta}{\kappa})^2$, $c_2 := \frac{3L}{2}$. We have almost surely:

$$\widehat{\mathcal{Z}}_{t+1} - \widehat{\mathcal{Z}}_t \leq \langle \mathbf{d}_t, \nabla f(\mathbf{y}_t) - \mathbf{g}_t \rangle + c_2\mathbb{U}_t - c_1\mathbb{S}_t, \tag{10}$$

where $\mathbb{S}_t := \frac{\|\mathbf{r}_t\|_2^2}{\min(\mathbf{v}_t)}$, $\mathbb{U}_t := \frac{\|\mathbf{r}_t\|_2^2}{\min(\mathbf{v}_t)^2}$.

Proof. Let $\bar{\mathbf{x}} \in \arg\min_{\mathbf{x}} F(\mathbf{x})$, where $F(\mathbf{x}) := f(\mathbf{x}) + h(\mathbf{x})$.

Define $\mathbf{a}_t := \mathbf{y}_t - \mathbf{g}_t \div \mathbf{v}_t$, and $Q_t := \langle \mathbf{d}_t, \nabla f(\mathbf{y}_t) - \mathbf{g}_t \rangle$, where $\mathbf{d}_t := \mathbf{x}_{t+1} - \mathbf{x}_t$.

Define $\widehat{\mathcal{Z}}_t := F(\mathbf{x}_t) - F(\bar{\mathbf{x}}) + \frac{1}{2}\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2_{\sigma_{t-1}(\mathbf{v}_t+L)}$, and $\mathcal{X}^t := \frac{1}{2}\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2_{\sigma_{t-1}(\mathbf{v}_t+L)}$.

Using the optimality of $\mathbf{x}_{t+1} \in \arg\min_{\mathbf{x}} h(\mathbf{x}) + \frac{1}{2}\|\mathbf{x} - \mathbf{a}_t\|^2_{\mathbf{v}_t}$, we have the following inequality:

$$h(\mathbf{x}_{t+1}) + \tfrac{1}{2}\|\mathbf{x}_{t+1} - \mathbf{a}_t\|^2_{\mathbf{v}_t} \leq h(\mathbf{x}_t) + \tfrac{1}{2}\|\mathbf{x}_t - \mathbf{a}_t\|^2_{\mathbf{v}_t}. \tag{11}$$

Given $f(\mathbf{x})$ is L-smooth, we have:

$$f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) + \langle \mathbf{x}_{t+1} - \mathbf{x}_t, \nabla f(\mathbf{x}_t) \rangle + \tfrac{L}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|_2^2. \tag{12}$$

Adding Inequalities (11) and (12) together yields:

$$
\begin{aligned}
& F(\mathbf{x}_{t+1}) - F(\mathbf{x}_t) - \tfrac{L}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|_2^2 \\
\leq\ & \langle \mathbf{x}_{t+1} - \mathbf{x}_t, \nabla f(\mathbf{x}_t)\rangle + \tfrac{1}{2}\|\mathbf{x}_t - \mathbf{a}_t\|_{\mathbf{v}_t}^2 - \tfrac{1}{2}\|\mathbf{x}_{t+1} - \mathbf{a}_t\|_{\mathbf{v}_t}^2 \\
\overset{①}{=}\ & \langle \mathbf{x}_{t+1} - \mathbf{x}_t, \nabla f(\mathbf{x}_t)\rangle + \tfrac{1}{2}\|\mathbf{x}_t - \mathbf{x}_{t+1}\|_{\mathbf{v}_t}^2 + \langle \mathbf{a}_t - \mathbf{x}_{t+1}, \mathbf{x}_{t+1} - \mathbf{x}_t\rangle_{\mathbf{v}_t} \\
\overset{②}{=}\ & \langle \mathbf{x}_{t+1} - \mathbf{x}_t, \nabla f(\mathbf{x}_t) - \mathbf{g}_t\rangle + \tfrac{1}{2}\|\mathbf{x}_t - \mathbf{x}_{t+1}\|_{\mathbf{v}_t}^2 + \langle \mathbf{y}_t - \mathbf{x}_{t+1}, \mathbf{x}_{t+1} - \mathbf{x}_t\rangle_{\mathbf{v}_t} \\
\overset{③}{=}\ & Q_t + \langle \mathbf{x}_{t+1} - \mathbf{x}_t, \nabla f(\mathbf{x}_t) - \nabla f(\mathbf{y}_t)\rangle + \tfrac{1}{2}\|\mathbf{x}_t - \mathbf{y}_t\|_{\mathbf{v}_t}^2 - \tfrac{1}{2}\|\mathbf{y}_t - \mathbf{x}_{t+1}\|_{\mathbf{v}_t}^2 \\
\overset{④}{\leq}\ & Q_t + L\|\mathbf{x}_{t+1} - \mathbf{x}_t\|\|\mathbf{x}_t - \mathbf{y}_t\| + \tfrac{1}{2}\|\mathbf{x}_t - \mathbf{y}_t\|_{\mathbf{v}_t}^2 - \tfrac{1}{2}\|\mathbf{y}_t - \mathbf{x}_{t+1}\|_{\mathbf{v}_t}^2 \\
\overset{⑤}{=}\ & Q_t + \sigma_{t-1}L\|\mathbf{x}_{t+1} - \mathbf{x}_t\|\|\mathbf{x}_t - \mathbf{x}_{t-1}\| \\
& + \tfrac{(\sigma_{t-1})^2}{2}\|\mathbf{x}_t - \mathbf{x}_{t-1}\|_{\mathbf{v}_t}^2 - \tfrac{1}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t - \sigma_{t-1}(\mathbf{x}_t - \mathbf{x}_{t-1})\|_{\mathbf{v}_t}^2 \\
=\ & Q_t + \sigma_{t-1}L\|\mathbf{x}_{t+1} - \mathbf{x}_t\|\|\mathbf{x}_t - \mathbf{x}_{t-1}\| - \tfrac{1}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|_{\mathbf{v}_t}^2 + \sigma_{t-1}\langle \mathbf{x}_{t+1} - \mathbf{x}_t, \mathbf{x}_t - \mathbf{x}_{t-1}\rangle_{\mathbf{v}_t} \\
\overset{⑥}{\leq}\ & Q_t + \tfrac{\sigma_{t-1}L}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|_2^2 + \tfrac{\sigma_{t-1}L}{2}\|\mathbf{x}_t - \mathbf{x}_{t-1}\|_2^2 \\
& - \tfrac{1}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|_{\mathbf{v}_t}^2 + \tfrac{\sigma_{t-1}}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|_{\mathbf{v}_t}^2 + \tfrac{\sigma_{t-1}}{2}\|\mathbf{x}_t - \mathbf{x}_{t-1}\|_{\mathbf{v}_t}^2 \\
=\ & Q_t + -\tfrac{1}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|_{\mathbf{v}_t}^2 + \underbrace{\tfrac{1}{2}\|\mathbf{x}_t - \mathbf{x}_{t-1}\|_{\sigma_{t-1}(\mathbf{v}_t+L)}^2}_{:=\mathcal{X}^t} + \tfrac{1}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|_{\sigma_{t-1}(\mathbf{v}_t+L)}^2, \qquad (13)
\end{aligned}
$$

where step ① uses the Pythagoras Relation as in Lemma A.1 that $\mathbf{x}_+ = \mathbf{x}_{t+1}$, $\mathbf{x} = \mathbf{x}_t$, $\mathbf{a} = \mathbf{a}_t$, and $\mathbf{v} = \mathbf{v}_t$; step ② uses $\mathbf{a}_t := \mathbf{y}_t - \mathbf{g}_t \div \mathbf{v}_t$; step ③ uses the definition of Q_t, and the Pythagoras Relation as in Lemma A.1 that $\mathbf{x}_+ = \mathbf{x}_{t+1}$, $\mathbf{x} = \mathbf{x}_t$, $\mathbf{a} = \mathbf{y}_t$, and $\mathbf{v} = \mathbf{v}_t$; step ④ uses L-smoothness of $f(\cdot)$; step ⑤ uses $\mathbf{y}_t - \mathbf{x}_t = \sigma_{t-1}(\mathbf{x}_t - \mathbf{x}_{t-1})$; step ⑥ uses $ab \leq \frac{a^2}{2} + \frac{b^2}{2}$ for all $a, b \in \mathbb{R}$, and $\langle \mathbf{a}, \mathbf{b}\rangle_{\mathbf{v}} \leq \frac{1}{2}\|\mathbf{a}\|_{\mathbf{v}}^2 + \frac{1}{2}\|\mathbf{b}\|_{\mathbf{v}}^2$ for all $\mathbf{a}, \mathbf{b}, \mathbf{v} \in \mathbb{R}^n$ with $\mathbf{v} \geq \mathbf{0}$.

Define $\widehat{\mathcal{Z}}_t := F(\mathbf{x}_t) - F(\bar{\mathbf{x}}) + \frac{1}{2}\|\mathbf{x}_t - \mathbf{x}_{t-1}\|_{\sigma_{t-1}(\mathbf{v}_t+L)}^2$. Given Inequality (13), we have the following inequalities for all $t \geq 0$ almost surely:

$$
\begin{aligned}
& \widehat{\mathcal{Z}}_{t+1} - \widehat{\mathcal{Z}}_t - Q_t \\
\leq\ & \tfrac{L}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|_2^2 - \tfrac{1}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|_{\mathbf{v}_t}^2 + \tfrac{1}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|_{\sigma_{t-1}(\mathbf{v}_t+L)}^2 + \mathcal{X}^{t+1} \\
\overset{①}{=}\ & \tfrac{L}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|_2^2 - \tfrac{1}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|_{\mathbf{v}_t}^2 + \tfrac{1}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|_{[\sigma_{t-1}\mathbf{v}_t + \sigma_{t-1}L + \sigma_t\mathbf{v}_{t+1} + \sigma_t L]}^2 \\
=\ & \tfrac{L + \sigma_{t-1}L + \sigma_t L}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|_2^2 + \tfrac{1}{2}\langle \sigma_{t-1}\mathbf{v}_t + \sigma_t\mathbf{v}_{t+1} - \mathbf{v}_t, \mathbf{d}_t \odot \mathbf{d}_t\rangle \\
\overset{②}{\leq}\ & \tfrac{3L}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|_2^2 - \tfrac{1}{2}(1-\theta)^2\|\mathbf{x}_{t+1} - \mathbf{x}_t\|_{\mathbf{v}_t}^2 \\
\overset{③}{\leq}\ & \underbrace{\tfrac{3L}{2}}_{:=c_2} \cdot \underbrace{\tfrac{1}{\min(\mathbf{v}_t)^2}\|\mathbf{v}_t \odot (\mathbf{x}_{t+1} - \mathbf{x}_t)\|_2^2}_{:=\mathbb{U}_t} - \underbrace{\tfrac{1}{2}(\tfrac{1-\theta}{\kappa})^2}_{:=c_1} \cdot \underbrace{\tfrac{1}{\min(\mathbf{v}_t)}\|\mathbf{v}_t \odot (\mathbf{x}_{t+1} - \mathbf{x}_t)\|_2^2}_{:=\mathbb{S}_t},
\end{aligned}
$$

where step ① uses the definition of $\mathcal{X}^t := \frac{1}{2}\|\mathbf{x}_t - \mathbf{x}_{t-1}\|_{\sigma_{t-1}(\mathbf{v}_t+L)}^2$; step ② uses $\sigma_t \leq 1$ for all $t \geq 0$, and $\sigma_t\mathbf{v}_{t+1} + (\sigma_{t-1} - 1)\mathbf{v}_t \leq -(1-\theta)^2\mathbf{v}_t$ for all $t \geq 0$ as shown in Lemma B.2**(b)**; step ③ uses the following two inequalities for all $\mathbf{d} \in \mathbb{R}^n$ with $\mathbf{d} = \mathbf{x}_{t+1} - \mathbf{x}_t$:

$$
\begin{aligned}
\|\mathbf{d}\|_2^2 &\leq \tfrac{1}{\min(\mathbf{v}_t)^2}\|\mathbf{v}_t \odot \mathbf{d}\|_2^2, \\
\|\mathbf{d}\|_{\mathbf{v}_t}^2\kappa^2 &\geq \|\mathbf{d}\|_{\mathbf{v}_t}^2\frac{\max(\mathbf{v}_t)^2}{\min(\mathbf{v}_t)^2} \geq \|\mathbf{d}\|_2^2\frac{\max(\mathbf{v}_t)^2}{\min(\mathbf{v}_t)} \geq \frac{\|\mathbf{v}_t \odot \mathbf{d}\|_2^2}{\min(\mathbf{v}_t)}.
\end{aligned}
$$

□

To sum the Lyapunov inequality, we need cumulative bounds for its two adaptive terms. The next lemma expresses both sums in terms of the terminal accumulator, preparing the boundedness arguments that follow.

Lemma B.4. *We define $\mathcal{V}_t$ as in Lemma B.1. We have the following results.*

(a) $\sum_{t=0}^{T} \mathbb{S}_t \le s_1 \mathcal{V}_{T+1}$, *where* $s_1 := \frac{2\dot{\kappa}}{\alpha}$.

(b) $\sum_{t=0}^{T} \mathbb{U}_t \le s_2 \sqrt{\mathcal{V}_{T+1}}$, *where* $s_2 := \frac{4\dot{\kappa}^2}{\alpha} \underline{\mathrm{v}}^{-1/2}$.

Proof. We define $\mathcal{V}_{t+1} := \sqrt{\underline{\mathrm{v}}^2 + (\alpha+\beta) M_t}$, where $M_t := \sum_{i=0}^{t} \|\mathbf{r}_i\|_2^2$.

Part (a). We derive the following results:

$$\begin{aligned} \sum_{t=0}^{T} \frac{\|\mathbf{r}_t\|_2^2}{\min(\mathbf{v}_t)} &= \sum_{t=0}^{T} \frac{\|\mathbf{r}_t\|_2^2}{\min(\mathbf{v}_{t+1})} \cdot \frac{\min(\mathbf{v}_{t+1})}{\min(\mathbf{v}_t)} \overset{①}{\le} \dot{\kappa} \cdot \sum_{t=0}^{T} \frac{\|\mathbf{r}_t\|_2^2}{\sqrt{\underline{\mathrm{v}}^2 + \alpha \sum_{j=0}^{t} \|\mathbf{r}_j\|_2^2}} \\ &\overset{②}{\le} \frac{2\dot{\kappa}}{\alpha} \cdot \sqrt{\underline{\mathrm{v}}^2 + \alpha \sum_{t=0}^{T} \|\mathbf{r}_t\|_2^2} \le \underbrace{\frac{2\dot{\kappa}}{\alpha}}_{:=s_1} \cdot \underbrace{\sqrt{\underline{\mathrm{v}}^2 + (\alpha+\beta) \sum_{t=0}^{T} \|\mathbf{r}_t\|_2^2}}_{:=\mathcal{V}_{T+1}}, \end{aligned} \tag{14}$$

where step ① uses Lemma B.1**(c)** that $\min(\mathbf{v}_{t+1})/\min(\mathbf{v}_t) \le \dot{\kappa}$, and Lemma B.1**(a)**; step ② uses Lemma A.5.

Part (b). We have the following results:

$$\begin{aligned} \sum_{t=0}^{T} \frac{\|\mathbf{r}_t\|_2^2}{\min(\mathbf{v}_t)^2} &= \sum_{t=0}^{T} \frac{\|\mathbf{r}_t\|_2^2}{\min(\mathbf{v}_{t+1})^2} \cdot \left(\frac{\min(\mathbf{v}_{t+1})}{\min(\mathbf{v}_t)}\right)^2 \overset{①}{\le} \frac{\dot{\kappa}^2}{\alpha} \cdot \sum_{t=0}^{T} \frac{\|\mathbf{r}_t\|_2^2}{\underline{\mathrm{v}}^2/\alpha + \sum_{j=0}^{t} \|\mathbf{r}_j\|_2^2} \\ &\overset{②}{\le} \frac{\dot{\kappa}^2}{\alpha} \cdot \frac{1}{1/4} \cdot \left(1 + \frac{1}{\underline{\mathrm{v}}^2/\alpha} \sum_{t=0}^{T} \|\mathbf{r}_t\|_2^2\right)^{1/4} = \frac{4\dot{\kappa}^2}{\alpha \cdot \underline{\mathrm{v}}^{1/2}} \cdot \left(\underline{\mathrm{v}}^2 + \alpha \sum_{t=0}^{T} \|\mathbf{r}_t\|_2^2\right)^{1/4} \\ &\overset{③}{\le} \underbrace{\frac{4\dot{\kappa}^2}{\alpha \cdot \underline{\mathrm{v}}^{1/2}}}_{:=s_2} \cdot \underbrace{\left(\underline{\mathrm{v}}^2 + (\alpha+\beta) \sum_{t=0}^{T} \|\mathbf{r}_t\|_2^2\right)^{1/4}}_{:=\sqrt{\mathcal{V}_{T+1}}}, \end{aligned}$$

where step ① uses Lemma B.1**(a)** and Lemma B.1**(c)**; step ② uses Lemma A.6 with $p = 1/4$, $A_t = \|\mathbf{r}_t\|_2^2$, and $c = \underline{\mathrm{v}}^2/\alpha$; step ③ uses $\beta \ge 0$.

□

B.2 Analysis for ALFPG-D

This subsection provides the convergence analysis of ALFPG-D.

For the full-batch method, the gradient-estimation error vanishes. Combining the preceding descent and summation estimates yields coupled bounds for the Lyapunov sequence and the metric accumulator; the next lemma closes these bounds uniformly in the iteration horizon.

Lemma B.5. *(*Boundedness of $\mathcal{Z}_t$ and $\mathcal{V}_t$*)* For all $t \ge 0$, we obtain: **(a)** It holds $\mathcal{Z}_t \le \overline{\mathcal{Z}}$ for some positive constant $\overline{\mathcal{Z}}$. **(b)** It holds $\mathcal{V}_t \le \overline{\mathrm{v}}$ for some positive constant $\overline{\mathrm{v}}$.

Proof. We define $\mathcal{V}_{t+1} := \sqrt{\underline{\mathrm{v}}^2 + (\alpha+\beta) M_t}$, where $M_t := \sum_{i=0}^{t} \|\mathbf{r}_i\|_2^2$.

We define $\mathbf{r}_t := \mathbf{v}_t \odot \mathbf{d}_t$, where $\mathbf{d}_t := \mathbf{x}_{t+1} - \mathbf{x}_t$.

Initially, for the full-batch, deterministic setting where $\nabla f(\mathbf{y}_t) = \mathbf{g}_t$, we obtain from Lemma B.3 that

$$0 \le \mathcal{Z}_t - \mathcal{Z}_{t+1} + \frac{c_2 \|\mathbf{r}_t\|_2^2}{\min(\mathbf{v}_t)^2} - \frac{c_1 \|\mathbf{r}_t\|_2^2}{\min(\mathbf{v}_t)} \tag{15}$$

Multiplying both sides of Inequality (15) by $\min(\mathbf{v}_t)$ yields:

$$0 \le -c_1 \|\mathbf{r}_t\|_2^2 + \min(\mathbf{v}_t)(\mathcal{Z}_t - \mathcal{Z}_{t+1}) + \frac{c_2 \|\mathbf{r}_t\|_2^2}{\min(\mathbf{v}_t)}.$$

Summing this inequality over t from $t = 0$ to T, we obtain:

$$\begin{aligned} 0 &\le -c_1 \sum_{t=0}^{T} \|\mathbf{r}_t\|_2^2 + \sum_{t=0}^{T} \min(\mathbf{v}_t)(\mathcal{Z}_t - \mathcal{Z}_{t+1}) + c_2 \sum_{t=0}^{T} \frac{\|\mathbf{r}_t\|_2^2}{\min(\mathbf{v}_t)} \\ &\overset{①}{\le} -c_1 \sum_{t=0}^{T} \|\mathbf{r}_t\|_2^2 + (\max_{t=0}^{T} \mathcal{Z}_t) \cdot \min(\mathbf{v}_T) + c_2 s_1 \mathcal{V}_{T+1} \\ &\overset{②}{\le} -\frac{c_1}{\alpha+\beta} \left(\mathcal{V}_{T+1}^2 - \underline{\mathrm{v}}^2\right) + (\max_{t=0}^{T} \mathcal{Z}_t) \cdot \mathcal{V}_{T+1} + c_2 s_1 \mathcal{V}_{T+1}, \end{aligned} \tag{16}$$

where step ① uses Lemma A.4 with $A_i = \min(\mathbf{v}_i)$ for all $i \in [T]$ with $A_1 \le A_2 \le \ldots \le A_T$, and $B_j = \mathcal{Z}_j$ for all $j \ge 0$, and Lemma B.4 that $\sum_{t=0}^{T} \mathbb{S}_t \le s_1 \mathcal{V}_{T+1}$; step ② uses the definition of $\mathcal{V}_T$, along with the facts that $\mathbf{v}_t \le \mathbf{v}_{t+1} \le \mathcal{V}_{t+1}$ and $\mathcal{Z}_i \ge 0$.

Notably, the inequality $\sum_{t=0}^{T} \min(\mathbf{v}_t)(\mathcal{Z}_t - \mathcal{Z}_{t+1}) \le \min(\mathbf{v}^T) \sum_{t=0}^{T} (\mathcal{Z}_t - \mathcal{Z}_{t+1})$ does not hold in general, as the sequence $\{\mathcal{Z}_t\}_{t=0}^{T}$ is not necessarily monotonic. To address this, we instead employ the alternative inequality which can be implied by Lemma A.4: $\sum_{t=0}^{T} \min(\mathbf{v}_t)(\mathcal{Z}_t - \mathcal{Z}_{t+1}) \le \min(\mathbf{v}_T) \max_{t=0}^{T} \mathcal{Z}_t$.

In view of Inequality (16), we have the following quadratic inequality for all $T \ge 0$:

$$\tfrac{c_1}{\alpha+\beta}(\mathcal{V}_{T+1})^2 \le \left(c_2 s_1 + \max_{t=0}^{T} \mathcal{Z}_t\right) \cdot \mathcal{V}_{T+1} + \tfrac{c_1}{\alpha+\beta} \underline{\mathrm{v}}^2.$$

Applying Lemma A.3 with $a = \frac{c_1}{\alpha+\beta}$, $b = c_2 s_1 + \max_{t=0}^{T} \mathcal{Z}_t$, $c = \frac{c_1}{\alpha+\beta}\underline{\mathrm{v}}^2$, and $x = \mathcal{V}_{T+1}$ yields:

$$\begin{aligned} \mathcal{V}_{T+1} &\le \sqrt{c/a} + b/a \;=\; \underline{\mathrm{v}} + \tfrac{\alpha+\beta}{c_1} \cdot \left(c_2 s_1 + \max_{t=0}^{T} \mathcal{Z}_t\right) \\ &= \underbrace{\underline{\mathrm{v}} + \tfrac{\alpha+\beta}{c_1} \cdot c_2 s_1}_{:=w_1} + \underbrace{\tfrac{\alpha+\beta}{c_1}}_{:=w_2} \cdot \max_{t=0}^{T} \mathcal{Z}_t, \end{aligned} \tag{17}$$

The upper bound for $\mathcal{V}_{T+1}$ is established in Inequality (17), but it depends on the unknown variable $(\max_{t=0}^{T} \mathcal{Z}_t)$.

Part (a). We now show that $(\max_{t=0}^{T} \mathcal{Z}_t)$ is always bounded above by a universal constant $\overline{\mathcal{Z}}$. Dropping the negative term $-\frac{c_1 \|\mathbf{r}_t\|_2^2}{\min(\mathbf{v}_t)}$ on the right-hand side of Inequality (15) and summing over t from $t = 0$ to T yields:

$$\begin{aligned} \mathcal{Z}_{T+1} &\le \mathcal{Z}_0 + c_2 \textstyle\sum_{t=0}^{T} \frac{\|\mathbf{r}_t\|_2^2}{\min(\mathbf{v}_t)^2} \\ &\overset{①}{\le} \mathcal{Z}_0 + c_2 s_2 \sqrt{\mathcal{V}_{T+1}} \\ &\overset{②}{\le} \mathcal{Z}_0 + c_2 s_2 \sqrt{w_1 + w_2 \max_{t=0}^{T} \mathcal{Z}_t} \\ &\overset{③}{\le} \underbrace{\mathcal{Z}_0 + c_2 s_2 \sqrt{w_1}}_{:=\dot{a}} + \underbrace{c_2 s_2 \sqrt{w_2}}_{:=\dot{b}} \cdot \sqrt{\max_{t=0}^{T} \mathcal{Z}_t} \qquad (18) \\ &\overset{④}{\le} \max(\mathcal{Z}_0, 2\dot{b}^2 + 2\dot{a}) \\ &\overset{⑤}{=} 2\dot{b}^2 + 2\dot{a}, \qquad (19) \end{aligned}$$

where step ① uses Lemma B.4**(b)**; step ② uses Inequality (17); step ③ uses $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ for all $a, b \ge 0$; step ④ uses Lemma A.7 with $a = \dot{a}$ and $b = \dot{b}$; step ⑤ uses the fact that $\dot{a} \ge \mathcal{Z}_0$. Thus one admissible explicit choice is

$$\overline{\mathcal{Z}} := 2\dot{b}^2 + 2\dot{a} = 2c_2^2 s_2^2 w_2 + 2\mathcal{Z}_0 + 2c_2 s_2 \sqrt{w_1}.$$

This constant bounds every $\mathcal{Z}_t$, including $\mathcal{Z}_0$, because $\dot{a} \ge \mathcal{Z}_0$. All quantities on the right were specified above and are independent of the iteration horizon.

This further leads to:

$$\begin{aligned} \mathcal{Z}_{T+1} &\le 2\dot{b}^2 + 2\dot{a} && (20) \\ &\le \tfrac{9L^2}{2} s_2^2 w_2 + 2\mathcal{Z}_0 + 3L s_2 \sqrt{w_1} && (21) \\ &\le \mathcal{O}(L^2 \underline{\mathrm{v}}^{-1}) + \mathcal{O}(\mathcal{Z}_0) + \mathcal{O}(L \underline{\mathrm{v}}^{-1/2}) \sqrt{\mathcal{O}(\underline{\mathrm{v}}) + \mathcal{O}(L)} := \overline{\mathcal{Z}} && (22) \\ &\le \mathcal{O}(1), && (23) \end{aligned}$$

where we use $s_1 := \frac{2\dot{\kappa}}{\alpha} = \mathcal{O}(1)$, $s_2 := \frac{4\dot{\kappa}^2}{\alpha} \underline{\mathrm{v}}^{-1/2} = \mathcal{O}(\underline{\mathrm{v}}^{-1/2})$, $c_1 := \frac{1}{2}(\frac{1-\theta}{\kappa})^2 = \mathcal{O}(1)$, $c_2 := \frac{3L}{2} = \mathcal{O}(L)$, $s_2 := \frac{4\dot{\kappa}^2}{\alpha \cdot \underline{\mathrm{v}}^{1/2}} = \mathcal{O}(\underline{\mathrm{v}}^{-1/2})$, $w_2 = \frac{\alpha+\beta}{c_1} = \mathcal{O}(1)$, $w_1 = \underline{\mathrm{v}} + \frac{\alpha+\beta}{c_1} \cdot c_2 s_1 = \mathcal{O}(\underline{\mathrm{v}}) + \mathcal{O}(L)$.

Part (b). We derive the following inequalities for all $T \geq 0$:

$$\begin{aligned}\mathcal{V}_{T+1} &\overset{①}{\leq} w_1 + w_2 \max_{t=0}^{T} \mathcal{Z}_t \\ &\overset{②}{\leq} w_1 + w_2 \overline{\mathcal{Z}} \\ &\leq \mathcal{O}(\underline{\mathrm{v}}) + \mathcal{O}(L) + \mathcal{O}(L^2 \underline{\mathrm{v}}^{-1}) + \mathcal{O}(\mathcal{Z}_0) + \mathcal{O}(L \underline{\mathrm{v}}^{-1/2}) \sqrt{\mathcal{O}(\underline{\mathrm{v}}) + \mathcal{O}(L)} := \overline{\mathrm{v}} \\ &\leq \mathcal{O}(1),\end{aligned}$$

where step ① uses Inequality (17); step ② uses the bound established in **Part (a)**.

□

B.2.1 Proof of Theorem 4.4

The metric boundedness established above makes the squared proximal displacements summable. We first record this bound and then use proximal optimality to convert it into the stationarity estimate of Theorem 4.4.

Proof. Define $\mathbf{d}_t := \mathbf{x}_{t+1} - \mathbf{x}_t$ and $\mathbf{r}_t := \mathbf{v}_t \odot \mathbf{d}_t$.

Part (a). We have the following inequalities:

$$\begin{aligned}\sum_{t=0}^{T} \|\mathbf{x}_{t+1} - \mathbf{x}_t\|_2^2 &\overset{①}{\leq} \tfrac{1}{\underline{\mathrm{v}}^2} \sum_{t=0}^{T} \min(\mathbf{v}_t)^2 \|\mathbf{x}_{t+1} - \mathbf{x}_t\|_2^2 \\ &\overset{②}{\leq} \tfrac{1}{\underline{\mathrm{v}}^2} \sum_{t=0}^{T} \|\mathbf{v}_t \odot (\mathbf{x}_{t+1} - \mathbf{x}_t)\|_2^2 \\ &\overset{③}{=} \tfrac{1}{\underline{\mathrm{v}}^2} \sum_{t=0}^{T} \|\mathbf{r}_t\|_2^2 = \tfrac{1}{\underline{\mathrm{v}}^2} M_T \\ &\overset{④}{\leq} \tfrac{1}{\underline{\mathrm{v}}^2} \tfrac{1}{\alpha} (\overline{\mathrm{v}}^2 - \underline{\mathrm{v}}^2) := \overline{\mathrm{X}},\end{aligned} \tag{24}$$

where step ① uses $\mathbf{v}_t \geq \underline{\mathrm{v}}$; step ② uses $\min(\mathbf{v})\|\mathbf{d}\| \leq \|\mathbf{v} \odot \mathbf{d}\|$ for all $\mathbf{v}, \mathbf{d} \in \mathbb{R}^n$ with $\mathbf{v} \geq \mathbf{0}$; step ③ uses the definition of $\mathbf{r}_t := \mathbf{v}_t \odot \mathbf{d}_t$; step ④ uses Lemma B.1**(a)** that $\underline{\mathrm{v}}^2 + \alpha M_t \leq (\mathbf{v}_{t+1})^2 \leq \overline{\mathrm{v}}^2$ for all t.

Part (b). First, by the first-order necessary condition of $\mathbf{x}_{t+1}$ that $\mathbf{x}_{t+1} \in \mathrm{Prox}_h(\mathbf{y} - \mathbf{g}_t \div \mathbf{v}_t; \mathbf{v}_t) = \arg\min_{\mathbf{x}} h(\mathbf{x}) + \frac{1}{2}\|\mathbf{x} - (\mathbf{y} - \mathbf{g}_t \div \mathbf{v}_t)\|_{\mathbf{v}_t}^2$, we have:

$$\mathbf{0} \in \partial h(\mathbf{x}_{t+1}) + \mathbf{g}_t + \mathbf{v}_t \odot (\mathbf{x}_{t+1} - \mathbf{y}_t). \tag{25}$$

Second, we obtain:

$$\begin{aligned}\|\nabla f(\mathbf{x}_{t+1}) + \partial h(\mathbf{x}_{t+1})\| &\overset{①}{\leq} \|\nabla f(\mathbf{x}_{t+1}) - \nabla f(\mathbf{y}_t) - \mathbf{v}_t \odot (\mathbf{x}_{t+1} - \mathbf{y}_t)\| \\ &\overset{②}{\leq} L\|\mathbf{y}_t - \mathbf{x}_{t+1}\| + \max(\mathbf{v}_t)\|\mathbf{y}_t - \mathbf{x}_{t+1}\| \\ &\overset{③}{=} (L + \max(\mathbf{v}_t)) \cdot \|\mathbf{x}_t + \sigma_{t-1}(\mathbf{x}_t - \mathbf{x}_{t-1}) - \mathbf{x}_{t+1}\| \\ &\overset{④}{\leq} (L + \overline{\mathrm{v}}) \cdot (\|\mathbf{x}_t - \mathbf{x}_{t+1}\| + \|\mathbf{x}_t - \mathbf{x}_{t-1}\|),\end{aligned} \tag{26}$$

where step ① uses the inclusion in (25) with $\mathbf{g}_t = \nabla f(\mathbf{y}_t)$, as in ALFPG-D; step ② uses the triangle inequality, the L-smoothness of $f(\cdot)$, and $\|\mathbf{v}_t \odot \mathbf{a}\| \leq \max(\mathbf{v}_t)\|\mathbf{a}\|$ for all $\mathbf{a} \in \mathbb{R}^n$; step ③ uses $\mathbf{y}_t = \mathbf{x}_t + \sigma_{t-1}(\mathbf{x}_t - \mathbf{x}_{t-1})$; step ④ uses $\mathbf{v}_t \leq \overline{\mathrm{v}}$, the triangle inequality, and $\sigma_{t-1} \leq 1$ for all t.

Third, we obtain the following results:

$$\begin{aligned}&\sum_{t=0}^{T} \|\partial h(\mathbf{x}_{t+1}) + \nabla f(\mathbf{x}_{t+1})\|_2^2 \\ \overset{①}{\leq}\ & 2(L + \overline{\mathrm{v}})^2 \cdot \textstyle\sum_{t=0}^{T} (\|\mathbf{x}_{t+1} - \mathbf{x}_t\|_2^2 + \|\mathbf{x}_t - \mathbf{x}_{t-1}\|_2^2) \\ =\ & 2(L + \overline{\mathrm{v}})^2 \cdot \{\textstyle\sum_{t=0}^{T} \|\mathbf{x}_{t+1} - \mathbf{x}_t\|_2^2 + \sum_{t=-1}^{T-1} \|\mathbf{x}_{t+1} - \mathbf{x}_t\|_2^2\} \\ =\ & 2(L + \overline{\mathrm{v}})^2 \cdot \{\|\mathbf{x}_{-1} - \mathbf{x}_0\|_2^2 - \|\mathbf{x}_{T+1} - \mathbf{x}_T\|_2^2 + 2\textstyle\sum_{t=0}^{T} \|\mathbf{x}_{t+1} - \mathbf{x}_t\|_2^2\} \\ \overset{②}{\leq}\ & 4(L + \overline{\mathrm{v}})^2 \cdot \textstyle\sum_{t=0}^{T} \|\mathbf{x}_{t+1} - \mathbf{x}_t\|_2^2 \\ \overset{③}{\leq}\ & 4(L + \overline{\mathrm{v}})^2 \cdot \overline{\mathrm{X}} = \mathcal{O}(1),\end{aligned} \tag{27}$$

where step ① uses Inequality (26); step ② uses the choice $\mathbf{x}_{-1} = \mathbf{x}_0$ as shown in Algorithm 1, and $-\|\mathbf{x}_{T+1} - \mathbf{x}_T\|^2 \leq 0$; step ③ uses Inequality (24).

Finally, using the inequality $\|\mathbf{a}\|_2^2 \geq \frac{1}{T+1}(\|\mathbf{a}\|_1)^2$ for all $\mathbf{a} \in \mathbb{R}^{T+1}$, we deduce from Inequality (27) that

$$\frac{1}{T+1}\sum_{t=0}^{T} \|\partial h(\mathbf{x}_{t+1}) + \nabla f(\mathbf{x}_{t+1})\| = \mathcal{O}(\tfrac{1}{\sqrt{T+1}}).$$

□

B.3 Analysis for ALFPG-FS

This subsection provides the convergence analysis of ALFPG-FS.

We fix $q \geq 1$. For all $t \geq 0$, we denote $r_t := \lfloor \frac{t}{q} \rfloor + 1$ [1], leading to $(r_t - 1)q \leq t \leq r_t q - 1$.

We first recall the SPIDER variance recursion from Fang et al. (2018). It relates the estimation error accumulated between refreshes to the motion of the extrapolated iterates.

Lemma B.6. *(*Lemma 1 in Fang et al. (2018)) The SPIDER estimator produces a stochastic gradient $\mathbf{g}_t$ that, for all t with $(r_t - 1)q + 1 \leq t \leq r_t q - 1$, we have: $\mathbb{E}[\|\mathbf{g}_t - \nabla f(\mathbf{y}_t)\|_2^2] - \mathbb{E}[\|\mathbf{g}_{t-1} - \nabla f(\mathbf{y}_{t-1})\|_2^2] \leq \frac{L^2}{b} Y_{t-1}$, where $Y_i := \mathbb{E}[\|\mathbf{y}_{i+1} - \mathbf{y}_i\|_2^2]$.

At a refresh iteration, $\mathbf{g}_t = \nabla f(\mathbf{y}_t)$ and the estimation error is zero.

Telescoping the SPIDER recursion within one refresh interval bounds the current estimation error. Combining that bound with the common Lyapunov inequality gives the stochastic descent estimate below.

Lemma B.7. *For any positive constant ϕ, we define $c_1 := \frac{1}{2}(\frac{1-\theta}{\kappa})^2$, $c_2' := \frac{(3+\phi)L}{2}$, $c_3 := \frac{L}{2\phi}\frac{q}{b}$. We define $Y_i := \mathbb{E}[\|\mathbf{y}_{i+1} - \mathbf{y}_i\|_2^2]$. For all t with $(r_t - 1)q \leq t \leq r_t q - 1$, we have:*

(a) $\mathbb{E}[\|\mathbf{g}_t - \nabla f(\mathbf{y}_t)\|_2^2] \leq \frac{L^2}{b}\sum_{i=(r_t-1)q}^{t-1} Y_i$.

(b) $\mathcal{Z}_{t+1} - \mathcal{Z}_t + \mathbb{E}[c_1\mathbb{S}_t - c_2'\mathbb{U}_t] \leq \frac{c_3}{q}\sum_{i=(r_t-1)q}^{t-1} Y_i$.

Proof. We define $c_1 := \frac{1}{2}(\frac{1-\theta}{\kappa})^2$, $c_2' := \frac{(3+\phi)L}{2}$, and $c_3 := \frac{L}{2\phi}\frac{q}{b}$, where $\phi > 0$ can be any constant.

We define $\mathcal{Z}_t := \mathbb{E}[F(\mathbf{x}_t) - F(\bar{\mathbf{x}}) + \frac{1}{2}\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2_{\sigma_{t-1}(\mathbf{v}_t + L)}]$.

We define $Y_i := \mathbb{E}[\|\mathbf{y}_{i+1} - \mathbf{y}_i\|_2^2]$.

We define $\mathbb{S}_t := \frac{\|\mathbf{r}_t\|_2^2}{\min(\mathbf{v}_t)}$, and $\mathbb{U}_t := \frac{\|\mathbf{r}_t\|_2^2}{\min(\mathbf{v}_t)^2}$.

Part (a). Telescoping the inequality $\mathbb{E}[\|\mathbf{g}_t - \nabla f(\mathbf{y}_t)\|_2^2] - \mathbb{E}[\|\mathbf{g}_{t-1} - \nabla f(\mathbf{y}_{t-1})\|_2^2] \leq \frac{L^2}{b}\mathbb{E}[\|\mathbf{y}_t - \mathbf{y}_{t-1}\|_2^2]$ (as stated in Lemma B.6) over t from $(r_t - 1)q + 1$ to t, where $t \leq r_t q - 1$, we obtain:

$$\begin{aligned} & \mathbb{E}[\|\mathbf{g}_t - \nabla f(\mathbf{y}_t)\|_2^2] \\ \leq\ & \mathbb{E}[\|\mathbf{g}^{(r_t-1)q} - \nabla f(\mathbf{y}^{(r_t-1)q})\|_2^2] + \tfrac{L^2}{b}\textstyle\sum_{i=(r_t-1)q+1}^{t} \mathbb{E}[\|\mathbf{y}_i - \mathbf{y}_{i-1}\|_2^2] \\ \overset{①}{=}\ & 0 + \tfrac{L^2}{b}\textstyle\sum_{i=(r_t-1)q}^{t-1} \underbrace{\mathbb{E}[\|\mathbf{y}_{i+1} - \mathbf{y}_i\|_2^2]}_{:=Y_i}, \end{aligned} \tag{28}$$

where step ① uses $\mathbf{g}_j = \nabla f(\mathbf{y}_j)$ when j is a multiple of q. Notably, Inequality (28) holds for every t of the form $t = (r_t - 1)q$, since at these points we have $\mathbf{g}_t = \nabla f(\mathbf{y}_t)$.

[1] For example, if $q = 3$ and $t \in \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$, then the corresponding values of r_t are $\{1, 1, 1, 2, 2, 2, 3, 3, 3, 4\}$.

Part (b). For every t, Lemma B.3 and Young's inequality give

$$\begin{aligned}\mathcal{Z}_{t+1} - \mathcal{Z}_t + c_1\mathbb{E}[\mathbb{S}_t] &\leq \tfrac{3L}{2}\mathbb{E}[\mathbb{U}_t] + \tfrac{\phi L}{2}\mathbb{E}\|\mathbf{d}_t\|^2 + \tfrac{1}{2\phi L}\mathbb{E}\|\nabla f(\mathbf{y}_t) - \mathbf{g}_t\|^2 \\ &\leq c_2'\mathbb{E}[\mathbb{U}_t] + \tfrac{c_3}{q}\textstyle\sum_{i=(r_t-1)q}^{t-1} Y_i, \end{aligned} \tag{29}$$

where $\phi > 0$, $c_2' = (3+\phi)L/2$, and $c_3 = Lq/(2\phi b)$. The second inequality uses $\|\mathbf{d}_t\|^2 \leq \mathbb{U}_t$ and Part (a). We may take $L > 0$ without loss of generality by enlarging a zero smoothness bound.

□

The stochastic descent estimate involves motion at the extrapolated points rather than at the proximal iterates. The next lemma converts this motion into the adaptive displacement sums already bounded in Lemma B.4.

Lemma B.8. *We define* $V_t := \min(\mathbf{v}_t)$, *and* $Y_i := \mathbb{E}[\|\mathbf{y}_{i+1} - \mathbf{y}_i\|_2^2]$. *We have:*

(a) $\mathbb{E}\Big[\sum_{t=0}^T V_t\|\mathbf{y}_{t+1} - \mathbf{y}_t\|_2^2\Big] \leq u_1\mathbb{E}[\mathcal{V}_{T+1}]$, *where* $u_1 := (8 + 2\dot{\kappa})s_1$.

(b) $\sum_{t=0}^T Y_t \leq u_2\mathbb{E}[\sqrt{\mathcal{V}_{T+1}}]$, *where* $u_2 := 10s_2$.

Proof. We define $V_t := \min(\mathbf{v}_t)$, and $Y_i := \mathbb{E}[\|\mathbf{y}_{i+1} - \mathbf{y}_i\|_2^2]$.

First, for all $\tau > 0$, we have the following results:

$$\begin{aligned}\|\mathbf{y}_{t+1} - \mathbf{y}_t\|_2^2 &\overset{①}{=} \|(\mathbf{x}_{t+1} + \sigma_t\mathbf{d}_t) - (\mathbf{x}_t + \sigma_{t-1}\mathbf{d}_{t-1})\|_2^2 \\ &= \|(1+\sigma_t)\mathbf{d}_t - \sigma_{t-1}\mathbf{d}_{t-1}\|_2^2 \\ &\overset{②}{\leq} [(1+\tau)\|(1+\sigma_t)\mathbf{d}_t\|_2^2 + (1+1/\tau)\|\sigma_{t-1}\mathbf{d}_{t-1}\|_2^2] \\ &\overset{③}{\leq} 4(1+\tau)\|\mathbf{d}_t\|_2^2 + (1+1/\tau)\|\mathbf{d}_{t-1}\|_2^2, \end{aligned} \tag{30}$$

where step ① uses $\mathbf{y}_{t+1} = \mathbf{x}_{t+1} + \sigma_t\mathbf{d}_t$; step ② uses $\|\mathbf{a}+\mathbf{b}\|_2^2 \leq (1+\tau)\|\mathbf{a}\|_2^2 + (1+1/\tau)\|\mathbf{b}\|_2^2$ for all $\tau > 0$; step ③ uses $\sigma_t \leq \theta < 1$.

Second, we obtain the following inequalities:

$$\begin{aligned}\min(\mathbf{v}_t)\|\mathbf{y}_{t+1} - \mathbf{y}_t\|_2^2 &\overset{①}{\leq} \min(\mathbf{v}_t)\cdot[4(1+\tau)\|\mathbf{d}_t\|_2^2 + (1+1/\tau)\|\mathbf{d}_{t-1}\|_2^2] \\ &\overset{②}{\leq} 4(1+\tau)\min(\mathbf{v}_t)\|\mathbf{d}_t\|_2^2 + (1+1/\tau)\dot{\kappa}\min(\mathbf{v}_{t-1})\|\mathbf{d}_{t-1}\|_2^2 \\ &\overset{③}{\leq} (4+\dot{\kappa})\left(\min(\mathbf{v}_t)\|\mathbf{d}_t\|_2^2 + \min(\mathbf{v}_{t-1})\|\mathbf{d}_{t-1}\|_2^2\right), \end{aligned} \tag{31}$$

where step ① uses Inequality (30); step ② uses $\min(\mathbf{v}_t) \leq \min(\mathbf{v}_{t-1})\dot{\kappa}$, as shown in Lemma B.1**(c)**; step ③ uses the choice $\tau = \frac{\dot{\kappa}}{4}$. Note that Inequality (31) also holds at $t = 0$ with $\mathbf{v}_{-1} = \mathbf{v}_0$ and $\mathbf{d}_{-1} = \mathbf{0}$.

Part (a). We have the following inequalities:

$$\begin{aligned}\textstyle\sum_{t=0}^T \min(\mathbf{v}_t)\|\mathbf{y}_{t+1} - \mathbf{y}_t\|_2^2 &\overset{①}{\leq} (4+\dot{\kappa})\left(\textstyle\sum_{t=0}^T \min(\mathbf{v}_t)\|\mathbf{d}_t\|_2^2 + \sum_{t=0}^T \min(\mathbf{v}_{t-1})\|\mathbf{d}_{t-1}\|_2^2\right) \\ &= (4+\dot{\kappa})\left(\textstyle\sum_{t=0}^T \min(\mathbf{v}_t)\|\mathbf{d}_t\|_2^2 + \sum_{t=-1}^{T-1} \min(\mathbf{v}_t)\|\mathbf{d}_t\|_2^2\right) \\ &\overset{②}{=} (4+\dot{\kappa})\left(\textstyle\sum_{t=0}^T \min(\mathbf{v}_t)\|\mathbf{d}_t\|_2^2 + \sum_{t=0}^{T-1} \min(\mathbf{v}_t)\|\mathbf{d}_t\|_2^2\right) \\ &\leq (8+2\dot{\kappa})\left(\textstyle\sum_{t=0}^T \min(\mathbf{v}_t)\|\mathbf{d}_t\|_2^2\right) \\ &\overset{③}{\leq} (8+2\dot{\kappa})\textstyle\sum_{t=0}^T \min(\mathbf{v}_t)\cdot\frac{1}{\min(\mathbf{v}_t)^2}\|\mathbf{v}_t \odot \mathbf{d}_t\|_2^2 \\ &\overset{④}{\leq} \underbrace{(8+2\dot{\kappa})\cdot s_1}_{:=u_1}\cdot\mathcal{V}_{T+1}, \end{aligned} \tag{32}$$

where step ① uses Inequality (31); step ② uses $\mathbf{d}_{-1} = \mathbf{x}_0 - \mathbf{x}_{-1} = \mathbf{0}$; step ③ uses $\min(\mathbf{v}_t)\|\mathbf{d}_t\| \leq \|\mathbf{v}_t \odot \mathbf{d}_t\|$; step ④ uses $\sum_{t=0}^T \mathbb{S}_t \leq s_1\mathcal{V}_{T+1}$ with $\mathbb{S}_t := \frac{\|\mathbf{r}_t\|_2^2}{\min(\mathbf{v}_t)}$, as shown in Lemma B.4**(a)**.

Part (b). We obtain the following inequalities:

$$\begin{aligned}\sum_{t=0}^{T}\|\mathbf{y}_{t+1}-\mathbf{y}_t\|_2^2 &\overset{①}{\le} \sum_{t=0}^{T}\left(8\|\mathbf{d}_t\|_2^2+2\|\mathbf{d}_{t-1}\|_2^2\right)\\ &\overset{②}{=} 8\left(\sum_{t=0}^{T}\|\mathbf{d}_t\|_2^2\right)+2\left(\sum_{t=1}^{T}\|\mathbf{d}_{t-1}\|_2^2\right)\\ &= 8\left(\sum_{t=0}^{T}\|\mathbf{d}_t\|_2^2\right)+2\left(\sum_{t=0}^{T-1}\|\mathbf{d}_t\|_2^2\right)\\ &\le 10\left(\sum_{t=0}^{T}\|\mathbf{d}_t\|_2^2\right)\\ &\overset{③}{\le} 10\left(\sum_{t=0}^{T}\tfrac{1}{\min(\mathbf{v}_t)^2}\|\mathbf{v}_t\odot\mathbf{d}_t\|_2^2\right)\\ &\overset{④}{\le} \underbrace{10\cdot s_2}_{:=u_2}\cdot\sqrt{\mathcal{V}_{T+1}},\end{aligned}$$

where step ① uses Inequality (30) with $\tau=1$; step ② uses $\mathbf{d}_{-1}=\mathbf{x}_0-\mathbf{x}_{-1}=\mathbf{0}$; step ③ uses $\min(\mathbf{v}_t)\|\mathbf{d}_t\|\le\|\mathbf{v}_t\odot\mathbf{d}_t\|$; step ④ uses $\sum_{t=0}^{T}\mathbb{U}_t\le s_2\sqrt{\mathcal{V}_{T+1}}$ with $\mathbb{U}_t:=\frac{\|\mathbf{r}_t\|_2^2}{\min(\mathbf{v}_t)^2}$, as shown in Lemma B.4**(b)**.

Taking expectations in the two pathwise bounds proves the stated results. □

We next count how often each motion term occurs in the sums over a refresh interval. In particular, part (c) controls the total SPIDER error without introducing the metric-growth factor appearing in the weighted bound of part (a).

Lemma B.9. *We define $Y_i:=\mathbb{E}[\|\mathbf{y}_{i+1}-\mathbf{y}_i\|_2^2]$, $V_i:=\min(\mathbf{v}_i)$, and $q':=q\grave{\kappa}^{q-1}$. For all t with $(r_t-1)q\le t\le r_tq-1$, we have:*

(a) $\sum_{j=(r_t-1)q}^{t}[V_j\sum_{i=(r_j-1)q}^{j-1}Y_i]\le q'\sum_{i=(r_t-1)q}^{t-1}V_iY_i$.

(b) $\sum_{j=(r_t-1)q}^{t}[\sum_{i=(r_j-1)q}^{j-1}Y_i]\le(q-1)\sum_{i=(r_t-1)q}^{t-1}Y_i$.

(c) $\sum_{t=0}^{T}[\sum_{i=(r_t-1)q}^{t-1}Y_i]\le(q-1)\sum_{i=0}^{T}Y_i$.

Proof. We define $V_t:=\min(\mathbf{v}_t)$, where $\{V_j\}_{j=0}^{\infty}$ is non-decreasing. We define $q':=q\grave{\kappa}^{q-1}$.

Bounding the Term $(t-(r_t-1)q)$**.** For any integer $t\ge0$, we derive the following inequalities:

$$t-(r_t-1)q \overset{①}{=} t-(\lfloor\tfrac{t}{q}\rfloor+1-1)q = t-\lfloor\tfrac{t}{q}\rfloor q \overset{②}{\le} q-1, \tag{33}$$

where step ① uses $r_t:=\lfloor\frac{t}{q}\rfloor+1$; step ② uses the fact that $t-\lfloor\frac{t}{q}\rfloor q\le q-1$ for all integer $t\ge0$ and $q\ge1$.

Part (a). For any t with $t\ge(r_t-1)q$, we have the following results:

$$\frac{\min(\mathbf{v}_t)}{\min(\mathbf{v}_{(r_t-1)q})} = \frac{\min(\mathbf{v}_{(r_t-1)q+1})}{\min(\mathbf{v}_{(r_t-1)q})}\cdot\frac{\min(\mathbf{v}_{(r_t-1)q+2})}{\min(\mathbf{v}_{(r_t-1)q+1})}\cdots\frac{\min(\mathbf{v}_t)}{\min(\mathbf{v}_{t-1})} \overset{①}{\le} \grave{\kappa}^{q-1}, \tag{34}$$

where step ① uses the fact that the product length is at most $([t]-[(r_t-1)q+1]+1)$ and Inequality (33).

For all t with $(r_t-1)q \le t \le r_t q-1$, we have:

$$\begin{aligned}
\textstyle\sum_{j=(r_t-1)q}^{t}\Big(V_j\cdot\sum_{i=(r_j-1)q}^{j-1}Y_i\Big) &\overset{①}{\le} \textstyle\sum_{j=(r_t-1)q}^{t}\Big(V_j\cdot\sum_{i=(r_j-1)q}^{t-1}Y_i\Big)\\
&\overset{②}{=} \textstyle\sum_{j=(r_t-1)q}^{t}\Big(V_j\cdot\sum_{i=(r_t-1)q}^{t-1}Y_i\Big)\\
&\overset{③}{=} \textstyle\sum_{i=(r_t-1)q}^{t-1}\Big(Y_i\cdot\sum_{j=(r_t-1)q}^{t}V_j\Big)\\
&\overset{④}{\le} \textstyle\sum_{i=(r_t-1)q}^{t-1}\Big(Y_i\cdot\sum_{j=(r_t-1)q}^{t}V_t\Big)\\
&\overset{⑤}{\le} qV_t\textstyle\sum_{i=(r_t-1)q}^{t-1}Y_i\\
&\overset{⑥}{\le} q(\grave{\kappa}^{q-1}V_{(r_t-1)q})\textstyle\sum_{i=(r_t-1)q}^{t-1}Y_i\\
&\overset{⑦}{\le} \underbrace{q\grave{\kappa}^{q-1}}_{:=q'}\cdot\textstyle\sum_{i=(r_t-1)q}^{t-1}V_iY_i,
\end{aligned}$$

where step ① uses $j\le t$ for all $j\in[(r_t-1)q,t]$; step ② uses $r_j=r_t$ for all $j\in[(r_t-1)q,t]$ with $t\in[(r_t-1)q,r_tq-1]$; step ③ uses the fact that $\sum_{j=\underline{j}}^{\overline{j}}(\mathbf{a}_j\sum_{i=\underline{i}}^{\overline{i}}\mathbf{b}_i)=\sum_{i=\underline{i}}^{\overline{i}}(\mathbf{b}_i\sum_{j=\underline{j}}^{\overline{j}}\mathbf{a}_j)$ for all $\underline{i}\le\overline{i}$ and $\underline{j}\le\overline{j}$; step ④ uses $V_j\le V_t$ as $j\le t$; step ⑤ uses $t-(r_t-1)q+1\le q$; step ⑥ uses Inequality (34); step ⑦ uses $i\ge(r_t-1)q$.

Part (b). For all t with $(r_t-1)q\le t\le r_tq-1$, we have:

$$\begin{aligned}
\textstyle\sum_{j=(r_t-1)q}^{t}\sum_{i=(r_j-1)q}^{j-1}Y_i &\overset{①}{=} \textstyle\sum_{i=(r_t-1)q}^{t-1}(t-i)Y_i\\
&\overset{②}{\le} ([t-1]-[(r_t-1)q]+1)\cdot\textstyle\sum_{i=(r_t-1)q}^{t-1}Y_i\\
&\overset{③}{\le} (q-1)\textstyle\sum_{i=(r_t-1)q}^{t-1}Y_i,
\end{aligned}$$

where step ① uses basic reduction; step ② uses $i\ge(r_t-1)q$; step ③ uses Inequality (33).

Part (c). Each Y_i can appear only at later indices in its own epoch, at most $q-1$ times. Thus, including a possibly incomplete final epoch, $\sum_{t=0}^{T}\sum_{i=(r_t-1)q}^{t-1}Y_i\le(q-1)\sum_{i=0}^{T}Y_i$.

□

The preceding estimates control the accumulated SPIDER error through the extrapolated motion. We now combine this control with pathwise descent and apply Jensen's inequality only after taking expectations, obtaining a closed first-moment bound for the random metric accumulator.

Lemma B.10. *(*Boundedness of $\mathcal{Z}_t$ and $\mathcal{V}_t$*)* Under Assumptions 3.1, 4.1, and 4.2, let the sequence be generated by ALFPG-FS from a fixed $\mathbf{x}_0\in\mathrm{dom}(F)$. There exist finite constants $\overline{\mathcal{Z}}>0$ and $\overline{\mathrm{v}}>0$, independent of t, such that, for all $t\ge0$,

(a) $\mathcal{Z}_t\le\overline{\mathcal{Z}}$.

(b) $\mathbb{E}[\mathcal{V}_t]\le\overline{\mathrm{v}}$.

Proof. Let $\mathbf{e}_t=\mathbf{g}_t-\nabla f(\mathbf{y}_t)$.

Define $\mathcal{V}_{t+1}:=\sqrt{\underline{\mathrm{v}}^2+(\alpha+\beta)M_t}$, where $M_t:=\sum_{i=0}^{t}\|\mathbf{r}_i\|_2^2$.

Define $\mathbb{S}_t:=\frac{\|\mathbf{r}_t\|_2^2}{\min(\mathbf{v}_t)}$, $\mathbb{U}_t:=\frac{\|\mathbf{r}_t\|_2^2}{\min(\mathbf{v}_t)^2}$. Let $c_1:=\frac{1}{2}\left(\frac{1-\theta}{\kappa}\right)^2$. Define $Y_i:=\mathbb{E}[\|\mathbf{y}_{i+1}-\mathbf{y}_i\|_2^2]$.

Define $u_2=10s_2$, where s_2 and u_2 are defined in Lemmas B.4(b) and B.8(b), respectively.

Define $\mathcal{Z}_t=\mathbb{E}[\widehat{\mathcal{Z}}_t]$, where $\widehat{\mathcal{Z}}_t:=F(\mathbf{x}_t)-F(\bar{\mathbf{x}})+\frac{1}{2}\|\mathbf{x}_t-\mathbf{x}_{t-1}\|^2_{\sigma_{t-1}(\mathbf{v}_t+L)}$. It holds that $\widehat{\mathcal{Z}}_0=\mathcal{Z}_0$.

Bounding the term $\mathbb{E}[\sum_{t=0}^{T}\|\mathbf{e}_t\|^2]$. We derive:

$$\begin{aligned}\mathbb{E}[\textstyle\sum_{t=0}^{T}\|\mathbf{e}_t\|^2] &\overset{①}{\leq} \tfrac{L^2}{b}\textstyle\sum_{t=0}^{T}\sum_{i=(r_t-1)_q}^{t-1} Y_i \\ &\overset{②}{\leq} \tfrac{L^2(q-1)}{b}\textstyle\sum_{i=0}^{T} Y_i \overset{③}{\leq} \tfrac{L^2(q-1)u_2}{b}\mathbb{E}[\sqrt{\mathcal{V}_{T+1}}],\end{aligned} \tag{35}$$

where step ① uses Lemma B.7(a); step ② uses Lemma B.9(c); step ③ uses Lemma B.8(b).

Pathwise bounds for the adaptive sums. We derive:

$$\mathbb{S}_t := \tfrac{\|\mathbf{r}_t\|^2}{\min(\mathbf{v}_t)} \geq \tfrac{\|\mathbf{r}_t\|^2}{\mathcal{V}_t} \overset{①}{=} \tfrac{(\mathcal{V}_{t+1}-\mathcal{V}_t)(\mathcal{V}_{t+1}+\mathcal{V}_t)}{(\alpha+\beta)\mathcal{V}_t} \geq \tfrac{2}{\alpha+\beta}(\mathcal{V}_{t+1}-\mathcal{V}_t), \tag{36}$$

where step ① uses $\mathcal{V}_{t+1}^2-\mathcal{V}_t^2=(\alpha+\beta)(M_t-M_{t-1})=(\alpha+\beta)\|\mathbf{r}_t\|^2$. Summing this inequality and invoking Lemma B.4(b) give the following bounds on every sample path:

$$\textstyle\sum_{t=0}^{T}\mathbb{S}_t \geq \tfrac{2}{\alpha+\beta}(\mathcal{V}_{T+1}-\underline{\mathrm{v}}), \qquad \sum_{t=0}^{T}\mathbb{U}_t \leq s_2\sqrt{\mathcal{V}_{T+1}}. \tag{37}$$

Pathwise descent inequality. We have from Lemma B.3:

$$\begin{aligned}\widehat{\mathcal{Z}}_{t+1}-\widehat{\mathcal{Z}}_t &\leq -\langle\mathbf{d}_t,\mathbf{e}_t\rangle+\tfrac{3L}{2}\mathbb{U}_t-c_1\mathbb{S}_t \\ &\overset{①}{\leq} \tfrac{L}{2}\|\mathbf{d}_t\|^2+\tfrac{1}{2L}\|\mathbf{e}_t\|^2+\tfrac{3L}{2}\mathbb{U}_t-c_1\mathbb{S}_t \\ &\overset{②}{\leq} \tfrac{L}{2}\mathbb{U}_t+\tfrac{1}{2L}\|\mathbf{e}_t\|^2+\tfrac{3L}{2}\mathbb{U}_t-c_1\mathbb{S}_t,\end{aligned}$$

where step ① uses Young's inequality; step ② uses $\|\mathbf{d}_t\|^2\leq\mathbb{U}_t$. Summing these pathwise inequalities yields:

$$\begin{aligned}\widehat{\mathcal{Z}}_{T+1}+c_1\textstyle\sum_{t=0}^{T}\mathbb{S}_t &\leq \mathcal{Z}_0+2L\textstyle\sum_{t=0}^{T}\mathbb{U}_t+\frac{1}{2L}\sum_{t=0}^{T}\|\mathbf{e}_t\|^2 \\ &\overset{①}{\leq} \mathcal{Z}_0+2Ls_2\sqrt{\mathcal{V}_{T+1}}+\tfrac{1}{2L}\textstyle\sum_{t=0}^{T}\|\mathbf{e}_t\|^2,\end{aligned}$$

where step ① uses the second inequality in (37). Taking expectations yields:

$$\begin{aligned}\mathcal{Z}_{T+1} &\leq -c_1\mathbb{E}\left[\textstyle\sum_{t=0}^{T}\mathbb{S}_t\right]+\mathcal{Z}_0+\mathbb{E}[2Ls_2\sqrt{\mathcal{V}_{T+1}}+\tfrac{1}{2L}\textstyle\sum_{t=0}^{T}\|\mathbf{e}_t\|^2] \\ &\overset{①}{\leq} -\tfrac{2c_1}{\alpha+\beta}(\mathbb{E}[\mathcal{V}_{T+1}]-\underline{\mathrm{v}})+\mathcal{Z}_0+\mathbb{E}[2Ls_2\sqrt{\mathcal{V}_{T+1}}]+\tfrac{1}{2L}\tfrac{L^2(q-1)u_2}{b}\mathbb{E}[\sqrt{\mathcal{V}_{T+1}}],\end{aligned}$$

where step ① uses Inequality (37).

Bounding the term $\mathbb{E}[\mathcal{V}_{T+1}]$. Set $x:=\mathbb{E}[\mathcal{V}_{T+1}]$ and $y:=\mathcal{Z}_{T+1}$. Using Jensen's inequality yields:

$$\tfrac{\alpha+\beta}{2c_1}\cdot y+x \leq A+B\sqrt{x}, \tag{38}$$

where $A:=\underline{\mathrm{v}}+\frac{(\alpha+\beta)\mathcal{Z}_0}{2c_1}$, and $B:=\frac{(\alpha+\beta)}{2c_1}\cdot(2Ls_2+\frac{L^2(q-1)u_2}{2Lb})$.

Dropping the non-negative term y in Inequality (38) and using $B\sqrt{x}\leq(x+B^2)/2$, we obtain:

$$x:=\mathbb{E}[\mathcal{V}_{T+1}] \leq \overline{\mathrm{v}} := 2A+B^2.$$

Dropping the non-negative term x and using $x\leq\overline{\mathrm{v}}$, we conclude that

$$y := \mathcal{Z}_{T+1} \leq \overline{\mathcal{Z}} := \tfrac{2c_1}{\alpha+\beta}\left(A+B\sqrt{\overline{\mathrm{v}}}\right).$$

□

B.3.1 Proof of Theorem 4.6

We now convert the first-moment metric bound into expected stationarity. The proof controls the displacement, estimator-error, and metric-scaled residual terms separately, and then counts all component-gradient evaluations.

Proof. Write $\mathbf{d}_t:=\mathbf{x}_{t+1}-\mathbf{x}_t$, $\mathbf{r}_t:=\mathbf{v}_t\odot\mathbf{d}_t$, and $\mathbf{e}_t:=\mathbf{g}_t-\nabla f(\mathbf{y}_t)$, with $\mathbf{d}_{-1}=\mathbf{r}_{-1}=\mathbf{0}$.

Let $s_2=\frac{4\dot{\kappa}^2}{\alpha\sqrt{\underline{\mathrm{v}}}}$, which is defined in Lemma B.4(b).

Let $u_2=10s_2$, which is defined in Lemma B.8(b).

Part (a). We derive:

$$\mathbb{E}\sum_{t=0}^{T}\|\mathbf{d}_t\|_2^2 \overset{①}{\le} s_2\mathbb{E}[\sqrt{\mathcal{V}_{T+1}}] \overset{②}{\le} s_2\sqrt{\mathbb{E}[\mathcal{V}_{T+1}]} \le s_2\sqrt{\bar{\mathrm{v}}} := D. \tag{39}$$

where step ① uses Lemma B.4(b) that $\sum_{t=0}^{T}\mathbb{U}_t \le s_2\sqrt{\mathcal{V}_{T+1}}$; step ② uses Jensen's inequality.

Furthermore, we derive:

$$\mathbb{E}[\textstyle\sum_{t=0}^{T}\|\mathbf{e}_t\|_2^2] \overset{①}{\le} \frac{L^2(q-1)u_2}{b}\mathbb{E}[\sqrt{\mathcal{V}_{T+1}}] \le \frac{L^2(q-1)u_2}{b}\sqrt{\bar{\mathrm{v}}} := E. \tag{40}$$

where step ① uses Inequality (35).

The optimality condition for the proximal subproblem implies

$$\mathbf{0} \in \partial h(\mathbf{x}_{t+1}) + \mathbf{g}_t + \mathbf{v}_t \odot (\mathbf{x}_{t+1} - \mathbf{y}_t).$$

Since $\mathbf{y}_t = \mathbf{x}_t + \sigma_{t-1}\mathbf{d}_{t-1}$, $0 \le \sigma_{t-1} \le \theta$, and $\sigma_{t-1}\mathbf{v}_t \le \theta\mathbf{v}_{t-1}$ coordinatewise for $t \ge 1$, smoothness of f yields

$$\begin{aligned}\mathcal{R}_{t+1} &\le \big\|\nabla f(\mathbf{x}_{t+1}) - \mathbf{g}_t - \mathbf{v}_t \odot (\mathbf{x}_{t+1} - \mathbf{y}_t)\big\| \\ &\le L\big(\|\mathbf{d}_t\| + \theta\|\mathbf{d}_{t-1}\|\big) + \|\mathbf{e}_t\| + \|\mathbf{r}_t\| + \theta\|\mathbf{r}_{t-1}\|.\end{aligned} \tag{41}$$

For $t = 0$, the terms with index -1 vanish.

By Cauchy–Schwarz and (39)–(40),

$$\mathbb{E}[\textstyle\sum_{t=0}^{T}\|\mathbf{d}_t\|] \le \sqrt{(T+1)D}, \qquad \mathbb{E}[\sum_{t=0}^{T}\|\mathbf{e}_t\|] \le \sqrt{(T+1)E}.$$

The accumulator identity $\mathcal{V}_{T+1}^2 - \underline{\mathrm{v}}^2 = (\alpha+\beta)\sum_{t=0}^{T}\|\mathbf{r}_t\|_2^2$ also gives

$$\mathbb{E}[\textstyle\sum_{t=0}^{T}\|\mathbf{r}_t\|_2] \le \sqrt{\frac{T+1}{\alpha+\beta}}\,\mathbb{E}\big[\sqrt{\mathcal{V}_{T+1}^2 - \underline{\mathrm{v}}^2}\big] \le \sqrt{\frac{T+1}{\alpha+\beta}}\,\mathbb{E}[\mathcal{V}_{T+1}] \le \sqrt{\frac{T+1}{\alpha+\beta}}\,\bar{\mathrm{v}}.$$

Summing (41) over $t = 0, \dots, T$, and noting that the shifted sums do not exceed their unshifted counterparts, we obtain

$$\tfrac{1}{T+1}\textstyle\sum_{t=0}^{T}\mathbb{E}[\mathcal{R}_{t+1}] \le \frac{C}{\sqrt{T+1}}, \qquad C := L(1+\theta)\sqrt{D} + \sqrt{E} + \frac{(1+\theta)\bar{\mathrm{v}}}{\sqrt{\alpha+\beta}}. \tag{42}$$

This constant is finite and independent of T. At least one $\bar{t} \in \{1, \dots, T+1\}$ satisfies $\mathbb{E}[\mathcal{R}_{\bar{t}}] \le C/\sqrt{T+1}$. If J is uniform on $\{1, \dots, T+1\}$ and independent of the algorithm, then $\mathbb{E}[\mathcal{R}_J] = (T+1)^{-1}\sum_{t=0}^{T}\mathbb{E}[\mathcal{R}_{t+1}]$. Hence both claims follow when $T+1 \ge C^2\epsilon^{-2}$.

Part (b). Let b denote the mini-batch size and q the refresh frequency of ALFPG-FS. Set $H := T+1$. When $\mathrm{mod}(t, q) = 0$, the full gradient $\nabla f(\mathbf{y}_t)$ is computed using N component-gradient evaluations. This occurs $F := \lceil H/q \rceil$ times. Otherwise, the SPIDER update evaluates the mini-batch at both $\mathbf{y}_t$ and $\mathbf{y}_{t-1}$, using at most $2b$ component-gradient evaluations. This occurs $H - F$ times. Hence the total number of component-gradient evaluations satisfies

$$\begin{aligned}NF + 2b(H-F) &\overset{①}{\le} N + \left(\frac{N}{q} + 2b\left(1 - \frac{1}{q}\right)\right)H \\ &\overset{②}{\le} N + 3\sqrt{N}\,H = \mathcal{O}\Big(N + \sqrt{N}\,\epsilon^{-2}\Big),\end{aligned}$$

where step ① uses $H/q \le F \le H/q + 1$; step ② uses $q = b = \lceil\sqrt{N}\rceil$ and the choice $H = \max\{1, \lceil C^2\epsilon^{-2}\rceil\}$ from Part (a), with C independent of N.

□

C Proof for Section 5

This appendix proves the guarantees in Section 5. After a shared adaptive-weight estimate, each variant is analyzed through preliminary bounds, a tracking-error recursion, proximal descent, cumulative moment estimates, a motion comparison, and the final stationarity argument.

We first establish estimates shared by ALFPG-MVR1 and ALFPG-MVR2, then prove the results for each variant. Throughout this appendix, Assumptions 3.1, 5.1, 5.2, 5.3, and 5.4 are in force; the analysis of ALFPG-MVR2 also uses Assumption 5.7.

Overview of the proof strategy for ALFPG-MVR1. For ALFPG-MVR1, we first separate the EMA tracking error into a predictable drift and a conditionally centered sampling term (Lemma C.3). The lagged adaptive weights allow conditional expectations to remove the centered term. Self-normalized sums and summation by parts then give logarithmic first- and second-moment bounds for the weighted tracking energy, together with a pathwise logarithmic bound for the normalized motion (Lemma C.5). Proximal descent transfers these estimates to the descent energy, and the accumulator identity yields the weighted motion bound in Lemma C.6. Finally, the gradient-discrepancy statistic is controlled by sampling noise and iterate motion. Cauchy–Schwarz removes the remaining random weights and converts these estimates into the noise-adaptive stationarity bound. Only smoothness of the expected loss is needed; the motion contribution remains controlled even when the noise vanishes.

Overview of the proof strategy for ALFPG-MVR2. For ALFPG-MVR2, the same-sample correction cancels the deterministic gradient drift in the tracking recursion. Individual smoothness controls the remaining gradient-difference innovation. Rescaling the tracking error by $\alpha_t^{1/4}$ gives a contracting energy recursion (Lemma C.9), from which logarithmic first- and second-moment bounds follow. For convex h, the motion estimate uses subgradient monotonicity, with the first proximal step handled separately. For domain-Lipschitz h, a telescoping potential instead couples motion to the tracking energy, without assuming bounded ambient subgradients (Lemma C.11). Proximal descent and the motion bridge then control the weighted cumulative displacements. Unlike MVR1, the discrepancy statistic compares two samples at the same point, so its expected squared magnitude is controlled by noise alone. Removing the random weights yields the stated rate and recovers the deterministic rate when $\sigma = 0$.

C.1 Shared Preliminary Estimates

Both expectation variants use the same running-minimum construction for the averaging weight. We first bound its variation; these estimates allow shifted weights to be compared and preserve contraction in the tracking-error recursions.

Lemma C.1. *(Adaptive-weight regularity) For $t \geq 1$, define $\rho_t = \min_{1\leq k\leq t}\left(\frac{1+\max_{1\leq i\leq k}\chi_i^2}{1+\sum_{i=1}^k \chi_i^2}\right)^p$, where $p \in (0,1)$ and $\rho_{-1} = \rho_0 := 1$. Then*

(a) $0 \leq \frac{1}{\rho_t} - \frac{1}{\rho_{t-1}} \leq p$.

(b) $\rho_{t-1} - \rho_t \leq p\rho_{t-1}^{1+1/p}$.

(c) $(1-\rho_t)\sqrt{\frac{\rho_{t-1}}{\rho_t}} \leq 1-(1-p)\rho_t$.

Proof. Define $D_t := \sum_{i=1}^t \chi_i^2$, $N_t := \max_{1\leq i\leq t}\chi_i^2$, where $D_0 = N_0 := 0$.

Define $u_t := \frac{1+D_t}{1+N_t} \geq 1$, $U_t := \max_{1\leq k\leq t} u_k$, where $u_0 = U_0 := 1$.

Since $\left(\frac{1+N_k}{1+D_k}\right)^p = u_k^{-p}$ and $f(x) = x^{-p}$ is decreasing on $(0,\infty)$, we have $\rho_t = U_t^{-p}$.

Inverse-weight increment. On the one hand, if $\chi_t^2 \leq N_{t-1}$, then $N_t = N_{t-1}$ and $u_t - u_{t-1} = \frac{\chi_t^2}{1+N_{t-1}} \leq 1$. On the other hand, if $\chi_t^2 > N_{t-1}$, then $N_t = \chi_t^2$ and $u_t = \frac{1+D_{t-1}+\chi_t^2}{1+\chi_t^2} = 1 + \frac{D_{t-1}}{1+\chi_t^2} \leq 1 + \frac{D_{t-1}}{1+N_{t-1}} \leq u_{t-1}+1$. Thus, $u_t \leq u_{t-1}+1 \leq U_{t-1}+1$. Since $U_t = \max\{U_{t-1}, u_t\}$, we have

$$0 \leq U_t - U_{t-1} \leq 1. \tag{43}$$

Part (a). Because $f(x) = x^p$ is increasing and concave on $[1,\infty)$, we derive:

$$0 \;\leq\; U_t^p - U_{t-1}^p \;\leq\; pU_{t-1}^{p-1}(U_t - U_{t-1}) \;\leq\; p,$$

where we used $U_{t-1} \geq 1$, $p-1 < 0$, and (43). Using $\rho_t = U_t^{-p}$, we have:

$$0 \;\leq\; \tfrac{1}{\rho_t} - \tfrac{1}{\rho_{t-1}} \;\leq\; p.$$

Part (b). Using the integral representation of $x \mapsto x^{-p}$ and (43), we obtain

$$\begin{aligned}\rho_{t-1} - \rho_t &= U_{t-1}^{-p} - U_t^{-p} = p\int_{U_{t-1}}^{U_t} z^{-p-1}\,dz \\ &\le p(U_t - U_{t-1})U_{t-1}^{-p-1} \le pU_{t-1}^{-p-1} = p\rho_{t-1}^{1+1/p}.\end{aligned}$$

This proves $\rho_{t-1} - \rho_t \le p\rho_{t-1}^{1+1/p}$.

Part (c). Using Part (a), we have the following equivalent inequalities:

$$\tfrac{\rho_{t-1}-\rho_t}{\rho_t\rho_{t-1}} \le p \;\Leftrightarrow\; \rho_{t-1}(1-p\rho_t) \le \rho_t \;\Leftrightarrow\; \tfrac{\rho_{t-1}}{\rho_t} \le \tfrac{1}{1-p\rho_t}.$$

Since $0 < \rho_t \le 1$ and $p \in (0,1)$, it follows that

$$(1-\rho_t)\sqrt{\tfrac{\rho_{t-1}}{\rho_t}} \le \tfrac{1-\rho_t}{\sqrt{1-p\rho_t}} \le \tfrac{1-\rho_t}{1-p\rho_t} = 1-(1-p)\rho_t - \tfrac{p(1-p)\rho_t^2}{1-p\rho_t} \le 1-(1-p)\rho_t.$$

□

C.2 Analysis for ALFPG-MVR1

Notation. Let $p = \lambda = 1/2$. Write $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_{t-1}]$.

Define $\boldsymbol{\epsilon}_t := \mathbf{g}_t - \nabla f(\mathbf{x}_t)$, $\mathbf{s}_t := \mathbf{m}_t - \nabla f(\mathbf{x}_t)$, $\mathbf{r}_t := \frac{\mathbf{x}_t - \mathbf{x}_{t+1}}{\gamma_t}$, $A_t := 1 + \sum_{i=1}^t \|\mathbf{r}_i\|^2$, $D_t := \sum_{i=1}^t \chi_i^2$, $N_t := \max_{1\le i\le t}\chi_i^2$, where $A_0 = 1$ and $D_0 = N_0 = 0$.

Set $\mathbf{g}_0 := \mathbf{g}_1$, consistently with $\mathbf{x}_0 = \mathbf{x}_1$ and $\xi_0 = \xi_1$, so $\chi_1 = 0$.

Set $\alpha_t = \rho_{t-1}$, $\beta_t = A_{t-1}^{-1/2}$, $\gamma_t = \theta\sqrt{\alpha_t}\beta_t$. Both α_t and γ_t are $\mathcal{F}_{t-1}$-measurable.

Set $\alpha_0 = 1$, $\gamma_0 = 0$, $\mathbf{r}_0 = \mathbf{0}$, and $\mathbf{s}_0 = -\nabla f(\mathbf{x}_1)$.

Write $U_T := \sum_{t=1}^T \alpha_t\|\mathbf{s}_t\|^2$, $W_T := \sum_{t=1}^T \gamma_t\|\mathbf{r}_t\|^2$, $H_T := \sum_{t=1}^T \frac{\gamma_t^2}{\alpha_t}\|\mathbf{r}_t\|^2$, $M_T := \sum_{t=1}^T \|\mathbf{r}_t\|^2 = A_T - 1$, $S_T := \sum_{t=1}^T \|\mathbf{s}_t\|^2$, $\ell_T := \ln(e+T)$.

We begin the MVR1 analysis with pathwise bounds for the estimator, the proximal residual, and successive adaptive weights. The convex and domain-Lipschitz cases require different residual arguments, but lead to the common estimates used below.

Lemma C.2. *(Preliminary bounds for MVR1) Let* $K_h := \begin{cases}\frac{1}{\theta}\|\mathbf{x}_1 - \widehat{\mathbf{x}}_1\|, & h \text{ convex},\\ \hat{H}, & \text{otherwise},\end{cases}$, $\widehat{\mathbf{x}}_1 := \mathrm{Prox}_{\theta h}(\mathbf{x}_1)$, $\widehat{C} := 2\widehat{G} + 2K_h$, $\hat{R} := 1 + \widehat{C}$, *and* $\kappa := 1 + \hat{R}$. *Almost surely, for every* $t \ge 1$,

(a) $\|\mathbf{m}_t\| \le \widehat{G}$, $\|\mathbf{s}_t\| \le 2\widehat{G}$.

(b) $\|\mathbf{r}_t\| \le \widehat{C}t$, $\frac{\|\mathbf{r}_t\|^2}{A_{t-1}} \le \hat{R}^2$.

(c) $\frac{\alpha_t}{\alpha_{t+1}} \le \frac{3}{2}$, $\frac{\beta_t}{\beta_{t+1}} \le \kappa$.

Proof. Define $\widehat{\mathbf{x}}_1 := \mathrm{Prox}_{\theta h}(\mathbf{x}_1)$.

Bounding the term $\|\mathbf{r}_1\|$ when $h(\cdot)$ is convex. For the first iteration, we have $\gamma_1 = \theta$ since $\alpha_1 = \beta_1 = 1$. By the nonexpansiveness of the proximal operator, $\|\mathbf{x}_2 - \widehat{\mathbf{x}}_1\| = \|\mathrm{Prox}_{\theta h}(\mathbf{x}_1 - \theta\mathbf{m}_1) - \mathrm{Prox}_{\theta h}(\mathbf{x}_1)\| \le \theta\|\mathbf{m}_1\|$. Therefore,

$$\|\mathbf{r}_1\| = \frac{\|\mathbf{x}_1 - \mathbf{x}_2\|}{\theta} \le \frac{\|\mathbf{x}_1 - \widehat{\mathbf{x}}_1\|}{\theta} + \frac{\|\widehat{\mathbf{x}}_1 - \mathbf{x}_2\|}{\theta} \le K_h + \|\mathbf{m}_1\| \le K_h + \widehat{G}. \tag{44}$$

Part (a-i). The stochastic momentum estimator is updated as $\mathbf{m}_t = (1-\alpha_t)\mathbf{m}_{t-1} + \alpha_t\mathbf{g}_t$, where $\alpha_t \in (0,1]$. Since the stochastic oracle is bounded, $\|\mathbf{g}_t\| \le \widehat{G}$ almost surely. Therefore, $\mathbf{m}_t$ is a convex combination of previous bounded estimators and stochastic gradients. By induction,

$$\|\mathbf{m}_t\| \le (1-\alpha_t)\|\mathbf{m}_{t-1}\| + \alpha_t\|\mathbf{g}_t\| \le \widehat{G}.$$

Part (a-ii). By the definition $\mathbf{s}_t = \mathbf{m}_t - \nabla f(\mathbf{x}_t)$, and Jensen's inequality together with the bounded stochastic oracle, we have $\|\nabla f(\mathbf{x}_t)\| = \|\mathbb{E}_t[\mathbf{g}_t]\| \le \mathbb{E}_t\|\mathbf{g}_t\| \le \widehat{G}$. Hence,

$$\|\mathbf{s}_t\| \le \|\mathbf{m}_t\| + \|\nabla f(\mathbf{x}_t)\| \le 2\widehat{G}.$$

Part (b-i). We next establish an upper bound for $\|\mathbf{r}_t\|$ when h is convex. Assume $t \ge 2$. Note that $\mathbf{x}_{t+1} = \mathrm{Prox}_{\gamma_t h}(\mathbf{x}_t - \gamma_t \mathbf{m}_t)$, and $\mathbf{r}_t = \frac{\mathbf{x}_t - \mathbf{x}_{t+1}}{\gamma_t}$. The first-order optimality condition of the proximal subproblem gives $\mathbf{r}_t - \mathbf{m}_t \in \partial h(\mathbf{x}_{t+1})$. Similarly, using the proximal update at iteration $t-1$, we have $\mathbf{r}_{t-1} - \mathbf{m}_{t-1} \in \partial h(\mathbf{x}_t)$.

Since h is convex, its subdifferential is monotone. Therefore,

$$\langle (\mathbf{r}_t - \mathbf{m}_t) - (\mathbf{r}_{t-1} - \mathbf{m}_{t-1}), \mathbf{x}_{t+1} - \mathbf{x}_t \rangle \ge 0.$$

Using $\mathbf{x}_{t+1} - \mathbf{x}_t = -\gamma_t \mathbf{r}_t$, we obtain

$$\langle (\mathbf{r}_t - \mathbf{m}_t) - (\mathbf{r}_{t-1} - \mathbf{m}_{t-1}), \mathbf{r}_t \rangle \le 0.$$

Expanding the inner product gives

$$\|\mathbf{r}_t\|^2 \;\le\; \langle \mathbf{r}_{t-1} + \mathbf{m}_t - \mathbf{m}_{t-1}, \mathbf{r}_t \rangle \;\overset{①}{\le}\; \|\mathbf{r}_t\|\|\mathbf{r}_{t-1} + \mathbf{m}_t - \mathbf{m}_{t-1}\| \le \|\mathbf{r}_t\| \left(\|\mathbf{r}_{t-1}\| + 2\widehat{G} \right),$$

where step ① uses the Cauchy–Schwarz inequality. If $\mathbf{r}_t = \mathbf{0}$, the claimed bound is immediate; otherwise, dividing by $\|\mathbf{r}_t\|$ gives:

$$\|\mathbf{r}_t\| \le \|\mathbf{r}_{t-1}\| + 2\widehat{G}. \tag{45}$$

Applying this recursion repeatedly gives

$$\|\mathbf{r}_t\| \;\le\; \|\mathbf{r}_1\| + 2(t-1)\widehat{G} \;\le\; K_h + \widehat{G} + 2(t-1)\widehat{G} \;\le\; \widehat{C}t. \tag{46}$$

Note that the inequality above also holds for $t = 1$.

We next establish an upper bound for $\|\mathbf{r}_t\|$ when $h(\cdot)$ is possibly nonconvex and domain-Lipschitz. Global optimality of the proximal update, with $\mathbf{x}_t$ as a comparison point, gives

$$h(\mathbf{x}_{t+1}) + \langle \mathbf{m}_t, \mathbf{x}_{t+1} - \mathbf{x}_t \rangle + \frac{\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2}{2\gamma_t} \le h(\mathbf{x}_t).$$

Since both $\mathbf{x}_t$ and $\mathbf{x}_{t+1}$ belong to $\mathrm{dom}(h)$, we obtain

$$\begin{aligned} \frac{\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2}{2\gamma_t} &\le h(\mathbf{x}_t) - h(\mathbf{x}_{t+1}) - \langle \mathbf{m}_t, \mathbf{x}_{t+1} - \mathbf{x}_t \rangle \\ &\le (K_h + \|\mathbf{m}_t\|)\|\mathbf{x}_{t+1} - \mathbf{x}_t\|. \end{aligned}$$

If $\mathbf{x}_{t+1} = \mathbf{x}_t$, then $\mathbf{r}_t = \mathbf{0}$. Otherwise, dividing by $\|\mathbf{x}_{t+1} - \mathbf{x}_t\|$ gives

$$\|\mathbf{r}_t\| = \frac{\|\mathbf{x}_{t+1} - \mathbf{x}_t\|}{\gamma_t} \le 2(K_h + \|\mathbf{m}_t\|) \le 2(K_h + \widehat{G}) = \widehat{C}. \tag{47}$$

Combining both cases gives $\|\mathbf{r}_t\| \le \widehat{C}t$ for all $t \ge 1$.

Part (b-ii). In view of Inequalities (45) and (47), we have: $\|\mathbf{r}_t\| \le \|\mathbf{r}_{t-1}\| + \widehat{C}$, for all $t \ge 1$. Combining with $A_{t-1} \ge 1 + \|\mathbf{r}_{t-1}\|^2$, we have:

$$\frac{\|\mathbf{r}_t\|^2}{A_{t-1}} \;\le\; \frac{(\|\mathbf{r}_{t-1}\| + \widehat{C})^2}{\|\mathbf{r}_{t-1}\|^2 + 1} \;\le\; 1 + \widehat{C}^2 \;\le\; \hat{R}^2.$$

Part (c-i). Lemma C.1(a) with $p = 1/2$ gives $\frac{1}{\rho_t} - \frac{1}{\rho_{t-1}} \le \frac{1}{2}$. Therefore,

$$\tfrac{\rho_{t-1}}{\rho_t} \le 1 + \tfrac{\rho_{t-1}}{2} \le \tfrac{3}{2}.$$

Part (c-ii). By definition, $\beta_t = A_{t-1}^{-1/2}$. Thus,

$$\frac{\beta_t}{\beta_{t+1}} \;=\; \sqrt{\frac{A_t}{A_{t-1}}} \;=\; \sqrt{1+\frac{\|\mathbf{r}_t\|^2}{A_{t-1}}} \;\le\; \sqrt{1+\hat{R}^2} \;\le \kappa.$$

□

The preliminary bounds control the size of the quantities entering the EMA recursion. We now separate its tracking error into contraction, iterate drift, and fresh sampling noise, retaining the conditionally mean-zero cross term for the later moment estimates.

Lemma C.3. *(Tracking-error recursion for MVR1) Define* $\mathcal{E}_t = 2\alpha_t\langle \mathbf{b}_t, \boldsymbol{\epsilon}_t\rangle$, $\dot{\mathcal{E}}_t = \alpha_t^2\|\boldsymbol{\epsilon}_t\|^2$, *and* $\ddot{\mathcal{E}}_t = \frac{2L^2\gamma_{t-1}^2}{\alpha_{t-1}}\|\mathbf{r}_{t-1}\|^2$, *where* $\mathbf{b}_t = (1-\alpha_t)(\mathbf{m}_{t-1} - \nabla f(\mathbf{x}_t))$.

(a) Then almost surely, $\|\mathbf{s}_t\|^2 \le (1-\alpha_t)\|\mathbf{s}_{t-1}\|^2 + \ddot{\mathcal{E}}_t + \dot{\mathcal{E}}_t + \mathcal{E}_t$.

(b) $\mathbb{E}_t[\mathcal{E}_t] = 0$, *and* $\mathcal{E}_t^2 \le 16\widehat{G}^2\dot{\mathcal{E}}_t$.

Proof. Define $\boldsymbol{\epsilon}_t := \mathbf{g}_t - \nabla f(\mathbf{x}_t)$, $\mathbf{s}_t = \mathbf{m}_t - \nabla f(\mathbf{x}_t)$.

Let $\mathbf{b}_t := (1-\alpha_t)(\mathbf{m}_{t-1} - \nabla f(\mathbf{x}_t))$, and $\boldsymbol{\delta}_t := (1-\alpha_t)(\nabla f(\mathbf{x}_{t-1}) - \nabla f(\mathbf{x}_t))$.

Exact decomposition of the tracking error. Recall that $\mathbf{m}_t = (1-\alpha_t)\mathbf{m}_{t-1} + \alpha_t\mathbf{g}_t$. We derive:

$$\begin{aligned}\mathbf{s}_t &= (1-\alpha_t)(\mathbf{m}_{t-1} - \nabla f(\mathbf{x}_t)) + \alpha_t\boldsymbol{\epsilon}_t \\ &= (1-\alpha_t)\mathbf{s}_{t-1} + (1-\alpha_t)(\nabla f(\mathbf{x}_{t-1}) - \nabla f(\mathbf{x}_t)) + \alpha_t\boldsymbol{\epsilon}_t \;=\; \mathbf{b}_t + \alpha_t\boldsymbol{\epsilon}_t. \end{aligned} \tag{48}$$

Bounding the deterministic drift $\|\boldsymbol{\delta}_t\|$. By the L-smoothness of f,

$$\|\boldsymbol{\delta}_t\| \;\le\; (1-\alpha_t)L\|\mathbf{x}_t - \mathbf{x}_{t-1}\| \;\overset{①}{\le}\; (1-\alpha_t)L\gamma_{t-1}\|\mathbf{r}_{t-1}\| \;\le\; L\gamma_{t-1}\|\mathbf{r}_{t-1}\|. \tag{49}$$

where step ① uses $\mathbf{x}_t - \mathbf{x}_{t-1} = -\gamma_{t-1}\mathbf{r}_{t-1}$.

Part (a). Since $\mathbf{b}_t = (1-\alpha_t)\mathbf{s}_{t-1} + \alpha_t(\frac{1}{\alpha_t}\boldsymbol{\delta}_t)$, due to the convexity of $\|\cdot\|_2^2$, we have:

$$\begin{aligned}\|\mathbf{b}_t\|^2 &\le (1-\alpha_t)\|\mathbf{s}_{t-1}\|^2 + \alpha_t\|(\tfrac{1}{\alpha_t}\boldsymbol{\delta}_t)\|^2 \;\overset{①}{\le}\; (1-\alpha_t)\|\mathbf{s}_{t-1}\|^2 + \tfrac{1}{\alpha_t}L^2\gamma_{t-1}^2\|\mathbf{r}_{t-1}\|^2 \\ &\overset{②}{\le} (1-\alpha_t)\|\mathbf{s}_{t-1}\|^2 + \tfrac{2L^2\gamma_{t-1}^2}{\alpha_{t-1}}\|\mathbf{r}_{t-1}\|^2, \end{aligned} \tag{50}$$

where step ① uses Inequality (49); step ② uses $\frac{1}{\alpha_t} \le \frac{2}{\alpha_{t-1}}$, which can be implied by Lemma C.2 that $\frac{\alpha_{t-1}}{\alpha_t} \le \frac{3}{2} < 2$.

On the other hand, we have:

$$\|\mathbf{s}_t\|^2 = \|\mathbf{b}_t + \alpha_t\boldsymbol{\epsilon}_t\|^2 = \|\mathbf{b}_t\|^2 + \dot{\mathcal{E}}_t + \mathcal{E}_t. \tag{51}$$

Combining Inequality (50) and Inequality (51) yields:

$$\|\mathbf{s}_t\|^2 \le (1-\alpha_t)\|\mathbf{s}_{t-1}\|^2 + \tfrac{2L^2\gamma_{t-1}^2}{\alpha_{t-1}}\|\mathbf{r}_{t-1}\|^2 + \dot{\mathcal{E}}_t + \mathcal{E}_t.$$

Part (b). Since $\mathbf{b}_t$ is $\mathcal{F}_{t-1}$-measurable and $\mathbb{E}_t[\boldsymbol{\epsilon}_t] = 0$, we have $\mathbb{E}_t[\mathcal{E}_t] = 2\alpha_t\langle \mathbf{b}_t, \mathbb{E}_t[\boldsymbol{\epsilon}_t]\rangle = 0$. Finally, $\mathcal{E}_t^2 = 4\alpha_t^2\langle \mathbf{b}_t, \boldsymbol{\epsilon}_t\rangle^2$. By Cauchy–Schwarz,

$$\mathcal{E}_t^2 \;\le\; 4\alpha_t^2\|\mathbf{b}_t\|^2\|\boldsymbol{\epsilon}_t\|^2 \;\overset{①}{\le}\; 4\alpha_t^2\cdot 4\widehat{G}^2\|\boldsymbol{\epsilon}_t\|^2 \;=\; 16\widehat{G}^2\dot{\mathcal{E}}_t.$$

where step ① uses $\|\mathbf{b}_t\| \le (1-\alpha_t)(\|\mathbf{m}_{t-1}\| + \|\nabla f(\mathbf{x}_t)\|) \le 2\widehat{G}$.

□

To connect tracking accuracy with optimization progress, we combine proximal descent with the preceding error recursion. The logarithmic correction absorbs the normalized motion term and yields a joint Lyapunov inequality.

Lemma C.4. *(Proximal descent and a joint Lyapunov inequality for MVR1) Let* $C_f := \theta^2\kappa^2(2L+1)^2$ *and* $Q_t := F(\mathbf{x}_t) + 2\|\mathbf{s}_t\|^2 - C_f \ln A_{t-1}$. *Then, almost surely,*

$$F(\mathbf{x}_{t+1}) - F(\mathbf{x}_t) + \frac{\gamma_t}{2}\|\mathbf{r}_t\|^2 \le \frac{\alpha_t}{2}\|\mathbf{s}_t\|^2 + \frac{1+L}{2}\frac{\gamma_t^2}{\alpha_t}\|\mathbf{r}_t\|^2, \tag{52}$$

$$Q_{t+1} - Q_t + \frac{\gamma_t}{2}\|\mathbf{r}_t\|^2 + \frac{\alpha_t}{2}\|\mathbf{s}_t\|^2 \le 2\dot{\mathcal{E}}_{t+1} + 2\mathcal{E}_{t+1}. \tag{53}$$

Proof. Define $\mathbf{s}_t := \mathbf{m}_t - \nabla f(\mathbf{x}_t)$ and $\mathbf{r}_t := (\mathbf{x}_t - \mathbf{x}_{t+1})/\gamma_t$.

Let $A_t := 1 + \sum_{i=1}^t \|\mathbf{r}_i\|^2$, where $A_0 := 1$.

Let $\mathbf{x}_{t+1} \in \arg\min_{\mathbf{y}} \Big\{ h(\mathbf{y}) + \langle \mathbf{m}_t, \mathbf{y} - \mathbf{x}_t\rangle + \frac{1}{2\gamma_t}\|\mathbf{y} - \mathbf{x}_t\|^2 \Big\}$.

By the optimality of $\mathbf{x}_{t+1}$, we have:

$$h(\mathbf{x}_{t+1}) \le h(\mathbf{x}_t) - \langle \mathbf{m}_t, \mathbf{x}_{t+1} - \mathbf{x}_t\rangle - \frac{1}{2\gamma_t}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2.$$

By L-smoothness of f,

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

Part (a). Adding the two inequalities gives

$$\begin{aligned} F(\mathbf{x}_{t+1}) - F(\mathbf{x}_t) &\le \langle \mathbf{x}_{t+1} - \mathbf{x}_t, \nabla f(\mathbf{x}_t) - \mathbf{m}_t\rangle - \tfrac{1}{2\gamma_t}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 + \tfrac{L}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \\ &= \langle \gamma_t\mathbf{r}_t, \mathbf{s}_t\rangle - \tfrac{\gamma_t}{2}\|\mathbf{r}_t\|^2 + \tfrac{L\gamma_t^2}{2}\|\mathbf{r}_t\|^2 \\ &\overset{①}{\le} \tfrac{\alpha_t}{2}\|\mathbf{s}_t\|^2 + \tfrac{\gamma_t^2}{2\alpha_t}\|\mathbf{r}_t\|^2 - \tfrac{\gamma_t}{2}\|\mathbf{r}_t\|^2 + \tfrac{L\gamma_t^2}{2}\|\mathbf{r}_t\|^2 \\ &\le \tfrac{\alpha_t}{2}\|\mathbf{s}_t\|^2 - \tfrac{\gamma_t}{2}\|\mathbf{r}_t\|^2 + \tfrac{(L+1)\gamma_t^2}{2\alpha_t}\|\mathbf{r}_t\|^2, \end{aligned}$$

where step ① uses weighted Young inequality $\gamma_t\langle \mathbf{r}_t, \mathbf{s}_t\rangle \le \frac{\alpha_t}{2}\|\mathbf{s}_t\|^2 + \frac{\gamma_t^2}{2\alpha_t}\|\mathbf{r}_t\|^2$.

Part (b). We further derive:

$$\begin{aligned} & F(\mathbf{x}_{t+1}) - F(\mathbf{x}_t) + \tfrac{\gamma_t}{2}\|\mathbf{r}_t\|^2 + \tfrac{\alpha_t}{2}\|\mathbf{s}_t\|^2 \\ \le\ & \alpha_t\|\mathbf{s}_t\|^2 + \tfrac{(L+1)\gamma_t^2}{2\alpha_t}\|\mathbf{r}_t\|^2 \\ \overset{①}{\le}\ & 2\alpha_{t+1}\|\mathbf{s}_t\|^2 + \tfrac{(L+1)\gamma_t^2}{2\alpha_t}\|\mathbf{r}_t\|^2 \\ \overset{②}{\le}\ & 2\left(\|\mathbf{s}_t\|^2 - \|\mathbf{s}_{t+1}\|^2 + \tfrac{2L^2\gamma_t^2}{\alpha_t}\|\mathbf{r}_t\|^2 + \dot{\mathcal{E}}_{t+1} + \mathcal{E}_{t+1}\right) + \tfrac{(L+1)\gamma_t^2}{2\alpha_t}\|\mathbf{r}_t\|^2 \\ =\ & 2\left(\|\mathbf{s}_t\|^2 - \|\mathbf{s}_{t+1}\|^2 + \dot{\mathcal{E}}_{t+1} + \mathcal{E}_{t+1}\right) + \left(\tfrac{(L+1)}{2} + 4L^2\right)\cdot\left(\tfrac{\gamma_t^2}{\alpha_t}\|\mathbf{r}_t\|^2\right), \end{aligned} \tag{54}$$

where step ① uses $\frac{\alpha_t}{\alpha_{t+1}} = \frac{\rho_{t-1}}{\rho_t} \le 1 + p = \frac{3}{2} < 2$; step ② uses Lemma C.3(a) that $\alpha_t\|\mathbf{s}_{t-1}\|^2 \le \|\mathbf{s}_{t-1}\|^2 - \|\mathbf{s}_t\|^2 + \ddot{\mathcal{E}}_t + \dot{\mathcal{E}}_t + \mathcal{E}_t$.

Furthermore, we have:

$$\begin{aligned} & \left(\tfrac{(L+1)}{2} + 4L^2\right)\cdot\left(\tfrac{\gamma_t^2}{\alpha_t}\|\mathbf{r}_t\|^2\right) \overset{①}{\le} \left(2L+1\right)^2\cdot\left(\theta^2\beta_t^2\|\mathbf{r}_t\|^2\right) \\ \overset{②}{\le}\ & \left(2L+1\right)^2\cdot\left(\theta^2\kappa^2\beta_{t+1}^2\|\mathbf{r}_t\|^2\right) = \left(2L+1\right)^2\cdot\left(\theta^2\kappa^2\right)\cdot\tfrac{\|\mathbf{r}_t\|^2}{1+\sum_{i=1}^t\|\mathbf{r}_i\|^2} \\ \overset{③}{\le}\ & \left(2L+1\right)^2\cdot\left(\theta^2\kappa^2\right)\cdot\big(\ln(A_t) - \ln(A_{t-1})\big) = C_f\big(\ln(A_t) - \ln(A_{t-1})\big), \end{aligned} \tag{55}$$

where step ① uses $\gamma_t = \theta\sqrt{\alpha_t}\beta_t$; step ② uses $\frac{\beta_t}{\beta_{t+1}} \le \kappa$; step ③ uses Lemma A.8(a).

Define $Q_t := F(\mathbf{x}_t) + 2\|\mathbf{s}_t\|^2 - C_f \ln A_{t-1}$. In view of Inequalities (54) and (55), we have:

$$Q_{t+1} - Q_t + \tfrac{\gamma_t}{2}\|\mathbf{r}_t\|^2 + \tfrac{\alpha_t}{2}\|\mathbf{s}_t\|^2 \le 2\dot{\mathcal{E}}_{t+1} + 2\mathcal{E}_{t+1}.$$

□

The one-step estimates must now be converted into cumulative bounds. We first control the discrepancy and normalized motion sums, and then derive the first two moments of the weighted tracking error; the second-moment bound will be needed for the motion estimate.

Lemma C.5. *(Logarithmic bounds for MVR1) For every $T \geq 1$, we have:*

(a) $V_T := \sum_{t=1}^{T} \alpha_t^2 \chi_t^2 \leq C_v \ell_T$ *almost surely, where* $C_v = (1 + 4\widehat{G}^2)^2 \ln(e + 4\widehat{G}^2)$.

(b) $H_T := \sum_{t=1}^{T} \frac{\gamma_t^2}{\alpha_t} \|\mathbf{r}_t\|^2 \leq C_h \ell_T$ *almost surely, where* $C_h = 3\kappa^2\theta^2 \ln(e + \widehat{C}^2)$.

(c) $\mathbb{E}[\sum_{t=1}^{T} \dot{\mathcal{E}}_t] = \mathbb{E}[\sum_{t=1}^{T} \alpha_t^2 \|\boldsymbol{\epsilon}_t\|^2] \leq C_e \ell_T$, *where* $C_e := \widehat{G}^2 + C_v$.

(d) Let $U_T := \sum_{t=1}^{T} \alpha_t \|\mathbf{s}_t\|^2$. *We have:* $\mathbb{E}[U_T] \leq C_u \ell_T$, *where* $C_u = 4L^2 C_h + 4C_v$.

(e) $\mathbb{E}[U_T^2] \leq C_u' \ell_T^2 \leq C_u^2 \ell_T^2$, *where* $C_u' := \big(12\widehat{G}^2 + 4L^2 C_h + 2C_v\big) C_u$. *Here, the simpler choice* $C_u' = C_u^2$ *is also valid.*

Proof. Define $D_t = \sum_{i=1}^{t} \chi_i^2$ and $N_t = \max_{1 \leq i \leq t} \chi_i^2$, where $\chi_t = \|\mathbf{g}_{t-1} - \mathbf{g}_t\|$.

Define $V_T = \sum_{t=1}^{T} \alpha_t^2 \chi_t^2$. Let $A_t := 1 + \sum_{i=1}^{t} \|\mathbf{r}_i\|^2$. Let $c_\chi := (1 + 4\widehat{G}^2)^2$.

Define $C_v = c_\chi \ln(e + 4\widehat{G}^2)$. Denote $\ell_T = \ln(e + T)$.

Bounding the term $\sum_{t=2}^{T+1} \alpha_t^2 \chi_t^2$. We derive:

$$\begin{aligned}
&\textstyle\sum_{t=2}^{T+1} \alpha_t^2 \chi_t^2 \overset{①}{\leq} (1 + 4\widehat{G}^2) \sum_{t=2}^{T+1} \frac{\chi_t^2}{1+D_{t-1}} \overset{②}{\leq} (1 + 4\widehat{G}^2)^2 \sum_{t=2}^{T+1} \frac{\chi_t^2}{1+D_t} \\
\overset{③}{\leq}\ & c_\chi \textstyle\sum_{t=2}^{T+1} (\ln(1 + D_t) - \ln(1 + D_{t-1})) \leq c_\chi \ln(1 + D_{T+1}) \overset{④}{\leq} c_\chi \ln(1 + T4\widehat{G}^2),
\end{aligned}$$

where step ① uses $\alpha_t^2 \leq \frac{1+N_{t-1}}{1+D_{t-1}} \leq \frac{1+4\widehat{G}^2}{1+D_{t-1}}$; step ② uses $\frac{1+D_t}{1+D_{t-1}} = 1 + \frac{\chi_t^2}{1+D_{t-1}} \leq 1 + 4\widehat{G}^2$, which follows from $\chi_t^2 \leq 4\widehat{G}^2$; step ③ uses Lemma A.8(a); step ④ uses $D_{T+1} = \sum_{t=2}^{T+1} \chi_t^2 \leq 4\widehat{G}^2 T$.

Part (a). We derive:

$$\begin{aligned}
\textstyle\sum_{t=1}^{T} \alpha_t^2 \chi_t^2 &\leq \textstyle\sum_{t=1}^{T} \frac{1+N_{t-1}}{1+D_{t-1}} \chi_t^2 \overset{①}{\leq} \sum_{t=1}^{T} \frac{1+4\widehat{G}^2}{1+D_t} \cdot \frac{1+D_t}{1+D_{t-1}} \cdot \chi_t^2 \\
&\overset{②}{\leq} (1 + 4\widehat{G}^2)^2 \textstyle\sum_{t=1}^{T} \frac{\chi_t^2}{1+D_t} \overset{③}{\leq} c_\chi \sum_{t=1}^{T} (\ln(1 + D_t) - \ln(1 + D_{t-1})) \\
&\overset{④}{=} c_\chi \ln(1 + D_T) \leq c_\chi \ln(1 + T4\widehat{G}^2) \overset{⑤}{\leq} c_\chi \ln(e + 4\widehat{G}^2) \ln(e + T),
\end{aligned}$$

where step ① uses $N_{t-1} \leq 4\widehat{G}^2$; step ② uses $\frac{1+D_t}{1+D_{t-1}} = 1 + \frac{\chi_t^2}{1+D_{t-1}} \leq 1 + 4\widehat{G}^2$; step ③ uses Lemma A.8(a); step ④ uses $D_0 = 0$; step ⑤ uses $\ln(1 + ab) \leq \ln(e + a) \ln(e + b)$ for all $a, b \geq 0$.

Part (b). We derive:

$$\begin{aligned}
H_T &= \textstyle\sum_{t=1}^{T} \frac{\gamma_t^2}{\alpha_t} \|\mathbf{r}_t\|^2 \overset{①}{=} \theta^2 \sum_{t=1}^{T} \beta_t^2 \|\mathbf{r}_t\|^2 \overset{②}{\leq} \kappa^2\theta^2 \sum_{t=1}^{T} \beta_{t+1}^2 \|\mathbf{r}_t\|^2 \\
&\overset{③}{=} \kappa^2\theta^2 \textstyle\sum_{t=1}^{T} \frac{\|\mathbf{r}_t\|^2}{1+\sum_{i=1}^{t} \|\mathbf{r}_i\|^2} \overset{④}{\leq} \kappa^2\theta^2 \sum_{t=1}^{T} (\ln(A_t) - \ln(A_{t-1})) \\
&\overset{④}{\leq} \kappa^2\theta^2 \ln(A_T) \overset{⑤}{\leq} \kappa^2\theta^2 \ln(1 + \widehat{C}^2 \textstyle\sum_{t=1}^{T} t^2) \overset{⑥}{\leq} \kappa^2\theta^2 \ln(1 + \widehat{C}^2 T^3) \\
&\leq 3\kappa^2\theta^2 \ln(e + \widehat{C}^2) \ln(e + T) = C_h \ell_T,
\end{aligned}$$

where step ① uses $\gamma_t = \theta\sqrt{\alpha_t}\beta_t$; step ② uses Lemma C.2 (c) that $\frac{\beta_t}{\beta_{t+1}} \leq \kappa$; step ③ uses $\beta_t = A_{t-1}^{-1/2}$, where $A_{t-1} = 1 + \sum_{i=1}^{t-1} \|\mathbf{r}_i\|^2$; step ④ uses Lemma A.8(a); step ⑤ uses Lemma C.2(b) that $\|\mathbf{r}_t\| \leq \widehat{C}t$, leading to $A_T = 1 + \sum_{t=1}^{T} \|\mathbf{r}_t\|^2 \leq 1 + \widehat{C}^2 \sum_{t=1}^{T} t^2$; step ⑥ uses $\sum_{t=1}^{T} t^2 = \frac{T(T+1)(2T+1)}{6} \leq T^3$.

Part (c). The first sample needs to be handled separately, because $\mathbf{g}_0 = \mathbf{g}_1$ implies $\chi_1 = 0$. Since $\alpha_1 = 1$, unbiasedness at the fixed initial point gives

$$\mathbb{E}[\dot{\mathcal{E}}_1] \;=\; \mathbb{E}\|\boldsymbol{\epsilon}_1\|^2 \;=\; \mathbb{E}\|\mathbf{g}_1\|^2 - \|\nabla f(\mathbf{x}_1)\|^2 \;\le\; \widehat{G}^2.$$

For $t \ge 2$, $\nabla f(\mathbf{x}_t) - \mathbf{g}_{t-1}$ and α_t are $\mathcal{F}_{t-1}$-measurable. Hence

$$\mathbb{E}_t[\chi_t^2] \;=\; \|\nabla f(\mathbf{x}_t) - \mathbf{g}_{t-1}\|^2 + \mathbb{E}_t\|\boldsymbol{\epsilon}_t\|^2 \;\ge\; \mathbb{E}_t\|\boldsymbol{\epsilon}_t\|^2.$$

Using the tower property before summing yields

$$\mathbb{E}[\textstyle\sum_{t=1}^T \dot{\mathcal{E}}_t] \;\le\; \widehat{G}^2 + \mathbb{E}\sum_{t=2}^T[\alpha_t^2\chi_t^2] \;\le\; \widehat{G}^2 + C_v\ell_T \le (\widehat{G}^2 + C_v)\ell_T \;=\; C_e\ell_T.$$

Part (d). Applying Lemma C.3(a) at iteration $t+1$ gives

$$\begin{aligned}\|\mathbf{s}_{t+1}\|^2 &\le (1-\alpha_{t+1})\|\mathbf{s}_t\|^2 + \ddot{\mathcal{E}}_{t+1} + \dot{\mathcal{E}}_{t+1} + \mathcal{E}_{t+1}\\ &\overset{①}{\le} (1-\tfrac{1}{2}\alpha_t)\|\mathbf{s}_t\|^2 + \ddot{\mathcal{E}}_{t+1} + \dot{\mathcal{E}}_{t+1} + \mathcal{E}_{t+1},\end{aligned}$$

where step ① uses $\frac{\alpha_t}{\alpha_{t+1}} \le \frac{3}{2} < 2$. Consequently,

$$\alpha_t\|\mathbf{s}_t\|^2 \;\le\; 2\left(\|\mathbf{s}_t\|^2 - \|\mathbf{s}_{t+1}\|^2 + \ddot{\mathcal{E}}_{t+1} + \dot{\mathcal{E}}_{t+1} + \mathcal{E}_{t+1}\right).$$

Summing over $t = 1, \dots, T$ and taking the expectation of both sides, we obtain

$$\begin{aligned}\mathbb{E}[U_T] &\le \mathbb{E}\left[2\left(\|\mathbf{s}_1\|^2 - \|\mathbf{s}_{T+1}\|^2 + \textstyle\sum_{t=2}^{T+1}\ddot{\mathcal{E}}_t + \sum_{t=2}^{T+1}\dot{\mathcal{E}}_t + \sum_{t=2}^{T+1}\mathcal{E}_t\right)\right]\\ &\overset{①}{\le} 2\left(\widehat{G}^2 + \mathbb{E}[\textstyle\sum_{t=2}^{T+1}\ddot{\mathcal{E}}_t + \sum_{t=2}^{T+1}\dot{\mathcal{E}}_t]\right)\\ &\overset{②}{\le} 2\left(\widehat{G}^2 + 2L^2C_h\ell_T + \mathbb{E}[\textstyle\sum_{t=2}^{T+1}\alpha_t^2\chi_t^2]\right)\\ &\overset{③}{\le} 2\left(\widehat{G}^2 + 2L^2C_h\ell_T + c_\chi\ln(e + 4\widehat{G}^2)\ell_T\right)\\ &\le 2\left(\widehat{G}^2 + 2L^2C_h + c_\chi\ln(e+4\widehat{G}^2)\right)\ell_T\\ &\le \left(4L^2C_h + 4c_\chi\ln(e+4\widehat{G}^2)\right)\ell_T = (4L^2C_h + 4C_v)\ell_T,\end{aligned}$$

where step ① $\mathbb{E}\|\mathbf{s}_1\|^2 = \mathbb{E}\|\mathbf{m}_1 - \nabla f(\mathbf{x}_1)\|^2 = \mathbb{E}\|\mathbf{g}_1 - \nabla f(\mathbf{x}_1)\|^2 = \mathbb{E}\|\mathbf{g}_1\|^2 - \|\nabla f(\mathbf{x}_1)\|^2 \le \widehat{G}^2$ (which follows from $\mathbf{m}_1 = \mathbf{g}_1$), and $\mathbb{E}\big[\sum_{t=2}^{T+1}\mathcal{E}_t\big] = \sum_{t=2}^{T+1}\mathbb{E}\big[\mathbb{E}_t[\mathcal{E}_t]\big] = 0$ (which follows from Lemma C.3(b) that $\mathbb{E}_t[\mathcal{E}_t] = 0$ and the tower property); step ② uses $\ddot{\mathcal{E}}_t = \frac{2L^2\gamma_{t-1}^2}{\alpha_{t-1}}\|\mathbf{r}_{t-1}\|^2$, leading to $\sum_{t=2}^{T+1}\ddot{\mathcal{E}}_t = 2L^2\sum_{t=1}^T\frac{\gamma_t^2}{\alpha_t}\|\mathbf{r}_t\|^2 = 2L^2H_T \le 2L^2C_h\ell_T$; step ③ uses $\mathbb{E}[\sum_{t=2}^{T+1}\alpha_t^2\chi_t^2] \le c_\chi\ln(1 + T4\widehat{G}^2) \le c_\chi\ln(e+4\widehat{G}^2)\ell_T$.

Part (e). Let $q_t := \|\mathbf{s}_t\|^2$, $w_t := \alpha_t q_t$, $U_0 := 0$, and $\eta_t := \ddot{\mathcal{E}}_t + \dot{\mathcal{E}}_t + \mathcal{E}_t$. By Lemma C.2(a), $0 \le q_t, w_t \le 4\widehat{G}^2$. For $t \ge 2$, conditional unbiasedness gives

$$\mathbb{E}_t[\chi_t^2] = \|\nabla f(\mathbf{x}_t) - \mathbf{g}_{t-1}\|^2 + \mathbb{E}_t\|\boldsymbol{\epsilon}_t\|^2 \ge \mathbb{E}_t\|\boldsymbol{\epsilon}_t\|^2.$$

Since $U_{t-1}\alpha_t^2$ is $\mathcal{F}_{t-1}$-measurable, $U_0 = 0$, and $\mathbb{E}_t[\mathcal{E}_t] = 0$, parts (a)–(b) imply

$$\mathbb{E}\left[\textstyle\sum_{t=1}^T U_{t-1}\eta_t\right] \;\le\; \mathbb{E}\left[U_T\left(2L^2H_T + V_T\right)\right] \;\le\; (2L^2C_h + C_v)\ell_T\,\mathbb{E}[U_T]. \tag{56}$$

Here the initial noise has zero weight, and $\sum_{t=1}^T\ddot{\mathcal{E}}_t = 2L^2H_{T-1} \le 2L^2H_T$, with $H_0 := 0$.

The tracking recursion yields

$$w_t \le (1-\alpha_t)(q_{t-1} - q_t) + \eta_t. \tag{57}$$

Set $P_t := U_{t-1}(1-\alpha_t)$. Since U_t is nondecreasing and α_t is nonincreasing, P_t is nondecreasing and $P_1 = 0$. Summation by parts therefore gives

$$\textstyle\sum_{t=1}^T P_t(q_{t-1} - q_t) = -P_Tq_T + \sum_{t=1}^{T-1}(P_{t+1} - P_t)q_t \le 4\widehat{G}^2U_T. \tag{58}$$

Multiplying the recursion by U_{t-1}, summing, and using (56), we obtain

$$\mathbb{E}\left[\sum_{t=1}^{T} U_{t-1} w_t\right] \le \left[4\widehat{G}^2 + (2L^2 C_h + C_v)\ell_T\right]\mathbb{E}[U_T]. \tag{59}$$

Finally, $U_T^2 = \sum_{t=1}^{T} w_t^2 + 2\sum_{t=1}^{T} U_{t-1} w_t$ and $w_t^2 \le 4\widehat{G}^2 w_t$. Using part (d) and $\ell_T \ge 1$, we conclude that

$$\begin{aligned}\mathbb{E}[U_T^2] &\le \left[12\widehat{G}^2 + (4L^2 C_h + 2C_v)\ell_T\right]\mathbb{E}[U_T] \\ &\le (12\widehat{G}^2 + 4L^2 C_h + 2C_v) C_u \ell_T^2 = C_u' \ell_T^2.\end{aligned}$$

Moreover, $2C_v - 12\widehat{G}^2 \ge 2 + 4\widehat{G}^2 + 32\widehat{G}^4 \ge 0$. Thus $12\widehat{G}^2 + 4L^2 C_h + 2C_v \le C_u$, which proves $C_u' \le C_u^2$.

□

We next use proximal descent to transfer the tracking-error bounds to the descent energy. A pathwise comparison with the residual accumulator then controls the weighted cumulative motion, which is the remaining ingredient for the stationarity proof.

Lemma C.6. *(Descent energy and motion bridge for MVR1) Let $\Delta := F(\mathbf{x}_1) - F_\star$, $\ell_T := \ln(e+T)$. Let C_h and C_u be defined in Lemma C.5. Then, for every integer $T \ge 1$,*

(a) *$\mathbb{E}[W_T] \le C_w \ell_T$ and $\mathbb{E}[W_T^2] \le C_w^2 \ell_T^2$, where $C_w := 2\Delta + C_u + (1+L)C_h$.*

(b) *$\mathbb{E}[\alpha_T M_T] \le C_m \ell_T^2$, where $C_m := \frac{C_w}{\theta} + \frac{C_w^2}{4\theta^2}$.*

Proof. Let $C_w := 2\Delta + C_u + (1+L)C_h$.

Part (a-i). Summing the first inequality of Lemma C.4 over $t = 1, \dots, T$ gives

$$F(\mathbf{x}_{T+1}) - F(\mathbf{x}_1) + \frac{1}{2} W_T \le \frac{1}{2} U_T + \frac{1+L}{2} H_T.$$

Since $F(\mathbf{x}_{T+1}) \ge F_\star$, we obtain

$$0 \le W_T \le 2\Delta + U_T + (1+L)H_T. \tag{60}$$

Taking expectations and applying Lemma C.5(b),(d), we have

$$\mathbb{E}[W_T] \;\le\; 2\Delta + C_u \ell_T + (1+L)C_h \ell_T \;\le\; \big(2\Delta + C_u + (1+L)C_h\big)\ell_T \;=\; C_w \ell_T, \tag{61}$$

where the second inequality uses $\ell_T \ge 1$.

Part (a-ii). Set the deterministic quantity

$$a_T := 2\Delta + (1+L)C_h \ell_T \ge 0.$$

Since $H_T \le C_h \ell_T$ almost surely, (60) yields $0 \le W_T \le a_T + U_T$. Using Lemma C.5(d),(e), we derive

$$\begin{aligned}\mathbb{E}[W_T^2] &\le a_T^2 + 2a_T \mathbb{E}[U_T] + \mathbb{E}[U_T^2] \;\le\; a_T^2 + 2a_T C_u \ell_T + C_u^2 \ell_T^2 \\ &= (a_T + C_u \ell_T)^2 \;=\; \left[2\Delta + \big(C_u + (1+L)C_h\big)\ell_T\right]^2 \;\le\; C_w^2 \ell_T^2,\end{aligned}$$

where the last inequality again uses $\ell_T \ge 1$. This proves both estimates in part (a).

Part (b). Let $A_t := 1 + \sum_{i=1}^{t} \|\mathbf{r}_i\|^2$, with $A_0 := 1$.

The running-minimum definition of ρ_t ensures that $\alpha_t = \rho_{t-1}$ is nonincreasing. Combining this property with $\gamma_t = \theta\sqrt{\alpha_t}/\sqrt{A_{t-1}}$, we obtain

$$\begin{aligned}W_T &= \theta \textstyle\sum_{t=1}^{T} \sqrt{\alpha_t}\, \frac{A_t - A_{t-1}}{\sqrt{A_{t-1}}} \;\ge\; \theta\sqrt{\alpha_T} \textstyle\sum_{t=1}^{T} \frac{A_t - A_{t-1}}{\sqrt{A_{t-1}}} \\ &\ge 2\theta\sqrt{\alpha_T} \textstyle\sum_{t=1}^{T} \big(\sqrt{A_t} - \sqrt{A_{t-1}}\big) \;=\; 2\theta\sqrt{\alpha_T}\big(\sqrt{1+M_T} - 1\big).\end{aligned} \tag{62}$$

Here the second inequality follows from

$$\frac{A_t - A_{t-1}}{\sqrt{A_{t-1}}} \;=\; \big(\sqrt{A_t} - \sqrt{A_{t-1}}\big)\left(1 + \sqrt{\tfrac{A_t}{A_{t-1}}}\right) \;\geq\; 2\big(\sqrt{A_t} - \sqrt{A_{t-1}}\big), \tag{63}$$

because $A_t \geq A_{t-1}$. Rearranging (62) gives

$$\sqrt{\alpha_T(1 + M_T)} \leq \sqrt{\alpha_T} + \frac{W_T}{2\theta}.$$

Both sides are nonnegative. Squaring and subtracting α_T therefore yields

$$\alpha_T M_T \leq \frac{\sqrt{\alpha_T}}{\theta} W_T + \frac{W_T^2}{4\theta^2} \leq \frac{W_T}{\theta} + \frac{W_T^2}{4\theta^2}, \tag{64}$$

where the last inequality uses $0 < \alpha_T \leq 1$. Taking expectations and applying part (a), we conclude that

$$\begin{aligned} \mathbb{E}[\alpha_T M_T] \;&\leq\; \tfrac{1}{\theta}\mathbb{E}[W_T] + \tfrac{1}{4\theta^2}\mathbb{E}[W_T^2] \\ &\leq\; \tfrac{C_w}{\theta}\ell_T + \tfrac{C_w^2}{4\theta^2}\ell_T^2 \;\leq\; \left(\tfrac{C_w}{\theta} + \tfrac{C_w^2}{4\theta^2}\right)\ell_T^2 \;=\; C_m \ell_T^2, \end{aligned}$$

where the third inequality uses $\ell_T \geq 1$.

□

The preceding lemmas bound the weighted tracking error and cumulative motion. We finish by estimating the inverse adaptive weight in terms of the noise variance, removing the weights by CauchySchwarz, and applying proximal optimality at the uniformly sampled output.

Theorem C.7. *(Noise-adaptive composite stationarity of* ALFPG-MVR1*) Suppose Assumptions 3.1, 5.1, 5.2, 5.3, and 5.4 hold, and run the MVR1 branch of Algorithm 2 with any fixed $\theta > 0$. Let $R \sim \mathrm{Unif}\{1, \dots, T\}$ be independent of all algorithmic randomness, and define $\mathcal{R}_t := \mathrm{dist}\big(0, \nabla f(\mathbf{x}_t) + \partial h(\mathbf{x}_t)\big)$, $\ell_T := \ln(e + T)$. Let C_h, C_u, and C_m be the constants in Lemmas C.5 and C.6, and set $C_R := \big((1 + L\theta)\sqrt{C_m} + \sqrt{C_u}\big) \cdot \big(3 + 2L^{1/2}C_h^{1/4}\big)$. Then, for every integer $T \geq 1$,*

$$\mathbb{E}[\mathcal{R}_{R+1}] \leq C_R\big(\tfrac{\ell_T^{5/4}}{T^{1/2}} + \tfrac{\sigma^{1/2}\ell_T}{T^{1/4}}\big). \tag{65}$$

Thus ALFPG-MVR1 *attains $\widetilde{\mathcal{O}}(T^{-1/2} + \sigma^{1/2}T^{-1/4})$ using one stochastic-gradient evaluation per iteration. Hence, the stochastic-gradient oracle complexity for obtaining $\mathbb{E}[\mathcal{R}_{R+1}] \leq \epsilon$ is $\widetilde{\mathcal{O}}(\epsilon^{-2} + \sigma^2\epsilon^{-4})$.*

Proof. Define $D_t := \sum_{i=1}^t \chi_i^2$, $N_t := \max_{1 \leq i \leq t} \chi_i^2$, $S_T := \sum_{t=1}^T \|\mathbf{s}_t\|^2$, with $D_0 = N_0 = 0$.

Define $B_T := 1 + 3(T-1)\sigma^2 + 2L^2 C_h \ell_T$. Let $\chi_1 = 0$.

Bounding the term $\mathbb{E}[\alpha_T^{-1}]$. For $t \geq 2$, set $\mathbf{z}_t := \nabla f(\mathbf{x}_t) - \mathbf{g}_{t-1}$. Since $\mathbf{z}_t$ is $\mathcal{F}_{t-1}$-measurable and $\mathbb{E}_t[\boldsymbol{\epsilon}_t] = 0$, $\mathbb{E}_t[\chi_t^2] = \|\mathbf{z}_t\|^2 + \mathbb{E}_t\|\boldsymbol{\epsilon}_t\|^2$. Moreover, $\mathbf{z}_t = \nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1}) - \boldsymbol{\epsilon}_{t-1}$. By L-smoothness, the variance bound, and $\mathbf{x}_t - \mathbf{x}_{t-1} = -\gamma_{t-1}\mathbf{r}_{t-1}$, we obtain

$$\begin{aligned} \mathbb{E}[\chi_t^2] \;&\leq\; \mathbb{E}[2\|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1})\|^2 + 2\|\boldsymbol{\epsilon}_{t-1}\|^2 + \|\boldsymbol{\epsilon}_t\|^2] \\ &\leq\; 2L^2\mathbb{E}[\gamma_{t-1}^2\|\mathbf{r}_{t-1}\|^2] + 3\sigma^2. \end{aligned}$$

Since $\chi_1 = 0$, summing over $t = 2, \dots, T$ yields:

$$\begin{aligned} \mathbb{E}[D_T] \;&=\; \mathbb{E}[\textstyle\sum_{t=1}^T \chi_t^2] \leq 3(T-1)\sigma^2 + 2L^2\mathbb{E}\left[\sum_{t=1}^{T-1}\gamma_t^2\|\mathbf{r}_t\|^2\right] \\ &\leq\; 3(T-1)\sigma^2 + 2L^2\mathbb{E}[H_T] \;\leq\; 3(T-1)\sigma^2 + 2L^2 C_h \ell_T \;=\; -1 + B_T. \end{aligned} \tag{66}$$

We further derive for $T \geq 2$:

$$\mathbb{E}[\alpha_T^{-1}] \;\overset{①}{\leq}\; \mathbb{E}\sqrt{1 + D_T} \;\overset{②}{\leq}\; \sqrt{1 + \mathbb{E}[D_T]} \leq B_T^{1/2}, \tag{67}$$

where step ① uses the running-minimum definition of α_T for all $T \geq 1$ that $\alpha_T^{-2} = \max_{1 \leq k \leq T-1} \frac{1 + D_k}{1 + N_k} \leq 1 + D_{T-1} \leq 1 + D_T$; step ② uses Jensen's inequality; step ③ uses Inequality (66). Inequality (67) follows directly from $\alpha_1 = 1$ and $D_1 = 0$.

Bounding the composite residual. Global optimality of the proximal update implies $\mathbf{r}_t - \mathbf{m}_t \in \partial h(\mathbf{x}_{t+1})$. Consequently, L-smoothness and $\mathbf{s}_t = \mathbf{m}_t - \nabla f(\mathbf{x}_t)$ give

$$\begin{aligned}\mathcal{R}_{t+1} &:= \operatorname{dist}\big(0, \nabla f(\mathbf{x}_{t+1}) + \partial h(\mathbf{x}_{t+1})\big) \\ &\le \|\nabla f(\mathbf{x}_{t+1}) + \mathbf{r}_t - \mathbf{m}_t\| \le \|\mathbf{r}_t\| + \|\mathbf{s}_t\| + \|\nabla f(\mathbf{x}_{t+1}) - \nabla f(\mathbf{x}_t)\| \\ &\le (1 + L\gamma_t)\|\mathbf{r}_t\| + \|\mathbf{s}_t\| \le (1 + L\theta)\|\mathbf{r}_t\| + \|\mathbf{s}_t\|, \end{aligned} \tag{68}$$

where $\gamma_t = \theta\sqrt{\alpha_t}\beta_t \le \theta$.

Removing the adaptive weights. We have:

$$\mathbb{E}[\sqrt{M_T}] = \mathbb{E}\big[\sqrt{\alpha_T M_T}\sqrt{\alpha_T^{-1}}\big] \overset{①}{\le} \sqrt{\mathbb{E}[\alpha_T M_T]}\sqrt{\mathbb{E}[\alpha_T^{-1}]} \le \sqrt{C_m}\,\ell_T \cdot B_T^{1/4}, \tag{69}$$

where step ① uses Cauchy–Schwarz inequality; step ② uses Lemma C.6(b) and Inequality (67);

On the other hand, we derive:

$$\mathbb{E}[\sqrt{S_T}] \overset{①}{\le} \mathbb{E}\big[\sqrt{U_T}\sqrt{\alpha_T^{-1}}\big] \overset{②}{\le} \sqrt{\mathbb{E}[U_T]}\sqrt{\mathbb{E}[\alpha_T^{-1}]} \le \sqrt{C_u}\,\ell_T^{1/2} B_T^{1/4}, \tag{70}$$

where step ① uses $\alpha_T S_T \le \sum_{t=1}^T \alpha_t \|\mathbf{s}_t\|^2 = U_T$ since $\alpha_t \ge \alpha_T$ for $t \le T$; step ② uses Cauchy–Schwarz inequality; step ③ uses Lemma C.5(d).

Establishing the noise-adaptive rate. Since R is uniform and independent of the algorithm, we derive:

$$\begin{aligned}\mathbb{E}[\mathcal{R}_{R+1}] &= \tfrac{1}{T}\textstyle\sum_{t=1}^T \mathbb{E}[\mathcal{R}_{t+1}] \overset{①}{\le} \tfrac{1+L\theta}{T}\mathbb{E}\textstyle\sum_{t=1}^T \|\mathbf{r}_t\| + \tfrac{1}{T}\mathbb{E}\textstyle\sum_{t=1}^T \|\mathbf{s}_t\| \\ &\overset{②}{\le} \tfrac{1+L\theta}{\sqrt{T}}\mathbb{E}[\sqrt{M_T}] + \tfrac{1}{\sqrt{T}}\mathbb{E}[\sqrt{S_T}] \overset{③}{\le} \tfrac{(1+L\theta)\sqrt{C_m}\ell_T + \sqrt{C_u}\ell_T^{1/2}}{\sqrt{T}} B_T^{1/4} \\ &\le \big((1+L\theta)\sqrt{C_m} + \sqrt{C_u}\big) \cdot \tfrac{\ell_T}{\sqrt{T}} B_T^{1/4} \\ &\overset{④}{\le} \big((1+L\theta)\sqrt{C_m} + \sqrt{C_u}\big) \cdot \tfrac{\ell_T}{\sqrt{T}} \cdot \Big(3 + 2L^{1/2}C_h^{1/4}\Big)\Big(\ell_T^{1/4} + T^{1/4}\sigma^{1/2}\Big) \end{aligned} \tag{71}$$

where step ① uses Inequality (68); step ② uses Cauchy–Schwarz inequality; step ③ uses Inequalities (69) and (70); step ④ uses $B_T^{1/4} \le (1 + 2L^2 C_h \ell_T)^{1/4} + \big(3(T-1)\sigma^2\big)^{1/4} \le \Big(3 + 2L^{1/2}C_h^{1/4}\Big)\Big(\ell_T^{1/4} + T^{1/4}\sigma^{1/2}\Big)$.

□

C.3 Analysis for ALFPG-MVR2

Notation. Let $p = 2/3$ and $\lambda = 1/4$. Write $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_{t-1}]$.

Define $\boldsymbol{\epsilon}_t = \mathbf{g}_t - \nabla f(\mathbf{x}_t)$, $\boldsymbol{\epsilon}_t^- = \mathbf{g}_t^- - \nabla f(\mathbf{x}_{t-1})$, $\mathbf{s}_t = \mathbf{m}_t - \nabla f(\mathbf{x}_t)$, $\mathbf{r}_t = (\mathbf{x}_t - \mathbf{x}_{t+1})/\gamma_t$, $A_t = 1 + \sum_{i=1}^t \|\mathbf{r}_i\|^2$, $D_t = \sum_{i=1}^t \chi_i^2$, $N_t = \max_{1\le i\le t} \chi_i^2$, where $A_{-1} = A_0 = 1$, $D_0 = N_0 = 0$.

Set $\alpha_0 = 1$, $\gamma_0 = 0$, $\mathbf{r}_0 = \mathbf{0}$, and $\mathbf{s}_0 = -\nabla f(\mathbf{x}_1)$. Set $\mathbf{g}_0 = \mathbf{g}_1^- = \mathbf{g}_1$, leading to $\chi_1 = 0$.

Define $P_t = \frac{\|\mathbf{s}_t\|^2}{\sqrt{\alpha_t}}$, $U_T = \sum_{t=1}^T \alpha_t P_t = \sum_{t=1}^T \sqrt{\alpha_t}\|\mathbf{s}_t\|^2$, $W_T = \sum_{t=1}^T \gamma_t \|\mathbf{r}_t\|^2$, $H_T = \sum_{t=1}^T \frac{\gamma_t^2}{\sqrt{\alpha_t}}\|\mathbf{r}_t\|^2$, $M_T = \sum_{t=1}^T \|\mathbf{r}_t\|^2 = A_T - 1$, $S_T = \sum_{t=1}^T \|\mathbf{s}_t\|^2$. Set $\ell_T = \ln(e + T)$.

For MVR2, the same-sample correction changes the estimator bounds, so the MVR1 preliminary estimates cannot be reused directly. We first control the estimator increments and proximal residuals; in the convex case, the initial accumulator ratio is handled separately from subsequent iterations.

Lemma C.8. *(Preliminary bounds for MVR2) Let* $K_h := \begin{cases}\theta^{-1}\|\mathbf{x}_1 - \widehat{\mathbf{x}}_1\|, & h \text{ convex}, \\ \hat{H}, & \text{otherwise}, \end{cases}$, $\widehat{\mathbf{x}}_1 := \operatorname{Prox}_{\theta h}(\mathbf{x}_1)$. *Almost surely, for every* $t \ge 1$ *and* $T \ge 1$,

(a) $\|\mathbf{m}_t\| \le 2\widehat{G}/\alpha_t$ *and* $\|\mathbf{m}_t - \mathbf{m}_{t-1}\| \le 4\widehat{G}$.

(b) $\|\mathbf{r}_t\| \le C_r t$, *where* $C_r = 2K_h + 4\widehat{G}$.

(c) *If* h *is convex,* $A_t/A_{t-1} \le \kappa$ *for all* $t \ge 1$, *where* $\kappa := \max(2 + 16\widehat{G}^2, 1 + C_r^2)$.

Proof. Let $\alpha_t = \rho_{t-1}$. Define $C_r := 2K_h + 4\widehat{G}$.

Part (a). The stochastic momentum estimator is updated as $\mathbf{m}_t = (1-\alpha_t)(\mathbf{m}_{t-1} - \mathbf{g}_t^-) + \mathbf{g}_t$. Since the stochastic oracle is bounded, we derive:

$$\begin{aligned}\|\mathbf{m}_t\| &\le (1-\alpha_t)\|\mathbf{m}_{t-1}\| + (2-\alpha_t)\widehat{G} \\ &\overset{①}{\le} (1-\alpha_t)\tfrac{2\widehat{G}}{\alpha_t} + (2-\alpha_t)\widehat{G} \;=\; \tfrac{2\widehat{G}}{\alpha_t} - \alpha_t\widehat{G} \;\le\; \tfrac{2\widehat{G}}{\alpha_t},\end{aligned}$$

where step ① uses $\alpha_t \le \alpha_{t-1}$ and $\|\mathbf{m}_{t-1}\| \le 2\widehat{G}/\alpha_{t-1}$; the initial case follows from $\mathbf{m}_0 = \mathbf{0}$. Furthermore,

$$\|\mathbf{m}_t - \mathbf{m}_{t-1}\| \;\le\; \alpha_t\|\mathbf{m}_{t-1}\| + (2-\alpha_t)\widehat{G} \;\le\; 4\widehat{G}.$$

Part (b-i). Suppose that h is convex. Since $\gamma_1 = \theta$ and $\mathbf{m}_1 = \mathbf{g}_1$, nonexpansiveness of the proximal operator gives

$$\|\mathbf{r}_1\| \;\le\; \frac{\|\mathbf{x}_1 - \widehat{\mathbf{x}}_1\|}{\theta} + \frac{\|\widehat{\mathbf{x}}_1 - \mathbf{x}_2\|}{\theta} \;\le\; K_h + \widehat{G}.$$

For $t \ge 2$, proximal optimality implies $\mathbf{r}_t - \mathbf{m}_t \in \partial h(\mathbf{x}_{t+1})$ and $\mathbf{r}_{t-1} - \mathbf{m}_{t-1} \in \partial h(\mathbf{x}_t)$. Monotonicity of ∂h, together with $\mathbf{x}_{t+1} - \mathbf{x}_t = -\gamma_t \mathbf{r}_t$, yields

$$\|\mathbf{r}_t\|^2 \;\le\; \langle \mathbf{r}_{t-1} + \mathbf{m}_t - \mathbf{m}_{t-1}, \mathbf{r}_t\rangle \;\le\; \big(\|\mathbf{r}_{t-1}\| + 4\widehat{G}\big)\|\mathbf{r}_t\|.$$

Therefore $\|\mathbf{r}_t\| \le \|\mathbf{r}_{t-1}\| + 4\widehat{G}$, and hence $\|\mathbf{r}_t\| \le C_r \cdot t$.

Part (b-ii). Suppose that h is nonconvex and Lipschitz on $\mathrm{dom}(h)$. Global optimality of the proximal update, with $\mathbf{x}_t$ as a comparison point, gives

$$\begin{aligned}\frac{\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2}{2\gamma_t} &\le h(\mathbf{x}_t) - h(\mathbf{x}_{t+1}) - \langle \mathbf{m}_t, \mathbf{x}_{t+1} - \mathbf{x}_t\rangle \\ &\le (K_h + \|\mathbf{m}_t\|)\|\mathbf{x}_{t+1} - \mathbf{x}_t\|.\end{aligned}$$

If $\mathbf{x}_{t+1} = \mathbf{x}_t$, then $\mathbf{r}_t = \mathbf{0}$. Otherwise, dividing by $\|\mathbf{x}_{t+1} - \mathbf{x}_t\|$ gives

$$\|\mathbf{r}_t\| \;\le\; 2(K_h + \|\mathbf{m}_t\|) \;\le\; 2K_h + 4\widehat{G}t^{2/3} \;\le\; C_r \cdot t.$$

Part (c). For $t \ge 2$, we derive:

$$\tfrac{A_t}{A_{t-1}} \;\overset{①}{=}\; \tfrac{A_{t-1} + \|\mathbf{r}_t\|^2}{A_{t-1}} \;\le\; 1 + \tfrac{\|\mathbf{r}_t\|^2}{A_{t-1}} \;\overset{②}{\le}\; 1 + \tfrac{(\|\mathbf{r}_{t-1}\| + 4\widehat{G})^2}{1 + \|\mathbf{r}_{t-1}\|^2} \;\overset{③}{\le}\; 1 + 1 + 16\widehat{G}^2,$$

where step ① uses the definition of $A_t := 1 + \sum_{i=1}^t \|\mathbf{r}_i\|^2$; step ② uses $\|\mathbf{r}_t\| \le \|\mathbf{r}_{t-1}\| + 4\widehat{G}$; step ③ uses $(a+b)^2 \le (1+a^2)(1+b^2)$ for all a, b. For $t = 1$, we have: $\frac{A_t}{A_{t-1}} = 1 + \|\mathbf{r}_1\|^2 \le 1 + C_r^2$. Taking both cases into consideration, we have $\frac{A_t}{A_{t-1}} \le \kappa := \max(2 + 16\widehat{G}^2, 1 + C_r^2)$.

□

We now exploit individual smoothness to control the difference of two gradients evaluated with the same sample. Rescaling the tracking error produces a contractive recursion with an explicit innovation term and a conditionally mean-zero cross term.

Lemma C.9. *(Tracking-error recursion for MVR2) Retain the notation introduced above, and define*

$$\boldsymbol{\delta}_t := \boldsymbol{\epsilon}_t - \boldsymbol{\epsilon}_t^-, \qquad \boldsymbol{\zeta}_t := \frac{\boldsymbol{\delta}_t + \alpha_t \boldsymbol{\epsilon}_t^-}{\alpha_t^{1/4}}, \qquad \varpi_t := (1-\alpha_t)\left(\frac{\alpha_{t-1}}{\alpha_t}\right)^{1/4},$$

$$\mathcal{E}_t := 2\varpi_t \langle \mathbf{z}_{t-1}, \boldsymbol{\zeta}_t \rangle, \qquad \mathcal{V}_t := \|\boldsymbol{\zeta}_t\|^2, \qquad \dot{\mathcal{E}}_t := \alpha_t^{3/2}\|\boldsymbol{\epsilon}_t^-\|^2, \quad \ddot{\mathcal{E}}_t := 11L^2 \frac{\gamma_{t-1}^2}{\sqrt{\alpha_{t-1}}}\|\mathbf{r}_{t-1}\|^2.$$

Then, almost surely, for every $t \geq 1$:

(a) *$\varpi_t^2 \leq 1-\alpha_t$, and $\mathcal{V}_t \leq \ddot{\mathcal{E}}_t + 2\dot{\mathcal{E}}_t$. In particular,*

$$P_t = \varpi_t^2 P_{t-1} + \mathcal{V}_t + \mathcal{E}_t \leq (1-\alpha_t)P_{t-1} + \ddot{\mathcal{E}}_t + 2\dot{\mathcal{E}}_t + \mathcal{E}_t. \tag{72}$$

(b) *$\mathbb{E}_t[\mathcal{E}_t] = 0$, and $\mathcal{E}_t^2 \leq 4P_{t-1}\mathcal{V}_t$.*

Proof. Define $\mathbf{s}_t = \mathbf{m}_t - \nabla f(\mathbf{x}_t)$, and $\mathbf{z}_t = \mathbf{s}_t/\alpha_t^{1/4}$, where $\mathbf{z}_0 = \mathbf{0}$ and $P_0 = 0$.

Define the random increment $\boldsymbol{\eta}_t := \boldsymbol{\epsilon}_t - (1-\alpha_t)\boldsymbol{\epsilon}_t^- = \boldsymbol{\delta}_t + \alpha_t\boldsymbol{\epsilon}_t^-$. Let $\alpha_t = \rho_{t-1}$.

Exact decomposition of the tracking error. Using the MVR2 update, we obtain

$$\begin{aligned}\mathbf{s}_t &= (1-\alpha_t)(\mathbf{m}_{t-1} - \mathbf{g}_t^-) + \mathbf{g}_t - \nabla f(\mathbf{x}_t) \\ &= (1-\alpha_t)\mathbf{s}_{t-1} + \boldsymbol{\epsilon}_t - (1-\alpha_t)\boldsymbol{\epsilon}_t^- \\ &= (1-\alpha_t)\mathbf{s}_{t-1} + \boldsymbol{\eta}_t.\end{aligned} \tag{73}$$

Dividing by $\alpha_t^{1/4}$ gives

$$\mathbf{z}_t = \varpi_t \mathbf{z}_{t-1} + \boldsymbol{\zeta}_t.$$

At $t=1$, $\alpha_1 = 1$ and $\boldsymbol{\delta}_1 = \mathbf{0}$, so this identity also holds under the convention $\mathbf{z}_0 = \mathbf{0}$.

Bounding the term $\mathcal{V}_t := \|\boldsymbol{\zeta}_t\|^2$. By individual smoothness and smoothness of the expected loss,

$$\begin{aligned}\|\boldsymbol{\delta}_t\| &= \|\mathbf{g}_t - \mathbf{g}_t^- - \nabla f(\mathbf{x}_t) + \nabla f(\mathbf{x}_{t-1})\| \leq \|\mathbf{g}_t - \mathbf{g}_t^-\| + \|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1})\| \\ &\leq 2L\|\mathbf{x}_t - \mathbf{x}_{t-1}\| = 2L\gamma_{t-1}\|\mathbf{r}_{t-1}\|.\end{aligned}$$

The last equality also holds at $t=1$ because $\mathbf{x}_0 = \mathbf{x}_1$ and $\gamma_0 = 0$. Consequently,

$$\|\boldsymbol{\eta}_t\|^2 = \|\boldsymbol{\delta}_t + \alpha_t\boldsymbol{\epsilon}_t^-\|^2 \leq 2\|\boldsymbol{\delta}_t\|^2 + 2\alpha_t^2\|\boldsymbol{\epsilon}_t^-\|^2 \leq 8L^2\gamma_{t-1}^2\|\mathbf{r}_{t-1}\|^2 + 2\alpha_t^2\|\boldsymbol{\epsilon}_t^-\|^2.$$

Dividing both sides by $\sqrt{\alpha_t}$ and using the definition of $\mathcal{V}_t = \frac{\|\boldsymbol{\eta}_t\|^2}{\sqrt{\alpha_t}}$ yields:

$$\mathcal{V}_t \leq \frac{8L^2\gamma_{t-1}^2}{\sqrt{\alpha_t}}\|\mathbf{r}_{t-1}\|^2 + 2\alpha_t^{3/2}\|\boldsymbol{\epsilon}_t^-\|^2 \overset{①}{\leq} 8L^2\sqrt{\tfrac{5}{3}}\frac{\gamma_{t-1}^2}{\sqrt{\alpha_{t-1}}}\|\mathbf{r}_{t-1}\|^2 + 2\dot{\mathcal{E}}_t \leq \ddot{\mathcal{E}}_t + 2\dot{\mathcal{E}}_t, \tag{74}$$

where step ① uses $\alpha_{t-1}/\alpha_t = \rho_{t-2}/\rho_{t-1} \leq 1 + 2/3$ by Lemma C.1(a).

Part (a). For all $t \geq 2$, we derive:

$$\varpi_t^2 \overset{①}{=} (1-\alpha_t)^2\sqrt{\frac{\alpha_{t-1}}{\alpha_t}} \overset{②}{\leq} \frac{(1-\alpha_t)^2}{\sqrt{1-\frac{2}{3}\alpha_t}} \overset{③}{\leq} 1-\alpha_t, \tag{75}$$

where step ① uses the definition of ϖ; step ② uses $\frac{1}{\alpha_t} - \frac{1}{\alpha_{t-1}} \leq \frac{2}{3}$, leading to $\frac{\alpha_{t-1}}{\alpha_t} \leq \frac{1}{1-\frac{2}{3}\alpha_t}$; step ③ uses $(1-a) \leq \sqrt{1-\frac{2}{3}a}$ for $a \in [0,1]$. Inequality (75) also holds at $t=1$ because $\varpi_1 = 0$.

Expanding the squared norm of the rescaled recursion yields

$$\begin{aligned}P_t &= \|\varpi_t\mathbf{z}_{t-1} + \boldsymbol{\zeta}_t\|^2 = \varpi_t^2 P_{t-1} + \|\boldsymbol{\zeta}_t\|^2 + 2\varpi_t\langle \mathbf{z}_{t-1}, \boldsymbol{\zeta}_t\rangle \\ &= \varpi_t^2 P_{t-1} + \mathcal{V}_t + \mathcal{E}_t \overset{①}{\leq} (1-\alpha_t)P_{t-1} + \ddot{\mathcal{E}}_t + 2\dot{\mathcal{E}}_t + \mathcal{E}_t,\end{aligned}$$

where step ① uses Inequality (74), $\varpi_t^2 \leq 1-\alpha_t$, and $P_{t-1} \geq 0$.

Part (b). By conditional unbiasedness, $\mathbb{E}_t[\boldsymbol{\epsilon}_t] = \mathbf{0}$, and $\mathbb{E}_t[\boldsymbol{\epsilon}_t^-] = \mathbf{0}$. Since α_t is $\mathcal{F}_{t-1}$-measurable,

$$\mathbb{E}_t[\boldsymbol{\zeta}_t] = \alpha_t^{-1/4}\mathbb{E}_t[\boldsymbol{\epsilon}_t - (1-\alpha_t)\boldsymbol{\epsilon}_t^-] = \mathbf{0}.$$

Moreover, ϖ_t and $\mathbf{z}_{t-1}$ are $\mathcal{F}_{t-1}$-measurable. Therefore,

$$\mathbb{E}_t[\mathcal{E}_t] = 2\varpi_t\langle \mathbf{z}_{t-1}, \mathbb{E}_t[\boldsymbol{\zeta}_t]\rangle = 0.$$

Finally, Cauchy–Schwarz and $\varpi_t^2 \le 1$ give

$$\mathcal{E}_t^2 = 4\varpi_t^2\langle \mathbf{z}_{t-1}, \boldsymbol{\zeta}_t\rangle^2 \le 4\varpi_t^2 P_{t-1}\mathcal{V}_t \le 4P_{t-1}\mathcal{V}_t.$$

□

The rescaled recursion is now coupled to proximal descent. The resulting Lyapunov inequality tracks both objective progress and estimator accuracy, with the normalized motion accumulated in the correction term.

Lemma C.10. *(Proximal descent and a joint Lyapunov inequality for MVR2) Let $C_f := (1+L)/2 + 22L^2$ and $\mathcal{Q}_t := F(\mathbf{x}_t) + 2P_t - C_f H_{t-1}$. Then, almost surely,*

$$F(\mathbf{x}_{t+1}) - F(\mathbf{x}_t) + \frac{\gamma_t}{2}\|\mathbf{r}_t\|^2 \le \frac{\sqrt{\alpha_t}}{2}\|\mathbf{s}_t\|^2 + \frac{1+L}{2}\frac{\gamma_t^2}{\sqrt{\alpha_t}}\|\mathbf{r}_t\|^2, \tag{76}$$

$$\mathcal{Q}_{t+1} - \mathcal{Q}_t + \frac{\gamma_t}{2}\|\mathbf{r}_t\|^2 + \frac{\sqrt{\alpha_t}}{2}\|\mathbf{s}_t\|^2 \le 4\dot{\mathcal{E}}_{t+1} + 2\mathcal{E}_{t+1}. \tag{77}$$

Proof. By global optimality of the proximal update and L-smoothness of f,

$$h(\mathbf{x}_{t+1}) \le h(\mathbf{x}_t) - \langle \mathbf{m}_t, \mathbf{x}_{t+1} - \mathbf{x}_t\rangle - \frac{\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2}{2\gamma_t},$$
$$f(\mathbf{x}_{t+1}) \le f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}_t\rangle + \frac{L}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2.$$

Part (a). Adding the preceding two inequalities and using $\mathbf{x}_{t+1} - \mathbf{x}_t = -\gamma_t\mathbf{r}_t$, we derive:

$$\begin{aligned} F(\mathbf{x}_{t+1}) - F(\mathbf{x}_t) &\le \gamma_t\langle \mathbf{r}_t, \mathbf{s}_t\rangle - \frac{\gamma_t}{2}\|\mathbf{r}_t\|^2 + \frac{L\gamma_t^2}{2}\|\mathbf{r}_t\|^2 \\ &\overset{①}{\le} \frac{\sqrt{\alpha_t}}{2}\|\mathbf{s}_t\|^2 + \frac{\gamma_t^2}{2\sqrt{\alpha_t}}\|\mathbf{r}_t\|^2 - \frac{\gamma_t}{2}\|\mathbf{r}_t\|^2 + \frac{L\gamma_t^2}{2}\|\mathbf{r}_t\|^2 \\ &\le \frac{\sqrt{\alpha_t}}{2}\|\mathbf{s}_t\|^2 - \frac{\gamma_t}{2}\|\mathbf{r}_t\|^2 + \frac{1+L}{2}\frac{\gamma_t^2}{\sqrt{\alpha_t}}\|\mathbf{r}_t\|^2, \end{aligned}$$

where step ① uses the weighted Young inequality $\gamma_t\langle \mathbf{r}_t, \mathbf{s}_t\rangle \le \frac{\sqrt{\alpha_t}}{2}\|\mathbf{s}_t\|^2 + \frac{\gamma_t^2}{2\sqrt{\alpha_t}}\|\mathbf{r}_t\|^2$.

Part (b). Since $\sqrt{\alpha_t}\|\mathbf{s}_t\|^2 = \alpha_t P_t$, part (a) gives

$$\begin{aligned} &F(\mathbf{x}_{t+1}) - F(\mathbf{x}_t) + \frac{\gamma_t}{2}\|\mathbf{r}_t\|^2 + \frac{\sqrt{\alpha_t}}{2}\|\mathbf{s}_t\|^2 \le \alpha_t P_t + \tfrac{1+L}{2}\tfrac{\gamma_t^2}{\sqrt{\alpha_t}}\|\mathbf{r}_t\|^2 \\ \overset{①}{\le}\ & 2\alpha_{t+1}P_t + \frac{1+L}{2}\frac{\gamma_t^2}{\sqrt{\alpha_t}}\|\mathbf{r}_t\|^2 \overset{②}{\le} 2(P_t - P_{t+1}) + C_f\frac{\gamma_t^2}{\sqrt{\alpha_t}}\|\mathbf{r}_t\|^2 + 4\dot{\mathcal{E}}_{t+1} + 2\mathcal{E}_{t+1}, \end{aligned}$$

where step ① uses $\alpha_t/\alpha_{t+1} \le 1 + 2/3 < 2$; step ② uses Lemma C.9(a) and $\ddot{\mathcal{E}}_{t+1} = 11L^2\gamma_t^2\|\mathbf{r}_t\|^2/\sqrt{\alpha_t}$. Finally, $H_t - H_{t-1} = \gamma_t^2\|\mathbf{r}_t\|^2/\sqrt{\alpha_t}$. Substituting the definition of $\mathcal{Q}_t$ proves (77). □

We next establish the cumulative estimates needed to sum the MVR2 descent inequality. The proof treats the convex and domain-Lipschitz motion bounds separately and controls second moments without assuming that the rescaled tracking error is uniformly bounded.

Lemma C.11. *(Logarithmic bounds for MVR2) Then, for every $T \ge 1$,*

(a) *$V_T := \sum_{t=1}^T \alpha_t^{3/2}\chi_t^2 \le C_v \ell_T$ almost surely, where $C_v := 2(1+4\widehat{G}^2)^2\ln(e+4\widehat{G}^2)$.*

(b) *Let $Q_T := \sum_{t=1}^T \dot{\mathcal{E}}_t$. We have: $\mathbb{E}[Q_T] \le C_q \ell_T$ and $\mathbb{E}[Q_T^2] \le C_q^2 \ell_T^2$, where $C_q := 8\widehat{G}^2 + 2C_v$.*

(c) *Let $H_T := \sum_{t=1}^T \frac{\gamma_t^2}{\sqrt{\alpha_t}}\|\mathbf{r}_t\|^2$. We have: $\mathbb{E}[H_T] \le C_h \ell_T$ and $\mathbb{E}[H_T^2] \le C_h^2\ell_T^2$, where $C_h := \begin{cases}\dot{C}_h, & h \text{ convex}\\ \ddot{C}_h, & \text{otherwise}\end{cases}$, $\dot{C}_h := 6\theta^2\kappa\ln(e+C_r)$, $\ddot{C}_h := \theta^2\big[C_A + B_h + 80(11L^2\theta^2+2C_q)\big]$, $C_A := 3+\ln(1+C_r^2)$, and $B_h := 8(K_h+\widehat{G})^2$.*

(d) *Let $U_T := \sum_{t=1}^T \sqrt{\alpha_t}\|\mathbf{s}_t\|^2 = \sum_{t=1}^T \alpha_t P_t$. Then $\mathbb{E}[U_T] \le C_u\ell_T$ and $\mathbb{E}[U_T^2] \le C_u^2\ell_T^2$. Here, $C_u := 20(11L^2C_h + 2C_q)$.*

Proof. Define $\chi_t := \|\mathbf{g}_{t-1} - \mathbf{g}_t^-\|$. Let $c_\chi := (1+4\widehat{G}^2)^2$.

Define $\alpha_t = \rho_{t-1}$, where $\rho_t := \min_{1\le k\le t}\left(\frac{1+\max_{1\le i\le k}\chi_i^2}{1+\sum_{i=1}^k\chi_i^2}\right)^{2/3}$.

Define $D_t := \sum_{i=1}^t \chi_i^2$, and $N_t := \max_{1\le i\le t}\chi_i^2$.

Part (a). We derive:

$$\begin{aligned} V_T &\le V_{T+1} \overset{①}{=} \textstyle\sum_{t=2}^{T+1}\alpha_t^{3/2}\chi_t^2 \overset{②}{\le} (1+4\widehat{G}^2)\sum_{t=2}^{T+1}\frac{\chi_t^2}{1+D_{t-1}} \\ &\overset{③}{\le} (1+4\widehat{G}^2)^2\textstyle\sum_{t=2}^{T+1}\frac{\chi_t^2}{1+D_t} \overset{④}{\le} c_\chi\big(\ln(1+D_{T+1}) - \ln(1+D_1)\big) \\ &\le c_\chi\ln(1+4\widehat{G}^2(T+1)) \overset{⑤}{\le} \underbrace{2c_\chi\ln(e+4\widehat{G}^2)}_{:=C_v}\ell_T. \end{aligned}$$

where step ① uses $\chi_1 = 0$; step ② uses $\alpha_t^{3/2} \le (1+N_{t-1})/(1+D_{t-1})$ and $N_{t-1} \le 4\widehat{G}^2$; step ③ uses $(1+D_t)/(1+D_{t-1}) \le 1+4\widehat{G}^2$; step ④ uses $\chi_t^2/(1+D_t) \le \ln(1+D_t) - \ln(1+D_{t-1})$; step ⑤ uses $\ln(1+a(T+1)) \le \ln(e+a)\ln(e+T+1) \le 2\ln(e+a)\ln(e+T)$ for all $a \ge 0$.

Part (b-i). Let $q_t := \dot{\mathcal{E}}_t = \alpha_t^{3/2}\|\boldsymbol{\epsilon}_t^-\|^2$ and $Q_t := \sum_{i=1}^t q_i$, where $Q_0 = 0$. Since $0 < \alpha_t \le 1$ and $\|\boldsymbol{\epsilon}_t^-\| \le 2\widehat{G}$, we have

$$0 \le q_t \le 4\widehat{G}^2, \qquad q_t^2 \le 4\widehat{G}^2 q_t.$$

For $t \ge 2$, recall that $\chi_t = \|\boldsymbol{\epsilon}_{t-1} - \boldsymbol{\epsilon}_t^-\|$. Since $\boldsymbol{\epsilon}_{t-1}$ is $\mathcal{F}_{t-1}$-measurable and $\mathbb{E}_t[\boldsymbol{\epsilon}_t^-] = \mathbf{0}$, we obtain

$$\mathbb{E}_t[\chi_t^2] = \|\boldsymbol{\epsilon}_{t-1}\|^2 + \mathbb{E}_t\|\boldsymbol{\epsilon}_t^-\|^2 \ge \mathbb{E}_t\|\boldsymbol{\epsilon}_t^-\|^2.$$

Because α_t is $\mathcal{F}_{t-1}$-measurable, $\mathbb{E}_t[q_t] \le \alpha_t^{3/2}\mathbb{E}_t[\chi_t^2]$ for all $t \ge 2$.

We obtain:

$$\begin{aligned} V_{T+1} &\le (1+4\widehat{G}^2)^2\ln(1+D_{T+1}) \overset{①}{\le} (1+4\widehat{G}^2)^2\ln(1+4\widehat{G}^2T) \\ &\le (1+4\widehat{G}^2)^2\ln(e+4\widehat{G}^2)\ell_T. \end{aligned}$$

where step ① uses $D_{T+1} = \sum_{t=2}^{T+1}\chi_t^2 \le 4\widehat{G}^2T$ since $\chi_1 = 0$ and $\chi_t^2 \le 4\widehat{G}^2$. Taking expectations, summing the conditional comparison over $t = 2,\dots,T+1$, and bounding q_1 separately therefore gives

$$\begin{aligned} \mathbb{E}[Q_T] &\le \mathbb{E}[Q_{T+1}] \le 4\widehat{G}^2 + \mathbb{E}\left[\textstyle\sum_{t=2}^{T+1}\alpha_t^{3/2}\chi_t^2\right] \\ &\le 4\widehat{G}^2 + \tfrac{C_v}{2}\ell_T \le \tfrac{C_v}{2} + \tfrac{C_v}{2}\ell_T \le C_v\ell_T. \end{aligned}$$

Part (b-ii). The identity $Q_t^2 - Q_{t-1}^2 = 2Q_{t-1}q_t + q_t^2$, together with $Q_0 = 0$, gives

$$\mathbb{E}[Q_{T+1}^2] = 2\mathbb{E}\left[\sum_{t=1}^{T+1} Q_{t-1}q_t\right] + \mathbb{E}\left[\sum_{t=1}^{T+1} q_t^2\right].$$

For $t \geq 2$, the measurability of $Q_{t-1}\alpha_t^{3/2}$ gives

$$\mathbb{E}[Q_{t-1}q_t] \;=\; \mathbb{E}\left[Q_{t-1}\alpha_t^{3/2}\mathbb{E}_t\|\boldsymbol{\epsilon}_t^-\|^2\right] \;\leq\; \mathbb{E}\left[Q_{t-1}\alpha_t^{3/2}\mathbb{E}_t[\chi_t^2]\right] \;=\; \mathbb{E}\left[Q_{t-1}\alpha_t^{3/2}\chi_t^2\right].$$

Consequently,

$$\begin{aligned}\mathbb{E}[Q_{T+1}^2] &\overset{①}{\leq} 4\widehat{G}^2\mathbb{E}[Q_{T+1}] + 2\mathbb{E}\left[\sum_{t=2}^{T+1} Q_{t-1}\alpha_t^{3/2}\chi_t^2\right] \overset{②}{\leq} 4\widehat{G}^2\mathbb{E}[Q_{T+1}] + 2\mathbb{E}[Q_{T+1}V_{T+1}] \\ &\overset{③}{\leq} (4\widehat{G}^2 + 2C_v\ell_T)\mathbb{E}[Q_{T+1}] \overset{④}{\leq} (4\widehat{G}^2 + 2C_v\ell_T)(4\widehat{G}^2 + C_v\ell_T) \\ &\leq 2(4\widehat{G}^2 + C_v\ell_T)^2 \leq C_q^2\ell_T^2, \end{aligned} \tag{78}$$

where step ① uses $q_t^2 \leq 4\widehat{G}^2 q_t$ and $Q_0 = 0$; step ② uses $Q_{t-1} \leq Q_{T+1}$; step ③ uses $V_{T+1} \leq C_v\ell_T$; step ④ uses $\mathbb{E}[Q_{T+1}] \leq 4\widehat{G}^2 + C_v\ell_T$.

Part (c-i). We consider the case when $h(\cdot)$ is convex. We have almost surely:

$$\begin{aligned} H_T &= \textstyle\sum_{t=1}^T \frac{\gamma_t^2}{\sqrt{\alpha_t}}\|\mathbf{r}_t\|^2 \overset{①}{=} \theta^2 \sum_{t=1}^T \frac{\|\mathbf{r}_t\|^2}{A_{t-1}} \overset{②}{\leq} \theta^2\kappa \sum_{t=1}^T \frac{\|\mathbf{r}_t\|^2}{A_t} \overset{③}{\leq} \theta^2\kappa\ln(A_T) \\ &\overset{④}{\leq} \theta^2\kappa\ln(1 + TC_r^2T^2) \;\leq\; \theta^2\kappa\ln(e + T^3)\ln(e + C_r^2) \;\leq\; 6\theta^2\kappa\ln(e + C_r)\ell_T, \end{aligned}$$

where step ① uses $\gamma_t = \theta\alpha_t^{1/4}\beta_t = \theta\alpha_t^{1/4}A_{t-1}^{-1/2}$; step ② uses $A_t/A_{t-1} \leq \kappa$; step ③ uses $\frac{\|\mathbf{r}_t\|^2}{A_t} \leq \ln A_t - \ln A_{t-1}$; step ④ uses $A_T = 1 + \sum_{t=1}^T \|\mathbf{r}_t\|^2$ and $\|\mathbf{r}_t\| \leq C_r t$. Letting $C_h = \dot{C}_h = 6\theta^2\kappa\ln(e + C_r)$, we directly conclude that

$$\mathbb{E}[H_T^2] \leq C_h^2\ell_T^2, \qquad \mathbb{E}[H_T] \leq C_h\ell_T.$$

Part (c-ii). We consider the case when $h(\cdot)$ is Lipschitz on $\mathrm{dom}(h)$ with constant K_h (convexity is not required). Define $A_t = 1 + \sum_{i=1}^t \|\mathbf{r}_i\|^2$, and $P_t = \frac{\|\mathbf{s}_t\|^2}{\sqrt{\alpha_t}}$, $C_r := 2K_h + 4\widehat{G}$, $C_A := 3 + \ln(1 + C_r^2)$, $B_h := 8(K_h + \widehat{G})^2$, $C_v := 2(1 + 4\widehat{G}^2)^2\ln(e + 4\widehat{G}^2)$, $C_q := 8\widehat{G}^2 + 2C_v$.

Define $v_t := \frac{\mathcal{V}_t}{A_{t-1}}$, $e_t := \frac{\mathcal{E}_t}{A_{t-1}}$, $\overline{B}_t := \sum_{i=1}^t v_i$, $\overline{Z}_t := \sum_{i=1}^t e_i$, with $\overline{B}_0 = \overline{Z}_0 = 0$.

- *A telescoping potential.* Define $\Psi_t := \frac{B_h + 8P_t}{A_{t-1}}$. Global optimality of the proximal update gives

$$\frac{\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2}{2\gamma_t} \;\leq\; h(\mathbf{x}_t) - h(\mathbf{x}_{t+1}) - \langle \mathbf{m}_t, \mathbf{x}_{t+1} - \mathbf{x}_t\rangle \;\leq\; (K_h + \|\mathbf{m}_t\|)\|\mathbf{x}_{t+1} - \mathbf{x}_t\|. \tag{79}$$

Thus $\|\mathbf{r}_t\| \leq 2(K_h + \|\mathbf{m}_t\|)$, also when $\mathbf{x}_{t+1} = \mathbf{x}_t$. Since $\|\mathbf{m}_t\| \leq \widehat{G} + \|\mathbf{s}_t\|$ and $\|\mathbf{s}_t\|^2 = \sqrt{\alpha_t}P_t \leq P_t$, we have $\|\mathbf{r}_t\|^2 \leq B_h + 8P_t$. Using $\|\mathbf{r}_t\|^2 = A_t - A_{t-1}$, we obtain

$$\frac{\|\mathbf{r}_t\|^2}{A_{t-1}} \;=\; \frac{\|\mathbf{r}_t\|^2}{A_t} + \|\mathbf{r}_t\|^2\left(\tfrac{1}{A_{t-1}} - \tfrac{1}{A_t}\right) \;\leq\; \ln A_t - \ln A_{t-1} + (B_h + 8P_t)\left(\tfrac{1}{A_{t-1}} - \tfrac{1}{A_t}\right).$$

On the other hand,

$$\Psi_{t+1} - \Psi_t = -(B_h + 8P_t)\left(\tfrac{1}{A_{t-1}} - \tfrac{1}{A_t}\right) + \tfrac{8(P_{t+1} - P_t)}{A_t}. \tag{80}$$

Adding these relations and using the tracking-error recursion yields

$$\tfrac{\|\mathbf{r}_t\|^2}{A_{t-1}} + \Psi_{t+1} - \Psi_t \leq \ln A_t - \ln A_{t-1} + 8\tfrac{\mathcal{V}_{t+1} + \mathcal{E}_{t+1}}{A_t}. \tag{81}$$

Summing over $t = 1, \ldots, T$, and using $\Psi_{T+1} \geq 0$, $\Psi_1 = B_h + 8P_1$, $P_1 = \mathcal{V}_1$, and $\mathcal{E}_1 = 0$, gives

$$H_T := \theta^2 \textstyle\sum_{t=1}^T \frac{\|\mathbf{r}_t\|^2}{A_{t-1}} \leq \theta^2\left[\ln A_T + B_h + 8\sum_{t=1}^{T+1} \frac{\mathcal{V}_t + \mathcal{E}_t}{A_{t-1}}\right]. \tag{82}$$

• *Bounding the cumulative innovations.* We first bound the nonnegative cumulative innovation $\overline{B}_{T+1}$. For $t \geq 2$, the innovation estimate gives

$$v_t \leq 11L^2\theta^2 \frac{\|\mathbf{r}_{t-1}\|^2}{A_{t-2}A_{t-1}} + \frac{2\dot{\mathcal{E}}_t}{A_{t-1}}.$$

Since $\mathcal{V}_1 = \dot{\mathcal{E}}_1$ and $A_{t-1} \geq 1$,

$$\begin{aligned} \overline{B}_{T+1} &\leq 11L^2\theta^2 \textstyle\sum_{t=1}^{T} \frac{\|\mathbf{r}_t\|^2}{A_{t-1}A_t} + 2Q_{T+1} \\ &= 11L^2\theta^2\big(1 - \tfrac{1}{A_T}\big) + 2Q_{T+1} \;\leq\; 11L^2\theta^2 + 2Q_{T+1}. \end{aligned}$$

The proof of part (b) established $\mathbb{E}[Q_{T+1}^2] \leq C_q^2 \ell_T^2$. Therefore, Cauchy–Schwarz and $\ell_T \geq 1$ yield

$$\mathbb{E}[\overline{B}_{T+1}^2] \leq \Big(11L^2\theta^2 + 2\sqrt{\mathbb{E}[Q_{T+1}^2]}\Big)^2 \leq (11L^2\theta^2 + 2C_q)^2 \ell_T^2. \tag{83}$$

• *A direct second-moment estimate.* We use partial sums as weights and summation by parts, as in the proof of Lemma C.5(e). For $t \geq 2$, the tracking-error recursion, $\varpi_t^2 \leq 1$, and $A_{t-1} \geq A_{t-2}$ imply

$$\frac{P_t}{A_{t-1}} \leq \frac{P_{t-1}}{A_{t-2}} + v_t + e_t.$$

Since $P_1 = \mathcal{V}_1$ and $e_1 = 0$, induction gives

$$0 \leq \frac{P_t}{A_{t-1}} \leq \overline{B}_t + \overline{Z}_t.$$

Moreover, A_{t-1} is $\mathcal{F}_{t-1}$-measurable, so $\mathbb{E}_t[e_t] = 0$. The pointwise estimate $\mathcal{E}_t^2 \leq 4P_{t-1}\mathcal{V}_t$ gives, for $t \geq 2$,

$$e_t^2 \leq 4\frac{P_{t-1}}{A_{t-1}} v_t \leq 4(\overline{B}_{t-1} + \overline{Z}_{t-1})v_t.$$

The last inequality also holds at $t = 1$ because $e_1 = 0$. Expanding the square and using conditional mean zero, we obtain

$$\mathbb{E}[\overline{Z}_{T+1}^2] = \mathbb{E}\textstyle\sum_{t=1}^{T+1} e_t^2 \leq 4\mathbb{E}\sum_{t=1}^{T+1}(\overline{B}_{t-1} + \overline{Z}_{t-1})v_t.$$

To control the right-hand side, summation by parts yields

$$\begin{aligned} 2\textstyle\sum_{t=1}^{T+1} \overline{B}_{t-1}v_t &= \overline{B}_{T+1}^2 - \sum_{t=1}^{T+1} v_t^2, \\ \textstyle\sum_{t=1}^{T+1} \overline{Z}_{t-1}v_t &= \overline{B}_{T+1}\overline{Z}_{T+1} - \sum_{t=1}^{T+1} \overline{B}_{t-1}e_t - \sum_{t=1}^{T+1} v_t e_t. \end{aligned}$$

Because $\mathbb{E}[\overline{B}_{t-1}e_t] = 0$, it follows that

$$\mathbb{E}[\overline{Z}_{T+1}^2] \;\leq\; 2\mathbb{E}[\overline{B}_{T+1}^2] + 4\mathbb{E}[\overline{B}_{T+1}\overline{Z}_{T+1}] - 4\mathbb{E}\textstyle\sum_{t=1}^{T+1} v_t e_t.$$

Cauchy–Schwarz bounds the two cross terms as follows:

$$\big|\mathbb{E}[\overline{B}_{T+1}\overline{Z}_{T+1}]\big| \;\leq\; \sqrt{\mathbb{E}[\overline{B}_{T+1}^2]\mathbb{E}[\overline{Z}_{T+1}^2]},$$

and

$$\Big|\mathbb{E}\textstyle\sum_{t=1}^{T+1} v_t e_t\Big| \;\leq\; \sqrt{\big(\mathbb{E}\sum_{t=1}^{T+1} v_t^2\big)\big(\mathbb{E}\sum_{t=1}^{T+1} e_t^2\big)} \;\leq\; \sqrt{\mathbb{E}[\overline{B}_{T+1}^2]\mathbb{E}[\overline{Z}_{T+1}^2]}.$$

Consequently, Young's inequality gives

$$\mathbb{E}[\overline{Z}_{T+1}^2] \;\leq\; 2\mathbb{E}[\overline{B}_{T+1}^2] + 8\sqrt{\mathbb{E}[\overline{B}_{T+1}^2]\mathbb{E}[\overline{Z}_{T+1}^2]} \;\leq\; 34\mathbb{E}[\overline{B}_{T+1}^2] + \tfrac{1}{2}\mathbb{E}[\overline{Z}_{T+1}^2].$$

Thus $\mathbb{E}[\overline{Z}_{T+1}^2] \leq 68\mathbb{E}[\overline{B}_{T+1}^2]$. Another application of Cauchy–Schwarz yields

$$\mathbb{E}[(\overline{B}_{T+1} + \overline{Z}_{T+1})^2] \leq (1 + \sqrt{68})^2\mathbb{E}[\overline{B}_{T+1}^2] \leq 100\mathbb{E}[\overline{B}_{T+1}^2]. \tag{84}$$

• *Establishing* $\mathbb{E}[H_T^2] \le C_h^2 \ell_T^2$ *and* $\mathbb{E}[H_T] \le C_h \ell_T$. We combine the preceding estimates to bound H_T. The preliminary estimate $\|\mathbf{r}_t\| \le C_r t$ gives

$$\ln A_T \le \ln(1 + C_r^2 T^3) \le C_A \ell_T.$$

By (82),

$$0 \;\le\; H_T \le \theta^2 \left[C_A \ell_T + B_h + 8(\overline{B}_{T+1} + \overline{Z}_{T+1}) \right].$$

Expanding the square and applying Cauchy–Schwarz, together with (83) and (84), we obtain

$$\begin{aligned} \sqrt{\mathbb{E}[H_T^2]} &\le \theta^2 \left[C_A \ell_T + B_h + 8\sqrt{\mathbb{E}[(\overline{B}_{T+1} + \overline{Z}_{T+1})^2]} \right] \\ &\le \theta^2 \left[C_A + B_h + 80(11L^2\theta^2 + 2C_q) \right] \ell_T \;=\; \ddot{C}_h \ell_T \;=\; C_h \ell_T. \end{aligned}$$

Squaring this inequality and using $\mathbb{E}[H_T] \le \sqrt{\mathbb{E}[H_T^2]}$, we conclude that

$$\mathbb{E}[H_T^2] \le C_h^2 \ell_T^2, \qquad \mathbb{E}[H_T] \le C_h \ell_T. \tag{85}$$

Part (d-i). Define $\mathcal{B}_t := \sum_{i=1}^t \mathcal{V}_i$ and $\mathcal{Z}_t := \sum_{i=1}^t \mathcal{E}_i$, where $\mathcal{B}_0 = \mathcal{Z}_0 = 0$.

The tracking recursion at iteration $t+1$ gives

$$\alpha_{t+1} P_t \le P_t - P_{t+1} + \mathcal{V}_{t+1} + \mathcal{E}_{t+1}.$$

Since $\alpha_t/\alpha_{t+1} \le 5/3 < 2$, summing over $t = 1, \dots, T$ and using $P_1 = \mathcal{V}_1$, $\mathcal{E}_1 = 0$, and $P_{T+1} \ge 0$, we obtain

$$\begin{aligned} U_T \;&\le\; 2\textstyle\sum_{t=1}^T \alpha_{t+1} P_t \;\le\; 2\left[P_1 - P_{T+1} + \sum_{t=2}^{T+1} (\mathcal{V}_t + \mathcal{E}_t) \right] \\ &=\; 2(\mathcal{B}_{T+1} + \mathcal{Z}_{T+1} - P_{T+1}) \le 2(\mathcal{B}_{T+1} + \mathcal{Z}_{T+1}). \end{aligned} \tag{86}$$

Furthermore,

$$\mathcal{B}_{T+1} \le \textstyle\sum_{t=1}^{T+1} (\ddot{\mathcal{E}}_t + 2\dot{\mathcal{E}}_t) = 11L^2 H_T + 2Q_{T+1}. \tag{87}$$

Conditional mean zero gives $\mathbb{E}[\mathcal{Z}_{T+1}] = 0$. Taking expectations in (86) and applying part (c) and the bound $\mathbb{E}[Q_{T+1}] \le C_v \ell_T$ proved in part (b-i), we obtain the more precise first-moment estimate

$$\mathbb{E}[U_T] \le 2\mathbb{E}[\mathcal{B}_{T+1}] \le 2(11L^2 C_h + 2C_v)\ell_T. \tag{88}$$

Part (d-ii). We again use partial sums and summation by parts, without assuming a uniform bound on P_t. The exact recursion gives the identity

$$\mathcal{B}_t + \mathcal{Z}_t = P_t + \textstyle\sum_{i=1}^t (1 - \varpi_i^2) P_{i-1} \ge P_t \ge 0. \tag{89}$$

By conditional mean zero,

$$\mathbb{E}[\mathcal{Z}_{T+1}^2] \;=\; \mathbb{E}\textstyle\sum_{t=1}^{T+1} \mathcal{E}_t^2 \le 4\mathbb{E}\sum_{t=1}^{T+1} P_{t-1}\mathcal{V}_t \;\le\; 4\mathbb{E}\sum_{t=1}^{T+1} (\mathcal{B}_{t-1} + \mathcal{Z}_{t-1})\mathcal{V}_t. \tag{90}$$

The partial-sum identities give

$$\begin{aligned} 2\textstyle\sum_{t=1}^{T+1} \mathcal{B}_{t-1}\mathcal{V}_t \;&=\; \mathcal{B}_{T+1}^2 - \sum_{t=1}^{T+1} \mathcal{V}_t^2, \\ \textstyle\sum_{t=1}^{T+1} \mathcal{Z}_{t-1}\mathcal{V}_t \;&=\; \mathcal{B}_{T+1}\mathcal{Z}_{T+1} - \sum_{t=1}^{T+1} \mathcal{B}_{t-1}\mathcal{E}_t - \sum_{t=1}^{T+1} \mathcal{V}_t\mathcal{E}_t. \end{aligned}$$

Here $\mathbb{E}[\mathcal{B}_{t-1}\mathcal{E}_t] = 0$ because $\mathcal{B}_{t-1}$ is predictable. Combining these identities with (90) and Cauchy–Schwarz gives

$$\begin{aligned} \mathbb{E}[\mathcal{Z}_{T+1}^2] &\le 2\mathbb{E}[\mathcal{B}_{T+1}^2] + 4\mathbb{E}[\mathcal{B}_{T+1}\mathcal{Z}_{T+1}] - 4\mathbb{E}\textstyle\sum_{t=1}^{T+1} \mathcal{V}_t\mathcal{E}_t \\ &\le 2\mathbb{E}[\mathcal{B}_{T+1}^2] + 8\sqrt{\mathbb{E}[\mathcal{B}_{T+1}^2]\mathbb{E}[\mathcal{Z}_{T+1}^2]} \\ &\le 34\mathbb{E}[\mathcal{B}_{T+1}^2] + \tfrac{1}{2}\mathbb{E}[\mathcal{Z}_{T+1}^2]. \end{aligned}$$

For the second inequality, the same-time product is bounded by

$$\left| \mathbb{E}\textstyle\sum_{t=1}^{T+1} \mathcal{V}_t\mathcal{E}_t \right| \le \sqrt{\left(\mathbb{E}\sum_{t=1}^{T+1} \mathcal{V}_t^2\right)\left(\mathbb{E}\sum_{t=1}^{T+1} \mathcal{E}_t^2\right)} \le \sqrt{\mathbb{E}[\mathcal{B}_{T+1}^2]\mathbb{E}[\mathcal{Z}_{T+1}^2]}.$$

Consequently,

$$\mathbb{E}[\mathcal{Z}_{T+1}^2] \le 68\mathbb{E}[\mathcal{B}_{T+1}^2], \qquad \sqrt{\mathbb{E}[(\mathcal{B}_{T+1}+\mathcal{Z}_{T+1})^2]} \le 10\sqrt{\mathbb{E}[\mathcal{B}_{T+1}^2]}. \tag{91}$$

Finally, (87), part (c), and the bound $\mathbb{E}[Q_{T+1}^2] \le C_q^2\ell_T^2$ in (78) imply

$$\sqrt{\mathbb{E}[\mathcal{B}_{T+1}^2]} \le 11L^2\sqrt{\mathbb{E}[H_T^2]} + 2\sqrt{\mathbb{E}[Q_{T+1}^2]} \le (11L^2C_h + 2C_q)\ell_T.$$

Since the right-hand side of (86) is nonnegative, squaring that inequality and invoking (91) yields

$$\sqrt{\mathbb{E}[U_T^2]} \le 2\sqrt{\mathbb{E}[(\mathcal{B}_{T+1}+\mathcal{Z}_{T+1})^2]} \le 20(11L^2C_h + 2C_q)\ell_T = C_u\ell_T.$$

Together with (88), or simply $\mathbb{E}[U_T] \le \sqrt{\mathbb{E}[U_T^2]}$, this proves both claims in part (d).

□

With the first two moments of the tracking and normalized motion sums available, we can bound the descent energy. The following pathwise comparison then transfers this control to the residual accumulator with the weight appropriate to MVR2.

Lemma C.12. *(Descent energy and motion bridge for MVR2) Let $\Delta := F(\mathbf{x}_1) - F_\star$ and $\ell_T := \ln(e+T)$. Let C_h and C_u be defined in Lemma C.11. Then, for every integer $T \ge 1$,*

(a) *Define $W_T := \sum_{t=1}^T \gamma_t\|\mathbf{r}_t\|^2$. We have $\mathbb{E}[W_T] \le C_w\ell_T$ and $\mathbb{E}[W_T^2] \le C_w^2\ell_T^2$, where $C_w := 2\Delta + C_u + (1+L)C_h$.*

(b) *$\mathbb{E}[\sqrt{\alpha_T}M_T] \le C_m\ell_T^2$, where $C_m := C_w/\theta + C_w^2/(4\theta^2)$.*

Proof. Define $W_T := \sum_{t=1}^T \gamma_t\|\mathbf{r}_t\|^2$. Let $L_+ = L + 1$.

Part (a-i). Summing the first inequality of Lemma C.10 over $t = 1, \dots, T$ gives $\frac{1}{2}W_T \le F(\mathbf{x}_1) - F(\mathbf{x}_{T+1}) + \frac{1}{2}U_T + \frac{1+L}{2}H_T$. Since $F(\mathbf{x}_{T+1}) \ge F_\star$, we obtain

$$0 \le W_T \le 2\Delta + U_T + L_+H_T. \tag{92}$$

Taking expectations and applying Lemma C.11(c),(d), we have

$$\mathbb{E}[W_T] \;\le\; 2\Delta + C_u\ell_T + L_+C_h\ell_T \;\le\; \big(2\Delta + C_u + L_+C_h\big)\ell_T \;=\; C_w\ell_T.$$

Part (a-ii). Squaring (92) gives

$$\begin{aligned}
\mathbb{E}[W_T^2] &\le 4\Delta^2 + \mathbb{E}[U_T^2] + L_+^2\mathbb{E}[H_T^2] + 4\Delta\mathbb{E}[U_T] + 4\Delta L_+\mathbb{E}[H_T] + 2L_+\mathbb{E}[U_TH_T] \\
&\overset{①}{\le} 4\Delta^2 + C_u^2\ell_T^2 + L_+^2C_h^2\ell_T^2 + 4\Delta C_u\ell_T + 4\Delta L_+C_h\ell_T + 2L_+\mathbb{E}[U_TH_T] \\
&\overset{②}{\le} 4\Delta^2 + C_u^2\ell_T^2 + L_+^2C_h^2\ell_T^2 + 4\Delta C_u\ell_T + 4\Delta L_+C_h\ell_T + 2L_+C_uC_h\ell_T^2 \\
&= (2\Delta + C_u\ell_T + L_+C_h\ell_T)^2 \le C_w^2\ell_T^2,
\end{aligned}$$

where step ① uses Lemma C.11; step ② uses Cauchy–Schwarz that $\mathbb{E}[U_TH_T] \le \sqrt{\mathbb{E}[U_T^2]}\sqrt{\mathbb{E}[H_T^2]} \le C_uC_h\ell_T^2$, where the last step uses Lemma C.11.

Part (b). We obtain

$$\begin{aligned}
W_T &:= \textstyle\sum_{t=1}^T \gamma_t\|\mathbf{r}_t\|^2 \overset{①}{=} \theta\sum_{t=1}^T \alpha_t^{1/4}\frac{A_t - A_{t-1}}{\sqrt{A_{t-1}}} \ge \theta\alpha_T^{1/4}\sum_{t=1}^T \frac{A_t - A_{t-1}}{\sqrt{A_{t-1}}} \\
&\overset{②}{\ge} 2\theta\alpha_T^{1/4}\textstyle\sum_{t=1}^T(\sqrt{A_t} - \sqrt{A_{t-1}}) = 2\theta\alpha_T^{1/4}(\sqrt{1+M_T} - 1),
\end{aligned} \tag{93}$$

where step ① uses $\gamma_t = \theta\alpha_t^{1/4}/\sqrt{A_{t-1}}$ and $\|\mathbf{r}_t\|^2 = A_t - A_{t-1}$; step ② uses the fact that $\frac{A_t - A_{t-1}}{\sqrt{A_{t-1}}} = (\sqrt{A_t} - \sqrt{A_{t-1}})\left(1 + \sqrt{\frac{A_t}{A_{t-1}}}\right) \ge 2(\sqrt{A_t} - \sqrt{A_{t-1}})$.

We have from Inequality (93): $W_T + 2\theta\alpha_T^{1/4} \geq 2\theta\alpha_T^{1/4}\sqrt{1+M_T}$. Squaring both nonnegative sides and dividing both sides by $4\theta^2$ gives $\frac{1}{4\theta^2}W_T^2 + \alpha_T^{1/2} + \frac{1}{\theta}W_T\alpha_T^{1/4} \geq \alpha_T^{1/2}(1+M_T)$, leading to:

$$\sqrt{\alpha_T}M_T \;\leq\; \tfrac{\alpha_T^{1/4}}{\theta}W_T + \tfrac{W_T^2}{4\theta^2} \;\leq\; \tfrac{W_T}{\theta} + \tfrac{W_T^2}{4\theta^2}.$$

Taking expectations and applying Part (a), we conclude that

$$\mathbb{E}[\sqrt{\alpha_T}M_T] \;\leq\; \tfrac{C_w}{\theta}\ell_T + \tfrac{C_w^2}{4\theta^2}\ell_T^2 \;\leq\; \left(\tfrac{C_w}{\theta} + \tfrac{C_w^2}{4\theta^2}\right)\ell_T^2 \;=\; C_m\ell_T^2.$$

□

We now assemble the MVR2 stationarity guarantee. Unlike MVR1, its discrepancy compares two samples at the same point, so the inverse-weight estimate depends directly on the noise variance; combining this estimate with the motion and tracking bounds yields the stated rate.

Theorem C.13. *(Noise-adaptive composite stationarity of* ALFPG-MVR2*) Suppose Assumptions 3.1, 5.1, 5.2, 5.3, 5.4, and 5.7 hold, and run the MVR2 branch of Algorithm 2 with any fixed $\theta > 0$. Let $R \sim \mathrm{Unif}\{1,\dots,T\}$ be independent of all algorithmic randomness, and define $\mathcal{R}_t := \mathrm{dist}\big(0, \nabla f(\mathbf{x}_t) + \partial h(\mathbf{x}_t)\big)$. Let C_u and C_m be the constants in Lemmas C.11 and C.12, and set $C_R := 2(1+L\theta)\sqrt{C_m} + 2\sqrt{C_u}$. Then, for every integer $T \geq 1$,*

$$\mathbb{E}[\mathcal{R}_{R+1}] \leq C_R\ell_T\big(\frac{1}{T^{1/2}} + \frac{\sigma^{1/3}}{T^{1/3}}\big). \tag{94}$$

Thus ALFPG-MVR2 *attains $\widetilde{\mathcal{O}}(T^{-1/2} + \sigma^{1/3}T^{-1/3})$; when $\sigma = 0$, this reduces to $\widetilde{\mathcal{O}}(T^{-1/2})$. Hence, the stochastic-gradient oracle complexity for obtaining $\mathbb{E}[\mathcal{R}_{R+1}] \leq \epsilon$ is $\widetilde{\mathcal{O}}(\epsilon^{-2} + \sigma\epsilon^{-3})$.*

Proof. Define $\ell_T := \ln(e+T)$, and $\mathcal{R}_t := \mathrm{dist}\big(0, \nabla f(\mathbf{x}_t) + \partial h(\mathbf{x}_t)\big)$.

Define $D_t = \sum_{i=1}^t \chi_i^2$ and $N_t = \max_{1\leq i\leq t}\chi_i^2$, where $\chi_1 = 0$.

Define the deterministic quantity $\mathcal{B}_T := 1 + 2(T-1)\sigma^2$.

Bounding the term $\mathbb{E}[\alpha_T^{-1/2}]$. For $T \geq 2$, the following running-maximum expression is nonempty. Using the initial discrepancy $\chi_1 = 0$, we derive:

$$\alpha_T^{-1/2} \;=\; \big(\max_{1\leq k\leq T-1}\tfrac{1+D_k}{1+N_k}\big)^{1/3} \;\leq\; \big(1+D_{T-1}\big)^{1/3} \;\leq\; \big(1+\textstyle\sum_{i=1}^{T-1}\chi_i^2\big)^{1/3}.$$

Taking the expectations yields:

$$\mathbb{E}[\alpha_T^{-1/2}] \;\leq\; \mathbb{E}[\big(1+\textstyle\sum_{i=1}^{T-1}\chi_i^2\big)^{1/3}] \;\overset{①}{\leq}\; \big(1+2(T-1)\sigma^2\big)^{1/3} \;=\; \mathcal{B}_T^{1/3}, \tag{95}$$

where step ① uses $\mathbb{E}[\chi_t^2] = \mathbb{E}\|\boldsymbol{\epsilon}_{t-1}\|^2 + \mathbb{E}\|\boldsymbol{\epsilon}_t^-\|^2 \leq 2\sigma^2$. For $T = 1$, Inequality (95) holds directly.

Bounding the composite residual. We derive:

$$\begin{aligned}
\mathcal{R}_{t+1} \;&:=\; \mathrm{dist}\big(0, \nabla f(\mathbf{x}_{t+1}) + \partial h(\mathbf{x}_{t+1})\big) \\
&\overset{①}{\leq}\; \|\nabla f(\mathbf{x}_{t+1}) + \mathbf{r}_t - \mathbf{m}_t\| \;\overset{②}{\leq}\; \|\mathbf{r}_t\| + \|\mathbf{s}_t\| + \|\nabla f(\mathbf{x}_{t+1}) - \nabla f(\mathbf{x}_t)\| \\
&\overset{③}{\leq}\; (1+L\gamma_t)\|\mathbf{r}_t\| + \|\mathbf{s}_t\| \;\overset{④}{\leq}\; (1+L\theta)\|\mathbf{r}_t\| + \|\mathbf{s}_t\|,
\end{aligned} \tag{96}$$

where step ① uses the global optimality of the proximal update that $\mathbf{r}_t - \mathbf{m}_t \in \partial h(\mathbf{x}_{t+1})$; step ② uses $\mathbf{s}_t = \mathbf{m}_t - \nabla f(\mathbf{x}_t)$; step ③ uses L-smoothness of $f(\cdot)$; step ④ uses $\gamma_t = \theta\alpha_t^{1/4}\beta_t \leq \theta$.

Removing the adaptive weights. We obtain

$$\mathbb{E}[\sqrt{M_T}] \;=\; \mathbb{E}\big[\sqrt{\sqrt{\alpha_T}M_T}\,\alpha_T^{-1/4}\big] \;\overset{①}{\leq}\; \sqrt{\mathbb{E}[\sqrt{\alpha_T}M_T]}\sqrt{\mathbb{E}[\alpha_T^{-1/2}]} \;\leq\; \sqrt{C_m}\ell_T\mathcal{B}_T^{1/6}, \quad (97)$$

where step ① uses Cauchy–Schwarz inequality; step ② uses Lemma C.12(b), and Inequality (95).

Another application of Cauchy–Schwarz and Lemma C.11(d) gives:

$$\mathbb{E}[\sqrt{S_T}] \;\overset{①}{\leq}\; \mathbb{E}[\sqrt{U_T}\,\alpha_T^{-1/4}] \;\overset{②}{\leq}\; \sqrt{\mathbb{E}[U_T]}\sqrt{\mathbb{E}[\alpha_T^{-1/2}]} \;\overset{③}{\leq}\; \sqrt{C_u\ell_T}\mathcal{B}_T^{1/6}, \quad (98)$$

where step ① uses $\sqrt{\alpha_T}S_T \leq \sum_{t=1}^T \sqrt{\alpha_t}\|\mathbf{s}_t\|^2 := U_T$ since $\alpha_t \geq \alpha_T$ for $t \leq T$; step ② uses Cauchy–Schwarz inequality; step ③ uses Inequality (95).

Establishing the noise-adaptive rate. Let $C_R := 2(1+L\theta)\sqrt{C_m} + 2\sqrt{C_u}$. Since R is uniform and independent of the algorithm, we derive:

$$\begin{aligned}
\mathbb{E}[\mathcal{R}_{R+1}] \;&=\; \tfrac{1}{T}\textstyle\sum_{t=1}^T \mathbb{E}[\mathcal{R}_{t+1}] \;\overset{①}{\leq}\; \tfrac{1+L\theta}{T}\mathbb{E}\textstyle\sum_{t=1}^T \|\mathbf{r}_t\| + \tfrac{1}{T}\mathbb{E}\textstyle\sum_{t=1}^T \|\mathbf{s}_t\| \\
&\leq\; \tfrac{1+L\theta}{\sqrt{T}}\mathbb{E}[\sqrt{M_T}] + \tfrac{1}{\sqrt{T}}\mathbb{E}[\sqrt{S_T}] \;\overset{②}{\leq}\; \tfrac{(1+L\theta)\sqrt{C_m}\ell_T + \sqrt{C_u}\ell_T^{1/2}}{\sqrt{T}}\mathcal{B}_T^{1/6} \\
&\overset{③}{\leq}\; \tfrac{C_R\ell_T}{2\sqrt{T}}(1 + 2T^{1/6}\sigma^{1/3}) \;\leq\; C_R\tfrac{\ell_T}{\sqrt{T}} + \tfrac{C_R\ell_T}{T^{1/3}}\sigma^{1/3},
\end{aligned}$$

where step ① uses Inequality (96); step ② uses Inequalities (97) and (98); step ③ uses $\mathcal{B}_T^{1/6} \leq 1 + \big(2(T-1)\sigma^2\big)^{1/6} \leq 1 + 2^{1/6}T^{1/6}\sigma^{1/3} \leq 1 + 2T^{1/6}\sigma^{1/3}$. □