COMPLEXITY AND LAST-ITERATE CONVERGENCE OF INEXACT RIEMANNIAN PROXIMAL MOMENTUM SHUFFLING WITH VARIANCE REDUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

We study nonconvex finite-sum composite optimization over compact embedded manifolds with a convex nonsmooth regularizer. We establish pathwise complexity and last-iterate convergence for methods combining full-pass variance reduction, projected momentum, and inexact tangent-space proximal steps. The dependence induced by sampling without replacement prevents a direct application of standard fresh-sampling error recursions. We introduce an epochwise estimator error–memory condition and a projected-momentum lifting principle, and verify them for projection-based SVRG, projection-based SARAH/SPIDER, projected ambient SAGA, and projected ambient SAG. Combined with a prescribed Fenchel-dual residual criterion for the tangent proximal subproblems, this yields a deterministic epochwise descent estimate that controls estimator error, momentum memory, shuffling, and inner inexactness simultaneously. For fixed finite-sum size, we obtain $O(\varepsilon^{-2})$ epoch and retraction complexity and total proximal-evaluation complexities of $O(\varepsilon^{-4})$, $O(\varepsilon^{-3}\log\varepsilon^{-1})$, and $O(\varepsilon^{-3})$ for FISTA, NFG, and accumulative-regularization inner solvers, respectively. Unlike stochastic analyses that guarantee stationarity in expectation, these bounds hold along every admissible shuffling trajectory. We further establish epochwise nonmonotone descent and relative-error estimates for the natural tangent-bundle lift of the objective. Under its ordinary Kurdyka–Łojasiewicz property, the full iterate trajectory has finite length and converges to a single critical point, without requiring a common deterministic limiting objective value across trajectories. The finite-length estimate further yields a pathwise $O(\varepsilon^{-1})$ best-iterate outer complexity. Under an additional KL exponent assumption, we obtain explicit last-iterate rates.

1 INTRODUCTION

Many problems in machine learning, statistics, signal processing, and scientific computing require optimizing a data-fitting objective over a structured manifold in the presence of a nonsmooth regularizer. Representative examples include sparse principal component analysis and sparse canonical correlation analysis (Chen et al., 2020b; d'Aspremont et al., 2007; Zou et al., 2006), sparse subspace clustering (Elhamifar & Vidal, 2013), and learning problems with normalization or orthogonality constraints (Chen et al., 2020a; Gao et al., 2018; Wen & Yin, 2013). When the smooth data-fitting term is formed from a collection of sample losses, these problems lead to the finite-sum composite model

$$\min_{x\in\mathcal{M}} \Phi(x) := F(x) + R(x), \qquad F(x) := \frac{1}{n}\sum_{j=1}^{n} f_j(x), \tag{1}$$

where $\mathcal{M} \subseteq \mathbb{R}^d$ is an embedded Riemannian submanifold, each f_j is smooth on a neighborhood of $\mathcal{M}$, and $R : \mathbb{R}^d \to \mathbb{R}$ is proper, closed, and convex.

A standard way to handle the nonsmooth term is the Riemannian proximal-gradient or ManPG paradigm (Chen et al., 2020a; 2024; Huang & Wei, 2022). At the current point, it constructs a proximal model in the tangent space, exactly minimizes this model, and maps the resulting tangent step back to the manifold through a retraction. This approach can attain the favorable $O(\varepsilon^{-2})$ retraction complexity for finding an ε-critical point. However, its application to finite-sum problems

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.

raises two additional computational issues. First, repeatedly evaluating the full gradient may be expensive, motivating stochastic or variance-reduced estimators. Second, the tangent-space proximal subproblem generally has no closed-form solution, so its computational cost cannot be ignored when it is solved only approximately.

Existing studies address these difficulties along three related directions.

Stochastic and variance-reduced Riemannian methods. For smooth finite-sum optimization on manifolds, numerous stochastic and variance-reduced methods have been developed, including Riemannian SVRG (Zhang et al., 2016; Sato et al., 2019), recursive and SPIDER-type estimators (Kasai et al., 2018; Zhang et al., 2018), SAGA-type methods (Babanezhad et al., 2019), and recursive-momentum methods (Han & Gao, 2021). These methods provide effective mechanisms for controlling stochastic-gradient errors, but their analyses are primarily designed for smooth objectives.

For nonsmooth Riemannian optimization, stochastic smoothing, augmented Lagrangian, splitting, and ADMM-type methods have also been proposed (Peng et al., 2023; Deng et al., 2025a; Jin et al., 2025; Deng et al., 2025b). These approaches avoid solving the tangent-space proximal subproblems that arise in ManPG-type methods, but their retraction-complexity orders are typically worse than $O(\varepsilon^{-2})$. This difference can be important because a retraction may require a QR factorization, a polar decomposition, or another nontrivial matrix operation. Closer to the present setting, Wang et al. (2022) incorporated stochastic and SpiderBoost-type gradient estimators into the ManPG framework and retained an $O(\varepsilon^{-2})$ retraction order for problems on the Stiefel manifold. This result demonstrates that variance reduction and tangent proximal steps can be combined without losing the favorable outer complexity. Its analysis, however, is tied to a particular estimator and manifold structure, assumes that every tangent-space proximal subproblem is solved exactly, and does not quantify the total work required by an implementable inner solver. It also establishes stationarity complexity rather than whole-sequence or last-iterate convergence.

Inexact Riemannian proximal methods. Inexact tangent-space proximal steps have been studied in several deterministic frameworks (He et al., 2026; Huang & Wei, 2023; Jiang et al., 2025; Xu et al., 2025; Zheng et al., 2025). In particular, He et al. (2026), Jiang et al. (2025), and Zheng et al. (2025) develop implementable inexactness criteria and derive outer and inner complexity guarantees for ManPG-type methods. These results show that the $O(\varepsilon^{-2})$ retraction order can be preserved when the tangent proximal subproblems are solved approximately. Nevertheless, these analyses use full-gradient information and therefore do not address the estimator errors and memory recursions arising in finite-sum variance-reduced methods. Moreover, some existing stopping conditions determine the required inner accuracy through quantities generated during the current subproblem solve. This makes it difficult to prescribe the tolerance before calling the inner solver and to combine the resulting inexactness error with a stochastic or variance-reduced estimator recursion.

More recently, Zhang (2026) combined variance reduction, projected momentum, and prescribed inexact tangent solves through a single-step conditional error-dissipation condition. The resulting iRPMVR framework retains the $O(\varepsilon^{-2})$ retraction order for expected squared model-criticality and quantifies the total inner work. Its estimator verification, however, relies on fresh stochastic sampling. It does not establish the corresponding pathwise complexity for full-pass shuffling, where dependence within a permutation prevents direct use of the same conditional variance recursions.

KL and last-iterate analyses. The Kurdyka–Łojasiewicz inequality has been widely used to strengthen subsequential convergence to whole-sequence convergence and to derive last-iterate rates. For Euclidean finite-sum optimization, KL analyses of random reshuffling and shuffling-momentum methods were developed in (Li et al., 2023; Liang & Xu, 2025); a nonsmooth normal-map-based reshuffling method was studied in (Qiu et al., 2025). These analyses already handle nonmonotone dynamics caused by shuffling, but they do not incorporate changing tangent spaces, projected momentum, variance-reduction memory, and inexact tangent proximal solves simultaneously.

For stochastic inexact proximal methods, Zhang (2026) obtains almost-sure finite length and whole-sequence convergence under the ordinary KL property of a natural shifted objective, but additionally assumes that its values converge almost surely to a deterministic constant. The explicit KL rates are established for the deterministic specialization. Thus, that stochastic result does not directly establish last-iterate convergence or pathwise rates for full-pass shuffling when different trajectories may converge to different critical values.

Table 1: Representative nonsmooth Riemannian complexity bounds for an ε-critical point. We treat n as a fixed problem constant and report dependence on the target accuracy only.

Method	Stoch.	#Retr	$\#\,\mathrm{prox}_R$	KL conv.
RALM (Deng et al., 2025a; Xu et al., 2025)	No	$O(\varepsilon^{-3})$	$O(\varepsilon^{-3})$	No
RADMM (Li et al., 2025)	No	$O(\varepsilon^{-4})$	$O(\varepsilon^{-4})$	No
RSG (det.) (Beck & Rosset, 2023; Peng et al., 2023)	No	$O(\varepsilon^{-3})$	$O(\varepsilon^{-3})$	No
RADA (Xu et al., 2026)	No	$O(\varepsilon^{-3})$	$O(\varepsilon^{-3})$	No
OADMM (Yuan, 2024)	No	$O(\varepsilon^{-3})$	$O(\varepsilon^{-3})$	Yes
RSG (sto.) (Peng et al., 2023)	Yes	$O(\varepsilon^{-5})$	$O(\varepsilon^{-5})$	No
StoManIAL (Deng et al., 2025a)	Yes	$\widetilde{O}(\varepsilon^{-3.5})$	$\widetilde{O}(\varepsilon^{-3.5})$	No
MARS-ADMM (Jin et al., 2025)	Yes	$\widetilde{O}(\varepsilon^{-3})$	$\widetilde{O}(\varepsilon^{-3})$	No
Smoothing (Deng et al., 2025b)	Yes	$O(\varepsilon^{-3})$	$O(\varepsilon^{-3})$	No
iRPDC-BB (Jiang et al., 2025)	No	$O(\varepsilon^{-2})$	$O(\varepsilon^{-4})$	No
iRPDC-NFG (Jiang et al., 2025)	No	$O(\varepsilon^{-2})$	$O(\varepsilon^{-3}\log\varepsilon^{-1})$	No
iRPDC-AR (Jiang et al., 2025)	No	$O(\varepsilon^{-2})$	$O(\varepsilon^{-3})$	No
IVManPL (He et al., 2026)	No	$O(\varepsilon^{-2})$	$O(\varepsilon^{-3})$	Yes
IManPL (Zheng et al., 2025)	No	$O(\varepsilon^{-2})$	$O(\varepsilon^{-3})$	No
iRPMVR-FISTA (Zhang, 2026)	Yes	$O(\varepsilon^{-2})$	$O(\varepsilon^{-4})$	a.s.[†]
iRPMVR-NFG (Zhang, 2026)	Yes	$O(\varepsilon^{-2})$	$O(\varepsilon^{-3}\log\varepsilon^{-1})$	a.s.[†]
iRPMVR-AR (Zhang, 2026)	Yes	$O(\varepsilon^{-2})$	$O(\varepsilon^{-3})$	a.s.[†]
iRPMS-FISTA (ours)	Yes	$O(\varepsilon^{-2})$ (KL) $O(\varepsilon^{-1})$	$O(\varepsilon^{-4})$ (KL) $O(\varepsilon^{-(r+1)})$	Yes
iRPMS-NFG (ours)	Yes	$O(\varepsilon^{-2})$ (KL) $O(\varepsilon^{-1})$	$O(\varepsilon^{-3}\log\varepsilon^{-1})$ (KL) $O(\varepsilon^{-(1+r/2)}\log\varepsilon^{-1})$	Yes
iRPMS-AR (ours)	Yes	$O(\varepsilon^{-2})$ (KL) $O(\varepsilon^{-1})$	$O(\varepsilon^{-3})$ (KL) $O(\varepsilon^{-(1+r/2)})$	Yes

[†]iRPMVR complexity targets expected squared criticality; its KL convergence is almost sure and assumes a deterministic limiting shifted-objective value. iRPMS bounds are pathwise; the KL bounds may have trajectory-dependent constants, and $r > 2$.

Computational gap and objective of this paper. The preceding results reveal a concrete tradeoff. Stochastic smoothing and splitting-type methods avoid tangent proximal subproblems but generally have worse retraction complexity. Exact variance-reduced ManPG-type methods retain the $O(\varepsilon^{-2})$ retraction order but leave the tangent proximal subproblems unaccounted for. Deterministic inexact ManPG methods quantify the inner work but do not exploit the finite-sum structure. The stochastic framework discussed above addresses both error sources in conditional expectation. What remains is a pathwise analysis of their joint effects under full-pass shuffling, including last-iterate convergence without a common deterministic limiting objective value.

Motivated by this gap, we ask whether one can develop a common theoretical analysis that simultaneously achieves the following goals:

(i) retain the $O(\varepsilon^{-2})$ retraction order of ManPG-type methods along each admissible shuffling trajectory;

(ii) accommodate a class of variance-reduced estimators rather than a single estimator-specific recursion;

(iii) permit full-pass shuffling and implementable inexact tangent proximal solves while quantifying the total inner work; and

(iv) establish whole-sequence and last-iterate convergence under an ordinary KL assumption, without a common deterministic limiting objective value.

This paper answers these questions through an estimator-agnostic convergence and complexity analysis. The common algorithmic realization performs one shuffled pass through the component functions in every epoch, combines a variance-reduced estimator with projected momentum, solves the tangent proximal model up to a prescribed Fenchel-dual residual, and retracts the resulting tangent step to the manifold. The objective is not to introduce a new variance-reduction recursion. Instead, we identify abstract estimator and memory conditions under which the same descent, complexity, and KL arguments apply, and then verify these conditions for four existing variance-reduction mechanisms.

Table 1 provides a structural comparison, as the problem classes and stationarity measures differ. For fixed n, iRPMS retains the $O(\varepsilon^{-2})$ epoch and retraction order while accounting for inexact tangent solves. Under the KL assumption, the best-iterate outer bound becomes $O(\varepsilon^{-1})$ along each trajectory.

The corresponding proximal-evaluation exponents improve on the non-KL bounds for $2 < r < 3$ with FISTA and $2 < r < 4$ with NFG or AR.

Main contributions. Our contributions are summarized as follows.

First, a pathwise epochwise theory for full-pass variance reduction. We develop an epochwise error–memory condition for Riemannian variance-reduced estimators coupled with projected momentum. The condition is formulated directly along a complete shuffled pass and does not rely on conditional unbiasedness or one-step variance recursions. This distinction is essential for full-pass shuffling, because conditioning on the within-epoch history restricts the next component to the remaining indices rather than producing a fresh sample from the full index set. Hence the single-step conditional error-dissipation recursion of Zhang (2026) does not directly yield the pathwise epochwise estimator bound needed here. A projected-momentum lifting argument reduces the verification to estimator-specific pathwise bounds, and we establish these bounds for projection-based SVRG, projection-based SARAH/SPIDER, projected ambient SAGA, and projected ambient SAG. Hence the four full-pass realizations, despite their different memory structures, are covered by the same epochwise analysis.

Second, pathwise outer and overall computational complexity with inexact tangent solves. Building on the prescribed Fenchel-dual residual criterion for the tangent proximal subproblem, we derive an epochwise descent relation that couples shuffling error, variance-reduction memory, projected momentum, and inner inexactness in a single deterministic estimate. For fixed finite-sum size n, this yields an $O(\varepsilon^{-2})$ epoch and retraction complexity along every admissible shuffling trajectory, together with total proximal-evaluation complexities $O(\varepsilon^{-4})$, $O(\varepsilon^{-3}\log\varepsilon^{-1})$, and $O(\varepsilon^{-3})$ for FISTA, NFG, and accumulative regularization, respectively. Although these accuracy orders match those in Zhang (2026), the guarantees are qualitatively different: there the stationarity bound is in expected squared criticality, whereas here the stated computational budgets produce an ε-critical point pathwise, without taking expectations over the shuffling order.

Third, deterministic whole-sequence convergence and KL rates for the complete shuffled trajectory. For the natural tangent-bundle lift $\Phi_0(x,\eta) := \Phi(x+\eta) + \iota_{T\mathcal{M}}(x,\eta)$, we establish an epochwise nonmonotone descent inequality together with a relative-error bound that incorporates delayed estimator memory and summable inner errors. Under the ordinary KL property of Φ_0, these estimates yield finite length and convergence of every within-epoch iterate sequence to the same critical point. Moreover, the finite-length estimate gives a pathwise $O(\varepsilon^{-1})$ best-iterate epoch and retraction complexity, and, under a KL exponent, explicit last-iterate rates for the lifted objective gap, the criticality-related quantities, and every inner iterate. These conclusions differ essentially from the stochastic KL analysis of Zhang (2026). There, whole-sequence convergence is almost sure and additionally requires the shifted-objective values to converge almost surely to a common deterministic constant, while the explicit algorithmic KL rates are established only for the deterministic specialization. Here, the convergence and rate results hold deterministically for each fixed shuffling order, and pathwise for random permutations, without requiring a common limiting objective value across trajectories.

This paper is organized as follows. Section 2 collects the geometric facts used in the analysis. Section 3 describes the full-pass shuffling realization and the common estimator condition. Section 4 presents the descent, complexity, KL convergence, and variance-reduction verification results. The appendices supply the necessary background, parameter choices, proofs of the full-pass results, and estimator-specific calculations.

2 Setup and basic facts

We impose the following assumptions on the geometry of $\mathcal{M}$ and the regularity of the objective functions.

Assumption 2.1 (Geometry and regularity)**.** The following conditions hold.

(i) $\mathcal{M} \subseteq \mathbb{R}^d$ is a compact embedded Riemannian submanifold of dimension $d - p$.

(ii) Each $f_j : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable, and its ambient gradient is Lipschitz continuous on a neighborhood of $\mathcal{M}$ with a common constant ℓ_f.

(iii) $R : \mathbb{R}^d \to \mathbb{R}$ is proper, closed, convex, and globally Lipschitz with constant ℓ_R.

(iv) $\mathrm{Retr} : T\mathcal{M} \to \mathcal{M}$ is a fixed smooth retraction, and $\mathrm{P}_{T_x\mathcal{M}}$ denotes the ambient orthogonal projection onto $T_x\mathcal{M}$.

A point $x \in \mathcal{M}$ is critical when

$$0 \in \operatorname{grad} F(x) + \partial R(x) + N_x\mathcal{M}. \tag{2}$$

The notation, geometric preliminaries, and KL property are recalled in Appendix A. Uniform bounds for the tangent constraint, retraction estimates, and normal lifting on the tangent bundle are given in Lemmas A.1, A.2, and A.3, respectively.

Full-pass shuffling compares component gradients in changing tangent spaces. The next lemma controls both a single projected difference and its accumulated pathwise counterpart.

Lemma 2.2 (Local and pathwise projection estimates)**.** *There are $\delta_T, L_T > 0$ such that, when $\|\eta\| \le \delta_T$,*

$$\|\operatorname{grad} f_j(\mathrm{Retr}_x(\eta)) - \mathrm{P}_{T_{\mathrm{Retr}_x(\eta)}\mathcal{M}} \operatorname{grad} f_j(x)\| \le L_T\|\eta\|. \tag{3}$$

If $x_{k+1} = \mathrm{Retr}_{x_k}(\eta_k)$ with $\|\eta_k\| \le \delta_T$, then

$$\|\operatorname{grad} f_j(x_i) - \mathrm{P}_{T_{x_i}\mathcal{M}} \operatorname{grad} f_j(x_0)\| \le L_T \sum_{k=0}^{i-1} \|\eta_k\|. \tag{4}$$

Both estimates hold with f_j replaced by F.

The local bound controls the recursive SARAH/SPIDER and projected-momentum errors, while the pathwise bound controls the SVRG snapshot error; the table methods use the analogous ambient-gradient path estimate.

3 THE iRPMS FRAMEWORK

3.1 INEXACT TANGENT PROXIMAL STEP

We first describe the implementable inexact solution of the tangent-space proximal subproblem and the residual criterion used to control its deviation from the exact ManPG step. Under Assumption 2.1, a smooth local defining map gives B_x with $T_x\mathcal{M} = \ker(B_x)$; see Lemma A.1. Given a tangent surrogate $m_i^t \in T_{x_i^t}\mathcal{M}$ for $\operatorname{grad} F(x_i^t)$, the exact ManPG step is

$$\widehat{\eta}_i^t = \arg\min_{\eta \in T_{x_i^t}\mathcal{M}} \left\{ \langle m_i^t, \eta\rangle + \frac{1}{2\alpha}\|\eta\|^2 + R(x_i^t + \eta) \right\}. \tag{5}$$

This constrained proximal problem usually has no closed form. Let

$$F_i^t(\eta) := \langle m_i^t, \eta\rangle + \frac{1}{2\alpha}\|\eta\|^2 + R(x_i^t + \eta), \qquad G_i^t(\lambda) := F_i^{t,*}(-B_{x_i^t}^\top \lambda). \tag{6}$$

The inner routine to solve the dual problem of (5), that is $\min_{\lambda \in \mathbb{R}^p} G_i^t(\lambda)$, returns λ_i^t satisfying the prescribed, directly checkable condition $\|\nabla G_i^t(\lambda_i^t)\| \le \Delta_t$. Its associated primal point and tangent step are

$$w_i^t = \operatorname{prox}_{\alpha R}\Big(x_i^t - \alpha(m_i^t + B_{x_i^t}^\top \lambda_i^t)\Big) - x_i^t, \qquad \eta_i^t = \mathrm{P}_{T_{x_i^t}\mathcal{M}} w_i^t. \tag{7}$$

The residual has a concrete geometric meaning: uniformly in t and i,

$$\|B_{x_i^t} w_i^t\| \le \Delta_t, \qquad \|w_i^t - \eta_i^t\| \le C_0\Delta_t, \qquad \|\eta_i^t - \widehat{\eta}_i^t\| \le C_1\sqrt{\Delta_t}. \tag{8}$$

Thus a dual stopping rule controls both tangent infeasibility and deviation from the exact proximal direction. Here C_0 and C_1 are the constants in Lemmas A.1 and B.1, respectively.

3.2 iRPMS UPDATE

We next assemble the full-pass shuffling scheme by combining a variance-reduced gradient estimator, projected momentum, and the inexact tangent proximal step described above, and then formulate the epochwise estimator condition used in the subsequent analysis. At epoch t, choose a permutation π_t

Algorithm 1: iRPMS (inexact Riemannian Proximal Momentum Shuffling)

1. Choose $\widetilde{x}^0 \in \mathcal{M}$, $0 < \alpha \le \bar{\alpha}$, $0 < \underline{\gamma} \le \gamma_i^t \le 1$, $0 \le \chi_i^t \le \bar{\chi} < 1$ with $\chi_0^t = 0$, and $0 \le \Delta_t \le \bar{\Delta}$. Set $\bar{m}_0^0 = 0$.
2. For epochs $t = 0, 1, \ldots$:
   (a) Set $x_0^t = \widetilde{x}^t$ and choose a permutation π_t of $\{1, \ldots, n\}$.
   (b) For $i = 0, \ldots, n-1$, do:
      i. Construct $u_i^t \in T_{x_i^t}\mathcal{M}$, a variance-reduced estimator of $\operatorname{grad} F(x_i^t)$, according to the chosen variance-reduction module and the current permutation, and set $m_i^t = \chi_i^t \bar{m}_i^t + (1 - \chi_i^t) u_i^t$.
      ii. Find λ_i^t such that $\|\nabla G_i^t(\lambda_i^t)\| \le \Delta_t$.
      iii. Form w_i^t and η_i^t by (7).
      iv. Update $x_{i+1}^t = \operatorname{Retr}_{x_i^t}(\gamma_i^t \eta_i^t)$ and $\bar{m}_{i+1}^t = \mathrm{P}_{T_{x_{i+1}^t}\mathcal{M}} m_i^t$.
   (c) Set $\widetilde{x}^{t+1} = x_n^t$ and $\bar{m}_0^{t+1} = \bar{m}_n^t$.

of $\{1, \ldots, n\}$ and set $x_0^t = \widetilde{x}^t$. For $i = 0, \ldots, n-1$, the chosen variance-reduction module forms an estimator $u_i^t \in T_{x_i^t}\mathcal{M}$ of $\operatorname{grad} F(x_i^t)$ according to the current permutation and the estimator-specific update. Projected momentum and the estimator are combined as

$$m_i^t = \chi_i^t \bar{m}_i^t + (1 - \chi_i^t) u_i^t, \qquad 0 \le \chi_i^t \le \bar{\chi} < 1, \qquad \chi_0^t = 0. \tag{9}$$

After computing (7), the algorithm performs

$$x_{i+1}^t = \operatorname{Retr}_{x_i^t}(\gamma_i^t \eta_i^t), \qquad \bar{m}_{i+1}^t = \mathrm{P}_{T_{x_{i+1}^t}\mathcal{M}}(m_i^t), \qquad 0 < \underline{\gamma} \le \gamma_i^t \le 1, \tag{10}$$

and closes the epoch with $\widetilde{x}^{t+1} = x_n^t$ and $\bar{m}_0^{t+1} = \bar{m}_n^t$. The concrete algorithm is stated in Algorithm 1, and is referred to as iRPMS (inexact Riemannian Proximal Momentum Shuffling).

Algorithm 1, with deterministic or random permutations, has three layers: a full-pass variance-reduced estimator, projected momentum, and an inexact tangent proximal step. The convergence proof uses only the estimator energy condition below, rather than an estimator-specific recursion. Let $e_i^t := m_i^t - \operatorname{grad} F(x_i^t)$ and $S_t^2 := \sum_{i=0}^{n-1} \|\eta_i^t\|^2$. The estimator requirements are collected in the following single assumption.

Assumption 3.1 (Epochwise estimator condition)**.** There are constants $M_u, c_e, \mu, \nu > 0$ and a nonnegative sequence $\{E_t\}$ with $E_0 = 0$ such that

(i) $\|u_i^t\| \le M_u$ for all $t \ge 0$ and $i = 0, \ldots, n-1$;

(ii) the projected-momentum errors satisfy

$$\sum_{i=0}^{n-1} \|e_i^t\|^2 + \mu E_t \le c_e(E_t - E_{t+1}) + \nu S_t^2. \tag{11}$$

Assumption 3.1 is verified for four variance-reduced estimators in Section 4.3. For accuracy complexity, we use the predictable rule

$$\Delta_t = \min\{\bar{\Delta}, \max\{\kappa S_{t-1}^2,\ \varsigma_t\}\}, \qquad S_{-1} := 0, \tag{12}$$

where κ is small enough for the history term to be absorbed by the preceding epoch decrease. The deterministic floor ς_t prevents the inner target from becoming unnecessarily small and therefore permits an a priori inner-work bound. For the KL analysis, the same two-term maximum uses a summable polynomial tail; the capped rule is

$$\Delta_t = \min\{\bar{\Delta}, \max\{\kappa_1 S_{t-1}^2,\ \kappa_2 (t - t_0)^{-r}\}\}, \qquad t > t_0 + 1, \quad r > 2. \tag{13}$$

Both tolerances are known before solving the current dual problem, unlike relative criteria involving the current exact or approximate tangent step (He et al., 2026; Zheng et al., 2025).

4 Main theoretical results

4.1 Descent, stationarity and complexity

We first establish the epochwise descent and stationarity properties of iRPMS, and then derive the resulting outer and overall inner complexity bounds. The next result condenses the descent

mechanism. Its full constants and proof appear in Theorem B.3. We write $\mathcal{L}_t := \Phi(\widetilde{x}^t) + c_e E_t/(2\tau)$ with $\tau > 0$.

Theorem 4.1 (Epochwise descent)**.** *Suppose Assumptions 2.1 and 3.1 hold. Let* $\{(x_i^t, \eta_i^t) : t \in \mathbb{N},\ i = 0, \dots, n-1\}$ *be generated by Algorithm 1, and let* $\widehat{\eta}_i^t$ *be defined by* (5)*. For sufficiently small fixed* $\alpha > 0$ *and a suitable* $\tau > 0$*, there exist constants* $a_1, a_2, a_3, C > 0$ *such that*

$$\mathcal{L}_{t+1} \le \mathcal{L}_t - a_1 S_t^2 - a_2 \sum_{i=0}^{n-1} \|\widehat{\eta}_i^t\|^2 - a_3 E_t + C\Delta_t. \tag{14}$$

Consequently, if $\sum_t \Delta_t < \infty$*, then* $S_t \to 0$*,* $E_t \to 0$*,* $\widehat{\eta}_i^t \to 0$*, and every accumulation point of the iterates is critical for* (1)*.*

The proof first compares the inexact tangent model with its exact minimizer using (8), then combines smooth retraction estimates with convexity of R. Young's inequality converts the cross term $\langle e_i^t, \eta_i^t \rangle$ into estimator energy, after which (11) telescopes. Compactness provides uniform geometric and boundedness estimates.

For complexity, we use the model-based criticality measure defined in (Zhang, 2026),

$$\Theta(x, \eta) := \sqrt{\operatorname{dist}^2(0, \nabla F(x+\eta) + \partial R(x+\eta) + N_x\mathcal{M}) + \|\eta\|^2}. \tag{15}$$

We call x ε-critical when $\Theta(x, \eta) \le \varepsilon$ for some $\eta \in T_x\mathcal{M}$. The following proposition relates this measure to the exact model step and estimator error.

Proposition 4.2 (Criticality estimate)**.** *Suppose Assumptions 2.1 and 3.1 hold. Let* $\{(x_i^t, \eta_i^t) : t \in \mathbb{N},\ i = 0, \dots, n-1\}$ *be generated by Algorithm 1 with* $\bar{\alpha} \le C_\alpha$ *as in Lemma B.2, and let* $\widehat{\eta}_i^t$ *be defined by* (5)*. Then, for every* $t \ge 0$ *and* $i = 0, \dots, n-1$*,*

$$\Theta(x_i^t, \widehat{\eta}_i^t)^2 \le C_\Theta \|\widehat{\eta}_i^t\|^2 + 2\|e_i^t\|^2, \qquad C_\Theta := 1 + 2(\ell_f + \alpha^{-1})^2. \tag{16}$$

The proof follows from Proposition D.1 in Appendix D.

The following theorem combines this criticality estimate with the predictable two-term tolerance and the three dual-solver bounds; its outer-complexity estimate is proved in Theorem D.2, and the overall proximal-evaluation bounds are completed in Appendix D.3.

Theorem 4.3 (Accuracy complexity for fixed n)**.** *Under the assumptions and parameter bounds of Theorem D.2, regard* n *and the algorithmic parameters as fixed problem constants. Let* $\varepsilon > 0$ *satisfy* $\varepsilon^2/C_{\mathrm{tol}} \le \bar{\Delta}$*. For* $t < T_\varepsilon$*, choose* (12) *with* $\varsigma_t = \varepsilon^2/C_{\mathrm{tol}}$*, where* κ *is below the absorption threshold and* $C_{\mathrm{tol}} > 0$ *is the constant specified in Theorem D.2. Then iRPMS produces an* ε*-critical point within* $O(\varepsilon^{-2})$ *epochs and, because one epoch has fixed length* n*,* $O(\varepsilon^{-2})$ *retractions. The total numbers of proximal evaluations of* R *required by FISTA, NFG, and accumulative regularization (AR) as the dual solver are, respectively,* $O(\varepsilon^{-4})$*,* $O(\varepsilon^{-3}\log\varepsilon^{-1})$*, and* $O(\varepsilon^{-3})$*.*

Remark 4.4 (Comparison with stochastic complexity guarantees). For fixed n, the retraction and overall proximal-evaluation complexity orders in Theorem 4.3 match those established in Zhang (2026). The corresponding stationarity guarantees, however, are different. The latter targets expected squared criticality, whereas Theorem 4.3 guarantees an ε-critical point within the stated computational budgets for every admissible sequence of permutations. Thus, our bounds hold deterministically for fixed shuffling orders and pathwise for random permutations, without taking expectations over the permutations.

4.2 KL WHOLE-SEQUENCE CONVERGENCE

We next study the asymptotic behavior of the full-pass iterates under the KL property of the natural shifted objective, including whole-sequence convergence and last-iterate rates. Since the proximal model is defined in the tangent space, we consider the shifted objective $\Phi_0(x, \eta) := \Phi(x+\eta) + \iota_{T\mathcal{M}}(x, \eta)$, where $T\mathcal{M} := \{(x, \eta) : x \in \mathcal{M},\ \eta \in T_x\mathcal{M}\}$. The optimality condition of the exact model step provides subgradient information at $x_i^t + \widehat{\eta}_i^t$, making Φ_0 suitable for the KL analysis. Set $z_t := (x_{n-1}^t, \widehat{\eta}_{n-1}^t)$. The following theorem gives the epochwise descent and relative-error estimates for Φ_0.

Theorem 4.5 (Shifted-objective estimates)**.** *Suppose Assumptions 2.1 and 3.1 hold. Let* $\{(x_i^t, \eta_i^t) : t \in \mathbb{N},\ i = 0, \dots, n-1\}$ *be generated by Algorithm 1 with* $\gamma_i^t = 1$*, and let* $\widehat{\eta}_i^t$ *be defined by* (5)*. Choose* $\rho_1, \rho_2 > 0$ *and the stepsize as in Proposition C.4(ii). With the constants defined there, the following statements hold.*

(i) For every $t \geq 1$*,*

$$\begin{aligned}\Phi_0(z_t) \leq \Phi_0(z_{t-1}) - \sigma_\eta \sum_{i=0}^{n-2} \|\widehat{\eta}_i^t\|^2 - \sigma' \|\widehat{\eta}_{n-1}^t\|^2 + b\|\widehat{\eta}_{n-1}^{t-1}\|^2 \\ - \kappa_e c_e E_{t+1} + (\kappa_e c_e - \kappa_e \mu - \rho_1 c_e) E_t + (\rho_1 c_e - \rho_1 \mu) E_{t-1} \\ - \sigma_S S_t^2 + \rho_1 \nu S_{t-1}^2 + n\bar{C}_\Delta \Delta_t + C_\Delta \Delta_{t-1},\end{aligned} \tag{17}$$

where $\sigma_\eta > 0$*,* $\sigma' > b$*, and* $\sigma_S > 0$*.*

(ii) There exists $c_\Phi > 0$ *such that, for every* $t \geq 0$*,*

$$\operatorname{dist}(0, \partial\Phi_0(z_t)) \leq c_\Phi\big(\|\widehat{\eta}_{n-1}^t\| + S_t + \sqrt{E_t}\big). \tag{18}$$

The proof is given in Propositions C.4 and C.5 in Appendix C. Combining these estimates with the tolerance rule yields the following KL convergence result.

Theorem 4.6 (KL sequential convergence and complexity)**.** *Suppose Assumptions 2.1 and 3.1 hold. Let the iterates be generated by Algorithm 1 with* $\gamma_i^t = 1$*, the tolerance rule* (13) *with* $r > 2$*, and the parameter conditions in Theorem C.6. If* Φ_0 *is KL at every cluster point of* $\{z_t : t \in \mathbb{N}\}$*, then*

$$\sum_{t=1}^{\infty} \left(\sum_{i=0}^{n-1} \|\widehat{\eta}_i^t\| + S_t + \sqrt{E_t} \right) < \infty, \tag{19}$$

and all inner-iterate sequences $\{x_i^t : t \in \mathbb{N}\}$ *converge to one critical point* x^**. With fixed* n*, the best-iterate bound reaches* ε*-criticality within* $O(\varepsilon^{-1})$ *epochs along each trajectory. Under* $\bar{\Delta} > 0$*,* $\kappa_2 > 0$*, and fixed positive tolerances independent of* ε *for the finitely many initial epochs, the FISTA, NFG, and AR proximal-evaluation bounds become* $O(\varepsilon^{-(r+1)})$*,* $O(\varepsilon^{-(1+r/2)} \log \varepsilon^{-1})$*, and* $O(\varepsilon^{-(1+r/2)})$*, respectively.*

The next theorem gives last-iterate rates under a KL exponent assumption; its proof is given in Appendix C.4.

Theorem 4.7 (KL convergence rates)**.** *Suppose the assumptions and parameter conditions of Theorem 4.6 hold. Let* x^* *be the critical point obtained therein, set* $z^* := (x^*, 0)$ *and* $\mathcal{R}_t := \sum_{i=0}^{n-1} \|\widehat{\eta}_i^t\| + S_t + \sqrt{E_t}$*, and suppose that* Φ_0 *has KL exponent* $\theta \in [0, 1)$ *at* z^**. Then the following statements hold.*

(i) If $\kappa_2 = 0$ *and* $\theta \in [0, 1/2]$*, there exist* $C > 0$ *and* $\rho \in (0, 1)$ *such that, for all sufficiently large* t *and* $i = 0, \dots, n-1$*,*

$$|\Phi_0(z_t) - \Phi_0(z^*)| \leq C\rho^t, \qquad \mathcal{R}_t \leq C\sqrt{n}\,\rho^t, \qquad \|x_i^t - x^*\| \leq C\rho^t. \tag{20}$$

(ii) If $\kappa_2 > 0$ *or* $\theta \in (1/2, 1)$*, define* $\bar{\theta} := \max\{\theta, r/[2(r-1)]\}$ *when* $\kappa_2 > 0$ *and* $\bar{\theta} := \max\{\theta, 1/2\}$ *otherwise. There exist* $C, C' > 0$ *such that, for all sufficiently large* t *and* $i = 0, \dots, n-1$*,*

$$|\Phi_0(z_t) - \Phi_0(z^*)| \leq Ct^{-1/(2\bar{\theta}-1)} + C\kappa_2^{r/[2\bar{\theta}(r-1)]} t^{-r/(2\bar{\theta})}, \tag{21}$$

$$\mathcal{R}_t \leq C\sqrt{n}\, t^{-1/(4\bar{\theta}-2)} + C\sqrt{n}\, \kappa_2^{r/[4\bar{\theta}(r-1)]} t^{-r/(4\bar{\theta})}, \tag{22}$$

$$\|x_i^t - x^*\| \leq C' t^{-\kappa}, \tag{23}$$

where $\kappa := (1-\bar{\theta})/(2\bar{\theta}-1)$ *when* $\kappa_2 = 0$*, and* $\kappa := \min\{(1-\bar{\theta})/(2\bar{\theta}-1), r(1-\bar{\theta})/(2\bar{\theta}), (r-2)/2\}$ *otherwise.*

Remark 4.8 (Comparison with stochastic KL convergence). Zhang (2026) establishes almost-sure finite length and whole-sequence convergence under the additional assumption that the shifted-objective values converge almost surely to a common deterministic constant. Its explicit algorithmic

KL rates are given for the deterministic specialization with $n = 1$. In contrast, Theorems 4.6 and 4.7 establish finite length, whole-sequence convergence, and last-iterate rates along every full-pass shuffling trajectory satisfying their respective assumptions, without requiring a common limiting objective value across trajectories. These conclusions are deterministic for fixed shuffling orders and hold pathwise for random permutations, owing to the epochwise pathwise descent established above.

4.3 Estimator instantiations and scope

We first establish a momentum-lifting lemma that reduces the verification of Assumption 3.1 to bounds for the base estimators.

Lemma 4.9 (Epochwise projected-momentum lifting)**.** *Suppose Assumption 2.1 holds. Let* $\{(x_i^t, u_i^t, m_i^t, \eta_i^t) : t \in \mathbb{N},\ i = 0, \dots, n-1\}$ *be generated by Algorithm 1, and define* $\delta_i^t := \|u_i^t - \operatorname{grad} F(x_i^t)\|$. *Suppose that:*

(i) The estimators u_i^t *are uniformly bounded and* $\|\eta_i^t\| \le \delta_T$ *for all* t, i, *where* δ_T *is from Lemma 2.2.*

(ii) There exist $C_u > 0$, $C_M \ge 0$, *and a nonnegative sequence* $\{M_t\}$ *such that, for every* $t \ge 0$ *and* $i = 0, \dots, n-1$,

$$\delta_i^t \le C_u(S_t + \sqrt{M_t}), \qquad \delta_0^t \le C_u\sqrt{M_t}, \qquad M_{t+1} \le C_M S_t^2. \tag{24}$$

Then, with $\rho = 1$, *there exist* $C_{\mathrm{lift}}, C_{\mathrm{next}} > 0$ *such that* $E_t := \|e_0^t\|^2 + \rho M_t$ *satisfies*

$$\sum_{i=0}^{n-1} \|e_i^t\|^2 \le C_{\mathrm{lift}}(E_t + S_t^2), \qquad E_{t+1} \le C_{\mathrm{next}} S_t^2. \tag{25}$$

Consequently, any $c_e > C_{\mathrm{lift}}$, $0 < \mu < c_e - C_{\mathrm{lift}}$, *and* $\nu \ge C_{\mathrm{lift}} + c_e C_{\mathrm{next}}$ *verify* (11). *If* $e_0^0 = M_0 = 0$, *then* $E_0 = 0$.

The proof is given in Appendix E.1. The local step bound follows from Lemma B.2 when $\bar{\alpha} \le C_\alpha$. The following theorem records the initialization, memory, and parameter choices for four base estimators. Its proof is in Appendix E.

Theorem 4.10 (Four full-pass variance-reduction verifications)**.** *Suppose Assumption 2.1 holds. For each estimator below, choose* M_u *according to the corresponding uniform bound and* $\bar{\alpha} \le C_\alpha$ *so that Lemma B.2 applies. Let the iterates and estimators be generated by Algorithm 1, with the local step bounds in Lemma B.2. Define*

$$C_{\mathrm{tab}} := \sqrt{2\max\{1, n^2\ell_f^2 C_R^2\}}, \qquad C_M^{\mathrm{tab}} := n^2\ell_f^2 C_R^2. \tag{26}$$

The following choices satisfy the base bounds (24)*:*

Estimator	*Initialization/update*	M_t	C_u	C_M
Projected SVRG	*snapshot;* (63)	0	$2L_T\sqrt{n}$	0
SARAH/SPIDER	*full-gradient reset;* (64)–(65)	0	$2L_T\sqrt{n}$	0
Ambient SAGA	$y_j^0 = \nabla f_j(x_0^0)$; (66)	$\sum_j \|y_j^t - \nabla f_j(x_0^t)\|^2$	$(1 + n^{-1/2})C_{\mathrm{tab}}$	C_M^{tab}
Ambient SAG	$y_j^0 = \nabla f_j(x_0^0)$; (67)	$\sum_j \|y_j^t - \nabla f_j(x_0^t)\|^2$	$(n^{-1} + n^{-1/2})C_{\mathrm{tab}}$	C_M^{tab}

For SAGA and SAG, keep the gradient table fixed within each epoch and update it at the end using the component gradients collected during that epoch. For SVRG and SARAH/SPIDER, use a full-gradient epoch reset. Under the displayed initializations, $e_0^0 = M_0 = 0$. *Take* $\rho = 1$ *as in Lemma 4.9, then take*

$$c_e > C_{\mathrm{lift}}, \qquad 0 < \mu < c_e - C_{\mathrm{lift}}, \qquad \nu \ge C_{\mathrm{lift}} + c_e C_{\mathrm{next}}. \tag{27}$$

These parameters verify Assumption 3.1 with $E_0 = 0$.

Remark 4.11 (Applicability to the four estimators). Theorem 4.10 verifies Assumption 3.1 for projection-based SVRG, projection-based SARAH/SPIDER, projected ambient SAGA, and projected ambient SAG, combined with the projected-momentum update in Algorithm 1. Consequently, the descent and complexity results in Theorems 4.1 and 4.3, as well as the KL convergence and rate results in Theorems 4.6 and 4.7, apply to all four full-pass realizations under their respective assumptions and parameter conditions.

AI USE STATEMENT

Generative AI tools were used for literature search and retrieval, language editing, and LaTeX formatting, including adaptation to the ICLR template. All AI-assisted edits were reviewed by the authors, who take full responsibility for the final content of the submission.

REPRODUCIBILITY STATEMENT

This is a theoretical paper. The standing assumptions and algorithm are stated in the main text. The appendices provide the notation, auxiliary results, explicit parameter choices, proofs of the full-pass conclusions, and estimator verifications. Previously established results are stated with references to their sources.

REFERENCES

P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, Princeton, NJ, 2008.

Hedy Attouch, Jérôme Bolte, and Benar Fux Svaiter. Convergence of descent methods for semi-algebraic and tame problems: Proximal algorithms, forward–backward splitting, and regularized gauss–seidel methods. *Mathematical Programming*, 137(1–2):91–129, 2013. doi: 10.1007/s10107-011-0484-9.

Reza Babanezhad, Issam H. Laradji, Alireza Shafaei, and Mark Schmidt. MASAGA: A linearly-convergent stochastic first-order method for optimization on manifolds. In *Machine Learning and Knowledge Discovery in Databases*, volume 11052 of *Lecture Notes in Computer Science*, pp. 344–359. Springer, 2019. doi: 10.1007/978-3-030-10928-8_21.

Heinz H. Bauschke and Patrick L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. CMS Books in Mathematics. Springer, Cham, 2 edition, 2017. doi: 10.1007/978-3-319-48311-5.

Amir Beck and Itay Rosset. A dynamic smoothing technique for a class of nonsmooth optimization problems on manifolds. *SIAM Journal on Optimization*, 33(3):1473–1493, 2023.

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. doi: 10.1137/080716542.

Jérôme Bolte, Aris Daniilidis, Adrian S. Lewis, and Masahiro Shiota. Clarke subgradients of stratifiable functions. *SIAM Journal on Optimization*, 18(2):556–572, 2007. doi: 10.1137/050639860.

Jérôme Bolte, Shoham Sabach, and Marc Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1–2):459–494, 2014. doi: 10.1007/s10107-013-0701-9.

Nicolas Boumal. *An Introduction to Optimization on Smooth Manifolds*. Cambridge University Press, Cambridge, UK, 2023.

Shixiang Chen, Shiqian Ma, Anthony Man-Cho So, and Tong Zhang. Proximal gradient method for nonsmooth optimization over the Stiefel manifold. *SIAM Journal on Optimization*, 30(1):210–239, 2020a. doi: 10.1137/18M122457X.

Shixiang Chen, Shiqian Ma, Lingzhou Xue, and Hui Zou. An alternating manifold proximal gradient method for sparse principal component analysis and sparse canonical correlation analysis. *INFORMS Journal on Optimization*, 2(3):192–208, 2020b. doi: 10.1287/ijoo.2019.0032.

Shixiang Chen, Shiqian Ma, Anthony Man-Cho So, and Tong Zhang. Nonsmooth optimization over the Stiefel manifold and beyond: Proximal gradient method and recent variants. *SIAM Review*, 66 (2):319–352, 2024. doi: 10.1137/24M1628578.

Alexandre d'Aspremont, Laurent El Ghaoui, Michael I. Jordan, and Gert R. G. Lanckriet. A direct formulation for sparse PCA using semidefinite programming. *SIAM Review*, 49(3):434–448, 2007. doi: 10.1137/050645506.

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

Kangkang Deng, Jiang Hu, Jiayuan Wu, and Zaiwen Wen. Oracle complexities of augmented lagrangian methods for nonsmooth composite optimization on a compact submanifold. *Mathematics of Operations Research*, 2025a. doi: 10.1287/moor.2024.0498.

Kangkang Deng, Zheng Peng, and Weihe Wu. Single-loop $\mathcal{O}(\varepsilon^{-3})$ stochastic smoothing algorithms for nonsmooth Riemannian optimization, 2025b. URL https://arxiv.org/abs/2505.09485.

Ehsan Elhamifar and René Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(11):2765–2781, 2013. doi: 10.1109/TPAMI.2013.57.

Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. SPIDER: Near-optimal non-convex optimization via stochastic path integrated differential estimator. In *Advances in Neural Information Processing Systems*, volume 31, pp. 689–699, 2018.

Bin Gao, Xin Liu, Xiaojun Chen, and Ya-xiang Yuan. A new first-order algorithmic framework for optimization problems with orthogonality constraints. *SIAM Journal on Optimization*, 28(1): 302–332, 2018. doi: 10.1137/16M1098759.

Andi Han and Junbin Gao. Riemannian stochastic recursive momentum method for non-convex optimization. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pp. 2505–2511. International Joint Conferences on Artificial Intelligence Organization, 2021. doi: 10.24963/ijcai.2021/345.

Hao He, Ruyu Liu, Yitian Qian, and Shaohua Pan. An inexact variable metric proximal linearization method for composite optimization on manifolds, 2026. URL https://doi.org/10.1007/s10589-026-00810-3.

Wen Huang and Ke Wei. Riemannian proximal gradient methods. *Mathematical Programming*, 194 (1–2):371–413, 2022. doi: 10.1007/s10107-021-01632-3.

Wen Huang and Ke Wei. An inexact riemannian proximal gradient method. *Computational Optimization and Applications*, 85(1):1–32, 2023. doi: 10.1007/s10589-023-00451-w.

Bo Jiang, Meng Xu, Xingju Cai, and Ya-Feng Liu. An inexact proximal framework for nonsmooth riemannian difference-of-convex optimization, 2025. URL https://arxiv.org/abs/2509.08561.

Jiachen Jin, Kangkang Deng, and Hongxia Wang. A single-loop stochastic Riemannian ADMM for nonsmooth optimization, 2025. URL https://arxiv.org/abs/2512.22750.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, pp. 315–323, 2013.

Hiroyuki Kasai, Hiroyuki Sato, and Bamdev Mishra. Riemannian stochastic recursive gradient algorithm. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2516–2524. PMLR, 2018.

Guanghui Lan, Yuyuan Ouyang, and Zhe Zhang. Optimal and parameter-free gradient minimization methods for convex and nonconvex optimization. *Mathematical Programming*, 2026. doi: 10.1007/s10107-026-02352-2.

Guoyin Li and Ting Kei Pong. Calculus of the exponent of kurdyka-łojasiewicz inequality and its applications to linear convergence of first-order methods. *Foundations of Computational Mathematics*, 18(5):1199–1232, 2018. doi: 10.1007/s10208-017-9366-8.

Jiaxiang Li, Shiqian Ma, and Tejes Srivastava. A riemannian alternating direction method of multipliers. *Mathematics of Operations Research*, 50(4):3222–3242, 2025. doi: 10.1287/moor.2023.0068.

Qia Li, Na Zhang, Junyu Feng, and Hanwei Yan. Proximal methods for structured nonsmooth optimization over Riemannian submanifolds, 2024. URL https://arxiv.org/abs/2411.15776. arXiv:2411.15776, version 4.

Xiao Li, Andre Milzarek, and Junwen Qiu. Convergence of random reshuffling under the Kurdyka–Łojasiewicz inequality. *SIAM Journal on Optimization*, 33(2):1092–1120, 2023. doi: 10.1137/21M1468048.

Yuqing Liang and Dongpo Xu. Last-iterate convergence of shuffling momentum gradient method under the kurdyka-lojasiewicz inequality. *Journal of Machine Learning Research*, 26(123):1–51, 2025. URL https://www.jmlr.org/papers/v26/24-1243.html.

Yurii Nesterov. *Lectures on Convex Optimization*. Springer, Cham, 2 edition, 2018. doi: 10.1007/978-3-319-91578-4.

Lam M. Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. SARAH: A novel method for machine learning problems using stochastic recursive gradient. In *International Conference on Machine Learning*, volume 70, pp. 2613–2621, 2017.

Zheng Peng, Weihe Wu, Jiang Hu, and Kangkang Deng. Riemannian smoothing gradient type algorithms for nonsmooth optimization problem on compact riemannian submanifold embedded in euclidean space. *Applied Mathematics & Optimization*, 88:85, 2023. doi: 10.1007/s00245-023-10061-x.

Junwen Qiu, Xiao Li, and André Milzarek. A new random reshuffling method for nonsmooth nonconvex finite-sum optimization. *Journal of Machine Learning Research*, 26(191):1–46, 2025.

R. Tyrrell Rockafellar and Roger J.-B. Wets. *Variational Analysis*. Springer, 1998. doi: 10.1007/978-3-642-02431-3.

Hiroyuki Sato, Hiroyuki Kasai, and Bamdev Mishra. Riemannian stochastic variance reduced gradient algorithm with retraction and vector transport. *SIAM Journal on Optimization*, 29(2):1444–1472, 2019. doi: 10.1137/17M1116787.

Mark Schmidt, Nicolas Le Roux, and Francis Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1–2):83–112, 2017. doi: 10.1007/s10107-016-1030-6.

Bokun Wang, Shiqian Ma, and Lingzhou Xue. Riemannian stochastic proximal gradient methods for nonsmooth optimization over the Stiefel manifold. *Journal of Machine Learning Research*, 23 (106):1–33, 2022.

Zaiwen Wen and Wotao Yin. A feasible method for optimization with orthogonality constraints. *Mathematical Programming*, 142(1–2):397–434, 2013. doi: 10.1007/s10107-012-0584-1.

Meng Xu, Bo Jiang, Ya-Feng Liu, and Anthony Man-Cho So. On the oracle complexity of a riemannian inexact augmented lagrangian method for nonsmooth composite problems over riemannian submanifolds. *Optimization Letters*, 2025. doi: 10.1007/s11590-025-02210-8.

Meng Xu, Bo Jiang, Ya-Feng Liu, and Anthony Man-Cho So. A Riemannian alternating descent ascent algorithmic framework for nonconvex-linear minimax problems on Riemannian manifolds. *Mathematics of Operations Research*, 2026. doi: 10.1287/moor.2025.1055.

Ganzhao Yuan. ADMM for nonsmooth composite optimization under orthogonality constraints, 2024. URL https://arxiv.org/abs/2405.15129.

Hongyi Zhang, Sashank J. Reddi, and Suvrit Sra. Riemannian SVRG: Fast stochastic optimization on riemannian manifolds. In *Advances in Neural Information Processing Systems*, volume 29, pp. 4592–4600, 2016.

Jingzhao Zhang, Hongyi Zhang, and Suvrit Sra. R-SPIDER: A fast riemannian stochastic optimization algorithm with curvature independent rate. *arXiv preprint arXiv:1811.04194*, 2018. doi: 10.48550/arXiv.1811.04194.

Na Zhang. An inexact Riemannian proximal momentum variance-reduced method: Complexity bounds and KL sequential convergence, 2026. URL https://arxiv.org/abs/2608.16355. arXiv:2608.16355.

Zhong Zheng, Xin Yu, Shiqian Ma, and Lingzhou Xue. A new inexact manifold proximal linear algorithm with adaptive stopping criteria, 2025. URL https://arxiv.org/abs/2508.19234.

Hui Zou, Trevor Hastie, and Robert Tibshirani. Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15(2):265–286, 2006. doi: 10.1198/106186006X113430.

A Notation, basic concepts, and auxiliary results

This appendix supplies the notation and background used in the main text. The notation agrees with Sections 2–3.

A.1 Euclidean and variational notation

We use $\mathbb{N} := \{0, 1, 2, \ldots\}$. The ambient space $\mathbb{R}^d$ is equipped with its Euclidean inner product $\langle \cdot, \cdot \rangle$ and norm $\| \cdot \|$. For matrix variables, these are the Frobenius inner product and norm. The same norm notation denotes the induced operator norm of a linear map. Its adjoint is denoted by A^*, or by $A^\top$ when a matrix representation is used. We write $B(x, \delta) := \{y : \|y - x\| < \delta\}$ and $\operatorname{dist}(x, S) := \inf_{y \in S} \|x - y\|$, with $\operatorname{dist}(x, \varnothing) := +\infty$. The symbol $B(x, \delta)$ denotes a ball and is distinct from the constraint matrix B_x defined below. For any set S, ι_S denotes its indicator, equal to zero on S and $+\infty$ elsewhere. For a nonempty closed convex set S, P_S denotes its Euclidean projection.

For a proper extended-real-valued function ϕ, let $\operatorname{dom}\phi := \{x : \phi(x) < +\infty\}$. We use the limiting subdifferential $\partial\phi$ of Rockafellar & Wets (1998); specifically, for $x \in \operatorname{dom}\phi$,

$$\widehat{\partial}\phi(x) := \left\{ v : \liminf_{\substack{y \to x \\ y \neq x}} \frac{\phi(y) - \phi(x) - \langle v, y - x \rangle}{\|y - x\|} \geq 0 \right\},$$

and $v \in \partial\phi(x)$ if there are sequences $x^k \to x$, $\phi(x^k) \to \phi(x)$, and $v^k \in \widehat{\partial}\phi(x^k)$ with $v^k \to v$. Both subdifferentials are empty outside $\operatorname{dom}\phi$. We write $\operatorname{dom}\partial\phi := \{x : \partial\phi(x) \neq \varnothing\}$. For a closed set S, its limiting normal cone is $N_S(x) := \partial\iota_S(x)$. On a smooth embedded manifold this is its usual normal space. In particular, $N_x\mathcal{M}$ denotes $N_{\mathcal{M}}(x)$, whereas $N_{T\mathcal{M}}(x, \eta)$ is a normal space in the product ambient space.

When ϕ is proper, closed, and convex, $\partial\phi$ is the convex subdifferential, its Fenchel conjugate is $\phi^*(y) := \sup_x\{\langle y, x\rangle - \phi(x)\}$, and, for $\alpha > 0$, $\operatorname{prox}_{\alpha\phi}(x) := \arg\min_y\{\phi(y) + \|y - x\|^2/(2\alpha)\}$; see Bauschke & Combettes (2017). Accordingly, ∂R denotes the convex subdifferential, whereas $\partial\Phi_0$ denotes the limiting subdifferential. Global ℓ_R-Lipschitz continuity of the convex function R implies $\|\xi\| \leq \ell_R$ for every $\xi \in \partial R(x)$.

The finite-sum objective F and the tangent model F_i^t are different functions, defined in (1) and (6), respectively. The superscript t indexes epochs; $i = 0, \ldots, n-1$ indexes tangent subproblems, while $x_n^t = x_0^{t+1}$ is the endpoint of epoch t. The exact model step is $\widehat{\eta}_i^t$, the computed tangent step is η_i^t, and w_i^t need not be tangent. The quantities e_i^t, S_t, and E_t retain their definitions in Section 3. Empty sums are zero, including when $n = 1$. Unless stated otherwise, constants may depend on the fixed problem, n, and the chosen parameters, but not on t or i.

A.2 Embedded geometry and the construction of B_x

We recall the embedded-manifold conventions of Absil et al. (2008); Boumal (2023). A smooth embedded submanifold $\mathcal{M} \subseteq \mathbb{R}^d$ of dimension $d - p$ admits, near each point, a smooth local defining

map $\psi : U \to \mathbb{R}^p$ such that $\mathcal{M} \cap U = \{x \in U : \psi(x) = 0\}$ and $D\psi(x)$ has full row rank on $\mathcal{M} \cap U$. Its tangent and normal spaces satisfy $T_x\mathcal{M} = \ker D\psi(x)$ and $N_x\mathcal{M} = (T_x\mathcal{M})^\perp$. The Riemannian metric is induced by the ambient inner product, so that $\operatorname{grad} h(x) = \mathrm{P}_{T_x\mathcal{M}} \nabla h(x)$ for a continuously differentiable ambient function h. In particular, $\nabla h(x) - \operatorname{grad} h(x) \in N_x\mathcal{M}$.

The tangent bundle $T\mathcal{M}$ introduced in Section 4.2 is a smooth embedded submanifold of $\mathbb{R}^d \times \mathbb{R}^d$. A smooth retraction $\mathrm{Retr} : T\mathcal{M} \to \mathcal{M}$ satisfies $\mathrm{Retr}_x(0) = x$ and $D\mathrm{Retr}_x(0) = \mathrm{Id}_{T_x\mathcal{M}}$. The projection field $x \mapsto \mathrm{P}_{T_x\mathcal{M}}$ is smooth. At the zero section, $T_{(x,0)}T\mathcal{M} = T_x\mathcal{M} \times T_x\mathcal{M}$ and $N_{T\mathcal{M}}(x, 0) = N_x\mathcal{M} \times N_x\mathcal{M}$.

By compactness, choose finitely many open sets V_ℓ, U_ℓ and smooth defining maps $\psi_\ell : U_\ell \to \mathbb{R}^p$, $\ell = 1, \dots, J$, such that $\overline{V}_\ell \subset U_\ell$, $\{V_\ell \cap \mathcal{M}\}_{\ell=1}^J$ covers $\mathcal{M}$, and each $D\psi_\ell$ has full row rank on $U_\ell \cap \mathcal{M}$. For each $x \in \mathcal{M}$, select one index $\ell(x)$ with $x \in V_{\ell(x)}$ and define

$$B_x := D\psi_{\ell(x)}(x) \in \mathbb{R}^{p\times d}, \qquad \ker B_x = T_x\mathcal{M}, \qquad \operatorname{range}(B_x^\top) = N_x\mathcal{M}. \tag{28}$$

The selection is fixed whenever charts overlap. No global smoothness of $x \mapsto B_x$ across chart changes is required: all estimates below are uniform over this finite family. This is the matrix used in the dual problem (6), not an additional algorithmic variable.

The following uniform geometric estimates are standard consequences of compactness and the embedded-manifold structure; see Absil et al. (2008); Boumal (2023). Their proofs are omitted.

Lemma A.1 (Uniform bounds for the tangent constraint)**.** *Under Assumption 2.1, there exist constants $0 < C_B \le \widetilde{C}$ such that every singular value of B_x lies in $[C_B, \widetilde{C}]$, uniformly over $x \in \mathcal{M}$. With $C_0 := C_B^{-1}$,*

$$\|B_x v\| \le \widetilde{C}\|v\|, \qquad \operatorname{dist}(v, T_x\mathcal{M}) = \|\mathrm{P}_{N_x\mathcal{M}} v\| \le C_0\|B_x v\| \tag{29}$$

for every $v \in \mathbb{R}^d$. Moreover, $\mathrm{P}_{N_x\mathcal{M}} = B_x^\dagger B_x$ and $\mathrm{P}_{T_x\mathcal{M}} = I - B_x^\dagger B_x$, where $B_x^\dagger$ is the Moore–Penrose inverse.

Lemma A.2 (Uniform local retraction estimates)**.** *Under Assumption 2.1, there are constants $\delta, C_R, C_Q, L_R, L_\Phi, L_F > 0$ such that, for every $x \in \mathcal{M}$ and $\eta \in T_x\mathcal{M}$ with $\|\eta\| \le \delta$,*

$$\|\mathrm{Retr}_x(\eta) - x\| \le C_R\|\eta\|, \tag{30}$$

$$\|\mathrm{Retr}_x(\eta) - x - \eta\| \le C_Q\|\eta\|^2, \tag{31}$$

$$|R(\mathrm{Retr}_x(\eta)) - R(x + \eta)| \le L_R\|\eta\|^2, \tag{32}$$

$$|\Phi(\mathrm{Retr}_x(\eta)) - \Phi(x + \eta)| \le L_\Phi\|\eta\|^2, \tag{33}$$

$$F(\mathrm{Retr}_x(\eta)) - F(x) - \langle \operatorname{grad} F(x), \eta\rangle \le \frac{L_F}{2}\|\eta\|^2. \tag{34}$$

The radius is chosen small enough that the shifted points and the line segments used in the smoothness estimates lie in the neighborhood of Assumption 2.1(ii).

Lemma A.3 (Normal lifting on the tangent bundle)**.** *Under Assumption 2.1, there exists $C_{\mathrm{tb}} > 0$ such that, for every $x \in \mathcal{M}$, $\eta \in T_x\mathcal{M}$, and $\zeta \in N_x\mathcal{M}$, one can choose $u \in \mathbb{R}^d$ satisfying $(u, \zeta) \in N_{T\mathcal{M}}(x, \eta)$ and $\|u - \zeta\| \le C_{\mathrm{tb}}\|\eta\|\,\|\zeta\|$. Consequently, $\|u\| \le (1 + C_{\mathrm{tb}}\|\eta\|)\|\zeta\|$.*

For the normal-lifting construction, see also Li et al. (2024).

Compactness gives a constant $M_f > 0$ such that $\|\nabla f_j(x)\| \le M_f$ for every $x \in \mathcal{M}$ and $j = 1, \dots, n$. To justify Lemma 2.2, write $P_x := \mathrm{P}_{T_x\mathcal{M}}$. The smooth projection field is uniformly Lipschitz on $\mathcal{M}$; hence $\operatorname{grad} f_j(x) = P_x \nabla f_j(x)$ is Lipschitz there with a common constant L_g. Since $\operatorname{grad} f_j(y) \in T_y\mathcal{M}$ and P_y is nonexpansive, $\|\operatorname{grad} f_j(y) - P_y \operatorname{grad} f_j(x)\| \le L_g\|y - x\|$. Taking $y = \mathrm{Retr}_x(\eta)$ proves the local estimate. Summing the retraction displacement bounds gives the pathwise estimate, with $L_T = L_g C_R$ after enlarging constants if needed and $\delta_T \le \delta$. Averaging over j yields the estimates for F. Only Lipschitz continuity of the ambient gradients is used.

A.3 KL PROPERTY, EXPONENT, AND UNIFORMIZATION

We recall the ordinary Kurdyka–Łojasiewicz (KL) property (Attouch et al., 2013; Bolte et al., 2007).

Definition A.4 (KL property and KL exponent)**.** A proper lower semicontinuous function $\phi : \mathbb{R}^N \to (-\infty, +\infty]$ has the KL property at $\bar{z} \in \mathrm{dom}\partial\phi$ if there exist $\varepsilon, \delta > 0$ and a continuous concave function $\varphi : [0, \varepsilon) \to [0, +\infty)$ such that $\varphi(0) = 0$, $\varphi \in C^1(0, \varepsilon)$, $\varphi' > 0$ on $(0, \varepsilon)$, and

$$\varphi'(\phi(z) - \phi(\bar{z}))\ \mathrm{dist}(0, \partial\phi(z)) \geq 1 \tag{35}$$

whenever $z \in B(\bar{z}, \delta)$ and $0 < \phi(z) - \phi(\bar{z}) < \varepsilon$. If one can take $\varphi(s) = cs^{1-\theta}$ for some $c > 0$ and $\theta \in [0, 1)$, then ϕ has KL exponent θ at $\bar{z}$. A KL function satisfies this property at every point of $\mathrm{dom}\partial\phi$.

Proper lower semicontinuous semialgebraic functions are KL functions (Bolte et al., 2007). The following standard uniformization is used when the cluster set contains more than one point; see Bolte et al. (2014, Lemma 6).

Lemma A.5 (Uniformized KL property)**.** *Let $\Omega \subseteq \mathrm{dom}\partial\phi$ be compact, let ϕ be proper and lower semicontinuous and constant on Ω, and suppose that ϕ has the KL property at every point of Ω. Then there exist $\varepsilon, \delta > 0$ and one continuous concave $\varphi : [0, \varepsilon) \to [0, +\infty)$ with $\varphi(0) = 0$, $\varphi \in C^1(0, \varepsilon)$, and $\varphi' > 0$, such that* (35) *holds for every $\bar{z} \in \Omega$ and every z satisfying $\mathrm{dist}(z, \Omega) < \delta$ and $0 < \phi(z) - \phi(\bar{z}) < \varepsilon$.*

The shifted objective Φ_0 is defined in Section 4.2. Since $F + R$ is continuous and $T\mathcal{M}$ is closed, Φ_0 is proper, lower semicontinuous, and continuous relative to $T\mathcal{M}$. Thus $z^k \to z$ in $T\mathcal{M}$ implies $\Phi_0(z^k) \to \Phi_0(z)$. Its KL property is the separate assumption in Theorem 4.6.

B Inexact tangent-solve estimates and epochwise descent

All variables in this section are those of Algorithm 1 and (5)–(10). Assumption 3.1(i), projection nonexpansiveness, and the convex combination in (9) imply

$$\|m_i^t\| \leq M_u \qquad (t \geq 0,\ i = 0, \ldots, n-1). \tag{36}$$

Indeed, $m_0^t = u_0^t$ because $\chi_0^t = 0$, and induction within each epoch gives the bound.

B.1 Residual and tangent-step estimates

The next two lemmas are the tangent-model estimates established in Zhang (2026), stated in epochwise notation without repeating their proofs. The model-comparison inequality is included for use below.

Lemma B.1 (Primal error and model comparison)**.** *Suppose Assumption 2.1 holds and the tangent subproblem is solved as in* (6)–(7)*, with $0 < \alpha \leq \bar{\alpha}$ and $0 \leq \Delta_t \leq \bar{\Delta}$. The residual estimates* (8) *hold, and in addition $\|w_i^t - \widehat{\eta}_i^t\| \leq C_2\sqrt{\Delta_t}$, where one may take $C_1 = 2\sqrt{\ell_R C_0 \bar{\alpha}}$ and $C_2 = C_1 + C_0\sqrt{\bar{\Delta}}$. For every $v \in T_{x_i^t}\mathcal{M}$,*

$$F_i^t(v) - F_i^t(\eta_i^t) \geq \frac{1}{2\alpha}\|v - \eta_i^t\|^2 - 2\ell_R C_0 \Delta_t. \tag{37}$$

These constants are independent of t and i; with the geometric and parameter bounds fixed, they are also independent of the particular $\alpha \in (0, \bar{\alpha}]$.

Lemma B.2 (Uniform tangent-step bounds)**.** *Suppose Assumptions 2.1 and 3.1(i) hold and let the iterates be generated by Algorithm 1. Then, for every $t \geq 0$ and $i = 0, \ldots, n-1$,*

$$\|\eta_i^t\| \leq (M_u + \ell_R)\alpha + 2\sqrt{\ell_R C_0 \bar{\Delta}}\sqrt{\alpha}, \qquad \|\widehat{\eta}_i^t\| \leq (M_u + \ell_R)\alpha. \tag{38}$$

Let δ and δ_T be the radii in Lemmas A.2 and 2.2, and set

$$C_\alpha := \left(\frac{\sqrt{\ell_R C_0 \bar{\Delta} + (M_u + \ell_R)\min\{\delta, \delta_T\}} - \sqrt{\ell_R C_0 \bar{\Delta}}}{M_u + \ell_R} \right)^2. \tag{39}$$

If $0 < \bar{\alpha} \leq C_\alpha$, then both $\|\eta_i^t\|$ and $\|\widehat{\eta}_i^t\|$ are at most $\min\{\delta, \delta_T\}$, uniformly in t, i.

Strong convexity of F_i^t makes its conjugate differentiable, and $\nabla G_i^t(\lambda) = -B_{x_i^t} w_i^t(\lambda)$, where $w_i^t(\lambda)$ is the proximal expression in (7). Thus the dual residual measures tangent infeasibility.

B.2 Proof of the epochwise descent and stationarity result

The next statement supplies the explicit coefficients behind Theorem 4.1.

Theorem B.3 (Detailed epochwise descent)**.** *Suppose Assumptions 2.1 and 3.1 hold. Let the iterates be generated by Algorithm 1, and assume* $0 < \alpha \le \bar{\alpha} \le \min\{C_\alpha, 1/(L_F + 2L_R)\}$. *Set* $\sigma_0 := 1/(2\alpha) - L_F/2 - L_R$. *For every* $t \ge 0$ *and* $\tau > 0$,

$$\begin{aligned}\Phi(\widetilde{x}^{t+1}) \le \Phi(\widetilde{x}^t) - \Big(\sigma_0 - \frac{\tau}{2}\Big)\sum_{i=0}^{n-1}\gamma_i^t\|\eta_i^t\|^2 - \frac{1}{2\alpha}\sum_{i=0}^{n-1}\gamma_i^t\|\widehat{\eta}_i^t\|^2 \\ + \frac{c_e}{2\tau}(E_t - E_{t+1}) - \frac{\mu}{2\tau}E_t + \frac{\nu}{2\tau}S_t^2 + 2\ell_R C_0 n\Delta_t.\end{aligned} \tag{40}$$

Proof. Write $c_\Delta := 2\ell_R C_0$. Taking $v = \widehat{\eta}_i^t$ in (37) and using strong convexity of F_i^t on $T_{x_i^t}\mathcal{M}$ give $F_i^t(\eta_i^t) \le F_i^t(0) - \|\widehat{\eta}_i^t\|^2/(2\alpha) + c_\Delta\Delta_t$. Expanding this inequality and using convexity of R at $x_i^t + \gamma_i^t\eta_i^t$, followed by the retraction estimates, yields

$$\Phi(x_{i+1}^t) \le \Phi(x_i^t) - \sigma_0\gamma_i^t\|\eta_i^t\|^2 - \frac{\gamma_i^t}{2\alpha}\|\widehat{\eta}_i^t\|^2 - \gamma_i^t\langle e_i^t, \eta_i^t\rangle + c_\Delta\Delta_t. \tag{41}$$

Here $(\gamma_i^t)^2 \le \gamma_i^t \le 1$ and Lemma B.2 justify the uniform retraction constants. Young's inequality gives $-\gamma_i^t\langle e_i^t, \eta_i^t\rangle \le (\tau/2)\gamma_i^t\|\eta_i^t\|^2 + \|e_i^t\|^2/(2\tau)$. Summing over the epoch and applying (11) proves (40). □

To obtain Theorem 4.1, choose $\tau = \sqrt{\nu/\underline{\gamma}}$ and α small enough that $a_1 := \underline{\gamma}(\sigma_0 - \tau/2) - \nu/(2\tau) > 0$. Then one may take $a_2 = \underline{\gamma}/(2\alpha)$, $a_3 = \mu/(2\tau)$, and $C = \max\{1, 2\ell_R C_0 n\}$ in (14). If $\sum_t \Delta_t < \infty$, summing this descent relation gives $S_t \to 0$, $E_t \to 0$, and $\widehat{\eta}_i^t \to 0$. Moreover, (11) implies $\|e_i^t\| \le \sqrt{c_e}\sqrt{E_t} + \sqrt{\nu}S_t \to 0$.

For any convergent subsequence $x_i^{t_k} \to x^*$, exact model optimality gives vectors $\xi_k \in \partial R(x_i^{t_k} + \widehat{\eta}_i^{t_k})$ and $\zeta_k \in N_{x_i^{t_k}}\mathcal{M}$ such that

$$0 = \operatorname{grad} F(x_i^{t_k}) + e_i^{t_k} + \alpha^{-1}\widehat{\eta}_i^{t_k} + \xi_k + \zeta_k. \tag{42}$$

Since $\|\xi_k\| \le \ell_R$, this equality also bounds ζ_k. Passing to a further subsequence, the closedness of the convex subdifferential graph and of the normal-space graph gives (2) at x^*. There are only finitely many within-epoch indices, so this argument covers every accumulation point of the full sequence of iterates.

C KL tools and proofs of the sequential results

We first state the deterministic KL tools, then verify their hypotheses for the shifted objective Φ_0 and the sequence z_t of Section 4.2.

C.1 Deterministic KL tools with memory and tails

We first present an abstract memory-and-tail KL argument, which will be used later to analyze the convergence of iRPMS.

Theorem C.1 (Deterministic KL principle with memory)**.** *Let* $\Psi : \mathbb{R}^N \to (-\infty, +\infty]$ *be proper, lower semicontinuous, and continuous relative to* $\operatorname{dom}\Psi$. *Let* $\{y^t\} \subset \operatorname{dom}\Psi$ *be bounded and satisfy* $\inf_t \Psi(y^t) > -\infty$. *Let* $\lambda_{\ell,t} \ge 0$, $\ell = 1, \dots, m$, *and* $\sigma_t \ge 0$, *and define* $\lambda_{\ell,-1} := 0$, $\mathcal{T}_t := \sum_{k=t+1}^{\infty}\sigma_k$, *and* $d_t := \sum_{\ell=1}^{m}(\lambda_{\ell,t+1} + \lambda_{\ell,t} + \lambda_{\ell,t-1})$. *Suppose the following conditions hold for all sufficiently large* t.

(i) There exist $a_{1,\ell} > 0$ *and* $a_{2,\ell}, a_{3,\ell} \in \mathbb{R}$ *with* $a_{2,\ell} + \max\{a_{3,\ell}, 0\} < a_{1,\ell}$ *such that*

$$\Psi(y^{t+1}) \le \Psi(y^t) - \sum_{\ell=1}^{m} a_{1,\ell}\lambda_{\ell,t+1}^2 + \sum_{\ell=1}^{m} a_{2,\ell}\lambda_{\ell,t}^2 + \sum_{\ell=1}^{m} a_{3,\ell}\lambda_{\ell,t-1}^2 + \sigma_{t+1}. \tag{43}$$

(ii) There exist $c_{\mathrm{re}} > 0$ *and* $q \in (0,1)$ *such that* $\operatorname{dist}(0, \partial\Psi(y^t)) \le c_{\mathrm{re}} d_t$ *and* $\sum_{t=0}^{\infty} \mathcal{T}_t^q < \infty$.

(iii) Ψ *has the KL property at every accumulation point of* $\{y^t\}$.

Then $\Psi(y^t)$ *converges to a finite value* Ψ_*, $\sum_{t=0}^{\infty} d_t < \infty$, *and every accumulation point is critical for* Ψ. *If, additionally,* $\|y^{t+1} - y^t\| \le c_s d_{t+1}$ *for some* $c_s > 0$ *and all sufficiently large* t, *then* $\{y^t\}$ *has finite length and converges to one point* y^* *with* $0 \in \partial\Psi(y^*)$.

Proof. Choose $b_{1,\ell}, b_{2,\ell} > 0$ such that $b_{2,\ell} > \max\{a_{3,\ell}, 0\}$ and $a_{2,\ell} + b_{2,\ell} < b_{1,\ell} < a_{1,\ell}$. Set $s := 1/(1-q)$, $Q(u,v) := \sum_\ell (b_{1,\ell} u_\ell^2 + b_{2,\ell} v_\ell^2)$, and $\mathcal{A}(y,u,v,\rho) := \Psi(y) + Q(u,v) + |\rho|^s$. Writing $\lambda_t = (\lambda_{1,t}, \dots, \lambda_{m,t})$, define $Y_t := (y^t, \lambda_t, \lambda_{t-1}, \mathcal{T}_t^{1/s})$ and $L_t := \mathcal{A}(Y_t)$. Coefficient separation, $\mathcal{T}_t - \mathcal{T}_{t+1} = \sigma_{t+1}$, and condition (ii) give

$$L_t - L_{t+1} \ge c d_t^2, \qquad \operatorname{dist}(0, \partial\mathcal{A}(Y_t)) \le C(d_t + \mathcal{T}_t^q) \tag{44}$$

for some $c, C > 0$ and all sufficiently large t. Since L_t is bounded below, it converges and $\sum_t d_t^2 < \infty$. Consequently, all memory terms and $\mathcal{T}_t$ vanish, and $\Psi(y^t)$ converges to the same finite limit Ψ_*. Lower semicontinuity and relative continuity imply that $\Psi = \Psi_*$ on the compact cluster set Ω of $\{y^t\}$. The relative-error estimate and closedness of the limiting subdifferential then give $0 \in \partial\Psi(\bar{y})$ for $\bar{y} \in \Omega$.

The function $\mathcal{A}$ is KL at every point of $\Omega \times \{0\}$. Indeed, for a positive total gap $H = h + Q(u,v) + |\rho|^s$, where $h := \Psi(y) - \Psi_*$, use the KL inequality of Ψ when $h \ge H/2$ and that of the added nonnegative terms otherwise. If φ is a local desingularizer of Ψ, one may use $2\varphi(H/2) + C_{\mathrm{aug}} H^{1-\theta_Q}$ for $\mathcal{A}$, with $C_{\mathrm{aug}} > 0$ sufficiently large and $\theta_Q := \max\{1/2, q\}$. Lemma A.5 therefore supplies one concave desingularizer $\widehat{\varphi}$ along the tail of $\{Y_t\}$. Set $\Gamma_t := L_t - \Psi_* \ge 0$. Whenever $\Gamma_t > 0$, (44) and concavity give

$$\widehat{\varphi}(\Gamma_t) - \widehat{\varphi}(\Gamma_{t+1}) \ge c' \frac{d_t^2}{d_t + \mathcal{T}_t^q} \ge c'(d_t - \mathcal{T}_t^q). \tag{45}$$

If $\Gamma_t = 0$, monotonicity and the first bound in (44) give $d_t = 0$. Summing (45) proves $\sum_t d_t < \infty$. The successive-iterate bound then gives finite length, and the limit is critical by the argument above. □

To make the KL-exponent transfer in the rate analysis explicit, we use the following local specialization of the block-separable KL-exponent calculus; see Li & Pong (2018, Theorem 3.3). We include a short proof to cover the present local setting and the endpoint case $\theta = 0$.

Lemma C.2 (KL exponent of a separable augmentation)**.** *Let* $\Psi : \mathbb{R}^N \to (-\infty, +\infty]$ *be proper and lower semicontinuous and have KL exponent* $\theta \in [0,1)$ *at* y^*. *Let* $Q(v) = \sum_{j=1}^{m_0} b_j v_j^2$ *with* $b_j > 0$, *and let* $s > 1$. *Then* $\widetilde{\Psi}(y,v,\rho) := \Psi(y) + Q(v) + |\rho|^s$ *has KL exponent* $\bar{\theta} = \max\{\theta, 1/2, 1 - 1/s\}$ *at* $(y^*, 0, 0)$. *Without the power term, the exponent is* $\max\{\theta, 1/2\}$.

Proof. Set $\widehat{Q}(v,\rho) := Q(v) + |\rho|^s$ and $\theta_Q := \max\{1/2, 1 - 1/s\}$. Comparing the largest nonzero summand with the sum gives $\|\nabla\widehat{Q}\| \ge c_Q \widehat{Q}^{\theta_Q}$ near the origin. Write $h := \Psi(y) - \Psi(y^*)$ and $H := h + \widehat{Q}(v,\rho)$. Choose $\varepsilon_0 \in (0,1)$ within the KL value neighborhood of Ψ, and take points sufficiently near $(y^*, 0, 0)$ with $0 < H < \varepsilon_0$. If $h \ge H/2$, then $0 < h \le H$ and the KL inequality for Ψ gives $\operatorname{dist}(0, \partial\Psi(y)) \ge c H^{\bar{\theta}}$. Otherwise, $\widehat{Q} > H/2$, so $\|\nabla\widehat{Q}\| \ge c H^{\bar{\theta}}$. The product rule $\partial\widetilde{\Psi}(y,v,\rho) = \partial\Psi(y) \times \{\nabla\widehat{Q}(v,\rho)\}$ proves the claim. Omitting ρ gives the tail-free case. □

The following result gives the deterministic rate consequence for polynomial tail perturbations.

Theorem C.3 (Deterministic KL rates with polynomial tails)**.** *Suppose the assumptions and successive-iterate bound of Theorem C.1 hold, and let* y^* *be its limiting point. Suppose* Ψ *has KL exponent* $\theta \in [0,1)$ *at* y^* *and* $0 \le \sigma_t \le C_\sigma t^{-r}$ *eventually, for some* $C_\sigma \ge 0$ *and* $r > 2$.

(i) If $C_\sigma = 0$ *and* $\theta \le 1/2$, *there exist* $C > 0$ *and* $\rho \in (0,1)$ *such that* $|\Psi(y^t) - \Psi(y^*)| + d_t + \|y^t - y^*\| \le C\rho^t$ *eventually.*

(ii) Otherwise, set $\bar{\theta} := \max\{\theta, r/[2(r-1)]\}$ *if* $C_\sigma > 0$*, and* $\bar{\theta} := \max\{\theta, 1/2\}$ *if* $C_\sigma = 0$*. There exist* $C, C' > 0$ *such that, for all sufficiently large* t*,*

$$|\Psi(y^t) - \Psi(y^*)| \le Ct^{-1/(2\bar{\theta}-1)} + CC_\sigma^{r/[2\bar{\theta}(r-1)]} t^{-r/(2\bar{\theta})}, \tag{46}$$

$$d_t \le Ct^{-1/(4\bar{\theta}-2)} + CC_\sigma^{r/[4\bar{\theta}(r-1)]} t^{-r/(4\bar{\theta})}, \tag{47}$$

and $\|y^t - y^*\| \le C't^{-\kappa}$*, where* $\kappa = (1-\bar{\theta})/(2\bar{\theta}-1)$ *if* $C_\sigma = 0$ *and* $\kappa = \min\{(1-\bar{\theta})/(2\bar{\theta}-1), r(1-\bar{\theta})/(2\bar{\theta}), (r-2)/2\}$ *otherwise.*

Constants are independent of t *but may depend on the fixed data and trajectory. Terms containing a positive power of* C_σ *vanish when* $C_\sigma = 0$*.*

Proof. Use the augmentation in the proof of Theorem C.1, choosing $q = r/[2(r-1)]$ when $C_\sigma > 0$ and omitting the tail variable when $C_\sigma = 0$. This is permissible since $\mathcal{T}_t \le CC_\sigma t^{1-r}$ and $q > 1/(r-1)$ in the first case. Set $\Gamma_t := L_t - \Psi(y^*)$. Lemma C.2 gives the stated exponent $\bar{\theta}$ of the augmented function. Thus, for all sufficiently large t, (44) and the KL inequality imply

$$\Gamma_t - \Gamma_{t+1} \ge a\Gamma_t^{2\bar{\theta}} - bC_\sigma^{r/(r-1)} t^{-r}, \qquad d_t^2 \le C(\Gamma_t - \Gamma_{t+1}), \tag{48}$$

where the perturbation term is absent when $C_\sigma = 0$. For $C_\sigma = 0$ and $\bar{\theta} = 1/2$, reducing a to a number in $(0,1)$ gives geometric decay of Γ_t and then of d_t. Since $|\Psi(y^t) - \Psi(y^*)| \le \Gamma_t + Cd_t^2 + \mathcal{T}_t$, summation of the successive-iterate bound proves Item (i), after enlarging the contraction factor if necessary.

For Item (ii), put $p := 2\bar{\theta} \in (1,2)$ and $K := C_\sigma^{r/(r-1)}$. A polynomial comparison gives $\Gamma_t \le Ct^{-1/(p-1)} + CK^{1/p}t^{-r/p}$. To verify it, set $w_t = At^{-1/(p-1)} + BK^{1/p}t^{-r/p}$. The mean-value theorem and Young's inequality $K^{1/p}t^{-r/p-1} \le \epsilon Kt^{-r} + C_\epsilon t^{-p/(p-1)}$ show that, for sufficiently large A, B, $w_t - w_{t+1} + bKt^{-r} \le aw_{t+1}^p$. As Γ_t is nonincreasing, (48) implies $\Gamma_{t+1} + a\Gamma_{t+1}^p \le \Gamma_t + bKt^{-r}$. Choose A also to ensure the initial comparison. Monotonicity of $u \mapsto u + au^p$ then proves $\Gamma_t \le w_t$ by induction. The second inequality in (48) yields (47). The preceding objective-gap bound yields (46), since $r/(2\bar{\theta}) \le r-1$ when $C_\sigma > 0$; for each fixed positive C_σ, the tail is absorbed in the stated bound.

For the iterate rate, take the power desingularizer $\widehat{\varphi}(u) = c_\varphi u^{1-\bar{\theta}}$ in (45) and sum from t onward to obtain

$$\sum_{j=t}^{\infty} d_j \le C\Gamma_t^{1-\bar{\theta}} + C\sum_{j=t}^{\infty} \mathcal{T}_j^q. \tag{49}$$

The last term is zero eventually when $C_\sigma = 0$, and otherwise is at most $CC_\sigma^q t^{-(r-2)/2}$. The bound for Γ_t and $\|y^t - y^*\| \le c_s \sum_{j=t}^{\infty} d_{j+1}$ now give the stated exponent κ. □

C.2 Constants and proof of the shifted-objective estimates

We now establish the shifted-objective estimates used in Theorem 4.5. The function Φ_0 and the endpoint pair z_t retain their main-text definitions. We first derive two single-step estimates and then accumulate them over a complete shuffled epoch.

Proposition C.4 (Constants for the shifted-objective estimate)**.** *Suppose Assumptions 2.1 and 3.1 hold and Algorithm 1 uses* $\gamma_i^t = 1$*. Choose* $\rho_1, \rho_2 > 0$ *and* $0 < \alpha \le \bar{\alpha} < \min\{C_\alpha, 1/\ell_f\}$*. Set* $a := 1/(2\alpha) - \ell_f/2 - \rho_2/2$*,* $b := 1/(4\alpha) + \ell_f^2\rho_1 + 2L_\Phi$*,* $\kappa_e := \rho_1 + 1/(2\rho_2)$*, and* $C_\Delta := \frac{C_1^2}{\alpha} + 2L_\Phi C_1^2 + \frac{C_1^2}{2\rho_1} + 2C_0\ell_R$*.*

(i) For $t \ge 0$ *and* $i = 0, \dots, n-1$*, the single-step bounds are*

$$\Phi(x_{i+1}^t) \le \Phi(x_i^t + \widehat{\eta}_i^t) + b\|\widehat{\eta}_i^t\|^2 + \rho_1\|e_i^t\|^2 + C_\Delta\Delta_t, \tag{50}$$

$$\Phi(x_i^t + \widehat{\eta}_i^t) \le \Phi(x_i^t) - a\|\widehat{\eta}_i^t\|^2 + \frac{1}{2\rho_2}\|e_i^t\|^2. \tag{51}$$

In particular, the first bound includes the step to $x_n^t = x_0^{t+1}$*.*

(ii) If, additionally,

$$0 < \bar{\alpha} < \min\left\{ C_\alpha, \frac{1}{2\ell_f + 2\rho_2 + 4\ell_f^2\rho_1 + 8L_\Phi + 16\kappa_e\nu} \right\}, \tag{52}$$

then, for every $t \geq 1$, (17) *holds with $\sigma_\eta := (a-b)/2$, $\sigma' := (a+b)/2$, $\sigma_S := (a-b)/4 - \kappa_e\nu$, and $\bar{C}_\Delta := C_\Delta + (a-b)C_1^2/2$. These constants satisfy $\sigma_\eta > 0$, $\sigma' > b$, and $\sigma_S > 0$.*

Proof. For Item (i), fix t, i and write $x := x_i^t$, $\eta := \eta_i^t$, $\widehat{\eta} := \widehat{\eta}_i^t$, $m := m_i^t$, and $e := e_i^t$. Taking $v = \widehat{\eta}$ in (37), adding the smoothness estimate for F, and using $\alpha < 1/\ell_f$ give

$$\Phi(x+\eta) \leq \Phi(x+\widehat{\eta}) + \langle \nabla F(x+\widehat{\eta}) - m, \eta - \widehat{\eta} \rangle + \frac{\|\widehat{\eta}\|^2 - \|\eta\|^2}{2\alpha} + 2C_0\ell_R\Delta_t.$$

Because $\eta - \widehat{\eta} \in T_x\mathcal{M}$ and $e = m - \operatorname{grad} F(x)$, Young's inequality bounds the inner product by $\ell_f^2\rho_1\|\widehat{\eta}\|^2 + \rho_1\|e\|^2 + C_1^2\Delta_t/(2\rho_1)$. Using $\|\widehat{\eta}\|^2 - \|\eta\|^2 \leq \frac{1}{2}\|\widehat{\eta}\|^2 + 2\|\widehat{\eta} - \eta\|^2$, $\|\eta\|^2 \leq 2\|\widehat{\eta}\|^2 + 2C_1^2\Delta_t$, and (33) yields (50) with the stated b and C_Δ. For (51), comparison with zero in the exact model gives $\langle m, \widehat{\eta}\rangle + \|\widehat{\eta}\|^2/(2\alpha) + R(x+\widehat{\eta}) \leq R(x)$. Add the smoothness estimate for $F(x+\widehat{\eta})$ and use $-\langle e, \widehat{\eta}\rangle \leq \|e\|^2/(2\rho_2) + (\rho_2/2)\|\widehat{\eta}\|^2$. This proves Item (i).

For Item (ii), (52) gives $a - b > 4\kappa_e\nu > 0$. Apply (50) to the last step of epoch $t-1$, using $x_n^{t-1} = x_0^t$, and combine it with (51) at $i = 0$. For $i = 1, \ldots, n-1$, combine (50) at $i-1$ with (51) at i. With $H_t := \sum_{i=0}^{n-1} \|\widehat{\eta}_i^t\|^2$, summation gives

$$\begin{aligned}\Phi_0(z_t) \leq \Phi_0(z_{t-1}) &- (a-b)H_t - b\|\widehat{\eta}_{n-1}^t\|^2 + b\|\widehat{\eta}_{n-1}^{t-1}\|^2 \\ &+ \kappa_e \sum_{i=0}^{n-1} \|e_i^t\|^2 + \rho_1\|e_{n-1}^{t-1}\|^2 + (n-1)C_\Delta\Delta_t + C_\Delta\Delta_{t-1},\end{aligned}$$

where the nonpositive term $-\rho_1\|e_{n-1}^t\|^2$ has been dropped. The calculation also holds for $n = 1$, with empty sums equal to zero. Bound the two error terms by Assumption 3.1 at t and $t-1$. In addition, Lemma B.1 gives $S_t^2 \leq 2H_t + 2nC_1^2\Delta_t$, and hence

$$-(a-b)H_t \leq -\frac{a-b}{2}H_t - \frac{a-b}{4}S_t^2 + \frac{n(a-b)C_1^2}{2}\Delta_t.$$

Substitution, the definitions of $\sigma_\eta, \sigma', \sigma_S$ and $\bar{C}_\Delta$, and $(n-1)C_\Delta \leq nC_\Delta$ give (17). Finally, $a - b > 4\kappa_e\nu$ implies $\sigma_\eta > 0$, $\sigma' > b$, and $\sigma_S > 0$. □

Proposition C.5 (Relative-error estimate)**.** *Under Assumption 2.1, the boundedness in* (36)*, and the local-step condition of Lemma B.2, there exists $c_0 > 0$ such that*

$$\operatorname{dist}(0, \partial\Phi_0(x_i^t, \widehat{\eta}_i^t)) \leq c_0\big(\|\widehat{\eta}_i^t\| + \|e_i^t\|\big) \tag{53}$$

for every $t \geq 0$ and $i = 0, \ldots, n-1$.

The base estimate (53) is established in Zhang (2026). Under Assumption 3.1, (11) gives $\|e_i^t\| \leq \sqrt{c_e}\sqrt{E_t} + \sqrt{\nu}S_t$. Taking $i = n-1$ in (53) therefore proves (18), for example with $c_\Phi = c_0 \max\{1, \sqrt{c_e}, \sqrt{\nu}\}$. Together with the preceding proof, this establishes Theorem 4.5.

C.3 Parameter conditions and proof of whole-sequence convergence

The next technical statement fixes the parameter choices cited in Theorem 4.6. It also records exactly which part of the tolerance rule is needed for convergence.

Theorem C.6 (Parameter conditions for KL convergence)**.** *Suppose Assumptions 2.1 and 3.1 hold, Algorithm 1 uses $\gamma_i^t = 1$, and the tolerance rule* (13) *is used, with $r > 2$ and $\kappa_1, \kappa_2 \geq 0$. Let $\rho_1, \rho_2 > 0$ and use the constants of Proposition C.4. Assume*

$$\begin{gathered}0 < \alpha \leq \bar{\alpha} < \min\left\{ C_\alpha, \frac{1}{2\ell_f + 2\rho_2 + 4\ell_f^2\rho_1 + 8L_\Phi + 16(\kappa_e + \rho_1)\nu} \right\}, \\ 0 \leq \kappa_1 < \frac{\sigma_S - \rho_1\nu}{n\bar{C}_\Delta + C_\Delta}.\end{gathered} \tag{54}$$

If Φ_0 is KL at every accumulation point of $\{z_t\}$, then (19) *and the whole-sequence convergence conclusion of Theorem 4.6 hold. For these two conclusions it is sufficient to require, eventually, $\Delta_t \le \kappa_1 S_{t-1}^2 + \kappa_2 (t-t_0)^{-r}$, in addition to $0 \le \Delta_t \le \bar{\Delta}$.*

Proof. The first condition in (54) implies (52) and $\sigma_S > \rho_1 \nu$. For all sufficiently large t, substituting the tolerance upper bound into (17) gives

$$\begin{aligned}\Phi_0(z_t) \le \Phi_0(z_{t-1}) &- \sigma_\eta \sum_{i=0}^{n-2} \|\widehat{\eta}_i^t\|^2 - \sigma' \|\widehat{\eta}_{n-1}^t\|^2 + b \|\widehat{\eta}_{n-1}^{t-1}\|^2 \\ &- \kappa_e c_e E_{t+1} + (\kappa_e c_e - \kappa_e \mu - \rho_1 c_e) E_t + (\rho_1 c_e - \rho_1 \mu) E_{t-1} \\ &- \sigma_S S_t^2 + (\rho_1 \nu + n \bar{C}_\Delta \kappa_1) S_{t-1}^2 + C_\Delta \kappa_1 S_{t-2}^2 + \sigma_t,\end{aligned} \tag{55}$$

where one may take $\sigma_t = \kappa_2 [n \bar{C}_\Delta (t-t_0)^{-r} + C_\Delta (t-1-t_0)^{-r}]$ for these sufficiently large indices. In particular, $0 \le \sigma_t \le C_\sigma \kappa_2 t^{-r}$ for a fixed $C_\sigma > 0$; the finitely many earlier tail terms can be defined arbitrarily as finite nonnegative numbers.

Apply Theorem C.1 with $y^t = z_t$, $\Psi = \Phi_0$, and the following four scalar memory sequences and coefficients:

ℓ	$\lambda_{\ell,t}$	$a_{1,\ell}$	$a_{2,\ell}$	$a_{3,\ell}$
1	S_t	σ_S	$\rho_1 \nu + n \bar{C}_\Delta \kappa_1$	$C_\Delta \kappa_1$
2	$\sqrt{E_{t+1}}$	$\kappa_e c_e$	$\kappa_e c_e - \kappa_e \mu - \rho_1 c_e$	$\rho_1 (c_e - \mu)$
3	$\|\widehat{\eta}_{n-1}^t\|$	σ'	b	0
4	$\left(\sum_{i=0}^{n-2} \|\widehat{\eta}_i^t\|^2\right)^{1/2}$	σ_η	0	0

The descent at time t in (55) is precisely (43) at time $t-1$. The four coefficient-separation margins are, respectively, $\sigma_S - \rho_1 \nu - (n \bar{C}_\Delta + C_\Delta) \kappa_1$, $\kappa_e \mu + \rho_1 \min\{c_e, \mu\}$, $\sigma' - b$, and σ_η, all strictly positive.

The relative-error quantity d_t in that theorem includes S_t, $\|\widehat{\eta}_{n-1}^t\|$, and $\sqrt{E_t}$, the last through $\lambda_{2,t-1}$. Thus (18) gives the required relative-error bound with this indexing. Compactness of $\mathcal{M}$ and Lemma B.2 imply boundedness of $\{z_t\}$. Its closure is a compact subset of $T\mathcal{M}$, where Φ_0 is continuous and bounded below. If $\kappa_2 > 0$, the polynomial tail satisfies $\sum_t \mathcal{T}_t^q < \infty$ for every $q \in (1/(r-1), 1)$; if $\kappa_2 = 0$, the tail vanishes eventually. All hypotheses of Theorem C.1 are therefore verified.

It follows that all four memory sequences are summable. Since $\sum_i \|\widehat{\eta}_i^t\| \le \sqrt{n-1} \lambda_{4,t} + \lambda_{3,t}$, this implies (19), with the first term zero when $n = 1$. Moreover, $\sum_{t=1}^{\infty} \sum_{i=0}^{n-1} \|x_{i+1}^t - x_i^t\| \le C_R \sqrt{n} \sum_{t=1}^{\infty} S_t < \infty$. Together with $x_n^t = x_0^{t+1}$, this proves that all within-epoch sequences converge to the same point $x^* \in \mathcal{M}$. Finally, $\widehat{\eta}_i^t \to 0$ and $e_i^t \to 0$. The exact-model optimality and closed-graph argument in (42) then show that x^* satisfies (2). □

C.4 Proof of Theorem 4.7

Use the four memory sequences just specified and the corresponding d_t. For $z_t = (x_{n-1}^t, \widehat{\eta}_{n-1}^t)$, the retraction bound and the epoch boundary identity give

$$\|z_{t+1} - z_t\| \le C_R \sqrt{n} (S_{t+1} + S_t) + \|\widehat{\eta}_{n-1}^{t+1}\| + \|\widehat{\eta}_{n-1}^t\| \le C_s d_{t+1} \tag{56}$$

for a fixed $C_s > 0$. Relative continuity of Φ_0 was verified above, and Φ_0 has the assumed exponent θ at $z^* = (x^*, 0)$. Theorem C.3 applies with perturbation coefficient zero when $\kappa_2 = 0$, and with a fixed multiple of κ_2 otherwise. This gives the lifted objective rates in (20) and (21).

The quantity $\mathcal{R}_t$ of Theorem 4.7 obeys $\mathcal{R}_t \le \sqrt{n}\, d_t$: the terms involving the exact steps are bounded by $\sqrt{n}(\lambda_{3,t} + \lambda_{4,t})$, and d_t also contains S_t and $\sqrt{E_t}$. The rate for d_t therefore proves the corresponding rates for $\mathcal{R}_t$, including (22). For every $i = 0, \dots, n-1$, $\|x_i^t - x^*\| \le \|x_{n-1}^t - x^*\| + C_R \sqrt{n} S_t$. When $\kappa_2 = 0$ and $\theta > 1/2$, $\kappa \le 1/(4\bar{\theta} - 2)$; when $\kappa_2 > 0$, also $\kappa \le r/(4\bar{\theta})$. Thus every nonzero term in the bound on S_t decays at least as fast as $t^{-\kappa}$, which transfers the rate to all x_i^t. In the unperturbed linear case, enlarging the contraction factor within $(0, 1)$ if necessary gives a common linear rate. This proves Theorem 4.7.

D PROOFS OF THE COMPLEXITY BOUNDS

Throughout this appendix, Θ is the measure in (15), and we abbreviate $\Theta_i^t := \Theta(x_i^t, \widehat{\eta}_i^t)$. The bounds below record dependence on the target accuracy for fixed n.

D.1 CRITICALITY AND CUMULATIVE ESTIMATES

Proposition D.1 (Pointwise certificate and cumulative bound)**.** *Suppose Assumptions 2.1 and 3.1 hold, let the iterates be generated by Algorithm 1, and assume* $0 < \alpha \le \bar{\alpha} \le C_\alpha$.

(i) The exact tangent model has the pointwise certificate

$$\operatorname{dist}\Big(0, \nabla F(x_i^t + \widehat{\eta}_i^t) + \partial R(x_i^t + \widehat{\eta}_i^t) + N_{x_i^t}\mathcal{M}\Big) \le (\ell_f + \alpha^{-1})\|\widehat{\eta}_i^t\| + \|e_i^t\|. \tag{57}$$

(ii) With C_Θ *as in* (16)*, for every integer* $T \ge 1$*,*

$$\sum_{t=0}^{T-1}\sum_{i=0}^{n-1}(\Theta_i^t)^2 \le C_\Theta \sum_{t=0}^{T-1}\sum_{i=0}^{n-1}\|\widehat{\eta}_i^t\|^2 + 2\nu\sum_{t=0}^{T-1} S_t^2. \tag{58}$$

Proof. Exact model optimality gives $m_i^t + \alpha^{-1}\widehat{\eta}_i^t + \xi_i^t + \zeta_i^t = 0$ for some $\xi_i^t \in \partial R(x_i^t + \widehat{\eta}_i^t)$ and $\zeta_i^t \in N_{x_i^t}\mathcal{M}$. Since $\operatorname{grad} F(x_i^t) - \nabla F(x_i^t) \in N_{x_i^t}\mathcal{M}$, the vector $\nabla F(x_i^t + \widehat{\eta}_i^t) - \nabla F(x_i^t) - e_i^t - \alpha^{-1}\widehat{\eta}_i^t$ belongs to the set in (57). Lipschitz continuity of ∇F proves Item (i). For Item (ii), squaring (57), using $(a+b)^2 \le 2a^2 + 2b^2$, and adding $\|\widehat{\eta}_i^t\|^2$ proves (16) with $C_\Theta = 1 + 2(\ell_f + \alpha^{-1})^2$. Summing (11), using $E_0 = 0$, and dropping the nonpositive memory terms gives $\sum_{t=0}^{T-1}\sum_i \|e_i^t\|^2 \le \nu\sum_{t=0}^{T-1} S_t^2$. Combining these bounds proves (58). □

D.2 OUTER COMPLEXITY AND THE KL BEST-ITERATE BOUND

For the tolerance rule (12), take $\kappa \ge 0$ and a deterministic sequence $0 \le \varsigma_t \le \bar{\Delta}$. Then $\varsigma_t \le \Delta_t \le \kappa S_{t-1}^2 + \varsigma_t$. The upper bound permits absorption of the history term; the lower bound is used separately to count inner work.

Theorem D.2 (Outer complexity with specified constants)**.** *Suppose Assumptions 2.1 and 3.1 hold and Algorithm 1 uses* (12)*, with the floor range above and* $S_{-1} = 0$*. Let* $a_* > 1$ *and assume*

$$0 < \alpha \le \bar{\alpha} < \min\left\{C_\alpha, \frac{1}{L_F + 2L_R + 2a_*\sqrt{\nu/\underline{\gamma}}}\right\}. \tag{59}$$

Set $\tau := \sqrt{\nu/\underline{\gamma}}$*,* $\sigma_0 := 1/(2\alpha) - L_F/2 - L_R$*,* $\widehat{\sigma} := \underline{\gamma}(\sigma_0 - \tau/2) - \nu/(2\tau)$*,* $c_\Delta := 2\ell_R C_0$*, and assume* $a_S := \widehat{\sigma} - c_\Delta n\kappa > 0$*. Define* $G_0 := \Phi(\widetilde{x}^0) - \inf_{x\in\mathcal{M}}\Phi(x)$ *and*

$$C_{\mathrm{comp}} := \left(\frac{2\alpha C_\Theta}{\underline{\gamma}} + \frac{2\nu}{a_S}\right)\max\{1, c_\Delta\}. \tag{60}$$

Then the following statements hold.

(i) For every $T \ge 1$*,*

$$\sum_{t=0}^{T-1}\sum_{i=0}^{n-1}(\Theta_i^t)^2 \le C_{\mathrm{comp}}\left(G_0 + n\sum_{t=0}^{T-1}\varsigma_t\right). \tag{61}$$

(ii) Set $C_{\mathrm{tol}} := 2C_{\mathrm{comp}}$ *and* $A_0 := \max\{1, C_{\mathrm{comp}}G_0/n\}$*. Given* $\varepsilon > 0$ *with* $\varepsilon^2/C_{\mathrm{tol}} \le \bar{\Delta}$*, set* $T_\varepsilon := \max\{1, \lceil 2A_0\varepsilon^{-2}\rceil\}$ *and* $\varsigma_t := \varepsilon^2/C_{\mathrm{tol}}$ *for* $0 \le t < T_\varepsilon$*. Then* $\min_{0\le t<T_\varepsilon,\, 0\le i<n}\Theta_i^t \le \varepsilon$*. No tolerance specification after this stopping horizon is needed for the stated complexity bound.*

Proof. Let $\mathcal{L}_t = \Phi(\widetilde{x}^t) + c_e E_t/(2\tau)$ as in the main text. The stepsize bound makes $\sigma_0 > \tau$ and $\widehat{\sigma} > 0$. With this τ, Theorem B.3 gives the descent coefficients $\widehat{\sigma}$, $\underline{\gamma}/(2\alpha)$, and $\mu/(2\tau)$ for S_t^2,

$\sum_i \|\widehat{\eta}_i^t\|^2$, and E_t, respectively. Use $\Delta_t \le \kappa S_{t-1}^2 + \varsigma_t$, sum to $T-1$, and shift the history term using $S_{-1} = 0$. Dropping a nonnegative terminal term yields

$$a_S \sum_{t=0}^{T-1} S_t^2 + \frac{\gamma}{2\alpha} \sum_{t=0}^{T-1} \sum_{i=0}^{n-1} \|\widehat{\eta}_i^t\|^2 + \frac{\mu}{2\tau} \sum_{t=0}^{T-1} E_t \le G_0 + c_\Delta n \sum_{t=0}^{T-1} \varsigma_t. \tag{62}$$

Here $\mathcal{L}_0 = \Phi(\widetilde{x}^0)$ and $\mathcal{L}_T \ge \inf_{\mathcal{M}} \Phi$. Combining (62) with (58) proves Item (i) for (60). Dividing by nT_ε gives

$$\min_{\substack{0 \le t < T_\varepsilon \\ 0 \le i < n}} (\Theta_i^t)^2 \le \frac{C_{\mathrm{comp}} G_0}{nT_\varepsilon} + \frac{C_{\mathrm{comp}}}{C_{\mathrm{tol}}} \varepsilon^2 \le \varepsilon^2.$$

This proves Item (ii). Since n is fixed, the horizon has order $O(\varepsilon^{-2})$ both in epochs and in retractions. □

The constant G_0 is the initial objective gap. A known lower bound $m_\Phi \le \inf_{\mathcal{M}} \Phi$ gives the conservative replacement $G_0 = \Phi(\widetilde{x}^0) - m_\Phi$; computing the global minimum is not required by Algorithm 1.

Corollary D.3 (KL best-iterate complexity)**.** *Under the conditions of Theorem C.6, there is a finite $C_{\mathrm{KL}} > 0$, independent of T and ε, such that $\min_{1 \le t \le T,\, 0 \le i < n} \Theta_i^t \le C_{\mathrm{KL}}/T$ for all $T \ge 1$. The constant may depend on the fixed trajectory.*

Proof. By (57) and (11), $\Theta_i^t \le (1 + \ell_f + \alpha^{-1}) \|\widehat{\eta}_i^t\| + \sqrt{c_e}\sqrt{E_t} + \sqrt{\nu} S_t$. Theorem C.6 gives $\sum_{t=1}^{\infty} \sum_{i=0}^{n-1} \Theta_i^t < \infty$. Averaging over nT terms proves the claim, after increasing the constant if necessary. This proves the best-iterate part of Theorem 4.6; no specified KL exponent is required. □

D.3 Inner solvers and total proximal work

We state the single-subproblem bounds under the initialization and regularization conventions used to count proximal evaluations. The underlying smooth convex optimization results are cited below.

Lemma D.4 (Uniform dual data and residual complexity)**.** *Under Assumption 2.1 and the local-step conditions above, G_i^t is convex and has gradient Lipschitz constant at most $L_D := \bar{\alpha} \widetilde{C}^2$. It has a minimizer $\lambda_i^{t,*}$ satisfying $\|\lambda_i^{t,*}\| \le \ell_R / C_B$. Initialize each dual solve at zero and let $0 < \Delta \le \bar{\Delta}$. The following bounds on proximal evaluations of R suffice to produce λ with $\|\nabla G_i^t(\lambda)\| \le \Delta$:*

(i) Smooth FISTA applied to G_i^t: $K_{\mathrm{FISTA}}(\Delta) = O(1 + \Delta^{-1})$.

(ii) NFG applied to the quadratically regularized dual problem, with regularization proportional to Δ: $K_{\mathrm{NFG}}(\Delta) = O(1 + \Delta^{-1/2} \log(1 + \Delta^{-1}))$.

(iii) Accumulative regularization (AR): $K_{\mathrm{AR}}(\Delta) = O(1 + \Delta^{-1/2})$.

The constants are uniform over the outer iterates and the prescribed tolerances, with the problem and algorithmic parameters fixed.

The uniform dual data follow from strong convexity of F_i^t: $\nabla F_i^{t,*}$ is α-Lipschitz, so the gradient Lipschitz constant of G_i^t is at most $\alpha \|B_{x_i^t}\|^2$. Exact model optimality supplies a multiplier with $B_{x_i^t}^\top \lambda_i^{t,*} = -\mathrm{P}_{N_{x_i^t}\mathcal{M}} \xi_i^t$ for some $\xi_i^t \in \partial R(x_i^t + \widehat{\eta}_i^t)$. Thus $\|\lambda_i^{t,*}\| \le \ell_R / C_B$. FISTA has zero nonsmooth term in the dual space (Beck & Teboulle, 2009; Nesterov, 2018); the NFG and AR implementations are the regularized-gradient procedures described in Jiang et al. (2025, Section 5.2), with the AR estimate based on Lan et al. (2026). A dual gradient evaluation uses one $\mathrm{prox}_{\alpha R}$ evaluation through (7). Residual checks change this count by at most a fixed factor. The zero initialization and the uniform multiplier bound in the lemma supply the common initial-distance bound; an arbitrary unbounded warm start is not covered by this count.

Completion of Theorem 4.3. For $t < T_\varepsilon$, $\Delta_t \ge \varsigma_t = \varepsilon^2 / C_{\mathrm{tol}} > 0$. Lemma D.4 therefore bounds the work of each tangent subproblem by $O(\varepsilon^{-2})$ for FISTA, $O(\varepsilon^{-1} \log \varepsilon^{-1})$ for NFG, and $O(\varepsilon^{-1})$ for AR. There are nT_ε such subproblems, with $T_\varepsilon = O(\varepsilon^{-2})$ by Theorem D.2. Multiplication gives the three proximal-evaluation orders in Theorem 4.3. □

KL proximal-evaluation bounds in Theorem 4.6. Suppose $\bar{\Delta} > 0$, $\kappa_2 > 0$, and the equality rule (13) is used. For $0 \le t \le t_0 + 1$, take fixed positive tolerances independent of ε. The initial solves then contribute a fixed finite amount of work. For all sufficiently large t, $\kappa_2(t - t_0)^{-r} \le \bar{\Delta}$, and hence the capped maximum satisfies $\Delta_t \ge \kappa_2(t - t_0)^{-r}$. Corollary D.3 gives a trajectorywise horizon $T_\varepsilon = O(\varepsilon^{-1})$. The three total inner-work bounds are thus, respectively, $O(\sum_{t=1}^{T_\varepsilon} t^r)$, $O(\sum_{t=1}^{T_\varepsilon} t^{r/2} \log(1 + t))$, and $O(\sum_{t=1}^{T_\varepsilon} t^{r/2})$, with finite prefixes absorbed into the constants. Substituting the horizon yields $O(\varepsilon^{-(r+1)})$, $O(\varepsilon^{-(1+r/2)} \log \varepsilon^{-1})$, and $O(\varepsilon^{-(1+r/2)})$. The lower bound on Δ_t is specific to the equality rule; a mere upper bound on the tolerance suffices for convergence but not for this total-work conclusion. □

E Estimator formulas and proofs of their verification

We provide the formulas referenced by Theorem 4.10 and prove the corresponding verification results. The errors e_i^t, δ_i^t, and the epoch quantity S_t keep their main-text meanings. All bounds hold for each fixed sequence of permutations.

E.1 Proof of Lemma 4.9

Proof. Uniform boundedness of u_i^t implies (36). For $i \ge 1$, the momentum recursion, projection nonexpansiveness, and Lemma 2.2, applied to $\gamma_{i-1}^t \eta_{i-1}^t$, give

$$\|e_i^t\| \le \chi_i^t \|e_{i-1}^t\| + \chi_i^t L_T \gamma_{i-1}^t \|\eta_{i-1}^t\| + (1 - \chi_i^t)\delta_i^t.$$

Using (24), $\gamma_{i-1}^t \le 1$, and $\|\eta_{i-1}^t\| \le S_t$, iteration yields $\|e_i^t\| \le \bar{\chi}^i \|e_0^t\| + C'(S_t + \sqrt{M_t})$, where $C' := (\bar{\chi} L_T + C_u)/(1 - \bar{\chi})$. With $\rho = 1$ and $E_t = \|e_0^t\|^2 + M_t$, squaring and summing give $\sum_{i=0}^{n-1} \|e_i^t\|^2 \le C_{\mathrm{lift}}(E_t + S_t^2)$ with $C_{\mathrm{lift}} := 3n \max\{1, (C')^2\}$.

Since $\chi_0^{t+1} = 0$, $e_0^{t+1} = u_0^{t+1} - \operatorname{grad} F(x_0^{t+1})$. Hence $\|e_0^{t+1}\| \le C_u \sqrt{M_{t+1}}$ and $E_{t+1} \le C_{\mathrm{next}} S_t^2$, where one may take $C_{\mathrm{next}} := \max\{1, (C_u^2 + 1)C_M\}$. If $c_e > C_{\mathrm{lift}}$ and $0 < \mu < c_e - C_{\mathrm{lift}}$, then $C_{\mathrm{lift}} + \mu < c_e$, so the preceding estimates imply (11) for every $\nu \ge C_{\mathrm{lift}} + c_e C_{\mathrm{next}}$. Finally, $e_0^0 = M_0 = 0$ gives $E_0 = 0$. □

E.2 Proof of Theorem 4.10

Proof. We first consider projection-based full-pass variants of SVRG and SARAH/SPIDER (Johnson & Zhang, 2013; Nguyen et al., 2017; Fang et al., 2018). For projected SVRG, compute $\operatorname{grad} F(x_0^t)$ at the beginning of epoch t and use

$$u_i^t := \operatorname{grad} f_{\pi_t(i+1)}(x_i^t) - \mathrm{P}_{T_{x_i^t}\mathcal{M}}\big(\operatorname{grad} f_{\pi_t(i+1)}(x_0^t) - \operatorname{grad} F(x_0^t)\big), \qquad i = 0, \dots, n-1. \tag{63}$$

For projected SARAH/SPIDER, set

$$u_0^t := \operatorname{grad} F(x_0^t), \tag{64}$$

and, for $i = 1, \dots, n-1$, define

$$u_i^t := \operatorname{grad} f_{\pi_t(i+1)}(x_i^t) - \mathrm{P}_{T_{x_i^t}\mathcal{M}}\big(\operatorname{grad} f_{\pi_t(i+1)}(x_{i-1}^t) - u_{i-1}^t\big). \tag{65}$$

The bound M_f from Appendix A.2 gives $\|u_i^t\| \le 3M_f$ for SVRG. For SARAH/SPIDER, $\|u_0^t\| \le M_f$ and $\|u_i^t\| \le \|u_{i-1}^t\| + 2M_f$, so $\|u_i^t\| \le (2n-1)M_f$. Thus M_u can be fixed before choosing $\bar{\alpha} \le C_\alpha$ in Lemma B.2.

For SVRG, adding and subtracting $\mathrm{P}_{T_{x_i^t}\mathcal{M}} \operatorname{grad} F(x_0^t)$ and applying (4) to the sampled component and to F give $\delta_i^t \le 2L_T\sqrt{n}\, S_t$, with $\delta_0^t = 0$. For SARAH/SPIDER, (3) gives $\delta_i^t \le \delta_{i-1}^t + 2L_T \gamma_{i-1}^t \|\eta_{i-1}^t\|$. Since $\delta_0^t = 0$, summation yields the same bound. Hence (24) holds for both methods with $M_t \equiv 0$, $C_u = 2L_T\sqrt{n}$, and $C_M = 0$. Their initial estimators are exact, so $e_0^0 = M_0 = 0$.

We next consider projected ambient, frozen-table full-pass variants of SAGA and SAG (Defazio et al., 2014; Schmidt et al., 2017). For each component, let $z_j^t \in \mathcal{M}$ be the stored point at the beginning of

epoch t and set $y_j^t := \nabla f_j(z_j^t)$. The table is kept fixed during the epoch. Projected ambient SAGA uses

$$v_i^t := \nabla f_{\pi_t(i+1)}(x_i^t) - y_{\pi_t(i+1)}^t + \frac{1}{n}\sum_{j=1}^n y_j^t, \qquad u_i^t := \mathrm{P}_{T_{x_i^t}\mathcal{M}} v_i^t, \tag{66}$$

whereas projected ambient SAG uses

$$v_i^t := \frac{1}{n}\big(\nabla f_{\pi_t(i+1)}(x_i^t) - y_{\pi_t(i+1)}^t\big) + \frac{1}{n}\sum_{j=1}^n y_j^t, \qquad u_i^t := \mathrm{P}_{T_{x_i^t}\mathcal{M}} v_i^t. \tag{67}$$

At the end of epoch t, set $z_{\pi_t(i+1)}^{t+1} := x_i^t$ and $y_{\pi_t(i+1)}^{t+1} := \nabla f_{\pi_t(i+1)}(x_i^t)$ for $i = 0, \dots, n-1$. The initialization is $z_j^0 = x_0^0$ and $y_j^0 = \nabla f_j(x_0^0)$.

Since all stored points belong to $\mathcal{M}$, $\|u_i^t\| \le 3M_f$ for SAGA. For SAG, v_i^t is an average of one current component gradient and $n-1$ stored gradients, so $\|u_i^t\| \le M_f$. Thus M_u can again be fixed before applying Lemma B.2.

Set $M_t := \sum_{j=1}^n \|y_j^t - \nabla f_j(x_0^t)\|^2$ and $d_{j,i}^t := \nabla f_j(x_i^t) - y_j^t$. For $j_i := \pi_t(i+1)$, projection nonexpansiveness gives $\delta_i^t \le \|d_{j_i,i}^t - n^{-1}\sum_j d_{j,i}^t\|$ for SAGA and $\delta_i^t \le \|n^{-1}d_{j_i,i}^t - n^{-1}\sum_j d_{j,i}^t\|$ for SAG. The ambient-gradient and retraction bounds imply

$$\sum_{j=1}^n \|d_{j,i}^t\|^2 \le 2M_t + 2n^2\ell_f^2 C_R^2 S_t^2 \le C_{\mathrm{tab}}^2(M_t + S_t^2). \tag{68}$$

Consequently, $\delta_i^t \le (1+n^{-1/2})C_{\mathrm{tab}}(S_t+\sqrt{M_t})$ for SAGA and $\delta_i^t \le (n^{-1}+n^{-1/2})C_{\mathrm{tab}}(S_t+\sqrt{M_t})$ for SAG. At $i=0$, the same constants give $\delta_0^t \le C_u\sqrt{M_t}$.

For $j = \pi_t(i+1)$, the end-of-epoch refresh and $x_0^{t+1} = x_n^t$ give $\|y_j^{t+1} - \nabla f_j(x_0^{t+1})\| \le \ell_f C_R\sqrt{n}\, S_t$. Since every component is visited exactly once, $M_{t+1} \le n^2\ell_f^2 C_R^2 S_t^2 = C_M^{\mathrm{tab}} S_t^2$. Thus (24) holds with the constants stated in Theorem 4.10. The initialization gives $M_0 = 0$ and $u_0^0 = \operatorname{grad} F(x_0^0)$, hence $e_0^0 = 0$.

We have therefore verified (24) for all four estimators with the choices in Theorem 4.10. Applying Lemma 4.9 and choosing c_e, μ, ν as in (27) proves Assumption 3.1 with $E_0 = 0$ for all four full-pass realizations. □