axioms

MDPI

Article

A Safeguarded Transport-Extrapolated Proximal Gradient Method on the Stiefel Manifold

Chuanjun Yan [1] and Lin Chen [2,*]

[1] School of Mathematical Sciences, Chongqing Normal University, Chongqing 401331, China
[2] National Center for Applied Mathematics in Chongqing, Chongqing Normal University, Chongqing 401331, China
* Correspondence: allenchen@cqnu.edu.cn

Abstract

Motivated by sparse principal component analysis (PCA), this paper proposes ATE-ManPG, a safeguarded transport-extrapolated proximal gradient method for minimizing the sum of a smooth function and a convex, possibly nonsmooth function over the Stiefel manifold. The method augments the manifold proximal gradient method (ManPG) with two ingredients: momentum obtained by transporting the previous proximal search direction to the current tangent space, protected by a monotone restart safeguard, and a bounded adaptive proximal parameter. For the exact idealized iteration, global convergence to stationary points and $O(1/\varepsilon^2)$ iteration complexity are proved; under a proximal Kurdyka–Łojasiewicz condition with exponent $1/2$ and a relative-error condition, both stated as explicit hypotheses, the convergence is R-linear. Experiments on a sparse-PCA model, validated by an external residual certificate at the returned point, show a $3.6\times$ mean iteration-count reduction relative to fixed-parameter ManPG; a factorial ablation attributes most of this reduction to the adaptive parameter, while transported momentum adds a further 6.4% on average, concentrated in one paired trial.

Keywords: Stiefel manifold; proximal gradient method; Nesterov extrapolation; Kurdyka–Łojasiewicz inequality; adaptive stepsize; iteration complexity

MSC: 90C30; 90C26; 65K05

1. Introduction

The Stiefel manifold $\mathrm{St}(n,r) = \{X \in \mathbb{R}^{n\times r} : X^\top X = I_r\}$ captures orthogonality constraints essential to sparse principal component analysis (PCA) [1], compressed modes [2], feature selection [3], and blind deconvolution [4]. This paper considers the composite problem

$$\min_{X\in\mathrm{St}(n,r)} F(X) := f(X) + h(X), \tag{1}$$

where f is smooth with L-Lipschitz gradient and h is convex and possibly nonsmooth.

Chen, Ma, So, and Zhang [5] proposed the manifold proximal gradient method (ManPG), a retraction-based algorithm that solves a strongly convex subproblem in the tangent space at each iteration via a semismooth Newton (SSN) method. ManPG converges globally to stationary points with $O(1/\varepsilon^2)$ iteration complexity and was selected for republication in the SIGEST section of *SIAM Review* in 2024 [6], testifying to its practical impact and its role in unifying the theory of nonsmooth Riemannian optimization.

Check for updates

Academic Editor: Giovanni Calvaruso

Received: 28 September 2026
Revised: 25 October 2026
Accepted: 27 October 2026
Published: 5 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1.1. Motivation and Related Work

Two lines of research motivate the present work: acceleration on manifolds and convergence under geometric conditions.

Accelerated proximal gradient on manifolds. Huang and Wei [7] extended the fast iterative shrinkage-thresholding algorithm (FISTA) [8,9] to the Stiefel manifold via inverse retraction and a periodic safeguard; their Riemannian proximal gradient (RPG) framework [10] gives $O(1/k)$ rates under retraction-convexity and local KL rates, but leaves the vanilla accelerated variant (V-ARPG) unanalyzed. Accelerated Riemannian proximal rates under curvature/convexity assumptions appear in Feng et al. [11]. Related Euclidean theory includes nonconvex acceleration [12,13] and adaptive restart [14]. Other Stiefel tools include Riemannian proximal Newton [15], inexact RPG [16], and recent inertial methods [17]. Overviews and orthogonality-constrained first-order frameworks are [18,19].

Penalty and constraint-handling on Stiefel manifolds. Smooth exact penalties for orthogonality constraints and $\ell_{2,1}$ problems [20–22], a penalty-free infeasible method with tangential and normal steps [23], constraint dissolving [24,25], and multiplier/parallel schemes [26,27] form alternative families; they are not drop-in replacements of ManPG tangent subproblems.

Adaptive ManPG. Chen et al. [5] already propose ManPG-Ada (their Algorithm 2). Adaptive proximal parameters are therefore not new here; we focus on transported-direction extrapolation with a monotone safeguard and on when a KL rate is preserved.

KL-based local linear convergence. The Kurdyka–Łojasiewicz (KL) inequality [28–30] yields linear rates when $\theta = 1/2$ and a relative-error bound holds. Liu, So, and Wu [31] prove $\theta = 1/2$ for smooth quadratics on $\mathrm{St}(n, r)$. Wang, Liu, and So [32] study robust ℓ_1-PCA ($\max \|D^\top Q\|_1$), not the penalized variance objective. Huang and Wei [10] obtain a Riemannian KL property with unspecified exponent. Semialgebraicity alone does not prescribe $\theta = 1/2$ for $-\mathrm{Tr}(X^\top A^\top A X) + \mu\|X\|_1$; see Remark 1.

1.2. Our Contributions

This paper studies transported-direction extrapolation of the ManPG search direction with monotone restart and a bounded adaptive proximal parameter. Contributions:

1. **Algorithm.** ATE-ManPG (adaptive transport-extrapolated ManPG, Algorithm 1): orthogonal-projection transport (no inverse retraction), bounded $\beta_k \le \beta_{\max}$, objective-based restart, adaptive $t \in [1/L, 10/L]$, and return of the work point where the residual is certified. One extra retraction and function evaluation per iteration relative to ManPG.
2. **Theory.** Sufficient decrease (Lemma 1), global subsequential stationarity (Theorem 1), $O(\varepsilon^{-2})$ iteration complexity (Theorem 2), and a *conditional* R-linear rate (Theorem 3) via a standard KL argument, uniform in adaptive $t \in [1/L, 10/L]$. We do not claim $\theta = 1/2$ for every sparse-PCA stationary point (Remark 1); relative error is verified under smoothness or sign compatibility (Proposition 1).
3. **Experiments.** ATE-ManPG is compared with ManPG, ManPG-Ada, TE-ManPG, AManPG, P-ARPG, and RPG on sparse PCA from archived scalar records with an external certificate. A harmonized factorial ablation shows adaptation explains most of the total speedup; transported momentum adds 6.4% mean iterations on top of adaptation (concentrated in one paired trial), with runtime overhead and occasional quality tradeoffs. Cap sensitivity covers $\beta_{\max} \in [0.2, 1]$.

https://doi.org/10.3390/axioms15110828

Algorithm 1 ATE-ManPG for solving (1)

Require: $X_0 \in \mathrm{St}(n,r)$, $t_0 = 1/L$, $t = t_0$, $t_{\max} = 10/L$, $\varepsilon_{\text{stop}} > 0$ (tolerance on $\|V_k\|_{\mathrm{F}}$), $\beta_{\max} > 0$ (default 0.5), optional floor $\alpha_{\text{tol}} \geq 0$

$D_{\text{prev}} \leftarrow 0, \quad \alpha_{\text{prev}} \leftarrow 1$
for $k = 0, 1, 2, \ldots$ **do**
  if $k = 0$ or $D_{\text{prev}} = 0$ **then**
    $\beta_k \leftarrow 0, \quad X_{\text{work}} \leftarrow X_k$
  else
    $\tilde{D} \leftarrow \mathrm{Proj}_{T_{X_k}\mathrm{St}(n,r)}(\alpha_{\text{prev}} D_{\text{prev}})$
    $\beta_k \leftarrow \min\left(\frac{k}{k+3}, \beta_{\max}\right)$ ▷ matches the archived implementation
    $X_{\text{work}} \leftarrow \mathrm{Retr}_{X_k}(\beta_k \tilde{D})$
    if $F(X_{\text{work}}) > F(X_k)$ **then** ▷ monotone restart safeguard
      $X_{\text{work}} \leftarrow X_k, \quad D_{\text{prev}} \leftarrow 0, \quad \beta_k \leftarrow 0$
    end if
  end if
  Compute $\nabla f(X_{\text{work}})$
  Solve (7) at X_{work} to obtain V_k
  if $\|V_k\|_{\mathrm{F}} \leq \varepsilon_{\text{stop}}$ **then return** X_{work} ▷ certify the returned point
  end if
  $\alpha \leftarrow 1$
  while $F(\mathrm{Retr}_{X_{\text{work}}}(\alpha V_k)) > F(X_{\text{work}}) - \frac{\alpha}{4t}\|V_k\|_{\mathrm{F}}^2$ and $\alpha > \alpha_{\text{tol}}$ **do**
    $\alpha \leftarrow 0.5\,\alpha$
  end while
  if the Armijo test fails at the final α (evaluated once more) **then return** status "linesearch failed"
  end if
  $X_{k+1} \leftarrow \mathrm{Retr}_{X_{\text{work}}}(\alpha V_k)$
  $D_{\text{prev}} \leftarrow V_k, \quad \alpha_{\text{prev}} \leftarrow \alpha$
  ▷ Adaptive proximal parameter t_k:
  if $\alpha = 1$ **then**
    $t \leftarrow \min(t_{\max}, t \cdot 1.02)$
  else if $\alpha = 0.5$ **then**
    $t \leftarrow t$
  else
    $t \leftarrow \max(t_0, t/1.1)$
  end if
end for

2. Preliminaries

2.1. Stiefel Manifold Geometry

The Stiefel manifold $\mathrm{St}(n,r) = \{X \in \mathbb{R}^{n\times r} : X^\top X = I_r\}$ is an embedded submanifold of $\mathbb{R}^{n\times r}$ of dimension $nr - r(r+1)/2$, compact when $n \geq r$. The tangent space at X is

$$T_X\mathrm{St}(n,r) = \{V \in \mathbb{R}^{n\times r} : X^\top V + V^\top X = 0\}. \tag{2}$$

Equipped with the Frobenius inner product $\langle A, B\rangle = \mathrm{Tr}(A^\top B)$, the orthogonal projection onto $T_X\mathrm{St}(n,r)$ is

$$\mathrm{Proj}_{T_X\mathrm{St}(n,r)}(Z) = Z - X\,\mathrm{sym}(X^\top Z), \tag{3}$$

where $\mathrm{sym}(A) = (A + A^\top)/2$. The projection is non-expansive: $\|\mathrm{Proj}_{T_X\mathrm{St}(n,r)}(Z)\|_F \leq \|Z\|_F$.

A retraction $\mathrm{Retr}_X : T_X\mathrm{St}(n,r) \to \mathrm{St}(n,r)$ satisfies $\mathrm{Retr}_X(0) = X$, and its derivative at the origin is the identity, $\frac{d}{dt}\mathrm{Retr}_X(tV)\big|_{t=0} = V$ for every tangent V. For the *polar* retraction

$$\mathrm{Retr}_X(V) = (X+V)(I + V^\top V)^{-1/2},$$

the matrix $I + V^\top V$ is positive definite for every tangent V, so the map is globally defined, and one has the global bounds (no radius restriction)

$$\|\mathrm{Retr}_X(V) - (X+V)\|_F \le \tfrac{1}{2}\|V\|_F^2, \tag{4}$$

$$\|\mathrm{Retr}_X(V) - X\|_F \le \|V\|_F. \tag{5}$$

Thus we may take $M_1 = 1/2$ and $M_2 = 1$ below; see also [33,34].

For a smooth function $f : \mathrm{St}(n,r) \to \mathbb{R}$, the Riemannian gradient is $\mathrm{grad} f(X) = \mathrm{Proj}_{T_X\mathrm{St}(n,r)}(\nabla f(X))$. A point $X^* \in \mathrm{St}(n,r)$ is *stationary* for (1) if

$$0 \in \mathrm{grad} f(X^*) + \mathrm{Proj}_{T_{X^*}\mathrm{St}(n,r)}(\partial h(X^*)), \tag{6}$$

where $\partial h(X^*)$ is the limiting subdifferential of h in the ambient space [5].

2.2. The ManPG Framework

At iteration k, ManPG [5] solves the strongly convex subproblem in the tangent space:

$$V_k = \mathrm{argmin}_{V \in T_{X_k}\mathrm{St}(n,r)} \langle \nabla f(X_k), V\rangle + \frac{1}{2t}\|V\|_F^2 + h(X_k + V), \tag{7}$$

where $t = 1/L$, and updates $X_{k+1} = \mathrm{Retr}_{X_k}(\alpha_k V_k)$ with Armijo backtracking on α_k.

The subproblem (7) is solved by semismooth Newton (SSN) on the dual system [5]: a symmetric multiplier $\Lambda \in \mathbb{R}^{r\times r}$ yields $q = r(r+1)/2$ variables; the dense code factors a $q \times q$ system ($O(r^6)$ with assembly), with conjugate gradient (CG) for larger r. For $h = \mu\|\cdot\|_1$ the prox is element-wise soft-thresholding. Warm starts give a few inner iterations, but the theory assumes *exact* minimizers (Section 4; Remark 3 quantifies how the inexact inner solves used in practice relate to that framework).

2.3. Vector Transport via Orthogonal Projection

Parallel transport costs $O(n^3)$ on $\mathrm{St}(n,r)$; Huang and Wei [7] use inverse retraction ($O(r^3)$ via a Lyapunov equation). We use orthogonal projection as vector transport [33]:

$$\Gamma_{X\to Y}(D) := \mathrm{Proj}_{T_Y\mathrm{St}(n,r)}(D), \qquad \|\Gamma_{X\to Y}(D)\|_F \le \|D\|_F,$$

at cost $O(nr^2)$ (one product and symmetrization). Boundedness, not isometry, is enough for our momentum step; no linear system is required.

2.4. Kurdyka–Łojasiewicz Inequality

Let Φ be a proper lower semicontinuous function. Following the standard local form [30], Φ satisfies the KL inequality at $z^* \in \mathrm{dom}\,\partial\Phi$ with exponent $\theta \in [0,1)$ if there exist $\eta > 0$, a neighborhood $\mathcal{U}$ of z^*, and $\delta > 0$ such that

$$\Phi(z^*) < \Phi(z) < \Phi(z^*) + \delta, \quad z \in \mathcal{U} \quad \Longrightarrow \quad \big(\Phi(z) - \Phi(z^*)\big)^\theta \le \eta\ \mathrm{dist}(0, \partial\Phi(z)). \tag{8}$$

On $\mathrm{St}(n,r)$ we work with the ambient composite $F = f + h$ and the projected stationarity measure $\mathrm{dist}(0, \mathrm{Proj}_{T_X\mathrm{St}}(\partial F(X)))$, writing "proximal KL" when (8) is used with that measure. The exponent controls the local rate: $\theta = 0$ finite termination; $\theta \in (0, 1/2]$ linear; $\theta \in (1/2, 1)$ sublinear [30]. The theoretical residual is $\|V_k\|_F$ (equivalently $\min_{i\le k}\|V_i\|_F$); experiments threshold the scaled squared residual $\|V_k\|_F^2/t^2$ against a tolerance proportional to nr (Section 5)—a different quantity from the norm accuracy ε in Theorem 2.

Remark 1 (Scope of exponent $1/2$)**.** *We do not claim that the penalized objective*

$$F(X) = -\mathrm{Tr}(X^\top A^\top A X) + \mu\|X\|_1$$

has KL exponent $1/2$ *at every stationary point. Explicitly, on* $\mathrm{St}(2,1) = S^1$ *with* $\mu = 1$ *and*

$$B = \frac{1}{250}\begin{pmatrix} 337 & 91 \\ 91 & 338 \end{pmatrix}, \quad x_* = (3/5, 4/5)^\top, \quad w = (-4/5, 3/5)^\top,$$

set $x(s) = x_* \cos s + w \sin s$*. Both coordinates stay positive for small* $|s|$*, so* $F(x) = -x^\top Bx + \|x\|_1$ *is smooth on that arc;* x_* *is stationary and*

$$F(x(s)) - F(x_*) = -\tfrac{1}{10}s^3 + O(s^4), \qquad |(F \circ x)'(s)| = \tfrac{3}{10}s^2 + O(|s|^3).$$

On the positive-gap side $s < 0$*,* $\sqrt{F - F^*}/\|\mathrm{grad} F\|$ *diverges, so no finite KL constant with exponent* $1/2$ *exists (local exponent* $2/3$*). This disproves an unrestricted stationary-point claim; it does not assert that every local minimizer has exponent* $2/3$*. Liu, So, and Wu [31] prove* $\theta = 1/2$ *for smooth quadratics* $\mathrm{Tr}(X^\top QX)$*; Wang, Liu, and So [32] analyze robust* ℓ_1*-PCA* ($\max \|D^\top Q\|_1$)*, not the penalized variance objective; Huang and Wei [10] obtain a Riemannian KL property for semialgebraic restrictions with unspecified exponent. Semialgebraicity guarantees some KL exponent, not* $1/2$*. Theorem 3 is therefore conditional: it applies at limit points where exponent* $1/2$ *and Assumption 1 are both verified (e.g., under suitable nondegeneracy on an identified smooth stratum).*

3. The ATE-ManPG Algorithm

3.1. Transport-Based Search-Direction Extrapolation

Classical FISTA forms $y_k = x_k + \beta_k(x_k - x_{k-1})$ [8,9]. On a manifold, $X_k - X_{k-1}$ need not lie in $T_{X_k}\mathrm{St}(n,r)$. We instead transport the previous *proximal search direction* V_{k-1} (already a tangent vector at the previous work point):

1. **Transport:** $\tilde{D}_{k-1} = \mathrm{Proj}_{T_{X_k}\mathrm{St}(n,r)}(\alpha_{k-1}V_{k-1})$.
2. **Extrapolate:** $X_k^{\mathrm{ext}} = \mathrm{Retr}_{X_k}(\beta_k \tilde{D}_{k-1})$.
3. **Solve:** At X_k^{ext}, solve the ManPG subproblem to obtain V_k, and retract to obtain X_{k+1}.

Even on a flat manifold this is *not* FISTA: transported-direction extrapolation behaves like $y_k = x_k + \beta_k(x_k - y_{k-1})$, not $x_k + \beta_k(x_k - x_{k-1})$.

Relative to Huang and Wei [7]: (i) no inverse retraction; (ii) bounded $\beta_k \le \beta_{\max}$ (default 0.5) as a design choice (FISTA's $\beta_k \uparrow 1$ stays finite); (iii) monotone restart if extrapolation increases F (Section 3.2). Related: [14].

3.2. Algorithm Description

Idealized iteration versus implementation. Section 4 analyzes the *exact idealized iteration*: exact subproblems, unrestricted backtracking ($\alpha_{\mathrm{tol}} = 0$), and no positive numerical stopping when infinite sequences are studied. A finite-floor implementation may return a linesearch-failure status; the theorems do not cover failed runs. The archived MATLAB code uses $\alpha_{\mathrm{tol}} = 10^{-4}$; here α_{tol} is the *Armijo step-size floor* of the backtracking loop, not the accuracy of the inner solver. The SSN subproblems are solved to an inner tolerance tied to the outer stopping rule: inner_tol $= \max(10^{-13}, \min(10^{-11}, 10^{-3}\epsilon t^2))$ on the squared dual residual, tightened to $\max(5 \times 10^{-16}, \min(10^{-14}, 10^{-5}\epsilon t^2))$ whenever the line search struggles. Here ϵ is the experimental tolerance on $\|V_k\|_{\mathrm{F}}^2/t^2$ of Section 5 (Algorithm 1 states the idealized norm test with $\varepsilon_{\mathrm{stop}}$). Remark 3 and (18) state in what sense this inexactness affects the theory. In that sense, the theoretical results support the implemented algorithm as follows: the safeguard, the extrapolation step, and the acceptance test are implemented

verbatim; under the relative inner stopping rule just described, inexactness degrades only the constants of Lemma 1 and Theorems 1–2 (Remark 3); and every reported run is validated by an external certificate—one additional SSN solve at the returned point—whose verdict is independent of the inner accuracy used during the run, with finite-floor failures reported as failures rather than successes. The archived code accepts a run only if an external residual certificate at the returned point passes. Failure is decided by the Armijo predicate after the loop, not merely by reaching the floor.

Output certificate. $\|V_k\|_F$ is computed at X_{work}, so we return X_{work}, not X_k.

Armijo. Trials are compared with $F(X_{\text{work}})$ (valid after restart). Lemma 1 gives $F(\text{Retr}_Y(\alpha V)) \leq F(Y) - \frac{\alpha}{t}\|V\|_F^2 + K\alpha^2\|V\|_F^2$, so $\frac{\alpha}{4t}\|V\|_F^2$ is guaranteed for $\alpha \leq 3/(4Kt)$; unrestricted halving accepts some $\alpha_k \geq \min\{1, 3/(8Kt)\}$.

Restart and $\beta_{\max}$. If extrapolation increases F, we set $X_{\text{work}} \leftarrow X_k$ and clear the momentum buffer. $\beta_{\max} = 0.5$ is a tuning default; Lemma 1 does not depend on it (Table 1).

The adaptive update of t_k uses only backtracking information already computed. The analysis of Section 4 is uniform over $t \in [1/L, 10/L]$; see Remark 7. We write t_k when the proximal parameter varies.

Table 1. Cap sensitivity for ATE-ManPG ($n = 500$, $r = 5$, $\mu = 0.8$, $\epsilon = 10^{-8}nr$, 10 trials; all certified 10/10).

$\beta_{\max}$	**0.2**	**0.5**	**0.8**	**1.0**
Iterations	205 ± 63	206 ± 70	204 ± 69	205 ± 69
Restarts/run	104	148	163	168

3.3. Comparison with Prior Accelerated Methods

AManPG (accelerated ManPG) [7]. Nesterov sequence $s_{k+1} = (1 + \sqrt{1 + 4s_k^2})/2$ and inverse-retraction extrapolation $y_{k+1} = \text{Retr}_{x_{k+1}}(\frac{1-s_k}{s_{k+1}} R_{x_{k+1}}^{-1}(x_k))$ (moving away from the old iterate), with a periodic safeguard that has a stationarity guarantee; no accelerated rate in that paper.

ARPG/P-ARPG (accelerated RPG variants) [10]. RPG rates as above; the vanilla variant V-ARPG is left open; the practical variant P-ARPG uses restarts.

Feng et al. [11]. Accelerated rates under geodesic convexity/curvature, different subproblem and safeguards.

ATE-ManPG. Differences: (i) projected *search-direction* extrapolation (no inverse retraction or iterate combination); (ii) objective-based restart every iteration; (iii) conditional KL rate for the safeguarded tangent-linearized ManPG iteration. We do not claim the first accelerated rate for all Riemannian proximal methods, nor a worst-case order improvement over ManPG.

4. Convergence Analysis

We establish the theoretical properties of the *exact idealized iteration* of Algorithm 1: exact subproblem minimizers, unrestricted backtracking ($\alpha_{\text{tol}} = 0$), and no positive numerical stopping when infinite sequences are studied. The analysis covers both a fixed parameter $t = 1/L$ and the adaptive rule of Section 3.2 with $t \in [1/L, 10/L]$ (Remark 7); the inexactness of the inner SSN solves used by the implementation is treated in Remark 3. It proceeds in four stages: sufficient decrease, global convergence, worst-case complexity, and improved complexity under the KL condition.

Standing assumptions and constants. f is the restriction of a smooth ambient function with L-Lipschitz gradient near $\text{St}(n, r)$; h is finite-valued convex on $\mathbb{R}^{n\times r}$. Model and retracted points lie in a compact tube; $\text{Lip}(h)$ is a Lipschitz constant of h there and M_Λ bounds Karush–Kuhn–Tucker (KKT) multipliers $\|\Lambda\|_2$ on that tube. Let $G := \max_{\text{St}} \|\nabla f\|_F$,

$K := L/2 + M_1(G + \mathrm{Lip}(h))$, $M_1 = 1/2$, $C_{\mathrm{gap}} := M_2\beta_{\max} = \beta_{\max}$, $D_k := \alpha_k V_k$, $D_{-1} := 0$, $F_{\mathrm{cl}} := \lim_k F(X_k)$. In Theorem 2, F^* may be any $F_{\mathrm{lb}} \le F$; in Theorem 3, $F^* := F_{\mathrm{cl}}$.

4.1. Sufficient Decrease

Lemma 1 (Sufficient decrease). *Let $\{X_k\}$ be generated by Algorithm 1, and let X_k^{ext} denote X_{work} at iteration k (so $X_k^{\mathrm{ext}} = X_k$ after a restart). Then $F(X_k^{\mathrm{ext}}) \le F(X_k)$ for every k. If $t > 0$ is fixed, backtracking accepts some*

$$\alpha_k \ge \bar{\alpha} := \min\Big\{1, \frac{3}{8Kt}\Big\} > 0 \tag{9}$$

and

$$F(X_{k+1}) - F(X_k) \le -\frac{\alpha_k}{4t}\|V_k\|_{\mathrm{F}}^2 \le -\frac{\bar{\alpha}}{4t}\|V_k\|_{\mathrm{F}}^2. \tag{10}$$

If $t_k \in [1/L, 10/L]$ varies, the same holds uniformly with

$$\bar{\alpha}_{\min} := \min\Big\{1, \frac{3}{8Kt_{\max}}\Big\}, \qquad a_{\min} := \frac{\bar{\alpha}_{\min}}{4t_{\max}} > 0, \tag{11}$$

so $\alpha_k \ge \bar{\alpha}_{\min}$ and $F(X_{k+1}) - F(X_k) \le -a_{\min}\|V_k\|_F^2$ for all k.

Proof. The restart safeguard yields $F(X_k^{\mathrm{ext}}) \le F(X_k)$.

Let $Y := X_k^{\mathrm{ext}}$ and $V := V_k$. Optimality of (7) and convexity of h give the *strong* model inequality

$$\langle \nabla f(Y), V\rangle + h(Y+V) - h(Y) \le -\frac{1}{t}\|V\|_{\mathrm{F}}^2, \tag{12}$$

because the first-order condition at V paired with V yields $\langle \nabla f(Y), V\rangle + \frac{1}{t}\|V\|^2 + \langle g, V\rangle = 0$ for some $g \in \partial h(Y+V)$, and $\langle g, V\rangle \ge h(Y+V) - h(Y)$ by convexity. (Testing the objective at V versus 0 gives the weaker bound with $\frac{1}{2t}\|V\|^2$; we use the stronger form.)

By L-Lipschitz continuity of ∇f along the polar path, the retraction bound (4) with $M_1 = 1/2$, and convexity of h,

$$f(\mathrm{Retr}_Y(\alpha V)) \le f(Y) + \alpha\langle \nabla f(Y), V\rangle + \tfrac{L}{2}\alpha^2\|V\|_{\mathrm{F}}^2 + GM_1\alpha^2\|V\|_{\mathrm{F}}^2, \tag{13}$$

$$h(\mathrm{Retr}_Y(\alpha V)) \le h(Y + \alpha V) + \mathrm{Lip}(h)M_1\alpha^2\|V\|_{\mathrm{F}}^2 \tag{14}$$

$$\le (1-\alpha)h(Y) + \alpha h(Y+V) + \mathrm{Lip}(h)M_1\alpha^2\|V\|_{\mathrm{F}}^2. \tag{15}$$

Summing and applying (12),

$$F(\mathrm{Retr}_Y(\alpha V)) \le F(Y) - \frac{\alpha}{t}\|V\|_{\mathrm{F}}^2 + K\alpha^2\|V\|_{\mathrm{F}}^2. \tag{16}$$

The Armijo target $F(\mathrm{Retr}_Y(\alpha V)) \le F(Y) - \frac{\alpha}{4t}\|V\|^2$ holds whenever $K\alpha \le 3/(4t)$, i.e., for all $\alpha \le 3/(4Kt)$. Halving from $\alpha = 1$ therefore accepts some $\alpha_k \ge \min\{1, 3/(8Kt)\} = \bar{\alpha}$, and

$$F(X_{k+1}) \le F(X_k^{\mathrm{ext}}) - \frac{\alpha_k}{4t}\|V_k\|_{\mathrm{F}}^2 \le F(X_k) - \frac{\alpha_k}{4t}\|V_k\|_{\mathrm{F}}^2. \tag{17}$$

For $t \le t_{\max}$ the uniform envelopes follow as stated. □

Remark 2. *Lemma 1 uses only the monotone restart (not decay of $\|V_k\|$), so there is no circular dependence on later theorems. When extrapolation is not monotone, momentum is discarded and the step reduces to ManPG at X_k.*

Remark 3 (Inexact inner solves). *The SSN solver returns a direction whose subproblem optimality condition (7) holds up to a dual residual E_k with $\|E_k\|_F =: \delta_k$, the quantity monitored by the SSN*

stopping test. Pairing the perturbed condition with V_k exactly as in the proof of Lemma 1 yields, in place of (12),

$$\langle \nabla f(Y), V\rangle + h(Y+V) - h(Y) \le -\frac{1}{t}\|V\|_{\mathrm{F}}^2 + \delta_k \|V\|_{\mathrm{F}}, \tag{18}$$

and hence $F(\mathrm{Retr}_Y(\alpha V)) \le F(Y) - \frac{\alpha}{t}\|V\|_{\mathrm{F}}^2 + K\alpha^2\|V\|_{\mathrm{F}}^2 + \alpha\delta_k\|V\|_{\mathrm{F}}$. If the inner solver is stopped by a relative rule

$$\delta_k \le c_0 \frac{\|V_k\|_{\mathrm{F}}}{t_k}, \qquad c_0 \in (0, \tfrac{3}{4}), \tag{19}$$

then the last term is absorbed as $c_0 \frac{\alpha}{t_k}\|V_k\|_{\mathrm{F}}^2$, the Armijo predicate holds for every α with $K\alpha \le \frac{3/4 - c_0}{t_k}$, and the conclusions of Lemma 1, Theorem 1, and Theorem 2 persist verbatim with $\bar{\alpha}$, $\bar{\alpha}_{\min}$, and $a_{\min}$ multiplied by $1 - \frac{4c_0}{3}$, all uniformly over $t \in [1/L, 10/L]$. The archived code enforces a rule of the form (19): an iterate is processed only while $\|V_k\|_{\mathrm{F}}^2/t_k^2 \ge \epsilon$, and the branch inner_tol $\le 10^{-3}\epsilon t_k^2$ *of Section 3.2 then gives $\delta_k \le \sqrt{10^{-3}}\, t_k\sqrt{\epsilon} \le 0.032\,\|V_k\|_{\mathrm{F}}$, i.e., (19) with $c_0 = 0.032\, t_k \le 0.32/L \le 0.16$ on the sparse-PCA instances of Section 5, where column normalization gives $L \ge 2$. Inexactness therefore degrades constants, not the qualitative guarantees. At tolerances so tight that the absolute floor 10^{-13} dominates the $10^{-3}\epsilon t_k^2$ branch, the relative guarantee (19) ceases to bind; this is precisely the regime of the high-precision failures discussed in Section 5.3. A complete inexact analysis in the spirit of [16] is left as future work (Section 5.4).*

4.2. Global Convergence and Worst-Case Complexity

Theorem 1 (Global convergence)**.** *Every limit point of $\{X_k\}$ generated by Algorithm 1 is a stationary point of (1). Moreover $\|V_k\|_F \to 0$ and $\|X_k^{\mathrm{ext}} - X_k\|_F \to 0$.*

Proof. From Lemma 1, $\{F(X_k)\}$ is nonincreasing. Compactness of $\mathrm{St}(n,r)$ and continuity of F imply $\{F(X_k)\}$ converges. Summing the descent inequality gives $\sum_{i=0}^{\infty}\|V_i\|_F^2 < \infty$, hence $\|V_k\|_F \to 0$. By (5) and non-expansiveness of projection,

$$\|X_k^{\mathrm{ext}} - X_k\|_{\mathrm{F}} \le M_2\beta_k\|\tilde{D}_{k-1}\|_{\mathrm{F}} \le C_{\mathrm{gap}}\|D_{k-1}\|_{\mathrm{F}} \le C_{\mathrm{gap}}\|V_{k-1}\|_{\mathrm{F}} \to 0. \tag{20}$$

Let X_* be a limit point with subsequence $\{X_{k_j}\} \to X_*$. Then also $X_{k_j}^{\mathrm{ext}} \to X_*$ and $X_{k_j}^{\mathrm{ext}} + V_{k_j} \to X_*$. The subproblem optimality condition at $Y_j := X_{k_j}^{\mathrm{ext}}$ is

$$0 \in \nabla f(Y_j) + \frac{1}{t_j}V_{k_j} + \partial h(Y_j + V_{k_j}) + \mathcal{A}_{Y_j}^*(\Lambda_j), \tag{21}$$

where $t_j \in [1/L, 10/L]$, $\mathcal{A}_X^*(\Lambda) = X\Lambda$ with $\Lambda = \Lambda^\top$. Since $\|\mathcal{A}_{Y_j}^*(\Lambda_j)\|_F = \|\Lambda_j\|_F$ and, choosing $g_j \in \partial h(Y_j + V_{k_j})$ in the inclusion,

$$\|\Lambda_j\|_{\mathrm{F}} \le \|\nabla f(Y_j)\|_{\mathrm{F}} + \frac{\|V_{k_j}\|_{\mathrm{F}}}{t_j} + \|g_j\|_{\mathrm{F}} \le G + L\|V_{k_j}\|_{\mathrm{F}} + \mathrm{Lip}(h), \tag{22}$$

the sequence $\{\Lambda_j\}$ is bounded. Passing to a further subsequence with $\Lambda_j \to \Lambda_*$ (symmetric) and $g_j \to g_*$, outer semicontinuity of the limiting subdifferential of the convex function h [35] yields $g_* \in \partial h(X_*)$, and continuity of ∇f gives

$$0 \in \nabla f(X_*) + \partial h(X_*) + \mathcal{A}_{X_*}^*(\Lambda_*). \tag{23}$$

Projecting onto $T_{X_*}\mathrm{St}(n,r)$ and noting $\mathrm{Proj}_{T_{X_*}\mathrm{St}}(\mathcal{A}_{X_*}^*(\Lambda_*)) = 0$ because Λ_* is symmetric yields the stationarity condition (6). □

https://doi.org/10.3390/axioms15110828

Theorem 2 (Worst-case complexity). *Algorithm 1 requires $O(\varepsilon^{-2})$ iterations to achieve* $\min_{i<k} \|V_i\|_F \le \varepsilon$.

Proof. From Lemma 1, telescoping yields $\min_{i<k} \|V_i\|_F^2 \le \frac{F(X_0)-F_{\mathrm{lb}}}{a_{\min}} \cdot \frac{1}{k}$ for any lower bound $F_{\mathrm{lb}} \le F$, giving $\min_{i<k} \|V_i\|_F = O(1/\sqrt{k})$, i.e., $O(\varepsilon^{-2})$ iterations. □

4.3. Conditional Linear Convergence Under the KL Condition

We state a conditional local rate. Its inputs are (i) proximal KL with exponent 1/2 at the limit set and (ii) a relative-error bound linking projected stationarity at the *accepted* point to previous steps. Neither hypothesis is verified here for all sparse-PCA instances (Remark 1). Relative error of this type is standard for proximal algorithms [29,30], but for nonsmooth h the subproblem supplies a subgradient at the affine model point $Y + V$, while the accepted iterate is $\mathrm{Retr}_Y(\alpha V)$; small distance between those points does not by itself transfer subgradients across a kink.

Assumption 1 (Lagged relative error). *There exist $b_1, b_2 > 0$ such that for all sufficiently large k,*

$$\mathrm{dist}\big(0, \mathrm{Proj}_{T_{X_k}\mathrm{St}(n,r)}(\partial F(X_k))\big) \le b_1 \|V_{k-1}\|_{\mathrm{F}} + b_2 \|D_{k-1}\|_{\mathrm{F}}. \tag{24}$$

Proposition 1 (Sufficient conditions for Assumption 1). *Let Y be a work point, V the exact subproblem direction, $Z = Y + V$, and $X = \mathrm{Retr}_Y(\alpha V)$ the accepted point with $\alpha \in (0, 1]$. Suppose there exist $g_Z \in \partial h(Z)$ and $g_X \in \partial h(X)$ with $\|g_X - g_Z\| \le c_h \|V\|_F$, and let M_Λ bound the KKT multiplier. Then*

$$\mathrm{dist}\big(0, \mathrm{Proj}_{T_X\mathrm{St}}(\partial F(X))\big) \le (2L + c_h + M_\Lambda)\, \|V\|_F. \tag{25}$$

In particular: (a) if h is C^1 with L_h-Lipschitz gradient near Y, Z, X, take $c_h = 2L_h$; (b) if $h = \mu\|\cdot\|_1$ and Z, X are sign-compatible on the nonzero coordinates of X (same signs; zeros of X admit $[-\mu, \mu]$), take $g_X = g_Z$ and $c_h = 0$. Numeric observation of a stable support is not a proof of (b).

Proof. The subproblem KKT system yields a symmetric Λ with $\|\Lambda\|_2 \le M_\Lambda$ and $g_Z \in \partial h(Z)$ such that $\nabla f(Y) + t^{-1}V + g_Z + Y\Lambda = 0$. Since $\alpha \le 1$ and $M_2 = 1$, $\|X - Y\| \le \|V\|_F$ and $\|X - Z\| \le 2\|V\|_F$. Using $\mathrm{Proj}_{T_X}(X\Lambda) = 0$, nonexpansiveness of projection, $\|\nabla f(X) - \nabla f(Y)\| \le L\|X - Y\|$, $1/t \le L$, and $\|g_X - g_Z\| \le c_h\|V\|_F$,

$$\mathrm{dist}(0, \mathrm{Proj}_{T_X}\partial F(X)) \le \|\mathrm{Proj}_{T_X}(\nabla f(X) + g_X)\| \le (2L + c_h + M_\Lambda)\|V\|_F.$$

Under unrestricted backtracking $\alpha \ge \bar{\alpha}_{\min} > 0$, so $\|V\|_F \le \bar{\alpha}_{\min}^{-1}\|D\|_F$ and Assumption 1 follows on the accepted step. □

Remark 4 (Scope of Assumption 1). *Without Proposition 1 (for example at a kink of h with incompatible signs), $\|V_k\|_F$ remains the primary residual in Theorem 2, and the rate below stays an abstract implication. We do not claim that Assumption 1 holds for every sparse-PCA instance.*

Theorem 3 (Conditional linear convergence under KL). *Let $F^* := F_{\mathrm{cl}} = \lim_k F(X_k)$. Assume the proximal KL inequality holds with exponent $\theta = 1/2$ on a neighborhood of the limit set of $\{X_k\}$ in the one-sided window $F^* < F(X) < F^* + \delta$, and that Assumption 1 holds. Then there exist k_0 and $\rho \in (0, 1)$ such that for $k \ge k_0$,*

$$\Delta_{k+1} \le \rho\, \Delta_k, \qquad \|V_k\|_F \le C\, \rho^{(k-k_0)/2}, \tag{26}$$

where $\Delta_k := F(X_k) - F^*$. Consequently $\min_{i \le k} \|V_i\|_F \le \varepsilon$ in $O(\log(1/\varepsilon))$ iterations.

Proof. Set $p_k := \|V_k\|_F$. By Lemma 1, $\Delta_k - \Delta_{k+1} \geq a\, p_k^2$ with $a := a_{\min} > 0$ uniform. KL with $\theta = 1/2$ and Assumption 1 give, for $k \geq k_0$,

$$\Delta_k^{1/2} \leq \eta\ \mathrm{dist}\big(0, \mathrm{Proj}_{T_{X_k}\mathrm{St}}(\partial F(X_k))\big) \leq \eta(b_1 p_{k-1} + b_2\|D_{k-1}\|_F) \leq \eta(b_1+b_2)\, p_{k-1}, \quad (27)$$

using $\|D_{k-1}\|_F = \alpha_{k-1} p_{k-1} \leq p_{k-1}$. In particular $\Delta_k \leq c_{\mathrm{KL}}^2 p_{k-1}^2$ with $c_{\mathrm{KL}} := \eta(b_1+b_2)$. Shifting the index, $\Delta_{k+1} \leq c_{\mathrm{KL}}^2 p_k^2$, while $\Delta_k - \Delta_{k+1} \geq a p_k^2$ implies $p_k^2 \leq (\Delta_k - \Delta_{k+1})/a$, so

$$\Delta_{k+1} \leq \frac{c_{\mathrm{KL}}^2}{a + c_{\mathrm{KL}}^2}\, \Delta_k =: \rho\, \Delta_k, \qquad \rho := \frac{c_{\mathrm{KL}}^2}{a + c_{\mathrm{KL}}^2} \in (0,1). \quad (28)$$

This is a standard one-step KL contraction: a momentum-energy Lyapunov term is not required once (24) is in force. R-linear decay of p_k follows from $a p_k^2 \leq \Delta_k - \Delta_{k+1} \leq \Delta_k \leq \Delta_{k_0} \rho^{k-k_0}$. □

Remark 5 (Uniformization and scope)**.** *If the proximal KL inequality holds uniformly on a compact neighborhood of the compact limit set intersected with a sublevel set, then $k_0 < \infty$. Under Theorem 3, $\|X_{k+1} - X_k\| \leq M_2\|V_k\| + C_{\mathrm{gap}}\|V_{k-1}\|$ is summable, so $\{X_k\}$ has finite length and converges to one stationary point. For ManPG, $\beta_k = 0$ and $X_{\mathrm{work}} = X_k$; the accepted step D_{k-1} remains the natural step measure.*

Remark 6 (Role of the bounded schedule)**.** $\beta_{\max} = 0.5$ *is a tuning default. Polar bounds $M_1 = 1/2$, $M_2 = 1$ are global; FISTA's $\beta_k \uparrow 1$ stays bounded; Lemma 1 has no $\beta_{\max}$ dependence. Restart, not the cap, gives monotonicity (Table 1). We do not claim that the cap prevents basin transitions. Restart enforces $F(X_{\mathrm{work}}) \leq F(X_k)$; objective monotonicity does not rule out basin transitions in a nonconvex problem.*

Remark 7 (Adaptive t)**.** $a_{\min}$, $\bar{\alpha}_{\min}$, *and $1/t \leq L$ are uniform on $t \in [1/L, 10/L]$, so Algorithm 1's adaptive rule preserves the results with those envelopes. Adaptation is not proved to improve the rate relative to fixed $t = 1/L$.*

In summary, the adaptive parameter $t_k \in [1/L, 10/L]$ is theoretically harmless (uniform constants, Remark 7); the theorems assume exact solves and unrestricted backtracking (Section 3.2), and Remark 3 records that the relative inner stopping rule used by the implementation degrades only the constants of those theorems.

5. Numerical Experiments

We evaluate ATE-ManPG on the sparse-PCA model $f(X) = -\mathrm{Tr}(X^\top A^\top A X)$, $h(X) = \mu\|X\|_1$, with $A \in \mathbb{R}^{50\times n}$ (columns normalized). Baselines:

- **ManPG** [5]: fixed $t = 1/L$, SSN subproblem, Armijo constant $1/(2t)$.
- **ManPG-Ada** [5]: the adaptive proximal-parameter variant of Chen et al. (their Algorithm 2), included as the natural adaptive baseline. Our implementation adapts t from the backtracking count (increase by 1.01 without backtracking, decrease by 1.01 otherwise) and, for comparability with the ATE family, caps $t \in [1/L, 10/L]$; we do not claim identical update constants to their pseudocode.
- **TE-ManPG** (transport-extrapolated ManPG): the fixed-t transported-direction variant (extrapolation and restart, no adaptation), reported in the ablation.
- **AManPG** [7]: Huang–Wei FISTA-style extrapolation with inverse retraction and a periodic safeguard. The supplied driver uses its own tangent semismooth subproblem routine (not the RPMWH routine of [10], the Riemannian proximal-mapping subproblem solver of Huang and Wei).

- **P-ARPG** and **RPG** [10]: practical accelerated and non-accelerated RPG, using the RPMWH subproblem model.

Certified protocol. Tables come from archived scalar records (method, dimensions, tolerance, seed, outer iterations, wall-clock time, termination status, restarts, external residual/feasibility at the returned point). Data matrices and returned iterates are not archived; endpoints require rerunning from the seed. Iteration counts are method-specific outer iterations (AManPG and P-ARPG may solve several subproblems per outer step). The external certificate is one SSN solve at returned X with $t = 1/L$ and tight inner tolerance; success requires residual below tolerance and feasibility $<10^{-6}$. Seeds are shared (trial t: $42 + 10{,}000\,t$). Random singular-value-decomposition (SVD) initial points; MATLAB R2024a, Windows 11, Core i7. Main tables: 10 trials; sweeps: 3; high-precision: 1 trial (seed 10042). Sub-second timings are implementation-specific.

5.1. Standard-Precision Comparison

Table 2 reports $\epsilon = 10^{-8}$, $n = 500$, $r = 5$, $\mu = 0.8$, 10 trials. All 60 runs shown in the table, together with the 10 additional TE-ManPG runs of the factorial ablation (Section 5.4), passed the external certificate (residual below tolerance at the returned point, feasibility $<10^{-6}$); every method certified 10/10.

Table 2. Standard-precision comparison ($n = 500$, $r = 5$, $\mu = 0.8$, $\epsilon = 10^{-8}nr$, 10 trials), generated from archived per-run records. All methods certified 10/10. Iter. speedup is the mean iteration count relative to ManPG; Solution F is the mean final certified objective over the trials. The proposed method ATE-ManPG is typeset in bold; boldface numbers indicate the best value in each column among the methods sharing the same subproblem model (ManPG, ManPG-Ada, ATE-ManPG). Methods marked † use a different subproblem model and are not ranked, and the Solution F column is not ranked.

Method	Iterations	Time (s)	Iter. Speedup	Solution F
ManPG	741 ± 296	0.070 ± 0.045	$1.00\times$	-20.062
ManPG-Ada	251 ± 72	$\mathbf{0.034 \pm 0.020}$	$2.95\times$	-20.062
ATE-ManPG	$\mathbf{206 \pm 70}$	0.051 ± 0.026	$\mathbf{3.59}\times$	-19.994
AManPG	140 ± 25	0.049 ± 0.032	$5.29\times$	-19.963 †
P-ARPG	107 ± 23	0.116 ± 0.047	$6.93\times$	-19.983 †
RPG	658 ± 255	0.286 ± 0.095	$1.13\times$	-20.091 †

ATE-ManPG shows a $3.59\times$ mean iteration reduction vs. fixed-parameter ManPG, but this is not a momentum-only effect. ManPG-Ada already gives $2.95\times$. In the harmonized ablation, adding transported momentum yields a further **6.4%** mean iteration reduction (220.4 → 206.3), concentrated in one paired trial (seed 90042: about 106 of 141 iterations saved, with a worse objective -18.75 vs. -19.43); the other nine pairs improve by about 1.8%. Mean runtime is higher with momentum (0.051 s vs. 0.0317 s adaptive-only; ATE slower in 9/10 pairs).

Table 2 is therefore best read jointly with the factorial ablation of Section 5.4 when attributing the improvement. Of the $3.59\times$ total mean iteration reduction relative to fixed-parameter ManPG, the pre-existing adaptive proximal parameter accounts for $2.95\times$ (ManPG-Ada vs. ManPG), and the harmonized ablation separates the two mechanisms on one implementation: adaptation alone gives $3.4\times$, transported extrapolation alone $1.6\times$, and their combination $3.6\times$, so the marginal contribution of extrapolation on top of adaptation is 6.4% in mean iterations. The difference between $3.59\times$ and $2.95\times$ in Table 2 should accordingly not be read as a pure momentum effect: that table compares total

methods whose adaptive constants differ (1.02/1.1 here against the 1.01-rule of ManPG-Ada), and part of that gap is tuning rather than extrapolation (Section 5.4).

The main advantages of ATE-ManPG over the closest methods, as demonstrated by these examples, are as follows. Compared with fixed-parameter ManPG, it is uniformly faster across settings: 3.59× fewer mean iterations here (741 → 206, all 10/10 certified), 2.9–3.9× across $n = 100$–1500 (Table 3), and 1.5–3.6× across μ (Table 4). Compared with the extrapolation-based AManPG and P-ARPG, it avoids inverse retractions (the transport is one orthogonal projection), enforces decrease of the objective at every iteration through the restart, and returns the point at which the residual is certified; on this test bed it also reaches a lower certified objective than either ($F = -19.994$ vs. -19.963 and -19.983 in Table 2) at a lower runtime than P-ARPG (0.051 s vs. 0.116 s), whereas P-ARPG needs fewer outer iterations under a different subproblem model and outer-step accounting. Compared with ManPG-Ada, its advantage is modest and instance-dependent, as quantified above. At standard precision, RPG attains a marginally lower objective (-20.091) at the highest iteration and time cost; at high precision it is the only method certified at all four tolerances (Section 5.3), so no single method dominates. The demonstrated strength of ATE-ManPG is the combination of ManPG-level guarantees, retraction independence, and certified monotone progress at standard precision.

Table 3. Dimension sweep: iterations vs. n ($r = 5$, $\mu = 0.8$, $\epsilon = 10^{-8}nr$, 3 trials). Parenthesized entries give certified/total runs. Boldface indicates the best value in each column.

n	100	200	500	800	1000	1500
ManPG	487 ± 87	528 ± 123	1021 ± 178	1474 ± 699	1487 ± 638	1337 ± 621
ManPG-Ada	177 ± 15	190 ± 31	326 ± 49	459 ± 184	$\mathbf{459 \pm 196}$	$\mathbf{457 \pm 168}$
ATE-ManPG	$\mathbf{126 \pm 11}$	$\mathbf{166 \pm 63}$	$\mathbf{287 \pm 51}$	$\mathbf{415 \pm 204}$	516 ± 156 (2/3)	462 ± 164
ATE speedup	3.9×	3.2×	3.6×	3.6×	2.9×	2.9×

Table 4. Iterations vs. μ ($n = 500$, $r = 5$, $\epsilon = 10^{-8}nr$, 3 trials; all certified 3/3). Boldface indicates the best value in each column.

μ	0.2	0.4	0.8	1.6	3.2
ManPG	1122 ± 602	768 ± 116	1021 ± 178	82 ± 29	36 ± 6
ManPG-Ada	$\mathbf{482 \pm 230}$	$\mathbf{315 \pm 41}$	326 ± 49	60 ± 15	31 ± 5
ATE-ManPG	486 ± 234	324 ± 75	$\mathbf{287 \pm 51}$	$\mathbf{41 \pm 8}$	$\mathbf{24 \pm 3}$
ATE speedup	2.3×	2.4×	3.6×	2.0×	1.5×

5.2. Dimension Scaling and Regularization

Tables 3 and 4 report iteration counts (3 trials, ManPG-Ada included). All runs passed the external certificate except one ATE-ManPG trial at $n = 1000$ (reported as 2/3), whose returned residual exceeded the tolerance after its internal stopping triggered.

Iteration reductions vs. ManPG are stable in n but vary with μ. These sweeps compare total methods (different tunings), not isolated momentum; at $n = 1000$ ATE certifies only 2/3 runs, so that column is not paired.

5.3. High-Precision Behavior

Table 5 reports a single trial (seed 10042) under a uniform iteration budget of 5000, with success decided *only* by the external certificate. All iteration counts are archived outer iterations (numbers of outer subproblem solves).

The certified picture differs from internal-stopping counts. At $\epsilon \leq 10^{-12}$ the SSN family (ManPG, TE, ATE) hits explicit Armijo failure before the tolerance (operative inner

floor 10^{-13} in the revised code). ManPG-Ada passes 10^{-12} and fails at 10^{-14}. P-ARPG safeguards out after 168 iterations with residual near 1.3×10^{-7}. AManPG's inner solver slows severely and did not finish the budget at $\epsilon \leq 10^{-12}$ (no coded timeout in the archives). RPG is certified at all four tolerances here.

Table 5. High-precision comparison: iterations vs. scaled squared-residual tolerance ($n = 500$, $r = 5$, $\mu = 0.8$, single trial, budget 5000). ✓: external certificate passed at the returned point. "LS fail": explicit Armijo failure triggered before the tolerance was reached (returned residual exceeds it). "safe.": early safeguard exit whose returned residual exceeds the tolerance. "—": did not complete the budget within the allotted time (see text).

ϵ	ManPG	ManPG-Ada	TE-ManPG	ATE-ManPG	AManPG	P-ARPG	RPG
$10^{-8}\,nr$	1054 ✓	327 ✓	725 ✓	297 ✓	171 ✓	125 ✓	824 ✓
$10^{-10}\,nr$	1245 ✓	372 ✓	852 ✓	337 ✓	191 ✓	168 ✓	1017 ✓
$10^{-12}\,nr$	1320 LS	412 ✓	885 LS	339 LS	—	168 SAFE.	1212 ✓
$10^{-14}\,nr$	1320 LS	434 LS	885 LS	339 LS	—	168 SAFE.	1409 ✓

The failure entries have identifiable causes, which are numerical rather than theoretical. The Armijo failures of the SSN family at $\epsilon \leq 10^{-12}nr$ reflect the interaction between the line search and the inner solver at very small tolerances: Lemma 1 guarantees acceptance of every sufficiently small α only for exact subproblems, whereas near the requested residual the decrease predicted by the computed direction, of order $\frac{\alpha}{4t}\|V_k\|_F^2$, falls below the perturbation induced by the inner dual-residual floor ($\sqrt{10^{-13}} \approx 3 \times 10^{-7}$), which biases V_k at a level that backtracking cannot repair—halving α shrinks the target linearly while the realizable decrease is already dominated by that bias. The adaptive tightening of inner_tol when the line search struggles (Section 3.2) targets precisely this interaction. At $\epsilon = 10^{-14}nr$ the predicted per-step decrease is of order 10^{-12}; the inner-residual bias then dominates the scale being verified, and the cancellation roundoff in evaluating differences of objective values of order 20 (absolute noise of order 10^{-14}) is no longer negligible, so certification at that level sits at the limit of double-precision arithmetic for this family of codes. The P-ARPG entries are not numerical failures: its safeguard exit is a designed termination that returns a point with residual $\approx 1.3 \times 10^{-7}$, above the tighter thresholds. AManPG's dashes are budget failures: its inner routine ties the inner effort to the outer tolerance, the cost per outer step grows as ϵ shrinks, and the run did not finish the budget within the allotted time. None of these entries contradicts the theory: Theorems 1–2 concern the exact idealized iteration, and the certified entries of ManPG-Ada and RPG at $\epsilon = 10^{-12}nr$ show that the practical floor is implementation-dependent rather than universal.

These counts are empirical observations under one implementation and budget, not a verification of asymptotic orders: a worst-case $O(\varepsilon^{-2})$ bound neither predicts individual trajectories nor is refuted by budget exhaustion, the table's ϵ thresholds a scaled *squared* residual (a factor-100 tighter squared tolerance is a factor-10 tighter norm tolerance), and every method can enjoy faster local behavior under additional regularity.

5.4. Factorial Ablation and Cap Sensitivity

Table 6 reports the factorial ablation requested to separate the two mechanisms: the four variants are produced by one implementation (identical SSN accuracy, stopping certificate, Armijo constant, adaptive rule, seeds), toggling only the momentum flag and the adaptive proximal parameter.

Table 6. Factorial ablation ($n = 500$, $r = 5$, $\mu = 0.8$, $\epsilon = 10^{-8}nr$, 10 trials; all certified 10/10). "Restarts" is the mean number of monotone-restart triggers per run.

Configuration	Iterations	Solution F	Restarts/Run
fixed t, no extrapolation (= ManPG)	741 ± 296	-20.062	—
fixed t, extrapolation	477 ± 187	-19.994	35
adaptive t, no extrapolation	220 ± 70	-20.062	—
adaptive t, extrapolation (ATE)	$\mathbf{206 \pm 70}$	-19.994	148

The attribution is as follows. Adaptation alone: 3.4× (741 → 220); momentum alone: 1.6× (741 → 477); combined: 3.6× (741 → 206). Momentum adds **6.4%** on top of the harmonized adaptive rule, concentrated in one paired trial (Section 5), with higher runtime and a worse objective there. ATE triggers restarts frequently (148 per ∼206 iterations). Our adaptive factors 1.02/1.1 are stronger than the 1.01-rule ManPG-Ada (220 vs. 251), so part of that gap is tuning.

The marginal effect of extrapolation is also strongly instance-dependent, and its concentration in one pair indicates a dependence on the problem instance and on the initial point. The restart statistics are consistent with this mechanism: extrapolation can help only while successive proximal directions remain correlated along a smooth arc of the objective, whereas near the ℓ_1 kinks of h the directions decorrelate and the monotone restart discards the momentum—ATE triggers restarts on 148 of roughly 206 iterations—so only instances whose trajectory stays on such an arc long enough for the extrapolation to compound benefit materially; the seed-90042 trial is the one such instance among the ten. The transported momentum should accordingly be regarded as a safeguarded option that is decisive for certain instances rather than a uniform accelerator: across the μ-sweep its marginal effect ranges from negligible ($\mu = 0.2$: 486 vs. 482 iterations) to moderate ($\mu = 1.6$: 41 vs. 60), and the worst-case theory of Section 4 claims no order improvement from it (Remark 6).

Table 1 varies the momentum cap. Mean iteration counts are flat across $\beta_{\max} \in \{0.2, 0.5, 0.8, 1.0\}$ (all within one standard deviation), while the restart frequency increases mildly with the cap. This confirms that 0.5 is a tuning default with no special theoretical role (Remark 6): the safeguard, not the numerical cap, controls the iteration.

Limitations. (i) Prove $\theta = 1/2$ and Assumption 1 for nondegenerate sparse-PCA subclasses, or state general-exponent rates; (ii) harmonize inexact SSN with the analysis [16], of which Remark 3 is a constant-level first step; (iii) restart-only/cap-only studies and additional applications; (iv) explain the quality tradeoff in the concentrated trial. Penalty and constraint-dissolving alternatives [20,23,25] remain uncompared.

6. Conclusions

ATE-ManPG is a safeguarded transport-extrapolated proximal gradient method on the Stiefel manifold: projected search-direction extrapolation, monotone restart, and a bounded adaptive proximal parameter, returning the work point where the residual is certified. For the exact idealized iteration we proved global subsequential stationarity and $O(\varepsilon^{-2})$ iteration complexity, and a *conditional* R-linear rate under explicit KL-1/2 and relative-error hypotheses (smoothness or sign compatibility). The penalized sparse-PCA objective has no universal exponent-1/2 property; we give an explicit counterexample. Any local order change from $O(\varepsilon^{-2})$ to $O(\log(1/\varepsilon))$ is thus a KL effect rather than a worst-case acceleration (Theorem 3); we do not claim a better worst-case order than ManPG. Experiments under a uniform external certificate show a 3.6× mean iteration reduction vs. fixed-parameter ManPG, mostly from the pre-existing adaptive parameter; transported momentum adds a small, instance-dependent 6.4% with runtime overhead and occasional quality tradeoffs.

https://doi.org/10.3390/axioms15110828

The distinctive construction is search-direction transport plus objective-based restart; the theory records when a KL rate is preserved.

Author Contributions: Conceptualization, L.C. and C.Y.; methodology, L.C. and C.Y.; software, C.Y.; validation, L.C. and C.Y.; formal analysis, L.C.; investigation, L.C. and C.Y.; writing—original draft preparation, L.C.; writing—review and editing, L.C. and C.Y.; supervision, L.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partially supported by the Science and Technology Research Program of Chongqing Municipal Education Commission (KJQN202000533, KJZD-M202300504), the Chongqing Natural Science Foundation Project (CSTB2024NSCQ-LZX0140, CSTB2025NSCQ-LZX0008), the National Science Foundation of China (12171063).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The MATLAB source code and archived per-run experimental records supporting the numerical results are available from the authors on reasonable request.

Acknowledgments: The authors thank colleagues for comments that improved the experimental protocol and the scope of the theoretical claims.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Jolliffe, I.T.; Trendafilov, N.T.; Uddin, M. A modified principal component technique based on the LASSO. *J. Comput. Graph. Stat.* **2003**, *12*, 531–547. [CrossRef]
2. Ozoliņš, V.; Lai, R.; Caflisch, R.; Osher, S. Compressed modes for variational problems in mathematics and physics. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 18368–18373. [CrossRef]
3. Yang, Y.; Shen, H.T.; Ma, Z.; Huang, Z.; Zhou, X. $\ell_{2,1}$-Norm Regularized Discriminative Feature Selection for Unsupervised Learning. In Proceedings of the IJCAI, Barcelona, Spain, 16–22 July 2011; pp. 1589–1594.
4. Zhang, Y.; Lau, Y.; Kuo, H.-W.; Cheung, S.; Pasupathy, A.; Wright, J. On the Global Geometry of Sphere-Constrained Sparse Blind Deconvolution. In Proceedings of the CVPR, Honolulu, HI, USA, 21–26 July 2017; pp. 4891–4899.
5. Chen, S.; Ma, S.; So, A.M.-C.; Zhang, T. Proximal gradient method for nonsmooth optimization over the Stiefel manifold. *SIAM J. Optim.* **2020**, *30*, 210–239. [CrossRef]
6. Chen, S.; Ma, S.; So, A.M.-C.; Zhang, T. Nonsmooth optimization over the Stiefel manifold and beyond: Proximal gradient method and recent variants. *SIAM Rev.* **2024**, *66*, 319–352. [CrossRef]
7. Huang, W.; Wei, K. An extension of fast iterative shrinkage-thresholding algorithm to Riemannian optimization for sparse principal component analysis. *Numer. Linear Algebra Appl.* **2022**, *29*, e2409. [CrossRef]
8. Beck, A.; Teboulle, M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.* **2009**, *2*, 183–202. [CrossRef]
9. Nesterov, Y. A method for solving a convex programming problem with convergence rate $O(1/k^2)$. *Sov. Math. Dokl.* **1983**, *27*, 372–376.
10. Huang, W.; Wei, K. Riemannian proximal gradient methods. *Math. Program.* **2022**, *194*, 371–413. [CrossRef]
11. Feng, S.; Jiang, Y.; Huang, W.; Ying, S. A Riemannian accelerated proximal gradient method. *arXiv* **2025**, arXiv:2509.21897.
12. Ghadimi, S.; Lan, G. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Math. Program.* **2016**, *156*, 59–99. [CrossRef]
13. Li, H.; Lin, Z. Accelerated proximal gradient methods for nonconvex programming. In Proceedings of the NeurIPS, Montréal, QC, Canada, 7–12 December 2015; pp. 379–387.
14. O'Donoghue, B.; Candès, E. Adaptive restart for accelerated gradient schemes. *Found. Comput. Math.* **2015**, *15*, 715–732. [CrossRef]
15. Si, W.; Absil, P.-A.; Huang, W.; Jiang, R.; Vary, S. A Riemannian proximal Newton method. *SIAM J. Optim.* **2024**, *34*, 654–681. [CrossRef]
16. Huang, W.; Wei, K. An inexact Riemannian proximal gradient method. *Comput. Optim. Appl.* **2023**, *85*, 1–32. [CrossRef]
17. Xu, J.; He, B.; Peng, Z.; Zhang, X. An improved proximal alternating linearized minimization algorithm and its inertial Bregman extension for structured nonconvex nonsmooth optimization problems on Riemannian manifold. *J. Comput. Appl. Math.* **2026**, *473*, 116886. [CrossRef]

18. Hu, J.; Liu, X.; Wen, Z.; Yuan, Y. A brief introduction to manifold optimization. *J. Oper. Res. Soc. China* **2020**, *8*, 199–248. [CrossRef]
19. Gao, B.; Liu, X.; Chen, X.; Yuan, Y. A new first-order algorithmic framework for optimization problems with orthogonality constraints. *SIAM J. Optim.* **2018**, *28*, 302–332. [CrossRef]
20. Xiao, N.; Liu, X.; Yuan, Y. Exact penalty function for $\ell_{2,1}$ norm minimization over the Stiefel manifold. *SIAM J. Optim.* **2021**, *31*, 3097–3126. [CrossRef]
21. Xiao, N.; Liu, X.; Yuan, Y. A class of smooth exact penalty function methods for optimization problems with orthogonality constraints. *Optim. Methods Softw.* **2022**, *37*, 1205–1241. [CrossRef]
22. Xiao, N.; Liu, X. Solving optimization problems over the Stiefel manifold by smooth exact penalty functions. *J. Comput. Math.* **2024**, *42*, 1246–1276. [CrossRef]
23. Liu, X.; Xiao, N.; Yuan, Y. A penalty-free infeasible approach for a class of nonsmooth optimization problems over the Stiefel manifold. *J. Sci. Comput.* **2024**, *99*, 30. [CrossRef]
24. Xiao, N.; Liu, X.; Toh, K. Dissolving constraints for Riemannian optimization. *Math. Oper. Res.* **2024**, *49*, 366–397. [CrossRef]
25. Hu, X.; Xiao, N.; Liu, X.; Toh, K. A constraint dissolving approach for nonsmooth optimization over the Stiefel manifold. *IMA J. Numer. Anal.* **2024**, *44*, 3717–3748. [CrossRef]
26. Wang, L.; Gao, B.; Liu, X. Multipliers correction methods for optimization problems over the Stiefel manifold. *CSIAM Trans. Appl. Math.* **2021**, *2*, 508–531. [CrossRef]
27. Gao, B.; Liu, X.; Yuan, Y. Parallelizable algorithms for optimization problems with orthogonality constraints. *SIAM J. Sci. Comput.* **2019**, *41*, A1949–A1983. [CrossRef]
28. Attouch, H.; Bolte, J.; Redont, P.; Soubeyran, A. Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka–Łojasiewicz inequality. *Math. Oper. Res.* **2010**, *35*, 438–457. [CrossRef]
29. Bolte, J.; Sabach, S.; Teboulle, M. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Math. Program.* **2014**, *146*, 459–494. [CrossRef]
30. Attouch, H.; Bolte, J.; Svaiter, B.F. Convergence of descent methods for semi-algebraic and tame problems: Proximal algorithms, forward–backward splitting, and regularized Gauss–Seidel methods. *Math. Program.* **2013**, *137*, 91–129. [CrossRef]
31. Liu, H.; So, A.M.-C.; Wu, W. Quadratic optimization with orthogonality constraints: Explicit Łojasiewicz exponents and linear convergence of retraction-based line-search and stochastic variance-reduced gradient methods. *Math. Program.* **2019**, *178*, 215–262. [CrossRef]
32. Wang, P.; Liu, H.; So, A.M.-C. Linear convergence of a proximal alternating minimization method with extrapolation for ℓ_1-norm principal component analysis. *SIAM J. Optim.* **2023**, *33*, 684–712. [CrossRef]
33. Absil, P.-A.; Mahony, R.; Sepulchre, R. *Optimization Algorithms on Matrix Manifolds*; Princeton University Press: Princeton, NJ, USA, 2008.
34. Boumal, N.; Absil, P.-A.; Cartis, C. Global rates of convergence for nonconvex optimization on manifolds. *IMA J. Numer. Anal.* **2019**, *39*, 1–33. Erratum in *IMA J. Numer. Anal.* **2020**, *40*, 2940. [CrossRef]
35. Rockafellar, R.T.; Wets, R.J.-B. *Variational Analysis*; Springer: Cham, Switzerland, 2009.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.