J. Nonlinear Var. Anal. 11 (2027), No. 2, pp. 271-295
Available online at http://jnva.biemdas.com
https://doi.org/10.23952/jnva.11.2027.2.03

ESCAPING STRICT SADDLE POINTS IN NONSMOOTH NONCONVEX OPTIMIZATION WITH PROXIMAL ALTERNATING LINEARIZED MINIMIZATION

JINGWEN FAN[1], TANXING WANG[1], DEREN HAN[1,2,*]

[1]School of Mathematical Sciences, Beihang University, Beijing 100191, China
[2]School of Mathematics and Computational Science, Xiangtan University, Xiangtan 411105, China

Abstract. This paper establishes the saddle-point escaping theory for the proximal alternating linearized minimization (PALM) algorithm in the context of nonsmooth and nonconvex optimization. Under the assumptions of bounded smoothness, weakly convexity, and a separable active manifold, we prove that the PALM can almost surely escape strict saddle points, provided that the coefficients of the proximal regularization terms are sufficiently large. We explicitly derive a lower bound for these coefficients, which is a constant depending on problem-specific parameters at the optimal point. The numerical experiments are displayed to show the theoretical results. The work provides the theoretical guarantee for the PALM to escape the strict saddle points and find superior solutions in challenging nonsmooth, nonconvex, and nonseparable problems.

Keywords. Active manifold; Center stable manifold theorem; Proximal alternating linearized minimization; Strict saddle points.
2020 MSC. 90C26, 90C30, 65K05.

1. INTRODUCTION

In nonconvex optimization, a fundamental challenge is to move beyond first-order stationary points (FOSP), probably locate a more meaningful local minima and thus provide higher-quality solutions. This pursuit is critical in many machine learning and signal processing applications where the objective function's landscape is often riddled with saddle points. For some special problems like sparsity-inducing regularizers [15], there may exist good property, which means the FOSPs of the problem can be the local minima or even the global minima. However, the challenge intensifies considerably when the objective function possesses both nonsmooth and nonseparable properties. Such problems are ubiquitous, arising in tensor decomposition, blind source separation, and structured matrix recovery, inseparable interaction terms [11, 14, 29]. For this important class of problems, although alternating minimization algorithms are widely used in practice, their theoretical behavior near saddle points—specifically, whether they can escape strict saddles—remains unclear, creating a significant gap between application and theoretical assurance.

*Corresponding author.
E-mail address: handr@buaa.edu.cn (D. Han).
Received 3 June 2026; Accepted 24 August 2026; Published online 20 November 2026.
©2027 Journal of Nonlinear and Variational Analysis

We consider the nonsmooth and nonconvex optimization problem of the form

$$\min_{x\in\mathbb{R}^m, y\in\mathbb{R}^n} \quad f(x,y) := r(x)+s(y)+H(x,y) \tag{1.1}$$

under the following assumptions.

Assumption 1.1. The functions H, r and s satisfy
(a) $\inf_{\mathbb{R}^n\times\mathbb{R}^m} f > -\infty$, $\inf_{\mathbb{R}^n} r > -\infty$ and $\inf_{\mathbb{R}^m} s > -\infty$.
(b) For any fixed y the function $x \to H(x,y)$ is $C^{1,1}_{L_1(y)}$, namely the partial gradient $\nabla_x H(x,y)$ is globally Lipschitz with moduli $L_1(y)$, that is

$$\|\nabla_x H(x_1,y)-\nabla_x H(x_2,y)\| \le L_1(y)\|x_1-x_2\|, \quad \forall x_1,x_2\in\mathbb{R}^n.$$

Likewise, for any fixed x, the function $y \to H(x,y)$ is assumed to be $C^{1,1}_{L_2(x)}$.
(c) For $i=1,2$, there exists $\lambda_i^-, \lambda_i^+ > 0$ such that

$$\begin{aligned} &\inf\{L_1(y^k) : k\in\mathbb{N}\} \ge \lambda_1^- \quad \text{and} \quad \inf\{L_2(x^k) : k\in\mathbb{N}\} \ge \lambda_2^-, \\ &\sup\{L_1(y^k) : k\in\mathbb{N}\} \le \lambda_1^+ \quad \text{and} \quad \sup\{L_2(x^k) : k\in\mathbb{N}\} \le \lambda_2^+. \end{aligned}$$

(d) ∇H is Lipschitz continuous on bounded subsets of $\mathbb{R}^n\times\mathbb{R}^m$. In other words, for each bounded subsets $B_1\times B_2$ of $\mathbb{R}^n\times\mathbb{R}^m$, there exists $M>0$ such that, for all $(x_i,y_i)\in B_1\times B_2, i=1,2$,

$$\|(\nabla_x H(x_1,y_1)-\nabla_x H(x_2,y_2), \nabla_y H(x_1,y_1)-\nabla_y H(x_2,y_2))\| \le M\|(x_1-x_2,y_1-y_2)\|. \tag{1.2}$$

(e) $r(x)$ and $s(y)$ are weakly convex functions with coefficients ρ_r and ρ_s respectively.

A prevalent and effective method for solving Problem (1.1) is the Proximal Alternating Linearized Minimization (PALM) algorithm [5]. Its iteration scheme is as follows:

$$\begin{cases} x_{k+1} = \arg\min\limits_{x\in\mathbb{R}^m}\left\{r(x)+\langle\nabla_x H(\tilde{x}_k,\tilde{y}_k), x-\tilde{x}_k\rangle + \dfrac{\mu_k}{2}\|x-\tilde{x}_k\|^2\right\} & (1.3)\\ y_{k+1} = \arg\min\limits_{y\in\mathbb{R}^n}\left\{s(y)+\langle\nabla_y H(x_{k+1},\tilde{y}_k), y-\tilde{y}_k\rangle + \dfrac{\sigma_k}{2}\|y-\tilde{y}_k\|^2\right\} & (1.4)\\ (\tilde{x}_{k+1},\tilde{y}_{k+1}) = (1-\alpha)(\tilde{x}_k,\tilde{y}_k)+\alpha(x_{k+1},y_{k+1}). & (1.5) \end{cases}$$

Here, we give the damped version of the original PALM, where $\alpha\in(0,1)$ is the combination coefficient. PALM has been proven to converge to a first-order stationary point under the Kurdyka-Łojasiewicz property. However, convergence to an FOSP is insufficient for many practical applications, as saddle points can also satisfy first-order conditions. A pivotal theoretical question thus arises:

Can PALM almost surely escape strict saddle points and under what conditions?

This question constitutes the central focus of our investigation.

Lots of scholars have studied the problem of escaping the strict saddle points. There are two main catagories including first-order and second-order methods. Second-order methods include nonsmooth trust-region method and curvilinear linesearch method [3, 4, 25]. However, second-order methods require Hessian-based computations, which are expansive for large-scale problems. Therefore, research on the first-order methods for escaping almost surely (a.s.) or with a high probability saddle points has gained attention.

Extensive research develops along several main paths. For smooth problems, gradient descent with a sufficiently small step size and almost sure random initialization has been shown to avoid strict saddle points [19, 24]. Anandkumar and Ge [1] proposed higher-order algorithm to handle the denenerated saddle point and proved that escaping the third-order stationary point is a NP-hard problem. However, a ill-conditioned example is constructed to clarify that the simple gradient descent method need take exponential time ($O(e^d)$) to escape saddle points a.s., where d is the dimension of the problem [10]. To address the issue that gradient descent may take exponential time to escape saddles in ill-conditioned settings, Jin, Ge, and others introduced stochastic perturbations, leading to perturbed gradient descent methods with polynomial-time complexity [13, 16], which incorporate the random perturbation with the gradient descent and achieve a polynomial complexity ($O(\log(d)^n)$) with high probability. Some accelerated schemes were also introduced since that [17, 23, 30] and reduced the complexity about precision from $O(1/\varepsilon^2)$ to $O(1/\varepsilon^{1.75})$. Stochastic techniques [7, 18] were combined for large-scale problems. Some explore the conditions for step size under Lipschitz assumption. The bound was extended from $1/L$ to $2/L$ in [26] and [28] gave novel adaptive gradient methods to escape saddle points. The above studied the all-at-once algorithms, which means the variables are updated at one time. Some alternating methods are explored recently. The alternating minimization and proximal alternating minimization methods were proved to converge to second-order stationary points a.s. with smoothness of the objective function [20]. Also, the perturbed version of alternating methods were also proposed to reduce complexity [22]. [6] expanded this concept to the problems with manifold constraints and proposed a new perturbed Riemannian gradient descent method. However, these analyses rely fundamentally on the smoothness of the objective function.

The gradient flow for smooth problems show that there is a zero-measure initial point set who gets stuck in the saddle points. However, the landscape for nonsmooth problems is more complex. Even simple subgradient descent can converge to a strict saddle with positive probability [9]. Many scholars studied the separable objective functions and utilize the property of envelope to explore the escapism of spliting methods. For a nonsmooth problem but with C^2-smooth envelope, it has been proved that with weak convexity, D-Y spliting method, which is designed to solve three-block separable problems, can escape saddle points a.s. [21]. To overcome the challenge of C^2-smoothness requirement, Davis et. al proposed the active manifold property, with which the proximal point algorithm (PPA), proximal gradient descent and proximal linear method can escape the strict saddle points (defined in [9]). It gave less restrictive assumptions, which is the requirement for C^1-smooth envelope. In the same year, they also introduced a perturbed version for PPA and also proved the polynomial iteration complexity [8], in where they gave an approximation algorithm to calculate the inexact gradient of Moreau envelope of the objective function. Recently, D-R spliting method under nonsmooth configuration was proved to converge to SOSP a.s. [2].

A clear pattern emerges from the literature: the analysis of saddle-point escape for nonsmooth problems heavily depends on the assumption of separability in the objective function (i.e., the form $f(x,y) = g(x) + h(y)$), or on a proximal structure that is tractable for the entire objective. However, a theoretical gap remains for the more general and practical class of problems featuring a nonseparable coupling term $H(x,y)$, as in Problem (1.1), a structure central to many

tensor decomposition and matrix sensing applications. Specifically, for alternating minimization algorithms like PALM applied to this challenging nonsmooth and nonseparable setting, no theoretical guarantees for escaping strict saddle points currently exist.

To bridge this gap, we introduce the key assumption of a separable active manifold. This assumption posits that near a critical point, the active manifolds for the variables x and y are decoupled. It serves as a vital bridge connecting the nonsmooth reality of the problem with the possibility of local smooth analysis, extending the active manifold concept to the multi-block, nonseparable context. Under this assumption and standard conditions, we analyze the damped PALM algorithm.

The main contributions of this work are as follows. First, we prove that under the separable active manifold assumption, the iteration map of the damped PALM algorithm is C^1-smooth in a neighborhood of any strict saddle point. This foundational regularity result unlocks the use of powerful tools from dynamical systems theory, specifically the Center Stable Manifold Theorem. Second, we derive an explicit, problem-dependent lower bound for the proximal regularization parameters (μ_k, σ_k). We demonstrate that when these parameters exceed this bound, the Jacobian of the PALM iteration map at any strict saddle point possesses an eigenvalue with magnitude strictly greater than one, rendering the fixed point unstable. An immediate consequence of the Center Stable Manifold Theorem is that the set of initial points from which the algorithm converges to a strict saddle has measure zero; in other words, PALM almost surely escapes strict saddle points. Third, we construct a concrete example illustrating that when the regularization parameters fall below our derived threshold, PALM can indeed converge to and become trapped at a strict saddle. This confirms the practical necessity of the sufficiently large regularization coefficients required by our theory. Finally, we show that our theoretical framework extends naturally to multi-block nonsmooth and nonseparable problems. To the best of our knowledge, this work provides the first theoretical guarantee for the almost sure escape from strict saddle points for the PALM algorithm applied to general nonsmooth, nonconvex, and nonseparable problems, thereby addressing a question in the convergence analysis of alternating minimization methods.

The rest of this paper is organized as follows. Section 2 reviews necessary preliminaries. Section 3 presents our main theoretical results on the smoothness of the iteration map and the proof of saddle-point escape. It also discusses the extension to multi-block problems. Section 4 provides a numerical example. Finally, Section 5 concludes the paper.

2. PRELIMINARIES

In this section, we review essential concepts and definitions from nonsmooth optimization and dynamical systems theory that are fundamental to our analysis. We denote by $\langle \cdot, \cdot \rangle$ the standard inner product, by $\| \cdot \|$ the Euclidean norm, by $\mathrm{Fix}(T)$ the set of fixed points of an operator T, and by $\lambda_{\min}(\cdot)$ the smallest eigenvalue of a matrix.

We begin with generalized derivatives for nonsmooth functions.

Definition 2.1 (Subdifferential)**.** Let $y \xrightarrow{f} x$ denote $y \to x$ and $f(y) \to f(x)$. The (limiting) subdifferential of a function $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ at a point $x \in \mathrm{dom}\, f$ is defined as

$$\partial f(x) := \left\{ v \in \mathbb{R}^n : \exists x^t \xrightarrow{f} x,\ v^t \to v, \text{ with } \liminf_{z \to x^t} \frac{f(z) - f(x^t) - \langle v^t, z - x^t \rangle}{\|z - x^t\|} \geq 0 \text{ for each } t \right\}.$$

An element $v \in \partial f(x)$ is called a subgradient. If f is differentiable at x, then $\partial f(x) = \{\nabla f(x)\}$. If f is convex, the definition reduces to the classical subdifferential:

$$\partial f(x) = \{v \in \mathbb{R}^n : f(z) \geq f(x) + \langle v, z - x\rangle \ \ \forall z \in \mathbb{R}^n\}.$$

For the structured objective function $f(x,y) = r(x) + s(y) + H(x,y)$ in Problem (1.1), the subdifferential enjoys a separable structure when H is smooth.

Proposition 2.1 (Subdifferential of the Structured Objective)**.** *If H is continuously differentiable, then, for all $(x,y) \in \mathbb{R}^n \times \mathbb{R}^m$,*

$$\partial f(x,y) = \big(\partial_x f(x,y),\ \partial_y f(x,y)\big) = \big(\nabla_x H(x,y) + \partial r(x),\ \nabla_y H(x,y) + \partial s(y)\big).$$

Definition 2.2 (Stationary Points)**.** A point x^* is called a first-order stationary point (FOSP) of function f if $0 \in \partial f(x^*)$. If f is twice differentiable at x^*, then x^* is a strict saddle point if it is an FOSP and $\lambda_{\min}(\nabla^2 f(x^*)) < 0$.

The following concepts allow a refined second-order analysis of functions that are not twice differentiable in the classical sense.

Definition 2.3 (Subderivatives)**.** Consider $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ and a point $\bar{x}$ with $f(\bar{x})$ finite.

- The subderivative of f at $\bar{x}$ in direction $u \in \mathbb{R}^d$ is

$$df(\bar{x})(u) := \liminf_{\substack{t \searrow 0 \\ u' \to u}} \frac{f(\bar{x} + tu') - f(\bar{x})}{t}.$$

- For $\bar{v} \in \mathbb{R}^d$, the critical cone at $\bar{x}$ is $C_f(\bar{x}, \bar{v}) := \{u \in \mathbb{R}^d : \langle \bar{v}, u\rangle = df(\bar{x})(u)\}$.
- For $\bar{u} \in \operatorname{dom} df(\bar{x})$, the parabolic subderivative at $\bar{x}$ in direction $\bar{u}$ with respect to w is

$$d^2 f(\bar{x})(\bar{u} \mid w) := \liminf_{\substack{t \searrow 0 \\ w' \to w}} \frac{f(\bar{x} + t\bar{u} + \frac{1}{2}t^2 w') - f(\bar{x}) - t\, df(\bar{x})(\bar{u})}{\frac{1}{2}t^2}.$$

When f is twice differentiable, $d^2 f(\bar{x})(u \mid w) = \langle \nabla^2 f(\bar{x})u, u\rangle$.

A key property that retains much of the tractability of convex functions is weak convexity.

Definition 2.4 (Weakly Convex Functions)**.** A function $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is ρ-weakly convex for some $\rho \geq 0$ if for all $x, y \in \operatorname{dom}(f)$ and any $\lambda \in [0,1]$,

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y) + \frac{\rho\lambda(1-\lambda)}{2}\|x - y\|^2.$$

Equivalently, f is ρ-weakly convex if $f(\cdot) + \frac{\rho}{2}\|\cdot\|^2$ is convex. Weak convexity ensures the well-definedness and useful properties of the proximal operator.

Definition 2.5 (Proximal Operator)**.** For a proper, closed, and ρ-weakly convex function $\phi : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ and a parameter $\mu > \rho$, the proximal operator is defined as

$$\operatorname{prox}_{\mu^{-1}\phi}(z) := \arg\min_{x \in \mathbb{R}^d} \left\{ \phi(x) + \frac{\mu}{2}\|x - z\|^2 \right\}.$$

The condition $\mu > \rho$ guarantees the strongly convexity of the objective, making the proximal operator single-valued and Lipschitz continuous.

The iteration steps (1.3) and (1.4) of the PALM algorithm are precisely evaluations of such proximal operators.

Now, we introduce active Manifolds and strict saddles in the nonsmooth setting. The active manifold concept, introduced by Davis et al. [9], provides a powerful framework for local smooth analysis of nonsmooth functions.

Definition 2.6 (Active Manifold)**.** Consider a closed, ρ-weakly convex function $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ and a set $\mathscr{M} \subseteq \mathbb{R}^d$ containing a critical point $\bar{x}$. Then $\mathscr{M}$ is called an active C^p-manifold around $\bar{x}$ if there exists a neighborhood U of $\bar{x}$ such that:

(i) $\mathscr{M} \cap U$ is a C^p-smooth embedded submanifold of $\mathbb{R}^d$, and the restriction $f|_{\mathscr{M}\cap U}$ is C^p-smooth.

(ii) (Sharpness) $\inf\{\|v\| : v \in \partial f(x),\, x \in U \setminus \mathscr{M}\} > 0$.

The tangent space of $\mathscr{M}$ at $\bar{x}$ is denoted by $T_{\mathscr{M}}(\bar{x})$.

The sharpness condition ensures that any sequence converging to $\bar{x}$ while staying off the manifold $\mathscr{M}$ must have subgradients bounded away from zero, forcing algorithms that drive the subgradient to zero to eventually identify and land on $\mathscr{M}$. For our multi-block problem, we require a stronger condition.

Definition 2.7 (Separable Active Manifold)**.** For the function $f(x,y) = r(x) + s(y) + H(x,y)$ in (1.1), we say that it admits a separable active manifold $\mathscr{M} = \mathscr{M}_x \times \mathscr{M}_y$ around a critical point $(\bar{x},\bar{y})$ if $\mathscr{M}_x \subseteq \mathbb{R}^m$ is an active manifold for $x \mapsto f(x,\bar{y})$ at $\bar{x}$, and $\mathscr{M}_y \subseteq \mathbb{R}^n$ is an active manifold for $y \mapsto f(\bar{x},y)$ at $\bar{y}$.

This separability is crucial for the PALM algorithm, as it allows the x- and y-update subproblems to independently identify their respective active manifolds, leading to a locally smooth iteration map. On an active manifold, we can define a second-order notion of a strict saddle.

Remark 2.1 (On the Cartesian Product Structure of the Active Manifold)**.** The assumption that the active manifold takes the Cartesian product form $\mathscr{M} = \mathscr{M}_x \times \mathscr{M}_y$ might initially appear restrictive, particularly for problems where variables x and y are highly coupled. However, it is crucial to distinguish between functional coupling and the geometry of nonsmoothness.

In our objective function $f(x,y) = r(x) + s(y) + H(x,y)$, the functional coupling between x and y is entirely absorbed by the smooth term $H(x,y)$. The active manifold, on the other hand, is strictly determined by the nonsmooth regularizers $r(x)$ and $s(y)$. In practical applications, such as structured sparse tensor decomposition, constraints and nonsmooth penalties (e.g., ℓ_1-norm for core sparsity) are typically applied in a block-separable manner. Because these nondifferentiable features occur independently for each block, the resulting active manifold at a stationary point naturally forms a Cartesian product $\mathscr{M}_x \times \mathscr{M}_y$. Furthermore, if a stationary point lies in a region where both r and s are locally smooth, the active manifold trivially reduces to the full space, which evidently satisfies the Cartesian product structure.

Definition 2.8 (Strict Saddle on an Active Manifold)**.** Let f be weakly convex and $\mathscr{M}$ be a C^2-active manifold around a critical point $\bar{x}$. Then $\bar{x}$ is a strict saddle of f if there exists a direction $u \in T_{\mathscr{M}}(\bar{x})$ such that the second-order subderivative along the manifold is negative: $d^2(f|_{\mathscr{M}})(\bar{x})(u) < 0$. If $f|_{\mathscr{M}}$ is twice differentiable, this is equivalent to $u^\top \nabla^2(f|_{\mathscr{M}})(\bar{x})u < 0$.

Our analysis relies on viewing the PALM iteration as a discrete dynamical system and applying a fundamental result from dynamical systems theory.

Theorem 2.1 (Center Stable Manifold Theorem [27]). *Let the origin be a fixed point of a C^1 local diffeomorphism $T : U \to \mathbb{R}^d$, where U is a neighborhood of the origin. Let $E^s \oplus E^c \oplus E^u$ be the invariant splitting of $\mathbb{R}^d$ associated with the generalized eigenspaces of $\nabla T(0)$ corresponding to eigenvalues of modulus less than, equal to, and greater than one, respectively. Then there exists a local C^1 embedded disk W^{cs}_{loc}, tangent to $E^s \oplus E^c$ at 0, and a neighborhood B of the origin such that $T(W^{cs}_{loc}) \cap B \subseteq W^{cs}_{loc}$. Moreover, if $T^k(x) \in B$ for all $k \geq 0$, then $x \in W^{cs}_{loc}$.*

A direct corollary is that if $\nabla T(0)$ is invertible and has an eigenvalue with modulus > 1 (i.e., an unstable direction), then the set of points whose iterates remain near the fixed point has measure zero. This motivates the following definition, which links the dynamics to optimization.

Definition 2.9 (Unstable Fixed Point). A fixed point $\bar{z}$ of a map $T : \mathbb{R}^d \to \mathbb{R}^d$ is called unstable if T is C^1 smooth around $\bar{z}$ and the Jacobian $\nabla T(\bar{z})$ has at least one eigenvalue with magnitude strictly greater than one.

In our context, T will be the iteration map of (damped) PALM. The Center Stable Manifold Theorem implies that if a strict saddle point $(\bar{x}, \bar{y})$ is an unstable fixed point of T, then the set of initial points from which PALM converges to $(\bar{x}, \bar{y})$ is of measure zero, the algorithm *almost surely escapes* the saddle.

With the C^1 smoothness of the iteration map, we will use the classical Inverse Function Theorem to give the explicit expression of the differential of PALM iteration map.

Theorem 2.2 (Inverse Function Theorem). *Let $F : \mathbb{R}^d \to \mathbb{R}^d$ be a C^1 mapping. If the Jacobian matrix $JF(a)$ at a point a is invertible, then there exist open neighborhoods U of a and V of $F(a)$ such that $F : U \to V$ is a C^1 diffeomorphism (i.e., bijective with a C^1 inverse). Moreover, $J_{F^{-1}}(F(a)) = (JF(a))^{-1}$.*

Theorem 2.3 (Neumann Series). *Let X be a Banach space and $T \in \mathscr{B}(X)$ a bounded linear operator with $\|T\| < 1$. Then the operator $I - T$ is invertible, and its inverse is given by the absolutely convergent series $(I-T)^{-1} = \sum_{n=0}^{\infty} T^n$, where the convergence is in the operator norm topology.*

3. Theoretical Proof

3.1. **Convergence of the Damped PALM.** To establish the escapism of the PALM algorithm, the bijectivity of the operator need to be clarified. With the C^1 smoothness of the PALM operator, the convex combination $\alpha T_{PALM} + (1-\alpha)I$ (I is the unit map) is bijective. Therefore, we begin by establishing the convergence properties of the damped PALM algorithm defined by iterations (1.3)–(1.5). The following lemma constructs a Lyapunov function and proves its sufficient decrease, which underpins the global convergence to a stationary point.

Lemma 3.1 (Sufficient Descent of the Lyapunov Function). *Let $\{(x_k, y_k, \tilde{x}_k, \tilde{y}_k)\}$ be a sequence generated by the damped PALM algorithm under Assumption 1.1. Then, for any $\varepsilon \in (0,1)$ and $\alpha \in (0,1)$ satisfying $1-\alpha < \sqrt{2(1-\varepsilon)/(1+\varepsilon)}$, there exist positive constants $\delta_1, C_1, C_2, C_3, C_4$ such that the following inequality holds:*

$$\begin{aligned} &f(x_{k+1}, y_{k+1}) + C_1\|x_{k+1} - \tilde{x}_{k+1}\|^2 + C_2\|y_{k+1} - \tilde{y}_{k+1}\|^2 \\ &\quad \leq f(x_k, y_k) + C_3\|x_k - \tilde{x}_k\|^2 + C_4\|y_k - \tilde{y}_k\|^2. \end{aligned} \tag{3.1}$$

Specifically,

$$C_1 = C_2 = (1+\varepsilon)\delta_1 = (1+\varepsilon)\frac{(2-\alpha)(3-\alpha)\lambda^+ + (\alpha^2-8\alpha+11)M}{4(1-\alpha)(1-\varepsilon)-2(1-\alpha)^3(1+\varepsilon)}$$
$$C_3 = C_4 = (1-\varepsilon)\delta_1 = (1-\varepsilon)\frac{(2-\alpha)(3-\alpha)\lambda^+ + (\alpha^2-8\alpha+11)M}{4(1-\alpha)(1-\varepsilon)-2(1-\alpha)^3(1+\varepsilon)}.$$

Proof. The proof proceeds by combining the optimality conditions from the proximal steps, the smoothness properties of H, and the damping update.

From the optimality of x_{k+1} in (1.3), we have for any x, in particular for $x = x_k$:

$$\begin{aligned} &r(x_{k+1}) + \langle \nabla_x H(\tilde{x}_k, \tilde{y}_k), x_{k+1} - \tilde{x}_k \rangle + \frac{\mu_k}{2}\|x_{k+1} - \tilde{x}_k\|^2 \\ \leq &r(x_k) + \langle \nabla_x H(\tilde{x}_k, \tilde{y}_k), x_k - \tilde{x}_k \rangle + \frac{\mu_k}{2}\|x_k - \tilde{x}_k\|^2. \end{aligned}$$

Rearranging terms gives

$$r(x_{k+1}) \leq r(x_k) + \langle x_{k+1} - x_k, \nabla_x H(\tilde{x}_k, \tilde{y}_k) \rangle + \frac{\mu_k}{2}\left(\|x_k - \tilde{x}_k\|^2 - \|x_{k+1} - \tilde{x}_k\|^2\right). \tag{3.2}$$

Similarly, from the optimality of y_{k+1} in (1.4) with $y = y_k$:

$$s(y_{k+1}) \leq s(y_k) + \langle y_{k+1} - y_k, \nabla_y H(x_{k+1}, \tilde{y}_k) \rangle + \frac{\sigma_k}{2}\left(\|y_k - \tilde{y}_k\|^2 - \|y_{k+1} - \tilde{y}_k\|^2\right). \tag{3.3}$$

Then, we give the bound of $H(x_{k+1}, y_{k+1})$. Using the block-wise Lipschitz smoothness of H (Assumption 1.1(b)), we have

$$H(x_{k+1}, y_{k+1}) \leq H(x_{k+1}, y_k) + \langle \nabla_y H(x_{k+1}, y_k), y_{k+1} - y_k \rangle + \frac{L_2(x_{k+1})}{2}\|y_{k+1} - y_k\|^2, \tag{3.4}$$
$$H(x_{k+1}, y_k) \leq H(x_k, y_k) + \langle \nabla_x H(x_k, y_k), x_{k+1} - x_k \rangle + \frac{L_1(y_k)}{2}\|x_{k+1} - x_k\|^2. \tag{3.5}$$

Adding (3.4) and (3.5) yields:

$$\begin{aligned} H(x_{k+1}, y_{k+1}) \leq &H(x_k, y_k) + \langle \nabla_x H(x_k, y_k), x_{k+1} - x_k \rangle + \langle \nabla_y H(x_{k+1}, y_k), y_{k+1} - y_k \rangle \\ &+ \frac{L_1(y_k)}{2}\|x_{k+1} - x_k\|^2 + \frac{L_2(x_{k+1})}{2}\|y_{k+1} - y_k\|^2. \end{aligned} \tag{3.6}$$

Summing (3.2), (3.3), and the inequality above for H, we obtain:

$$\begin{aligned} f(x_{k+1}, y_{k+1}) \leq &f(x_k, y_k) + \langle x_{k+1} - x_k, \nabla_x H(\tilde{x}_k, \tilde{y}_k) - \nabla_x H(x_k, y_k) \rangle \\ &+ \langle y_{k+1} - y_k, \nabla_y H(x_{k+1}, \tilde{y}_k) - \nabla_y H(x_{k+1}, y_k) \rangle \\ &+ \frac{\mu_k}{2}\left(\|x_k - \tilde{x}_k\|^2 - \|x_{k+1} - \tilde{x}_k\|^2\right) + \frac{\sigma_k}{2}\left(\|y_k - \tilde{y}_k\|^2 - \|y_{k+1} - \tilde{y}_k\|^2\right) \\ &+ \frac{L_1(y_k)}{2}\|x_{k+1} - x_k\|^2 + \frac{L_2(x_{k+1})}{2}\|y_{k+1} - y_k\|^2. \end{aligned}$$

From the damping step (1.5), $\tilde{x}_{k+1} = (1-\alpha)\tilde{x}_k + \alpha x_{k+1}$, we have

$$\tilde{x}_{k+1} - x_{k+1} = (1-\alpha)(\tilde{x}_k - x_{k+1}) = (1-\alpha)(\tilde{x}_k - x_k) + (1-\alpha)(x_k - x_{k+1}). \tag{3.7}$$

From (3.7), $\|x_{k+1}-\tilde{x}_{k+1}\|^2 = (1-\alpha)^2\|x_{k+1}-\tilde{x}_k\|^2$, and we can derive a bound for $\|x_{k+1}-x_k\|^2$.

$$\begin{aligned}
&\|x_{k+1}-x_k\|^2\\
=&\frac{1}{(1-\alpha)^2}\|\tilde{x}_{k+1}-x_{k+1}-(1-\alpha)(\tilde{x}_k-x_k)\|^2\\
=&\frac{1}{(1-\alpha)^2}\left(\|\tilde{x}_{k+1}-x_{k+1}\|^2+(1-\alpha)^2\|\tilde{x}_k-x_k\|^2-2(1-\alpha)\langle\tilde{x}_{k+1}-x_{k+1},\tilde{x}_k-x_k\rangle\right)\\
\le&\frac{1}{(1-\alpha)^2}\left(\|\tilde{x}_{k+1}-x_{k+1}\|^2+(1-\alpha)^2\|\tilde{x}_k-x_k\|^2+2(1-\alpha)\|\tilde{x}_{k+1}-x_{k+1}\|\,\|\tilde{x}_k-x_k\|\right)\\
\le&\frac{1}{(1-\alpha)^2}\left(\|\tilde{x}_{k+1}-x_{k+1}\|^2+(1-\alpha)^2\|\tilde{x}_k-x_k\|^2+(1-\alpha)(\|\tilde{x}_{k+1}-x_{k+1}\|^2+\|\tilde{x}_k-x_k\|^2)\right)\\
=&\frac{2-\alpha}{(1-\alpha)^2}\|\tilde{x}_{k+1}-x_{k+1}\|^2+\frac{2-\alpha}{1-\alpha}\|\tilde{x}_k-x_k\|^2.
\end{aligned}$$

The first inequality uses the Cauchy-Schwarz inequality and the second one uses $2ab \le a^2+b^2$. Similarly,

$$\|y_{k+1}-y_k\|^2 \le \frac{2-\alpha}{(1-\alpha)^2}\|\tilde{y}_{k+1}-y_{k+1}\|^2+\frac{2-\alpha}{1-\alpha}\|\tilde{y}_k-y_k\|^2. \tag{3.8}$$

Next, we bound the gradient difference terms. Let M be the Lipschitz constant of ∇H on a bounded set containing the iterates (Assumption 1.1(d)). Using Young's inequality $ab \le \frac{s}{2}a^2+\frac{1}{2s}b^2$ for any $s>0$, we have

$$\begin{aligned}
\langle x_{k+1}-x_k,\nabla_x H(\tilde{x}_k,\tilde{y}_k)-\nabla_x H(x_k,y_k)\rangle &\le \frac{s_1}{2}\|x_{k+1}-x_k\|^2+\frac{1}{2s_1}\|\nabla_x H(\tilde{x}_k,\tilde{y}_k)-\nabla_x H(x_k,y_k)\|^2\\
&\le \frac{s_1}{2}\|x_{k+1}-x_k\|^2+\frac{M^2}{2s_1}\left(\|x_k-\tilde{x}_k\|^2+\|y_k-\tilde{y}_k\|^2\right).
\end{aligned}$$

The last line uses the joint Lipschitz condition (1.2). Similarly,

$$\begin{aligned}
&\langle y_{k+1}-y_k,\nabla_y H(x_{k+1},\tilde{y}_k)-\nabla_y H(x_{k+1},y_k)\rangle\\
\le&\frac{s_2}{2}\|y_{k+1}-y_k\|^2+\frac{1}{2s_2}\|\nabla_y H(x_{k+1},\tilde{y}_k)-\nabla_y H(x_{k+1},y_k)\|^2\\
\le&\frac{s_2}{2}\|y_{k+1}-y_k\|^2+\frac{M^2}{2s_2}\|y_k-\tilde{y}_k\|^2.
\end{aligned}$$

We now substitute the bounds for $\|x_{k+1}-x_k\|^2$, $\|y_{k+1}-y_k\|^2$, and the gradient difference terms into the combined inequality for $f(x_{k+1},y_{k+1})$. We also use the boundedness of the partial Lipschitz constants (Assumption 1.1(c)): $L_1(y_k) \le \lambda_1^+$, $L_2(x_{k+1}) \le \lambda_2^+$. Let $\lambda^+ =$

$\max(\lambda_1^+, \lambda_2^+)$. After extensive but straightforward algebra, we arrive at

$$\begin{aligned} f(x_{k+1}, y_{k+1}) \leq & f(x_k, y_k) + \left[\frac{\mu_k}{2} + \frac{\lambda^+}{2} \cdot \frac{2-\alpha}{1-\alpha} + \frac{s_1}{2} \cdot \frac{2-\alpha}{1-\alpha} + \frac{M^2}{2s_1}\right] \|x_k - \tilde{x}_k\|^2 \\ & + \left[\frac{\sigma_k}{2} + \frac{\lambda^+}{2} \cdot \frac{2-\alpha}{1-\alpha} + \frac{s_2}{2} \cdot \frac{2-\alpha}{1-\alpha} + \frac{M^2}{2s_1} + \frac{M^2}{2s_2}\right] \|y_k - \tilde{y}_k\|^2 \\ & - \left[\frac{\mu_k}{(1-\alpha)^2} - \frac{\lambda^+}{2} \cdot \frac{2-\alpha}{(1-\alpha)^2} - \frac{s_1}{2} \cdot \frac{2-\alpha}{(1-\alpha)^2}\right] \|x_{k+1} - \tilde{x}_{k+1}\|^2 \\ & - \left[\frac{\sigma_k}{(1-\alpha)^2} - \frac{\lambda^+}{2} \cdot \frac{2-\alpha}{(1-\alpha)^2} - \frac{s_2}{2} \cdot \frac{2-\alpha}{(1-\alpha)^2}\right] \|y_{k+1} - \tilde{y}_{k+1}\|^2. \end{aligned} \tag{3.9}$$

To obtain the desired form (3.1), we choose s_1 and s_2 to balance terms. A convenient choice is $s_1 = s_2 = \frac{(1-\alpha)M}{2-\alpha}$. Then $\frac{M^2}{s_1} = \frac{M(2-\alpha)}{1-\alpha}$. Substituting these into (3.9) yields

$$\begin{aligned} & f(x_{k+1}, y_{k+1}) + \left[\frac{\mu_k}{(1-\alpha)^2} - \frac{(2-\alpha)\lambda^+}{2(1-\alpha)^2} - \frac{M}{2(1-\alpha)}\right] \|x_{k+1} - \tilde{x}_{k+1}\|^2 \\ & + \left[\frac{\sigma_k}{(1-\alpha)^2} - \frac{(2-\alpha)\lambda^+}{2(1-\alpha)^2} - \frac{M}{2(1-\alpha)}\right] \|y_{k+1} - \tilde{y}_{k+1}\|^2 \\ \leq & f(x_k, y_k) + \left[\frac{\mu_k}{2} + \frac{(2-\alpha)\lambda^+}{2(1-\alpha)} + \frac{(3-2\alpha)M}{2(1-\alpha)}\right] \|x_k - \tilde{x}_k\|^2 \\ & + \left[\frac{\sigma_k}{2} + \frac{(2-\alpha)\lambda^+}{2(1-\alpha)} + \frac{(5-3\alpha)M}{2(1-\alpha)}\right] \|y_k - \tilde{y}_k\|^2. \end{aligned} \tag{3.10}$$

We now want to find a positive constant δ_1 and ensure the coefficients satisfy the pattern in (3.1) with $C_1 = C_2 = (1+\varepsilon)\delta_1$ and $C_3 = C_4 = (1-\varepsilon)\delta_1$ for all $\varepsilon \in (0,1)$. This requires

$$\frac{\mu_k}{(1-\alpha)^2} - \frac{(2-\alpha)\lambda^+}{2(1-\alpha)^2} - \frac{M}{2(1-\alpha)} = (1+\varepsilon)\delta_1, \tag{3.11}$$

$$\frac{\mu_k}{2} + \frac{(2-\alpha)\lambda^+}{2(1-\alpha)} + \frac{(5-3\alpha)M}{2(1-\alpha)} = (1-\varepsilon)\delta_1. \tag{3.12}$$

The expressions for σ_k are similar. Adding $\frac{1}{2}$ times (3.11) to (3.12) (after multiplying (3.12) by $\frac{1}{(1-\alpha)^2}$ to align terms) eliminates μ_k

$$\left(\frac{(2-\alpha)\lambda^+}{2(1-\alpha)^3} + \frac{(5-3\alpha)M}{2(1-\alpha)^3}\right) + \left(\frac{(2-\alpha)\lambda^+}{4(1-\alpha)^2} + \frac{M}{4(1-\alpha)}\right) = \left(\frac{1-\varepsilon}{(1-\alpha)^2} - \frac{1+\varepsilon}{2}\right)\delta_1.$$

Solving for δ_1 gives

$$\delta_1 = \frac{(2-\alpha)(3-\alpha)\lambda^+ + (\alpha^2 - 8\alpha + 11)M}{4(1-\alpha)(1-\varepsilon) - 2(1-\alpha)^3(1+\varepsilon)}. \tag{3.13}$$

For $\delta_1 > 0$, $\alpha^2 - 8\alpha + 11 > 0$ for all $\alpha \in (0,1)$. We only require the denominator positive. $4(1-\alpha)(1-\varepsilon) > 2(1-\alpha)^3(1+\varepsilon)$, which simplifies to $1-\alpha < \sqrt{\frac{2(1-\varepsilon)}{1+\varepsilon}}$. This gives the condition on α stated in the lemma.

Finally, we get the expression of C_1, C_2, C_3, C_4 from (3.13). This completes the proof, establishing the sufficient descent property with explicitly defined constants. □

Lemma 3.2 (Relative Error Bound)**.** *Under the conditions of Lemma 3.1, for the sequence generated by damped PALM, there exists a constant $C > 0$ such that*

$$\left\| \begin{pmatrix} \nabla_x H(x_{k+1}, y_{k+1}) - \nabla_x H(\tilde{x}_k, \tilde{y}_k) - \mu_k(x_{k+1} - \tilde{x}_k) \\ \nabla_y H(x_{k+1}, y_{k+1}) - \nabla_y H(x_{k+1}, \tilde{y}_k) - \sigma_k(y_{k+1} - \tilde{y}_k) \end{pmatrix} \right\| \leq C\left(\|x_{k+1} - \tilde{x}_{k+1}\| + \|y_{k+1} - \tilde{y}_{k+1}\|\right).$$

Consequently, $\operatorname{dist}(0, \partial f(x_{k+1}, y_{k+1})) \to 0$ *as* $k \to \infty$.

Proof. Define the function $\Phi(x, y, \tilde{x}, \tilde{y}) = f(x, y) + \delta_1(\|x - \tilde{x}\|^2 + \|y - \tilde{y}\|^2)$, where δ_1 is from Lemma 3.1. Its partial subdifferentials are:

$$\begin{aligned} \partial_x \Phi &= \partial r(x) + \nabla_x H(x, y) + 2\delta_1(x - \tilde{x}), \\ \partial_y \Phi &= \partial s(y) + \nabla_y H(x, y) + 2\delta_1(y - \tilde{y}), \\ \partial_{\tilde{x}} \Phi &= 2\delta_1(\tilde{x} - x), \quad \partial_{\tilde{y}} \Phi = 2\delta_1(\tilde{y} - y). \end{aligned}$$

From the optimality conditions of the proximal steps (1.3) and (1.4), we have:

$$0 \in \partial r(x_{k+1}) + \nabla_x H(\tilde{x}_k, \tilde{y}_k) + \mu_k(x_{k+1} - \tilde{x}_k), \tag{3.14}$$

$$0 \in \partial s(y_{k+1}) + \nabla_y H(x_{k+1}, \tilde{y}_k) + \sigma_k(y_{k+1} - \tilde{y}_k). \tag{3.15}$$

We now construct a specific subgradient $v \in \partial\Phi(x_{k+1}, y_{k+1}, \tilde{x}_{k+1}, \tilde{y}_{k+1})$. Let

$$\begin{aligned} v_x &= \nabla_x H(x_{k+1}, y_{k+1}) - \nabla_x H(\tilde{x}_k, \tilde{y}_k) - \mu_k(x_{k+1} - \tilde{x}_k) + 2\delta_1(x_{k+1} - \tilde{x}_{k+1}), \\ v_y &= \nabla_y H(x_{k+1}, y_{k+1}) - \nabla_y H(x_{k+1}, \tilde{y}_k) - \sigma_k(y_{k+1} - \tilde{y}_k) + 2\delta_1(y_{k+1} - \tilde{y}_{k+1}), \\ v_{\tilde{x}} &= 2\delta_1(\tilde{x}_{k+1} - x_{k+1}), \quad v_{\tilde{y}} = 2\delta_1(\tilde{y}_{k+1} - y_{k+1}). \end{aligned}$$

It is straightforward to verify, using (3.14) and (3.15), that $(v_x, v_y, v_{\tilde{x}}, v_{\tilde{y}}) \in \partial\Phi(x_{k+1}, y_{k+1}, \tilde{x}_{k+1}, \tilde{y}_{k+1})$.

Now, we bound the norm of this subgradient. Using the Lipschitz continuity of ∇H (Assumption 1.1(d)) and the identity (3.7) (and its counterpart for y), we have

$$\begin{aligned} \|\nabla_x H(x_{k+1}, y_{k+1}) - \nabla_x H(\tilde{x}_k, \tilde{y}_k)\| &\leq M\sqrt{\|x_{k+1} - \tilde{x}_k\|^2 + \|y_{k+1} - \tilde{y}_k\|^2} \\ &= M\sqrt{1/(1-\alpha)^2\|x_{k+1} - \tilde{x}_{k+1}\|^2 + 1/(1-\alpha)^2\|y_{k+1} - \tilde{y}_{k+1}\|^2} \\ &= \frac{M}{1-\alpha}\sqrt{\|x_{k+1} - \tilde{x}_{k+1}\|^2 + \|y_{k+1} - \tilde{y}_{k+1}\|^2}. \end{aligned}$$

Similarly, $\|\nabla_y H(x_{k+1}, y_{k+1}) - \nabla_y H(x_{k+1}, \tilde{y}_k)\| \leq \frac{M}{1-\alpha}\|y_{k+1} - \tilde{y}_{k+1}\|$.

Using these bounds, we can find a constant C' such that

$$\|(v_x, v_y, v_{\tilde{x}}, v_{\tilde{y}})\| \leq C'\left(\|x_{k+1} - \tilde{x}_{k+1}\| + \|y_{k+1} - \tilde{y}_{k+1}\|\right).$$

Since $(v_x, v_y) \in \partial_x\Phi \times \partial_y\Phi$ and by Proposition 2.1, the distance from (v_x, v_y) to $\partial f(x_{k+1}, y_{k+1})$ is at most $\|(v_{\tilde{x}}, v_{\tilde{y}})\| = 2\delta_1\sqrt{\|x_{k+1} - \tilde{x}_{k+1}\|^2 + \|y_{k+1} - \tilde{y}_{k+1}\|^2}$, the first result follows. The limit $\operatorname{dist}(0, \partial f(x_k, y_k)) \to 0$ is then a direct consequence of Lemma 3.1, which implies $\|x_k - \tilde{x}_k\| \to 0$ and $\|y_k - \tilde{y}_k\| \to 0$ with the lower bound of f in Assumption 1.1(a). □

With Lemma 3.1 and Lemma 3.2, we can prove the global convergence to FOSP of the sequence generated by damped PALM using the KŁproperty of the objective function, which is a standard process [12, 31], and we omit here.

3.2. **Escaping Strict Saddle Points.** We now prove the main result on escaping strict saddle points. The proof leverages the separable active manifold assumption, the Center Stable Manifold Theorem, and an analysis of the Jacobian of the PALM iteration map at a strict saddle.

First, the relationship between the fixed point of operator PALM and FOSP of the Problem (1.1) is clarified.

Proposition 3.1. *For Problem (1.1), the fixed point of PALM and the FOSP of the problem is equivalent.*

Proof. If (x^*, y^*) is the fixed point of PALM, then from iteration (1.3) and (1.4), we have

$$\partial r(x^*) + \nabla_x H(x^*, y^*) = 0$$
$$\partial s(x^*) + \nabla_y H(x^*, y^*) = 0.$$

With Proposition 2.1, the fixed point is the FOSP of Problem (1.1).

If (x^*, y^*) is the FOSP of Problem (1.1), it is straightforward that (x^*, y^*) is the solution of the form

$$x^* = \arg\min_{x\in\mathbb{R}^m} \left\{ r(x) + \langle \nabla_x H(x^*, y^*), x - x^* \rangle + \frac{\mu_k}{2} \|x - x^*\|^2 \right\}$$
$$y^* = \arg\min_{y\in\mathbb{R}^n} \left\{ s(y) + \langle \nabla_y H(x^*, y^*), y - y^* \rangle + \frac{\sigma_k}{2} \|y - y^*\|^2 \right\},$$

which is the fixed point of operator PALM. □

Using Proposition 3.1, we can analyze the fixed point to clarify the escaping of saddle points.

3.2.1. *Smoothness of the Iteration Map on the Active Manifold.* Let (x^*, y^*) be a strict saddle point and $\mathcal{M} = \mathcal{M}_x \times \mathcal{M}_y$ be the separable active manifold around it (Definition 2.7). By Lemma 3.2 and the sharpness property of the active manifold, the iterates (x_k, y_k) will eventually lie on $\mathcal{M}$ for all sufficiently large k.

Proposition 3.2 (Manifold identification)**.** *Consider the Problem (1.1), and let (x^*, y^*) be a critical point of f. Under Assumptions 1.1, for all (x, y) near (x^*, y^*), $T_{PALM}(x, y)$ lies in $\mathcal{M}$, where T_{PALM} is the operator expression for the algorithm mentioned above.*

Proof. With the conclusion in Lemma 3.2 $\operatorname{dist}(0, \partial f(x_{k+1}, y_{k+1})) \to 0$ and the C^2 smoothness of manifold, the sharpness condition implies $x_k, y_k \in \mathcal{M}$ for all sufficient large k. □

On $\mathcal{M}$, the functions r and s restrict to C^2 functions, which we denote as R_1 on $\mathcal{M}_x$ and R_2 on $\mathcal{M}_y$. The constrained PALM subproblems become smooth. We analyze the smooth iteration map $T : \mathcal{M} \to \mathcal{M}$ defined by one iteration of (damped) PALM, starting from $(x, y) \in \mathcal{M}$ near (x^*, y^*).

Consider the x-update subproblem on $\mathcal{M}_x$ for a given $(u, v) \in \mathcal{M}$:

$$T_1(u, v) = \arg\min_{x\in\mathcal{M}_x} \left\{ R_1(x) + \langle \nabla_x H(u, v), x - u \rangle + \frac{\mu}{2} \|x - u\|^2 \right\}. \tag{3.16}$$

Then, we investigate the smoothness of the operator T_{PALM}. Using Lemma 1 in [8], the smoothness of the proximal operator is clear. The next theorem states the smoothness of T_{PALM}. Here we add restricts on μ_k and σ_k to make the subproblems strong convex and have the unique solutions.

Assumption 3.1. The coefficients of regularization terms μ_k and σ_k satisfy

$$\mu_k > \rho_r, \quad \sigma_k > \rho_s. \tag{3.17}$$

Theorem 3.1 (Smooothness of iteration map T_{PALM})**.** *Consider the optimization Problem (1.1). Let (x^*, y^*) be a critical point of f. Under Assumptions 1.1 and 3.1, the map T_{PALM} is C^1-smooth.*

Proof. We set the first step of iteration is set as operation T_1, and the second step of iteration is set as operation T_2. It is known that operators T_1 and T_2 are continuous, then $T = T_2 \circ T_1$ is continuous. We prove the C^1 smoothness of operator $T_{PALM} = T_2 \circ T_1$. From the strong convexity of the subproblem and Proposition 1 in [8], we can get the C^1-smoothness for $T_1(u,v)$ with respective to u. To prove $T_1(x_k, y_k)$ is C^1 smooth for v, we need that $\tilde{x}$ varies continuously for y in the neighboor of y_k. With the iteration scheme, we have $x(y) = \text{prox}_{\mu_k^{-1} r}(x_k - \mu_k^{-1} \nabla_x H(x_k, y))$. With the Assumption 1.1(d), we can get the Lipschitz continuity of operator $x_k - \mu_k^{-1} \nabla_x H(x_k, y)$. It is known that the proximal operator is Lipschitz continous. The Lipschitz continuity of $x(y)$ for y can be obtained by the chain rule of Lipschitz continous operators. □

Let $R_1(x)$ and $R_2(x)$ are the smooth version, which are align with $r(x)$ and $s(y)$ in the manifold $\mathscr{M}_x$ and $\mathscr{M}_y$. Let $G_i(x) = 0$, $F_j(y) = 0$ define the local constraints for $\mathscr{M}_x$ and $\mathscr{M}_y$ respectively. For subproblem 1, we have the Lagrangian of subproblem 1:

$$\mathscr{L}((u,v), x, \lambda) = H(u,v) + \langle \nabla_x H(u,v), x - u \rangle + R_1(x) + \frac{\mu}{2} \|x - u\|^2 \sum_{i \geq 1} \lambda_i G_i(x).$$

At the optimal solution $\tilde{x} = T_1(u,v)$ with multipliers $\tilde{\lambda}$, the KKT conditions hold

$$\begin{cases} 0 = \nabla R_1(\tilde{x}) + \nabla_x H(u,v) + \mu(\tilde{x} - u) + \sum_i \tilde{\lambda}_i \nabla G_i(\tilde{x}), \\ 0 = G_i(\tilde{x}). \quad \forall i. \end{cases}$$

The expression w.r.t y is similar.

$$\begin{cases} 0 = \nabla R_2(\tilde{y}) + \nabla_y H(u,v) + \mu(\tilde{y} - v) + \sum_i \tilde{\beta}_j \nabla F_j(\tilde{y}), \\ 0 = G_i(\tilde{y}). \quad \forall i. \end{cases}$$

3.2.2. *Jacobian Computation and Instability Analysis.* We now compute the derivative of T at $(\tilde{x}, \tilde{y})$, which is a FOSP of the problem. Differentiating the KKT system, we obtain

$$\begin{cases} \nabla^2 R_1(\tilde{x}) \dfrac{\partial \tilde{x}}{\partial x} + \nabla^2_{xx} H(x,y) + \mu \dfrac{\partial \tilde{x}}{\partial x} - \mu I + \sum_{i \geq 1} \lambda_i \nabla^2 G_i(\tilde{x}) \dfrac{\partial \tilde{x}}{\partial x} = 0, \\ \nabla^2 R_1(\tilde{x}) \dfrac{\partial \tilde{x}}{\partial y} + \nabla^2_{xy} H(x,y) + \mu \dfrac{\partial \tilde{x}}{\partial y} + \sum_{i \geq 1} \lambda_i \nabla^2 G_i(\tilde{x}) \dfrac{\partial \tilde{x}}{\partial y} = 0, \\ \nabla^2 R_2(\tilde{y}) \dfrac{\partial \tilde{y}}{\partial x} + \nabla^2_{yx} H(\tilde{x},y) \dfrac{\partial \tilde{x}}{\partial x} + \sigma \dfrac{\partial \tilde{y}}{\partial x} + \sum_{j \geq 1} \beta_j \nabla^2 F_j(\tilde{y}) \dfrac{\partial \tilde{y}}{\partial x} = 0, \\ \nabla^2 R_2(\tilde{y}) \dfrac{\partial \tilde{y}}{\partial y} + \nabla^2_{yx} H(\tilde{x},y) \dfrac{\partial \tilde{x}}{\partial y} + \sigma \dfrac{\partial \tilde{y}}{\partial y} + \nabla^2_{yy} H(\tilde{x},y) - \sigma I + \sum_{j \geq 1} \beta_j \nabla^2 F_j(\tilde{y}) \dfrac{\partial \tilde{y}}{\partial y} = 0. \end{cases}$$

Set

$$\Gamma_1 = \begin{bmatrix} \nabla^2 R_1(\tilde{x}) + \mu I + \sum_{i\geq 1} \lambda_i \nabla^2 G_i(\tilde{x}) & 0 \\ \nabla^2_{yx} H(\tilde{x}, y) & \nabla^2 R_2(\tilde{y}) + \sigma I + \sum_{j\geq 1} \beta_j \nabla^2 F_j(\tilde{y}) \end{bmatrix},$$

$$\Gamma_2 = \begin{bmatrix} \mu I - \nabla^2_{xx} H(x,y) & -\nabla^2_{xy} H(x,y) \\ 0 & \sigma I - \nabla^2_{yy} H(\tilde{x}, y) \end{bmatrix}.$$

We have

$$\Gamma_1 \begin{bmatrix} \frac{\partial \tilde{x}}{\partial x} & \frac{\partial \tilde{x}}{\partial y} \\ \frac{\partial \tilde{y}}{\partial x} & \frac{\partial \tilde{y}}{\partial y} \end{bmatrix} = \Gamma_2.$$

Now, we need to clarify that Γ_1 is nonsingular in the tangent space. Set orthogonal projection matrices U and V, $Q = \begin{bmatrix} U & 0 \\ 0 & V \end{bmatrix}$, and $\tilde{\Gamma}_1 := Q\Gamma_1 Q$. It can be shown that $\tilde{\Gamma}_1$ is nonsingular. $\forall \begin{pmatrix} u \\ v \end{pmatrix}$ such that $\tilde{\Gamma}_1 \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, we have

$$U\left(\nabla^2 R_1(\tilde{x}) + \mu I + \sum_{i\geq 1} \lambda_i \nabla^2 G_i(\tilde{x})\right) Uu = 0,$$

$$U\nabla^2_{yx} H(\tilde{x}, y) Vu + V\left(\nabla^2 R_2(\tilde{y}) + \sigma I + \sum_{j\geq 1} \beta_j \nabla^2 F_j(\tilde{y})\right) Vv = 0.$$

$U\left(\nabla^2 R_1(\tilde{x}) + \mu I + \sum_{i\geq 1} \lambda_i \nabla^2 G_i(\tilde{x})\right) U$ is positive definite, so $u = 0$. Similarly,

$$V\left(\nabla^2 R_2(\tilde{y}) + \sigma I + \sum_{j\geq 1} \beta_j \nabla^2 F_j(\tilde{y})\right) V$$

is positive definite, so $v = 0$. Here μ and σ are required to be large enough (set as $\mu > \rho$ and $\mu > \rho$) to ensure the nonsingularity of Γ_1. The invertibility is guaranteed here. Therefore, in the tangent space, the Jacobian is $D := \begin{bmatrix} \frac{\partial \tilde{x}}{\partial x} & \frac{\partial \tilde{x}}{\partial y} \\ \frac{\partial \tilde{y}}{\partial x} & \frac{\partial \tilde{y}}{\partial y} \end{bmatrix} = (\Gamma_1)^{-1}\Gamma_2$.

Theorem 3.2 (Unstable Fixed Point at Strict Saddle Points)**.** *Consider problem* (1.1) *under the PALM algorithm. Let* (x^*, y^*) *be a strict saddle point of the problem, which corresponds to a fixed point of the PALM operator* T_{PALM}*. Then, there exists a positive threshold* $t_0 > 0$ *such that if the proximal regularization coefficients satisfy* $\mu_k > t_0$ *and* $\sigma_k > t_0$*,* (x^*, y^*) *is an unstable fixed point of* T_{PALM}*.*

Proof. We substitute (x, y) and $\tilde{x}, \tilde{y}$ with the strict saddle point (x^*, y^*), and denote

$$L := \begin{bmatrix} L_{11} & L_{12} \\ L_{21} & L_{22} \end{bmatrix}$$
$$= \begin{bmatrix} \nabla^2 R_1(x^*) + \nabla^2_{xx} H(x^*, y^*) + \sum_{i\geq 1} \lambda_i^* \nabla^2 G_i(x^*) & \nabla^2_{xy} H(x^*, y^*) \\ \nabla^2_{yx} H(x^*, y^*) & \nabla^2 R_2(y^*) + \nabla^2_{yy} H(x^*, y^*) + \sum_{j\geq 1} \beta_j^* \nabla^2 F_j(y^*) \end{bmatrix}.$$

We know that QLQ has a strictly negative eigenvalue, where Q is the orthogonal projection matrix onto the tangent space at (x^*, y^*). In the tangent space, the projected Jacobian matrix

satisfies

$$
\begin{aligned}
&QDQ\\
:=&Q\begin{bmatrix} L_{11}+\mu I-\nabla^2_{xx}H(x^*,y^*) & 0\\ L_{21} & L_{22}+\sigma I-\nabla^2_{yy}H(x^*,y^*)\end{bmatrix}^{-1}\begin{bmatrix}\mu I-\nabla^2_{xx}H(x^*,y^*) & -L_{12}\\ 0 & \sigma I-\nabla^2_{yy}H(x^*,y^*)\end{bmatrix}Q\\
=&Q\left(I-\begin{bmatrix} L_{11}+\mu I-\nabla^2_{xx}H(x^*,y^*) & 0\\ L_{21} & L_{22}+\sigma I-\nabla^2_{yy}H(x^*,y^*)\end{bmatrix}^{-1}L\right)Q.
\end{aligned}
$$

Let $S_1=\nabla^2R_1(x^*)+\sum_{i\geq 1}\lambda_i^*\nabla^2G_i(x^*)$, $S_2=\nabla^2R_2(y^*)+\sum_{j\geq 1}\beta_j^*\nabla^2F_j(y^*)$, $H_{xx}=\nabla^2_{xx}H(x^*,y^*)$, $H_{yy}=\nabla^2_{yy}H(x^*,y^*)$, and $H_{yx}=\nabla^2_{yx}H(x^*,y^*)$. Then $L_{11}=S_1+H_{xx}$, $L_{22}=S_2+H_{yy}$, and $L_{12}=H_{xy}$. By the symmetry of the Hessian, L is symmetric, which implies $L_{21}=L_{12}^T$.

Set $A=\begin{bmatrix}M & 0\\ L_{12}^T & N\end{bmatrix}$, where $M=S_1+\mu I$ (operating on $\mathrm{Im}(U)$) and $N=S_2+\sigma I$ (on $\mathrm{Im}(V)$). Given the weak convexity conditions, there exist $\mu_*,\sigma_*>0$ such that for $\mu>\mu_*$ and $\sigma>\sigma_*$, $M\succ 0$ on $\mathrm{Im}(U)$ and $N\succ 0$ on $\mathrm{Im}(V)$. We define $\kappa=\max(\mu_*,\sigma_*,0)$.

The inverse block matrix is given by $A^{-1}=\begin{bmatrix}M^{-1} & 0\\ -N^{-1}L_{12}^TM^{-1} & N^{-1}\end{bmatrix}$. To simplify the notation, we restrict all operators to the subspace $\mathrm{Im}(U)\oplus\mathrm{Im}(V)$ (where Q acts as the identity mapping), and define $\rho=\max(\|S_1\|,\|S_2\|,\kappa)$.

Since L is symmetric and (x^*,y^*) is a strict saddle point, L possesses a strictly negative eigenvalue $\alpha<0$. Thus, there exists a unit vector $w=\begin{bmatrix}p\\ q\end{bmatrix}$ with $\|w\|^2=1$ such that $Lw=\alpha w$, which implies $w^TLw=-|\alpha|<0$.

We investigate the parameter-dependent matrix operator $D(t)=I-A(t)^{-1}L$ by setting $\mu=\sigma=t>\rho$. We expand M^{-1} and N^{-1} using the Neumann series

$$M^{-1}=\frac{1}{t}I-\frac{1}{t^2}S_1+R_M(t),\quad N^{-1}=\frac{1}{t}I-\frac{1}{t^2}S_2+R_N(t),$$

where $\|R_M(t)\|\leq C_1/t^3$ and $\|R_N(t)\|\leq C_2/t^3$ for some constants $C_1,C_2>0$. Applying this expansion to the quadratic form $f(t):=w^TA^{-1}Lw$, we decompose it as $f(t)=g(w)+h(w)$, where $g(w)=w^T\,\mathrm{diag}(M^{-1},N^{-1})Lw$ and $h(w)=-q^TN^{-1}L_{12}^TM^{-1}(L_{11}p+L_{12}q)$.

Substituting the Neumann expansions into $g(w)$ yields

$$
\begin{aligned}
g(w)&=\frac{1}{t}w^TLw-\frac{1}{t^2}\left[p^TS_1(L_{11}p+L_{12}q)+q^TS_2(L_{12}^Tp+L_{22}q)\right]+\mathscr{O}(1/t^3)\\
&=-\frac{|\alpha|}{t}-\frac{\beta}{t^2}+\mathscr{O}(1/t^3),
\end{aligned}
$$

where $\beta=p^TS_1L_{11}p+p^TS_1L_{12}q+q^TS_2L_{12}^Tp+q^TS_2L_{22}q$ is bounded by a constant $|\beta|\leq\Delta:=4\rho\|L\|$.

Similarly, analyzing $h(w)$ yields the leading cross-term

$$h(w)=-\frac{1}{t^2}q^TL_{12}^T(L_{11}p+L_{12}q)+\mathscr{O}(1/t^3)=-\frac{\eta}{t^2}+\mathscr{O}(1/t^3),$$

where $\eta=q^TL_{12}^TL_{11}p+\|L_{12}q\|^2$ is bounded by $|\eta|\leq\Gamma:=2\|L\|^2$.

Combining these evaluations provides the precise asymptotic expression for the quadratic form, we get

$$f(t) = w^T A^{-1} L w = -\frac{|\alpha|}{t} - \frac{\delta}{t^2} + \mathscr{O}(1/t^3),$$

where $\delta = \beta + \eta$ is bounded by $|\delta| \leq \Delta + \Gamma$. The remainder term is rigorously bounded by K/t^3 for a specific constant $K > 0$ derived from the matrix norms.

Since $D(t)$ is non-symmetric, we decompose $D(t)$ as $D(t) = M_0(t) + E(t)$, where $M_0(t) = I - \frac{1}{t}L$ represents the unperturbed symmetric leading-order matrix, and $E(t) = \frac{1}{t}L - A^{-1}L = \mathscr{O}(1/t^2)$ acts as a non-symmetric perturbation.

Since $Lw = \alpha w$, the unperturbed symmetric matrix $M_0(t)$ has a real eigenvalue $\mu_0(t) = 1 - \frac{\alpha}{t} = 1 + \frac{|\alpha|}{t} > 1$ with the corresponding eigenvector w. According to standard non-symmetric eigenvalue perturbation theory, the first-order correction to a simple eigenvalue is determined by the term $\frac{v^T E(t) w}{v^T w}$, where v and w are the respective left and right eigenvectors of the unperturbed matrix $M_0(t)$. Crucially, because the unperturbed matrix $M_0(t)$ is symmetric , its left and right eigenvectors exactly coincide (i.e., $v = w$). Given that w is a normalized unit vector ($w^T w = 1$), their inner product in the denominator evaluates to 1. Consequently, the first-order correction seamlessly reduces to the Rayleigh quotient $w^T E(t) w$, allowing us to write the true perturbed eigenvalue $\lambda(t)$ of $D(t)$ as

$$\begin{aligned}
\lambda(t) &= \mu_0(t) + w^T E(t) w + \mathscr{O}(\|E(t)\|^2) \\
&= w^T (M_0(t) + E(t)) w + \mathscr{O}(1/t^4) \\
&= w^T D(t) w + \mathscr{O}(1/t^4) \\
&= 1 - f(t) + \mathscr{O}(1/t^4).
\end{aligned}$$

Substituting our rigorous expansion for $f(t) = -\frac{|\alpha|}{t} - \frac{\delta}{t^2} + \mathscr{O}(1/t^3)$, and absorbing the $\mathscr{O}(1/t^4)$ term into the cubic remainder, we obtain the true eigenvalue lower bound

$$\lambda(t) \geq 1 + \frac{|\alpha|}{t} - \frac{|\delta|}{t^2} - \frac{K'}{t^3},$$

where K' is a redefined absolute constant for the remainder.

To guarantee that the algorithm is locally unstable at the strict saddle point, it suffices to prove that this true eigenvalue strictly exceeds 1, which requires

$$\frac{|\alpha|}{t} > \frac{|\delta|}{t^2} + \frac{K'}{t^3} \iff |\alpha| t^2 - |\delta| t - K' > 0.$$

The positive root of this quadratic inequality is explicitly given by

$$t_+ = \frac{|\delta| + \sqrt{|\delta|^2 + 4|\alpha| K'}}{2|\alpha|}. \tag{3.18}$$

By choosing the proximal parameter threshold $t_0 = \max(\rho, t_+, \rho_r, \rho_s)$, for any $t > t_0$, the inequality $|\alpha| t^2 - |\delta| t - K' > 0$ strictly holds. Consequently, the true perturbed eigenvalue $\lambda(t)$ of the non-symmetric Jacobian $D(t)$ strictly exceeds 1.

Since $D(t)$ has a real eigenvalue with a magnitude greater than 1, the strict saddle point (x^*, y^*) is an unstable fixed point of the PALM operator. □

3.2.3. *Application of the Center Stable Manifold Theorem.* Since T is C^1 near the unstable fixed point (x^*,y^*), and its Jacobian has an eigenvalue with magnitude >1, the Center Stable Manifold Theorem (Theorem 2.1) applies. It implies that the set of initial points in a neighborhood of (x^*,y^*) whose iterates under T converge to (x^*,y^*) is contained within a local center-stable manifold, which has Lebesgue measure zero. Consequently, for random initialization, the damped PALM algorithm almost surely does not converge to the strict saddle point (x^*,y^*).

We state the main theorem consolidating these results.

Theorem 3.3 (Almost Sure Escape from Strict Saddles)**.** *Consider Problem* (1.1) *under Assumption 1.1 and the separable active manifold assumption (Definition 2.7). Let (x^*,y^*) be a strict saddle point. Then, there exists an explicit threshold $t_0>0$ given by* (3.18) *such that if the damped PALM algorithm is run with proximal parameters $\mu_k,\sigma_k>\max(t_0,\rho_r,\rho_s)$ for all sufficiently large k, and the damping factor α satisfies $1-\alpha<\sqrt{2(1-\varepsilon)/(1+\varepsilon)}$, the following holds: The set of initial points (x_0,y_0) from which the algorithm converges to (x^*,y^*) has Lebesgue measure zero. Hence, the PALM escapes the strict saddle points a.s.*

3.3. **Proof for Three-block Problems.** The theoretical framework developed for the two-block problem naturally extends to the three-block problem. Set

$$\Gamma_1=\begin{bmatrix}\nabla^2R_1(\tilde{x})+\mu I+\sum_{i\geq1}\lambda_i\nabla^2G_i(\tilde{x}) & 0 & 0\\ \nabla^2_{yx}H(\tilde{x},y,z) & \nabla^2R_2(\tilde{y})+\sigma I+\sum_{j\geq1}\beta_j\nabla^2F_j(\tilde{y}) & 0\\ \nabla^2_{zx}H(\tilde{x},\tilde{y},z) & \nabla^2_{zy}H(\tilde{x},\tilde{y},z) & \nabla^2R_3(\tilde{z})+\theta I+\sum_{l\geq1}\gamma_l\nabla^2P_l(\tilde{z})\end{bmatrix},$$

and

$$\Gamma_2=\begin{bmatrix}\mu I-\nabla^2_{xx}H(x,y,z) & -\nabla^2_{xy}H(x,y,z) & -\nabla^2_{xz}H(x,y,z)\\ 0 & \sigma I-\nabla^2_{yy}H(\tilde{x},y,z) & -\nabla^2_{yz}H(\tilde{x},y,z)\\ 0 & 0 & \theta I-\nabla^2_{zz}H(\tilde{x},\tilde{y},z)\end{bmatrix}.$$

According to the first-order optimality conditions, we have

$$\Gamma_1\begin{bmatrix}\frac{\partial\tilde{x}}{\partial x} & \frac{\partial\tilde{x}}{\partial y} & \frac{\partial\tilde{x}}{\partial z}\\ \frac{\partial\tilde{y}}{\partial x} & \frac{\partial\tilde{y}}{\partial y} & \frac{\partial\tilde{y}}{\partial z}\\ \frac{\partial\tilde{z}}{\partial x} & \frac{\partial\tilde{z}}{\partial y} & \frac{\partial\tilde{z}}{\partial z}\end{bmatrix}=\Gamma_2.$$

To ensure the iteration map is a local diffeomorphism (a prerequisite for the Center Stable Manifold Theorem), the Jacobian matrix must be invertible. With sufficiently large μ, σ, θ, in the tangent space, Γ_1 is guaranteed to be nonsingular. Furthermore, since Γ_2 is a block upper-triangular matrix, choosing sufficiently large penalty parameters ensures its diagonal blocks are strictly positive definite. Thus, Γ_2 is also nonsingular, and the Jacobian matrix is well-defined as $D:=\begin{bmatrix}\frac{\partial\tilde{x}}{\partial x} & \frac{\partial\tilde{x}}{\partial y} & \frac{\partial\tilde{x}}{\partial z}\\ \frac{\partial\tilde{y}}{\partial x} & \frac{\partial\tilde{y}}{\partial y} & \frac{\partial\tilde{y}}{\partial z}\\ \frac{\partial\tilde{z}}{\partial x} & \frac{\partial\tilde{z}}{\partial y} & \frac{\partial\tilde{z}}{\partial z}\end{bmatrix}=(\Gamma_1)^{-1}\Gamma_2$. Next, we prove that the maximum magnitude of the eigenvalues of D in the tangent space is strictly greater than 1 at any strict saddle point.

Substitute (x,y,z) and $\tilde{x},\tilde{y},\tilde{z}$ with the strict saddle point (x^*,y^*,z^*). Denote

$$L:=\begin{bmatrix}L_{11} & L_{12} & L_{13}\\ L_{21} & L_{22} & L_{23}\\ L_{31} & L_{32} & L_{33}\end{bmatrix}=\begin{bmatrix}L_{11} & \nabla^2_{xy}H^* & \nabla^2_{xz}H^*\\ \nabla^2_{yx}H^* & L_{22} & \nabla^2_{yz}H^*\\ \nabla^2_{zx}H^* & \nabla^2_{zy}H^* & L_{33}\end{bmatrix},$$

where $L_{11} = \nabla^2 R_1(x^*) + \nabla^2_{xx} H^* + \sum_{i\geq 1} \lambda_i^* \nabla^2 G_i(x^*)$, $L_{22} = \nabla^2 R_2(y^*) + \nabla^2_{yy} H^* + \sum_{j\geq 1} \beta_j^* \nabla^2 F_j(y^*)$, and $L_{33} = \nabla^2 R_3(z^*) + \nabla^2_{zz} H^* + \sum_{l\geq 1} \gamma_l^* \nabla^2 P_l(z^*)$. Let Q be the orthogonal projection matrix onto the tangent space at (x^*, y^*, z^*). By definition of a strict saddle point, the symmetric matrix QLQ possesses a strictly negative eigenvalue $\xi < 0$. Thus, there exists a unit eigenvector $w = (p,q,r)^T \in \mathrm{Im}(Q)$ ($\|w\|^2 = 1$) corresponding to ξ, such that $Lw = \xi w$.

Let $H_{xx} = \nabla^2_{xx} H^*$, $H_{yy} = \nabla^2_{yy} H^*$, $H_{zz} = \nabla^2_{zz} H^*$. We define $S_1 = L_{11} - H_{xx}$, $S_2 = L_{22} - H_{yy}$, and $S_3 = L_{33} - H_{zz}$. We set $\mu = \sigma = \theta = t$ and choose $t > \rho := \max\{\|S_1\|, \|S_2\|, \|S_3\|, \kappa\}$ to ensure the invertibility of the block components

$$M = S_1 + tI, \quad N = S_2 + tI, \quad P = S_3 + tI.$$

In the tangent space (where Q acts as the identity operator), the derivative matrix $D(t)$ evaluated at the saddle point can be compactly expressed as

$$D(t) = I - \begin{bmatrix} M & 0 & 0 \\ L_{21} & N & 0 \\ L_{31} & L_{32} & P \end{bmatrix}^{-1} L = I - \Gamma_1^{-1} L.$$

The block lower-triangular matrix Γ_1 and its exact inverse are given by

$$\Gamma_1 = \begin{bmatrix} M & 0 & 0 \\ L_{21} & N & 0 \\ L_{31} & L_{32} & P \end{bmatrix}, \quad \Gamma_1^{-1} = \begin{bmatrix} M^{-1} & 0 & 0 \\ -N^{-1} L_{21} M^{-1} & N^{-1} & 0 \\ -P^{-1} L_{31} M^{-1} + P^{-1} L_{32} N^{-1} L_{21} M^{-1} & -P^{-1} L_{32} N^{-1} & P^{-1} \end{bmatrix}.$$

Let $Lw = [u_1, u_2, u_3]^T$, where $u_1 = L_{11}p + L_{12}q + L_{13}r$, $u_2 = L_{21}p + L_{22}q + L_{23}r$, and $u_3 = L_{31}p + L_{32}q + L_{33}r$. We analyze the parameter-dependent quadratic form $f(t) := w^T \Gamma_1^{-1} Lw$. Utilizing the Neumann series expansion for $t > \rho$, we have

$$M^{-1} = \frac{1}{t} I - \frac{1}{t^2} S_1 + \mathscr{O}(t^{-3}), \quad N^{-1} = \frac{1}{t} I - \frac{1}{t^2} S_2 + \mathscr{O}(t^{-3}), \quad P^{-1} = \frac{1}{t} I - \frac{1}{t^2} S_3 + \mathscr{O}(t^{-3}).$$

Substituting these expansions into the expression for $f(t)$ and collecting terms up to $\mathscr{O}(t^{-2})$ yields

$$\begin{aligned}
p^T M^{-1} u_1 &= \frac{1}{t} p^T u_1 - \frac{1}{t^2} p^T S_1 u_1 + \mathscr{O}(t^{-3}), \\
-q^T N^{-1} L_{21} M^{-1} u_1 &= -\frac{1}{t^2} q^T L_{21} u_1 + \mathscr{O}(t^{-3}), \\
q^T N^{-1} u_2 &= \frac{1}{t} q^T u_2 - \frac{1}{t^2} q^T S_2 u_2 + \mathscr{O}(t^{-3}), \\
-r^T P^{-1} L_{31} M^{-1} u_1 &= -\frac{1}{t^2} r^T L_{31} u_1 + \mathscr{O}(t^{-3}), \\
r^T P^{-1} L_{32} N^{-1} L_{21} M^{-1} u_1 &= \mathscr{O}(t^{-3}), \\
-r^T P^{-1} L_{32} N^{-1} u_2 &= -\frac{1}{t^2} r^T L_{32} u_2 + \mathscr{O}(t^{-3}), \\
r^T P^{-1} u_3 &= \frac{1}{t} r^T u_3 - \frac{1}{t^2} r^T S_3 u_3 + \mathscr{O}(t^{-3}).
\end{aligned}$$

Summing these components, we obtain the asymptotic expansion as follows:

$$f(t) = w^T \Gamma_1^{-1} Lw = \frac{1}{t}(p^T u_1 + q^T u_2 + r^T u_3) - \frac{\beta}{t^2} + \mathscr{O}(t^{-3}) = \frac{\xi}{t} - \frac{\beta}{t^2} + \mathscr{O}(t^{-3}),$$

where

$$\beta = p^T S_1 u_1 + q^T L_{21} u_1 + q^T S_2 u_2 + r^T L_{31} u_1 + r^T L_{32} u_2 + r^T S_3 u_3.$$

Given $\|w\| = 1$ and $\|u_i\| \le \|L\|$, the coefficient β is strictly bounded in magnitude by $\max|\beta| \le \Delta := 3\rho\|L\| + 3\|L\|^2$. The higher-order remainder is uniformly bounded by $|\mathscr{O}(t^{-3})| \le K/t^3$ with $K = 3(\rho^2\|L\| + \|L\|^3)$.Notice that the perturbed Jacobian $D(t) = I - \Gamma_1^{-1} L$ is non-symmetric. However, as $t \to \infty$, the leading-order unperturbed matrix $M_0(t) = I - \frac{1}{t}L$ is perfectly symmetric. Consequently, its left and right eigenvectors associated with the eigenvalue $1 - \frac{\xi}{t}$ coincide exactly as w. According to standard non-symmetric eigenvalue perturbation theory, because $w^T w = 1$, the first-order correction to the true eigenvalue $\lambda(t)$ of $D(t)$ seamlessly reduces to the Rayleigh quotient $f(t)$. Thus, we obtain

$$\lambda(t) = 1 - f(t) + \mathscr{O}(t^{-3}) = 1 - \frac{\xi}{t} + \frac{\beta}{t^2} + \mathscr{O}(t^{-3}).$$

The rest part of the proof is similar to the proof for the two-block problem.

The proof extends straightforwardly to k blocks, which are prevalent in applications such as tensor factorization and multi-layer sparse coding. Consider the following nonsmooth, nonconvex multi-block optimization problem

$$\min_{x_1 \in \mathbb{R}^{d_1}, \ldots, x_m \in \mathbb{R}^{d_m}} \quad F(x_1, \ldots, x_m) := \sum_{i=1}^{m} r_i(x_i) + H(x_1, \ldots, x_m), \tag{3.19}$$

where each r_i is a proper, closed, and weakly convex regularizer with weak convexity coefficients ρ_{r_i}, and H is a smooth coupling function satisfying block-wise Lipschitz gradient conditions analogous to Assumption 1.1. The multi-block Proximal Alternating Linearized Minimization (PALM) algorithm updates each block sequentially

$$\begin{cases} x_i^{k+1} = \arg\min\limits_{x_i \in \mathbb{R}^{d_i}} \left\{ r_i(x_i) + \langle \nabla_{x_i} H(x_1^{k+1}, \ldots, x_{i-1}^{k+1}, \hat{x}_i^k, \cdots, \hat{x}_m^k), x_i - \hat{x}_i^{k,i} \rangle + \dfrac{\mu_i^k}{2} \|x_i - \hat{x}_i^k\|^2 \right\}, \\ \hat{x}_i^{k+1} = (1-\alpha)\hat{x}_i^k + \alpha x_i^{k+1}, \end{cases}$$

where $\alpha \in (0,1)$ is the damping parameter.

Under the natural extension of the separable active manifold assumption to multiple blocks, we establish the following result.

Theorem 3.4 (Almost Sure Escape for Multi-block PALM)**.** *Consider Problem* (3.19) *under assumptions analogous to Assumption 1.1 and the multi-block separable active manifold assumption. Let* $\mathbf{x}^* = (x_1^*, \ldots, x_m^*)$ *be a strict saddle point. Then, there exists an explicit threshold* $t_0 > 0$ *(depending on problem constants and the geometry at* $\mathbf{x}^*$*) such that if the proximal parameters* $\mu_i^k > t_0$ *for all blocks i and all sufficiently large k, and the damping factor* α *is chosen appropriately, the following holds: The set of initial points from which the multi-block damped PALM algorithm converges to* $\mathbf{x}^*$ *has Lebesgue measure zero. Consequently, with random initialization, the algorithm almost surely escapes the strict saddle points.*

3.4. **Proof Sketch.** The proof follows the same three-pillar structure as the two-block case. The key point is proving that at a strict saddle, the Jacobian has an eigenvalue exceeding one when regularization is sufficiently large, and applying the Center Stable Manifold Theorem.

The matrix L is the symmetric reduced Hessian of the Lagrangian on $\mathscr{M}$ at x^*. The strict saddle condition implies L has at least one negative eigenvalue $\xi < 0$. Thus, there exists a unit

eigenvector $w = (w_1; \ldots; w_m) \in T_{\mathscr{M}}(x^*)$ with $\|w\|^2 = 1$ corresponding to ξ, such that $Lw = \xi w$. We set all regularization parameters equal to a large value t, which means $\mu_1 = \cdots = \mu_m = t$. By choosing $t > \rho := \max_i\{\|S_i\|, \kappa\}$, we ensure the diagonal block components $M_i = S_i + tI$ are strictly positive definite, which guarantees the invertibility of the block lower-triangular matrix A. The core of the proof is to analyze the parameter-dependent quadratic form $f(t) := w^T A^{-1} L w$. Using the block lower triangular structure of A, we compute A^{-1} via forward substitution. For large t, we expand each diagonal block's inverse using the Neumann series as $M_i^{-1} = t^{-1}I - t^{-2}S_i + \mathscr{O}(t^{-3})$. Substituting these expansions and performing careful bookkeeping yields the asymptotic relation for the Rayleigh quotient

$$f(t) := w^T A^{-1} L w = \frac{\xi}{t} - \frac{\beta}{t^2} + \mathscr{O}(t^{-3}),$$

where β is a constant (depending on w, L, and the S_i) bounded by $|\beta| \le \Delta$ for some $\Delta > 0$, and the higher-order remainder $\mathscr{O}(t^{-3})$ is bounded in magnitude by K/t^3 for some $K > 0$. Crucially, while the Jacobian $D(t) = I - A^{-1}L$ is non-symmetric, its leading-order unperturbed matrix $I - \frac{1}{t}L$ is perfectly symmetric. Thus, its left and right eigenvectors associated with the eigenvalue $1 - \frac{\xi}{t}$ coincide exactly as w. According to non-symmetric eigenvalue perturbation theory, because $w^T w = 1$, the first-order correction to the true eigenvalue $\lambda(t)$ of $D(t)$ seamlessly reduces to the Rayleigh quotient $f(t)$. Therefore, the true perturbed eigenvalue is given by

$$\lambda(t) = 1 - f(t) + \mathscr{O}(t^{-3}) = 1 - \frac{\xi}{t} + \frac{\beta}{t^2} + \mathscr{O}(t^{-3}).$$

Since $\xi < 0$, the dominant term in the perturbation for large t is $-\frac{\xi}{t} > 0$. To ensure the algorithm is locally unstable at the strict saddle point, it suffices to require $\lambda(t) > 1$, which translates to

$$-\frac{\xi}{t} + \frac{\beta}{t^2} + \mathscr{O}(t^{-3}) > 0 \iff |\xi| t^2 - |\beta| t - K > 0.$$

This quadratic inequality strictly holds for all $t > t_+$, where

$$t_+ = \max\left(\rho, \frac{|\beta| + \sqrt{|\beta|^2 + 4|\xi|K}}{2|\xi|}\right).$$

With T being C^1 near the unstable fixed point x^*, the Center Stable Manifold Theorem (Theorem 2.1) implies that the set of initial points converging to x^* is contained in a local center-stable manifold, which has measure zero. Thus, multi-block damped PALM almost surely escapes the strict saddle point x^*.

The analysis for $m > 2$ blocks introduces no essential mathematical difficulty beyond increased notational complexity. The matrix A maintains a block lower triangular structure, and its inverse can be computed recursively via block forward substitution. The asymptotic expansion generalizes directly, with the leading term always being ξ/t where $\xi = \mathbf{w}^\top L \mathbf{w} < 0$. The threshold $t_0 = \max(\rho_{r_i}, t_+, \rho)$ continues to provide an explicit, problem-dependent sufficient condition for escape. The necessity of this condition is again demonstrated by examples analogous to the two-block case (see Section 4).

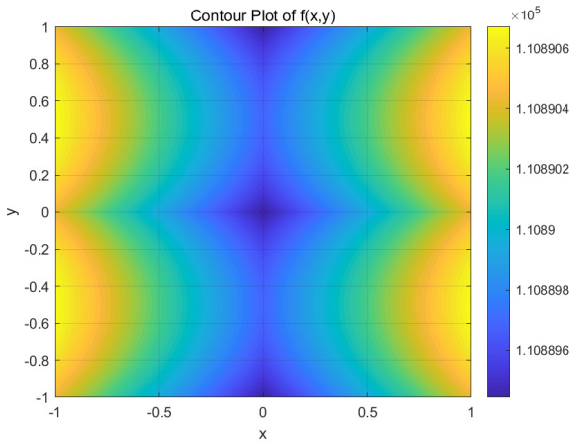


FIGURE 1. Surface plot of the example function $f(x,y)$ defined in (4.1).

4. AN EXAMPLE: VERIFYING THE NECESSITY OF LARGE REGULARIZATION

The theoretical analysis in Section 3 establishes a condition for the damped PALM algorithm to escape strict saddle points. The proximal regularization parameters μ,σ must exceed a problem-dependent threshold t_0. Here, we construct a concrete example demonstrating that when μ,σ are chosen below a certain value, PALM can indeed converge to and become trapped at a strict saddle point, even when initialized outside the (measure-zero) stable manifold associated with the saddle's subgradient flow. This example serves two purposes. It confirms the practical necessity of the theoretical condition derived in Theorem 3.3, and provides intuition about the role of regularization in stabilizing the algorithm's update.

4.1. Problem Specification and Geometric Analysis.

Consider the following nonsmooth, nonconvex, and nonseparable function

$$f(x,y) = (xy-1)^2 - y^2 + |x| + |y|. \tag{4.1}$$

Here, $r(x) = |x|$, $s(y) = |y|$, and $H(x,y) = (xy-1)^2 - y^2$. The function H is smooth in a bounded subset of (x,y), while r and s are nonsmooth (ℓ_1 norms) and weakly convex.

The point $(x^*,y^*) = (0,0.5)$ is a strict saddle point. To see this, note that on the manifold $\{x = 0\}$ (which is the active manifold for the regularizer $|x|$ at $x^* = 0$), the restricted function reduces to $f|_{\{x=0\}}(y) = 1 - y^2 + |y|$. Its derivative w.r.t. y for $y > 0$ is $-2y+1$, which vanishes at $y^* = 0.5$. Furthermore, the second derivative along this manifold at $(0,0.5)$ is $d^2(f|_{\mathscr{M}})(0,0.5) = -2 < 0$. Therefore, according to Definition 2.8, $(0,0.5)$ is a strict saddle point. The subgradient flow of this function (Fig. 2) confirms that the stable set of the saddle $(0,0.5)$ is a one-dimensional line (a set of measure zero in $\mathbb{R}^2$). $(0,0)$ is a local minimum point. We aim to see if the algorithm can escape the saddle points and reach a better solution with finite iterations.

4.2. Connecting to the Theoretical Threshold t_0.

Theorem 3.3 states that escape is guaranteed if $\mu,\sigma > t_0$, where t_0 depends on the local geometry at (x^*,y^*). For our example function (4.1), we can estimate the key quantities that constitute t_0.

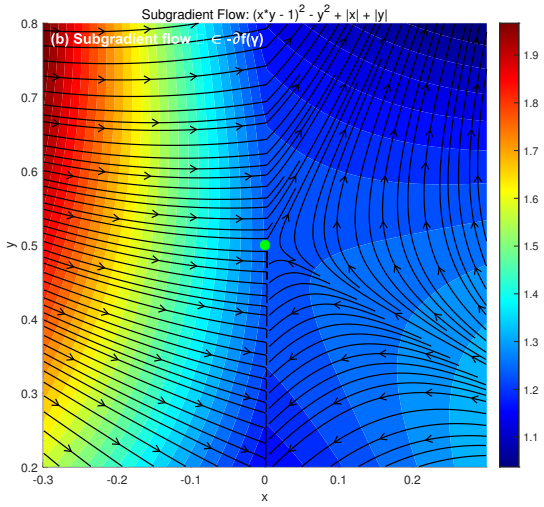
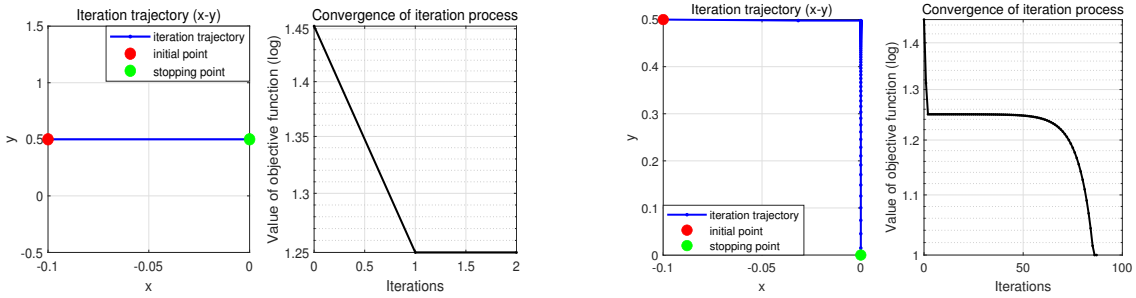


FIGURE 2. Subgradient flow of $f(x,y)$. The stable manifold leading to the saddle point $(0,0.5)$ is the line coming from a specific point in the negative orthant of x and finally to the point $(0,0.5)$, which has measure zero. Most initial points, like $(-0.1,0.5)$, are not on this manifold and should flow away from the saddle under gradient dynamics.

- **Reduced Hessian (H) and its negative eigenvalue (ξ)**: On the active manifold $\mathscr{M} = \{0\} \times \mathbb{R}$, the reduced Hessian H is simply the scalar $d^2(f|_{\mathscr{M}})(0,0.5) = -2$. Thus, $\xi = -2$ and $|\xi| = 2$.
- **Hessian of the smooth regularizer parts (S_1, S_2)**: For $r(x) = |x|$, its second derivative on the active manifold (away from $x = 0$) is 0, so $S_1 = 0$. For $s(y) = |y|$ at $y^* = 0.5 > 0$, the function is locally y, so $S_2 = 0$.
- **Partial smoothness coefficients L_1 and L_2**: In this subset, we have the partial smoothness coefficient for x is $2y^2 \leqslant 0.5$. The similar bound for y is $2x^2 \leqslant 0.02$.
- **Coefficients β and K**: The constant β involves interactions between S_i, L, and the eigenvector w. With $S_1 = S_2 = 0$, β simplifies. A detailed calculation (omitted for brevity) yields that $|\beta|$ and K are moderate constants depending on $\|L\| = 2$.
- **Smoothness coefficient M:** We control the iteration in bounded subset $[-0.1,0] \times [0,0.5]$. The Hessian matrix is $\begin{bmatrix} 2y^2 & 4xy-2 \\ 4xy-2 & 2x^2 \end{bmatrix}$. The maximum eigenvalue can be expressed as $(x^2+y^2) + \sqrt{(x^2+y^2)^2 + 12x^2y^2 - 16xy + 4}$, which is smaller than 3 in the subset.

Substituting $|\xi| = 2$ and the order-of-magnitude estimates for β and K into (3.18) suggests a finite, positive threshold t_0 on the order of 10^1 for this problem. While computing the exact t_0 requires the precise eigenvector w and is involved, the key insight is that a non-trivial lower bound exists.

4.3. **Numerical Verification: Sticking vs. Escaping.** We now test the damped PALM algorithm on Problem (4.1) with two different regularization settings, using an initial point $(-0.1,0.5)$ which is not on the stable manifold of the saddle (see Fig. 2).



(A) PALM with $\mu = \sigma = 10$. Iterates converge to the saddle $(0, 0.5)$. The objective value plateaus at the saddle value $f(0, 0.5) = 1.25$.

(B) PALM with $\mu = \sigma = 30$. Iterates escape the saddle and converge to a local minimum $(0, 0)$ with $f(0, 0) = 1$.

FIGURE 3. Trajectories of the damped PALM algorithm from the same initial point $(-0.1, 0.5)$ under different regularization strengths. These numerical results validate the necessity of the condition $\mu, \sigma > t_0$ for escaping strict saddles.

Case 1: Small Regularization ($\mu = \sigma = 10$). As shown in Fig. 3a, when the regularization parameters are set to 10 (which is below the theoretical threshold t_0 for this problem), the PALM iterates converge directly to the strict saddle point $(0, 0.5)$. The algorithm fails to escape, and the objective value stagnates at the saddle value. This confirms that the condition $\mu, \sigma > t_0$ is not merely sufficient; violating it can lead to algorithmic failure.

Case 2: Large Regularization ($\mu = \sigma = 30$). In contrast, Fig. 3b shows that when μ, σ are increased to 30, the algorithm behavior changes dramatically. The iterates diverge from the neighborhood of the saddle point $(0, 0.5)$ and eventually reaches the local minima $(0, 0)$, achieving a lower objective value. The algorithm successfully escapes the strict saddle.

4.4. **Discussion and Interpretation.** The example provides compelling evidence for the sufficiency of sufficiently large proximal regularization in PALM to ensure escape from strict saddles.

- When μ, σ is small, the contraction introduced by the proximal term is too weak to overcome the negative curvature at the saddle. The update map may have a stable direction along the saddle's descending manifold, causing the algorithm to get stuck.
- When μ, σ is large, the strong convexity from the proximal term dominates the local geometry. It makes the iteration map unstable in the direction of negative curvature, effectively "repelling" the iterates from the saddle, as predicted by our instability analysis in Section 3.2.2.

Furthermore, the observed critical value (between 10 and 30 in our experiment) aligns with the order of magnitude expected from the theoretical estimate of t_0, which depends on $|\xi| = 2$. While the precise t_0 may be difficult to compute analytically for complex problems, its existence and qualitative effect are clearly demonstrated.

This example completes the narrative of our theoretical work. Section 3 proves that large regularization is sufficient for almost sure escape. This section demonstrates, via a concrete problem, that it is also indeed neccessary. Together, they provide a robust and validated theoretical foundation for the parameter tuning of PALM in nonsmooth, nonconvex optimization. Note that our theoretical analysis reveals the threshold for escaping a saddle point is determined by

its local geometric properties. To balance convergence speed and escape capability in practice, one can initially use small coefficients to quickly locate a stationary point. Upon identifying a saddle point, its local geometry can be used to calculate the exact threshold, guiding the subsequent adjustment of the coefficients. Intuitively, it is reasonable to reduce the step size and make the next updates more carefully to navigate away from the saddle region and achieve a superior solution.

5. Conclusion

In this work, we established the theoretical guarantees for the Proximal Alternating Linearized Minimization (PALM) algorithm to escape strict saddle points in nonconvex and nonseparable optimization problems. We demonstrated that, under the assumptions of separable active manifolds and weakly convex structures, the PALM algorithm can escape strict saddle points almost surely, provided that the proximal regularization coefficients are sufficiently large. Furthermore, under the strict saddle property of the objective function, the algorithm was guaranteed to converge to a local minimum. We also provided a concrete example to illustrate the absolute necessity of these large regularization coefficients. This highlights a distinctive feature of PALM as a specialized alternating minimization scheme, contrasting with prior works that primarily rely on active manifold identification without requiring specific thresholds for the proximal parameters.

Our analysis currently assumes a uniform regularization coefficient across all coordinate blocks, resulting in a single theoretical threshold. Extending this framework to accommodate distinct coefficients for individual blocks and deriving their respective thresholds represents a compelling direction for future research. Additionally, while our proposed framework provides strict saddle point escape guarantees for a broad class of problems, we acknowledge a structural limitation regarding the coupling term $H(x,y)$. In certain practical applications, the coupling term itself might be nonsmooth. Handling a nonseparable and nonsmooth $H(x,y)$ within an alternating update scheme presents profound theoretical challenges. In future work, we plan to investigate whether exploiting the geometric properties of prox-regular functions can help decouple these perturbation bounds and extend the applicability of alternating algorithms.

Acknowledgements
The work of Deren Han was supported by the National Natural Science Foundation of China (NSFC) under Grant 12131004. The work of Tanxing Wang was supported by the China Postdoctoral Science Foundation under Grant Number 2026M794650 and the Fundamental Research Funds for the Central Universities under Grant 501QYJC2026109002.

References

[1] A. Anandkumar, R. Ge, Efficient approaches for escaping higher order saddle points in non-convex optimization, Conf. Learn. Theory (2016), 81-102.
[2] F. Atenas, Weakly convex douglas-rachford splitting avoids strict saddle points, Optim. Online (2024).
[3] D. Avdiukhin, G. Yaroslavtsev, Escaping saddle points with compressed sgd, Adv. Neural Inf. Process. Syst. 34 (2021), 10273-10284.
[4] A. Bodard, M. Ahookhosh, P. Patrinos, Second-order methods for provably escaping strict saddle points in composite nonconvex and nonsmooth optimization, arXiv preprint arXiv:2506.22332 (2025).

[5] J. Bolte, S. Sabach, M. Teboulle, Proximal alternating linearized minimization for nonconvex and nonsmooth problems, Math. Program. 146 (2014), 459-494.
[6] C. Criscitiello, N. Boumal, Efficiently escaping saddle points on manifolds, Adv. Neural Inf. Process. Syst. 32 (2019).
[7] H. Daneshmand, J. Kohler, A. Lucchi, T. Hofmann, Escaping saddles with stochastic gradients, Int. Conf. Mach. Learn. (2018), 1155-1164.
[8] D. Davis, M. Diaz, D. Drusvyatskiy, Escaping strict saddle points of the moreau envelope in nonsmooth optimization, SIAM J. Optim. 32 (2022), 1958-1983.
[9] D. Davis, D. Drusvyatskiy, Proximal methods avoid active strict saddles of weakly convex functions, Found. Comput. Math. 22 (2022), 561-606.
[10] S. S. Du, C. Jin, J. D. Lee, M. I. Jordan, A. Singh, B. Poczos, Gradient descent can take exponential time to escape saddle points, Adv. Neural Inf. Process. Syst. 30 (2017).
[11] A. Frandsen, R. Ge, Optimization landscape of tucker decomposition, Math. Program. 193 (2022), 687-712.
[12] X. Gao, X. Cai, D. Han, A Gauss–Seidel type inertial proximal alternating linearized minimization for a class of nonconvex optimization problems, J. Glob. Optim. 76 (2020), 863-887.
[13] R. Ge, F. Huang, C. Jin, Y. Yuan, Escaping from saddle points—online stochastic gradient for tensor decomposition, Conf. Learn. Theory (2015), 797-842.
[14] X. Gong, M. J. Mohlenkamp, T. R. Young, The optimization landscape for fitting a rank-2 tensor with a rank-1 tensor, SIAM J. Appl. Dyn. Syst. 17 (2018), 1432-1477.
[15] J. Guan, A. M.-C. So, l_1-norm rank-one symmetric matrix factorization has no spurious second-order stationary points, arXiv preprint arXiv:2410.05025 (2024).
[16] C. Jin, R. Ge, P. Netrapalli, S. M. Kakade, M. I. Jordan, How to escape saddle points efficiently, Int. Conf. Mach. Learn. (2017), 1724-1732.
[17] C. Jin, P. Netrapalli, M. I. Jordan, Accelerated gradient descent escapes saddle points faster than gradient descent, Conf. Learn. Theory (2018), 1042-1085.
[18] C. Jin, P. Netrapalli, R. Ge, S. M. Kakade, M. I. Jordan, On nonconvex optimization for machine learning: Gradients, stochasticity, and saddle points, J. ACM 68 (2021), 1-29.
[19] J. D. Lee, M. Simchowitz, M. I. Jordan, B. Recht, Gradient descent only converges to minimizers, Conf. Learn. Theory (2016), 1246-1257.
[20] Q. Li, Z. Zhu, G. Tang, Alternating minimizations converge to second-order optimal solutions, Int. Conf. Mach. Learn. (2019), 3935-3943.
[21] Y. Liu, W. Yin, An envelope for davis–yin splitting and strict saddle-point avoidance, J. Optim. Theory Appl. 181 (2019), 567-587.
[22] S. Lu, M. Hong, Z. Wang, Pa-gd: On the convergence of perturbed alternating gradient descent to second-order stationary points for structured nonconvex optimization, Int. Conf. Mach. Learn. (2019), 4134-4143.
[23] M. O'Neill, S. J. Wright, Behavior of accelerated gradient methods near critical points of nonconvex functions, Math. Program. 176 (2019), 403-427.
[24] I. Panageas, G. Piliouras, X. Wang, First-order methods almost always avoid saddle points: The case of vanishing step-sizes, Adv. Neural Inf. Process. Syst. 32 (2019).
[25] S. Reddi, M. Zaheer, S. Sra, B. Poczos, F. Bach, R. Salakhutdinov, A. Smola, A generic approach for escaping saddle points, Int. Conf. Artif. Intell. Stat. (2018), 1233-1242.
[26] H. Schaeffer, S. G. McCalla, Extending the step-size restriction for gradient descent to avoid strict saddle points, SIAM J. Math. Data Sci. 2 (2020), 1181-1197.
[27] M. Shub, Global stability of dynamical systems, Springer (2013).
[28] M. Staib, S. Reddi, S. Kale, S. Kumar, S. Sra, Escaping saddle points with adaptive gradient methods, Int. Conf. Mach. Learn. (2019), 5956-5965.
[29] M. Ubl, M. Hale, K. Yazdani, Linear regularizers enforce the strict saddle property, Proc. AAAI Conf. Artif. Intell. 37 (2023), 10017-10024.
[30] C. Zhang, T. Li, Escape saddle points by a simple gradient-descent based algorithm, Adv. Neural Inf. Process. Syst. 34 (2021), 8545-8556.
[31] H. Zhang, B. Wen, Z. Zha, B. Zhang, Y. Tang, G. Yu, W. Du, Accelerated PALM for nonconvex low-rank matrix recovery with theoretical analysis, IEEE Trans. Circuits Syst. Video Technol. 34 (2023), 2304-2317.