Foundations of Computational Mathematics
https://doi.org/10.1007/s10208-021-09516-w

FOUNDATIONS OF COMPUTATIONAL MATHEMATICS
The Journal of the Society for the Foundations of Computational Mathematics

Check for updates

Author Proof

Proximal Methods Avoid Active Strict Saddles of Weakly Convex Functions

Damek Davis[1] · **Dmitriy Drusvyatskiy**[2]

Received: 22 January 2020 / Revised: 16 February 2021 / Accepted: 8 April 2021
© SFoCM 2021

Abstract

We introduce a geometrically transparent strict saddle property for nonsmooth functions. This property guarantees that simple proximal algorithms on weakly convex problems converge only to local minimizers, when randomly initialized. We argue that the strict saddle property may be a realistic assumption in applications, since it provably holds for generic semi-algebraic optimization problems.

Keywords Strict saddle · Proximal gradient · Proximal point · Center stable manifold theorem · Semi-algebraic

Mathematics Subject Classification 65K05 · 65K10 · 90C30

1 Introduction

Nonconvex optimization techniques are increasingly playing a major role in modern signal processing, high-dimensional statistics, and machine learning. A driving theme, fully supported by empirical evidence, is that simple algorithms often work well in highly nonconvex and even nonsmooth settings. Gradient descent, for example, often

Communicated by Michael Overton.

D. Drusvyatskiy: Research of Drusvyatskiy was supported by the NSF DMS 1651851 and CCF 1740551 awards.

✉ Dmitriy Drusvyatskiy
ddrusv@uw.edu
http://www.sites.math.washington.edu/ ddrusv/

Damek Davis
http://www.people.orie.cornell.edu/dsd95/

1 School of Operations Research and Information Engineering, Cornell University, Ithaca, NY 14850, USA

2 Department of Mathematics, University of Washington, Seattle, WA 98195, USA

SPI Journal: 10208 Article No.: 9516 TYPESET DISK LE CP Disp.:2021/4/28 Pages: 46 Layout: Small-Ex

uncorrected proof

Author Proof

finds points with small objective value, despite existence of many highly suboptimal critical points. A growing body of literature provides one compelling explanation for this phenomenon. Namely, typical smooth objective functions provably satisfy the *strict saddle property*, meaning each critical point is either a local minimizer or has a direction of strictly negative curvature (e.g., [6,28,29,61,62]). For such functions, randomly initialized gradient-type methods provably converge to local minimizers, escaping all strict saddle points [35,51]. Moreover, stochastically perturbed gradient methods escape strict saddles efficiently, indeed, in polynomial time [22,27,33].

Smoothness of the objective plays a crucial role in the existing literature on saddle avoidance. Some extensions to constrained optimization do exist. The papers [15,27,63] investigate saddle point avoidance for the problem of minimizing a smooth functions over a smooth manifold. The works [30,44,49] propose algorithms for minimizing a smooth objective over a closed convex set. At each step of these algorithms, one must minimize a nonconvex quadratic over a certain convex set (an NP hard problem in general). The work [4] proposes a polynomial time first-order algorithm for minimizing a smooth objective over linear inequality constraints.[1] At each step of this algorithm, one identifies the "active linear constraints" and then attempts to find a "second-order stationary point" of the loss in the restricted subspace.

Though impressive in scope, existing work has yet to answer the following question:

> Do simple algorithms on typical nonsmooth and nonconvex optimization problems converge only to local minimizers?

This question as stated is purposefully vague, since "simple algorithms" and "typical optimization problems" can be interpreted in multiple ways. Let us try to formalize both ideas. First, if one believes that gradient descent is a canonical first-order method for smooth minimization, it is natural to focus on three concrete algorithms for nonsmooth and nonconvex problems: the proximal point [42,43,46,55], proximal gradient [5,48], and proximal linear [9,20,21,40,47] methods. This is the path we follow in the current work.

The latter issue, identifying a typical optimization problem, is more subtle. To motivate our approach, let us revisit the following question: why is the strict saddle property a reasonable assumption for smooth minimization? A first compelling reason is that the property holds in practice for specific problems of interest. There is, however, a more classical justification, one rooted in stability to perturbations. Namely, consider the task of minimizing a smooth function f on $\mathbb{R}^d$. Then, for a full measure set of perturbations $v \in \mathbb{R}^d$, the perturbed function $x \mapsto f(x) - \langle v, x\rangle$ is guaranteed to satisfy the strict saddle property—a direct consequence of Sard's theorem. Consequently, in a precise mathematical sense, the strict saddle property holds *generically* in smooth optimization. This justification does not suggest one can omit verification of the strict saddle property in concrete circumstances, but it does suggest that the strict saddle property is widely prevalent.

Seeking to identify a similarly reasonable class of nonsmooth objectives on which simple algorithms converge to local minimizers, the current paper accomplishes the following.

[1] This work appeared concurrently with our manuscript.

Uncorrected Proof

SPI Journal: 10208 Article No.: 9516 TYPESET DISK LE CP Disp.:2021/4/28 Pages: 46 Layout: Small-Ex

Author Proof

> We formulate natural geometric conditions, guaranteeing the proximal point, proximal gradient, and proximal linear algorithms escape all saddle points. Moreover, the proposed conditions are generic: they hold for almost all linear perturbations of weakly convex and semi-algebraic problems.

The class of optimization problems we consider is broad. A function f is called ρ-*weakly convex* if the assignment $x \mapsto f(x) + \frac{\rho}{2}\|x\|^2$ is convex for some $\rho > 0$.[2] Common examples include pointwise maxima of smooth functions and all compositions of Lipschitz convex functions with smooth maps. For detailed contemporary examples, we refer the reader to [13,16,17,23,32]. The genericity guarantee applies to semi-algebraic functions,[3] and more broadly, to those that are definable in an o-minimal structure—a virtually exhaustive function class in applications.

1.1 The Role of Curvature

To motivate our core geometric conditions, we revisit the role that curvature plays in saddle-point avoidance. Setting the stage for the rest of the paper, consider the task of minimizing a weakly convex function f on $\mathbb{R}^d$. First-order optimality conditions show that any local minimizer $\bar{x}$ of f satisfies the *criticality condition*:

$$df(\bar{x})(v) \geq 0 \quad \text{for all } v \in \mathbb{R}^d,$$

where $df(\bar{x})(v)$ denotes the directional derivative of f at $\bar{x}$ in direction v (see Definition 2.1). Conversely, sufficient conditions for local optimality at a critical point $\bar{x}$ require a close look at the second-order variations of f along particular directions, namely those where the directional derivative is zero. Mirroring the smooth setting, one may naively call a critical point $\bar{x}$ a strict saddle if there exists a direction v such that $df(\bar{x})(v) = 0$ and f decreases quadratically along v. This definition, however, is insufficient for saddle avoidance: simple examples show that typical algorithms can converge to such saddle points from a positive measure of initial conditions.

> Negative curvature alone does not guarantee escape from saddle points.

To illustrate what can go wrong, consider the example

$$\min_{(x,y)\in\mathbb{R}^2} f(x, y) = (|x| + |y|)^2 - 2x^2, \tag{1.1}$$

the graph of which is shown in Fig. 1a. First, observe that the origin meets the conditions of the candidate "strict saddle" definition. Indeed, f is differentiable at the origin and the origin is a critical point. Moreover, f decreases quadratically along all directions making a small angle with the x-axis. Next, we turn to algorithm dynamics. Figure 1b depicts the subgradient flow $-\dot{\gamma}(t) \in \partial f(\gamma(t))$. From the picture we

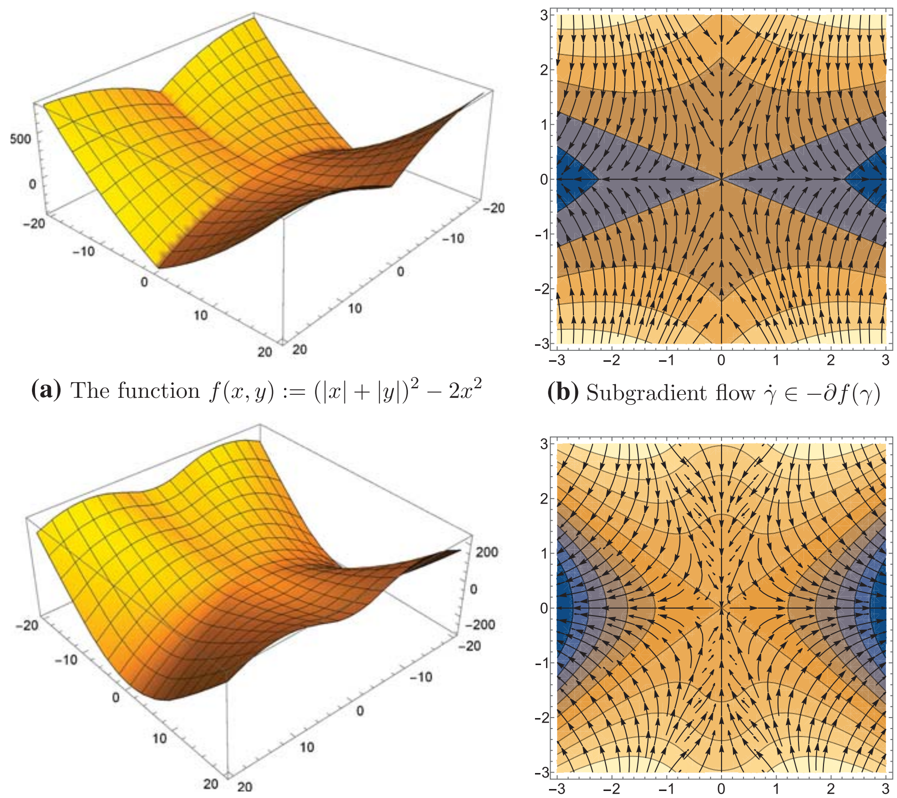
[2] Weakly convex functions also go by other names such as lower-C^2, uniformly prox-regularity, paraconvex, and semiconvex. We refer the reader to the seminal works on the topic [2,50,53,56,58].

[3] A function is called semi-algebraic if its graph decomposes into a finite union of sets, each defined by finitely many polynomial inequalities.

uncorrected proof

SPI Journal: 10208 Article No.: 9516 TYPESET DISK LE CP Disp.:2021/4/28 Pages: 46 Layout: Small-Ex

Author Proof



(a) The function $f(x, y) := (|x| + |y|)^2 - 2x^2$ **(b)** Subgradient flow $\dot{\gamma} \in -\partial f(\gamma)$

(c) Envelope $\hat{f}(z) := \min_{z' \in \mathbb{R}^2} \{f(z') + 3\|z' - z\|^2\}$ **(d)** Gradient flow $\dot{\gamma} = -\nabla \hat{f}(\gamma)$

Fig. 1 The function f in (1.1), its Moreau envelope, and their subgradient flows. **a** The function $f(x, y) := (|x| + |y|)^2 - 2x^2$. **b** Subgradient flow $\dot{\gamma} \in -\partial f(\gamma)$. **c** Envelope $\hat{f}(z) := \min_{z' \in \mathbb{R}^2} \left\{ f(z') + 3\|z' - z\|^2 \right\}$. **d** Gradient flow $\dot{\gamma} = -\nabla \hat{f}(\gamma)$

Uncorrected

find a positive measure cone, surrounding the y-axis and consisting of origin-attracted initial conditions. Moreover, we show in Appendix B that a typical algorithm—the proximal point method—initialized anywhere within this cone also converges to the origin, illustrating the inadequacy of the definition. While this argument shows that negative curvature is insufficient for nonsmooth optimization, it can be made even more definitive by smoothing the problem at hand. Namely, an alternative view of the proximal point method recognizes that the dynamics of the algorithm coincide with the dynamics of gradient descent on a C^1 smooth approximation of f, called the *Moreau envelope* (see Sect. 2.3). The smooth envelope, whose graph and gradient flow are shown in Fig. 1c, d, has the same cone of directions of second-order negative curvature as f, but despite its smoothness and negative curvature, gradient descent cannot escape the origin. The problem persists under a variety of different choices of the step-size. Note that there is no contradiction with the saddle avoidance property of gradient descent on smooth functions, since the envelope is not C^2, but merely C^1-

SPI Journal: 10208 Article No.: 9516 TYPESET DISK LE CP Disp.:2021/4/28 Pages: 46 Layout: Small-Ex

Author Proof

smooth around the origin. Although this example appears damning at first, it is highly unstable, since small linear tilts of the function do not exhibit the same pathological behavior around critical points—a direct consequence of the forthcoming results.

1.2 The Role of the Active Manifold

We have seen that negative curvature alone is insufficient for saddle avoidance. We argue here that in addition to negative curvature, one must make a structural assumption on the way nonsmoothness manifests. To illustrate and contrast with example (1.1) above, consider:

$$\min_{(x,y)\in\mathbb{R}^2} \; g(x, y) = |x| - y^2. \tag{1.2}$$

The graph of g is shown in Fig. 2a, while its subgradient flow appears in Fig. 2b. Looking at the figure, we see that the subgradient flow of g sharply contrasts with that of the pathological example (1.1). Indeed, while both functions have directions of negative curvature, the set of origin-attracted initial conditions of the flow $-\partial g$ is simply the x-axis—a measure zero set. This favorable behavior of g arises because its nonsmoothness manifests in a structured way: its unique critical point $\bar{z}$ (the origin) lies on a smooth manifold $\mathcal{M}$ (the y-axis). The function g then varies smoothly along $\mathcal{M}$ and sharply normal to $\mathcal{M}$ meaning:

$$\inf\{\|v\| : v \in \partial g(z),\; z \in U \setminus \mathcal{M}\} > 0,$$

where U is some neighborhood of $\bar{z}$. Such "active manifolds" have classical roots in optimization and serve as a far reaching generalization of "active sets" in nonlinear programming. Important references include both the original works [1,10–12,24–26] and the more recent work on identifiable surfaces [64], UV-decomposition [38], partial smoothness [39], and cone reducibility [8]. Here, we most closely follow the framework developed in [19].

1.3 Escape from Saddles by the Center Stable Manifold Theorem

Formalizing the favorable behavior of example (1.2), we will call a critical point $\bar{x}$ of a function g a *strict saddle* whenever (i) g admits an active manifold containing $\bar{x}$, and (ii) the function g decreases quadratically along some direction v satisfying $dg(\bar{x})(v) = 0$. A function g is said to have the *strict saddle property* if each of its critical points is either a local minimizer or a strict saddle.[4] Though it may seem that this definition is stringent at first, the strict saddle property is in a precise mathematical sense generic. Namely, it follows from [18] that given any semi-algebraic weakly convex function g, the perturbed function $g_v(x) = g(x) - \langle v, x\rangle$ has the strict saddle property for almost all $v \in \mathbb{R}^d$.[5] In particular, almost all linear perturbations of the function f in

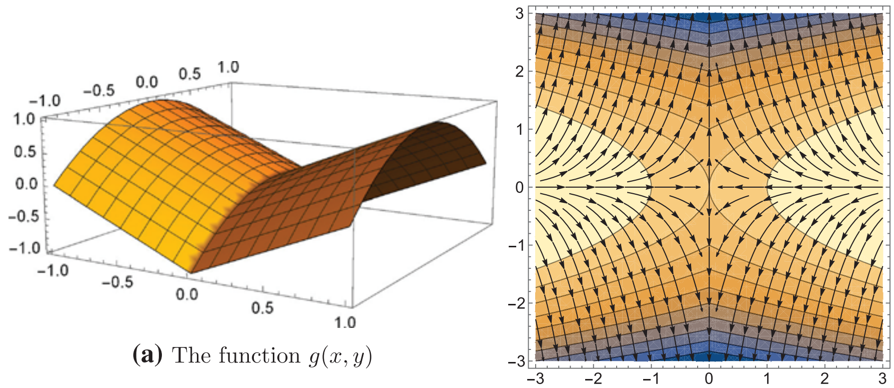
[4] Perhaps more appropriate would be the terms *active strict saddle* and the *active strict saddle property*. For brevity, we omit the word "active."

[5] Weak convexity is not essential here, provided one modifies the definitions appropriately. Moreover, this guarantee holds more generally for functions definable in an o-minimal structure.

Uncorrected Proof

SPI Journal: 10208 Article No.: 9516 TYPESET DISK LE CP Disp.:2021/4/28 Pages: 46 Layout: Small-Ex

Author Proof



(a) The function $g(x, y)$

(b) Subgradient flow $\dot{\gamma} \in -\partial g(\gamma)$

Fig. 2 The function $g(x, y) = |x| - y^2$ has an active manifold at the origin. **a** The function $g(x, y)$. **b** Subgradient flow $\dot{\gamma} \in -\partial g(\gamma)$

Table 1 The three algorithms with the update $S(x) = \operatorname{argmin}_y f_x(y)$; we assume h is convex, r is weakly convex, and both g and F are smooth

Algorithm	Objective	Update function $f_x(y)$
Prox-point	$r(x)$	$r(y) + \frac{1}{2\mu}\|y - x\|^2$
Prox-gradient	$g(x) + r(x)$	$g(x) + \langle \nabla g(x), y - x \rangle + r(y) + \frac{1}{2\mu}\|y - x\|^2$
Prox-linear	$h(F(x)) + r(x)$	$h(F(x) + \nabla F(x)(y - x)) + r(y) + \frac{1}{2\mu}\|y - x\|^2$

(1.1) do have the strict saddle property. That being said, it is important to note that under more nuanced perturbations, the strict saddle property may fail. For example, the classical NP-complete problem of checking copositivity of a matrix $A \in \mathbb{R}^{d \times d}$ amounts to verifying if the origin $\bar{x} = 0$ locally minimizes the quadratic $x^T Ax$ over the nonnegative orthant $\mathbb{R}^d_+$. It is straightforward to see that this constrained problem does not admit an active manifold at $\bar{x}$ for any matrix A.

With the definition of a strict saddle at hand, we can now outline the main results of the paper. As in the smooth setting, first explored in the seminal paper [35], our arguments will be based on the center stable manifold theorem. Namely, we will interpret the three simple minimization algorithms as fixed point iterations

$$x_{k+1} = S(x_k) \qquad \text{for some maps } S \colon \mathbb{R}^d \to \mathbb{R}^d.$$

Table 1 lists the maps $S(\cdot)$ for the proximal point, proximal gradient, and proximal linear algorithms. In each case, the fixed points of $S(\cdot)$ are precisely the critical points of the minimization problem.

To put our guarantees in context, it will be useful to recall the center stable manifold theorem. To this end, suppose that the iteration map $S(\cdot)$ is C^1-smooth on a neighborhood of some fixed point $\bar{x}$. Then, $\bar{x}$ is called an *unstable fixed point* of S if the Jacobian $\nabla S(\bar{x})$ has at least one eigenvalue whose magnitude is strictly greater than

Uncorrected Proof

SPI Journal: 10208 Article No.: 9516 TYPESET DISK LE CP Disp.:2021/4/28 Pages: 46 Layout: Small-Ex

Author Proof

one. The center stable-manifold theorem [60, Theorem III.7] guarantees the following: if $\bar{x}$ is an unstable fixed point of S and the Jacobian $\nabla S(\bar{x})$ is invertible, then almost all initializers x in a neighborhood U of $\bar{x}$ generate iterates $\{S^k(x)\}_{k\geq 1}$ that eventually escape the neighborhood. More precisely, the theorem guarantees that the set of initial conditions

$$\left\{x \in U : S^k(x) \in U \text{ for all } k \geq 1\right\}$$

has zero Lebesgue measure. All that is needed to globalize this guarantee is to ensure that the preimage $S^{-1}(V)$ of any measure zero set V is itself measure zero. Then, for almost all initial conditions $x \in \mathbb{R}^d$, the limit $\lim_{k\to\infty} S^k(x)$, when it exists, is not an unstable fixed point of S. A straightforward way to ensure that the inverse S^{-1} respects null sets is by introducing the relaxation map:

$$T(x) := (1-\alpha)x + \alpha S(x). \tag{1.3}$$

Both T and S have the same fixed points, and any fixed point $\bar{x}$ at which $\nabla S(\bar{x})$ has a real eigenvalue strictly greater than one is an unstable fixed point of T. Moreover, if the map S is Lipschitz, then the inverse T^{-1} preserves null-sets for sufficiently small $\alpha \in (0, 1)$.

1.4 The Main Results

We can now summarize our main results:

> We show that around each strict saddle of the problem, each of the iterations maps $S(\cdot)$ in Table 1 is C^1 smooth. Moreover, if $\bar{x}$ is a strict saddle, then the Jacobian $\nabla S(\bar{x})$ has a real eigenvalue strictly greater than one.

From this result, the center stable manifold theorem guarantees that iteration (1.3) locally escapes strict saddles. Seeking to globalize the guarantees, we compute the global Lipschitz constants for the proximal point and proximal gradient methods. We deduce that, when randomly initialized, the relaxed iterations (1.3) for both the proximal point and proximal gradient methods converge to local minimizers of weakly convex functions, provided they have the strict saddle property. On the other hand, without placing further restrictions on the problem data, we are unable to compute the global Lipschitz constant of the map $S(\cdot)$ corresponding to the proximal linear algorithm. We leave it as an intriguing open question to determine Lipschitz properties of the proximal linear update.

The outlined results may seem surprising at first: the optimization problem is non-smooth and yet we prove the iteration maps $S(\cdot)$ are C^1-smooth around any strict saddle. The reason is transparent and derives from the interplay between the active manifold and weak convexity. Take the proximal point method, for example. The very definition of the active manifold guarantees that the fixed point iteration $S(\cdot)$ maps an entire neighborhood $\mathcal{X}$ around an strict saddle $\bar{x}$ *into* the active manifold $\mathcal{M}$. Consequently, for all $x \in \mathcal{X}$, the update $S(x)$ can be realized as a minimizer of a smooth

SPI Journal: 10208 Article No.: 9516 TYPESET DISK LE CP Disp.:2021/4/28 Pages: 46 Layout: Small-Ex

Author Proof

function over the active manifold:

$$S(x) = \operatorname*{argmin}_{y \in \mathcal{M}} \; f(y) + \frac{1}{2\mu}\|y - x\|^2. \tag{1.4}$$

Weak convexity, in turn, ensures that $S(\bar{x})$ satisfies a quadratic growth condition for the problem (1.4), which by classical perturbation theory guarantees that $S(\cdot)$ is C^1-smooth on a neighborhood of $\bar{x}$. It only remains to argue that the negative curvature of the objective function at $\bar{x}$ implies that the Jacobian $\nabla S(\bar{x})$ has at least one real eigenvalue greater than one. Though this computation is straightforward for the proximal point method, it becomes more interesting (and surprising) for the proximal gradient and proximal linear algorithms.

Roadmap The outline of the paper is as follows. Section 2 is a self-contained presentation of the necessary preliminaries for formalizing the ideas of the introduction. Then, in Sects. 3, 4, and 5 we directly analyze the iteration maps for the proximal point, proximal gradient, and proximal linear algorithms. Section 6 establishes iterate convergence of the relaxed schemes (1.3) under the Kurdyka–Łojasiewicz property.

2 Preliminaries

Throughout, we follow standard notation in convex and variational analysis, as set out, for example, in the monographs [14,45,54,57]. We consider a Euclidean space $\mathbb{R}^d$ endowed with an inner product $\langle \cdot, \cdot \rangle$ and the induced norm $\|x\| = \sqrt{\langle x, x \rangle}$. The unit sphere in $\mathbb{R}^d$ will be denoted by $\mathbb{S}^{d-1}$. For any function $f \colon \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$, the *domain* and *epigraph* are the sets

$$\operatorname{dom} f = \{x \in \mathbb{R}^d : f(x) < \infty\}, \qquad \operatorname{epi} f = \{(x, r) \in \mathbb{R}^d \times \mathbb{R} : r \geq f(x)\},$$

respectively. The function f is called *closed* if $\operatorname{epi} f$ is a closed set. For any set $\mathcal{M} \subset \mathbb{R}^d$, the indicator function $\delta_{\mathcal{M}}$ evaluates to zero on $\mathcal{M}$ and to $+\infty$ off it. For any function $f \colon \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ and a set $\mathcal{M} \subset \mathbb{R}^d$, we define the restriction $f_{\mathcal{M}} := f + \delta_{\mathcal{M}}$. Throughout, the symbol $o(r)$ will denote any univariate function satisfying $o(r)/r \to 0$ as $r \searrow 0$.

Consider a differentiable mapping $F(x) = (F_1(x), \ldots, F_m(x))$ from $\mathbb{R}^d$ to $\mathbb{R}^m$. Throughout, the symbol $\nabla F(x) \in \mathbb{R}^{m \times d}$ will denote the Jacobian matrix, whose ij'th entry is given by $\frac{d}{dx_j} F_i(x)$. Thus, row i of $\nabla F(x)$ is the gradient of the coordinate function $F_i(x)$. In the particular case $m = 1$, we will treat $\nabla F(x)$ either as a column or as a row vector, depending on context. For a C^2-smooth function $g \colon \mathbb{R}^d \times \mathbb{R}^n \to \mathbb{R}^m$, we partition the Hessian as follows:

$$\nabla^2 g(x, y) = \begin{bmatrix} \nabla_{xx} g(x, y) & \nabla_{xy} g(x, y) \\ \nabla_{yx} g(x, y) & \nabla_{yy} g(x, y) \end{bmatrix}$$

uncorrected proof

SPI Journal: 10208 Article No.: 9516 TYPESET DISK LE CP Disp.:2021/4/28 Pages: 46 Layout: Small-Ex

Author Proof

2.1 Subdifferentials and Subderivatives

The following definition records the standard first- and second-order differential constructions, which we will use in the paper. After the definition, we will comment on the role of each construction. For further details we refer the reader to [57, Definitions 8.1, 8.3, 13.59].

Definition 2.1 (*Subdifferential and subderivatives*) Consider a function $f\colon \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ and a point $\bar{x}$ with $f(\bar{x})$ finite. Then, the *subdifferential* of f at $\bar{x}$, denoted $\partial f(\bar{x})$, consists of all vectors v satisfying

$$f(x) \geq f(\bar{x}) + \langle v, x - \bar{x}\rangle + o(\|x - \bar{x}\|) \qquad \text{as } x \to \bar{x}.$$

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

$$df(\bar{x})(\bar{u}) := \liminf_{\substack{t \searrow 0 \\ u \to \bar{u}}} \frac{f(\bar{x} + tu) - f(\bar{x})}{t}.$$

The *critical cone of f at $\bar{x}$ for $\bar{v} \in \mathbb{R}^d$* is

$$C_f(\bar{x}, \bar{v}) := \{u \in \mathbb{R}^d : \langle \bar{v}, u\rangle = df(\bar{x})(u)\}.$$

The *parabolic subderivative of f at $\bar{x}$ for $\bar{u} \in \operatorname{dom} df(\bar{x})$ with respect to $\bar{w}$* is

$$d^2 f(\bar{x})(\bar{u}|\bar{w}) = \liminf_{\substack{t \searrow 0 \\ w \to \bar{w}}} \frac{f(\bar{x} + t\bar{u} + \frac{1}{2}t^2 w) - f(\bar{x}) - df(\bar{x})(\bar{u})}{\frac{1}{2}t^2}.$$

We now comment on these definitions, in order. First, a vector v lies in the subdifferential $\partial f(\bar{x})$ precisely when the affine function $x \mapsto f(\bar{x}) + \langle v, x - \bar{x}\rangle$ minorizes f up to first order near $\bar{x}$. The definition reduces to familiar objects in classical circumstances. For example, differentiability of f at $\bar{x}$ implies the set $\partial f(\bar{x})$ is a singleton, containing only the gradient $\nabla f(\bar{x})$. Convexity of f too entails a simplification, wherein $\partial f(\bar{x})$ reduces to the subdifferential of convex analysis.

While the subdifferential encodes the set of approximate affine minorants, the subderivative measures the maximal instantaneous rate of decrease of f in direction $\bar{u}$. Like the subdifferential, the subderivative reduces to familiar objects in classical circumstances. For example, if f is locally Lipschitz at $\bar{x}$, then one may set $u = \bar{u}$ in its defining expression. Simplifying further, if f is differentiable at $\bar{x}$, we recover the directional derivative expression $df(\bar{x})(\bar{u}) = \langle \nabla f(\bar{x}), \bar{u}\rangle$. Finally, if f is convex, then the subderivative reduces to the support function of the subdifferential

$$df(\bar{x})(\bar{u}) = \sup\{\langle \bar{u}, v\rangle : v \in \partial f(\bar{x})\},$$

highlighting the dual roles of the subdifferential and subderivative constructions.

Uncorrected proof

For smooth losses, necessary optimality conditions entail vanishing gradients, while sufficient optimality conditions follow from second-order growth properties of f. Similar characterizations persist in the nonsmooth setting. In particular, the subderivative and the subdifferential feature in first-order necessary optimality conditions, where the (dual) criticality condition $0 \in \partial f(\bar{x})$ is equivalent to the (primal) nonnegativity condition

$$df(\bar{x})(u) \geq 0 \qquad \text{for all } u \in \mathbb{R}^d. \tag{2.1}$$

A point $\bar{x}$ satisfying these first-order necessary conditions (2.1) is thus called *critical* for f. Sufficient optimality conditions, on the other hand, make use of second-order variations of f. Namely, suppose that a point $\bar{x}$ is critical for f and consider a direction $\bar{u} \in \mathbb{R}^d$. There are two possibilities to consider. On the one hand, if $df(\bar{x})(\bar{u}) > 0$, then f must locally increase in direction $\bar{u}$. On the other hand, if $df(\bar{x})(\bar{u}) = 0$, then we must examine second-order variations of f to determine local optimality. Such directions of ambiguity for the subderivative make up the critical cone $C_f(\bar{x}, 0)$. For these directions, we must look to the parabolic derivative $d^2 f(\bar{x})(\bar{u}|\bar{w})$, a measurement of the second-order variation of f along a parabolic arc with tangent direction $\bar{u}$ and second-order variation $\bar{w}$. This construction too simplifies when f is C^2 smooth at $\bar{x}$, reducing to the familiar second-order variation:

$$d^2 f(\bar{x})(\bar{u}|\bar{w}) = \langle \nabla^2 f(\bar{x})\bar{u}, \bar{u} \rangle.$$

This relation suggests second-order optimality conditions for nonsmooth problems. Although we will not appeal to such conditions directly in this work, we record them here for completeness. If $\bar{x}$ is a local minimizer of f, then $df(\bar{x})(u) \geq 0$ for all $u \in \mathbb{R}^n$, and moreover $\inf_{w \in \mathbb{R}^n} d^2 f(\bar{x})(u|w) \geq 0$ for any nonzero $u \in C_f(\bar{x}, 0)$. Complementing this necessary condition, a large class of functions, those that are *parabolically regular,* may also be endowed with a sufficient optimality condition. Namely, if $df(\bar{x})(u) \geq 0$ for all $u \in \mathbb{R}^n$ and $\inf_{w \in \mathbb{R}^n} d^2 f(\bar{x})(u|w) > 0$ for any nonzero $u \in C_f(\bar{x}, 0)$, then $\bar{x}$ is a local minimizer of f. We refer the reader to [8] or [57, Theorem 13.66] for details.

2.2 Smooth Minimization on a Manifold

The main results of this work exploit local smooth features of nonsmooth optimization problems (c.f. Definition 2.6). In the presence of these features, the constructions of Definition 2.1 locally simplify. Before moving to the general setting, we thus interpret the various derivative constructions in the classical setting of minimizing a C^2-smooth function f on a C^2-smooth manifold $\mathcal{M}$. To that end, we first recall the definition of a manifold.

Definition 2.2 (*Smooth manifold*) A subset $\mathcal{M} \subset \mathbb{R}^n$ is a C^p *manifold of dimension* r *around* $\bar{x} \in \mathcal{M}$ if there is an open neighborhood U around $\bar{x}$ and a mapping G from $\mathbb{R}^n$ to $\mathbb{R}^{n-r}$ such that following hold: G is C^p-smooth, the derivative $\nabla G(\bar{x})$ has full rank, and equality holds:

$$\mathcal{M} \cap U = \{x \in U : G(x) = 0\}.$$

Author Proof
uncorrected proof

SPI Journal: 10208 Article No.: 9516 TYPESET DISK LE CP Disp.:2021/4/28 Pages: 46 Layout: Small-Ex

Author Proof

We call $G = 0$ the *local defining equations* for $\mathcal{M}$ around $\bar{x}$. The *tangent space* to $\mathcal{M}$ at $\bar{x}$ is $T_{\mathcal{M}}(\bar{x}) := \ker \nabla G(\bar{x})$ and the *normal space* to $\mathcal{M}$ at $\bar{x}$ is $N_{\mathcal{M}}(\bar{x}) := \operatorname{range} \nabla G(\bar{x})^*$.

Turning to the classical setting, consider the optimization problem

$$\min_{y \in \mathbb{R}^d} \; f(y) \quad \text{subject to } y \in \mathcal{M}. \tag{2.2}$$

Fix a point $\bar{y} \in \mathcal{M}$ and suppose that both the function f is C^2-smooth around $\bar{y}$ and $\mathcal{M}$ is a C^2-smooth manifold around $\bar{y}$. Due to local smoothness, the subdifferential admits the simple expression:

$$\partial f_{\mathcal{M}}(\bar{y}) = \nabla f(\bar{y}) + N_{\mathcal{M}}(\bar{y}).$$

Recall that we use the shorthand $f_{\mathcal{M}} := f + \delta_{\mathcal{M}}$. From this expression, we see that a point $\bar{y} \in \mathcal{M}$ is first-order critical for the problem (2.2) precisely when the inclusion holds:

$$0 \in \nabla f(\bar{y}) + N_{\mathcal{M}}(\bar{y}). \tag{2.3}$$

This inclusion can be equivalently stated in terms of the Lagrangian function. Namely, let $G = 0$ be the local defining equations for $\mathcal{M}$ around $\bar{y}$ and define the Lagrangian function

$$\mathcal{L}(y, \lambda) := f(y) + \langle G(y), \lambda \rangle.$$

Then, (2.3) amounts to existence of a (unique) multiplier vector $\bar{\lambda} \in \mathbb{R}^m$ satisfying $0 = \nabla_y \mathcal{L}(\bar{y}, \bar{\lambda})$. Next, assuming $\bar{y}$ is critical, second-order necessary conditions read

$$\langle \nabla^2_{yy} \mathcal{L}(\bar{y}, \bar{\lambda}) u, u \rangle \geq 0 \quad \text{for all } u \in T_{\mathcal{M}}(\bar{y}). \tag{2.4}$$

Conversely, second-order sufficient conditions read

$$\langle \nabla^2_{yy} \mathcal{L}(\bar{y}, \bar{\lambda}) u, u \rangle > 0 \quad \text{for all } 0 \neq u \in T_{\mathcal{M}}(\bar{y}). \tag{2.5}$$

It is well known that the sufficient condition (2.5) implies more than just local minimality; namely, (2.5) holds if and only if there exists $\alpha > 0$ such that

$$f(y) - f(\bar{y}) \geq \alpha \|y - \bar{y}\|^2, \qquad \text{for all } y \in \mathcal{M} \text{ near } \bar{y}. \tag{2.6}$$

Any point $\bar{y}$ satisfying (2.6) is called a *strong local minimizer* of f on $\mathcal{M}$.

The Lagrangian conditions (2.4) and (2.5) may be succinctly expressed through parabolic subderivatives of $f_{\mathcal{M}}(y)$, yielding a form independent of the choice of local defining equations $G = 0$. In particular, a quick computation shows that for any $u \in T_{\mathcal{M}}(\bar{y})$, the function $w \mapsto d^2 f(\bar{y})(u|w)$ is constant on its domain.[6] Dropping

[6] The domain of $d^2 f_{\mathcal{M}}(\bar{y})(u|\cdot)$ consists of w satisfying $(\langle \nabla^2 G_1(\bar{y})u, u\rangle, \ldots, \langle \nabla^2 G_{n-r}(\bar{y})u, u\rangle) = -\nabla G(\bar{y})w$, where G_i are the coordinate functions of G.

SPI Journal: 10208 Article No.: 9516 TYPESET DISK LE CP Disp.:2021/4/28 Pages: 46 Layout: Small-Ex

uncorrected proof

Author Proof

the dependence on w, the equation then holds:

$$d^2 f_{\mathcal{M}}(\bar{y})(u) = \langle \nabla^2_{yy}\mathcal{L}(\bar{y}, \bar{\lambda})u, u\rangle \qquad \text{for all } u \in T_{\mathcal{M}}(\bar{y}).$$

The use of (2.5) goes far beyond verifying local optimality; indeed, this condition plays a fundamental role in certifying solution stability under small perturbations. To illustrate, consider the value function of the parametric family

$$\varphi(x) = \inf_y \; \{f(x, y) : y \in \mathcal{M}\}, \tag{$\mathcal{P}_x$}$$

where f is C^2-smooth and $\mathcal{M} \subset \mathbb{R}^d$ is a closed set. Let $\bar{y}$ be a minimizer of $\mathcal{P}_{\bar{x}}$ for a fixed parameter $\bar{x}$, and suppose that $\mathcal{M}$ is a C^2-smooth manifold around $\bar{y}$. Let $G = 0$ be the local defining equations for $\mathcal{M}$ around $\bar{y}$ and define the parametric Lagrangian function

$$\mathcal{L}(x, y, \lambda) = f(x, y) + \langle G(y), \lambda\rangle.$$

Since $\bar{y}$ solves $\mathcal{P}_{\bar{x}}$, there is a multiplier vector $\bar{\lambda}$ satisfying $0 = \nabla_y \mathcal{L}(\bar{x}, \bar{y}, \bar{\lambda})$.

The following perturbation result will form the core of our arguments. In short: both the value function $\varphi(x)$ and the minimizer of $\mathcal{P}_x$ vary smoothly with x, provided two mild conditions hold (level-boundedness and quadratic growth). Moreover, the derivatives of both the value function and the solution maps can be computed explicitly. For details and a much more general perturbation result, see [59, Theorem 3.1].

Theorem 2.3 (Perturbation analysis) *Suppose that the following two properties hold.*

1. **(Level-boundedness)** *There exists a number $\gamma > \varphi(\bar{x})$ and a neighborhood $\mathcal{X}$ of $\bar{x}$ such that the set*

$$\bigcup_{x \in \mathcal{X}} \{y \in \mathcal{M} : f(x, y) \le \gamma\} \qquad \textit{is bounded.}$$

2. **(Quadratic growth)** *The point $\bar{y}$ is a strong local minimizer and a unique global minimizer of $\mathcal{P}_{\bar{x}}$.*

Define the partial Hessian matrices

$$H_{xx} = \nabla^2_{xx}\mathcal{L}(\bar{x}, \bar{y}, \bar{\lambda}), \qquad H_{xy} = \nabla^2_{xy}\mathcal{L}(\bar{x}, \bar{y}, \bar{\lambda}), \qquad H_{yy} = \nabla^2_{yy}\mathcal{L}(\bar{x}, \bar{y}, \bar{\lambda}),$$

and the quantities

$$\eta(h) = \min_{v \in T_{\mathcal{M}}(\bar{y})} \; \langle H_{xx}h, h\rangle + 2\langle H_{xy}v, h\rangle + \langle H_{yy}v, v\rangle,$$

$$\Phi(h) = \operatorname*{argmin}_{v \in T_{\mathcal{M}}(\bar{y})} \; \langle H_{xx}h, h\rangle + 2\langle H_{xy}v, h\rangle + \langle H_{yy}v, v\rangle.$$

uncorrected proof

SPI Journal: 10208 Article No.: 9516 TYPESET DISK LE CP Disp.:2021/4/28 Pages: 46 Layout: Small-Ex

Then, for every x near $\bar{x}$, the problem $\mathcal{P}_x$ admits a unique solution $y(x)$, which varies C^1-smoothly and admits the first-order expansion

$$\bar{y}(\bar{x}+h) = \bar{y} + \Phi(h) + o(\|h\|) \qquad \text{as } h \to 0.$$

Moreover, the function φ is C^2-smooth around $\bar{x}$ and admits the second-order expansion

$$\varphi(\bar{x}+h) = \varphi(\bar{x}) + \langle \nabla_x f(\bar{x},\bar{y}), h\rangle + \tfrac{1}{2}\eta(h) + o(\|h\|^2) \qquad \text{as } h \to 0.$$

The two assumptions of the theorem play different roles. The level-boundedness property ensures that the solutions of the perturbed problems $\mathcal{P}_x$ lie in a compact set around $\bar{y}$. The quadratic growth property in turn ensures smoothness of both the solution map and the value function. In what follows, we will apply this result several times. Both conditions will follow in all cases from the next simple lemma.

Lemma 2.4 (Sufficient conditions for level boundedness) *Consider a closed function $\varphi\colon \mathbb{R}^d \times \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ and fix a point $\bar{x} \in \mathbb{R}^d$. Suppose there exists $\alpha > 0$ such that for all x near $\bar{x}$, the function $\varphi(x,\cdot)$ is α-strongly convex and its minimizer $y(x)$ varies continuously. Then, $y(x)$ is a strong global minimizer of $\varphi(x,\cdot)$ for all x near $\bar{x}$. Moreover, there exists a neighborhood $\mathcal{X}$ of $\bar{x}$ such that for any real $\gamma > \varphi(\bar{x}, y(\bar{x}))$, the set*

$$\bigcup_{x\in\mathcal{X}} \{y \in \mathbb{R}^n : \varphi(x,y) \le \gamma\} \qquad \text{is bounded.}$$

Proof Strong convexity ensures there is a neighborhood $\mathcal{X}$ of $\bar{x}$ such that for any $x \in \mathcal{X}$, the estimate holds:

$$\varphi(x, y(x)) + \frac{\alpha}{2}\|y - y(x)\|^2 \le \varphi(x,y) \qquad \forall y \in \mathbb{R}^n, \tag{2.7}$$

showing $y(x)$ is a strong global minimizer of $\varphi(x,\cdot)$. Shrinking $\mathcal{X}$ if necessary, we may assume that $y(\cdot)$ also varies continuously on $\mathcal{X}$. Choose any $\delta > 0$. Then, by shrinking $\mathcal{X}$ again and by leveraging both closedness of φ and continuity of y, we may ensure that

$$\|y(x) - y(\bar{x})\| \le \delta \qquad \text{and} \qquad \varphi(x, y(x)) \ge \varphi(\bar{x}, y(\bar{x})) - \delta \qquad \text{for all } x \in \mathcal{X}. \tag{2.8}$$

The proof will now follow quickly from the bound (2.8). Indeed, consider any points $x \in \mathcal{X}$ and $y \in \mathbb{R}^d$ satisfying $\varphi(x,y) \le \gamma$. Then, (2.7) yields

$$\|y - y(x)\| \le \sqrt{\frac{2(\gamma - \varphi(x, y(x)))}{\alpha}}.$$

Applying (2.8) then gives the uniform bound

Uncorrected Proof
Author Proof

SPI Journal: 10208 Article No.: 9516 TYPESET DISK LE CP Disp.:2021/4/28 Pages: 46 Layout: Small-Ex

Author Proof

$$\|y - y(\bar{x})\| \leq \|y(x) - y(\bar{x})\| + \sqrt{\frac{2(\gamma - \varphi(x, y(x)))}{\alpha}} \leq \delta + \sqrt{\frac{2(\gamma + \delta - \varphi(\bar{x}, y(\bar{x})))}{\alpha}},$$

completing the proof. □

2.3 Weak Convexity and the Moreau Envelope

In general, the little-o error term in the definition of $\partial f(\bar{x})$ (Definition 2.1) may depend both on the base point $\bar{x}$ and on the subgradient v. In this work, we focus on a particular class of functions for which the error in approximation is uniform. Namely, we focus on the class of ρ-*weakly convex* functions $f: \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$, meaning those for which the assignment $x \mapsto f(x) + \frac{\rho}{2}\|x\|^2$ defines a convex function. Subgradients of a ρ-weakly convex function f automatically yield a uniform lower bound:

$$f(y) \geq f(x) + \langle v, y - x\rangle - \frac{\rho}{2}\|y - x\|^2, \qquad \forall x, y \in \mathbb{R}^d,\ v \in \partial f(x). \tag{2.9}$$

A useful feature of weakly convex functions is that they admit a smooth approximation that preserves critical points. Setting the notation, fix a ρ-weakly convex function $f: \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ and a parameter $\mu < \rho^{-1}$. Define the *Moreau envelope* and the *proximal point* map, respectively:

$$f_\mu(x) = \inf_{y \in \mathbb{R}^d} \left\{ f(y) + \frac{1}{2\mu}\|y - x\|^2 \right\},$$

$$\text{prox}_{\mu f}(x) = \operatorname*{argmin}_{y \in \mathbb{R}^d} \left\{ f(y) + \frac{1}{2\mu}\|y - x\|^2 \right\}.$$

We will use a few basic properties of these two constructions, summarized below.

Lemma 2.5 (Moreau envelope and the proximal point map) *Consider a ρ-weakly convex function $f: \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ and fix a parameter $\mu < \rho^{-1}$. Then, the following are true.*

1. *The envelope f_μ is C^1-smooth with its gradient given by*

$$\nabla f_\mu(x) = \mu^{-1}(x - \text{prox}_{\mu f}(x)). \tag{2.10}$$

2. *The envelope $f_\mu(\cdot)$ is μ^{-1}-smooth and $\frac{\rho}{1-\mu\rho}$-weakly convex meaning:*

$$-\frac{\rho}{2(1 - \mu\rho)}\|x' - x\|^2 \leq f_\mu(x') - f_\mu(x) - \langle \nabla f_\mu(x), x' - x\rangle \leq \frac{1}{2\mu}\|x' - x\|^2, \tag{2.11}$$

for all $x, x' \in \mathbb{R}^d$.

3. *The proximal map $\text{prox}_{\mu f}(\cdot)$ is $\frac{1}{1-\mu\rho}$-Lipschitz continuous and the gradient map ∇f_μ is Lipschitz continuous with constant $\max\{\mu^{-1}, \frac{\rho}{1-\mu\rho}\}$.*

SPI Journal: 10208 Article No.: 9516 TYPESET DISK LE CP Disp.:2021/4/28 Pages: 46 Layout: Small-Ex

uncorrected proof

Author Proof

4. *The critical points of* f *and* f_μ *coincide. In particular, they are exactly the fixed points of the proximal map* $\mathrm{prox}_{\mu f}$.

Proof Claim 1 follows, for example, from [53, Theorem 4.4]. The left-hand side of (2.11) is proved in [53, Theorem 5.2]. To see the right-hand side, observe

$$
\begin{aligned}
f_\mu(x') &\le f(\mathrm{prox}_{\mu f}(x)) + \frac{1}{2\mu}\|\mathrm{prox}_{\mu f}(x) - x'\|^2 \\
&= f_\mu(x) + \frac{1}{2\mu}\left(\|\mathrm{prox}_{\mu f}(x) - x'\|^2 - \|x - \mathrm{prox}_{\mu f}(x)\|^2\right) \\
&= f_\mu(x) + \langle \mu^{-1}(x - \mathrm{prox}_{\mu f}(x)), x' - x\rangle + \frac{1}{2\mu}\|x - x'\|^2.
\end{aligned}
$$

Thus, claim 2 holds. The result [53, Theorem 4.4] shows that $\mathrm{prox}_{\mu f}(\cdot)$ is Lipschitz continuous with constant $\frac{1}{1-\mu\rho}$. Lipschitz continuity of $\nabla f_\mu(\cdot)$ with constant $\max\{\mu^{-1}, \frac{\rho}{1-\mu\rho}\}$ follows from (2.11) and Alexandrov's theorem [57, Theorem 13.51]. Thus, claim 3 holds. Claim 4 is immediate from (2.10) and the observation that the function $y \mapsto f(y) + \frac{1}{2\mu}\|y - x\|^2$ is strongly convex for any x. □

2.4 Active Manifolds

The nonsmooth behavior of sets and functions arising in applications is typically far from pathological and instead manifests in highly structured ways. Formalizing this perspective we will assume that nonsmoothness, in a certain localized sense, only occurs along an "active manifold." This notion, introduced in [39] under the name of partial smoothness and rooted in the earlier works [1,10–12,24–26,64], extends the concept of *active sets* in nonlinear programming far beyond the classical setting. In this work, we will take the related perspective developed in [19], since it will be most expedient for our purpose.

Before giving the formal definition, we provide some intuition. Taking a geometric view, we will assume that each critical point of a function f lies on a smooth manifold $\mathcal{M}$, and that the objective varies smoothly along the manifold, but sharply off of it. For example, consider Fig. 2a: there, the function $f(x, y) = |x| - y^2$ admits the active manifold $\mathcal{M} = \{0\} \times \mathbb{R}$ around its unique critical point (the origin). From an algorithmic point of view, active manifolds are the sets that typical algorithms (e.g., proximal point, proximal gradient [31], and dual averaging [37]) identify in finite time. Active manifolds also play a central role for sensitively analysis, providing a path to reduce such questions to the smooth setting. In particular, reasonable conditions guarantee that the active manifold is smoothly traced out by critical points of slight perturbations of the problem. We are now ready to state the formal definition.[7]

Definition 2.6 (*Active manifold*) Consider a closed weakly convex function $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ and fix a set $\mathcal{M} \subseteq \mathbb{R}^d$ containing a critical point $\bar{x}$ of f. Then, $\mathcal{M}$ is called

[7] What we call an *active manifold* here is called an *identifiable manifold* in [19]—the reference we most closely follow. The term active is more evocative in the context of the current work.

uncorrected proof

SPI Journal: 10208 Article No.: 9516 TYPESET DISK LE CP Disp.:2021/4/28 Pages: 46 Layout: Small-Ex

Author Proof

an active C^p-manifold around $\bar{x}$ if there exist a neighborhood U around $\bar{x}$ satisfying the following.

- **(smoothness)** The set $\mathcal{M} \cap U$ is a C^p-smooth manifold and the restriction of f to $\mathcal{M} \cap U$ is C^p-smooth.
- **(sharpness)** The lower bound holds:

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

If f admits an active manifold around a critical point $\bar{x}$, then it must be locally unique: any two active manifolds at $\bar{x}$ must coincide on a neighborhood of $\bar{x}$ [19, Proposition 2.4, Proposition 10.10].[8] Moreover, the critical cone $C_f(\bar{x}, 0)$ coincides with the tangent space $T_{\mathcal{M}}(\bar{x})$ [19, Proposition 10.8]. With the definition of the active manifold in hand, we can now introduce the strict saddle property for nonsmooth functions.[9]

Definition 2.7 (*Strict saddles*) Consider a weakly convex function $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$. Then, we say that a critical point $\bar{x}$ is a *strict saddle* of f if there exists a C^2-active manifold $\mathcal{M}$ of f at $\bar{x}$ and the inequality $d^2 f_{\mathcal{M}}(\bar{x})(u) < 0$ holds for some vector $u \in T_{\mathcal{M}}(\bar{x})$. If every critical point of f is either a local minimizer or a strict saddle, then we say that f satisfies the *strict saddle property*.

Looking at Fig. 2a, we see that the function $f(x, y) = |x| - y^2$ indeed has the strict saddle property: the restriction of f to the axis $\mathcal{M} = \{0\} \times \mathbb{R}$, namely $f_{\mathcal{M}}(0, t) = -t^2$, certainly has directions of negative curvature. Figure 2b depicts the subgradient flow generated by this function. Notice that the set of initial conditions attracted to the origin has measure zero. This observation suggests that typical algorithms are also unlikely to stall at the strict saddle point, an observation made precise by the forthcoming results.

The curvature condition in the definition of the strict saddle pertains only to negative curvature of the restriction of f to $\mathcal{M}$. One may instead ask whether existence of directions of negative curvature for f alone suffice. The answer turns out to be yes.

Theorem 2.8 ([18, Corollary 4.15]) *Consider a closed weakly convex function $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ that admits a C^3-active manifold $\mathcal{M}$ around a critical point $\bar{x}$. Then, it holds:*

$$d^2 f(\bar{x})(u \mid w) \geq d^2 f_{\mathcal{M}}(\bar{x})(u) \qquad \textit{for all } u \in C_f(\bar{x}, 0),\ w \in \mathbb{R}^d.$$

A natural question is whether we expect the strict saddle property to hold typically. One supporting piece of evidence is that the property holds under generic linear perturbations of semialgebraic problems.[10] This is almost immediate from guarantees

[8] Note that due to the convention $\inf_{\emptyset} = +\infty$, the entire space $\mathcal{M} = \mathbb{R}^d$ is the active manifold for a globally C^p-smooth function f around any of its critical points.

[9] Better terminology would be the terms *active strict saddle* and the *active strict saddle property*. To streamline the notation, we omit the word active, as it should be clearly understood from context.

[10] A function is semi-algebraic if its graph can be written as a finite union of sets each cut out by finitely many polynomial inequalities.

uncorrected proof

SPI Journal: 10208 Article No.: 9516 TYPESET DISK LE CP Disp.:2021/4/28 Pages: 46 Layout: Small-Ex

in [18, Theorem 4.16], though this conclusion is not explicitly stated in the theorem statement. We state this guarantee below and provide a quick proof in Sect. A for completeness.

Theorem 2.9 (Strict saddle property is generic). *Consider a closed, weakly convex, semi-algebraic function $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$. Then, for a full Lebesgue measure set of perturbations $v \in \mathbb{R}^d$, the perturbed function $x \mapsto f(x) - \langle v, x\rangle$ has the strict saddle property.*

2.5 The Center Stable Manifold Theorem

In this work, we will show that a variety of simple algorithms escape strict saddle points. To prove results of this type, we will interpret algorithms as fixed point iterations of a nonlinear map $T : \mathbb{R}^d \to \mathbb{R}^d$, having certain favorable properties. As in the smooth setting of [35], the core of our arguments will be based on the center stable manifold theorem.

Theorem 2.10 (The Center Stable Manifold Theorem [60, Theorem III.7]) *Let the origin be a fixed point of the C^1 local diffeomorphism $T : U \to \mathbb{R}^d$ where U is a neighborhood of the origin in $\mathbb{R}^d$. Let $E^s \oplus E^c \oplus E^u$ be the invariant splitting of $\mathbb{R}^d$ into the generalized eigenspaces of the Jacobian $\nabla T(0)$ corresponding to eigenvalues of absolute value less than one, equal to one, and greater than one. Then, there exists a local T invariant C^1 embedded disk W^{cs}_{loc}, tangent to $E^s \oplus E^c$ at 0 and a neighborhood B around zero such that $T(W^{cs}_{loc}) \cap B \subseteq W^{cs}_{loc}$. In addition, if $T^k(x) \in B$ for all $k \geq 0$, then $x \in W^{cs}_{loc}$.*

An immediate consequence of this theorem is the following: if $\nabla T(0)$ is invertible and has at least one eigenvalue of magnitude greater than one, then there exists a neighborhood B of the origin such that the set

$$\{x \in B : T^k(x) \in B \text{ for all } k \geq 0\},$$

has measure zero. This fact motivates the following key definition.

Definition 2.11 (*Unstable fixed points*) A fixed point $\bar{x}$ of a map $T : \mathbb{R}^d \to \mathbb{R}^d$ is called *unstable* if T is C^1-smooth around $\bar{x}$ and the Jacobian $\nabla T(\bar{x})$ has an eigenvalue of magnitude strictly greater than one.

To globalize the guarantees of the center stable manifold theorem, we will need to impose global regularity properties on T. In this work, we will require the map T to be a *lipeomorphism*, namely, we require that T is globally Lipschitz and its inverse T^{-1} is a well-defined globally Lipschitz map. The following corollary is now immediate. Its proof closely follows the presentation in [34, Theorem 2].

Corollary 2.12 *Let $T : \mathbb{R}^d \to \mathbb{R}^d$ be a lipeomorphism and let $\mathcal{U}_T$ consist of all unstable fixed points x of T at which the Jacobian $\nabla T(x)$ is invertible. Then, the set of initial*

SPI Journal: 10208 Article No.: 9516 TYPESET DISK LE CP Disp.:2021/4/28 Pages: 46 Layout: Small-Ex

Author Proof
uncorrected proof

Author Proof

conditions attracted by such fixed points

$$W := \left\{ x \in \mathbb{R}^d : \lim_{k\to\infty} T^k(x) \in \mathcal{U}_T \right\}$$

has zero Lebesgue measure.

Proof For every $\bar{x} \in \mathcal{U}_T$ there exists a neighborhood U of $\bar{x}$ such that $T: U \to \mathbb{R}^d$ is a local diffeomorphism. Thus, the center stable manifold theorem shows there exists an open neighborhood $B_{\bar{x}}$ of $\bar{x}$ so that $S_{\bar{x}} := \bigcap_{k=0}^{\infty} T^{-k}(B_{\bar{x}})$ is contained in a measure zero set. In particular, $S_{\bar{x}}$ itself is measure zero.

Now observe that $\mathcal{U}_T \subseteq \bigcup_{\bar{x}\in\mathcal{U}_T} B_{\bar{x}}$ is an open cover of $\mathcal{U}_T$. Since $\mathbb{R}^d$ is second countable, this cover has a countable subcover $\mathcal{U}_T \subseteq \bigcup_{i=1}^{\infty} B_{\bar{x}_i}$. Observe the inclusion $W \subseteq \bigcup_{i=1}^{\infty}\bigcup_{j=0}^{\infty} T^{-j}(S_{\bar{x}_i})$. Since T is a lipeomorphism, the right-hand side is a countable union of measure zero sets, and therefore, W has measure zero. □

To verify that a map T is a lipeomorphism, we will appeal to the following standard sufficient condition. We provide a quick proof for completeness.

Lemma 2.13 *Let $H: \mathbb{R}^d \to \mathbb{R}^d$ be a Lipschitz continuous map with constant $\lambda < 1$. Then, $I + H$ is invertible and $(I + H)^{-1}: \mathbb{R}^d \to \mathbb{R}^d$ is Lipschitz continuous with constant* $(1-\lambda)^{-1}$.

Proof To show that $(I + H)$ is invertible, we must show that for every $u \in \mathbb{R}^d$, the equation $u = H(x) + x$ has a unique solution $x(u) \in \mathbb{R}^d$. Equivalently, we must show that for every $u \in \mathbb{R}^d$, the mapping

$$\zeta_u(x) := u - H(x)$$

has a unique fixed point. This is immediate from Banach's fixed point theorem since $\zeta_u(\cdot)$ is strictly contractive.

To show that $(I + H)^{-1}$ is Lipschitz, choose arbitrary $u, v \in \mathbb{R}^d$ and define $x := (I + H)^{-1}(u)$ and $y := (I + H)^{-1}(v)$. We then compute

$$\begin{aligned}\|u - v\| &= \|(I + H)(x) - (I + H)(y)\| \geq \|x - y\| - \|H(x) - H(y)\| \\ &\geq (1-\lambda)\|x - y\|,\end{aligned}$$

where we have used the reverse triangle inequality and Lipschitz continuity of H. Rearranging completes the proof. □

While the iteration mappings S of Sect. 1.3 can be Lipschitz, they are usually not invertible. Thus, to ensure Lipschitz invertibility, we will consider damped fixed point iterations, as summarized in the following elementary lemma. We provide a quick proof for completeness.

Lemma 2.14 (Damped fixed point iterations). *Consider a map $S: \mathbb{R}^d \to \mathbb{R}^d$ and fix a damping parameter* $\alpha \in (0, 1)$. *Define the map*

$$T(x) = (1-\alpha)x + \alpha \cdot S(x).$$

SPI Journal: 10208 Article No.: 9516 TYPESET DISK LE CP Disp.:2021/4/28 Pages: 46 Layout: Small-Ex

uncorrected proof

Then, the following are true.

1. *The fixed points of T and S coincide.*
2. *If S is differentiable at $\bar{x}$ and the Jacobian $\nabla S(\bar{x})$ has a real eigenvalue strictly greater than one, then $\bar{x}$ is an unstable fixed point of T.*
3. *If the map S is continuous and the iterates generated by the process $x_{k+1} = T(x_k)$ converge to some point $\bar{x}$, then $\bar{x}$ must be a fixed point of S.*
4. *If the map $I - S$ is L-Lipschitz, then T is a lipeomorphism for any $\alpha \in (0, L^{-1})$.*

Proof Claims 1 and 2 follow directly from algebraic manipulations. Claim 4 follows immediately from Lemma 2.13 by writing $T = I + H$ with $H = \alpha(S - I)$. To see claim 3, suppose that T is continuous and that x_k converge to some point $\bar{x}$. Then, we deduce

$$T(\bar{x}) = T\left(\lim_{k\to\infty} x_k\right) = \lim_{k\to\infty} T(x_k) = \lim_{k\to\infty} x_{k+1} = \bar{x}.$$

Therefore, $\bar{x}$ is a fixed point of T. Using claim 1, we deduce that $\bar{x}$ is a fixed point of S. □

3 The Proximal Point Method

We now turn to the saddle escape properties of the proximal-point method. Fixing the problem at hand, we consider

$$\min_{x\in\mathbb{R}^d} f(x),$$

where $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is a ρ-weakly convex function that is bounded from below. For a fixed $\mu < \rho^{-1}$, the classical proximal-point method is precisely the fixed point iteration

$$x_{t+1} = \mathrm{prox}_{\mu f}(x_t).$$

Key to our analysis is the equivalence between this algorithm and gradient descent on the Moreau envelope. This equivalence follows from (2.10), which quickly yields the description

$$x_{k+1} = x_k - \mu \cdot \nabla f_\mu(x_k).$$

The saddle escape properties of the proximal point method thus flow from the strict saddle properties of the Moreau envelope. Indeed, the following theorem shows that when f admits a C^2 active manifold around a critical point $\bar{x}$, the envelope f_μ is automatically C^2-smooth near $\bar{x}$. Moreover, if $\bar{x}$ is a strict saddle of f, then it is also a strict saddle of f_μ. Consequently, any strict saddle point of f is an unstable fixed point of the proximal map $\mathrm{prox}_{\mu f}(\cdot)$.

SPI Journal: 10208 Article No.: 9516 TYPESET DISK LE CP Disp.:2021/4/28 Pages: 46 Layout: Small-Ex

Author Proof
uncorrected proof

Author Proof

Theorem 3.1 (Saddle points of the Moreau envelope). *Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be a closed and ρ-weakly convex function and let $\bar{x}$ be any critical point of f. Suppose that f admits a C^2 active manifold $\mathcal{M}$ at $\bar{x}$. Then, for any $\mu < \rho^{-1}$, the Moreau envelope f_μ is C^2-smooth around $\bar{x}$ and its Hessian satisfies*

$$\min_{h \in \mathbb{S}^{d-1} \cap T_{\mathcal{M}}(\bar{x})} \langle \nabla^2 f_\mu(\bar{x})h, h\rangle \leq \min_{h \in \mathbb{S}^{d-1} \cap T_{\mathcal{M}}(\bar{x})} d^2 f_{\mathcal{M}}(\bar{x})(h). \tag{3.1}$$

Consequently, if $\bar{x}$ is a strict saddle point of f, then $\bar{x}$ is both a strict saddle point of f_μ and an unstable fixed point of the proximal map $\mathrm{prox}_{\mu f}(\cdot)$. *Moreover,* $\nabla \mathrm{prox}_{\mu f}(\bar{x})$ *has a real eigenvalue that is strictly greater than one.*

Proof It is well known (for example, from [31]) that for all x near $\bar{x}$, the inclusion $\mathrm{prox}_{\mu f}(x) \in \mathcal{M}$ holds. From this inclusion, we will be able to view the proximal subproblem through the lens of the perturbation result in Theorem 2.3. For the sake of completeness, however, let us first quickly verify the claim. Consider a sequence $x_i \to \bar{x}$ and observe the inclusion $\nabla f_\mu(x_i) \in \partial f(\mathrm{prox}_{\mu f}(x_i))$. Since the gradient ∇f_μ is continuous, we deduce the limits $\mathrm{prox}_{\mu f}(x_i) \to \bar{x}$ and $\nabla f_\mu(x_i) \to 0$. Therefore, by definition of the active manifold, we have $\mathrm{prox}_{\mu f}(x_i) \in \mathcal{M}$ for all sufficiency large indices i, proving the claim.

Turning to the perturbation result, let $F : \mathbb{R}^d \to \mathbb{R}$ be any C^2-smooth function agreeing with f on a neighborhood of $\bar{x}$ in $\mathcal{M}$.[11] Applying the claim, we find that the equality

$$f_\mu(x) = \min_{y \in \mathcal{M}} \left\{ F(y) + \frac{1}{2\mu}\|y - x\|^2 \right\},$$

holds for all x near $\bar{x}$. Our goal is to apply the perturbation result (Theorem 2.3) with $f(x, y) := F(y) + \frac{1}{2\mu}\|y - x\|^2$ and $\varphi(x) := f_\mu(x)$. To that end, we now verify the assumptions of Theorem 2.3. First, we verify the quadratic growth condition: since we have chosen $\mu < \rho^{-1}$, it follows that for every $x \in \mathbb{R}^d$ the function $y \mapsto f(x) + \frac{1}{2\mu}\|y - x\|^2$ is strongly convex with constant $\mu^{-1} - \rho$. Next, we verify the level boundedness condition: since the minimizer $y(x) := \mathrm{prox}_{\mu f}(x)$ of this function varies continuously and satisfies $y(\bar{x}) = \bar{x}$, the conditions of Lemma 2.4 are satisfied. Therefore, the assumptions of Theorem 2.3 are valid.

We now apply Theorem 2.3. To that end, let $G = 0$ be the defining equation of $\mathcal{M}$ around $\bar{x}$ and define the parametric Lagrangian function

$$\mathcal{L}(x, y, \lambda) := F(y) + \frac{1}{2\mu}\|y - x\|^2 + \langle G(y), \lambda\rangle.$$

Since $\bar{x}$ is critical for f, the equality $\bar{x} = \mathrm{prox}_{\mu f}(\bar{x})$ holds. Consequently, $y(\bar{x}) = \bar{x}$ minimizes the function $y \mapsto F(y) + \frac{1}{2\mu}\|y - \bar{x}\|^2$ on $\mathcal{M}$. Therefore, first-order

[11] For example, let F be a C^2 function defined on a neighborhood U of $\bar{x}$ that agrees with f on $U \cap \mathcal{M}$. Using a partition of unity (e.g., [36, Lemma 2.26]), one can extend F from a slightly smaller neighborhood to be C^2 on all of $\mathbb{R}^d$.

Author Proof

optimality conditions guarantee there exists a multiplier vector $\bar{\lambda}$ satisfying

$$0 = \nabla_y \mathcal{L}(\bar{x}, \bar{x}, \bar{\lambda}) = \nabla F(\bar{x}) + \sum_{i \geq 1} \bar{\lambda}_i G_i(\bar{x}),$$

where $G_i(\cdot)$ are the coordinate functions of $G(\cdot)$. Appealing to Theorem 2.3, we learn both that f_μ is C^2-smooth around $\bar{x}$ and that its Hessian satisfies

$$\langle \nabla^2 f_\mu(\bar{x})h, h\rangle = \min_{u \in T_{\mathcal{M}}(\bar{x})} \langle H_{xx}h, h\rangle + 2\langle H_{xy}u, h\rangle + \langle H_{yy}u, u\rangle, \tag{3.2}$$

where the Hessian matrices are given by

$$H_{xx} = \mu^{-1} I, \qquad H_{xy} = -\mu^{-1} I, \qquad H_{yy} = \nabla^2 F(\bar{x}) + \sum_{i=1}^{m} \bar{\lambda}_i \nabla^2 G_i(\bar{x}) + \mu^{-1} I.$$

Thus, rearranging (3.2) and setting $D := \nabla^2 F(\bar{x}) + \sum_{i=1}^m \bar{\lambda}_i \nabla^2 G_i(\bar{x})$, we have

$$\langle \nabla^2 f_\mu(\bar{x})h, h\rangle = \min_{u \in T_{\mathcal{M}}(\bar{x})} \left\{ \langle Du, u\rangle + \mu^{-1}\|h - u\|^2 \right\}.$$

Therefore, we arrive at the estimate

$$\begin{aligned} \min_{h \in \mathbb{S}^{d-1} \cap T_{\mathcal{M}}(\bar{x})} \langle \nabla^2 f_\mu(\bar{x})h, h\rangle &= \min_{u \in T_{\mathcal{M}}(\bar{x})} \min_{h \in \mathbb{S}^{d-1} \cap T_{\mathcal{M}}(\bar{x})} \left\{ \langle Du, u\rangle + \mu^{-1}\|h - u\|^2 \right\} \\ &\leq \min_{h \in \mathbb{S}^{d-1} \cap T_{\mathcal{M}}(y_0)} \langle Dh, h\rangle = \min_{h \in \mathbb{S}^{d-1} \cap T_{\mathcal{M}}(\bar{x})} d^2 f_{\mathcal{M}}(\bar{x})(h), \end{aligned}$$

thereby verifying (3.1). If $\bar{x}$ is a strict saddle point of f, then (3.1) implies that $\nabla^2 f_\mu(\bar{x})$ has a strictly negative eigenvalue. From the expression $\mathrm{prox}_{\mu f} = I - \mu \nabla f_\mu$, we therefore deduce that the Jacobian of $\mathrm{prox}_{\mu f}$ at $\bar{x}$ has at least one real eigenvalue that is strictly greater than one. Consequently, $\bar{x}$ is an unstable fixed point of $\mathrm{prox}_{\mu f}$. □

Even if the proximal mapping has an unstable fixed-point, it often fails to meet the conditions of the center stable manifold theorem (Theorem 2.10). Indeed, the proximal mapping is generally not injective, even near critical points. To remedy this issue, we instead analyze a slightly damped version of the proximal point method

$$x_{k+1} = (1 - \alpha)x_k + \alpha \cdot \mathrm{prox}_{\mu f}(x_k),$$

where $\alpha \in (0, 1)$ is a fixed constant. Reinterpreting this algorithm in terms of the Moreau envelope, we arrive at the recurrence

$$x_{k+1} = x_k - (\alpha\mu) \cdot \nabla f_\mu(x_k). \tag{3.3}$$

Thus, the role of damping is clear: it still induces gradient descent on the Moreau envelope, but with a stepsize slightly below the "theoretically optimal" step μ. This is

uncorrected proof

SPI Journal: 10208 Article No.: 9516 TYPESET DISK LE CP Disp.:2021/4/28 Pages: 46 Layout: Small-Ex

Author Proof

entirely in line with the saddle point escape guarantees for gradient descent in smooth minimization [35].

Theorem 3.2 (Proximal point method: global escape). *Let $f: \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be a closed and ρ-weakly convex function satisfying the strict saddle property. Choose a constant $\mu < \rho^{-1}$ and a damping parameter $\alpha \in (0, \min\{1, (\mu\rho)^{-1} - 1\})$. With these choices, consider the algorithm*

$$x_{k+1} = (1-\alpha)x_k + \alpha \cdot \mathrm{prox}_{\mu f}(x_k). \tag{3.4}$$

Then, for almost all initializers x_0, the following holds: if the limit of $\{x_k\}_{k\geq 0}$ exists, it must be a local minimizer of f.

Proof Define the map $S := \mathrm{prox}_{\mu f}(x_k)$. Lemma 2.5 guarantees that the map $I - S = \mu \nabla f_\mu$ is Lipschitz continuous with constant $\max\{1, \frac{\mu\rho}{1-\mu\rho}\}$. Taking into account the range of α and applying Lemma 2.14 and Theorem 3.1, we may deduce the following three properties: (1) T is a lipeomorphism, (2) the limit of the sequence x_k, if it exists, must be a critical point of f, and (3) if a critical point of f is not a local minimum, then it is an unstable fixed point of T. An application of Corollary 2.12 completes the proof. □

4 The Proximal Gradient Method

We now turn to the saddle escape properties of the proximal gradient method. Fixing the problem at hand, we consider

$$\min_{x\in\mathbb{R}^d} \; f(x) = g(x) + r(x), \tag{4.1}$$

where $g: \mathbb{R}^d \to \mathbb{R}$ is a C^2-smooth function with β-Lipschitz gradient and $r: \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is a closed and ρ-weakly convex function. We assume throughout that f is bounded from below. For this problem, the proximal gradient method takes the form

$$x_{k+1} = \mathrm{prox}_{\mu r}\left(x_k - \mu \nabla g(x_k)\right).$$

Unlike the proximal point algorithm, the proximal gradient algorithm may not correspond to gradient descent on a smooth envelope of the problem. Still, as the following theorem shows, the iteration mapping is C^1 smooth near $\bar{x}$ whenever f admits a C^2 active manifold around a critical point $\bar{x}$. Moreover, if $\bar{x}$ is a strict saddle point of f, then $\bar{x}$ is an unstable fixed point of the iteration mapping

Theorem 4.1 (Unstable fixed points of the prox-gradient map). *Consider the optimization problem (4.1) and let $\bar{x}$ be any critical point of f. Suppose that f admits a C^2 active manifold $\mathcal{M}$ at $\bar{x}$. Then, for any $\mu \in (0, \rho^{-1})$, the proximal-gradient map*

$$S(x) := \mathrm{prox}_{\mu r}\left(x - \mu \nabla g(x)\right)$$

SPI Journal: 10208 Article No.: 9516 TYPESET DISK LE CP Disp.:2021/4/28 Pages: 46 Layout: Small-Ex

is C^1-smooth on a neighborhood of $\bar{x}$. Moreover, if $\bar{x}$ is a strict saddle point of f, then $\nabla S(\bar{x})$ has a real eigenvalue that is strictly greater than one.

Proof It is well known (for example, from [31]) that for all x near $\bar{x}$, the point $S(x)$ lies in $\mathcal{M}$. From this inclusion, we will be able to view the proximal subproblem through the lens of the perturbation result in Theorem 2.3. For the sake of completeness, however, we provide a quick proof. Indeed, consider a sequence $x_i \to \bar{x}$ and set $y_i = S(x_i)$. Then, by definition of the proximal gradient map, we have $0 \in \nabla g(x_i) + \mu^{-1}(y_i - x_i) + \partial r(y_i)$, and therefore

$$\begin{aligned}\operatorname{dist}(0, \partial f(y_i)) = \operatorname{dist}(-\nabla g(y_i), \partial r(y_i)) &\leq \operatorname{dist}(-\nabla g(x_i), \partial r(y_i)) + \beta\|y_i - x_i\| \\ &\leq (\mu^{-1} + \beta)\|y_i - x_i\|.\end{aligned}$$

Since $S(\cdot)$ is continuous and $S(\bar{x}) = \bar{x}$, we deduce $y_i \to \bar{x}$ and therefore $\operatorname{dist}(0, \partial f(y_i)) \to 0$. Therefore, the points y_i lie in $\mathcal{M}$ for all sufficiently large indices i, proving the claim.

Turning to the perturbation result, let $R\colon \mathbb{R}^d \to \mathbb{R}$ be any C^2-smooth function agreeing with r on a neighborhood of $\bar{x}$ in $\mathcal{M}$. Applying the claim, we find that for x near $\bar{x}$, the point $S(x)$ uniquely minimizes problem

$$\min_{y \in \mathcal{M}} \left\{ g(x) + \langle \nabla g(x), y - x\rangle + R(y) + \frac{1}{2\mu}\|y - x\|^2 \right\}. \qquad (\mathcal{P}_x)$$

Our goal is to apply the perturbation result (Theorem 2.3) with $f(x, y) := g(x) + \langle \nabla g(x), y - x\rangle + R(y) + \frac{1}{2\mu}\|y - x\|^2$. To that end, we now verify the assumptions of Theorem 2.3. First, we verify the quadratic growth condition: since we have chosen $\mu < \rho^{-1}$, it follows that for every $x \in \mathbb{R}^d$ the function $y \mapsto f(x, y)$ is strongly convex with the constant $\mu^{-1} - \rho$. Next, we verify the level-boundedness condition: since the minimizer $S(x)$ clearly varies continuously and satisfies $S(\bar{x}) = \bar{x}$, the conditions of Lemma 2.4 are satisfied. Therefore, the assumptions of Theorem 2.3 are valid.

We now apply Theorem 2.3. To that end, let $G = 0$ be the defining equation of $\mathcal{M}$ around $\bar{x}$ and define the parametric Lagrangian function

$$\mathcal{L}(x, y, \lambda) = g(x) + \langle \nabla g(x), y - x\rangle + R(y) + \frac{1}{2\mu}\|y - x\|^2 + \sum_{i \geq 1} \lambda_i G_i(y),$$

where $G_i(\cdot)$ are the coordinate functions of G. Clearly $y(\bar{x}) = \bar{x}$ minimizes $f(\bar{x}, \cdot)$ on $\mathcal{M}$. Therefore, first-order optimality conditions guarantee there exists a multiplier vector $\bar{\lambda}$ satisfying

$$0 = \nabla_y \mathcal{L}(\bar{x}, \bar{x}, \bar{\lambda}) = \nabla g(\bar{x}) + \nabla R(\bar{x}) + \sum_{i \geq 1} \bar{\lambda}_i G_i(\bar{x}).$$

Appealing to Theorem 2.3, we learn that the solution map $S(\cdot)$ is C^1-smooth around $\bar{x}$ with

$$\nabla S(\bar{x})h = \operatorname*{argmin}_{v \in T_{\mathcal{M}}(\bar{x})} \; 2\langle H_{xy} v, h\rangle + \langle H_{yy} v, v\rangle, \qquad (4.2)$$

Author Proof
uncorrected proof

SPI Journal: 10208 Article No.: 9516 TYPESET DISK LE CP Disp.:2021/4/28 Pages: 46 Layout: Small-Ex

Author Proof

where the Hessian matrices are given by

$$H_{xy} = \nabla^2 g(\bar{x}) - \mu^{-1} I, \qquad H_{yy} = \nabla^2 R(\bar{x}) + \mu^{-1} I + \sum_{i=1}^{p} \bar{\lambda}_i \nabla^2 G_i(\bar{x}).$$

We now simplify the expression (4.2). To that end, let W be the orthogonal projection onto $T_{\mathcal{M}}(\bar{x})$ and define the linear maps $\overline{H_{yy}} \colon T_{\mathcal{M}}(\bar{x}) \to T_{\mathcal{M}}(\bar{x})$ and $\overline{H_{xy}} \colon T_{\mathcal{M}}(\bar{x}) \to T_{\mathcal{M}}(\bar{x})$ by setting $\overline{H_{yy}} = W H_{yy} W$ and $\overline{H_{xy}} = W H_{xy} W$, respectively. Since $\bar{x}$ is a strong local minimizer of $\mathcal{P}_{\bar{x}}$, the map $\overline{H_{yy}}$ is positive definite, and hence invertible. Solving (4.2) then yields the expression

$$\nabla S(\bar{x})h = -\overline{H}_{yy}^{-1} \overline{H_{xy}}^{\top} h \qquad \text{for all } h \in T_{\mathcal{M}}(\bar{x}).$$

Note that $\overline{H_{xy}}^{\top}$ is a symmetric matrix, so we drop the "$\top$" throughout.

Let us now verify that if $\bar{x}$ is a strict saddle of f, then $\nabla S(\bar{x})$ has a real eigenvalue that is greater than one. To this end, observe that $\gamma \in \mathbb{R}$ is a real eigenvalue of $\nabla S(\bar{x})$ with an associated eigenvector $v \in T_{\mathcal{M}}(\bar{x})$ if and only if

$$\nabla S(\bar{x})v = \gamma v \qquad \Longleftrightarrow \qquad -\overline{H}_{yy}^{-1} \overline{H_{xy}} v = \gamma v \qquad \Longleftrightarrow \qquad (\gamma \overline{H}_{yy} + \overline{H_{xy}})v = 0.$$

In particular, if the matrix $\gamma \overline{H}_{yy} + \overline{H_{xy}}$ is singular, then γ is an eigenvalue of $\nabla S(\bar{x})$. To prove such a γ exists, we will examine the following ray of symmetric matrices

$$\{\gamma \overline{H}_{yy} + \overline{H_{xy}} : \gamma \geq 1\}.$$

Beginning with the end point, the strict saddle property shows that

$$\overline{H}_{yy} + \overline{H_{xy}} = W \left(\nabla^2 g(\bar{x}) + \nabla^2 R(\bar{y}) + \sum_i \bar{\lambda}_i \nabla^2 G_i(\bar{x}) \right) W.$$

has a strictly negative eigenvalue. On the other hand, the direction of the ray $\overline{H}_{yy}$ is a positive definite matrix. Therefore, by continuity of the minimal eigenvalue function, there exists some $\gamma > 1$ such that the matrix $\gamma \overline{H}_{yy} + \overline{H_{xy}}$ is singular, as claimed. □

Similar to the proximal point method, the proximal gradient mapping fails to meet the conditions of the center stable manifold theorem (Theorem 2.10), since it generally lacks invertibility. Therefore, as before we will analyze a slightly damped version of the process, and prove the following theorem.

Theorem 4.2 (Proximal gradient method: global escape). *Consider the optimization problem* (4.1) *and suppose that f has the strict saddle property. Choose any constant $\mu \in (0, \rho^{-1})$ and a damping parameter $\alpha \in (0, 1)$ satisfying*

$$\alpha \cdot \left(\mu\beta + (1 + \mu\beta) \max \left\{ 1, \tfrac{\mu\rho}{1-\mu\rho} \right\} \right) < 1.$$

uncorrected proof

Author Proof

Consider the algorithm

$$x_{k+1} = (1-\alpha)x_k + \alpha \cdot \operatorname{prox}_{\mu r}\left(x_k - \mu \nabla g(x_k)\right). \tag{4.3}$$

Then, for almost all initializers x_0, the following holds: if the limit of $\{x_k\}_{k\geq 0}$ exists, it must be a local minimizer of f.

Proof Define the maps $S = \operatorname{prox}_{\mu r}(I - \mu\nabla g)$. We successively rewrite

$$\begin{aligned} I - S &= (I - \mu\nabla g) - \operatorname{prox}_{\mu r}(I - \mu\nabla g) + \mu\nabla g \\ &= \mu \cdot \nabla r_\mu \circ (I - \mu\nabla g) + \mu\nabla g. \end{aligned}$$

Lemma 2.5 implies that the map $I - S$ is Lipschitz continuous with constant $\mu\beta + (1+\mu\beta)\max\left\{1, \frac{\mu\rho}{1-\mu\rho}\right\}$. Taking into account the range of α and applying Lemma 2.14 and Theorem 4.1, we may deduce the following three properties: (1) T is a lipeomorphism, (2) the limit of the sequence x_k, if it exists, must be a critical point for f, and (3) if a critical point of f is not a local minimum, then it is an unstable fixed point of T. An application of Corollary 2.12 then completes the proof. □

5 The Proximal Linear Method

We now turn to the saddle escape properties of the proximal linear method, a generalization of the proximal point and proximal gradient methods. Setting the stage, consider the composite optimization problem

$$\min_x \; f(x) = h(F(x)) + r(x), \tag{5.1}$$

where $F\colon \mathbb{R}^d \to \mathbb{R}^m$ is a C^2-smooth map, $h\colon \mathbb{R}^d \to \mathbb{R}$ is convex, and $r\colon \mathbb{R}^d \to \mathbb{R}\cup\{\infty\}$ is ρ-weakly convex. As is standard in the literature, we will assume that there exists a constant $\beta > 0$ satisfying

$$|h(F(y)) - h(F(x) + \nabla F(x)(y-x))| \leq \frac{\beta}{2}\|y-x\|^2, \qquad \forall x, y \in \mathbb{R}^d. \tag{5.2}$$

These assumptions then easily imply that f is weakly convex with constant $\beta + \rho$.

With the stage set, we now slightly refine the notion of a strict saddle, adapting it to the compositional nature of the problem. This refinement intuitively asks that the active manifold for f at a critical point $\bar{x}$ is induced by active manifolds of h and r. Similar conditions have appeared elsewhere, for example, in [19,39,40]. To describe the condition formally, we will also revise the definition of an active manifold, allowing us to discuss active manifolds of $h(\cdot)$ and $r(\cdot)$ at noncritical points. The revision is intuitive, requiring just a linear tilt of the functions:

- Consider a set $\mathcal{R} \subset \mathbb{R}^d$, a point $x \in \mathcal{R}$, and a subgradient $v \in \partial r(x)$. We will say that $\mathcal{R}$ is a C^2 *active manifold of r at x for v* if $\mathcal{R}$ is a C^2 active manifold of the tilted function $r - \langle v, \cdot\rangle$ at x in the sense of Definition 2.6.

uncorrected proof

SPI Journal: 10208 Article No.: 9516 TYPESET DISK LE CP Disp.:2021/4/28 Pages: 46 Layout: Small-Ex

Author Proof

We may likewise define the active manifold of h at z for $w \in \partial h(z)$, based on a tilting of h by w. Coupling these definitions, we arrive at the active manifold concept for the composite problem (5.1).

Definition 5.1 (*Composite active manifold*) Consider the compositional problem (5.1) and let $\bar{x}$ be a critical point of f. Fix arbitrary vectors $\bar{w} \in \partial h(F(\bar{x}))$ and $\bar{v} \in \partial r(\bar{x})$ satisfying

$$0 \in \nabla F(\bar{x})^* \bar{w} + \bar{v}. \tag{5.3}$$

Suppose the following hold.

1. There exist C^2-smooth manifolds $\mathcal{R} \subset \mathbb{R}^d$ and $\mathcal{H} \subset \mathbb{R}^m$ containing $\bar{x}$ and $F(\bar{x})$, respectively, and satisfying the transversality condition:

$$\nabla F(\bar{x}) \left[T_{\mathcal{R}}(\bar{x})\right] + T_{\mathcal{H}}(F(\bar{x})) = \mathbb{R}^m. \tag{5.4}$$

2. $\mathcal{R}$ is an active manifold of r at $\bar{x}$ for $\bar{v}$ and $\mathcal{H}$ is an active manifold of h at $F(\bar{x})$ for $\bar{w}$.

Then, we will call $\mathcal{M} := \mathcal{R} \cap F^{-1}(\mathcal{H})$ a *composite* C^2 *active manifold* for the problem (5.1) at $\bar{x}$. If in addition the inequality $d^2 f_{\mathcal{M}}(\bar{x})(u) < 0$ holds for some vector $u \in T_{\mathcal{M}}(\bar{x})$, then we will call $\bar{x}$ a *composite strict saddle point*.

This definition has several important subtleties. First, the set $\mathcal{M} := \mathcal{R} \cap F^{-1}(\mathcal{H})$ is indeed a C^2-smooth manifold around $\bar{x}$, due to the classical transversality condition (5.4), a central fact in differential geometry [36, Theorem 6.30]. Next, the vectors $\bar{v}$ and $\bar{w}$ do exist. This follows since $\bar{x}$ is first-order critical for f:

$$0 \in \nabla F(\bar{x})^* \partial h(F(\bar{x})) + \partial r(\bar{x}).$$

Beyond existence, the vectors $\bar{v} \in \partial r(\bar{x})$ and $\bar{w} \in \partial h(F(\bar{x}))$ are in fact the unique elements satisfying (5.3), a second consequence of transversality. To see this, we state (5.4) in dual terms as

$$(\nabla F(\bar{x})^*)^{-1} N_{\mathcal{R}}(\bar{x}) \cap N_{\mathcal{H}}(F(\bar{x})) = \{0\}. \tag{5.5}$$

Considering another pair $v \in \partial r(\bar{x})$ and $w \in \partial h(F(\bar{x}))$ satisfying (5.3), we deduce

$$0 = \nabla F^*(\bar{x})(\bar{w} - w) + (\bar{v} - v).$$

To conclude $v = \bar{v}$ and $w = \bar{w}$, we use (5.5) and simply recall that span $\partial h(F(\bar{x})) = N_{\mathcal{H}}(F(\bar{x}))$ and span $\partial r(\bar{x}) = N_{\mathcal{R}}(\bar{x})$, as shown in [19, Proposition 10.12]. Finally, collecting these facts together, it follows from the chain rule [19, Proposition 5.1] that $\mathcal{M}$ is an active manifold of f at $\bar{x}$ in the sense of Definition 2.7.

A natural question is whether we expect the composite strict saddle property to hold typically. One supporting piece, of evidence, analogous to Theorem 2.9, is that the property holds under generic linear perturbations of semialgebraic composite problems. This result quickly follows from [18, Theorem 5.2]. We provide a proof sketch in Sect. A.

uncorrected proof

Author Proof

Theorem 5.2 (Strict saddle property is generic). *Consider the composite problem* (5.1), *where h, r, and F are in addition semi-algebraic. Then, for a full Lebesgue measure set of perturbations* $(y, v) \in \mathbb{R}^m \times \mathbb{R}^d$, *the problem*

$$\min_x \; h(F(x) + y) + r(x) - \langle v, x \rangle$$

has the composite strict saddle property.

Turning to our central task, we aim to analyze the saddle escape properties of the proximal linear method:

$$x_{k+1} = \operatorname*{argmin}_y \; h(F(x_k) + \nabla F(x_k)(y - x_k)) + r(y) + \frac{1}{2\mu}\|y - x_k\|^2.$$

To analyze this method, we prove the following theorem, showing that any strict saddle point of the composite problem (5.1) is an unstable fixed point of proximal linear update.

Theorem 5.3 (Unstable fixed points of the proximal linear map). *Consider the composite problem* (5.1) *and let* $\bar{x}$ *be any critical point of* f. *Suppose the problem admits a composite* C^2 *active manifold* $\mathcal{M}$ *at* $\bar{x}$. *Then, for any* $\mu \in (0, \rho^{-1})$, *the proximal linear map*

$$S(x) := \operatorname*{argmin}_y \; h(F(x) + \nabla F(x)(y - x)) + r(y) + \frac{1}{2\mu}\|y - x\|^2. \tag{5.6}$$

is C^1*-smooth on a neighborhood of* $\bar{x}$. *Moreover, if* $\bar{x}$ *is a composite strict saddle point, then the Jacobian* $\nabla S(\bar{x})$ *has a real eigenvalue strictly greater than one.*

In most ways, the proof mirrors that of Theorem 3.2. There is, however, an important complication: we must move beyond the perturbation result of Theorem 2.3 and instead analyze a parametric family of optimization problems where both the objective and *the constraints* depend on a perturbation parameter. Therefore, we will rely on the following generalization of Theorem 2.3. For details and a much more general perturbation result, see [59, Theorem 4.2].

Theorem 5.4 (Perturbation analysis). *Consider the family of optimization problems*

$$\min_y \; f(x, y) \quad \text{subject to} \quad G(x, y) = 0 \tag{$\mathcal{Q}_x$}$$

Fix a point $\bar{x}$ *and a minimizer* $\bar{y}$ *of* $\mathcal{Q}_{\bar{x}}$, *and suppose the following hold.*

1. **(Nondegeneracy)** *The function* $f(\cdot, \cdot)$ *and the map* $G(\cdot, \cdot)$ *are* C^2*-smooth near* $(\bar{x}, \bar{y})$, *and the Jacobian* $\nabla_y G(\bar{x}, \bar{y})$ *is surjective.*
2. **(Level-boundedness)** *There exists a neighborhood* $\mathcal{X}$ *of* $\bar{x}$ *and a number* γ *greater than the minimal value of* $\mathcal{Q}_{\bar{x}}$ *such that the set*

$$\bigcup_{x \in \mathcal{X}} \{y \in Y(x) : f(x, y) \leq \gamma\} \quad \text{is bounded,}$$

Uncorrected Proof

SPI Journal: 10208 Article No.: 9516 TYPESET DISK LE CP Disp.:2021/4/28 Pages: 46 Layout: Small-Ex

Author Proof

where $Y(x) := \{y : G(x, y) = 0\}$ denotes the set of feasible points for $\mathcal{Q}_x$.

3. **(Quadratic growth)** *The point $\bar{y}$ is a strong local minimizer and a unique global minimizer of $\mathcal{Q}_{\bar{x}}$.*

Define the parametric Lagrangian function

$$\mathcal{L}(x, y, \lambda) = f(x, y) + \langle G(x, y), \lambda\rangle.$$

Fix the multiplier vector $\bar{\lambda}$ satisfying $0 = \nabla_y \mathcal{L}(\bar{x}, \bar{y}, \bar{\lambda})$ and define the Hessian matrices

$$H_{xx} = \nabla^2_{xx}\mathcal{L}(\bar{x}, \bar{y}, \bar{\lambda}), \qquad H_{xy} = \nabla^2_{xy}\mathcal{L}(\bar{x}, \bar{y}, \bar{\lambda}), \qquad H_{yy} = \nabla^2_{yy}\mathcal{L}(\bar{x}, \bar{y}, \bar{\lambda}).$$

Then, for every x near $\bar{x}$, the problem $\mathcal{Q}_x$ admits a unique solution $y(x)$, which varies C^1-smoothly. Moreover, its directional derivative in direction h given by

$$\begin{aligned} \nabla y(\bar{x})h \quad &= \quad \operatorname*{argmin}_v \; 2\langle H_{xy}v, h\rangle + \langle H_{yy}v, v\rangle \\ &\qquad \text{s.t.} \quad \nabla_x G(\bar{x}, \bar{y})h + \nabla_y G(\bar{x}, \bar{y})v = 0. \end{aligned} \tag{5.7}$$

With these tools in hand, we now prove Theorem 5.3.

Proof of Theorem 5.3 Let $\bar{v}, \bar{w}, \mathcal{H}, \mathcal{R}$, and $\mathcal{M}$ be the vectors and manifolds specified in Definition 5.1. It is known from [40, Theorem 4.11] that for all x near $\bar{x}$, the inclusions hold:

$$S(x) \in \mathcal{M} \qquad \text{and} \qquad F(x) + \nabla F(x)(S(x) - x) \in \mathcal{H}.$$

From this inclusion, we will be able to view the proximal subproblem through the lens of the perturbation result in Theorem 5.4. For the sake of completeness, however, we provide a quick proof. Indeed, consider a sequence $x_i \to \bar{x}$ and define $z_i = F(x_i) + \nabla F(x_i)(S(x_i) - x_i)$. Then, appealing to the optimality conditions of the proximal linear subproblem, we deduce that there exist vectors $v_i \in \partial r(x_i)$ and $w_i \in \partial h(z_i)$ satisfying $\frac{1}{\mu}(x_i - S(x_i)) = \nabla F(x_i)^* w_i + v_i$. Since $S(\cdot)$ is continuous and h is Lipschitz, the vectors w_i and v_i are bounded. Passing to a subsequence, we may assume that w_i and v_i converge to some $w \in \partial h(F(\bar{x}))$ and $v \in \partial r(\bar{x})$, respectively, and moreover, that

$$0 \in \nabla F(\bar{x})^* w + v.$$

We therefore deduce $w = \bar{w}$ and $v = \bar{v}$. Taking into account that $\mathcal{R}$ is a C^2-active manifold at $\bar{x}$ for $\bar{v}$ and $\mathcal{H}$ is a C^2-active manifold at $F(\bar{x})$ for $\bar{w}$, we deduce $S(x_i) \in \mathcal{R}$ and $z_i \in \mathcal{H}$ for all large indices i, proving the claim.

Turning to the perturbation result, let $\hat{h}\colon \mathbb{R}^m \to \mathbb{R}$ be any C^2-smooth function agreeing with h on a neighborhood of $F(\bar{x})$ in $\mathcal{H}$, and let $\hat{r}\colon \mathbb{R}^d \to \mathbb{R}$ be any C^2-smooth function agreeing with r on a neighborhood of $\bar{x}$ in $\mathcal{R}$. Applying the claim, we find that for x near $\bar{x}$, we may write

SPI Journal: 10208 Article No.: 9516 TYPESET DISK LE CP Disp.:2021/4/28 Pages: 46 Layout: Small-Ex

$$S(x) \quad = \quad \underset{y}{\operatorname{argmin}} \ \hat{h}\left(F(x)+\nabla F(x)(y-x)\right)+\hat{r}(y)+\frac{1}{2\mu}\|y-x\|^2 \qquad (5.8)$$
$$\text{s.t.} \quad F(x)+\nabla F(x)(y-x)\in\mathcal{H} \quad \text{and} \quad y\in\mathcal{R}$$

Our goal is to apply the perturbation result (Theorem 5.4) to the parametric family (5.8). To this end, let $\omega = 0$ be the local defining equations of $\mathcal{H}$ around $F(\bar{x})$ and let $\eta = 0$ be the local defining equation of $\mathcal{R}$ around $\bar{x}$. We can now place (5.8) in the setting of Theorem 5.4 by setting

$$f(x, y) = \hat{h}\left(F(x)+\nabla F(x)(y-x)\right)+\hat{r}(y)+\frac{1}{2\mu}\|y-x\|^2$$

and

$$G(x, y) := (G^{\mathcal{H}}(x, y), G^{\mathcal{R}}(x, y)) := (\omega(F(x)+\nabla F(x)(y-x)), \eta(y)).$$

For these functions, we now verify the assumptions of Theorem 2.3. First, the nondegeneracy property follows from the transversality condition (5.4). Second, we verify the quadratic growth condition: since we have chosen $\mu < \rho^{-1}$, it follows that for every $x \in \mathbb{R}^d$ the function $y \mapsto f(x, y)$ is strongly convex with the constant $\mu^{-1} - \rho$. Finally, we verify the level-boundedness condition: since the minimizer $S(x)$ clearly varies continuously and satisfies $S(\bar{x}) = \bar{x}$, the conditions of Lemma 2.4 are satisfied. Therefore, the assumptions of Theorem 2.3 are valid. In particular, we learn that the solution map $S(\cdot)$ is C^1-smooth around $\bar{x}$.

Computing the Jacobian of the solution mapping will occupy the remainder of the proof. To that end, define the parametric Lagrangian

$$\mathcal{L}(x, y, \lambda) = f(x, y) + \langle G(x, y), \lambda\rangle.$$

Localizing, the identification properties then entail that $y = \bar{x}$ is a minimizer of the problem (5.8) corresponding to $x = \bar{x}$. We conclude there exists a Lagrange multiplier vector $\bar{\lambda} = (\bar{\lambda}^{\mathcal{H}}, \bar{\lambda}^{\mathcal{R}})$ satisfying $0 = \nabla_y \mathcal{L}(\bar{x}, \bar{x}, \bar{\lambda})$, a fact we will return to after a few calculations.

We now compute the first-order variations of f and G. To simplify notation, we adopt two conventions. First, we align the notation of gradients and Jacobians, viewing every gradient as a row vector. Second, we let the symbol $\nabla^2 F[x; v]$ denote the $m \times d$ matrix whose ith row equals $v^\top \nabla^2 F_i(x)$. Then, defining the map

$$\zeta(x, y) = F(x) + \nabla F(x)(y - x),$$

a quick computation shows

$$\nabla_y \zeta(x, y) = \nabla F(x) \qquad \text{and} \qquad \nabla_x \zeta(x, y) = \nabla^2 F[x, y - x].$$

SPI Journal: 10208 Article No.: 9516 TYPESET DISK LE CP Disp.:2021/4/28 Pages: 46 Layout: Small-Ex

Author Proof
uncorrected proof

Therefore, using the chain rule, we compute the first-order variations

$$\nabla_x G^{\mathcal{H}}(x,y) = \nabla\omega\,(\zeta(x,y)) \cdot \nabla^2 F[x, y-x]$$
$$\nabla_y G^{\mathcal{H}}(x,y) = \nabla\omega(\zeta(x,y)) \cdot \nabla F(x)$$
$$\nabla_x G^{\mathcal{R}}(x,y) = 0$$
$$\nabla_y G^{\mathcal{R}}(x,y) = \nabla\eta(y)$$
$$\nabla_x f(x,y) = \nabla\hat{h}\,(\zeta(x,y)) \cdot \nabla^2 F[x, y-x] + \mu^{-1}(x-y)^\top$$
$$\nabla_y f(x,y) = \nabla\hat{h}\,(\zeta(x,y)) \cdot \nabla F(x) + \nabla\hat{r}(y) + \mu^{-1}(y-x)^\top.$$

From these variations we deduce $\nabla_x G(\bar{x},\bar{x}) = 0$ and therefore the constraint in (5.7) simply amounts to the inclusion

$$\begin{aligned} v \in \ker \nabla_y G(\bar{x},\bar{x}) &= \Big(\ker \nabla\eta(\bar{x})\Big) \cap \Big(\ker(\nabla\omega(F(\bar{x})) \cdot \nabla F(\bar{x}))\Big) \\ &= T_{\mathcal{R}}(\bar{x}) \cap \nabla F(\bar{x})^{-1} T_{\mathcal{H}}(F(\bar{x})) = T_{\mathcal{M}}(\bar{x}). \end{aligned} \tag{5.9}$$

In particular, formula (5.7) reduces to

$$\nabla S(\bar{x})h = \operatorname*{argmin}_{v \in T_{\mathcal{M}}(\bar{x})} \; 2\langle H_{xy}v, h\rangle + \langle H_{yy}v, v\rangle, \tag{5.10}$$

To find an explicit solution, we mirror the analysis of the proximal gradient method. We let W be the orthogonal projection onto $T_{\mathcal{M}}(\bar{x})$ and define the linear maps $\overline{H_{yy}} : T_{\mathcal{M}}(\bar{x}) \to T_{\mathcal{M}}(\bar{x})$ and $\overline{H_{xy}} : T_{\mathcal{M}}(\bar{x}) \to T_{\mathcal{M}}(\bar{x})$ by setting $\overline{H_{yy}} = WH_{yy}W$ and $\overline{H_{xy}} = WH_{xy}W$, respectively. Since $\bar{x}$ is a strong local minimizer of (5.7), the map $\overline{H_{yy}}$ is positive definite and invertible. Solving (5.7) then yields the expression

$$\nabla S(\bar{x})h = -\overline{H}_{yy}^{-1}\overline{H_{xy}}^\top h \qquad \text{for all } h \in T_{\mathcal{M}}(\bar{x}).$$

Let us now verify that if $\bar{x}$ is a composite strict saddle of f, then $\nabla S(\bar{x})$ has a real eigenvalue that is greater than one. To this end, observe that $\gamma \in \mathbb{R}$ is an eigenvalue of $\nabla S(\bar{x})$ with an associated eigenvector $v \in T_{\mathcal{M}}(\bar{x})$ if and only if

$$\nabla S(\bar{x})v = \gamma v \quad \iff \quad -\overline{H}_{yy}^{-1}\overline{H_{xy}}^\top v = \gamma v \quad \iff \quad (\gamma\overline{H}_{yy} + \overline{H_{xy}}^\top)v = 0.$$

In particular, if the matrix $\gamma\overline{H}_{yy} + \overline{H_{xy}}^\top$ is singular, then γ is an eigenvalue of $\nabla S(\bar{x})$. To prove such a $\gamma \geq 1$ exists, we will show that $\overline{H}_{xy}$ is self-adjoint, and then, we will examine the following ray of symmetric matrices

$$\{\gamma\overline{H}_{yy} + \overline{H_{xy}}^\top : \gamma \geq 1\}.$$

Beginning with the end point, we will show that the matrix $\overline{H}_{yy} + \overline{H_{xy}}^\top$ has a strictly negative eigenvalue. On the other hand, we already know the direction of the ray

SPI Journal: 10208 Article No.: 9516 TYPESET DISK LE CP Disp.:2021/4/28 Pages: 46 Layout: Small-Ex
Author Proof
uncorrected proof

Author Proof

$\overline{H}_{yy}$ is a positive definite matrix. Therefore, by continuity of the minimal eigenvalue function, there will exist some $\gamma > 1$ such that the matrix $\gamma \overline{H}_{yy} + \overline{H_{xy}}$ is singular, as claimed.

To this end, we now compute the second-order variations.

$$\begin{aligned}
\nabla_{xy} G_i^{\mathcal{H}}(x, y)v &= \nabla^2 F[x; v]^\top \nabla \omega_i(\zeta(x, y))^\top \\
&\quad + \nabla^2 F[x; y - x]^\top \nabla^2 \omega_i(\zeta(x, y)) \nabla F(x) v \\
\nabla_{yy} G_i^{\mathcal{H}}(x, y)v &= \nabla F(x)^\top \nabla^2 \omega_i(\zeta(x, y)) \nabla F(x) v \\
\nabla_{xy} f(x, y)v &= \nabla^2 F[x; v]^\top \nabla \hat{h}(\zeta(x, y))^\top \\
&\quad + \nabla^2 F[x; y - x] \nabla^2 \hat{h}(\zeta(x, y)) \nabla F(x) v - \mu^{-1} v \\
\nabla_{yy} f(x, y)v &= \nabla F(x)^\top \nabla^2 \hat{h}(\zeta(x, y)) \nabla F(x) v + \nabla^2 \hat{r}(y) v + \mu^{-1} v.
\end{aligned}$$

A quick computation then shows that $\nabla_{xy} f(\bar{x}, \bar{x})$ and $\nabla_{xy} G_i^{\mathcal{H}}(\bar{x}, \bar{x})$ are self-adjoint operators. Consequently, we obtain $H_{xy} = H_{xy}^\top$ and the expression

$$\begin{aligned}
(H_{yy} + H_{xy}^\top)v &= \nabla F(\bar{x})^\top \nabla^2 \hat{h}(F(\bar{x})) \nabla F(\bar{x}) v + \nabla^2 \hat{r}(\bar{x}) v + \nabla^2 F[\bar{x}; v]^\top \nabla \hat{h}(F(\bar{x}))^\top \\
&\quad + \sum_{i \geq 1} \bar{\lambda}_i^{\mathcal{H}} \left(\nabla F(\bar{x})^\top \nabla^2 \omega_i(F(\bar{x})) \nabla F(\bar{x}) v + \nabla^2 F[\bar{x}; v]^\top \nabla \omega_i(F(\bar{x}))^\top \right) \\
&\quad + \sum_{i \geq 1} \bar{\lambda}_i^{\mathcal{R}} \nabla^2 \eta_i(y) v.
\end{aligned}$$

To prove that $H_{yy} + H_{xy}^\top$ has a strictly negative eigenvalue, we will show that it coincides with the Hessian of the Lagrangian of the problem:

$$\min_x \ \hat{h}(F(x)) + \hat{r}(x) \qquad \text{subject to} \qquad \omega(F(x)) = 0, \eta(x) = 0.$$

Indeed, define the Lagrangian function

$$\mathcal{L}_0(x, \lambda) = \hat{h}(F(x)) + \hat{r}(x) + \sum_{i \geq 1} \lambda_i^{\mathcal{H}} \omega(F(x)) + \sum_{i \geq 1} \lambda_i^{\mathcal{R}} \eta(x).$$

A quick computation shows

$$\begin{aligned}
\nabla^2 (\hat{h} \circ F)(x) v &= \nabla F(x)^\top \nabla^2 \hat{h}(F(x)) \nabla F(x) v + \nabla^2 F[x, v]^\top \nabla \hat{h}(F(x))^\top \\
\nabla^2 (\omega_i \circ F)(x) v &= \nabla F(x)^\top \nabla^2 \omega_i(F(x)) \nabla F(x) v + \nabla^2 F[x, v]^\top \nabla \omega_i(F(x))^\top
\end{aligned}$$

and therefore the equality

$$\nabla^2 \mathcal{L}_0(\bar{x}, \bar{\lambda}) = H_{yy} + H_{xy}^\top.$$

The composite strict saddle property guarantees that the matrix $\nabla^2 \mathcal{L}_0(\bar{x}, \bar{\lambda})$ has a strictly negative eigenvalue, completing the proof. □

uncorrected proof

Author Proof

In line with the previous sections, one could ask whether a damped and randomly initialized proximal linear method almost surely escapes all composite strict saddle points. An immediate obstacle is that the global Lipschitz constant of the proximal linear map $S(\cdot)$ defined in (5.6) seems unclear, and therefore, we are unable to find an appropriate damping parameter. Instead we will settle for a local escape guarantee supplied by the center stable manifold theorem. We leave it as an intriguing open question to obtain global escape guarantees for the damped proximal linear algorithm.

A first difficulty in applying the center stable manifold theorem is that the Jacobian $\nabla S(\bar{x})$ at the saddle point $\bar{x}$ may not be invertible. Consequently, we will damp the proximal linear method, forcing the update to be a local diffeomorphism. To compute an appropriate threshold for the damping parameter, we will need to estimate the operator norm of $\nabla S(\bar{x})$. This is the content of the following lemma.

Lemma 5.5 (The slope at the critical points). *Consider the composite optimization problem (5.1) and choose any $\mu \in (0, (\rho + 2\beta)^{-1})$. Then, for all points $x \in \mathbb{R}^d$ and all critical points $\bar{x} \in \mathbb{R}^d$, the proximal linear map $S(\cdot)$ defined in (5.6) satisfies*

$$\|S(x) - \bar{x}\| \leq \left(1 + \sqrt{\frac{2\beta\mu}{1 - \mu\beta - \mu\rho}}\right) \cdot \max\left\{1, \frac{\mu\rho + \mu\beta}{1 - \mu\rho - 2\mu\beta}\right\} \cdot \|x - \bar{x}\|.$$

Proof To simplify notation, define the map

$$\zeta(x, y) = F(x) + \nabla F(x)(y - x).$$

Set $\gamma := \mu^{-1} - \beta$, fix an arbitrary point $x \in \mathbb{R}^d$, and define

$$x^+ := S(x) \quad \text{and} \quad \hat{x} := \operatorname{prox}_{f/\gamma}(x).$$

Using strong convexity of the prox-linear and proximal subproblems and the estimate (5.2), we successively compute

$$\begin{aligned}
h(\hat{x}) + r(\hat{x}) + \frac{\gamma}{2}\|\hat{x} - x\|^2 &\leq h(x^+) + r(x^+) + \frac{\gamma}{2}\|x^+ - x\|^2 - \frac{\gamma - \rho - \beta}{2}\|x^+ - \hat{x}\|^2 \\
&\leq h(\zeta(x, x^+)) + r(x^+) + \frac{\gamma + \beta}{2}\|x^+ - x\|^2 - \frac{\gamma - \rho - \beta}{2}\|x^+ - \hat{x}\|^2 \\
&\leq h(\zeta(x, \hat{x})) + r(\hat{x}) + \frac{\gamma + \beta}{2}\|\hat{x} - x\|^2 - (\gamma - \rho)\|x^+ - \hat{x}\|^2 \\
&\leq h(\hat{x}) + r(\hat{x}) + \frac{\gamma + 2\beta}{2}\|\hat{x} - x\|^2 - (\gamma - \rho)\|x^+ - \hat{x}\|^2.
\end{aligned}$$

Rearranging yields the estimate

$$(\gamma - \rho)\|x^+ - \hat{x}\|^2 \leq 2\beta\|\hat{x} - x\|^2 = 2\beta\gamma^{-2}\|\nabla f_{1/\gamma}(x)\|^2.$$

SPI Journal: 10208 Article No.: 9516 TYPESET DISK LE CP Disp.:2021/4/28 Pages: 46 Layout: Small-Ex

Author Proof

Therefore, using Lipschitz continuity of the gradient $\nabla f_{1/\gamma}$ (Lemma 2.5) and the triangle inequality yields

$$\begin{aligned}\|x^+ - \bar{x}\| &\leq \left(\gamma^{-1} + \sqrt{\frac{2\beta\gamma^{-2}}{\gamma - \rho}}\right) \cdot \max\left\{\gamma, \frac{\rho + \beta}{1 - \gamma^{-1}(\rho + \beta)}\right\} \cdot \|x - \bar{x}\| \\ &= \left(1 + \sqrt{\frac{2\beta\mu}{1 - \mu\beta - \mu\rho}}\right) \cdot \max\left\{1, \frac{\mu\rho + \mu\beta}{1 - \mu\rho - 2\mu\beta}\right\} \cdot \|x - \bar{x}\|,\end{aligned}$$

as claimed. □

We are now ready to deduce that the damped proximal linear method almost locally escapes any composite strict saddle point.

Theorem 5.6 (Proximal linear method: local escape). *Consider the composite problem (5.1) and let $\bar{x}$ be any composite strict saddle point. Choose any constant $\mu \in (0, (\rho + 2\beta)^{-1})$ and a damping parameter $\alpha \in (0, 1)$ satisfying*

$$\alpha \cdot \left(1 + \left(\left(1 + \sqrt{\frac{2\beta\mu}{1 - \mu\beta - \mu\rho}}\right) \cdot \max\left\{1, \frac{\mu\rho + \mu\beta}{1 - \mu\rho - 2\mu\beta}\right\}\right)\right) < 1.$$

Define the damped proximal linear update

$$T(x) = (1 - \alpha)x + \alpha S(x),$$

where $S(\cdot)$ is the proximal linear map defined in (5.6). Then, there exists a neighborhood U of $\bar{x}$ such that the set of initial conditions

$$\{x \in U : S^k(x) \in U \text{ for all } k \geq 0\}$$

has zero Lebesgue measure.

Proof First, using Theorem 5.3 and Lemma 2.14, we deduce that $\bar{x}$ is an unstable fixed point of $\bar{x}$. Let us next verify that T is a local diffeomorphism around $\bar{x}$. To see this, observe

$$\nabla T(\bar{x}) = I - \alpha(I - \nabla S(\bar{x})).$$

Using Theorem 5.5, we deduce $\alpha\|I - \nabla S(\bar{x})\|_{\text{op}} < 1$ and therefore T is invertible. An application of the center stable manifold theorem (Theorem 2.10) completes the proof. □

uncorrected proof

SPI Journal: 10208 Article No.: 9516 TYPESET DISK LE CP Disp.:2021/4/28 Pages: 46 Layout: Small-Ex

Author Proof

6 Convergence of Relaxed Descent Methods

Thus far, all of our escape theorems made an assumption that the iterate sequence generated by the algorithms converges. In this section, we verify this assumption for the damped proximal point, proximal gradient, and proximal linear methods. Taking a general view, we see that the iterative methods of this paper can be understood within a broad family of damped model-based algorithms for minimizing a function f. These algorithms construct iterates $x_0, x_1 \dots$ by repeatedly minimizing a local model $f_x(\cdot)$ of the function and moving in the direction of its minimizer. More specifically, in the section we suppose that there exist constant $\rho, \eta, \beta > 0$ such that the the following properties hold:

(A1) The function $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is closed and ρ-weakly convex.
(A2) For all $x \in \mathbb{R}^d$ there exists a closed η-weakly convex function $f_x : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ satisfying

$$|f(y) - f_x(y)| \leq \frac{\beta}{2}\|y - x\|^2 \qquad \text{for all } y \in \mathbb{R}^d.$$

Under these assumptions, we will study how the following algorithm behaves: given iterates $x_0, \dots, x_t$ define

$$\begin{aligned} y_t &= \operatorname*{argmin}_{y \in \mathbb{R}^d} \left\{ f_{x_t}(y) + \frac{\tau}{2}\|y - x_t\|^2 \right\} \\ x_{t+1} &= (1 - \alpha)x_t + \alpha y_t, \end{aligned} \qquad (\mathcal{MBA})$$

where $\tau > 0$ and $\alpha > 0$ are fixed constants, determined below.

To analyze this algorithm, we rely on the seminal paper [3]. There, the authors identified three conditions, guaranteeing global convergence of a sequence $\{z_t\}$ of "algorithm iterates" to a critical point of a closed function $g : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$. Namely, they assume there exist $a, b > 0$ such that the following holds:

(B1) **(Sufficient Decrease.)** For each $t \in \mathbb{N}$, we have

$$g(z_{t+1}) + a\|z_{t+1} - z_t\|^2 \leq g(z_t)$$

(B2) **(Relative Error Conditions.)** For each $t \in \mathbb{N}$ there exists $w_{t+1} \in \partial g(z_{t+1})$ such that

$$\|w_{t+1}\| \leq b\|z_{t+1} - z_t\|$$

(B3) **(Continuity Condition.)** There exists a subsequence $\{z_{t_j}\}$ and $\tilde{z}$ such that

$$z_{t_j} \to \tilde{z} \text{ and } g(z_{t_j}) \to g(\tilde{z}), \qquad \text{as } j \to \infty.$$

The above assumptions alone may not guarantee that z_t converges to a critical point of g. Instead, the authors of [3] restrict their focus to the broad class of functions satisfying the *Kurdyka–Łojasiewicz property*.

uncorrected proof

SPI Journal: 10208 Article No.: 9516 TYPESET DISK LE CP Disp.:2021/4/28 Pages: 46 Layout: Small-Ex

Author Proof

Definition 6.1 (KŁ *Function*) Let $g\colon \mathbb{R}^d \to \mathbb{R}\cup\{\infty\}$ be a closed function. We say that g has the *Kurdyka–Łojasiewicz (KL) property* at a point $\bar{x}$, where $\partial g(\bar{x})$ is nonempty, if there exists $\varepsilon \in (0, +\infty]$, a neighborhood U of $\bar{x}$, and a continuous convex function $\varphi\colon [0, \varepsilon) \to \mathbb{R}_+$ satisfying

1. $\varphi(0) = 0$,
2. φ is C^1 on $(0, \varepsilon)$ with $\varphi' > 0$, and
3. the KŁ inequality

$$\mathrm{dist}(0, \partial g(x)) \geq \frac{1}{\varphi'(g(x) - g(\bar{x}))},$$

holds for all $x \in U$ satisfying $g(\bar{x}) < g(x) < g(\bar{x}) + \varepsilon$.

If g satisfies the KŁ property at each point x, with $\partial g(x) \neq \emptyset$, then g is called a *KŁ function.*

The class of KŁ functions is broad, containing all closed semialgebraic functions and more broadly any functions definable in an o-minimal structure, as shown in the pioneering work [7]. Under these assumptions we have the following theorem from [3, Theorem 2.9].

Theorem 6.2 *Let $g\colon \mathbb{R}^d \to \mathbb{R}\cup\{\infty\}$ be a closed function. Consider a sequence x_t that satisfies (B1), (B2), and (B3). If g satisfies the KŁ property at some cluster point $\tilde{x}$, then $\tilde{x}$ is a critical point of g, the entire sequence x_k converges to $\tilde{x}$, and the sequence x_t has finite length*

$$\sum_{t=0}^{\infty} \|x_{t+1} - x_t\| < +\infty.$$

In the remainder of this section, we will verify assumption (B1), (B2), and (B3) for the sequence $\{z_t\} = \{x_t\}$ and the Moreau envelope $g := f_{1/\hat{\rho}}$, where $\hat{\rho}$ will be chosen in a moment. Since the critical points of f and $f_{1/\hat{\rho}}$ agree, the result will imply convergence to critical points of f. To do so, we employ one final assumption.

(A3) For every $\hat{\rho} > 0$, the Moreau envelope $f_{1/\hat{\rho}}$ is a KŁ function.

Although assumption (A3) may appear hard to verify, it holds whenever f is semialgebraic since in this case $f_{1/\hat{\rho}}$ is also semialgebraic. More generally, the analogous statement holds if f is definable in an o-minimal structure. The following is the main result of this section.

Theorem 6.3 (Convergence of relaxed model-based methods). *Suppose that* $\alpha \in (0, 1]$, *that* $\tau > \max\{\eta, 2\rho, \frac{4\beta+\rho+\eta}{2}\}$, *and that assumptions (A1) and (A2) hold. Then, for all $T \geq 0$, we have*

$$\min_{t=0,\ldots,T} \|\nabla f_{1/\hat{\rho}}(x_t)\| \leq \sqrt{\frac{f_{1/\hat{\rho}}(x_0) - \inf f}{\frac{\alpha(2\hat{\rho}-\rho-\eta-\beta)}{2\hat{\rho}(\hat{\rho}+\tau-\rho-\eta)}(T+1)}}.$$

uncorrected proof

SPI Journal: 10208 Article No.: 9516 TYPESET DISK LE CP Disp.:2021/4/28 Pages: 46 Layout: Small-Ex

Author Proof

where $\hat{\rho} = (1/2)\tau + (1/4)(\rho + \eta)$. Moreover, if (A3) also holds and the sequence $\{x_t\}$ has a cluster point $\bar{x}$, then $\bar{x}$ is critical for f and the entire sequence $\{x_t\}$ converges to $\bar{x}$. Moreover, the sequence $\{x_t\}$ has finite length.

$$\sum_{t=0}^{\infty} \|x_{t+1} - x_t\| < +\infty.$$

This result is new and may be of independent interest. In particular, the conclusion of the theorem extends the convergence guarantees for the proximal linear method developed in [52] to all relaxed model-based algorithms.

6.1 Proof of Theorem 6.3

We are free to choose the parameter $\hat{\rho}$ defining the Moreau envelope. To this end, we will need the existence of a parameter $\hat{\rho}$, satisfying the following inequalities.

Lemma 6.4 *Under the assumptions of Theorem 6.3, it holds that $\hat{\rho} > \rho$ and*

1. $\tau - \hat{\rho} - \beta > 0$,
2. $2\hat{\rho} - \rho - \eta - \beta > 0$,
3. $\hat{\rho} + \tau - \rho - \eta > 0$,
4. $1 - \frac{2\hat{\rho}-\rho-\eta-\beta}{\hat{\rho}+\tau-\rho-\eta} > 0$.

Proof Note that $\hat{\rho} > \tau/2 > \rho > 0$ and that $\hat{\rho} = \tau - \beta - \varepsilon/2$ for $\varepsilon = (2\tau - 4\beta - \rho - \eta)/2 > 0$. To prove the first inequality, notice that $\tau - \hat{\rho} - \beta = \varepsilon/2 > 0$. To prove the second inequality, notice that

$$2\hat{\rho} - \rho - \eta - \beta = 2\tau - 4\beta - \rho - \eta - \varepsilon = \varepsilon > 0.$$

To prove the third inequality, observe

$$\hat{\rho} + \tau - \rho - \eta > \hat{\rho} + (\hat{\rho} + \beta) - \rho - \eta \geq 2\hat{\rho} - \rho - \eta - \beta > 0,$$

where the first and second inequalities follow from items 6.4 and 6.4, respectively. To prove the fourth inequality, we compute

$$1 - \frac{2\hat{\rho} - \rho - \eta - \beta}{\hat{\rho} + \tau - \rho - \eta} = \frac{\beta + \tau - \hat{\rho}}{\hat{\rho} + \tau - \rho - \eta} = \frac{2\beta + \varepsilon/2}{\hat{\rho} + \tau - \rho - \eta} > 0,$$

as desired. □

Throughout the rest of this section, we fix a constant $\hat{\rho}$ satisfying the conditions of Lemma 6.4. Critical to our proof is the following lemma, comparing the proximal point

$$\hat{x}_t := \text{prox}_{f/\hat{\rho}}(x_t)$$

uncorrected proof

SPI Journal: 10208 Article No.: 9516 TYPESET DISK LE CP Disp.:2021/4/28 Pages: 46 Layout: Small-Ex

Author Proof

to the "approximately proximal point" y_t. A closely related estimate appeared in [16, Lemma 4.2], driving the convergence analysis of that paper.

Lemma 6.5 *It holds that*

$$\|\hat{x}_t - y_t\|^2 \le \|\hat{x}_t - x_t\|^2 - \frac{2\hat{\rho} - \rho - \eta - \beta}{\hat{\rho} + \tau - \rho - \eta}\|\hat{x}_t - x_t\|^2 - \frac{\tau - \hat{\rho} - \beta}{\hat{\rho} + \tau - \rho - \eta}\|x_t - y_t\|^2.$$

Proof Since the function $y \mapsto f(y) + \frac{\hat{\rho}}{2}\|y - x_t\|^2$ is $(\hat{\rho} - \rho)$-strongly convex and $\hat{x}_t$ is its minimizer, we have

$$\frac{\hat{\rho} - \rho}{2}\|\hat{x}_t - y_t\|^2 \le \left(f(y_t) + \frac{\hat{\rho}}{2}\|y_t - x_t\|^2\right) - \left(f(\hat{x}_t) + \frac{\hat{\rho}}{2}\|\hat{x}_t - x_t\|^2\right).$$

Consequently, using the double-sided model property (A2), we find

$$\frac{\hat{\rho} - \rho}{2}\|\hat{x}_t - y_t\|^2 \le f_{x_t}(y_t) - f_{x_t}(\hat{x}_t) + \frac{\hat{\rho} + \beta}{2}\|x_t - y_t\|^2 - \frac{\hat{\rho} - \beta}{2}\|\hat{x}_t - x_t\|^2. \quad (6.1)$$

Since the function $y \mapsto f_{x_t}(y) + \frac{\tau}{2}\|y - x_t\|^2$ is $(\tau - \eta)$-strongly convex and y_t is its minimizer, we have

$$f_{x_t}(y_t) - f_{x_t}(\hat{x}_t) \le \frac{\tau}{2}\|\hat{x}_t - x_t\|^2 - \frac{\tau}{2}\|y_t - x_t\|^2 - \frac{\tau - \eta}{2}\|y_t - \hat{x}_t\|^2.$$

Combining this estimate with (6.1), we compute

$$\begin{aligned}\frac{\hat{\rho} - \rho}{2}\|\hat{x}_t - y_t\|^2 &\le \frac{\tau}{2}\|\hat{x}_t - x_t\|^2 - \frac{\tau}{2}\|y_t - x_t\|^2 - \frac{\tau - \eta}{2}\|y_t - \hat{x}_t\|^2 \\ &\quad + \frac{\hat{\rho} + \beta}{2}\|x_t - y_t\|^2 - \frac{\hat{\rho} - \beta}{2}\|\hat{x}_t - x_t\|^2 \\ &= \frac{\beta + \tau - \hat{\rho}}{2}\|\hat{x}_t - x_t\|^2 + \frac{\hat{\rho} + \beta - \tau}{2}\|x_t - y_t\|^2 - \frac{\tau - \eta}{2}\|y_t - \hat{x}_t\|^2.\end{aligned}$$

Rearranging, we conclude

$$\frac{\hat{\rho} + \tau - \rho - \eta}{2}\|\hat{x}_t - y_t\|^2 \le \frac{\beta + \tau - \hat{\rho}}{2}\|\hat{x}_t - x_t\|^2 + \frac{\hat{\rho} + \beta - \tau}{2}\|x_t - y_t\|^2.$$

Dividing both sides by $\frac{\hat{\rho}+\tau-\rho-\eta}{2}$, we achieve the result:

$$\begin{aligned}\|\hat{x}_t - y_t\|^2 &\le \frac{\beta + \tau - \hat{\rho}}{\hat{\rho} + \tau - \rho - \eta}\|\hat{x}_t - x_t\|^2 + \frac{\hat{\rho} + \beta - \tau}{\hat{\rho} + \tau - \rho - \eta}\|x_t - y_t\|^2 \\ &= \|\hat{x}_t - x_t\|^2 - \left(1 - \frac{\beta + \tau - \hat{\rho}}{\hat{\rho} + \tau - \rho - \eta}\right)\|\hat{x}_t - x_t\|^2 + \frac{\hat{\rho} + \beta - \tau}{\hat{\rho} + \tau - \rho - \eta}\|x_t - y_t\|^2 \\ &= \|\hat{x}_t - x_t\|^2 - \frac{2\hat{\rho} - \rho - \eta - \beta}{\hat{\rho} + \tau - \rho - \eta}\|\hat{x}_t - x_t\|^2 - \frac{\tau - \hat{\rho} - \beta}{\hat{\rho} + \tau - \rho - \eta}\|x_t - y_t\|^2.\end{aligned}$$

This completes the proof of the lemma. □

uncorrected proof

SPI Journal: 10208 Article No.: 9516 TYPESET DISK LE CP Disp.:2021/4/28 Pages: 46 Layout: Small-Ex

Author Proof

The following lemma verifies the Assumption (B1).

Lemma 6.6 (Sufficient Decrease) *We have*

$$f_{1/\hat{\rho}}(x_{t+1}) \leq f_{1/\hat{\rho}}(x_t) - \frac{\hat{\rho}(\tau - \hat{\rho} - \beta)}{2\alpha(\hat{\rho} + \tau - \rho - \eta)}\|x_{t+1} - x_t\|^2$$
$$- \frac{\alpha(2\hat{\rho} - \rho - \eta - \beta)}{2\hat{\rho}(\hat{\rho} + \tau - \rho - \eta)}\|\nabla f_{1/\hat{\rho}}(x_t)\|^2.$$

In particular, $f_{1/\hat{\rho}}$ and $\{x_t\}$ satisfy (B1). Moreover, for all $T \geq 0$, we have

$$\min_{t=0,\ldots,T} \|\nabla f_{1/\hat{\rho}}(x_t)\|^2 \leq \frac{1}{T+1}\sum_{t=0}^{T} \|\nabla f_{1/\hat{\rho}}(x_t)\|^2 \leq \frac{f_{1/\hat{\rho}}(x_0) - \inf f}{\frac{\alpha(2\hat{\rho}-\rho-\eta-\beta)}{2\hat{\rho}(\hat{\rho}+\tau-\rho-\eta)}(T+1)}$$

Proof We successively compute

$$\begin{aligned}
f_{1/\hat{\rho}}(x_{t+1}) &= f(\hat{x}_{t+1}) + \frac{\hat{\rho}}{2}\|\hat{x}_{t+1} - x_{t+1}\|^2 \\
&\leq f(\hat{x}_t) + \frac{\hat{\rho}}{2}\|\hat{x}_t - x_{t+1}\|^2 \\
&= f(\hat{x}_t) + \frac{\hat{\rho}}{2}\|(1-\alpha)(\hat{x}_t - x_t) + \alpha(\hat{x}_t - y_t)\|^2 \\
&\leq f(\hat{x}_t) + \frac{\hat{\rho}(1-\alpha)}{2}\|\hat{x}_t - x_t\|^2 + \frac{\hat{\rho}\alpha}{2}\|\hat{x}_t - y_t\|^2 \\
&\leq f(\hat{x}_t) + \frac{\hat{\rho}}{2}\|\hat{x}_t - x_t\|^2 \\
&\quad - \frac{\hat{\rho}\alpha}{2}\left(\frac{2\hat{\rho} - \rho - \eta - \beta}{\hat{\rho} + \tau - \rho - \eta}\|\hat{x}_t - x_t\|^2 + \frac{\tau - \hat{\rho} - \beta}{\hat{\rho} + \tau - \rho - \eta}\|x_t - y_t\|^2\right) \\
&\leq f_{1/\hat{\rho}}(x_t) - \frac{\hat{\rho}\alpha(\tau - \hat{\rho} - \beta)}{2(\hat{\rho} + \tau - \rho - \eta)}\|x_t - y_t\|^2 - \frac{\alpha(2\hat{\rho} - \rho - \eta - \beta)}{2\hat{\rho}(\hat{\rho} + \tau - \rho - \eta)}\|\nabla f_{1/\hat{\rho}}(x_t)\|^2,
\end{aligned} \tag{6.2}$$

where (6.2) follows from Lemma 6.5, and the final inequality follows since $\hat{\rho}(x_t - \hat{x}_t) = \nabla f_{1/\hat{\rho}}(x_t)$. To get the descent inequality, it remains to write $x_t - y_t = (x_{t+1} - x_t)/\alpha$. Finally, the bound on the average gradient norm follows by induction. □

The following lemma verifies the Assumption (B2).

Lemma 6.7 (Relative Error). *It holds*

$$\|\nabla f_{1/\hat{\rho}}(x_{t+1})\| \leq \left(\max\left\{\hat{\rho}, \frac{\rho}{1-\rho/\hat{\rho}}\right\} + \frac{\hat{\rho}}{\alpha}\frac{1}{1 - \sqrt{\left(1 - \frac{2\hat{\rho}-\rho-\eta-\beta}{\hat{\rho}+\tau-\rho-\eta}\right)}}\right)\|x_{t+1} - x_t\|.$$

In particular, $f_{1/\hat{\rho}}$ and $\{x_t\}$ satisfy (B2).

SPI Journal: 10208 Article No.: 9516 TYPESET DISK LE CP Disp.:2021/4/28 Pages: 46 Layout: Small-Ex

uncorrected proof

Author Proof

Proof We have

$$\|\nabla f_{1/\hat{\rho}}(x_{t+1})\| \leq \|\nabla f_{1/\hat{\rho}}(x_t)\| + \max\left\{\hat{\rho}, \frac{\rho}{1-\rho/\hat{\rho}}\right\} \|x_{t+1} - x_t\|.$$

Thus, we want to bound

$$\|\nabla f_{1/\hat{\rho}}(x_t)\| = \hat{\rho}\|\hat{x}_t - x_t\|$$

by a multiple of $\|x_{t+1} - x_t\|$. This follows by Lemma 6.5:

$$\|\hat{x}_t - x_t\| \leq \|\hat{x}_t - y_t\| + \|y_t - x_t\| \leq \sqrt{\left(1 - \frac{2\hat{\rho} - \rho - \eta - \beta}{\hat{\rho} + \tau - \rho - \eta}\right)}\|x_t - \hat{x}_t\| + \|y_t - x_t\|$$

Rearranging and using the definition $x_t - y_t = (x_{t+1} - x_t)/\alpha$, it holds

$$\begin{aligned}\|\hat{x}_t - x_t\| &\leq \frac{1}{1 - \sqrt{\left(1 - \frac{2\hat{\rho}-\rho-\eta-\beta}{\hat{\rho}+\tau-\rho-\eta}\right)}}\|y_t - x_t\| \\ &= \frac{1}{\alpha}\frac{1}{1 - \sqrt{\left(1 - \frac{2\hat{\rho}-\rho-\eta-\beta}{\hat{\rho}+\tau-\rho-\eta}\right)}}\|x_{t+1} - x_t\|.\end{aligned}$$

The proof is complete. as desired. □

Finally, we can dispense with Assumption (B3), which is a simple consequence of the continuity of $f_{\hat{\rho}}$.

Lemma 6.8 (Continuity Condition). *The function $f_{\hat{\rho}}$ and the sequence $\{x_t\}$ satisfy (B3).*

Acknowledgements We thank John Duchi for his insightful comments on an early version of the manuscript. We also thank the anonymous referees for numerous suggestions that have improved the readability of the paper.

A Proofs of Theorems 2.9 and 5.2

In this section, we prove Theorem 2.9. We should note that Theorem 2.9, appropriately restated, holds much more broadly beyond the weakly convex function class. To simplify the notational overhead, however, we impose the weak convexity assumption, throughout.

We will require some basic notation from variational analysis; for details, we refer the reader to [57]. A set-valued map $F\colon \mathbb{R}^d \rightrightarrows \mathbb{R}^m$ assigns to each point $x \in \mathbb{R}^d$ a set $F(x)$ in $\mathbb{R}^m$. The graph of F is defined by

$$\text{gph } F := \{(x, v) : v \in F(x)\}.$$

uncorrected proof

Author Proof

A map $F\colon \mathbb{R}^d \rightrightarrows \mathbb{R}^m$ is called *metrically regular at* $(\bar{x}, \bar{v}) \in \operatorname{gph} F$ if there exists a constant $\kappa > 0$ such that the estimate holds:

$$\operatorname{dist}(x, F^{-1}(v)) \leq \kappa \operatorname{dist}(v, F(x))$$

for all x near $\bar{x}$ and all v near $\bar{v}$. If the graph $\operatorname{gph} F$ is a C^1-smooth manifold around $(\bar{x}, \bar{v})$, then metric regularity at $(\bar{x}, \bar{v})$ is equivalent to the condition [57, Theorem 9.43(d)]:[12]

$$(0, u) \in N_{\operatorname{gph} F}(\bar{x}, \bar{v}) \quad \Longrightarrow \quad u = 0. \tag{A.1}$$

We begin with the following lemma.

Lemma A.1 (Subdifferential metric regularity in smooth minimization). *Consider the optimization problem*

$$\min_{x \in \mathbb{R}^d} f(x) \quad \textit{subject to} \quad x \in \mathcal{M},$$

where $f\colon \mathbb{R}^d \to \mathbb{R}$ is a C^2-smooth function and $\mathcal{M}$ is a C^2-smooth manifold. Let $\bar{x} \in \mathcal{M}$ satisfy the criticality condition $0 \in \partial f_{\mathcal{M}}(\bar{x})$ and suppose that the subdifferential map $\partial f_{\mathcal{M}}\colon \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is metrically regular at $(\bar{x}, 0)$. Then, the guarantee holds:

$$\inf_{u \in \mathbb{S}^{d-1} \cap T_{\mathcal{M}}(\bar{x})} d^2 f_{\mathcal{M}}(\bar{x})(u) \neq 0. \tag{A.2}$$

Proof First, appealing to (A.1), we conclude that the implication holds:

$$(0, u) \in N_{\operatorname{gph} \partial f_{\mathcal{M}}}(\bar{x}, 0) \quad \Longrightarrow \quad u = 0. \tag{A.3}$$

Let us now interpret the condition (A.3) in Lagrangian terms. To this end, let $G = 0$ be the local defining equations for $\mathcal{M}$ around $\bar{x}$. Define the Lagrangian function

$$\mathcal{L}(x, \lambda) = f(x) + \langle G(x), \lambda \rangle,$$

and let $\bar{\lambda}$ be the unique Lagrange multiplier vector satisfying $\nabla_x \mathcal{L}(\bar{x}, \bar{\lambda}) = 0$. According to [41, Corollary 2.9], we have the following expression:

$$(0, u) \in N_{\operatorname{gph} \partial f_{\mathcal{M}}}(\bar{x}, 0) \quad \Longleftrightarrow \quad u \in T_{\mathcal{M}}(\bar{x}) \quad \text{and} \quad Lu \in N_{\mathcal{M}}(\bar{x}), \tag{A.4}$$

where $L := \nabla^2_{xx} \mathcal{L}(\bar{x}, \bar{\lambda})$ denotes the Hessian of the Lagrangian. Combining (A.3) and (A.4), we deduce that the only vector $u \in T_{\mathcal{M}}(\bar{x})$ satisfying $Lu \in N_{\mathcal{M}}(\bar{x})$ is the zero vector $u = 0$.

Now for the sake of contradiction, suppose that (A.2) fails. Then, the quadratic form $Q(u) = \langle Lu, u \rangle$ is nonnegative on $T_{\mathcal{M}}(\bar{x})$ and there exists $0 \neq \bar{u} \in T_{\mathcal{M}}(\bar{x})$ satisfying $Q(\bar{u}) = 0$. We deduce that $\bar{u}$ minimizes $Q(\cdot)$ on $T_{\mathcal{M}}(\bar{x})$, and therefore, the inclusion $L\bar{u} \in N_{\mathcal{M}}(\bar{x})$ holds, a clear contradiction. □

[12] We should note that metric regularity of F at $(\bar{x}, \bar{v})$ is equivalent to (A.1) for an arbitrary set-valued map F with closed graph, provided we interpret $N_{\operatorname{gph} F}(\bar{x}, \bar{v})$ as the limiting normal cone [57, Definition 6.3].

uncorrected proof

SPI Journal: 10208 Article No.: 9516 TYPESET DISK LE CP Disp.:2021/4/28 Pages: 46 Layout: Small-Ex

The following corollary for active manifolds will now quickly follow.

Corollary A.2 (Subdifferential metric regularity and active manifolds). *Consider a closed and weakly convex function* $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$. *Suppose that* f *admits a* C^2*-smooth active manifold around a critical point* $\bar{x}$ *and that the subdifferential map* $\partial f : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ *is metrically regular at* $(\bar{x}, 0)$. *Then,* $\bar{x}$ *is either a strong local minimizer of* f *or satisfies the curvature condition* $d^2 f_{\mathcal{M}}(\bar{x})(u) < 0$ *for some* $u \in T_{\mathcal{M}}(\bar{x})$.

Proof The result [19, Proposition 10.2] implies that gph ∂f coincides with gph $\partial f_{\mathcal{M}}$ on a neighborhood of $(\bar{x}, 0)$. Therefore, the subdifferential map $\partial f_{\mathcal{M}} : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is metrically regular at $(\bar{x}, 0)$. Using Lemma A.1, we obtain the guarantee:

$$\inf_{u \in \mathbb{S}^{d-1} \cap T_{\mathcal{M}}(\bar{x})} d^2 f_{\mathcal{M}}(\bar{x})(u) \neq 0.$$

If the infimum is strictly negative, the proof is complete. Otherwise, the infimum is strictly positive. In this case, $\bar{x}$ is a strong local minimizer of $f_{\mathcal{M}}$, and therefore by [19, Proposition 7.2] a strong local minimizer of f. □

We are now ready for the proofs of Theorems 2.9 and 5.2.

Proof of Theorem 2.9 The result [18, Corollary 4.8] shows that for almost all $v \in \mathbb{R}^d$, the function $g(x) := f(x) - \langle v, x \rangle$ has at most finitely many critical points. Moreover each such critical point $\bar{x}$ lies on some C^2 active manifold $\mathcal{M}$ of g and the subdifferential map $\partial g : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is metrically regular at $(\bar{x}, 0)$. Applying Corollary A.2 to g for such generic vectors v, we deduce that every critical point $\bar{x}$ of g is either a strong local minimizer or a strict saddle of g. The proof is complete. □

Proof of Theorem 5.2 The proof is identical to that of Theorem 2.9 with [18, Theorem 5.2] playing the role of [18, Corollary 4.8]. □

B Pathological Example

Theorem B.1 *Consider the following function*

$$f(x, y) = \frac{1}{2}(|x| + |y|)^2 - \frac{\rho}{2}x^2$$

Assume that $\lambda > \rho$. *Define a mapping* $T : \mathbb{R}^d \to \mathbb{R}$ *by the following formula.*

$$S(x, y) = \begin{cases} 0 & \text{if } (x, y) = 0; \\ \left(0, \frac{\lambda}{1+\lambda} y\right) & \text{if } |x| \leq \frac{1}{1+\lambda}|y|; \\ \left(\frac{\lambda}{1+\lambda-\rho} x, 0\right) & \text{if } |y| \leq \frac{1}{1+\lambda-\rho}|x|, \end{cases}$$

Author Proof
uncorrected proof

SPI Journal: 10208 Article No.: 9516 TYPESET DISK LE CP Disp.:2021/4/28 Pages: 46 Layout: Small-Ex

Author Proof

and if $\frac{1}{(1+\lambda-\rho)}|x| < |y| < (1+\lambda)|x|$, *we have*

$$S(x,y) = \begin{cases} \frac{\lambda}{(1+\lambda)(1+\lambda-\rho)-1}\begin{bmatrix}(1+\lambda) & -1 \\ -1 & (1+\lambda-\rho)\end{bmatrix}\begin{bmatrix}x \\ y\end{bmatrix} & \textit{if } \operatorname{sign}(x) = \operatorname{sign}(y); \\ \frac{\lambda}{(1+\lambda)(1+\lambda-\rho)-1}\begin{bmatrix}(1+\lambda) & 1 \\ 1 & (1+\lambda-\rho)\end{bmatrix}\begin{bmatrix}x \\ y\end{bmatrix} & \textit{if } \operatorname{sign}(x) \neq \operatorname{sign}(y). \end{cases}$$

Then, $\operatorname{prox}_{(1/\lambda)f}(x,y) = S(x,y)$.

Proof Let us denote the components of $S(x,y)$ by $(x_+, y_+) = S(x,y)$. By first-order optimality conditions, we have $\operatorname{prox}_{(1/\lambda)f}(x,y) = (x_+, y_+)$ if and only if

$$\lambda(x-(1-(1/\lambda)\rho)x_+, y-y_+) \in$$
$$\begin{cases} \{x_+ + \operatorname{sign}(x_+)|y_+|\} \times \{\operatorname{sign}(y_+)|x_+| + y_+\} & \text{if } x_+ \neq 0 \text{ and } y_+ \neq 0; \\ ([-1,1]y_+) \times \{y_+\} & \text{if } x_+ = 0 \text{ and } y_+ \neq 0; \\ \{x_+\} \times ([-1,1]x_+) & \text{if } x_+ \neq 0 \text{ and } y_+ = 0; \\ \{0\} \times \{0\} & \text{if } x_+ = 0 \text{ and } y_+ = 0. \end{cases}$$

Let us show that (x_+, y_+) indeed satisfies this inclusion.

1. If $(x,y) = 0$, then $x_+ = y_+ = 0$, and the pair satisfies the inclusion.
2. If $|x| \leq \frac{1}{1+\lambda}|y|$ and $y \neq 0$, then $x_+ = 0$, $y_+ = \frac{\lambda}{1+\lambda}y$, and

$$\lambda(x-(1-(1/\lambda)\rho)x_+, y-y_+) = \lambda\left(x, \frac{1}{1+\lambda}y\right) \in ([-1,1]y_+) \times \{y_+\}.$$

   Thus, the pair satisfies the inclusion.
3. If $|y| \leq \frac{1}{1+\lambda-\rho}|x|$ and $x \neq 0$, then $x_+ = \frac{\lambda}{(1+\lambda-\rho)}x$, $y_+ = 0$, and

$$\begin{aligned} &\lambda(x-(1-(1/\lambda)\rho)x_+, y-y_+) \\ &= \lambda\left(x - \frac{\lambda-\rho}{(1+\lambda-\rho)}x, y\right) \in \{x_+\} \times ([-1,1]x_+). \end{aligned}$$

For the remaining two cases, let us assume that $\frac{1}{(1+\lambda-\rho)}|x| < |y| < (1+\lambda)|x|$.

4. If $\operatorname{sign}(x) = \operatorname{sign}(y)$, let $s = \operatorname{sign}(x)$ and note that

$$\begin{aligned} \begin{bmatrix}x_+ \\ y_+\end{bmatrix} &= \frac{\lambda}{(1+\lambda)(1+\lambda-\rho)-1}\begin{bmatrix}(1+\lambda) & -1 \\ -1 & (1+\lambda-\rho)\end{bmatrix}\begin{bmatrix}x \\ y\end{bmatrix} \\ &= \frac{s\lambda}{(1+\lambda)(1+\lambda-\rho)-1}\begin{bmatrix}(1+\lambda)|x| - |y| \\ -|x| + (1+\lambda-\rho)|y|\end{bmatrix} \end{aligned}$$

SPI Journal: 10208 Article No.: 9516 TYPESET DISK LE CP Disp.:2021/4/28 Pages: 46 Layout: Small-Ex

uncorrected proof

Author Proof

From this equation we learn $\operatorname{sign}(x_+) = \operatorname{sign}(y_+) = s$. Inverting the matrix, we also learn

$$\lambda \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} (1+\lambda-\rho) & 1 \\ 1 & (1+\lambda) \end{bmatrix} \begin{bmatrix} x_+ \\ y_+ \end{bmatrix} = \begin{bmatrix} x_+ + \lambda(1-\rho/\lambda)x_+ + y_+ \\ x_+ + y_+ + \lambda y_+ \end{bmatrix}$$
$$= \begin{bmatrix} x_+ + \operatorname{sign}(x_+)|y_+| + \lambda(1-\rho/\lambda)x_+ \\ \operatorname{sign}(y_+)|x_+| + y_+ + \lambda y_+ \end{bmatrix}.$$

Thus, the pair satisfies the inclusion.

5. If $\operatorname{sign}(x) \neq \operatorname{sign}(y)$, let $s = \operatorname{sign}(x)$ and note that

$$\begin{bmatrix} x_+ \\ y_+ \end{bmatrix} = \frac{\lambda}{(1+\lambda)(1+\lambda-\rho)-1} \begin{bmatrix} (1+\lambda) & 1 \\ 1 & (1+\lambda-\rho) \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$
$$= \frac{s\lambda}{(1+\lambda)(1+\lambda-\rho)-1} \begin{bmatrix} (1+\lambda)|x| - |y| \\ |x| - (1+\lambda-\rho)|y| \end{bmatrix}.$$

From this equation we learn $\operatorname{sign}(x_+) \neq \operatorname{sign}(y_+)$. Inverting the matrix we also learn

$$\lambda \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} (1+\lambda-\rho) & -1 \\ -1 & (1+\lambda) \end{bmatrix} \begin{bmatrix} x_+ \\ y_+ \end{bmatrix} = \begin{bmatrix} x_+ + \lambda(1-\rho/\lambda)x_+ - y_+ \\ -x_+ + y_+ + \lambda y_+ \end{bmatrix}$$
$$= \begin{bmatrix} x_+ + \operatorname{sign}(x_+)|y_+| + \lambda(1-\rho/\lambda)x_+ \\ \operatorname{sign}(y_+)|x_+| + y_+ + \lambda y_+ \end{bmatrix}.$$

Thus, the pair satisfies the inclusion.

Therefore, the proof is complete. □

Corollary B.2 (Convergence to Saddles). *Assume the setting of Theorem B.1. Let* $\alpha \in (0, 1]$ *and define the operator* $T := (1-\alpha)I + \alpha S$ *on* $\mathbb{R}^2$*. Then, the cone* $\mathcal{K} = \{(x, y) \colon |x| \le (1+\lambda)^{-1} y\}$ *satisfies* $T\mathcal{K} \subseteq \mathcal{K}$*. Moreover, for any* $(x, y) \in \mathcal{K}$*, it holds that* $T^k(x, y) = ((1-\alpha)^k x, (1-\alpha(1-\lambda(1+\lambda)^{-1}))^k y)$ *linearly converges to the origin as k tends to infinity.*

Proof Since $\mathcal{K}$ is convex, it suffices to show that $S\mathcal{K} \subseteq \mathcal{K}$. This follows from Theorem B.1. □

References

1. F. Al-Khayyal and J. Kyparisis. Finite convergence of algorithms for nonlinear programs and variational inequalities. *J. Optim. Theory Appl.*, 70(2):319–332, 1991.
2. P. Albano and P. Cannarsa. Singularities of semiconcave functions in Banach spaces. In *Stochastic analysis, control, optimization and applications*, Systems Control Found. Appl., pages 171–190. Birkhäuser Boston, Boston, MA, 1999.
3. H. Attouch, J. Bolte, and B.F. Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized gauss–seidel methods. *Mathematical Programming*, 137(1-2):91–129, 2013.

uncorrected proof

SPI Journal: 10208 Article No.: 9516 TYPESET DISK LE CP Disp.:2021/4/28 Pages: 46 Layout: Small-Ex

Author Proof

4. D. Avdiukhin, c. Jin, and G. Yaroslavtsev. Escaping saddle points with inequality constraints via noisy sticky projected gradient descent. *Optimization for Machine Learning Workshop*, 2019.
5. A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, 2(1):183–202, 2009.
6. S. Bhojanapalli, B. Neyshabur, and N. Srebro. Global optimality of local search for low rank matrix recovery. In *Advances in Neural Information Processing Systems*, pages 3873–3881, 2016.
7. J. Bolte, A. Daniilidis, A. Lewis, and M. Shiota. Clarke subgradients of stratifiable functions. *SIAM Journal on Optimization*, 18(2):556–572, 2007.
8. J.F. Bonnans and A. Shapiro. *Perturbation Analysis of Optimization Problems*. Springer, New York, 2000.
9. J.V. Burke. Descent methods for composite nondifferentiable optimization problems. *Math. Programming*, 33(3):260–279, 1985.
10. J.V. Burke. On the identification of active constraints. II. The nonconvex case. *SIAM J. Numer. Anal.*, 27(4):1081–1103, 1990.
11. J.V. Burke and J.J. Moré. On the identification of active constraints. *SIAM J. Numer. Anal.*, 25(5):1197–1211, 1988.
12. P.H. Calamai and J.J. Moré. Projected gradient methods for linearly constrained problems. *Math. Prog.*, 39(1):93–116, 1987.
13. V. Charisopoulos, Y. Chen, D. Davis, M. Díaz, L. Ding, and D. Drusvyatskiy. Low-rank matrix recovery with composite optimization: good conditioning and rapid convergence. *Foundations of Computational Mathematics*, pages 1–89, 2021.
14. F.H. Clarke, Yu. Ledyaev, R.I. Stern, and P.R. Wolenski. *Nonsmooth Analysis and Control Theory*. Texts in Math. 178, Springer, New York, 1998.
15. C. Criscitiello and N. Boumal. Efficiently escaping saddle points on manifolds. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.
16. D. Davis and D. Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019.
17. D. Drusvyatskiy. The proximal point method revisited. *SIAG/OPT Views and News*, 26(2), 2018.
18. D. Drusvyatskiy, A.D. Ioffe, and A.S. Lewis. Generic minimizing behavior in semialgebraic optimization. *SIAM Journal on Optimization*, 26(1):513–534, 2016.
19. D. Drusvyatskiy and A.S. Lewis. Optimality, identifiablity, and sensitivity. *Math. Program.*, 147(1-2, Ser. A):467–498, 2014.
20. D. Drusvyatskiy and A.S. Lewis. Error bounds, quadratic growth, and linear convergence of proximal methods. *Mathematics of Operations Research*, 43(3):919–948, 2018.
21. D. Drusvyatskiy and C. Paquette. Efficiency of minimizing compositions of convex functions and smooth maps. *Mathematical Programming*, 178(1-2):503–558, 2019.
22. S.S. Du, C. Jin, J.D. Lee, M.I. Jordan, A. Singh, and B. Poczos. Gradient descent can take exponential time to escape saddle points. In *Advances in neural information processing systems*, pages 1067–1077, 2017.
23. J.C. Duchi and F. Ruan. Stochastic methods for composite and weakly convex optimization problems. *SIAM Journal on Optimization*, 28(4):3229–3259, 2018.
24. J.C. Dunn. On the convergence of projected gradient processes to singular critical points. *J. Optim. Theory Appl.*, 55(2):203–216, 1987.
25. M.C. Ferris. Finite termination of the proximal point algorithm. *Math. Program.*, 50(3, (Ser. A)):359–366, 1991.
26. S.D. Flåm. On finite convergence and constraint identification of subgradient projection methods. *Math. Program.*, 57:427–437, 1992.
27. R. Ge, F. Huang, C. Jin, and Y. Yuan. Escaping from saddle points—online stochastic gradient for tensor decomposition. In *Conference on Learning Theory*, pages 797–842, 2015.
28. R. Ge, C. Jin, and Y. Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1233–1242. JMLR. org, 2017.
29. R. Ge, J.D. Lee, and T. Ma. Matrix completion has no spurious local minimum. In *Advances in Neural Information Processing Systems*, pages 2973–2981, 2016.
30. N. Hallak and M. Teboulle. Finding second-order stationary points in constrained minimization: A feasible direction approach. *Journal of Optimization Theory and Applications*, 186(2):480–503, 2020.
31. W.L. Hare and A.S. Lewis. Identifying active manifolds. *Algorithmic Oper. Res.*, 2(2):75–82, 2007.

uncorrected proof

SPI Journal: 10208 Article No.: 9516 TYPESET DISK LE CP Disp.:2021/4/28 Pages: 46 Layout: Small-Ex

Author Proof

32. C. Jin, P. Netrapalli, and M. Jordan. What is local optimality in nonconvex-nonconcave minimax optimization? In *International Conference on Machine Learning*, pages 4880–4889. PMLR, 2020.
33. R. Jin, C.and Ge, P. Netrapalli, S.M. Kakade, and M.I. Jordan. How to escape saddle points efficiently. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1724–1732. JMLR. org, 2017.
34. J.D. Lee, I. Panageas, G. Piliouras, M. Simchowitz, M.I. Jordan, and B. Recht. First-order methods almost always avoid strict saddle points. *Math. Program.*, 176(1-2):311–337, 2019.
35. J.D. Lee, M. Simchowitz, M.I. Jordan, and B. Recht. Gradient descent only converges to minimizers. In *Conference on learning theory*, pages 1246–1257, 2016a.
36. J.M. Lee. Smooth manifolds. In *Introduction to Smooth Manifolds*, pages 1–31. Springer, 2013.
37. Sangkyun Lee and Stephen J Wright. Manifold identification in dual averaging for regularized stochastic online learning. *Journal of Machine Learning Research*, 13(Jun):1705–1744, 2012.
38. C. Lemaréchal, F. Oustry, and C. Sagastizábal. The U-lagrangian of a convex function. *Trans. Amer. Math. Soc.*, 352:711–729, 1996.
39. A.S. Lewis. Active sets, nonsmoothness, and sensitivity. *SIAM J. Optim.*, 13(3):702–725 (electronic) (2003), 2002.
40. A.S. Lewis and S.J. Wright. A proximal method for composite minimization. *Math. Program.*, pages 1–46, 2015.
41. A.S. Lewis and S. Zhang. Partial smoothness, tilt stability, and generalized Hessians. *SIAM Journal on Optimization*, 23(1):74–94, 2013.
42. B. Martinet. Régularisation d'inéquations variationnelles par approximations successives. *Rev. Française Informat. Rech. Opérationnelle*, 4(Sér. R-3):154–158, 1970.
43. B. Martinet. Détermination approchée d'un point fixe d'une application pseudo-contractante. Cas de l'application prox. *C. R. Acad. Sci. Paris Sér. A-B*, 274:A163–A165, 1972.
44. A. Mokhtari, A. Ozdaglar, and A. Jadbabaie. Escaping saddle points in constrained optimization. In *Advances in Neural Information Processing Systems*, pages 3629–3639, 2018.
45. B.S. Mordukhovich. *Variational analysis and generalized differentiation. I*, volume 330 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 2006. Basic theory.
46. J.-J. Moreau. Proximité et dualité dans un espace hilbertien. *Bull. Soc. Math. France*, 93:273–299, 1965.
47. Yu. Nesterov. Modified Gauss–Newton scheme with worst case guarantees for global performance. *Optimisation Methods and Software*, 22(3):469–483, 2007.
48. Yu. Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161, 2013.
49. M. Nouiehed, J.D. Lee, and M. Razaviyayn. Convergence to second-order stationarity for constrained non-convex optimization. arXiv preprint 1810.02024, 2018.
50. E.A. Nurminskii. The quasigradient method for the solving of the nonlinear programming problems. *Cybernetics*, 9(1):145–150, 1973.
51. I. Panageas and G. Piliouras. Gradient descent only converges to minimizers: Non-isolated critical points and invariant regions. arXiv preprint 1605.00405, 2016.
52. E. Pauwels. The value function approach to convergence analysis in composite optimization. *Operations Research Letters*, 44(6):790–795, 2016.
53. R.A. Poliquin and R.T. Rockafellar. Prox-regular functions in variational analysis. *Trans. Amer. Math. Soc.*, 348:1805–1838, 1996.
54. R.T. Rockafellar. *Convex analysis*. Princeton Mathematical Series, No. 28. Princeton University Press, Princeton, N.J., 1970.
55. R.T. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM J. Control Optimization*, 14(5):877–898, 1976.
56. R.T. Rockafellar. Favorable classes of Lipschitz-continuous functions in subgradient optimization. In *Progress in nondifferentiable optimization*, volume 8 of *IIASA Collaborative Proc. Ser. CP-82*, pages 125–143. Int. Inst. Appl. Sys. Anal., Laxenburg, 1982.
57. R.T. Rockafellar and R.J-B. Wets. *Variational Analysis*. Grundlehren der mathematischen Wissenschaften, Vol 317, Springer, Berlin, 1998.
58. S. Rolewicz. On paraconvex multifunctions. In *Third Symposium on Operations Research (Univ. Mannheim, Mannheim, 1978), Section I*, volume 31 of *Operations Res. Verfahren*, pages 539–546. Hain, Königstein/Ts., 1979.

uncorrected proof

SPI Journal: 10208 Article No.: 9516 TYPESET DISK LE CP Disp.:2021/4/28 Pages: 46 Layout: Small-Ex

Author Proof

59. A. Shapiro. Second order sensitivity analysis and asymptotic theory of parametrized nonlinear programs. *Mathematical Programming*, 33(3):280–299, 1985.
60. M. Shub. *Global stability of dynamical systems*. Springer Science & Business Media, 2013.
61. J. Sun, Q. Qu, and J. Wright. When are nonconvex problems not scary? arXiv preprint 1510.06096, 2015.
62. J. Sun, Q. Qu, and J. Wright. A geometric analysis of phase retrieval. *Foundations of Computational Mathematics*, 18(5):1131–1198, 2018.
63. Y. Sun, N. Flammarion, and M. Fazel. Escaping from saddle points on riemannian manifolds. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.
64. S.J. Wright. Identifiable surfaces in constrained optimization. *SIAM J. Control Optim.*, 31:1063–1079, 1993.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

uncorrected proof

Journal: 10208
Article: 9516

Author Proof

Author Query Form

Please ensure you fill out your response to the queries raised below and return this form along with your corrections

Dear Author

During the process of typesetting your article, the following queries have arisen. Please check your typeset proof carefully against the queries listed below and mark the necessary changes either directly on the proof/online grid or in the 'Author's response' area provided below

Query	Details required	Author's response
1.	Please confirm if the corresponding author is correctly identified.	

uncorrected proof