SADDLE AVOIDANCE, ASYMPTOTIC NORMALITY, AND EXPONENTIAL ACCELERATION IN NONSMOOTH OPTIMIZATION

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Liwei Jiang

May 2024

SADDLE AVOIDANCE, ASYMPTOTIC NORMALITY, AND EXPONENTIAL ACCELERATION IN NONSMOOTH OPTIMIZATION

Liwei Jiang, Ph.D.
Cornell University 2024

Optimization-based algorithms are the foundation for empirically successful methods in modern fields, such as artificial intelligence and data science. Although classical optimization theory provides guarantees for functions with smoothness or convexity, a significant portion of modern problems do not possess any of these. Despite the worst-case examples where efficient algorithms are unavailable, typical nonsmoothness arises with a “partly smooth” structure, meaning that they are well-behaved relative to a smooth “active manifold.”

This thesis develops and analyzes first-order algorithms based on the aforementioned nonsmooth structure. We first develop two regularity conditions describing how subgradients interact with active manifolds and then show that they hold for a broad and generic class of functions. With these cornerstones, we demonstrate that when randomly perturbed or equipped with stochastic noise, subgradient methods only converge to minimizers of generic, Clarke regular semialgebraic problems. When convergence to a certain minimizer is known, we demonstrate that stochastic (projected) subgradient methods have asymptotic normality, making them asymptotically optimal algorithms in the locally minimax sense of Hájek and Le Cam.

These findings culminate with a new first-order algorithm—`NTDescent`—which exhibits local nearly linear convergence on typical nonsmooth functions with quadratic growth. The convergence rate of `NTDescent` depends only on the function’s intrinsic quantities but not the problem’s underlying dimension.

BIOGRAPHICAL SKETCH

Liwei Jiang was born in 1997. He grew up in Xuzhou, Jiangsu, China. He graduated from Nanjing University in 2019, where he studied in the Department of Mathematics and received an undergraduate degree in Statistics. Liwei then began his Ph.D. in the School of Operations Research & Information Engineering at Cornell University, where his interests switched to continuous optimization. Upon completing his doctoral studies, he will spend a year as a postdoctoral researcher at the Georgia Institute of Technology, followed by a tenure-track Assistant Professor appointment in the School of Industrial Engineering at Purdue University in West Lafayette.

This thesis is dedicated to my family and friends.

ACKNOWLEDGEMENTS

I want to express my gratitude to my Ph.D. advisor, Damek Davis. During our years of collaboration, I have consistently drawn inspiration from his knowledge, energy, creativity, and research taste. He devoted countless hours to discussing and working on problems with me, always patient in answering my questions, even the most seemingly trivial ones. He taught me how to be a great researcher, and I will carry this forward into my future research career. Thank you, Damek.

I want to thank my thesis committee members: Yudong Chen, Adrian Lewis, and Katya Scheinberg. I have taken classes with all of you, and you introduced me to optimization and data science, where I found my research interest. I am very grateful for all the wealth of knowledge I gained from you and all the support I received as I navigated the academic job market.

I wish to express my gratitude to my informal mentor, Dmitriy Drusvyatskiy. His research style and mathematical elegance have influenced me since our first project on saddle point avoidance. In addition, I want to thank all of my research collaborators: Lijun Ding, Zhihui Zhu, Qing Qu, Zhen Qin, and Jinxin Zhou.

I would like to thank lots of folks at Cornell University. I am very grateful to Shane Henderson, who not only shared the same hobby as me but also gave me courage, guidance, and support. I want to thank Soroosh Shafiee for his valuable insights, advice, and feedback during my academic job search. I would like to thank a lot of students at Cornell. To Vasileios Charisopoulos, Mateo Díaz, Lijun Ding, Tao Jiang, Billy Jin, Cathy Meng, Trang Tran, and Miaolan Xie, for years of great seminar discussions and research collaboration.

I also want to thank the outstanding staff members at Cornell ORIE, Henry Lam, Luiz Vieira, Odrick Palmer, Mingming Wang, and Onnolee Wierson. Thank you for your help and for making the department a nice place to grow and thrive.

I am grateful to the Cornell Lindseth Climbing Center, where my passion for climbing ignited. During my Ph.D. studies, climbing has become my primary hobby, enabling me to accomplish feats I never dreamed possible. I especially want to thank Jessica Synder, Kyle Kuang, Shiyu Cai, and Justin Tan, whom I met when I was a beginner in the gym. I would like to express my gratitude to my friends who have frequently joined me for outdoor climbing trips, listed in order of the number of days we spent at crags: Kyle Kuang, Yingxi Li, Shiyu Cai, Nathan Cho, B. Park, Tucker Swenson, Cong Chen, Lucas Wang, Hajong Lee, Matt Huang, Luk Payne, Jason Schibler, James Oh, Mansur Zhussupbekov, Keqin He, Caroline Sun, Melanie Gao, Allan Zhou, Kenny Cai, and Claire Chang. Those experiences have given me some of my most treasured memories. I also want to express my special thanks to my non-climber friends, Atlas Wu, Shipu Zhao, Xinyi Cai, Bob Yu, and Yiwei Li. Ithaca wouldn't be the same without you.

I am filled with deep gratitude for my partner, Yingxi, for loving me, standing by me, and forgiving me. Your companionship has awakened a bravery within me that I never knew existed, empowering me to dream beyond what I thought possible and attempt things I never imagined trying.

Lastly, I want to thank my family for their unconditional love, support, and patience. I am profoundly indebted to you all, as I wouldn't achieve anything without your presence.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Optimization-based algorithms are the foundation for empirically successful methods in modern fields, such as artificial intelligence and data science. Although classical optimization theory provides guarantees for functions with smoothness or convexity, modern problems often do not possess either of these characteristics. For example, industry-backed solvers, such as TensorFlow and PyTorch, now routinely train nonsmooth and nonconvex deep networks for modern machine learning problems using (stochastic) first-order methods. The widespread empirical success of these methods underscores the need for a better understanding of nonsmooth and nonconvex optimization.

While nonsmooth functions such as the Cantor function and Weierstrass function can be pathological, it is rare to see such behavior in practice. For example [1–7], many nonsmooth and nonconvex functions are “partly smooth”, which entails the existence of a smooth “active manifold” containing the critical point of the function along which the function is smooth, and off of which the function grows sharply. Partial smoothness ensures that the nonsmooth objective functions are well-behaved near their critical points and enables us to extend results from classical smooth optimization theory to nonsmooth settings.

Equipped with the “partial smoothness” structure, we study three fundamental aspects of nonsmooth optimization: avoiding saddle points, asymptotically optimal algorithms, and fast local convergence. In the smooth setting, these aspects are already well-understood:

- Randomly initialized gradient descent almost always escapes strict saddle points [8, 9]. Consequently, gradient descent provably converges to local minimizers for typical smooth objective functions satisfying the strict saddle property, meaning each critical point is either a local minimizer or a strict saddle point (e.g., [10–14]).
- The running average of stochastic gradient descent sequence exhibits asymptotical normality [15] with optimal covariance matrix [16], and thus stochastic gradient descent with averaging is an asymptotically optimal algorithm in the locally minimax sense of Hájek and Le Cam [17, 18];
- Gradient descent with constant stepsize converges linearly when initialized near a minimizer with positive definite Hessian [19]. As a result, one only needs $C\log(1/\varepsilon)$ iterations to achieve a function gap of size ε. Here, C depends on the condition number of the objective function but not the problem's underlying dimension.

These three results rely on a crucial fact about smooth functions: they can be well approximated by their first or second-order Taylor expansion. Consequently, linear dynamical systems can approximate and help us understand gradient descent dynamics. However, this type of argument breaks down for nonsmooth problems because, even with convexity, nonsmooth functions can only be approximated by their local linearization from below. Therefore, the lack of Taylor approximation presents a challenge in generalizing the above results to subgradient-based first-order methods.

This thesis addresses this challenge by developing and analyzing new techniques and algorithms for nonsmooth optimization problems. We note that if a function admits an active manifold, its restriction onto this active manifold is smooth. Therefore, a high-level idealized algorithmic idea is first to identify the active manifold and then study the smooth dynamics of iterative methods along the manifold, thereby generalizing classical results in smooth optimization to nonsmooth problems. While appealing, the smooth dynamics are not available in practice because subgradient-based algorithms do not identify the active manifold. Instead, they oscillate around the active manifold indefinitely. Nevertheless, the key insight of this thesis is that one may still deduce favorable properties of nonsmooth optimization algorithms by connecting the behavior of functions on and off the active manifold. We now describe our main contributions.

- In Chapter 3, we introduce four compatibility conditions between two sets, which generalizes classical stratification theory by Whitney [20–22], Kuo [23], and Verdier [24]. We then apply these compatibility conditions to the epigraphs and active manifolds of partly smooth functions and derive regularity conditions that quantify how subgradients interact with active manifolds.

Out of the four regularity conditions, the (*b*) and strong-(*a*) regularity conditions are cornerstones for the new analysis and algorithms in the following chapters. They enable us to link iterations with their projections onto the active manifolds, leading to the following algorithmic consequences:

- In Chapter 4, we generalize saddle point avoidance results for gradient descent in smooth optimization to nonsmooth optimization. We show that the stochastic or randomly perturbed subgradient method almost always escapes the strict saddle point. As a consequence, the iterates only converge to minimizers of weakly convex "typical functions," which are built from concrete structured examples or unstructured linear perturbations;
- In Chapter 5, we extend the classical asymptotic normality result for stochastic smooth optimization by Polyak and Juditsky [15] to stochastic nons-

mooth/constrained optimization. We prove that the “simplest” online first-order method – stochastic (projected) subgradient method has asymptotic normality with the optimal covariance matrix, and hence it is an asymptotically optimal algorithm in the locally minimax sense of Hájek and Le Cam [17, 18];

- In Chapter 6, we present a first-order method for a broad class of nonsmooth functions with quadratic growth. The algorithm is parameter-free and locally converges nearly linearly, meaning that to achieve a function gap of size ε, one needs at most $C\log^3(1/\varepsilon)$ first-order oracle evaluations. Here, C depends on the objective function's intrinsic quantities but not the problem's underlying dimension. Moreover, the algorithm's memory cost and per-iteration complexity have the same order as the standard subgradient method.

Our presentation assumes a certain familiarity with nonsmooth analysis and differential geometry. For convenience, we have compiled most of the necessary notation and background in Chapter 2.

This thesis is based on the following research projects:

- Chapter 3 and 4 are based on joint work with Damek Davis and Dmitriy Drusvyatskiy [25].
- Chapter 5 is based on joint work with Damek Davis and Dmitriy Drusvyatskiy [26].
- Chapter 6 is based on joint work with Damek Davis [27].

CHAPTER 2

PRELIMINARIES

2.1 Notation

Throughout, we let $\mathbf{E}$ and $\mathbf{Y}$ denote Euclidean spaces with inner products denoted by $\langle \cdot, \cdot \rangle$ and the induced norm $\|x\| = \sqrt{\langle x, x \rangle}$. The symbol $\mathbf{B}$ will stand for the closed unit ball in $\mathbf{E}$, while $B_r(x)$ will denote the closed ball of radius r around a point x. The closure of any set $Q \subset \mathbf{E}$ will be denoted by $\operatorname{cl} Q$, while its convex hull will be denoted by $\operatorname{conv} Q$. The relative interior of a convex set Q will be written as $\operatorname{ri} Q$. The lineality space of any convex cone is the linear subspace $\operatorname{lin}(Q) := Q \cap -Q$.

For any function $f\colon \mathbf{E} \to \mathbb{R} \cup \{+\infty\}$, the *domain*, *graph*, and *epigraph* are defined as

$$\begin{aligned} \operatorname{dom} f &:= \{x \in \mathbf{E} : f(x) < \infty\}, \\ \operatorname{gph} f &:= \{(x, f(x)) \in \mathbf{E} \times \mathbb{R} : x \in \operatorname{dom} f\}, \\ \operatorname{epi} f &:= \{(x, r) \in \mathbf{E} \times \mathbb{R} : r \geq f(x)\}, \end{aligned}$$

respectively. We say that f is closed if $\operatorname{epi} f$ is a closed set, or equivalently if f is lower-semicontinuous at every point in its domain. If $\mathcal{M}$ is some subset of $\mathbf{E}$, the symbol $f|_{\mathcal{M}}$ denotes the restriction of f to $\mathcal{M}$ and we set $\operatorname{gph} f|_{\mathcal{M}} := (\operatorname{gph} f) \cap (\mathcal{M} \times \mathbb{R})$. We call a function $h\colon \mathbb{R}^d \to \mathbb{R}$ *sublinear* if its epigraph is a closed convex cone, and in that case we define

$$\operatorname{lin}(h) := \{x \in \mathbb{R}^d\colon h(x) = -h(-x)\}$$

to be its *lineality space.* The graph of h restricted to $\operatorname{lin}(h)$ is precisely the lineality space of $\operatorname{epi} h$.

Given a mapping $F\colon \mathbb{R}^d \to \mathbb{R}^m$ and a point $\bar{x} \in \mathbb{R}^d$, we define

$$\mathrm{lip}_F(\bar{x}) := \limsup_{\substack{x,x' \to \bar{x} \\ x \neq x'}} \frac{\|F(x) - F(x')\|}{\|x - x'\|}.$$

Given a mapping $F\colon \mathbb{R}^d \to \mathbb{R}^{m\times n}$ into the space of $m \times n$ matrices and a point $x \in \mathbb{R}^d$ then we define

$$\mathrm{lip}_F^{\mathrm{op}}(\bar{x}) := \limsup_{\substack{x,x' \to \bar{x} \\ x \neq x'}} \frac{\|F(x) - F(x')\|_{\mathrm{op}}}{\|x - x'\|},$$

where $\|\cdot\|_{\mathrm{op}}$ denotes the operator norm defined on $\mathbb{R}^{m\times n}$.

The *distance* and the *projection* of a point $x \in \mathbf{E}$ onto a set $Q \subset \mathbf{E}$ are

$$d(x, Q) := \inf_{y \in Q} \|y - x\| \qquad \text{and} \qquad P_Q(x) := \operatorname*{argmin}_{y \in Q} \|y - x\|,$$

respectively. Note that the function $\mathrm{dist}(\cdot, \mathcal{X})$ is 1-Lipschitz for any set $\mathcal{X}$. For any set $\mathcal{X} \subseteq \mathbb{R}^d$, all $\bar{x} \in \mathcal{X}$, all $x \in \mathbb{R}^d$, and all $y \in P_{\mathcal{X}}(x)$, we have

$$\|y - \bar{x}\| \leq 2\|x - \bar{x}\|.$$

We denote the diameter of a set $\mathcal{X}$ by

$$\mathrm{diam}(\mathcal{X}) = \sup_{x,y \in \mathcal{X}} \|x - y\|.$$

The indicator function of a set Q, denoted by $\delta_Q\colon \mathbf{E} \to \mathbb{R} \cup \{\infty\}$, is defined to be zero on Q and $+\infty$ off it. The *gap* between any two closed cones $U, V \subset \mathbf{E}$ is defined as

$$\Delta(U, V) := \sup\{\mathrm{dist}(u, V) : u \in U,\ \|u\| = 1\}.$$

2.2 Nonsmooth analysis

Nonsmooth functions will play a central role in this thesis. We follow standard terminology and notation of nonsmooth and variational analysis, following mostly closely the monograph of Rockafellar-Wets [28]. Other influential treatments of the subject include [29–32].

Normal cones and Clarke regularity The symbol "$o(h)$ as $h \to 0$" stands for any univariate function $o(\cdot)$ satisfying $o(h)/h \to 0$ as $h \searrow 0$. The *Fréchet normal cone* to a set $Q \subset \mathbf{E}$ at a point $x \in \mathbf{E}$, denoted $\hat{N}_Q(x)$, consists of all vectors $v \in \mathbf{E}$ satisfying

$$\langle v, y - x\rangle \leq o(\|y - x\|) \quad \text{as} \quad y \to x \text{ in } Q. \tag{2.2.1}$$

The *limiting normal cone* to Q at $x \in Q$, denoted by $N_Q(x)$, consists of all vectors $v \in \mathbf{E}$ for which there exist sequences $x_i \in Q$ and $v_i \in \hat{N}_Q(x_i)$ satisfying $(x_i, v_i) \to (x, v)$. The *Clarke normal cone* is the closed convex hull $N^c_Q(x) = \operatorname{cl}\operatorname{conv} N_Q(x)$. Thus the inclusions

$$\hat{N}_Q(x) \subset N_Q(x) \subset N^c_Q(x), \tag{2.2.2}$$

hold for all $x \in Q$. The set Q is called *Clarke regular* at $\bar{x} \in Q$ if Q is locally closed around $\bar{x}$ and equality $N^c_Q(\bar{x}) = \hat{N}_Q(\bar{x})$ holds. In this case, all inclusions in (2.2.2) hold as equalities.

Prox-regularity. A particularly large class of Clarke regular sets consists of those called prox-regular. Following [33,34], a locally closed set $Q \subset \mathbf{E}$ is called *prox-regular at* $\bar{x} \in Q$ if the projection $P_Q(x)$ is a singleton set for all points x near $\bar{x}$. Equivalently [33, Theorem 1.3], a locally closed set Q is prox-regular at $\bar{x} \in Q$ if and only if there exist constants $\epsilon, \rho > 0$ satisfying

$$\langle v, y - x\rangle \leq \frac{\rho}{2}\|y - x\|^2,$$

for all $y, x \in Q \cap B_\epsilon(\bar{x})$ and all normal vectors $v \in N_Q(x) \cap \epsilon\mathbf{B}$. If Q is prox-regular at $\bar{x}$, then the projection $P_Q(\cdot)$ is automatically locally Lipschitz continuous around $\bar{x}$ [33, Theorem 1.3]. Common examples of prox-regular sets are convex sets and C^2 manifolds, as well as sets cut out by finitely many C^2 inequalities under transversality conditions [35]. Prox-regular sets are closely related to proximally smooth sets [34] and sets with positive reach [36].

Subdifferentials. Generalized gradients of functions can be defined through the normal cones to epigraphs. Namely, consider a function $f \colon \mathbf{E} \to \mathbb{R} \cup \{\infty\}$ and a point $x \in \operatorname{dom} f$. The *Fréchet*, *limiting*, and *Clarke* subdifferentials of f at x are defined, respectively, as

$$\begin{aligned} \hat{\partial} f(x) &:= \{v \in \mathbf{E} : (v, -1) \in \hat{N}_{\operatorname{epi} f}(x, f(x))\}, \\ \partial f(x) &:= \{v \in \mathbf{E} : (v, -1) \in N_{\operatorname{epi} f}(x, f(x))\}, \\ \partial_c f(x) &:= \{v \in \mathbf{E} : (v, -1) \in N^c_{\operatorname{epi} f}(x, f(x))\}. \end{aligned} \tag{2.2.3}$$

Explicitly, the inclusion $v \in \hat{\partial} f(x)$ amounts to requiring the lower-approximation property:

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

Moreover, a vector v lies in $\partial f(x)$ if and only if there exist sequences $x_i \in \mathbf{E}$ and Fréchet subgradients $v_i \in \hat{\partial} f(x_i)$ satisfying $(x_i, f(x_i), v_i) \to (x, f(x), v)$ as $i \to \infty$. If f is locally Lipschitz continuous around x, then equality $\partial_c f(x) = \operatorname{conv} \partial f(x)$ holds. A point $\bar{x}$ satisfying $0 \in \partial f(x)$ is called *critical* for f, while a point satisfying $0 \in \partial_c f(x)$ is called *Clarke critical*. The distinction disappears for subdifferentially regular functions. We say that f is *subdifferentially regular* at $x \in \operatorname{dom} f$ if the epigraph of f is Clarke regular at $(x, f(x))$.

The three subdifferentials defined in (2.2.3) fail to capture the horizontal normals to the epigraph—meaning those of the form $(v, 0)$. Such horizontal normals play an important role in variational analysis, particularly for developing subdifferential calculus rules. Consequently, we define the *limiting* and *Clarke horizon subdifferentials*, respectively, by:

$$\begin{aligned} \partial^\infty f(x) &:= \{v \in \mathbf{E} : (v, 0) \in N_{\operatorname{epi} f}(x, f(x))\}, \\ \partial^\infty_c f(x) &:= \{v \in \mathbf{E} : (v, 0) \in N^c_{\operatorname{epi} f}(x, f(x))\}. \end{aligned} \tag{2.2.4}$$

Weak convexity. A function $f\colon \mathbf{E} \to \mathbb{R} \cup \{\infty\}$ is called *ρ-weakly convex* if the quadratically perturbed function $x \mapsto f(x) + \frac{\rho}{2}\|x\|^2$ is convex. Weakly convex functions are subdifferentially regular. Indeed, the subgradients of a ρ-weakly convex function yield quadratic minorants, meaning

$$f(y) \geq f(x) + \langle v, y - x \rangle - \frac{\rho}{2}\|y - x\|^2$$

all points $x, y \in \operatorname{dom} f$ and all subgradients $v \in \partial f(x)$. The epigraph of any weakly convex function is a prox-regular set at each point. A primary example of weakly convex functions consists of compositions of Lipschitz convex functions with smooth maps [37, 38].

Semialgebraicity. We call a set $\mathcal{X} \subseteq \mathbb{R}^d$ *semialgebraic* if it is the union of finitely many sets defined by finitely many polynomial inequalities. Likewise, we call a function $f\colon \mathbb{R}^d \to \mathbb{R}$ semialgebraic if its graph $\operatorname{gph}(f) = \{(x, f(x))\colon x \in \mathbb{R}^d\}$ is semialgebraic.

2.3 Differential geometry

In this section, we introduce basic definitions and properties of manifolds.

Manifolds. We next set forth some basic notation when dealing with smooth embedded submanifolds of $\mathbf{E}$. Throughout the thesis, all smooth manifolds $\mathcal{M}$ are assumed to be embedded in $\mathbf{E}$ and we consider the tangent and normal spaces to $\mathcal{M}$ as subspaces of $\mathbf{E}$. Thus, a set $\mathcal{M} \subset \mathbf{E}$ is a *C^p manifold* (with $p \geq 1$) if around any point $x \in \mathcal{M}$ there exists an open neighborhood $U \subset \mathbf{E}$ and a C^p-smooth map F from U to some Euclidean space $\mathbf{Y}$ such that the Jacobian $\nabla F(x)$ is surjective and equality $\mathcal{M} \cap U = F^{-1}(0)$ holds. Then the tangent and normal spaces to $\mathcal{M}$ at x are simply $T_{\mathcal{M}}(x) := \operatorname{Null}(\nabla F(x))$ and

$N_{\mathcal{M}}(x) := (T_{\mathcal{M}}(x))^{\perp}$, respectively. Note that for C^p manifolds $\mathcal{M}$ with $p \geq 1$, the projection $P_{\mathcal{M}}$ is C^{p-1}-smooth on a neighborhood of each point x in $\mathcal{M}$, and is C^p smooth on the tangent space $T_{\mathcal{M}}(x)$ [39]. Moreover, the inclusion $\operatorname{range}(\nabla P_{\mathcal{M}}(x)) \subseteq T_{\mathcal{M}}(x)$ holds for all x near $\mathcal{M}$ and the equality $\nabla P_{\mathcal{M}}(x) = P_{T_{\mathcal{M}}(x)}$ holds for all $x \in \mathcal{M}$.

Covariant gradient and Hessian. Let $\mathcal{M} \subset \mathbf{E}$ be a C^p-manifold for some $p \geq 1$. Then a function $f\colon \mathcal{M} \to \mathbb{R}$ is called C^p-smooth around a point $x \in \mathcal{M}$ if there exists a C^p function $\hat{f}\colon U \to \mathbb{R}$ defined on an open neighborhood U of x and that agrees with f on $U \cap \mathcal{M}$. Then the *covariant gradient of f at x* is defined to be the vector $\nabla_{\mathcal{M}} f(x) := P_{T_{\mathcal{M}}(x)}(\nabla \hat{f}(x))$. When f and $\mathcal{M}$ are C^2-smooth, the *covariant Hessian of f at x* is defined to be the unique self-adjoint bilinear form $\nabla^2_{\mathcal{M}} f(x)\colon T_{\mathcal{M}}(x) \times T_{\mathcal{M}}(x) \to \mathbb{R}$ satisfying

$$\langle \nabla^2_{\mathcal{M}} f(x)u, u\rangle = \frac{d^2}{dt^2} f(P_{\mathcal{M}}(x + tu)) \mid_{t=0} \qquad \text{for all } u \in T_{\mathcal{M}}(x).$$

If $\mathcal{M}$ is C^3-smooth, then we can identify $\nabla^2_{\mathcal{M}} f(x)$ with the matrix $P_{T_{\mathcal{M}}(x)} \nabla^2 \hat{f}(x) P_{T_{\mathcal{M}}(x)}$.

2.4 Active manifolds and active strict saddles

Critical points of typical nonsmooth functions lie on a certain manifold that captures the activity of the problem in the sense that critical points of slight linear tilts of the function do not leave the manifold. Such active manifolds have been modeled in a variety of ways, including identifiable surfaces [1], partial smoothness [2], $\mathcal{UV}$-structures [3, 4], $g \circ F$ decomposable functions [5], and minimal identifiable sets [6].

In this thesis, we adopt the following formal model of activity, explicitly used in [6], where the only difference is that we focus on the Clarke subdifferential instead of the limiting one.

Definition 2.4.1 (Active manifold)**.** Consider a function $f\colon \mathbf{E} \to \mathbb{R} \cup \{\infty\}$ and fix a set $\mathcal{M} \subseteq \mathbf{E}$ containing a point $\bar{x}$ satisfying $0 \in \partial_c f(\bar{x})$. Then $\mathcal{M}$ is called an *active C^p-manifold around $\bar{x}$* if there exists a constant $\epsilon > 0$ satisfying the following.

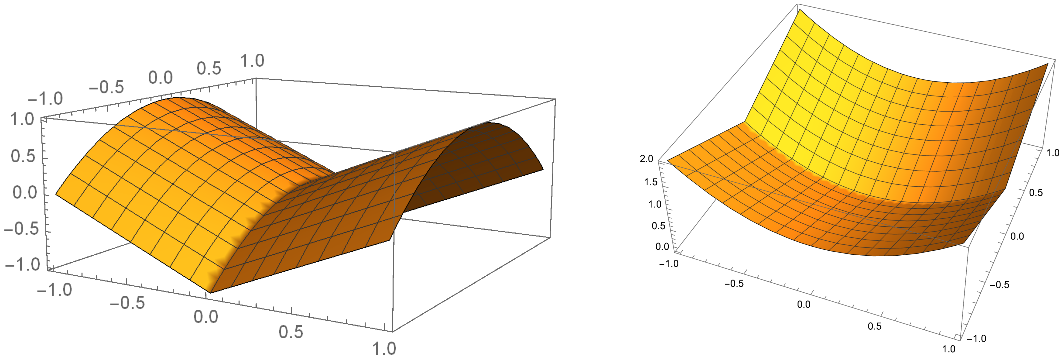
- **(smoothness)** The set $\mathcal{M}$ is a C^p-smooth manifold near $\bar{x}$ and the restriction of f to $\mathcal{M}$ is C^p-smooth near $\bar{x}$.
- **(sharpness)** The lower bound holds:

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

where we set $U = \{x \in B_\epsilon(\bar{x}) : |f(x) - f(\bar{x})| < \epsilon\}$.

The sharpness condition simply means that the subgradients of f must be uniformly bounded away from zero at points off the manifold that are sufficiently close to $\bar{x}$ in distance and in function value. The localization in function value can be omitted for example if f is weakly convex or if f is continuous on its domain; see [6] for details.

Two examples of the active manifold can be found in Figure 2.1, where we have a saddle point for f_1 and a minimizer for f_2. More examples can be found in Chapter 3.



(a) The function $f_1(x, y) = |x| - y^2$

(b) The function $f_2(x, y) = |x| + y^2$

Figure 2.1: The y-axis is an active manifold for both functions.

Intuitively, the active manifold has the distinctive feature that the function grows linearly in normal directions to the manifold; see Figure 4.1a for an illustration. This is summarized by the following theorem from [40, Theorem D.2].

Proposition 2.4.2 (Identification implies sharpness)**.** *Suppose that a closed function* $f\colon \mathbf{E} \to \mathbb{R} \cup \{\infty\}$ *admits an active manifold* $\mathcal{M}$ *at a point* $\bar{x}$ *satisfying* $0 \in \hat{\partial} f(\bar{x})$*. Then there exist constants* $c, \epsilon > 0$ *such that*

$$f(x) - f(P_{\mathcal{M}}(x)) \geq c \cdot \operatorname{dist}(x, \mathcal{M}), \qquad \forall x \in B_{\epsilon}(\bar{x}). \tag{2.4.1}$$

Notice that there is a nontrivial assumption $0 \in \hat{\partial} f(\bar{x})$ at play in Proposition 2.4.2. Indeed, under the weaker inclusion $0 \in \partial_c f(\bar{x})$ the growth condition (2.4.1) may easily fail, as the univariate example $f(x) = -|x|$ shows. It is worthwhile to note that under the assumption $0 \in \hat{\partial} f(\bar{x})$, the active manifold is locally unique around $\bar{x}$ [6, Proposition 8.2].

Active manifolds are useful because they allow to reduce many questions about nonsmooth functions to a smooth setting. In particular, the notion of a strict saddle point of smooth functions naturally extends to a nonsmooth setting. The following definition is taken from [41]. See Figure 4.1 for an illustration.

Definition 2.4.3 (Active strict saddle)**.** Fix an integer $p \geq 2$ and consider a closed function $f\colon \mathbf{E} \to \mathbb{R} \cup \{\infty\}$ and a point $\bar{x}$ satisfying $0 \in \partial_c f(\bar{x})$. We say that $\bar{x}$ is a C^p *strict active saddle point of* f if f admits a C^p active manifold $\mathcal{M}$ at $\bar{x}$ such that the inequality $\langle \nabla^2_{\mathcal{M}} f(\bar{x}) u, u \rangle < 0$ holds for some $u \in T_{\mathcal{M}}(\bar{x})$.

It is often convenient to think about active manifolds of slightly tilted functions. Therefore, we say that $\mathcal{M}$ is an *active* C^p *manifold of* f *at* $\bar{x}$ *for* $v \in \partial_c f(\bar{x})$ if $\mathcal{M}$ is an active C^p manifold for the tilted function $x \mapsto f(x) - \langle v, x \rangle$ at $\bar{x}$. Active manifolds for sets are defined through their indicator functions. Namely a set $\mathcal{M} \subset Q$ is *an active*

C^p manifold of Q at $\bar{x} \in Q$ for $v \in N_Q^c(\bar{x})$ if it is an active C^p manifold of the indicator function δ_Q at $\bar{x}$ for v.

CHAPTER 3

THE FOUR FUNDAMENTAL REGULARITY CONDITIONS

This chapter introduces compatibility conditions between two sets, motivated by the works of Whitney [20–22], Kuo [23], and Verdier [24]. Our discussion builds on the recent survey of Trotman [42]. We illustrate the definitions with examples and prove basic relations between them. It is important to note that these classical works focused on compatibility conditions between smooth manifolds, wherein primal (tangent) and dual (normal) based characterizations are equivalent. In contrast, it will be more expedient to base definitions on normal vectors instead of tangents. The reason is that when applied to epigraphs, such conditions naturally imply some regularity properties for the subgradients, which underpin all algorithmic consequences in this thesis.

3.1 Definitions and basic properties

Throughout this section, we fix two sets $\mathcal{X}$ and $\mathcal{Y}$ and a point $\bar{x} \in \mathcal{Y}$. The reader should keep in mind the most important setting when $\mathcal{Y}$ is a smooth manifold contained in the closure of $\mathcal{X}$. The phenomena we study are naturally one-sided, and therefore we will deal with variational conditions that differ only in the choice of the orientation of the inequalities. With this in mind, in order to simplify notation, we let $\diamond$ stand for any of the symbols in $\{\leq, =, \geq\}$. We begin with the extensions of the two classical conditions of Whitney [21, 22].

Definition 3.1.1 (Whitney conditions)**.** Fix two sets $\mathcal{X}, \mathcal{Y} \subset \mathbf{E}$.

1. We say that $\mathcal{X}$ is *(a)-regular along* $\mathcal{Y}$ if for any sequence $x_i \in \mathcal{X}$ converging to a point $y \in \mathcal{Y}$ and any sequence of normals $v_i \in N_{\mathcal{X}}(x_i)$, every limit point of v_i lies in $N_{\mathcal{Y}}(y)$.

2. We say that $\mathcal{X}$ is $(b_{\leq})$-*regular along* $\mathcal{Y}$ if the estimate

$$\langle v, y - x\rangle \leq o(\|y - x\|) \tag{3.1.1}$$

holds for all $x \in \mathcal{X}$, $y \in \mathcal{Y}$, and all $v \in N_{\mathcal{X}}(x) \cap \mathbf{B}$. Properties $(b_{\geq})$ and $(b_{=})$ are defined analogously with the inequality in (3.1.1) replaced by $\geq$ and $=$, respectively.

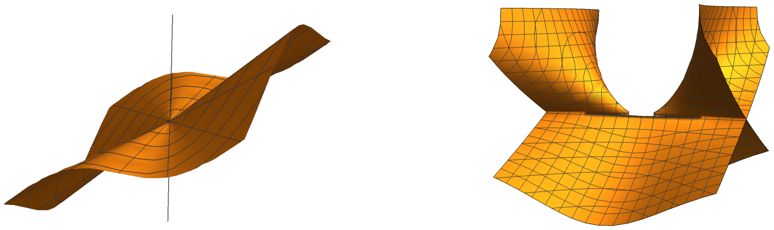
More generally, we say that $\mathcal{X}$ *is regular along* $\mathcal{Y}$ *near a point* $\bar{x} \in \mathcal{Y}$, in any of the above senses, if there exists a neighborhood U of $\bar{x}$ such that $\mathcal{X} \cap U$ is regular along $\mathcal{Y} \cap U$.

Both conditions (a) and $(b_{\diamond})$ are geometrically transparent. Condition (a) simply asserts that "limits of normals to $\mathcal{X}$ are normal to $\mathcal{Y}$"—clearly a desirable property. Figure 3.1a illustrates how condition (a) may fail using the classical example of the Cartan umbrella $\mathcal{X} = \{(x, y, z) : z(x^2 + y^2) = x^3\}$, which is not (a)-regular along the z-axis near the origin. Explicitly, condition $(b_{\leq})$ means that for any sequences $x_i \in \mathcal{X}$ and $y_i \in \mathcal{Y}$ converging to the same point, the condition

$$\limsup_{i\to\infty} \left\langle v_i, \frac{y_i - x_i}{\|y_i - x_i\|}\right\rangle \leq 0,$$

holds, where $v_i \in N_{\mathcal{X}}(x_i)$ are arbitrary unit normal vectors. That is, the angle between the rays spanned by $x_i - y_i$ and any normal vector $v_i \in N_{\mathcal{X}}(x_i)$ becomes obtuse in the limit as $x_i \in \mathcal{X}$ and $y_i \in \mathcal{Y}$ tend to the same point. Conditions $(b_{=})$ and $(b_{\geq})$ have analogous interpretations, with the word obtuse replaced by acute and ninety degrees, respectively. Note that when $\mathcal{X}$ is a smooth manifold, the normal cone $N_{\mathcal{X}}(x)$ is a linear subspace, and therefore all three versions of property $(b_{\diamond})$ are equivalent. On the other hand, a prox-regular set $\mathcal{X}$ is $(b_{\leq})$-regular along any subset $\mathcal{Y}$. Moreover, semismooth sets $\mathcal{X}$ in the sense of [43, 44] are $(b_{=})$-regular along any singleton set $\mathcal{Y} := \{\bar{x}\}$ contained in $\mathcal{X}$.

We will use the following simple lemma frequently. It states that whenever $\mathcal{Y}$ is



(a) $x^3 = z(x^2 + y^2)$ (b) $y^2 = x^2z^2 - z^3$

Figure 3.1: Illustrations of conditions (a) and (b).

contained in $\mathcal{X}$, condition (a) simply amounts to the inclusion of normal cones, $N_{\mathcal{X}}(\bar{x}) \subseteq N_{\mathcal{Y}}(\bar{x})$.

Lemma 3.1.1 (Inclusion of normal cones)**.** *Consider two sets $\mathcal{Y} \subseteq \mathcal{X} \subseteq \mathbf{E}$. Then $\mathcal{X}$ is (a)-regular along $\mathcal{Y}$ at $\bar{x}$ if and only if the inclusion $N_{\mathcal{X}}(y) \subseteq N_{\mathcal{Y}}(y)$ holds for all $y \in \mathcal{Y}$.*

Proof. Suppose first that the inclusion $N_{\mathcal{X}}(y) \subseteq N_{\mathcal{Y}}(y)$ holds for all $y \in \mathcal{Y}$. Consider a sequence $x_i \xrightarrow{\mathcal{X}} y$ and vectors $v_i \in N_{\mathcal{X}}(x_i)$ converging to some vector v. Then we deduce $v \in N_{\mathcal{X}}(y) \subseteq N_{\mathcal{Y}}(y)$, as claimed. Conversely, suppose that $\mathcal{X}$ is (a)-regular along $\mathcal{Y}$. Note that the inclusion $\hat{N}_{\mathcal{X}}(y) \subset \hat{N}_{\mathcal{Y}}(y)$ holds trivially for any $y \in \mathcal{Y}$. For any vector $v \in N_{\mathcal{X}}(y)$, by definition, there exists a sequence $x_i \xrightarrow{\mathcal{X}} y$ and vectors $v_i \in \hat{N}_{\mathcal{X}}(x_i)$ converging to $\bar{v}$. Condition (a) therefore guarantees $\bar{v} \in N_{\mathcal{Y}}(y)$, as claimed. □

The following lemma shows that condition $(b_{\leq})$ implies condition (a) for any sets $\mathcal{X}$ and $\mathcal{Y}$. Moreover, it is classically known that there exist smooth manifolds $\mathcal{X}$ and $\mathcal{Y}$ that satisfy condition $(b_{\leq})$ but not (a); see e.g. [42]. Therefore $(b_{\leq})$ is strictly stronger than (a).

Lemma 3.1.2. *The implication $(b_{\leq}) \implies (a)$ holds for any sets $\mathcal{X}$ and $\mathcal{Y}$. Moreover, the implication $(b_{\geq}) \implies (a)$ holds if $\hat{N}_{\mathcal{Y}}(\bar{x})$ is a linear subspace.*

Proof. Suppose that $\mathcal{X}$ is $(b_{\leq})$-regular along $\mathcal{Y}$. Consider a sequence $x_i \in \mathcal{X}$ converging to a point $y \in \mathcal{Y}$ and vectors $v_i \in N_{\mathcal{X}}(x_i)$ converging to some vector v. It suffices to argue that the inclusion $v \in \hat{N}_{\mathcal{Y}}(y)$ holds. To this end, consider an arbitrary sequence $y_j \in \mathcal{Y} \setminus \{y\}$ converging to y. Passing to a subsequence, we may suppose that the unit vectors $\frac{y_j - y}{\|y_j - y\|}$ converge. For each j we may choose an index i_j satisfying $\|x_{i_j} - y\| \leq \frac{\|y_j - y\|}{j}$. Straightforward algebraic manipulations directly imply

$$\lim_{j\to\infty} \frac{\langle v, y_j - y\rangle}{\|y_j - y\|} \leq \limsup_{j\to\infty} \frac{\langle v_{i_j}, y_j - x_{i_j}\rangle}{\|y_j - x_{i_j}\|} \leq 0,$$

where the last inequality follows from $(b_{\leq})$-regularity. Thus, v lies in $\hat{N}_{\mathcal{Y}}(\bar{x})$, as claimed. The proof of the implication $(b_{\geq}) \Rightarrow (a)$ when $\hat{N}_{\mathcal{Y}}(\bar{x})$ is a linear subspace is analogues.

□

Notice that condition (a) does not specify the rate at which the gap $\Delta(N_{\mathcal{X}}(x_i), N_{\mathcal{Y}}(y))$ tends to zero as $x_i \in \mathcal{X}$ tends to y. A natural strengthening of the condition, introduced by Verdier [24] in the smooth category, requires the gap to be linearly bounded by $\|x_i - y\|$, with a coefficient that is uniform over all $y \in \mathcal{Y}$.[1] Condition (b) can be similarly strengthened. The following definition records the resulting two properties.

Definition 3.1.2 (Strong (a) and strong (b))**.** Consider two sets $\mathcal{X}, \mathcal{Y}$ in $\mathbf{E}$.

1. We say that $\mathcal{X}$ is *strongly (a)-regular along* $\mathcal{Y}$ if there exists a constant $C > 0$ satisfying

$$\Delta(N_{\mathcal{X}}(x), N_{\mathcal{Y}}(y)) \leq C \cdot \|x - y\|, \tag{3.1.2}$$

for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$.

2. We say that $\mathcal{X}$ is *strongly* $(b_{\leq})$*-regular along* $\mathcal{Y}$ if there exists a constant $C > 0$ satisfying

$$\langle v, y - x\rangle \leq C\|x - y\|^2, \tag{3.1.3}$$

[1] What we call strong (a) is often called condition (w), the Verdier condition, or the Kuo-Verdier (kw) condition in the stratification literature.

for all $x \in \mathcal{X}$, $y \in \mathcal{Y}$, and all vectors $v \in N_{\mathcal{X}}(x) \cap \mathbf{B}$. Properties strong $(b_{\geq})$ and strong $(b_{=})$ are defined analogously with the inequality in (3.1.3) replaced by $\geq$ and $=$, respectively.

More generally, we say that $\mathcal{X}$ *is regular along* $\mathcal{Y}$ *near a point* $\bar{x} \in \mathcal{Y}$, in any of the above senses, if there exists a neighborhood U of $\bar{x}$ such that $\mathcal{X} \cap U$ is regular along $\mathcal{Y} \cap U$.

Summarizing, we have defined four fundamental regularity conditions quantifying the compatibility of two sets $\mathcal{X}$ and $\mathcal{Y}$. The most important situation for our purposes is when $\mathcal{Y}$ is a smooth manifold contained in $\mathcal{X}$. The algorithmic importance of these conditions becomes clear when we interpret what they mean for epigraphs of functions. With this in mind, for the rest of the section, we fix a closed function $f\colon \mathbf{E} \to \mathbb{R} \cup \{\infty\}$, a set $\mathcal{M} \subset \operatorname{dom} f$, and a point $\bar{x} \in \mathcal{M}$.

Definition 3.1.3 (Condition $(b_{\diamond})$ for functions)**.** We say that f is (a)*-regular along* $\mathcal{M}$ *near* $\bar{x}$ if the epigraph of f is (a)-regular along $\operatorname{gph} f|_{\mathcal{M}}$ near $(\bar{x}, f(\bar{x}))$. Conditions $(b_{\diamond})$, strong (a), and strong $(b_{\diamond})$ are defined similarly.

Our immediate goal is to interpret regularity of a function f along $\mathcal{M}$ in purely analytic terms. We begin with conditions $(b_{\diamond})$ and strong $(b_{\diamond})$. To this end, we will need the following simple lemma.

Lemma 3.1.3 (Regularity of the domain)**.** *Suppose that f is locally Lipschitz continuous on its domain. If f is $(b_{\diamond})$-regular along $\mathcal{M}$ near $\bar{x}$, then the domain of f is $(b_{\diamond})$-regular along $\mathcal{M}$ near $\bar{x}$. Analogous statements hold for strong $(b_{\diamond})$.*

Proof. Suppose that f is $(b_{\diamond})$-regular along $\mathcal{M}$ near $\bar{x}$. For any $x \in \operatorname{dom} f$ and $y \in \mathcal{M}$ set $X = (x, f(x))$ and $Y = (y, f(y))$. Then for any unit vector $v \in N_{\operatorname{dom} f}(x)$, the vector

$V = (v, 0)$ satisfies the inclusion $V \in N_{\operatorname{epi} f}(X)$ and therefore we may write

$$\left\langle v, \frac{y - x}{\|y - x\|} \right\rangle = \left\langle V, \frac{Y - X}{\|Y - X\|} \right\rangle \cdot \frac{\|Y - X\|}{\|y - x\|}.$$

Using $(b_\diamond)$-regularity of $\operatorname{epi} f$ along $\mathcal{Y}$ and local Lipschitz continuity of f on its domain immediately guarantees that $\operatorname{dom} f$ is $(b_\diamond)$-regular along $\mathcal{M}$ near $\bar{x}$. The analogous statement for strong $(b_\diamond)$ follows from the same argument. □

The following result interprets $(b_\diamond)$-regularity of a function in purely analytic terms.

Theorem 3.1.4 (From geometry to analysis)**.** *Suppose that f is locally Lipschitz continuous on its domain. Then the following are true.*

1. **(condition** (b)**)** *f is $(b_\leq)$-regular along $\mathcal{M}$ near $\bar{x}$ if and only if there exists $\epsilon > 0$ such that the estimates*

$$\frac{f(x) + \langle v, y - x\rangle - f(y)}{\sqrt{1 + \|v\|^2}} \leq o(\|y - x\|), \tag{3.1.4}$$

$$\left\langle \frac{w}{\|w\|}, y - x \right\rangle \leq o(\|y - x\|), \tag{3.1.5}$$

hold for all $x \in \operatorname{dom} f \cap B_\epsilon(\bar{x})$, $y \in \mathcal{M} \cap B_\epsilon(\bar{x})$, $v \in \partial f(x)$, and $w \in \partial^\infty f(x)$.

2. **(strong** (b)**)** *f is strongly $(b_\leq)$-regular along $\mathcal{M}$ near $\bar{x}$ if and only if there exists a constant $\epsilon 0$ such that the estimate holds:*

$$\frac{f(x) + \langle v, y - x\rangle - f(y)}{\sqrt{1 + \|v\|^2}} \leq O(\|y - x\|^2), \tag{3.1.6}$$

$$\left\langle \frac{w}{\|w\|}, y - x \right\rangle \leq O(\|y - x\|^2), \tag{3.1.7}$$

hold for all $x \in \operatorname{dom} f \cap B_\epsilon(\bar{x})$, $y \in \mathcal{M} \cap B_\epsilon(\bar{x})$, $v \in \partial f(x)$, and $w \in \partial^\infty f(x)$.

Analogous equivalences hold for $(b_=)$ and $(b_\geq)$, along with their strong variants, by replacing the inequalities in (3.1.4)-(3.1.7) *by = and ≤, respectively.*

Proof. Throughout the proof, set $\mathcal{X} = \operatorname{epi} f$ and $\mathcal{Y} := \operatorname{gph} f|_{\mathcal{M}}$. We will use capital letters X, Y, and $\bar{X}$ to denote the lifted points $(x, f(x))$, $(y, f(y))$, and $(\bar{x}, f(\bar{x}))$, respectively. We will use the relationship for any point $x \in \operatorname{dom} f$ [28, Theorem 8.9]:

$$N_{\mathcal{X}}(X) \text{ coincides with the union of } \quad \mathbb{R}_{++}(\partial f(x) \times \{-1\}) \quad \text{and} \quad \partial^{\infty} f(x) \times \{0\}. \tag{3.1.8}$$

By definition, f is $(b_{\leq})$-regular along $\mathcal{M}$ near $\bar{x}$ if and only if for any the estimate

$$\langle V, Y - (x, r)\rangle \leq o(\|(x, r) - Y\|) \cdot \|V\|, \tag{3.1.9}$$

holds for all $x \in \operatorname{dom} f$ and $y \in \mathcal{M}$ with (x, r) and Y sufficiently close to $\bar{X}$, and for all $V \in N_{\mathcal{X}}(x, r)$. Let us look at the two cases $r = f(x)$ and $r > f(x)$. In the former case $r = f(x)$, condition (3.1.9) is formally equivalent to the two conditions (3.1.4) and (3.1.5). In the latter case $r > f(x)$, the expression (3.1.9) becomes

$$\langle w, y - x\rangle \leq o(\|(x, r) - Y\|) \cdot \|w\|$$

for all $w \in N_{\operatorname{dom} f}(x)$. Clearly, this is implied by $(b_{\leq})$ regularity of $\operatorname{dom} f$ along $\mathcal{M}$ near $\bar{x}$. The claimed equivalence for $(b_{\leq})$-regularity now follows immediately from Lemma 3.1.3. The rest of the equivalence follow from an analogous argument. □

The conditions in Theorem 3.1.4 are particularly transparent when f is Lipschitz continuous near $\bar{x}$. Then $\partial^{\infty} f(\bar{x})$ consists only of the zero vector and $\partial f(x)$ is nonempty and uniformly bounded near $\bar{x}$. Therefore, conditions $(b_{\leq})$ and strong $(b_{\leq})$, respectively, are equivalent to the two properties

$$f(y) \geq f(x) + \langle v, y - x\rangle + o(\|y - x\|)$$

$$f(y) \geq f(x) + \langle v, y - x\rangle + O(\|y - x\|^2)$$

as x and $y \in \mathcal{M}$ tend to $\bar{x}$ and $v \in \partial f(x)$ is arbitrary. In words, condition $(b_{\leq})$ ensures a restricted lower Taylor approximation property as x and $y \in \mathcal{M}$ tend to $\bar{x}$ and $v \in \partial f(x)$

are arbitrary. Strong (b)-regularity, in turn, replaces the little-o term with the squared norm $O(\|x - y\|^2)$. In particular, this holds automatically if f is weakly convex. When $\mathcal{M} = \{\bar{x}\}$ is a single point, condition ($b_=$) reduces to generalized differentiability in the sense of Norkin [45] and is closely related to the semismoothness property of Mifflin [44].

Condition ($b_\leq$) becomes particularly useful algorithmically when the inclusion $0 \in \hat{\partial} f(\bar{x})$ holds and $\mathcal{M}$ is a C^1 active manifold of f around $\bar{x}$. Indeed, condition ($b_\leq$) along with the sharp growth guarantee of Theorem 2.4.2 then imply that there exists a constant $\mu > 0$ such that the estimate

$$\langle v, x - P_{\mathcal{M}}(x)\rangle \geq \mu \cdot \mathrm{dist}(x, \mathcal{M}), \tag{3.1.10}$$

holds for all $x \in \mathrm{dom}\, f$ near $\bar{x}$ and for all $v \in \partial f(x)$. In words, this means that negative subgradients of f at x always point towards the active manifold. The angle condition (3.1.10) together with strong (a) regularity will form the core of the algorithmic developments. For ease of reference, we record a slight generalization of the angle condition (3.1.10) when f is not necessarily locally Lipschitz around $\bar{x}$ and can even be infinite-valued.

Corollary 3.1.5 (Proximal aiming)**.** *Consider a closed function $f\colon \mathbf{E} \to \mathbb{R} \cup \{\infty\}$ that admits an active C^1-manifold $\mathcal{M}$ at a point $\bar{x}$ satisfying $0 \in \hat{\partial} f(\bar{x})$. Suppose that f is locally Lipschitz continuous on its domain and that f is ($b_\leq$)-regular along $\mathcal{M}$ near $\bar{x}$. Then, there exists a constant $\mu > 0$ such that the estimate*

$$\langle v, x - P_{\mathcal{M}}(x)\rangle \geq \mu \cdot \mathrm{dist}(x, \mathcal{M}) - \sqrt{1 + \|v\|^2} \cdot o(\mathrm{dist}(x, \mathcal{M})), \tag{3.1.11}$$

holds for all $x \in \mathrm{dom}\, f$ near $\bar{x}$ and for all $v \in \partial f(x)$. Moreover, if f is locally Lipschitz around $\bar{x}$, the same statement holds with $\partial f(x)$ replaced by $\partial_c f(x)$ and with the negative term omitted in (3.1.11).[2]

[2] The last claim follows immediately from (3.1.11) by possibly increasing $\mu > 0$ and taking convex combinations of limiting subgradients, all of which are uniformly bounded.

Next, we move on to interpreting conditions (a) and strong (a) in analytic terms. We will focus on the most interesting setting when $\mathcal{M}$ is a smooth manifold and the restriction of f to $\mathcal{M}$ is smooth near $\bar{x}$. In particular, we will make use of the following observation in our arguments: the tangent space to $\mathcal{Y} := \operatorname{gph} f|_{\mathcal{M}}$ at $Y := (y, f(y))$ is:

$$T_{\mathcal{Y}}(Y) = \{(u, \langle \nabla_{\mathcal{M}} f(y), u \rangle) : u \in T_{\mathcal{M}}(y)\}. \tag{3.1.12}$$

Lemma 3.1.4 (Regularity of the domain)**.** *Suppose that f is locally Lipschitz continuous on its domain, $\mathcal{M}$ is a C^1 manifold around $\bar{x}$, and the restriction of f to $\mathcal{M}$ is C^1-smooth near $\bar{x}$. If f is (a)-regular along $\mathcal{M}$ near $\bar{x}$, then the domain of f is (a)-regular along $\mathcal{M}$ near $\bar{x}$. Analogous statement holds for strong (a)-regularity.*

Proof. Throughout the proof, set $\mathcal{Y} = \operatorname{gph} f|_{\mathcal{M}}$. Suppose first that f is (a)-regular along $\mathcal{M}$ near $\bar{x}$. Note the inclusion $N_{\operatorname{dom} f}(y) \times \{0\} \subseteq N_{\operatorname{epi} f}(y, f(y))$ for all y near $\bar{x}$. Using Lemma 3.1.1, we therefore conclude $N_{\operatorname{dom} f}(y) \times \{0\} \subseteq N_{\mathcal{Y}}(y, f(y))$. The desired inclusion $N_{\operatorname{dom} f}(y) \subset N_{\mathcal{M}}(y)$ now follows immediately from (3.1.12).

Finally, suppose that f is strongly (a)-regular along $\mathcal{M}$ near $\bar{x}$. Fix points $x \in \operatorname{dom} f$ and $y \in \mathcal{M}$ near $\bar{x}$ and as before define $X = (x, f(x))$ and $Y = (y, f(y))$. Then condition (a) implies that there exists a constant $C > 0$ such that for any $v \in N_{\operatorname{dom} f}(\bar{x})$ there is a vector $(w_1, w_2) \in N_{\mathcal{Y}}(Y)$ satisfying $\|(v, 0) - (w_1, w_2)\| \le C\|X - Y\|$. It follows easily from the description (3.1.12) that the inclusion $w_1 + w_2 \nabla_{\mathcal{M}} f(y) \in N_{\mathcal{M}}(y)$ holds, and therefore

$$\operatorname{dist}(v, N_{\mathcal{M}}(y)) \le \|v - w_1 - w_2 \nabla_{\mathcal{M}} f(y)\| \le C(1 + \|\nabla_{\mathcal{M}} f(y)\|)\|X - Y\|.$$

Since f is locally Lipschitz continuous on its domain, there exists $C' > 0$ satisfying $\|\nabla_{\mathcal{M}} f(y)\| \le C'$ and $\|X - Y\| \le C'\|x - y\|$ for all $x \in \operatorname{dom} f$ and $y \in \mathcal{M}$ near $\bar{x}$. Thus $\operatorname{dom} f$ is strongly (a)-regular along $\mathcal{M}$ at $\bar{x}$, as claimed. □

The following theorem reinterprets conditions conditions (a) and strong (a) in entirely analytic terms.

Theorem 3.1.6 (From geometry to analysis)**.** *Suppose that f is locally Lipschitz continuous on its domain, $\mathcal{M}$ is a C^1 manifold around $\bar{x}$, and the restriction of f to $\mathcal{M}$ is C^1-smooth near $\bar{x}$. The following claims are true.*

1. **(condition** (a)**)** *f is (a)-regular along $\mathcal{M}$ near $\bar{x}$ if and only if the inclusions hold:*

$$P_{T_{\mathcal{M}}(x)}(\partial f(x)) \subseteq \{\nabla_{\mathcal{M}} f(x)\} \qquad \text{and} \qquad \partial^{\infty} f(x) \subseteq N_{\mathcal{M}}(x). \tag{3.1.13}$$

for all $x \in \mathcal{M}$ near $\bar{x}$.

2. **(strong** (a)**)** *f is strongly (a)-regular along $\mathcal{M}$ near $\bar{x}$ if and only if there exist constants $C, \epsilon > 0$ satisfying:*

$$\|P_{T_{\mathcal{M}(y)}}(v - \nabla_{\mathcal{M}} f(y))\| \leq C\sqrt{1 + \|v\|^2}\|x - y\|, \tag{3.1.14}$$

$$\|P_{T_{\mathcal{M}(y)}}(w)\| \leq C\|w\| \cdot \|x - y\|, \tag{3.1.15}$$

for all $x \in \operatorname{dom} f \cap B_\epsilon(\bar{x})$ and $y \in \mathcal{M} \cap B_\epsilon(\bar{x})$, $v \in \partial f(x)$, and $w \in \partial^\infty f(x)$.

Proof. The proof is similar to that of Theorem 3.1.4. Throughout, set $\mathcal{X} = \operatorname{epi} f$ and $\mathcal{Y} := \operatorname{gph} f|_{\mathcal{M}}$. We will use capital letters X, Y, and $\bar{X}$ to denote the lifted points $(x, f(x))$, $(y, f(y))$, and $(\bar{x}, f(\bar{x}))$, respectively. We also recall the relationship for any point $x \in \operatorname{dom} f$ [28, Theorem 8.9]:

$$N_{\mathcal{X}}(X) \text{ coincides with the union of } \quad \mathbb{R}_{++}(\partial f(x) \times \{-1\}) \quad \text{and} \quad \partial^\infty f(x) \times \{0\}. \tag{3.1.16}$$

Lemma 3.1.1 implies that f is (a)-regular along $\mathcal{M}$ near $\bar{x}$ if and only if the inclusion $N_{\mathcal{X}}(X) \subset N_{\mathcal{Y}}(X)$ holds for all x near $\bar{x}$, or equivalently $\langle N_{\mathcal{X}}(X), V\rangle = \{0\}$ for all $V \in T_{\mathcal{Y}}(X)$. In light of (3.1.12) and (3.1.16), this happens if and only if

$$\langle \partial f(x) - \nabla_{\mathcal{M}} f(x), u\rangle \subseteq \{0\} \qquad \text{and} \qquad \langle \partial^\infty f(x), u\rangle = \{0\} \qquad \forall u \in T_{\mathcal{M}}(x),$$

which is clearly equivalent to (3.1.13).

Next, by definition f is strongly (a)-regular along $\mathcal{M}$ near $\bar{x}$ if and only if there exists a constant C such that

$$\langle U, V\rangle \leq C\|U\| \cdot \|V\| \cdot \|(x, r) - Y\| \tag{3.1.17}$$

for all $(x, r) \in \mathcal{X}$ and $Y \in \mathcal{Y}$ sufficiently close to $\bar{X}$, and for all $U \in N_{\mathcal{X}}((x, r))$ and $V \in T_{\mathcal{Y}}(Y)$. Let us interpret (3.1.17) in two cases, $r = f(x)$ and $r > f(x)$. In the former case $r = f(x)$, in light of (3.1.16) and local Lipschitz continuity of f on its domain, condition (3.1.17) simplifies to

$$\langle v - \nabla_{\mathcal{M}} f(y), u\rangle \leq C' \sqrt{\|v\|^2 + 1} \cdot \|u\| \cdot \|x - y\|, \tag{3.1.18}$$

$$\langle w, u\rangle \leq C'\|w\| \cdot \|u\| \cdot \|x - y\|. \tag{3.1.19}$$

holding for some constant C', for all $x \in \operatorname{dom} f$ and $y \in \mathcal{M}$ sufficiently close to $\bar{x}$, and for all $u \in T_{\mathcal{M}}(y)$, $v \in \partial f(x)$, and $w \in \partial^{\infty} f(x)$. In the case $r > f(x)$, taking into account the equality $N_{\mathcal{X}}(x, r) = N_{\operatorname{dom} f} \times \{0\}$, we see that (3.1.17) reduces to

$$\langle w, u\rangle \leq C'\|w\| \cdot \|u\| \cdot \sqrt{\|x - y\|^2 + (r - f(y))^2}$$

holding for all $w \in N_{\operatorname{dom} f}(x)$. Clearly, this is implied by $\operatorname{dom} f$ being strongly (a)-regular along $\mathcal{M}$ at $\bar{x}$. In particular, taking into account Lemma 3.1.4 we see that this condition holds automatically if f is strongly (a) regular along $\mathcal{M}$ at $\bar{x}$. The claimed equivalence for strong (a) regularity follows immediately. □

Again the conditions in Theorem 3.1.6 become particularly transparent when f is Lipschitz continuous near $\bar{x}$. Then conditions (a) and strong (a), respectively, are equivalent to

$$P_{T_{\mathcal{M}(y)}}(\partial f(y)) = \{\nabla_{\mathcal{M}} f(y)\}$$

$$\|P_{T_{\mathcal{M}(y)}}(\partial f(x) - \nabla_{\mathcal{M}} f(y))\| = O(\|x - y\|)$$

holding as $x \to \bar{x}$ and $y \in \mathcal{M}$ tend to $\bar{x}$. In words, condition (a) is equivalent to the projection $P_{T_{\mathcal{M}}(y)}(\partial f(y))$ reducing to a a single point—the covariant gradient $\nabla_{\mathcal{M}} f(y)$. This type of property is called the projection formula in [46]. Strong (a) provides a "stable improvement" over the projection formula wherein the deviation $\partial f(x) - \nabla_{\mathcal{M}} f(y)$ in tangent directions $T_{\mathcal{M}}(y)$ is linearly bounded by $\|x - y\|$, for points $x \in \mathbf{E}$ and $y \in \mathcal{M}$ near $\bar{x}$.

The rest of the chapter is devoted to exploring the relationship between the four basic regularity conditions, presenting examples, proving calculus rules, and justifying that these conditions hold "generically" along active manifolds. Section 4.2 will use these conditions to analyze subgradient-type algorithms. This chapter is based on the works [25, 47].

3.2 Relation between the four conditions

The goal of this section is to explore the relationship between the four regularity conditions. Recall that Lemma 3.1.2 already established the implication $(b_{\leq}) \Rightarrow (a)$. More generally, the goal of this section is to show in reasonable settings the string of implications:

$$\boxed{(a) \quad \Leftarrow \quad (b_{=}) \quad \Leftarrow \quad \text{strong}\,(a) \quad \Leftarrow \quad \text{strong}\,(b_{=})}. \tag{3.2.1}$$

Before passing to formal statements, we require some preparation. Namely, the task of verifying conditions $(b_{\diamond})$, strong (a), and strong $(b_{\diamond})$ requires considering arbitrary points $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, which are a priori unrelated. We now show that it essentially suffices to set y to be the projection of x onto $\mathcal{Y}$, or more generally a retraction of x onto $\mathcal{Y}$. In this way, we may remove one degree of flexibility for the question of verification. We begin by defining the projected variants of conditions $(b_{\diamond})$, strong (a), and strong

$(b_\diamond)$.

We begin by defining retractions onto a set $\mathcal{Y}$, with the nearest point projection being the primary example. The added flexibility will be useful once we pass to functions.

Definition 3.2.1 (Retractions)**.** A map $\pi\colon \mathbf{E} \to \mathbf{E}$ is a *retraction* onto a set $\mathcal{Y} \subset \mathbf{E}$ near a point $\bar{x} \in \mathcal{Y}$ if

1. the inclusion $\pi(x) \in \mathcal{X}$ holds for all x near $\bar{x}$,
2. there exists a constant $C \geq 0$ such that the inequality $\|x - \pi(x)\| \leq C \cdot \mathrm{dist}(x, \mathcal{Y})$ holds for all x near $\bar{x}$.

If π is C^p-smooth near $\bar{x}$, we call π a *C^p-smooth retraction*.

Next, we define the projected conditions.

Definition 3.2.2 (Projected conditions)**.** Fix two sets $\mathcal{X}, \mathcal{Y} \subset \mathbf{E}$, a point $\bar{x} \in \mathcal{Y}$, and a retraction π onto $\mathcal{Y}$. We say that $\mathcal{X}$ is $(b_\diamond^\pi)$*-regular* along $\mathcal{Y}$ at $\bar{x}$ if it satisfies condition $(b_\diamond)$ in the restricted setting $y_i = \pi(x_i)$. Conditions *strong* (a^π) and *strong* $(b_\diamond^\pi)$ are defined analogously.

The following theorem allows one to reduce the question of verifying regularity conditions to the setting $y \in \pi(x)$.

Theorem 3.2.3. *Fix two sets $\mathcal{X}, \mathcal{Y} \subset \mathbf{E}$, a point $\bar{x} \in \mathcal{Y}$, and a C^1-smooth retraction π onto $\mathcal{Y}$. Suppose moreover that $\mathcal{Y}$ is a C^1-smooth manifold near $\bar{x}$. Then the equivalences hold:*

1. *strong* $(a) \;\Leftrightarrow\;$ *strong* (a^π)

2. (a) *and* $(b^{\pi}_{\diamond}) \;\Leftrightarrow\; (b_{\diamond})$

Moreover, if π *is* C^2*-smooth, then the implication holds:*

$$\textit{strong } (a^{\pi}) \textit{ and strong } (b^{\pi}_{\diamond}) \qquad \Rightarrow \qquad \textit{strong } (b_{\diamond}).$$

Proof. Suppose that $\mathcal{X}$ is strongly (a^{π})-regular regular along $\mathcal{Y}$ near $\bar{x}$. Thus there exists a constant $C_1 > 0$ such that

$$\Delta\left(N_{\mathcal{X}}(x), N_{\mathcal{Y}}(\pi(x))\right) \leq C_1\|x - \pi(x)\|, \tag{3.2.2}$$

for all $x \in \mathcal{X}$ sufficiently close to $\bar{x}$. On the other hand, since π is a retraction onto $\mathcal{Y}$, there exists some constant $C' > 0$ satisfying $\|x - \pi(x)\| \leq C' \cdot \|x - y\|$ for all $\in \mathcal{X}$ and $y \in \mathcal{Y}$ near $\bar{x}$. Moreover, since $\mathcal{Y}$ is a C^1-smooth manifold, there exists a constant $C_2 > 0$ such that

$$\begin{aligned} \Delta\left(N_{\mathcal{Y}}(\pi(x)), N_{\mathcal{Y}}(y)\right) &\leq C_2\|\pi(x) - y\| \\ &\leq C_2\left(\|\pi(x) - x\| + \|x - y\|\right) \\ &\leq (1 + C')C_2\|x - y\|. \end{aligned} \tag{3.2.3}$$

Combining (3.2.2) and (3.2.3), and using the triangle inequality, we conclude $\Delta(N_{\mathcal{X}}(x), N_{\mathcal{Y}}(y)) \leq (C_1C' + (1 + C')C_2)\,\|x - y\|$, for all $x \in \mathcal{X}, y \in \mathcal{Y}$ sufficiently close to $\bar{x}$. Thus $\mathcal{X}$ is strongly (a)-regular along $\mathcal{Y}$ at $\bar{x}$ as claimed.

Next, suppose that $\mathcal{X}$ is both (a) and $(b^{\pi}_{\diamond})$ regular along $\mathcal{Y}$ near $\bar{x}$. Let $x_i \in \mathcal{X}$ and $y_i \in \mathcal{Y}$ be sequences converging to some point y near $\bar{x}$ and let $v_i \in N_{\mathcal{X}}(x_i)$ be arbitrary. Let us write

$$\left\langle v_i, \frac{y_i - x_i}{\|y_i - x_i\|}\right\rangle = \left\langle v_i, \frac{\pi(x_i) - x_i}{\|y_i - x_i\|}\right\rangle + \left\langle v_i, \frac{y_i - \pi(x_i)}{\|y_i - x_i\|}\right\rangle. \tag{3.2.4}$$

We analyze each term on the right side separately. To this end, observe

$$\left\langle v_i, \frac{\pi(x_i) - x_i}{\|y_i - x_i\|}\right\rangle = \left\langle v_i, \frac{\pi(x_i) - x_i}{\|\pi(x_i) - x_i\|}\right\rangle \cdot \frac{\|\pi(x_i) - x_i\|}{\|y_i - x_i\|}.$$

Therefore, the accumulation points of $\left\langle v_i, \frac{\pi(x_i)-x_i}{\|y_i-x_i\|} \right\rangle$ inherit the sign of the accumulation points of $\left\langle v_i, \frac{\pi(x_i)-x_i}{\|\pi(x_i)-x_i\|} \right\rangle$.

Next, moving on since the retraction π is C^1-smooth near $\bar{x}$, we deduce

$$\limsup_{i\to\infty} \left|\left\langle v_i, \frac{y_i - \pi(x_i)}{\|y_i - x_i\|} \right\rangle\right| \leq \limsup_{i\to\infty} \left|\left\langle v_i, \frac{\nabla\pi(x_i)(y_i - x_i)}{\|y_i - x_i\|} \right\rangle\right|. \tag{3.2.5}$$

Passing to a subsequence, we may assume $\frac{y_i-x_i}{\|y_i-x_i\|}$ tends to some vector $w \in \mathbf{E}$ and that v_i converge to some vector v. Observe that since π maps points into $\mathcal{Y}$, the range of $\nabla\pi(y)$ is contained in the tangent space $T_{\mathcal{Y}}(y)$. Noting that condition (a) guarantees $v \in N_{\mathcal{Y}}(y)$, we deduce that the right-side of (3.2.5) is zero. Thus condition $(b_\diamond)$ holds.

Next, suppose that π is C^2-smooth and that $\mathcal{X}$ is both strongly (a^π)-regular and strongly $(b^\pi_\diamond)$-regular along $\mathcal{Y}$ near $\bar{x}$. Note that we already proved that strong (a^π) implies strong (a). We return to the decomposition:

$$\left\langle v_i, \frac{y_i - x_i}{\|y_i - x_i\|^2} \right\rangle = \left\langle v_i, \frac{\pi(x_i) - x_i}{\|y_i - x_i\|^2} \right\rangle + \left\langle v_i, \frac{y_i - \pi(x_i)}{\|y_i - x_i\|^2} \right\rangle. \tag{3.2.6}$$

and analyze each term separately. To this end, we may write

$$\left\langle v_i, \frac{\pi(x_i) - x_i}{\|y_i - x_i\|^2} \right\rangle = \left\langle v_i, \frac{\pi(x_i) - x_i}{\|\pi(x_i) - x_i\|^2} \right\rangle \frac{\|\pi(x_i) - x_i\|^2}{\|y_i - x_i\|^2}.$$

Therefore, the accumulation points of $\left\langle v_i, \frac{\pi(x_i)-x_i}{\|y_i-x_i\|^2} \right\rangle$ inherit the sign of the accumulation points of $\left\langle v_i, \frac{\pi(x_i)-x_i}{\|\pi(x_i)-x_i\|^2} \right\rangle$. Next, since π is C^2 smooth, we compute

$$\limsup_{i\to\infty} \left|\left\langle v_i, \frac{y_i - \pi(x_i)}{\|y_i - x_i\|^2} \right\rangle\right| \leq \limsup_{i\to\infty} \frac{1}{\|y_i - x_i\|} \cdot \left|\left\langle v_i, \frac{\nabla\pi(y_i)(y_i - x_i)}{\|y_i - x_i\|} \right\rangle\right|. \tag{3.2.7}$$

Since $w_i := \frac{\nabla\pi(y_i)(y_i-x_i)}{\|y_i-x_i\|}$ is tangent to $\mathcal{Y}$ at y_i, strong (a) regularity implies that the right side of (3.2.7) is finite. We thus conclude that $\mathcal{X}$ is strongly $(b_\diamond)$ regular along $\mathcal{Y}$ near $\bar{x}$, as claimed. □

With Theorem 3.2.3 at hand, we may now establish the remaining implications in (3.2.1), beginning with strong $(b_\geq)$ implies strong (a).

Proposition 3.2.4 (Strong $(b_{\geq})$ implies strong (a))**.** *Consider a C^3 manifold $\mathcal{Y}$ that is contained in a set $\mathcal{X} \subset \mathbf{E}$. Suppose that $\mathcal{X}$ is prox-regular at a point $\bar{x} \in \mathcal{Y}$. Then the following implication holds:*

$$\text{strong } (b_{\geq}) \quad \Rightarrow \quad \text{strong } (a).$$

Proof. Suppose that $\mathcal{X}$ is strongly $(b_{\geq})$-regular along $\mathcal{Y}$ near $\bar{x}$. In light of Theorem 3.2.3, it suffices to prove that the strong (a^{π}) condition holds for C^2-smooth retraction. We will use the projection $\pi := P_{\mathcal{Y}}$, which is indeed a C^2-smooth retraction onto $\mathcal{Y}$ since $\mathcal{Y}$ is a C^3 manifold. Thus, there exist constants $\epsilon, L > 0$ satisfying

$$||P_{\mathcal{Y}}(y + h) - P_{\mathcal{Y}}(y) - \nabla P_{\mathcal{Y}}(y)h|| \leq L||h||^2, \tag{3.2.8}$$

for all $y \in B_{\epsilon}(\bar{x})$ and $h \in \epsilon\mathbf{B}$. Fix now two points $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ and a unit vector $v \in N_{\mathcal{X}}(x)$. Clearly, we may suppose $v \notin N_{\mathcal{Y}}(y)$, since otherwise the claim is trivially true. Define the normalized vector $w := -\frac{P_{T_{\mathcal{Y}}(y)}(v)}{||P_{T_{\mathcal{Y}}(y)}(v)||}$. Noting the equality $\nabla P_{\mathcal{Y}}(y) = P_{T_{\mathcal{Y}}(y)}$ and appealing to (3.2.8), we deduce the estimate

$$||P_{\mathcal{Y}}(y - \alpha w) - (y - \alpha w)|| \leq L||\alpha w||^2 = L\alpha^2,$$

for all $y \in B_{\epsilon}(\bar{x})$ and $\alpha \in (0, \epsilon)$. Shrinking $\epsilon > 0$, prox-regularity yields the estimate

$$\langle v, P_{\mathcal{Y}}(y - \alpha w) - x\rangle \leq \frac{\rho}{2}||x - P_{\mathcal{Y}}(y - \alpha w)||^2,$$

for some constant $\rho > 0$. Therefore, we conclude

$$\begin{aligned}\alpha||P_{T_{\mathcal{Y}}(y)}v|| = -\alpha\,\langle v, w\rangle = \langle v, x - y\rangle + \langle v, P_{\mathcal{Y}}(y - \alpha w) - x\rangle + \langle v, (y - \alpha w) - P_{\mathcal{Y}}(y - \alpha w)\rangle \\ \leq C||x - y||^2 + \frac{\rho}{2}||x - P_{\mathcal{Y}}(x - \alpha w)||^2 + L\alpha^2,\end{aligned}$$

where the last inequality follows from the strong $(b_{\geq})$ condition. Note that the middle term is small:

$$||P_{\mathcal{Y}}(y-\alpha w)-x||^2 \leq 2||P_{\mathcal{Y}}(y-\alpha w)-(y-\alpha w)||^2+2||y-\alpha w-x||^2 \leq 2L^2\alpha^4+4||y-x||^2+4\alpha^2.$$

Thus, we have

$$\alpha\|P_{T_{\mathcal{Y}}(y)}v\| \leq C\|x-y\|^2 + \rho L^2\alpha^4 + 2\rho\|x-y\|^2 + 2\rho\alpha^2 + L\alpha^2.$$

Dividing both sides by α and setting $\alpha = \|x-y\|$ completes the proof. □

Next we prove the last implication, strong $(a) \Rightarrow (b_=)$, in the definable category. This result thus generalizes the theorems of Kuo [48], Verdier [24], and Ta Le Loi [49]. The proof technique we present is different from those in the earlier works on the subject and will be based on an application of the Kurdyka-Łojasiewicz inequality [46].

Theorem 3.2.5 (Strong (a) implies (b))**.** *Fix two definable sets $\mathcal{X}, \mathcal{Y} \subset \mathbf{E}$ and a point $\bar{x} \in \mathcal{Y}$. Suppose in addition that $\mathcal{Y}$ is a C^2-smooth manifold around $\bar{x}$ and that $\mathcal{X}$ is a locally closed set. Then the following implication holds:*

$$\text{strong } (a) \quad \Rightarrow \quad (b_=).$$

We note that the theorem may easily fail for general C^∞-manifolds $\mathcal{X}$ and $\mathcal{Y}$, without some extra "tameness" assumption such as definability. See the discussion in [49] for details.

Proof. Suppose that $\mathcal{X}$ is strongly (a)-regular along $\mathcal{Y}$ near $\bar{x}$. In light of Theorem 3.2.3, it suffices to show that $\mathcal{X}$ is (b^π)-regular along $\mathcal{Y}$ near $\bar{x}$. To this end, define the function

$$g(x,v) = |\langle v, x - P_{\mathcal{Y}}(x)\rangle| + \delta_{\mathrm{cl}\,\mathcal{X}}(x).$$

Fix a compact neighborhood U of $\{\bar{x}\} \times \mathbb{B}$. Then the KL-inequality [46, Theorem 11] ensures that there exists $\eta > 0$ and a continuous function $\psi\colon [0,\eta) \to \mathbb{R}$ satisfying $\psi(0) = 0$ and $\psi'(0) = 0$ such that

$$g(x,v) \leq \psi(\mathrm{dist}(0, \partial g(x,v)))). \tag{3.2.9}$$

for any $(x, v) \in U$ with $g(x, v) \leq \eta$. It suffices now to show that $\mathrm{dist}(0, \partial g(x, v))$ is linearly upper bounded by $\mathrm{dist}(x, \mathcal{Y})$ for all $x \in \mathcal{X}$ near $\bar{x}$ and all unit vectors $v \in N_{\mathcal{X}}(x)$. To this end, fix any point (x, v). Clearly, we may assume $g(x, v) \neq 0$, since otherwise there is nothing to prove. We compute

$$\partial g(x, v) = \{(I - \nabla P_Y(x))v + N_{\mathrm{cl}\,\mathcal{X}}(x)\} \times \{x - P_{\mathcal{Y}}(x)\}.$$

Therefore as long as $v \in N_{\mathcal{X}}(x)$ we have

$$\mathrm{dist}(0, \partial g(x, v)) \leq \|\nabla P_{\mathcal{Y}}(x)v\| + \mathrm{dist}(x, \mathcal{Y}). \tag{3.2.10}$$

Since $\mathcal{Y}$ is a C^2-manifold near $\bar{x}$, there exists a constant $L > 0$ such that the inequality $\|\nabla P_{\mathcal{Y}}(x)\| \leq L$ holds for all x near $\bar{x}$. Further, let $C > 0$ be the constant from the defining property (3.1.2) of strong (a) regularity. Thus, as long as $x \in \mathcal{X}$ is sufficiently close to $\bar{x}$, there exists a vector $w \in N_{\mathcal{Y}}(P_{\mathcal{Y}}(x))$ satisfying $\|v - w\| \leq C\mathrm{dist}(x, \mathcal{Y})$. Therefore, continuing with (3.2.10) we deduce

$$\mathrm{dist}(0, \partial g(x, v)) \leq \|\nabla P_{\mathcal{Y}}(x)w\| + (1 + CL)\mathrm{dist}(x, \mathcal{Y}).$$

To complete the proof, note that $\nabla P_{\mathcal{Y}}(x)w = 0$ since $\mathrm{range}(\nabla P_{\mathcal{Y}}(x)) \subseteq T_{\mathcal{Y}}(P_{\mathcal{Y}}(x))$. □

3.3 Basic examples

Having a clear understanding of how the four regularity conditions are related, we now present a few interesting examples of sets that are regular along a distinguished submanifold. More interesting examples can be constructed with the help of calculus rule, discussed at the end of the section. We begin with the following simple example showing that any convex cone is regular along its lineality space.

Proposition 3.3.1 (Cones along the lineality space)**.** *Let $\mathcal{X} \subset \mathbf{E}$ be a convex cone and let $\mathcal{Y} = \mathrm{lin}(X)$ denote its lineality space. Then $\mathcal{X}$ is both strongly* (*a*) *and strongly* ($b_=$) *regular along $\mathcal{Y}$.*

Proof. Strong (*a*) regularity follows from the inclusion $N_{\mathcal{X}}(x) \subset N_{\mathcal{Y}}(y)$ holding for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Next, fix any points $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ and a vector $v \in N_{\mathcal{X}}(x)$. Strong ($b_=$) regularity follows from the equality $\langle v, x - y\rangle = 0$, which is straightforward to verify. □

More interesting examples may be constructed as diffeomorphic images of cones around points in the lineality space. Following [5], a set $\mathcal{X} \subset \mathbf{E}$ is said to be *C^p-cone reducible around a point $\bar{x} \in \mathcal{X}$* if there exist a closed convex cone K in some Euclidean space $\mathbf{Y}$, open neighborhoods U of $\bar{x}$ and V of the origin in $\mathbf{Y}$, and a diffeomorphism $\varphi \colon U \to V$ satisfying $\varphi(\bar{x}) = 0$ and $\mathcal{X} \cap U = \varphi^{-1}(K \cap V)$. In this case, it follows from [2, Theorem 4.2] that the set $\mathcal{M} = \varphi^{-1}(\mathrm{lin}(K) \cap V)$ is an active manifold for $\mathcal{X}$ at $\bar{x}$ for any $v \in \mathrm{ri}\, N_{\mathcal{X}}(\bar{x})$. Common examples of sets that are cone reducible around each of their points are polyhedral sets, the cone of positive semidefinite matrices, the Lorentz cone, and any set cut out by smooth nonlinear inequalities with linearly independent gradients. It is straightforward to see that conditions (*a*) and ($b_\diamond$) are preserved under C^1 diffeomorphisms, while strong (*a*) and strong ($b_\diamond$) are preserved under C^2 diffeomorphisms. The following is therefore an immediate consequence of Proposition 3.3.1.

Corollary 3.3.2 (Cone reducible sets are regular along the active manifold)**.** *Suppose that a set $\mathcal{X}$ is C^2 cone reducible to K by $\varphi \colon U \to V$ around $\bar{x}$. Then $\mathcal{X}$ is strongly* (*a*) *and strongly* ($b_=$)*-regular along $\varphi^{-1}(\mathrm{lin}(K) \cap V)$ near $\bar{x}$.*

The next proposition shows that any convex set is strongly (a)-regular along any affine space contained in it.

Proposition 3.3.3 (Affine subsets of convex sets)**.** *Consider a convex set $\mathcal{X} \subset \mathbf{E}$ and a subset $\mathcal{Y} \subset \mathcal{X}$ that is locally affine around a point $\bar{x} \in \mathcal{Y}$. Then $\mathcal{X}$ is strongly* (a)*-regular along $\mathcal{Y}$ near $\bar{x}$.*

Proof. Translating the sets we may suppose $\bar{x} = 0$ and therefore that $\mathcal{Y}$ coincides with a linear subspace near the origin. Fix now points $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ and a unit vector $v \in N_{\mathcal{X}}(x)$. Clearly, we may suppose $v \notin N_{\mathcal{Y}}(y)$, since otherwise the claim is trivially true. Define the normalized vector $w := -\frac{P_{\mathcal{Y}}(v)}{\|P_{\mathcal{Y}}(v)\|}$. The for all $y \in \mathcal{Y}$ near $\bar{x}$ and all small $\alpha > 0$, using the linearity of the projection $P_{\mathcal{Y}}$ we compute

$$\alpha\|P_{T_{\mathcal{Y}}(y)}v\| = \alpha\|P_{\mathcal{Y}}v\| = -\alpha\,\langle v, w\rangle = \langle v, x - y\rangle + \langle v, P_{\mathcal{Y}}(y - \alpha w) - x\rangle \le \|x - y\|,$$

where the last inequality follows from convexity of $\mathcal{X}$. This completes the proof. □

Not surprisingly, the conclusion of Theorem 3.3.3 can easily fail if $\mathcal{X}$ is prox-regular (instead of convex) or if $\mathcal{Y}$ is a smooth manifold (instead of affine). This is the content of the following example.

Example 3.3.1 (Failure of strong (a)-regularity)**.** Define $\mathcal{X}$ to be the epigraph of the function $f(x, y) = \max\{0, y - x^2\}$ and set $\mathcal{Y}$ to be the x-axis $Y = \mathbb{R} \times \{0\} \times \{0\}$. Consider the sequence $y_k = (1/k, 0, 0))$ in $\mathcal{Y}$ and $x_k = (1/k, 1/k^2, 0)$ in $\mathcal{X}$ converging to the origin. Fix the sequence of normal vectors $v_k = (-2/k, 1, -1) \in N_X(x_k)$ and note $N_{\mathcal{Y}}(y_k) = \{0\} \times \mathbb{R} \times \mathbb{R}$. A quick computation shows

$$\Delta\Big(\frac{v_k}{\|v_k\|}, N_{\mathcal{Y}}(y_k)\Big) = \frac{2/k}{\sqrt{2 + 4/k^2}} \ge \frac{2}{k\sqrt{6}} = \frac{2}{\sqrt{6}}\sqrt{\|x_k - y_k\|}.$$

Therefore $\mathcal{X}$ is not strongly (a)-regular along $\mathcal{Y}$ near $\bar{x}$.

Strong (a)-regularity fails in the above example "by a square root factor in the distance to $\mathcal{Y}$." The following theorem shows a surprising fact: the estimate (3.1.2) is guaranteed to hold up to a square root for any prox-regular set along a smooth submanifold.

Since we will not use this result and the proof is very similar to that of Proposition 3.2.4, we have placed the argument in the appendix.

Proposition 3.3.4 (Strong (a) up to square root)**.** *Consider a C^3 manifold $\mathcal{Y}$ that is contained in a set $\mathcal{X} \subset \mathbf{E}$. Suppose that $\mathcal{X}$ is prox-regular around a point $\bar{x} \in \mathcal{Y}$. Then there exists a constant $C > 0$ satisfying*

$$\Delta(N_{\mathcal{X}}(x), N_{\mathcal{Y}}(y)) \leq C \cdot \sqrt{\|x - y\|}, \tag{3.3.1}$$

for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ sufficiently close to $\bar{x}$.

The following example connects $(b_=)$-regularity to inner-semicontinuity of the normal cone map. Recall that a set-valued map $F \colon \mathbf{E} \rightrightarrows \mathbf{Y}$ is an assignment of points $x \in \mathbf{E}$ to subsets $F(x) \subset \mathbf{Y}$. The map F is called *inner-semicontinuous* at $\bar{x} \in \mathbf{E}$ if for any vector $\bar{y} \in F(\bar{x})$ and any sequence $x_i \to \bar{x}$, there exists a sequence $y_i \in F(x_i)$ converging to $\bar{y}$.

Proposition 3.3.5 (Condition (b) and inner semicontinuity)**.** *Consider a set $\mathcal{X}$ and a subset $\mathcal{Y} \subset \mathcal{X}$. Suppose that $\mathcal{X}$ is prox-regular at some point $\bar{x} \in \mathcal{Y}$ and that that the normal cone map $N_{\mathcal{X}}$ is inner-semicontinuous on $\mathcal{Y}$ near $\bar{x}$. Then $\mathcal{X}$ is $(b_=)$-regular along $\mathcal{Y}$ near $\bar{x}$.*

Proof. Consider sequences $x_i \in \mathcal{X}$ and $y_i \in \mathcal{Y}$ converging to a point $y \in \mathcal{Y}$ near $\bar{x}$. Let $v_i \in N_{\mathcal{X}}(x_i)$ be arbitrary unit normal vectors. Passing to a subsequence we may assume that v_i converge to some unit normal vector $\bar{v} \in N_{\mathcal{X}}(y)$. By inner semicontinuity, there exist unit vectors $w_i \in N_{\mathcal{X}}(y_i)$ converging to $\bar{v}$. Define the unit vectors $u_i := \frac{x_i - y_i}{\|x_i - y_i\|}$. Prox-regularity of $\mathcal{X}$ therefore guarantees $\langle v_i, u_i \rangle \geq -\frac{\rho}{2}\|x_i - y_i\|$ and $\langle w_i, u_i \rangle \leq \frac{\rho}{2}\|x_i - y_i\|$. We conclude

$$-\frac{\rho}{2}\|x_i - y_i\| \leq \langle v_i, u_i \rangle = \langle w_i, u_i \rangle + \langle v_i - w_i, u_i \rangle \leq \frac{\rho}{2}\|x_i - y_i\| + \|v_i - w_i\|.$$

Noting that the left and right sides both tend to zero completes the proof. □

In particular, any proximally smooth set is $(b_=)$-regular along any of its partly smooth submanifolds in the sense of Lewis [2].

3.4 Preservation of regularity under preimages by transversal maps

More interesting examples may be constructed through calculus rules. The next theorem shows that the four regularity conditions are preserved by taking preimages of smooth maps under a transversality condition.

Theorem 3.4.1 (Smooth preimages)**.** *Consider a C^1-map $F\colon \mathbf{Y} \to \mathbf{E}$ and an arbitrary point $\bar{x} \in \mathbf{Y}$. Let $\mathcal{X}, \mathcal{Y} \subset \mathbf{E}$ be two locally closed sets with $\mathcal{Y}$ Clarke regular and containing $F(\bar{x})$. Suppose that the transversality condition holds:*

$$N_{\mathcal{Y}}(F(\bar{x})) \cap \operatorname{Null}(\nabla F(\bar{x})^*) = \{0\}. \tag{3.4.1}$$

Then the following are true.

1. *If $\mathcal{X}$ is (a)-regular along $\mathcal{Y}$ at $F(\bar{x})$ then $F^{-1}(\mathcal{X})$ is (a)-regular along $F^{-1}(\mathcal{Y})$ at $\bar{x}$.*
2. *If $\mathcal{X}$ is (a)-regular and $(b_\diamond)$-regular along $\mathcal{Y}$ at $F(\bar{x})$, then $F^{-1}(\mathcal{X})$ is $(b_\diamond)$-regular along $F^{-1}(\mathcal{Y})$ at $\bar{x}$.*

If in addition F is C^2-smooth, then the following are true.

3 *If $\mathcal{X}$ is strongly (a)-regular along $\mathcal{Y}$, then $F^{-1}(\mathcal{X})$ is strongly (a)-regular along $F^{-1}(\mathcal{Y})$ at $\bar{x}$.*

4 *If $\mathcal{X}$ is both (a)-regular and strongly $(b_\diamond)$-regular along $\mathcal{Y}$ at $F(\bar{x})$, then $F^{-1}(\mathcal{X})$ is strongly $(b_\diamond)$-regular along $F^{-1}(\mathcal{Y})$ at $\bar{x}$.*

Proof. Notice that the transversality condition (3.4.1) is stable under perturbation of $\bar{x}$. In particular, it straightforward to see that there exists a constant $\tau > 0$ and a neighborhood U of $\bar{x}$ satisfying

$$\|\nabla F(y)^* v\| \geq \tau \|v\| \qquad \text{for all } y \in F^{-1}(\mathcal{Y}) \cap U,\ v \in N_{\mathcal{Y}}(F(y)).$$

Moreover, shrinking U, we may assume that F is ℓ-Lipschitz continuous on U. We prove the theorem in the order: $(1), (3), (2), (4)$.

Claim 1: Suppose that $\mathcal{X}$ is (a)-regular along $\mathcal{Y}$ is at $F(\bar{x})$. Then, shrinking $\eta, \tau > 0$ and U, we may ensure:

$$\|\nabla F(x)^* v\| \geq \tau \|v\| \qquad \text{for all } x \in F^{-1}(\mathcal{X}) \cap U,\ v \in N_{\mathcal{X}}(F(x)). \tag{3.4.2}$$

Transversality and Clarke regularity of $\mathcal{Y}$ imply [28, Theorem 10.6]

$$N_{F^{-1}(\mathcal{Y})}(y) = \nabla F(y)^* N_{\mathcal{Y}}(F(y)) \qquad \text{and} \qquad N_{F^{-1}(\mathcal{X})}(x) \subset \nabla F(x)^* N_{\mathcal{X}}(F(x)) \tag{3.4.3}$$

for all $y \in F^{-1}(\mathcal{Y})$ and $x \in F^{-1}(\mathcal{X})$ sufficiently close to $\bar{x}$.

Consider now a sequence $x_i \in F^{-1}(\mathcal{X})$ converging to a point $y \in F^{-1}(\mathcal{Y})$ near $\bar{x}$ and a sequence of unit normal vectors $w_i \in N_{F^{-1}(\mathcal{X})}(x_i)$ converging to some vector w. Using (3.4.3), we may write $w_i = \nabla F(x_i)^* v_i$ for some vectors $v_i \in N_{\mathcal{X}}(F(x_i))$. Note that due to (3.4.2), the sequence v_i is bounded. Indeed, the norm of v_i is upper bounded by a constant that is independent of x_i and y_i. Therefore passing to a subsequence we may suppose v_i converges to some vector v. Since $\mathcal{X}$ is (a)-regular along $\mathcal{Y}$ at $F(\bar{x})$, the inclusion $v \in N_{\mathcal{Y}}(F(y))$ holds. Therefore using (3.4.3) we deduce

$w = \lim_{i\to\infty} \nabla F(x_i)^* v_i = \nabla F(y)^* v \in N_{F^{-1}(\mathcal{Y})}(F(y))$. Thus $F^{-1}(\mathcal{X})$ is (a)-regular along $F^{-1}(\mathcal{Y})$ near $\bar{x}$.

Before moving on to the next three claims, note that each of them implies condition (a) and therefore we can be sure that the expressions (3.4.2) and (3.4.3) hold. Therefore for the rest of the proof, we will fix sequences x_i, v_i, and w_i as in the proof of condition (a), and we let $y_i \in F^{-1}(\mathcal{Y})$ be an arbitrary sequence near $\bar{x}$.

Claim 3: Suppose that F is C^2-smooth and that $\mathcal{X}$ is strongly (a)-regular along $\mathcal{Y}$ at $F(\bar{x})$. Let $C > 0$ be the corresponding constant in (3.1.2). Shrinking U we may assume ∇F is L-Lipschitz continuous on U. We successively compute

$$\begin{aligned}
\operatorname{dist}(w_i, N_{F^{-1}(\mathcal{Y})}(y_i)) &= \operatorname{dist}(\nabla F(x_i)^* v_i, N_{F^{-1}(\mathcal{Y})}(y_i)) \\
&\leq \|\nabla F(x_i) - \nabla F(y_i)\|_{\mathrm{op}}\|v_i\| + \operatorname{dist}(\nabla F(y_i)^* v_i, N_{F^{-1}(\mathcal{Y})}(y_i)) \qquad (3.4.4) \\
&= \|\nabla F(x_i) - \nabla F(y_i)\|_{\mathrm{op}}\|v_i\| + \operatorname{dist}(\nabla F(y_i)^* v_i, \nabla F(y_i)^* N_{\mathcal{Y}}(F(y_i))) \\
& \qquad (3.4.5) \\
&\leq \|\nabla F(x_i) - \nabla F(y_i)\|_{\mathrm{op}}\|v_i\| + \|\nabla F(y_i)\|_{\mathrm{op}}\operatorname{dist}(v_i, N_{\mathcal{Y}}(F(y_i))) \qquad (3.4.6) \\
&\leq L\|v_i\|\|x_i - y_i\| + C\ell\|v_i\|\|F(x_i) - F(y_i)\| \qquad (3.4.7) \\
&\leq (L + C\ell^2)\|v_i\|\|x_i - y_i\| \\
&\leq (L + C\ell^2)\tau^{-1}\|\nabla F(x_i)^* v_i\|\|x_i - y_i\| \qquad (3.4.8) \\
&= (L + C\ell^2)\tau^{-1}\|w_i\|\|x_i - y_i\|,
\end{aligned}$$

where (3.4.4) follows from the triangle inequality, (3.4.5) follows from (3.4.3), the estimate (3.4.7) follows from strong (a)-regularity, and (3.4.8) follows from (3.4.2). Thus $F^{-1}(\mathcal{X})$ is strongly (a)-regular along $F^{-1}(\mathcal{Y})$ near $\bar{x}$.

Setting the stage for the remainder of the proof, we compute

$$\langle w_i, y_i - x_i \rangle = \langle v_i, F(y_i) - F(x_i) \rangle - \langle v_i, F(y_i) - F(x_i) - \nabla F(x_i)(y_i - x_i) \rangle. \qquad (3.4.9)$$

Claim 2: Suppose that $\mathcal{X}$ is (a)-regular and $(b_\diamond)$-regular along $\mathcal{Y}$ near $F(\bar{x})$. Dividing (3.4.9) though by $\|x_i - y_i\|$ and taking into account that F is C^1-smooth, we deduce that the limit points of $\langle w_i, \frac{y_i - x_i}{\|y_i - x_i\|} \rangle$ inherit the sign from the limit points of $\langle v_i, \frac{F(y_i) - F(x_i)}{\|F(y_i) - F(x_i)\|} \rangle$. Thus $F^{-1}(\mathcal{X})$ is $(b_\diamond)$-regular along $F^{-1}(\mathcal{Y})$ near $\bar{x}$.

Claim 4: This is completely analogous to the proof of $(b_\diamond)$-regularity, except we divide (3.4.9) though by $\|x_i - y_i\|^2$ and pass to the limit. □

3.5 Preservation of regularity under spectral lifts

In this section, we study the prevalence of the four regularity conditions in eigenvalue problems. We begin with some notation. The symbol $\mathbf{S}^n$ will denote the Euclidean space of symmetric matrices, endowed with the trace inner product $\langle A, B \rangle = \operatorname{tr}(AB)$ and the induced Frobenius norm $\|A\| = \sqrt{\operatorname{tr}(A^2)}$. The symbol $O(n)$ will denote the set of $n \times n$ orthogonal matrices. The eigenvalue map $\lambda \colon \mathbf{S}^n \to \mathbb{R}^n$ assigns to every matrix X its ordered list of eigenvalues

$$\lambda_1(X) \geq \lambda_2(X) \geq \ldots \geq \lambda_n(X).$$

The following class of sets will be the subject of the study.

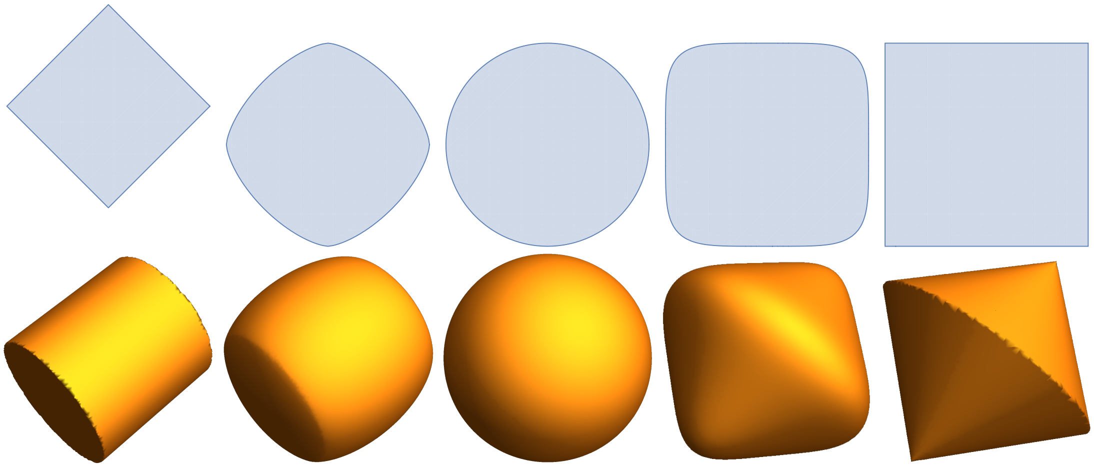
Definition 3.5.1. A set $\mathcal{X} \subset \mathbb{R}^n \to \overline{\mathbb{R}}$ is called *symmetric* if it satisfies

$$\pi\mathcal{X} \subset \mathcal{X} \qquad \text{for all } \pi \in \Pi(n).$$

Definition 3.5.2. A set $Q \subset \mathbf{S}^n$ is called *spectral* if it satisfies

$$UQU^T \subset Q \qquad \text{for all } U \in O(n).$$

Thus a set in $\mathbb{R}^n$ is symmetric if it is invariant under reordering of the coordinates. For example, all ℓ_p-norm balls, the nonnegative orthant, and the unit simplex are symmetric. A set in $\mathbf{S}^n$ is spectral if it is invariant under conjugation of its argument by orthogonal matrices. Spectral sets are precisely those that can be written as $\lambda^{-1}(\mathcal{X})$ for some symmetric set $\mathcal{X} \subset \mathbb{R}^n$. See figure 3.2 for an illustration.



(a) $p = 1$ (b) $p = 1.5$ (c) $p = 2$ (d) $p = 5$ (e) $p = \infty$

Figure 3.2: Unit ℓ_p balls in $\mathbb{R}^2$ (top row) and unit balls of Schatten ℓ_p-norms $\|A\|_p = \|\lambda(A)\|_p$ over $\mathbf{S}^2$ (bottom row).

A prevalent theme in variational analysis is that a variety of geometric properties of a symmetric set $\mathcal{X}$ and those of its induced spectral set $\lambda^{-1}(\mathcal{X})$ are in one-to-one correspondence. Notable examples include convexity [50, 51], smoothness [52, 53], prox-regularity [54], and partial smoothness [55]. In this section, we add to this list the four regularity conditions. The key idea of the arguments is to pass through the projected conditions (Definition 3.2.2) and then invoke Theorem 3.2.3.

We will use the following expressions for the normal cone and the projection map to spectral sets $\lambda^{-1}(\mathcal{X})$:

$$\begin{aligned} P_{\lambda^{-1}(\mathcal{X})}(X) &= \{U\mathrm{Diag}(w)U^T : w \in P_{\mathcal{X}}(\lambda(X)),\ U \in O_X\} \\ N_{\lambda^{-1}(\mathcal{X})}(X) &= \{U\mathrm{Diag}(y)U^T : y \in N_{\mathcal{X}}(\lambda(X)),\ U \in O_X\} \end{aligned}. \tag{3.5.1}$$

where for any matrix X, we define the set of diagonalizing matrices

$$O_X := \{U \in O(n) : X = U\mathrm{Diag}(\lambda(X))U^T\}.$$

The expression for the proximal map was established in [56] while the normal cone formula was proved in [57]. An elementary proof of the subdifferential formula appears in [56].

Theorem 3.5.3 (Spectral preservation of projected regularity)**.** *Let $\bar{X} \in \mathbf{S}^n$ be a symmetric matrix and set $\bar{x} = \lambda(\bar{X})$. Consider two locally closed symmetric sets $\mathcal{X}, \mathcal{Y} \subseteq \mathbb{R}^n$ such that $\mathcal{Y}$ contains $\bar{x}$. Let π and Π be the nearest-point projections onto $\mathcal{Y}$ and $\lambda^{-1}(\mathcal{Y})$, respectively. Then the following are true.*

1. *If $\mathcal{X}$ is (a)-regular along $\mathcal{Y}$ near $\bar{x}$, then $\lambda^{-1}(\mathcal{X})$ is (a)-regular along $\lambda^{-1}(\mathcal{Y})$ near $\bar{X}$.*
2. *If $\mathcal{Y}$ is prox-regular at $\bar{x}$ and $\mathcal{X}$ is strongly (a^π)-regular along $\mathcal{Y}$ near $\bar{x}$, then $\lambda^{-1}(\mathcal{Y})$ is prox-regular at $\bar{X}$ and $\mathcal{X}$ is strongly (a^Π)-regular along $\lambda^{-1}(\mathcal{X})$ near $\bar{X}$. The analogous statement holds for $(b^\pi_\diamond)$ and strong $(b^\pi_\diamond)$ conditions.*

Proof. The result for (a)-regularity holds trivially from (3.5.1). Suppose now that $\mathcal{Y}$ is prox-regular at $\bar{x}$. Then the work [54] guarantees that $\lambda^{-1}(\mathcal{Y})$ is prox-regular at $\bar{X}$. As preparation for the rest of the proof, consider an arbitrary matrix $X \in \lambda^{-1}(\mathcal{X})$ near $\bar{X}$ and a normal vector $V \in N_{\lambda^{-1}(\mathcal{X})}(X)$ with unit Frobenius length. We may then write

$$V = U\mathrm{Diag}(v)U^T,$$

for some unit vector $v \in N_{\mathcal{X}}(\lambda(X))$ and orthogonal matrix $U \in O_X$. Setting $Y := \Pi(X)$ and using (3.5.1), we may write

$$Y = U\mathrm{Diag}(\pi(\lambda(X)))U^T.$$

Notice that because the coordinates of $\lambda(X)$ are decreasing and $\mathcal{Y}$ is symmetric, the coordinates of $\pi(\lambda(X))$ are also decreasing; otherwise, one may reorder $\pi(\lambda(X))$ and find a vector closer to $\lambda(X)$ in $\mathcal{Y}$. Consequently, we have

$$\lambda(Y) = \pi(\lambda(X)) \qquad \text{and} \qquad U \in O_Y. \tag{3.5.2}$$

Suppose now that $\mathcal{X}$ is strongly (a^π)-regular along $\mathcal{Y}$ near $\lambda(\bar{X})$ and let C be the corresponding constant in (3.1.2). Thus there exists $w \in N_{\mathcal{Y}}(\pi(\lambda(X)))$ satisfying

$$\|v - w\| = \mathrm{dist}(v, N_{\mathcal{Y}}(P_{\mathcal{Y}}(\lambda(X))) \leq C\|\lambda(X) - \pi(\lambda(X))\| = C\|X - Y\|,$$

where the last equation follows X and Y being simultaneously diagonalizable. Taking into account (3.5.1) and (3.5.2), we deduce that $W := U\mathrm{Diag}(w)U^T$ lies in $N_{\lambda^{-1}(\mathcal{Y})}(Y)$. Therefore we compute

$$\mathrm{dist}(V, N_{\lambda^{-1}(\mathcal{Y})}(Y)) \leq \|V - W\| = \|v - w\| \leq C\|X - Y\|.$$

Thus $\lambda^{-1}(\mathcal{X})$ is strongly (a^π)-regular along $\lambda^{-1}(\mathcal{Y})$ near $\bar{X}$, as claimed.

Next moving onto conditions $(b^\pi_\diamond)$ and strong $(b^\pi_\diamond)$, we compute

$$\langle V, X - Y\rangle = \langle v, \lambda(X) - \pi(\lambda(X))\rangle.$$

The claimed results now follow immediately by noting $\|\lambda(X) - \pi(\lambda(X))\| = \|X - Y\|$. □

Combining Theorems 3.5.3, 3.2.3, and spectral preservation of smoothness [55] yields the main result of the section.

Proposition 3.5.4 (Spectral Lifts)**.** *Let $\bar{X} \in \mathbf{S}^n$ be a symmetric matrix and set $\bar{x} = \lambda(\bar{X})$. Consider two locally closed symmetric sets $\mathcal{X}, \mathcal{Y} \subseteq \mathbb{R}^n$ such that $\mathcal{Y}$ contains $\bar{x}$. Then the following are true.*

1. *If $\mathcal{Y}$ is a C^2-smooth manifold at $\bar{x}$ and $\mathcal{X}$ is strongly (a)-regular along $\mathcal{Y}$ near $\bar{x}$, then $\lambda^{-1}(\mathcal{Y})$ is a C^2-smooth manifold at $\bar{X}$ and $\mathcal{X}$ is strongly (a)-regular along $\lambda^{-1}(\mathcal{X})$ near $\bar{X}$. The analogous statement holds for $(b_\diamond)$.*
2. *If $\mathcal{Y}$ is a C^3-smooth manifold at $\bar{x}$ and $\mathcal{X}$ is both strongly (a) and strongly (b) regular along $\mathcal{Y}$ near $\bar{x}$, then $\lambda^{-1}(\mathcal{Y})$ is a C^3-smooth manifold at $\bar{X}$ and $\mathcal{X}$ is both strongly (a) and strongly (b) regular along $\lambda^{-1}(\mathcal{X})$ near $\bar{X}$.*

Proof. This follows directly by combining Theorems 3.5.3, 3.2.3, and spectral preservation of smoothness [55, Theorem 2.7] yields the main result of the section. □

All the results in this section extend in a standard way (e.g. [53]) to orthogonally invariant sets of *rectangular matrices* $X \in \mathbb{R}^{m \times n}$. Namely, one only needs to replace (i) eigenvalues $\lambda_i(X)$ with singular values $\sigma_i(X)$, (ii) symmetric sets $\mathcal{X}$ with absolutely symmetric sets (i.e. those invariant under all *signed permutations* of coordinates), and (iii) spectral sets $\mathcal{Q}$ with those that are in variant under the map $X \mapsto UXV^\top$ for any orthogonal matrices $U \in O(m)$ and $V \in O(n)$.

3.6 Regularity of functions along manifolds

The previous sections developed basic examples and calculus rules for the four basic regularity conditions. In this section we interpret these results for functions through their epigraphs. We begin with the following lemma, which follows directly from Propositions 3.3.1, 3.3.3, and 3.3.5.

Lemma 3.6.1 (Basic examples)**.** *Consider a function $f\colon \mathbf{E} \to \mathbb{R} \cup \{\infty\}$, a set $\mathcal{M} \subset$ dom f, and a point $\bar{x} \in \mathcal{M}$. The following statements are true.*

1. *If f is a sublinear function and $\mathcal{M} = \{x : f(x) = -f(-x)\}$ is its lineality space, then f is both strongly (a) and strongly $(b_=)$ regular along $\mathcal{M}$ near $\bar{x}$.*
2. *If f is convex, $\mathcal{M}$ is locally affine near $\bar{x}$, and f restricted to $\mathcal{M}$ is an affine function near $\bar{x}$, then f is strongly (a)-regular along $\mathcal{M}$ near $\bar{x}$.*
3. *If f is weakly convex and locally Lipschitz near $\bar{x}$ and the subdifferential map $x \mapsto \partial f(x)$ is inner-semicontinuous on $\mathcal{M}$ near $\bar{x}$, then f is $(b_=)$-regular along $\mathcal{M}$ near $\bar{x}$.*

The baic calculus rule established in Theorem 3.4.1 yields the following chain rule.

Theorem 3.6.2 (Chain rule)**.** *Consider a C^p-smooth map $c\colon \mathbf{Y} \to \mathbf{E}$ and a closed function $h\colon \mathbf{E} \to \mathbb{R} \cup \{\infty\}$. Fix a set $\mathcal{M} \subset \mathbf{E}$ and a point $\bar{x}$ with $c(\bar{x}) \in \mathcal{M}$. Suppose that $\mathcal{M}$ is a C^1 manifold around $c(\bar{x})$, the restriction $h|_{\mathcal{M}}$ is C^1-smooth near $\bar{x}$, and transversality holds:*

$$N_{\mathcal{M}}(c(\bar{x})) \cap \mathrm{Null}\,(\nabla c(\bar{x})^*) = \{0\}. \tag{3.6.1}$$

Define the composition $f(x) = h(c(x))$ and the set $\mathcal{L} := c^{-1}(\mathcal{M})$. The following are true.

1. *If h is (a)-regular along $\mathcal{M}$ near $c(\bar{x})$ then f is (a)-regular along $\mathcal{L}$ near $\bar{x}$.*
2. *If h is (a)-regular and $(b_\diamond)$-regular along $\mathcal{M}$ near $c(\bar{x})$, then f is $(b_\diamond)$-regular along $\mathcal{L}$ near $\bar{x}$.*

If in addition $\mathcal{M}$ is a C^2 manifold around $c(\bar{x})$ and the restriction $h|_{\mathcal{M}}$ is C^2-smooth near $\bar{x}$, then the following are true.

3 *If h is strongly (a)-regular along $\mathcal{M}$, then f is strongly (a)-regular along $\mathcal{L}$ near $\bar{x}$.*

4 *If h is both (a)-regular and strongly $(b_\diamond)$-regular along $\mathcal{M}$ at $c(\bar{x})$, then f is strongly $(b_\diamond)$-regular along $\mathcal{L}$ near $\bar{x}$.*

Proof. First, the transversality condition (3.6.1) classically guarantees that $\mathcal{L}$ is a smooth manifold around $\bar{x}$ with the same order of smoothness as $\mathcal{M}$. Moreover, for any $x \in \mathcal{L}$, we may write $f(x) = h(c(x)) = (h|_{\mathcal{M}} \circ c)(x)$. Therefore the restriction of f to $\mathcal{L}$ has the same order of smoothness as $h|_{\mathcal{M}}$. Next, observe that we may write $\operatorname{epi} f = \{(x, r) : (c(x), r) \in \operatorname{epi} h\}$. Thus in the notation of Theorem 3.4.1, setting $\mathcal{X} = \operatorname{epi} h$, $\mathcal{Y} = \operatorname{gph} h|_{\mathcal{M}}$, and $F(x, r) = (c(x), r))$, we may write

$$\operatorname{epi} f = F^{-1}(\mathcal{X}) \qquad \text{and} \qquad \operatorname{gph} f|_{\mathcal{L}} = F^{-1}(\mathcal{Y}).$$

A quick computation shows that the transversality condition (3.4.1) follows from (3.6.1). An application of Theorem 3.4.1 completes the proof. □

An interesting class of examples where the chain rule is useful consists of decomposable functions [5], which serve as functional analogues of cone reducible sets. A function $f\colon \mathbf{E} \to \mathbb{R} \cup \{\infty\}$ is called *properly C^p decomposable at $\bar{x}$ as $h \circ c$* if on a neighborhood of $\bar{x}$ it can be written as

$$f(x) = f(\bar{x}) + h(c(x)),$$

for some C^p-smooth mapping $c\colon \mathbf{E} \to \mathbf{Y}$ satisfying $c(\bar{x}) = 0$ and some proper, closed sublinear function $h\colon \mathbf{Y} \to \mathbb{R}$ satisfying the transversality condition:

$$\operatorname{lin}(h) + \operatorname{Range}(\nabla c(\bar{x})) = \mathbf{Y}.$$

It is shown in [5, p 683] that if $f\colon \mathbf{E} \to \mathbb{R} \cup \{\infty\}$ is properly C^p decomposable at $\bar{x}$ as $h \circ c$, then the set $\mathcal{L} = c^{-1}(\text{lin}(h))$ is a C^p-active manifold around $\bar{x}$ for any subgradient $v \in \text{ri}\,\partial f(\bar{x})$. The following is immediate from Lemma 3.6.1 and Theorem 3.6.2.

Corollary 3.6.3 (Decomposable functions are regular)**.** *Suppose that a function f is properly C^1 decomposable as $h \circ c$ around $\bar{x}$ and define $\mathcal{L} = c^{-1}(\text{lin}(h))$. Then f is both* (a) *and* $(b_=)$ *regular along $\mathcal{L}$ near $\bar{x}$. Moreover, if f is properly C^2-decomposable as $h \circ c$ around $\bar{x}$, then f is strongly* (a) *and strongly* $(b_=)$ *regular along $\mathcal{L}$ near $\bar{x}$.*

The chain rule can be used to obtain a variety of other calculus rules, including the sum rule. To see this, note that regularity of functions f_i along sets $\mathcal{M}_i$ directly implies regularity of the separable function $f(y_1, \ldots, y_k) = \sum_{i=1}^k f_i(y_i)$ along the product set $\prod_{i=1}^k \mathcal{M}_i$. Then a general sum rule for $f(x) = \sum_{i=1}^k f_i(x)$ follows from applying the chain rule (Theorem 3.6.2) to the decomposition $f(x) = h(c(x))$ with the linear map $c(x) = (x, \ldots, x)$ and the separable function $h(y_1, \ldots, y_k) = \sum_{i=1}^n f_i(y_i)$. For the sake of brevity, we leave details for the reader.

We end the section with an extension of the material in Section 3.5 to the functional setting. Namely, a function $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is called *symmetric* if equality $f(\pi x) = f(x)$ holds for all $x \in \mathbb{R}^n$ and all $\pi \in \Pi(n)$. A function $f\colon \mathbf{S}^n \to \mathbb{R} \cup \{\infty\}$ is called *spectral* if it satisfies $F(UXU^\top) = F(X)$ for all $X \in \mathbf{S}^n$ and all $U \in O(n)$. It is straightforward to see that any spectral function F decomposes as $F = f \circ \lambda$ for some symmetric function f. Explicitly, we may take f as the diagonal restriction $f(x) = (F \circ \text{Diag})(x)$. The subdifferentials of F and f are related by the expressions [57]:

$$\partial F(X) = \{U \text{Diag}(y) U^T : y \in \partial f(\lambda(X)),\ U \in O_X\} \quad . \tag{3.6.2}$$

where for any matrix X, we define the set of diagonalizing matrices

$$O_X := \{U \in O(n) : X = U \text{Diag}(\lambda(X)) U^T\}.$$

The following theorem shows that the regularity of a symmetric function f is inherited by the spectral function $F = f \circ \lambda$.

Theorem 3.6.4 (Spectral Lifts)**.** *Consider a symmetric function $f\colon \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ and let $\mathcal{M}$ be a symmetric C^2-manifold containing $\bar{x}$. Suppose that f is locally Lipschitz continuous around $\bar{x}$ and the restriction $f|_{\mathcal{M}}$ is C^2-smooth near $\bar{x}$. Fix now a matrix $\bar{X}$ satisfying $\bar{x} := \lambda(\bar{X})$. Then if f is (a)-regular along $\mathcal{M}$ around $\bar{x}$, then $f \circ \lambda$ is (a)-regular along $\lambda^{-1}(\mathcal{M})$ near $\bar{X}$. The analogous statement holds for strong (a)-regularity and $(b_\diamond)$-regularity. If $\mathcal{M}$ is in addition C^3 smooth, then the analogous statement holds for strong $(b_\diamond)$-regularity.*

Proof. First, [55, Theorem 2.7] shows that $\mathcal{M}$ is a C^p manifold (with $p \geq 2$) around $\bar{x}$ if and only if $\lambda^{-1}(\mathcal{M})$ is a C^p manifold around $\bar{X}$. The analogous statement is true for the restriction of f to $\mathcal{M}$ and for the restriction of $f \circ \lambda$ to $\lambda^{-1}(\mathcal{M})$.

The claim about (a)-regularity follows immediately from Lemma 3.1.1. The main idea for verifying the rest of the properties is to instead focus on the analogous conditions with respect to the retraction π onto $\operatorname{gph} F|_{\lambda^{-1}(\mathcal{M})}$ defined by the expression

$$\pi(X, r) = (P_{\lambda^{-1}(\mathcal{M})}(X), F(P_{\lambda^{-1}(\mathcal{M})}(X))).$$

To this end, suppose that f is strongly (a)-regular near $\bar{x}$. We claim that $\operatorname{epi} F$ is strongly (a^π) regular along $\operatorname{gph} f|_{\mathcal{M}}$ near $(\bar{X}, F(\bar{X}))$. To see this, consider a matrix $X \in \mathbf{S}^n$ near $\bar{X}$ and set $Y = P_{\lambda^{-1}(\mathcal{M})}(X)$. Let $Z \in \partial F(X)$ be arbitrary. Exactly the same argument as in the proof of Theorem 3.5.3 shows that there exists a matrix $W \in \partial F(Y)$ satisfying $\|Z - W\|_F \leq C\|X - Y\|_F$, where C is a fixed constant independent of X and Y. It follows immediately that $\operatorname{epi} F$ is strongly (a^π) regular along $\operatorname{gph} f|_{\mathcal{M}}$ near $(\bar{X}, F(\bar{X}))$. An application of Theorem 3.2.3 therefore guarantees that F is strongly (a) regular along $\lambda^{-1}(\mathcal{M})$ near $\bar{X}$. The claims about $(b_\diamond)$ and strong $(b_\diamond)$ properties follow similarly by using the characterization in Theorem 3.1.4 and arguing regularity with respect to the retraction π. We leave the details for the reader. □

3.7 Generic regularity along active manifolds

How can one justify the use of a particular regularity condition? One approach, highlighted in the previous sections, is to verify the conditions for certain basic examples and then show that they are preserved under transverse smooth deformations. Stratification theory adapts another viewpoint, wherein a regularity condition between two manifolds is considered acceptable if reasonable sets (e.g. semialgebraic, subanalytic or definable) can always be partitioned into finitely many smooth manifolds so that the regularity condition holds along any two "adjacent" manifolds. See the survey [42] for an extensive discussion.

To formalize this viewpoint, we begin with a definition of a stratification.

Definition 3.7.1 (Stratification)**.** A *C^p-stratification* ($p \geq 1$) of a set $Q \subset \mathbf{E}$ is a partition of Q into finitely many C^p manifolds, called *strata*, such that any two strata $\mathcal{X}$ and $\mathcal{Y}$ satisfy the implication:

$$\mathcal{Y} \cap \operatorname{cl} \mathcal{X} \neq \emptyset \quad \Longrightarrow \quad \mathcal{Y} \subset \operatorname{cl} \mathcal{X}.$$

A stratum $\mathcal{Y}$ is said to be *adjacent* to a stratum $\mathcal{X}$ if the inclusion $\mathcal{Y} \subset \operatorname{cl} \mathcal{X}$ holds. If the strata are definable in some o-minimal structure, the stratification is called *definable*.

Thus a stratification of Q is simply a partition of Q into smooth manifolds so that the closure of any stratum is a union of strata. Stratifications such that any pair of adjacent strata are strongly (a)-regular are called Verdier stratifications.

Definition 3.7.2. A C^p *Verdier stratification* ($p \geq 1$) of a set $Q \subset \mathbf{E}$ is a C^p stratification of Q such that any stratum $\mathcal{X}$ is strongly (a)-regular along any stratum $\mathcal{Y}$ contained in $\operatorname{cl} \mathcal{X}$.

It is often useful to refine stratifications. To this end, a stratification is *compatible* with a collection of sets $Q_1, \ldots, Q_k$ if for every index i, every stratum $\mathcal{M}$ is either contained in Q_i or is disjoint from it. The following theorem, due to Ta Le Loi [49], shows that definable sets admit a Verdier stratification, which is compatible with any finite collection of definable sets.

Theorem 3.7.3 (Verdier stratification)**.** *For any $p \geq 1$, any definable set $Q \subset \mathbf{E}$ admits a definable C^p Verdier stratification. Moreover, given finitely many definable subsets $Q_1, \ldots, Q_k$, we may ensure that the Verdier stratification of Q is compatible with $Q_1, \ldots, Q_k$.*

The analogous theorem for condition ($b_=$) (and therefore condition (a)) was proved earlier; see the discussion in [58]. The strong ($b_=$) condition does not satisfy such decomposition properties. It can fail even relative to a single point of a definable set in $\mathbb{R}^2$, as Example 3.7.1 shows. Nonetheless, as we have seen in previous sections, it does hold in a number of interesting settings in optimization (e.g. for cone reducible sets along the active manifold).

Example 3.7.1 (Strong (b) is not generic)**.** Define the curve $\gamma(t) = (t, t^{3/2})$ in $\mathbb{R}^2$. Let $\mathcal{X}$ be the graph of γ and let $\mathcal{Y}$ be the origin in $\mathbb{R}^2$. Then a quick computation shows that a unit normal $u(t) \in N_{\mathcal{X}}(\gamma(t))$ is given by $(-\frac{2}{3}\sqrt{t}, 1)/\sqrt{1 + \frac{4}{9}t}$ and therefore

$$\left\langle u(t), \frac{\gamma(t)}{\|\gamma(t)\|^2} \right\rangle = \frac{t^{3/2}}{3(t^2 + t^3)\sqrt{1 + \frac{4}{9}t}} \to \infty \qquad \text{as} \quad t \to 0.$$

Therefore, the strong condition ($b_=$) fails for the pair (X, Y) at the origin.

Applying Theorem 3.7.3, to epigraphs immediately yields the following.

Theorem 3.7.4 (Verdier stratification of a function)**.** *Consider a definable function $f\colon \mathbf{E} \to \mathbb{R} \cup \{\infty\}$ that is continuous on its domain. Then for any $p > 0$, there exists a partition of* $\operatorname{dom} f$ *into finitely many C^p-smooth manifolds such that f is C^p-smooth on each manifold $\mathcal{M}$, and f is strongly (a)-regular and $(b_=)$-regular along any manifold $\mathcal{M}$.*

Proof. We first form a nonvertical stratification $\{\mathcal{M}_i\}$ of $\operatorname{gph} f$, guaranteed to exist by [46]. Choose any integer $p \geq 2$. Restratifying using Theorem 3.7.3 yields a nonvertical C^p-Verdier stratification $\{\mathcal{K}_j\}$ of $\operatorname{gph} f$. Let $\mathcal{X}_j$ denote the image of $\mathcal{K}_j$ under the canonical projection $(x, r) \mapsto x$. As explained in [46], each set $\mathcal{X}_j$ is a C^p-smooth manifolds, the function f restricted to $\mathcal{X}_j$ is C^p-smooth, and equality $\operatorname{gph} f|_{\mathcal{X}_j} = \mathcal{K}_j$ holds.

Consider now an arbitrary stratum $\mathcal{K}_j$. It remains to verify that $\operatorname{epi} f$ is strongly (a)-regular along $\mathcal{K}_j$. This follows immediately from the fact that there are finitely many strata and that the inclusion $N_{\operatorname{epi} f}(X) \subset N_{\mathcal{K}_l}(X)$ holds for any index l and any $X \in \mathcal{K}_l$. □

In this thesis, we will be interested in sets that are regular along a particular manifold—the active one. Theorem 3.7.3 quickly implies that critical points of "generic" definable functions lie on an active manifold along which the objective function is strongly (a)-regular.

Theorem 3.7.5 (Regularity at critical points of generic functions)**.** *Consider a closed definable function $f\colon \mathbf{E} \to \mathbb{R} \cup \{\infty\}$. Then for almost every direction $v \in \mathbf{E}$ in the sense of Lebesgue measure, the perturbed function $f_v := f(x) - \langle v, x\rangle$ has at most finitely many limiting critical points, each lying on a unique C^p-smooth active manifold and along which the function f_v is strongly (a)-regular.*

This theorem is a special case of a more general result that applies to structured problems of the form

$$\min_x \; g(x) + h(x) \tag{3.7.1}$$

for definable functions g and h. Algorithms that utilize this structure, such as the proximal subgradient method, generate a sequence that may convergence to *composite Clarke critical points* $\bar{x}$, meaning those satisfying

$$0 \in \partial_c g(\bar{x}) + \partial_c h(\bar{x}).$$

This condition is typically weaker than $0 \in \partial_c(g+h)(\bar{x})$. Points $\bar{x}$ satisfying the stronger inclusion $0 \in \partial g(\bar{x}) + \partial h(\bar{x})$ will be called *composite limiting critical*.

The following theorem shows that under a reasonably rich class of perturbations, the problem (3.7.1) admits no extraneous composite limiting critical points. Moreover each of the functions involved admits an active manifold along which the function is strongly (a)-regular. The proof is a small modification of [59, Theorem 5.2].

Theorem 3.7.6 (Regularity at critical points of generic functions)**.** *Consider closed definable functions $g\colon \mathbf{E} \to \mathbb{R} \cup \{\infty\}$ and $h\colon \mathbf{E} \to \mathbb{R} \cup \{\infty\}$ and define the parametric family of problems*

$$\min_x \; f_{y,v}(x) = g(x) - \langle v, x\rangle + h(x+y) \tag{3.7.2}$$

Define the tilted function $g_v(x) = g(x) - \langle v, x\rangle$. Then there exists an integer $N > 0$ such that for almost all parameters (v, y) in the sense of Lebesgue measure, the problem (3.7.2) *has at most N composite Clarke critical points. Moreover, for any limiting composite critical point $\bar{x}$, there exists a unique vector*

$$\bar{\lambda} \in \partial h(\bar{x} + y) \qquad \textit{satisfying } -\bar{\lambda} \in \partial g_v(\bar{x}),$$

and the following properties are true.

1. *The inclusions* $\bar{\lambda} \in \hat{\partial} h(\bar{x}+y)$ *and* $-\bar{\lambda} \in \hat{\partial} g_v(\bar{x})$ *hold.*

2. g_v *admits a* C^p *active manifold* $\mathcal{M}$ *at* $\bar{x}$ *for* $-\bar{\lambda}$ *and* h *admits a* C^p *active manifold* $\mathcal{K}$ *at* $\bar{x}+y$ *for* $\bar{\lambda}$*, and the two manifolds intersect transversally:*

$$N_{\mathcal{K}}(\bar{x}) \cap N_{\mathcal{M}}(\bar{x}) = \{0\}.$$

3. $\bar{x}$ *is either a local minimizer of* $f_{y,v}$ *or a* C^p *strict active saddle point of* $f_{y,v}$*.*

4. g_v *is strongly* (a)*-regular along* $\mathcal{M}$ *at* $\bar{x}$ *and* h *is strongly* (a)*-regular along* $\mathcal{K}$ *at* $\bar{x}+y$*.*

Proof. All the claims, except for 3 and 4, are proved in [59]; note, that in that work, active manifolds are defined using the limiting subdifferential, but exactly the same arguments apply under the more restrictive Definition 2.4.1. Claim 3 is proved in [60, Theorem 5.2][3]; it is a direct consequence of the classical Sard's theorem and existence of stratifications. Claim 4 follows from a small modification to the proof of [59]. Namely, the first-bullet point in the proof may be replaced by "g is C^p-smooth and strongly (a) regular on $X_i^j(\widehat{U_i})$ and h is C^p-smooth and strongly (a)-regular on $F_i^j(\widehat{U_i})$". □

[3] weak convexity is invoked in the theorem statement but is not necessary for the result.

CHAPTER 4

AVOIDING SADDLE POINTS IN SUBGRADIENT-BASED ALGORITHMS

4.1 Introduction

The subgradient method is the workhorse procedure for finding minimizers of Lipschitz continuous functions f on $\mathbb{R}^n$. One common variant, and the one we focus on here, proceeds using the update

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k) + \alpha_k \nu_k, \tag{4.1.1}$$

for some sequence $\alpha_k > 0$ and a mean zero noise vector ν_k chosen by the user. As long as ν_k is absolutely continuous with respect to the Lebesgue measure, the algorithm will only encounter points at which f is differentiable, and therefore the recursion (4.1.1) is well defined. The typical choice of α_k, and one that is well-grounded in theory, is proportional to $k^{-\gamma}$ for $\gamma \in (1/2, 1)$. The subgradient method is core to a wide array of tasks in computational mathematics and applied sciences, such as in statistics, machine learning, control, and signal processing. Despite its ubiquity and the striking simplicity of the evolution equation (4.1.1), the following question remains open.

Is there a broad class of nonsmooth and nonconvex functions for which the subgradient dynamics (4.1.1) are sure to converge only to local minimizers?

In order to better situate the question, let us look at the analogous question for smooth functions, where the answer is entirely classical. Indeed, the seminal work of Pemantle [61] shows that the subgradient method applied to a Morse function either diverges or converges to a local minimizer. Conceptually, the nondegeneracy of the Hessian stipulated by the Morse assumption ensures that around every extraneous critical point, the function admits a direction of negative curvature. Such directions ensure that the stochastic process (4.1.1) locally escapes any neighborhood of the extraneous critical point. Aside from being generic, the Morse assumption or rather the slightly weaker strict saddle property is known to hold for a wealth of concrete statistical estimation and learning problems, as shown for example in [10–14]. Going beyond smooth functions requires new tools. In particular, a positive answer is impossible for general Lipschitz functions, since generic (in Baire sense) Lipschitz functions may have highly oscillatory derivatives [62, 63]. Therefore one must isolate some well-behaved function class to make progress. In this chapter, we focus on Lipschitz functions that are semialgebraic, or more generally definable in an o-minimal structure [58]. The class of definable functions is virtually exhaustive in contemporary applications of optimization and has been the subject of intensive research over the past decade. The following is an informal statement of one of our main results.

Theorem 4.1.1 (Informal)**.** *Let f be a function that is Lipschitz continuous, subdifferentially regular, and is definable in some o-minimal structure. Then for a full-measure set of vectors $v \in \mathbb{R}^n$, the subgradient method applied to the perturbed function $f_v(x) = f(x) - \langle v, x\rangle$ either diverges or converges to a local minimizer of f_v.*

Subdifferential regularity is a common assumption in nonsmooth analysis [28, 31] and is in particular valid for weakly convex functions. Weakly convex functions are those for which the assignment $x \mapsto f(x) + \frac{\rho}{2}\|x\|^2$ is convex for some $\rho \in \mathbb{R}$; equivalently, these are exactly the functions whose epigraph has positive reach in the sense of Federer [36]. This function class is broad and includes convex functions, smooth functions with Lipschitz continuous gradient, and any function of the form $f(x) = h(c(x)) + r(x)$, where h is a Lipschitz convex function, r is a convex function taking values in $\mathbb{R} \cup \{\infty\}$, and $c(\cdot)$ is a smooth map with Lipschitz Jacobian. Classical literature highlights the importance of such composite functions in optimization [5, 63–66], while recent advances in statistical learning and signal processing have further reinvigorated the problem class. For example, nonlinear least squares, phase retrieval [67–69], robust principal component analysis [70, 71], and adversarial learning [72, 73] naturally lead to composite/weakly convex problems. We refer the reader to the recent expository articles [37,74] for more details on this problem class and its numerous applications.

Though our arguments make heavy use of subdifferential regularity, we conjecture that the conclusion of Theorem 4.1.1 is valid without this assumption. We note in passing that in the smooth setting, the noiseless gradient method ($v_k \equiv 0$) applied to a Morse function is also known to converge only to local minimizers, as long as it is initialized outside of a certain Lebesgue null set [75, 76]. It is unclear how to extend this class of results to the nonsmooth setting, without explicitly incorporating noise injection v_t as we do here.

4.1.1 Main ingredients of the proof.

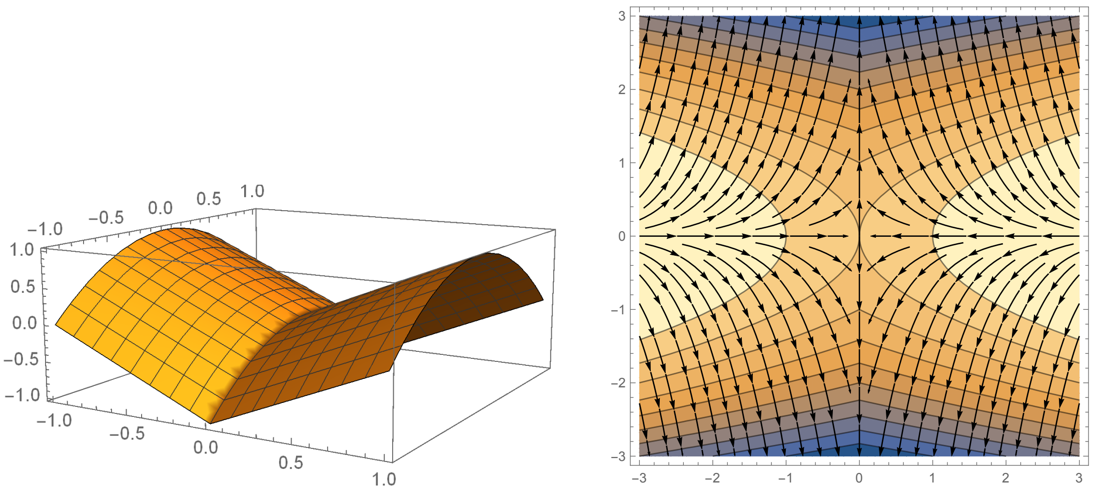
As the starting point, let us recall the baseline guarantee from [60] for the subgradient method when applied to a semialgebraic function f, or more generally one definable in an o-minimal structure. The main result of [77] shows that for such functions, almost surely, every limit point $\bar{x}$ of the subgradient sequence $\{x_k\}$ is Clarke critical. Explicitly, this means that the zero vector lies in the Clarke subdifferential

$$\partial_c f(\bar{x}) = \mathrm{conv}\left\{\lim_{i\to\infty} \nabla f(y_i) : y_i \in \mathrm{dom}\,(\nabla f), y_i \to \bar{x}\right\}.$$

Therefore, our task reduces to isolating geometric conditions around extraneous Clarke critical points which facilitate local escape of the subgradient sequence.

The main difficulty in contrast to the smooth setting is that there is no simple analogue of the Morse lemma that can reduce a nonsmooth function to a common functional form by a diffeomorphism. Instead, a fundamentally different idea is required. Our arguments focus on a certain smooth manifold that captures the "nonsmooth activity" of the function near a critical point. Formal models of such manifolds have appeared throughout the optimization literature, notably in [1–6]. Following [6], a smooth embedded submanifold $\mathcal{M}$ of $\mathbb{R}^d$ is called *active* for f at $\bar{x}$ if (*i*) the restriction of f to $\mathcal{M}$ is smooth near $\bar{x}$, and (*ii*) the subgradients $w \in \partial_c f(x)$ are uniformly bounded away from zero at all points $x \in \mathbb{R}^n \setminus \mathcal{M}$ near $\bar{x}$. For subdifferentially regular functions, such manifolds are geometrically distinctive in that f varies smoothly along $\mathcal{M}$ and sharply in directions normal to $\mathcal{M}$. As an illustration, Figure 4.1a depicts a nonsmooth function, having the y-axis as the active manifold around the the critical point (origin). A critical point $\bar{x}$ is called an *active strict saddle* if f decreases quadratically along some smooth path in the active manifold $\mathcal{M}$ emanating from $\bar{x}$. Returning to Figure 4.1, the origin is indeed an active strict saddle since f has negative curvature along the y-axis at the origin. Our focus on active manifolds and active saddles is justified because these structures are in a sense generic for definable functions. Indeed, the earlier work [59, 60] shows that for a definable function f, there exists a full-measure set of perturbations $v \in \mathcal{V}$ such that every critical point $\bar{x}$ of the tilted function $f_v(x) = f(x) - \langle v, x \rangle$ lies on a unique active manifold and is either a local minimizer or an active strict saddle.

The importance of the active manifold for subgradient dynamics is best illustrated in continuous time by looking at the trajectories of the differential inclusion $\dot{\gamma} \in -\partial_c f(\gamma)$. Returning to the running example, Figure 4.1b shows that the set of initial conditions that are attracted to the critical point by subgradient flow (x-axis) has zero measure. It appears therefore that although the subgradient method never reaches the active manifold, it nonetheless inherits desirable properties from the function along the manifold, e.g.,



(a) The function $f(x, y)$ (b) Subgradient flow $\dot{\gamma} \in -\partial_c f(\gamma)$

Figure 4.1: The y-axis is an active manifold for the function $f(x, y) = |x| - y^2$ at the origin.

saddle point avoidance. In this chapter, we rigorously verify this general phenomenon.

Our central observation is that under two mild regularity conditions on f, which we will describe shortly, the subgradient dynamics can be understood as an inexact Riemannian gradient method on the restriction of f to $\mathcal{M}$. Explicitly, we will find that the "shadow sequence" $y_k = P_{\mathcal{M}}(x_k)$, satisfies the recursion

$$y_{k+1} = y_k - \alpha_k(\nabla_{\mathcal{M}} f(y_k) + P_{T_{\mathcal{M}(y_k)}}(v_k)) + o(\alpha_k), \tag{4.1.2}$$

near $\bar{x}$, where $P_{\mathcal{M}}(\cdot)$ is the nearest-point projection onto $\mathcal{M}$, $P_{T_{\mathcal{M}(y_k)}}$ is the projection onto the tangent space of $\mathcal{M}$ at y_k, and $\nabla_{\mathcal{M}} f$ denotes the *covariant gradient* of f along $\mathcal{M}$. [1] The "smooth" dynamic equation (4.1.2) will allow us to prove that the shadow iterates y_k eventually escape from any small neighborhood around an active strict saddle $\bar{x}$ of f, so does x_k.

The validity of (4.1.2) relies on two regularity properties of f that we developed in Chapter 3. We describe them here for the reader's convenience.

[1] The covariant gradient $\nabla_{\mathcal{M}} f(y)$ is the projection onto $T_{\mathcal{M}}(y)$ of $\nabla \hat{f}(y)$ where $\hat{f}$ is any C^1 smooth function defined on a neighborhood U of $\bar{x}$ and that agrees with f on $U \cap \mathcal{M}$.

Regularity property I: aiming towards the manifold. The first condition we require is simply that near the critical point, subgradients are well aligned with directions pointing towards the nearest point on the manifold. Formally, we model this condition with the *proximal aiming* inequality:

$$\langle v, x - P_{\mathcal{M}}(x)\rangle \geq c \cdot \mathrm{dist}(x, \mathcal{M}) \qquad \text{for all } x \text{ near } \bar{x} \text{ and } v \in \partial f_c(x). \tag{4.1.3}$$

for some constant $c > 0$. It is not hard to see that if f is subdifferentially regular and $\mathcal{M}$ is its active manifold, then proximal aiming (4.1.3) is implied by the (b)-regularity condition:

$$f(y) \geq f(x) + \langle v, y - x\rangle + o(\|y - x\|) \qquad \text{as } x \to \bar{x},\ y \xrightarrow{\mathcal{M}} \bar{x},\ \text{with } v \in \partial_c f(x). \tag{4.1.4}$$

We refer readers to Theorem 3.1.4 and Corollary 3.1.5 for details. This estimate stipulates that subgradients $v \in \partial_c f(x)$ yield affine minorants of f up to first-order near $\bar{x}$, but only when comparing points x and $y \in \mathcal{M}$. This condition is automatically true for weakly convex functions and holds in much broader settings as we saw in Chapter 3.

Regularity property II: subgradients on and off the manifold. The second regularity property posits that subgradients on and off the manifold are aligned in tangent directions up to a linear error, that is, there exists $C > 0$ satisfying

$$\|P_{T_{\mathcal{M}(y)}}(\partial_c f(x) - \nabla_{\mathcal{M}} f(y))\| \leq C \cdot \|x - y\| \qquad \text{for all } x \in \mathbf{R}^d \text{ and } y \in \mathcal{M} \text{ near } \bar{x}. \tag{4.1.5}$$

Whenever (4.1.5) holds, we say that f is *strongly (a)-regular along* $\mathcal{M}$. We refer readers to Theorem 3.1.6 for details.

Reassuringly, typical functions, whether built from concrete structured examples or from unstructured linear perturbations, admit an active manifold around each critical point along which the objective function is both (b) and strongly (a) regular.

In the final stages of completing the work [47] which this chapter is based on, we became aware of the concurrent and independent work [78]. The two papers, share similar core ideas, rooted in strong (a) regularity and proximal aiming. However, the proof of the main result in [78]—avoidance of saddle points—fundamentally relies on a claimed equivalence in [79, Theorem 4.1], which is known to be false. The most recent draft on arxiv takes a different approach that does not rely on [79, Theorem 4.1].

4.1.2 Outline of the chapter.

Section 4.2 introduces the algorithms that we study in the chapter and the relevant assumptions. Section 4.3 discusses the two pillars of our algorithmic development (aiming and strong (a)-regularity) and the dynamics of the shadow iteration. Section 4.4 presents the main results of the chapter on saddle-point avoidance. Most of the technical proofs appear in Sections 4.5 and 4.6, respectively.

4.2 Algorithm and main assumptions

In this chapter, we introduce our main algorithmic consequences of the strong (a) and $(b_\diamond)$ regularity properties developed in the previous sections. Setting the stage, throughout we consider a minimization problem

$$\min_{x\in\mathbb{R}^d} f(x), \tag{4.2.1}$$

where $f\colon \mathbb{R}^d \to \mathbb{R}\cup\{+\infty\}$ is a closed function. The function f may enforce constraints or regularization; it may also be the population loss of a stochastic optimization problem. In order to simultaneously model algorithms which exploit such structure, we take a fairly abstract approach, assuming access to a *generalized gradient mapping* for f:

$$G\colon \mathbb{R}_{++} \times \operatorname{dom} f \times \mathbb{R}^d \to \mathbb{R}^d$$

We then consider the following stochastic method: given $x_0 \in \mathbb{R}^d$, we iterate

$$x_{k+1} = x_k - \alpha_k G_{\alpha_k}(x_k, \nu_k), \tag{4.2.2}$$

where $\alpha_k > 0$ is a control sequence and ν_k is stochastic noise. We will place relevant assumptions on the noise ν_k later in Section 4.3. The most important example of (4.2.2), valid for locally Lipschitz functions f, is the stochastic subgradient method:

$$x_{k+1} = x_k - \alpha_k(w_k + \nu_k) \qquad \text{where } w_k \in \partial_c f(x_k),$$

In this case, the mapping G satisfies

$$G_\alpha(x, \nu) \in \partial_c f(x) + \nu \qquad \text{for all } x, \nu \in \mathbb{R}^d \text{ and } \alpha > 0. \tag{4.2.3}$$

More generally, G may represent a stochastic projected gradient method or a stochastic proximal gradient method—two algorithms we examine in detail in Section 4.2.1.

The purpose of this chapter is to understand how iteration (4.2.2) is affected by the existence of "active manifolds" $\mathcal{M}$ contained within the domain of f. For this, we posit a tight interaction between G and the active manifold $\mathcal{M}$, described in the following assumption.

Assumption A (Strong (a) and aiming)**.** Fix a point $\bar{x} \in \operatorname{dom} f$. We suppose that there exist constants $C, \mu > 0$, a neighborhood $\mathcal{U}$ of $\bar{x}$, and a C^3 manifold $\mathcal{M} \subseteq \operatorname{dom} f$ containing $\bar{x}$ such that the following hold for all $\nu \in \mathbb{R}^d$ and $\alpha > 0$, where we set $\mathcal{U}_f := \mathcal{U} \cap \operatorname{dom} f$.

(A1) **(Local Boundedness)** We have

$$\sup_{x \in \mathcal{U}_f} \|G_\alpha(x, \nu)\| \leq C(1 + \|\nu\|).$$

(A2) **(Strong (a))** The function f is C^2 on $\mathcal{M}$ and for all $x \in \mathcal{U}_f$, we have

$$\|P_{T_{\mathcal{M}}(P_{\mathcal{M}}(x))}(G_\alpha(x, \nu) - \nabla_{\mathcal{M}} f(P_{\mathcal{M}}(x)) - \nu)\| \leq C(1 + \|\nu\|)^2(\mathrm{dist}(x, \mathcal{M}) + \alpha).$$

(A3) **(Proximal Aiming)** For $x \in \mathcal{U}_f$ tending to $\bar{x}$, we have

$$\langle G_\alpha(x, \nu) - \nu, x - P_{\mathcal{M}}(x)\rangle \geq \mu \cdot \mathrm{dist}(x, \mathcal{M}) - (1 + \|\nu\|)^2(o(\mathrm{dist}(x, \mathcal{M})) + C\alpha).$$

Some comments are in order. Assumption (A1) is similar to classical Lipschitz assumptions and ensures the steplength can only scale linearly in $\|\nu\|$. Assumption (A2) is the natural analogue of strong (a) regularity for the operator $G_\alpha(x, \nu)$. It ensures that the shadow sequence $y_k = P_{\mathcal{M}}(x_k)$ locally remains an inexact stochastic Riemannian gradient sequence with implicit retraction. Assumption (A3) ensures that after subtracting the noise from $G_{\alpha_k}(x_k, \nu_k)$, the update direction $x_{k+1} - x_k$ locally points towards the manifold $\mathcal{M}$. We will later show that this ensures the iterates x_k approach the manifold $\mathcal{M}$ at a controlled rate. Finally we note in passing that the power of $(1 + \|\nu\|)$ in the above expressions must be at least 2 for common iterative algorithms to satisfy Assumption A; one may also take higher powers, but this requires higher moment bounds on $\|\nu_k\|$. Before making these results precise in Section 4.3, we first formalize our statements about the subgradient method and introduce several examples.

The rest of the section is devoted to examples of algorithms satisfying Assumption A.

4.2.1 Stochastic subgradient method

The most immediate example of operator G arises from the subgradient method applied to a locally Lipschitz function f. In this setting, any measurable selection $s \colon \mathbb{R}^d \to \mathbb{R}$ of

$\partial_c f(x)$ gives rise to a mapping

$$G_\alpha(x, v) = s(x) + v, \tag{4.2.4}$$

which is independent of α. Then Algorithm (4.2.2) is the classical stochastic subgradient method:

$$x_{k+1} = x_k - \alpha_k(s(x_k) + v_k). \tag{4.2.5}$$

Let us place the following assumption on f, which we will shortly show implies Assumption A.

Assumption B (Assumptions for the subgradient mapping)**.** Let $f\colon \mathbb{R}^d \to \mathbb{R}$ be a function that is locally Lipschitz continuous around a point $\bar{x} \in \mathbb{R}^d$. Let $\mathcal{M} \subseteq \mathcal{X}$ be a C^3 manifold containing $\bar{x}$ and suppose that f is C^2 on $\mathcal{M}$ near x.

(B1) **(Strong (a))** The function f is strongly (a)-regular along $\mathcal{M}$ near $\bar{x}$.

(B2) **(Proximal aiming)** There exists $\mu > 0$ such that the inequality holds

$$\langle v, x - P_{\mathcal{M}}(x)\rangle \geq \mu \cdot \operatorname{dist}(x, \mathcal{M}) \qquad \text{for all } x \text{ near } \bar{x} \text{ and } v \in \partial_c f(x). \tag{4.2.6}$$

Note that Corollary 3.1.5 shows that the aiming condition (B2) holds as long as $\mathcal{M}$ is an active manifold for f at $\bar{x}$ satisfying $0 \in \hat{\partial} f(\bar{x})$ and f is $(b_{\leq})$-regular along $\mathcal{M}$ near $\bar{x}$. The following proposition follows immediately from Corollary 3.1.5.

Proposition 4.2.1 (Subgradient method)**.** *Assumption B implies Assumption A with the map G defined in* (4.2.4).

Thus, all three properties arise from reasonable assumptions on the function f, as discussed in the previous sections. Moreover, for definable functions, they hold generically, as the following corollary shows. Indeed, this is a direct consequence of Theorem 3.7.6.

Corollary 4.2.2. *Suppose that $f\colon \mathbb{R}^d \to \mathbb{R}$ is locally Lipschitz and definable in o-minimal structure. Then there exists a finite N such that for a generic set of $v \in \mathbb{R}^d$ the tilted function $f_v(x) := f(x) - \langle v, x\rangle$ has at most N Clarke critical points. Moreover, each limiting critical point $\bar{x}$ is in fact Fréchet critical and satisfies the following.*

1. *The function f and the subgradient mapping* (4.2.5) *satisfy Assumption A at $\bar{x}$ with respect to some C^3 active manifold $\mathcal{M}$.*
2. *The limiting critical point $\bar{x}$ is either a local minimizer or an active strict saddle point of f.*

4.2.2 Stochastic projected subgradient method

Throughout this section let $g\colon \mathbb{R}^d \to \mathbb{R}$ be a locally Lipschitz function and let $\mathcal{X}$ be a closed set and consider the constrained minimization problem

$$\min f(x) := g(x) + \delta_{\mathcal{X}}(x).$$

A classical algorithm for solving this problem is known as the stochastic projected subgradient method. Each iteration of the method updates

$$x_{k+1} \in P_{\mathcal{X}}(x_k - \alpha_k(v_k + \nu_k)) \qquad \text{where } v_k \in \partial_c g(x_k) \tag{4.2.7}$$

This algorithm can be reformulated as an instance of (4.2.2). Indeed, let $s_{\mathcal{X}}\colon \mathbb{R}^d \to \mathbb{R}^d$ be a measurable selection of $P_{\mathcal{X}}$, let $s_g\colon \mathbb{R}^d \to \mathbb{R}^d$ be a measurable selection of $\partial_c g$, and define the generalized gradient mapping

$$G_\alpha(x, \nu) := \frac{x - s_{\mathcal{X}}(x - \alpha(s_g(x) + \nu))}{\alpha} \qquad \text{for all } x \in \mathbb{R}^d, \nu \in \mathbb{R}^d, \alpha > 0. \tag{4.2.8}$$

Evidently, the update rule (4.2.2) reduces to (4.2.7).

In order to ensure Assumption A for the stochastic projected subgradient method, we introduce the following assumptions on g and $\mathcal{X}$.

Assumption C (Assumptions for the projected gradient mapping)**.** Let $f := g + \delta_{\mathcal{X}}$, where $\mathcal{X}$ is a closed set and $g\colon \mathbb{R}^d \to \mathbb{R}$ is a locally Lipschitz continuous function. Fix $\bar{x} \in \mathbb{R}^d$ and let $\mathcal{M} \subseteq \mathcal{X}$ be a C^3 manifold containing $\bar{x}$ and suppose that f is C^2 on $\mathcal{M}$ near $\bar{x}$.

(C1) **(Strong (a))** The function g and set $\mathcal{X}$ are strongly (a)-regular along $\mathcal{M}$ at $\bar{x}$.

(C2) **(Proximal aiming)** There exists $\mu > 0$ such that the inequality holds

$$\langle v, x - P_{\mathcal{M}}(x)\rangle \geq \mu \cdot \operatorname{dist}(x, \mathcal{M}) \qquad \text{for all } x \in \mathcal{X} \text{ near } \bar{x} \text{ and } v \in \partial_c g(x). \quad (4.2.9)$$

(C3) **(Condition (b))** The set $\mathcal{X}$ is $(b_{\leq})$-regular along $\mathcal{M}$ at $\bar{x}$.

Note that Corollary 3.1.5 shows that the aiming condition (C2) holds as long as $\mathcal{M}$ is an active manifold for f at $\bar{x}$ satisfying $0 \in \hat{\partial} f(\bar{x})$ and f is $(b_{\leq})$-regular along $\mathcal{M}$ at $\bar{x}$.[2] The following proposition shows that Assumption C is sufficient to ensure Assumption A; we defer the proof to Appendix 7.1.2 since it's fairly long.

Proposition 4.2.3 (Projected subgradient method)**.** *Assumption C implies Assumption A for the map G defined in* (4.2.8).

Given this proposition, an immediate question is whether Assumption C holds generically under for problems that are definable in an o-minimal structure. The following

[2] Corollary 3.1.5 shows that there exists a constant $c > 0$ such that for any $\delta > 0$, the estimate

$$\langle v, x - P_{\mathcal{M}}(x)\rangle \geq (c - \delta\sqrt{1 + \|v\|^2}) \cdot \operatorname{dist}(x, \mathcal{M}), \qquad (4.2.10)$$

holds for all $x \in \mathcal{X}$ near $\bar{x}$ and for all $v \in \partial f(x)$. In particular, due to the inclusion $\hat{\partial} g(x) + \hat{N}_{\mathcal{X}}(x) \subset \hat{\partial} f(x)$, we may choose any $v \in \hat{\partial} g(x)$ in (4.2.10). Therefore, taking into account that g is locally Lipschitz, we deduce that there is a constant μ such that $\langle v, x - P_{\mathcal{M}}(x)\rangle \geq \mu \cdot \operatorname{dist}(x, \mathcal{M})$ for all $x \in \mathcal{X}$ near $\bar{x}$ and for all $v \in \hat{\partial} g(x)$. Taking limits and convex hulls, the same statement holds for all $v \in \partial_c g(x)$.

corollary, which is an immediate consequence of Proposition 4.2.3, Theorem 3.7.6, and Corollary 3.1.5, shows that the answer is yes.

Corollary 4.2.4. *Suppose that $f = g + \delta_{\mathcal{X}}$, where $\mathcal{X} \subseteq \mathbb{R}^d$ is closed and $g\colon \mathbb{R}^d \to \mathbb{R}$ is locally Lipschitz, and both $\mathcal{X}$ and g are definable in an o-minimal structure. Then there exists a finite N such that for a generic set of $v, w \in \mathbb{R}^d$ the tilted function $f_{v,w}(x) := g(x + w) + \delta_{\mathcal{X}}(x) - \langle v, x\rangle$ has at most N composite Clarke critical points. Moreover, each composite limiting critical point $\bar{x}$ is in fact Fréchet critical and satisfies the following,*

1. *The function f and the projected subgradient mapping G define in* (4.2.8) *satisfy Assumption A at $\bar{x}$ with respect to some C^3 active manifold $\mathcal{M}$.*
2. *The composite limiting critical point $\bar{x}$ is either a local minimizer or an active strict saddle point of f.*

In the above corollary, the qualification *composite critical points*, as defined in Theorem 3.7.6, is important, since the projected subgradient method is only known to converge to such points.

4.2.3 Proximal gradient method

Throughout this section let $g\colon \mathbb{R}^d \to \mathbb{R}$ be a C^1 function and let $h\colon \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a closed function. We then consider the minimization problem

$$\min_{x\in\mathbb{R}^d} f(x) := g(x) + h(x).$$

A classical algorithm for solving this problem is the stochastic proximal gradient method. Each iteration of the method solves the proximal problem:

$$x_{k+1} \in \operatorname*{argmin}_{x\in\mathbb{R}^d} \left\{ h(x) + \langle \nabla g(x_k) + \nu_k, x - x_k\rangle + \frac{1}{2\alpha_k}\|x - x_k\|^2 \right\}. \tag{4.2.11}$$

This algorithm can be reformulated as an instance of (4.2.2). Indeed, let $s\colon \mathbb{R}_{++} \times \mathbb{R}^d \to \mathbb{R}^d$ be a measurable selection of the proximal map $(x, \alpha) \mapsto \operatorname{argmin}_y\{h(y) + \frac{1}{2\alpha}\|y - x\|^2\}$ and consider the mapping G defined by

$$G_\alpha(x, v) = \frac{x - s_\alpha(x - \alpha(\nabla g(x) + v))}{\alpha} \qquad \text{for all } x \in \mathbb{R}^d,\ v \in \mathbb{R}^d \text{ and } \alpha > 0. \tag{4.2.12}$$

Evidently, the update rule (4.2.2) is equivalent to (4.2.11).

In order to ensure Assumption A for the stochastic proximal gradient method, we introduce the following assumptions on g and h.

Assumption D (Assumptions for the proximal gradient mapping)**.** Let $f := g + h$, where $g\colon \mathbb{R}^d \to \mathbb{R}$ is C^1 and $h\colon \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is closed. Denote $\mathcal{X} := \operatorname{dom} h$ and let $\mathcal{M} \subseteq \mathcal{X}$ be a C^3 manifold containing some point $\bar{x}$ and suppose that f is C^2 on $\mathcal{M}$ near $\bar{x}$.

(D1) **(Lipschitz gradient/boundedness)** The gradient ∇g Lipschitz near $\bar{x}$. Moreover, there exists $C > 0$ such that $\|\nabla g(x)\| \leq C(1 + \|x\|)$ for all $x \in \mathcal{X}$.

(D2) **(Lipschitz proximal term)** The function h is Lipschitz on $\mathcal{X}$.

(D3) **(Strong (a))** The function h is strongly (a)-regular along $\mathcal{M}$ at $\bar{x}$.

(D4) **(Proximal Aiming)** There exists $\mu > 0$ such that the inequality holds

$$\langle v, x - P_{\mathcal{M}}(x)\rangle \geq \mu \cdot \operatorname{dist}(x, \mathcal{M}) - (1 + \|v\|)o(\operatorname{dist}(x, \mathcal{M})) \tag{4.2.13}$$

for all $x \in \operatorname{dom} h$ near $\bar{x}$ and $v \in \partial f(x)$.

Note that Corollary 3.1.5 shows that the aiming condition (D4) holds as long as $\mathcal{M}$ is an active manifold for f at $\bar{x}$ satisfying $0 \in \hat{\partial} f(\bar{x})$ and f is $(b_\leq)$-regular along $\mathcal{M}$ at $\bar{x}$. The following proposition shows that Assumption D is sufficient to ensure Assumption A. The proof of the Proposition appears in Appendix 7.1.3

Proposition 4.2.5 (Proximal gradient method)**.** *If assumption D holds at $\bar{x} \in \operatorname{dom} f$, then f and G satisfy Assumption A at $\bar{x}$.*

The following corollary, which is an immediate consequence of Proposition 4.2.5 and Theorem 3.7.5, shows that assumption D is automatically true for definable problems.

Corollary 4.2.6. *Suppose that $f = g + h_0 + \delta_{\mathcal{X}}$, where $\mathcal{X} \subseteq \mathbb{R}^d$, g is a C^1 function with Lipschitz gradient, the function $h_0 \colon \mathbb{R}^d \to \mathbb{R}$ is Lipschitz on $\mathcal{X}$, and we define $h := h_0 + \delta_{\mathcal{X}}$. Suppose that g, h_0, and $\mathcal{X}$ are definable in an o-minimal structure. Then there exists a finite N such that for a full measure set of $v, w \in \mathbb{R}^d$, the tilted function $f_{v,w} := g(x + w) + h_0(x + w) + \delta(x) - \langle v, x \rangle$ has at most N composite Clarke critical points $\bar{x}$. Moreover, each composite limiting critical point $\bar{x}$ is in fact composite Fréchet critical and satisfies the following.*

1. *The function f and the proximal gradient mapping* (4.2.12) *satisfy Assumption A at $\bar{x}$ with respect to some active manifold $\mathcal{M}$.*
2. *The critical point $\bar{x}$ is either a local minimizer or an active strict saddle point of f.*

Thus, we find that Assumption A is satisfied for common iterative mappings, under reasonable assumptions, and is even automatic for certain generic classes of functions. In the next several sections, we turn our attention to the algorithmic consequences of theses assumptions.

4.3 The two pillars

Assumption A at a point $\bar{x}$ guarantees two useful behaviors, provided the iterates $\{x_k\}$ of iteration (4.2.2) remain in a small ball around $\bar{x}$. First x_k must approach the manifold $\mathcal{M}$ containing $\bar{x}$ at a controlled rate, a consequence of the proximal aiming condition. Second the shadow $y_k = P_{\mathcal{M}}(x_k)$ of the iterates along the manifold form an approximate Riemannian stochastic gradient sequence with an implicit retraction. Moreover, the approximation error of the sequence decays with $\operatorname{dist}(x_k, \mathcal{M})$ and α_k, quantities that quickly tend to zero.

The formal statements of our results crucially require local arguments and frequently refer to the following stopping time: given an index $k \geq 1$ and a constant $\delta > 0$, define

$$\tau_{k,\delta} := \inf\{j \geq k \colon x_j \notin B_\delta(\bar{x})\}. \tag{4.3.1}$$

Note that the stopping time implicitly depends on $\bar{x}$, a point at which Assumption A is satisfied. In the statements of our result, the point $\bar{x}$ will always be clear from the context. Second, we make the following standing assumption on α_k and ν_k. We assume they are in force throughout the rest of the sections.

Assumption E (Standing assumptions)**.** Assume the following.

(E1) The map G is measurable.

(E2) There exist constants $c_1, c_2 > 0$ and $\gamma \in (1/2, 1]$ such that

$$\frac{c_1}{k^\gamma} \leq \alpha_k \leq \frac{c_2}{k^\gamma}.$$

(E3) $\{\nu_k\}$ is a martingale difference sequence w.r.t. to the increasing sequence of σ-fields

$$\mathcal{F}_k = \sigma(x_j \colon j \leq k \text{ and } \nu_j \colon j < k),$$

and there exists a function $q\colon \mathbb{R}^d \to \mathbb{R}_+$ that is bounded on bounded sets with

$$\mathbb{E}[\nu_k \mid \mathcal{F}_k] = 0 \qquad \text{and} \qquad \mathbb{E}[\|\nu_k\|^4 \mid \mathcal{F}_k] < q(x_k).$$

We let $\mathbb{E}_k[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_k]$ denote the conditional expectation.

(E4) The inclusion $x_k \in \operatorname{dom} f$ holds for all $k \geq 1$.

All items in Assumption E are standard in the literature on stochastic approximation methods and mirror those found in [77, Assumption C]. The only exception is the fourth moment bound on $\|\nu_k\|$, which stipulates that ν_k has slightly lighter tails. This bound appears to be necessary for the setting we consider. We now turn to the first pillar.

4.3.1 Pillar I: Aiming towards the manifold

The following proposition ensures the sequence x_k approaches the manifold. The proof appears in Section 4.5.1.

Proposition 4.3.1. *Suppose that f satisfies Assumption A at $\bar{x}$. Let $\gamma \in (1/2, 1]$ and assume $c_1 \geq 32/\mu$ if $\gamma = 1$. Then for all $k_0 \geq 1$ and sufficiently small $\delta > 0$, there exists a constant C, such that the following hold with stopping time $\tau_{k_0,\delta}$ defined in* (4.3.1)*:*

1. *There exists a random variable $V_{k_0,\delta}$ such that*

   (a) *The limit holds:*

$$\frac{k^{2\gamma-1}}{\log(k+1)^2}\operatorname{dist}^2(x_k, \mathcal{M})1_{\tau_{k_0,\delta}>k} \xrightarrow{a.s.} V_{k_0,\delta}.$$

   (b) *The sum is almost surely finite:*

$$\sum_{k=1}^{\infty} \frac{k^{\gamma-1}}{\log(k+1)^2}\operatorname{dist}(x_k, \mathcal{M})1_{\tau_{k_0,\delta}>k} < +\infty.$$

2. *We have*

   (a) *The expected squared distance satisfies:*

$$\mathbb{E}[\mathrm{dist}^2(x_k, \mathcal{M})1_{\tau_{k_0,\delta}>k}] \leq C\alpha_k \qquad \textit{for all } k \geq 1.$$

   (b) *The tail sum is bounded:*

$$\mathbb{E}\left[\sum_{i=k}^{\infty} \alpha_i \mathrm{dist}(x_i, \mathcal{M})1_{\tau_{k_0,\delta}>i}\right] \leq C \sum_{i=k}^{\infty} \alpha_i^2 \qquad \textit{for all } k \geq 1.$$

We note that Part 1b of the proposition holds not only almost surely, but also in expectation, which is a stronger statement in general. Now we turn our attention to Pillar II: the shadow iteration.

4.3.2 Pillar II: The shadow iteration

Next we study the evolution of the shadow $y_k = P_{\mathcal{M}}(x_k)$ along the manifold, showing that y_k is locally an inexact Riemannian stochastic gradient sequence with error that asymptotically decays as x_k approaches the manifold. Consequently, we may control the error using Proposition 4.3.1. The proof appears in Section 4.5.2

Proposition 4.3.2. *Suppose that f satisfies Assumption A at $\bar{x}$. Then for all $k_0 \geq 1$ and sufficiently small $\delta > 0$, there exists a constant C, such that the following hold with stopping time $\tau_{k_0,\delta}$ defined in* (4.3.1)*: there exists a sequence of $\mathcal{F}_{k+1}$-measurable random vectors $E_k \in \mathbb{R}^d$ such that*

1. *The shadow sequence*

$$y_k = \begin{cases} P_{\mathcal{M}}(x_k) & \textit{if } x_k \in B_{2\delta}(\bar{x}) \\ \bar{x} & \textit{otherwise.} \end{cases}$$

satisfies $y_k \in B_{4\delta}(\bar{x}) \cap \mathcal{M}$ for all k and the recursion holds:

$$\boxed{y_{k+1} = y_k - \alpha_k \nabla_{\mathcal{M}} f(y_k) - \alpha_k P_{T_{\mathcal{M}}(y_k)}(\nu_k) + \alpha_k E_k \qquad \text{for all } k \geq 1.} \tag{4.3.2}$$

Moreover, for such k, we have $\mathbb{E}_k[P_{T_{\mathcal{M}(y_k)}}(\nu_k)] = 0$.

2. *Let $\gamma \in (1/2, 1]$ and assume that $c_1 \geq 32/\mu$ if $\gamma = 1$.*

   (a) *We have the following bounds for $k_0 \leq k \leq \tau_{k_0,\delta} - 1$:*

      i. $\|E_k\| 1_{\tau_{k_0,\delta} > k} \leq C(1 + \|\nu_k\|)^2(\operatorname{dist}(x_k, \mathcal{M}) + \alpha_k) 1_{\tau_{k_0,\delta} > k}$

      ii. $\max\{\mathbb{E}_k[\|E_k\|] 1_{\tau_{k_0,\delta} > k}, \mathbb{E}_k[\|E_k\|^2] 1_{\tau_{k_0,\delta} > k}\} \leq C$.

      iii. $\mathbb{E}[\|E_k\|^2] 1_{\tau_{k_0,\delta} > k} \leq C\alpha_k$

   (b) *The following sums are finite*

      i. $\sum_{k=1}^{\infty} \frac{k^{\gamma-1}}{\log(k+1)^2} \max\{\|E_k\| 1_{\tau_{k_0,\delta} > k}, \mathbb{E}_k[\|E_k\|] 1_{\tau_{k_0,\delta} > k}\} < +\infty$

      ii. $\sum_{k=1}^{\infty} \frac{k^{\gamma-1}}{\log(k+1)^2} \max\{\|E_k\|^2 1_{\tau_{k_0,\delta} > k}, \mathbb{E}_k[\|E_k\|^2] 1_{\tau_{k_0,\delta} > k}\} < +\infty$

   (c) *The tail sum is bounded*

$$\mathbb{E}\left[1_{\tau_{k_0,\delta} = \infty} \sum_{i=k}^{\infty} \alpha_i \|E_k\|\right] \leq C \sum_{i=k}^{\infty} \alpha_i^2 \qquad \text{for all } k \geq 1.$$

With the two pillars we separate our study of the sequence x_k into two orthogonal components: In the tangent/smooth directions, we study the sequence y_k, which arises from an inexact gradient method with rapidly decaying errors and is amenable to the techniques of smooth optimization. In the normal/nonsmooth directions, we steadily approach the manifold, allowing us to infer strong properties of x_k from corresponding properties for y_k.

4.4 Avoiding saddle points

In this section, we ask whether x_k can converge to points $\bar{x}$ at which $\nabla^2_{\mathcal{M}} f(\bar{x})$ has at least one strictly negative eigenvalue. We call such points *strict saddle points*, and when $\mathcal{M}$ is in addition an active manifold for f, then we call such points *active strict saddle points*. We use a well-known technique in the stochastic approximation literature: isotropic noise injection [61, 80–82].

Let us briefly describe this technique. Fix a point $p \in \mathbb{R}^d$ and consider a C^2 mapping $F_p \colon \mathbb{R}^d \to \mathbb{R}^d$ with an unstable zero at p, meaning $\nabla F_p(p)$ has an eigenvalue with a strictly positive real part. Then a well-known result of Pemantle [61] states that, with probability 1, the following perturbed iteration cannot converge to p:

$$\begin{Bmatrix} \text{Sample } \xi_k \sim \text{Unif}(B_1(0)) \\ \text{Set } Y_{k+1} = Y_k + \alpha_k F_p(Y_k) + \alpha_k \xi_k \end{Bmatrix}. \tag{4.4.1}$$

As stated, the result of [61] does not shed light on the iteration (4.2.2). Nevertheless, in light of (4.3.2), the shadow iteration y_k does satisfy an iteration similar to (4.4.1) with mapping

$$F_p(y) = -\nabla(f \circ P_{\mathcal{M}})(y),$$

which under reasonable assumptions is locally C^2 near p and satisfies $F_p(y) = -\nabla_{\mathcal{M}} f(y)$ and $\nabla F_p(y) = -\nabla^2_{\mathcal{M}} f(y)$ for all $y \in \mathcal{M}$ near p. Moreover, if p is an active strict saddle of f, then $\nabla^2_{\mathcal{M}} f(p)$ has a strictly negative eigenvalue, so p is an “unstable zero” of F_p. Thus, we might reasonably expect y_k to converge to p only with probability zero. If this is the case, we can then lift the argument to x_k, showing that if x_k converges to p, then so does y_k—a probability zero event. This is the strategy we will apply in what follows, taking into account the additional error term E_k in the shadow iteration (4.3.2), a key technical issue that we have so far ignored.

In order to formalize the above strategy, we prove the following extension of the main result of [61] which takes into account the relationship between x_k and y_k described above. The proof, which we defer to Section 4.6.1, draws on the techniques of [61, 80–83].

Theorem 4.4.1 (Nonconvergence)**.** *Fix $c_1, c_2 > 0$ and let $S \subseteq \mathbb{R}^d$. Suppose for any $p \in S$, there exists a ball $B_{\epsilon_p}(p)$ centered at p and a C^2 mapping $F_p \colon B_{\epsilon_p}(p) \to \mathbb{R}^d$ that vanishes at p and has a symmetric Jacobian $\nabla F_p(p)$ that has at least one positive eigenvalue. Suppose $\{X_k\}_{k=1}^{\infty}$ is a stochastic process and for any k_0, $p \in S$, and $\delta > 0$ define the stopping time:*

$$\tau_{k_0,\delta}(p) = \inf\{k \geq k_0 \colon X_k \notin B_\delta(p)\}.$$

Suppose that for any $p \in S$, $k_0 \geq 1$, and all sufficiently small $\delta_p \leq \epsilon_p$ the following hold: there exists $c_3, c_4 > 0$ possibly depending on p, but not on δ_p and ϵ_p, such that on the event $\Omega_0 = \{\tau_{k_0,\delta_p}(p) = \infty\}$, we have

1. ***(Local iteration.)** There exists a process $\{Y_k \colon k \geq k_0\} \subseteq B_{\epsilon_p/2}(p)$ satisfying*

   $$Y_{k+1} = Y_k + \alpha_k F_p(Y_k) + \alpha_k \xi_k + \alpha_k E_k \tag{4.4.2}$$

   for error sequence $\{E_k\}$, noise sequence $\{\xi_k\}$, and deterministic stepsize sequence $\{\alpha_k\}$ that are square summable, but not summable.

2. ***(Noise Conditions.)** Let $\mathcal{F}_k$ be the sigma algebra generated by $X_{k_0}, \ldots, X_k$ and $Y_{k_0}, \ldots, Y_k$. Define W_p to be the subspace of eigenvectors of $\nabla F_p(p)$ with positive eigenvalues. Then we have*

   (a) $\mathbb{E}[\xi_k \mid \mathcal{F}_k] = 0$.

   (b) $\limsup_k \mathbb{E}[\|\xi_k\|^4 \mid \mathcal{F}_k] \leq c_3$.

   (c) $\mathbb{E}[|\langle \xi_k, w\rangle| \mid \mathcal{F}_k] \geq c_4$ *for $k \geq k_0$ and all unit norm $w \in W_p$.*

3. ***(Error Conditions.)***

   (a) *We have* $\limsup_k \mathbb{E}[1_{\Omega_0}\|E_k\|^4 \mid \mathcal{F}_k] < \infty$.

   (b) *For all* $n \geq k_0$, *we have* $\mathbb{E}\left[1_{\Omega_0} \sum_{k=n}^{\infty} \alpha_k \|E_k\|\right] = O_{k_0}\left(\sum_{k=n}^{\infty} \alpha_k^2\right)$.

Then $P(\lim_{k\to\infty} X_k \in S) = 0$.

Looking at the theorem, recursion condition (4.4.2) is clearly modeled on the shadow sequence of Proposition 4.3.2. Moreover, the error condition 3b on E_k precisely matches 2c. Finally, the noise ξ_k is modeled on $P_{T_{\mathcal{M}(y_k)}}(v_k)$ in the shadow iteration, which is mean zero and has bounded fourth moment. Condition 2c is not automatic for all noise distributions and requires that v_k has nontrivial mass in all directions of negative curvature for f.

Given Theorem 4.4.1, we now ask: can x_k converge to critical points $\bar{x}$ at which $\nabla^2_{\mathcal{M}} f(\bar{x})$ has a strict negative eigenvalue? In the following theorem we show that the answer is no, provided that we choose the noise v_k according to the following assumption:

Assumption F (Uniform noise)**.** There exists $r > 0$ such that $v_k \sim \text{Unif}(B_r(0))$ for all k.

The proof of the theorem appears in Section 4.6.2.

Theorem 4.4.2 (Nonconvergence to strict saddle point)**.** *Let* $S \subseteq \mathbb{R}^d$ *and suppose that Assumption A holds at each point* $\bar{x} \in S$, *where each manifold is* C^4. *Let* $\mathcal{M}$ *be the manifold associated to a point* $\bar{x} \in S$ *and suppose that* $\nabla^2_{\mathcal{M}} f(\bar{x})$ *has a strictly negative eigenvalue. Suppose that* v_k *satisfies Assumption F. In addition, suppose that* $\gamma \in (\frac{1}{2}, 1)$. *Then*

$$P\left(\lim_{k\to\infty} x_k \in S\right) = 0. \tag{4.4.3}$$

Note that the theorem applies to arbitrary sets S, making no assumptions on countability/isolatedness. Second the result does not preclude the *limit points* of x_k from lying in S. Thus, the result is useful only when x_k is known to converge.

We now examine two applications of the above theorem for the projected and proximal subgradient methods. The following corollary provides sufficient conditions for the projected subgradient method to avoid active strict saddle points. We place the proof in Appendix 7.1.4.

Corollary 4.4.3 (Projected subgradient methods)**.** *Suppose that $f = g + \delta_{\mathcal{X}}$, where $g\colon \mathbb{R}^d \to \mathbb{R}$ is locally Lipschitz and $\mathcal{X} \subseteq \mathbb{R}^d$ is closed. Let S consist of points x satisfying $0 \in \hat{\partial} f(x)$ and that are C^4 active strict saddle points of f. Suppose the following hold for all $x \in S$ with associated active manifold $\mathcal{M}_x$:*

1. *The function g and the set $\mathcal{X}$ are strongly (a)-regular along $\mathcal{M}_x$ at x.*
2. *The function g is weakly convex around x or $(b_{\leq})$-regular along $\mathcal{M}_x$ at x.*
3. *The set $\mathcal{X}$ is prox-regular at x or $(b_{\leq})$-regular along $\mathcal{M}_x$ at x.*

Suppose that v_k satisfies Assumption F. Then the iterates of the stochastic projected subgradient method (4.2.7) *satisfy*

$$P\left(\lim_{k\to\infty} x_k \in S\right) = 0.$$

Next we analyze the the proximal gradient method. Recall that the paper [60] showed that randomly initialized proximal gradient methods avoid active strict saddles of weakly convex functions. The following Corollary shows that the same behavior holds for perturbed proximal gradient methods beyond the weakly convex class. We place the proof in Appendix 7.1.5.

Corollary 4.4.4 (Proximal gradient methods)**.** *Suppose that $f = g + h$, where $h\colon \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is closed and Lipschitz on its domain $\mathcal{X} := \operatorname{dom} h$ and $g\colon \mathbb{R}^d \to \mathbb{R}$ is C^1 with Lipschitz continuous gradient on $\mathcal{X}$. Let S consist of points x satisfying $0 \in \hat{\partial} f(x)$ and that are C^4 active strict saddle points of f. Suppose that for all $x \in S$ with associated active manifold $\mathcal{M}_x$, the function f is strong (a)-regular and $(b_{\leq})$-regular along $\mathcal{M}_x$ at x. Suppose that ν_k satisfies Assumption F. Then the iterates of the stochastic proximal gradient method* (4.2.7) *satisfy*

$$P\left(\lim_{k\to\infty} x_k \in S\right) = 0.$$

4.4.1 Consequences for generic semialgebraic functions

The results we have presented so far show that the perturbed projected subgradient and the proximal gradient method cannot converge to Fréchet active strict saddle points, provided that x_k converges and various regularity properties hold. Although the convergence of x_k and the required regularity properties may seem stringent, they are in a precise sense generic. Indeed, the genericity of the regularity properties was already addressed in Section 3.7. Convergence also holds generically: it is known that all limit points of the stochastic subgradient method, the stochastic projected subgradient method, and the stochastic proximal method are (composite) Clarke critical points, as long as f is a semialgebraic function [77, Corollary 6.4.]. Thus, since generic semialgebraic functions have only finitely many (composite) Clarke critical points and one can show (with small effort) that the set of limit points of each algorithm is connected, it follows that the entire sequence x_k must converge on generic problems (if the sequence remains bounded). Thus we have the following three corollaries, whose proofs we place in Appendix 7.1.6.

Corollary 4.4.5 (Subgradient method on generic semialgebraic functions)**.** *Let $f\colon \mathbb{R}^d \to$*

$\mathbb{R}$ be a locally Lipschitz semialgebraic function. Then for a full measure set of v the following is true for the tilted function $f_v(x) := f(x) - \langle v, x\rangle$: Let $\{x_k\}_{k\in\mathbb{N}}$ be generated by the subgradient method 4.2.5 on f_v. Suppose that v_k satisfies Assumption F. Then on the event $\{x_k\}_{k\in\mathbb{N}}$ is bounded, almost surely we have only two possibilities

1. *x_k converges to a local minimizer $\bar{x}$ of f_v.*
2. *x_k converges to a Clarke critical point of f_v*

Thus, if f is Clarke regular, the sequence x_k must converge to a local minimizer of f_v.

Corollary 4.4.6 (Projected subgradient method on generic semialgebraic functions)**.** *Let $f = g + \delta_{\mathcal{X}}$, where $\mathcal{X} \subseteq \mathbb{R}^d$ semialgebraic and closed and $g\colon \mathbb{R}^d \to \mathbb{R}$ is locally Lipschitz and semialgebraic. Then for a full measure set of $v, w \in \mathbb{R}^d$ the following is true for the tilted function $f_{v,w}(x) := g(x + w) + \delta_{\mathcal{X}}(x) - \langle v, x\rangle$. Let $\{x_k\}_{k\in\mathbb{N}}$ be generated by the projected subgradient method 4.2.8. Suppose that v_k satisfies Assumption F. Then on the event $\{x_k\}_{k\in\mathbb{N}}$ is bounded, almost surely we have only two possibilities*

1. *x_k converges to a local minimizer $\bar{x}$ of $f_{v,w}$.*
2. *x_k converges to a composite Clarke critical point of $f_{v,w}$.*

Thus, if g and $\mathcal{X}$ are Clarke regular, the sequence x_k converges to a local minimizer of $f_{v,w}$.

Corollary 4.4.7 (Proximal gradient method on generic semialgebraic functions)**.** *Suppose that $f = g + h_0 + \delta_{\mathcal{X}}$, where $\mathcal{X} \subseteq \mathbb{R}^d$, g is a C^1 function with Lipschitz gradient on $\mathcal{X}$, the function $h_0\colon \mathbb{R}^d \to \mathbb{R}$ is Lipschitz on $\mathcal{X}$, and we define $h := h_0 + \delta_{\mathcal{X}}$. Then for a full measure set of $v, w \in \mathbb{R}^d$ the following is true for the tilted function*

$f_{v,w} := g(x + w) + h_0(x + w) + \delta(x) - \langle v, x\rangle$. Let $\{x_k\}_{k\in\mathbb{N}}$ be generated by the proximal gradient method 4.2.12. Suppose that v_k satisfies Assumption F. Then on the event $\{x_k\}_{k\in\mathbb{N}}$ is bounded, almost surely we have only two possibilities

1. *x_k converges to a local minimizer $\bar{x}$ of $f_{v,w}$.*
2. *x_k converges to a composite Clarke critical point of $f_{v,w}$.*

Thus, if h_0 and $\mathcal{X}$ are Clarke regular, the sequence x_k converges to a local minimizer of $f_{v,w}$.

In short, the main conclusion of the above three theorems is

On generic regular semialgebraic functions, perturbed
subgradient/proximal methods converge only to local minimizers

We note in passing that the results hold verbatim if one replaces the word "semialgebraic" with "definable in an *o*-minimal structure," throughout.

4.5 Proofs of the two pillars

Throughout this section, we let $\mathbb{E}_k[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_k]$ denote the conditional expectation. We now present the proofs of the two pillars.

4.5.1 Proof of Proposition 4.3.1: aiming towards the manifold

Throughout the proof, we let C denote a constant depending on k_0 and δ, which may change from line to line. Choose $\delta \leq \min\{1, \frac{c_1\mu}{12\gamma}\}$, satisfying $B_\delta(\bar{x}) \subseteq \mathcal{U}$ where $\mathcal{U}$ is the neighborhood in which Assumption A holds. Define $Q := \max\{\sup_{x \in B_\delta} q(x), 1\}$. By shrinking δ slightly, we can assume that the little o term in (A3) satisfies

$$o(\mathrm{dist}(x, \mathcal{M})) \leq \frac{\mu}{4(1+Q)} \mathrm{dist}(x, \mathcal{M}) \qquad \text{for all } x \in B_\delta(\bar{x}).$$

Now define: $D_k := \mathrm{dist}(x_k, \mathcal{M})$ for all $k \geq 0$. We prove a recurrence relation satisfied by the sequence D_k. To that end, define $v_k = G_{\alpha_k}(x_k, \nu_k)$ and observe that in the event $A_k := \{\tau_{k_0,\delta} > k\}$, we have

$$\begin{aligned}
D_{k+1}^2 &\leq \|x_{k+1} - P_{\mathcal{M}}(x_k)\|^2 \\
&= \|x_k - \alpha_k v_k - P_{\mathcal{M}}(x_k)\|^2 \\
&= \|x_k - P_{\mathcal{M}}(x_k)\|^2 - 2\alpha_k \langle v_k, x_k - P_{\mathcal{M}}(x_k)\rangle + \alpha_k^2 \|v_k\|^2 \\
&\leq D_k^2 - 2\alpha_k \mu D_k + 2\alpha_k (1 + \|v_k\|)^2 o(D_k) \\
&\qquad - 2\alpha_k \langle v_k, x_k - P_{\mathcal{M}}(x_k)\rangle + \underbrace{C(1 + \|v_k\|)^2}_{:=B_k} \alpha_k^2,
\end{aligned} \tag{4.5.1}$$

where the second inequality follows from the proximal aiming and local boundedness properties of G; see Assumption A. This inequality will allow us to prove all parts of the result.

Indeed, let us prove Part 1. To that end, first note that the bound $\mathbb{E}_k[\|v_k\|^2]1_{A_k} \leq q(x_k)1_{A_k} \leq Q$ implies that there exists $C > 0$ such that

$$\mathbb{E}_k[B_k]1_{A_k} \leq C,$$

meaning the conditional expectation is bounded for all k. Moreover, by our choice of δ,

$$\mathbb{E}_k[(1 + \|v\|)^2 o(D_k) 1_{A_k}] \leq \frac{\mu}{2} D_k 1_{A_k}.$$

Thus, for each k, we have

$$\mathbb{E}_k[D_{k+1}^2 1_{A_{k+1}}] \leq \mathbb{E}_k[D_{k+1}^2 1_{A_k}]$$

$$
\begin{aligned}
&\le D_k^2 1_{A_k} - \alpha_k \mu D_k 1_{A_k} + \mathbb{E}_k[B_k] 1_{A_k} \alpha_k^2 - 2\alpha_k \langle \mathbb{E}_k[\nu_k], x_k - P_{\mathcal{M}}(x_k) \rangle 1_{A_k} \\
&\le D_k^2 1_{A_k} - \alpha_k \mu D_k 1_{A_k} + C\alpha_k^2 \\
&\le (1 - (\alpha_k/2)\mu) D_k^2 1_{A_k} - (\alpha_k/2)\mu D_k 1_{A_k} + C\alpha_k^2, \qquad (4.5.2)
\end{aligned}
$$

where the first inequality follows from $1_{A_{k+1}} \le 1_{A_k}$; the second inequality follows from $\mathcal{F}_k$-measurability of A_k; and the fourth inequality follows since $D_k 1_{A_k} \ge D_k^2 1_{A_k}$ (recall $\delta \le 1$). Now apply Lemma 7.1.6 with the sequences $X_k := D_k^2 1_{A_k}, Y_k := \alpha_k \mu D_k 1_{A_k}$, and $Z_k := C\alpha_k^2$ and deduce that $(k^{2\gamma-1}/\log(k+1)^2) D_k^2$ almost surely converges to a finite valued random variable and the following sum is finite:

$$
\sum_{k=1}^{\infty} \frac{k^{2\gamma-1}\alpha_k}{\log(k+1)^2} \mu D_k 1_{A_k} < +\infty.
$$

Recalling that $\alpha_k \ge c_1/k^\gamma$, we get the claimed summability result.

Next we prove Part 2. To that end, take expectation of (4.5.2) and use the law of total expectation to deduce that for some $C > 0$, we have

$$
\begin{aligned}
\mathbb{E}[D_{k+1}^2 1_{A_k}] &\le (1 - \mu\alpha_k/2)\mathbb{E}[D_k^2 1_{A_k}] - (\alpha_k/2)\mu\mathbb{E}[D_k 1_{A_k}] + C\alpha_k^2 \\
&\le (1 - \mu c_1 k^{-\gamma}/2)\mathbb{E}[D_k^2 1_{A_k}] - (\alpha_k/2)\mu\mathbb{E}[D_k 1_{A_k}] + Ck^{-2\gamma}
\end{aligned}
$$

To prove part 2a, simply apply Lemma 7.1.8 applied with sequence $s_k = \mathbb{E}[D_k^2 1_{A_k}]$ and constants $c = \mu c_1/2$ and C. To prove part 2b, sum the above inequality from n to infinity to get

$$
\sum_{k=n}^{\infty} (\alpha_k/2)\mu\mathbb{E}[D_k 1_{A_k}] \le \mathbb{E}[D_n^2 1_{A_n}] + C\sum_{k=n}^{\infty} \alpha_k^2 \le Cn^{-\gamma} + C\sum_{k=n}^{\infty} \alpha_k^2,
$$

where the second inequality follows from Part 2a. Noting that $n^{-\gamma} = O(\sum_{k=n}^{\infty} \alpha_k^2)$ proves the result.

4.5.2 Proof of Proposition 4.3.2: the shadow iteration

Throughout the proof we let C denote a constant depending on k_0 and δ, but not on k, which may change from line to line. We assume δ is small enough that the conclusions of Proposition 4.3.1 hold; that $B_{4\delta}(\bar{x}) \subseteq \mathcal{U}$ where $\mathcal{U}$ is the neighborhood in which Assumption A holds; and that $P_{\mathcal{M}}$ and $\nabla P_{\mathcal{M}}$ are Lipschitz continuous on $B_{4\delta}(\bar{x})$. Write $\tau = \tau_{k_0,\delta}$ and fix index $k \geq 1$. Finally, recall that $P_{\mathcal{M}}$ is C^2 on $\mathcal{U}$ and $\nabla P_{\mathcal{M}}(x) = P_{T_{\mathcal{M}}(x)}$ for all $x \in \mathcal{M}$.

Let us first prove that $y_k \in B_{4\delta}(\bar{x})$. Clearly, we need only consider the case $x \in B_{2\delta}(\bar{x})$. In this case,

$$\|y_k - \bar{x}\| \leq \|y_k - x_k\| + \|x_k - \bar{x}\| \leq 2\|x_k - \bar{x}\| \leq 4\delta,$$

where the final inequality follows since $\bar{x} \in \mathcal{M}$. Therefore, we always have $\|y_k - \bar{x}\| \leq 4\delta$.

Next, let us define the error sequence E_k in the shadow iteration. To that end, denote $T_k := T_{\mathcal{M}}(y_k)$ and

$$w_k := y_k - \alpha_k \nabla_{\mathcal{M}} f(y_k) - \alpha_k P_{T_k}(v_k)$$

Then with error sequence $E_k := (y_{k+1} - w_k)/\alpha_k$, the claimed recursion is trivially true. Thus, in the remainder of the proof, we bound E_k.

Turning to the bound, we first note that throughout the proof, we must separate the analysis into two cases: $x_{k+1} \in B_{2\delta}(\bar{x})$ and $x_{k+1} \notin B_{2\delta}(\bar{x})$. In the second case, the following preliminary observation will be useful:

<u>Claim:</u> Suppose that in the event $\{\tau > k\}$ it holds that $x_{k+1} \notin B_{2\delta}(\bar{x})$. Then there exists $C > 0$ such that

$$\|y_{k+1} - y_k\| \leq 4\delta \leq C\|x_{k+1} - x_k\|. \tag{4.5.3}$$

Proof. First notice that

$$\|x_{k+1} - x_k\| \geq \|x_{k+1} - \bar{x}\| - \|x_k - \bar{x}\| \geq 2\delta - \delta \geq \delta.$$

Therefore, the result trivially holds since $\|y_{k+1} - y_k\| \leq 4\delta$. □

With the preliminaries set, we now bound $\|E_k\|$. To that end, in what follows we assume we are in the event $\{\tau > k\}$ where $k \geq k_0$. In this event, our strategy will be to bound the terms R_1 and R_2 in the following decomposition:

$$\begin{aligned}\|E_k\| &= \|(y_{k+1} - w_k)/\alpha_k\| \\ &\leq \underbrace{\|y_{k+1} - y_k - P_{T_k}(y_{k+1} - y_k)\|/\alpha_k}_{:=R_1} + \underbrace{\|P_{T_k}(y_{k+1} - y_k)/\alpha_k + \nabla f_{\mathcal{M}}(y_k) + P_{T_k}(v_k)\|}_{:=R_2}.\end{aligned} \tag{4.5.4}$$

In our bounds of these terms, we frequently use the following bound: there exists $C > 0$ such that

$$\|x_{k+1} - x_k\| \leq \alpha_k \|G_{\alpha_k}(x_k, v_k)\| \leq C(1 + \|v_k\|)\alpha_k. \tag{4.5.5}$$

We now bound R_1 and R_2 separately.

The following claim bounds R_1.

Claim: There exists $C > 0$ such that

$$R_1 1_{\tau>k} \leq C(1 + \|v_k\|)^2 \alpha_k 1_{\tau>k}. \tag{4.5.6}$$

Proof. We consider two cases. First suppose $x_{k+1} \in B_{2\delta}(\bar{x})$. Let $C > 0$ be a local Lipschitz constant of $\nabla P_{\mathcal{M}}$ and $P_{\mathcal{M}}$. Then it follows that vector $y_{k+1} - y_k = P_{\mathcal{M}}(x_{k+1}) - P_{\mathcal{M}}(x_k)$ is nearly tangent to the manifold at y_k:

$$\|y_{k+1} - y_k - P_{T_k}(y_{k+1} - y_k)\| \leq C\|y_{k+1} - y_k\|^2 \leq C^3\|x_{k+1} - x_k\|^2.$$

Thus, taking into account (4.5.5), we have for some $C > 0$, the bound:

$$R_1 \leq C(1 + \|v_k\|)^2 \alpha_k,$$

as desired. Now suppose that $x_{k+1} \notin B_{2\delta}(\bar{x})$. Therefore, there exists $C > 0$ such that

$$\|y_{k+1} - y_k - P_{T_k}(y_{k+1} - y_k)\| \leq 2\|y_{k+1} - y_k\| \leq C\|x_{k+1} - x_k\| \leq \frac{C^2}{\delta}\|x_{k+1} - x_k\|^2,$$

where the first inequality follows since $\|P_{T_k}\| \leq 1$ and the second and third inequalities follow from Claim 4.5.2. Thus taking into account (4.5.5), we again have for some $C > 0$, the bound:

$$R_1 \leq C(1 + \|v_k\|)^2 \alpha_k,$$

Thus, putting together both bounds on R_1, the result follows. □

The following claim bounds R_2.

Claim: There exists $C > 0$ such that

$$R_2 1_{\tau > k} \leq C(1 + \|v_k\|)^2(\operatorname{dist}(x_k, \mathcal{M}) + \alpha_k) 1_{\tau > k}. \tag{4.5.7}$$

Proof. To bound R_2, we first simplify:

$$\begin{aligned} R_2 &= \|P_{T_k}(y_{k+1} - y_k)/\alpha_k + \nabla_{\mathcal{M}} f(y_k) + P_{T_k}(v_k)\| \\ &\leq \|P_{T_k}(y_{k+1} - x_{k+1})/\alpha_k\| + \|P_{T_k}(x_k - y_k)/\alpha_k\| + \|P_{T_k}(x_{k+1} - x_k)/\alpha_k + \nabla_{\mathcal{M}} f(y_k) + P_{T_k}(v_k)\| \\ &\leq \|P_{T_k}(y_{k+1} - x_{k+1})/\alpha_k\| + C(1 + \|v_k\|)^2(\operatorname{dist}(x_k, \mathcal{M}) + \alpha), \end{aligned} \tag{4.5.8}$$

where the second inequality follows from by Assumption A and the inclusion $x_k - y_k \in N_{\mathcal{M}}(y_k)$, which implies that $P_{T_k}(x_k - y_k) = 0$. We now bound the term $\|P_{T_k}(y_{k+1} - x_{k+1})/\alpha_k\|$.

First suppose that $x_{k+1} \in B_{2\delta}(\bar{x})$ and note that $y_{k+1} \in B_{4\delta}(\bar{x}) \cap \mathcal{M} \subseteq \mathcal{U} \cap \mathcal{M}$. Let $C' > 0$ be a local Lipschitz constant of ∇P_M and $P_{\mathcal{M}}$. Then for some $C > 0$ larger than

C', we have

$$\begin{aligned}\|P_{T_k}(y_{k+1} - x_{k+1})/\alpha_k\| &\leq \|(P_{T_{k+1}} - P_{T_k})(y_{k+1} - x_{k+1})/\alpha_k\| \\ &\leq C'\|y_{k+1} - y_k\|\text{dist}(x_{k+1}, \mathcal{M})/\alpha_k \\ &\leq (C')^2\|x_{k+1} - x_k\|(\text{dist}(x_k, \mathcal{M}) + \|x_{k+1} - x_k\|)/\alpha_k \\ &\leq C^3(1 + \|v_k\|)\text{dist}(x_k, \mathcal{M}) + C^4(1 + \|v_k\|)^2\alpha_k,\end{aligned}$$

where the first inequality follows from $x_{k+1} - y_{k+1} \in N_{\mathcal{M}}(y_{k+1})$, which implies $P_{T_{k+1}}(y_{k+1} - x_{k+1}) = 0$; the second inequality follows from Lipschitz continuity of $\nabla P_{\mathcal{M}}(y) = P_{T_{\mathcal{M}(y)}}$ in y; the third inequality follows from Lipschitz continuity of $P_{\mathcal{M}}$ and Lipschitz continuity of $\text{dist}(\cdot, \mathcal{M})$; and the fourth inequality follows from (4.5.5). Plugging this bound into (4.5.8), yields that for some $C > 0$, we have

$$R_2 \leq C(1 + \|v_k\|)^2(\text{dist}(x_k, \mathcal{M}) + \alpha_k),$$

as desired.

Now suppose that $x_{k+1} \notin B_{2\delta}(\bar{x})$. Then, there exists $C > 0$ such that

$$\begin{aligned}\|P_{T_k}(y_{k+1} - x_{k+1})/\alpha_k\| &\leq \|P_{T_k}(y_{k+1} - x_k)\|/\alpha_k + \|P_{T_k}(x_k - x_{k+1})\|/\alpha_k \\ &\leq 2\delta/\alpha_k + \|x_k - x_{k+1}\|/\alpha_k \\ &\leq (1 + C)\|x_k - x_{k+1}\|/\alpha_k \\ &\leq \frac{(1 + C)C}{\delta\alpha_k}\|x_k - x_{k+1}\|^2 \\ &\leq \frac{(1 + C)C^3}{\delta}(1 + \|v_k\|)^2\alpha_k\end{aligned}$$

where first inequality follows from the triangle inequality; the second inequality follows since $x_k \in B_{2\delta}(\bar{x})$ and $y_{k+1} = \bar{x}$; the third and fourth third inequalities follow from Claim 4.5.2; and the fifth follows from (4.5.5). Thus, in this case, we find that there exists $C > 0$ with

$$R_2 \leq C(1 + \|v_k\|)^2(\text{dist}(x_k, \mathcal{M}) + \alpha_k).$$

Therefore, putting together both bounds on R_2, the result follows. □

Now we prove Part 2a. Beginning with subpart 2(a)i, we find that by Claim 4.5.2 and 4.5.2, we have that for some $C > 0$, the bound

$$\|E_k\|1_{\tau>k} \le R_1 1_{\tau>k} + R_2 1_{\tau>k} \le C(1 + \|\nu_k\|)^2(\mathrm{dist}(x_k, \mathcal{M}) + \alpha_k)1_{\tau>k}, \tag{4.5.9}$$

as desired. Turning to Part 2(a)ii, first note that that $\mathrm{dist}(x_k, \mathcal{M})1_{\tau>k} \le \delta$. Thus, the bound will follow if the conditional expectation of $(1 + \|\nu_k\|)^4$ is bounded whenever $x_k \in B_\delta(\bar{x})$. This holds by assumption, since

$$\mathbb{E}_k[\|\nu_k\|^4]1_{\tau>k} \le \sup_{x \in B_\delta(\bar{x})} q(x) < \infty.$$

Finally, we prove Part 2(a)iii. Again using the boundedness of the conditional fourth moment of $\|\nu_k\|1_{\tau>k}$, we find that there exists a $C > 0$ such that

$$\mathbb{E}_k[\|E_k\|^2 1_{\tau>k}] \le C\mathrm{dist}^2(x_k, \mathcal{M})1_{\tau>k} + C\alpha_k^2 1_{\tau>k}, \tag{4.5.10}$$

where the first inequality follows from Jensen's inequality and the second inequality follows from (4.5.9). Consequently, there exists $C' > 0$ such that

$$\mathbb{E}[\|E_k\|^2 1_{\tau>k}] = \mathbb{E}[\mathbb{E}_k\|E_k\|^2 1_{\tau>k}] \le C\mathbb{E}[\mathrm{dist}^2(x_k, \mathcal{M})1_{\tau>k}] + C\alpha_k^2 \le C'\alpha_k,$$

where the third inequality follows from Part 2a of Proposition 4.3.1. This prove Part 2a.

Now we prove Part 2b, beginning with Part 2(b)i. To that end, define $F_k = \frac{k^{\gamma-1}}{\log(k+1)^2}\|E_k\|1_{\tau>k}$. Recall that by the conditional Borel-Cantelli theorem (Lemma 7.1.2), the sequence F_k is summable whenever $\mathbb{E}_k[F_k]$ is summable. Thus, we first upper bound $\mathbb{E}_k[F_k]$ by a summable sequence: there exists $C > 0$ such that

$$\begin{aligned}\mathbb{E}_k[F_k] &\le C\frac{k^{\gamma-1}}{\log(k+1)^2}(\mathrm{dist}(x_k, \mathcal{M}) + \alpha_k)1_{\tau>k} \\ &\le C\frac{k^{\gamma-1}}{\log(k+1)^2}\mathrm{dist}(x_k, \mathcal{M})1_{\tau>k} + C\frac{c_2}{k\log(k+1)^2},\end{aligned}$$

where the first inequality follows from (4.5.10) and the second inequality follows by definition of α_k. By Part 1 of Proposition 4.3.1, it follows that we have upper bounded $\mathbb{E}_k[F_k]$ by a summable sequence. Therefore, it follows that F_k is summable, as desired. This proves part 2(b)i.

Now we prove part 2(b)ii. The conditional expectation is summable by Part 2(a)iii, since

$$\sum_{k=k_0}^{\infty} \frac{k^{\gamma-1}}{\log(k+1)^2}\mathbb{E}[\|E_k\|^2 1_{\tau>k}] \le C \sum_{k=k_0}^{\infty} \frac{k^{-1}}{\log(k+1)^2} < +\infty.$$

By conditional Borel-Cantelli theorem (Lemma 7.1.2), we also have that

$$\sum_{k=k_0}^{\infty} \frac{k^{\gamma-1}}{\log(k+1)^2}\|E_k\|^2 1_{\tau>k} < +\infty,$$

as desired.

Now we prove Part 2c. To that end, note that there exists $C > 0$ such that

$$\mathbb{E}[\alpha_k\|E_k\|1_{\tau>k}] = \mathbb{E}[\alpha_k\mathbb{E}_k[\|E_k\|1_{\tau>k}]] \le C\mathbb{E}[\alpha_k \mathrm{dist}(x_k, \mathcal{M})1_{\tau>k} + \alpha_k^2 1_{\tau>k}].$$

where the inequality follows from (4.5.10). Thus, the result follows by Part 2b of Proposition 4.3.1.

4.6 Proofs of the main theorems

In this section, we prove the remaining theorems.

4.6.1 Proof of Theorem 4.4.1: nonconvergence of stochastic process

We begin by recalling and slightly reframing Proposition 3 in [61]. This result provides a Lyapunov function, which we will use to show that each local process Y_k escapes a local neighborhood of each $p \in S$.

Proposition 4.6.1 (Lyapunov Function)**.** *Fix $p \in \mathbb{R}^d$ and suppose $F\colon \mathbb{R}^d \to \mathbb{R}^d$ is a C^2 mapping that is zero at p and has a symmetric Jacobian $\nabla F(p)$. Suppose that $\nabla F(p)$ has at least one positive eigenvalue and let W denote the subspace of eigenvectors of $\nabla F(p)$ with positive eigenvalues. Then, there exists a matrix $A \in \mathbb{R}^{d\times d}$ with* $\mathrm{range}(A^T) = W$*, a ball $\mathcal{B}$ centered at p, and a C^2 mapping $\Phi\colon \mathcal{B} \to \mathbb{R}^d$ with $\Phi(p) = p$ and $\nabla\Phi(p) = I_d$ such that the function $\eta\colon \mathcal{B} \to \mathbb{R}$ defined as*

$$\eta(v) = \|A(\Phi(v) - p)\|_2$$

satisfies the following condition: There exists $c, c' > 0$ such that

$$\eta(v + \epsilon F(v)) \geq (1 + c\epsilon)\eta(v) - c'\epsilon^2 \qquad \textit{for } v \textit{ in } \mathcal{B} \textit{ and all sufficiently small } \epsilon.$$

In particular, we have

$$\eta'(v; F(v)) \geq c\eta(v) \qquad \textit{for all } v \in \mathcal{B}.$$

Turning to the proof of Theorem 4.4.1, we begin with a covering argument: For any $p \in S$, choose ϵ_p small enough that both the conditions of Theorem 4.4.1 and Proposition 4.6.1 hold in $B_{\epsilon_p}(p)$ for F_p. Let $\delta_p \leq \epsilon_p$ and $c_3, c_4, c_5 > 0$ be the associated constants. Clearly, the union $\cup_{p\in S} B_{\delta_p}(p)$ is an open cover of set S, and therefore there exists a countable index set $\Lambda \subset S$ such that $S \subset \cup_{p\in\Lambda} B_{\delta_p}(p)$. Therefore, to prove Theorem 4.4.1, it suffices to show that

$$P\left(X_k \in B_{\delta_p}(p), \forall k \geq k_0\right) = 0 \qquad \text{for all } k_0 \geq K_p. \tag{4.6.1}$$

To this end, fix $p \in \Lambda$ and $k_0 \geq K_p$. Let $F = F_p$ denote the local mapping in Condition 1 of Theorem 4.4.1. In addition, let $\eta = \eta_p$, denote the mapping associated to F, guaranteed to exist by Theorem 4.6.1.[3] Furthermore, recall the stopping time $\tau_{k_0} = \tau_{k_0,\delta_p}(p)$, defined as $\tau_{k_0,\delta_p}(p) = \inf\{k \geq k_0 \colon X_k \notin B_{\delta_p}(p)\}$. Note that (4.6.1) holds if $P(\tau_{k_0} = \infty) = 0$.

Our strategy is as follows. We first prove that on the event $\{\tau_{k_0} = \infty\}$, we have $\eta(Y_k) \to 0$ almost surely. Then we show that $P(\{\tau_{k_0} = \infty\} \cap \{\eta(Y_k) \to 0\}) = 0$. This will imply that $P(\{\tau_{k_0} = \infty\}) = 0$ and the proof will be complete. These two claims are subjects of the following two subsections.

4.6.1.1 Claim: On the event $\{\tau_{k_0} = \infty\}$, we have $\eta(Y_k) \to 0$

To prove this claim, note that the following hold for almost all sample paths in the event $\{\tau_{k_0} = \infty\}$:

1. The sequence Y_k is bounded.
2. Define $\beta_k = \sum_{i=0}^{k-1} \alpha_i$. Then for each $T > 0$, the limit holds:

$$\lim_{n\to\infty} \left(\sup_{k \colon 0 \leq \beta_k - \beta_n \leq T} \left\| \sum_{i=n}^{k-1} \alpha_i \cdot (\xi_i + E_k) \right\| \right) = 0. \tag{4.6.2}$$

   Indeed, note that by Condition 3b of the Theorem, it suffices to show $M_k = \sum_{i=0}^{k} \alpha_i \xi_i$ converges almost surely, since then it is a Cauchy sequence. To prove that M_k converges, note that $\sum_i \alpha_i^2 < \infty$ and $\limsup \mathbb{E}[\|\xi_k\|^2 \mid \mathcal{F}_k] < \infty$, so M_k is a martingale. Moreover,

$$\sup_{k\geq 0} \mathbb{E}\left[\|M_k\|\right]^2 \leq \sup_{k\geq 0} \mathbb{E}\left[\|M_k\|^2\right] \leq c_3^{\frac{1}{2}} \sum_{i\geq 0} \alpha_i^2 < \infty. \tag{4.6.3}$$

   Standard martingale theory then shows that M_k converges almost surely (Theorem 4.2.11 in [85]). Therefore, (4.6.2) holds almost surely.

[3] Note that strictly speaking we should extend F to all $\mathbb{R}^d$, for example, by a partition of unity [84, Lemma 2.26]. Since the argument that follows is local, we omit this discussion for simplicity.

These conditions match those of [81, Theorem 1.2]. Consequently, by this result it holds that the set of limit points of Y_k is almost surely invariant under the mapping $\Theta_t \colon B_{\epsilon_p/2}(p) \to \mathbb{R}^d$, defined as the time-$t$ map of the ODE $\dot{\gamma}(t) = F(\gamma(t))$. Thus, for any x' in limit set of Y_k, we have $\Theta_t(x') \in \overline{B_{\epsilon_p/2}(p)}$ for all $t \geq 0$. Consequently, by Proposition 4.6.1, we have

$$\eta'(\Theta_t(x'); F(\Theta_t(x'))) \geq c\eta(\Theta_t(x')) \qquad \text{for all } t \geq 0. \tag{4.6.4}$$

Therefore, by integrating η' with respect to t, we have for all $t \geq 0$, the bound

$$\eta(\Theta_t(x')) = \eta(\Theta_0(x')) + \int_0^t \eta'(\Theta_s(x'); F(\Theta_s(x')))ds \geq \eta(\Theta_0(x')) + \int_0^t c\eta(\Theta_s(x'))ds.$$

Thus, by Gronwall's inequality [86] it holds that

$$\eta(\Theta_t(x')) \geq e^{ct}\eta(\Theta_0(x')) = e^{ct}\eta(x') \qquad \text{for all } t \geq 0.$$

Now observe that since $\Theta_t(x') \in \overline{B_{\epsilon_p/2}(p)}$, the quantity $\eta(\Theta_t(x'))$ is bounded for all $t \geq 0$. Consequently, we must have $\eta(x') = 0$. Thus, we have shown that for all limits points x' of Y_k, we have $\eta(x') = 0$. Since η is continuous in $\overline{B_{\epsilon_p/2}(p)}$, we must therefore have $\eta(Y_k) \to 0$.

4.6.1.2 Claim: We have $P(\{\tau_{k_0} = \infty\} \cap \{\eta(Y_k) \to 0\}) = 0$.

We begin by stating the following straightforward extension of [83, Theorem 4.1].

Lemma 4.6.1. *Let $\{\zeta_k\}_k$ be a nonnegative sequence of random variables adapted to a filtration $\{\mathcal{F}_k\}$ satisfying the following recurrence almost surely on an $\mathcal{F}_\infty$-measurable set Ω_0:*

$$\zeta_{k+1} \geq \zeta_k + \alpha_k(e_{k+1} + r_{k+1} + \hat{r}_{k+1}) \qquad \textit{for all } k \geq k_0.$$

where $\{\alpha_k\}$ is a square-summable, but not summable sequence. Assume that $\{e_k\}_k$, $\{r_k\}$, and $\{\hat{r}_k\}_k$ are $\mathcal{F}_k$ measurable and satisfy

$$\mathbb{E}[e_{k+1} \mid \mathcal{F}_k] = 0; \qquad \sum_{k=1}^{\infty} r_k^2 < +\infty$$

$$\limsup_k \mathbb{E}[e_{k+1}^2 \mid \mathcal{F}_k] < \infty \qquad \liminf_k \mathbb{E}[|e_{k+1}| \mid \mathcal{F}_k] > 0,$$

almost surely on Ω_0. Assume that for $n \geq k_0$, we have

$$\mathbb{E}\left[1_{\Omega_0} \sum_{k=n}^{\infty} \alpha_k |\hat{r}_{k+1}|\right] = O\left(\sum_{k=n}^{\infty} \alpha_k^2\right).$$

Then we have $P(\Omega_0 \cap \{\zeta_k \to 0\}) = 0$.

Proof. Without loss of generality we may assume $k_0 = 0$. Following [83, Theorem 4.1] (itself based on [80, Page 401]) it suffices to work in the case where there exist fixed constants μ and $C > 0$ such that almost surely on the whole probability space, we have

$$\mathbb{E}[e_{k+1} \mid \mathcal{F}_k] = 0 \qquad \text{and} \qquad \limsup_k \mathbb{E}[e_{k+1}^2 \mid \mathcal{F}_k] < C$$
$$\liminf_k \mathbb{E}[|e_{k+1}| \mid \mathcal{F}_k] > \mu > 0 \qquad \text{and} \qquad \sum_{k=1}^{\infty} r_k^2 < C.$$

Now define the nonnegative residual sequence:

$$\alpha_k U_{k+1} = \zeta_{k+1} - \zeta_k - \alpha_k(e_{k+1} + r_{k+1} + \hat{r}_{k+1})$$

Notice that for all $k \geq 0$, we have

$$\zeta_k = \left[\zeta_0 + \sum_{j=0}^{k} \alpha_j(e_{j+1} + r_{j+1} + \hat{r}_{j+1} + U_{j+1})\right] \qquad \text{on } G := \Omega_0 \cap \{\zeta_k \to 0\}.$$

Therefore, on G, we have

$$-\zeta_0 = \left[\sum_{j=0}^{\infty} \alpha_j(e_{j+1} + r_{j+1} + \hat{r}_{j+1} + U_{j+1})\right].$$

Then as argued the proof of [83, Theorem 4.1] it suffices by Theorem A of [80] (included as Lemma 7.1.3 in the Appendix) to show that

$$\mathbb{E}\left[1_G \sum_{k=n}^{\infty} \alpha_k |U_{k+1} + \hat{r}_{j+1}|\right] = o\left(\left(\sum_{k=n}^{\infty} \alpha_k^2\right)^{1/2}\right),$$

Clearly, it suffices to bound the series $\mathbb{E}\left[1_G \sum_{j=K}^{\infty} \alpha_j U_{j+1}\right]$ (which consists of nonnegative terms), since by assumption, we have

$$\mathbb{E}\left[1_G \sum_{k=n}^{\infty} \alpha_k |\hat{r}_{j+1}|\right] = O\left(\sum_{j=n}^{\infty} \alpha_j^2\right) = o\left(\left(\sum_{j=n}^{\infty} \alpha_j^2\right)^{1/2}\right).$$

To that end, note that for all $k, n \geq 0$, we have

$$\zeta_{n+k} = \left[\zeta_n + \sum_{j=n}^{n+k} \alpha_j (e_{j+1} + r_{j+1} + \hat{r}_{j+1} + U_{j+1})\right]$$

Hence on G, we may let k tend to infinity, yielding:

$$-\zeta_n = \sum_{j=n}^{\infty} \alpha_j (e_{j+1} + r_{j+1} + \hat{r}_{j+1} + U_{j+1}).$$

Thus, on the event G, we have

$$\sum_{j=n}^{\infty} \alpha_j U_{j+1} = -\zeta_n - \sum_{j=n}^{\infty} \alpha_j (e_{j+1} + r_{j+1} + \hat{r}_{j+1})$$

Therefore, we find that

$$\begin{aligned}\mathbb{E}\left[1_G \sum_{j=n}^{\infty} \alpha_j U_{j+1}\right] &\leq -\mathbb{E}\left[\zeta_n\right] - \mathbb{E}\left[1_G \sum_{j=n}^{\infty} \alpha_j (e_{j+1} + r_{j+1} + \hat{r}_{j+1})\right] \\ &\leq \left|\mathbb{E}\left[1_G \sum_{j=n}^{\infty} \alpha_j (e_{j+1} + r_{j+1})\right]\right| + o\left(\left(\sum_{j=n}^{\infty} \alpha_j^2\right)^{1/2}\right).\end{aligned}$$

where the second inequality follows from nonnegativity of ζ_n and our assumptions on $\hat{r}_{j+1}$. Thus, to complete the bound of $\mathbb{E}[1_G \sum_{j=K}^{\infty} \alpha_j U_{j+1}]$ we must show that

$$\left|\mathbb{E}\left[1_G \sum_{j=n}^{\infty} \alpha_j (e_{j+1} + r_{j+1})\right]\right| = o\left(\left(\sum_{j=n}^{\infty} \alpha_j^2\right)^{1/2}\right).$$

The above bound follows by the exact same argument as [83, Theorem 4.1], which we reproduce for completeness: First let $G_n = \mathbb{E}[1_G \mid \mathcal{F}_n]$, recall that G is $\mathcal{F}_\infty$ measurable and that G_n converges to 1_G almost surely in L^p for every $p \geq 1$, e.g., $\mathbb{E}[(G_n - 1_G)^2] \to 0$. Turning to the bound, we have

$$\begin{aligned}
&\left|\mathbb{E}\left[1_G \sum_{j=n}^{\infty} \alpha_j(e_{j+1} + r_{j+1})\right]\right| \\
&\leq \left|\mathbb{E}\left[(1_G - G_n) \sum_{j=n}^{\infty} \alpha_j(e_{j+1} + r_{j+1})\right]\right| + \left|\mathbb{E}\left[G_n \sum_{j=n}^{\infty} \alpha_j(e_{j+1} + r_{j+1})\right]\right| \\
&\leq \mathbb{E}[(1_G - G_n)^2]^{1/2} \underbrace{\left(\mathbb{E}\left[\left(\sum_{j=n}^{\infty} \alpha_j(e_{j+1} + r_{j+1})\right)^2\right]\right)^{1/2}}_{=:R_1} + \underbrace{\mathbb{E}\left[\sum_{j=n}^{\infty} \alpha_j |r_{j+1}|\right]}_{=:R_2}.
\end{aligned}$$

The proof will be complete if $R_1 = O\left(\sum_{j=n}^{\infty} \alpha_j^2\right)^{1/2}$ and $R_2 = o\left(\sum_{j=n}^{\infty} \alpha_j^2\right)^{1/2}$. Let us first bound R_2:

$$R_2 \leq \left(\sum_{j=n}^{\infty} \alpha_j^2\right)^{1/2} \mathbb{E}\left[\sum_{j=n}^{\infty} r_{j+1}^2\right]^{1/2} = o\left(\left(\sum_{j=n}^{\infty} \alpha_j^2\right)^{1/2}\right),$$

where the last inequality follows from the bound $\sum_{k=1}^{\infty} r_{k+1}^2 < C$. Now we bound R_1:

$$\begin{aligned}
R_1 &\leq \left(\mathbb{E}\left[\left(\sum_{j=n}^{\infty} \alpha_j e_{j+1}\right)^2\right]\right)^{1/2} + \left(\mathbb{E}\left[\left(\sum_{j=n}^{\infty} \alpha_j r_{j+1}\right)^2\right]\right)^{1/2} \\
&\leq \left(\mathbb{E}\left[\sum_{j=n}^{\infty} \alpha_j^2 \mathbb{E}[e_{j+1}^2 \mid \mathcal{F}_k]\right]\right)^{1/2} + \left(\sum_{j=n}^{\infty} \alpha_j^2\right)^{1/2} \left(\mathbb{E}\left[\sum_{j=n}^{\infty} r_{j+1}^2\right]\right)^{1/2} = O\left(\left(\sum_{j=n}^{\infty} \alpha_j^2\right)^{1/2}\right).
\end{aligned}$$

Therefore, the proof is complete. □

Now we apply the above Lemma. To that end, we state a few simplifications and facts to be used below. First, throughout the proof, we let C be a positive constant that changes from line to line. Second, we simplify notation and let τ denote $\tau_{k_0,\delta}$. Third, we recall the bound $\frac{c_1}{k^\gamma} \leq \alpha_k \leq \frac{c_2}{k^\gamma}$. Fourth, the function η is weakly convex and Lipschitz continuous on $B_{\epsilon_p}(p)$. Fifth, the Jacobian $\nabla\Phi$ is $\mathrm{Lip}_{\nabla\Phi}$-Lipschitz in $B_{\epsilon_p}(p)$. Sixth, we note that for sufficiently large k, we have the following on $\{\tau = \infty\}$: $Y_k + \alpha_k F(Y_k) \in B_{\epsilon_p}(p)$. We may assume without loss of generality that these assertions hold for all $k \geq 1$. Finally, we note that by shrinking ϵ_p, if necessary, we can assume that on the event $\{\tau = \infty\}$, we have

$$s_{\min}(A) \liminf_k \inf_{w \in W \cap \mathbb{S}^{d-1}} \mathbb{E}[|\langle w, \xi_k \rangle| \mid \mathcal{F}_k] - \epsilon_p \limsup_k \mathbb{E}[\|\xi_k\| \mid \mathcal{F}_k]\|A\|_{\mathrm{op}} \mathrm{Lip}_{\nabla\Phi}$$
$$\geq c_4 s_{\min}(A) - \epsilon_p c_3^{1/4} \|A\|_{\mathrm{op}} \mathrm{Lip}_{\nabla\Phi} > 0 \quad (4.6.5)$$

where c_4 and c_3 are independent of ϵ_p and δ_p, A is defined in Proposition 4.6.1, and $s_{\min}(A)$ denotes the minimal nonzero singular value of A.

Now let $s \colon B_{\epsilon_p}(p) \to \mathbb{R}^d$ be a selection of $\partial\eta$ defined as follows: for all $y \in B_{\epsilon_p}(p)$,

- If $\eta(y) \neq 0$, then η is differentiable at Y, so set $s(y) = \nabla\eta(y)$.
- If $\eta(y) = 0$, then η is nondifferentiable, so we choose subgradient

  $$s(Y) = \nabla\Phi(y)^\top A^\top u \in \partial\eta(y)$$

  where $u \in \mathbb{S}^{d-1}$ satisfies $\|A^\top u\| = \|A\|_{\mathrm{op}} > 0$.

Next, consider the event $\Omega_0 = \{\tau = \infty\}$. Then by the boundedness of $s(Y_k + \alpha_k F(Y_k))$ and the weak convexity of η on $B_{\epsilon_p}(p)$, there exists $C > 0$ such that

$$\begin{aligned}\eta(Y_{k+1}) &\geq \eta(Y_k + \alpha_k F(Y_k)) + \langle s(Y_k + \alpha_k F(Y_k)), \alpha_k E_k + \alpha_k \xi_k \rangle - C\|\alpha_k E_k + \alpha_k \xi_k\|^2 \\ &\geq \eta(Y_k + \alpha_k F(Y_k)) + \langle s(Y_k + \alpha_k F(Y_k)), \alpha_k \xi_k \rangle - C\|\alpha_k E_k + \alpha_k \xi_k\|^2 - C\alpha_k\|E_k\| \\ &\geq (1 + c\alpha_k)\eta(Y_k) + \langle s(Y_k + \alpha_k F(Y_k)), \alpha_k \xi_k \rangle - C\|\alpha_k E_k + \alpha_k \xi_k\|^2 - C\alpha_k\|E_k\| - C\alpha_k^2.\end{aligned} \quad (4.6.6)$$

Now define four sequences:

$$\zeta_k := \eta(Y_k); \quad e_{k+1} := \langle s(Y_k + \alpha_k F(Y_k)), \xi_k \rangle; \quad r_{k+1} := -C\alpha_k \left(1 + \|E_k + \xi_k\|^2\right); \quad \hat{r}_{k+1} := -C\|E_k\|$$

and observe that on Ω_0, we have

$$\zeta_{k+1} \geq \zeta_k + \alpha_k(e_{k+1} + r_{k+1} + \hat{r}_{k+1}).$$

Now we must verify the assumptions of the Lemma. We begin with $\hat{r}_{k+1}$. To that end, observe that

$$\mathbb{E}\left[1_{\Omega_0}\sum_{k=n}^{\infty}\alpha_k\hat{r}_{k+1}\right] = O\left(\sum_{k=n}^{\infty}\alpha_k^2\right),$$

by our assumption on $\|E_k\|$. Next we prove square summability of r_{k+1} on Ω_0: Indeed, observe

$$r_{k+1}^2 \leq C\alpha_k^2(\|\xi_k\|^4 + \|E_k\|^4 + 1).$$

Moreover both $\limsup_k \mathbb{E}_k[\|\xi_k\|^4 \mid \mathcal{F}_k] < \infty$ and $\limsup_k \mathbb{E}_k[\|E_k\|^4 \mid \mathcal{F}_k] < \infty$ are bounded on Ω_0. Therefore, by conditional Borel-Cantelli Lemma 7.1.2, we have

$$\sum_{k=1}^{\infty} r_{k+1}^2 < +\infty.$$

almost surely on Ω_0.

Finally we prove that e_k has the desired properties. First note that we have

$$\mathbb{E}[e_{k+1} \mid \mathcal{F}_k] = 0 \qquad \text{and} \qquad \limsup_k \mathbb{E}[e_{k+1}^2 \mid \mathcal{F}_k] < \infty.$$

on Ω_0. Indeed, this follows since $\limsup_k \mathbb{E}[\|\xi_k\|^4 \mid \mathcal{F}_k] < \infty$ almost surely and and $Y_k + \alpha_k F(Y_k) \in B_{\epsilon_p}(p)$ on Ω_0. Next, since η is globally Lipschitz on $B_{\epsilon_p}(p)$, we have that $s(Y_k + \alpha_k F(Y_k))$ is uniformly bounded. Thus,

$$\limsup_k \mathbb{E}[e_{k+1}^2 \mid \mathcal{F}_k] \leq \limsup_k \mathbb{E}[\|s(Y_k + \alpha_k F(Y_k))\|^2\|\xi_k\|^2 \mid \mathcal{F}_k] < \infty,$$

on Ω_0, as desired.

Now we prove that $\liminf \mathbb{E}[|e_{k+1}| \mid \mathcal{F}_k]$ is positive on Ω_0. To that end, recall that the mapping Φ satisfies $\nabla\Phi(p) = I_d$. Turning to the proof, there are two cases to consider.

First suppose that $\eta(Y_k + \alpha_k F(Y_k)) \neq 0$. Then η is differentiable at $Y_k + \alpha_k F(Y_k)$. Now define $u_k := \frac{A(\Phi(Y_k+\alpha_k F(Y_k))-p)}{\|A(\Phi(Y_k+\alpha_k F(Y_k))-p)\|}$ and note that

$$\begin{aligned} s(Y_k + \alpha_k F(Y_k)) = \nabla\eta(Y_k + \alpha_k F(Y_k)) &= \nabla\Phi(Y_k + \alpha_k F(Y_k))^\top A^\top u_k \\ &= A^\top u_k + (\nabla\Phi(Y_k + \alpha_k F(Y_k)) - \nabla\Phi(p))^\top A^\top u_k \\ &\in A^\top u_k + \epsilon_p \|A\|_{\mathrm{op}} \mathrm{Lip}_{\nabla\Phi} B_1(0), \end{aligned}$$

where the inclusion follows since $Y_k + \alpha_k F(Y_k) \in B_{\epsilon_p}(p)$. Let $s_{\min}(A)$ denote the minimal nonzero singular value of A and notice that since $u_k \in \mathbb{S}^{d-1} \cap \mathrm{range}(A)$, we have that $w_k := A^T u_k$ satisfies and

$$w_k \in W \qquad \text{and} \qquad \|w_k\| \geq s_{\min}(A) > 0.$$

Therefore, it follows that on the event Ω_0, we have

$$\begin{aligned} \mathbb{E}[|e_{k+1}| \mid \mathcal{F}_k] &= \mathbb{E}[|\langle s(Y_k + \alpha_k F(Y_k)), \xi_k\rangle| \mid \mathcal{F}_k] \\ &\geq \mathbb{E}[|\langle w_k, \xi_k\rangle| \mid \mathcal{F}_k] - \epsilon_p \mathbb{E}[\|\xi_k\| \mid \mathcal{F}_k] \|A\|_{\mathrm{op}} \mathrm{Lip}_{\nabla\Phi} \\ &\geq s_{\min}(A) \inf_{w \in W \cap \mathbb{S}^{d-1}} \mathbb{E}[|\langle w, \xi_k\rangle| \mid \mathcal{F}_k] - \epsilon_p \mathbb{E}[\|\xi_k\| \mid \mathcal{F}_k] \|A\|_{\mathrm{op}} \mathrm{Lip}_{\nabla\Phi} \end{aligned}$$

We now consider the case $\eta(Y_k + \alpha_k F(Y_k)) = 0$. In this case, there exists $u_k \in \mathbb{S}^{d-1}$ such that $\|A^\top u_k\| = \|A\|_{\mathrm{op}}$ and

$$s(Y_k + \alpha_k F(Y_k)) = \nabla\Phi(Y_k + \alpha_k F(Y_k))^\top A^\top u_k \in A^\top u_k + \epsilon_p \|A\|_{\mathrm{op}} \mathrm{Lip}_{\nabla\Phi} B_1(0),$$

Recall $\mathrm{range}(A^\top) = W$. Thus, we have that the vector $w_k := A^\top u_k$ is in W and $\|w_k\| = \|A\|_{\mathrm{op}} > 0$. Thus, for all $v \in \mathbb{R}^d$, we have

$$|\langle s(Y_k + \alpha_k F(Y_k)), v\rangle| = \langle \nabla\Phi(Y_k + \alpha_k F(Y_k))^\top A^\top u_k, v\rangle \geq \langle w_k, v\rangle - \epsilon_p \mathrm{Lip}_{\nabla\Phi} \|A\|_{\mathrm{op}} \|v\|.$$

Taking $v = \xi_k$, we obtain

$$\mathbb{E}[|\langle s(Y_k + \alpha_k F(Y_k)), \xi_k\rangle| \mid \mathcal{F}_k] \geq \|A\|_{\mathrm{op}} \inf_{w \in W \cap \mathbb{S}^{d-1}} \mathbb{E}[|\langle w, \xi_k\rangle| \mid \mathcal{F}_k] - \epsilon_p \mathbb{E}[\|\xi_k\| \mid \mathcal{F}_k] \|A\|_{\mathrm{op}} \mathrm{Lip}_{\nabla\Phi}$$

Thus, putting both cases together, we find that on the event Ω_0, we have

$$\liminf_k \mathbb{E}[|e_{k+1}| \mid \mathcal{F}_k] \geq s_{\min}(A) \liminf_k \inf_{w \in W \cap \mathbb{S}^{d-1}} \mathbb{E}[|\langle w, \xi_k \rangle| \mid \mathcal{F}_k] - \epsilon_p \limsup_k \mathbb{E}[\|\xi_k\| \mid \mathcal{F}_k] \|A\|_{\mathrm{op}} \mathrm{Lip}_{\nabla\Phi} > 0,$$

where the last inequality follows from (4.6.5).

4.6.2 Proof of Theorem 4.4.2: nonconvergence to saddle points

In this section, prove Theorem 4.4.2 by verifying that the iterates $\{x_k\}_{k \in \mathbb{N}}$ satisfy the conditions of Theorem 4.4.1. We begin with some notation. To this end, observe that there exists $\epsilon > 0$ such that the function $f_{\mathcal{M}} \colon B_{2\epsilon}(\bar{x}) \to \mathbb{R}$, defined as the composition

$$f_{\mathcal{M}} := f \circ P_{\mathcal{M}} \tag{4.6.7}$$

is C^2 and satisfies

$$\nabla f_{\mathcal{M}}(x) = \nabla_{\mathcal{M}} f(x) \qquad \text{and} \qquad \nabla^2 f_{\mathcal{M}}(x) = \nabla^2_{\mathcal{M}} f(x)$$

for all $x \in B_{2\epsilon}(\bar{x}) \cap \mathcal{M}$. Moreover, we may also assume that the projection map $P_{\mathcal{M}} \colon B_{2\epsilon}(\bar{x}) \to \mathbb{R}^d$ is C^2, in particular, Lipschitz with Lipschitz Jacobian. Throughout the proof, we assume that $\delta \leq \epsilon/4$ is small enough that conclusions of Propositions 4.3.1 and 4.3.2 are valid; we shrink δ several further times throughout the proof. In addition, we let C denote a constant depending on k_0 and δ, which may change from line to line.

Now, denote stopping time (4.3.1) by $\tau := \tau_{k_0,\delta}$ and the noise bound by $Q := \sup_{x \in B_\delta(\bar{x})} q(x)$. Observe that by Proposition 4.3.2, the shadow sequence y_k satisfies $y_k \in B_{4\delta}(x_k) \cap \mathcal{M} \subseteq B_\epsilon(\bar{x}) \cap \mathcal{M}$ and recursion holds:

$$y_{k+1} = y_k - \alpha_k \nabla f_{\mathcal{M}}(y_k) - \alpha_k P_{T_{\mathcal{M}}(y_k)}(\nu_k) + \alpha_k E_k.$$

In addition, defining

$$f^* := \inf_{x \in B_\epsilon(\bar{x})} f_{\mathcal{M}}(x),$$

we have the bound $f^* 1_{\tau > k} \leq f(y_k) 1_{\tau > k}$ for all k. We now turn to the proofs.

To that end, fix a point $p \in S$ with associated manifold $\mathcal{M}$ and neighborhood $\mathcal{U}$. Let ϵ_p be small enough that $B_{\epsilon_p}(\bar{x}) \subseteq \mathcal{U}$ and define the C^2 mapping $F_p \colon B_{\epsilon_p}(p) \to \mathbb{R}^d$ by:

$$F_p(y) = -\nabla f_{\mathcal{M}}(y),$$

where $f_{\mathcal{M}} := f \circ P_{\mathcal{M}}$. Note that the mapping F is indeed C^2, since $\mathcal{M}$ is a C^4 manifold, and hence, $f_{\mathcal{M}}$ is C^3. Moreover, since $\nabla F(p) = -\nabla^2_{\mathcal{M}} f(p)$, the mapping F_p has at least one eigenvector with positive eigenvalue. In addition, the subspace W_p spanned by such eigenvectors is contained in $T_{\mathcal{M}}(p)$.

Turning to the proof, define $X_k = x_k$ for all $k \geq 1$. We now construct the sequences Y_k, ξ_k, and E_k and show they satisfy the assumptions of the theorem. Beginning with Y_k, recall that by Proposition 4.3.2, for all $k \geq 1$ and all sufficiently small $\delta > 0$, the sequence

$$Y_k := \begin{cases} P_{\mathcal{M}}(X_k) & \text{if } x_k \in B_{2\delta}(\bar{x}) \\ p & \text{otherwise.} \end{cases}, \tag{4.6.8}$$

satisfies $Y_k \in B_{4\delta}(\bar{x}) \cap \mathcal{M}$ and the recursion

$$Y_{k+1} = Y_k - \alpha_k \nabla f_{\mathcal{M}}(y_k) - \alpha_k \xi_k + \alpha_k E_k \qquad \text{for all } k \geq 1.$$

where $\xi_k := P_{T_{\mathcal{M}(Y_k)}}(\nu_k)$ and E_k is an error sequence. Moving to E_k, let us show that the error sequence satisfies the assumptions of the theorem. To that end, Proposition 4.3.2 shows that for δ sufficiently small, there exists $C > 0$ such that for all $n \geq k_0$, we have

$$\mathbb{E}\left[1_{\tau_{k_0,\delta} = \infty} \sum_{k=n}^{\infty} \alpha_k \|E_k\|\right] \leq C \sum_{k=n}^{\infty} \alpha_k^2.$$

Moreover, by the Part 2(a)i from Proposition 4.3.2, the sequence $\|E_k\| 1_{\tau_{k_0,\delta} > k}$ is bounded above by a bounded sequence that almost surely converges to zero:

$$\|E_k\| 1_{\tau_{k_0,\delta} > k} \leq C(1 + \|\nu_k\|)^2 (\text{dist}(x_k, \mathcal{M}) + \alpha_k) 1_{\tau_{k_0,\delta} > k} \leq C(1 + r)^2 (\delta + \alpha_k),$$

Thus, on the event $\{\tau_{k_0,\delta} = \infty\}$, we have

$$\limsup_k 1_{\Omega_0}\mathbb{E}[\|E_k\|^4 \mid \mathcal{F}_k] \leq \limsup_k \mathbb{E}[\|E_k\|^4 1_{\tau_{k_0,\delta}>k} \mid \mathcal{F}_k] \leq \left(C(1+r)^2(\delta + \alpha_k)\right)^4.$$

Therefore, Y_k and E_k satisfy the conditions 1 and 3 of Theorem 4.4.1 for all sufficiently small δ_p satisfying $\delta_p \leq \epsilon_p/8$.

To conclude the proof, we now show that Condition 2 of Theorem 4.4.1 is satisfied. To that end, clearly $\|\xi_k\| = \|P_{T_k}(\nu_k)\| \leq r =: c_3$ for all $k \geq k_0$. In addition, we have that

$$\mathbb{E}\left[\xi_k \mid \mathcal{F}_k\right] = P_{T_k}(\mathbb{E}\left[\nu_k \mid X_{k_0}, \ldots, X_k\right]) = 0.$$

Indeed, this follows from two facts: first Y_k is a measurable function of X_k; and second the noise sequence ν_k is mean zero and independent of $X_{k_0}, \ldots, X_k$. Finally, we must show that ξ_k has positive correlation with the unstable subspace W_p.

To prove correlation with the unstable subspace, recall that there exists $C' > 0$ such that the mapping $x \mapsto P_{T_{\mathcal{M}}(x)}$ is C'-Lipschitz mapping on $\mathcal{M} \cap B_{\epsilon_p}(p)$. In addition, we have that $W_p \subseteq T_{\mathcal{M}}(p)$. Therefore, since $Y_k \in \mathcal{M} \cap B_{\epsilon_p}(p)$ for all $k \geq k_0$, we have the following bound for all $w \in W \cap \mathbb{S}^{d-1}$:

$$\begin{aligned}
\mathbb{E}[|\langle \xi_k, w\rangle| \mid \mathcal{F}_k] &= \mathbb{E}[|\langle \nu_k, P_{T_{\mathcal{M}(Y_k)}}w\rangle| \mid \mathcal{F}_k] \\
&\geq \mathbb{E}[|\langle \nu_k, w\rangle| \mid \mathcal{F}_k] - r\|(P_{T_{\mathcal{M}(Y_k)}} - P_{T_{\mathcal{M}(p)}})w\| \\
&\geq rc_d - rC'\|Y_k - p\|,
\end{aligned}$$

where c_d is a constant dependent only on d since $\nu_k \sim \mathrm{Unif}(B_r(0))$. By slightly shrinking ϵ_p if needed, we can ensure that $\inf_{x \in B_{\epsilon_p}(p)}\{rc_d - rC'\|x - p\|\} > (1/2)rc_d =: c_4$, as desired.

CHAPTER 5

ASYMPTOTIC NORMALITY AND OPTIMALITY IN STOCHASTIC NONSMOOTH/CONSTRAINED OPTIMZATION

5.1 Introduction

Polyak and Juditsky [15] famously showed that the stochastic gradient method for minimizing smooth and strongly convex functions enjoys a central limit theorem: the error between the running average of the iterates and the minimizer, normalized by the square root of the iteration counter, converges to a normal random vector. Moreover, the asymptotic covariance matrix is in a precise sense "optimal" among any estimation procedure. A long-standing open question is whether similar guarantees – asymptotic normality and optimality – exist for nonsmooth optimization and, more generally, for equilibrium problems. In this part, we obtain such guarantees under mild conditions that hold both in concrete circumstances (e.g. nonlinear programming) and under generic linear perturbations.

The types of problems we will consider are best modeled as stochastic variational inequalities. Setting the stage, consider the task of finding a solution $x^\star$ of the inclusion

$$0 \in \mathop{\mathbb{E}}_{z\sim\mathcal{P}}[A(x,z)] + N_{\mathcal{X}}(x). \tag{5.1.1}$$

Here, $\mathcal{P}$ is a probability distribution accessible only through sampling, $A(\cdot,z)$ is a smooth map for almost every $z \sim \mathcal{P}$, and $N_{\mathcal{X}}(x)$ denotes the normal cone to a closed set $\mathcal{X}$. Stochastic variational inequalities (5.1.1) are ubiquitous in contemporary optimization. For example, optimality conditions for constrained optimization problems

$$\min_{x} \mathop{\mathbb{E}}_{z\sim\mathcal{P}} f(x,z) \qquad \text{subject to } x \in \mathcal{X},$$

fit into the framework (5.1.1) by setting $A(x,z) = \nabla f(x,z)$ in (5.1.1). More generally still, Nash equilibria $x^\star = (x_1^\star, \ldots, x_m^\star)$ of stochastic games are solutions of the system

$$x_j^\star \in \operatorname*{argmin}_{x_j \in \mathcal{X}_j} \mathop{\mathbb{E}}_{z \sim \mathcal{P}} [f_j(x,z)] \qquad \text{for all } j = 1, \ldots, m,$$

where f_j and $\mathcal{X}_j$, respectively, are the loss function and the strategy set of player j. First order optimality conditions for these k coupled inclusions can be modeled as (5.1.1) by setting $[A(x,z)]_j := \nabla_{x_j} f_j(x,z)$ and $\mathcal{X} := \mathcal{X}_1, \ldots, \mathcal{X}_m$.

There are two standard strategies for solving (5.1.1): sample average approximation (SAA) and the stochastic forward-backward algorithm (SFB). The former proceeds by drawing a batch of samples $z_1, z_2, \ldots, z_k \overset{\text{iid}}{\sim} \mathcal{P}$ and finding a solution x_k to the empirical approximation

$$0 \in \frac{1}{k} \sum_{i=1}^{k} [A(x, z_i)] + N_{\mathcal{X}}(x). \tag{5.1.2}$$

In contrast, the stochastic forward-backward (SFB) algorithm proceeds in an online manner, drawing a single sample $z_k \sim \mathcal{P}$ in each iteration k and declaring the next iterate x_{k+1} as

$$x_{k+1} \in P_{\mathcal{X}}(x_k - \alpha_k \cdot A(x_k, z_k)). \tag{5.1.3}$$

Here, $P_{\mathcal{X}}(\cdot)$ denotes the nearest-point projection onto $\mathcal{X}$. In the case of constrained optimization, $A(x,z) = \nabla f(x,z)$ is the gradient of some loss function $f(x,z)$, and the process (5.1.3) reduces to the stochastic projected gradient algorithm. Online algorithms like SFB are usually preferable to SAA since each iteration is inexpensive and can be performed online, whereas SAA requires solving the auxiliary optimization problem (5.1.2). Although the asymptotic distribution of the SAA estimators is by now well-understood [87–89], our understanding of the asymptotic performance of the SFB iterates is limited in nonsmooth and constrained settings. The goal of this part is to fill this gap. The main result is the following.

> Under reasonable assumptions, the running average of the SFB iterates exhibits the same asymptotic distribution as SAA. Moreover, both SAA and SFB are asymptotically optimal in a locally minimax sense of Hájek and Le Cam [17, 18].

We next describe our results, and their consequences, in some detail. Namely, it is classically known (e.g. [87–89]) that the asymptotic performance of SAA (5.1.2) is strongly influenced by the sensitivity of the solution $x^\star$ to perturbations of the left-hand-side of (5.1.1). In order to isolate this effect, let $S(v)$ consist of solutions x to the perturbed system

$$v \in \mathop{\mathbb{E}}_{z\sim\mathcal{P}}[A(x,z)] + N_{\mathcal{X}}(x).$$

Throughout, we will assume that the solutions $S(v)$ vary smoothly near $x^\star$. More precisely, we will assume that the graph of S locally around $(0, x^\star)$ coincides with the graph of some smooth map $\sigma(\cdot)$. In the language of variational analysis [90], the map $\sigma(\cdot)$ is called a smooth localization of S around $(0, x^\star)$. It is known that this assumption holds in a variety of concrete circumstances and under generic linear perturbations of semialgebraic problems [59].

Let us next provide the context and state our results. It is known from [87, 88] that under mild assumptions, the solutions x_k of SAA (5.1.2) are asymptotically normal:

$$\sqrt{k}(x_k - x^\star) \xrightarrow{D} \mathsf{N}(0, \nabla\sigma(0)\cdot \mathrm{Cov}(A(x^\star, z))\cdot \nabla\sigma(0)^\top). \tag{5.1.4}$$

Thus the Jacobian of the solution map $\nabla\sigma(0)$ appears in the asymptotic covariance of the SAA estimator. In fact, we will argue that this is unavoidable. Our first contribution is that we prove that the asymptotic performance of SAA is locally minimax optimal—in the sense of Hájek and Le Cam [17, 18]—among all estimation procedures. Roughly speaking, this means that for any estimation procedure that outputs $\hat{x}_k$ based on k samples, there exists a sequence of perturbations $\mathcal{P}_k$ with $\frac{d\mathcal{P}_k}{d\mathcal{P}} = 1 + O(k^{-1/2})$, such that the performance of $\hat{x}_k$ on the perturbed sequence of problems is asymptotically no better than the performance of SAA on the perturbed problems. We note that the analogous lower bound for stochastic nonlinear programming was obtained earlier in [16], and our arguments are motivated by the techniques therein. The fact that the SFB algorithm for smooth problems is asymptotically optimal was proved in [91, Theorem 5.6] by verifying that the averaged iterates of SFB have the same asymptotic limit as SAA; we follow a similar argument here.

Aside from the lower bound, the main result of this chapter is to show that under reasonable assumptions, the running average of the SFB iterates enjoys the same asymptotics as (5.1.4) and is thus asymptotically optimal.

The guarantees we develop are already interesting for stochastic nonlinear programming:

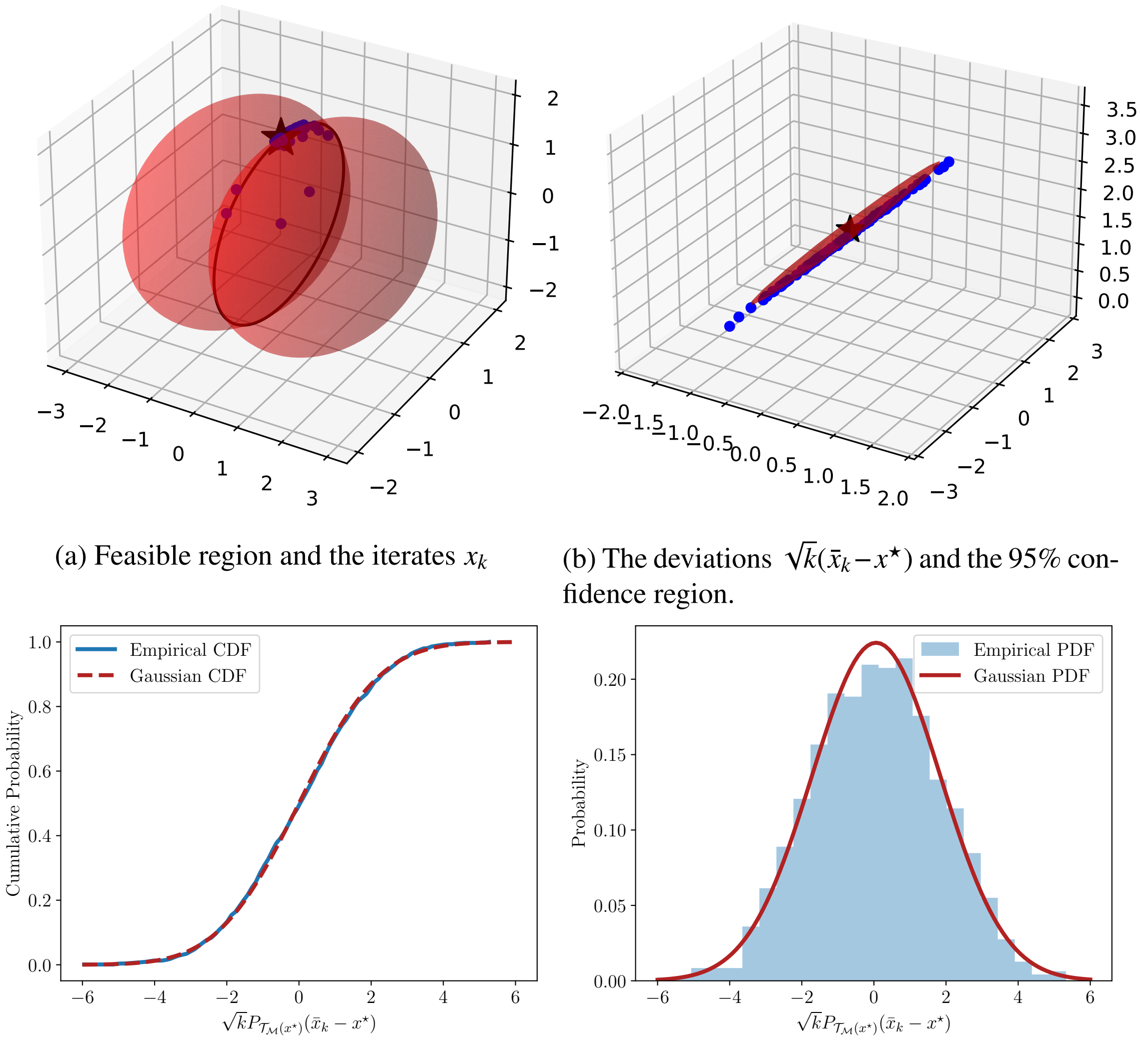
$$\min_x \; f(x) = \mathop{\mathbb{E}}_{z\sim\mathcal{P}}[f(x,z)] \qquad \text{subject to} \qquad g_i(x) \le 0 \qquad \forall i = 1,\ldots,m. \tag{5.1.5}$$

Here each g_i is a smooth function and the map $x \mapsto f(x,z)$ is smooth for a.e. $z \sim \mathcal{P}$. The optimality conditions for this problem can be modeled as the variational inequality (5.1.1) under the identification $A(x,z) = \nabla f(x,z)$ and $\mathcal{X} = \{x : g_i(x) \le 0 \; \forall i = 1,\ldots,m\}$. The stochastic forward-backward algorithm then becomes the stochastic projected gradient method. Our results imply that under the three standard conditions—linear independence of active gradients, strict complementarity, and strong second-order sufficiency—the running average of the SFB iterates $\bar{x}_k = \frac{1}{k}\sum_{i=1}^k x_i$ is asymptotically normal and optimal:

$$\sqrt{k}(\bar{x}_k - x^\star) \xrightarrow{D} N\left(0, \nabla\sigma(0)\cdot \mathrm{Cov}(\nabla f(x^\star,z))\cdot\nabla\sigma(0)\right).$$

Moreover, as is classically known, the Jacobian $\nabla\sigma(0)$ admits an explicit description as

$$\nabla\sigma(0) = (P_{\mathcal{T}}\nabla^2_{xx}\mathcal{L}(x^\star,y^\star)P_{\mathcal{T}})^\dagger, \tag{5.1.6}$$



(a) Feasible region and the iterates x_k

(b) The deviations $\sqrt{k}(\bar{x}_k - x^\star)$ and the 95% confidence region.

(c) Empirical vs Gaussian CDF

(d) Histogram vs Gaussian density.

Figure 5.1: The stochastic projected gradient method for minimizing $\mathbb{E}_g[-x_1 + \langle g, x\rangle]$ over the intersection of two balls centered around $(-1, 0, 0)$ and $(1, 0, 0)$ of radius two. The expectation is taken over a Gaussian $g \sim N(0, I)$. The optimal solution $(0, 0, \sqrt{3})$ (marked with a star) lies on the active manifold $\mathcal{M}$, which is a circle depicted in black. The figure on the top left depicts the iterates generated by a single run of the process initialized at the origin with stepsize $\eta_k = k^{-3/4}$ and executed for 1000 iterations. The figure on the top right depicts the rescaled deviations $\sqrt{k}(\bar{x}_K - x^\star)$ taken over 100 runs with $K = 10^6$. The two figures clearly show that the iterates rapidly approach the active manifold and asymptotically the deviations $\sqrt{k}(\bar{x}_k - x^\star)$ are supported only along the tangent space to $\mathcal{M}$ at $x^\star$. The two figures on the second row show the histogram and the empirical CDF, respectively, of the tangent components $\sqrt{k}P_{T_{\mathcal{M}}(x^\star)}(\bar{x}_k - x^\star)$, overlaid with the analogous functions for a Gaussian.

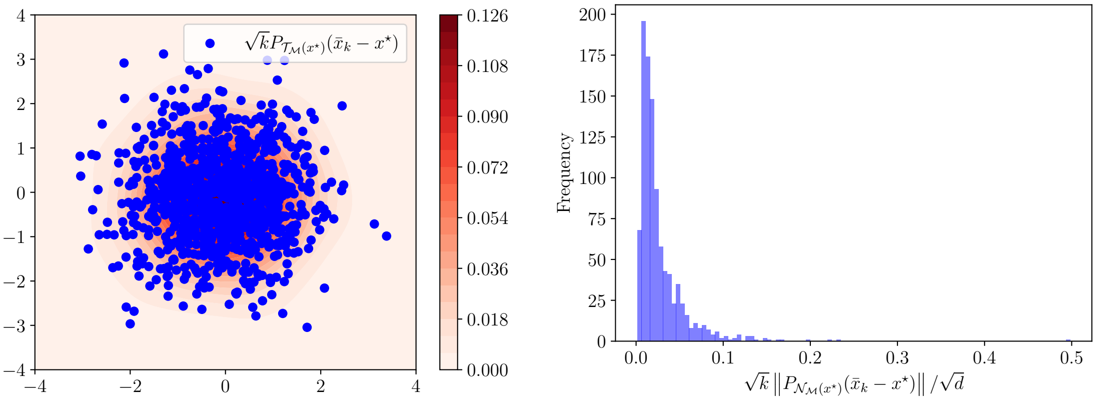
where $\nabla^2_{xx}\mathcal{L}(x^\star, y^\star)$ is the Hessian of the Lagrangian function, the symbol $\dagger$ denotes the Moore-Penrose pseudoinverse, and $P_{\mathcal{T}}$ is the projection onto the linear subspace $\{\nabla g_i(x^\star)\}^\perp_{i\in\mathcal{I}}$ and $\mathcal{I} = \{i : g_i(x^\star) = 0\}$ is the set of active indices. An illustrative example of the announced result is depicted in Figure 5.1, which plots the performance of the projected stochastic gradient method for minimizing a linear function over the intersection of two balls. A further illustration for a nonconvex problem of sparse recovery is depicted in Figure 5.2. This result may be surprising in light of the existing literature. Namely, Duchi and Ruan [16] uncover a striking gap between the estimation quality of SAA and at least one standard online method, called dual averaging [92, 93], for stochastic nonlinear optimization. Indeed, even for the problem of minimizing the expectation of a linear function over a ball, the dual averaging method exhibits a suboptimal asymptotic covariance [16, Section 5.2].[1] In contrast, we see that the stochastic projected gradient method is asymptotically optimal.

Let us now return to the general problem (5.1.1) and the stochastic forward-backward algorithm (5.1.3). In order to derive the claimed asymptotic guarantees for SFB, we will impose a few extra assumptions. First, in addition to assuming that $\sigma(\cdot)$ is smooth near the origin, we will assume that there exists a neighborhood U of the origin such that $\sigma(U)$ is a smooth manifold. This assumption is mild, since it holds automatically for example if the matrix $\nabla\sigma(\cdot)$ has constant rank on a neighborhood of the origin. With these assumptions, the set $\mathcal{M} = \sigma(U)$ is an *active manifold* around $\bar{x}$ [6]. Returning to the case of stochastic nonlinear programming, the active manifold is simply the zero-set of the active inequalities

$$\mathcal{M} = \{x : g_i(x) = 0 \ \ \forall i \in \mathcal{I}\}.$$

See Figure 5.1 for an illustration.

[1] In contrast, in the special case that $\mathcal{X}$ is polyhedral and convex, the dual averaging method is optimal [16].



(a) Kernel density estimation on tangent deviations.

(b) Histogram of normal deviations.

Figure 5.2: The stochastic projected gradient method for minimizing $\mathbb{E}_{(a,b)}[(\langle a, x\rangle - b)^2]$ over the ℓ_0 ball $\mathcal{X} = \{x \colon \|x\|_0 \leq 2\}$. Here $a \sim N(0, I)$ and $b = \langle a, x^\star\rangle + g$ where $g \sim N(0, 1)$ and $x^\star := e_1 + e_2$, the sum of the first two standard basis vectors; in this example, $d = 20$. The optimal solution $x^\star$ lies on the active manifold $\mathcal{M} = \mathrm{span}\{e_1, e_2\}$. The figure on the left depicts a kernel density estimation of the rescaled deviations $\sqrt{K} \cdot P_{\mathcal{T}_{\mathcal{M}}(x^\star)}(\bar{x}_K - x^\star)$ taken over 1000 runs of SGD (Gaussian kernel, bandwidth .5); here, the method is initialized at the origin with stepsize $\eta_k = k^{-3/4}$ and ran for $K = 10^6$ iterations. The figure on the right depicts the rescaled normal deviations $\| \sqrt{K} \cdot P_{\mathcal{N}_{\mathcal{M}}(x^\star)}(\bar{x}_K - x^\star)\|/\sqrt{d}$. Taken together, the figures again show that the iterates rapidly approach the active manifold and asymptotically the deviations $\sqrt{k}(\bar{x}_k - x^\star)$ are supported only along the tangent space to $\mathcal{M}$ at $x^\star$.

The main idea of our argument is to relate the nonsmooth dynamics of SFB to a smooth stochastic approximation algorithm on $\mathcal{M}$, which is similar to the techniques used in saddle avoidance results. More precisely, we will show that under mild conditions, the shadow sequence $y_k := P_{\mathcal{M}}(x_k)$ along the manifold $\mathcal{M}$ behaves smoothly up to a small error

$$y_{k+1} = y_k - \alpha_k P_{T_{\mathcal{M}}(y_k)}(A(y_k, z_k)) + o(\alpha_k), \tag{5.1.7}$$

where $T_{\mathcal{M}}(y_k)$ denotes the tangent space of $\mathcal{M}$ at y_k. We note that in the constrained optimization setting, the iteration (5.1.7) becomes an inexact Riemannian gradient method on the restriction of f to $\mathcal{M}$. Consequently, we may build on the techniques of Polyak and Juditsky [15] to obtain the asymptotics of the shadow sequence y_k, and then infer information about the original iterates x_k.

The validity of (5.1.7) relies on two regularity conditions – (b)-regularity and strong (a)-regularity, which were introduced in Chapter 3 and applied in Chapter 4. We refer readers to the end of section 4.1.1 for a detailed discussion. As we see below, both regularity conditions hold automatically for stochastic nonlinear programming.

5.1.1 Outline of the chapter.

The outline of the rest of the chapter is as follows. Section 5.2 discusses nonlinear programming and l_1-regulization, two main examples in this chapter. Section 5.3 introduces the notation of smoothly invertible maps. The existence of smooth localizations $\sigma(\cdot)$ is a central assumption of this chapter. Section 5.4 develops asymptotic convergence guarantees for SAA, which motivate much of the subsequent sections. Section 5.5 presents the classes of algorithms that we consider. Section 5.6 states the main result on asymptotic normality of iterative methods. Section 5.7 present shows that SAA and SFB are both asymptotically local minimax optimal in the sense of Hájek and Le Cam. This chapter is based on the work [26].

5.2 Examples: nonlinear programming and l_1-regularization

Although (b) and strong (a)-regularity conditions for functions were defined in Chapter 3. We refer readers to Theorem 3.1.4 and Theorem 3.1.6. The two regularity conditions easy extend to sets through their indicator functions. Namely, we say that a set $Q \subset \mathbb{R}^d$ is $(b_{\leq})$-regular (respectively strongly (a)-regular) along a C^1 manifold $\mathcal{M} \subset Q$ at $\bar{x} \in \mathcal{M}$ if the indicator function δ_Q is $(b_{\leq})$-regular (respectively strongly (a)-regular) along $\mathcal{M}$ at $\bar{x}$.

Chapter 3 presents a wide array of functions that admit active manifolds along which both conditions ($b_{\leq}$) and strong (a) hold. Here, we discuss in detail examples of nonlinear programming and l_1-regularization.

Example 5.2.1 (Nonlinear programming)**.** Consider the problem of nonlinear programming

$$\begin{aligned} \min_{x} \ & f(x) \\ \text{s.t. } & g_i(x) \leq 0 \qquad \text{for } i = 1, \ldots, m \\ & g_i(x) = 0 \qquad \text{for } i = m+1, \ldots, n, \end{aligned} \tag{5.2.1}$$

where f and g_i are C^p-smooth functions on $\mathbb{R}^d$. Let $\mathcal{X}$ denotes the set of all feasible points to the problem. Consider now a point $\bar{x} \in \mathcal{X}$ that is critical for the function $f + \delta_{\mathcal{X}}$ and define the active index set

$$\mathcal{I} = \{i : g_i(\bar{x}) = 0\}.$$

Suppose the following is true:

- **(LICQ)** the gradients $\{\nabla g_i(\bar{x})\}_{i \in \mathcal{I}}$ are linearly independent.

Then the set

$$\mathcal{M} = \{x : g_i(x) = 0 \ \forall i \in \mathcal{I}\}$$

is a C^p smooth manifold locally around $\bar{x}$. Moreover, all three functions f, $\delta_{\mathcal{X}}$, and $f + \delta_{\mathcal{X}}$ are ($b_{\leq}$)-regular and strongly (a)-regular along $\mathcal{M}$ near $\bar{x}$. In order to ensure that $\mathcal{M}$ is an active manifold of $f + \delta_{\mathcal{X}}$, an extra condition is required. Define the Lagrangian function

$$\mathcal{L}(x, y) := f(x) + \sum_{i=1}^{n+m} y_i g_i(x).$$

Criticality of $\bar{x}$ and LICQ ensures that there exists a (unique) Lagrange multiplier vector $\bar{y} \in \mathbb{R}^m_+ \times \mathbb{R}^n$ satisfying $\nabla_x \mathcal{L}(\bar{x}, \bar{y}) = 0$ and $\bar{y}_i = 0$ for all $i \notin \mathcal{I}$. Suppose the following standard assumption is true:

- **(Strict complementarity)** $\bar{y}_i > 0$ for all $i \in \mathcal{I} \cap \{1, \ldots, m\}$.

Then $\mathcal{M}$ is indeed an active C^p manifold for $f + \delta_{\mathcal{X}}$ at $\bar{x}$.

Example 5.2.2 (ℓ_1-regularization)**.** *Consider the stochastic optimization problem with ℓ_1 regularization*

$$\min_x \; g(x) = f(x) + \lambda \|x\|_1, \tag{5.2.2}$$

where $f(x) = \mathbb{E}_{z \in \mathcal{P}}[f(x, z)]$ is a C^p-smooth function in $\mathbb{R}^d$. Consider now a point $\bar{x} \in \mathbb{R}^d$ that is critical for the function g and define the index set $\mathcal{I} = \{i \colon \bar{x}_i = 0\}$. Then the set

$$\mathcal{M} = \{x \colon x_i = \bar{x}_i, \; \forall i \in \mathcal{I}\}$$

is an affine space, hence a smooth manifold. Note that the definition of criticality ensures that $0 \in \partial g(\bar{x})$, so we always have

$$-(\nabla f(x))_i \in [-\lambda, \lambda], \qquad \forall i \in \mathcal{I}.$$

Suppose the following condition is true:

- **(Strict complementarity)** $-(\nabla f(x))_i \in (-\lambda, \lambda)$ *for all i* $\in \mathcal{I}$.

Then $\mathcal{M}$ is indeed an active C^p manifold for g at $\bar{x}$. Moreover, $(b_{\leq})$-regularity and strong (a)-regularity hold trivially for g along $\mathcal{M}$ at $\bar{x}$.

5.3 Smoothly invertible maps and active manifolds

Performance of statistical estimation procedures strongly depends on the sensitivity of the problem to perturbation. A variety of estimation problems can in turn be modeled as the task of solving an inclusion $0 \in F(x)$ for some set-valued map F, whose values we can only approximate by sampling. We next review basic perturbation theory based on the inverse/implicit function theorem paradigm, while closely following the monograph [90].

A *set-valued map* $F\colon \mathbb{R}^d \rightrightarrows \mathbb{R}^m$ is an assignment that maps a point $x \in \mathbb{R}^d$ to a set $F(x) \subset \mathbb{R}^m$. Set-valued maps always admit a set-valued inverse:

$$F^{-1}(y) = \{x : y \in F(x)\}.$$

The domain and graph of F are defined, respectively, as

$$\operatorname{dom} F := \{x : F(x) \neq \emptyset\} \qquad \text{and} \qquad \operatorname{gph} F := \{(x, y) : y \in F(x)\}.$$

We will be interested in the sensitivity of the solutions to the system $v \in F(x)$ with respect to perturbations of the left-hand-side v, or equivalently, the variational behavior of the map $v \mapsto F^{-1}(v)$. In particular, we will be interested in settings when the graph of F^{-1} coincides locally around a base point (v, x) with a graph of a smooth map. This is the content of the following definition.

Definition 5.3.1 (Smooth invertibility)**.** Consider a set-valued map $F\colon \mathbb{R}^d \rightrightarrows \mathbb{R}^m$ and a pair $(\bar{x}, \bar{v}) \in \operatorname{gph} F$. We say that F is C^p *invertible around* $(\bar{x}, \bar{v})$ *with inverse* $\sigma(\cdot)$ if there exists a single-valued C^p-smooth map $\sigma(\cdot)$ and a neighborhood U of $(\bar{v}, \bar{x})$ satisfying

$$U \cap \operatorname{gph} F^{-1} = U \cap \operatorname{gph} \sigma.$$

The definition might seem odd at first: there is nothing “smooth” about F, and yet we require the graph of F^{-1} to coincide with a graph of a smooth function near $(\bar{v}, \bar{x})$. On the contrary, we will see that in a variety of settings this assumption is indeed valid. In particular, smooth invertibility is typical in nonlinear programing.

Example 5.3.1 (Nonlinear programming)**.** Returning to Example (5.2.1) with $p \geq 2$, define the set-valued map

$$F(x) = \nabla f(x) + N_{\mathcal{X}}(x).$$

Then it is classically known that F is C^{p-1} invertible at $(\bar{x}, 0)$ if and only if the matrix

$$\Sigma := P_{T_{\mathcal{M}}(\bar{x})} \nabla^2_{xx} \mathcal{L}(\bar{x}, \bar{y}) P_{T_{\mathcal{M}}(\bar{x})}$$

is nonsingular on $T_{\mathcal{M}}(\bar{x})$. In this case, the Jacobian of the inverse map is $\nabla\sigma(0) = \Sigma^{\dagger}$, where † denotes the Moore-Penrose pseudoinverse. It is worthwhile to note that Σ can be equivalently written as $P_{T_{\mathcal{M}}(\bar{x})} \nabla^2_{\mathcal{M}} f(\bar{x}) P_{T_{\mathcal{M}}(\bar{x})}$.

Example 5.3.2 (ℓ_1-regularization)**.** *Returning to Example* (5.2.2) *with $p \geq 2$, define the set-valued map $F(x) = \nabla f(x) + \lambda \partial(\| \cdot \|_1)(x)$. Then F is C^{p-1} invertible at* $(\bar{x}, 0)$ *if and only if the matrix $\Sigma := P_{T_{\mathcal{M}}(\bar{x})} \nabla^2 f(\bar{x}) P_{T_{\mathcal{M}}(\bar{x})}$ is nonsingular on $T_{\mathcal{M}}(\bar{x})$.*

Smooth invertibility is closely tied to active manifolds, and Example 5.3.1 and Example 5.3.2 are simple consequences. Indeed the following much more general statement is true. This result follows from a standard argument combining active manifolds and the implicit function theorem. The proof appears in Section 7.2.1 of the supplementary document. We will require a mild assumption on the considered functions. Following [64, Definition 2.1] a function f is called *subdifferentially continuous* at a point $\bar{x}$ if for any sequences $(x_i, v_i) \in \operatorname{gph} \partial f$ converging to some pair $(\bar{x}, \bar{v}) \in \operatorname{gph} \partial f$, the function values $f(x_i)$ converge to $f(\bar{x})$. In particular, functions that are continuous on their domains and closed convex functions are subdifferentially continuous.

Theorem 5.3.2 (Smooth Invertibility and Active Manifolds)**.** *Consider the map*

$$F(x) := A(x) + \partial f(x),$$

where $A \colon \mathbb{R}^d \to \mathbb{R}^d$ is C^p-smooth and $f \colon \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is subdifferentially continuous near a point $\bar{x}$. Suppose that f admits a C^{p+1} active manifold $\mathcal{M}$ at some point $\bar{x}$ for

$-A(\bar{x}) \in \hat{\partial} f(\bar{x})$. Let $G(x) = 0$ be any C^{p+1}-smooth local defining equations for $\mathcal{M}$ near $\bar{x}$ and let $\hat{f}$ be a C^{p+1}-smooth function that agrees with f on a neighborhood of $\bar{x}$ in $\mathcal{M}$. Define the map

$$\mathcal{H}(x, y) := A(x) + \nabla \hat{f}(x) + \nabla G(x)^\top y.$$

Then there exists a unique multiplier vector $\bar{y}$ satisfying the condition $0 = \mathcal{H}(\bar{x}, \bar{y})$. Moreover, F is C^p-invertible around $(0, \bar{x})$ with inverse $\sigma(\cdot)$ if and only if the matrix

$$\Sigma := P_{T_{\mathcal{M}}(\bar{x})} \nabla_x \mathcal{H}(\bar{x}, \bar{y}) P_{T_{\mathcal{M}}(\bar{x})}$$

is nonsingular on $T_{\mathcal{M}}(\bar{x})$, and in this case equality $\nabla \sigma(0) = \Sigma^\dagger$ holds.

5.4 Asymptotic normality of SAA

Before analyzing the asymptotic performance of stochastic approximation algorithms, it is instructive to first recall guarantees for sample average approximation (SAA), where the assumptions, conclusions, and arguments are much simpler to state. This is the content of this section: we derive the asymptotic distribution of the SAA estimator for nonsmooth problems.Throughout the section we focus on the problem of finding a point $x^\star$ satisfying the variational inclusion:

$$0 \in A(x) + H(x) \qquad \text{where} \qquad A(x) = \mathop{\mathbb{E}}_{z \sim \mathcal{P}} A(x, z). \tag{5.4.1}$$

Here $H\colon \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is a set-valued map with closed graph, $\mathcal{P}$ is a fixed probability distribution on some measure space $(\mathcal{Z}, \mathcal{F})$, and $A\colon \mathbb{R}^d \times \mathcal{Z} \to \mathbb{R}^d$ is a measurable map. We will impose the following assumption throughout the rest of the section.

Assumption G. The map $F := A + H$ is C^1-smoothly invertible near $(0, \bar{x})$ with inverse $\sigma(\cdot)$.

The SAA approach to solving (5.4.1) proceeds as follows. Let $S = \{z_1, \ldots, z_k\}$ be i.i.d samples drawn from $\mathcal{P}$ and let x_k be a solution of the problem

$$0 \in A_S(x) + H(x) \qquad \text{where} \qquad A_S(x) := \frac{1}{k}\sum_{i=1}^{k} A(x, z_i), \tag{5.4.2}$$

assuming one exists. We will now show that the solutions of sample average approximations are asymptotically normal with covariance $\nabla\sigma(0) \cdot \mathrm{Cov}(A(\bar{x}, z)) \cdot \nabla\sigma(0)^\top$. Though variants of this result are known [87–89], we provide a short proof in Section 7.2.2 of the supplementary document highlighting the use of the solution map $\sigma(\cdot)$. To this end, we impose the following standard assumption.

Assumption H (Integrability and smoothness)**.** Suppose that there exists a neighborhood U around $\bar{x}$ satisfying the following.

1. For almost every z, the map $A(\cdot, z)$ is differentiable at every $x \in U$.
2. The second moment bounds hold:

$$\sup_{x \in U} \mathop{\mathbb{E}}_{z \sim \mathcal{P}} \|A(x, z)\|^2 < \infty \qquad \text{and} \qquad \sup_{x \in U} \mathop{\mathbb{E}}_{z \sim \mathcal{P}} \left[\|\nabla A(x, z)\|_{\mathrm{op}}^2\right] < \infty.$$

The following theorem shows that as long as x_k eventually stay in a sufficiently small neighborhood of $\bar{x}$, the error $\sqrt{k}(x_k - \bar{x})$ is asymptotically normal with covariance $\nabla\sigma(0) \cdot \mathrm{Cov}(M(\bar{x}, z)) \cdot \nabla\sigma(0)^\top$. Verifying that the problem (5.4.2) admits solutions x_k that are sufficiently close to $\bar{x}$ is a separate and well-studied topic and we do not discuss it here.

Theorem 5.4.1 (Sample average approximation)**.** *Suppose that Assumptions G and H hold. In particular, there exist $\epsilon_1, \epsilon_2 > 0$ and a C^1-smooth map $\sigma\colon B_{\epsilon_1}(0) \to B_{\epsilon_2}(\bar{x})$ with*

$$\operatorname{gph}\sigma = (B_{\epsilon_1}(0) \times B_{\epsilon_2}(\bar{x})) \cap \operatorname{gph} F^{-1}.$$

Suppose moreover that there exists a square integrable function $L(z)$ satisfying

$$\|\nabla A(x_1, z) - \nabla A(x_2, z)\| \leq L(z)\|x_1 - x_2\| \qquad \forall x_1, x_2 \in B_{\epsilon_2}(\bar{x}). \tag{5.4.3}$$

Shrinking ϵ_2, if necessary, let us ensure that $\epsilon_2 \leq \min\left\{\frac{\mathrm{lip}(\sigma)^{-1}}{2\mathbb{E}L}, \sqrt{\frac{\epsilon_1}{2\mathbb{E}L}}\right\}$. *Let* $S = \{z_1, \ldots, z_k\}$ *be i.i.d samples drawn from* $\mathcal{P}$ *and let* x_k *be a measurable selection of the solutions* (5.4.2) *such that* $\Pr[x_k \in B_{\epsilon_2}(\bar{x})] \to 1$ *as k tends to infinity. Then the expansion holds:*

$$\sqrt{k}(x_k - \bar{x}) = -\nabla\sigma(0) \cdot \sqrt{k}(A(\bar{x}) - A_S(\bar{x})) + o_P(1),$$

and therefore

$$\sqrt{k}(x_k - \bar{x}) \xrightarrow{D} \mathsf{N}(0, \nabla\sigma(0) \cdot \mathrm{Cov}(A(\bar{x}, z)) \cdot \nabla\sigma(0)^\top). \tag{5.4.4}$$

Note that Theorem 5.4.1 assumes existence of a measurable selection of the solutions (5.4.2) such that $\Pr[x_k \in B_{\epsilon_2}(\bar{x})] \to 1$ as k tends to infinity. This is a very mild assumption and follows for example from uniform convergence and smooth invertibility.

Theorem 5.4.2 (Existence of measurable selections). *Suppose that F is C^1-smoothly invertible around near* $(0, \bar{x})$ *and that $A(x, z)$ and $\nabla A(x, z)$ converge uniformly on some ball around $\bar{x}$, that is there exists $\epsilon > 0$ such that*

$$\sup_{x \in B_\epsilon(\bar{x})} \|\nabla A_S(x) - \nabla A(x)\| = o_p(1) \qquad \textit{and} \qquad \sup_{x \in B_\epsilon(\bar{x})} \|A_S(x) - A(x)\| = o_p(1).$$

Then there exists $\delta > 0$ and a measurable selection of the solutions (5.4.2) *such that* $\Pr[x_k \in B_\delta(\bar{x})] \to 1$ *as k tends to infinity.*

Proof. Standard results on the implicit function theorem (see proof of [90, Theorem 3G.3]) imply that there exist sufficiently small $\varepsilon_2, \varepsilon_3 > 0$ such that whenever $\sup_{x \in B_\epsilon(\bar{x})} \|A_S(x) - A(x)\| < \varepsilon_2$ and $\sup_{x \in B_\epsilon(\bar{x})} \|\nabla A_S(x) - \nabla A(x)\| < \varepsilon_2$, the map $A_S + H$ is guaranteed to be smoothly invertible on $B_{\varepsilon_3}(\bar{x}) \times B_{\varepsilon_3}(0)$. In particular, the solution

$x_k \in B_{\varepsilon_3}(\bar{x})$ of (5.4.2) exists and is unique. To see measurability of x_k, observe that the maps A_S and H are both measurable [28, Exercise 14.9] and therefore so is the map $D(S) = \text{gph}(A_S + H)$. Notice now that $x_k(S)$ uniquely satisfies the inclusion $(x_k(S), 0) \in D(S)$. Therefore, [28, Theorem 14.16] implies that x_k is measurable. □

Our goal in the rest of the chapter is to show that a simple online algorithm, namely the stochastic forward backward (SFB) method, under reasonable assumptions enjoys the same guarantees as (5.4.4) for SAA. Moreover, in the final section of the chapter (Section 5.7), we will show that this performance is best possible among any estimation procedure in a local minimax sense, and therefore both SAA and SFB are asymptotically local minimax optimal.

5.5 Stochastic approximation: assumptions & examples

We now move to stochastic approximation algorithms, and in this section set forth the algorithms we will consider and the relevant assumptions. This section can be viewed as a generalization of Section 4.2, where everything is stated specifically for optimization problems rather than for finding zeros of set-valued maps. The concrete examples we will present will all be geared toward' solving variational inclusions, but the specifics of this problem class are somewhat distracting. Therefore we will instead only isolate the essential ingredients that are needed for our results to take hold. Setting the stage, our goal is to find a point x satisfying the inclusion

$$0 \in F(x), \tag{5.5.1}$$

where $F\colon \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is a set-valued map. Throughout, we fix one such solution $\bar{x}$ of (5.5.1). We will assume that in a certain sense, the problem (5.5.1) is “variationally smooth". That is, there exists a distinguished manifold $\mathcal{M}$—the active manifold in concrete examples—containing $\bar{x}$ and such that the map $x \mapsto P_{T_{\mathcal{M}}(x)}F(x)$ is single-valued and C^1-smooth on $\mathcal{M}$ near $\bar{x}$. The following assumption makes this precise.

Assumption I (Smooth reduction)**.** Suppose that there exists a C^p manifold $\mathcal{M} \subset \mathbb{R}^d$ such that the following properties are true.

(I1) The map $F_{\mathcal{M}} \colon \mathcal{M} \to \mathbb{R}^d$ defined by

$$F_{\mathcal{M}}(x) := P_{T_{\mathcal{M}}(x)}F(x)$$

is single-valued on some neighborhood of $\bar{x}$ in $\mathcal{M}$.

(I2) There exists a neighborhood U of $(\bar{x}, 0)$ such that

$$U \cap \operatorname{gph} F = U \cap \operatorname{gph}(F_{\mathcal{M}} + N_{\mathcal{M}}).$$

We note that smooth invertibility of F can be easily characterized in terms of the covariant Jacobian $\nabla_{\mathcal{M}}F_{\mathcal{M}}(\bar{x})$. This is the content of the following lemma.

Lemma 5.5.1 (Jacobian of the solution map)**.** *The map F is C^p-smoothly invertible around $(\bar{x}, 0)$ with localization $\sigma(\cdot)$ if and only if the linear map $P_{T_{\mathcal{M}}(\bar{x})}\nabla F_{\mathcal{M}}(\bar{x})P_{T_{\mathcal{M}}(\bar{x})}$ is nonsingular on $T_{\mathcal{M}}(\bar{x})$. In this case, the Jacobian of the localization is given by*

$$\nabla\sigma(0) = (P_{T_{\mathcal{M}}(\bar{x})}\nabla_{\mathcal{M}}F_{\mathcal{M}}(\bar{x})P_{T_{\mathcal{M}}(\bar{x})})^{\dagger}.$$

Proof. Let Φ be a smooth extension of F to a neighborhood $V \subset \mathbb{R}^d$ of $\bar{x}$. In light of Assumption (I2), the graphs of F and $\Phi + N_{\mathcal{M}}$ coincide near $(\bar{x}, 0)$, and therefore we can focus on existence of smooth localizations of $(\Phi + N_{\mathcal{M}})^{-1}$. Applying Lemma 7.2.2 with $\bar{y} = 0$, we see that $\Phi + N_{\mathcal{M}}$ is C^p-smoothly invertible around $(\bar{x}, 0)$ if and only if the linear map $P_{T_{\mathcal{M}}(\bar{x})}\nabla\Phi(\bar{x})P_{T_{\mathcal{M}}(\bar{x})}$ is nonsingular on $T_{\mathcal{M}}(\bar{x})$. In this case, the Jacobian of the localization is given by $\nabla\sigma(0) = (P_{T_{\mathcal{M}}(\bar{x})}\nabla\Phi(\bar{x})P_{T_{\mathcal{M}}(\bar{x})})^{\dagger}$. Noting the equality $\nabla F_{\mathcal{M}}(\bar{x})P_{T_{\mathcal{M}}(\bar{x})} = \nabla_{\mathcal{M}}\Phi(\bar{x})P_{T_{\mathcal{M}}(\bar{x})}$ completes the proof. □

The stochastic approximation algorithms we consider assume access to a *generalized gradient mapping*:

$$G\colon \mathbb{R}_{++} \times \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d.$$

Given $x_0 \in \mathbb{R}^d$, the algorithm iterates the update

$$x_{k+1} = x_k - \alpha_k G_{\alpha_k}(x_k, \nu_k), \tag{5.5.2}$$

where $\alpha_k > 0$ is a control sequence and ν_k is stochastic noise. We will place relevant assumptions on the noise ν_k later in Section 5.6.

We make two assumptions on G. The first (Assumption J) is similar to classical Lipschitz assumptions and ensures the steplength can only scale linearly in $\|\nu\|$.

Assumption J (Steplength)**.** We suppose that there exists a constant $C > 0$ and a neighborhood $\mathcal{U}$ of $\bar{x}$ such that the estimate

$$\sup_{x \in \mathcal{U}_F} \|G_\alpha(x, \nu)\| \leq C(1 + \|\nu\|),$$

holds for all $\nu \in \mathbb{R}^d$ and $\alpha > 0$, where we set $\mathcal{U}_F := \mathcal{U} \cap \operatorname{dom} F$.

The second assumption makes precise the relationship between the mapping G and $F_{\mathcal{M}}$.

Assumption K (Strong (a) and aiming)**.** We suppose that there exist constants $C, \mu > 0$ and a neighborhood $\mathcal{U}$ of $\bar{x}$ such that the following hold for all $\nu \in \mathbb{R}^d$ and $\alpha > 0$, where we set $\mathcal{U}_F := \mathcal{U} \cap \operatorname{dom} F$.

(K1) **(Tangent comparison)** For all $x \in \mathcal{U}_F$, we have

$$\|P_{T_{\mathcal{M}}(P_{\mathcal{M}}(x))}(G_\alpha(x, \nu) - F(P_{\mathcal{M}}(x)) - \nu)\| \leq C(1 + \|\nu\|)^2(\operatorname{dist}(x, \mathcal{M}) + \alpha).$$

(K2) **(Proximal Aiming)** For $x \in \mathcal{U}_F$, we have

$$\langle G_\alpha(x,\nu) - \nu, x - P_{\mathcal{M}}(x)\rangle \geq \mu \cdot \mathrm{dist}(x,\mathcal{M}) - (1+\|\nu\|)^2(o(\mathrm{dist}(x,\mathcal{M})) + C\alpha).$$

Some comments are in order. Assumption (K1) ensures that the direction of motion $G_{\alpha_k}(x_k,\nu_k)$ approximates well $F_{\mathcal{M}}(P_{\mathcal{M}}(x))$ in tangent directions to the manifold $\mathcal{M}$. Assumption (K2) ensures that after subtracting the noise from $G_{\alpha_k}(x_k,\nu_k)$, the update direction $x_k - x_{k+1}$ locally points towards the manifold $\mathcal{M}$. Note that the little-o term in (K2) depends only on $\mathrm{dist}(x,\mathcal{M})$ and not on α. We will later show that this ensures the iterates x_k approach the manifold $\mathcal{M}$ at a controlled rate.

5.5.1 Examples of stochastic approximation for variation inclusions

The rest of the section is devoted to examples of algorithms satisfying Assumptions J and K. In all cases, we will consider the task of solving the variational inclusion

$$0 \in A(x) + \partial g(x) + \partial f(x). \tag{5.5.3}$$

Here $A\colon \mathbb{R}^d \to \mathbb{R}^d$ is any single-valued continuous map, $f\colon \mathbb{R}^d \to \overline{\mathbb{R}}$ is a closed function, and $g\colon \mathbb{R}^d \to \overline{\mathbb{R}}$ is a closed function that is bounded from below.[2] As explained in the introduction, variational inclusions encompass a variety of problems, most-notably first-order optimality conditions for nonlinear programming and Nash equilibria of games. In order to identify (5.5.3) with (5.5.1), we define the set-valued map F to be

$$F(x) := A(x) + \partial g(x) + \partial f(x).$$

Throughout, we fix a point $x^\star$ satisfying the inclusion (5.5.3).

[2] In particular, $\mathrm{prox}_{\alpha f}(x)$ is nonempty for all $x \in \mathbb{R}^d$ and all $\alpha > 0$.

A classical algorithm for problem (5.5.3) is the stochastic forward-backward iteration, which proceeds by taking "forward-steps" on $A + \partial g$ and proximal steps on f. Specifically, given a current iterate x_t, the algorithm performs the update

$$\begin{Bmatrix} \text{Choose } w_t \in \partial g(x_t) \\ \text{Choose } x_{t+1} \in \operatorname{prox}_{\alpha_t f}(x_t - \alpha_t(A(x_t) + w_t + \nu_t)) \end{Bmatrix}, \tag{5.5.4}$$

where ν_t is a noise sequence. The operator $G_\alpha(x, \nu)$ corresponding to this algorithm is simply

$$G_\alpha(x, \nu) := \frac{x - s_f(x - \alpha(A(x) + s_g(x) + \nu))}{\alpha},$$

where $s_g(x)$ is any selection of the subdifferential $\partial g(x)$ and $s_f(x)$ is any selection of the proximal map $\operatorname{prox}_{\alpha f}(x)$. The goal of this section is to verify Assumption K for this operator under a number of reasonable assumptions on A, g, and f.

In particular, the local boundedness condition J for G is widely used in the literature, with a variety of sufficient conditions known. The following lemma describes a number of such conditions, which we will use in what follows. The proof appear in Section 7.2.3 of the supplementary document.

Lemma 5.5.2 (Local boundedness)**.** *Suppose that $A(\cdot)$ and $s_g(\cdot)$ are locally bounded around $\bar{x}$. Then Assumption J is guaranteed to hold in any of the following settings.*

1. *f is the indicator function of a closed set $\mathcal{X}$.*
2. *f is convex and the function $x \mapsto \operatorname{dist}(0, \partial f(x))$ is bounded on $\operatorname{dom} f$ near $\bar{x}$.*
3. *f is Lipschitz continuous on $\operatorname{dom} g \cap \operatorname{dom} f$.*

We next verify Assumption K in a number of reasonable settings; all proofs appear in the supplementary document. In particular, it will be useful to note the following expression for $F_{\mathcal{M}}$. We will use this lemma throughout the section, without explicit reference.

Lemma 5.5.3 (Local tangent reduction)**.** *Suppose that f and g are Lipschitz continuous on their domains, A is C^p-smooth, $f+g$ admits an active C^{p+1} manifold at $x^\star$ for $-A(x^\star)$, and f and g are both C^{p+1}-smooth and strongly (a) regular along $\mathcal{M}$ near $x^\star$. Then Assumption I holds and $F_{\mathcal{M}}$ admits the simple form*

$$F_{\mathcal{M}}(x) = P_{T_{\mathcal{M}(x)}}(A(x)) + \nabla_{\mathcal{M}} g(x) + \nabla_{\mathcal{M}} f(x), \tag{5.5.5}$$

for all $x \in \mathcal{M}$ near $x^\star$.

5.5.1.1 Stochastic forward algorithm ($f = 0$)

We begin with the simplest case of (5.5.3) where $f = 0$. In this case, the iteration (5.5.2) reduces to a pure stochastic forward algorithm and the map G takes the simple form

$$G_\alpha(x, v) := A(x) + s_g(x) + v,$$

which is independent of α. Let us introduce the following assumption on the problem data.

Assumption L (Assumptions for the forward algorithm)**.** Suppose that $f = 0$ and that both $g(\cdot)$ and $A(\cdot)$ are Lipschitz continuous around $\bar{x}$. Suppose that $\mathcal{M} \subseteq \mathcal{X}$ is a C^p-smooth manifold for g at $\bar{x}$.

(L1) **(Strong (a))** The function g is strongly (a)-regular along $\mathcal{M}$ at $\bar{x}$.

(L2) **(Proximal aiming)** There exists $\mu > 0$ such that the inequality holds:

$$\langle A(\bar{x}) + v, x - P_{\mathcal{M}}(x)\rangle \geq \mu \cdot \operatorname{dist}(x, \mathcal{M}) \qquad \text{for all } x \text{ near } \bar{x} \text{ and } v \in \partial g(x). \tag{5.5.6}$$

Note that Corollary 3.1.5 shows that the aiming condition (L2) holds as long as the inclusion $-A(\bar{x}) \in \hat{\partial} g(\bar{x})$ holds, $\mathcal{M}$ is an active manifold for g at $\bar{x}$ for $v = -A(\bar{x})$, and g is

$(b_{\leq})$-regular along $\mathcal{M}$ at $\bar{x}$. The following proposition shows that Assumption L suffices to ensure Assumption K. The proof appears in the supplementary document.

Proposition 5.5.4 (Forward method)**.** *Assumption L implies Assumption K.*

The following is now immediate.

Corollary 5.5.5 (Active manifolds)**.** *Suppose $f = 0$ and that both $g(\cdot)$ and $A(\cdot)$ are Lipschitz continuous around $\bar{x}$. Suppose moreover that the inclusion $-A(\bar{x}) \in \hat{\partial} g(\bar{x})$ holds, that g admits a C^2 active manifold around $\bar{x}$ for $\bar{v} = -A(\bar{x})$, and that g is both $(b)_{\leq}$-regular and strongly (a)-regular along $\mathcal{M}$ at $\bar{x}$. Then Assumption K holds.*

5.5.1.2 Stochastic projected forward algorithm ($f = \delta_{\mathcal{X}}$)

Next, we focus on the particular instance of (5.5.3) where f is an indicator function of a closed set $\mathcal{X}$. In this case, the iteration (5.5.2) reduces to a stochastic projected forward algorithm and the map G takes the form

$$G_{\alpha}(x, v) := \frac{x - s_{\mathcal{X}}(x - \alpha(A(x) + s_g(x) + v))}{\alpha},$$

where $s_{\mathcal{X}}(x)$ is a selection of the projection map $P_{\mathcal{X}}(x)$. In order to ensure Assumption K for the stochastic projected forward method, we introduce the following assumption.

Assumption M (Assumptions for the projected gradient mapping)**.** Suppose that f is the indicator function of a closed set $\mathcal{X}$ and both $g(\cdot)$ and $A(\cdot)$ are Lipschitz continuous around $\bar{x}$. Let $\mathcal{M} \subseteq \mathcal{X}$ be a C^2 manifold containing $\bar{x}$ and suppose that f is C^2 on $\mathcal{M}$ near $\bar{x}$.

(M1) **(Strong (a))** The function g and set $\mathcal{X}$ are strongly (a)-regular along $\mathcal{M}$ at $\bar{x}$.

(M2) **(Proximal aiming)** There exists $\mu > 0$ such that the inequality holds

$$\langle A(\bar{x}) + v, x - P_{\mathcal{M}}(x)\rangle \geq \mu \cdot \operatorname{dist}(x, \mathcal{M}) \qquad \forall\, x \in \mathcal{X} \text{ near } \bar{x} \text{ and } v \in \partial g(x). \tag{5.5.7}$$

(M3) **(Condition (b))** The set $\mathcal{X}$ is $(b_{\leq})$-regular along $\mathcal{M}$ at $\bar{x}$.

Note that a similar argument as Corollary 3.1.5 shows that the aiming condition (M2) holds as long as the inclusion $-A(\bar{x}) \in \hat{\partial}(g + f)(\bar{x})$ holds, $\mathcal{M}$ is an active manifold of $g + f$ at $\bar{x}$ for $v = -A(\bar{x})$, and g is $(b_{\leq})$-regular along $\mathcal{M}$ at $\bar{x}$.

The following proposition shows that Assumption M is sufficient to ensure Assumption K.

Proposition 5.5.6 (Projected forward method)**.** *If Assumptions J and M hold, then so does Assumption K.*

The following is now immediate.

Corollary 5.5.7 (Active manifolds)**.** *Suppose that f is the indicator function of a closed set $\mathcal{X}$ and both $g(\cdot)$ and $A(\cdot)$ are Lipschitz continuous around $\bar{x}$. Suppose moreover the inclusion $-A(\bar{x}) \in \hat{\partial}(g + f)(\bar{x})$ holds, $g + f$ admits a C^2 active manifold around $\bar{x}$ for the vector $\bar{v} = -A(\bar{x})$, and both g and f are $(b_{\leq})$-regular and strongly (a)-regular along $\mathcal{M}$ at $\bar{x}$. Then Assumption K holds.*

5.5.1.3 Stochastic forward-backward method ($g = 0$)

Finally, we focus on the particular instance of (5.5.3) where $g = 0$. In this case, the iteration (5.5.2) reduces to a stochastic forward-backward algorithm and the map G becomes

$$G_{\alpha}(x, v) := \frac{x - s_f(x - \alpha(A(x) + v))}{\alpha},$$

In order to ensure Assumption K for the stochastic proximal gradient method, we introduce the following assumptions.

Assumption N (Assumptions for the forward-backward method)**.** Suppose $g = 0$ and $f(\cdot)$ and $A(\cdot)$ are Lipschitz continuous on $\operatorname{dom} f$ near $\bar{x}$. Suppose moreover that there exists a C^2 manifold $\mathcal{M} \subset \mathcal{X}$ containing $\bar{x}$ and such that f is C^2-smooth on $\mathcal{M}$ near $\bar{x}$.

(N1) **(Strong (a))** The function f is strongly (a)-regular along $\mathcal{M}$ at $\bar{x}$.

(N2) **(Proximal Aiming)** There exists $\mu > 0$ such that the inequality

$$\langle A(\bar{x}) + v, x - P_{\mathcal{M}}(x)\rangle \geq \mu \cdot \operatorname{dist}(x, \mathcal{M}) - (1 + \|v\|)o(\operatorname{dist}(x, \mathcal{M})) \qquad (5.5.8)$$

holds for all $x \in \operatorname{dom} f$ near $\bar{x}$ and $v \in \hat{\partial} f(x)$.

Note that Corollary 3.1.5 shows that the aiming condition (N2) holds as long as the inclusion $-A(\bar{x}) \in \hat{\partial} f(\bar{x})$ holds, $\mathcal{M}$ is an active manifold for f at $\bar{x}$ for $v = -A(\bar{x})$, and f is $(b_{\leq})$-regular along $\mathcal{M}$ at $\bar{x}$.

The following proposition shows that Assumption N is sufficient to ensure Assumption K.

Proposition 5.5.8 (Forward-backward method)**.** *If Assumptions J and N hold, then so does Assumption K.*

The following is now immediate.

Corollary 5.5.9 (Active manifolds)**.** *Suppose $g = 0$ and both f and $A(\cdot)$ are Lipschitz continuous on* $\operatorname{dom} f$ *near $\bar{x}$. Suppose moreover the inclusion $-A(\bar{x}) \in \hat{\partial} f(\bar{x})$ holds. Suppose that f admits a C^2 active manifold around $\bar{x}$ for $\bar{v} = -A(\bar{x})$ and f is both $(b)_{\leq}$-regular and strongly (a)-regular along $\mathcal{M}$ at $\bar{x}$. Then Assumption K holds.*

5.6 Asymptotic normality

Next, we impose two assumptions on the step-size α_k and the noise sequence ν_k. The first is standard and is summarized next.

Assumption O (Standing assumptions)**.** Assume the following.

(O1) The map G is measurable.

(O2) There exist constants $c_1, c_2 > 0$ and $\gamma \in (1/2, 1]$ such that

$$\frac{c_1}{k^\gamma} \leq \alpha_k \leq \frac{c_2}{k^\gamma}.$$

(O3) $\{\nu_k\}$ is a martingale difference sequence w.r.t. to the increasing sequence of σ-fields

$$\mathcal{F}_k = \sigma(x_j \colon j \leq k \text{ and } \nu_j \colon j < k),$$

and there exists a function $q \colon \mathbb{R}^d \to \mathbb{R}_+$ that is bounded on bounded sets with

$$\mathbb{E}[\nu_k \mid \mathcal{F}_k] = 0 \qquad \text{and} \qquad \mathbb{E}[\|\nu_k\|^4 \mid \mathcal{F}_k] < q(x_k).$$

We let $\mathbb{E}_k[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_k]$ denote the conditional expectation.

(O4) The inclusion $x_k \in \operatorname{dom} F$ holds for all $k \geq 1$.

All items in Assumption E are standard in the literature on stochastic approximation methods and mirror for example those found in [77, Assumption C]. The only exception is the fourth moment bound on $\|\nu_k\|$, which stipulates that ν_k has slightly lighter tails. This bound appears to be necessary for the setting we consider.

To prove our asymptotic normality results, we impose a further assumption on the noise sequence ν_k, which also appears in [16, Assumption D]. Before stating it, as motivation, consider the stochastic variational inequality (5.5.3) given by:

$$0 \in A(x) + \partial f(x) + \partial g(x) \qquad \text{where} \qquad A(x) = \mathop{\mathbb{E}}_{z \sim \mathcal{P}} A(x, z).$$

Then the noise ν_k in the algorithm (5.5.4) takes the form

$$\nu_k = A(x_k; z_k) - A(x_k).$$

Equivalently, we may decompose the right-hand-side as

$$\nu_k = \underbrace{A(\bar{x}; z_k) - A(\bar{x})}_{=:\nu_k^{(1)}} + \underbrace{(A(x_k; z_k) - A(\bar{x}; z_k)) + (A(\bar{x}) - A(x_k))}_{=:\nu_k^{(2)}(x_k)},$$

The two components $\nu_k^{(1)}$ and $\nu_k^{(2)}(x_k)$ are qualitatively different in the following sense. On one hand, the sum $\frac{1}{\sqrt{k}} \sum_{i=1}^{k} \nu_i^{(1)}$ clearly converges to a zero-mean normal vector as long as the covariance $\mathrm{Cov}(A(\bar{x}, z))$ exists. On the other hand, $\nu_k^{(2)}(x_k)$ is small in the sense that $\mathbb{E}_k \|\nu_k^{(2)}(x_k)\|^2 \leq 2 \cdot \mathbb{E}_z[L(z)^2] \cdot \|x_k - \bar{x}\|^2$, where $L(z)$ is a Lipschitz constant of $A(\cdot, z)$. With this example in mind, we introduce the following assumption on the noise sequence.

Assumption P. Fix a point $\bar{x} \in \operatorname{dom} F$ at which Assumption I holds and let U be a matrix whose column vectors form an orthogonal basis of $T_{\mathcal{M}}(\bar{x})$. We suppose the noise sequence has decomposable structure $\nu_k = \nu_k^{(1)} + \nu_k^{(2)}(x_k)$, where $\nu_k^{(2)} \colon \operatorname{dom} F \to \mathbb{R}^d$ is a random function satisfying

$$\mathbb{E}_k[\|U^\top \nu_k^{(2)}(x)\|^2] \leq C\|x - \bar{x}\|^2 \qquad \text{for all } x \in \operatorname{dom} F \text{ near } \bar{x},$$

and some $C > 0$. In addition, we suppose that for all $x \in \operatorname{dom} F$, we have $\mathbb{E}_k[\nu_k^{(1)}] = \mathbb{E}_k[\nu_k^{(2)}(x)] = 0$ and the following limit holds:

$$\frac{1}{\sqrt{k}} \sum_{i=1}^{k} U^\top \nu_i^{(1)} \xrightarrow{D} N(0, U^\top \Sigma U).$$

for some symmetric positive semidefinite matrix Σ.

Note that Assumption P only requires that $v_k^{(1)}$ and $v_k^{(2)}$ have zero conditional mean, which is weaker than being independent of the previous iterates. We are now ready to state the main result of this chapter—asymptotic normality for stochastic approximation algorithms.

Theorem 5.6.1 (Asymptotic Normality)**.** *Suppose that Assumption I, J, K, E, and P hold. Suppose that $\gamma \in (\frac{1}{2}, 1)$ and that the sequence x_k generated by the process* (4.2.2) *converges to $\bar{x}$ with probability one. Suppose that there exists a constant $\mu > 0$ satisfying*

$$\langle \nabla_{\mathcal{M}} F_{\mathcal{M}}(\bar{x}) v, v \rangle \geq \mu \|v\|^2 \qquad \textit{for all} \qquad v \in T_{\mathcal{M}}(\bar{x}). \tag{5.6.1}$$

Then F is C^p-smoothly invertible around $(\bar{x}, 0)$ with inverse $\sigma(\cdot)$ and the average iterate $\bar{x}_k = \frac{1}{k}\sum_{i=1}^k x_i$ admits the expansion

$$\sqrt{k}(\bar{x}_k - \bar{x}) = -\frac{1}{\sqrt{k}} \sum_{i=1}^{k} U(U^\top \nabla_{\mathcal{M}} F_{\mathcal{M}}(\bar{x}) U)^{-1} U^\top v_i^{(1)} + o_P(1),$$

and hence

$$\sqrt{k}(\bar{x}_k - \bar{x}) \xrightarrow{D} N\left(0, \nabla\sigma(0) \cdot \Sigma \cdot \nabla\sigma(0)^\top\right).$$

Moreover, $\nabla\sigma(0)$ can be equivalently written as $\nabla\sigma(0) = (P_{T_{\mathcal{M}}(\bar{x})} \nabla_{\mathcal{M}} F_{\mathcal{M}}(\bar{x}) P_{T_{\mathcal{M}}(\bar{x})})^\dagger$.

The conclusion of this theorem is surprising: although the sequence x_k never reaches the manifold, the limiting distribution of $\sqrt{k}(\bar{x}_k - \bar{x})$ is supported on the tangent space $T_{\mathcal{M}}(\bar{x})$. Thus asymptotically, the "directions of nonsmoothness," which are normal to $\mathcal{M}$, are quickly "averaged out." When $\|G_{\alpha_k}(x_k, v_k)\|$ is bounded away from 0 for all k, this means that x_k must oscillate across the manifold, instead of approaching it from one direction.

5.6.1 Asymptotic normality in nonlinear programming

As a simple illustration of Theorem 5.6.1, we now spell out the consequence for the stochastic projected gradient method for stochastic nonlinear programming, already discussed in Example 5.2.1. Namely, consider the problem (5.2.1) and let $\bar{x}$ be a local minimizer. Suppose that g_i are C^3-smooth near $\bar{x}$ and f takes the form $f(x) = \mathbb{E}_{z\sim\mathcal{P}} f(x, z)$ for some probability distribution $\mathcal{P}$ and each function $f(\cdot, z)$ is C^3-smooth near $\mathcal{X}$. Consider the following stochastic projected gradient method:

$$\begin{aligned} &\text{Sample: } z_k \sim P \\ &\text{Update: } x_{k+1} \in P_{\mathcal{X}}(x_k - \alpha_k \nabla f(x_k; z_k)). \end{aligned} \tag{5.6.2}$$

In order to understand the asymptotics of the algorithm, as in Example 5.2.1, let $\bar{y}$ be the Lagrange multiplier vector and suppose that LICQ and strict complementarity holds. Suppose moreover the second-order sufficient conditions: there exists $\mu > 0$ such that

$$w^\top \left[\nabla^2_{xx}\mathcal{L}(\bar{x}, \bar{y})\right] w \geq \mu\|w\|^2 \qquad \text{for all } w \in T_{\mathcal{M}}(\bar{x}). \tag{5.6.3}$$

Note that, as explained in Example 5.3.1, this condition is simply the requirement that the covariant Hessian of $f := f_0 + \delta_{\mathcal{X}}$

$$\nabla^2_{\mathcal{M}} f(\bar{x}) = P_{T_{\mathcal{M}}(\bar{x})} \nabla^2_{xx}\mathcal{L}(x^\star, y^\star) P_{T_{\mathcal{M}}(\bar{x})}$$

is positive definite on $T_{\mathcal{M}}(\bar{x})$. Finally, to ensure our noise sequence

$$\begin{aligned} \nu_k &= \nabla f(x_k; z_k) - \nabla f(x_k) \\ &= \underbrace{\nabla f(\bar{x}; z_k) - \nabla f_0(\bar{x})}_{=:\nu_k^{(1)}} + \underbrace{(\nabla f(x_k; z_k) - \nabla f(\bar{x}; z_k) + \nabla f(\bar{x}) - \nabla f(x_k))}_{=:\nu_k^{(2)}(x_k)}, \end{aligned}$$

satisfies Assumptions E and P, we assume the stochasticity is sufficiently well-behaved:

(R) **(Stochastic Gradients)** As a function of x, the fourth moment

$$x \in \mathcal{X} \mapsto \mathbb{E}_{z\sim\mathcal{P}}[\|\nabla f(x;z) - \nabla f(x)\|^4]$$

is bounded on bounded sets. Moreover, there exists $C > 0$ such that

$$\mathbb{E}_{z\sim\mathcal{P}}[\|\nabla f(x;z) - \nabla f(\bar{x};z)\|^2] \leq C\|x - \bar{x}\|^2 \qquad \text{for all } x \in \mathcal{X}.$$

Finally, the gradients $P_{T_{\mathcal{M}}(\bar{x})}\nabla f(\bar{x};z)$ have finite covariance $\Sigma = \mathrm{Cov}(P_{T_{\mathcal{M}}(\bar{x})}\nabla f(\bar{x};z))$.

With these assumptions in hand, we have the following asymptotic normality result for nonlinear programming—a direct corollary of Theorem 5.6.1.

Corollary 5.6.2 (Asymptotic normality in nonlinear programming)**.** *Suppose that LICQ, strict complementary, second-order sufficient conditions, and Assumption* (R) *hold. Suppose that $\gamma \in (\frac{1}{2}, 1)$ and consider the iterates x_k generated by the stochastic projected gradient method* (5.6.2)*. Then if x_k converges to $\bar{x}$ with probability 1, the average iterate $\bar{x}_k = \frac{1}{k}\sum_{i=1}^{k} x_i$ admits the expansion*

$$\sqrt{k}(\bar{x}_k - \bar{x}) = -\frac{1}{\sqrt{k}}\sum_{i=1}^{k} U(U^\top \nabla^2_{xx}\mathcal{L}(\bar{x},\bar{y})U)^{-1}U^\top v_i^{(1)} + o_P(1),$$

where the columns of U form an orthonormal basis of $T_{\mathcal{M}}(\bar{x})$. Consequently, asymptotic normality holds:

$$\sqrt{k}(\bar{x}_k - \bar{x}) \xrightarrow{d} N\left(0, \nabla\sigma(0)\cdot \mathrm{Cov}(\nabla f(\bar{x};z))\cdot\nabla\sigma(0)^\top\right),$$

where $\nabla\sigma(0) = (P_{T_{\mathcal{M}}(\bar{x})}\nabla^2_{xx}\mathcal{L}(\bar{x},\bar{y})P_{T_{\mathcal{M}}(\bar{x})})^\dagger$.

As stated in the introduction, this appears to be the first asymptotic normality guarantee for the standard stochastic projected gradient method in general nonlinear programming problems with C^3 data, even in the convex setting. Finally we note that even for simple optimization problems, dual averaging procedures can achieve suboptimal convergence [16]. This is surprising since such methods identify the active manifold [94] (also see [16, Section 4.1]), while projected stochastic gradient methods do not.

Example 5.6.1. *It is instructive to look at three problem formulations for sparse recovery:*

$$\min_{x} \; \mathbb{E}[f(x,z)] + \lambda\|x\|_1, \tag{regularized}$$

$$\min_{\|x\|_1 \le A} \; \mathbb{E}[f(x,z)], \tag{l_1 constraint}$$

$$\min_{|supp(x)| \le s} \; \mathbb{E}[f(x,z)]. \tag{l_0 constraint}$$

Problem (regularized) *is typically solved by the stochastic proximal algorithm, while* (l_1 constraint) *and* (l_0 constraint) *are solved by the stochastic projected gradient method. Both methods are trivially examples of the algorithm* (5.5.4) *that we have studied in the section. Let us now look at the asymptotic covariance of these methods corresponding to the three problems. To this end, let $x^\star$ denote the optimal solution for the three problems and suppose that $\|x^\star\|_1 = A$ and $|supp(x^\star)| = s$. Without loss of generality suppose moreover $supp(x^\star) = \{1, \ldots, s\}$. In all cases, under the regularity conditions discussed in the section, the asymptotic covariance of the average iterate is*

$$\nabla\sigma(0) \cdot \mathrm{Cov}(\nabla f(x^\star, z)) \cdot \nabla\sigma(0)^\top.$$

Thus the only distinction is in Jacobian of the solution map $\nabla\sigma(0)$. It is straightforward to see that the active manifold (under strict complementarity) for (regularized) *and* (l_0 constraint) *is*

$$\mathcal{M}_{1,3} = \mathbb{R}^s \times \{0\}^{d-s},$$

while the active manifold for (l_1 constraint) *is*

$$\mathcal{M}_2 = \mathcal{M}_{1,3} \cap \left\{ x : \sum_{i=1}^{s} |x_i| = A \right\}.$$

Because the active manifold $\mathcal{M}$ in all cases is piecewise linear, an application of Theorem 5.3.2 yields the expression:

$$\nabla\sigma(0) = (P_{T_{\mathcal{M}}(x^\star)}\mathbb{E}[\nabla^2 f(x^\star, z)]P_{T_{\mathcal{M}}(x^\star)})^\dagger.$$

For the problem (regularized) *and* (l_0 constraint)*, the tangent space is simply* $T_{\mathcal{M}_{1,3}}(x^\star) = \mathbb{R}^s \times \{0\}^{d-s}$*, while for* ($l_1$ constraint)*, the tangent space is smaller*

$$T_{\mathcal{M}_2}(x^\star) = \left\{ v \in T_{\mathcal{M}_{1,3}}(x^\star) : \sum_{i=1}^{s} \text{sign}(x_i^\star) v_i = 0 \right\}.$$

In particular, the asymptotic covariance corresponding to (l_1 constraint) *is no larger in the Loewner order than that of* (regularized) *and* (l_0 constraint). *Consequently, the formulation* (l_1 constraint) *may be preferable when* $A = \|x^\star\|_1$ *is known.*

5.7 Asymptotic optimality of SAA and SFB

In this section, we show that the asymptotic covariance in (5.4.4) is the best possible among all estimators of $\bar{x}$, and therefore both SAA and SFB are asymptotically optimal. Namely, we will lower-bound the performance of *any* estimation procedure for finding a solution of an adversarially-chosen sequence of small perturbations of the target problem. In order to specify this sequence, define the set

$$\mathcal{G} := \{g \colon \mathcal{Z} \to \mathbb{R}^d : \mathop{\mathbb{E}}_{z\sim\mathcal{P}}[g(z)] = 0, \ \mathop{\mathbb{E}}_{z\sim\mathcal{P}}\|g(z)\|^2 < \infty\}.$$

Fix now a function $g \in \mathcal{G}$ and an arbitrary C^3-smooth function $h \colon \mathbb{R} \to [-1, 1]$ such that its first three derivatives are bounded and $h(t) = t$ for all $t \in [-1/2, 1/2]$. Now for each $u \in \mathbb{R}^d$, define a new probability distribution $\mathcal{D}^u$ whose density is given by

$$d\mathcal{P}_u(z) := \frac{1 + h(u^\top g(z))}{C(u)}\, d\mathcal{P}(z), \tag{5.7.1}$$

where $C(u)$ is the normalizing constant $C(u) := 1 + \int h(u^\top g(z))\, d\mathcal{P}(z)$. Thus each vector $u \in \mathbb{R}^d$ induces the perturbed problem

$$0 \in L(x,u) + H(x) \qquad \text{where} \qquad L(x,u) = \mathop{\mathbb{E}}_{z\sim\mathcal{P}_u} A(x,z). \tag{5.7.2}$$

Reassuringly, the following lemma shows that map $(x,u) \mapsto L(x,u)$ is C^1 near $(\bar{x}, 0)$. All proofs of results in this section appear in Section 7.2 of the supplement.

Lemma 5.7.1. *The map* $(x,u) \mapsto L(x,u)$ *is* C^1 *near* $(\bar{x},0)$ *with partial derivatives*

$$\nabla_x L(\bar{x},0) = \nabla A(\bar{x}) \qquad \textit{and} \qquad \nabla_u L(x,0) = \mathop{\mathbb{E}}_{z\sim\mathcal{P}} A(\bar{x},z) g(z)^\top.$$

The family of problems (5.7.2) would not be particularly useful if their solution would vary wildly in u. On the contrary, the following lemma shows that for all small u, each problem (5.7.2) admits a unique solution in U, which moreover varies smoothly in u. We will use the following standard notation. A map $\sigma(\cdot)$ is called a *localization* of a set-valued map $F(\cdot)$ around a pair $(\bar{u},\bar{v}) \in \operatorname{gph} F$ if the two sets, $\operatorname{gph}\sigma$ and $\operatorname{gph} F$, coincide locally around $(\bar{u},\bar{v})$.

Lemma 5.7.2 (Derivative of the solution map)**.** *The solution map*

$$S(u) = \{x : 0 \in L(x,u) + H(x)\}.$$

admits a single-valued localization $s(\cdot)$ *around around* $(0,\bar{x})$ *that is differentiable at* 0 *with Jacobian*

$$\nabla s(0) = -\nabla\sigma(0) \cdot \mathop{\mathbb{E}}_{z\sim\mathcal{P}} [A(\bar{x},z) g(z)^\top].$$

In light of Lemma 5.7.2, for all small u, we define the solution $\bar{x}_u := s(u)$. The following theorem provides an asymptotic lower bound on the performance of any estimator when applied to the problems within our parametric family. We let $\mathbb{E}_{P_u^k}$ denote the expectation with respect to k i.i.d. observations $z_i \sim \mathcal{P}_u$.

Theorem 5.7.3 (Local minimax)**.** *Let $\mathcal{L}\colon \mathbb{R}^d \to [0,\infty)$ be symmetric, quasiconvex, and lower semicontinuous, let $\widehat{x}_k\colon \mathcal{Z}^k \to U$ be a sequence of estimators, and set $g(z) := A(\bar{x}, z) - A(\bar{x})$. Then the inequality*

$$\sup_{\mathcal{I}\subset\mathbb{R}^d,\, |\mathcal{I}|<\infty} \liminf_{k\to\infty} \max_{u\in\mathcal{I}} \mathbb{E}_{P^k_{u/\sqrt{k}}}[\mathcal{L}(\sqrt{k}(\widehat{x}_k - \bar{x}_{u/\sqrt{k}}))] \geq \mathbb{E}[\mathcal{L}(Z)] \tag{5.7.3}$$

holds, where $Z \sim \mathsf{N}(0, \nabla\sigma(0)\cdot \mathrm{Cov}(A(\bar{x}, z))\cdot \nabla\sigma(0)^\top)$.

In particular, applying Theorem 5.7.3 with quadratics $\mathcal{L}$ yields a lower bound on the achievable covariance among any estimator. We will now show that both SAA and SFB fulfill (5.7.3) with equality, and therefore in a precise sense *asymptotically minimax optimal*. Note that we already know that the asymptotic covariance $\sigma(0)\cdot \mathrm{Cov}(A(\bar{x}, z))\cdot \nabla\sigma(0)^\top$ is achieved by both SAA (Theorem 5.4.1) and SFB (Theorem 5.6.1) when applied to the *fixed problem* $u = 0$. It remains therefore to argue that $\sqrt{k}(\widehat{x}_k - \bar{x}_u)$ along the perturbed sequence of problems is asymptotically independent of u. This is the content of the following theorem.

Theorem 5.7.4 (Tightness of SAA)**.** *Under the same assumptions as Theorem 5.4.1, the sample average approximation estimator $\hat{x}_k := x_k$ satisfies* (5.7.3) *with equality for any bounded continuous function $\mathcal{L}\colon \mathbb{R}^d \to [0,\infty)$.*

SFB enjoys completely analogous results, which we summarize next.

Theorem 5.7.5 (Tightness of SFB)**.** *Suppose the same setting as Theorem 5.6.1 and that $v_i^{(1)} = A(\bar{x}, z_i) - A(\bar{x})$ with $z_i \overset{\text{iid}}{\sim} \mathcal{P}$ and such that $\mathbb{E}\|v_i^{(1)}\|^2 < \infty$. Then the average* *iterate* $\hat{x}_k = \frac{1}{k}\sum_{i=1}^{k} x_i$ *satisfies* (5.7.3) *with equality for any bounded continuous function* $\mathcal{L}\colon \mathbb{R}^d \to [0,\infty)$.

Theorem 5.7.6 (Tightness of SFB for nonlinear programming)**.** *Under the same assumptions as Corollary 5.6.2, the average iterate* $\hat{x}_k := \frac{1}{k}\sum_{i=1}^{k} x_i$ *satisfies* (5.7.3) *with equality for any bounded continuous function* $\mathcal{L}\colon \mathbb{R}^d \to [0,\infty)$.

We note that asymptotic optimality of SFB for smooth problems was proved in [91, Theorem 5.6], and the proof we present of the three theorems above is an adaptation of the argument therein.

CHAPTER 6

A LOCAL NEARLY LINEARLY CONVERGENT FIRST-ORDER METHOD

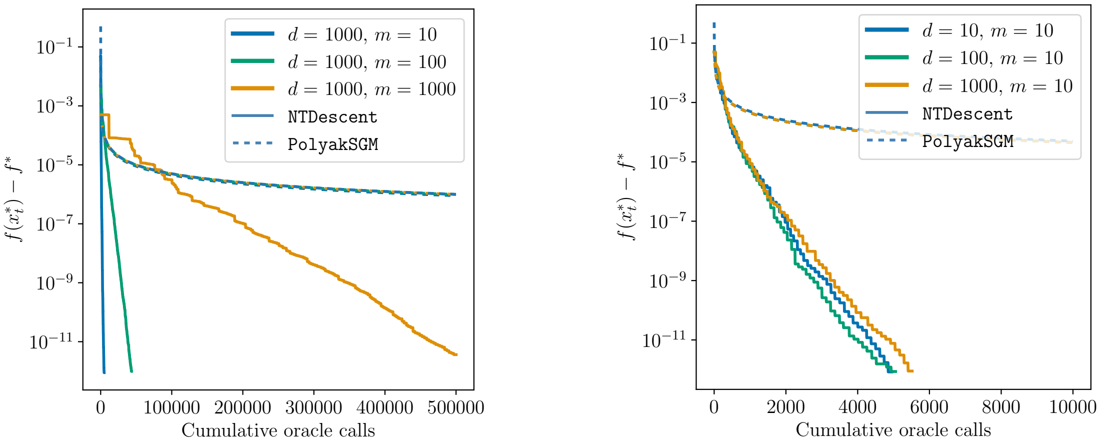
6.1 Introduction

Slow sublinear convergence of first-order methods in nonsmooth optimization is often illustrated with the following simple strongly convex function:

$$f(x) = \max_{1\leq i\leq m} x_i + \frac{1}{2}\|x\|^2 \qquad \text{for some } m \leq d \text{ and all } x \in \mathbb{R}^d. \tag{6.1.1}$$

For example, consider the subgradient method applied to f, which generates iterates x_k. Since f is strongly convex, classical results dictate that $f(x_k) - \inf f = O(k^{-1})$. On the other hand, under proper initialization and an adversarial first-order oracle, there is a matching lower bound for the first m iterations: $f(x_k) - \inf f \geq (2m)^{-1}$ for all $k \leq m$; see [95,96]. Beyond the subgradient method, the lower bound also holds for any algorithm whose kth iterate lies within the linear span of the initial iterate and past $k-1$ computed subgradients. Thus, one must make more than m first-order *oracle calls* to f, i.e., function and subgradient evaluations, before possibly seeing improved convergence behavior.

While such methods make little progress when $k \leq m$, this behavior may or may not continue for $k \gg m$. On one extreme, the subgradient method continues to converge slowly even when equipped with the popular Polyak stepsize (`PolyakSGM`) [97]; see dashed lines in Figure 6.1. On the opposite extreme, more sophisticated algorithms such as the center of gravity method or the ellipsoid method converge linearly, but their complexity scales with the dimension of the problem, a necessary consequence of the linear rate of convergence; see the discussion in [96, Chapter 2].



(a) (b)

Figure 6.1: Comparison of `NTDescent` with `PolyakSGM` on (6.1.1). Left: we fix d and vary m; Right: we fix m and vary d. For both algorithms, the value $f(x_t^*)$ denotes the best function value seen after t oracle evaluations.

A natural question is whether a first-order method exists whose behavior lies between these two extremes, at least for nonsmooth functions f satisfying regularity conditions at local minimizers. Regularity conditions often take the form of growth – linear or quadratic – away from minimizers. Well-known results show that subgradient methods converge linearly on nonsmooth functions with linear (also called *sharp*) growth [97]. On the other hand, in smooth convex optimization, quadratic growth entails linear convergence of gradient methods. However, to our knowledge, no parallel result exists for nonsmooth functions with quadratic growth. Thus, in this chapter, we ask

> is there a locally nearly linearly convergent method for nonsmooth functions with quadratic growth whose rate of convergence and region of rapid local convergence solely depends on f?

Let us explain the qualifiers “nearly” and “solely depends on f.” First, the qualifier “nearly” signifies that the method locally achieves a function gap of size ε using at most, say, $O(C_f \log^3(1/\varepsilon))$ first-order oracle evaluations of f, where C_f depends on f. Second, the qualifier "solely depends on the function" signifies that C_f and the size of the region of local convergence do not depend on the dimension of the problem but instead depend only on the function f through intrinsic quantities, such as Lipschitz and quadratic growth constants.

In this chapter, we positively answer the above question for a class of nonsmooth optimization problems with quadratic growth. The method we develop is called *Normal Tangent Descent* (`NTDescent`). We formally describe `NTDescent` in Section 6.1.7. For now, we illustrate the performance of `NTDescent` on f from (6.1.1) in Figure 6.1. In both plots, we see `NTDescent` improves on the performance of `PolyakSGM`, measured in terms of oracle calls, which is a fair basis for comparison since both `PolyakSGM` and `NTDescent` perform a similar amount of computation per oracle call. Figure 6.1b also shows that the performance of `NTDescent` is dimension independent. We highlight that `NTDescent` achieved this performance without any parameter tuning. Indeed, our central theoretical guarantees for `NTDescent` (Theorem 6.1.1) do not require the user to set any parameters.

The problem class on which `NTDescent` succeeds consists of locally Lipschitz nonsmooth functions with quadratic growth and a certain *smooth substructure* at local minimizers. Importantly, we do not assume the problems under consideration are convex, though convexity entails improved guarantees. Two example classes with such smooth substructure include (i) "generic" semialgebraic functions and (ii) properly C^p decomposable loss functions satisfying strict complementarity and quadratic growth conditions [5]. A semialgebraic function is one whose graph is the finite union of intersections of polynomial inequalities. Semialgebraic functions (more generally *tame* [98] functions) model most problems of interest in applications. When f is semialgebraic, we will show that for a full Lebesgue measure set of $w \in \mathbb{R}^d$, the tilted function $f_w \colon x \mapsto f(x) + w^\top x$ has quadratic growth and the desired smooth substructure at each local minimizer, explaining the qualifier "generic." This fact follows from combining results of [7, 25]. On the other hand, a properly C^p decomposable function is one that decomposes near local minimizers as a composition of a positively homogeneous convex function with a smooth mapping that maps the minimizer to the origin. Decomposable functions appear often in practice, e.g., in eigenvalue and data fitting problems. An important subclass of decomposable functions consists of so-called "max-of-smooth" functions, which are the maximum of finitely many smooth functions that satisfy certain regularity conditions at minimizers, e.g., f in (6.1.1).

The precise smooth substructure used in this chapter was recently identified in [25], where it was shown to be available in decomposable and generic semialgebraic problems. Since it is available in many problems of interest, throughout this introduction, we call this combination of quadratic growth and smooth substructure *typical structure* and call functions possessing this combined structure *typical.* We present the formal structure in Section 6.3. At the heart of this structure is a distinguished smooth manifold $\mathcal{M}$ – called the *active manifold* – containing a local minimizer of interest. We formally define the active manifold concept in Definition 2.4.1, but at a high level, the two crucial characteristics are that (i) along the manifold, the function f is smooth and (ii) normal to the manifold, the function grows sharply. For example, Figure 6.2 depicts the nonsmooth function $f(u, v) = u^2 + |v|$ for which the u-axis plays the role of $\mathcal{M}$. Section 6.1.5 will examine this function and explain how we use its typical structure in `NTDescent`. This example also has the smooth substructure developed in several seminal works in the optimization literature, including those found in work on identifiable surfaces [1], partly smooth functions [2], $\mathcal{VU}$-structures [3, 99], and minimal identifiable sets [6]. However, crucial to the analysis of `NTDescent` are two further properties introduced in [25], called *strong* (a)-*regularity* and ($b_{\leq}$)-*regularity*. Strong (a)-regularity roughly states that the function is smooth in tangent directions to the manifold up to an error term, which is linear in the distance to the manifold. On the other hand, ($b_{\leq}$)-regularity is a one-sided uniform semismoothness [100] property that holds automatically when f is (weakly) convex. Both properties hold for the function in Figure 6.2 and for the function in (6.1.1), where the active manifold is the subspace in which the first m variables take on the same value: $\mathcal{M} = \{x \in \mathbb{R}^d \colon x_1 = x_2 = \ldots x_m\}$.

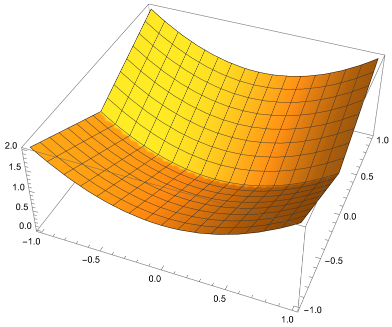


Figure 6.2: The function $f(u, v) = u^2 + |v|$ has typical structure.

Before turning to the description of `NTDescent`, we point out that similar smooth substructure has been used in the analysis first-order methods in nonsmooth optimization, most famously for functions with $\mathcal{VU}$-structure [3,99] and more recently for max-of-smooth functions.[1] For $\mathcal{VU}$ functions, so-called “bundle-methods,” [101,102] which possess an inner-outer loop structure, have been shown to converge superlinearly with respect to the number of outer-loop steps [99]; see also the survey [103]. These methods have excellent empirical performance, but a complete account of their inner-loop complexity remains elusive. On the other hand, in a recent work, Han and Lewis proposed a first-order method – Survey Descent – that converges linearly on certain strongly convex max-of-smooth objectives, stepping beyond the classical smooth setting [104]. The method shows favorable performance beyond the max-of-smooth class, e.g., on cer-

[1] Though they also benefit from smooth substructure, *proximal-methods* do not fall within the oracle model of first-order methods considered in this chapter. Thus, we omit them from our discussion.

tain eigenvalue optimization problems, but no theoretical justification for this success is available. We discuss Survey Descent in more detail in Section 6.7.1. We now motivate `NTDescent`.

6.1.1 Motivation: Goldstein's conceptual subgradient method

To motivate `NTDescent` and the role of smooth substructure, let us set the stage: consider the nonsmooth optimization problem:

$$\text{minimize}_{x\in\mathbb{R}^d}\ f(x),$$

where $f\colon \mathbb{R}^d \to \mathbb{R}$ is a locally Lipschitz function, which is not necessarily convex. The algorithm developed in this chapter assumes *first-order oracle access* to f [95, 96, 105]. In particular, at every $x \in \mathbb{R}^d$, we must be able to evaluate $f(x)$ and retrieve an element of the *Clarke subdifferential* $\partial f(x)$. Informally, the Clarke subdifferential is comprised of convex combinations of limits of gradients taken at nearby points; a formal definition appears in Section 5.2. The Clarke subdifferential reduces to the familiar objects in classical settings. For example, when f is C^1, the Clarke subdifferential reduces to the singleton mapping $\{\nabla f\}$. In addition, when f is convex, the Clarke subdifferential reduces to the subdifferential in the sense of convex analysis.

The starting point of this chapter is the classical conceptual subgradient method of Goldstein [106]. The core object in this method is the Goldstein subdifferential:

$$\partial_\sigma f(x) := \operatorname{conv}\left(\bigcup_{y\in\overline{B}_\sigma(x)} \partial_c f(y)\right) \qquad \text{for all } x \in \mathbb{R}^d \text{ and } \sigma > 0. \tag{6.1.2}$$

This subdifferential is the convex hull of all Clarke subgradients of f taken at points inside the ball of radius σ. Its importance arises from the following descent property proved in [106]: fix $\sigma > 0$ and $x \in \mathbb{R}^d$ and let w denote the minimal norm element of $\partial_\sigma f(x)$. Then

$$f\left(x - \sigma\frac{w}{\|w\|}\right) \leq f(x) - \sigma\|w\| \qquad \text{if } w \neq 0. \tag{6.1.3}$$

This property motivates Goldstein's conceptual subgradient method, which iterates:

$$x_{k+1} = x_k - \sigma\frac{w_k}{\|w_k\|} \qquad \text{where} \qquad w_k = \operatorname*{argmin}_{w \in \partial_\sigma f(x_k)} \|w\|. \tag{6.1.4}$$

This algorithm is remarkable since it is a descent method for any Lipschitz function and even converges at a sublinear rate. Indeed, a quick appeal to (6.1.3) yields

$$\min_{k=0,\ldots,K-1} \|w_k\| \leq \varepsilon \qquad \text{holds when} \qquad K \geq \frac{f(x_0) - \min f}{\sigma\varepsilon}.$$

While this exact variant of the Goldstein method is not necessarily implementable, recent work has devised approximate versions with similar sublinear convergence properties [107, 108].

The algorithm introduced in this chapter approximately implements the method (6.1.4). This chapter aims to prove that the method is locally nearly linearly convergent on typical nonsmooth functions. We must resolve two issues for this problem class to develop such a method. First, we must develop rapidly convergent algorithms that approximately compute the minimal norm element of the Goldstein subdifferential. Second, we must devise an appropriate regularity property that ensures the proposed method converges nearly linearly. We will discuss both of these properties in turn, beginning with a regularity property that relates the decrement in (6.1.3) to the function gap.

6.1.2 Linear convergence via a gradient inequality

Observe that if the bound

$$\sigma \|w_k\| \geq \eta(f(x_k) - \min f)$$

holds for some $\eta > 0$ and all $k > 0$, then the Goldstein method (6.1.4) converges linearly to a minimizer of f. A potential issue with this inequality is that the vector w_k is zero whenever σ is larger than the distance of x_k to the nearest critical point of f; thus, the algorithm may stall whenever x_k is near enough to a minimizer. Thus, we propose a relaxation of the property that allows σ to depend on x_k.

Indeed, we will provide conditions under which the following bound holds near a local minimizer $\bar{x}$ of f: there exists a constant $\eta > 0$ and a function $\sigma \colon \mathbb{R}^d \to \mathbb{R}_+$ such that for all x near $\bar{x}$, we have

$$\sigma(x)\mathrm{dist}(0, \partial_{\sigma(x)} f(x)) \geq \eta(f(x) - f(\bar{x})). \tag{6.1.5}$$

throughout, we will refer to this bound as a *gradient inequality*, due to its similarity to the Kurdyka-Łojasiewicz (KL) gradient inequality [46]. The KL inequality requires that a suitable nonlinear reparameterization $\psi \colon \mathbb{R} \to \mathbb{R}$ of the function gap is bounded by the minimal norm Clarke subgradient for all x near $\bar{x}$:

$$\mathrm{dist}(0, \partial_c f(x)) \geq \psi(f(x) - f(\bar{x})).$$

In recent years, the KL inequality has played a key role in establishing convergence and rates of convergence for proximal methods in nonsmooth optimization and in continuous time analogs of the subgradient method; see, e.g., [46, 109–112].

To illustrate, let us specialize to the semialgebraic setting, where the desingularization function ψ is known to take the form $\psi(r) = r^\theta$ for $\theta \in [0, 1)$. The work [113, Theorem 2] initiated the study of convergence of proximal methods in this setting, showing that the proximal point method asymptotically converges to its limit point, which is critical but not necessarily optimal. The method convergence in finitely many steps when

$\theta = 0$, locally converges linearly when $\theta \in (0, 1/2]$, and locally converges at the rate $k^{\frac{-(1-\theta)}{2\theta-1}}$ when $\theta \in (1/2, 1)$. Further works such as [109, 111] generalized the techniques to related proximal methods. Passing to continuous time, one is interested in the convergence of the trajectory of subdifferential inclusion satisfying $\dot{x}(t) \in -\partial_c f(x(t))$ at almost every t. Here, the rates of convergence exactly parallel those in the proximal methods as shown in [114, Theorem 4.7].[2] In contrast to the proximal and continuous-time settings, we do not know whether the KL inequality alone allows one to design a locally linearly convergent discrete-time subgradient method, except in the setting where $\theta = 0$ (i.e., f is *sharp*) and f is convex [97] or *weakly convex* [116]; weakly convex functions form a broad class of nonconvex functions that includes all compositions of Lipschitz convex functions with smooth mappings. When $\theta > 0$, to the best of our knowledge, the best rate proved in the literature for any subgradient type method is $k^{\frac{-(1-\theta)}{2\theta}}$ [117]; this result is only known to hold for convex functions.[3]

A well-known property of the KL inequality is its prevalence: it holds at each critical point of an arbitrary lower-semicontinuous semialgebraic function f [46]. We will show that the gradient inequality (6.1.5) is also prevalent in the sense that it holds for the problems above with typical structure. In this way, the conceptual method (6.1.4) with varying $\sigma_k := \sigma(x_k)$ will locally converge linearly on such problems. The reader may wonder whether we can or must find the precise value $\sigma(x_k)$. We will show that for typical problems, an appropriate σ_k may be found through a line search procedure.

[2] These rates were shown only for "lower-C^2" semialgebraic losses, but extend to locally Lipschitz semialgebraic functions via the semialgebraic "chain rule" proved in [115].

[3] The results stated in [117] pertain to functions with Hölder growth; thus, to prove the results stated in the paragraph, we must use the following known fact: functions satisfying the KL inequality with exponent θ have Hölder growth with exponent $1/(1-\theta)$, which follows from the proof of [118, Theorem 3.7].

6.1.3 Approximately implementing Goldstein's method

The gradient inequality ensures that the conceptual Goldstein method converges linearly, provided the stepsize σ is chosen adaptively. To move beyond the conceptual setting, we must develop strategies for approximating the minimal norm element of $\partial_\sigma f(x)$ for $\sigma > 0$ and $x \in \mathbb{R}^d$. Suppose we have such a method and denote it by $\texttt{MinNorm}(x, \sigma)$. Then, the method of this chapter iterates:

$$x_{k+1} = x_k - \sigma_k \frac{w_k}{\|w_k\|} \qquad \text{and} \qquad w_k = \texttt{MinNorm}(x_k, \sigma_k) \tag{6.1.6}$$

for an appropriate sequence $\sigma_k > 0$. We will discuss and develop two different implementations of $\texttt{MinNorm}(x, \sigma)$ in this chapter. Given $x \in \mathbb{R}^d$ and $\sigma > 0$, both methods iteratively construct a sequence of Clarke subgradients $g_0, \ldots, g_{T-1}$ taken at points in the ball $\overline{B}_\sigma(x)$ and then output a "small" convex combination $w \in \operatorname{conv}\{g_0, \ldots, g_{T-1}\}$, which satisfies the descent condition

$$f\left(x - \sigma \frac{w}{\|w\|}\right) \leq f(x) - \frac{\sigma}{8}\|w\|. \tag{6.1.7}$$

The oracle complexity of $\texttt{MinNorm}(x, \sigma)$ is then T function/subgradient evaluations, and we hope to ensure that T is relatively small, for example, a constant or at most

$$T = O\left(\log\left(\Delta_{x,\sigma}^{-1}\right)\right) \qquad \text{where } \Delta_{x,\sigma} := \operatorname{dist}(0, \partial_\sigma f(x)).$$

Provided that T is on this order, that f satisfies the gradient inequality (6.1.5), and that σ_k is chosen appropriately, the iterate x_k will satisfy $f(x_k) - f(\bar{x}) \leq \varepsilon$ after at most $O(\log^2(1/\varepsilon))$ iterations, a nearly linear rate of convergence. This complexity ignores the cost of choosing an appropriate stepsize σ_k, but we will show that in typical problems, we can find appropriate σ_k with at most $O(\log(1/\varepsilon))$ function/subgradient evaluations.

We know of two $\texttt{MinNorm}$ type methods in the literature, but their complexity is either too large or useful only in low dimensions problems. For example, the works [107,108] introduced such a method for general locally Lipschitz functions. However, the complexity of the method is $T = O(1/\Delta_{x,\sigma})$ – too large for our purposes. On the other hand, the work [108] also introduced a method tailored to low-dimensional weakly convex functions. However, the method is based on cutting plane techniques, so its complexity scales linearly with dimension: $T = O(d\log(1/\Delta_{x,\sigma}))$.

While existing `MinNorm` methods are slow for general Lipschitz functions, we show that the aforementioned typical structure allows us to develop `MinNorm` methods that accelerate in a neighborhood of the minimizer. Our approach is based on a decomposition of a neighborhood of the minimizer into two regions: one where the method of [107, 108] is applicable, and another region where a novel `MinNorm` method may be applied.

6.1.4 The normal and tangent regions

In this chapter, we use the active manifold $\mathcal{M}$ to split the space of (x, σ) for x nearby the minimizer $\bar{x} \in \mathcal{M}$ into two sets where fast `MinNorm` methods are available. We call the first set the *normal region.* This region consists of points whose normal distance $\mathrm{dist}(x, \mathcal{M})$ is larger than a multiple of the squared tangential distance $\|P_{\mathcal{M}}(x) - \bar{x}\|^2$, together with stepsizes σ proportional to a multiple of the normal distance:

$$\begin{cases} \frac{a_1}{2}\mathrm{dist}(x, \mathcal{M}) \leq \sigma \leq a_1 \mathrm{dist}(x, \mathcal{M}); \\ a_2^2\|P_{\mathcal{M}}(x) - \bar{x}\|^2 \leq \mathrm{dist}(x, \mathcal{M}), \end{cases}$$

for problem dependent constants $a_1, a_2 \in (0, 1)$; see Theorem 6.4.3 for more details. We will show that in this region, we have $\Delta_{x,\sigma} = \Omega(1)$, so the `MinNorm` method of [107,108] terminates with descent in finitely many steps.

On the other hand, we call the second set the *tangent region.* This set consists of points whose squared tangential distance is larger than a multiple of the normal distance, together with stepsizes σ proportional to a multiple of the tangential distance:

$$\begin{cases} \frac{a_2}{2}\|P_{\mathcal{M}}(x) - \bar{x}\| \le \sigma \le a_2\|P_{\mathcal{M}}(x) - \bar{x}\|; \\ \frac{\operatorname{dist}(x,\mathcal{M})}{\sigma} \le 2a_2\|P_{\mathcal{M}}(x) - \bar{x}\|, \end{cases}$$

where a_1 and a_2 are as in the normal region. We will propose a new `MinNorm` method for this region, which terminates rapidly. We note that we provide a range of valid σ rather than a single value in both cases since we aim to estimate σ with a line search.

6.1.5 A simple example

Before rigorously describing the `MinNorm` methods in detail, let us provide intuition on the regions and the principles of the methods through the following simple function of two variables $f(x) = u^2 + |v|$, where $x := (u, v) \in \mathbb{R}^2$. This function has a unique minimizer at $\bar{x} = (0, 0)$. Here, the u-axis is the active manifold $\mathcal{M}$. Along the manifold, f is smooth and grows quadratically, while off the manifold, f grows sharply; see Figure 6.2 for a plot of the function. Figure 6.3 plots the set of x such that there exists $\sigma > 0$ with (x, σ) in the normal and tangent regions for f, respectively (with $a_2 = 1/8$). The manifold $\mathcal{M}$ induces a decomposition of f into smooth $f_{\mathcal{U}}(u, v) = u^2$ and nonsmooth $f_{\mathcal{V}}(u, v) = |v|$ components. In particular, denoting that $x = (u, v)$, we have

$$f(x) = f_{\mathcal{U}}(x) + f_{\mathcal{V}}(x) = \|P_{\mathcal{M}}(x) - \bar{x}\|^2 + \operatorname{dist}(x, \mathcal{M}). \tag{6.1.8}$$

This decomposition shows that $f_{\mathcal{V}}$ is dominant in the normal region, while $f_{\mathcal{U}}$ is dominant in the tangent region. Likewise, as we will argue momentarily, the minimal norm Goldstein subgradient $w_\sigma \in \partial_\sigma f(x)$ satisfies $\|w_\sigma\| \ge \|\nabla f_{\mathcal{V}}(x)\|$ in the normal region, while $\|w_\sigma\| = \Omega(\|\nabla f_{\mathcal{U}}(x)\|)$ in the tangent region. Several consequences follow from this observation. First, in the normal region, the `MinNorm` method of [107, 108] will terminate in finitely many steps due to the lower bound $\|w_\sigma\| \geq 1$. On the other hand, in the tangent region, $\|w_\sigma\|$ can be much smaller, so we must introduce a new method to generate descent. Finally, assuming these approximations are accurate, the gradient inequality (6.1.6) quickly follows: in the normal region, we have

$$\sigma\|w_\sigma\| = \Omega(\mathrm{dist}(x, \mathcal{M})) = \Omega(f(x)),$$

while in the tangent region, we have

$$\sigma\|w_\sigma\| = \Omega(\|P_{\mathcal{M}}(x) - \bar{x}\|\|\nabla f_{\mathcal{U}}(x)\|) = \Omega(\|P_{\mathcal{M}}(x) - \bar{x}\|^2) = \Omega(f(x)).$$

Though it follows from immediate calculations in this example, in the more general setting, the following consequence of quadratic growth will be crucial in establishing a similar bound: $\|\nabla f_{\mathcal{U}}(x)\| = \Theta(\|P_{\mathcal{M}}(x) - \bar{x}\|)$.

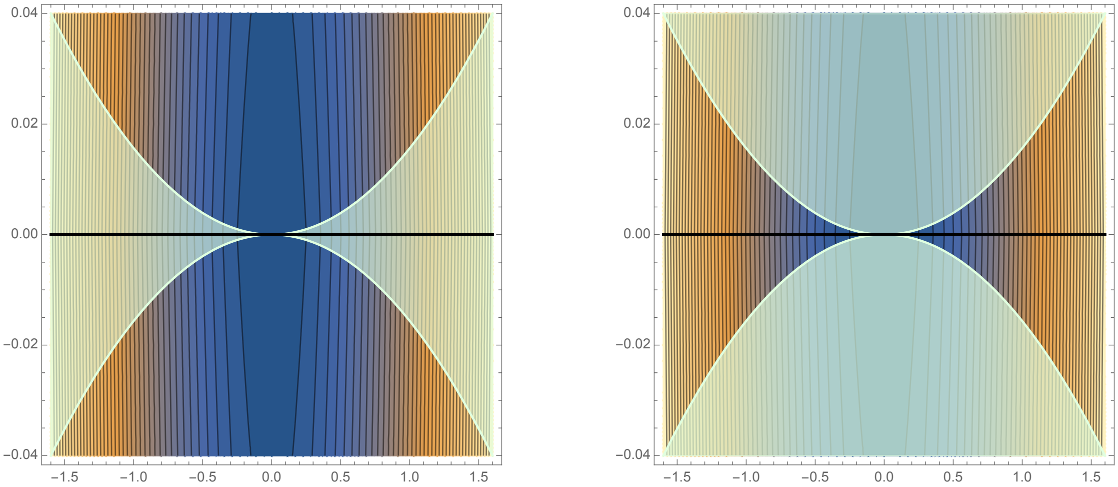


Figure 6.3: Contour plots for $f(u, v) = u^2 + |v|$ together with $\mathcal{M}$, shown in black. The light green regions consist of $x := (u, v) \in \mathbb{R}^2$ such that there exists $\sigma > 0$ with (x, σ) in the tangent (left) and normal (right) regions.

Now, to lower bound $\|w_\sigma\|$ we use the following fact: $\nabla f_{\mathcal{U}}(u, v)$ is tangent to $\mathcal{M}$, while $\nabla f_{\mathcal{V}}(u, v)$ is normal to $\mathcal{M}$ when $v \neq 0$. Thus, to lower bound $\|w_\sigma\|$ in the normal region, we will lower bound the size of the normal component of w_σ. Indeed, since $\sigma < \operatorname{dist}(x, \mathcal{M})$, all points $x' \in B_\sigma(x)$ are on the same side of $\mathcal{M}$. Therefore, the normal component of w_σ is an average of *identical* gradients $\nabla f_{\mathcal{V}}(x') = \nabla f_{\mathcal{V}}(x)$. Likewise, in the tangent region, we lower bound the tangent component of w_σ. Indeed, since $\sigma < \|\bar{x} - P_{\mathcal{M}}(x)\|/8$, the projection onto $\mathcal{M}$ of all points $x' \in B_\sigma(x)$ are on the same side of the origin. Thus, the tangent component of w_σ is an average of nearly identical gradients $\nabla f_{\mathcal{U}}(x') \approx \nabla f_{\mathcal{U}}(x)$, yielding the lower bound. We prove a more general form of these lower bounds in Lemma 6.4.1 and Lemma 6.4.2, which follow from a similar argument.

Turning to algorithms, we have noted that the `MinNorm` method of [107,108] may be used in the normal region. In the tangent region, we are unsure how to design a method that can quickly recover w_σ. Instead of searching for w_σ directly, we take a slightly different perspective in the tangent region: we seek a vector $g \in \partial_\sigma f(x)$ with "small" normal component, meaning:

$$\|P_N(g)\| = O(\|\nabla f_{\mathcal{U}}(x)\|^2)$$

where N is the normal space to $\mathcal{M}$, i.e., $\mathcal{M}^\perp$. Intuitively, when g has a small normal component, the nonsmooth part $f_{\mathcal{V}}$ minimally changes along a gradient step. On the other hand, if g is sufficiently aligned with $\nabla f_{\mathcal{U}}$, the smooth part $f_{\mathcal{U}}$ decreases at an appropriate rate; we prove this in a more general setting in Lemma 6.5.2.

Why might one expect such a g to be available in the tangent region? The reason is that the gradient of the smooth component is itself a Goldstein subgradient. Indeed, for points near the origin and in the tangent region, the tangential distance is much larger than the normal distance. Thus, the reflection of any point (u, v) across the manifold $\mathcal{M}$ is contained in $B_\sigma(x)$, which immediately implies gradient of the smooth component is an element of Goldstein subdifferential:

$$\nabla f_{\mathcal{U}}(u, v) = \frac{1}{2}\nabla f(u, v) + \frac{1}{2}\nabla f(u, -v) \in \partial_\sigma f(x). \tag{6.1.9}$$

While the inclusion (6.1.9) illustrates one way to construct such a g, we cannot hope for perfect symmetry in general problems.

Instead, a central insight of this chapter is that a similar approximate reflection exists in problems with typical structure. To illustrate, consider Figure 6.4. This figure depicts a point x in the tangent region together with the result of a normalized gradient step:

$$x_+ := x - \sigma\frac{\nabla f(x)}{\|\nabla f(x)\|}.$$

As can be seen from the figure, x_+ is an approximate reflection of x across the u-axis, which "flips the sign" of the nonsmooth component of ∇f: $\nabla f_{\mathcal{V}}(x) = -\nabla f_{\mathcal{V}}(x_+)$. Thus, in this setting, one may "cancel out" the nonsmooth component by a simple averaging:

$$\nabla f_{\mathcal{U}}(x) \approx \frac{1}{2}\nabla f(x) + \frac{1}{2}\nabla f(x_+).$$

While seemingly crude, we will show this strategy generalizes to typical functions. An important distinction with the general setting is that a single averaging step alone will no longer suffice. Nevertheless, we show that by iterating this process, we can geometrically shrink the normal component of the Goldstein gradient, eventually yielding descent.

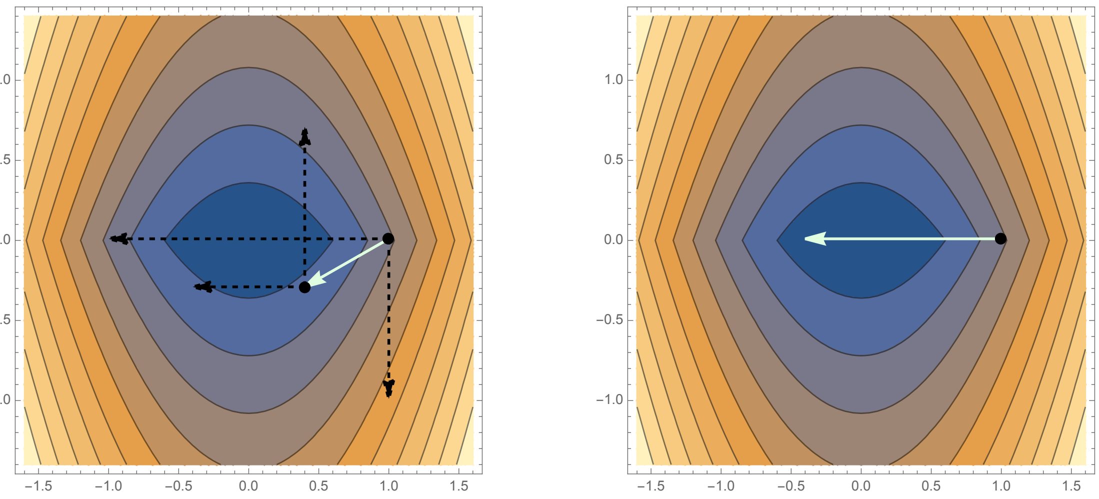


Figure 6.4: Contour plots for $f(u, v) = u^2 + |v|$. Left: The point $x = (1, .1)$ together with the approximate reflection $x_+ = x - .3\frac{\nabla f(x)}{\|\nabla f(x)\|}$ across the u axis. The solid light green arrow is parallel to the negative gradient direction $-\nabla f(x)$. The dashed arrows denote the orthogonal decomposition of $-\nabla f(x)$, respectively $-\nabla f(x_+)$, into the vectors $-\nabla f_{\mathcal{U}}(x)$ and $-\nabla f_{\mathcal{V}}(x)$, respectively $-\nabla f_{\mathcal{U}}(x_+)$ and $-\nabla f_{\mathcal{V}}(x_+)$. From the plot, we see $\nabla f_{\mathcal{V}}(x) = -\nabla f_{\mathcal{V}}(x_+)$. Right: The point x with estimate $-\frac{1}{2}(\nabla f(x) + \nabla f(x_+))$ of the vector $-\nabla f_{\mathcal{U}}(x)$.

6.1.6 Two `MinNorm` methods: `NDescent` and `TDescent`

To generalize the strategy outlined in the previous section, we will prove that the minimal norm Goldstein subgradients of typical problems similarly split into tangent and normal components just as in Section 6.1.5. Then, we introduce two `MinNorm` type methods for "normal" and "tangent" steps.

For (x, σ) in the normal region, we use a small modification of the `MinNorm` type method of [108]. We call this method *Normal Descent* (`NDescent`) and describe it in Algorithm 1. As in the simple example above, we will show that `NDescent` must terminate with an approximately minimal norm Goldstein subgradient in finitely many steps, provided σ lies within an appropriate range. We will show that this subgradient is a descent direction satisfying (6.1.7).

Algorithm 1 `NDescent`(x, g, σ, T)

```
**Set** $g_0 = g$ and $t = 0$.
**while** $T - 1 \geq t$, $\|g_t\| > 0$, and $\frac{\sigma}{8}\|g_t\| \geq f(x) - f\left(x - \sigma \frac{g_t}{\|g_t\|}\right)$ **do**
    Choose any $r$ satisfying $0 < r < \sigma\|g_t\|$.
    Sample $\zeta_t$ uniformly from $\mathbb{B}_r(g_t)$.
    Choose $y_t$ uniformly at random in the segment $\left[x, x - \sigma \frac{\zeta_t}{\|\zeta_t\|}\right]$.
    Choose $\hat{g}_t \in \partial_c f(y_t)$.
    $g_{t+1} = \operatorname{argmin}_{z \in [g_t, \hat{g}_t]} \|z\|_2$.
    $t = t + 1$.
**end while**
**return** $g_t$.
```

We illustrate the principle behind `NDescent` as follows. Suppose we are given a vector $g \in \partial_\sigma f(x)$ not satisfying the descent condition, i.e., with $u := \frac{g}{\|g\|}$, we have

$$f\left(x - \sigma u\right) - f(x) \geq -\frac{\|g\|}{8}.$$

Then by Lebourg mean value theorem [31, Theorem 2.4] (provided that f is differentiable along the line segment between $[x, x']$, which can be ensured by adding a small perturbation to g; we ignore this in our discussion), we may assume that

$$f\left(x - \sigma u\right) - f(x) = \sigma \int_0^1 - \left\langle \nabla f\left(x - \sigma t u\right), u \right\rangle dt = -\sigma \left\langle v, u \right\rangle,$$

where $v := \int_0^1 \nabla f\left(x - tu\right) dt \in \partial_\sigma f(x)$. Consequently, $\langle v, g \rangle \leq \|g\|^2/8$. While it is not possible to compute v, we can compute a *random* element of the Goldstein subdifferential, satisfying the same inequality in expectation. Indeed, defining $v' = \nabla f(y)$ where y is uniformly sampled from the line segment $[x, x - \sigma u]$ (with end points x and $x - \sigma u$), we have $\langle \mathbb{E}_y[v'], g \rangle \leq \|g\|^2/8$. Based on this bound, a quick calculation shows that the minimal norm element g_+ of the line segment $[g, v']$ satisfies the bound

$$\mathbb{E}_y \|g_+\|^2 \leq \|g\|^2 - \frac{\|g\|^4}{16L^2},$$

where L is the Lipschitz constant of function f on the ball $B_{2\sigma}(x)$. Moreover $g_+ \in \partial_\sigma f(x)$. Thus, repeating this process yields a decreasing sequence of Goldstein subgradients, which tend to zero as long as the descent condition is not met. In general, the norms of the subgradients generated by this process decay at a rate of $1/k$. However, we will prove that $\operatorname{dist}(0, \partial_\sigma f(x))$ is bounded below by a fixed constant when (x, σ) is in the normal region described in Section 6.1.4. Consequently, the loop must exist in finite time with descent (with high probability), for otherwise, we will have found a subgradient norm strictly smaller than $\operatorname{dist}(0, \partial_\sigma f(x))$; see Proposition 6.5.1. Readers interested in the formal calculations may consult [107, 108].

On the other hand, for (x, σ) in the tangent region, we develop a new `MinNorm` type method, which likewise relies on an approximate reflection property. We call this method *Tangent Descent* (`TDescent`) and present it in Algorithm 2. Given an input point x, stepsize $\sigma > 0$, and initial subgradient $g_0 \in \partial_c f(x)$, `TDescent` repeats the following steps

$$\text{Choose: } \hat{g}_k \in \partial_c f\left(x - \sigma \frac{g_k}{\|g_k\|}\right);$$

$$\text{Update: } g_{k+1} = \operatorname*{argmin}_{g \in [g_k, \hat{g}_k]} \|g\|,$$

until it achieves descent $f(x - \sigma \frac{g_k}{\|g_k\|}) \leq f(x) - \frac{\sigma}{8}\|g_k\|$ or runs over budget.

Algorithm 2 `TDescent`(x, g, σ, T)

```
Set $g_0 = g$ and $t = 0$.
while $T - 1 \geq t$, $\|g_t\| > 0$, and $\frac{\sigma}{8}\|g_t\|_2 \geq f(x) - f\left(x - \sigma \frac{g_t}{\|g_t\|}\right)$ do
    Choose $\hat{g}_t \in \partial_c f(x - \sigma \frac{g_t}{\|g_t\|})$.
    $g_{t+1} = \operatorname{argmin}_{z \in [g_t, \hat{g}_t]} \|z\|$.
    $t = t + 1$.
end while
return $g_t$.
```

The motivation for this method is that for typical problems, the step $x - \sigma \frac{g_k}{\|g_k\|}$ is locally an approximate reflection across $\mathcal{M}$ that "flips" the normal component of the Goldstein subgradient. Indeed, let $y := P_{\mathcal{M}}(x)$ denote the projection of x onto $\mathcal{M}$ and let $N := N_{\mathcal{M}}(y)$ denote the normal space to $\mathcal{M}$ at y; see Section 5.2 for a precise definition of these concepts. Then we will prove that for all k, we have

$$\langle P_N g_k, \hat{g}_k \rangle \leq -C\|P_N g_k\| + O(\|y - \bar{x}\|^2),$$

for some $C > 0$, provided σ lies within an appropriate range. This inequality ensures that each step of the `TDescent` geometrically decreases the "normal component" of g_k, until we arrive at a Goldstein subgradient with normal component on the order of $O(\|y - \bar{x}\|^2)$; see Section 6.5.2. Moreover, given $g \in \partial_\sigma f(x)$ satisfying

$$\|P_N(g)\| \leq C_3\|y - \bar{x}\|^2$$

for a particular problem dependent constant $C_3 > 0$, we will prove the descent condition

$$f\left(x - \sigma\frac{g}{\|g\|}\right) \leq f(x) - \frac{\sigma\|g\|}{8}$$

holds; see Lemma 6.5.2. Combining these two facts shows that `TDescent` will rapidly terminate with descent.

6.1.7 The `NTDescent` algorithm

We call the main algorithm of this chapter *Normal Tangent Descent* (`NTDescent`) and present it in Algorithm 4. At a high level, the method is an approximate implementation of Goldstein's conceptual subgradient method as in (6.1.6), using `NDescent` and `TDescent` as `MinNorm` type methods. As input it takes three parameters: an initial point x; a sequence of grid-sizes $\{G_k\}$ for the line search on σ; and a sequence of budgets $\{T_k\}$ for the `MinNorm` type methods `NDescent` and `TDescent`. Later, we will show that the user may set $T_k = G_k = k + 1$ for all $k \geq 0$.

Algorithm 3 `linesearch`(x, g, s, G, T)

```
1: Set $v_0 = g$.
2: for $i = 0, \dots, G-1$ do
3:     $\sigma_i = 2^{-(G-i)}$.
4:     $u_i = \texttt{TDescent}(x, v_i, \sigma_i, T)$.
5:     $v_{i+1} = \texttt{NDescent}(x, u_i, \sigma_i, T)$.
6: end for
7: $\tilde{x} := \operatorname{argmin}\{f(x')\colon x' \in \{x\} \cup \{x - \sigma_i \frac{v_{i+1}}{\|v_{i+1}\|} : \sigma_i \leq \frac{\|v_{i+1}\|}{s}, i = 0, \dots, G-1\}\}$.
8: return $\tilde{x}$.
```

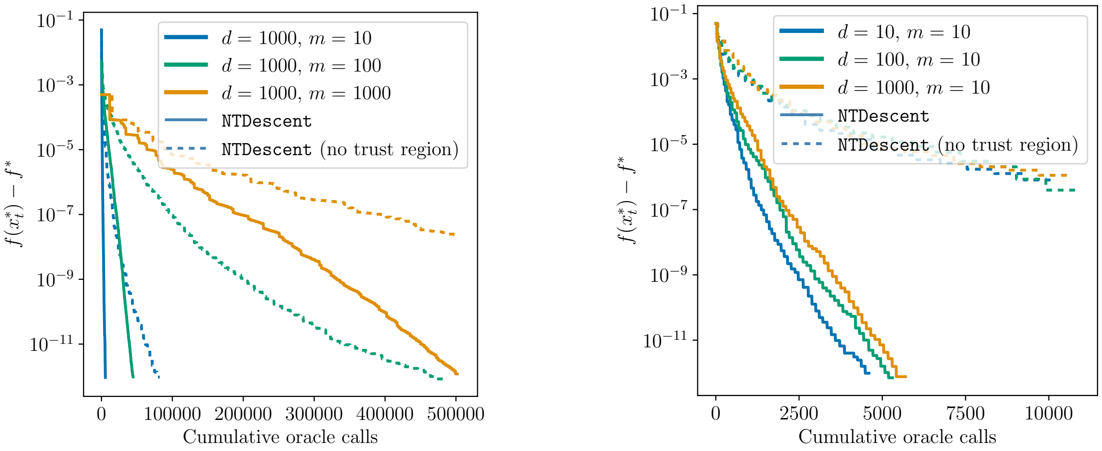
Algorithm 4 `NTDescent`$(x, g, c_0, \{G_k\}, \{T_k\})$

```
Require: $g \neq 0$, $c_0 \in (0, 1]$
1: Set $x_0 = x$ and $g_0 = g$.
2: for $k = 0, 1, \dots$ do
3:     $x_{k+1} = \texttt{linesearch}(x_k, g_k, \max\{\|g_k\|, c_0\|g_0\|\}, G_k, T_k)$.
4:     Choose $g_{k+1} \in \partial_c f(x_{k+1})$.
5: end for
```

The workhorse of `NTDescent` is the line search procedure in Algorithm 3 (`linesearch`). Let us briefly comment on the structure of this method. Lines 2 through 6 of Algorithm 3 implement a line search on σ. Line 7 chooses the Goldstein subgradient that provides the most descent while enforcing the trust-region constraint $\sigma_i \leq \frac{\|v_{i+1}\|}{s}$. Line 7 also ensures the `NTDescent` is a descent method. Within the line search procedure, we evaluate `TDescent` and `NDescent` a total of G times each. Not all calls to `TDescent` and `NDescent` will succeed with descent within the allotted budget T. Still, we will show that for typical problems, at least one will generate sufficient descent provided x_k is close enough to a local minimizer and T is sufficiently large. The reason at least one will succeed with descent is that given any x sufficiently near the solution and parameters G and T sufficiently large, `linesearch` will find a σ such that (x, σ) is in either the normal or tangent region described in Section 6.1.4. The line search allows the possibility that σ is as large as $1/2$, which might force x_{k+1} to leave the region surrounding the minimizer $\bar{x}$. This concern is what motivates the somewhat unusual structure of the line search method wherein the `MinNorm`-type methods are nested. Indeed, on the one hand, the nesting ensures the norms of the Goldstein subgradients $\|v_{i+1}\|$ are decaying as σ_i increases. On the other hand, the trust region constraint ensures that σ_i is not chosen too large, which we need for two technical reasons in our analysis: (i) it prevents x_{k+1} from leaving a small neighborhood around the minimizer where our regularity assumptions hold; (ii) we can only ensure `TDescent` terminates quickly when $\sigma \leq \delta_{\text{Grid}}$, for a certain radius δ_{Grid} defined in Lemma 6.5.7, which may be substantially smaller than $1/2$.

Computationally, it may seem desirable to drop the trust region constraint. Figure 6.5 shows this may not be the case. We suspect the reason is two-fold: First, the trust region constraint allows us to cut off a range of σ from our search, which might otherwise waste oracle calls; indeed, since $\|v_{i+1}\|$ is nonincreasing in i, and σ_i is increasing, once the trust region is violated, it will be violated for all larger i. Second, although we may take longer steps by disabling the trust region constraint, the amount of descent we expect is on the order of $\Omega(\sigma_i\|v_{i+1}\|)$. Thus, since the norms $\|v_{i+1}\|$ are nonincreasing, larger stepsizes σ_i do not necessarily translate to larger descent.

Finally, we comment on our motivation for choosing the scaling $s_k = \max\{\|g_k\|, c_0\|g_0\|\}$ in the trust region constraint. First, note that it is possible to prove, using identical techniques, that the `NTDescent` converges when one replaces s_k by any positive sequence bounded from above and below by positive constants. For our particular choice of s_k, the term $c_0\|g_0\|$ ensures the sequence is bounded below, while the local Lipschitz continuity of f ensures that s_k is bounded above. Second, we wish for the trust region constraint to be unaffected by rescalings of f. Our choice of s_k guarantees scaling invariance since the subgradients of af are simply the subgradients of f scaled by a for any positive constant a. One might introduce other schemes for choosing s, but we did not explore such strategies. Finally, we found that performance of NTDescent is relatively insensitive to the choice of $c_0 > 0$, and any $c_0 \in \{10^{-i} : i = 0, 2, 4, 6\}$ yielded adequate performance; see Figures 6.6e and 6.6f.



(a)

(b)

Figure 6.5: Comparison of NTDescent on Problem (6.1.1) with the trust region constraint in Line 6 of Algorithm 3 removed. Left: we fix d and vary m; Right: we fix m and vary d. We invite the reader to compare these plots with Figure 6.1.

6.1.8 Main convergence guarantees for NTDescent

The main contribution of this chapter is a local, nearly linear convergence rate for NTDescent. The local rate holds under a key structural assumption – Assumption Q – which formalizes the typical structure concept and mirrors the simple function's structure considered in Section 6.1.5. While we formally describe Assumption Q in Section 6.3, for now, we mention that it holds for max-of-smooth and properly C^p decomposable functions, provided the local minimizer $\bar{x}$ is a strong local minimizer that satisfies a strict complementarity condition; this class includes the max-of-smooth setting considered in [104]. Assumption Q also holds for generic linear tilts of semialgebraic functions: if f is semialgebraic, then for a full Lebesgue measure set of $w \in \mathbb{R}^d$, Assumption Q holds at every local minimizer $\bar{x}$ of the tilted function $f_w \colon x \mapsto f(x) + w^\top x$. We now present the theorem.

Theorem 6.1.1 (Main convergence theorem)**.** *Let $f \colon \mathbb{R}^d \to \mathbb{R}$ satisfy Assumption Q at a local minimizer $\bar{x} \in \mathbb{R}^d$. Fix scalar $c_0 \in (0, 1]$, budget $\{T_k\}$ and grid size $\{G_k\}$ sequences satisfying*

$$\min\{T_k, G_k\} \geq k + 1 \qquad \text{for all } k \geq 0.$$

Suppose that for initial point $x_0 \in \mathbb{R}^d$, there exists a subgradient $g_0 \in \partial_c f(x_0)$ such that $g_0 \neq 0$. Consider iterates $\{x_k\}$ generated by $\texttt{NTDescent}(x_0, g_0, c_0, \{G_k\}, \{T_k\})$. For any $q, k_0, C > 0$, let $E_{k_0,q,C}$ denote the event:

$$f(x_k) - f(\bar{x}) \leq \max\{(f(x_{k_0}) - f(\bar{x}))q^{k-k_0}, Cq^k\} \text{ for all } k \geq k_0.$$

Then there exists $q \in (0, 1)$, $C, C' > 0$, and a neighborhood U of $\bar{x}$ depending solely on f such that for any failure probability $p \in (0, 1)$ and all $k_0 \geq C' \max\{\log(1/p), 1\}$, we have

$$P(E_{k_0,q,C} \mid x_{k_0} \in U) \geq 1 - p,$$

provided $P(x_{k_0} \in U) > 0$. Moreover, if f is convex, we have

$$P(E_{k_0,q,C}) \geq 1 - p.$$

The theorem, justified in Theorems 6.6.3 and 6.6.5, bounds the function gap and distance by a quantity that geometrically decays in k. Let us examine the local complexity. Recall that each outer iteration of `NTDescent` requires at most $2T_kG_k$ first-order oracle evaluations. Thus, if $T_k = G_k = k + 1$ for all $k \geq 0$, the total number of oracle evaluations of K steps of `NTDescent` is at most $O(K^3)$. In other words, the local complexity of achieving an ε optimal solution is $O(\log^3(1/\varepsilon))$ for all sufficiently small $\varepsilon > 0$, where the big-O notation hides terms depending on the local conditioning of f; see Lemma 6.6.6. Therefore, the theorem establishes a local nearly linear convergence rate for `NTDescent`.

6.1.9 Outline

The outline of this chapter is as follows. In Section 5.2, we present notation and basic constructions. This section describes a key structure – the active manifold – and cannot be skipped. In Section 6.2, we present the sublinear convergence guarantees, which will be helpful in the convex setting. This section also introduces key properties of the `NDescent` method, which will be used later in the chapter. In Section 6.3, we introduce our central structural assumption – Assumption Q – and show that it is satisfied for the generic semialgebraic and decomposable problem classes. In Section 6.4, we show that Assumption Q implies the gradient inequality (6.1.5). In Section 6.5, we show that the `TDescent` and `NDescent` methods terminate rapidly under appropriate conditions. In Section 6.6, we use the gradient inequality (6.1.5) and Assumption Q to prove that `NTDescent` locally nearly linearly converges. Finally, in Section 6.7 we provide a brief numerical illustration. This chapter is based on the work [27].

6.2 Global sublinear convergence of `NTDescent`

The main goal of this chapter is to show that `NTDescent` locally converges nearly linearly for "typical" nonsmooth optimization problems. A natural question is whether `NTDescent` also possesses global nonasymptotic convergence guarantees. In this section, we prove two such guarantees: First, for arbitrary Lipschitz functions, we analyze the rate at which $\mathrm{dist}(0, \partial_{\sigma_i} f(x_k))$ tends to zero as a function of k. Second, for convex Lipschitz functions, we analyze the rate at which $f(x_k)$ tends to $\inf f$.

In the proofs of this section, the `TDescent` loop is ignored as we can only prove it terminates with descent near the minimizer. Instead, the global convergence guarantees follow from the properties of NDescent. Thus, our analysis follows that of [108], where a nearly identical MinNorm method was introduced. The main difference between the NDescent and the method of [108] lies in the perturbation radius in Line 3 of Algorithm 1: while the radius of NDescent can be computed with access only to $\sigma\|g_t\|$, the radius in [108] requires knowledge of the Lipschitz constant of f, which we do not assume. Finally, we mention that [108] did not consider convergence rates for convex problems.

Before stating the main result, we recall three key Lemmas that underlie the proof. The first lemma shows that the vectors u_i and v_i generated by linesearch are Goldstein subgradients of decreasing norm.

Lemma 6.2.1 (Properties of linesearch)**.** *Let $f\colon \mathbb{R}^d \to \mathbb{R}$ be a locally Lipschitz function. Fix $x \in \mathbb{R}^d$, subgradient $g \in \partial_c f(x)$, budget T, and grid size G. Let u_i and v_i be generated by* linesearch(x, g, G, T). *Then*

$$u_i, v_{i+1} \in \partial_{\sigma_i} f(x) \qquad \text{and} \qquad \|v_{i+1}\| \leq \|u_i\| \leq \|v_i\| \tag{6.2.1}$$

for all $i = 0, \dots, G-1$.

Proof. The proof follows by induction. We prove the base case only since the induction is straightforward. First note that the inclusion $v_0 \in \partial_c f(x)$ implies that $u_0 \in \partial_{\sigma_0} f(x)$, since TDescent constructs u_0 as a convex combinations of subgradients evaluated in the ball $\overline{B}_{\sigma_0}(\bar{x})$. Likewise, due to the argmin operation on line 4 of Algorithm 2, the subgradients generated by TDescent are decreasing in norm. Consequently, we have $\|u_0\| \leq \|v_0\|$. A similar argument shows that $v_1 \in \partial_{\sigma_0} f(x)$ and $\|v_1\| \leq \|u_0\|$. This completes the proof. □

The following lemma shows that when f is convex, the minimal norm Goldstein subgradient may be used to bound the function values. Since it follows from a standard argument, we place the proof in Appendix 7.3.1.

Lemma 6.2.2 (Subgradient inequality)**.** *Suppose $f\colon \mathbb{R}^d \to \mathbb{R}$ is a continuous convex function. Let $x, y \in \mathbb{R}^d$. Let L denote a Lipschitz constant for f on the ball $B_{2\sigma}(x)$. Then*

$$f(x) - f(y) \leq \|x - y\| \mathrm{dist}(0, \partial_\sigma f(x)) + 2\sigma L.$$

The final lemma provides conditions under which `NDescent` terminates with descent with high probability. The result is closely related to [108, Corollary 2.6], , but we take extra care to analyze the perturbation radius in Line 3 of Algorithm 1.

Lemma 6.2.3 (`NDescent` loop terminates with descent)**.** *Let f be a locally Lipschitz function. Fix initial point $x \in \mathbb{R}^d$, radius $\sigma > 0$, subgradient $g \in \partial_\sigma f(x)$, and failure probability $p \in (0, 1)$. Furthermore, let L be a Lipschitz constant of f on the ball $B_{2\sigma}(x)$. Suppose that*

$$\sigma \leq \frac{\mathrm{dist}(0, \partial_\sigma f(x))}{\sqrt{128}L}; \qquad \textit{and} \qquad T \geq \left\lceil \frac{64L^2}{\mathrm{dist}^2(0, \partial_\sigma f(x))} \right\rceil \lceil 2\log(1/p) \rceil .$$

Define $g_+ := \texttt{NDescent}(x, g, \sigma, T)$. Then $\|g_+\| \neq 0$ and the point $x_+ := x - \sigma \frac{g_+}{\|g_+\|}$ satisfies

$$f(x_+) \leq f(x) - \frac{\sigma \mathrm{dist}(0, \partial_\sigma f(x))}{8} \qquad \textit{with probability at least } 1 - p.$$

Proof. First note that $g_+ \in \partial_\sigma f(x)$, so $\|g_+\| \geq \mathrm{dist}(0, \partial_\sigma f(x)) > 0$. Now, observe that `NDescent` is precisely [108, Algorithm 1] with a different bound on the perturbation radius r. Indeed, in [108, Algorithm 1], r must satisfy

$$r < \|g_t\| \sqrt{1 - \left(1 - \frac{\|g_t\|^2}{128L^2}\right)^2}$$

for all $t \geq 0$. We now show that the constraint $r \leq \sigma \|g_t\|$ implies the above bound. To that end, define the univariate function $h\colon a \mapsto \sqrt{1 - (1 - \frac{a^2}{128L^2})^2}$. Then h is increasing in a for $a \leq L$. Moreover, for $a \in [0, L]$, we have $h(a) \geq \frac{a}{\sqrt{128}L}$. Consequently, since

$$\operatorname{dist}(0, \partial_\sigma f(x)) \leq \|g_t\| \leq L$$

for all $t \leq T$, we have

$$r < \sigma\|g_t\| \leq \frac{\operatorname{dist}(0, \partial_\sigma f(x))\|g_t\|}{\sqrt{128}L} \leq h(\operatorname{dist}(0, \partial_\sigma f(x)))\|g_t\| \leq h(\|g_t\|)\|g_t\|.$$

Thus, the proof is a direct application of [108, Corollary 2.6]. □

Given these lemmata, we are ready to state and prove our main sublinear convergence guarantee.

Theorem 6.2.4 (Sublinear convergence)**.** *Let $f\colon \mathbb{R}^d \to \mathbb{R}$ be a locally Lipschitz function. Fix initial point $x_0 \in \mathbb{R}^d$ and subgradient $g_0 \in \partial_c f(x_0)$. Assume that $g_0 \neq 0$. Let $L \in \mathbb{R} \cup \{+\infty\}$ be any Lipschitz constant of f over the widened sublevel set*

$$S := \{x + u\colon f(x) \leq f(x_0) \text{ and } u \in \overline{B}(x)\}.$$

Fix a scalar $c_0 \in (0, 1]$, budget sequence $\{T_k\}$, grid size sequence $\{G_k\}$, and failure probability $p \in (0, 1)$. Let $\{x_k\}$ be generated by $\texttt{NTDescent}(x, g, c_0, \{G_k\}, \{T_k\})$. Then for all $K > 0$, the following holds with probability at least $1-p$: Define $G := \min_{K \leq k \leq 2K-1} G_k$ and $T := \min_{K \leq k \leq 2K-1} T_k$. Then for all $i \leq G$, the following bound holds with $\sigma_i := 2^{-(G-i)}$:

$$\min_{K \leq k \leq 2K-1} \operatorname{dist}(0, \partial_{\sigma_i} f(x_k)) \leq \max\left\{\frac{8(f(x_K) - \inf f)}{\sigma_i K}, \frac{16L\sqrt{2\log(KG/p)}}{\sqrt{T}}, \sqrt{128}L\sigma_i\right\}.$$

Finally, suppose that f is convex and $D := \operatorname{diam}(\{x \in \mathbb{R}^d\colon f(x) \leq f(x_0)\}) < +\infty$. Then

$$f(x_{2K-1}) - \inf f \leq \min_{i \leq G}\left\{D \max\left\{\frac{8(f(x_K) - \inf f)}{\sigma_i K}, \frac{16L\sqrt{2\log(KG/p)}}{\sqrt{T}}, \sqrt{128}L\sigma_i\right\} + 2L\sigma_i\right\}. \tag{6.2.2}$$

Proof. Let us assume that $L < +\infty$; otherwise, the result is trivial. Fix $K > 0$ and $i \leq G$. Define

$$\epsilon_i := \max\left\{\frac{16L\sqrt{2\log(KG/p)}}{\sqrt{T}}, \sqrt{128}L\sigma_i\right\}.$$

For every $K \leq k \leq 2K-1$, define

$$x_{k,i} := x_k - \sigma_i \frac{v_{i+1}}{\|v_{i+1}\|}, \qquad \text{where } v_{i+1} := \texttt{NDescent}(x_k, u_i, \sigma_i, T_k),$$

and u_i appear in the definition of $\texttt{linesearch}(x_k, g_k, \max\{\|g_k\|, c_0\|g_0\|\}, G_k, T_k)$; see Algorithm 3. Note that $v_{i+1} \in \partial_{\sigma_i} f(x_k)$ by Lemma 6.2.1. Thus, in the event $\{\text{dist}(0, \partial_{\sigma_i} f(x_k)) \geq \epsilon_i\}$, we have

1. $x_{k,i}$ is well-defined since $v_{i+1} \neq 0$;
2. the trust region constraint $\sigma_i \leq \frac{\|v_{i+1}\|}{s}$ is satisfied for $s = \max\{\|g_k\|, c_0\|g_0\|\}$ (in Algorithm 3); indeed,

   $$\frac{\|v_{i+1}\|}{s} \geq \frac{\text{dist}(0, \partial_{\sigma_i} f(x_k))}{s} \geq \frac{\sqrt{128}L\sigma_i}{s} \geq \sigma_i,$$

   where the final inequality follows from the bound $s \leq L$, a consequence of the inclusion $x_0 \subseteq \text{int}\, S$ and the Lipschitz continuity of f on S.

Finally, for every $K \leq k \leq 2K-1$, define

$$A_{k,i} := \left\{f(x_{k,i}) - f(x_k) \geq -\frac{\sigma_i \text{dist}(0, \partial_{\sigma_i} f(x_k))}{8}\right\} \cap \{\text{dist}(0, \partial_{\sigma_i} f(x_k)) \geq \epsilon_i\}.$$

Now we apply Lemma 6.2.3.

To that end, observe that since $f(x_k)$ is nonincreasing and $\sigma_i \leq 1/2$, every iterate x_k satisfies $B_{2\sigma_i}(x_k) \subseteq S$. Consequently, L is a Lipschitz constant of f on $B_{2\sigma_i}(x_k)$. Therefore, by Lemma 6.2.3, for every $K \leq k \leq 2K-1$, we have

$$P(A_{k,i}) \leq P(A_{k,i} \mid \text{dist}(0, \partial_{\sigma_i} f(x_k)) \geq \epsilon_i) \leq \frac{p}{GK}. \tag{6.2.3}$$

Thus, by a union bound, with probability at least $1 - \frac{p}{G}$, at least one of the following must hold at every index $K \leq k \leq 2K - 1$:

$$f(x_{k,i}) - f(x_k) \leq -\frac{\sigma_i \mathrm{dist}(0, \partial_{\sigma_i} f(x_k))}{8} \qquad \text{or} \qquad \mathrm{dist}(0, \partial_{\sigma_i} f(x_k)) \leq \epsilon_i.$$

If $\mathrm{dist}(0, \partial_{\sigma_i} f(x_k)) \leq \epsilon_i$ for some k satisfying $K \leq k \leq 2K - 1$, then the result follows. On the other hand, suppose that for all $K \leq k \leq 2K - 1$, we have $\mathrm{dist}(0, \partial_{\sigma_i} f(x_k)) > \epsilon_i$; in particular, we have $\mathrm{dist}(0, \partial_{\sigma_i} f(x_k)) > \sqrt{128} L\sigma_i$. Therefore, with probability at least $1 - \frac{p}{G}$, we must have

$$f(x_{k+1}) \leq f(x_{k,i}) \leq f(x_k) - \frac{\sigma_i \mathrm{dist}(0, \partial_{\sigma_i} f(x_k))}{8}, \qquad \text{for all } K \leq k \leq 2K - 1.$$

where the first inequality follows since the trust region constraint is satisfied for $x_{k,i}$. Iterating this inequality, we have with probability at least $1 - \frac{p}{G}$, the bound

$$\min_{K \leq k \leq 2K-1} \mathrm{dist}(0, \partial_{\sigma_i} f(x_k)) \leq \frac{1}{K} \sum_{k=K}^{2K-1} \mathrm{dist}(0, \partial_{\sigma_i} f(x_k)) \leq \frac{8(f(x_K) - f(x_{2K}))}{\sigma_i K}.$$

This proves the result for i. A union bound over i yields the bound for minimal norm Goldstein subgradient for all $i \leq G$.

To prove (6.2.2), fix an $i \leq G$ and let k_i be the index that attains the minimum. Then

$$f(x_{2K-1}) - \inf f \leq f(x_{k_i}) - \inf f \leq \mathrm{dist}(x_{k_i}, \mathcal{X}_*) \min_{K \leq k \leq 2K-1} \mathrm{dist}(0, \partial_{\sigma_i} f(x_k)) + 2\sigma_i L,$$

where the first inequality follows since $f(x_k)$ is nonincreasing and the second inequality follows from Lemma 6.2.2. The proof then follows from the upper bound $\mathrm{dist}(x_{k_i}, \mathcal{X}) \leq D$. □

The theorem provides bounds on the minimal norm Goldstein subgradient within any window of indices $K \leq k \leq 2K - 1$. Let us briefly investigate the setting $T_k = k + 1$ for all $k \geq 0$. In this case, the theorem implies that with probability at least $1 - p$, we have

$$\min_{K \leq k \leq 2K-1} \mathrm{dist}(0, \partial_{\sigma_i} f(x_k)) \leq \max\left\{\frac{8(f(x_K) - \inf f)}{\sigma_i K}, \frac{16L\sqrt{2\log(KG/p)}}{\sqrt{2K}}, \sqrt{128} L\sigma_i\right\}$$

for all $i \leq G$. Let us now suppose G is large enough that there exists $i \leq G$ satisfying $(1/2)K^{-1/2} \leq \sigma_i \leq K^{-1/2}$, e.g., we may assume $G_k = \Omega(\log(k^{1/2}))$ for all $k > 0$. Then, we find that at most $O(KTG) = O(K^2G)$ first-order oracle evaluations are needed to find a point x_k satisfying

$$\mathrm{dist}(0, \partial_{K^{-1/2}} f(x_k)) = \tilde{O}(K^{-1/2}),$$

where $\tilde{O}$ hides logarithmic terms in G, K and p. Let's consider two settings for G_k.

1. **Setting 1: $G_k = O(\log(k^{1/2}))$.** In this case, `NTDescent` finds a point x_k satisfying $\mathrm{dist}(0, \partial_\varepsilon f(x_k)) \leq \varepsilon$ using at most $\tilde{O}(\varepsilon^{-4})$ first-order oracle evaluations.
2. **Setting 2: $G_k = k + 1$.** In this case, `NTDescent` finds a point x_k satisfying $\mathrm{dist}(0, \partial_\varepsilon f(x_k)) \leq \varepsilon$ using at most $\tilde{O}(\varepsilon^{-6})$ first-order oracle evaluations.

The complexity of Setting 1 is more favorable than the complexity of Setting 2. Nevertheless, when we establish our local rapid convergence guarantees, we will work in Setting 2, which has more favorable local convergence properties. Before moving on, we note that the above guarantees likewise apply in the convex setting, namely `NTDescent` finds a point x_k with $f(x_k) - f^* \leq \varepsilon$ using at most $\tilde{O}(\varepsilon^{-4})$, respectively $\tilde{O}(\varepsilon^{-6})$, first-order oracle evaluations in Setting 1, respectively Setting 2.

In addition to the nonasymptotic guarantees of Theorem 6.2.4, the reader may wonder whether a given limit point $\bar{x}$ of `NTDescent` is Clarke critical, meaning $0 \in \partial_c f(\bar{x})$. We prove that this is indeed the case under a bounded sublevel set condition. We place the proof in Appendix 7.3.3 since it follows a similar line of reasoning as Theorem 6.2.4.

Corollary 6.2.5 (Limiting points are Clarke critical)**.** *Let $f \colon \mathbb{R}^d \to \mathbb{R}$ be a locally Lipschitz function. Fix initial point $x_0 \in \mathbb{R}^d$ and subgradient $g_0 \in \partial_c f(x_0)$. Assume that $g_0 \neq 0$. Suppose the sublevel set $\{x \colon f(x) \leq f(x_0)\}$ is bounded. Fix scalar $c_0 \in (0, 1]$, budget sequence $\{T_k\}$, grid size sequence $\{G_k\}$ such that $\{G_k\}$ tends to infinity and $T_k \geq k$.*

Let $\{x_k\}$ *be generated by* `NTDescent`$(x, g, c_0, \{G_k\}, \{T_k\})$*. Then, with probability one, all the limiting points of* $\{x_k\}$ *are Clarke critical.*

This concludes our sublinear convergence guarantees for `NTDescent`. In the following section, we describe the key structural assumptions needed to ensure that `NTDescent` locally rapidly converges.

6.3 Main assumption, examples, and consequences

This section introduces our key structural assumption – Assumption Q. In Section 6.3.1, we show that Assumption Q holds for generic semialgebraic functions and certain properly C^p decomposable functions. Then, in Section 6.3.2, we extract several key consequences of Assumption Q. These consequences will be instrumental in proving the gradient inequality (6.1.5) and rapid convergence of `NTDescent`. We now turn to the assumption.

Assumption Q. Function $f\colon \mathbb{R}^d \to \mathbb{R}$ is locally Lipschitz with local minimizer $\bar{x} \in \mathbb{R}^d$.

(Q1) **(Quadratic Growth)** There exists $\gamma > 0$ such that

$$f(x) - f(\bar{x}) \geq \frac{\gamma}{2}\|x - \bar{x}\|^2 \qquad \text{for all } x \text{ near } \bar{x}.$$

(Q2) **(Active Manifold)** Function f admits a C^4-smooth active manifold $\mathcal{M}$ around $\bar{x}$.

(Q3) **(Strong-(a) regularity)** There exists $C_{(a)} > 0$ such that

$$\|P_{T_{\mathcal{M}}(y)}(v - \nabla_{\mathcal{M}} f(y))\| \leq C_{(a)}\|x - y\| \qquad \text{for all } x \in \mathbb{R}^d,\ v \in \partial_c f(x), \text{ and } y \in \mathcal{M} \text{ near } \bar{x}.$$

(Q4) **(($b_{\leq}$)-regularity)** The following inequality holds

$$f(y) \geq f(x) + \langle v, y - x\rangle + o(\|y - x\|) \qquad \text{as } y \xrightarrow{\mathcal{M}} \bar{x} \text{ and } x \to \bar{x} \text{ with } v \in \partial_c f(x),$$

where $o(\cdot)$ is any univariate function satisfying $\lim_{t\to 0} o(t)/t = 0$.

Some comments are in order. Assumption (Q1) is a classical regularity condition that ensures local linear convergence of gradient methods for smooth convex functions. Assumptions (Q2), (Q3), and (Q4) describe the interaction of f and a distinguished smooth manifold $\mathcal{M}$. Assumption (Q2) requires $\mathcal{M}$ to be an active manifold for f around $\bar{x}$ in the sense of Definition 2.4.1. In particular, along the manifold $\mathcal{M}$, the function f is C^4 smooth with covariant gradient $\nabla_{\mathcal{M}} f$; see Section 5.2 for a definition. Assumption (Q3) shows that in tangent directions, the covariant gradient along the manifold approximates the subgradients of f up to a linear error. Finally, Assumption (Q4) is a restricted lower smoothness property, showing that linear models of f off the manifold are underapproximators of f on the manifold up to first-order. Note that the property is automatic if f is weakly convex, meaning the mapping $x \mapsto f(x) + \frac{\rho}{2}\|x\|^2$ is convex for some $\rho \geq 0$. The weakly convex class is broad and contains all compositions of convex functions with smooth mappings that have Lipschitz Jacobians; see the survey [119] for an introduction. The last two assumptions were extensively studied in Chapter 3. We refer readers to Theorem 3.1.4 and Theorem 3.1.6 for details.

In the following section, we provide examples of functions satisfying Assumption Q.

6.3.1 Examples of Assumption Q

This section shows that the problems above satisfy Assumption Q. The most important example is the class of generic semialgebraic functions. The following theorem is essentially contained in [7, 25], but we provide a proof for completeness.

Theorem 6.3.1 (Generic semialgebraic functions)**.** *Consider a locally Lipschitz semialgebraic function $f\colon \mathbb{R}^d \to \mathbb{R}$. Then, for a full Lebesgue measure set of $w \in \mathbb{R}^d$, the tilted* *function $f_w\colon x \mapsto f(x) + w^\top x$ satisfies Assumption Q at every local minimizer.*

Proof. The proof is a consequence of [25, Theorem 3.31] and [7, Corollary 4.8, Theorem 4.16]. A combination of Corollary 4.8 and Theorem 4.16 in [7] shows that for a full Lebesgue measure set of $w \in \mathbb{R}^d$, the following hold: every local minimizer $\bar{x}$ of f_w lies on a C^4 active manifold $\mathcal{M}$, verifying (Q2); and the quadratic growth condition (Q1) holds at $\bar{x}$. Next, [25, Theorem 3.31] shows that f_w also satisfies the strong (a) property (Q3) along $\mathcal{M}$; applying [25, Theorem 3.11 and Theorem 3.4], we deduce that f_w also satisfies the $(b_{\leq})$-regularity property (Q4) along $\mathcal{M}$ at $\bar{x}$. □

Turning to our second class, we introduce so-called *properly C^p decomposable* functions, originally proposed and analyzed in [5]. At a high level, the class consists of functions that are locally the composition of a sublinear function with a smooth mapping, which together satisfy a transversality condition.

Definition 6.3.2 (Decomposable functions)**.** A function $f\colon \mathbb{R}^d \to \mathbb{R}$ is called *properly C^p decomposable at $\bar{x}$ as $h \circ c$* if near $\bar{x}$ it can be written as

$$f(x) = f(\bar{x}) + h(c(x))$$

for some C^p-smooth mapping $c\colon \mathbb{R}^d \to \mathbb{R}^m$ satisfying $c(\bar{x}) = 0$ and some proper, closed sublinear function $h\colon \mathbb{R}^m \to \mathbb{R}$ satisfying the transversality condition:

$$\operatorname{lin}(h) + \operatorname{range}(\nabla c(\bar{x})) = \mathbb{R}^m.$$

The following theorem shows that decomposable functions satisfy Assumption Q near local minimizers if they satisfy a strict complementarity condition and a quadratic growth bound. The proof is a consequence of results found in works [2, 5, 6, 25].

Theorem 6.3.3 (Properly decomposable functions)**.** *Consider a locally Lipschitz function $f\colon \mathbb{R}^d \to \mathbb{R}$. Let $\bar{x}$ be a local minimizer of f and suppose that f is properly C^4 decomposable at $\bar{x}$. Furthermore, suppose that*

1. **(Strict Complementarity)** *We have that* $0 \in \operatorname{ri} \partial_c f(\bar{x})$.
2. **(Quadratic growth)** *There exists* $\gamma > 0$ *such that*

$$f(x) - f(\bar{x}) \geq \frac{\gamma}{2}\|x - \bar{x}\|^2 \qquad \text{for all } x \text{ near } \bar{x}.$$

Then f satisfies Assumption Q at $\bar{x}$.

Proof. To set the notation for the proof, recall that since f is properly C^4 decomposable, there exist functions h and c satisfying the conditions of Definition 6.3.2. The discussion in [5, p. 683-4] then shows that the set

$$\mathcal{M} := c^{-1}(\operatorname{lin}(h))$$

is a so-called C^4 *manifold of partial smoothness* for f around $\bar{x}$ in the sense of Lewis [2]. Moreover, f is prox-regular at $\bar{x}$ for 0 in the sense of [64, Definition 1.1], since by definition it is *strongly amenable* [64, Definition 2.4] at $\bar{x}$; see [64, Proposition 2.5]. Thus, according to [120, Theorem 4.10], partial smoothness, prox-regularity, and strict complementarity ensure that the sharpness condition of Definition 2.4.1 holds. Consequently, $\mathcal{M}$ is a C^4 smooth active manifold around $\bar{x}$, verifying (Q2). In addition, [25, Corollary 3.24] ensures that f satisfies the (Q3) and (Q4) properties along $\mathcal{M}$. □

A popular class of decomposable objectives arises from the pointwise maxima of smooth functions that satisfy an affine independence property. For example, this class was considered in the work of Han and Lewis [104]. As an immediate corollary of Theorem 6.3.3, we show that such functions satisfy Assumption Q.

Corollary 6.3.4 (Max-of-smooth functions)**.** *Consider a locally Lipschitz function f and a family of C^4 smooth functions $f_i \colon \mathbb{R}^d \to \mathbb{R}$ indexed by a finite set $i \in I$. Fix a local minimizer $\bar{x}$ of f and suppose the set $\{\nabla f_i(\bar{x})\}_{i \in I}$ is affinely independent. Suppose furthermore that f is locally expressible as*

$$f(x) := \max_{i \in I} f_i(x) \qquad \text{for all } x \text{ near } \bar{x}.$$

Then, provided the strict complementarity and quadratic growth conditions of Theorem 6.3.3 hold, the function f satisfies Assumption Q at $\bar{x}$.

Proof. To prove the result, note that the affine independence property is simply a restatement of the transversality condition of Definition 6.3.2 for the smooth mapping $x \mapsto (f_i(x))_{i \in I}$ and the sublinear function $y \mapsto \max_{i \in I} y_i$. □

We now turn our attention to the key consequences of Assumption Q.

6.3.2 Key consequences of Assumption Q

The following proposition summarizes the key consequences of Assumption Q. The proof of the result is straightforward but technical, so we place it in Appendix 7.3.2.

Proposition 6.3.5 (Consequences of Assumption Q)**.** *Suppose f satisfies Assumption Q at $\bar{x}$. Then there exists $\delta_{\mathrm{A}} > 0$ such that on the ball $B_{2\delta_{\mathrm{A}}}(\bar{x})$, the projection operator $P_{\mathcal{M}}$ is C^3 with Lipschitz Jacobian and the smooth extension $f_{\mathcal{M}} := f \circ P_{\mathcal{M}}$ is C^3 with Lipschitz gradient. Moreover, the following bounds hold:*

1. **(Quadratic growth)** *The quadratic growth bound* (Q1) *holds throughout $\overline{B}_{2\delta_{\mathrm{A}}}(\bar{x})$.*

2. **(Smoothness of $P_{\mathcal{M}}$)** *For all $x \in B_{\delta_{\mathrm{A}}}(\bar{x})$ and $x' \in B_{2\delta_{\mathrm{A}}}(\bar{x})$, we have*

$$\|P_{\mathcal{M}}(x') - P_{\mathcal{M}}(x) - P_{T_{\mathcal{M}}(P_{\mathcal{M}}(x))}(x' - x)\| \leq C_{\mathcal{M}}(\mathrm{dist}^2(x, \mathcal{M}) + \|x - x'\|^2), \quad (6.3.1)$$

where $C_{\mathcal{M}} := 2\mathrm{lip}^{\mathrm{op}}_{\nabla P_{\mathcal{M}}}(\bar{x})$.

3. **(Bounds on $\nabla_{\mathcal{M}} f$)** *For all $x \in B_{\delta_{\mathrm{A}}}(\bar{x})$, we have*

$$\frac{\gamma}{2}\|P_{\mathcal{M}}(x) - \bar{x}\| \leq \|\nabla_{\mathcal{M}} f(P_{\mathcal{M}}(x))\| \leq \beta\|P_{\mathcal{M}}(x) - \bar{x}\|, \quad (6.3.2)$$

where $\beta := 2\mathrm{lip}_{\nabla f_{\mathcal{M}}}(\bar{x})$.

4. **(Consequence of strong (a))** *For all $x \in B_{\delta_{\mathrm{A}}}(\bar{x})$ and $\sigma \leq \delta_{\mathrm{A}}$, we have*

$$\sup_{g \in \partial_\sigma f(x)} \|P_{T_{\mathcal{M}}(P_{\mathcal{M}}(x))}(g - \nabla_{\mathcal{M}} f(P_{\mathcal{M}}(x)))\| \leq C_{(a)}(\mathrm{dist}(x, \mathcal{M}) + \sigma); \quad (6.3.3)$$

$$\sup_{g \in \partial_\sigma f(x)} \|P_{T_{\mathcal{M}}(P_{\mathcal{M}}(x))} g\| \leq C_{(a)}(\mathrm{dist}(x, \mathcal{M}) + \sigma) + \beta\|P_{\mathcal{M}}(x) - \bar{x}\|; \quad (6.3.4)$$

$$\sup_{g, g' \in \partial_\sigma f(x)} \|P_{T_{\mathcal{M}}(P_{\mathcal{M}}(x))}(g - g')\| \leq 2C_{(a)}(\mathrm{dist}(x, \mathcal{M}) + \sigma). \quad (6.3.5)$$

5. **(Aiming)** *For all $x \in B_{\delta_{\mathrm{A}}}(\bar{x})$ and all $v \in \partial_c f(x)$, we have*

$$\langle v, x - P_{\mathcal{M}}(x)\rangle \geq \mu\, \mathrm{dist}(x, \mathcal{M}), \quad (6.3.6)$$

where $\mu := \frac{1}{4} \liminf_{x' \xrightarrow{\mathcal{M}^c} \bar{x}} \mathrm{dist}(0, \partial_c f(x'))$.

6. **(Subgradient bound)** *For all $x \in B_{\delta_{\mathrm{A}}}(\bar{x})$ and $\sigma \leq \delta_{\mathrm{A}}$, we have*

$$\sup_{g \in \partial_\sigma f(x)} \|g\| \leq L,$$

where $L := 2\mathrm{lip}_f(\bar{x})$

7. **(Function gap)** *For all $x \in B_{\delta_{\mathrm{A}}}(\bar{x})$, we have*

$$f(x) - f(\bar{x}) \leq L\mathrm{dist}(x, \mathcal{M}) + \frac{\beta}{2}\|P_{\mathcal{M}}(x) - \bar{x}\|^2. \quad (6.3.7)$$

Let us briefly comment on the result. Item 2 provides a crucial smoothness property of the projection operator of $\mathcal{M}$. Item 3 shows that the Riemannian gradient of f is proportional to the distance of the projection y to $\bar{x}$. Item 4 shows how the Goldstein subgradients inherit the strong (a) property (Q3) of Assumption Q. Indeed, Equation (6.3.4) shows that Goldstein subgradients are "small" in tangent directions and Equation (6.3.5) shows Goldstein subgradients vary in an approximate Lipschitz fashion in tangent directions. Item 5 shows that the subgradients of f off of the manifold have a constant level of alignment with the normal vector $x - P_{\mathcal{M}}(x)$, i.e., the direction $-v$ "aims" towards the manifold. Note that $\mu > 0$ due to the active manifold Assumption (Q2). The proof of Item 5 is based on Assumptions (Q2) and (Q4); a similar result appears in [121, Theorem D.2]. Item 6 provides a bound on the Goldstein subgradients of f near $\bar{x}$; we will appeal to this bound throughout the analysis without referencing this proposition. Finally, Item 7 decomposes the function gap into two terms: the distance to the manifold and the squared distance of the projection to the solution. The proof relies on the smoothness of f along the manifold. Note that the trivial upper bound $L\|x - \bar{x}\|$ for the gap can be weaker than (6.3.7).

This concludes our discussion of Assumption Q. The following three sections establish further consequences: the gradient inequality (6.1.5) (Section 6.4); rapid local convergence of `NDescent` and `TDescent` (Section 6.5); and rapid local convergence of `NTDescent` (Section 6.6). We use the notation and results introduced in Proposition 6.3.5 in all three sections. Finally, the statements of the results in Section 6.4 and 6.5 contain several parameters/radii, which we will use in Section 6.6 to determine the region of near linear convergence and the oracle complexity for `NTDescent`. For the readers' convenience, we have listed these parameters in Table 6.1.

Parameter	Definition
D_1	$\frac{\mu}{8(\mu+L)}$
D_2	$\frac{\mu}{2}$
C_1	$\frac{\gamma}{4}$
C_2	$\min\left\{\frac{\gamma}{8C_{(a)}}, \frac{\min\{1,1/\delta_{\mathrm{A}}\}}{2}\right\}$
C_3	$\frac{C_1^2}{8L}$
C_4	$\min\left\{\frac{\beta}{C_{(a)}(1+\delta_{\mathrm{A}})}, \frac{\min\{\mu/\delta_{\mathrm{A}},C_3D_2/\beta\}}{4(1+(1+\delta_{\mathrm{A}})C_{\mathcal{M}})(\mu+L))}, \frac{1}{2}\right\}$
C_5	$\min\left\{\frac{\beta}{2C_{(a)}}, \frac{C_3D_2}{32C_{(a)}\beta}, C_4, \frac{C_2}{4}\right\}$
δ_{GI}	$\min\left\{\frac{\delta_A}{4}, \frac{D_1}{C_{\mathcal{M}}}\right\}$
$\delta_{\mathtt{ND}}$	$\min\left\{\delta_{\mathrm{GI}}, \frac{D_2}{D_1L\sqrt{128}}\right\}$
δ_{Grid}	$\min\left\{\frac{\delta_{\mathrm{A}}}{2}, \frac{1}{C_{\mathcal{M}}(D_1^{-1}+1)}, \frac{\mu}{8(C_{(a)}+\beta)}\right\}$

Table 6.1: Parameters used throughout Sections 6.4 and 6.5.

6.4 Verifying the gradient inequality (6.1.5) under Assumption Q

In this section, we establish the gradient inequality (6.1.5) for functions satisfying Assumption Q. Throughout the section, we assume that Assumption Q is in force. We also use the notation in Proposition 6.3.5.

We present the formal statement and the gradient inequality (6.1.5) in Theorem 6.4.3, which appears at the end of this section. The proof is a consequence of the two lemmata. In the first lemma, we prove a constant-sized lower bound for $\operatorname{dist}(0, \partial_\sigma f(x))$, whenever σ is sufficiently small. The proof of this bound relies on the active manifold assumption (Q2) and the aiming inequality (6.3.6). A consequence of the argument is that all elements of $\partial_\sigma f(x)$ are correlated with the normal direction $x - P_{\mathcal{M}}(x) \in N_{\mathcal{M}}(P_{\mathcal{M}}(x))$. Later in Proposition 6.5.1, we will also show that Algorithm 1 (`NDescent`) terminates rapidly when σ is in the region, motivating the name Normal Descent. We now turn to the lemma.

Lemma 6.4.1 (Lower bound on Goldstein subgradients I)**.** *Define*

$$D_1 := \frac{\mu}{8(\mu + L)}; \qquad D_2 := \frac{\mu}{2}; \qquad \textit{and} \qquad \delta_{\mathrm{GI}} := \min\left\{\frac{\delta_{\mathrm{A}}}{4}, \frac{D_1}{C_{\mathcal{M}}}\right\}.$$

Then for all $x \in B_{\delta_{\mathrm{GI}}}(\bar{x})$ *and* $0 < \sigma \leq D_1 \mathrm{dist}(x, \mathcal{M})$*, we have*

$$\mathrm{dist}(0, \partial_\sigma f(x)) \geq D_2.$$

Proof. We begin with some preliminary bounds. Fix $x \in B_{\delta_{\mathrm{GI}}}(\bar{x})$ and $\sigma > 0$ satisfying the lemma assumptions. We observe that

$$\sigma \leq D_1 \mathrm{dist}(x, \mathcal{M}) \leq \mathrm{dist}(x, \mathcal{M}) \leq \|x - \bar{x}\| \leq \delta_{\mathrm{GI}},$$

where the second inequality follows since $D_1 \leq 1$ and the third follows since $\bar{x} \in \mathcal{M}$. Consequently,

$$\begin{aligned} LC_{\mathcal{M}}(\sigma^2 + \mathrm{dist}^2(x, \mathcal{M})) &\leq \delta_{\mathrm{GI}} LC_{\mathcal{M}}(\sigma + \mathrm{dist}(x, \mathcal{M})) \\ &\leq 2L\delta_{\mathrm{GI}} C_{\mathcal{M}} \mathrm{dist}(x, \mathcal{M}) \\ &\leq 2LD_1 \mathrm{dist}(x, \mathcal{M}), \end{aligned} \tag{6.4.1}$$

where the first inequality follows from the bound $\max\{\sigma, \mathrm{dist}(x, \mathcal{M})\} \leq \delta_{\mathrm{GI}}$ and the second follows from the bound $\sigma \leq \mathrm{dist}(x, \mathcal{M})$. We now turn to the proof.

Now, let $x' \in \overline{B}_\sigma(x) \subseteq B_{\delta_{\mathrm{A}}}(\bar{x})$ and observe that by aiming condition (6.3.6),

$$\langle v, x' - P_{\mathcal{M}}(x') \rangle \geq \mu \mathrm{dist}(x', \mathcal{M}) \qquad \text{for all } v \in \partial_c f(x').$$

We claim that $\langle v, x - P_{\mathcal{M}}(x) \rangle \geq D_2 \mathrm{dist}(x, \mathcal{M})$ for all $v \in \partial_c f(x')$. Indeed, for all $v \in \partial_c f(x')$ we may upper bound the inner product as follows:

$$\begin{aligned} &\langle v, x' - P_{\mathcal{M}}(x') \rangle \\ &\leq \langle v, x - P_{\mathcal{M}}(x) \rangle + \|v\| \|x' - P_{\mathcal{M}}(x') - x - P_{\mathcal{M}}(x)\| \end{aligned}$$

$$\le \langle v, x - P_{\mathcal{M}}(x)\rangle + L\|(I - P_{T_{\mathcal{M}}(P_{\mathcal{M}}(x))})(x - x')\| + LC_{\mathcal{M}}(\sigma^2 + \mathrm{dist}^2(x, \mathcal{M}))$$

$$\le \langle v, x - P_{\mathcal{M}}(x)\rangle + 3LD_1 \mathrm{dist}(x, \mathcal{M}),$$

where the second inequality follows from the bound $\|v\| \le L$ and Item 2 of Proposition 6.3.5; and the third inequality follows from $\|x-x'\| \le \sigma \le D_1\mathrm{dist}(x, \mathcal{M})$ and (6.4.1). Consequently, for all $v \in \partial_c f(x')$, we have

$$\begin{aligned}\langle v, x - P_{\mathcal{M}}(x)\rangle &\ge \mu \mathrm{dist}(x', \mathcal{M}) - 3LD_1\mathrm{dist}(x, \mathcal{M}) \\ &\ge \mu\mathrm{dist}(x, \mathcal{M}) - \mu\sigma - 3LD_1\mathrm{dist}(x, \mathcal{M}) \\ &\ge \mu(1 - D_1(1 + 3L/\mu))\mathrm{dist}(x, \mathcal{M}) \\ &= D_2\mathrm{dist}(x, \mathcal{M}), \qquad (6.4.2)\end{aligned}$$

where the second inequality follows from 1-Lipschitz continuity of $\mathrm{dist}(\cdot, \mathcal{M})$; and the final inequality follows from the bound $D_1 \le \frac{1}{2(1+3L/\mu)}$. This proves the claim.

Now, fix $g \in \partial_\sigma f(x)$. By definition of $\partial_\sigma f(x)$, there exists a family of coefficients $\lambda_i \in [0, 1]$, points $x_i \in \overline{B}_\sigma(x) \subseteq B_{\delta_{\mathrm{A}}}(\bar{x})$, and subgradients $g_i \in \partial_c f(x_i)$ indexed by a finite set $i \in I$ such that $\sum_{i\in I} \lambda_i = 1$ and $g = \sum_{i\in I} \lambda_i g_i$. Thus, by (6.4.2), we have

$$\langle g, x - P_{\mathcal{M}}(x)\rangle = \sum_{i\in I} \lambda_i \langle g_i, x - P_{\mathcal{M}}(x)\rangle \ge D_2\mathrm{dist}(x, \mathcal{M}).$$

Therefore, we have

$$\|g\| \ge \frac{\langle g, x - P_{\mathcal{M}}(x)\rangle}{\mathrm{dist}(x, \mathcal{M})} \ge D_2,$$

as desired. □

In the second lemma, we provide a lower bound for $\mathrm{dist}(0, \partial_\sigma f(x))$ on the order of $\|P_{\mathcal{M}}(x) - \bar{x}\|$, provided $\sigma = O(\|P_{\mathcal{M}}(x) - \bar{x}\|)$. The proof of this bound relies on quadratic growth (Q1) and strong (a)-regularity (Q3). A consequence of the argument is that the minimal norm element of $\partial_\sigma f(x)$ is close to the tangent vector $\nabla_{\mathcal{M}} f(P_{\mathcal{M}}(x)) \in$

$T_{\mathcal{M}}(P_{\mathcal{M}}(x))$. Later in Proposition 6.5.6, we will also show that Algorithm 2 (`TDescent`) terminates rapidly when σ is in the region, motivating the name Tangent Descent. We now turn to the lemma.

Lemma 6.4.2 (Lower bound on Goldstein subgradients II)**.** *Define*

$$C_1 := \frac{\gamma}{4}; \qquad \textit{and} \qquad C_2 := \min\left\{\frac{\gamma}{8C_{(a)}}, \frac{\min\{1, 1/\delta_{\mathrm{A}}\}}{2}\right\}.$$

Then for all $x \in B_{\delta_{\mathrm{A}}}(\bar{x})$ and $\sigma \geq 0$ satisfying

$$\max\{\operatorname{dist}(x, \mathcal{M}), \sigma\} \leq C_2\|P_{\mathcal{M}}(x) - \bar{x}\|,$$

we have

$$\|P_{T_{\mathcal{M}(P_{\mathcal{M}}(x))}}(g)\| \geq C_1\|P_{\mathcal{M}}(x) - \bar{x}\| \qquad \textit{for all } g \in \partial_\sigma f(x).$$

Proof. Note that the term $1/\delta_{\mathrm{A}}$ in the definition of C_2 is unnecessary; however, it will be crucial in the proof of Theorem 6.4.3. Turning to the proof, fix $x \in B_{\delta_{\mathrm{A}}}(\bar{x})$ and $\sigma \geq 0$ satisfying the lemma assumptions. Define $y = P_{\mathcal{M}}(x)$. Note that

$$\sigma \leq C_2\|y - \bar{x}\| \leq 2C_2\|x - \bar{x}\| \leq \delta_{\mathrm{A}}.$$

Thus, by (6.3.3), for all $g \in \partial_\sigma f(x)$, we have

$$\|P_{T_{\mathcal{M}}(y)}(g - \nabla_{\mathcal{M}} f(y))\| \leq C_{(a)}(\operatorname{dist}(x, \mathcal{M}) + \sigma) \leq \frac{\gamma}{4}\|y - \bar{x}\|.$$

In addition, by (6.3.2), we have $\|\nabla_{\mathcal{M}} f(y)\| \geq \frac{\gamma}{2}\|y - \bar{x}\|$. Therefore, for all $g \in \partial_\sigma f(x)$, we have

$$\|P_{T_{\mathcal{M}}(y)}(g)\| \geq \|\nabla_{\mathcal{M}} f(y)\| - C_{(a)}(\operatorname{dist}(x, \mathcal{M}) + \sigma) \geq \frac{\gamma}{4}\|y - \bar{x}\|,$$

as desired. □

Given these lemmata, we are ready to establish the gradient inequality (6.1.5). The following theorem verifies the bound

$$\sigma \operatorname{dist}(0, \partial_\sigma f(x)) \geq \eta(f(x) - f(\bar{x})),$$

for some $\eta > 0$ provided x is sufficiently near $\bar{x}$ and (x, σ) lies within one of two regions, described in Item 1 and Item 2 of Theorem 6.4.3. Item 1 and Item 2 roughly correspond to the regions considered in Lemma 6.4.1 and Lemma 6.4.2, respectively. Comparing with the statement of the gradient inequality (6.1.5), we see that gradient inequality of Theorem 6.4.3 does not require knowledge of an explicit function $\sigma(x)$. Instead, we need only find some σ proportional to $D_1 \operatorname{dist}(x, \mathcal{M})$ or $C_2\|P_{\mathcal{M}}(x) - \bar{x}\|$ up to a factor of, say, 2. Later in Proposition 6.6.1, we show that this flexibility allows us to find an appropriate σ through the `linesearch` procedure.

Theorem 6.4.3 (Gradient inequality)**.** *Suppose that function f satisfies Assumption Q at $\bar{x} \in \mathbb{R}^d$. For any constants $a_1 \in (0, D_1]$ and $a_2 \in (0, C_2]$, we have*

$$\sigma \operatorname{dist}(0, \partial_\sigma f(x)) \geq \min\left\{\frac{\gamma a_2}{8 \max\{4La_2^2, \beta\}}, \frac{\mu a_1}{4 \max\{2L, \beta/a_2^2\}}\right\}(f(x) - f(\bar{x})),$$

whenever $x \in B_{\delta_{\mathrm{GI}}}(\bar{x})$ and $\sigma > 0$ satisfy Item 1 or Item 2:

1. *(a)* $\frac{a_1}{2}\operatorname{dist}(x, \mathcal{M}) \leq \sigma \leq a_1 \operatorname{dist}(x, \mathcal{M})$*;*

   (b) $a_2^2\|P_{\mathcal{M}}(x) - \bar{x}\|^2 \leq \operatorname{dist}(x, \mathcal{M})$.

2. *(a)* $\frac{a_2}{2}\|P_{\mathcal{M}}(x) - \bar{x}\| \leq \sigma \leq a_2\|P_{\mathcal{M}}(x) - \bar{x}\|$*;*

   (b) $\frac{\operatorname{dist}(x, \mathcal{M})}{\sigma} \leq 2a_2\|P_{\mathcal{M}}(x) - \bar{x}\|$.

Moreover, for any $x \in B_{\delta_{\mathrm{GI}}}(\bar{x}) \backslash \{\bar{x}\}$, there exists $\sigma > 0$ such that Item 1 or Item 2 is satisfied.

Proof. We first show that for any $x \in B_{\delta_{\mathrm{GI}}}(\bar{x})\backslash\{\bar{x}\}$, there exists $\sigma > 0$ such that either Item 1 or Item 2 is satisfied. We consider two cases. First, suppose $x \in \mathcal{M}$. Then Item 2 is trivially satisfied for $\sigma = a_2\|P_{\mathcal{M}}(x) - \bar{x}\|$. Second, suppose $x \notin \mathcal{M}$ and Item 1 cannot be satisfied for any $\sigma > 0$. In this case, we have

$$\operatorname{dist}(x, \mathcal{M}) \leq a_2^2\|P_{\mathcal{M}}(x) - \bar{x}\|^2 = 2a_2\sigma\|P_{\mathcal{M}}(x) - \bar{x}\| \qquad \text{with } \sigma := a_2\|P_{\mathcal{M}}(x) - \bar{x}\|/2.$$

Thus, Item 2 is satisfied.

Now we prove the gradient inequality is satisfied whenever σ satisfies Item 1 or Item 2. Let us suppose that Item 1 holds for some $x \in B_{\delta_{\mathrm{GI}}}(\bar{x})$ and $\sigma > 0$. From (6.3.7), we have the bound:

$$\begin{aligned}
\frac{1}{\max\{2L, \beta/a_2^2\}}(f(x) - f(\bar{x})) &\leq \frac{1}{\max\{2L, \beta/a_2^2\}}\left(L\operatorname{dist}(x, \mathcal{M}) + \frac{\beta}{2}\|P_{\mathcal{M}}(x) - \bar{x}\|^2\right) \\
&\leq \frac{1}{2}\left(\operatorname{dist}(x, \mathcal{M}) + a_2^2\|P_{\mathcal{M}}(x) - \bar{x}\|^2\right) \\
&\leq \operatorname{dist}(x, \mathcal{M}).
\end{aligned}$$

Now observe that the assumptions of Lemma 6.4.1 are satisfied since $x \in B_{\delta_{\mathrm{GI}}}(\bar{x})$, $a_1 \leq D_1$, and x and σ satisfy Item 1. Therefore, we have

$$\sigma\operatorname{dist}(0, \partial_\sigma f(x)) \geq \sigma D_2 \geq \frac{\mu a_1}{4}\operatorname{dist}(x, \mathcal{M}) \geq \frac{\mu a_1}{4\max\{2L, \beta/a_2^2\}}(f(x) - f(\bar{x})),$$

as desired.

Next, let us suppose that Item 2 holds for some $x \in B_{\delta_{\mathrm{GI}}}(\bar{x})$ and $\sigma > 0$. From (6.3.7), we have the bound:

$$\begin{aligned}
\frac{1}{\max\{4La_2^2, \beta\}}(f(x) - f(\bar{x})) &\leq \frac{1}{\max\{4La_2^2, \beta\}}\left(L\operatorname{dist}(x, \mathcal{M}) + \frac{\beta}{2}\|P_{\mathcal{M}}(x) - \bar{x}\|^2\right) \\
&\leq \frac{1}{2}\left(\frac{\operatorname{dist}(x, \mathcal{M})}{2a_2^2} + \|P_{\mathcal{M}}(x) - \bar{x}\|^2\right) \\
&\leq \frac{1}{2}\left(\frac{\operatorname{dist}(x, \mathcal{M})\|P_{\mathcal{M}}(x) - \bar{x}\|}{2a_2\sigma} + \|P_{\mathcal{M}}(x) - \bar{x}\|^2\right)
\end{aligned}$$

$$\leq \|P_{\mathcal{M}}(x) - \bar{x}\|^2.$$

Now observe that since $a_2 \leq C_2$ and x and σ satisfy Item 2, we have

$$\sigma \leq C_2\|P_{\mathcal{M}}(x) - \bar{x}\| \leq 2C_2\delta_{\mathrm{GI}} \leq (1/\delta_{\mathrm{A}})(\delta_{\mathrm{A}}/4) \leq 1,$$

where we use the bound $C_2 \leq 1/2\delta_{\mathrm{A}}$. Consequently, we have

$$\operatorname{dist}(x, \mathcal{M}) \leq 2\sigma C_2\|P_{\mathcal{M}}(x) - \bar{x}\| \leq C_2\|P_{\mathcal{M}}(x) - \bar{x}\|.$$

Therefore, $\max\{\operatorname{dist}(x, \mathcal{M}), \sigma\} \leq C_2\|P_{\mathcal{M}}(x) - \bar{x}\|$, so the conditions of Lemma 6.4.2 are satisfied (recall $\delta_{\mathrm{GI}} \leq \delta_{\mathrm{A}}$). Thus, let g denote the minimal norm element of $\partial_\sigma f(x)$ and let us apply Lemma 6.4.2:

$$\operatorname{dist}(0, \partial_\sigma f(x)) = \|g\| \geq \|P_{T_{\mathcal{M}}(P_{\mathcal{M}}(x))}(g)\| \geq \frac{\gamma}{4}\|P_{\mathcal{M}}(x) - \bar{x}\|.$$

Consequently, we have

$$\sigma\operatorname{dist}(0, \partial_\sigma f(x)) \geq \frac{\sigma\gamma}{4}\|P_{\mathcal{M}}(x) - \bar{x}\| \geq \frac{\gamma a_2}{8\max\{4La_2^2, \beta\}}(f(x) - f(\bar{x})),$$

where the last inequality follows from $\sigma \geq \frac{a_2}{2}\|P_{\mathcal{M}}(x) - \bar{x}\|$. This completes the proof. □

Remark 1. *Note that $a_1, a_2 \in (0, 1)$ as claimed in Section 6.1.4, where we introduced the* normal and tangent regions *appearing in the statement of Theorem 6.4.3.*

This concludes the proof of the gradient inequality (6.1.5) under Assumption Q. In Section 6.6, we will use the gradient inequality to establish rapid local convergence of `NTDescent`. Before proving that, the following section analyzes `TDescent` and `NDescent` methods.

6.5 Rapid termination of `NDescent` and `TDescent` under Assumption Q

In this section, we analyze the `NDescent` and `TDescent` methods, showing that both methods rapidly terminate with descent in appropriate regions. Throughout the section, we assume that Assumption Q is in force. We also use the results and notation of Proposition 6.3.5, Table 6.1, Lemma 6.4.1, and Lemma 6.4.2.

The main results of this section are Propositions 6.5.1 and 6.5.6, which analyze `NDescent` and `TDescent`, respectively. Proposition 6.5.1 shows that `NDescent` terminates with descent in a constant number of iterations within the region considered in Item 1 of Theorem 6.4.3. Proposition 6.5.6 shows that `TDescent` either terminates with descent in $O(\log^{-1}(f(x) - f(\bar{x})))$ iterations or $f(x) - f(\bar{x})$ is already exponentially small in T within the region considered in Item 2 of Theorem 6.4.3. These lemmata will be the basis of our main convergence theorem – Theorem 6.6.3 – appearing in Section 6.6.

6.5.1 Analysis of `NDescent`

The following proposition shows that `NDescent` locally terminates in finitely many iterations whenever σ is sufficiently small. The result is a simple consequence of Lemmas 6.2.3 and 6.4.1.

Proposition 6.5.1 (`NDescent` loop terminates with descent)**.** *Define a radius*

$$\delta_{\mathrm{ND}} := \min\left\{\delta_{\mathrm{GI}}, \frac{D_2}{D_1 L\sqrt{128}}\right\}.$$

Then for all $x \in B_{\delta_{\mathrm{ND}}}(\bar{x})$, radii $\sigma > 0$ with $\sigma \leq D_1 \mathrm{dist}(x, \mathcal{M})$, subgradients $g \in \partial_\sigma f(x)$, *failure probabilities* $p \in (0,1)$ *and budgets* $T > 0$ *satisfying*

$$T \geq \left\lceil \frac{64L^2}{D_2^2} \right\rceil \lceil 2\log(1/p) \rceil,$$

the point $x_+ := \mathtt{NDescent}(x, g, \sigma, T)$ *satisfies*

$$f(x_+) \leq f(x) - \frac{\sigma \mathrm{dist}(0, \partial_\sigma f(x))}{8} \qquad \textit{with probability at least } 1 - p.$$

Proof. Fix $x \in B_{\delta_{\mathtt{ND}}}(\bar{x})$ and $\sigma > 0$ satisfying the lemma assumptions. Observe that

$$\sigma \leq D_1 \mathrm{dist}(x, \mathcal{M}) \leq D_1 \delta_{\mathtt{ND}} \leq \min\left\{\delta_{\mathrm{GI}}, \frac{D_2}{L\sqrt{128}}\right\},$$

where the final inequality follows from the bound $D_1 \leq 1$; see Lemma 7.3.2. Thus, by Lemma 6.4.1, we have $\mathrm{dist}(0, \partial_\sigma f(x)) \geq D_2$ (recall $\delta_{\mathtt{ND}} \leq \delta_{\mathrm{GI}}$). Consequently,

$$\sigma \leq \frac{D_2}{L\sqrt{128}} \leq \frac{\mathrm{dist}(0, \partial_\sigma f(x))}{L\sqrt{128}}.$$

Therefore, σ and T satisfy the assumptions of Lemma 6.2.3. Hence, the desired descent condition is guaranteed with probability at least $1 - p$. □

We now turn to the analysis of the `TDescent` step.

6.5.2 Analysis of `TDescent`

In this section, we analyze `TDescent`, proving two main results. First, we prove Proposition 6.5.6, which shows that `TDescent` terminates rapidly. Second, in Lemma 6.5.8, we show that the trust region constraint in Line 7 of Algorithm 3 (`linesearch`) prevents long steps. Thus, once the method enters a sufficiently small neighborhood of $\bar{x}$, it cannot leave.

We begin with descent Proposition 6.5.6, which relies on four technical lemmata that analyze the structure of Goldstein subgradients when σ is sufficiently small and

x is sufficiently near $\bar{x}$: Lemma 6.5.2 states that elements of Goldstein subdifferential with small normal components are descent directions. Lemmas 6.5.3 and 6.5.4 show that normalized subgradient steps approximately reflect points across the active manifold. Lemma 6.5.5 uses the approximate reflection property to show that `TDescent` geometrically decreases the normal component of the input subgradient, ensuring that we rapidly find a descent direction. We now turn to the Lemmata.

6.5.2.1 Descent with small normal part

The first lemma shows that Goldstein subgradients with small normal components are descent directions.

Lemma 6.5.2 (Descent with small normal part)**.** *Define*

$$C_3 := \frac{C_1^2}{8L}.$$

Then for all $x \in B_{\delta_A}(\bar{x})$, $\sigma > 0$, and $g \in \partial_\sigma f(x) \backslash \{0\}$ satisfying

1. $\max\{\operatorname{dist}(x, \mathcal{M}), \sigma\} \leq \frac{C_2}{4}\|P_{\mathcal{M}}(x) - \bar{x}\|$*;*

2. $\|P_{N_{\mathcal{M}}(P_{\mathcal{M}}(x))}(g)\| \leq C_3\|P_{\mathcal{M}}(x) - \bar{x}\|^2$,

we have

$$f\left(x - \sigma\frac{g}{\|g\|}\right) \leq f(x) - \frac{\sigma\|g\|}{8}.$$

Proof. We begin with preliminary notation and bounds. We fix $x \in B_{\delta_A}(\bar{x})$ and subgradient $g \in \partial_\sigma f(x)\backslash\{0\}$. We define $y := P_{\mathcal{M}}(x)$, $T := T_{\mathcal{M}}(y)$, and $N := N_{\mathcal{M}}(y)$. We observe that

$$\sigma \leq \frac{C_2}{4}\|y - \bar{x}\| \leq \frac{C_2}{2}\|x - \bar{x}\| \leq C_2\delta_A \leq \delta_A,$$

where the final inequality follows since $C_2 \leq 1$; see Lemma 6.4.2. We now turn to the proof.

The starting point of the proof is Lebourg's mean value Theorem [31, Theorem 2.4], which ensures that there exists $v \in \partial_\sigma f(x)$ such that

$$f\left(x - \sigma\frac{g}{\|g\|}\right) - f(x) = \left\langle v, -\sigma\frac{g}{\|g\|}\right\rangle = -\frac{\sigma}{\|g\|}\left\langle v, P_T(g)\right\rangle - \frac{\sigma}{\|g\|}\left\langle v, P_N(g)\right\rangle.$$

In what follows, we will show that the first term satisfies $\langle v, P_T(g)\rangle \geq \frac{3}{8}\|g\|^2$, while the second term satisfies $|\langle v, P_N(g)\rangle| \leq \frac{1}{8}\|g\|^2$, yielding the result.

Indeed, beginning with $|\langle v, P_N(g)\rangle|$, we note that

$$\|P_N(g)\| \leq C_3\|P_{\mathcal{M}}(x) - \bar{x}\|^2 \leq \frac{C_3}{C_1^2}\|g\|^2 = \frac{1}{8L}\|g\|^2, \tag{6.5.1}$$

where the second inequality follow from Lemma 6.4.2. Consequently, we have the bound $|\langle v, P_N(g)\rangle| \leq L\|P_N(g)\| \leq \frac{1}{8}\|g\|^2$, where we first inequality relies on the estimate $\|v\| \leq L$; see Item 6 of Proposition 6.3.5.

Next, we prove a lower bound on $\langle v, P_T(g)\rangle$. Since $v \in \partial_\sigma f(x)$,

$$\|P_T(v - g)\| \leq 2C_{(a)}(\mathrm{dist}(x, \mathcal{M}) + \sigma) \leq C_2C_{(a)}\|P_{\mathcal{M}}(x) - \bar{x}\| \leq \frac{C_2C_{(a)}}{C_1}\|g\| \leq \frac{1}{2}\|g\|.$$

where the first inequality follows from (6.3.5); the second by assumption; the third follows from Lemma 6.4.2; and the fourth follows from the bound $C_2 \leq \frac{C_1}{2C_{(a)}}$. Therefore,

$$\|P_T(v) - g\| \leq \|P_T(v - g)\| + \|P_N(g)\| \leq \frac{1}{2}\|g\| + \frac{1}{8L}\|g\|^2 \leq \frac{5}{8}\|g\|,$$

where the second inequality follows from (6.5.1) and the third follows from the bound $\|g\| \leq L$. Consequently, we have the bound

$$\langle v, P_T(g)\rangle = \langle P_T(v), g\rangle \geq \|g\|^2 - \|P_T(v) - g\|\|g\| \geq \frac{3}{8}\|g\|^2.$$

This completes the proof. □

Note that the proof implies a slightly stronger bound than claimed, namely that we have $f(x - \sigma g/\|g\|) \leq f(x) - \sigma\|g\|/4$. To maintain symmetry with Proposition 6.5.1, however, we use the constant $1/8$ throughout.

6.5.2.2 The approximate reflection property

The next two lemmata prove the approximate reflection property described in the introduction. The lemmas roughly show that normalized subgradient steps approximately "flip the sign" of the normal component of the subgradient near the manifold; see Section 6.1.5 for more intuition. The first lemma proves the approximate reflection property up to a tolerance depending on the distance to the manifold and σ. This lemma will be used again in the proofs of Lemma 6.5.4 and Lemma 6.5.7.

Lemma 6.5.3 (Approximate reflection inequality, general case)**.** *For all $x \in B_{\delta_{\mathrm{A}}/2}(\bar{x}), \sigma \in (0, \delta_{\mathrm{A}}/2], g \in \partial_\sigma f(x)\backslash\{0\}$ and $\hat{g} \in \partial_c f\left(x - \sigma\frac{g}{\|g\|}\right)$, we have*

$$\begin{aligned}&\langle P_{N_{\mathcal{M}}(P_{\mathcal{M}}(x))}(\hat{g}), g\rangle \\ &\qquad \leq -\mu\|P_{N_{\mathcal{M}}(P_{\mathcal{M}}(x))}g\| + \frac{(\mu+L)\|g\|\mathrm{dist}(x,\mathcal{M})}{\sigma} + \frac{(\mu+L)\|g\|C_{\mathcal{M}}(\mathrm{dist}^2(x,\mathcal{M})+\sigma^2)}{\sigma}.\end{aligned} \tag{6.5.2}$$

Proof. We begin with preliminary notation and bounds. We fix $x \in B_{\delta_{\mathrm{A}}/2}(\bar{x})$ and subgradient $g \in \partial_\sigma f(x)\backslash\{0\}$. We define $y := P_{\mathcal{M}}(x)$, $T := T_{\mathcal{M}}(y)$, and $N := N_{\mathcal{M}}(y)$. Finally, define $u := \frac{g}{\|g\|}$. Note that since $x \in B_{\delta_{\mathrm{A}}/2}(\bar{x})$ and $\sigma \leq \delta_{\mathrm{A}}/2$, we have $x - \sigma u \in B_{\delta_{\mathrm{A}}}(\bar{x})$.

Therefore, by the aiming inequality (6.3.6), we have

$$\underbrace{\langle \hat{g}, x - \sigma u - P_{\mathcal{M}}(x - \sigma u)\rangle}_{=:A} \geq \underbrace{\mu\|x - \sigma u - P_{\mathcal{M}}(x - \sigma u)\|}_{=:B}.$$

We aim to simplify this inequality with (6.3.1). To that end, first note that

$$\|(x - \sigma u - P_{\mathcal{M}}(x - \sigma u)) - (x - P_{\mathcal{M}}(x) - \sigma P_N(u))\| = \|P_{\mathcal{M}}(x - \sigma u) - P_{\mathcal{M}}(x) + \sigma P_T(u)\|$$

$$\le C_{\mathcal{M}}(\mathrm{dist}^2(x, \mathcal{M}) + \sigma^2). \qquad (6.5.3)$$

Consequently, we have

$$A \ge B \ge \mu\|x - P_{\mathcal{M}}(x) - \sigma P_N(u)\| - \mu C_{\mathcal{M}}(\mathrm{dist}^2(x, \mathcal{M}) + \sigma^2) \ge \sigma\mu\|P_N(u)\| - \mu S,$$

where $S := \mathrm{dist}(x, \mathcal{M}) + C_{\mathcal{M}}(\mathrm{dist}^2(x, \mathcal{M}) + \sigma^2)$. In addition, by (6.5.3) we have

$$\langle \hat{g}, (x - \sigma u - P_{\mathcal{M}}(x - \sigma u)) + \sigma P_N u\rangle \le LS.$$

Therefore, we have

$$\langle \hat{g}, \sigma P_N(u)\rangle = -A + \langle \hat{g}, (x - \sigma u - P_{\mathcal{M}}(x - \sigma u)) + \sigma P_N u\rangle \le -\sigma\mu\|P_N(u)\| + (\mu + L)S. \tag{6.5.4}$$

Inequality (6.5.2) then follows by multiplying both sides of inequality (6.5.4) by $\|g\|/\sigma$.

□

The second lemma is an application of Lemma 6.5.3 nearby the manifold.

Lemma 6.5.4 (Approximate reflection inequality near the manifold)**.** *Define*

$$C_4 := \min\left\{\frac{\beta}{C_{(a)}(1 + \delta_{\mathrm{A}})}, \frac{\min\{\mu/\delta_{\mathrm{A}}, C_3 D_2/\beta\}}{4(1 + (1 + \delta_{\mathrm{A}})C_{\mathcal{M}})(\mu + L)}, \frac{1}{2}\right\}.$$

Then for all $x \in B_{\delta_{\mathrm{A}}/2}(\bar{x})$, $\sigma > 0$, and $g \in \partial_\sigma f(x)\backslash\{0\}$ satisfying

$$\max\left\{\frac{\mathrm{dist}(x, \mathcal{M})}{\sigma}, \sigma\right\} \le C_4\|P_{\mathcal{M}}(x) - \bar{x}\|,$$

we have

$$\langle P_{N_{\mathcal{M}}(P_{\mathcal{M}}(x))}(\hat{g}), g\rangle \le -D_2\|P_{N_{\mathcal{M}}(P_{\mathcal{M}}(x))}g\| + \frac{C_3 D_2}{2}\|P_{\mathcal{M}}(x) - \bar{x}\|^2 \qquad \textit{for all } \hat{g} \in \partial_c f\left(x - \sigma\frac{g}{\|g\|}\right).$$

Proof. We begin with preliminary notation and bounds. We fix $x \in B_{\delta_{\mathrm{A}}/2}(\bar{x})$ and subgradient $g \in \partial_\sigma f(x)\backslash\{0\}$. We define $y := P_{\mathcal{M}}(x)$, $T := T_{\mathcal{M}}(y)$, and $N := N_{\mathcal{M}}(y)$. We observe that

$$\sigma \le C_4\|y - \bar{x}\| \le 2C_4\|x - \bar{x}\| \le C_4\delta_A \le \delta_{\mathrm{A}}/2.$$

Finally, we have

$$S := \operatorname{dist}(x, \mathcal{M}) + C_{\mathcal{M}}(\operatorname{dist}^2(x, \mathcal{M}) + \sigma^2) \leq \sigma C_4(1 + C_{\mathcal{M}}(1 + \delta_{\mathrm{A}}))\|y - \bar{x}\|. \qquad (6.5.5)$$

where the inequality follows from the bound $\operatorname{dist}(x, \mathcal{M}) \leq \|x - \bar{x}\| \leq \delta_{\mathrm{A}}$.

We now apply inequality (6.5.2):

$$\begin{aligned}
\langle P_N \hat{g}, g \rangle &\leq -\mu\|P_N g\| + \frac{(\mu + L)\|g\|S}{\sigma} \\
&\leq -\mu\|P_N g\| + (1 + (1 + \delta_{\mathrm{A}})C_{\mathcal{M}})(\mu + L)C_4\|g\|\|y - \bar{x}\| \\
&\leq -\mu\|P_N g\| + (1 + (1 + \delta_{\mathrm{A}})C_{\mathcal{M}})(\mu + L)C_4(\|P_T(g)\| + \|P_N(g)\|)\|y - \bar{x}\| \\
&\leq -\frac{\mu}{2}\|P_N g\| + \frac{C_3 D_2}{4\beta}\|P_T(g)\|\|y - \bar{x}\|,
\end{aligned}$$

where the second inequality follows from (6.5.5); the third inequality follows from triangle inequality; and the fourth inequality follows from the bound

$$(1 + (1 + \delta_{\mathrm{A}})C_{\mathcal{M}})(\mu + L)C_4\|y - \bar{x}\| \leq \frac{\mu/\delta_{\mathrm{A}}}{4} \cdot (2\|x - \bar{x}\|) \leq \frac{\mu/\delta_{\mathrm{A}}}{4}\delta_{\mathrm{A}} \leq \mu/2.$$

The proof will be complete if we can show that

$$\|P_T(g)\| \leq 2\beta\|y - \bar{x}\|.$$

To that end, we have

$$\|P_T(g)\| \leq C_{(a)}(\operatorname{dist}(x, \mathcal{M}) + \sigma) + \beta\|y - \bar{x}\| \leq (C_4 C_{(a)}(1 + \delta_{\mathrm{A}}) + \beta)\|y - \bar{x}\| \leq 2\beta\|y - \bar{x}\|,$$

where the first inequality follows from (6.3.4); the second inequality follows from the lemma assumptions and the bound $\operatorname{dist}(x, \mathcal{M}) \leq C_4\sigma\|y - \bar{x}\| \leq C_4\delta_{\mathrm{A}}\|y - \bar{x}\|$; and the third inequality follows from the bounds on C_4. This completes the proof. □

6.5.2.3 The normal component shrinks geometrically

The following lemma shows that every step of `TDescent` geometrically shrinks the normal component of the subgradient up to a tolerance of $O(\|P_{\mathcal{M}}(x) - \bar{x}\|^2)$.

Lemma 6.5.5 (Normal component shrinks geometrically)**.** *Define*

$$C_5 := \min\left\{\frac{\beta}{2C_{(a)}}, \frac{C_3 D_2}{32 C_{(a)}\beta}, C_4, \frac{C_2}{4}\right\}.$$

Then for all $x \in B_{\delta_{\mathrm{A}}/2}(\bar{x})$*,* $\sigma > 0$*,* $g \in \partial_\sigma f(x) \backslash \{0\}$*, and* $\hat{g} \in \partial_c f(x - \sigma\frac{g}{\|g\|}) \backslash \{0\}$ *satisfying*

1. $\|P_{N_{\mathcal{M}}(P_{\mathcal{M}}(x))} g\| \geq C_3 \|P_{\mathcal{M}}(x) - \bar{x}\|^2$*;*
2. $\max\left\{\frac{\mathrm{dist}(x,\mathcal{M})}{\sigma}, \sigma\right\} \leq C_5 \|P_{\mathcal{M}}(x) - \bar{x}\|$,

the vector $g' = \mathrm{argmin}_{h \in [g,\hat{g}]} \|h\|$ *satisfies:*

$$\|P_{N_{\mathcal{M}}(P_{\mathcal{M}}(x))}(g')\|^2 \leq \left(1 - \frac{3D_2^2}{64L^2}\right)\|P_{N_{\mathcal{M}}(P_{\mathcal{M}}(x))} g\|^2.$$

Proof. We begin with preliminary notation and bounds. We fix $x \in B_{\delta_{\mathrm{A}}/2}(\bar{x})$ and subgradient $g \in \partial_\sigma f(x) \backslash \{0\}$. We define $y := P_{\mathcal{M}}(x)$, $T := T_{\mathcal{M}}(y)$, and $N := N_{\mathcal{M}}(y)$. We observe two bounds. First, we have

$$\sigma \leq C_5 \|y - \bar{x}\| \leq 2C_5 \|x - \bar{x}\| \leq C_5 \delta_{\mathrm{A}} \leq 1.$$

where the final inequality follows since $C_5 \leq C_2/4 \leq 1/(8\delta_{\mathrm{A}})$. Second, we have

$$\mathrm{dist}(x, \mathcal{M}) \leq C_5 \sigma \|y - \bar{x}\| \leq C_5 \|y - \bar{x}\|, \tag{6.5.6}$$

since $\sigma \leq 1$. We now turn to the proof.

Consider the optimal weight $\lambda' := \mathrm{argmin}_{\lambda \in [0,1]} \|g + \lambda(\hat{g} - g)\|$. By definition we have $g' = g + \lambda'(\hat{g} - g)$. Moreover, a quick calculation shows that

$$\lambda' = \max\left\{\min\left\{-\frac{\langle g, \hat{g} - g\rangle}{\|\hat{g} - g\|^2}, 1\right\}, 0\right\}.$$

We claim that the following bound holds on λ':

$$\underbrace{-\frac{\langle P_N(g), \hat{g} - g\rangle}{8L^2}}_{=:\lambda_1} \leq \lambda' \leq -\underbrace{\frac{3\langle P_N(g), \hat{g} - g\rangle}{2\|P_N(\hat{g} - g)\|^2}}_{=:\lambda_2}. \tag{6.5.7}$$

Note that (6.5.7) is an immediate consequence of the following bound:

$$0 \leq -\frac{1}{2}\langle P_N(g), \hat{g} - g\rangle \leq -\langle g, \hat{g} - g\rangle \leq -\frac{3}{2}\langle P_N(g), \hat{g} - g\rangle. \tag{6.5.8}$$

Indeed, if (6.5.8) holds, then $\lambda' = \min\left\{-\frac{\langle g, \hat{g}-g\rangle}{\|\hat{g}-g\|^2}, 1\right\}$. Thus, we obtain the upper bound

$$\lambda' \leq -\frac{\langle g, \hat{g} - g\rangle}{\|\hat{g} - g\|^2} \leq -\frac{3}{2}\frac{\langle P_N(g), \hat{g} - g\rangle}{\|g - \hat{g}\|^2} \leq -\frac{3}{2}\frac{\langle P_N(g), \hat{g} - g\rangle}{\|P_N(g - \hat{g})\|^2} = \lambda_2.$$

Likewise, we obtain the lower bound

$$\lambda' = \min\left\{-\frac{\langle g, \hat{g} - g\rangle}{\|\hat{g} - g\|^2}, 1\right\} \geq \min\left\{-\frac{\langle g, \hat{g} - g\rangle}{4L^2}, 1\right\} = -\frac{\langle g, \hat{g} - g\rangle}{4L^2} \geq -\frac{\langle P_N(g), \hat{g} - g\rangle}{8L^2} = \lambda_1,$$

where the first inequality follows from the bound $\|\hat{g} - g\|^2 \leq 2(\|\hat{g}\|^2 + \|g\|^2) \leq 4L^2$; and the second equality follows from the bound $|\langle g, \hat{g} - g\rangle| \leq \|g\|\|\hat{g} - g\| \leq 2L^2$. Thus, we now prove (6.5.8).

To that end, note that (6.5.8) is equivalent to the following bound:

$$|\langle P_T(g), \hat{g} - g\rangle| \leq \frac{-\langle P_N(g), \hat{g} - g\rangle}{2}. \tag{6.5.9}$$

Therefore, we first bound $|\langle P_T(g), \hat{g} - g\rangle|$:

$$\begin{aligned}
|\langle P_T(g), \hat{g} - g\rangle| &\leq \|P_T(g)\|\|P_T(\hat{g} - g)\| \\
&\leq 2C_{(a)}(\mathrm{dist}(x, \mathcal{M}) + \sigma)(C_{(a)}(\mathrm{dist}(x, \mathcal{M}) + \sigma) + \beta\|y - \bar{x}\|) \\
&\leq 4C_{(a)}C_5(2C_{(a)}C_5 + \beta)\|y - \bar{x}\|^2 \\
&\leq \frac{C_3 D_2}{4}\|y - \bar{x}\|^2,
\end{aligned}$$

where the second inequality follows from (6.3.4) and (6.3.5); the third inequality follows from (6.5.6) and the bound $\sigma \leq C_5\|y - \bar{x}\|$; and the fourth inequality follows from the definition of C_5. To complete the proof of (6.5.9), we show that $\frac{C_3 D_2}{4}\|y - \bar{x}\|^2 \leq -\frac{1}{2}\langle P_N(g), \hat{g} - g\rangle$:

$$\frac{C_3 D_2}{2}\|y - \bar{x}\|^2 \leq D_2\|P_N(g)\| - \frac{C_3 D_2}{2}\|y - \bar{x}\|^2$$

$$
\begin{aligned}
&\leq -\langle P_N(g), \hat{g}\rangle \\
&\leq -\langle P_N(g), \hat{g} - g\rangle, \qquad (6.5.10)
\end{aligned}
$$

where the first inequality follows from the assumption $\frac{D_2}{2}\|P_N(g)\| \geq \frac{C_3 D_2}{2}\|y - \bar{x}\|^2$; the second inequality follows from Lemma 6.5.4 (recall $C_5 \leq C_4$ and $x \in B_{\delta_A/2}(\bar{x})$); and the third inequality follows from $\langle P_N(g), g\rangle = \|P_N(g)\|^2 \geq 0$. Thus, the equivalent bounds (6.5.9) and (6.5.8) hold. Consequently, Equation (6.5.7) holds.

Now, we turn to the contraction argument. Consider the function $r\colon \mathbb{R} \to \mathbb{R}$ satisfying

$$
r(\lambda) = \|P_N(g)\|^2 + 2\lambda \langle P_N(g), \hat{g} - g\rangle + \lambda^2 \|P_N(\hat{g} - g)\|^2 \qquad \text{for all } \lambda \in \mathbb{R}.
$$

Observe that

$$
\|P_N(g')\|^2 = \|P_N(g)\|^2 + 2\lambda' \langle P_N(g), \hat{g} - g\rangle + (\lambda')^2 \|P_N(\hat{g} - g)\|^2 = r(\lambda').
$$

Therefore, by convexity of r and (6.5.7), we have

$$
\|P_N(g')\|^2 = r(\lambda') \leq \max_{\lambda \in [\lambda_1, \lambda_2]} r(\lambda) \leq \max\{r(\lambda_1), r(\lambda_2)\}.
$$

To complete the proof, we show each term in the "max" is bounded by $\left(1 - \frac{3D_2^2}{64L^2}\right)\|P_N(g)\|^2$.

To show this, we will use the following consequence of (6.5.10):

$$
-\langle P_N(g), \hat{g} - g\rangle \geq D_2 \|P_N(g)\| - \frac{C_3 D_2}{2}\|y - \bar{x}\|^2 \geq \frac{D_2}{2}\|P_N(g)\|, \qquad (6.5.11)
$$

where the final inequality follows from the assumption $\frac{C_3 D_2}{2}\|y - \bar{x}\|^2 \leq \frac{D_2}{2}\|P_N(g)\|$. Indeed, first, observe that

$$
\begin{aligned}
r(\lambda_2) &= \|P_N(g)\|^2 - \frac{3}{4}\frac{\langle P_N(g), \hat{g} - g\rangle^2}{\|P_N(\hat{g} - g)\|^2} \\
&\leq \left(1 - \frac{3D_2^2}{16\|P_N(\hat{g} - g)\|^2}\right)\|P_N(g)\|^2
\end{aligned}
$$

$$\leq \left(1 - \frac{3D_2^2}{64L^2}\right) \|P_N(g)\|^2,$$

where the first inequality from (6.5.11) and the second inequality follows from the bound $\|P_N(\hat{g} - g)\|^2 \leq \|\hat{g} - g\|^2 \leq 4L^2$. Likewise, observe that

$$\begin{aligned} r(\lambda_1) &= \|P_N(g)\|^2 - \frac{\langle P_N(g), \hat{g} - g\rangle^2}{4L^2} + \frac{\langle P_N(g), \hat{g} - g\rangle^2 \|P_N(\hat{g} - g)\|^2}{64L^4} \\ &\leq \|P_N(g)\|^2 - \frac{\langle P_N(g), \hat{g} - g\rangle^2}{4L^2} + \frac{\langle P_N(g), \hat{g} - g\rangle^2}{16L^2} \\ &\leq \left(1 - \frac{3D_2^2}{64L^2}\right) \|P_N(g)\|^2, \end{aligned}$$

where the first inequality follows from the bound $\|P_N(\hat{g} - g)\|^2 \leq \|\hat{g} - g\|^2 \leq 4L^2$; and the second inequality follows from (6.5.11). Therefore, the proof is complete. □

`TDescent` **terminates with descent.** The following proposition is the main result of this section. It shows that `TDescent` must either terminate with descent or $f(x) - f(\bar{x})$ is already exponentially small in T.

Proposition 6.5.6 (`TDescent` loop terminates with descent)**.** *Fix $T \in \mathbb{N}$. Then for all $x \in B_{\delta_A/2}(\bar{x})$, $v \in \partial_\sigma f(x)$, and $\sigma > 0$ satisfying*

$$\max\left\{\frac{\operatorname{dist}(x, \mathcal{M})}{\sigma}, \sigma\right\} \leq C_5 \|P_{\mathcal{M}}(x) - \bar{x}\|,$$

at least one of the following holds:

1. *we have*
$$f(x) - f(\bar{x}) \leq \frac{(C_5^2 L + \beta)L}{C_3}\left(1 - \frac{3\mu^2}{256L^2}\right)^{T/2};$$

2. *the vector $g :=$ `TDescent`(x, v, σ, T) satisfies $\|g\| > 0$ and*
$$f\left(x - \sigma\frac{g}{\|g\|}\right) \leq f(x) - \frac{\sigma \operatorname{dist}(0, \partial_\sigma f(x))}{8}.$$

Proof. We begin with preliminary notation and bounds. We fix $x \in B_{\delta_A/2}(\bar{x})$ and subgradient $v \in \partial_\sigma f(x)$. We define $y := P_{\mathcal{M}}(x)$, and $N := N_{\mathcal{M}}(y)$. Observe that

$$\sigma \leq C_5\|y - \bar{x}\| \leq 2C_5\|x - \bar{x}\| \leq C_5\delta_A \leq 1,$$

where the final inequality follows by definition of $C_5 \leq C_2/4 \leq 1/(8\delta_A)$. In addition, since $C_5 \leq C_2/4$, we have $\sigma \leq (C_2/4)\|y - \bar{x}\|$ and

$$\text{dist}(x, \mathcal{M}) \leq \sigma C_5\|y - \bar{x}\| \leq \frac{C_2}{4}\|y - \bar{x}\|,$$

where the final inequality follows from $\sigma \leq 1$. Consequently,

$$\max\{\sigma, \text{dist}(x, \mathcal{M})\} \leq \frac{C_2}{4}\|y - \bar{x}\|. \tag{6.5.12}$$

We now turn to the proof.

Turning to the proof, note that since $x \in B_{\delta_A/2}(\bar{x})$, Lemma 6.4.2 and (6.5.12) ensure that

$$\text{dist}(0, \partial_\sigma f(x)) \geq C_1\|y - \bar{x}\| > 0.$$

Thus, if $\texttt{TDescent}(x, v, \sigma, T)$ terminates at $t < T$, then Item 2 must hold. For the remainder of the proof, we suppose that $\texttt{TDescent}(x, v, \sigma, T)$ terminates at the final iteration $t = T$ and that Item 2 does not hold. In this case, Lemma 6.5.2 and (6.5.12) ensure that the iterates g_t of $\texttt{TDescent}(x, v, \sigma, T)$ satisfy $\|P_N(g_t)\| > C_3\|y - \bar{x}\|^2$ for all $0 \leq t \leq T - 1$. Therefore, since $x \in B_{\delta_A/2}(\bar{x})$, $\max\{\text{dist}(x, \mathcal{M})/\sigma, \sigma\} \leq C_5\|y - \bar{x}\|$, and $\|P_N(g_t)\| > C_3\|y - \bar{x}\|^2$, Lemma 6.5.5, yields the contraction:

$$\|P_N(g_{t+1})\|^2 \leq \left(1 - \frac{3D_2^2}{64L^2}\right)\|P_N(g_t)\|^2, \qquad \text{for all } 0 \leq t \leq T - 1.$$

Unfolding this contraction, we see that g_T is an exponentially small Goldstein subgradient:

$$\|P_N(g_T)\| \leq \left(1 - \frac{3D_2^2}{64L^2}\right)^{T/2} \|P_N(g_0)\|.$$

As a result, the projection y is nearby $\bar{x}$:

$$\|y - \bar{x}\|^2 \leq \frac{\|P_N(g_T)\|}{C_3} \leq \frac{\|P_N(g_0)\|}{C_3}\left(1 - \frac{3D_2^2}{64L^2}\right)^{T/2} \leq \frac{L}{C_3}\left(1 - \frac{3D_2^2}{64L^2}\right)^{T/2}. \qquad (6.5.13)$$

Consequently,

$$\begin{aligned} f(x) - f(\bar{x}) &\leq L\text{dist}(x, \mathcal{M}) + \beta\|y - \bar{x}\|^2 \\ &\leq (C_5^2 L + \beta)\|y - \bar{x}\|^2 \\ &\leq \frac{(C_5^2 L + \beta)L}{C_3}\left(1 - \frac{3D_2^2}{64L^2}\right)^{T/2}, \end{aligned}$$

where the first inequality follows from (6.3.7) (recall $x \in B_{\delta_{\mathrm{A}}/2}(\bar{x})$); the second inequality follows since $\text{dist}(x, \mathcal{M}) \leq \sigma C_5\|y - \bar{x}\| \leq C_5^2\|y - \bar{x}\|^2$; and the third inequality follows from (6.5.13). The proof then follows from the identity $D_2 = \frac{\mu}{2}$. □

6.5.2.4 The “trust region” constraint prevents long steps

Before ending this section, we must establish one final technical result for `TDescent`. Namely, in Lemma 6.5.8, we show that for appropriate σ, `TDescent` eventually generates small subgradients on the order of $O(\|x - \bar{x}\|)$. This property is intuitive because $\text{dist}(0, \partial_\sigma f(x)) = 0$ whenever $\sigma \geq \|x - \bar{x}\|$. This property will help us ensure that the iterates of `NTDescent` (Algorithm 4) cannot leave sufficiently small neighborhoods of $\bar{x}$. Indeed, since the subgradients v_{i+1} generated by Algorithm 3 (`linesearch`) are decreasing in norm, we will show that the trust region constraint $\sigma_i \leq \frac{\|v_{i+1}\|}{s}$ in Line 7 of Algorithm 3 must eventually be violated for large i. This ensures large σ_i are never chosen.

To prove this claim, we first refine the approximate reflection property in Lemma 6.5.4. Compared to Lemma 6.5.4, the following lemma deals with a different range of parameters. We place the proof in Appendix 7.3.4 as it follows from a similar line of reasoning as Lemma 6.5.4.

Lemma 6.5.7 (Approximate reflection across manifold, large steps)**.** *Define*

$$\delta_{\mathrm{Grid}} := \min\left\{\frac{\delta_{\mathrm{A}}}{2}, \frac{1}{C_{\mathcal{M}}(D_1^{-1}+1)}, \frac{\mu}{8(C_{(a)}+\beta)}\right\}.$$

Then for all $x \in B_{\delta_{\mathrm{Grid}}}(\bar{x})$, $\sigma > 0$, and $g \in \partial_\sigma f(x)\backslash\{0\}$ satisfying

$$D_1^{-1}\mathrm{dist}(x, \mathcal{M}) \leq \sigma \leq \delta_{\mathrm{Grid}},$$

we have

$$\langle \hat{g}, g\rangle \leq -D_2\|g\| + 2D_2\|P_{T_{\mathcal{M}}(P_{\mathcal{M}}(x))}(g)\| \qquad \textit{for all } \hat{g} \in \partial_c f\left(x - \sigma\frac{g}{\|g\|}\right).$$

Finally, we prove that `TDescent` eventually generates small subgradients.

Lemma 6.5.8 (`TDescent` yields small subgradients)**.** *Fix $T \in \mathbb{N}$. Then for all $x \in B_{\delta_{\mathrm{Grid}}}(\bar{x})$, $\sigma > 0$, and $g \in \partial_\sigma f(x)\backslash\{0\}$ satisfying*

$$D_1^{-1}\mathrm{dist}(x, \mathcal{M}) \leq \sigma \leq \delta_{\mathrm{Grid}},$$

the vector $g' :=$ `TDescent`(x, g, σ, T) satisfies

$$\|g'\| \leq \max\left\{\left(1 - \frac{\mu^2}{64L^2}\right)^{T/2}\|g\|, 4C_{(a)}\sigma + 4(C_{(a)} + 2\beta)\|x - \bar{x}\|, \frac{8(f(x) - f(\bar{x}))}{\sigma}\right\}.$$

Proof. We begin with preliminary notation and bounds. We fix $x \in B_{\delta_{\mathrm{Grid}}}(\bar{x})$ and subgradient $g \in \partial_\sigma f(x)\backslash\{0\}$. We define $y := P_{\mathcal{M}}(x)$ and $T := T_{\mathcal{M}}(y)$. We also define $c := C_{(a)}(\mathrm{dist}(x, \mathcal{M}) + \sigma) + \beta\|y - \bar{x}\|$. We have the following two bounds: First, we have

$$c \leq C_{(a)}(\|x - \bar{x}\| + \sigma) + 2\beta\|x - \bar{x}\| \leq C_{(a)}\sigma + (C_{(a)} + 2\beta)\|x - \bar{x}\|. \tag{6.5.14}$$

Second, by (6.3.4), we have

$$\|P_T(v)\| \leq c \qquad \text{for all } v \in \partial_\sigma f(x). \tag{6.5.15}$$

We now turn to the proof.

Note that the result holds automatically if $g' = 0$. Thus, we first consider the case where `TDescent` terminates in descent, meaning

$$f(x_+) - f(x) \leq -\frac{\sigma\|g'\|}{8} \qquad \text{where } x_+ := x - \sigma\frac{g'}{\|g'\|}.$$

Since $\sigma \leq \delta_{\text{Grid}} \leq \delta_{\text{A}}/2$ and $x \in B_{\delta_{\text{A}}/2}(\bar{x})$, it follows that $x_+ \in B_{\delta_{\text{A}}}(\bar{x})$. Thus, by Item 1 of Proposition 6.3.5, we have

$$f(x_+) \geq f(\bar{x}) + \frac{\gamma}{2}\|x - \bar{x}\|^2 \geq f(\bar{x}).$$

Consequently, we have

$$f(\bar{x}) - f(x) \leq -\frac{\sigma\|g'\|}{8}.$$

Rearranging then gives the upper bound $\|g'\| \leq \frac{8(f(x)-f(\bar{x}))}{\sigma}$, as desired.

Let us suppose that `TDescent` does not terminate with descent or $g' = 0$. In this case, the iterates $g_0, \ldots, g_T$ of `TDescent`(x, g, σ, T) exist and satisfy $g_t \in \partial_\sigma f(x)$ for all $t \leq T$. We consider two cases.

Case 1. Now suppose $\|g_t\| \leq 4c$ for some t satisfying $0 \leq t \leq T$. Since $\|g_t\|$ is a decreasing sequence, it follows that $\|g'\| = \|g_T\| \leq 4c$. Recalling (6.5.14), yields the bound

$$\|g'\| \leq 4c \leq 4C_{(a)}\sigma + 4(C_{(a)} + 2\beta)\|x - \bar{x}\|,$$

as desired.

Case 2. Next suppose that for all $0 \leq t \leq T$ we have $4c < \|g_t\|$. In this case, Lemma 6.5.7 shows that for all $t \leq T$, we have

$$\langle \hat{g}_t, g_t \rangle \leq -\frac{\mu}{2}\|g_t\| + \mu\|P_T g_t\| \leq -\frac{\mu}{2}\|g_t\| + \mu c \leq -\frac{\mu}{4}\|g_t\|. \tag{6.5.16}$$

We use this bound to prove a one-step geometric improvement bound for $\|g_t\|^2$. To that end, fix any $t \leq T-1$ and define the weight $\lambda := \frac{\mu\|g_t\|}{16L^2}$ and the vector $g_\lambda := g_t + \lambda(\hat{g}_t - g_t)$. Notice that $\lambda \in [0,1]$, since

$$\lambda = \frac{\mu\|g_t\|}{16L^2} \leq \frac{\mu}{16L} \leq 1,$$

where the first equation follows since $g_t \in \partial_\sigma f(x)$ and the second follows since $L \geq \mu$; see Lemma 7.3.2. Thus

$$\begin{aligned}
\|g_{t+1}\|^2 \leq \|g_\lambda\|^2 &= \|g_t\|^2 + 2\lambda\,\langle g_t, \hat{g}_t - g_t\rangle + \lambda^2\|\hat{g}_t - g_t\|^2 \\
&\leq \|g_t\|^2 + 2\lambda\,\langle g_t, \hat{g}_t\rangle - 2\lambda\|g_t\|^2 + 4L^2\lambda^2 \\
&\leq \|g_t\|^2 - \frac{\lambda\mu}{2}\|g_t\| + 4L^2\lambda^2 \\
&= \left(1 - \frac{\mu^2}{64L^2}\right)\|g_t\|^2,
\end{aligned}$$

where the first inequality follows by definition of g_{t+1}; the second inequality follows from the fact that L is a local Lipschitz constant of f near $\bar{x}$; and the third inequality follows from (6.5.16). Thus, to complete the proof, unfold this recursion to get the bound

$$\|g'\| = \|g_T\| \leq \left(1 - \frac{\mu^2}{64L^2}\right)^{T/2}\|g_0\|^2,$$

as desired. □

6.6 Rapid local convergence of `NTDescent`

This Section presents our main convergence guarantees for the `NTDescent` method under Assumption Q. The main results of the section are Theorem 6.6.3 and Theorem 6.6.5, which analyze the nonconvex and convex settings, respectively. In the nonconvex setting, we prove that iterates of `NTDescent` locally nearly linearly converge, provided some iterate reaches a sufficiently small neighborhood of $\bar{x}$. In the convex setting, we strengthen this guarantee, showing that for any initial starting point x_0 and any failure probability p, there exists some index K_p after which `NTDescent` nearly converges linearly with probability at least $1 - p$. Both results result from the local one-step improvement bound of Proposition 6.6.1. This proposition shows that with high probability, the following hold locally for `linesearch`: its output is near its input, and the function gap geometrically decreases whenever it is larger than a quantity exponentially small in the inner loop budget and the grid size. The former property will help ensure that the iterates of `NTDescent` do not escape a local neighborhood of $\bar{x}$.

6.6.1 Assumptions and notation

This section assumes the following assumptions and notations are in force. We assume that

1. the budget T_k and grid size G_k satisfy $\min\{T_k, G_k\} \geq k + 1$ for all $k \geq 0$.
2. We fix an initial we an initial point $x_0 \in \mathbb{R}^d$ and $g_0 \in \partial_c f(x_0)$. We assume that $g_0 \neq 0$. We assume Assumption Q is in force at a point $\bar{x} \in \mathbb{R}^d$ and use the notation of Proposition 6.3.5 throughout. We let $\{x_k\}$ denote the sequence of iterates generated by `NTDescent`$(x_0, g_0, c_0, \{G_k\}, \{T_k\})$ when applied to f.

Turning to notation, we now summarize in Table 6.2 the main constants used in this section.

In the following, we lower and upper bound the trust region parameter in `linesearch`:

$$s_{\mathrm{lb}} \leq \max\{\|g_k\|, c_0\|g_0\|\} \leq L, \tag{6.6.1}$$

where the lower bound follows by definition, and the upper bound follows from Part 6 of Proposition 6.3.5. In addition, we apply Theorem 6.4.3 with the constants a_1, a_2. These

Parameter	Definition
s_{lb}	$c_0 \lVert g_0 \rVert$
a_1	$\min\{D_1, D_2/L\}$
a_2	$\frac{\min\{C_1/L, C_5\}}{2}$
$\delta_{\mathtt{LS}}$	$\min\left\{\frac{\delta_{\mathrm{A}}}{2}, \delta_{\mathrm{GI}}, \delta_{\mathtt{ND}}, \delta_{\mathrm{Grid}}, \frac{1}{2(a_1+2a_2)}, \frac{\gamma D_1^2 \min\{\delta_{\mathrm{Grid}}/2, 1/4\}^2}{2L}, 1\right\}$
C_6	$\max\left\{1, \frac{8(C_{(a)}+2\beta+2C_{(a)}D_1^{-1})}{s_{\mathrm{lb}}}, 2D_1^{-1}, \frac{4\gamma D_1}{s_{\mathrm{lb}}}\right\}$
$\epsilon_{1,T}$	$\max\left\{\frac{(C_5^2 L+\beta)L}{C_3}\left(1-\frac{3\mu^2}{256L^2}\right)^{T/2}, \left(1-\frac{\mu^2}{64L^2}\right)^{T/2} L\right\}$
$\epsilon_{2,G}$	$\max\left\{\frac{L}{\min\{1,a_1\}} + \frac{\beta}{2\min\{1,a_1\}a_2^2}, 8C_{(a)}, L\right\} 2^{-G}$
ρ	$1 - \frac{1}{8}\min\left\{\frac{\gamma a_2}{8\max\{4La_2^2, \beta\}}, \frac{\mu a_1}{4\max\{2L, \beta/a_2^2\}}\right\}$

Table 6.2: Parameters used throughout Section 6.6; see also Table 6.1.

constants are derived from the parameters D_1, D_2, C_1, and C_5 which are defined in Lemmas 6.4.1, 6.4.2, and 6.5.5 respectively. We also define a neighborhood $B_{\delta_{\mathtt{LS}}}(\bar{x})$ for which `linesearch` results in geometric improvement. Here, the radius $\delta_{\mathtt{LS}}$ is derived from the parameters δ_{A}, δ_{GI}, $\delta_{\mathtt{ND}}$, δ_{Grid}, and γ which appear in Proposition 6.3.5 and Lemmas 6.4.1, 6.5.1, 6.5.7, and 6.5.8. In addition, the constant C_6 will appear in an upper bound on the steplength of `linesearch`.

We then define three terms $\epsilon_{1,T}$, $\epsilon_{2,G}$, and ρ which appear in our convergence rate analysis. These terms are defined for all $T, G > 0$ and are derived from the parameters C_5, C_3, a_1, a_2, L, β, $C_{(a)}$, γ, and μ which appear in Lemma 6.5.2, Lemma 6.5.5, Proposition 6.3.5, and Assumption Q.

Finally, in the following propositions, the constant $\rho \in (0, 1)$ plays the role of a local contraction factor, while the terms $\epsilon_{1,T}$ and $\epsilon_{2,G}$ are upper bounds for function gap of `NTDescent`.

We now turn to the one-step improvement argument.

6.6.2 One step improvement

The following proposition presents our one-step improvement bound.

Proposition 6.6.1 (One step improvement)**.** *Assume the assumptions of Section 6.6.1 are satisfied. Recall the notation in Table 6.2. Then the following holds for all $x \in B_{\delta_{\mathrm{LS}}}(\bar{x})$, subgradients $g \in \partial_c f(x)$, and grid sizes $G > \lceil \log_2(1/\delta_{\mathrm{Grid}}) \rceil$: Fix a scalar $s \in [s_{\mathrm{lb}}, L]$, a failure probability $p \in (0, 1)$ and budget T satisfying*

$$T \geq \left\lceil \frac{256L^2}{\mu^2} \right\rceil \lceil 2\log(1/p) \rceil .$$

Then with probability at least $1 - p$, the point $\tilde{x} = \texttt{linesearch}(x, g, s, G, T)$ satisfies

1. $f(\tilde{x}) - f(\bar{x}) \leq \max\{\rho(f(x) - f(\bar{x})), \epsilon_{1,T}, \epsilon_{1,G}\}$;
2. $\|\tilde{x} - x\| \leq C_6 \max\left\{\epsilon_{1,T}/s_{\mathrm{lb}}, \epsilon_{2,G}/s_{\mathrm{lb}}, \sqrt{2(f(x) - f(\bar{x}))/\min\{s_{\mathrm{lb}}, \gamma\}}\right\}$,

Proof. We fix $x \in B_{\delta_{\mathrm{LS}}}(\bar{x})$, define $y := P_{\mathcal{M}}(x)$, and choose a subgradient $g \in \partial_c f(x)$. Throughout, we may freely use the results of Proposition 6.3.5 since $\delta_{\mathrm{LS}} \leq \delta_{\mathrm{A}}$. We will first establish the first item of the proposition. To that end, let us assume that

$$f(x) - f(\bar{x}) > \max\{\epsilon_{1,T}, \epsilon_{2,G}\};$$

otherwise, the proof is trivial. In this case, we claim that x must satisfy either Item 1 or Item 2 of Theorem 6.4.3 for at least one σ_i with $i \leq G - 1$. To derive a contradiction, suppose that both items are unsatisfied for x with any choice of σ_i with $i = 0, \ldots, G - 1$. We will show that neither Item 1b nor its complement can be satisfied, leading to a contradiction.

Throughout the following argument, we will use the following bound:

$$\max\{a_1 \mathrm{dist}(x, \mathcal{M}), a_2\|y - \bar{x}\|\} \leq (a_1 + 2a_2)\,\delta_{\mathrm{LS}} \leq \frac{1}{2} = \sigma_{G-1}.$$

Now suppose that Item 1b holds, i.e., $a_2^2\|y-\bar{x}\|^2 \leq \operatorname{dist}(x,\mathcal{M})$. Then by assumption, Item 1a must fail for any σ_i. We claim that this failure ensures that $\sigma_0 > a_1\operatorname{dist}(x,\mathcal{M})$. Indeed, if $\sigma_0 \leq a_1\operatorname{dist}(x,\mathcal{M})$, we must have

$$\sigma_0 \leq (a_1/2)\operatorname{dist}(x,\mathcal{M}) \leq a_1\operatorname{dist}(x,\mathcal{M}) \leq \sigma_{G-1},$$

since σ_0 cannot satisfy Item 1a. Thus, there exists some $j \leq G-1$ such that $\sigma_j = 2^j\sigma_0$ satisfies Item 1a, a contradiction. Therefore, we have

$$\sigma_0 > a_1\operatorname{dist}(x,\mathcal{M}) \geq a_1a_2^2\|y-\bar{x}\|^2.$$

In this case, by (6.3.7), we have

$$f(x)-f(\bar{x}) \leq L\operatorname{dist}(x,\mathcal{M}) + \frac{\beta}{2}\|y-\bar{x}\|^2 \leq \left(\frac{L}{a_1}+\frac{\beta}{2a_2^2a_1}\right)\sigma_0 \leq \epsilon_{2,G},$$

which is a contradiction. Therefore, Item 1b cannot hold, so we have $a_2^2\|y-\bar{x}\|^2 > \operatorname{dist}(x,\mathcal{M})$.

Next, for the sake of contradiction, suppose that there exists σ_i satisfying Item 2a. In this case, since $\sigma_i \geq (a_2/2)\|y-\bar{x}\|$, we have

$$\operatorname{dist}(x,\mathcal{M}) < a_2^2\|y-\bar{x}\|^2 \leq 2a_2\sigma_i\|y-\bar{x}\|,$$

i.e., σ_i also satisfies Item 2b, which is a contradiction. Therefore no σ_i satisfies Item 2a. We claim that this ensures $\sigma_0 > a_2\|y-\bar{x}\|$. Indeed, if $\sigma_0 \leq a_2\|y-\bar{x}\|$, we must have

$$\sigma_0 \leq (a_2/2)\|y-\bar{x}\| \leq a_0\|y-\bar{x}\| \leq \sigma_{G-1},$$

since σ_0 cannot satisfy Item 2a. Thus, there exists some $j \leq G-1$ such that $\sigma_j = 2^j\sigma_0$ satisfies Item 2a, a contradiction. Therefore, we have

$$\sigma_0 > a_2\|y-\bar{x}\| \geq \sqrt{\operatorname{dist}(x,\mathcal{M})}.$$

In this case, by (6.3.7), we have

$$f(x)-f(\bar{x}) \leq L\operatorname{dist}(x,\mathcal{M}) + \frac{\beta}{2}\|y-\bar{x}\|^2 \leq \left(L+\frac{\beta}{2a_2^2}\right)\sigma_0^2 \leq \epsilon_{2,G},$$

which is a contradiction. Therefore, there must exist σ_i satisfying either Item 1 or Item 2 of Theorem 6.4.3.

Let us now fix a σ_i satisfying either Item 1 or Item 2 of Theorem 6.4.3. Then, by Theorem 6.4.3, we have the bound

$$\sigma_i \mathrm{dist}(0, \partial_{\sigma_i} f(x)) \geq 8(1-\rho)(f(x) - f(\bar{x})).$$

In what follows, we will use the above bound to prove that with probability at least $1-p$, we have $f(\tilde{x}) - f(\bar{x}) \leq \rho(f(x) - f(\bar{x}))$ whenever $f(x) - f(\bar{x}) > \max\{\epsilon_{1,T}, \epsilon_{2,G}\}$.

Contraction case 1: normal step. We first suppose that there exists σ_i satisfying Item 1. In the interest of analyzing $v_{i+1} \in \partial_{\sigma_i} f(x)$, let us show that x, σ_i, and T satisfy the conditions of Proposition 6.5.1: First $x \in B_{\delta_{\mathtt{ND}}}(\bar{x})$ since $\delta_{\mathrm{LS}} \leq \delta_{\mathtt{ND}}$. Second, by Item 1a of Theorem 6.4.3, we have

$$0 < \sigma_i \leq a_1 \mathrm{dist}(x, \mathcal{M}) \leq D_1 \mathrm{dist}(x, \mathcal{M}). \tag{6.6.2}$$

Finally, from the definition $D_2 = \mu/2$, it follows that T satisfies the conditions of Proposition 6.5.1. Therefore, with probability at least $1-p$, we have

$$f\left(x - \sigma_i \frac{v_{i+1}}{\|v_{i+1}\|}\right) - f(\bar{x}) \leq f(x) - f(\bar{x}) - \frac{\sigma_i}{8}\mathrm{dist}(0, \partial_{\sigma_i} f(x)) \leq \rho(f(x) - f(\bar{x})).$$

Next, we show that v_{i+1} and σ_i satisfy the trust region condition $\sigma_i \leq \frac{\|v_{i+1}\|}{s}$. To that end, note that the conditions of Lemma 6.4.1 are met: We have $x \in B_{\delta_{\mathrm{GI}}}(\bar{x})$ since $\delta_{\mathrm{LS}} \leq \delta_{\mathrm{GI}}$. We also have bound $\sigma_i \leq D_1 \mathrm{dist}(x, \mathcal{M})$ from (6.6.2). Therefore, it follows that the minimal norm Goldstein subgradient is lower bounded: $\mathrm{dist}(0, \partial_{\sigma_i} f(x)) \geq D_2$. Consequently, we have

$$\sigma_i \leq a_1 \mathrm{dist}(x, \mathcal{M}) \leq \frac{D_2 \delta_{\mathrm{LS}}}{s} \leq \frac{\mathrm{dist}(0, \partial_{\sigma_i} f(x))\delta_{\mathrm{LS}}}{s} \leq \frac{\|v_{i+1}\|}{s},$$

where the second inequality follows from the definition of in Table 6.2 and the inequality $s \leq L$; and the fourth inequality follows from the bound $\delta_{\mathrm{LS}} \leq 1$. Therefore, since the trust region constraint $\sigma_i \leq \frac{\|v_{i+1}\|}{s}$ is satisfied, the following holds with probability at least $1 - p$:

$$f(\tilde{x}) - f(\bar{x}) \leq f\left(x - \sigma_i \frac{v_{i+1}}{\|v_{i+1}\|}\right) - f(\bar{x}) \leq \rho(f(x) - f(\bar{x})).$$

Thus, the first item of the proposition follows.

Contraction case 2: tangent step. Next, we suppose that there exists σ_i satisfying Item 2 of Theorem 6.4.3. In the interest of analyzing $u_i \in \partial_{\sigma_i} f(x)$, let us show that x, σ_i, and T satisfy the conditions of Proposition 6.5.6: $x \in B_{\delta_{\mathrm{A}}/2}(\bar{x})$ since $\delta_{\mathrm{LS}} \leq \delta_{\mathrm{A}}/2$. Second, by Item 2a of Theorem 6.4.3, we have

$$\sigma_i \leq a_2\|y - \bar{x}\| \leq C_5\|y - \bar{x}\|.$$

Finally, by Item 2b of Theorem 6.4.3, we have

$$\operatorname{dist}(x, \mathcal{M})/\sigma_i \leq 2a_2\|y - \bar{x}\| \leq C_5\|y - \bar{x}\|.$$

Therefore, since $f(x) - f(\bar{x}) > \epsilon_{1,T}$, Proposition 6.5.6 implies that

$$f\left(x - \sigma_i \frac{u_i}{\|u_i\|}\right) - f(\bar{x}) \leq f(x) - f(\bar{x}) - \frac{\sigma_i}{8}\operatorname{dist}(0, \partial_{\sigma_i} f(x)) \leq \rho(f(x) - f(\bar{x})).$$

Next, we show that u_i and σ_i satisfy the trust region condition $\sigma_i \leq \frac{\|u_i\|}{s}$. To show this, we first note that σ_i and x satisfy the conditions of Lemma 6.4.2: First $x \in B_{\delta_{\mathrm{A}}/2}(\bar{x})$ since $\delta_{\mathrm{LS}} \leq \delta_{\mathrm{A}}/2$. Second, by Item 2a of Theorem 6.4.3, we have

$$\sigma_i \leq a_2\|y - \bar{x}\| \leq C_2\|y - \bar{x}\|.$$

Finally, by Item 2 of Theorem 6.4.3, we have

$$\operatorname{dist}(x, \mathcal{M}) \leq 2a_2\sigma_i\|y - \bar{x}\| \leq 2a_2^2\|y - \bar{x}\|^2 \leq C_2\|y - \bar{x}\|,$$

where the third inequality follows from the bounds $\|y-\bar{x}\| \leq 2\delta_{\text{LS}} \leq 1/a_2$ and $a_2 \leq C_2/2$ (recall that $C_5 \leq C_2$). Therefore, by Lemma 6.4.2 we have $\|u_i\| \geq \|P_{T_{\mathcal{M}}(y)} u_i\| \geq C_1\|y-\bar{x}\|$. Consequently, we have

$$\sigma_i \leq a_2\|y-\bar{x}\| \leq \frac{C_1\|y-\bar{x}\|}{s} \leq \frac{\|u_i\|}{s},$$

where the second inequality follows from the definition of a_2 in Table 6.2 and the inequality $s \leq L$. To complete the proof, observe that $v_{i+1} = u_i$: since the sufficient descent condition is met, namely $f(x - \sigma_i u_i/\|u_i\|) \leq f(x) - \sigma\|u_i\|$, `NDescent` terminates at the first iteration. Therefore, we must have

$$f(\tilde{x}) - f(\bar{x}) \leq f\left(x - \sigma_i \frac{v_{i+1}}{\|v_{i+1}\|}\right) - f(\bar{x}) \leq \rho(f(x) - f(\bar{x})),$$

as desired.

Having proved the desired contraction $f(\tilde{x}) - f(\bar{x}) \leq \rho(f(x) - f(\bar{x}))$, we now turn to the bound on $\|\tilde{x} - x\|$.

Stepsize bound. We now no longer assume that $f(x) - f(\bar{x}) > \max\{\epsilon_{2,G}, \epsilon_{1,T}\}$. We claim that we have

$$\max_{0\leq i\leq G-1} \{\sigma_i \colon \sigma_i \leq \|v_{i+1}\|/s\} \leq C_6 \max\left\{\epsilon_{1,T}/s_{\text{lb}}, \epsilon_{2,G}/s_{\text{lb}}, \sqrt{2(f(x)-f(\bar{x}))/\min\{s_{\text{lb}}, \gamma\}}\right\}. \tag{6.6.3}$$

Note that inequality (6.6.3) immediately yields the second item of the proposition since

$$\|\tilde{x} - x\| \leq \max_{0\leq i\leq G-1} \{\sigma_i \colon \sigma_i \leq \|v_{i+1}\|/s\}.$$

To prove (6.6.3), we will apply Lemma 6.5.8.

To that end, first note that $x \in B_{\delta_{\text{Grid}}}(\bar{x})$ since $\delta_{\text{LS}} \leq \delta_{\text{Grid}}$. Next, we verify that there exists an index i such that σ_i satisfies a slightly stronger version of the assumptions of

Lemma 6.5.8. Indeed, recall that by the quadratic growth condition (Q1), we have the bound

$$\operatorname{dist}(x, \mathcal{M}) \leq \|x - \bar{x}\| \leq \sqrt{2(f(x) - f(\bar{x}))/\gamma}. \tag{6.6.4}$$

Thus, to satisfy the assumptions of Lemma 6.5.8, we prove that there exists i such that

$$R_x := D_1^{-1} \sqrt{2(f(x) - f(\bar{x}))/\gamma} \leq \sigma_i \leq \delta_{\text{Grid}}. \tag{6.6.5}$$

Indeed, first notice that $\sigma_0 \leq \delta_{\text{Grid}}$ since $G \geq \lceil \log_2(1/\delta_{\text{Grid}}) \rceil$. Thus, if $\sigma_0 \geq R_x$, the bound (6.6.5) holds for σ_0. If instead $\sigma_0 < R_x$, we have

$$\sigma_0 < R_x \leq D_1^{-1} \sqrt{2L\delta_{\text{LS}}/\gamma} \leq \min\{\delta_{\text{Grid}}/2, 1/4\} \leq \min\{\delta_{\text{Grid}}, 1/2\} \leq 1/2 = \sigma_{G-1},$$

where the second inequality follows since $\|x - \bar{x}\| < \delta_{\text{LS}}$ and f is L–Lipschitz continuous on $B_{\delta_{\text{LS}}}(\bar{x})$; and the third inequality follows since $\delta_{\text{LS}} \leq \gamma D_1^2 \min\{\delta_{\text{Grid}}/2, 1/4\}^2/(2L)$. Thus, there exists i such that $\sigma_i \in [\min\{\delta_{\text{Grid}}/2, 1/4\}, \min\{\delta_{\text{Grid}}, 1/2\}]$. Since $\min\{\delta_{\text{Grid}}/2, 1/4\} \geq R_x$, inequality (6.6.5) follows.

Now let i_* be the minimal such index such that (6.6.5) is satisfied for $i = i_*$. If $i_* \neq 0$, the bound $\sigma_{i_*-1} \leq R_x$ holds. In particular, $\sigma_{i_*} \leq 2R_x$. Therefore, considering the cases $i_* = 0$ and $i_* \neq 0$ separately, we have

$$R_x \leq \sigma_{i_*} \leq \max\{\sigma_0, 2R_x\}. \tag{6.6.6}$$

Now we bound the step length $\|x - \tilde{x}\|$ by considering two cases.

First suppose that $\sigma_{i_*} > \|u_{i_*}\|/s$. In this case, (6.2.1) ensures $\sigma_{i_*} > \|v_{i_*+1}\|/s$. Then, since σ_i is increasing in i, we have

$$\begin{aligned}\max_{0 \leq i \leq G-1} \{\sigma_i \colon \sigma_i \leq \|v_{i+1}\|/s\} \leq \sigma_{i_*} &\leq \max\{\sigma_0, 2R_x\} \\ &\leq C_6 \max\left\{\epsilon_{2,G}/s_{\text{lb}}, \sqrt{2(f(x) - f(\bar{x}))/\min\{s_{\text{lb}}, \gamma\}}\right\}.\end{aligned}$$

$$\leq C_6 \max\left\{\epsilon_{1,T}/s_{\text{lb}}, \epsilon_{2,G}/s_{\text{lb}}, \sqrt{2(f(x) - f(\bar{x}))/\min\{s_{\text{lb}}, \gamma\}}\right\},$$

which verifies (6.6.3). We now consider the alternative case.

Next suppose that $\sigma_{i_*} \leq \|u_{i_*}\|/s$. We consider two subcases. First, suppose that the following bound also holds:

$$\|u_{i_*}\| \leq \frac{8(f(x) - f(\bar{x}))}{\sigma_{i_*}}. \tag{6.6.7}$$

Then, since $\sigma_{i_*} \geq R_x$, we have

$$\|u_{i_*}\| \leq \sqrt{32\gamma D_1^2(f(x) - f(\bar{x}))}.$$

Second, suppose that (6.6.7) does not hold. Let us apply Lemma 6.5.8 to $\sigma = \sigma_{i_*}$:

$$\begin{aligned}
\|u_{i_*}\| &\leq \max\left\{\left(1 - \frac{\mu^2}{64L^2}\right)^{T/2} L, 4C_{(a)}\max\{\sigma_0, 2R_x\} + 4(C_{(a)} + 2\beta)\|x - \bar{x}\|\right\} \\
&\leq \max\left\{\left(1 - \frac{\mu^2}{64L^2}\right)^{T/2}, 8C_{(a)}\sigma_0, 8(C_{(a)} + 2\beta)\|x - \bar{x}\| + 16C_{(a)}R_x\right\} \\
&\leq \max\left\{1 \cdot \epsilon_{1,T}, 1 \cdot \epsilon_{2,G}, 8(C_{(a)} + 2\beta + 2C_{(a)}D_1^{-1})\sqrt{2(f(x) - f(\bar{x}))/\gamma}\right\} \\
&\leq sC_6 \max\{\epsilon_{1,T}/s_{\text{lb}}, \epsilon_{2,G}/s_{\text{lb}}, \sqrt{2(f(x) - f(\bar{x}))/\gamma}\},
\end{aligned}$$

where first inequality follows from Lemma 6.5.8 and bound (6.6.6); the second inequality follows from the bound: $\max\{a, b\} + c \leq a + b + c \leq 2\max\{a, b + c\}$ for all $a, b, c \geq 0$; the third inequality follows by definition of $\epsilon_{1,T}$, $\epsilon_{2,G}$ and R_x, and (6.6.4); and the last inequality follows since $C_6 \geq \max\{1, 8(C_{(a)} + 2\beta + 2C_{(a)}D_1^{-1})/s_{\text{lb}}\}$.

Therefore, as long as $\sigma_{i_*} \leq \|u_{i_*}\|/s$, we have

$$\begin{aligned}
\|u_{i_*}\|/s &\leq \max\left\{C_6 \max\{\epsilon_{1,T}/s_{\text{lb}}, \epsilon_{2,G}/s_{\text{lb}}, \sqrt{2(f(x) - f(\bar{x}))/\gamma}\}, \sqrt{32\gamma D_1^2(f(x) - f(\bar{x}))/s_{\text{lb}}^2}\right\} \\
&\leq C_6 \max\left\{\epsilon_{1,T}/s_{\text{lb}}, \epsilon_{2,G}/s_{\text{lb}}, \sqrt{2(f(x) - f(\bar{x}))/\gamma}\right\} \\
&\leq C_6 \max\left\{\epsilon_{1,T}/s_{\text{lb}}, \epsilon_{2,G}/s_{\text{lb}}, \sqrt{2(f(x) - f(\bar{x}))/\min\{s_{\text{lb}}, \gamma\}}\right\},
\end{aligned}$$

where second inequality follows from the bound $C_6 \geq 4\gamma D_1/(s_{\text{lb}})$; and the third inequality follows from the bound (6.6.4). To complete the proof of (6.6.3), recall that by (6.2.1), for all $j > i_*$, we have $\|v_j\| \leq \|u_{i_*}\|$. Consequently,

$$\begin{aligned}
&\max_{0\leq i\leq G-1} \{\sigma_i \colon \sigma_i \leq \|v_{i+1}\|/s\} \\
&\leq \max\{\sigma_{i_*}, \|u_{i_*}\|/s\} \\
&= \|u_{i_*}\|/s \\
&\leq C_6 \max\left\{\epsilon_{1,T}/s_{\text{lb}}, \epsilon_{2,G}/s_{\text{lb}}, \sqrt{2(f(x) - f(\bar{x}))/\min\{s_{\text{lb}}, \gamma\}}\right\},
\end{aligned}$$

which verifies (6.6.3). □

6.6.3 Main convergence theorems

We are now ready to prove the main results of this chapter. This section aims to prove that an event of the following form occurs with high probability.

Definition 6.6.2 ($E_{k_0,q,C}$)**.** For any $k_0 > 0$, $q \in (0, 1)$ and $C > 0$, let $E_{k_0,q,C}$ denote the event that for all $k \geq k_0$, we have the following two bounds:

$$\begin{aligned}
f(x_k) - f(\bar{x}) &\leq \max\{(f(x_{k_0}) - f(\bar{x}))q^{k-k_0}, Cq^k\}; \\
\|x_k - \bar{x}\|^2 &\leq \frac{2}{\gamma} \max\{(f(x_{k_0}) - f(\bar{x}))q^{k-k_0}, Cq^k\}.
\end{aligned}$$

We will lower bound the probability of the event $E_{k_0,q,C}$ in both nonconvex and convex settings for a particular choice of k_0, q, and C. In the nonconvex setting, our result will lower bound the conditional probability of $E_{k_0,q,C}$, given that iterate x_{k_0} enters a sufficiently small neighborhood of $\bar{x}$. To prove the result, we will iterate the one-step improvement bound of Proposition 6.6.1. In the convex setting, we will lower bound the unconditional probability of $E_{k_0,q,C}$. To prove this result, we will combine the conditional result with the sublinear convergence guarantee of Theorem 6.2.4.

Before turning to the proofs, we introduce the main parameters common to nonconvex and convex settings.

Parameter	Definition
C'	$\frac{2048L^2}{\mu^2}$
C	$\max\left\{\frac{(C_5^2L+\beta)L}{C_3}, L, \frac{L}{\min\{1,a_1\}} + \frac{\beta}{2\min\{1,a_1\}a_2^2}, 8C_{(a)}\right\}$
q	$\max\left\{\rho, \sqrt{1-\frac{3\mu^2}{256L^2}}, \frac{1}{2}\right\}$
$\delta_{\mathtt{NTD}}$	$\min\left\{\frac{\delta_{\mathrm{LS}}}{4}, \frac{\delta_{\mathrm{LS}}^2\min\{s_{\mathrm{lb}},\gamma\}(1-q^{1/2})^2}{32LC_6^2}, \frac{\delta_{\mathrm{LS}}s_{\mathrm{lb}}(1-q)}{4LC_6}\right\}$
K_0	$\left\lceil\max\left\{\log_q\left(\frac{\delta_{\mathrm{LS}}^2\min\{s_{\mathrm{lb}},\gamma\}(1-q^{1/2})^2}{32CC_6^2}\right), \log_q\left(\frac{\delta_{\mathrm{LS}}s_{\mathrm{lb}}(1-q)}{4CC_6}\right), \log_2\left(\frac{1}{\delta_{\mathrm{Grid}}}\right)\right\}\right\rceil$

Table 6.3: Parameters used throughout Section 6.6.3; see also Tables 6.1 and 6.2.

6.6.3.1 The nonconvex setting.

The following theorem is our main convergence theorem in the nonconvex setting.

Theorem 6.6.3 (Main Theorem: Nonconvex Setting)**.** *Assume the assumptions outlined at the start of Section 6.6 are satisfied. Recall the notation of Table 6.3. Fix a failure probability $p \in (0,1)$ and an index $k_0 \geq \max\{K_0, C'\log(C'/p)\}$. Suppose $P(x_{k_0} \in B_{\delta_{\mathtt{NTD}}}(\bar{x})) > 0$. Then,*

$$P(E_{k_0,q,C} \mid x_{k_0} \in B_{\delta_{\mathtt{NTD}}}(\bar{x})) \geq 1-p.$$

Proof. We begin with preliminary notation and bounds. Fix $k_0 \geq \max\{K_0, C'\log(C'/p)\}$ and for all $k \geq k_0$, define the quantity

$$R_k := \max\{(f(x_{k_0}) - f(\bar{x}))q^{k-k_0}, Cq^k\}.$$

Note that whenever $x_{k_0} \in B_{\delta_{\mathtt{NTD}}}(\bar{x})$ we have the bound

$$R_k \leq \max\{L\delta_{\mathtt{NTD}}q^{k-k_0}, Cq^k\}, \tag{6.6.8}$$

since f is L-Lipschitz continuous on $B_\delta(\bar{x})$.

Next, we prove that

$$\max\{\epsilon_{1,T_k}, \epsilon_{2,G_k}\} \leq R_{k+1} \qquad \text{for all } k \geq 0. \tag{6.6.9}$$

Indeed, beginning with ϵ_{1,T_k}, we have

$$\begin{aligned}
\epsilon_{1,T_k} &= \max\left\{\frac{(C_5^2 L + \beta)L}{C_3}\left(1 - \frac{3\mu^2}{256L^2}\right)^{T_k/2}, \left(1 - \frac{\mu^2}{64L^2}\right)^{T_k/2} L\right\} \\
&\leq C \max\left\{\left(1 - \frac{3\mu^2}{256L^2}\right)^{\frac{T_k}{2}}, \left(1 - \frac{\mu^2}{64L^2}\right)^{\frac{T_k}{2}}\right\} \\
&\leq Cq^{k+1} \leq R_{k+1},
\end{aligned}$$

where the first and second inequalities follow from the definitions of C and q together with the lower bound $T_k \geq k+1$. Turning to ϵ_{2,G_k}, we have

$$\epsilon_{2,G_k} = \max\left\{\frac{L}{\min\{1, a_1\}} + \frac{\beta}{2\min\{1, a_1\}a_2^2}, 8C_{(a)}, L\right\} 2^{-G_k} \leq C2^{-G_k} \leq Cq^{k+1} \leq R_{k+1},$$

where the first and second inequalities follow from the definition of C and q together with the lower bound $G_k \geq k+1$. Thus (6.6.9) holds.

Finally, we analyze the quantity

$$D_{k_0,\delta_{\mathtt{NTD}}} := \sum_{k=k_0}^{\infty} C_6 \max\left\{\sqrt{2R_k/\gamma'}, R_{k+1}/s_{\mathrm{lb}}\right\} \qquad \text{where } \gamma' := \min\{s_{\mathrm{lb}}, \gamma\}.$$

We claim in particular that

$$D_{k_0,\delta_{\mathtt{NTD}}} + \delta_{\mathtt{NTD}} \leq \delta_{\mathtt{LS}}/2. \tag{6.6.10}$$

Since $\delta_{\mathtt{NTD}} \leq \delta_{\mathtt{LS}}/4$, it suffices to prove $D_{k_0,\delta_{\mathtt{NTD}}} \leq \delta_{\mathtt{LS}}/4$. To that end, we have

$$\begin{aligned}
D_{k_0,\delta_{\mathtt{NTD}}} &= \sum_{k=k_0}^{\infty} C_6 \max\left\{\sqrt{2R_k/\gamma'}, R_{k+1}/s_{\mathrm{lb}}\right\} \\
&\leq \sum_{k=k_0}^{\infty} C_6 \max\left\{\sqrt{2\max\{L\delta_{\mathtt{NTD}}q^{k-k_0}, Cq^k\}/\gamma'}, \max\{L\delta_{\mathtt{NTD}}q^{k-k_0}/\gamma', Cq^k\}/s_{\mathrm{lb}}\right\}
\end{aligned}$$

$$\leq C_6 \max\left\{\frac{\sqrt{2L\delta_{\mathtt{NTD}}}}{\sqrt{\gamma'}(1-q^{1/2})}, \frac{\sqrt{2Cq^{k_0}}}{\sqrt{\gamma'}(1-q^{1/2})}, \frac{L\delta_{\mathtt{NTD}}}{s_{\mathrm{lb}}(1-q)}, \frac{Cq^{k_0}}{s_{\mathrm{lb}}(1-q)}\right\}$$
$$\leq \frac{\delta_{\mathtt{LS}}}{4},$$

where the first inequality follows from the bounds (6.6.8) and the bound $R_{k+1} \leq R_k$; the second inequality follows by summing the infinite series; and the third inequality follows from the definitions of K_0 and $\delta_{\mathtt{NTD}}$ together with the bound $k_0 \geq K_0$. This proves (6.6.10).

We now turn to the proof. Consider the following sequence defined for all $k \geq k_0$:

$$b_k := \delta_{\mathtt{NTD}} + \sum_{j=k_0}^{k-1} C_6 \max\left\{\sqrt{2R_k/\gamma'}, R_{k+1}/s_{\mathrm{lb}}\right\}.$$

Note that (6.6.10) ensures that $b_k \leq \delta_{\mathtt{LS}}/2$ for all $k \geq k_0$. Now, define the event

$$F_{k_0} := \{x_{k_0} \in B_{\delta_{\mathtt{NTD}}}(\bar{x})\}.$$

In addition, define the following decreasing sequence of events

$$A_k := \bigcap_{j=k_0}^{k} \{f(x_j) - f(\bar{x}) \leq R_j \text{ and } \|x_j - \bar{x}\| \leq b_j\}.$$

We claim that

$$P(A_{k+1} \mid A_k \cap F_{k_0}) \geq 1 - \exp(-T_k/C') \qquad \text{for all } k \geq k_0. \tag{6.6.11}$$

Indeed, Proposition 6.6.1 implies that conditioned on $A_k \cap F_{k_0}$, the following four inequalities are satisfied with probability at least $1 - \exp(-T_k/C')$:

1. $f(x_k) - f(\bar{x}) \leq R_k$
2. $\|x_k - \bar{x}\| \leq b_k$;
3. $\|x_{k+1} - x_k\| \leq C_6 \max\left\{\epsilon_{1,T_k}/s_{\mathrm{lb}}, \epsilon_{2,G_k}/s_{\mathrm{lb}}, \sqrt{2(f(x_k) - f(\bar{x}))/\gamma'}\right\}$;
4. $f(x_{k+1}) - f(\bar{x}) \leq \max\{\rho(f(x_k) - f(\bar{x})), \epsilon_{1,T_k}, \epsilon_{2,G_k}\}$.

(Note that in applying the Proposition 6.6.1, we use the scalar $s = \max\{\|g_k\|, c_0\|g_0\|\}$ and the inclusion $s \in [s_{\mathrm{lb}}, L]$, which was proved (6.6.1).) Thus, the bound (6.6.11) will follow by induction if we can prove that whenever the above four inequalities hold, we have $\|x_{k+1} - \bar{x}\| \leq b_{k+1}$ and $f(x_{k+1}) - f(\bar{x}) \leq R_{k+1}$.

To that end, we first prove $\|x_{k+1} - \bar{x}\| \leq b_{k+1}$. Indeed,

$$\begin{aligned}
\|x_{k+1} - \bar{x}\| &\leq \|x_{k+1} - x_k\| + \|x_k - \bar{x}\| \\
&\leq C_6 \max\left\{\epsilon_{1,T_k}/s_{\mathrm{lb}}, \epsilon_{2,G_k}/s_{\mathrm{lb}}, \sqrt{2(f(x_k) - f(\bar{x}))/\gamma'}\right\} + b_k \\
&\leq C_6 \max\left\{\sqrt{2R_k/\gamma'}, R_{k+1}/s_{\mathrm{lb}}\right\} + b_k = b_{k+1}, \qquad (6.6.12)
\end{aligned}$$

where the second inequality follows from Proposition 6.6.1; and the third inequality follows from the bound (6.6.9). Next, we prove the bound on $f(x_{k+1}) - f(\bar{x}) \leq R_{k+1}$. Indeed,

$$\begin{aligned}
f(x_{k+1}) - f(\bar{x}) &\leq \max\{\rho(f(x_k) - f(\bar{x})), \epsilon_{1,T_k}, \epsilon_{2,G_k}\} \\
&\leq \max\{\rho \max\{(f(x_{k_0}) - f(\bar{x}))q^{k-k_0}, Cq^k\}, \epsilon_{1,T_k}, \epsilon_{2,G_k}\} \\
&\leq R_{k+1},
\end{aligned}$$

where the final inequality follows from (6.6.9) and the bound $\rho \leq q$. Consequently, the bound (6.6.11) holds. Moreover, due to the bound $T_k \geq k + 1$, we have

$$P(A_{k+1} \mid A_k \cap F_{k_0}) \geq 1 - \exp(-T_k/C') \geq 1 - \exp(-(k+1)/C'). \qquad (6.6.13)$$

Now we relate A_k to $E_{k_0,q,C}$. To that end, by the conditional law of total probability, for all $k \geq k_0$, we have

$$P(A_{k+1} \mid F_{k_0}) \geq P(A_{k+1} \mid A_k \cap F_{k_0})P(A_k \mid F_{k_0}) \geq P(A_k \mid F_{k_0}) - \exp(-(k+1)/C').$$

Therefore, for all $k \geq k_0$, we have

$$P(A_k \mid F_{k_0}) \geq P(A_{k_0} \mid F_{k_0}) - \sum_{j=k_0+1}^{\infty} \exp(-j/C') = 1 - \frac{\exp(-\frac{k_0+1}{C'})}{1 - \exp(-\frac{1}{C'})} \geq 1 - p,$$

where the equality follows since $P(A_{k_0} \mid F_{k_0}) = 1$; and the final inequality follows by definition of $k_0 \geq C' \log(C'/p)$. Now recall that $\sup_{k \geq k_0} b_k \leq \delta_{\mathrm{LS}}/2$. Therefore, defining the event

$$E'_{k_0,q,C} := \{f(x_k) - f(\bar{x}) \leq R_k \text{ for all } k \geq k_0 \text{ and } x_k \in B_{\delta_{\mathrm{LS}}}(\bar{x})\},$$

we have

$$P(E'_{k_0,q,C} \mid F_{k_0}) \geq \lim_{k \to \infty} P(A_k \mid F_{k_0}) \geq 1 - p.$$

Next, recall that since $\delta_{\mathrm{LS}} \leq \delta_{\mathrm{A}}$, the quadratic growth bound (Q1)

$$\|x_k - \bar{x}\|^2 \leq \frac{2}{\gamma}(f(x_k) - f(\bar{x})) \leq \frac{2}{\gamma} R_k$$

holds for every $k \geq k_0$ within the event $E'_{k_0,q,C}$. Thus, $E_{k_0,q,C} \supseteq E'_{k_0,q,C}$. Therefore, we have

$$P(E_{k_0,q,C} \mid F_{k_0}) \geq P(E'_{k_0,q,C} \mid F_{k_0}) \geq 1 - p,$$

as desired. □

6.6.3.2 The convex setting.

Now, we turn to the convex setting. Our goal is to prove a lower bound on $P(E_{k_0,q,C})$ for q and C chosen as in Table 6.3 and all sufficiently large k_0. Before stating the result, we recall a simple fact about convex functions satisfying Assumption Q. A similar result appears in [122, Section 2.4], but for completeness we provide a proof in Appendix 7.3.6.

Lemma 6.6.4. *In addition to the assumption set out at the start of the section, suppose that function f is convex. Then for all $a > 0$, we have*

$$\{x \in \mathbb{R}^d \colon f(x) - f(\bar{x}) \leq a\} \subseteq \overline{B}_{r_a}(\bar{x}) \qquad \textit{where } r_a := \max\left\{\frac{2a}{\gamma\delta_A}, \sqrt{\frac{2a}{\gamma}}\right\}.$$

In particular, f has bounded sublevel sets.

We now turn to our main theorem.

Theorem 6.6.5 (Main Theorem: Convex setting)**.** *Assume the assumptions of Section 6.6.1 are satisfied. Recall the notation of Table 6.3. In addition, suppose that function f is convex. Consider the bounded set*

$$S := \{x + u\colon f(x) \le f(x_0) \text{ and } u \in \overline{B}(x)\}.$$

Let L' be a Lipschitz constant of f on S. Define the constants

$$a := \min\left\{\frac{\gamma\delta_{\mathrm{A}}\delta_{\mathtt{NTD}}}{4}, \frac{\gamma\delta_{\mathtt{NTD}}^2}{8}\right\}; \qquad \text{and} \qquad b := \inf_{\alpha\in(0,1)} \frac{\left(64L'\sqrt{\frac{2}{\alpha}}\right)^{\frac{2}{(1-\alpha)}}}{\left(\frac{a}{\mathrm{diam}(S)}\right)^{\frac{2\alpha}{(1-\alpha)}}}.$$

Finally, define

$$K_1 := \left\lceil \frac{4\mathrm{diam}^2(S)}{a^2} \min\left\{16^2(f(x_0) - \inf f)^2, \frac{b}{4}, 2048L'^2\log\left(\frac{2}{p}\right), 128(L')^2\right\} + \frac{(4L')^2}{a^2}\right\rceil.$$

Then, for every failure probability $p \in (0, 1)$, we have

$$P(E_{k_0,q,C}) \ge 1 - p \qquad \text{for all } k_0 \ge \max\left\{K_0, C'\log\left(\frac{2C'}{p}\right), 2K_1 - 1\right\}.$$

Proof. Theorem 6.6.3 shows that

$$P(E_{k_0,q,C} \mid x_{k_0} \in B_{\delta_{\mathtt{NTD}}}(\bar{x})) \ge 1 - p/2 \qquad \text{for all } k_0 \ge \max\left\{K_0, C'\log\left(\frac{2C'}{p}\right)\right\}. \quad (6.6.14)$$

We claim that

$$P(x_{k_0} \in B_{\delta_{\mathtt{NTD}}}(\bar{x})) \ge 1 - p/2 \qquad \text{for all } k_0 \ge 2K_1 - 1. \quad (6.6.15)$$

Note that this yields the proof since in that case

$$P(E_{k_0,q,C}) \ge P(E_{k_0,q,C} \mid x_{k_0} \in U)P(x_{k_0} \in B_{\delta_{\mathtt{NTD}}}(\bar{x})) \ge 1 + p^2/4 - p \ge 1 - p,$$

for all $k_0 \ge \max\left\{K_0, C'\log\left(\frac{2C'}{p}\right), 2K_1 - 1\right\}$.

Observe that (6.6.15) will follow if

$$P(f(x_{k_0}) - f(\bar{x}) \le a) \ge 1 - p/2 \qquad \text{for all } k_0 \ge 2K_1 - 1. \tag{6.6.16}$$

Indeed, by Lemma 6.6.4, we have.

$$\{x \in \mathbb{R}^d : f(x) - f(\bar{x}) \le a\} \subseteq \overline{B}_{\delta_{\mathtt{NTD}}/2}(\bar{x}) \subseteq B_{\delta_{\mathtt{NTD}}}(\bar{x}).$$

To prove (6.6.16), we apply Theorem 6.2.4. To that end, note that $\{x \in \mathbb{R}^d : f(x) \le f(x_0)\}$ and the widened sublevel set S are indeed bounded, due to Lemma 6.6.4. Therefore D and the Lipschitz constant L' of f on S are finite. Now observe $G := \min_{K_1 \le k \le 2K_1 - 1}\{G_k\} \ge K_1$ since $G_k \ge k + 1$ for all k. Thus, there exists $i \le G$ such that

$$(1/2)K_1^{-1/2} \le \sigma_i \le K_1^{-1/2}.$$

Therefore, applying Theorem 6.2.4 (in particular (6.2.2)) with this σ_i, we have

$$f(x_{2K_1-1}) - f(\bar{x}) \le D \max\left\{\frac{16(f(x_{K_1}) - \inf f)}{K_1^{1/2}}, \frac{16L'\sqrt{2\log(2K_1^2/p)}}{K_1^{1/2}}, \frac{\sqrt{128}L'}{K_1^{1/2}}\right\} + \frac{2L'}{K_1^{1/2}} \tag{6.6.17}$$

with probability at least $1 - p/2$. Thus, to complete the proof, we show that the left-hand side of (6.6.17) is smaller than a. Indeed, it is straightforward to check that

$$\max\left\{\frac{2L'}{K_1^{1/2}}, \frac{16D(f(x_{K_1}) - \inf f)}{K_1^{1/2}}, \frac{\sqrt{128}DL'}{K_1^{1/2}}\right\} \le \frac{a}{2}.$$

Thus, the proof will follow if

$$\frac{16DL'\sqrt{2\log(2K_1^2/p)}}{K_1^{1/2}} \le \frac{a}{2}. \tag{6.6.18}$$

We perform this calculation in Appendix 7.3.7. Thus, the proof is complete. □

6.6.3.3 Local oracle complexity.

Thus, we have established a local nearly linear convergence rate for `NTDescent`. To understand the overall complexity of the method, we must derive an upper bound on the contraction factor q. The following lemma, which is proved in Appendix 7.3.8, provides one that depends on a worst-case condition number of f.

Lemma 6.6.6. *Suppose without loss of generality that $\delta_{\mathrm{A}} \leq 1$. Define the condition number*

$$\kappa = \frac{\max\{L, \beta, C_{(a)}\}}{\min\{\gamma, \mu\}}.$$

Then there exists a universal constant $\eta > 0$ independent of f such that

$$q \leq 1 - \frac{\eta}{\kappa^8(1 + C_{\mathcal{M}})^2}.$$

where q is defined as in Table 6.3.

With this upper bound on q, it is straightforward to derive a local complexity estimate for `NTDescent`: the method locally produces a point $\hat{x}$ satisfying $f(\hat{x}) - f(\bar{x}) \leq \varepsilon$ with at most

$$O\left((\kappa^8(1 + C_{\mathcal{M}})^2 \log(1/\varepsilon))^3\right),$$

first-order oracle evaluations. This bound may be pessimistic since we did not attempt to optimize the constants C_i or a_i. We leave the improvement of this complexity as an intriguing open question.

Before moving to a brief numerical illustration, we explain how Theorem 6.1.1 from the introduction follows from the above results.

Remark 2 (Establishing Theorem 6.1.1)**.** Theorem 6.1.1 from the introduction immediately follows from Theorems 6.6.3 and 6.6.5. Indeed, first the event $E_{k_0,q,C}$ from Theorems 6.6.3 and 6.6.5 contains the corresponding event $E_{k_0,q,C}$ from Theorem 6.1.1 for particular q and C, which depend solely on f. Second, from the statement of theorems, we see that the neighborhood of local nearly linear convergence, $B_{\delta_{\mathtt{NTD}}}(\bar{x})$, depends solely on f.

6.7 Numerical illustration

In this section, we briefly illustrate the numerical performance of `NTDescent` on two nonsmooth objective functions, borrowed from [123–126]. In both experiments, we compare `NTDescent` to the subgradient method with the popular Polyak stepsize (`PolyakSGM`) [97], which iterates

$$x_{k+1} = x_k - \frac{f(x_k) - \inf f}{\|w_k\|^2} w_k \qquad \text{for some } w_k \in \partial_c f(x_k).$$

In the first example, $\inf f$ is known, in the second, we estimate $\inf f$ from multiple runs of `NTDescent`. We compare against the subgradient method because it is a simple first-order method with strong convergence guarantees in convex [97] and nonconvex settings [115]. Importantly, `PolyakSGM` accesses the objective solely through function and subgradient evaluations. Thus, we compare the accuracy achieved by `PolyakSGM` and `NTDescent` after a fixed number of oracle calls, i.e., evaluations of $\partial_c f$.

Let us comment on the implementation of `NTDescent`. First, in all experiments, unless otherwise noted, we do not tune parameters of `NTDescent`. Instead, we simply choose scaling constant $c_0 = 10^{-6}$ and loop size parameters

$$T_k = k + 1 \qquad \text{and} \qquad G_k = \min\{k + 1, \lceil \log_2(10^{-16}) \rceil\} \quad \text{for all } k \geq 0.$$

Second, we attempt to save first-order oracle calls by breaking the loop on Lines 2 through 6 of Algorithm 3 whenever we find that $\sigma_i > \|v_{i+1}\|/s$. Since σ_i is increasing in i and $\|v_{i+1}\|$ is nonincreasing in i, this does not affect the iterates x_k of `NTDescent`;

see Lemma 6.2.1. Finally, in all problems, we initialize `NTDescent` and `PolyakSGM` at a random vector az where z is sampled from the uniform distribution on the unit sphere. For all problems, we use $a = 1$ unless otherwise noted. Note that in the problems of Section 6.7.1 and 6.7.2, the solution is known, while in the problem of Section 6.7.3, the solution is unknown.

The purpose of this section is not to argue that `NTDescent` is a substitute for standard subgradient methods in most problems. Instead, we only wish to point out some scenarios where standard first-order methods are known to perform poorly, yet `NTDescent` asymptotically accelerates. We are also not arguing that `NTDescent` has fast global rates: indeed, we previously mentioned that the `NTDescent`'s global rate is $O(\epsilon^{-6})$ which is much worse than `PolyakSGM`'s $O(\epsilon^{-2})$ rate for general convex problems. In practice, one could devise schemes that couple `NTDescent` with `PolyakSGM`, eventually switching to `NTDescent` when it begins to outperform `PolyakSGM`. While we leave a more thorough numerical study to future work, the reader may download and run our PyTorch [127] implementation of `NTDescent` at the following url: https://github.com/COR-OPT/ntd.py

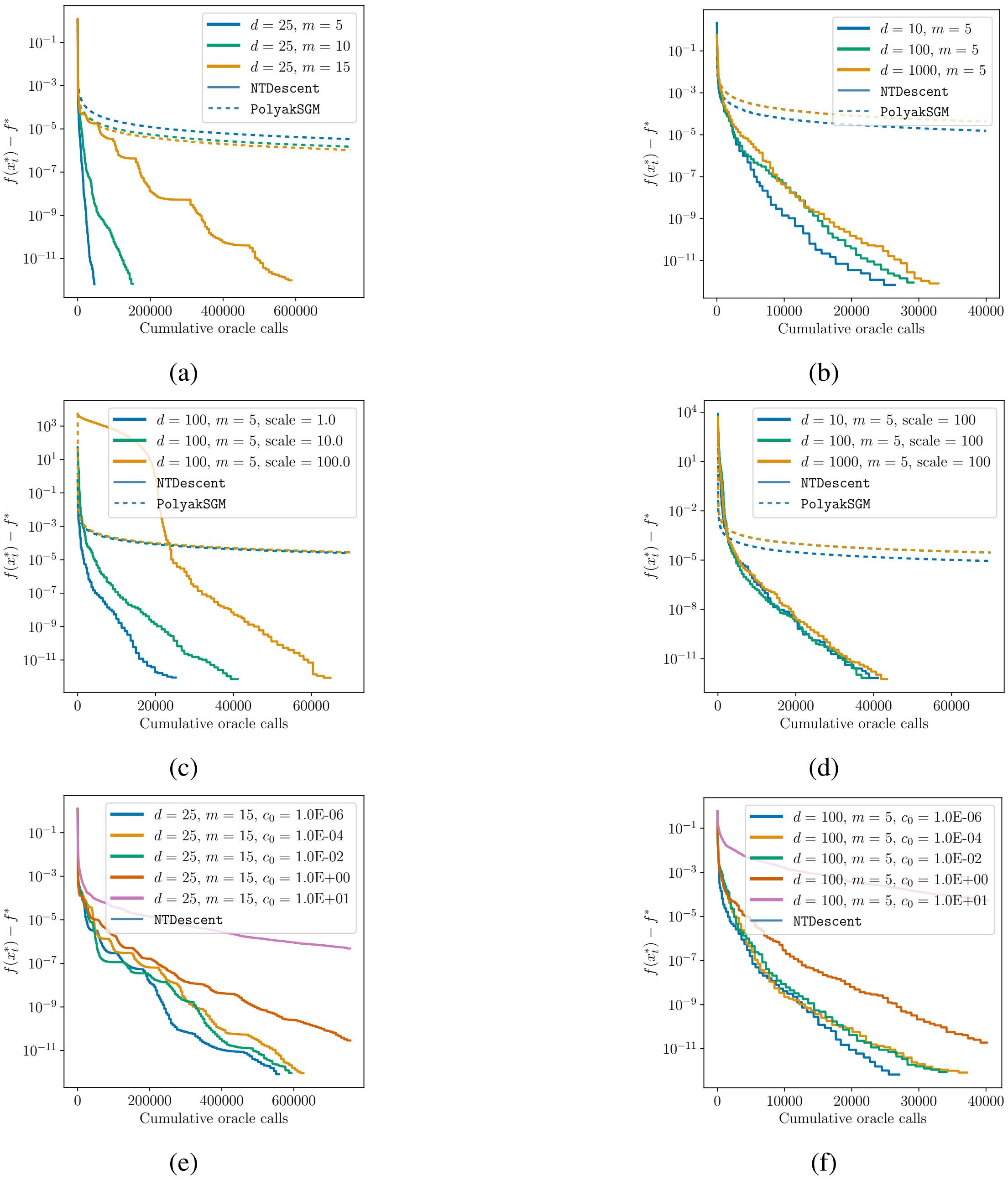
We now turn to the examples.

6.7.1 A max-of-smooth function

In this example, f takes the following form

$$f(x) = \max_{i=1,\ldots,m} \left\{ g_i^\top x + \frac{1}{2} x^T H_i x \right\}, \tag{6.7.1}$$

where we generate a random vector $\lambda \in \mathbb{R}^m$ in $\{\lambda > 0 \colon \sum_{i=1}^m \lambda_i = 0\}$, a random positive semi-definite matrix H_i, and a random vector g_i satisfying that $\sum_{i=1}^m \lambda_i g_i = 0$. In this case, one can show that with probability 1, f satisfies Assumption Q at its unique minimizer 0.



(a)

(b)

(c)

(d)

(e)

(f)

Figure 6.6: Comparison of `NTDescent` with `PolyakSGM` on (6.7.1). For both algorithms, the value $f(x_t^*)$ denotes the best function seen after t oracle evaluations. See text for description.

In Figure 6.6 we plot the performance of `NTDescent` and `PolyakSGM` for multiple pairs of (d, m), varying initialization scale, a slight modification of `NTDescent` that allows longer steps, and varying scales c_0. We begin with Figure 6.6a and Figure 6.6b. Figure 6.6a shows that the performance of `NTDescent` depends on m. On the other hand, Figure 6.6b shows `NTDescent` performance is independent of d, as expected. Both plots show that `NTDescent` asymptotically outperforms `PolyakSGM`. Turning to initialization, Figure 6.6c shows the result of initializing `NTDescent` at a random vector az, where z is uniformly drawn from the sphere and a is a scale parameter satisfying $a \in \{1, 10, 100\}$. Clearly, `NTDescent` is affected by the initialization scale but surpasses `PolyakSGM` after 30000 oracle calls. While we expect `NTDescent` to converge slowly when far from minimizers, we introduce a simple strategy to mitigate this behavior.

Adaptive grid strategy. Briefly, suppose we run `linesearch` the full G steps without exiting (via the violation of the trust region constraint). Then we simply continue the `linesearch` loop trying $\sigma_{-1} = 10\sigma_0$, $\sigma_{-2} = 10\sigma_{-1}$, and so on, until we violate the trust region constraint or σ_{-i} exceeds a predefined threshold.

Figure 6.6d shows the result of this strategy with a predefined threshold ∞, showing that it compensates for poor initialization quality. Finally in Figure 6.6e and 6.6f, we show the effect of changing the c_0 input to `NTDescent`. It appears `NTDescent` is relatively insensitive to c_0, and smaller choices generally result in better performance. This motivates our default choice $c_0 = 10^{-6}$ in the remainder of the experiments.

Before turning to our second experiment, we briefly mention two alternative methods – Prox-linear [66, 128–131] and Survey Descent [104] – which could be applied to this problem. In order to explain these algorithms, let us write $f = \max_{i=1,\ldots,m}\{f_i\}$, where the f_i are the quadratic function from (6.7.1).

Prox-linear method. Given a point $x \in \mathbb{R}^d$, the Prox-linear update x_+ solves

$$x_+ = \operatorname*{argmin}_{y \in \mathbb{R}^d} \max_{i=1,\ldots,m} \{f_i(x) + \langle \nabla f_i(x), y - x \rangle\} + \frac{\rho}{2}\|y - x\|^2.$$

One may show that x_+ geometrically improves on x; see [132]. However, in contrast to `NTDescent`, the prox-linear method requires that the components f_i are known. This is stronger than the first-order oracle model considered in this chapter. Thus, we do not compare `NTDescent` with prox-linear.

Survey Descent. The Survey Descent method is a multi-point generalization of gradient descent designed for max-of-smooth functions. Rather than maintaining a single iterate sequence, the Survey Descent maintains a *survey* S of points, meaning a collection of points $\{s_i\}_{i=1}^m$ at which f is differentiable. A single iteration of the Survey Descent method then aims to produce a new survey $S^+ = \{s_i^+\}_{i=1}^m$ satisfying

$$\begin{aligned}
s_i^+ := \operatorname*{argmin}_{x \in \mathbb{R}^d} & \left\| x - \left(s_i - \frac{1}{L}\nabla f(s_i) \right) \right\|^2 \\
& \text{subject to: } f(s_j) + \langle \nabla f(s_j), x - s_j \rangle + \frac{L}{2}\|x - s_j\|^2 \leq f(s_i) + \langle \nabla f(s_i), x - s_i \rangle \quad \forall j \neq i.
\end{aligned}$$

Here, L is an upper bound on the Lipschitz constant of ∇f_i for all $i = 1, \ldots, m$. In [104], Han and Lewis study linear convergence of Survey Descent on max-of-smooth functions under the conditions of Corollary 6.3.4. Given a survey S, they show that the updated survey S^+ geometrically improves on S (in an appropriate sense) whenever the following conditions are satisfied: (i) all elements of the survey S are near $\bar{x}$; (i) the survey S is *valid*, meaning there exists a permutation a on $[m]$ such that

$$f_{a(i)}(s_i) = f(s_i) \qquad \text{and} \qquad \partial_c f(s_i) = \{\nabla f_{a(i)}(s_i)\} \qquad \text{for all } i = 1, \ldots, m.$$

To estimate the number of components m and find a valid initial survey S sufficiently close to $\bar{x}$, Han and Lewis suggest an empirical procedure based on running a nonsmooth variant of BFGS [124] for several iterations. After running BFGS, they suggest

(i) computing an estimate $\hat{m}$ of m from a singular value decomposition of the computed gradients and (ii) building the survey from $\hat{m}$ past iterates in such a way that the computed gradients form an affine independent set. From the numerical illustration in [104], Survey Descent performs well on several small problems. However, since the initialization procedure and implementation of Survey Descent are somewhat sophisticated, we leave a detailed comparison between `NTDescent` and Survey Descent for future work.

6.7.2 A matrix sensing problem

In this example, f takes the following form

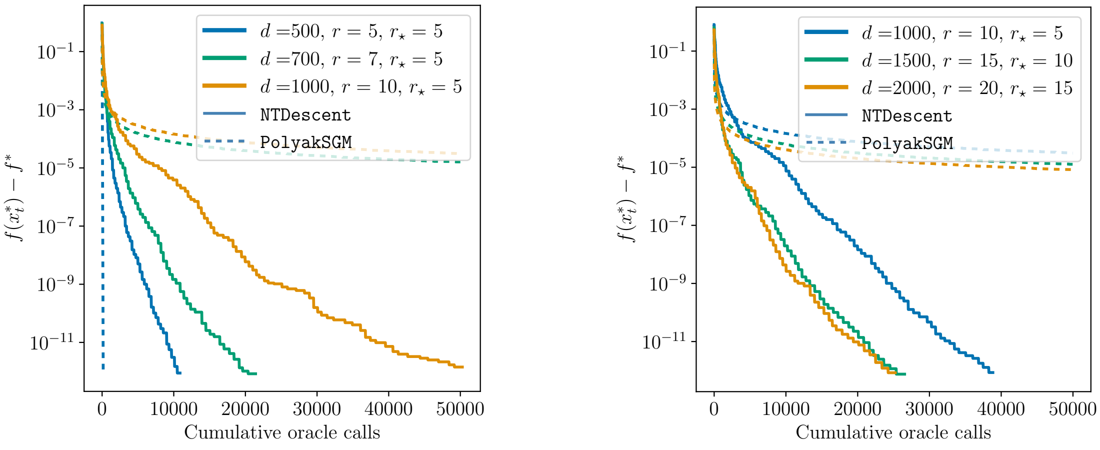
$$f(X) = \frac{1}{n}\|\mathcal{A}(XX^T) - \mathcal{A}(M_\star)\|_1,$$

where $M_\star \in \mathbb{R}^{N\times N}$ is an unknown positive semidefinite matrix of rank $r_\star$ that we wish to recover from known linear measurements $\mathcal{A}(M_\star)$; the linear operator $\mathcal{A}\colon \mathbb{R}^{N\times N} \to \mathbb{R}^n$ takes the form $Y \mapsto (a_i^T Y a_i - b_i^T Y b_i)_{i=1}^n$, for $n \in N$, where $a_i, b_i \in \mathbb{R}^d$ are random vectors sampled from a standard multivariate normal distribution; and the decision variable is a tall and skinny matrix $X \in \mathbb{R}^{N\times r}$, where in general we allow $r \neq r_\star$. This optimization problem appears in various signal processing applications and is known as *quadratic sensing* [133]. Note that this objective does not satisfy Assumption Q since the solution set is not isolated.

We consider two settings in this section: the exact setting $r = r_\star$ and the overparameterized setting $r > r_\star$. In the exact setting [134] showed that if $n = \Omega(Nr)$, the objective f is sharp, meaning $f(x) = \Omega(\mathrm{dist}(x, \operatorname{argmin} f))$ and that `PolyakSGM` converges linearly whenever the initial iterate is sufficiently close to the set of minimizers. In the overparameterized setting, we are not aware of similar guarantees since exisiting works with nonsmooth loss [135,136] all require Gaussian sensing matrices. In practice,

$r_\star$ is unknown, so the overparameterized setting will likely be encountered.

In Figure 6.7, we plot the performance of `NTDescent` and `PolyakSGM` in two experiments. In Figure 6.7a we use base dimensions $N = 100$, optimal rank $r_\star = 5$, and varying overparameterization $r \in \{r_\star, r_\star + 2, r_\star + 5\}$. In Figure 6.7b we use base dimensions $N = 100$, varying optimal rank $r_\star \in \{5, 10, 15\}$, and fixed overparameterization $r = r_\star + 5$. In both experiments, we fix $n = 4Nr_\star$. Note that the dimension of the decision variable X varies across each run since $d = Nr$ and r vary. As can be seen from the plot, `PolyakSGM` outperforms `NTDescent` in the exact setting. This is expected since f is a sharp function on which `PolyakSGM` is known to perform well. On the other hand, when $r > r_\star$, we find that both methods slow down. However, `NTDescent` converges nearly linearly, while `PolyakSGM` converges sublinearly. To the best of our knowledge, white-box algorithms based on preconditioning idea [137, 138] have a local linear convergence rate. However, all black-box algorithms in overparametrized settings rely on small initialization and early stopping to achieve linear rate [136, 139–144].



(a)

(b)

Figure 6.7: Comparison of `NTDescent` with `PolyakSGM` on (6.7.1). In both plots, the base dimension is $N = 100$. Left: fixed optimal rank $r_\star = 5$ and varying overparameterization $r \in \{5, 7, 10\}$; Right: varying $r_\star \in \{5, 10, 15\}$, fixed overparameterization $r = r_\star + 5$. For both algorithms, the value $f(x_t^*)$ denotes the best function seen after t oracle evaluations.

6.7.3 An eigenvalue product function

In this example, we aim to optimize a function $\tilde{f}$ that takes the following form

$$\tilde{f}(X) = \log E_K(A \odot X),$$

where A is a fixed positive semi-definite data matrix, $E_K(Y)$ denotes the product of K largest eigenvalues of a symmetric matrix $Y \in \mathbb{S}^N$, and $\odot$ denotes the Hadamard (entrywise) matrix product, subject to the constraint that X is positive semi-definite and its diagonal entries are 1. This example is a nonconvex relaxation of an entropy minimization problem arising in an environmental application [125, 126]. In our experiments, we choose A as in [126]: A is the leading $N \times N$ submatrix of a 63×63 covariance matrix, scaled so that the largest entry is 1. As suggested by [125], we reformulate this problem as an unconstrained optimization problem using a Burer-Monteiro type factorization

$$\min_{V \in \mathbb{R}^{N \times N}} f(V) = \tilde{f}(c(V)c(V)^\top), \tag{6.7.2}$$

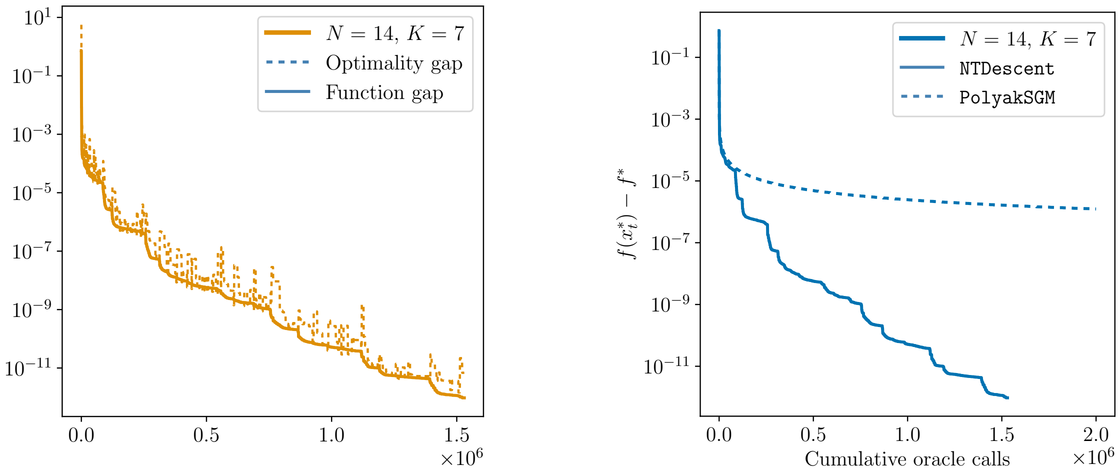
where $c \colon \mathbb{R}^{N \times N} \to \mathbb{S}^N$ satisfies $c(V) = \mathrm{Diag}([\mathrm{diag}(VV^\top)]^{-1/2})V$ for all $V \in \mathbb{R}^{N \times N}$. Here, the mapping $\mathrm{diag}(\cdot)$ takes a matrix an $N \times N$ matrix A to the N dimensional vector with ith entry A_{ii}. On the other hand, the mapping $\mathrm{Diag}(\cdot)$ takes an N dimensional vector v to the $N \times N$ diagonal matrix with ith diagonal entry v_i. A formula for the subgradient of f may be found [125]. We do not attempt to verify that f satisfies the full Assumption Q. Instead, we point out that under a "transversality condition," function f admits an active manifold at local minimizers [124].

Turning to the experiment, we consider the case where $N = 14$ and $K = 7$. In this example, the optimal function value $\inf f$ is not known. Thus, we run `NTDescent` from four random initial starting points. We terminate each run of `NTDescent` when a certain "optimality gap" R_k satisfies $R_k \leq 10^{-12}$. We denote the minimal function value achieved across all four runs by f^*. Let us now define and motivate the optimality gap.

For iteration k in Algorithm 4, define

$$R_k := \min\left\{\max\{\sigma_i^{(k)}, \|v_{i+1}^{(k)}\|^2\} \colon \sigma_i^{(k)} \le \|v_{i+1}^{(k)}\|\right\},$$

where $\sigma_i^{(k)}$ and $v_{i+1}^{(k)}$ are computed in Lines 2 through 6 of Algorithm 3 at iteration k. Provided that x_k is sufficiently close to a point $\bar{x}$ at which function f satisfies Assumption Q, it is possible to show that R_k satisfies $f(x_k) - f(\bar{x}) \lesssim R_k$. This is illustrated in Figure 6.8a: there, the optimality gap closely tracks the estimated function gap, when approximating by $\inf f$ by f^*. In Figure 6.8b, we compare the performance of `NTDescent` on the three runs which did not achieve function value f^* before termination. In all three cases, we see similar performance. Next, for each run of `NTDescent`, we also run `PolyakSGM` from the same initial starting point, estimating $\inf f$ by f^*. We see that `NTDescent` outperforms `PolyakSGM`.



(a)

(b)

Figure 6.8: Numerical performance on (6.7.2). Left: the close relationship between the "optimality gap" and function gap; Right: comparison of `PolyakSGM` and `NTDescent` from three initial starting points. For both algorithms, the value $f(x_t^*)$ denotes the best function seen after t oracle evaluations. See text for details.

CHAPTER 7

APPENDICES

7.1 Proofs for saddle avoidance

7.1.1 Proof of Proposition 3.3.4

Since X is a C^3 manifold, the projection P_Y is C^2-smooth. Therefore, there exist constants $\epsilon, L > 0$ satisfying

$$\|P_Y(y+h) - P_Y(y) - \nabla P_Y(y)h\| \leq L\|h\|^2 \tag{7.1.1}$$

for all $y \in B_\epsilon(\bar{x})$ and $h \in \epsilon\mathbb{B}$. Fix now two points $x \in X$ and $y \in Y$ and a unit vector $v \in N_X(x)$. Clearly, we may suppose $v \notin N_Y(y)$, since otherwise the claim is trivially true. Define the normalized vector $w := -\frac{P_{T_Y(y)}(v)}{\|P_{T_Y(y)}(v)\|}$. Noting the equality $\nabla P_Y(y) = P_{T_Y(y)}$ and appealing to (7.1.1), we deduce the estimate

$$\|P_Y(y-\alpha w) - (y - \alpha w)\| \leq L\|\alpha w\|^2 = L\alpha^2,$$

for all $y \in B_\epsilon(\bar{x})$ and $\alpha \in (0, \epsilon)$. Shrinking $\epsilon > 0$, prox-regularity yields the estimate

$$\langle v, P_Y(y-\alpha w) - x\rangle \leq \frac{\rho}{2}\|x - P_Y(y-\alpha w)\|^2,$$

for some constant $\rho > 0$. Therefore, we conclude

$$\begin{aligned}\alpha\|P_{T_Y(y)}v\| = -\alpha\langle v, w\rangle &= \langle v, x-y\rangle + \langle v, P_Y(y-\alpha w) - x\rangle + \langle v, (y-\alpha w) - P_Y(y-\alpha w)\rangle \\ &\leq \|x-y\| + \frac{\rho}{2}\|x - P_Y(x-\alpha w)\|^2 + L\alpha^2.\end{aligned}$$

Note that the middle term is small:

$$\|P_Y(y-\alpha w)-x\|^2 \leq 2\|P_Y(y-\alpha w)-(y-\alpha w)\|^2 + 2\|y-\alpha w - x\|^2 \leq 2L^2\alpha^4 + 4\|y-x\|^2 + 4\alpha^2.$$

Thus, we have

$$\alpha\|P_{T_{Y(y)}}v\| \leq \|x-y\| + \rho L^2\alpha^4 + 2\rho\|x-y\|^2 + 2\rho\alpha^2 + L\alpha^2.$$

Dividing both sides by α and setting $\alpha = \sqrt{\|x-y\|}$ completes the proof of (3.3.1).

7.1.2 Proof of Proposition 4.2.3: the projected gradient method

Choose $\epsilon > 0$ small enough that the following hold for all $x \in B_\epsilon(\bar{x}) \cap \mathcal{X}$. First (4.2.9) holds. Second we require that for some $L > 0$, we have

$$\|P_{T_{\mathcal{M}}(P_{\mathcal{M}}(x))}(s_g(x) - \nabla_{\mathcal{M}}g(P_{\mathcal{M}}(x))\| \leq L\text{dist}(x, \mathcal{M}), \tag{7.1.2}$$

$$\|P_{T_{\mathcal{M}}(z)}(u)\| \leq L\|x-z\|, \tag{7.1.3}$$

for all $u \in N_{\mathcal{X}}(x)$ of unit norm and all $z \in B_\epsilon(\bar{x}) \cap \mathcal{M}$, a consequence of (C1). Third, given an arbitrary $\delta \in (0,1)$ we may choose $\epsilon > 0$ so small so that

$$\langle z, x - x' \rangle \geq -o(\|x - x'\|) \tag{7.1.4}$$

for all $z \in N_{\mathcal{X}}(x)$ of unit norm, and $x' \in \mathcal{M} \cap B_\epsilon(\bar{x})$—a consequence of (C3). We will fix $x \in B_{\epsilon/2}(\bar{x}) \cap \mathcal{X}$ and arbitrary $\alpha > 0$ and $v \in \mathbb{R}^d$, and choose an arbitrary $y \in P_{\mathcal{M}}(x)$. Define

$$w = G_\alpha(x, v) - v - s_g(x) \qquad \text{and} \qquad x_+ = s_{\mathcal{X}}(x - \alpha(s_g(x) + v)).$$

Note the inclusion $w \in N_{\mathcal{X}}(x^+)$. Next, to verify (A1), we compute

$$\begin{aligned}\alpha\|G_\alpha(x, v)\| &= \|x - s_{\mathcal{X}}(x - \alpha(s_g(x) + v))\| \\ &\leq \text{dist}_{\mathcal{X}}(x - \alpha(s_g(x) + v)) + \alpha\|s_g(x) + v\| \\ &\leq 2\alpha\|s_g(x) + v\| = O(\alpha(1 + \|v\|)).\end{aligned}$$

Thus there exists a constant $C > 0$ satisfying

$$\max\{\|w\|, \|G_\alpha(x, v)\|\} \leq C(1 + \|v\|) \qquad \text{and} \qquad \|x_+ - x\| \leq C(1 + \|v\|)\alpha.$$

We will use these estimates often in the proof. Finally, we let C be a constant independent of x, α and v, which changes from line to line.

<u>Assumption (A2):</u> Suppose first that $x_+ \in B_\epsilon(\bar{x})$. Using (7.1.3), we compute

$$\begin{aligned} \|P_{T_{\mathcal{M}}(P_{\mathcal{M}}(x))} w\| &\leq L\|w\|\|x_+ - P_{\mathcal{M}}(x)\| \\ &\leq L\|w\|(\|x_+ - x\| + \text{dist}(x, \mathcal{M})) \\ &\leq C(1 + \|v\|)^2\alpha + C(1 + \|v\|)\text{dist}(x, \mathcal{M}). \end{aligned} \tag{7.1.5}$$

On the other hand, if $x_+ \notin B_\epsilon(\bar{x})$, then we compute

$$\|P_{T_{\mathcal{M}}(P_{\mathcal{M}}(x))} w\| \leq \|w\| \leq \frac{2}{\epsilon}\|w\|\|x_+ - x\| \leq C(1 + \|v\|)^2\alpha. \tag{7.1.6}$$

In either case, Assumption (A2) now follows since from (7.1.2) we have

$$\|P_{T_{\mathcal{M}}(P_{\mathcal{M}}(x))}(s_g(x) - \nabla_{\mathcal{M}} g(P_{\mathcal{M}}(x))\| \leq C\text{dist}(x, \mathcal{M}),$$

as we had to show.

<u>Assumption (A3):</u> We write the decomposition

$$\langle G_\alpha(x, v) - v, x - y\rangle = \underbrace{\langle s_g(x), x - y\rangle}_{R_1} + \underbrace{\langle w, x_+ - y\rangle}_{R_2} + \underbrace{\langle w, x - x_+\rangle}_{R_3}. \tag{7.1.7}$$

The aiming condition (C2) ensures

$$R_1 \geq \mu \cdot \text{dist}(x, \mathcal{M}). \tag{7.1.8}$$

We next look at two cases. Suppose first $x_+ \in B_\epsilon(\bar{x})$ and therefore $\|x_+ - x\| \geq \epsilon/2$. Using the inclusion $w \in N_{\mathcal{X}}(x_+)$ and Assumption (C3), we compute

$$R_2 \geq -\|w\| \cdot o(\|x_+ - y\|) \geq -\|w\| \cdot (o(\|y - x\|) + \|x - x_+\|)$$

$$\geq -C(1+\|v\|)^2(o(\operatorname{dist}(x,\mathcal{M}))+\alpha). \tag{7.1.9}$$

Next, the Cauchy–Schwarz inequality implies

$$|R_3| = \|w\|\|x - x_+\| \leq C(\alpha(1+\|v\|)^2). \tag{7.1.10}$$

Combining (7.1.7)-(7.1.10) yields the claimed bound (A3).

Suppose now on the contrary that $x_+ \notin B_\epsilon(\bar{x})$ and therefore $\|x-y\| \leq \|x-x_+\|$. We thus deduce $R_2 + R_3 = \langle w, x-y\rangle \geq -\|w\|\|x-y\| \geq -C\alpha(1+\|v\|)^2$ holds. Combining this estimate with (7.1.7) and (7.1.8) verifies the claim (A3).

7.1.3 Proof of Proposition 4.2.5: the proximal gradient method

Let $\epsilon \in (0,1)$ be small enough such that the following hold for all $x \in B_\epsilon(\bar{x}) \cap \operatorname{dom} f$. First, (4.2.13) holds and therefore:

$$\langle \nabla g(x) + v, x - P_{\mathcal{M}}(x)\rangle \geq \mu \cdot \operatorname{dist}(x,\mathcal{M}) - (1+\|v\|)o(\operatorname{dist}(x,\mathcal{M})), \tag{7.1.11}$$

for all $v \in \hat{\partial} h(x)$. Second we require that for some $L > 0$, we have

$$\|P_{T_{\mathcal{M}}(P_{\mathcal{M}}(x))}(u - \nabla_{\mathcal{M}} h(P_{\mathcal{M}}(x))\| \leq L\sqrt{1+\|u\|^2} \cdot \operatorname{dist}(x,\mathcal{M}) \tag{7.1.12}$$

for all $u \in \partial h(x)$, a consequence of strong (a) regularity. Third, we assume that $\nabla_{\mathcal{M}} f$ is L-Lipschitz on $B_\epsilon(\bar{x}) \cap \mathcal{M}$. Fourth, we assume that $\nabla g(\cdot)$ is L-Lipschitz. Shrinking ϵ we may moreover assume $\epsilon \leq \frac{\mu}{4L}$. Finally, we may also assume that the assignments $P_{\mathcal{M}}$ is L-Lipschitz on $B_\epsilon(\bar{x})$ and that the map $x \mapsto P_{T_{\mathcal{M}}(P_{\mathcal{M}}(x))}(\cdot)$ is L-Lipschitz on $B_\epsilon(\bar{x})$ with respect to the operator norm.

Fix $x \in B_{\epsilon/2}(\bar{x}) \cap \operatorname{dom} f$ and $v \in \mathbb{R}^d$ and set $y := P_{\mathcal{M}}(x)$. We define the vectors

$$w = G_\alpha(x,v) - \nabla g(x) - v \qquad \text{and} \qquad x_+ = s_\alpha(x - \alpha(\nabla g(x) + v)).$$

Claim: We have $w \in \hat{\partial} h(x_+)$ and there exists a constant C independent of x, v, α, such that the following bounds hold:

$$\max\{\|G_\alpha(x, v)\|, \|w\|\} \leq C(1 + \|v\|); \qquad \text{and} \qquad \|x_+ - x\| \leq C(1 + \|v\|)\alpha.$$

Proof. Beginning with the inclusion, first-order optimality conditions imply that w is a Fréchet subgradient:

$$w = \frac{x - \alpha(\nabla g(x) + v) - x_+}{\alpha} \in \hat{\partial} h(x_+),$$

as desired. First, we bound $\|x_+ - x\|$: Let $v = \nabla g(x) + v$ and observe from the very definition of x^+ that there exists $C > 0$ such that

$$\frac{1}{2\alpha}\|x_+ - x\|^2 \leq h(x) - h(x_+) - \langle v, x_+ - x\rangle \leq C\|x_+ - x\| + \|v\|\|x_+ - x\|.$$

Consequently, we have $\|x_+ - x\| \leq (2C + 2\|v\|)\alpha \leq 2(2C + \|v\|)\alpha$, as desired. Second, the bound on $G_\alpha(x, v)$ follows trivially from the computation

$$\|G_\alpha(x, v)\| = \|x_+ - x\|/\alpha \leq 2(2C + \|v\|).$$

Finally, we bound $\|w\|$ using the estimate

$$\|w\| = \|x - x_+\|/\alpha + \|\nabla g(x) + v\| \leq 4(2C + \|v\|),$$

as desired. □

We will use the estimates in the claim often in the proof. Finally, we let C be a constant independent of x, α and v, which changes from line to line.

Assumption (A2): First suppose $x_+ \in B_\epsilon(\bar{x})$. Using the triangle inequality, we write

$$\|P_{T_{\mathcal{M}}(P_{\mathcal{M}}(x))}(G_\alpha(x, v) - \nabla_{\mathcal{M}} f(P_{\mathcal{M}}(x)) - v)\|$$
$$= \|P_{T_{\mathcal{M}}(P_{\mathcal{M}}(x))}(w + \nabla g(x) - \nabla g(P_{\mathcal{M}}(x)) - \nabla_{\mathcal{M}} h(P_{\mathcal{M}}(x)))\|$$

$$\leq \underbrace{\|P_{T_{\mathcal{M}}(P_{\mathcal{M}}(x))}(w - \nabla_{\mathcal{M}} h(P_{\mathcal{M}}(x_+)))\|}_{R_1} + \underbrace{\|\nabla g(x) - \nabla g(P_{\mathcal{M}}(x))\|}_{R_2} + \underbrace{\|\nabla_{\mathcal{M}} h(P_{\mathcal{M}}(x))) - \nabla_{\mathcal{M}} h(P_{\mathcal{M}}(x_+)))\|}_{R_3}.$$

Taking into account that the assignment $x \mapsto P_{T_{\mathcal{M}}(P_{\mathcal{M}}(x))}(\cdot)$ is Lipschitz with respect to the operator norm, the estimate (7.1.12) implies

$$\begin{aligned}
R_1 &\leq L\sqrt{1 + \|w\|^2} \cdot \mathrm{dist}(x_+, \mathcal{M}) + L\|x - x_+\|\|w - \nabla_{\mathcal{M}} h(P_{\mathcal{M}}(x_+))\| \\
&\leq C(1 + \|v\|)\mathrm{dist}(x_+, \mathcal{M}) + L(1 + \|v\|)^2\alpha \\
&\leq C(1 + \|v\|)(\mathrm{dist}(x, \mathcal{M}) + C\|x - x_+\|) + L(1 + \|v\|)^2\alpha \\
&\leq C(1 + \|v\|)\mathrm{dist}(x, \mathcal{M}) + C(1 + \|v\|)^2\alpha.
\end{aligned}$$

Moreover, clearly we have $R_2 \leq C\mathrm{dist}(x, \mathcal{M})$ and $R_3 \leq C\|x - x_+\| \leq (1 + \|v\|)\alpha$. Condition (A2) follows immediately.

Now suppose that $x_+ \notin B_\epsilon(\bar{x})$, and therefore $\|x_+ - x\| \geq \epsilon/2$. Then, we may write

$$\begin{aligned}
\|P_{T_{\mathcal{M}}(P_{\mathcal{M}}(x))}(G_\alpha(x, v) - v - \nabla f_{\mathcal{M}}(P_{\mathcal{M}}(x)))\| &\leq \|G_\alpha(x, v)\| + \|v\| + \|\nabla f_{\mathcal{M}}(P_{\mathcal{M}}(x))\| \\
&\leq \frac{2}{\epsilon}(\|G_\alpha(x, v)\| + \|v\| + \|\nabla f_{\mathcal{M}}(P_{\mathcal{M}}(x))\|)\|x - x_+\| \\
&\leq C(1 + \|v\|)^2\alpha,
\end{aligned}$$

as desired.

Assumption (A3): We begin with the decomposition

$$\begin{aligned}
\langle G_\alpha(x, v) - v, x - y\rangle = &\underbrace{\langle \nabla g(x_+) + w, x_+ - P_{\mathcal{M}}(x_+)\rangle}_{R_1} \\
&+ \underbrace{\langle \nabla g(x) - \nabla g(x_+), x - y\rangle}_{R_2} + \underbrace{\langle \nabla g(x_+) + w, (x - P_{\mathcal{M}}(x)) - (x_+ - P_{\mathcal{M}}(x_+))\rangle}_{R_3}.
\end{aligned}$$

We now bound the two terms on the right in the case $x_+ \in B_\epsilon(\bar{x})$. Using (3.1.10), we estimate

$$R_1 \geq \mu \cdot \mathrm{dist}(x_+, \mathcal{M}) - (1 + \|v\|)o(\mathrm{dist}(x_+, \mathcal{M}))$$

$$\geq \mu \cdot (\mathrm{dist}(x, \mathcal{M}) - \|x - x_+\|) - (1 + \|v\|)(o(\mathrm{dist}(x, \mathcal{M})) + \|x - x_+\|)$$
$$\geq \mu \cdot \mathrm{dist}(x, \mathcal{M}) - (1 + \|v\|)^2(o(\mathrm{dist}(x, \mathcal{M})) + C\alpha).$$

Next, we compute

$$|R_2| \leq \|\nabla g(x) - \nabla g(x_+)\| \cdot \mathrm{dist}(x, \mathcal{M}) \leq 2L\epsilon \cdot \mathrm{dist}(x, \mathcal{M}) \leq \frac{\mu}{2}\mathrm{dist}(x, \mathcal{M}).$$

Next using Lipschitz continuity of the map $I - P_{\mathcal{M}}$ on $B_\epsilon(\bar{x})$, we deduce

$$|R_3| \leq (1 + L)\|\nabla g(x_+) + w\| \cdot \|x - x_+\| \leq C(1 + \|v\|)^2\alpha.$$

The claimed proximal aiming condition follows immediately.

Let us look now at the case $x_+ \notin B_\epsilon(\bar{x})$, and therefore $\mathrm{dist}(x, \mathcal{M}) \leq \frac{\epsilon}{2} \leq \|x - x^+\|$. Then we compute

$$\begin{aligned}\langle G_\alpha(x, v) - v, x - P_{\mathcal{M}}(x)\rangle &\geq -\mathrm{dist}(x, \mathcal{M}) \cdot \|G_\alpha(x, v) - v\| \\ &= \mathrm{dist}(x, \mathcal{M}) - \mathrm{dist}(x, \mathcal{M})(1 + \|G_\alpha(x, v) - v\|) \\ &\geq \mathrm{dist}(x, \mathcal{M}) - C\|x - x_+\|(1 + C(1 + \|v\|)) \\ &\geq \mathrm{dist}(x, \mathcal{M}) - C(1 + \|v\|)^2\alpha,\end{aligned}$$

as desired. The proof is complete.

7.1.4 Proof of Corollary 4.4.3: avoiding active strict saddle via projected subgradient method

By Proposition 4.2.3 we need only show that Assumption C holds. To that end, note that Assumptions(C1) and (C3) hold by assumption. Next we prove (C2). Note that if g satisfies ($b_\leq$) along $\mathcal{M}$, then (C2) holds by Corollary 3.1.5. Next, suppose that g is weakly convex around x. In this case, since each $x \in S$ is Fréchet critical and $\mathcal{M}_x$ is an active manifold, it follows by Proposition 2.4.2 that for some $\mu > 0$, we have

$$g(y) - g(P_{\mathcal{M}_x}(y)) \geq \mu \text{dist}(y, \mathcal{M}),$$

near x. Consequently, for all $v \in \partial_c g(x)$, we have

$$\langle v, y - P_{\mathcal{M}_x}(y) \rangle \geq g(y) - g(P_{\mathcal{M}_x}(y)) - O(\|x - y\|^2) \geq (\mu/2)\text{dist}(x, \mathcal{M}),$$

for all y near x, verifying (C2).

7.1.5 Proof of Corollary 4.4.4: avoiding active strict saddle via proximal gradient method

By Proposition 4.2.5, we need only show that Assumption D holds. Note that (D1), (D2), and (D3) hold by assumption. Thus, we need only verify (D4), which is immediate from ($b_{\leq}$)-regularity and Corollary 3.1.5.

7.1.6 Proofs of Corollaries 4.4.5, 4.4.6, and 4.4.7: saddle point avoidance for generic semialgebraic problems.

We first claim that the collection of limit points for all three methods is a connected set of composite Clarke critical points. To that end, note that by [77, Theorem 6.2/Corollary 6.4], we know that for each method, on the event the sequence x_k is bounded, all limit points are composite Clarke critical. We claim that the set of limit points is in fact connected. Indeed, by [145, Lemma 5(iii)], this will follow if

$$\lim_{k \to 0} \|x_{k+1} - x_k\| = \lim_{k \to 0} \|\alpha_k G_{\alpha_k}(x_k, \nu_k)\| = 0.$$

This in turn follows from [77, Lemma A.4, A.5, and A.6], which shows that $G_{\alpha_k}(x_k, \nu_k) = w_k + \xi_k$, where w_k is bounded and $\sum_{k=1}^{\infty} \alpha_k \xi_k$ exists almost surely. Consequently, we have $\|\alpha_k G_{\alpha_k}(x_k, \nu_k)\| = \alpha_k \|w_k + \xi_k\| \to 0$ almost surely, as desired.

Next we claim that the sequence x_k converges for all three methods. Indeed, by Corollaries 4.2.2, 4.2.4, and 4.2.6, it follows that each of the set of composite Clarke critical points for all three problems is finite for generic semialgebraic problems. Therefore, since the set of limit points of x_k is connected and discrete, it follows that on the event the sequence x_k is bounded, it must converge to a composite Clarke critical point.

To wrap up the proof, suppose that x_k converges to a composite limiting critical point. Then by Corollaries 4.2.2, 4.2.4, and 4.2.6 for any of the three methods, every composite limiting critical point of f is a composite Fréchet critical point which is either a local minimizer or an active strict saddle point at which Assumption A holds along the active manifold. By Theorem 4.4.2, the sequence x_k can converge to the such active strict saddle points only with probability zero. Therefore, the limit point must be a local minimizer, as desired.

7.1.7 Sequences and Stochastic Processes

7.1.7.1 Lemmas from other works.

Lemma 7.1.1 (Robbins-Siegmund [146])**.** *Let $A_k, B_k, C_k, D_k \geq 0$ be non-negative random variables adapted to the filtration $\{\mathcal{F}_k\}$ and satisfying*

$$\mathbb{E}[A_{k+1} \mid \mathcal{F}_k] \leq (1 + B_k)A_k + C_k - D_k.$$

Then on the event $\{\sum_k B_k < \infty, \sum_k C_k < \infty\}$, there is a random variable $A_\infty < \infty$ such that $A_k \xrightarrow{a.s.} A_\infty$ and $\sum_k D_k < \infty$ almost surely.

Lemma 7.1.2 (Conditional Borel-Cantelli [147]). *Let $\{X_n : n \geq 1\}$ be a sequence of nonnegative random variables defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and $\{\mathcal{F}_n : n \geq 0\}$ be a sequence of sub-σ-algebras of $\mathcal{F}$. Let $M_n = \mathbb{E}[X_n \mid \mathcal{F}_{n-1}]$ for $n \geq 1$. If $\{\mathcal{F}_n : n \geq 0\}$ is nondecreasing, i.e., it is a filtration, then $\sum_{n=1}^{\infty} X_n < \infty$ almost surely on $\{\sum_{n=1}^{\infty} M_n < \infty\}$.*

Lemma 7.1.3 ([80, Theorem A]). *Let $\{\mathcal{F}_k\}$ be a filtration and let $\{\epsilon_k\}$ be a sequence of random variables adapted to $\{\mathcal{F}_k\}$ satisfying for all k the bound*

$$\mathbb{E}[\epsilon_{k+1}^2 \mid \mathcal{F}_k] < \infty \qquad \textit{and} \qquad \mathbb{E}[\epsilon_{k+1} \mid \mathcal{F}_k] = 0.$$

Let $\{\Phi_k\}_k$ be another sequence of random variables adapted to $\{\mathcal{F}_k\}$. Let $\{c_k\}$ be a deterministic sequence that is square summable but not summable. Suppose that the following hold almost surely on an event H:

- *We have the Marcinkiewick-Zygmund conditions:*

$$\limsup_k \ \mathbb{E}[\epsilon_{k+1}^2 \mid \mathcal{F}_k] < \infty \qquad \textit{and} \qquad \liminf_k \ \mathbb{E}[|\epsilon_{k+1}| \mid \mathcal{F}_k] > 0.$$

- *There exists sequences of random variables $\{r_k\}$ and $\{R_k\}$, adapted to $\{\mathcal{F}_k\}$ such that $\Phi_k = r_k + R_k$ and*

$$\sum_k \|r_k\|^2 < \infty \qquad \textit{and} \qquad \mathbb{E}\left[1_H \sum_{k=K}^{\infty} c_k |R_k|\right] = o\left(\left(\sum_{k=K}^{\infty} c_k^2\right)^{1/2}\right).$$

Then on H the series $\sum_{k=1}^{\infty} c_k(\Phi_k + \epsilon_k)$ converges almost surely to a finite random variable L. Moreover, for any $p \in \mathbb{N}$ and any $\mathcal{F}_p$-measurable random variable Y we have

$$P(H \cap (L = Y)) = 0.$$

Lemma 7.1.4 ([148, Exercise 5.3.35]). *Let M_k be an L^2 martingale adapted to a filtration $\{\mathcal{F}_k\}$ and let $b_k \uparrow \infty$ be a positive deterministic sequence. Then if*

$$\sum_{k \geq 1} b_k^{-2} \mathbb{E}\left[(M_k - M_{k-1})^2 \mid \mathcal{F}_{k-1}\right] < +\infty,$$

we have $b_n^{-1} M_n \xrightarrow{a.s.} 0$.

Lemma 7.1.5 (Kronecker Lemma)**.** *Suppose* $\{x_k\}_k$ *is an infinite sequence of real number such that the sum* $\sum_{k=1}^{\infty} x_k$ *exists and is finite. Then for any divergent positive nondecreasing sequence* $\{b_k\}$*, we have*

$$\lim_{K \to \infty} \frac{1}{b_K} \sum_{k=1}^{K} b_k x_k = 0.$$

7.1.7.2 Lemmas used in Chapter 4

We will use the following two Lemmas on sequences. The proof of the following Lemma may be found in Appendix 7.1.8.

Lemma 7.1.6. *Fix* $k_0 \in \mathbb{N}, c > 0$*, and* $\gamma \in (1/2, 1]$*. Suppose that* $\{X_k\}, \{Y_k\}$*, and* $\{Z_k\}$ *are nonnegative random variables adapted to a filtration* $\{\mathcal{F}_k\}$*. Suppose the relationship holds:*

$$\mathbb{E}[X_{k+1} \mid \mathcal{F}_k] \leq (1 - ck^{-\gamma}) X_k - Y_k + Z_k \qquad \text{for all } k \geq k_0.$$

Assume furthermore that $c \geq 6$ *if* $\gamma = 1$*. Define the constants* $a_k := \frac{k^{2\gamma-1}}{\log^2(k+1)}$*. Then there exists a random variable* $V < \infty$ *such that on the event* $\{\sum_{k=1}^{\infty} a_{k+1} Z_k < +\infty\}$*, the following is true:*

1. *The limit holds*

$$a_k X_k \xrightarrow{a.s.} V.$$

2. *The sum is finite*

$$\sum_{k=1}^{\infty} a_{k+1} Y_k < +\infty.$$

The proof of the following Lemma may be found in Appendix 7.1.8.1.

Lemma 7.1.7. *Fix $k_0 \in \mathbb{N}$, $c, C > 0$, and $\gamma \in (1/2, 1]$. Suppose that $\{s_k\}_k$ is a nonnegative sequence satisfying*

$$s_k \le \frac{c}{12\gamma} \qquad \text{and} \qquad s_{k+1}^2 \le s_k^2 - ck^{-\gamma} s_k + Ck^{-2\gamma}, \qquad \text{for all } k \ge k_0,$$

Then, there exists a constant C_{ub} depending only on c, C, γ and k_0 such that

$$s_k \le C_{\mathsf{ub}} k^{-\gamma}, \qquad \forall k \ge 1.$$

The proof of the following Lemma may be found in Appendix 7.1.8.2.

Lemma 7.1.8. *Fix $k_0 \in \mathbb{N}$, $c, C > 0$, and $\gamma \in (1/2, 1]$. Suppose that $\{s_k\}_k$ is a nonnegative sequence satisfying*

$$s_{k+1} \le (1 - ck^{-\gamma}) s_k + Ck^{-2\gamma}, \qquad \text{for all } k \ge k_0,$$

Assume furthermore that $c \ge 16$ if $\gamma = 1$. Then, there exists a constant C_{ub} depending only on c, C, γ and k_0 such that

$$s_k \le C_{\mathsf{ub}} k^{-\gamma}, \qquad \forall k \ge 1.$$

7.1.8 Proof of Lemma 7.1.6

Proof. For all $k \ge 0$, define $a_k := \frac{k^{2\gamma-1}}{\log(k+1)^2}$ and observe that

$$\mathbb{E}[a_{k+1} X_{k+1} \mid \mathcal{F}_k] \le a_{k+1}(1 - ck^{-\gamma}) X_k - a_{k+1} Y_k + a_{k+1} Z_k \qquad \text{for all } k \ge k_0.$$

Thus, the result will follow from Robbins-Siegmund Lemma 7.1.1 if $a_{k+1}(1 - ck^{-\gamma}) \le a_k$ for all sufficiently large k. To that end, notice that for sufficiently large k, we have

$$\left(\frac{k+1}{k}\right)^{2\gamma-1} \le 1 + \frac{2(2\gamma-1)}{k}.$$

Therefore,

$$\frac{a_{k+1}}{a_k} \leq 1 + \frac{2(2\gamma - 1)}{k} \qquad \text{for all sufficiently large } k.$$

Now we deal separately with the cases $\gamma < 1$ and $\gamma = 1$. First suppose that $\gamma < 1$. Then there exists a constant $C' > 0$ such that

$$\frac{1}{1 - ck^{-\gamma}} \geq 1 + \frac{C'}{k^\gamma}, \qquad \text{for all sufficiently large } k.$$

Consequently, $a_{k+1}/a_k \leq (1 - ck^{-\gamma})^{-1}$ for all sufficiently large k, as desired.

Now assume that $\gamma = 1$. Then we compute

$$\frac{1}{1 - ck^{-1}} \geq 1 + \frac{c}{2k}, \qquad \text{for all sufficiently large } k.$$

Consequently, $a_{k+1}/a_k \leq (1 - ck^{-1})^{-1}$ for all large k, provided that $c \geq 6$. □

7.1.8.1 Proof of Lemma 7.1.7

Proof. It suffices to exhibit $C'_{\text{ub}} > 0$ such that

$$s_k \leq C'_{\text{ub}} k^{-\gamma} \qquad \text{for all sufficiently large } k \geq k_0.$$

To that end, choose k_1 large enough that the following two bounds hold:

1. $C'_{\text{ub}} := \max\left\{\frac{ck_1^\gamma}{12\gamma}, 2\sqrt{C}, \frac{4C}{c}\right\} = \frac{ck_1^\gamma}{12\gamma}$
2. $\left(\frac{k_1+1}{k_1}\right)^{2\gamma} \leq \min\left\{2, 1 + \frac{3\gamma}{k_1}\right\}$.

Then by assumption, we have

$$k^{2\gamma} s_{k+1}^2 \leq k^{2\gamma} s_k^2 - ck^\gamma s_k + C \qquad \text{for all } k \geq k_1.$$

Denoting $t_k := k^\gamma s_k$, we obtain the following bound for all $k \geq k_1$:

$$t_{k+1}^2 \leq \left(\frac{k+1}{k}\right)^{2\gamma} (t_k^2 - ct_k + C) \leq \left(\frac{k_1+1}{k_1}\right)^{2\gamma} (t_k^2 - ct_k + C). \tag{7.1.13}$$

Thus the claim will follow if $t_k \leq C'_{\mathrm{ub}}$ for all $k \geq k_1$. We prove the claim by induction. First the case $k = k_1$ holds by definition of C'_{ub}. Now suppose $t_k \leq C'_{\mathrm{ub}}$ for some $k \geq k_1$ and consider two cases

First suppose $t_k \in [0, \frac{1}{2}C'_{\mathrm{ub}}]$. By (7.1.13) and definition of C'_{ub}, we have

$$\begin{aligned} t_{k+1}^2 &\leq \left(\frac{k_1+1}{k_1}\right)^{2\gamma} (t_k^2 + C) \\ &\leq \left(\frac{k_1+1}{k_1}\right)^{2\gamma} \left(\frac{1}{4}C'^2_{\mathrm{ub}} + \frac{1}{4}C'^2_{\mathrm{ub}}\right) \\ &\leq C'^2_{\mathrm{ub}}. \end{aligned}$$

Second, suppose $t_k \in [\frac{1}{2}C'_{\mathrm{ub}}, C'_{\mathrm{ub}}]$. By (7.1.13) and definition of C'_{ub}, we have

$$\begin{aligned} t_{k+1}^2 &\leq \left(\frac{k_1+1}{k_1}\right)^{2\gamma} (t_k^2 - ct_k + C) \\ &\leq \left(\frac{k_1+1}{k_1}\right)^{2\gamma} \left(C'^2_{\mathrm{ub}} - \frac{cC'_{\mathrm{ub}}}{2} + C\right) \\ &\leq \left(\frac{k_1+1}{k_1}\right)^{2\gamma} \left(C'^2_{\mathrm{ub}} - \frac{cC'_{\mathrm{ub}}}{4}\right) \\ &= C'_{\mathrm{ub}} \left(\frac{k_1+1}{k_1}\right)^{2\gamma} \left(C'_{\mathrm{ub}} - \frac{c}{4}\right) \end{aligned}$$

We claim that $\left(\frac{k_1+1}{k_1}\right)^{2\gamma} (C'_{\mathrm{ub}} - \frac{c}{4}) \leq C'_{\mathrm{ub}}$. Indeed, we have

$$\begin{aligned} &\left(\frac{k_1+1}{k_1}\right)^{2\gamma} \left(C'_{\mathrm{ub}} - \frac{c}{4}\right) \\ &\leq \left(1 + \frac{3\gamma}{k_1}\right)\left(C'_{\mathrm{ub}} - \frac{c}{4}\right) \\ &\leq C'_{\mathrm{ub}} + \frac{3\gamma C'_{\mathrm{ub}}}{k_1} - \frac{c}{4} \\ &\leq C'_{\mathrm{ub}} + \frac{c}{4k_1^{1-\gamma}} - \frac{c}{4} \\ &\leq C'_{\mathrm{ub}}, \end{aligned}$$

as desired. This completes the induction. □

7.1.8.2 Proof of Lemma 7.1.8

Proof. It suffices to exhibit $C_{\mathsf{ub}} > 0$ such that

$$s_k \le C_{\mathsf{ub}} k^{-\gamma} \qquad \text{for all sufficiently large } k \ge k_0.$$

To that end, choose k_1 large enough that the following two bounds hold:

1. $(\frac{k+1}{k})^{\gamma} \le 1 + \frac{2\gamma}{k} \le 2$ for all $k \ge k_1$.
2. $k_1^{1-\gamma} \ge \frac{16\gamma}{c}$ if $\gamma \in (\frac{1}{2}, 1)$.

Now let $t_k = s_k k^{\gamma}$, then we rewrite the above inequality as

$$t_{k+1} \le \left(\frac{k+1}{k}\right)^{\gamma} \left[(1 - ck^{-\gamma})t_k + \frac{C}{k^{\gamma}}\right], \qquad \text{for all } k \ge k_0. \tag{7.1.14}$$

Let $C_{\mathsf{ub}} = \max\{s_{k_1} k_1^{\gamma}, 4C, \frac{8C}{c}\}$. By definition of C_{ub}, we know that

$$t_{k_1} = s_{k_1} k_1^{\gamma} \le C_{\mathsf{ub}}.$$

For the induction step, we consider two cases.

First suppose $t_k \in [0, \frac{1}{4}C_{\mathsf{ub}}]$. By (7.1.14) and definition of C_{ub}, we have

$$\begin{aligned} t_{k+1} &\le \left(\frac{k_0+1}{k_0}\right)^{\gamma} (t_k + C) \\ &\le \left(\frac{k_0+1}{k_0}\right)^{\gamma} \left(\frac{1}{4}C_{\mathsf{ub}} + \frac{1}{4}C_{\mathsf{ub}}\right) \\ &\le C_{\mathsf{ub}}. \end{aligned}$$

Second, suppose $t_k \in [\frac{1}{4}C_{\mathsf{ub},k_0,\bar{x}}, C_{\mathsf{ub},k_0,\bar{x}}]$. By (7.1.14) and definition of $\tilde{C}_{\mathsf{ub},k_0,\bar{x}}$, we have

$$\begin{aligned} t_{k+1} &\le \left(\frac{k+1}{k}\right)^{\gamma} \left(t_k - \frac{ct_k}{k^{\gamma}} + \frac{C}{k^{\gamma}}\right) \\ &\le \left(\frac{k+1}{k}\right)^{\gamma} \left(C_{\mathsf{ub}} - \frac{cC_{\mathsf{ub}}}{4k^{\gamma}} + \frac{C}{k^{\gamma}}\right) \end{aligned}$$

$$\begin{aligned}
&\leq \left(\frac{k+1}{k}\right)^{\gamma} \left(C_{\mathrm{ub}} - \frac{cC_{\mathrm{ub}}}{8k^{\gamma}}\right) \\
&= C_{\mathrm{ub}} \left(\frac{k+1}{k}\right)^{2\gamma} \left(1 - \frac{c}{8k^{\gamma}}\right)
\end{aligned}$$

We claim that $(\frac{k+1}{k})^{\gamma} (1 - \frac{c}{8k^{\gamma}}) \leq 1$. Indeed, we have

$$\begin{aligned}
&\left(\frac{k+1}{k}\right)^{\gamma} \left(1 - \frac{c}{8k^{\gamma}}\right) \\
&\leq \left(1 + \frac{2\gamma}{k}\right)\left(1 - \frac{c}{8k^{\gamma}}\right) \\
&\leq 1 + \frac{2\gamma}{k} - \frac{c}{8k^{\gamma}}
\end{aligned}$$

When $\gamma = 1$, $1 + \frac{2\gamma}{k} - \frac{c}{8k^{\gamma}}$ by our assumption on c. When $\gamma \in (\frac{1}{2}, 1)$, $1 + \frac{2\gamma}{k} - \frac{c}{8k^{\gamma}} \leq 1$ by our choice of k_0. This completes the induction. □

7.2 Proofs for Asymptotic normality and optimality

This supplement contains all the missing proofs justifying the results in Chapter 5. The supplement proceeds linearly through the sections.

7.2.1 Proof of Theorem 5.3.2

The proof of Theorem 5.3.2 follows from two lemmas. The first allows one to reduce the sensitivity analysis of the inclusion $v \in A(x) + \partial f(x)$ to an entirely smooth setting. More precisely, the following basic result, proved in [149, Proposition 10.12], shows that as soon as f admits an active manifold, the graph of the subdifferential ∂f admits a smooth description.

Lemma 7.2.1 (Smooth reduction)**.** *Let f be a subdifferentially continuous function that* *admits a C^2 active manifold $\mathcal{M}$ at a point $\bar{x}$ for a vector $\bar{w} \in \hat{\partial} f(\bar{x})$. Then equality holds:*

$$\operatorname{gph} \partial f = \operatorname{gph} \partial(f + \delta_{\mathcal{M}}) \qquad \textit{locally around} \qquad (\bar{x}, \bar{w}).$$

Note that letting $\hat{f}$ be a C^2-smooth function that agrees with f on $\mathcal{M}$ near $\bar{x}$, we may write

$$\partial(f + \delta_{\mathcal{M}})(x) = \partial(\hat{f} + \delta_{\mathcal{M}})(x) = \nabla \hat{f}(x) + N_{\mathcal{M}}(x).$$

Thus, under the same assumptions as in Lemma 7.2.1, equality :

$$\operatorname{gph} \partial f = \operatorname{gph} (\nabla \hat{f} + N_{\mathcal{M}}) \qquad \text{holds locally around} \qquad (\bar{x}, \bar{w}).$$

It follows from the lemma, that we may now focus on variational inclusions of the form $v \in \Phi(x) + N_{\mathcal{M}}(x)$, where Φ and $\mathcal{M}$ are smooth. Perturbation theory of such inclusions is entirely classical and is summarized in the following lemma.

Lemma 7.2.2 (Smooth variational inequality)**.** *Consider a set-valued map*

$$F(x) = \Phi(x) + N_{\mathcal{M}}(x) \tag{7.2.1}$$

and the let $\bar{x}$ be a point satisfying $0 \in F(\bar{x})$. Suppose that $\Phi \colon \mathbb{R}^d \to \mathbb{R}^d$ is a C^p-smooth map and $\mathcal{M} \subset \mathbb{R}^d$ is a C^{p+1}-smooth manifold around $\bar{x}$. Let $G(x) = 0$ be any C^{p+1}-smooth local defining equations for $\mathcal{M}$ and define the map

$$\mathcal{H}(x, y) = \Phi(x) + \nabla G(x)^\top y.$$

Then there exists a unique vector $\bar{y}$ satisfying the condition $0 = \mathcal{H}(\bar{x}, \bar{y})$. Moreover, F is C^p-invertible around $(0, \bar{x})$ with inverse $\sigma(\cdot)$ if and only if $P_{T_{\mathcal{M}}(\bar{x})} \nabla_x \mathcal{H}(\bar{x}, \bar{y}) P_{T_{\mathcal{M}}(\bar{x})}$ is nonsingular on $T_{\mathcal{M}}(\bar{x})$. In this case, equality holds:

$$\nabla \sigma(0) = (P_{T_{\mathcal{M}}(\bar{x})} \nabla_x \mathcal{H}(\bar{x}, \bar{y}) P_{T_{\mathcal{M}}(\bar{x})})^\dagger.$$

Proof. The existence of $\bar{y}$ follows from the expression $N_{\mathcal{M}}(\bar{x}) = \mathrm{range}(\nabla G(\bar{x})^{\top})$, while uniqueness follows from surjectivity of $\nabla G(\bar{x})$.

We first prove the backward implication and derive the claimed expression for the Jacobian. Suppose that $P_{T_{\mathcal{M}}(\bar{x})}\nabla_x \mathcal{H}(\bar{x}, \bar{y})P_{T_{\mathcal{M}}(\bar{x})}$ is indeed nonsingular on $T_{\mathcal{M}}(\bar{x})$. Then there exists $\epsilon > 0$ such that for any $v \in \epsilon\mathbb{B}$ and $x \in B_{\epsilon}(\bar{x})$, the inclusion $v \in \Phi(x) + N_{\mathcal{M}}(x)$ holds if and only if there exists y satisfying

$$\begin{Bmatrix} v = \Phi(x) + \nabla G(x)^{\top} y \\ 0 = G(x) \end{Bmatrix}. \tag{7.2.2}$$

Treating the right-hand-side as a mapping of (x, y), its Jacobian at $(\bar{x}, \bar{y})$ is given by

$$\begin{bmatrix} \nabla_x \mathcal{H}(\bar{x}, \bar{y}) & \nabla G(\bar{x})^{\top} \\ \nabla G(\bar{x}) & 0 \end{bmatrix}. \tag{7.2.3}$$

A quick computation shows that this matrix is invertible since $P_{T_{\mathcal{M}}(\bar{x})}\nabla_x \mathcal{H}(\bar{x}, \bar{y})P_{T_{\mathcal{M}}(\bar{x})}$ is nonsingular on $T_{\mathcal{M}}(\bar{x})$. Therefore, the inverse function theorem ensures that for all small v, the system (7.2.3) admits a unique solution $\sigma(v)$ near $\bar{x}$, and which varies C^p smoothly in v. Inverting (7.2.3) yields the expression for the Jacobian $\nabla\sigma(0) = (P_{T_{\mathcal{M}}(\bar{x})}\nabla_x \mathcal{H}(\bar{x}, \bar{y})P_{T_{\mathcal{M}}(\bar{x})})^{\dagger}$.

Conversely, suppose that F is C^p-smoothly invertible around $(0, \bar{x})$ with inverse $\sigma(\cdot)$. Fix a vector $v \in \mathbb{R}^d$. Then for all sufficiently small $t > 0$ there exists a unique vector $y(t) \in \mathbb{R}^d$ satisfying

$$tv = \Phi(\sigma(tv)) + \nabla G(\sigma(tv))^{\top} y(t).$$

Subtracting the equation $0 = \Phi(\bar{x}) + \nabla G(\bar{x})^{\top}\bar{y}$ and projecting both sides to $P_{T_{\mathcal{M}}(\bar{x})}$ yields

$$P_{T_{\mathcal{M}}(\bar{x})}v = P_{T_{\mathcal{M}}(\bar{x})}\left[\frac{\Phi(\sigma(tv)) - \Phi(\bar{x})}{t}\right] + P_{T_{\mathcal{M}}(\bar{x})}\left[\frac{\nabla G(\sigma(tv))^{\top}}{t} y(t)\right].$$

It is straightforward to see that $y(t)$ is continuous, and therefore the right-side tends to $P_{T_{\mathcal{M}}(\bar{x})}\nabla_x \mathcal{H}(\bar{x}, \bar{y})\nabla\sigma(0)v$ as t tends to zero. Summarizing, since v is arbitrary, we have shown the matrix identity

$$P_{T_{\mathcal{M}}(\bar{x})} = P_{T_{\mathcal{M}}(\bar{x})} \nabla_x \mathcal{H}(\bar{x}, \bar{y}) \nabla \sigma(0).$$

Taking into account that the range of $\nabla\sigma(0)$ is contained in $T_{\mathcal{M}}(\bar{x})$, it follows that the range of $P_{T_{\mathcal{M}}(\bar{x})} \nabla_x \mathcal{H}(\bar{x}, \bar{y}) P_{T_{\mathcal{M}}(\bar{x})}$ must be equal to $T_{\mathcal{M}}(\bar{x})$. Therefore $P_{T_{\mathcal{M}}(\bar{x})} \nabla_x \mathcal{H}(\bar{x}, \bar{y}) P_{T_{\mathcal{M}}(\bar{x})}$ must be nonsingular on $T_{\mathcal{M}}(\bar{x})$, as claimed. □

We are now ready to complete the proof of Theorem 5.3.2. Namely, Lemma 7.2.1 together with continuity of $A(\cdot)$ directly imply that locally around $(\bar{x}, 0)$, the graph of F coincides with the graph of the map

$$x \mapsto \Phi(x) + N_{\mathcal{M}}(x),$$

where we set $\Phi(x) = A(x) + \nabla \hat{f}(x)$. Lemma 7.2.2 directly implies that F is C^p-invertible around $(0, \bar{x})$ with inverse $\sigma(\cdot)$ if and only if $\Sigma = P_{T_{\mathcal{M}}(\bar{x})} \nabla_x \mathcal{H}(\bar{x}, \bar{y}) P_{T_{\mathcal{M}}(\bar{x})}$ is nonsingular on $T_{\mathcal{M}}(\bar{x})$. In this case, equality $\nabla\sigma(0) = \Sigma^{\dagger}$ holds.

7.2.2 Proof of Theorem 5.4.1

Define the random vectors $w_k := A(x_k) - A_S(x_k)$ and define the events

$$\mathcal{E}_k := \{x_k \in B_{\epsilon_2}(\bar{x})\} \qquad \text{and} \qquad \mathcal{Z}_k = \{w_k \in B_{\epsilon_1}(0)\}.$$

Note that $1_{\mathcal{E}_k} \xrightarrow{p} 1$ by our assumptions. The very definition of x_k implies the inclusion $w_k \in (A + H)(x_k)$. Therefore, the equality holds:

$$\sqrt{k}(x_k - \bar{x}) 1_{\mathcal{E}_k \cap \mathcal{Z}_k} = \sqrt{k}[\sigma(w_k 1_{\mathcal{E}_k \cap \mathcal{Z}_k}) - \sigma(0)]. \tag{7.2.4}$$

Our task therefore reduces to computing the asymptotics of $\sqrt{k} w_k 1_{\mathcal{E}_k \cap \mathcal{Z}_k}$ and then performing a first-order expansion.

Let us first show $1_{\mathcal{Z}_k} \xrightarrow{p} 1$. A first-order expansion of $A(\cdot)$ and $A(\cdot, z_i)$ around $\bar{x}$ yields

$$\begin{aligned} &\|w_k - \underbrace{(A(\bar{x}) - A_S(\bar{x}))}_{O_P(1/\sqrt{k})} + \underbrace{(\nabla A(\bar{x}) - \nabla A_S(\bar{x}))(x_k - \bar{x})}_{O_P(1/\sqrt{k})}\|1_{\mathcal{E}_k} \\ &\quad \le \tfrac{1}{2}\Big(\mathbb{E}L + \underbrace{\frac{1}{k}\sum_{i=1}^{k} L(z_i)}_{\mathbb{E}L + O_P(1/\sqrt{k})}\Big)\|x_k - \bar{x}\|^2 1_{\mathcal{E}_k}, \end{aligned} \tag{7.2.5}$$

where the statements in brackets follow from the central limit theorem. Rearranging, yields

$$\|w_k 1_{\mathcal{E}_k}\| \le \epsilon_2^2 \mathbb{E}L + O_P(1/\sqrt{k}) \le \frac{\epsilon_1}{2} + O_P(1/\sqrt{k}).$$

We conclude $1_{\mathcal{Z}_k \cap \mathcal{E}_k} \xrightarrow{p} 1$. Taking into account $1_{\mathcal{E}_k} \xrightarrow{p} 1$, we deduce $1_{\mathcal{Z}_k} \xrightarrow{p} 1$, as claimed.

Next, we show $\|x_k - \bar{x}\| = O_P(1/\sqrt{k})$. Observe that (7.2.4) directly implies $\|x_k - \bar{x}\| 1_{\mathcal{E}_k \cap \mathcal{Z}_k} \le \mathrm{lip}(\sigma)\|w_k 1_{\mathcal{E}_k \cap \mathcal{Z}_k}\|$. Combining this with (7.2.5), after multiplying through by $1_{\mathcal{Z}_k}$, we deduce

$$\mathrm{lip}(\sigma)^{-1}\|x_k - \bar{x}\| 1_{\mathcal{E}_k \cap \mathcal{Z}_k} = O_P(1/\sqrt{k}) + (\mathbb{E}L + O_P(1/\sqrt{k}))\|x_k - \bar{x}\|^2 1_{\mathcal{E}_k \cap \mathcal{Z}_k}.$$

Rearranging, yields

$$(\mathrm{lip}(\sigma)^{-1} - \mathbb{E}L\|x_k - \bar{x}\|) \cdot \|x_k - \bar{x}\| 1_{\mathcal{E}_k \cap \mathcal{Z}_k} = O_P(1/\sqrt{k}).$$

Noting that the coefficient on the left hand side is bounded below by $1/2$ in the event $\mathcal{E}_k$, we deduce $\|x_k - \bar{x}\| 1_{\mathcal{E}_k \cap \mathcal{Z}_k} = O_P(1/\sqrt{k})$. Taking into account $1_{\mathcal{E}_k \cap \mathcal{Z}_k} \xrightarrow{p} 1$, we deduce $\|x_k - \bar{x}\| = O_P(1/\sqrt{k})$ as claimed.

Finally, multiplying (7.2.5) through by $\sqrt{k}$ yields

$$\begin{aligned} \sqrt{k} w_k 1_{\mathcal{E}_k} = &\underbrace{\sqrt{k}(A(\bar{x}) - A_S(\bar{x}))1_{\mathcal{E}_k}}_{\xrightarrow{D} \mathrm{N}(0, \mathrm{Cov}(A(\bar{x}, z))} + \underbrace{(\nabla A(\bar{x}) - \nabla A_S(\bar{x}))\sqrt{k}(x_k - \bar{x})1_{\mathcal{E}_k}}_{\xrightarrow{p} 0} \\ &+ \underbrace{\frac{1}{2}\left(\mathbb{E}L(z) + \frac{1}{k}\sum_{i=1}^{k} L(z_i)\right) O(\sqrt{k}\|x_k - \bar{x}\|^2) 1_{\mathcal{E}_k}}_{\xrightarrow{p} 0}, \end{aligned}$$

where the claimed limits follow from the central limit theorem and the fact that $\sqrt{k}\|x_k - \bar{x}\|$ is bounded in probability. Multiplying through by $1_{\mathcal{Z}_k}$, we deduce

$$\sqrt{k}w_k 1_{\mathcal{E}_k \cap \mathcal{Z}_k} = \sqrt{k}(A(\bar{x}) - A_S(\bar{x}))1_{\mathcal{E}_k \cap \mathcal{Z}_k} + o_p(1).$$

Thus returning to (7.2.4) and using a first-order expansion, we obtain

$$\sqrt{k}(x_k - \bar{x})1_{\mathcal{E}_k \cap \mathcal{Z}_k} = \sqrt{k}\nabla\sigma(0)(A(\bar{x}) - A_S(\bar{x}))1_{\mathcal{E}_k \cap \mathcal{Z}_k} + o_p(1).$$

The proof is complete upon removing $1_{\mathcal{E}_k \cap \mathcal{Z}_k}$ from both sides, which can be done by noting that $1_{\mathcal{Z}_k \cap \mathcal{E}_k} \xrightarrow{p} 1$ and $\sqrt{k}(x_k - \bar{x}) = O_P(1)$ and $\sqrt{k}(A(\bar{x}) - A_S(\bar{x})) = O_P(1)$.

7.2.3 Proof of Lemma 5.5.2

Throughout the proof, let $\epsilon > 0$ and L be such that

$$\max\{\|s_g(x)\|, \|A(x)\|\} \leq L \qquad \forall x \in B_\epsilon(\bar{x}).$$

To see Claim (1), for all x sufficiently close to $\bar{x}$, we compute

$$\begin{aligned}
\alpha\|G_\alpha(x,\nu)\| &= \|x - s_f(x - \alpha(A(x) + s_g(x) + \nu))\| \\
&\leq \operatorname{dist}_{\mathcal{X}}(x - \alpha(A(x) + s_g(x) + \nu)) + \alpha\|A(x) + s_g(x) + \nu\| \\
&\leq 2\alpha\|A(x) + s_g(x) + \nu\| \\
&= 2\alpha(2L + \|\nu\|).
\end{aligned}$$

Throughout the rest of the proof, we set $x_+ := x - \alpha G_\alpha(x,\nu)$. We now verify Claim 2. To this end, suppose that f is convex and by increasing L we may ensure $\operatorname{dist}(0, \partial f(x))\} \leq L$ for all $x \in B_\epsilon(\bar{x}) \cap \operatorname{dom} f$. Choose a vector $z \in \partial f(x)$ of minimal length. Algebraic manipulations show $x = \operatorname{prox}_{\alpha f}(x + \alpha z)$. Since $\operatorname{prox}_{\alpha f}$ is nonexpansive, for all $x \in B_\epsilon(\bar{x})$, we deduce

$$\alpha\|G_\alpha(x,\nu)\| = \|x - \operatorname{prox}_{\alpha f}(x - \alpha(A(x) + s_g(x) + \nu))\|$$

$$
\begin{aligned}
&\le \alpha\|A(x) + s_g(x) + z + v\| \\
&\le \alpha(3L + \|v\|),
\end{aligned}
$$

as claimed.

We next verify Claim 3. To this end, suppose that f is L-Lipschitz continuous on $\operatorname{dom} g \cap \operatorname{dom} f$. Set $w := A(x) + s_g(x) + v$ and observe that the very definition of x_+ ensures

$$
\begin{aligned}
\frac{1}{2\alpha}\|x_+ - x\|^2 &\le f(x) - f(x_+) - \langle w, x_+ - x\rangle \\
&\le L\|x_+ - x\| + \|w\|\|x_+ - x\|.
\end{aligned}
$$

Consequently, for all $x \in B_\epsilon(\bar{x})$ we have $\alpha^{-1}\|x_+ - x\| \le 2(L + \|w\|) \le 2(3C + \|v\|)$, as desired.

7.2.4 Proof of Lemma 5.5.3

Strong (a)-regularity implies the equalities, $P_{T_{\mathcal{M}}(x)}\partial f(x) = \{\nabla_{\mathcal{M}} f(x)\}$ and $P_{T_{\mathcal{M}}(x)}\partial g(x) = \{\nabla_{\mathcal{M}} g(x)\}$, for all $x \in \mathcal{M}$ near $x^\star$. The claimed expression (5.5.5) follows immediately. Next, Lemma 7.2.1 implies that $\operatorname{gph}\partial(f+g)$ coincides with $\operatorname{gph}[\nabla_{\mathcal{M}}(f+g)+N_{\mathcal{M}}]$ around $(x^\star, -A(x^\star))$. Thus locally around $(x^\star, 0)$ equalities hold:

$$
\operatorname{gph}[A + \partial(f+g)] = \operatorname{gph}[A + \nabla_{\mathcal{M}} f + \nabla_{\mathcal{M}} g + N_{\mathcal{M}}] = \operatorname{gph}(F_{\mathcal{M}} + N_{\mathcal{M}}),
$$

as claimed.

7.2.5 Proof of Proposition 5.5.4

Proof. Fix $x \in \mathcal{X}$, $\alpha > 0$, and $v \in \mathbb{R}^d$. Assumption J holds trivially, and follows for example from Lemma 5.5.2(1). In order to verify Assumption (K1), we compute

$$\|P_{T_{\mathcal{M}}(P_{\mathcal{M}}(x))}(G_\alpha(x, v) - F(P_{\mathcal{M}}(x)) - v)\| = \|P_{T_{\mathcal{M}}(P_{\mathcal{M}}(x))}(A(x) + s_g(x) - F(P_{\mathcal{M}}(x)))\|.$$

Therefore, for x sufficiently close to $\bar{x}$ we may upper bound the right-hand-side as

$$\begin{aligned}&\|P_{T_{\mathcal{M}}(P_{\mathcal{M}}(x))}[s_g(x) - \nabla_{\mathcal{M}} g(P_{\mathcal{M}}(x))] + P_{T_{\mathcal{M}}(P_{\mathcal{M}}(x))}[A(x) - A(P_{\mathcal{M}}(x))]\| \\ &\leq C \cdot \operatorname{dist}(x, \mathcal{M}),\end{aligned}$$

where the inequality follows from (L1) and local Lipschitz continuity of $A(\cdot)$. Thus Assumption (K1) holds. Finally, to see Assumption (K2), we compute

$$\begin{aligned}\langle G_\alpha(x, v) - v, x - P_{\mathcal{M}}(x)\rangle &= \langle A(\bar{x}) + s_g(x), x - P_{\mathcal{M}}(x)\rangle + \langle A(x) - A(\bar{x}), x - P_{\mathcal{M}}(x)\rangle \\ &\geq \mu \cdot \operatorname{dist}(x, \mathcal{M}) - \|A(x) - A(\bar{x})\| \cdot \operatorname{dist}(x, \mathcal{M}) \\ &\geq \tfrac{\mu}{2} \cdot \operatorname{dist}(x, \mathcal{M}),\end{aligned}$$

for all x sufficiently close to $\bar{x}$. The proof is complete. □

7.2.6 Proof of Proposition 5.5.6

Choose $\epsilon > 0$ small enough that the following hold for all $x \in B_\epsilon(\bar{x}) \cap \mathcal{X}$. First (5.5.7) holds. In particular, since $A(\cdot)$ is locally Lipschitz near $\bar{x}$, we may be sure that

$$\langle A(x) + v, x - P_{\mathcal{M}}(x)\rangle \geq \frac{\mu}{2} \cdot \operatorname{dist}(x, \mathcal{M}), \tag{7.2.6}$$

for all $v \in \partial g(x)$. Second we require that for some $L > 0$, we have

$$\|P_{T_{\mathcal{M}}(P_{\mathcal{M}}(x))}(s_g(x) - \nabla_{\mathcal{M}} g(P_{\mathcal{M}}(x))\| \leq L \cdot \operatorname{dist}(x, \mathcal{M}), \tag{7.2.7}$$

$$\|P_{T_{\mathcal{M}(z)}}(u)\| \leq L\|x - z\|, \tag{7.2.8}$$

for all $u \in N_{\mathcal{X}}(x)$ of unit norm and all $z \in B_{\epsilon}(\bar{x}) \cap \mathcal{M}$, a consequence of (M1). Third, we may choose $\epsilon > 0$ so small so that

$$\langle z, x - x' \rangle \geq o(\|x - x'\|) \tag{7.2.9}$$

for all $z \in N_{\mathcal{X}}(x)$ of unit norm, and $x' \in \mathcal{M} \cap B_{\epsilon}(\bar{x})$—a consequence of (M3). We will fix $x \in B_{\epsilon/2}(\bar{x}) \cap \mathcal{X}$ and arbitrary $\alpha > 0$ and $\nu \in \mathbb{R}^d$, and choose an arbitrary $y \in P_{\mathcal{M}}(x)$. Define

$$w = G_{\alpha}(x, \nu) - \nu - A(x) - s_g(x) \qquad \text{and} \qquad x_+ = s_{\mathcal{X}}(x - \alpha(A(x) + s_g(x) + \nu)).$$

Note the inclusion $w \in N_{\mathcal{X}}(x^+)$. Moreover, shrinking $\epsilon > 0$ Assumption J directly implies

$$\max\{\|w\|, \|G_{\alpha}(x, \nu)\|\} \leq C(1 + \|\nu\|) \qquad \text{and} \qquad \|x_+ - x\| \leq C(1 + \|\nu\|)\alpha,$$

for some constant $C > 0$. We will use these estimates often in the proof. Finally, we let C be a constant independent of x, α and ν, which changes from line to line.

Assumption (K1): Suppose first that $x_+ \in B_{\epsilon}(\bar{x})$. Using (7.2.8), we compute

$$\begin{aligned}
\|P_{T_{\mathcal{M}(y)}} w\| &\leq L\|w\|\|x_+ - y\| \\
&\leq L\|w\|(\|x_+ - x\| + \operatorname{dist}(x, \mathcal{M})) \\
&\leq C(1 + \|\nu\|)^2\alpha + C(1 + \|\nu\|)\operatorname{dist}(x, \mathcal{M}). && (7.2.10)
\end{aligned}$$

On the other hand, if $x_+ \notin B_{\epsilon}(\bar{x})$, then we compute

$$\|P_{T_{\mathcal{M}(y)}} w\| \leq \|w\| \leq \frac{2}{\epsilon}\|w\|\|x_+ - x\| \leq C(1 + \|\nu\|)^2\alpha. \tag{7.2.11}$$

In either case, Assumption (K1) now follows since from (7.2.7) we have

$$\|P_{T_{\mathcal{M}(y)}}(A(x) + s_g(x) - F(y)\| \leq \|P_{T_{\mathcal{M}(y)}}(\nabla_{\mathcal{M}} g(y) - s_g(x))\|$$

$$+ \|P_{T_{\mathcal{M}(y)}}[A(x) - A(y)]\|$$

$$\leq C\mathrm{dist}(x, \mathcal{M}),$$

as we had to show.

Assumption (K2): We write the decomposition

$$\langle G_\alpha(x, v) - v, x - y\rangle = \underbrace{\langle A(x) + s_g(x), x - y\rangle}_{R_1} + \underbrace{\langle w, x_+ - y\rangle}_{R_2} + \underbrace{\langle w, x - x_+\rangle}_{R_3}. \tag{7.2.12}$$

The estimate (7.2.6) gives

$$R_1 \geq \tfrac{\mu}{2} \cdot \mathrm{dist}(x, \mathcal{M}). \tag{7.2.13}$$

We next look at two cases. Suppose first $x_+ \in B_\epsilon(\bar{x})$. Using the inclusion $w \in N_{\mathcal{X}}(x_+)$ and (7.2.9), we compute

$$\begin{aligned} R_2 &\geq \|w\| \cdot o(\|x_+ - y\|) \geq \|w\| \cdot (o(\|y - x\|) - \|x - x_+\|) \\ &\geq -C(1 + \|v\|)^2(o(\mathrm{dist}(x, \mathcal{M})) + \alpha). \end{aligned} \tag{7.2.14}$$

Next, the Cauchy–Schwarz inequality implies

$$|R_3| \leq \|w\|\|x - x_+\| \leq C(\alpha(1 + \|v\|)^2). \tag{7.2.15}$$

Combining (7.2.12)-(7.2.15) yields the claimed bound (K2).

Suppose now on the contrary that $x_+ \notin B_\epsilon(\bar{x})$ and therefore $\|x - y\| \leq \|x - x_+\|$. We thus deduce $R_2 + R_3 = \langle w, x - y\rangle \geq -\|w\|\|x - y\| \geq -C\alpha(1 + \|v\|)^2$ holds. Combining this estimate with (7.2.12) and (7.2.13) verifies the claim (K2).

7.2.7 Proof of Proposition 5.5.8

Let $\epsilon \in (0, 1)$ be small enough such that the following hold for all $x \in B_\epsilon(\bar{x}) \cap \mathrm{dom}\, f$. First (5.5.8) holds and therefore taking into account Lipschitz continuity of $A(\cdot)$, we may equivalently write

$$\langle A(x) + v, x - P_{\mathcal{M}}(x)\rangle \geq \frac{\mu}{2} \cdot \operatorname{dist}(x, \mathcal{M}) - (1 + \|v\|)o(\operatorname{dist}(x, \mathcal{M})), \tag{7.2.16}$$

for all $v \in \hat{\partial} f(x)$. Second we require that for some $L > 0$, we have

$$\|P_{T_{\mathcal{M}}(P_{\mathcal{M}}(x))}(u - \nabla_{\mathcal{M}} f(P_{\mathcal{M}}(x))\| \leq L\sqrt{1 + \|u\|^2} \cdot \operatorname{dist}(x, \mathcal{M}) \tag{7.2.17}$$

for all $u \in \partial f(x)$, a consequence of strong (a) regularity. Third, we assume that $\nabla_{\mathcal{M}} f$ is L-Lipschitz on $B_\epsilon(\bar{x}) \cap \mathcal{M}$. Fourth, we assume that $A(\cdot)$ is L-Lipschitz. Shrinking ϵ we may moreover assume $\epsilon \leq \frac{\mu}{8L}$. Finally, we may also assume that the maps $x \mapsto P_{\mathcal{M}}(x)$ and $x \mapsto P_{T_{\mathcal{M}}(P_{\mathcal{M}}(x))}$ are Lipschitz continuous on $B_\epsilon(\bar{x})$.

Fix $x \in B_{\epsilon/2}(\bar{x}) \cap \operatorname{dom} f$ and $v \in \mathbb{R}^d$ and set $y := P_{\mathcal{M}}(x)$. We define the vectors

$$w = G_\alpha(x, v) - A(x) - v \qquad \text{and} \qquad x_+ = s_f(x - \alpha(A(x) + v)).$$

Simple algebraic manipulations show the inclusion $w \in \hat{\partial} f(x_+)$. Moreover, Assumption J directly implies

$$\max\{\|w\|, \|G_\alpha(x, v)\|\} \leq C(1 + \|v\|) \qquad \text{and} \qquad \|x_+ - x\| \leq C(1 + \|v\|)\alpha.$$

We will use these estimates often in the proof. Finally, we let C be a constant independent of x, α and v, which changes from line to line.

Assumption (K1): First suppose $x_+ \in B_\epsilon(\bar{x})$. Using the triangle inequality, we write

$$\begin{aligned}
&\|P_{T_{\mathcal{M}}(P_{\mathcal{M}}(x))}(G_\alpha(x, v) - F(P_{\mathcal{M}}(x)) - v)\| \\
&= \|P_{T_{\mathcal{M}}(P_{\mathcal{M}}(x))}(w + A(x) - A(P_{\mathcal{M}}(x)) - \nabla_{\mathcal{M}} f(P_{\mathcal{M}}(x)))\| \\
&\leq \underbrace{\|P_{T_{\mathcal{M}}(P_{\mathcal{M}}(x))}(w - \nabla_{\mathcal{M}} f(P_{\mathcal{M}}(x_+)))\|}_{R_1} + \underbrace{\|A(x) - A(P_{\mathcal{M}}(x))\|}_{R_2} + \underbrace{\|\nabla_{\mathcal{M}} f(P_{\mathcal{M}}(x))) - \nabla_{\mathcal{M}} f(P_{\mathcal{M}}(x_+)))\|}_{R_3}.
\end{aligned}$$

Using the triangle inequality and the estimate (7.2.17) we deduce

$$R_1 \leq \|P_{T_{\mathcal{M}}(P_{\mathcal{M}}(x_+))}(w - \nabla_{\mathcal{M}} f(P_{\mathcal{M}}(x_+)))\| + \|P_{T_{\mathcal{M}}(P_{\mathcal{M}}(x_+))} - P_{T_{\mathcal{M}}(P_{\mathcal{M}}(x))}\|_{\text{op}} \cdot \|w - \nabla_{\mathcal{M}} f(P_{\mathcal{M}}(x_+))\|$$

$$
\begin{aligned}
&\leq C(1+\|w\|)\cdot \operatorname{dist}(x_+,\mathcal{M}) + C\|x-x_+\|\cdot(1+\|w\|)\\
&\leq C(1+\|v\|)\operatorname{dist}(x_+,\mathcal{M}) + C(1+\|v\|)^2\alpha\\
&\leq C(1+\|v\|)(\operatorname{dist}(x,\mathcal{M}) + C\|x-x_+\|) + C(1+\|v\|)^2\alpha\\
&\leq C(1+\|v\|)\operatorname{dist}(x,\mathcal{M}) + C(1+\|v\|)^2\alpha.
\end{aligned}
$$

Moreover, clearly we have $R_2 \leq C\operatorname{dist}(x,\mathcal{M})$ and $R_3 \leq C\|x-x_+\| \leq (1+\|v\|)\alpha$. Condition (K1) follows immediately.

Now suppose that $x_+ \notin B_\epsilon(\bar{x})$, and therefore $\|x_+ - x\| \geq \epsilon/2$. Then, we may write

$$
\begin{aligned}
\|P_{T_{\mathcal{M}}(P_{\mathcal{M}}(x))}(G_\alpha(x,v) - v - \nabla f_{\mathcal{M}}(P_{\mathcal{M}}(x)))\| &\leq \|G_\alpha(x,v)\| + \|v\| + \|\nabla f_{\mathcal{M}}(P_{\mathcal{M}}(x))\|\\
&\leq \frac{2}{\epsilon}(\|G_\alpha(x,v)\| + \|v\| + \|\nabla f_{\mathcal{M}}(P_{\mathcal{M}}(x))\|)\|x-x_+\|\\
&\leq C(1+\|v\|)^2\alpha,
\end{aligned}
$$

as desired.

Assumption (K2): We begin with the decomposition

$$
\begin{aligned}
\langle G_\alpha(x,v) - v, x-y\rangle = &\underbrace{\langle A(x_+) + w, x_+ - P_{\mathcal{M}}(x_+)\rangle}_{R_1}\\
&+ \underbrace{\langle A(x) - A(x_+), x-y\rangle}_{R_2} + \underbrace{\langle A(x_+) + w, (x - P_{\mathcal{M}}(x)) - (x_+ - P_{\mathcal{M}}(x_+))\rangle}_{R_3}.
\end{aligned}
$$

We now bound the two terms on the right in the case $x_+ \in B_\epsilon(\bar{x})$. Using (7.2.16), we estimate

$$
\begin{aligned}
R_1 &\geq \tfrac{\mu}{2}\cdot \operatorname{dist}(x_+,\mathcal{M}) - (1+\|w\|)o(\operatorname{dist}(x_+,\mathcal{M}))\\
&\geq \tfrac{\mu}{2}\cdot(\operatorname{dist}(x,\mathcal{M}) - \|x-x_+\|) - (1+\|v\|)(o(\operatorname{dist}(x,\mathcal{M})) + \|x-x_+\|)\\
&\geq \tfrac{\mu}{2}\cdot \operatorname{dist}(x,\mathcal{M}) - (1+\|v\|)^2(o(\operatorname{dist}(x,\mathcal{M})) + C\alpha).
\end{aligned}
$$

Next, we compute

$$
|R_2| \leq \|A(x) - A(x_+)\|\cdot \operatorname{dist}(x,\mathcal{M}) \leq 2L\epsilon\cdot \operatorname{dist}(x,\mathcal{M}) \leq \frac{\mu}{4}\operatorname{dist}(x,\mathcal{M}).
$$

Next using Lipschitz continuity of the map $I - P_{\mathcal{M}}$ on $B_\epsilon(\bar{x})$, we deduce

$$|R_3| \leq C\|A(x_+) + w\| \cdot \|x - x_+\| \leq C(1 + \|v\|)^2\alpha.$$

The claimed proximal aiming condition follows immediately with μ replaced by $\mu/4$.

Let us look now at the case $x_+ \notin B_\epsilon(\bar{x})$, and therefore $\operatorname{dist}(x, \mathcal{M}) \leq \frac{\epsilon}{2} \leq \|x - x^+\|$. Then we compute

$$\begin{aligned}\langle G_\alpha(x, v) - v, x - P_{\mathcal{M}}(x)\rangle &\geq -\operatorname{dist}(x, \mathcal{M}) \cdot \|G_\alpha(x, v) - v\| \\ &= \operatorname{dist}(x, \mathcal{M}) - \operatorname{dist}(x, \mathcal{M})(1 + \|G_\alpha(x, v) - v\|) \\ &\geq \operatorname{dist}(x, \mathcal{M}) - C\|x - x_+\|(1 + C(1 + \|v\|)) \\ &\geq \operatorname{dist}(x, \mathcal{M}) - C(1 + \|v\|)^2\alpha,\end{aligned}$$

as desired. The proof is complete.

7.2.8 Proof of Theorem 5.6.1

7.2.8.1 The two pillars of the proof of Theorem 5.6.1

We begin by outlining the main ingredients of the proof. Namely, Assumption K at a point $\bar{x}$ guarantees two useful behaviors, provided the iterates $\{x_k\}$ of algorithm (5.5.2) remain in a small ball around $\bar{x}$. First x_k must approach the manifold $\mathcal{M}$ containing $\bar{x}$ at a controlled rate, a consequence of the proximal aiming condition. Second the shadow $y_k = P_{\mathcal{M}}(x_k)$ of the iterates along the manifold form an approximate Riemannian stochastic gradient sequence with an implicit retraction. Moreover, the approximation error of the sequence decays with $\operatorname{dist}(x_k, \mathcal{M})$ and α_k, quantities that quickly tend to zero.

The formal statements summarizing these two modes of behavior require local arguments. Consequently, we will frequently refer to the following stopping time: given an index $k \geq 1$ and a constant $\delta > 0$, define

$$\tau_{k,\delta} := \inf\{j \geq k \colon x_j \notin B_\delta(\bar{x})\}. \tag{7.2.18}$$

Note that the stopping time implicitly depends on $\bar{x}$, a point at which Assumption K is satisfied. The following proposition shows that sequence x_k rapidly approaches the manifold. It was proved in Section 4.3.1 specifically for optimization problems rather than for finding zeros of set-valued maps; the argument in this more general setting is identical.

Proposition 7.2.3 (Pillar I: aiming towards the manifold)**.** *Suppose that Assumptions I, J, K, and E hold. Let $\gamma \in (1/2, 1]$ and assume $c_1 \geq 32/\mu$ if $\gamma = 1$. Then for all $k_0 \geq 1$ and sufficiently small $\delta > 0$, there exists a constant C, such that the following hold with stopping time $\tau_{k_0,\delta}$ defined in* (7.2.18)*:*

1. *There exists a random variable $V_{k_0,\delta}$ such that*

   (a) The limit holds:

   $$\frac{k^{2\gamma-1}}{\log(k+1)^2}\operatorname{dist}^2(x_k, \mathcal{M})1_{\tau_{k_0,\delta}>k} \xrightarrow{a.s.} V_{k_0,\delta}.$$

   (b) The sum is almost surely finite:

   $$\sum_{k=1}^{\infty} \frac{k^{\gamma-1}}{\log(k+1)^2}\operatorname{dist}(x_k, \mathcal{M})1_{\tau_{k_0,\delta}>k} < +\infty.$$

2. *The following are true.*

   (a) The expected squared distance satisfies:

   $$\mathbb{E}[\operatorname{dist}^2(x_k, \mathcal{M})1_{\tau_{k_0,\delta}>k}] \leq C\alpha_k \qquad \textit{for all } k \geq 1.$$

(b) *The tail sum is bounded:*

$$\mathbb{E}\left[\sum_{i=k}^{\infty} \alpha_i \mathrm{dist}(x_i, \mathcal{M}) 1_{\tau_{k_0,\delta} > i}\right] \le C \sum_{i=k}^{\infty} \alpha_i^2 \qquad \textit{for all } k \ge 1.$$

Next, we study the evolution of the shadow $y_k = P_{\mathcal{M}}(x_k)$ along the manifold, showing that y_k is locally an inexact Riemannian stochastic gradient sequence with an error that asymptotically decays as x_k approaches the manifold. Consequently, we may control the error using Proposition 7.2.3. The following proposition was proved in Section 4.3.2 specifically for optimization problems rather than for finding zeros of set-valued maps; the argument in this more general setting is identical.

Proposition 7.2.4 (Pillar II: the shadow iteration)**.** *Suppose that Assumptions I, J, K,E hold. Then for all $k_0 \ge 1$ and sufficiently small $\delta > 0$, there exists a constant C, such that the following hold with stopping time $\tau_{k_0,\delta}$ defined in* (7.2.18)*: there exists a sequence of $\mathcal{F}_{k+1}$-measurable random vectors $E_k \in \mathbb{R}^d$ such that*

1. *The shadow sequence*

$$y_k = \begin{cases} P_{\mathcal{M}}(x_k) & \textit{if } x_k \in B_{2\delta}(\bar{x}) \\ \bar{x} & \textit{otherwise.} \end{cases}$$

satisfies $y_k \in B_{4\delta}(\bar{x}) \cap \mathcal{M}$ for all k and the recursion holds:

$$\boxed{y_{k+1} = y_k - \alpha_k F_{\mathcal{M}}(y_k) - \alpha_k P_{T_{\mathcal{M}}(y_k)}(\nu_k) + \alpha_k E_k \qquad \textit{for all } k \ge 1.} \qquad (7.2.19)$$

Moreover, for such k, we have $\mathbb{E}_k[P_{T_{\mathcal{M}}(y_k)}(\nu_k)] = 0$.

2. *Let $\gamma \in (1/2, 1]$ and assume that $c_1 \ge 32/\mu$ if $\gamma = 1$.*

   (a) *We have the following bounds for $k_0 \le k \le \tau_{k_0,\delta} - 1$:*

      i. $\max\{\mathbb{E}_k[\|E_k\| 1_{\tau_{k_0,\delta} > k}, \mathbb{E}_k[\|E_k\|^2 1_{\tau_{k_0,\delta} > k}]\} \le C.$

      ii. $\mathbb{E}[\|E_k\|^2 1_{\tau_{k_0,\delta} > k}] \le C\alpha_k.$

iii. The sum is finite:

$$\sum_{k=1}^{\infty} \frac{k^{\gamma-1}}{\log(k+1)^2} \max\{\|E_k\| 1_{\tau_{k_0,\delta}>k}, \mathbb{E}_k[\|E_k\|] 1_{\tau_{k_0,\delta}>k}\} < +\infty.$$

(b) The tail sum is bounded

$$\mathbb{E}\left[1_{\tau_{k_0,\delta}=\infty} \sum_{i=k}^{\infty} \alpha_i \|E_i\|\right] \leq C \sum_{i=k}^{\infty} \alpha_i^2 \qquad \textit{for all } k \geq 1.$$

With the two pillars we separate our study of the sequence x_k into two orthogonal components: In the tangent/smooth directions, we study the sequence y_k, which arises from an inexact gradient method with rapidly decaying errors and is amenable to the techniques of smooth optimization. In the normal/nonsmooth directions, we steadily approach the manifold, allowing us to infer strong properties of x_k from corresponding properties for y_k.

We now outline common notation and conventions used in the rest of the proof. We let $\mathcal{U}$ be the neighborhood of $\bar{x}$ where the standing assumptions hold. Shrinking $\mathcal{U}$, we may assume that the projection map $P_{\mathcal{M}}$ is C^2 in $\mathcal{U}$, and in particular $P_{\mathcal{M}}$ is Lipschitz with Lipschitz Jacobian. Throughout, the proof we shrink $\mathcal{U}$ several times, when needed.

Now, denote stopping time (7.2.18) by $\tau := \tau_{k_0,\delta}$ and the noise bound by $Q := \sup_{x \in B_\delta(\bar{x})} q(x)$. Observe that by Proposition 7.2.4, the shadow sequence y_k satisfies $y_k \in B_{4\delta}(x_k) \cap \mathcal{M} \subseteq B_\epsilon(\bar{x}) \cap \mathcal{M}$ and recursion (7.2.19) holds. In addition, we let C denote a constant depending on k_0 and δ, which may change from line to line.

7.2.8.2 Rates near strong local minimizers

As the first step, we obtain a fast rate of convergence under the growth condition (5.6.1); this rate is comparable to the convergence rate of SGD for minimizing smooth strongly convex functions. To this end, we first need the following Lemma ensuring that $F_{\mathcal{M}}$ has sufficient curvature in $B_{2\delta}(\bar{x})$.

Lemma 7.2.5 (Curvature)**.** *The estimate*

$$\langle F_{\mathcal{M}}(y), y - \bar{x}\rangle \geq \frac{\mu}{2}\|y - \bar{x}\|^2,$$

holds for all $x \in \mathcal{M}$ sufficiently close to $\bar{x}$.

Proof. Let Φ be a smooth extension of $F_{\mathcal{M}}$ to a neighborhood of $U \subset \mathbb{R}^d$ of $\bar{x}$. Consider an arbitrary sequence $x_i \in \mathcal{M}$ converging to $\bar{x}$. Passing to a subsequence, we may assume that the unit vectors $\frac{x_i - \bar{x}}{\|x_i - \bar{x}\|}$ converge to some unit vector $w \in T_{\mathcal{M}}(\bar{x})$. Let $H_i := \int_0^1 \nabla\Phi(\bar{x} + \tau(x_i - \bar{x}))\, d\tau$ denote the average Jacobian between $\bar{x}$ and x_i. Note that H_i clearly tends to $\nabla\Phi(\bar{x})$ as i tends to infinity. The fundamental theorem of calculus yields

$$\frac{\langle \Phi(x_i) - \Phi(\bar{x}), x_i - \bar{x}\rangle}{\|x_i - \bar{x}\|^2} = \left\langle H_i\left(\frac{x_i - \bar{x}}{\|x_i - \bar{x}\|}\right), \frac{x_i - \bar{x}}{\|x_i - \bar{x}\|}\right\rangle \underset{i\to\infty}{\longrightarrow} \langle \nabla\Phi(\bar{x})w, w\rangle \geq \mu.$$

Since $x_i \in \mathcal{M}$ was an arbitrary sequence converging to $\bar{x}$, the result follows. □

Next, we obtain a familiar one-step improvement guarantee.

Lemma 7.2.6 (One-step improvement)**.** *For all sufficiently small δ, there exists a constant C such that for any $k \geq k_0$, we have*

$$\mathbb{E}[\|y_{k+1} - \bar{x}\|^2\, 1_{\tau > k}] \leq \left(1 - \frac{\alpha_k \mu}{2}\right)\mathbb{E}[\|y_k - \bar{x}\|^2\, 1_{\tau>k}] + C\alpha_k^2. \tag{7.2.20}$$

Proof. Expanding $\|y_{k+1} - \bar{x}\|^2$, we obtain

$$\begin{aligned}
&\|y_{k+1} - \bar{x}\|^2\, 1_{\tau>k}\\
&= \|y_k - \alpha_k F_{\mathcal{M}}(y_k) - \alpha_k P_{T_{\mathcal{M}}(y_k)}(\nu_k) + \alpha_k E_k - \bar{x}\|^2\, 1_{\tau>k}\\
&= \|y_k - \alpha_k F_{\mathcal{M}}(y_k) + \alpha_k E_k - \bar{x}\|^2\, 1_{\tau>k} + \alpha_k^2\, \|P_{T_{\mathcal{M}}(y_k)}(\nu_k)\|^2\, 1_{\tau>k}
\end{aligned}$$

$$
\begin{aligned}
&- 2\alpha_k \langle y_k - \alpha_k F_{\mathcal{M}}(y_k) + \alpha_k E_k - \bar{x}, P_{T_{\mathcal{M}}(y_k)}(\nu_k)\rangle\, 1_{\tau>k} \\
&= \underbrace{\|y_k - \alpha_k F_{\mathcal{M}}(y_k) - \bar{x}\|^2\, 1_{\tau>k}}_{P_1} + \alpha_k^2\, \|E_k\|^2\, 1_{\tau>k} + 2\alpha_k\, \underbrace{\langle y_k - \alpha_k F_{\mathcal{M}}(y_k) - \bar{x}, E_k\rangle\, 1_{\tau>k}}_{P_2} \\
&+ \alpha_k^2\, \|P_{T_{\mathcal{M}(y_k)}}(\nu_k)\|^2\, 1_{\tau>k} - 2\alpha_k\, \underbrace{\langle y_k - \alpha_k F_{\mathcal{M}}(y_k) + \alpha_k E_k - \bar{x}, P_{T_{\mathcal{M}(y_k)}}(\nu_k)\rangle\, 1_{\tau>k}}_{P_3}. \qquad (7.2.21)
\end{aligned}
$$

Using Lemma 7.2.5, we may bound P_1 as

$$
\begin{aligned}
P_1 &= (\|y_k - \bar{x}\|^2 - 2\alpha_k \langle F_{\mathcal{M}}(y_k), y_k - \bar{x}\rangle + \alpha_k^2\, \|\nabla f_{\mathcal{M}}(y_k)\|^2)\, 1_{\tau>k} \\
&\leq ((1 - \alpha_k\mu)\, \|y_k - \bar{x}\|^2 + C\alpha_k^2)\, 1_{\tau>k}
\end{aligned}
$$

Next, using Proposition 7.2.4 (2(a)i) and Assumption E, we see that $\mathbb{E}_k[\|E_k\|^2\, 1_{\tau>k}]$ and $\mathbb{E}_k[\|P_{T_{\mathcal{M}(y_k)}}(\nu_k)\|^2 1_{\tau>k}]$ are bounded by a numerical constant. It remains to bound P_2 and P_3. Beginning with the former, using Young's inequality, we compute

$$
P_2 \leq \frac{\mu\|y_k - \alpha_k F_{\mathcal{M}}(y_k) - \bar{x}\|^2 1_{\tau>k}}{8} + \frac{2\|E_k\|^2 1_{\tau>k}}{\mu} = \frac{\mu P_1}{8} + \frac{2\|E_k\|^2 1_{\tau>k}}{\mu}.
$$

Next, again using Young's inequality, we bound the conditional expectation of P_3 as follows:

$$
\mathbb{E}_k[P_3] = \alpha_k \mathbb{E}_k[\langle E_k, P_{T_{\mathcal{M}(y_k)}}(\nu_k)\rangle\, 1_{\tau>k}] \leq \frac{\alpha_k \mathbb{E}_k\|E_k\|^2 1_{\tau>k}}{2} + C\alpha_k.
$$

Thus, returning to Lemma 7.2.21 and using Proposition 7.2.4(2(a)ii), we arrive at the estimate:

$$
\mathbb{E}[\|y_{k+1} - \bar{x}\|^2\, 1_{\tau>k}] \leq (1 - \alpha_k\mu/2)\mathbb{E}[\|y_k - \bar{x}\|^2\, 1_{\tau>k}] + C\alpha_k^2.
$$

This completes the proof. □

Next, we can iterate the recursion to ensure a fast rate of convergence of $\|y_k - \bar{x}\|$. As a byproduct, we also obtain estimates on the size of the errors E_k. To simplify notation, we write $\tau_{k_0} := \tau_{k_0,\delta}$, since we will consider several values of k_0. Under these conventions, we have the following Proposition, which will be useful in ensuring summability of certain sequences.

Lemma 7.2.7. *There exists $C > 0$ such that*

1. $\mathbb{E}[\|y_k - \bar{x}\|^2 \, 1_{\tau_{k_0} > k}] \leq C/k^\gamma$ *for all* $k \geq 1$.
2. $\sum_{k=1}^{\infty} \frac{1}{\sqrt{k}} \|y_k - \bar{x}\|^2 < \infty$ *almost surely.*
3. $\frac{1}{\sqrt{n}} \sum_{k=1}^{n} \|y_k - \bar{x}\|^2 \to 0$ *almost surely.*
4. $\sum_{k=1}^{\infty} \frac{1}{\sqrt{k}} \|E_k\| < +\infty$ *almost surely.*
5. $\frac{1}{\sqrt{n}} \sum_{k=1}^{n} \|E_k\| < +\infty$ *almost surely.*

Proof. Part 1 follows immediately from Lemmas 7.2.6 and 7.2.19 by setting $s_k = \mathbb{E}[\|y_k - \bar{x}\|^2 \, 1_{\tau_{k_0} > k}]$. We now prove Part 2. By Part 1, we have

$$\mathbb{E}\left[\sum_{k=1}^{\infty} \frac{1}{\sqrt{k}} \|y_k - \bar{x}\|^2 \, 1_{\tau_{k_0} > k}\right] \leq \sum_{k=1}^{\infty} \frac{C}{k^{\gamma + \frac{1}{2}}} < \infty.$$

Therefore, $\sum_{k=1}^{\infty} \frac{1}{\sqrt{k}} \|y_k - \bar{x}\|^2 \, 1_{\tau_{k_0} > k}$ is finite almost surely. Taking into account that $x_k \to \bar{x}$ almost surely, the sum $\sum_{k=1}^{\infty} \frac{1}{\sqrt{k}} \|y_k - \bar{x}\|^2$ must be finite almost surely. Part 3 now follows immediately follows from Kronecker lemma 7.2.16

Next, we prove Part 4. By Proposition 7.2.4(2(a)iii), we know that the error sequence E_k almost surely satisfies

$$\sum_{k=1}^{\infty} \frac{1}{\sqrt{k}} \|E_k\| 1_{\tau_{k_0} > k} < +\infty.$$

Since $x_k \to \bar{x}$ almost surely, we deduce that almost surely we have $\sum_{k=1}^{\infty} \frac{1}{\sqrt{k}} \|E_k\| < +\infty$, as desired. Part 5 follows from Kronecker lemma 7.2.16. □

7.2.8.3 Completing the proof of Theorem 5.6.1

We now turn to the proof of Theorem 5.6.1. To this end, we introduce an additional sequence

$$z_k := P_{\bar{x} + T_{\mathcal{M}}(\bar{x})}(y_k). \tag{7.2.22}$$

Evidently, for all δ sufficiently small, z_k closely approximates y_k. Indeed, due to the smoothness of $\mathcal{M}$, there exists $C > 0$ such that

$$\|y_k - z_k\| 1_{\tau_{k_0} > k} \leq C\|y_k - \bar{x}\|^2 1_{\tau_{k_0} > k}. \tag{7.2.23}$$

The next result states that it suffices to study the distribution of $\frac{1}{\sqrt{n}} \sum_{k=1}^{n} (z_k - \bar{x})$.

Lemma 7.2.8 (Reduction to an auxiliary sequence)**.** *The equation holds:*

$$\frac{1}{\sqrt{n}} \sum_{k=1}^{n} (x_k - \bar{x}) = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} (z_k - \bar{x}) + o(1).$$

Proof. Note that

$$\frac{1}{\sqrt{n}} \sum_{k=1}^{n} (x_k - \bar{x}) = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} (z_k - \bar{x}) + \frac{1}{\sqrt{n}} \sum_{k=1}^{n} (x_k - y_k) + \frac{1}{\sqrt{n}} \sum_{k=1}^{n} (y_k - z_k).$$

By Lemma 7.2.12 (4), the result will follow once we show that

$$\frac{1}{\sqrt{n}} \sum_{k=1}^{n} (x_k - y_k) \to 0 \qquad \text{and} \qquad \frac{1}{\sqrt{n}} \sum_{k=1}^{n} (y_k - z_k) \to 0,$$

almost surely. To that end, we recall that Proposition 7.2.3(1b) guarantees that almost surely we have

$$\sum_{k=1}^{\infty} \frac{1}{\sqrt{k}} \|x_k - y_k\| \, 1_{\tau_{k_0} > k} < +\infty$$

Since $x_k \to \bar{x}$ almost surely, for almost every sample path, we can find a k_0 such that $\tau_{k_0} = \infty$. Therefore, almost surely we have $\sum_{k=1}^{\infty} \frac{\|x_k - y_k\|}{\sqrt{k}} < \infty$. Applying Kronecker lemma 7.2.16, almost surely we have

$$\frac{1}{\sqrt{n}} \sum_{k=1}^{n} \|x_k - y_k\| \to 0,$$

which implies $\frac{1}{\sqrt{n}} \sum_{k=1}^{n} (x_k - y_k) \to 0$. On the other hand, we have by Lemma 7.2.7 and inequality (7.2.23), that

$$\sum_{k=1}^{\infty} \frac{1}{\sqrt{k}} \|y_k - z_k\| \, 1_{\tau_{k_0} > k} \leq \sum_{k=1}^{\infty} \frac{C}{\sqrt{k}} \|y_k - \bar{x}\|^2 \, 1_{\tau_{k_0} > k} < +\infty$$

Again since for almost every sample path we may find k_0 such that $\tau_{k_0} = \infty$, we have that $\sum_{k=1}^{\infty} \frac{1}{\sqrt{k}} \|y_k - z_k\| < +\infty$, as desired. □

In light of Lemma 7.2.8, it suffices now to study the asymptotic of $\frac{1}{\sqrt{n}} \sum_{k=1}^{n}(z_k - \bar{x})$. This is the content of following lemma. Notice that in the lemmas, we state the asymptotic covariance matrix in a different equivalent form to that appearing in Theorem 5.6.1, and which is more convenient for computation.

Lemma 7.2.9. *The expansion holds:*

$$\frac{1}{\sqrt{n}} \sum_{k=1}^{n}(z_k - \bar{x}) = -\frac{1}{\sqrt{k}} \sum_{i=1}^{k}(U^\top \nabla_{\mathcal{M}} F_{\mathcal{M}}(\bar{x}) U)^{-1} U^\top v_i^{(1)} + o_P(1),$$

and therefore $\frac{1}{\sqrt{n}} \sum_{k=1}^{n}(z_k - \bar{x})$ *converges in distribution to*

$$N\left(0, U(U^\top \nabla F_{\mathcal{M}}(\bar{x}) U)^{-1} U^\top \cdot \Sigma \cdot U(U^\top \nabla F_{\mathcal{M}}(\bar{x}) U)^{-1} U^\top\right),$$

where U *is a matrix whose column vectors form an orthonormal basis of* $T_{\mathcal{M}}(\bar{x})$*.*

Proof. Recall that $UU^\top$ is the orthogonal projection onto $T_{\mathcal{M}}(\bar{x})$. Therefore, we may write $z_k = \bar{x} + UU^\top(y_k - \bar{x})$. Moreover, subtracting $\bar{x}$ from both sides of (7.2.19) and multiplying by $U^\top$, we have

$$\begin{aligned} U^\top(y_{k+1} - \bar{x}) &= U^\top(y_k - \bar{x}) - \alpha_k U^\top F_{\mathcal{M}}(y_k) - \alpha_k U^\top P_{T_{\mathcal{M}(y_k)}}(v_k) + \alpha_k U^\top E_k \\ &= U^\top(y_k - \bar{x}) - \alpha_k U^\top \nabla_{\mathcal{M}} F_{\mathcal{M}}(\bar{x}) UU^\top(y_k - \bar{x}) \\ &\quad - \alpha_k (U^\top F_{\mathcal{M}}(y_k) - U^\top \nabla_{\mathcal{M}} F_{\mathcal{M}}(\bar{x}) UU^\top(y_k - \bar{x})) \\ &\quad - \alpha_k U^\top P_{T_{\mathcal{M}(\bar{x})}}(v_k) - \alpha_k (U^\top P_{T_{\mathcal{M}(y_k)}}(v_k) - U^\top P_{T_{\mathcal{M}(\bar{x})}}(v_k)) + \alpha_k U^\top E_k. \end{aligned}$$

Define $\Delta_k = U^\top(y_k - \bar{x})$, $H = U^\top \nabla_{\mathcal{M}} F_{\mathcal{M}}(\bar{x}) U$, $\zeta_k = U^\top P_{T_{\mathcal{M}(y_k)}}(v_k) - U^\top P_{T_{\mathcal{M}(\bar{x})}}(v_k)$, and

$$R(y) = U^\top F_{\mathcal{M}}(y) - U^\top \nabla_{\mathcal{M}} F_{\mathcal{M}}(\bar{x}) UU^\top(y - \bar{x}).$$

By our assumption, for every vector z the matrix H satisfies

$$\langle Hz, z\rangle = \langle \nabla_{\mathcal{M}} F_{\mathcal{M}}(\bar{x})Uz, Uz\rangle \geq \sigma\|Uz\|^2 = \sigma\|z\|^2.$$

Consequently H is a strongly monotone matrix. Note moreover the equality $U^\top P_{T_{\mathcal{M}}(\bar{x})}(\nu_k) = U^\top UU^\top \nu_k = U^\top \nu_k$. Thus we can rewrite the update of Δ_k as

$$\Delta_{k+1} = \Delta_k - \alpha_k H\Delta_k - \alpha_k U^\top \nu_k - \alpha_k \left(R(y_k) + \zeta_k - U^\top E_k\right).$$

In the remainder of the proof, we study the asymptotics of $\frac{1}{\sqrt{n}}\sum_{k=1}^n \Delta_n$, which readily imply the claimed result using the expression $\frac{1}{\sqrt{n}}\sum_{k=1}^n (z_k - \bar{x}) = \frac{1}{\sqrt{n}}\sum_{k=1}^n U\Delta_k$. We note that our proof closely mirrors [150, Theorem 2]. Define the matrices

$$B_k^n = \alpha_k \sum_{i=k}^n \prod_{j=k+1}^i (I - \alpha_j H) \qquad \text{and} \qquad A_k^n = B_k^n - H^{-1}.$$

Polyak and Juditsky [15, Lemma 2] show that $\bar{\Delta}_n = \frac{1}{n}\sum_{k=1}^n \Delta_k$ satisfies the equality

$$\begin{aligned}\sqrt{n}\bar{\Delta}_n &= \frac{1}{\sqrt{n}}\sum_{k=1}^n H^{-1}U^\top \nu_k \\ &\quad + \frac{1}{\sqrt{n}}\sum_{k=1}^n A_k^n U^\top \nu_k + \frac{1}{\sqrt{n}}\sum_{k=1}^n B_k^n[R(y_k) + \zeta_k - U^\top E_k] + O\left(\frac{1}{\sqrt{n}}\right),\end{aligned}$$

where $\sup_{k,n}\max\{\left\|B_k^n\right\|, \left\|A_k^n\right\|\} < +\infty$ and $\lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^n \left\|A_k^n\right\| = 0$. Equivalently, after expanding $\nu_k = \nu_k^{(1)} + \nu_k^{(2)}(x_k)$, we obtain

$$\begin{aligned}\sqrt{n}\bar{\Delta}_n &= \frac{1}{\sqrt{n}}\sum_{k=1}^n H^{-1}U^\top \nu_k^{(1)} \\ &\quad + \frac{1}{\sqrt{n}}\sum_{k=1}^n A_k^n U^\top \nu_k^{(1)} + \frac{1}{\sqrt{n}}\sum_{k=1}^n B_k^n[R(y_k) + \zeta_k - U^\top E_k + \nu_k^{(2)}(x_k)] + O\left(\frac{1}{\sqrt{n}}\right).\end{aligned}$$

Assumption P ensures that the sum $\frac{1}{\sqrt{n}}\sum_{k=1}^n H^{-1}U^\top \nu_k^{(1)}$ converges in distribution to

$$N\left(0, (U^\top \nabla_{\mathcal{M}} F_{\mathcal{M}}(\bar{x})U)^{-1}U^\top \Sigma U(U^\top \nabla_{\mathcal{M}} F_{\mathcal{M}}(\bar{x})U)^{-1}\right).$$

Thus the theorem will be proved once we show that the other sums in our expression for $\sqrt{n}\bar{\Delta}_n$ converge to 0 almost surely. We do so in the following sequence of claims.

Claim: We have that

$$\frac{1}{\sqrt{n}}\sum_{k=1}^{n} A_k^n U^\top \nu_k^{(1)} \xrightarrow{\text{a.s.}} 0.$$

Proof. Using that $\sup_{k,n}\|A_k^n\| < \infty$ and that $\mathbb{E}\|\nu_k\|^2 1_{\tau_{k_0}>k}$ is bounded, we deduce

$$\begin{aligned}
\mathbb{E}\left[\left\|\frac{1}{\sqrt{n}}\sum_{k=1}^{n} A_k^n U^\top \nu_k^{(1)}\right\|^2 1_{\tau_{k_0}>k}\right] &= \frac{1}{n}\sum_{k=1}^{n}\mathbb{E}\left[\left\|A_k^n U^\top \nu_k^{(1)} 1_{\tau_{k_0}>k}\right\|^2\right] \\
&\le \frac{C}{n}\sum_{k=1}^{n}\|A_k^n\| \\
&\to 0.
\end{aligned}$$

Thus $\frac{1}{\sqrt{n}}\sum_{k=1}^{n} A_k^n U^\top \nu_k 1_{\tau_{k_0}>k}$ is a L^2-bounded martingale. By [85, Theorem 4.4.6], we know that $\frac{1}{\sqrt{n}}\sum_{k=1}^{n} A_k^n U^\top \nu_k 1_{\tau_{k_0}>k} \xrightarrow{L^2} 0$. On the other hand, by [85, Theorem 4.2.11], $\frac{1}{\sqrt{n}}\sum_{k=1}^{n} A_k^n U^\top \nu_k 1_{\tau_{k_0}>k}$ converges almost surely. Therefore, since for almost every sample path there exists k_0 such that $\tau_{k_0} = \infty$, we have $\frac{1}{\sqrt{n}}\sum_{k=1}^{n} A_k^n U^\top \nu_k \xrightarrow{\text{a.s.}} 0$, as desired. □

Claim: We have that

$$\frac{1}{\sqrt{n}}\sum_{k=1}^{n} B_k^n U^\top R(y_k) \xrightarrow{\text{a.s.}} 0.$$

Proof. Let Φ be a smooth extension of $F_{\mathcal{M}}$ to a neighborhood $U \subset \mathbb{R}^d$ of $\bar{x}$. We then deduce

$$\begin{aligned}
R(y) &= U^\top(\Phi(y) - \Phi(\bar{x}) - \nabla\Phi(\bar{x})UU^\top(y-\bar{x})) \\
&= U^\top\nabla\Phi(\bar{x})(I - UU^\top)(y-\bar{x}) + O(\|y-\bar{x}\|^2) \\
&= U^\top\nabla\Phi(\bar{x})P_{N_{\mathcal{M}}(\bar{x})}(y-\bar{x}) + O(\|y-\bar{x}\|^2)
\end{aligned}$$

Since $\mathcal{M}$ is C^2-smooth, it follows immediately that $\|P_{N_{\mathcal{M}}(\bar{x})}(y-\bar{x})\| \le O(\|y-\bar{x}\|^2)$ as $y \in \mathcal{M}$ tends to $\bar{x}$. Thus, we have $\|R(y)\| = O(\|y-\bar{x}\|^2)$. In addition, by our assumption that $x_k \xrightarrow{\text{a.s.}} \bar{x}$, we have $y_k \xrightarrow{\text{a.s.}} \bar{x}$. Consequently, there exists a constant C depending on sample path such that $\|R(y_k)\| \leq C \|y_k - \bar{x}\|^2$ almost surely. Uniform boundedness of B_k^n and Lemma 7.2.7 therefore implies $\frac{1}{\sqrt{n}} \sum_{k=1}^n B_k^n U^\top R(y_k) \xrightarrow{\text{a.s.}} 0$. □

Claim: We have that

$$\frac{1}{\sqrt{n}} \sum_{k=1}^{n} B_k^n \zeta_k \xrightarrow{\text{a.s.}} 0.$$

Proof. For $k \geq 1$, define truncated variables $\zeta_k^{(k_0)} = \zeta_k 1_{\tau_{k_0} > k}$. Note that suffices to show that

$$\frac{1}{\sqrt{n}} \sum_{k=1}^{n} B_k^n \zeta_k^{(k_0)} \xrightarrow{\text{a.s.}} 0,$$

since on every sample path there exists a k_0 such that $\tau_{k_0} = \infty$. Thus, we will work with these truncated variables throughout.

Turning to the proof, we first show that $\frac{1}{\sqrt{n}} \sum_{k=1}^n \zeta_k^{(k_0)} \xrightarrow{P} 0$ and $\frac{1}{\sqrt{n}} \sum_{k=1}^n A_k^n \zeta_k^{(k_0)} \xrightarrow{P} 0$. Recall that $\zeta_k = U^\top P_{T_{\mathcal{M}}(y_k)}(v_k) - U^\top P_{T_{\mathcal{M}}(\bar{x})}(v_k)$, so we have

$$\mathbb{E}\left[\zeta_k^{(k_0)} \mid \mathcal{F}_k\right] = \mathbb{E}\left[\zeta_k \mid \mathcal{F}_k\right] 1_{\tau_{k_0} > k} = 0.$$

Since $x \mapsto P_{T_{\mathcal{M}}(x)}$ is locally Lipschitz on a neighborhood of $\bar{x}$ in $\mathcal{M}$, we have the following bound for some $C > 0$ and all sufficiently small δ:

$$\left\|\zeta_k^{(k_0)}\right\|^2 \leq C \|y_k - \bar{x}\|^2 1_{\tau_{k_0} > k},$$

In particular, it holds that

$$\mathbb{E}\left[\left\|\zeta_k^{(k_0)}\right\|^2 \mid \mathcal{F}_k\right] \leq C^2 \|y_k - \bar{x}\|^2 1_{\tau_{k_0} > k}.$$

Combining with Lemma 7.2.7(1), we know that $\zeta_k^{(k_0)}$ is a martingale difference sequence and almost surely,

$$\sum_{k=1}^{\infty} \frac{1}{k} \mathbb{E}\left[\left\|\zeta_k^{(k_0)}\right\|^2 \mid \mathcal{F}_k\right] \leq C^2 \sum_{k=1}^{\infty} \frac{1}{k} \|y_k - \bar{x}\|^2 < \infty.$$

Therefore, by Lemma 7.2.15, we have

$$\frac{1}{\sqrt{n}} \sum_{k=1}^{n} \zeta_k^{(k_0)} \xrightarrow{\text{a.s.}} 0.$$

In particular, it holds that $\frac{1}{\sqrt{n}} \sum_{k=1}^{n} \zeta_k^{(k_0)} \xrightarrow{P} 0$.

Next we show that for any $k_0 < \infty$, we have $n^{-1/2} \sum_{k=1}^{n} A_k^n \zeta_k^{(k_0)} \xrightarrow{P} 0$. To see this, note that by Lemma 7.2.7, there exists $C' > 0$ such that

$$\mathbb{E}\left[\left\|\zeta_k^{(k_0)}\right\|^2\right] \leq C\mathbb{E}\left[\|y_k - \bar{x}\|^2 \, 1_{\tau_{k_0}>k}\right] \leq C'\alpha_k. \tag{7.2.24}$$

Hence, the following limit holds

$$\mathbb{E}\left[\left\|\frac{1}{\sqrt{n}} \sum_{k=1}^{n} A_k^n \zeta_k^{(k_0)}\right\|^2\right] = \frac{1}{n} \sum_{k=1}^{n} \mathbb{E}\left[\left\|A_k^n \zeta_k^{(k_0)}\right\|^2\right] \leq \frac{C'\alpha_k}{n} \sum_{k=1}^{n} \|A_k^n\|^2 \leq \frac{C'\alpha_k \sup_{k,n} \|A_k^n\|}{n} \sum_{k=1}^{\infty} \|A_k^n\| \to 0,$$

where the first equality follows from the martingale difference property and the second inequality follows from the boundedness of moments of $\zeta_k^{(k_0)}$. Consequently, we have shown that

$$\frac{1}{\sqrt{n}} \sum_{k=1}^{n} A_k^n \zeta_k^{(k_0)} \xrightarrow{L^2} 0,$$

which implies that $\frac{1}{\sqrt{n}} \sum_{k=1}^{n} A_k^n \zeta_k^{(k_0)} \xrightarrow{P} 0$.

We have therefore proved that $\frac{1}{\sqrt{n}} \sum_{k=1}^{n} B_k^n \zeta_k^{(k_0)} \xrightarrow{P} 0$. We now show that $\frac{1}{\sqrt{n}} \sum_{k=1}^{n} B_k^n \zeta_k^{(k_0)}$ converges almost surely. Since the almost sure limits and limits in probability agree when both exist, this will complete the proof.

To this end, define the sequence

$$Z_{n,k_0} = \sum_{k=1}^{n} B_k^n \zeta_k^{(k_0)}.$$

The result follows if we can prove that for any finite k_0, the sequence $n^{-1/2} Z_{n,k_0}$ almost surely converges. To that end, note that $B_k^{n+1} - B_k^n = \alpha_k \prod_{i=k+1}^{n} (I - \alpha_i H)$. Thus, defining

$$W_k^n = \prod_{i=k}^{n} (I - \alpha_i H), \qquad V_{n,k_0} = \sum_{k=1}^{n} \alpha_k W_{k+1}^{n+1} \zeta_k^{(k_0)},$$

we deduce that V_{n,k_0} is $\mathcal{F}_{n+1}$ measurable and Z_{n,k_0} admits the decomposition:

$$Z_{n,k_0} = Z_{n-1,k_0} + V_{n-1,k_0} + \alpha_n \zeta_n^{(k_0)} = \sum_{k=1}^{n-1} V_{k,k_0} + \sum_{k=1}^{n} \alpha_k \zeta_k^{(k_0)}.$$

Note that the sum $\sum_{k=1}^{n} \alpha_k \zeta_k^{(k_0)}$ is a square-integrable martingale with summable squared increments, so it converges almost surely [85, Theorem 4.2.11]. As a result, we have the following limit $n^{-1/2} \sum_{k=1}^{n} \alpha_k \zeta_k^{(k_0)} \xrightarrow{\text{a.s.}} 0$. It thus suffices to show that $n^{-1/2} \sum_{k=1}^{n-1} V_{k,k_0}$ converges almost surely.

To that end, let λ denote the smallest eigenvalue of H. Then we have

$$\mathbb{E}\left[\|V_{n,k_0}\|^2\right] = \sum_{k=1}^{n} \alpha_k^2 \left\|W_{k+1}^{n+1}\right\|^2 \mathbb{E}\left[\left\|\zeta_k^{(k_0)}\right\|^2\right] \leq C' \sum_{k=1}^{n} \alpha_k^3 \left\|W_{k+1}^{n+1}\right\|^2, \tag{7.2.25}$$

where the inequality follows from the bound $\mathbb{E}\left[\left\|\zeta_k^{(k_0)}\right\|^2\right] \leq C'\alpha_k$ (see Equation (7.2.24)). The result [, Lemma 1 (part 3)] shows that there exist constants $\beta > 0$ and $K < \infty$ such that for all k and $t \geq k$, the estimate holds:

$$\left\|W_{k+1}^{n+1}\right\|^2 \leq K \exp\left(-\beta \sum_{i=k+1}^{n} \alpha_i\right).$$

Plugging this estimate into (7.2.25), exactly the same proof as that of [150, Lemma A.7] with $\rho = 3$ shows that there exists some constant C such that

$$\mathbb{E}\left[\|V_{n,k_0}\|^2\right] \leq \frac{C \log n}{n^{2\gamma}}.$$

Hence, for any $\epsilon > 0$, we can find some C such that

$$\mathbb{E}\left[\|V_{n,k_0}\|^2\right] \leq \frac{C}{n^{2\gamma-\epsilon}}.$$

Now define $T_{n,k_0} = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} V_{k,k_0}$. We claim that T_{n,k_0} almost surely has finite length. Indeed, for any $\epsilon > 0$ there exists $C, C' > 0$ such that

$$\mathbb{E}\left[\|T_{n,k_0} - T_{n+1,k_0}\|\right] \leq \left|\frac{1}{\sqrt{n+1}} - \frac{1}{\sqrt{n}}\right| \sum_{k=1}^{n} \mathbb{E}\left[\|V_{k,k_0}\|\right] + \frac{1}{\sqrt{n+1}} \mathbb{E}\left[\|V_{n+1,k_0}\|\right]$$

$$\leq \frac{C}{n^{\frac{3}{2}}} \sum_{k=1}^{n} \frac{1}{k^{\gamma-\epsilon}} + \frac{1}{\sqrt{n}} \frac{1}{n^{\gamma-\epsilon}}$$
$$\leq \frac{C'}{n^{\gamma+1/2-\epsilon}}.$$

Since $\gamma \in (\frac{1}{2}, 1)$, we therefore have $\sum_n \mathbb{E}[\|T_{n,k_0} - T_{n+1,k_0}\|] < \infty$. Consequently, the sum is finite almost surely: $\sum_n \|T_{n,k_0} - T_{n+1,k_0}\| < +\infty$. This implies that $T_{n,k_0} = n^{-1/2} \sum_{k=1}^n V_{k,k_0}$ converges almost surely. Recalling the definition of V_{k,k_0}, we find that $n^{-1/2} Z_{n,k_0}$ almost surely converges, which completes the proof. □

Claim: We have that

$$\frac{1}{\sqrt{n}} \sum_{k=1}^{n} B_k^n \nu_k^{(2)}(x_k) \xrightarrow{\text{a.s.}} 0.$$

Proof. This may be proved by argument that mirrors Claim 7.2.8.3. Indeed, observe that the sequence $\xi_k = \nu_k^{(2)}(x_k) 1_{\tau > k_0}$ is a martingale difference sequence, the bounds hold for some $C > 0$

$$\mathbb{E}_k[\|\xi_k\|^2] \leq C\|x_k - \bar{x}\|^2 1_{\tau > k_0} \qquad \text{and} \qquad \mathbb{E}[\|\xi_k\|^2] \leq C\alpha_k,$$

and $\sum_{k=1}^\infty \frac{1}{k} \mathbb{E}_k[\|\xi_k\|^2] \leq \sum_{k=1}^\infty \frac{1}{k} \|x_k - \bar{x}\|^2 1_{\tau > k_0} < +\infty$. Only these facts for $\zeta_k^{(k_0)}$ were used to prove Claim 7.2.8.3. □

Claim: We have that

$$\frac{1}{\sqrt{n}} \sum_{k=1}^{n} B_k^n U^\top E_k \xrightarrow{\text{a.s.}} 0.$$

Proof. This follows immediately from Lemma 7.2.7(5 and the fact that $\sup_{k,n} \left\|B_k^n\right\| < \infty$.

□

Taking these claims into account, the proof is complete. □

7.2.9 Proof of Lemma 5.7.1

We first consider the normalizing constant $C(u) = 1 + \int h(u^\top g(z))\,d\mathcal{P}(z)$. The dominated convergence theorem[1] implies that $C(\cdot)$ is twice differentiable with $\nabla C(u) = \int h'(u^\top g(z))g(z)^\top\,d\mathcal{P}(z)$ and $\nabla^2 C(u) = \int h''(u^\top g(z))g(z)g(z)^\top\,d\mathcal{P}(z)$. Moreover, the dominated convergence theorem[2] implies that $A(x)$ is C^1-smooth with $\nabla A(x) = \mathbb{E}_{z\sim\mathcal{P}}\nabla A(x,z)$. Thus it now suffices to argue that $\hat{L}(x,u) := \int h(u^\top g(z))A(x,z)\,d\mathcal{P}(z)$ is C^1-smooth. An application of the dominated convergence theorem in u directly implies that $\hat{L}(x,u)$ is differentiable in u with $\nabla_u\hat{L}(x,u) = \int h'(u^\top g(z))A(x,z)g(z)^\top\,d\mathcal{P}(z)$ and moreover $\nabla_u\hat{L}(x,u)$ is continuous in (x,u). Similarly, the dominated convergence theorem[3] implies that $\hat{L}(x,u)$ is differentiable in x with $\nabla_x\hat{L}(x,u) = \int h(u^\top g(z))\nabla A(x,z)\,d\mathcal{P}(z)$ and $\nabla_x\hat{L}(x,u)$ is continuous in (x,u). Thus $L(\cdot,\cdot)$ is C^1-smooth near $(\bar{x},u)$. Observe that the expression $\nabla_x L(\bar{x},0) = \nabla A(\bar{x})$ follows trivially since $L(x,0) \equiv A(x)$ for all x. To see the expression for $\nabla_u L(x,0)$, observe that $\nabla C(0) = 0$ and $\nabla^2 C(0) = 0$ and therefore $C(u) = 1 + o(\|u\|^2)$. It follows immediately that $\nabla_u L(x,0) = \nabla_u\hat{L}(x,0)$ for all x, thereby completing the proof.

7.2.10 Proof of Lemma 5.7.2

Assumption G ensures that the map $A+H$ is C^1 invertible around $(\bar{x},0)$ with some inverse $\sigma(\cdot)$. Define now the linearization $\Psi(x) := A(\bar{x}) + \nabla A(\bar{x})(x-\bar{x})$ of A at $\bar{x}$. Invoking [90, Theorem 2B.10], we deduce that the map $\Psi + H$ is also C^1 invertible around $(0,\bar{x})$ with inverse $\hat{\sigma}$ and which satisfies $\nabla\hat{\sigma}(0) = \nabla\sigma(0)$. Note that in light of Lemma 5.7.1, we may equivalently write Ψ as $\Psi(x) := L(\bar{x},0) + \nabla_x L(\bar{x},0)(x-\bar{x})$. Applying [90, Theorem

[1] using that h' and h'' are bounded and $\mathbb{E}_{z\sim\mathcal{P}}\|g(z)\|^2 < \infty$

[2] using that there is a neighborhood U of $\bar{x}$ such that $\sup_{x\in U}\|\nabla A(x,z)\|^2$ is integrable.

[3] using that h is bounded and there is a neighborhood U of $\bar{x}$ such that $\sup_{x\in U}\|\nabla A(x,z)\|^2$ is integrable.

2D.6], we deduce that the map $S(u)$ admits a single-valued localization $s(\cdot)$ around $(0,\bar{x})$ that is differentiable at 0 and satisfies $\nabla s(0) = -\nabla\hat{\sigma}(0) \circ \nabla_u L(\bar{x}, 0)$. An application of Lemma 5.7.1 completes the proof.

7.2.11 Proof of Theorem 5.7.3

The proof of Theorem 5.7.3 will be based on the local minimax theorem of Hájek and Le Cam [17, Theorem 6.6.2], which is summarized in Section 7.2.15. We aim to apply Theorem 7.2.22 as follows. For each u, we take $(\Omega_k, \mathcal{F}_k, Q_{k,u})$ to be the k-fold product of the probability spaces $(\Omega, \mathcal{F}, \mathcal{P}_{u/\sqrt{k}})$ and set $\Gamma_k(u) = s(u/\sqrt{k})$. It was shown in [16, Lemma 8.3] that the the sequence $\{\Omega_k, \mathcal{F}_k, Q_{k,u}\}_{u\in\mathbb{R}^d}$ is locally asymptotically normal with precision $V = \mathbb{E}_{z\sim\mathcal{P}}[g(z)g(z)^\top]$. Moreover, the following lemma establishes regularity of the sequence Γ_k.

Lemma 7.2.10. *The sequence $\Gamma_k \colon \mathbb{R}^d \to \mathbb{R}^d$ is regular at zero with derivative $\dot{\Gamma} := -\nabla\sigma(0)\cdot \mathbb{E}_{z\sim\mathcal{P}}[g(z)A(\bar{x},z)^\top]$.*

Proof. Using Lemma 5.7.2, a first-order expansion of $s(\cdot)$ around $\bar{x}$ yields

$$\sqrt{k}(\Gamma_k(u) - \Gamma_k(0)) = \sqrt{k}(s(u/\sqrt{k}) - \bar{x}) = -\nabla\sigma(0)\cdot \mathop{\mathbb{E}}_{z\sim\mathcal{P}}[g(z)A(\bar{x},z)^\top]u + \frac{o(k^{-1/2})}{k^{-1/2}}.$$

Letting k tends to infinity completes the proof. □

We now apply Theorem 7.2.22. Let $\mathcal{L} \colon \mathbb{R}^d \to [0,\infty)$ be symmetric, quasiconvex, and lower semicontinuous, and $\widehat{x}_k \colon \mathcal{Z}^k \to \mathbb{R}^d$ be a sequence of estimators. Set

$$g(z) := A(\bar{x},z) - A(\bar{x}),$$

$$\Sigma := \mathop{\mathbb{E}}_{z\sim\mathcal{P}}[g(z)g(z)^\top],$$

$$K := \nabla\sigma(0).$$

Applying Theorem 7.2.22 yields

$$\sup_{\mathcal{I}\subset\mathbb{R}^d,\,|\mathcal{I}|<\infty} \liminf_{k\to\infty} \max_{u\in\mathcal{I}} \mathbb{E}_{P^k_{u/\sqrt{k}}}[\mathcal{L}(\sqrt{k}(\widehat{x_k} - \bar{x}_{u/\sqrt{k}}))] \geq \mathbb{E}[\mathcal{L}(Z_\lambda)] \tag{7.2.26}$$

where $Z_\lambda \sim \mathsf{N}(0, K\Sigma(\Sigma + \lambda I)^{-1}\Sigma^\top K^\top)$ for any $\lambda > 0$. Basic linear algebra shows

$$\lim_{\lambda\downarrow 0} \Sigma(\Sigma + \lambda I)^{-1}\Sigma = \Sigma.$$

A straightforward argument based on the monotone convergence theorem (see e.g. [91, Section 5.1.2]) therefore implies that the right side of (7.2.26) tends to $\mathbb{E}[\mathcal{L}(Z)]$ as $\lambda \downarrow 0$, where $Z \sim \mathsf{N}(0, W\Sigma W^\top)$. The proof is complete.

7.2.12 Proof of Theorems 5.7.4, 5.7.5, and 5.7.6

The proof is the same for all three theorems. Namely, we aim to apply Theorem 7.2.23. To this end, set $Q_{k,u} = P^k_{u/\sqrt{k}}$ and $\Gamma_k(u) = s(u/\sqrt{k})$. Set $Z_k := -\frac{1}{\sqrt{k}}\sum_{i=1}^k g(z_i)$. It is shown in [16, Lemma 8.3] that the the following expansion holds:

$$\log\frac{dQ_{k,u}}{dQ_{k,0}} = u^\top Z_k - \frac{1}{2}u^\top V u + o_{Q_{k,0}}(1) \tag{7.2.27}$$

with $V := \mathbb{E}_{\mathcal{P}} g(z)g(z)^\top = \mathrm{Cov}(A(\bar{x}, z))$. Each of Theorem 5.4.1, Theorem 5.6.1, and Corollary 5.6.2 yields the expansion

$$\sqrt{k}(x_k - \bar{x}) = \underbrace{-\nabla\sigma(0)}_{=:W} Z_k + o_{Q_{k,0}}(1),$$

Note that by Lemma 5.7.2, Γ_k is regular at zero with derivative

$$\dot{\Gamma} = \nabla\sigma(0) = -\nabla\sigma(0)\cdot \mathop{\mathbb{E}}_{z\sim\mathcal{P}}[A(\bar{x}, z)g(z)^\top] = -\nabla\sigma(0)\cdot\mathrm{Cov}(A(\bar{x}, z)).$$

Finally, observe the equalities

$$\dot{\Gamma} = WV \qquad \text{and} \qquad WVW^\top = \nabla\sigma(0)\cdot\mathrm{Cov}(A(\bar{x}, z))\nabla\sigma(0)^\top.$$

Theorem 7.2.23 thus ensures that

$$\sqrt{k}(x_k - \Gamma_k(u)) \overset{u}{\rightsquigarrow} \mathsf{N}(0, \nabla\sigma(0) \cdot \mathrm{Cov}(A(\bar{x}, z))\nabla\sigma(0)^\top). \tag{7.2.28}$$

Let $\mathcal{L} \colon \mathbb{R}^d \to \mathbb{R}$ be any bounded continuous function and $Z \sim \mathsf{N}(0, \nabla\sigma(0) \cdot \mathrm{Cov}(A(\bar{x}, z))\nabla\sigma(0)^\top)$. Then (7.2.31) directly implies that for every finite subset $\mathcal{I} \subset \mathbb{R}^d$, we have

$$\lim_{k\to\infty} \max_{u\in\mathcal{I}} \mathbb{E}_{Q_{k,u}}[\mathcal{L}(\sqrt{k}(T_k - \Gamma_k(u)))] = \max_{u\in\mathcal{I}} \lim_{k\to\infty} \mathbb{E}_{Q_{k,u}}[\mathcal{L}(\sqrt{k}(T_k - \Gamma_k(u)))] = \mathbb{E}[\mathcal{L}(Z)].$$

Hence

$$\sup_{\mathcal{I}\subset\mathbb{R}^d, |\mathcal{I}|<\infty} \liminf_{k\to\infty} \max_{u\in\mathcal{I}} \mathbb{E}_{Q_{k,u}}[\varphi(\sqrt{k}(T_k - \Gamma_k(u)))] = \mathbb{E}[\mathcal{L}(Z)],$$

thereby demonstrating equality in (5.7.3) whenever $\mathcal{L}$ is bounded and continuous.

7.2.13 Proofs of Theorems 5.7.5 and 5.7.5

The proofs of these two theorems are identical to the proof of Theorem 5.7.4.

7.2.14 Auxiliary facts about sequences of random variables.

Definition 7.2.11. Let $\{X_k\}_{k\geq 1}$ and X be random vectors in $\mathbb{R}^d$ defined on a probability space $(\Omega, \mathcal{F}, P)$.

1. *X_k converges almost surely to X*, denoted $X_k \xrightarrow{a.s.} X$ if for almost every $\omega \in \Omega$, the vector $X_k(\omega)$ converges to $X(\omega)$.
2. *X_k converges in probability to X*, denoted $X_k \xrightarrow{p} X$, if for every $\epsilon > 0$, we have $\lim_{k\to\infty} P(\|X_k - X\| \leq \epsilon) \to 1$.

3. *X_k converges in distribution to X*, denoted $X_k \xrightarrow{D} X$ if for every bounded continuous function $G\colon \mathbb{R}^d \to \mathbb{R}$, one has $\lim_{k\to\infty} \mathbb{E}G(x_k) = \mathbb{E}G(X)$.

4. X_k is *bounded in probability*, denoted $X_k = O_p(1)$, if for every $\epsilon > 0$, there exist M_ϵ such that $P(\|X_n\| > M_\epsilon) < \epsilon$ for all sufficiently large indices k.

Lemma 7.2.12. *Let $\{X_k\}_{k\geq 1}$, $\{Y_k\}_{k\geq 1}$, X, and Y be random vectors in some Euclidean space and let a, b be deterministic. The following statements are true.*

1. *The implications hold:*

$$X_k \xrightarrow{a.s.} X \qquad \Longrightarrow \qquad X_k \xrightarrow{p} X \qquad \Longrightarrow \qquad X_k \xrightarrow{D} X \qquad \Longrightarrow \qquad X_k = O_p(1).$$

2. *If $X_k \xrightarrow{p} X$ and $Y_k \xrightarrow{p} Y$, then $aX_k + bY_k \xrightarrow{p} aX + bY$ and $X_kY_k \xrightarrow{p} XY$. The analogous statement holds for almost sure convergence.*

3. *If $X_n = O_p(1)$ and $Y_k \xrightarrow{p} 0$, then $X_kY_k \xrightarrow{p} 0$.*

4. *(Slutsky I) If $X_k \xrightarrow{D} X$ and $Y_k \xrightarrow{p} a$, then $X_k + Y_k \xrightarrow{D} X + a$ and $X_kY_K \xrightarrow{D} aX$.*

5. *(Slutsky II) If $X_k \xrightarrow{p} X$ and $Y_k \xrightarrow{p} Y$, then $X_k + Y_k \xrightarrow{p} X + Y$ and $X_kY_K \xrightarrow{p} XY$.*

6. *If $X_kY_k \xrightarrow{D} X$ and $Y_k \xrightarrow{p} c$, then $X_k \xrightarrow{D} X/c$, as long as $c \neq 0$.*

7. *(Delta Method) Suppose that $\sqrt{k}(X_k - \mu) \xrightarrow{d} \mathsf{N}(0, \Sigma)$ for some $\mu \in \mathbb{R}^d$ and some matrix $\Sigma \in \mathbb{R}^{d\times d}$, then we have $\sqrt{k}(g(X_k) - g(\mu)) \xrightarrow{d} \mathsf{N}(0, \nabla g(\mu)\Sigma\nabla g(\mu)^\top)$ for any map $g\colon \mathbb{R}^d \to \mathbb{R}^m$ that is differentiable at μ.*

Lemma 7.2.13 (Robbins-Siegmund [146])**.** *Let $A_k, B_k, C_k, D_k \geq 0$ be non-negative random variables adapted to the filtration $\{\mathcal{F}_k\}$ and satisfying*

$$\mathbb{E}[A_{k+1} \mid \mathcal{F}_k] \leq (1 + B_k)A_k + C_k - D_k.$$

Then on the event $\{\sum_k B_k < \infty, \sum_k C_k < \infty\}$, there is a random variable $A_\infty < \infty$ such that $A_k \xrightarrow{a.s.} A_\infty$ and $\sum_k D_k < \infty$ almost surely.

Lemma 7.2.14 (Conditional Borel-Cantelli [147]). *Let $\{X_n : n \geq 1\}$ be a sequence of nonnegative random variables defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and $\{\mathcal{F}_n : n \geq 0\}$ be a sequence of sub-σ-algebras of $\mathcal{F}$. Let $M_n = \mathbb{E}[X_n \mid \mathcal{F}_{n-1}]$ for $n \geq 1$. If $\{\mathcal{F}_n : n \geq 0\}$ is nondecreasing, i.e., it is a filtration, then $\sum_{n=1}^{\infty} X_n < \infty$ almost surely on $\{\sum_{n=1}^{\infty} M_n < \infty\}$.*

Lemma 7.2.15 ([148, Exercise 5.3.35]). *Let M_k be an L^2 martingale adapted to a filtration $\{\mathcal{F}_k\}$ and let $b_k \uparrow \infty$ be a positive deterministic sequence. Then if*

$$\sum_{k \geq 1} b_k^{-2} \mathbb{E}\left[(M_k - M_{k-1})^2 \mid \mathcal{F}_{k-1}\right] < +\infty,$$

we have $b_n^{-1} M_n \xrightarrow{a.s.} 0$.

Lemma 7.2.16 (Kronecker Lemma). *Suppose $\{x_k\}_k$ is an infinite sequence of real number such that the sum $\sum_{k=1}^{\infty} x_k$ exists and is finite. Then for any divergent positive nondecreasing sequence $\{b_k\}$, we have*

$$\lim_{K \to \infty} \frac{1}{b_K} \sum_{k=1}^{K} b_k x_k = 0.$$

The proofs of the following three lemmas may be found in Section 7.1.7.2.

Lemma 7.2.17. *Fix $k_0 \in \mathbb{N}, c > 0$, and $\gamma \in (1/2, 1]$. Suppose that $\{X_k\}, \{Y_k\}$, and $\{Z_k\}$ are nonnegative random variables adapted to a filtration $\{\mathcal{F}_k\}$. Suppose the relationship holds:*

$$\mathbb{E}[X_{k+1} \mid \mathcal{F}_k] \leq (1 - ck^{-\gamma}) X_k - Y_k + Z_k \qquad \text{for all } k \geq k_0.$$

Assume furthermore that $c \geq 6$ if $\gamma = 1$. Define the constants $a_k := \frac{k^{2\gamma-1}}{\log^2(k+1)}$. Then there exists a random variable $V < \infty$ such that on the event $\{\sum_{k=1}^{\infty} a_{k+1} Z_k < +\infty\}$, the following is true:

1. *The limit holds*

$$a_k X_k \xrightarrow{a.s.} V.$$

2. *The sum is finite*

$$\sum_{k=1}^{\infty} a_{k+1} Y_k < +\infty.$$

Lemma 7.2.18. *Fix $k_0 \in \mathbb{N}$, $c, C > 0$, and $\gamma \in (1/2, 1]$. Suppose that $\{s_k\}_k$ is a nonnegative sequence satisfying*

$$s_k \leq \frac{c}{12\gamma} \qquad \text{and} \qquad s_{k+1}^2 \leq s_k^2 - ck^{-\gamma} s_k + Ck^{-2\gamma}, \qquad \text{for all } k \geq k_0,$$

Then, there exists a constant C_{ub} depending only on c, C, γ and k_0 such that

$$s_k \leq C_{\mathsf{ub}} k^{-\gamma}, \qquad \forall k \geq 1.$$

Lemma 7.2.19. *Fix $k_0 \in \mathbb{N}$, $c, C > 0$, and $\gamma \in (1/2, 1]$. Suppose that $\{s_k\}_k$ is a nonnegative sequence satisfying*

$$s_{k+1} \leq (1 - ck^{-\gamma}) s_k + Ck^{-2\gamma}, \qquad \text{for all } k \geq k_0,$$

Assume furthermore that $c \geq 16$ if $\gamma = 1$. Then, there exists a constant C_{ub} depending only on c, C, γ and k_0 such that

$$s_k \leq C_{\mathsf{ub}} k^{-\gamma}, \qquad \forall k \geq 1.$$

7.2.15 Background on local asymptotic minimax

In this section, we review mostly standard results in asymptotic statistics, primarily focusing on Hájek-Le Cam minimax theorem. We begin with several standard definitions, following the classical text [151]. Henceforth, we fix a sequence of parametric statistical models $\{Q_{k,u} \mid u \in \mathbb{R}^d\}$, where $Q_{k,u}$ is a probability measure on $(\Omega_k, \mathbf{S}_k)$ such that

$Q_{k,u} \ll Q_{k,0}$ for each $k \in \mathbf{N}$ and $u \in \mathbb{R}^d$. We write either $X_k \overset{u}{\rightsquigarrow} X$ or $X_k \overset{u}{\rightsquigarrow} \mathsf{D}$ to indicate that a sequence of random vectors $X_k \colon \Omega_k \to \mathbb{R}^m$ converges in distribution to a random vector $X \sim \mathsf{D}$ with respect to $Q_{k,u}$, i.e., $\lim_{k\to\infty} \mathbb{E}_{Q_{k,u}}[\varphi(X_k)] = \mathbb{E}_{X\sim\mathsf{D}}[\varphi(X)]$ for every bounded continuous function $\varphi \colon \mathbb{R}^m \to \mathbb{R}$. Notice that the limiting distribution D must not depend on u.

Definition 7.2.20 (Local asymptotic normality)**.** *The sequence* $\{Q_{k,u} \mid u \in \mathbb{R}^d\}$ *is* locally asymptotically normal (LAN) with precision V at zero *if there exist a sequence of random vectors* $Z_k \colon \Omega_k \to \mathbb{R}^d$ *and a positive semidefinite matrix* $V \in \mathbb{R}^{d\times d}$ *such that* $Z_k \overset{0}{\rightsquigarrow} \mathsf{N}(0, V)$ *and, for each* $u \in \mathbb{R}^d$*,*

$$\log \frac{dQ_{k,u}}{dQ_{k,0}} = u^\top Z_k - \frac{1}{2} u^\top V u + o_{Q_{k,0}}(1). \tag{7.2.29}$$

Definition 7.2.21 (Regular mapping sequence)**.** *A sequence of mappings* $\Gamma_k \colon \mathbb{R}^d \to \mathbb{R}^n$ *is* regular with derivative $\dot\Gamma$ at zero *if there exists a matrix* $\dot\Gamma \in \mathbb{R}^{n\times d}$ *satisfying*

$$\lim_{k\to\infty} \sqrt{k}(\Gamma_k(u) - \Gamma_k(0)) = \dot\Gamma u \quad \textit{for all } u \in \mathbb{R}^d.$$

Note that given any function $\psi \colon \mathbb{R}^d \to \mathbb{R}^n$ that is differentiable at zero, the induced mapping sequence $\Gamma_k \colon \mathbb{R}^d \to \mathbb{R}^n$ given by $\Gamma_k(u) = \psi(u/\sqrt{k})$ is clearly regular with derivative $\dot\Gamma = \nabla\psi(0)$ at zero. This will be the primary example of a regular mapping sequence.

Equipped with the preceding definitions, we are ready to state the following version of the Hájek-Le Cam minimax theorem [151, Theorem 3.11.5].

Theorem 7.2.22 (Local asymptotic minimax bound)**.** *Let* $\{Q_{k,u} \mid u \in \mathbb{R}^d\}$ *be locally asymptotically normal with precision* V *at zero,* $\Gamma_k \colon \mathbb{R}^d \to \mathbb{R}^n$ *be a regular mapping sequence with derivative* $\dot\Gamma$ *at zero, and* $\mathcal{L} \colon \mathbb{R}^n \to [0, \infty)$ *be symmetric, quasiconvex, and lower semicontinuous function. Then, for any sequence of estimators* $T_k \colon \Omega_k \to \mathbb{R}^n$*,* *we have*

$$\sup_{\mathcal{I}\subset\mathbb{R}^d,\,|\mathcal{I}|<\infty} \liminf_{k\to\infty} \max_{u\in\mathcal{I}} \mathbb{E}_{Q_{k,u}}[\mathcal{L}(\sqrt{k}(T_k - \Gamma_k(u)))] \geq \mathbb{E}[\mathcal{L}(Z)], \tag{7.2.30}$$

where $Z \sim \mathsf{N}(0, \dot{\Gamma}(V + \lambda I)^{-1}\dot{\Gamma}^\top)$ *for any* $\lambda > 0$*; if* V *is invertible, then* (7.2.30) *also holds with* $Z \sim \mathsf{N}(0, \dot{\Gamma}V^{-1}\dot{\Gamma}^\top)$.

Next, we'll need the following lemma that provides sufficient conditions for establishing a kind of uniform asymptotic normality of a statistical estimator. This is a small modification of [18, Lemma 8.14]. For the sake of completeness, we repeat here a short proof as it appears in [91, Proof of Lemma 5.15].

Theorem 7.2.23 (Asymptotic equivariance)**.** *Fix a sequence of estimators* $T_k\colon \Omega_k \to \mathbb{R}^n$*, a sequence of parametric statistical models* $\{Q_{k,u} \mid u \in \mathbb{R}^d\}$*, and a regular mapping sequence* Γ_k *with derivative* $\dot{\Gamma}$ *at zero. Suppose that there exists a sequence of random vectors* $Z_k\colon \Omega_k \to \mathbb{R}^d$ *with* $Z_k \overset{0}{\rightsquigarrow} \mathsf{N}(0, V)$*, a positive semidefinite matrix* $V \in \mathbb{R}^{d\times d}$*, a matrix* W*, and a vector* $x^\star \in \mathbb{R}^n$ *such that the following expansions hold:*

$$(\textbf{LAN}) \qquad \log\frac{dQ_{k,u}}{dQ_{k,0}} = u^\top Z_k - \frac{1}{2}u^\top Vu + o_{Q_{k,0}}(1) \qquad \forall u \in \mathbb{R}^d,$$

$$(\textbf{Normality for } Q_{k,0}) \qquad \sqrt{k}(T_k - \Gamma_k(0)) = WZ_k + o_{Q_{k,0}}(1).$$

Then T_k *are asymptotically equivariant-in-law with respect to* $\{Q_{k,u} \mid u \in \mathbb{R}^d\}$ *for estimating* $x^\star$*, that is*

$$\sqrt{k}(T_k - \Gamma_k(u)) \overset{u}{\rightsquigarrow} \mathsf{N}((WV - \dot{\Gamma})u, WVW^\top) \qquad \forall u \in \mathbb{R}^d. \tag{7.2.31}$$

Proof. Let $\bar{Z} \sim \mathsf{N}(0, \Sigma)$, fix $u \in \mathbb{R}^d$, and consider the affine map

$$\varphi(z) := \begin{pmatrix} W \\ u^\top \end{pmatrix} z + \begin{pmatrix} 0 \\ -\frac{1}{2}u^\top Vu \end{pmatrix}.$$

Then clearly $\varphi(Z_k) \overset{0}{\rightsquigarrow} \varphi(\bar{Z})$ and hence the continuous mapping theorem [18, see Theorems 2.3 and 2.7] implies

$$\begin{pmatrix} \sqrt{k}(T_k - \Gamma_k(0)) \\ \log \frac{dQ_{k,u}}{dQ_{k,0}} \end{pmatrix} \overset{0}{\rightsquigarrow} \begin{pmatrix} W\bar{Z} \\ u^\top \bar{Z} - \frac{1}{2}u^\top Vu \end{pmatrix} \sim \mathsf{N}\left(\begin{pmatrix} 0 \\ -\frac{1}{2}u^\top Vu \end{pmatrix}, \begin{pmatrix} WVW^{-\top} & WVu \\ u^\top VW^{-\top} & u^\top Vu \end{pmatrix} \right).$$

Applying Le Cam's Third Lemma [18, Example 6.7], we thus conclude

$$\sqrt{k}(T_k - \Gamma_k(0)) \overset{u}{\rightsquigarrow} \mathsf{N}(WVu, WVW^\top). \tag{7.2.32}$$

On the other hand, taking into account that Γ_k is a regular mapping sequence with derivative $\dot{\Gamma}$ at zero, we deduce $\sqrt{k}(\Gamma_k(u) - \Gamma_k(0)) \to \dot{\Gamma}u$ as $k \to \infty$. Combining this with (7.2.32) therefore yields (7.2.31), as claimed. □

7.3 Proofs for Normal Tangent Descent (NTD)

7.3.1 Proof of Lemma 6.2.2

Let g denote the minimal norm element of $\partial_\sigma f(x)$. Write g as a convex combination of subgradients: $g = \sum_{i=1}^n \lambda_i g_i$ where $\sum_{i=1}^n \lambda_i = 1$ and $g_i \in \partial f(x_i)$ for some $x_i \in B_\sigma(x)$ and $n > 0$. Then

$$\begin{aligned} f(x) &\leq f\left(\sum_{i=1}^n \lambda_i x_i + \sum_{i=1}^n \lambda_i (x - x_i)\right) \\ &\leq \sum_{i=1}^n \lambda_i f(x_i) + L\sigma \\ &\leq f(y) + \sum_{i=1}^n \langle \lambda_i g_i, x_i - y\rangle + L\sigma \\ &\leq f(y) + \langle g, x - y\rangle + \sum_{i=1}^n \lambda_i \langle g_i, x_i - x\rangle + L\sigma \\ &\leq f(y) + \operatorname{dist}(0, \partial_\sigma f(x))\|x - y\| + 2L\sigma, \end{aligned}$$

as desired.

7.3.2 Proof of Proposition 6.3.5

We begin with preliminary notation and bounds. First, since $\mathcal{M}$ is C^4 smooth, the projection $P_{\mathcal{M}}$ is C^3 smooth near $\bar{x}$. Second, since f is C^3 smooth along $\mathcal{M}$ near $\bar{x}$, the composition $f_{\mathcal{M}} := f \circ P_{\mathcal{M}}$ is also C^3 smooth near $\bar{x}$. Third, the constant μ is positive due to the active manifold assumption. Fourth, choose $\delta > 0$ small enough that the following hold:

1. $\nabla P_{\mathcal{M}}$ is $C_{\mathcal{M}}$-Lipschitz on $B_\delta(\bar{x})$;
2. $\nabla f_{\mathcal{M}}$ is β-Lipschitz on $B_\delta(\bar{x})$;
3. $\nabla^2 f_{\mathcal{M}}$ is ρ-Lipschitz on $B_\delta(\bar{x})$ in the operator norm, where $\rho := 2\mathrm{lip}^{\mathrm{op}}_{\nabla^2 f_{\mathcal{M}}}(\bar{x})$;
4. f is L-Lipschitz on $B_\delta(\bar{x})$;
5. the quadratic growth bound (Q1) holds:
$$f(x) - f(\bar{x}) \geq \frac{\gamma}{2}\|x - \bar{x}\|^2 \qquad \text{for all } x \in \overline{B}_\delta(\bar{x});$$
6. the strong (a) bound (Q3) holds:
$$\|P_{T_{\mathcal{M}}(y)}(v - \nabla_{\mathcal{M}} f(y))\| \leq C_{(a)}\|x - y\| \tag{7.3.1}$$
for all $x \in \overline{B}_\delta(\bar{x})$, $v \in \partial f(x)$, and $y \in \mathcal{M} \cap \overline{B}_\delta(\bar{x})$.
7. the $(b_\leq)$ regularity bound (Q4) holds:
$$f(x') \geq f(x) + \langle v, x' - x\rangle - \frac{\mu}{2}\|x - \hat{x}\| \tag{7.3.2}$$
for all $x \in B_\delta(\bar{x})$, $v \in \partial f(x)$, and $x' \in B_\delta(\bar{x}) \cap \mathcal{M}$.

8. the sharpness condition holds:

$$\mathrm{dist}(0, \partial f(x)) > 2\mu \qquad \text{for all } x \in B_\delta(\bar{x}) \backslash \mathcal{M}.$$

Given these bounds, let us define

$$\delta_{\mathrm{A}} := \frac{1}{2} \min\left\{\delta, \frac{9\gamma}{16\rho}, \frac{\mu}{2(C_{(a)} + 2\beta + 2C_{\mathcal{M}}L)}\right\}.$$

For this choice of δ_{A}, Item 1 holds automatically. We now prove the remaining items.

7.3.2.1 Item 2: Smoothness of $P_{\mathcal{M}}$.

Fix $x' \in B_{2\delta_{\mathrm{A}}}(\bar{x})$ and $x \in B_{\delta_{\mathrm{A}}}(x)$. Observe that $P_{\mathcal{M}}(x) \in B_{2\delta_{\mathrm{A}}}(\bar{x})$ and we have the inclusion $x - P_{\mathcal{M}}(x) \in N_{\mathcal{M}}(P_{\mathcal{M}}(x))$. Consequently, we have

1. $P_{T_{\mathcal{M}}(P_{\mathcal{M}}(x))}(x) = P_{T_{\mathcal{M}}(P_{\mathcal{M}}(x))}(P_{\mathcal{M}}(x))$;
2. $P_{\mathcal{M}}(x) = P_{\mathcal{M}}(P_{\mathcal{M}}(x))$;
3. $\nabla P_{\mathcal{M}}(P_{\mathcal{M}}(x)) = P_{T_{\mathcal{M}}(P_{\mathcal{M}}(x))}$.

Therefore, we have

$$\begin{aligned}
&\|P_{\mathcal{M}}(x') - P_{\mathcal{M}}(x) - P_{T_{\mathcal{M}}(P_{\mathcal{M}}(x))}(x' - x)\| \\
&= \|P_{\mathcal{M}}(x') - P_{\mathcal{M}}(P_{\mathcal{M}}(x)) - \nabla P_{\mathcal{M}}(P_{\mathcal{M}}(x))(x' - P_{\mathcal{M}}(x))\| \\
&\leq \frac{C_{\mathcal{M}}}{2}\|x' - P_{\mathcal{M}}(x)\|^2 \\
&\leq C_{\mathcal{M}}(\|x' - x\|^2 + \mathrm{dist}^2(x, \mathcal{M})),
\end{aligned}$$

where the first inequality follows from Lipschitz continuity of $\nabla P_{\mathcal{M}}$ on $B_{2\delta_{\mathrm{A}}}(\bar{x}) \subseteq B_\delta(\bar{x})$.

7.3.2.2 Item 3: Bounds on $\nabla_{\mathcal{M}} f$

Recall that $P_{\mathcal{M}}(x) \in B_{2\delta_{\mathrm{A}}}(\bar{x})$ whenever $x \in B_{\delta_{\mathrm{A}}}(\bar{x})$. Thus, below, we prove that

$$\frac{\gamma}{2}\|y - \bar{x}\| \leq \|\nabla f_{\mathcal{M}}(y)\| \leq \beta\|y - \bar{x}\| \qquad \text{for all } y \in B_{2\delta_{\mathrm{A}}}(\bar{x}) \cap \mathcal{M}.$$

This is equivalent to the claimed bound since $\nabla f_{\mathcal{M}}(y) = \nabla_{\mathcal{M}} f(y)$ for all $y \in B_{2\delta_{\mathrm{A}}}(\bar{x}) \cap \mathcal{M}$.

Let us first prove the claimed upper bound. Due to the inequality,

$$f_{\mathcal{M}}(x) - f_{\mathcal{M}}(\bar{x}) \geq \frac{\gamma}{2}\|P_{\mathcal{M}}(x) - \bar{x}\|^2 \qquad \text{for all } x \in B_{\delta}(\bar{x}),$$

it follows that $\bar{x}$ is a local minimizer of $f_{\mathcal{M}}$. Consequently, $\nabla f_{\mathcal{M}}(\bar{x}) = 0$. Thus, since β is a local Lipschitz constant of $\nabla f_{\mathcal{M}}$ on $B_{\delta}(\bar{x})$, we have

$$\|\nabla f_{\mathcal{M}}(y)\| \leq \beta\|y - \bar{x}\| \qquad \text{for all } y \in B_{\delta}(\bar{x}) \cap \mathcal{M}.$$

Since $2\delta_{\mathrm{A}} \leq \delta$, this proves the claimed upper bound.

Next, we prove the claimed lower bound. It suffices to establish the following convexity inequality:

$$f_{\mathcal{M}}(y) + \langle \nabla f_{\mathcal{M}}(y), \bar{x} - y\rangle \leq f_{\mathcal{M}}(\bar{x}) \qquad \text{for all } y \in B_{2\delta_{\mathrm{A}}}(\bar{x}) \cap \mathcal{M}. \tag{7.3.3}$$

Indeed, if this inequality holds, we have

$$\langle \nabla f_{\mathcal{M}}(y), y - \bar{x}\rangle \geq f_{\mathcal{M}}(y) - f_{\mathcal{M}}(\bar{x}) \geq \frac{\gamma}{2}\|y - \bar{x}\|^2 \qquad \text{for all } y \in B_{2\delta_{\mathrm{A}}}(\bar{x}) \cap \mathcal{M},$$

and the desired result follows from Cauchy-Schwarz.

To that end, observe that since $\nabla f_{\mathcal{M}}(\bar{x}) = 0$ and $\nabla^2 f_{\mathcal{M}}$ is ρ-Lipschitz in $B_{2\delta_{\mathrm{A}}}(\bar{x})$, we have

$$f_{\mathcal{M}}(y) \leq f_{\mathcal{M}}(\bar{x}) + \frac{1}{2}\langle \nabla^2 f_{\mathcal{M}}(\bar{x})(y - \bar{x}), y - \bar{x}\rangle + \frac{\rho}{6}\|y - \bar{x}\|^3 \qquad \text{for all } y \in B_{2\delta_{\mathrm{A}}}(\bar{x}).$$

Consequently, we have the lower bound on the quadratic form: for all $y \in B_{2\delta_{\mathrm{A}}}(\bar{x}) \cap \mathcal{M}$, we have

$$\begin{aligned}\frac{1}{2}\left\langle \nabla^2 f_{\mathcal{M}}(\bar{x})(y-\bar{x}), (y-\bar{x})\right\rangle &\geq f_{\mathcal{M}}(y) - f_{\mathcal{M}}(\bar{x}) - \frac{\rho}{6}\|y-\bar{x}\|^3 \\ &\geq \frac{\gamma}{2}\|y-\bar{x}\|^2 - \frac{\rho}{6}\|y-\bar{x}\|^3 \\ &\geq \frac{3\gamma}{8}\|y-\bar{x}\|^2, \end{aligned} \tag{7.3.4}$$

where the second inequality follows from the quadratic growth bound and the third follows from the bound $\|y-\bar{x}\| \leq 2\delta_{\mathrm{A}} \leq \frac{3\gamma}{4\rho}$. Therefore, for all $y \in \mathcal{M} \cap B_{2\delta_{\mathrm{A}}}(\bar{x})$, we have

$$\begin{aligned} f_{\mathcal{M}}(\bar{x}) &\geq f_{\mathcal{M}}(y) + \langle \nabla f_{\mathcal{M}}(y), \bar{x}-y\rangle + \frac{1}{2}\left\langle \nabla^2 f_{\mathcal{M}}(y)(\bar{x}-y), (\bar{x}-y)\right\rangle - \frac{\rho}{6}\|y-\bar{x}\|^3 \\ &\geq f_{\mathcal{M}}(y) + \langle \nabla f_{\mathcal{M}}(y), \bar{x}-y\rangle + \frac{1}{2}\left\langle \nabla^2 f_{\mathcal{M}}(\bar{x})(\bar{x}-y), (\bar{x}-y)\right\rangle - \frac{2\rho}{3}\|y-\bar{x}\|^3 \\ &\geq f_{\mathcal{M}}(y) + \langle \nabla f_{\mathcal{M}}(y), \bar{x}-y\rangle + \frac{3\gamma}{8}\|y-\bar{x}\|^2 - \frac{2\rho}{3}\|y-\bar{x}\|^3 \\ &\geq f_{\mathcal{M}}(y) + \langle \nabla f_{\mathcal{M}}(y), \bar{x}-y\rangle, \end{aligned}$$

where the first and second inequalities follow by Lipschitz continuity of $\nabla^2 f_{\mathcal{M}}$; the third inequality follows from (7.3.4); and the fourth inequality follows from the bound $\|y-\bar{x}\| \leq 2\delta_{\mathrm{A}} \leq \frac{9\gamma}{16\rho}$. This completes the proof.

7.3.2.3 Item 4: Consequences of strong (a)-regularity

Fix $x \in B_{\delta_{\mathrm{A}}}(\bar{x})$ and $\sigma \leq \delta_{\mathrm{A}}$. Recall that $y := P_{\mathcal{M}}(x) \in B_{2\delta_{\mathrm{A}}}(\bar{x})$ since $x \in B_{\delta_{\mathrm{A}}}(\bar{x})$. Fix $g \in \partial_\sigma f(x)$. By definition of $\partial_\sigma f(x)$, there exists a family of coefficients $\lambda_i \in [0,1]$, points $x_i \in \overline{B}_\sigma(x) \subseteq \overline{B}_\delta(\bar{x})$, and subgradients $g_i \in \partial f(x_i)$ indexed by a finite set $i \in I$ such that $\sum_{i\in I}\lambda_i = 1$ and $g = \sum_{i\in I}\lambda_i g_i$. Therefore, by averaging the strong (a) bound (7.3.1) over g_i, we find that

$$\|P_{T_{\mathcal{M}(y)}}(g - \nabla_{\mathcal{M}} f(y))\| \leq \sum_{i\in I}\lambda_i\|P_{T_{\mathcal{M}(y)}}(g_i - \nabla_{\mathcal{M}} f(y))\|$$

$$\leq \sum_{i \in I} \lambda_i C_{(a)} \|x_i - y\|.$$

$$\leq C_{(a)}(\text{dist}(x, \mathcal{M}) + \sigma).$$

Since g was arbitrary, it follows that for all $x \in B_{\delta_{\text{A}}}(\bar{x})$ and $\sigma \leq \delta_{\text{A}}$, we have

$$\sup_{g \in \partial_\sigma f(x)} \|P_{T_{\mathcal{M}}(y)}(g - \nabla_{\mathcal{M}} f(y))\| \leq C_{(a)}(\text{dist}(x, \mathcal{M}) + \sigma). \tag{7.3.5}$$

Now, we apply this bound to establish the two remaining inequalities.

Indeed, first observe that for all $x \in B_{\delta_{\text{A}}}(\bar{x})$ and $\sigma \leq \delta_{\text{A}}$, we have

$$\sup_{g \in \partial_\sigma f(x)} \|P_{T_{\mathcal{M}}(y)} g\| \leq \|\nabla_{\mathcal{M}} f(y)\| + C_{(a)}(\text{dist}(x, \mathcal{M}) + \sigma) \leq \beta\|y - \bar{x}\| + C_{(a)}(\text{dist}(x, \mathcal{M}) + \sigma),$$

where the first inequality follows from (7.3.5) and the second inequality follows from Item 3. This proves the first claimed bound. Second, observe that for all $x \in B_{\delta_{\text{A}}}(\bar{x})$ and $\sigma \leq \delta_{\text{A}}$, we have

$$\begin{aligned} \sup_{g, g' \in \partial_\sigma f(x)} \|P_{T_{\mathcal{M}}(y)}(g - g')\| &\leq \sup_{g \in \partial_\sigma f(x)} \|P_{T_{\mathcal{M}}(y)}(g - \nabla_{\mathcal{M}} f(y))\| + \sup_{g' \in \partial_\sigma f(x)} \|P_{T_{\mathcal{M}}(y)}(g' - \nabla_{\mathcal{M}} f(y))\| \\ &\leq 2C_{(a)}(\text{dist}(x, \mathcal{M}) + \sigma). \end{aligned}$$

where the second inequality follows from (7.3.5). This completes the proof.

7.3.2.4 Item 5: Aiming inequality

Consider a point $x \in B_{\delta_{\text{A}}}(\bar{x})$, let $\kappa = 2\mu$, and define

$$\hat{x} \in \operatorname*{argmin}_{x' \in \overline{B}_{2\delta_{\text{A}}}(\bar{x})} \{f(x') + \kappa\|x' - x\|\}.$$

We claim that $\hat{x} \in \mathcal{M} \cap B_{2\delta_{\text{A}}}(\bar{x})$. Indeed, first note that by definition of $\hat{x}$ and the inclusion $\hat{x} \in \overline{B}_{2\delta_{\text{A}}}(\bar{x})$, we have

$$\|\hat{x} - x\| \leq \frac{f(\bar{x}) - f(\hat{x})}{\kappa} + \|\bar{x} - x\| \leq \|\bar{x} - x\| < \delta_{\text{A}},$$

where the second inequality follows since $\bar{x}$ is a minimizer of f on $B_{2\delta_{\rm A}}(\bar{x})$, a consequence of quadratic growth. Thus, by the triangle inequality, we have $\hat{x} \in B_{2\delta_{\rm A}}(\bar{x})$. By Fermat's rule, we, therefore, have the inclusion:

$$0 \in \partial(f + \kappa\|\cdot - x\|)(\hat{x}) \subseteq \partial f(\hat{x}) + \kappa\overline{B}.$$

If $\hat{x} \notin \mathcal{M}$, then $\mathrm{dist}(0, \partial f(\hat{x})) > \kappa$, contradicting the above inclusion. Therefore, we have $\hat{x} \in \mathcal{M} \cap B_{2\delta_{\rm A}}(\bar{x})$.

Turning to the aiming inequality, apply the $(b_{\leq})$-regularity bound (7.3.2) to $\hat{x}$:

$$f(\hat{x}) \geq f(x) + \langle v, \hat{x} - x\rangle - \varepsilon\|x - \hat{x}\| \geq f(\hat{x}) + \langle v, \hat{x} - x\rangle + (\kappa - \varepsilon)\|x - \hat{x}\|,$$

where we define $\varepsilon := \mu/2$. Consequently, we have

$$\langle v, x - P_{\mathcal{M}}(x)\rangle \geq (\kappa - \varepsilon)\|x - \hat{x}\| + \langle v, \hat{x} - P_{\mathcal{M}}(x)\rangle \qquad \text{for all } v \in \partial f(x). \tag{7.3.6}$$

We now bound the term $\langle v, \hat{x} - P_{\mathcal{M}}(x)\rangle$: By the conclusion of Item 2, we have

$$\|P_{\mathcal{M}}(\hat{x}) - P_{\mathcal{M}}(x) - P_{T_{\mathcal{M}}(P_{\mathcal{M}}(x))}(\hat{x} - x)\| \leq C_{\mathcal{M}}(\|x - \hat{x}\|^2 + \mathrm{dist}^2(x, \mathcal{M})) \leq 2C_{\mathcal{M}}\|x - \hat{x}\|^2,$$

where the second inequality follows since $\hat{x} \in \mathcal{M}$. Thus, we have

$$\begin{aligned}
|\langle v, \hat{x} - P_{\mathcal{M}}(x)\rangle| &\leq |\langle v, P_{T_{\mathcal{M}}(P_{\mathcal{M}}(x))}(\hat{x} - x)\rangle| + 2C_{\mathcal{M}}\|v\|\|x - \hat{x}\|^2 \\
&\leq \|P_{T_{\mathcal{M}}(P_{\mathcal{M}}(x))}v\|\|\hat{x} - x\| + 2C_{\mathcal{M}}L\|x - \hat{x}\|^2 \\
&\leq (C_{(a)}\mathrm{dist}(x, \mathcal{M}) + \beta\|P_{\mathcal{M}}(x) - \bar{x}\|)\|\hat{x} - x\| + 2C_{\mathcal{M}}L\|x - \hat{x}\|^2 \\
&\leq (C_{(a)}\delta_{\rm A} + 2\beta\delta_{\rm A} + 2C_{\mathcal{M}}L\delta_{\rm A})\|\hat{x} - x\| \\
&\leq \varepsilon\|\hat{x} - x\|.
\end{aligned}$$

where the second inequality follows from Item 4 and the third inequality follows from the inclusion $P_{\mathcal{M}}(x) \in B_{2\delta_{\rm A}}(\bar{x})$. Therefore, plugging this bound into (7.3.6), we arrive at

$$\langle v, x - P_{\mathcal{M}}(x)\rangle \geq (\kappa - 2\varepsilon)\|x - \hat{x}\| \geq \mu\mathrm{dist}(x, \mathcal{M}),$$

as desired.

7.3.2.5 Item 6: Bounding subgradients

Fix $x \in B_{\delta_{\mathrm{A}}}(\bar{x})$, $\sigma \le \delta_{\mathrm{A}}$, and $g \in \partial_\sigma f(x)$. By definition of $\partial_\sigma f(x)$, there exists a family of coefficients $\lambda_i \in [0, 1]$, points $x_i \in \overline{B}_\sigma(x) \subseteq \overline{B}_\delta(\bar{x})$, and subgradients $g_i \in \partial f(x_i)$ indexed by a finite set $i \in I$ such that $\sum_{i \in I} \lambda_i = 1$ and $g = \sum_{i \in I} \lambda_i g_i$. Recall that by Lipschitz continuity of f on $B_\delta(\bar{x})$, we have $\|g_i\| \le L$ for $i \in I$. Therefore,

$$\|g\| \le \sum_{i \in I} \lambda_i \|g_i\| \le L,$$

as desired.

7.3.2.6 Item 7: Bounding the function gap

Fix a point $x \in B_{\delta_{\mathrm{A}}}(\bar{x})$ and recall that $P_{\mathcal{M}}(x) \in B_{2\delta_{\mathrm{A}}}(\bar{x})$. Then by Lipschitz continuity of f on $B_\delta(\bar{x})$, we have

$$f(x) - f(P_{\mathcal{M}}(\bar{x})) \le L\mathrm{dist}(x, \mathcal{M}).$$

Next, arguing as in the proof of Item 3, we find that $\nabla f_{\mathcal{M}}(\bar{x}) = 0$. Thus, since $\nabla f_{\mathcal{M}}$ is β-Lipschitz on $B_\delta(\bar{x})$, we have

$$f(P_{\mathcal{M}}(x)) - f(\bar{x}) = f_{\mathcal{M}}(P_{\mathcal{M}}(x)) - f(\bar{x}) \le \langle \nabla f_{\mathcal{M}}(\bar{x}), P_{\mathcal{M}}(x) - \bar{x} \rangle + \frac{\beta}{2}\|P_{\mathcal{M}}(x) - \bar{x}\|^2 = \frac{\beta}{2}\|P_{\mathcal{M}}(x) - \bar{x}\|^2.$$

By putting both bounds together, we have

$$f(x) - f(\bar{x}) = f(x) - f(P_{\mathcal{M}}(x)) + f(P_{\mathcal{M}}(x)) - f(\bar{x}) \le L\mathrm{dist}(x, \mathcal{M}) + \frac{\beta}{2}\|P_{\mathcal{M}}(x) - \bar{x}\|^2,$$

as desired.

7.3.3 Proof of Corollary 6.2.5

We begin with the following known Lemma, which immediately follows from [106, Proposition 2.8]

Lemma 7.3.1. *Let $f\colon \mathbb{R}^d \to \mathbb{R}$ be a locally Lipschitz function. Suppose that there exists sequences $x_k \to \bar{x}$, $\tau_k \to 0$, and $g_k \in \partial_{\tau_k} f(x_k)$ with $\|g_k\| \to 0$. Then $\bar{x}$ is a Clarke critical point.*

Now we turn to the proof of the Corollary. Since f has bounded initial sublevel set, the following widened sublevel set is bounded:

$$S := \{x + u\colon f(x) \leq f(x_0) \text{ and } u \in \overline{B}(x)\}.$$

Thus, there exists $L > 0$ such that f is L-Lipschitz on S. In addition, ∂f is uniformly bounded by L on int S.

We begin with a claim.

Claim: Fix $i > 0$ and define $\tau_i := 2^{-i}$. Let $s_k := \max\{\|g_k\|, c_0\|g_0\|\}$ be the trust region parameter used in Algorithm 3 and define $\epsilon_{i,k} := \sqrt{128}L\tau_i$. Then, with probability one, the event

$$E_k^{(i)} = \left\{\operatorname{dist}(0, \partial_{\tau_i} f(x_k)) > \epsilon_{i,k} \text{ and } f(x_{k+1}) > f(x_k) - \frac{\tau_i \operatorname{dist}(0, \partial_{\tau_i} f(x_k))}{8}\right\}$$

cannot happen infinitely often, i.e.,

$$P\left(\cap_{T=1}^{\infty} \cup_{k=T}^{\infty} E_k^{(i)}\right) = 0.$$

Proof: We prove that $P(E_k^{(i)})$ is summable in k. Indeed, first, note that $P(E_k^{(i)}) = 0$ when $P(\operatorname{dist}(0, \partial_{\tau_i} f(x_k)) > \epsilon_{i,k}) = 0$. On the other hand, suppose $P(\operatorname{dist}(0, \partial_{\tau_i} f(x_k)) > \epsilon_{i,k}) > 0$. Now we upper bound $P(E_k^{(i)})$ for all G_k satisfying $G_k \geq i$. For such $G := G_k$, the radius $\tau_i = \sigma_{G-i}$ is among those considered in Algorithm 3. Moreover, since $s_k \leq L$

(recall $x_k \in \operatorname{int}(S)$), the radius satisfies the trust region constraint: $\sigma_{G-i} = \tau_i \leq \epsilon_{i,k}/s_k \leq \operatorname{dist}(0, \partial_{\sigma_{G-i}} f(x))/s_k$. Therefore, if `NDescent` terminates with descent at the $(G-i)$-th level in Algorithm 3, it follows that

$$f(x_{k+1}) > f(x_k) - \frac{\tau_i \operatorname{dist}(0, \partial_{\tau_i} f(x_k))}{8}.$$

We estimate the probability of this success with Lemma 6.2.3: there exist $C > 0$ depending on $\epsilon_{i,k}$ and for all $k \geq i$, we have

$$\begin{aligned} P(E_k^{(i)}) &\leq P\left(f(x_{k+1}) > f(x_k) - \frac{\tau_i \operatorname{dist}(0, \partial_{\tau_i} f(x_k))}{8} \middle| \operatorname{dist}(0, \partial_{\tau_i} f(x_k)) > \epsilon_{i,k}\right) \\ &\leq \exp(-Ck). \end{aligned}$$

Therefore, $P(E_k^{(i)})$ is summable in k. The result then follows from Borel–Cantelli lemma. ■

By the claim and a union bound, we know that with probability one, for any fixed i, $E_k^{(i)}$ cannot happen infinitely often. Now, suppose that a subsequence $\{x_{k_l}\}$ (where $k_l \geq l$ is strictly increasing in l) converges to a point $\bar{x}$. We note that the sequence $\{f(x_k)\}$ is bounded below: Indeed, since x_{k_l} converges and f is continuous, it follows $\{f(x_{k_l})\}$ is bounded below by a constant $c \in \mathbb{R}$. Consequently, since $\{f(x_k))\}$ is nonincreasing and $k_l \geq l$, it follows that $c \leq f(x_{k_l}) \leq f(x_l)$ and for every $l > 0$, as desired. As a result, the following inequalities cannot be valid simultaneously infinitely often:

$$\operatorname{dist}(0, \partial_{\tau_i} f(x_{k_l})) > \epsilon_{i,k} \text{ and } f(x_{k_l+1}) \leq f(x_{k_l}) - \frac{\tau_i \operatorname{dist}(0, \partial_{\tau_i} f(x_{k_l}))}{8}.$$

Therefore, $\operatorname{dist}(0, \partial_{\tau_i} f(x_{k_l})) > \epsilon_{i,k}$ cannot happen infinitely often. Consequently, we can find a sequence of increasing indices j_i such that

$$\operatorname{dist}(0, \partial_{\tau_i} f(x_{j_i})) \leq \epsilon_{i,k} \qquad \text{and } x_{j_i} \to \bar{x}.$$

Since $\epsilon_{i,k} \to 0$ as $k \to \infty$, Lemma 7.3.1, shows that $\bar{x}$ is Clarke critical.

7.3.4 Proof of Lemma 6.5.7

We begin with preliminary notation and bounds. We fix $x \in B_{\delta_{\mathrm{Grid}}}(\bar{x})$ and subgradient $g \in \partial_\sigma f(x)\backslash\{0\}$. We define $y := P_{\mathcal{M}}(x)$, $T := T_{\mathcal{M}}(y)$, and $N := N_{\mathcal{M}}(y)$. We have the following two bounds: First, we have

$$(\mu + L)C_{\mathcal{M}}(D_1 \mathrm{dist}(x, \mathcal{M}) + \sigma) \leq (\mu + L)C_{\mathcal{M}}\delta_{\mathrm{Grid}}(D_1 + 1) = \frac{\mu}{8}C_{\mathcal{M}}(D_1^{-1} + 1)\delta_{\mathrm{Grid}} \leq \frac{\mu}{8}. \tag{7.3.7}$$

Second, we have

$$C_{(a)}(\mathrm{dist}(x, \mathcal{M}) + \sigma) + \beta\|y - \bar{x}\| \leq 2C_{(a)}\delta_{\mathrm{Grid}} + 2\beta\delta_{\mathrm{Grid}} \leq \frac{\mu}{4}. \tag{7.3.8}$$

We now turn to the proof.

By Lemma 6.5.3 (which is applicable since $x \in B_{\delta_{\mathrm{A}}/2}(\bar{x})$ and $\sigma \leq \delta_{\mathrm{Grid}} \leq \delta_{\mathrm{A}}/2$), we have

$$\left\langle \hat{g}, \sigma\frac{P_N g}{\|g\|} \right\rangle \leq -\sigma\mu\frac{\|P_N g\|}{\|g\|} + (\mu + L)\mathrm{dist}(x, \mathcal{M}) + (\mu + L)C_{\mathcal{M}}(\mathrm{dist}^2(x, \mathcal{M}) + \sigma^2).$$

Rearranging, we find that

$$\begin{aligned}
\langle P_N\hat{g}, g\rangle &\leq -\mu\|P_N g\| + \frac{(\mu + L)\|g\|\mathrm{dist}(x, \mathcal{M})}{\sigma} + \frac{(\mu + L)\|g\|C_{\mathcal{M}}(\mathrm{dist}^2(x, \mathcal{M}) + \sigma^2)}{\sigma} \\
&\leq -\mu\|P_N g\| + \frac{\mu}{8}\|g\| + (\mu + L)C_{\mathcal{M}}(D_1\mathrm{dist}(x, \mathcal{M}) + \sigma) \cdot \|g\| \\
&\leq -\mu\|P_N g\| + \frac{\mu}{4}\|g\|,
\end{aligned}$$

where the second inequality follows from the assumption $D_1^{-1}\mathrm{dist}(x, \mathcal{M}) \leq \sigma$ and the third follows from (7.3.7). Now observe that

$$\langle P_T\hat{g}, g\rangle \leq \|P_T\hat{g}\|\|g\| \leq (C_{(a)}(\mathrm{dist}(x, \mathcal{M}) + \sigma) + \beta\|y - \bar{x}\|) \cdot \|g\| \leq \frac{\mu}{4}\|g\|,$$

where second inequality follows from (6.3.4) and the third inequality follows from (7.3.8). Therefore,

$$\langle \hat{g}, g\rangle = \langle P_N\hat{g}, g\rangle + \langle P_T\hat{g}, g\rangle \leq -\mu\|P_N g\| + \frac{\mu}{2}\|g\| \leq -\frac{\mu}{2}\|g\| + \mu\|P_T(g)\|,$$

as desired.

7.3.5 Proof of $\mu \leq L$

Lemma 7.3.2. *We have that $\mu \leq L$.*

Proof. Indeed,

$$\mu = \frac{1}{4} \liminf_{x' \xrightarrow{\mathcal{M}^c} \bar{x}} \operatorname{dist}(0, \partial f(x)) \leq \limsup_{x \to \bar{x}} \operatorname{dist}(0, \partial f(x)) \leq L,$$

by Proposition 6.3.5. □

7.3.6 Proof of Lemma 6.6.4

We fix $a > 0$. Note that the claimed inclusion is a consequence of the following bound:

$$f(x) - f(\bar{x}) \geq \frac{\gamma}{2} \min\{\delta_A, \|x - \bar{x}\|\}\|x - \bar{x}\| \qquad \text{for all } x \in \mathbb{R}^d. \tag{7.3.9}$$

Here, we provide a proof for completeness.

To that end, we remind the reader that Assumption Q is in force. Consequently, by Item 1 of Proposition 6.3.5, we have:

$$f(x) - f(\bar{x}) \geq \frac{\gamma}{2}\|x - \bar{x}\|^2 \qquad \text{for all } x \in \overline{B}_{\delta_A}(\bar{x}).$$

Thus, if $x \in B_{\delta_A}(\bar{x})$, bound (7.3.9) is immediate. On the other hand, suppose that we have $x \in \mathbb{R}^d \backslash B_{\delta_A}(\bar{x})$. Define the curve $x_t \colon t \mapsto (1 - t)x + t\bar{x}$. Choose $t_0 \in [0, 1]$ such that $x_{t_0} \in \operatorname{bdry} B_{\delta_A}(\bar{x})$. Then, by Jensen's inequality, we have

$$(1-t_0)f(x) \geq f(x_{t_0}) - t_0 f(\bar{x}) \geq (1-t_0)f(\bar{x}) + \frac{\gamma}{2}\|x_{t_0} - \bar{x}\|^2 = (1-t_0)f(\bar{x}) + \frac{\gamma(1 - t_0)}{2}\|x - \bar{x}\|\|x_{t_0} - \bar{x}\|.$$

Consequently, since $\|x_{t_0} - \bar{x}\| = \delta_A$, we have

$$f(x) - f(\bar{x}) \geq \frac{\gamma\delta_A}{2}\|x - \bar{x}\| \geq \frac{\gamma}{2}\min\{\delta_A, \|x - \bar{x}\|\}\|x - \bar{x}\|,$$

as desired. This completes the proof.

7.3.7 Proof of (6.6.18)

Let us expand the left-hand-side of (6.6.18):

$$\frac{16L'\sqrt{2\log(2K_1^2/p)}}{K_1^{1/2}} \leq \underbrace{\frac{16DL'\sqrt{2\log(K_1^2)}}{K_1^{1/2}}}_{=:A} + \underbrace{\frac{16DL'\sqrt{2\log(2/p)}}{K_1^{1/2}}}_{=:B}.$$

Note that $B \leq a/4$ by definition of K_1. Consequently, the proof will follow if $A \leq a/4$. To that end, for any $\alpha \in (0,1)$, we have

$$A = \frac{16DL'\sqrt{2\log(K_1^2)}}{K_1^{1/2}} = \frac{16DL'\sqrt{2\log(K_1^{2\alpha})/\alpha}}{K_1^{1/2}} \leq \frac{16DL'\sqrt{2/\alpha}}{K_1^{(1-\alpha)/2}},$$

Therefore, we have $A \leq a/4$ whenever

$$K_1 \geq \inf_{\alpha\in(0,1)} \frac{\left(64DL'\sqrt{\frac{2}{\alpha}}\right)^{\frac{2}{(1-\alpha)}}}{a^{\frac{2}{(1-\alpha)}}} = \frac{D^2}{a^2}\inf_{\alpha\in(0,1)} \frac{\left(64L'\sqrt{\frac{2}{\alpha}}\right)^{\frac{2}{(1-\alpha)}}}{(\frac{a}{D})^{\frac{2\alpha}{(1-\alpha)}}} = \frac{D^2}{a^2}b.$$

This lower bound holds by definition of K_1. Consequently, $A \leq a/4$. Therefore, the proof is complete.

7.3.8 Proof of Lemma 6.6.6

Throughout this section, we use the symbol $a \lesssim b$ to mean that $a \leq \eta b$ for a fixed numerical constant η independent of f. In addition, we use the bound on the condition number: $\kappa \geq 1$, since $\mu \leq L$; see Lemma 7.3.5.

Turning to the bound, we wish to upper bound q.

$$q = \max\left\{\rho, \sqrt{1 - \frac{3\mu^2}{256L^2}}, \frac{1}{2}\right\}.$$

First note that

$$1 - \sqrt{1 - \frac{3\mu^2}{256L^2}} \gtrsim \frac{\mu^2}{L^2} \geq \frac{1}{\kappa^2}.$$

Next, we upper bound ρ. To that end, we must bound the constants a_1 and a_2, which rely on the somewhat involved constants C_4 and C_5. Thus, we first lower bound C_4:

$$\begin{aligned} C_4 &= \min\left\{\frac{\beta}{C_{(a)}(1+\delta_{\mathrm{A}})}, \frac{\min\{\mu/\delta_{\mathrm{A}}, C_3 D_2/\beta\}}{4(1+(1+\delta_{\mathrm{A}})C_{\mathcal{M}})(\mu+L))}, \frac{1}{2}\right\} \\ &\gtrsim \min\left\{\frac{\beta}{C_{(a)}}, \frac{\mu}{L(1+C_{\mathcal{M}})}, \frac{\gamma^2\mu}{L^2\beta(1+C_{\mathcal{M}})}\right\} \\ &\geq \frac{1}{\kappa^3(1+C_{\mathcal{M}})}, \end{aligned}$$

where we use the bounds $\mu \leq L$, $C_3 \gtrsim \gamma^2/L$, and $D_2 \gtrsim \mu$. Turning to C_5, we have:

$$\begin{aligned} C_5 &= \min\left\{\frac{\beta}{2C_{(a)}}, \frac{C_3 D_2}{32C_{(a)}\beta}, C_4, \frac{C_2}{4}\right\} \\ &\gtrsim \min\left\{\frac{\beta}{C_{(a)}}, \frac{\gamma^2\mu}{LC_{(a)}\beta}, \frac{1}{\kappa^3(1+C_{\mathcal{M}})}, \frac{\gamma}{C_{(a)}}\right\} \\ &\geq \frac{1}{\kappa^3(1+C_{\mathcal{M}})}, \end{aligned}$$

where we again use $C_3 \gtrsim \gamma^2/L$, and $D_2 \gtrsim \mu$. Therefore, we have the lower bound for a_2:

$$a_2 = \frac{\min\{C_1/L, C_5\}}{2} \gtrsim \min\left\{\frac{\gamma^2}{L^2}, \frac{1}{\kappa^3(1+C_{\mathcal{M}})}\right\} \gtrsim \frac{1}{\kappa^3(1+C_{\mathcal{M}})}.$$

In addition, we have the upper bound:

$$a_2 = \frac{\min\{C_1/L, C_5\}}{2} \leq C_4/2 \leq 1/4.$$

Finally to lower bound a_1, we have

$$a_1 = \min\{D_1, D_2/L\} \gtrsim \min\left\{\frac{\mu}{L}, \frac{\gamma}{L}\right\} \gtrsim \frac{1}{\kappa},$$

where we use the bound $D_1 \gtrsim \mu/L$ and $D_2/L \gtrsim \mu/L$.

Now we upper bound ρ by providing a lower bound on $1 - \rho$.

$$\begin{aligned}
1 - \rho &= \frac{1}{8} \min\left\{ \frac{\gamma a_2}{8 \max\{4La_2^2, \beta\}}, \frac{\mu a_1}{4 \max\{2L, \beta / a_2^2\}} \right\} \\
&\gtrsim \min\left\{ \frac{\gamma}{La_2}, \frac{\gamma a_2}{\beta}, \frac{\mu a_1}{L}, \frac{\mu a_1 a_2^2}{\beta} \right\} \\
&\gtrsim \min\left\{ \frac{\gamma}{L}, \frac{\gamma}{\kappa^3 (1 + C_{\mathcal{M}}) \beta}, \frac{\mu}{\kappa L}, \frac{\mu}{\beta \kappa^7 (1 + C_{\mathcal{M}})^2} \right\} \\
&\gtrsim \frac{1}{\kappa^8 (1 + C_{\mathcal{M}})^2}
\end{aligned}$$

Putting all these bounds together, we find that:

$$1 - q \gtrsim \min\left\{ \frac{1}{\kappa^8 (1 + C_{\mathcal{M}})^2}, \frac{1}{\kappa^2} \right\} \geq \frac{1}{\kappa^8 (1 + C_{\mathcal{M}})^2},$$

as desired.

BIBLIOGRAPHY

[1] S. J. Wright, "Identifiable surfaces in constrained optimization," *SIAM Journal on Control and Optimization*, vol. 31, no. 4, pp. 1063–1079, 1993.

[2] A. S. Lewis, "Active sets, nonsmoothness, and sensitivity," *SIAM Journal on Optimization*, vol. 13, no. 3, pp. 702–725, 2002.

[3] C. Lemaréchal, F. Oustry, and C. Sagastizábal, "The $\mathcal{U}$-lagrangian of a convex function," *Transactions of the American mathematical Society*, vol. 352, no. 2, pp. 711–729, 2000.

[4] R. Mifflin and C. Sagastizábal, "A VU-algorithm for convex minimization," *Mathematical programming*, vol. 104, no. 2, pp. 583–608, 2005.

[5] A. Shapiro, "On a class of nonsmooth composite functions," *Mathematics of Operations Research*, vol. 28, no. 4, pp. 677–692, 2003.

[6] D. Drusvyatskiy and A. S. Lewis, "Optimality, identifiability, and sensitivity," *Mathematical Programming*, vol. 147, no. 1, pp. 467–498, 2014. Citations refer to long version arXiv:1207.6628.

[7] D. Drusvyatskiy, A. D. Ioffe, and A. S. Lewis, "Generic minimizing behavior in semialgebraic optimization," *SIAM Journal on Optimization*, vol. 26, no. 1, pp. 513–534, 2016.

[8] J. D. Lee, M. Simchowitz, M. I. Jordan, and B. Recht, "Gradient descent only converges to minimizers," in *Conference on learning theory*, pp. 1246–1257, PMLR, 2016.

[9] I. Panageas and G. Piliouras, "Gradient descent only converges to minimizers: Non-isolated critical points and invariant regions," *arXiv preprint arXiv:1605.00405*, 2016.

[10] S. Bhojanapalli, B. Neyshabur, and N. Srebro, "Global optimality of local search for low rank matrix recovery," in *Advances in Neural Information Processing Systems*, pp. 3873–3881, 2016.

[11] 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*, pp. 1233–1242, JMLR. org, 2017.

[12] R. Ge, J. Lee, and T. Ma, “Matrix completion has no spurious local minimum,” in *Advances in Neural Information Processing Systems*, pp. 2973–2981, 2016.

[13] J. Sun, Q. Qu, and J. Wright, “When are nonconvex problems not scary?,” *arXiv preprint arXiv:1510.06096*, 2015.

[14] J. Sun, Q. Qu, and J. Wright, “A geometric analysis of phase retrieval,” *Foundations of Computational Mathematics*, vol. 18, no. 5, pp. 1131–1198, 2018.

[15] B. T. Polyak and A. B. Juditsky, “Acceleration of stochastic approximation by averaging,” *SIAM journal on control and optimization*, vol. 30, no. 4, pp. 838–855, 1992.

[16] J. C. Duchi and F. Ruan, “Asymptotic optimality in stochastic optimization,” *The Annals of Statistics*, vol. 49, no. 1, pp. 21–48, 2021.

[17] L. Le Cam, L. M. LeCam, and G. L. Yang, *Asymptotics in statistics: some basic concepts*. Springer Science & Business Media, 2000.

[18] A. W. Van der Vaart, *Asymptotic statistics*, vol. 3. Cambridge university press, 2000.

[19] Y. Nesterov *et al.*, *Lectures on convex optimization*, vol. 137. Springer, 2018.

[20] H. Whitney, “Elementary structure of real algebraic varieties,” in *Hassler Whitney Collected Papers*, pp. 456–467, Springer, 1992.

[21] H. Whitney, “Local properties of analytic varieties,” in *Hassler Whitney Collected Papers*, pp. 497–536, Springer, 1992.

[22] H. Whitney, “Tangents to an analytic variety,” in *Hassler Whitney Collected Papers*, pp. 537–590, Springer, 1992.

[23] T.-C. Kuo, “Characterizations of v-sufficiency of jets,” *Topology*, vol. 11, no. 1, pp. 115–131, 1972.

[24] J.-L. Verdier, “Stratifications de whitney et théoreme de bertini-sard,” *Inventiones mathematicae*, vol. 36, no. 1, pp. 295–312, 1976.

[25] D. Davis, D. Drusvyatskiy, and L. Jiang, “Active manifolds, stratifications, and convergence to local minima in nonsmooth optimization,” *arXiv preprint arXiv:2108.11832*, 2023.

[26] D. Davis, D. Drusvyatskiy, and L. Jiang, "Asymptotic normality and optimality in nonsmooth stochastic approximation," *arXiv preprint arXiv:2301.06632*, 2023.

[27] D. Davis and L. Jiang, "A nearly linearly convergent first-order method for nonsmooth functions with quadratic growth," *arXiv preprint arXiv:2205.00064*, 2022.

[28] R. T. Rockafellar and R. J.-B. Wets, *Variational analysis*, vol. 317. Springer Science & Business Media, 2009.

[29] B. S. Mordukhovich, *Variational analysis and generalized differentiation I: Basic theory*, vol. 330. Springer Science & Business Media, 2006.

[30] J.-P. Penot, *Calculus without derivatives*, vol. 266. Springer Science & Business Media, 2012.

[31] F. H. Clarke, Y. S. Ledyaev, R. J. Stern, and P. R. Wolenski, *Nonsmooth analysis and control theory*, vol. 178. Springer Science & Business Media, 2008.

[32] J. Borwein and A. S. Lewis, *Convex analysis and nonlinear optimization: theory and examples*. Springer Science & Business Media, 2010.

[33] R. Poliquin, R. Rockafellar, and L. Thibault, "Local differentiability of distance functions," *Transactions of the American mathematical Society*, vol. 352, no. 11, pp. 5231–5249, 2000.

[34] F. H. Clarke, R. Stern, and P. Wolenski, "Proximal smoothness and the lower-c2 property," *J. Convex Anal*, vol. 2, no. 1-2, pp. 117–144, 1995.

[35] R. A. Poliquin and R. T. Rockafellar, "A calculus of prox-regularity," *J. Convex Anal*, vol. 17, no. 1, pp. 203–210, 2010.

[36] H. Federer, "Curvature measures," *Transactions of the American Mathematical Society*, vol. 93, no. 3, pp. 418–491, 1959.

[37] D. Drusvyatskiy, "The proximal point method revisited," *SIAG/OPT Views and News*, vol. 26, pp. 1–8, 2017.

[38] D. Davis and D. Drusvyatskiy, "Stochastic model-based minimization of weakly convex functions," *SIAM Journal on Optimization*, vol. 29, no. 1, pp. 207–239, 2019.

[39] S. A. Miller and J. Malick, “Newton methods for nonsmooth convex minimization: connections among-lagrangian, riemannian newton and sqp methods,” *Mathematical programming*, vol. 104, no. 2, pp. 609–633, 2005.

[40] D. Davis, D. Drusvyatskiy, and V. Charisopoulos, “Stochastic algorithms with geometric step decay converge linearly on sharp functions,” *arXiv preprint arXiv:1907.09547*, 2019.

[41] D. Davis and D. Drusvyatskiy, “Active strict saddles in nonsmooth optimization,” *arXiv preprint arXiv:1912.07146*, 2019.

[42] D. Trotman, “Stratification theory,” in *Handbook of Geometry and Topology of Singularities I*, pp. 243–273, Springer, 2020.

[43] H. Gfrerer and V. Outrata, “On a semismooth* newton method for solving generalized equations,” *SIAM Journal on Optimization*, vol. 31, no. 1, pp. 489–517, 2021.

[44] R. Mifflin, “Semismooth and semiconvex functions in constrained optimization,” *SIAM Journal on Control and Optimization*, vol. 15, no. 6, pp. 959–972, 1977.

[45] V. Norkin, “Generalized-differentiable functions,” *Cybernetics*, vol. 16, no. 1, pp. 10–12, 1980.

[46] J. Bolte, A. Daniilidis, A. Lewis, and M. Shiota, “Clarke subgradients of stratifiable functions,” *SIAM Journal on Optimization*, vol. 18, no. 2, pp. 556–572, 2007.

[47] D. Davis, D. Drusvyatskiy, and L. Jiang, “Active manifolds, stratifications, and convergence to local minima in nonsmooth optimization,” *arXiv preprint arXiv:2108.11832*, 2021.

[48] T.-C. Kuo, “The ratio test for analytic whitney stratifications,” in *Proceedings of Liverpool Singularities—Symposium I*, pp. 141–149, Springer, 1971.

[49] T. Lê Loi, “Verdier and strict thom stratifications in o-minimal structures,” *Illinois Journal of Mathematics*, vol. 42, no. 2, pp. 347–356, 1998.

[50] A. S. Lewis, “Convex analysis on the hermitian matrices,” *SIAM Journal on Optimization*, vol. 6, no. 1, pp. 164–177, 1996.

[51] C. Davis, "All convex invariant functions of hermitian matrices," *Archiv der Mathematik*, vol. 8, no. 4, pp. 276–278, 1957.

[52] A. S. Lewis, "Derivatives of spectral functions," *Mathematics of Operations Research*, vol. 21, no. 3, pp. 576–588, 1996.

[53] A. S. Lewis and H. S. Sendov, "Nonsmooth analysis of singular values. part i: Theory," *Set-Valued Analysis*, vol. 13, no. 3, pp. 213–241, 2005.

[54] A. Daniilidis, A. Lewis, J. Malick, and H. Sendov, "Prox-regularity of spectral functions and spectral sets," *Journal of Convex Analysis*, vol. 15, no. 3, pp. 547–560, 2008.

[55] A. Daniilidis, D. Drusvyatskiy, and A. S. Lewis, "Orthogonal invariance and identifiability," *SIAM Journal on Matrix Analysis and Applications*, vol. 35, no. 2, pp. 580–598, 2014.

[56] D. Drusvyatskiy and C. Paquette, "Variational analysis of spectral functions simplified," *J. Convex Anal.*, vol. 25, no. 1, pp. 119–134, 2018.

[57] A. S. Lewis, "Nonsmooth analysis of eigenvalues," *Mathematical Programming*, vol. 84, no. 1, pp. 1–24, 1999.

[58] L. van den Dries and C. Miller, "Geometric categories and o-minimal structures," *Duke Math. J.*, vol. 84, no. 2, pp. 497–540, 1996.

[59] D. Drusvyatskiy, A. D. Ioffe, and A. S. Lewis, "Generic minimizing behavior in semialgebraic optimization," *SIAM J. Optim.*, vol. 26, no. 1, pp. 513–534, 2016.

[60] D. Davis and D. Drusvyatskiy, "Proximal methods avoid active strict saddles of weakly convex functions," *Foundations of Computational Mathematics*, pp. 1–46, 2021.

[61] R. Pemantle *et al.*, "Nonconvergence to unstable points in urn models and stochastic approximations," *The Annals of Probability*, vol. 18, no. 2, pp. 698–712, 1990.

[62] J. Borwein and X. Wang, "Lipschitz functions with maximal clarke subdifferentials are generic," *Proceedings of the American Mathematical Society*, vol. 128, no. 11, pp. 3221–3229, 2000.

[63] R. T. Rockafellar, "Favorable classes of lipschitz continuous functions in subgradient optimization," 1981.

[64] R. Poliquin and R. Rockafellar, "Prox-regular functions in variational analysis," *Transactions of the American Mathematical Society*, vol. 348, no. 5, pp. 1805–1838, 1996.

[65] R. A. Poliquin and R. T. Rockafellar, "Amenable functions in optimization," *Nonsmooth optimization: methods and applications (Erice, 1991)*, pp. 338–353, 1992.

[66] R. Fletcher, "A model algorithm for composite nondifferentiable optimization problems," in *Nondifferential and Variational Techniques in Optimization*, pp. 67–76, Springer, 1982.

[67] D. Davis, D. Drusvyatskiy, and C. Paquette, "The nonsmooth landscape of phase retrieval," *IMA Journal of Numerical Analysis*, vol. 40, no. 4, pp. 2652–2695, 2020.

[68] J. C. Duchi and F. Ruan, "Solving (most) of a set of quadratic equalities: Composite optimization for robust phase retrieval," *Information and Inference: A Journal of the IMA*, vol. 8, no. 3, pp. 471–529, 2019.

[69] Y. C. Eldar and S. Mendelson, "Phase retrieval: Stability and recovery guarantees," *Applied and Computational Harmonic Analysis*, vol. 36, no. 3, pp. 473–494, 2014.

[70] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?," *Journal of the ACM (JACM)*, vol. 58, no. 3, pp. 1–37, 2011.

[71] V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A. S. Willsky, "Rank-sparsity incoherence for matrix decomposition," *SIAM Journal on Optimization*, vol. 21, no. 2, pp. 572–596, 2011.

[72] C. Jin, P. Netrapalli, and M. Jordan, "What is local optimality in nonconvex-nonconcave minimax optimization?," in *International conference on machine learning*, pp. 4880–4889, PMLR, 2020.

[73] D. M. Ostrovskii, A. Lowy, and M. Razaviyayn, "Efficient search of first-order nash equilibria in nonconvex-concave smooth min-max problems," *SIAM Journal on Optimization*, vol. 31, no. 4, pp. 2508–2538, 2021.

[74] D. Drusvyatskiy and D. Davis, "Subgradient methods under weak convexity and tame geometry," *SIAG/OPT Views and News*, vol. 28, pp. 1–10, 2020.

[75] J. Lee, M. Simchowitz, M. Jordan, and B. Recht, "Gradient descent only converges to minimizers," in *Conference on learning theory*, pp. 1246–1257, 2016a.

[76] J. Lee, I. Panageas, G. Piliouras, M. Simchowitz, M. Jordan, and B. Recht, "First-order methods almost always avoid strict saddle points," *Math. Program.*, vol. 176, pp. 311–337, July 2019.

[77] D. Davis, D. Drusvyatskiy, S. Kakade, and J. D. Lee, "Stochastic subgradient method converges on tame functions," *Foundations of computational mathematics*, vol. 20, no. 1, pp. 119–154, 2020.

[78] P. Bianchi, W. Hachem, and S. Schechtman, "Stochastic subgradient descent escapes active strict saddles," *arXiv preprint arXiv:2108.02072*, 2021.

[79] M. Benaïm, J. Hofbauer, and S. Sorin, "Stochastic approximations and differential inclusions," *SIAM Journal on Control and Optimization*, vol. 44, no. 1, pp. 328–348, 2005.

[80] O. Brandière and M. Duflo, "Les algorithmes stochastiques contournent-ils les pièges ?," *Annales de l'I.H.P. Probabilités et statistiques*, vol. 32, no. 3, pp. 395–427, 1996.

[81] M. Benaim, "A dynamical system approach to stochastic approximations," *SIAM Journal on Control and Optimization*, vol. 34, no. 2, pp. 437–472, 1996.

[82] M. Benaïm, "Dynamics of stochastic approximation algorithms," in *Seminaire de probabilites XXXIII*, pp. 1–68, Springer, 1999.

[83] O. Brandiere, "Some pathological traps for stochastic approximation," *SIAM journal on control and optimization*, vol. 36, no. 4, pp. 1293–1314, 1998.

[84] S. Lee, S. J. Wright, and L. Bottou, "Manifold identification in dual averaging for regularized stochastic online learning.," *Journal of Machine Learning Research*, vol. 13, no. 6, 2012.

[85] R. Durrett, *Probability: theory and examples*, vol. 49. Cambridge university press, 2019.

[86] T. H. Gronwall, “Note on the derivatives with respect to a parameter of the solutions of a system of differential equations,” *Annals of Mathematics*, pp. 292–296, 1919.

[87] A. J. King and R. T. Rockafellar, “Asymptotic theory for solutions in statistical estimation and stochastic programming,” *Mathematics of Operations Research*, vol. 18, no. 1, pp. 148–162, 1993.

[88] A. Shapiro, “Asymptotic properties of statistical estimators in stochastic programming,” *The Annals of Statistics*, vol. 17, no. 2, pp. 841–858, 1989.

[89] J. Dupacová and R. Wets, “Asymptotic behavior of statistical estimators and of optimal solutions of stochastic optimization problems,” *The annals of statistics*, vol. 16, no. 4, pp. 1517–1549, 1988.

[90] A. L. Dontchev and R. T. Rockafellar, *Implicit functions and solution mappings*, vol. 543. Springer, 2009.

[91] J. Cutler, M. Díaz, and D. Drusvyatskiy, “Stochastic approximation with decision-dependent distributions: asymptotic normality and optimality,” *arXiv preprint arXiv:2207.04173*, 2022.

[92] Y. Nesterov, “Primal-dual subgradient methods for convex problems,” *Mathematical programming*, vol. 120, no. 1, pp. 221–259, 2009.

[93] L. Xiao, “Dual averaging method for regularized stochastic learning and online optimization,” *Advances in Neural Information Processing Systems*, vol. 22, 2009.

[94] S. Lee and S. J. Wright, “Manifold identification in dual averaging for regularized stochastic online learning,” *Journal of Machine Learning Research*, vol. 13, no. 55, pp. 1705–1744, 2012.

[95] Y. Nesterov, *Introductory lectures on convex optimization*, vol. 87 of *Applied Optimization*. Kluwer Academic Publishers, Boston, MA, 2004. A basic course.

[96] S. Bubeck *et al.*, “Convex optimization: Algorithms and complexity,” *Foundations and Trends® in Machine Learning*, vol. 8, no. 3-4, pp. 231–357, 2015.

[97] B. T. Polyak, “Minimization of unsmooth functionals,” *USSR Computational Mathematics and Mathematical Physics*, vol. 9, no. 3, pp. 14–29, 1969.

[98] A. Ioffe, "An invitation to tame optimization," *SIAM J. Optim.*, vol. 19, no. 4, pp. 1894–1917, 2009.

[99] R. Mifflin and C. Sagastizábal, "A $\mathcal{VU}$-algorithm for convex minimization," *Mathematical Programming*, vol. 104, no. 2, pp. 583–608, 2005.

[100] R. Mifflin, "Semismooth and semiconvex functions in constrained optimization," *SIAM J. Control Optim.*, vol. 15, no. 6, pp. 959–972, 1977.

[101] P. Wolfe, "A method of conjugate subgradients for minimizing nondifferentiable functions," in *Nondifferentiable optimization*, pp. 145–173, Springer, 1975.

[102] C. Lemarechal, "An extension of davidon methods to non differentiable problems," in *Nondifferentiable optimization*, pp. 95–109, Springer, 1975.

[103] W. d. Oliveira and C. Sagastizábal, "Bundle methods in the xxist century: A bird's-eye view," *Pesquisa Operacional*, vol. 34, pp. 647–670, 2014.

[104] X. Han and A. S. Lewis, "Survey descent: A multipoint generalization of gradient descent for nonsmooth optimization," *arXiv preprint arXiv:2111.15645*, 2021.

[105] A. Nemirovsky and D. Yudin, *Problem complexity and method efficiency in optimization*. A Wiley-Interscience Publication, John Wiley & Sons, Inc., New York, 1983. Translated from the Russian and with a preface by E. R. Dawson, Wiley-Interscience Series in Discrete Mathematics.

[106] A. Goldstein, "Optimization of lipschitz continuous functions," *Mathematical Programming*, vol. 13, no. 1, pp. 14–22, 1977.

[107] J. Zhang, H. Lin, S. Jegelka, S. Sra, and A. Jadbabaie, "Complexity of finding stationary points of nonconvex nonsmooth functions," *Proceedings of Machine Learning Research*, pp. 11173–11182, 2020.

[108] D. Davis, D. Drusvyatskiy, Y. T. Lee, S. Padmanabhan, and G. Ye, "A gradient sampling method with complexity guarantees for general lipschitz functions," *arXiv preprint arXiv:2112.06969*, 2022.

[109] H. Attouch, J. Bolte, P. Redont, and A. Soubeyran, "Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka-Łojasiewicz inequality," *Mathematics of operations research*, vol. 35, no. 2, pp. 438–457, 2010.

[110] 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*, vol. 137, no. 1, pp. 91–129, 2013.

[111] J. Bolte, S. Sabach, and M. Teboulle, "Proximal alternating linearized minimization for nonconvex and nonsmooth problems," *Mathematical Programming*, vol. 146, no. 1, pp. 459–494, 2014.

[112] Y. Xu and W. Yin, "A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion," *SIAM Journal on imaging sciences*, vol. 6, no. 3, pp. 1758–1789, 2013.

[113] H. Attouch and J. Bolte, "On the convergence of the proximal algorithm for nonsmooth functions involving analytic features," *Mathematical Programming*, vol. 116, pp. 5–16, 2009.

[114] J. Bolte, A. Daniilidis, and A. Lewis, "The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems," *SIAM J. Optim.*, vol. 17, no. 4, pp. 1205–1223 (electronic), 2006.

[115] D. Davis, D. Drusvyatskiy, S. Kakade, and J. D. Lee, "Stochastic subgradient method converges on tame functions," *Foundations of computational mathematics*, vol. 20, no. 1, pp. 119–154, 2020.

[116] D. Davis, D. Drusvyatskiy, K. J. MacPhee, and C. Paquette, "Subgradient methods for sharp weakly convex functions," *Journal of Optimization Theory and Applications*, vol. 179, pp. 962–982, 2018.

[117] P. R. Johnstone and P. Moulin, "Faster subgradient methods for functions with hölderian growth," *Mathematical Programming*, vol. 180, no. 1-2, pp. 417–450, 2020.

[118] D. Drusvyatskiy, A. D. Ioffe, and A. S. Lewis, "Nonsmooth optimization using taylor-like models: error bounds, convergence, and termination criteria," *Mathematical Programming*, vol. 185, pp. 357–383, 2021.

[119] D. Davis and D. Drusvyatskiy, "Subgradient methods under weak convexity and tame geometry," *SIAG/OPT Views and News*, vol. 28, no. 1, pp. 1–10, 2020.

[120] A. S. Lewis and S. Zhang, "Partial smoothness, tilt stability, and generalized hessians," *SIAM Journal on Optimization*, vol. 23, no. 1, pp. 74–94, 2013.

[121] D. Davis, D. Drusvyatskiy, and V. Charisopoulos, “Stochastic algorithms with geometric step decay converge linearly on sharp functions,” *arXiv preprint arXiv:1907.09547*, 2019.

[122] J. Bolte, T. P. Nguyen, J. Peypouquet, and B. W. Suter, “From error bounds to the complexity of first-order descent methods for convex functions,” *Mathematical Programming*, vol. 165, pp. 471–507, 2017.

[123] A. Lewis and C. Wylie, “A simple newton method for local nonsmooth optimization,” *arXiv preprint arXiv:1907.11742*, 2019.

[124] A. S. Lewis and M. L. Overton, “Nonsmooth optimization via quasi-newton methods,” *Mathematical Programming*, vol. 141, no. 1, pp. 135–163, 2013.

[125] S. Burer and J. Lee, “Solving maximum-entropy sampling problems using factored masks,” *Mathematical Programming*, vol. 109, no. 2, pp. 263–281, 2007.

[126] K. M. Anstreicher and J. Lee, “A masked spectral bound for maximum-entropy sampling,” in *mODa 7—Advances in Model-Oriented Design and Analysis*, pp. 1–12, Springer, 2004.

[127] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, *et al.*, “Pytorch: An imperative style, high-performance deep learning library,” *Advances in neural information processing systems*, vol. 32, 2019.

[128] J. V. Burke, “Descent methods for composite nondifferentiable optimization problems,” *Mathematical Programming*, vol. 33, no. 3, pp. 260–279, 1985.

[129] P. J. Enright and B. A. Conway, “Discrete approximations to optimal trajectories using direct transcription and nonlinear programming,” *Journal of Guidance, Control, and Dynamics*, vol. 15, no. 4, pp. 994–1002, 1992.

[130] S. J. Wright, “Convergence of an inexact algorithm for composite nonsmooth optimization,” *IMA journal of numerical analysis*, vol. 10, no. 3, pp. 299–321, 1990.

[131] Y.-X. Yuan, “On the superlinear convergence of a trust region algorithm for nonsmooth optimization,” *Mathematical Programming*, vol. 31, no. 3, pp. 269–285, 1985.

[132] D. Drusvyatskiy and A. Lewis, "Error bounds, quadratic growth, and linear convergence of proximal methods," *To appear in Math. Oper. Res., arXiv:1602.06661*, 2016.

[133] Y. Chen, Y. Chi, and A. J. Goldsmith, "Exact and stable covariance estimation from quadratic sampling via convex programming," *IEEE Transactions on Information Theory*, vol. 61, no. 7, pp. 4034–4059, 2015.

[134] 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*, vol. 21, no. 6, pp. 1505–1593, 2021.

[135] L. Ding, L. Jiang, Y. Chen, Q. Qu, and Z. Zhu, "Rank overspecified robust matrix recovery: Subgradient method and exact recovery," *arXiv preprint arXiv:2109.11154*, 2021.

[136] J. Ma and S. Fattahi, "Global convergence of sub-gradient method for robust matrix recovery: Small initialization, noisy measurements, and over-parameterization," *Journal of Machine Learning Research*, vol. 24, no. 96, pp. 1–84, 2023.

[137] G. Zhang, S. Fattahi, and R. Y. Zhang, "Preconditioned gradient descent for overparameterized nonconvex burer–monteiro factorization with global optimality certification," *Journal of Machine Learning Research*, vol. 24, no. 163, pp. 1–55, 2023.

[138] X. Xu, Y. Shen, Y. Chi, and C. Ma, "The power of preconditioning in overparameterized low-rank matrix sensing," in *International Conference on Machine Learning*, pp. 38611–38654, PMLR, 2023.

[139] Y. Li, T. Ma, and H. Zhang, "Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations," in *Conference On Learning Theory*, pp. 2–47, PMLR, 2018.

[140] D. Stöger and M. Soltanolkotabi, "Small random initialization is akin to spectral learning: Optimization and generalization guarantees for overparameterized low-rank matrix reconstruction," *Advances in Neural Information Processing Systems*, vol. 34, pp. 23831–23843, 2021.

[141] L. Ding, Z. Qin, L. Jiang, J. Zhou, and Z. Zhu, "A validation approach to over-parameterized matrix and image recovery," *arXiv preprint arXiv:2209.10675*, 2022.

[142] L. Jiang, Y. Chen, and L. Ding, "Algorithmic regularization in model-free overparametrized asymmetric matrix factorization," *SIAM Journal on Mathematics of Data Science*, vol. 5, no. 3, pp. 723–744, 2023.

[143] J. Jin, Z. Li, K. Lyu, S. S. Du, and J. D. Lee, "Understanding incremental learning of gradient descent: A fine-grained analysis of matrix sensing," in *International Conference on Machine Learning*, pp. 15200–15238, PMLR, 2023.

[144] M. Soltanolkotabi, D. Stöger, and C. Xie, "Implicit balancing and regularization: Generalization and convergence guarantees for overparameterized asymmetric matrix sensing," in *The Thirty Sixth Annual Conference on Learning Theory*, pp. 5140–5142, PMLR, 2023.

[145] J. Bolte, S. Sabach, and M. Teboulle, "Proximal alternating linearized minimization for nonconvex and nonsmooth problems," *Mathematical Programming*, vol. 146, pp. 459–494, Aug 2014.

[146] H. Robbins and D. Siegmund, "A convergence theorem for non negative almost supermartingales and some applications," in *Optimizing methods in statistics*, pp. 233–257, Elsevier, 1971.

[147] L. H. Y. Chen, "A Short Note on the Conditional Borel-Cantelli Lemma," *The Annals of Probability*, vol. 6, no. 4, pp. 699 – 700, 1978.

[148] A. Dembo, "Lecture notes on probability theory: Stanford statistics 310," *Accessed October*, vol. 1, p. 2016, 2016.

[149] D. Drusvyatskiy and A. S. Lewis, "Optimality, identifiability, and sensitivity," *arXiv:1207.6628v1*, 2012.

[150] H. Asi and J. C. Duchi, "Stochastic (approximate) proximal point methods: Convergence, optimality, and adaptivity," *SIAM Journal on Optimization*, vol. 29, no. 3, pp. 2257–2290, 2019.

[151] A. W. Van Der Vaart and J. A. Wellner, "Weak convergence and empirical processes: with applications to statistics," 1996.