THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

École doctorale : MSTII - Mathématiques, Sciences et technologies de l'information, Informatique
Spécialité : Mathématiques Appliquées
Unité de recherche : Laboratoire Jean Kuntzmann

Optimisation non-lisse structurée : identification proximale, convergence locale rapide et applications

Structured nonsmooth optimization: proximal identification, fast local convergence, and applications

Présentée par :

Gilles BAREILLES

Direction de thèse :

Jerome MALICK CNRS	Directeur de thèse
Franck IUTZELER MAITRE DE CONFERENCES, Université Grenoble Alpes	Co-directeur de thèse

Rapporteurs :

CLAUDIA ALEJANDRA SAGASTIZABAL
Professeur, Universidade Estadual de Campinas
JALAL FADILI
Professeur des Universités, ENSI CAEN

Thèse soutenue publiquement le **2 décembre 2022**, devant le jury composé de :

JERÔME MALICK Directeur de recherche, CNRS DELEGATION ALPES	Directeur de thèse
CLAUDIA ALEJANDRA SAGASTIZABAL Professeur, Universidade Estadual de Campinas	Rapporteure
JALAL FADILI Professeur des Universités, ENSI CAEN	Rapporteur
JEAN-CHARLES GILBERT Directeur de recherche, INRIA CENTRE DE PARIS	Examinateur
MATHURIN MASSIAS Chargé de recherche, CENTRE INRIA DE LYON	Examinateur
NADIA BRAUNER Professeur des Universités, UNIVERSITE GRENOBLE ALPES	Présidente
FRANCK IUTZELER Maître de conférences HDR, UNIVERSITE GRENOBLE ALPES	Co-directeur de thèse

Invités :

CLAUDE LEMARECHAL
Directeur de recherche émérite, INRIA CENTRE GRENOBLE RHONE-ALPES

PREFACE

This thesis deals with the optimization of structured nonsmooth functions, which appear for example in machine learning and signal processing. In particular, we consider matrix functions which feature eigenvalue functions and the nuclear norm. Our approach consists in exploiting the structure of these nonsmooth functions to design algorithms that converge fast – at the speed of Newton's method, thus yielding high precision estimates of nonsmooth minimizers.

More precisely, the nondifferentiability points of structured functions organize in smooth manifolds, such that the nonsmooth function is smooth along the manifold and nonsmooth across it. In this thesis, we propose optimization algorithms that *detect* and *exploit* these structure manifolds. The two key ingredients in our approach are *(i)* subtle geometrical properties of the proximal operator, and *(ii)* algorithmic tools from smooth constrained programming. We operate without assuming knowledge of the optimal structure manifold, and use variational analysis tools to cover both convex and nonconvex settings.

We first consider the minimization of the sum of a smooth function and a nonsmooth function, which encompasses sparse regression problems such as the "lasso". We show that the proximal-gradient operator, well-known for its minimization properties, also identifies relevant structure manifolds. We propose an algorithm that combines this identification information with tools from Riemannian optimization, and prove that it converges locally quadratically. We illustrate numerically this fast convergence on classical learning problems.

We also consider the minimization of the composition between a smooth map and a nonsmooth function. This setting encompasses the minimization of the largest eigenvalue of a smoothly parameterized symmetric matrix. We introduce and characterize a proximal identification tool that detects relevant structure manifolds near arbitrary points. We propose an algorithm that combines this tool with Newton steps for smooth constrained minimization. We prove that, when started near a minimizer, the algorithm exactly identifies its optimal manifold and converges quadratically. We compare our algorithm with state-of-the-art algorithms for nonsmooth optimization.

Thus, the proximal identification procedures proposed in this thesis detect efficiently the relevant manifolds of additive and composite nonsmooth functions. The obtained algorithms are carefully implemented in the Julia language and are released as open-source packages.

RÉSUMÉ

Cette thèse traite de l'optimisation de fonctions non-différentiables structurées, qui apparaissent notamment en apprentissage statistique et en traitement du signal. En particulier, nous considérons des fonctions matricielles, qui mettent en jeu les valeurs propres ou la norme nucléaire. Notre approche consiste à exploiter la structure de ces fonctions non-différentiables pour développer des algorithmes qui convergent rapidement – à la vitesse de la méthode de Newton – et qui retournent ainsi des estimations précises des solutions.

Plus précisément, les points de non-différentiabilité des fonctions structurées s'organisent en variétés différentiables, qui captent les directions de différentiabilité dans l'espace tangent et les directions de non-différentiabilité dans l'espace normal. Dans cette thèse, nous proposons des algorithmes qui *détectent* et *exploitent* ces variétés de structure. Les deux outils clés pour notre approche sont *(i)* des propriétés géométriques fines de l'opérateur proximal, et *(ii)* les méthodes algorithmiques de l'optimisation sous contraintes. Nous raisonnons sans supposer connue la variété optimale, et utilisons les outils de l'analyse variationnelle pour couvrir simultanément les cas convexes et non-convexes.

Nous considérons d'abord la minimisation de la somme d'une fonction différentiable et d'une fonction non-différentiable, cadre qui inclut notamment les problèmes de régression parcimonieuse tels que le "lasso". Nous montrons que l'opérateur gradient-proximal, connu pour ses propriétés de minimisation, identifie aussi les variétés de non-différentiabilité pertinentes. Nous proposons un algorithme qui combine ce résultat d'identification avec des outils de l'optimisation Riemannienne, et montrons qu'il converge localement quadratiquement. Cette convergence rapide est illustrée en pratique sur des problèmes d'apprentissage classiques.

Nous considérons ensuite la minimisation de la composition entre une application différentiable et une fonction non-différentiable. Ce cadre couvre notamment la minimisation de la valeur propre maximale d'une matrice symétrique paramétrée. Nous introduisons et caractérisons un outil d'identification proximale, qui détecte les variétés de non-différentiabilité autour de tout point. Nous montrons que cet outil peut être combiné avec des itérations de Newton de l'optimisation différentiable sous contrainte. Nous démontrons que l'algorithme obtenu détecte localement la variété d'un minimiseur et converge quadratiquement. Nous comparons notre algorithme avec l'état de l'art pour l'optimisation non-différentiable.

Ainsi, les procédés d'identification proximale proposés dans cette thèse sont à même de détecter efficacement les variétés pertinentes des fonctions non-différentiables additives et composites. Les algorithmes obtenus, ainsi que les ressorts numériques sur lesquels ils reposent, sont mis à disposition de la communauté sous forme de paquets Julia open source.

REMERCIEMENTS

Merci à Claudia et Jalal d'avoir accepté de rapporter cette thèse, vous dont les travaux ont contribué à guider mes pas pendant ces trois années. Merci Claudia pour tes retours positifs et ton enthousiasme précieux. Merci Jalal pour ton suivi et pour la richesse de nos discussions, j'apprends toujours beaucoup de nos échanges. Merci, Mathurin et Jean-Charles, d'avoir accepté d'examiner cette thèse. Vos travaux m'ont apporté beaucoup d'outils. Merci enfin Nadia d'avoir accepté de présider la soutenance, j'en suis très heureux.

Merci Franck, pour avoir proposé ce sujet de stage très opportun, mais aussi et surtout pour ta capacité à partager ta passion et ta vision des maths. Merci Jérôme, pour ton exigence, ton humour et ta bienveillance sans failles. J'en apprends toujours quand on discute, que ce soit de fonctions spectrales ou de chanteurs des années 80. Vous m'avez guidé de façon complémentaire, et j'ai apprécié la liberté progressive que vous m'avez accordé au fil du temps. Merci pour votre confiance.

Scientifiquement, je dois beaucoup aux collègues du LJK. Je pense notamment à Anatoli, Roland, Panayotis et Sergueï, dont j'ai beaucoup appris. Merci aussi à Yassine, Mytia, Yu-Guan, Waïss, Victor, Ieva, Anatole, François et Omar pour des discussions et des retours scientifiques toujours au top.

Je pense aussi à mes amis et collègues de Grenoble, qui ont été tour à tour compagnons de randonnée, d'escalade, de course d'arrête improvisée, de chœur, de grimpe, de ski, de soirée, de bureau, de musique, de discussion, de coloc, de GR20 en 8 jours devenus 7.5 et j'en passe. Merci donc à Yassine, Mytia, Yu-Guan, Sélim, Waïss, Victor, Nils, Eloa, Julien, Anatole, Hubert, Hyppolithe, François, Kévin, Ieva, Xiao, Thibault, Simon, Aurélien et tout ceux que je ne cite pas mais qui se reconnaîtront.

La musique a été une part importante de ma vie à Grenoble. Un immense merci à François, pour micromégas, à Daniel, pour ces quelques jours à Saou, à Stéphane et la bande à Pierrot, pour la semaine aux Roux. Merci infiniment, Christine, pour avoir réussi à m'embarquer dans toutes ces aventures riches de partage, d'amitié et d'estime. J'en profite pour saluer Îves, Benoît et Léa, avec qui tout à commencé.

Je pense aussi à mes amis d'école : François, Antoine, Ali, les Joyeux Lurons, Benjamin et Tristan – abondance !, Arthur, Juliette, Damien, Damien et Mocia.

Enfin, je réserve ces derniers mots à ma famille. À mes parents et à mon frère, qui ont su accompagner mes pas, à défaut de toujours les comprendre. Et À Alice, qui sait accompagner mes pas, et m'aide à les comprendre.

CONTENTS

1

INTRODUCTION

This section serves as an overview of the content of this manuscript. We first briefly introduce the rich domain of nonsmooth optimization. We then lay down the lines of research our work inherits from. We finally give a detailed summary of the contributions of this thesis.

1.1 NONSMOOTH OPTIMIZATION: MOTIVATIONS & FIRST ORDER METHODS

We consider optimization problems of the form

$$\min_{x \in \mathbb{R}^n} F(x) \qquad \text{with} \qquad F : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$$

where F is *not* differentiable everywhere. Throughout this thesis, we call such functions *nonsmooth*.[1] Such nonsmooth optimization problems abound in machine learning, signal processing, and operations research among other fields.

In this introductory section, we first review the main sources of nonsmoothness in applications. Then, we turn our attention to two kinds of first-order methods for minimizing nonsmooth functions. These methods are the counterparts of the gradient method, when the gradient of the objective does not exist everywhere.

1.1.1 *Faces of nonsmoothness*

In applications, nonsmoothness is produced by some special operations. We review three families of nonsmoothness: implicit, chosen, and in-between.

IMPLICIT NONSMOOTHNESS. Nonsmoothness often appears implicitly: the function to minimize is defined implicitly as the result of another computation. Typically, it is the solution of an optimization subproblem: for a given $x \in \mathbb{R}^n$,

$$F(x) = \sup_{u \in U} h(x, u). \tag{1.1}$$

Such nonsmooth functions appear in particular in the decomposition of large-scale or complex problems with, e.g., Lagrangian relaxation, Benders decomposition, or resources decomposition; see Briant et al. (2008) for applications in combinatorial optimization and Bonnans et al. (2006, Sec. 8.2-3) for a review of decomposition schemes. Such nonsmoothness also emerges naturally

A part of my team, DAO https://dao-ljk.imag.fr, *works on distributionally robust optimization, which produces problem of type* (1.1); *see Kuhn et al.* (2019) *for applications in machine learning.*

1 Depending on the literature, "smooth" mean functions which are either differentiable, $\mathcal{C}^\infty$, or have Lipschitz continuous gradient. Here, we will use smooth for functions differentiable everywhere, out of coherence with the term "nonsmooth".

in robust and risk-averse optimization, see e.g., the monograph Ben-Tal et al. (2009).

If $h(\cdot, u)$ is convex for all $u \in U$, then F is convex as well. In that case, for a given x, solving problem (1.3) provides the function value $F(x)$ *and* a subgradient $v \in \partial F(x)$, both potentially noisy if the subproblem is solved approximately. This subgradient gives some local first-order information on the function, akin to the gradient of smooth functions. We come back to this setting in Section 1.1.2.

This is the so-called "black-box" setting.

CHOSEN NONSMOOTHNESS. In sharp contrast with implicit nonsmoothness, many optimization problems explicitly incorporate some *chosen* nonsmoothness. This is notably the case for inverse problems stemming from statistical learning and image processing applications, which formalize as nonsmooth (convex) optimization problems; see e.g., the monograph Scherzer et al. (2009). There, the nonsmoothness is willingly added to the problem because it ensures some "low complexity" of the solutions.

A popular example is that of linear regression with ℓ_1-norm regularization (Tibshirani, 1996), called lasso in the machine learning community:

$$\min_{x \in \mathbb{R}^n} \frac{1}{2}\|Ax - b\|_2^2 + \lambda \|x\|_1. \tag{lasso}$$

Adding a multiple of the ℓ_1-norm to the smooth "data-fitting" term ensures that the learned vector is sparse, and that it benefits from improved statistical properties (Candès et al., 2006). In the same spirit, Bach (2008); Candès and Recht (2013) follow the seminal work Fazel et al. (2001) and consider matrix learning problems regularized with the trace norm (the sum of the singular value of the matrix). The obtained solution has again a low complexity — its rank is low, and it benefits from improved statistical guarantees.

This setting encompasses e.g., matrix completion and recommender systems tasks; see e.g., Koren et al. (2009).

More generally, many inverse problems formalize as the minimization of

$$F(x) = f(x) + g(x), \tag{1.2}$$

where f is a smooth function and g is a nonsmooth function; see the review paper Vaiter et al. (2015). In such problems, nonsmoothness is crucial: it is chosen (and sometimes crafted) to enforce a low complexity property on minimizers. This explicit nonsmoothness can also simplify optimization.

The nonsmoothness of these problems is usually localized in one or several simpler elements of the split objective function. When simple enough, this nonsmooth function admits a tractable "proximity operator", akin to a gradient step in smooth optimization. Handling each element separately then allows to build optimization methods minimizing nonsmooth objective functions such as e.g., Forward Backward for (1.2), or Douglas-Rachford, ADMM; see Bauschke and Combettes (2017). We return to the proximity operator and this setting (1.2) in Section 1.1.2.

IN BETWEEN. It also happens that the nonsmoothness is not chosen while still being explicit. This is the case when the nonsmooth problem writes as

$$F(x) = g \circ c(x), \tag{1.3}$$

where g is a nonsmooth function and c a smooth mapping. This setting encompasses a wide range of problems and applications listed in Shapiro (2003); Lewis and Wright (2016) e.g., penalty functions of nonlinear programming,

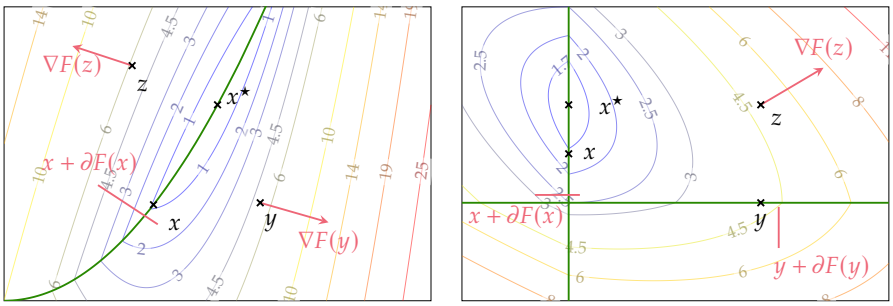


Figure 1.1: Illustration of the level lines and the subdifferentials of two nonsmooth functions of the plane. They are introduced in Example 1.1, we plot F_1 on the left, and F_2 on the right. The subdifferential is displayed at point x, y, and z. Depending on the smoothness of the function, it is either a single vector, the gradient of the function (e.g., for z), or a full set (e.g., for x).

robust regression which includes phase synchronization, and optimal control of helicopters (Apkarian et al., 2004).

In particular, we consider in this thesis the maximal eigenvalue function:

$$F(x) = \lambda_{\max} \circ c(x),$$

where $c(x)$ is a symmetric real matrix that depends smoothly on x. We place this function in the class of problems with explicit nonsmoothness since $\lambda_{\max}$ is a convex function, with known subdifferential and second-order information (Shapiro and Fan, 1995). This function appears in problems stemming from applications in control, learning or operations research such as matrix completion, community detection, phase retrieval. For instance, semidefinite programs with constant trace can be written in such a form (Helmberg and Rendl, 2000, Sec. 2-3).

A common element in problems (1.2) and (1.3) is the explicit nonsmoothness, which allows at a given point to compute more than one arbitrary subgradient: the full subdifferential, and even some second-order information, is available. We illustrate this idea with two nonsmooth functions in the following example, which will accompany us throughout the introduction.

Example 1.1 (Examples with explicit nonsmoothness). Throughout the introduction, we illustrate our discussion on two functions from $\mathbb{R}^2$ to $\mathbb{R}$ which admit an explicit nonsmoothness:

$$F_1(x) = 10(x_1 - 1)^2 + 5|x_1^2 - x_2|,$$

and F_2 a two dimensional lasso function. Both functions are illustrated in (1.1), where we show the level lines and the subdifferential at three points x, y and z where F is smooth or nonsmooth. □

1.1.2 *Basic nonsmooth algorithms*

SUBGRADIENT METHODS. At nonsmooth points, the notion of "subdifferential" replaces and generalizes that of gradient (Hiriart-Urruty and Lemaréchal, 1993; Rockafellar and Wets, 1998). This set contains all "subgradients" at a given point, that is, roughly speaking, the slopes of all linear tangent under-approximations

of the function at that point. Figure 1.1 shows the subdifferential of two simple functions at smooth and nonsmooth points.

One subgradient provides only partial first-order information.

The simplest algorithm to minimize a nonsmooth function is the subgradient method, which mimics the gradient descent method: $x_{k+1} = x_k - \gamma_k v_k / \|v_k\|$, with $v_k \in \partial F(x_k)$ and $\gamma_k > 0$ such that γ_k vanishes and $\sum \gamma_k$ diverges (Nesterov, 2018, Th. 3.2.2). This method is simple, but converges at a sublinear rate (Beck, 2017, Th. 8.13). More efficient and practical algorithms in the convex case include the bundle method (Hiriart-Urruty and Lemaréchal, 1993, Chap. XV), where one iteratively refines and leverages a model of the nonsmooth function by successive black box oracle calls. The bundle algorithm benefits from theoretical convergence guarantees to critical points, is backed by several decades of research, and is used in some important industrial applications (Hechme-Doukopoulos et al., 2010). Leveraging the fact that a Lipschitz function is almost everywhere differentiable, a gradient sampling algorithm was shown to converge to critical points,[2] albeit with a strong iteration cost; see the review article Burke et al. (2020). Besides, it was noted in Lewis and Overton (2013) that the quasi-Newton BFGS method converges well — often linearly — on nonsmooth functions, even though no theoretical guarantees are available.

PROXIMAL METHODS. Another central notion in nonsmooth optimization is the proximity operator of a nonsmooth function. This operator is defined, for $v \in \mathbb{R}^n$ and a proximal step $\gamma > 0$, by

$$\operatorname{prox}_{\gamma g}(x) = \operatorname*{arg\,min}_{u \in \mathbb{R}^n} \left\{ g(u) + \frac{1}{2\gamma} \|u - x\|^2 \right\}. \tag{1.4}$$

The proxmial operator thus uses the full first-order information, which is more than one (arbitrary) subgradient.

We discuss this definition and some characterizations in Section 2.2. This operator acts as an *implicit* (subgradient) step. As such, it has the nice property of ensuring some minimal functional descent between input and output point; see Lemma A.5 for a precise statement and references. The proximal point algorithm, defined by $x_{k+1} = \operatorname{prox}_{\gamma g}(x_k)$, thus generates a sequence of points which converge to minimizers, albeit with a slow sublinear rate; see Güler (1991) for precise results in the convex case.

However, the proximal point algorithm is rather impractical in general. Indeed, each iteration consists in applying the proximal operator to the current iterate, which amounts to solving one optimization problem. Interestingly, bundle methods can be seen as an implementable form of approximate proximal point algorithm (Correa and Lemaréchal, 1993).

See the numerous examples of `proximity-operator.net`

For simple enough functions, the proximal operator is easy to compute, and sometimes even available in closed form; see e.g., Example 1.2. This includes popular functions such as the ℓ_1 norm and the trace norm (Bach et al., 2012). This remark is key in building prox-like operators for certain nonsmooth functions which do not admit an explicit prox but have some additional structure, as we discuss in the next section.

2 Convergence occurs with probability one, if the iterative process never encounters a nondifferentiable point.

Example 1.2 (Closed-form proximity operator). The ℓ_1 norm and the maximum function are nonsmooth but simple enough to have explicit proximity operators. Indeed, for $y \in \mathbb{R}^n$, their coordinate-wise expression is

$$[\mathrm{prox}_{\gamma\|\cdot\|_1}(y)]_i = \begin{cases} y_i + \gamma & \text{if } y_i < -\gamma \\ 0 & \text{if } |y_i| \leq \gamma \\ y_i - \gamma & \text{if } y_i > \gamma \end{cases} \qquad \left[\mathrm{prox}_{\gamma \max}(y)\right]_i = \begin{cases} s & \text{if } y_i > s \\ y_i & \text{else} \end{cases}$$

where s is the unique real number such that $\sum_{\{i:y_i>s\}}(y_i - s) = \gamma$. □

1.2 STRUCTURED NONSMOOTHNESS

As shown by their wide use in machine learning and signal processing, the first order methods discussed above work well in practice. However, they reach their limit when one requires minimizers *with high precision*. This is relevant for instance in learning problems (1.2), where the low complexity of the solutions is valuable for practitioners: giving guarantees on this low complexity requires having a high precision estimation on the nonsmooth minimizer. This is also relevant in control applications (1.3), where the precision of the minimizer affects the performance of the ensuing controller.

In this section, we look at two types of structure in nonsmooth functions: first, additive and composite expressions in Section 1.2.1, and second, the existence of a smooth substructure in Section 1.2.2. We discuss the algorithmic improvements that each of these two structures bring on the first-order methods presented above. This sets the tone for the discussions of the next sections, where we present how combining these two structures lead to locally fast nonsmooth algorithms.

This presentation of this section draws from Sagastizábal (2011).

1.2.1 *Additive and composite structure*

When the objective function to minimize is explicit, the nonsmoothness can often be isolated in one simpler component. We consider in this thesis two type of functions: *additive* nonsmooth functions (1.2), which write as the sum of a smooth and a nonsmooth function, and *composite* functions (1.3), that is functions that write as a composition of a smooth mapping and a nonsmooth function. For example, the functions illustrated in Fig. 1.1 both write as the sum of a smooth and a nonsmooth function, and as well as a composition between a smooth mapping and a nonsmooth function.

We now detail how the first-order methods can be improved with this additional information.

ADDITIVE FUNCTIONS. Bundle methods are able to exploit the additive structure of (1.2): the model of F combines a subgradient-based model, specific to the nonsmooth function g, with a second-order Taylor expansion of the smooth term f (Lemaréchal et al., 2007).

If, in addition, the nonsmooth term admits an explicit proximal operator, one can compute the *proximal-gradient* operator of the objective function, defined as

Proximal gradient

$$x_{k+1} = \mathrm{prox}_{\gamma g}(x_k - \gamma \nabla f(x_k)).$$

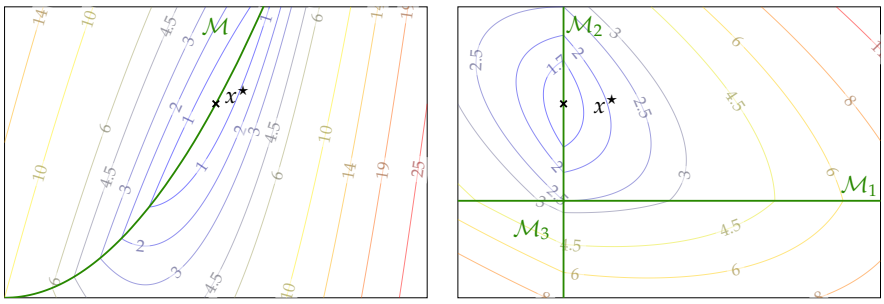


Figure 1.2: Illustration of the smooth substructure(s) of functions F_1 and F_2, introduced in Example 1.1. The smooth substructure manifolds are represented in green; their expression is given in Example 1.3.

Accelerated proximal gradient / FISTA

Iterating this operator yields a first order algorithm akin to the proximal point algorithm: in the convex case, it converges from any starting point at a sublinear rate (Beck, 2017, Th. 10.15, 10.21). It is popular to equip the proximal gradient with a computationally cheap inertial term (Nesterov, 1983), sometimes called "Nesterov acceleration", which improves its sublinear rate for convex functions. This is the setting of Chapter 3, in which we introduce a *Newton* acceleration of the proximal gradient.

Composite bundle

COMPOSITE FUNCTIONS. When the function writes as a composition (1.3), bundle methods can also be adapted (Sagastizábal, 2013). Again, one builds a model of the (simpler) nonsmooth term and use derivatives of the smooth mapping, which results in a finer model of the whole objective function. In specific cases, notably when the nonsmooth term is piece-wise linear, it was observed that the sequence of serious steps converges at a superlinear rate. This is the kind of result we aim at in this thesis.

Proxlinear method

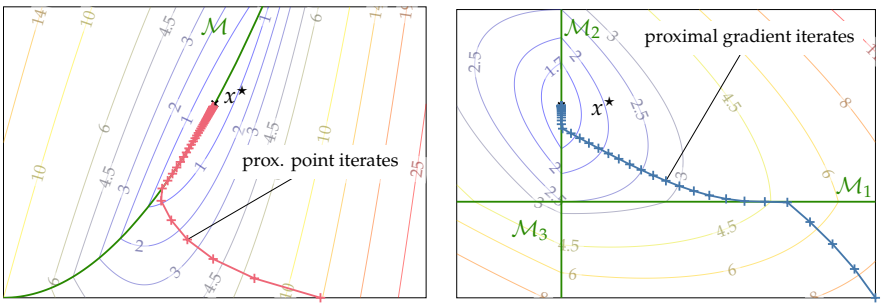
One can also consider *proxlinear* methods (Lewis and Wright, 2016). Each iteration writes $x_{k+1} = \arg\min_{u\in\mathbb{R}^n}\{g\left(c(x_k) + \mathrm{Jac}_c(x_k)\cdot(u - x_k)\right) + \frac{\mu}{2}\|u - x_k\|^2\}$, where Jac_c denotes the Jacobian of c. Note the resemblance with the proximity operator, the only difference lying in the linearization of the smooth element c. In this sense, the proxlinear operator can be seen as the counterpart of the proximal gradient operator for composite functions. This subproblem is rarely explicit, its solution needs to be approximated. In a weakly convex setting, Drusvyatskiy and Paquette (2019) analyze the complexity of the method: they show that, when the nonsmooth subproblem is smoothed and solved by a gradient-based method, the global complexity is sublinear, and just slightly worse than that of gradient descent. While this approach fully uses the composite nature of the problem, the smoothing of the subproblem alters the nice smooth substructure induced by the nonsmoothness, which we now discuss.

1.2.2 *Smooth substructure*

In many examples of interest, including (1.2), and (1.3) under some geometrical assumptions, the nonsmooth function to minimize usually locally exhibits a *smooth substructure*. More precisely, there is near a point a smooth submanifold such that the function is

(i) smooth when restricted to the manifold, and

(ii) nonsmooth in all directions normal to the manifold.



(a) Proximal point iterates **(b)** Proximal gradient iterates

Figure 1.3: Illustration of the identification of the proximal point (left pane) and the proximal gradient (right pane) on functions introduced in Example 1.1: the iterates eventually reach and remain on the smooth substructure of the minimizer.

A. Lewis formalizes this idea with the notion of *partial smoothness* (Lewis, 2002), which will be central in our work; see the precise definition in Section 2.4. Many classes of functions can be shown to be partly smooth under some natural assumptions, including problems (1.2) and (1.3), among other examples in Lewis (2002, Sec. 3). We illustrate in the next example and Fig. 1.2 such smooth substructures on two simple functions.

Example 1.3 (Smooth substructure). We illustrate the notion of smooth substructure on functions F_1 and F_2, introduced in Example 1.1. The partial smoothness manifolds presented here are displayed in Fig. 1.2.

Function F_1 admits a unique partial smoothness manifold $\mathcal{M} = \{x \in \mathbb{R}^2 : x_1^2 = x_2\}$: either $x \in \mathcal{M}$, and F_1 is partly smooth at x relative to $\mathcal{M}$, or $x \in \mathbb{R}^2 \setminus \mathcal{M}$, and F_1 is smooth at x (or partly smooth relative to the manifold $\mathbb{R}^2$).

Function F_2 is nonsmooth at points for which one coordinate is null. F_2 admits several partial smoothness manifolds: $\mathcal{M}_1 = \mathbb{R} \times \{0\}$, $\mathcal{M}_2 = \{0\} \times \mathbb{R}$, and $\mathcal{M}_3 = \{0\} \times \{0\}$. The partial smoothness of a point is given by its support e.g., at point $x = (0, t)$ with $t \neq 0$, the function is partly smooth relative to $\mathcal{M}_2$.□

IDENTIFICATION OF PROX-BASED ALGORITHMS. When F admits a smooth substructure, the proximity operator (1.4) gets an additional property: it *maps neighborhoods of a minimizer to its smooth substructure manifold*. For instance, the proximity operator of the ℓ_1-norm, recalled in Example 1.2, sets any coordinate small in absolute value to zero. As a result, the output of the prox is *exactly* a nondifferentiable point of the objective function. We illustrate this behavior in the forthcoming Fig. 1.4, and discuss it precisely in Section 2.4.

Operator identification

As a consequence, the iterates of the proximal point algorithm end up *exactly* on the smooth manifold, as illustrated in Fig. 1.3a. This behavior is the so-called *identification* property of the proximal point algorithm (Daniilidis et al., 2006, Th. 28). Actually, most prox-based algorithms also identify the structure of minimizers; see e.g., the review paper Iutzeler and Malick (2020, Sec. 4-5). For instance, Fig. 1.3b illustrates the identification of the proximal-gradient algorithm on a lasso problem.

Algorithmic identification

The *proximal identification* mechanism is strong: it exactly maps points to the structure manifold. Other identification mechanisms are known; for instance, the proxlinear method implicitly "detects" the smooth substructure of minimizers, even though its iterates never actually lie exactly on it; see Lewis and Wright (2016, Sec. 4.5). In Chapter 4, we will develop an *indirect identification*

for composite functions, following a similar idea. Besides, the general topic of identification of smooth substructure has received attention in the literature; see e.g., Burke and Moré (1988); Wright (1993); Drusvyatskiy and Lewis (2014); Lewis et al. (2022). In the section and the rest of the thesis, the notions of smooth substructure and identification will be useful to refine the analysis of existing algorithms, and to design Newton-type algorithms.

1.3 TOWARDS FAST ALGORITHMS

In this section, we discuss two approaches to build locally fast algorithms for nonsmooth optimization. They rely on the nonsmooth function admitting a smooth substructure, as discussed in the previous section.

1.3.1 *Better understanding and tuning of existing algorithms*

Consider the additive problem (1.2). With the help of partial smoothness and the proximal identification discussed above, it becomes possible to analyze the behavior of the proximal-gradient algorithm *after* identification. The nonsmoothness eventually traps the iterates of the method in the optimal manifold — that of the minimizer, after which the algorithm enters a somewhat smooth regime. Indeed, Liang et al. (2014) show that the convergence rate of the proximal-gradient algorithm improves from sublinear to linear after identification, when the smooth manifold is an affine subspace. This reveals a similarity with the (local) convergence speed of gradient descent in smooth optimization (Nesterov, 2018, Th. 2.1.15). In addition, Liang et al. (2014) propose a better tuning of the method parameters after identification, and show the surprising fact that this fine-tuned proximal-gradient can be faster than the popular accelerated proximal-gradient method. Similar properties are established for operator splitting methods, such as Douglas-Rachford or Alternating Direction of Multiplier Method (Liang et al., 2017b). A delicate point here is that the time of identification is unknown in general, making the moment to tune the method hard to decide in practice. Although it is possible to derive bounds on the maximum number of iterations necessary to reach the optimal manifold in specific cases (Nutini et al., 2019; Liang et al., 2017a), these bounds are conservative, and they rely on quantities delicate to estimate. In this thesis, we will pay a special attention to deriving algorithms that *do not* rely on the identification time or unknown quantities.

1.3.2 *Towards Newton methods*

We now discuss how to explicitly leverage both the composite nature of the function and the smooth substructure of its minimizers. The main idea stems from the following observation. Assume that the smooth substructure of a minimizer is known, then minimizing the *nonsmooth* function simplifies into minimizing its *smooth* restriction constrained on the *smooth* manifold. The smoothness of the latter problem makes it easier to deal with; one can employ efficient Newton-type methods for the reduced smooth problems. The major difficulty is the same as in Section 1.3.1: the crucial information of the optimal smooth substructure is not known beforehand. We illustrate this idea on the nonsmooth functions illustrated before.

Example 1.4 (Nonsmooth to smooth constrained). We illustrate on the running examples, introduced in Example 1.1, how the knowledge of the optimal

manifold allows us to reduce the nonsmooth minimization problem into a smooth constrained problem.

Function F_1 features only one smooth substructure manifold $\mathcal{M}$, introduced in Example 1.3. Knowing that the minimizer belongs to that manifold simplifies the minimization problem into the following reduced problem:

$$\min_{x\in\mathbb{R}^2} 10(x_1-1)^2 \quad \text{s.t.} \quad x_1^2 - x_2 = 0$$

The lasso example has minimizer $x^\star = (0,1)$, with smooth substructure $\mathcal{M}^\star = \{0\}\times\mathbb{R}$. Knowledge of this optimal manifolds simplifies the problem into the following smooth reduced problem:

$$\min_{x\in\mathbb{R}^2} \frac{1}{2}\|Ax-b\|_2^2 + \lambda x_2 \quad \text{s.t.} \quad x_1 = 0.$$

Ndiaye et al. (2017) proposed identification rules tailored to the lasso problem. These rules opened the way to very fast solvers that leverage the smooth substructure of the minimizers; see e.g., Massias et al. (2018); Bertrand et al. (2022). □

This idea is not new, and has been investigated by several authors; let us review the main approaches. All existing algorithms iteratively perform

(i) a smooth substructure detection operation, and

(ii) an efficient Newton-type step relative to the detected smooth substructure.

The case of the maximum of smooth functions was pioneered by Womersley and Fletcher (1986), which proposes an algorithm that *(i)* detects structure by comparing the activity of the smooth functions at different points, and *(ii)* employs Sequential Quadratic Programming (SQP) steps. The minimization of the maximum eigenvalue of a parametrized matrix is approached in a similar fashion by Noll and Apkarian (2005); Helmberg et al. (2014). The authors propose to *(i)* detect the smooth substructure, which corresponds to the multiplicity of the maximal eigenvalue, and *(ii)* employ SQP steps. Both methods locally exhibit a quadratic convergence speed if the minimizer structure is correctly detected; see Womersley and Fletcher (1986, p. 515) and Noll and Apkarian (2005, Th. 1).

The $\mathcal{VU}$-algorithm Mifflin and Sagastizábal (2005) uses similar ideas for the general class of convex functions. Relying on the proximal point interpretation of the bundle method, the authors propose to *(i)* gather (approximate) smooth substructure information from each bundle serious step. This information is then used by *(ii)* taking (approximate) "$\mathcal{U}$-Newton" steps (Lemaréchal et al., 2000) on the corresponding subspace. When the minimizer structure is correctly identified, the serious steps of the method converge locally superlinearly; see Mifflin and Sagastizábal (2005, Th. 15). Note that the presence of a number of null steps between each serious step, hard to control in theory, makes the analysis of the full iterate sequence more delicate.

The efficiency of the above-mentioned methods hinges on the assumption that they correctly identify the smooth substructure of the minimizer. There are no identification guarantees for the methods of Womersley and Fletcher (1986); Noll and Apkarian (2005); the $\mathcal{VU}$-algorithm only benefits from such a guarantee for a certain subclass of convex functions (Daniilidis et al., 2009). We conclude by noting that this question of identification reveals a combinatorial aspect of nonsmooth optimization, reminiscent of constraint identification in

the field of smooth constrained optimization. This is the main difficulty we will face too, in this thesis.

1.4 STRUCTURE OF THIS THESIS

We first give a high level overview of the ideas and organization of this thesis. Then, we provide a detailed summary of each chapter.

1.4.1 *Problematic, approach, and philosophy*

The objective of this thesis is to provide fast algorithms for minimizing structured nonsmooth functions. Specifically, we seek algorithms with the following properties:

(i) *guaranteed identification*, without any prior information on the minimizer structure.

(ii) *guaranteed fast local convergence*, with either superlinear or quadratic speed.

Building on the rich existing work presented before, our approach consists in *detecting* and *leveraging* the smooth substructure of nonsmooth objective functions.

We detect structure for specific functions, which display an additive or composite structure, and whose nonsmooth element admits a simple proximity operator. We use this proximity operator as a *structure oracle*. Indeed, we note a simple but crucial observation: the (nonsmooth) output of a proximity operator step often comes explicitly *with its structure*, at no additional cost. For instance, the prox of the trace norm sets singular values of the input matrix to 0 after a simple test. Tracking the number of singular values set to zero by the prox thus gives the *exact* rank of the output matrix, which encodes its smooth substructure. We illustrate these ideas in two contexts in this thesis. We consider in Chapter 3 the additive case, where the proximal gradient operator provides structure information while also minimizing the function. In Chapter 4 we study the more intricate composition setting, and build an identification tool based on the prox of the outer nonsmooth function.

We leverage nonsmooth structure and make efficient steps by following the philosophy of Section 1.3.2. If the structure of the minimizer were known, the nonsmooth minimization problem would reduce into a smooth constrained optimization problem. The major difficulty is that we never know the structure of the minimizer *a priori*, or during the course of the algorithms. We thus develop new tools to use smooth substructure *adaptively*, without ever assuming it is optimal: at each iteration, a new candidate manifold is obtained by a proximal identification procedure, which is leveraged by taking a Newton step on the smooth reduced subproblem. This is the main technical difference between our work and the literature reviewed in the previous section. The Newton steps on the reduced problems are taken from Riemannian optimization in Chapter 3, and from nonlinear programming in Chapters 4 and 5.

1.4.2 *Detailed contributions*

We now outline precisely the setting and contributions of each chapter. Throughout the thesis, we pay a special attention to illustrate our results with reproducible experiments and precise figures.

CHAPTER 2: PRELIMINARIES. This preliminary chapter provides the mathematical foundations of this thesis. We first present the necessary tools of variational analysis and Riemannian optimization. We then introduce the proximity operator and recall (and slightly extend) a classical characterization result, helpful in a nonconvex setting. We proceed with the notion of partial smoothness, and recall how it allows capturing the local identification behavior of the proximity operator near structured minimizers. This property is the foundation of the main intuitions and results of this thesis. We illustrate it in Fig. 1.4.

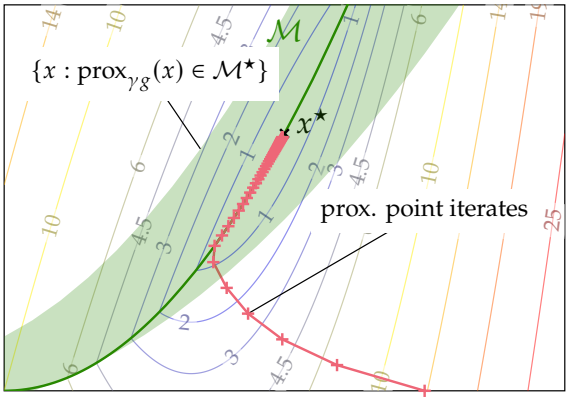


Figure 1.4: Illustration on a simple nonsmooth function of partial smoothness, and of the operator and algorithmic identification properties of the proximal operator.

This chapter is based on Bareilles et al. (2022b).

CHAPTER 3: A NEWTON METHOD FOR NONSMOOTH ADDITIVE MINIMIZATION. We first consider problems of the form (1.2): $\min_{x\in\mathbb{R}^n} f(x) + g(x)$, where f is smooth and g is a nonsmooth function that admits an explicit proximity operator. We take as examples for g two popular nonsmooth functions stemming from machine learning and signal processing applications: the ℓ_1-norm and the trace norm.

We begin by a precise study of the identification behavior of the proximal gradient operator. In a nonconvex setting, we give precise conditions which guarantee that the operator locally maps points to a given smooth substructure; see Theorem 3.1. In particular, we take special care to prove the result both with practical stepsizes and near arbitrary (structured) points, rather than small enough stepsizes and near minimizers.

We then introduce and study an optimization algorithm which alternates between a proximal gradient step, providing a structure candidate, and a Riemannian (truncated) Newton step, providing superlinear convergence near minimizers; see Fig. 1.5 for an illustration. We show that the method converges globally to critical points (Theorem 3.2), and that when the critical point is a qualified minimizer, the algorithm identifies its smooth substructure and converges locally superlinearly (Theorem 3.3). We illustrate this behavior on ℓ_1-norm and trace norm regularized regression problems.

This chapter is based on Bareilles et al. (2022a).

CHAPTER 4: LOCAL NEWTON METHOD FOR NONSMOOTH COMPOSITE MINIMIZATION. In this chapter, we turn to nonsmooth optimization problems of the form (1.3): $\min_{x\in\mathbb{R}^n} F(x) = g(c(x))$, where $c : \mathbb{R}^n \to \mathbb{R}^m$ is a smooth mapping and $g : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ is a nonsmooth function, possibly nonconvex, which admits an explicit proximity operator. We illustrate our developments on two

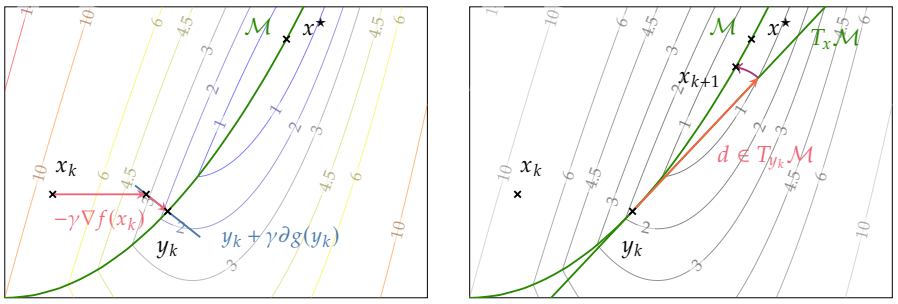


Figure 1.5: Proximal gradient steps, attracted to nonsmoothness (left pane) and Riemannian steps, that converge fast to the minimizer on that subspace (right pane).

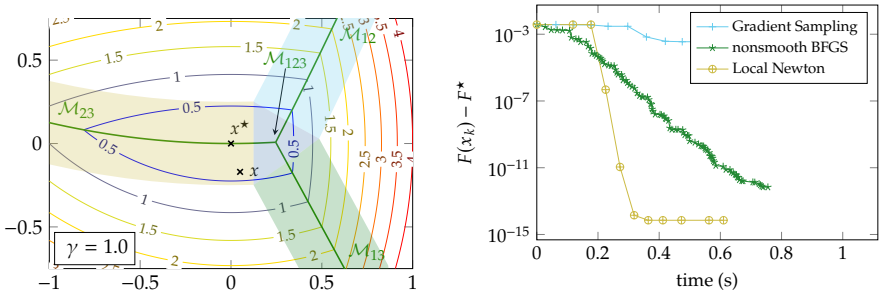
problems: the pointwise maximum of smooth real-valued functions c_i, and the maximum eigenvalue of a parametrized symmetric real matrix function c:

$$F(x) = \max_{i=1,\dots,m} (c_i(x)) \qquad \text{and} \qquad F(x) = \lambda_{\max}(c(x)). \tag{1.5}$$

We first show that the proximity operator *of the simple nonsmooth function* g can provide the exact structure of minimizers *of the full function* $g \circ c$. The prox parameter plays a crucial role here and should thus be selected carefully — neither too small nor too big. We derive precise and implementable bounds that help selection in practice (Theorem 4.4). Identification areas are illustrated in Fig. 1.6a.

We then combine this proximal identification procedure with sequential quadratic programming steps on the identified subspace. We show that, when started near a minimizer, decreasing geometrically the prox parameter ensures eventual identification of the minimizer structure, and local quadratic convergence (Theorem 4.7).

We illustrate numerically the behavior of our method on problems shown in Eq. (1.5), using exact second-order information. We show that it compares favorably with existing methods; see e.g., Fig. 1.6b for the minimization of the maximum eigenvalue of a parameterized matrix. This shows the benefits of exploiting substructure, when possible.



(a) Areas of detection of structure for the identification procedure of Chapter 4, on a maximum of three smooth functions.

(b) Illustration of algorithms minimizing a maximum eigenvalue function.

Figure 1.6: Illustration of structure detection and quadratic convergence.

CHAPTER 5: TOWARDS A GLOBAL NEWTON METHOD FOR NONSMOOTH COMPOSITE MINIMIZATION. In this last chapter, we consider the task of building a Newton algorithm that retains the nice guarantees of the local Newton algorithm of Chapter 4 when started *from arbitrary points*. We outline the challenges associated with the globalization of both the structure detection and the structure exploitation steps of the algorithm. We report partial results with a linesearch SQP algorithm for minimizing composite functions (1.1). We show that the linesearch may prevent the algorithm from reaching the local quadratic convergence speed, and prove that a second-order correction term fixes this issue (Theorem 5.6). We also introduce an optimality condition for a pair (point, smooth substructure), and propose a way to detect whether a structure manifold is optimal in sharp directions (Lemma 5.9). With these elements, we propose a heuristic algorithm and illustrate numerically its identification and local quadratic rate.

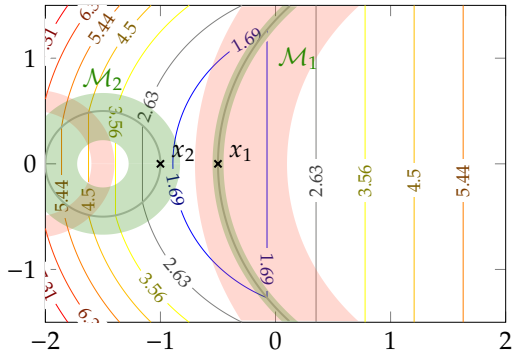


Figure 1.7: Illustration of the areas of detection of structure for the proximity operator of F (red), and the identification tool of Chapter 4 (green), on a maximum of three smooth function. The non-minimizing point x_2 may trap the local method of Chapter 4.

1.5 WORK NOT INCLUDED IN THIS THESIS

To conclude the introduction, I give below short summaries of the two papers not included in this thesis, and outline some implementation work not detailed here either.

1.5.1 *Theoretical contributions*

I detail here two research projects that are not presented in this thesis.

INTERPLAY BETWEEN ACCELERATION AND IDENTIFICATION. In this project, we study the interplay between inertial acceleration and structure identification for the proximal gradient algorithm. We report and analyze several cases where this interplay has negative effects on the algorithm behavior (iterates oscillation, loss of structure identification, etc.). We present a generic method that tames acceleration when structure identification may be at stake. Under a natural geometric condition, our method retains the convergence rate of the accelerated proximal gradient. We show empirically that the proposed method is more stable in terms of subspace identification compared to the accelerated proximal gradient method while keeping a similar functional decrease.

This project was realized with Franck Iutzeler during an internship that preceeded the PhD. It touches on the same themes as this thesis, but the huge jump forward was to introduce second-order information and to deal with identification.

The associated publication is:

> G. Bareilles, F. Iutzeler. On the Interplay between Acceleration and Identification for the Proximal Gradient algorithm. *Computational Optimization and Applications*, 2020.

RANDOMIZED AND ASYNCHRONOUS PROGRESSIVE HEDGING. In this project, we study the Progressive Hedging algorithm, a popular decomposition method for solving multi-stage stochastic optimization problems. A computational bottleneck of this algorithm is that all scenario subproblems have to be solved at each iteration. We introduce randomized versions of the Progressive Hedging algorithm able to produce new iterates as soon as a single scenario subproblem is solved. Building on the relation between Progressive Hedging and monotone operators, we leverage recent results on randomized fixed point methods to derive and analyze the proposed methods.

This project was realized at the beginning of my PhD. I was glad to be part of a team work, which was a nice way to collaborate with and learn from other PhD students. This project led to the following publication:

> G. Bareilles, Y. Laguel, D. Grishchenko, F. Iutzeler, J. Malick. Randomized Progressive Hedging methods for Multi-stage Stochastic Programming. *Annals of Operations Research*, 2020.

1.5.2 *Algorithmic contributions*

Implementing known methods and testing ideas numerically has been a major source of inspiration and intuition for me. Following the principles of reproducible research, all the algorithms presented or used in this thesis are accompanied by easy-to-use and free open-source Julia (Bezanson et al., 2017) implementations.

In this spirit, I released the following toolboxes for the proposed algorithms:

- `StructuredSolvers.jl` – proximal gradient algorithm and its acceleration with inertia (*à la* Nesterov) and Riemannian methods, introduced in Chapter 3. The numerical experiments of Section 3.6 make use of this toolbox.
- `RandomizedProgressiveHedging.jl` – this toolbox contains easy-to-use implementations of the algorithms studied in the above-mentioned paper, and shows the practical interest of randomized algorithms, notably in a parallel context.
- `LocalCompositeNewton.jl` – this package contains the algorithm introduced in Chapter 4 and the code used to reproduce the experiments. The baselines are implemented in a dedicated toolbox, presented below.

It was instructive to implement existing methods for nonsmooth optimization. I collected this code in the following packages:

- `NonSmoothSolvers.jl` – algorithms for blackbox nonsmooth optimization: nonsmooth BFGS with a specific line search (Lewis and Overton,

2013), Gradient Sampling (Burke et al., 2020), and the $\mathcal{VU}$-algorithm (Mifflin and Sagastizábal, 2005).

- `QuadProgSimplex.jl` – minimizing quadratic functions over the simplex (Wolfe, 1976; Kiwiel, 1986). These methods serve as subroutines for the bundle methods and gradient sampling algorithms implemented `NonSmoothSolvers.jl`.

All these packages are available on my webpage: `gbareilles.fr`.

2

PRELIMINARIES

In this chapter, we introduce the main mathematical notions which will be used in our developments. Our notation and terminology follow closely those of the monographs Rockafellar and Wets (1998) for nonsmooth optimization and Absil et al. (2009a); Boumal (2022) for Riemannian optimization.

For this chapter, $g : \mathbb{R}^n \to \bar{\mathbb{R}} = \mathbb{R} \cup \{+\infty\}$ is a function, possibly nonsmooth and nonconvex.

2.1 VARIATIONAL ANALYSIS IN A NUTSHELL

We begin with the basic notions of variational analysis used throughout this thesis.

A function g is *proper* when $g(x) < +\infty$ for at least one $x \in \mathbb{R}^n$ and $g(x) > -\infty$ for all $x \in \mathbb{R}^n$. Besides, we say that g is *lower semi-continuous at point* $\bar{x}$ when, for all $\varepsilon > 0$, there exists a neighborhood $\mathcal{N}_{\bar{x}}$ of $\bar{x}$ such that any x in $\mathcal{N}_{\bar{x}}$ satisfies $g(x) > g(\bar{x}) - \varepsilon$. This function is *lower semi-continuous* when this property holds at any point, or equivalently when the epigraph of g is closed.

SUBGRADIENTS. Several types of subgradients exist for nonconvex functions, in contrast with the convex setting. We introduce and discuss the types of subgradients that will be used throughout this thesis, following Rockafellar and Wets (1998, Chap. 8.B).

Consider a point $\bar{x}$ with $g(\bar{x})$ finite. The set of *regular (or Fréchet) subgradients*

$$\widehat{\partial} g(\bar{x}) \triangleq \left\{ v : g(x) \geq g(\bar{x}) + \langle v, x - \bar{x} \rangle + o(\|x - \bar{x}\|) \text{ for all } x \in \mathbb{R}^n \right\}$$

is closed and convex, but the subdifferential mapping $\widehat{\partial} g(\cdot)$ may not be outer semi-continuous (Rockafellar and Wets, 1998, Th. 8.6, Prop. 8.7). To overcome this problem, the set of *(general or limiting) subgradients* is defined as

$$\partial g(\bar{x}) \triangleq \left\{ \lim_r v_r : v_r \in \widehat{\partial} g(x_r), x_r \to \bar{x}, g(x_r) \to g(\bar{x}) \right\}.$$

The limiting subdifferential is by design outer semi-continuous:

$$\limsup_{x \to \bar{x}} \partial g(x) = \{ u : \exists x_r \to \bar{x}, \exists u_r \to u \text{ with } u_r \in \partial g(x_r) \} \subset \partial g(\bar{x}),$$

which is an attractive property to study sequences of points whose subgradients converge.

A function g is *(Clarke) regular* at $\bar{x}$ when the regular and limiting subdifferentials at $\bar{x}$ coincide (Rockafellar and Wets, 1998, Def. 7.25, Cor. 8.11). This is notably the case for convex functions where the two above definitions coincide with the convex subdifferential (Rockafellar and Wets, 1998, Prop. 8.12).

OPTIMALITY CONDITION. The subdifferential allows to formulate necessary optimality conditions: any local minimizer $x^\star$ of g satisfies the *generalized Fermat rule* (Rockafellar and Wets, 1998, Th. 10.1):

$$0 \in \partial g(x^\star).$$

A point satisfying these conditions is called a *critical point*. When F is convex, critical points coincide with (global) minimizers. In a nonconvex setting, a critical point can also be a local maximum or a saddle point.

2.2 PROXIMITY OPERATOR

A central tool to tackle non-differentiable functions is the *proximity operator*, introduced by the founding work of Jean Jacques Moreau (Moreau, 1965). The proximity operator of a function g with step $\gamma > 0$ at point $y \in \mathbb{R}^n$ is defined as

$$\operatorname{prox}_{\gamma g}(y) \triangleq \operatorname*{arg\,min}_{u \in \mathbb{R}^n} \left\{ g(u) + \frac{1}{2\gamma}\|u - y\|^2 \right\}.$$

While this operator is always well-defined as a set-valued mapping, it gains properties when g is prox-regular and prox-bounded. We quickly introduce these two notions and provide a result on the uniqueness and characterization of the prox operator, which is important in our developments.

Note that, though computing the proximal operator of an arbitrary function is difficult, it comes easy for some relevant cases such as the ℓ_1-norm and the maximum function, or some matrix functions such as the trace-norm and the $\lambda_{\max}$ functions (see Sections 3.2 and 4.2.1).

PROX-REGULARITY AND PROX-BOUNDEDNESS. A function g is *prox-regular* at a point $\bar{y}$ for a subgradient $\bar{v}$ if g is finite and locally lower semi-continuous at $\bar{y}$ with $\bar{v} \in \partial g(\bar{y})$, and there exist $r > 0$ and $\varepsilon > 0$ such that

$$g(y') \geq g(y) + \langle v, y' - y \rangle - \frac{r}{2}\|y' - y\|^2$$

whenever $v \in \partial g(y)$, $\|y - \bar{y}\| < \varepsilon$, $\|y' - \bar{y}\| < \varepsilon$, $\|v - \bar{v}\| < \varepsilon$, and $g(y) < g(\bar{y}) + \varepsilon$. When this holds for all $\bar{v} \in \partial g(\bar{y})$, we say that g is prox-regular at $\bar{y}$ (Rockafellar and Wets, 1998, Def. 13.27).

A function g is *prox-bounded* if there exists $R \geq 0$ such that the function $g + \frac{R}{2}\|\cdot\|^2$ is bounded below. The corresponding *threshold* (of prox-boundedness) is the smallest $r_{pb} \geq 0$ such that $g + \frac{R}{2}\|\cdot\|^2$ is bounded below for all $R > r_{pb}$. In this case, $g + \frac{R}{2}\|\cdot - \bar{y}\|^2$ is bounded below for any $\bar{y}$ and $R > r_{pb}$ (Rockafellar and Wets, 1998, Def. 1.23, Th. 1.25).

CHARACTERIZATION OF THE PROXIMITY OPERATOR. We can now recall a relevant result on the characterization of proximal points, Theorem 1 of Hare and Sagastizábal (2009). At points where g is prox-regular and prox-bounded, this result guarantees that the proximity operator is unique and locally Lipschitz, as well as give a complete characterization by its first-order optimality condition.

Theorem 2.1 (Nonconvex prox characterization). *Suppose that the lower semi-continuous function g is prox-regular at $\bar{x}$ for $\bar{v} \in \partial g(\bar{x})$ with parameter r_{pr}, and prox-bounded with threshold r_{pb}. Then, for any $\gamma < \min(r_{pr}^{-1}, r_{pb}^{-1})$ and all y near $\bar{x} + \gamma\bar{v}$, the proximal operator is:*

- *single-valued and locally Lipschitz continuous;*
- *uniquely determined by the relation*

$$x = \operatorname{prox}_{\gamma g}(y) \Leftrightarrow y \in x + \gamma \partial g(x).$$

We will need a slightly stronger version of this result in our developments, that describes a situation with the additional knowledge of a pair of points linked by the proximal operator. This result will prove useful for taking large steps γ in Chapter 3.

This is the only contribution of this chapter

Lemma 2.2 (Nonconvex prox characterization). *Consider a function $g : \mathbb{R}^n \to \mathbb{R}$, a pair of points $\bar{x}, \bar{y}$ and a step length $\bar{\gamma} > 0$ such that $\bar{x} = \operatorname{prox}_{\bar{\gamma} g}(\bar{y})$ and g is r prox-regular at $\bar{x}$ for subgradient $\bar{v} \triangleq (\bar{y} - \bar{x})/\bar{\gamma}$.*

Then, for any $\gamma \in (0, \min(1/r, \bar{\gamma}))$, there exists a neighborhood $\mathcal{N}_{\bar{y}}$ of $\bar{y}$ over which $\operatorname{prox}_{\gamma g}$ is single-valued and $(1 - \gamma r)^{-1}$-Lipschitz continuous. Furthermore, there holds

$$x = \operatorname{prox}_{\gamma g}(y) \Leftrightarrow y \in x + \gamma \partial g(x)$$

for $y \in \mathcal{N}_{\bar{y}}$ and x near $\bar{x}$ in the sense $\|x - \bar{x}\| < \varepsilon$, $|g(x) - g(\bar{x})| < \varepsilon$ and $\|(y - x)/\gamma - \bar{v}\| < \varepsilon$.

Proof. One can easily check that prox-regularity of g at $\bar{x}$ for subgradient $\bar{v}$ is equivalent to prox-regularity of function $\tilde{g}$ around 0 for subgradient 0, with $\tilde{g} = g(\cdot + \bar{x}) - \langle \bar{v}, \cdot \rangle - g(\bar{x})$ and a change of variable $\tilde{x} = x - \bar{x}$. Similarly, $\bar{x} = \operatorname{prox}_{\bar{\gamma} g}(\bar{y})$ is characterized by its global optimality condition

$$g(x) + \frac{1}{2\bar{\gamma}}\|x - \bar{y}\| > g(\bar{x}) + \frac{1}{2\bar{\gamma}}\|\bar{x} - \bar{y}\|^2 \qquad \text{for all } x \neq \bar{x},$$

which we may write as

$$g(x) > g(\bar{x}) + \langle \bar{v}, x - \bar{x} \rangle - \frac{1}{2\bar{\gamma}}\|x - \bar{x}\|^2 \qquad \text{for all } x \neq \bar{x}.$$

Under that same change of variable, since $\tilde{g}(0) = 0$, this optimality condition rewrites as

$$\tilde{g}(\tilde{x}) > -\frac{1}{2\bar{\gamma}}\|\tilde{x}\|^2 \qquad \text{for all } \tilde{x} \neq 0.$$

We may thus apply Theorem 4.4 from Poliquin and Rockafellar (1996) to get the claimed result on $\tilde{g}$, which transfers back to g as our change of function and variable is bijective. We thus obtain that for $\gamma \in (0, \min(1/r, \bar{\gamma}))$, on a neighborhood $\mathcal{N}_{\bar{y}}$ of $\bar{y}$, $\operatorname{prox}_{\gamma g}$ is single-valued, $(1 - \gamma r)^{-1}$-Lipschitz continuous and $\operatorname{prox}_{\gamma g}(y) = [I + \gamma T]^{-1}(y)$, where T denotes the g-attentive ε-localization of ∂g at $\bar{x}$. Taking y near $\bar{y}$ and x near $\bar{x}$ such that $\|x - \bar{x}\| < \varepsilon$, $|g(x) - g(\bar{x})| < \varepsilon$ and $\|(y - x)/\gamma - \bar{v}\| < \varepsilon$ allows to identify the localization of $\partial g(x)$ with $\partial g(x)$, so that

$$\frac{y - x}{\gamma} \in \partial g(x) \Leftrightarrow \frac{y - x}{\gamma} \in T(x) \Leftrightarrow (I + \gamma T)(x) = y \Leftrightarrow x = \operatorname{prox}_{\gamma g}(y).$$

Note that the proof of Poliquin and Rockafellar (1996, Th. 4.4) includes a minor error relative to the Lipschitz constant computation, we report here a corrected value. □

Remark 2.1 . Note here that the condition $\gamma \in (0, \min(1/r, \bar{\gamma}))$ may be misleading since for the second part of the result, the conditions $\|(y-x)/\gamma - \bar{v}\| < \varepsilon$ and $y \in \mathcal{N}_{\bar{y}}$ also have to be fulfilled. This means that the quantity

$$\|\bar{y} - \bar{x}\| \left(\frac{1}{\gamma} - \frac{1}{\bar{\gamma}} \right)$$

also has to be small. This can be done either a) by taking γ sufficiently close to $\bar{\gamma}$ (which may not be possible since γ has to be smaller than $1/r$); or b) when $\|\bar{y} - \bar{x}\|$ is sufficiently small, i.e., around fixed points of the proximal operator.△

2.3 BASICS OF RIEMANNIAN OPTIMIZATION

In this section, we introduce the tools of Riemannian optimization used in this thesis. We refer the reader to monographs Absil et al. (2009a) and Boumal (2022) for detailed presentations.

SUBMANIFOLDS. A subset $\mathcal{M}$ of $\mathbb{R}^n$ is said to be a *p-dimensional $\mathcal{C}^2$-submanifold* of $\mathbb{R}^n$ around $\bar{x} \in \mathcal{M}$ if there exists a $\mathcal{C}^2$ *manifold-defining map* $h : \mathbb{R}^n \to \mathbb{R}^{n-p}$ with a surjective derivative at $\bar{x} \in \mathcal{M}$ that satisfies for all x close enough to $\bar{x}$: $x \in \mathcal{M} \Leftrightarrow h(x) = 0$.

A basic tool to investigate manifolds is the notion of *smooth curves*. A smooth curve on $\mathcal{M}$ is a $\mathcal{C}^2$ application $\gamma : I \subset \mathbb{R} \to \mathcal{M} \subset \mathbb{R}^n$, where I is an open interval containing 0. At each point $x \in \mathcal{M}$, the *tangent space*, noted $T_x\mathcal{M}$, can be defined as the velocities of all smooth curves passing by x at 0:

$$T_x\mathcal{M} \triangleq \{c'(0) \mid c : I \to \mathcal{M} \text{ is a smooth curve around } 0 \text{ and } c(0) = x\}.$$

The tangent space is a p-dimensional space containing *tangent vectors*. Each tangent space $T_x\mathcal{M}$ is equipped with a scalar product $\langle \cdot, \cdot \rangle_x : T_x\mathcal{M} \times T_x\mathcal{M} \to \mathbb{R}$, and the associated norm $\| \cdot \|_x$. In many cases, the tangent metric varies smoothly with x, making the manifold *Riemannian*. In this thesis, we always use the ambient space scalar product to define the scalar product on tangent spaces; we will thus drop the subscript in the tangent scalar product and norm notations. Related to the tangent space, we will also consider the *normal space* $N_x\mathcal{M}$ at $x \in \mathcal{M}$, defined as the orthogonal space to $T_x\mathcal{M}$ in $\mathbb{R}^n$, and the *tangent bundle manifold* defined by:

$$T\mathcal{B} \triangleq \bigcup_{x \in \mathcal{M}} (x, T_x\mathcal{M}).$$

Note also that both tangent and normal spaces at $x \in \mathcal{M}$ admit explicit expressions from the differential of a local manifold-defining map:

$$T_x\mathcal{M} = \operatorname{Ker} \mathrm{D}\, h(x) \qquad N_x\mathcal{M} = \operatorname{Im} \mathrm{D}\, h(x)^*.$$

In this thesis, $\operatorname{dist}^{geo}_{\mathcal{M}}(x, y)$ *denotes the* geodesic *distance between two points on* $\mathcal{M}$*, while* $\operatorname{dist}_{\mathcal{M}}(x)$ *is the* Euclidean *distance from* x *to* $\mathcal{M}$*.*

A *metric* on $\mathcal{M}$ can be defined as the minimal length over all curves joining two points $x, y \in \mathcal{M}$, i.e., $\operatorname{dist}^{\text{geo}}_{\mathcal{M}}(x, y) \triangleq \inf_{c \in C_{x,y}} \int_0^1 \|c'(t)\|_{c(t)} \mathrm{d}t$, where $C_{x,y}$ is the set of $[0, 1] \to \mathcal{M}$ smooth curves c such that $c(0) = x$, $c(1) = y$. The minimizing curves generalize the notion of straight line between two points to manifolds. The constant speed parametrization of any minimizing curve is called a *geodesic*.

RIEMANNIAN GRADIENT AND HESSIAN. Let $f : \mathcal{M} \to \mathbb{R}$. The *Riemannian differential* of f at x is the linear operator $\mathrm{D}\, f(x) : T_x\mathcal{M} \to \mathbb{R}$ defined by

$\mathrm{D}f(x)[\eta] \triangleq \frac{\mathrm{d}}{\mathrm{d}t} f \circ c(t)\big|_{t=0}$, where c is a smooth curve such that $c(0) = x$ and $c'(0) = \eta$. In turn, the *Riemannian gradient* $\operatorname{grad} f(x)$ is the unique vector of $T_x\mathcal{M}$ such that, for any tangent vector η, $\mathrm{D}f(x)[\eta] = \langle \operatorname{grad} f(x), \eta \rangle$. If $\operatorname{grad} f(x)$ exists, a first order Taylor development can be formulated. Let $(x, \eta) \in T\mathcal{B}$ and c denote a smooth curve passing by x, with velocity η at 0; then, for t near 0,

$$f \circ c(t) = f(x) + t\langle \operatorname{grad} f(x), \eta \rangle + o(t).$$

In order to define second-order objects, we first introduce the notions of derivation for vector fields and of acceleration for curves. Consider a curve $c : I \to \mathcal{M}$ and a smooth vector field Z on c, i.e., a smooth map such that $Z(t) \in T_{c(t)}\mathcal{M}$ for $t \in I$. The *covariant derivative* of Z on the curve c, denoted $\frac{\mathrm{D}}{\mathrm{d}t} Z : I \to \cup_{x \in \mathcal{M}} T_x\mathcal{M}$, is defined by $\frac{\mathrm{D}}{\mathrm{d}t} Z(t) \triangleq \operatorname{proj}_{c(t)} Z'(t)$, where $Z'(t)$ denotes the derivative in the ambient space $\mathbb{R}^n$ and proj_x corresponds to the orthogonal projector from $\mathbb{R}^n$ to $T_x\mathcal{M}$. The *acceleration* of a curve c is defined as the covariant derivative of its velocity: $c''(t) \triangleq \frac{\mathrm{D}}{\mathrm{d}t} c'(0)$.

The *Riemannian Hessian* of f at x is the linear operator $\operatorname{Hess} f(x) : T_x\mathcal{M} \to T_x\mathcal{M}$ defined, for $\eta \in T_x\mathcal{M}$, by the relation $\operatorname{Hess} f(x)[\eta] \triangleq \frac{\mathrm{D}}{\mathrm{d}t} \operatorname{grad} f(c(t))\big|_{t=0}$, where c is a smooth curve such that $c(0) = x$ and $c'(0) = \eta$. Equivalently, we have $\langle \operatorname{Hess} f(x)[\eta], \eta \rangle = \frac{\mathrm{d}^2}{\mathrm{d}t^2} f \circ \gamma(t)\Big|_{t=0}$, where γ is a geodesic such that $\gamma(0) = x$, $\gamma'(0) = \eta$. A second order Taylor development can now be formulated. Let $(x, \eta) \in T\mathcal{B}$ and c be a smooth curve such that $c(0) = x$, $c'(0) = \eta$. Then, for t near 0,

$$f \circ c(t) = f(x) + t\langle \operatorname{grad} f(x), \eta \rangle + \frac{t^2}{2}\langle \operatorname{Hess} f(x)[\eta], \eta \rangle + \frac{t^2}{2}\langle \operatorname{grad} f(x), c''(0) \rangle + o(t^2).$$

Remark 2.2 (Euclidean to Riemannian gradient, Hessian). If $f : \mathcal{M} \to \mathbb{R}$ has a smooth extension on $\mathbb{R}^n$, the Riemannian gradient and Hessian can be computed from their Euclidean counterparts: for a smooth function $\bar{f} : \mathbb{R}^n \to \mathbb{R}$ that coincides with f on $\mathcal{M}$,

$$\operatorname{grad} f(x) = \operatorname{proj}_x(\nabla \bar{f}(x)), \tag{2.1}$$

and, for $\bar{G} : \mathbb{R}^n \to \mathbb{R}^n$ a smooth mapping that coincides with $\operatorname{grad} f$ on $\mathcal{M}$,

$$\operatorname{Hess} f(x)[\eta] = \operatorname{proj}_x\left(\mathrm{D}\bar{G}(x)[\eta]\right). \tag{2.2}$$

△

PRACTICAL TAYLOR DEVELOPMENTS WITH RETRACTIONS. Riemannian optimization methods require a way to produce curves on $\mathcal{M}$ given a point x and a tangent vector η. While a geodesic curve passing at (x, η) is attractive as the generalization of the straight line to manifolds, it usually has a prohibitive computational cost. We thus use *retractions*, i.e., approximations of geodesics, defined on a manifold $\mathcal{M}$ as smooth maps $\mathrm{R} : T\mathcal{B} \to \mathcal{M}$ such that

$$\mathrm{R}_x(0) = x \quad \text{and} \quad \mathrm{D}\mathrm{R}_x(0) : T_x\mathcal{M} \to T_x\mathcal{M} \text{ is the identity map: } \mathrm{D}\mathrm{R}_x(0)[v] = v,$$

where, for each $x \in \mathcal{M}$, $\mathrm{R}_x : T_x\mathcal{M} \to \mathcal{M}$ is defined as the restriction of R at x, so that $\mathrm{R}_x(v) = \mathrm{R}(x, v)$. A *second-order retraction* is a retraction R such that, for all $(x, \eta) \in T\mathcal{B}$, the curve $c(t) = \mathrm{R}_x(t\eta)$ has zero acceleration at 0: $c''(0) = 0$. Thus

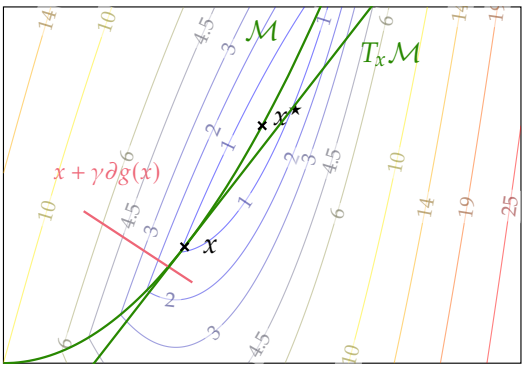


Figure 2.1: Illustration of partial smoothness on function $g(x) = 10(x_1 - 1)^2 + 5|x_1^2 - x_2|$. The function is smooth along $\mathcal{M}$, nonsmooth across, and the tangent space at $x \in \mathcal{M}$ is perpendicular to the subdifferential.

$t \mapsto \mathrm{R}_x(t\eta)$ is a practical curve passing by (x, η) at 0, and provides a similar development as above: for t near 0,

$$f \circ \mathrm{R}_x(t\eta) = f(x) + t\langle \operatorname{grad} f(x), \eta\rangle + \frac{t^2}{2}\langle \operatorname{Hess} f(x)[\eta], \eta\rangle + o(t^2\|\eta\|^2). \tag{2.3}$$

2.4 PARTIAL SMOOTHNESS AND PROXIMAL IDENTIFICATION

In this section, we introduce the key notion of *partial smoothness*. We then illustrate how it allows to characterize the identification behavior of the proximal operator, and of the corresponding proximal point algorithm.

This notion will be central in Chapters 3 and 4, which focus on the local analysis of algorithms near (structured) minimizers.

PARTIAL SMOOTHNESS. The concept of partial smoothness, introduced in Lewis (2002), formalizes the idea that a nonsmooth function g is locally smooth along a manifold and nonsmooth across it. We illustrate it on Figure 2.1 and Example 2.1.

A function g is ($\mathcal{C}^2$-)*partly smooth* at a point $\bar{y}$ relative to a set $\mathcal{M}^g$ containing $\bar{y}$ if $\mathcal{M}^g$ is a $\mathcal{C}^2$ manifold around $\bar{y}$ and if

- (smoothness) the restriction of g to $\mathcal{M}^g$ is a $\mathcal{C}^2$ function near $\bar{y}$;
- (regularity) g is regular at all points $y \in \mathcal{M}^g$ near $\bar{y}$, with $\partial g(y) \neq \emptyset$;
- (sharpness) the affine span of $\partial g(\bar{y})$ is a translate of $N_{\bar{y}}\mathcal{M}^g$;
- (sub-continuity) the mapping ∂g restricted to $\mathcal{M}^g$ is continuous at $\bar{y}$.

Example 2.1 . We detail each requirement of partial smoothness on the function $g(x) = 10(x_1 - 1)^2 + 5|x_1^2 - x_2|$; see Fig. 2.1 for an illustration. At any $x \in \mathcal{M} = \{x \in \mathbb{R}^2 : x_1^2 - x_2 = 0\}$, g is partly smooth relative to $\mathcal{M}$: (i) its restriction to $\mathcal{M}$, $g\big|_{\mathcal{M}} = 10(x_1 - 1)^2$ is smooth, (ii) the function is convex, thus (Clarke) regular everywhere. The subdifferential at $x \in \mathcal{M}$ writes

$$\partial g(x) = 20\begin{pmatrix} x_1 - 1 \\ 0 \end{pmatrix} + 5\operatorname{Conv}\left(\begin{pmatrix} 2x_1 \\ -1 \end{pmatrix}, \begin{pmatrix} -2x_1 \\ 1 \end{pmatrix}\right),$$

it is thus (iii) perpendicular to the tangent space to $\mathcal{M}$ at x (i.e., parallel to the normal space), (iv) and continuous. □

TYPICAL USE IN NONSMOOTH ANALYSIS. Combining partial smoothness and prox-regularity ensures that the proximal operator smoothly locates the structure manifold of qualified minimizers. This result is formalized in the following theorem, from Daniilidis et al. (2006, Th. 28).

Here, we have an operator *view of identification: the prox maps a neighborhood of the minimizer to the manifold.*

Proposition 2.3 (Prox locates manifold at minimizer). *Consider g a lower semi-continuous function on $\mathbb{R}^n$, and $\bar{x}$ a critical point $0 \in \partial g(\bar{x})$. Suppose that g is both prox-bounded and prox-regular at $\bar{x}$, and partly-smooth relative to $\mathcal{M}$ at $\bar{x}$.*

Take $\gamma > 0$. If $0 \in \operatorname{ri} \partial g(\bar{x})$ and γ is small enough, then the proximal operator $y \mapsto \operatorname{prox}_{\gamma g}(y)$ is $\mathcal{C}^1$ and $\mathcal{M}$-valued near $\bar{x}$.

We note that, throughout this thesis, the assumption of partial smoothness goes in hand with that of prox-regularity. One reason is that, in this case, uniqueness of the partial smoothness manifold is guaranteed locally. We derive a self-contained proof of this folklore result, that parallels Hare and Lewis (2004, Corollary 4.2, Example 7.1) for partly smooth sets.

Proposition 2.4 (Uniqueness of manifold). *Consider a function g, two manifolds $\mathcal{M}_1, \mathcal{M}_2$ and a point $\bar{x} \in \mathcal{M}_1 \cap \mathcal{M}_2$ such that g is r-prox-regular at $\bar{x}$ and partly-smooth relative to both manifolds $\mathcal{M}_1$ and $\mathcal{M}_2$. Then, near $\bar{x}$, $\mathcal{M}_1 = \mathcal{M}_2$.*

Proof. For the sake of eventual contradiction, let (x_k) denote any sequence converging to $\bar{x}$ such that $x_k \in \mathcal{M}_1 \setminus \mathcal{M}_2$ for all k. Since g is prox-regular, Rockafellar and Wets (1998, Prop. 13.37) tell us that there is $\bar{\gamma} > 0$ such that $\bar{x} = \operatorname{prox}_{\bar{\gamma} g}(\bar{y})$ for some $\bar{y} \in \bar{x} + \bar{\gamma} \operatorname{ri} \partial g(\bar{x}) \in \mathbb{R}^n$ (since g is partly smooth, $\partial g(\bar{x})$ has non-empty relative interior, and $\bar{y}$ can be taken as $\bar{x} + \bar{\gamma}\bar{v}$ for any $\bar{v} \in \operatorname{ri} \partial g(\bar{x})$ by reasoning as in the proof of Hare and Sagastizábal (2009, Th. 4)).

We can thus select a sequence $v_k \in \partial g(x_k)$ converging to $\bar{v} = (\bar{y} - \bar{x})/\bar{\gamma} \in \operatorname{ri} \partial g(\bar{x})$ and define $y_k = x_k + \gamma v_k$ for some $\gamma \in (0, \bar{\gamma})$. It is immediate to see that the sequence (y_k) converges to $y^\gamma = (1 - (\gamma/\bar{\gamma}))\bar{x} + (\gamma/\bar{\gamma})\bar{y}$ and that y^γ can be made arbitrarily close to $\bar{y}$ by taking γ close to $\bar{\gamma}$. Thus, we can consider that, properly choosing γ, y_k reaches any neighborhood of $\bar{y}$ in a finite number of iterations.

Lemma 2.2 then indicates that for k large enough, we have $x_k = \operatorname{prox}_{\gamma g}(y_k)$. Furthermore, Proposition 2.3 applied with $f = 0$, $\mathcal{M} = \mathcal{M}_2$ shows that $\operatorname{prox}_{\gamma g}$ is $\mathcal{M}_2$-valued near $\bar{y}$ which implies that $x_k = \operatorname{prox}_{\gamma g}(y_k) \in \mathcal{M}_2$ for large k which contradicts x_k being in $\mathcal{M}_1 \setminus \mathcal{M}_2$. □

ALGORITHMIC IDENTIFICATION OF THE PROXIMAL POINT ALGORITHM. We now informally illustrate the algorithmic identification of the proximity operator. Consider a proper closed convex function g. The proximal point algorithm, defined for some $\gamma \in \mathbb{R}_+^*$ and $x_0 \in \mathbb{R}^n$ as

$$x_{k+1} = \operatorname{prox}_{\gamma g}(x_k),$$

generates sequences that converge to critical points; see e.g., Attouch and Bolte (2009, Th. 1) for subanalytic functions. If the limit point $\bar{x}$ satisfies a qualification condition $0 \in \operatorname{ri} \partial g(\bar{x})$ and γ is small enough, Proposition 2.3 provides the existence of a neighborhood $\mathcal{N}_{\bar{x}}$ over which the proximity operator is $\mathcal{M}$-valued. Since the iterates converge to $\bar{x}$, they all belong to the identification neighborhood $\mathcal{N}_{\bar{x}}$ after some finite (but unknown) time. Therefore, the sequence of iterates is $\mathcal{M}$-valued after some finite time, which provides an *algorithmic* view of identification.

Here we take an algorithmic *view of identification: the iterates eventually belong to the minimizer manifold.*

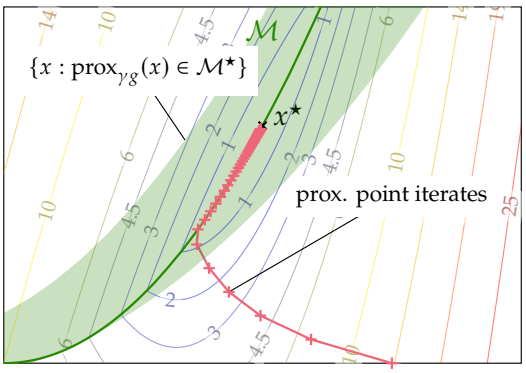


Figure 2.2: Illustration of the local identification of the proximal operator (Proposition 2.3) and finite time identification of proximal point algorithm. The minimizer is $x^\star = (1, 1)$, with structure manifold $\mathcal{M}^\star = \{x \in \mathbb{R}^2 : x_1^2 = x_2\}$. The green area shows the *operator identification* property of the prox: the operator maps a neighborhood of the minimizer to $\mathcal{M}^\star$. The red iterates illustrate the *algorithmic identification* of the proximal point algorithm: the iterates eventually belong to $\mathcal{M}^\star$.

We can now return to Figure 1.3a of the introductory chapter, recalled and enriched here as Fig. 2.2. The nonsmooth function is:

$$g(x) = 10(x_1 - 1)^2 + 5|x_1^2 - x_2|.$$

A direct computation yields an explicit formula for the proximity operator:

$$\mathrm{prox}_{\gamma g}(x) = \begin{cases} (\frac{x[1]+20\gamma}{1+30\gamma}, x_2 + 5\gamma) & \text{if } x_2 \leq \left(\frac{x_1+20\gamma}{1+30\gamma}\right)^2 - 5\gamma \\ (t, t^2) & \text{if } \left(\frac{x_1+20\gamma}{1+30\gamma}\right)^2 - 5\gamma \leq x_2 \leq \left(\frac{x_1+20\gamma}{1+10\gamma}\right)^2 + 5\gamma \\ (\frac{x_1+20\gamma}{1+10\gamma}, x_2 - 5\gamma) & \text{if } x_2 \geq \left(\frac{x_1+20\gamma}{1+10\gamma}\right)^2 + 5\gamma \end{cases},$$

where t minimizes the function $\varphi(t) = 10(t - x_1)^2 + 1/(2\gamma)\left((t - x_1)^2 + (t^2 - x_2)\right)^2$.

Note that the set of points mapped to $\mathcal{M}$, fully represented in Fig. 2.2, is actually more than a neighborhood of the minimizer. We return on this topic in Chapter 3.

3

A NEWTON METHOD FOR NONSMOOTH ADDITIVE MINIMIZATION

This chapter incorporates material from Bareilles et al. (2022b)

3.1 INTRODUCTION

In this chapter, we consider the nonsmooth optimization problem

$$\min_{x \in \mathbb{R}^n} F(x) \triangleq f(x) + g(x), \tag{$\mathcal{P}$}$$

where f is a smooth differentiable function, and g is a nonsmooth function. Throughout the chapter, we illustrate our developments on two nonsmooth functions stemming from machine learning and signal processing applications: the ℓ_1 norm

$$g : x \mapsto \|x\|_1 = \sum_{i=1}^{n} |x_i| \tag{3.1}$$

and the nuclear norm

$$g : x \mapsto \|x\|_* = \|\sigma(x)\|_1, \tag{3.2}$$

where $\sigma(x)$ denotes the singular values of x. In these cases and many others, the *nonsmooth* objective function presents a *smooth* substructure, which involves smooth submanifold on which the function is locally smooth.

In the sense of partial smoothness, see Section 2.4.

A critical aspect is the requirement that the proximal operator of g can be computed explicitly, and that it outputs a representation of the submanifold of the output point. In this setting, first-order methods to minimize F are the (accelerated) proximal gradient algorithms; see Beck (2017, Chap. 10) for a general review of these methods and their analysis. In nondegenerate cases, the iterates produced by these algorithms eventually reach the optimal submanifold (i.e., the manifold which contains the minimizer): it is the (algorithmic) *identification* property of proximal algorithms, discussed in previous chapters. We note that the identification of the final manifold happens after a finite *but unknown* number of iterations, which can be estimated only in specific cases.

This property is illustrated on Fig. 3.1, and discussed for the proximal point in Section 2.4

We propose in this chapter a *Newton acceleration* of the proximal gradient algorithm that adaptively uses identification. Our algorithm consists of two main ingredients:

i) structure *identification*, relying on the explicit proximal gradient operator to extract structure information, and

ii) structure *exploitation*, based on Riemannian Newton steps on the identified manifolds.

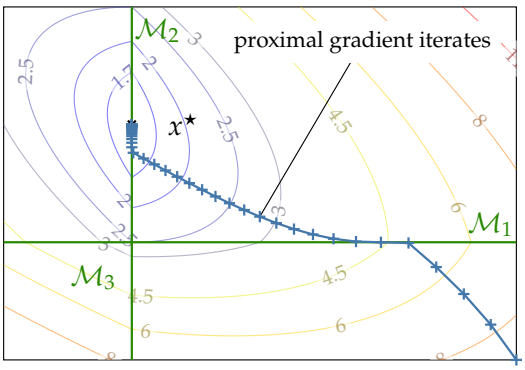


Figure 3.1: Illustration of the typical level lines of a Lasso objective and the eventual (algorithmic) identification of the proximal gradient: the iterates belong to the smooth substructure $\mathcal{M}^\star = \{x \in \mathbb{R}^2 : x_1 = 0\}$ of the minimizer $x^\star = (0, 1)$ in finite time.

We present a convergence analysis showing superlinear convergence of the resulting algorithm, under some natural qualification assumptions but without prior knowledge on the final optimal submanifold. Finally, we provide numerical illustrations showing the interests of the proposed Newton acceleration on typical structure-inducing regularized problems (sparse logistic regression and low-rank least-squares). Along the way, our study reveals results that have some interest on their own, in particular: we refine the smoothness properties of the proximal gradient operator around structured points, and we formalize complementary properties on line searches in Riemannian optimization.

Let us finally note that the Newton acceleration of the proximal gradient that we propose here should not be confused with proximal-Newton schemes such as Lee et al. (2014); Becker et al. (2019); Aravkin et al. (2022). These methods essentially replace the gradient step by a (quasi-)Newton step before applying a proximity operator. They therefore do not use second order information on g, which is crucial to obtain quadratic local rates. Besides, we choose the term "Newton acceleration" to emphasize the similarity with the celebrated Nesterov acceleration (Nesterov, 1983). Indeed both methods add an acceleration step after the proximal gradient iteration. But, unlike Nesterov's method where the acceleration is provided by an inertial step, the Newton acceleration comes from a second-order step on a smooth substructure, as we detail in this chapter.

OUTLINE OF THIS CHAPTER. First, in Section 3.2 we recall the relevant properties of the ℓ_1-norm (3.1) and nuclear norm (3.2). In Section 3.3, we show the structure identification properties of the proximal gradient near structured points. Then, we introduce in Section 3.4 our template algorithm, alternating a proximal gradient step with a Riemannian update on the identified manifold and show its convergence and identification properties. In Section 3.5, we specify the implementation of efficient Riemannian Newton-type methods and illustrate their performances in Section 3.6. Some material used in our proofs has been deferred to Appendices A.1 and A.2; most of these results are well-known and just recalled here, but some seem to be less-known or not precisely treated in the literature.

3.2 EXAMPLES OF STRUCTURE MANIFOLDS, PROXIMAL OPERATOR AND RIEMANNIAN DERIVATIVES

In this section, we illustrate on the ℓ_1 norm (3.1) and nuclear norm (3.2) the notions of Riemannian and nonsmooth optimization. We first present the structure submanifolds associated with these functions and the related Riemannian objects. Then, we describe the proximity operator of the ℓ_1 and nuclear norm, detail their partial smoothness and prox-regularity properties, and finally give their Riemannian derivatives.

See Section 2.3
See Sections 2.1, 2.2 and 2.4

SUBMANIFOLDS AND RELATED OBJECTS.

Example 3.1 (Fixed coordinate-sparsity subspaces). We consider the manifold

$$\mathcal{M}_I \triangleq \{x \in \mathbb{R}^n : x_i = 0 \text{ for } i \in I\}, \tag{3.3}$$

where $I \subset \{1, \dots, n\}$. This manifold is actually a vector space and all related notions have simple expressions, as follows.

The tangent space at any point identifies with the manifold itself: $T_x\mathcal{M}_I = \mathcal{M}_I$. The orthogonal projection of a vector $d \in \mathbb{R}^n$ on the tangent space writes $\mathrm{proj}_x(d)$, where $[\mathrm{proj}_x(d)]_i$ is d_i if $i \notin I$, and null otherwise. The map $\mathrm{R}_x(\eta) = x + \eta$ defines a second-order retraction.

Given a function f defined on the ambient space, the Riemannian gradient and Hessian-vector product of the restriction of F to $\mathcal{M}_I$ are obtained from their Euclidean counterparts by a simple projection: for $(x, \eta) \in T\mathcal{B}$,

$$\mathrm{grad}\, f(x) = \mathrm{proj}_x(\nabla f(x)) \qquad \mathrm{Hess}\, f(x)[\eta] = \mathrm{proj}_x(\nabla^2 f(x)[\eta]). \qquad \square$$

Example 3.2 (Fixed rank matrices). We consider the manifold of fixed-rank matrices

$$\mathcal{M}_r \triangleq \{x \in \mathbb{R}^{m\times n} : \mathrm{rank}(x) = r\}, \tag{3.4}$$

for which we refer to Boumal (2022, Sec. 7.5). A rank-r matrix $x \in \mathcal{M}_r$ is represented as $x = U\Sigma V^\top$, where $U \in \mathbb{R}^{m\times r}$, $V \in \mathbb{R}^{n\times r}$, $\Sigma \in \mathbb{R}^{r\times r}$ such that $U^\top U = I_n$, $V^\top V = I_m$ and Σ is a diagonal matrix with positive entries. Such a decomposition can be obtained by computing the singular value decomposition of the matrix x. Using this representation, a tangent vector $\eta \in T_x\mathcal{M}_r$ writes

$$\eta = UMV^\top + U_pV^\top + UV_p^\top,$$

where $M \in \mathbb{R}^{r\times r}$, $U_p \in \mathbb{R}^{m\times r}$, $V_p \in \mathbb{R}^{n\times r}$ such that $U^\top U_p = 0$, $V^\top V_p = 0$. The orthogonal projection of a vector $d \in \mathbb{R}^{m\times n}$ onto $T_x\mathcal{M}_r$ writes $\mathrm{proj}_x(d) = d - U^\top dV$.

Given a function f defined on the ambient space, a Riemannian gradient and Hessian-vector product of f restricted to $\mathcal{M}_r$ can be obtained from their Euclidean counterparts: for $x, \eta \in T\mathcal{B}$, and with $P_U^\top = I_m - UU^\top$, $P_V^\top = I_n - VV^\top$.

$$\mathrm{grad}\, f(x) = \mathrm{proj}_x(\nabla f(x))$$
$$\mathrm{Hess}\, f(x)[\eta] = \mathrm{proj}_x(\nabla^2 f(x)[\eta]) + \left[P_U^\top \nabla f(x) V_p \Sigma^{-1}\right] V^\top + U\left[P_V^\top \nabla f(x)^\top U_p \Sigma^{-1}\right]^\top.$$

$\square$

PROXIMITY OPERATOR AND STRUCTURE MANIFOLDS.

Example 3.3 (ℓ_1 norm). In the context of Example 3.1, we consider the ℓ_1 norm (3.1). This function is convex, thus prox-regular at every point with $r = 0$. Its proximity operator admits a closed form:

$$[\mathrm{prox}_{\gamma\|\cdot\|_1}(y)]_i = \begin{cases} y_i + \gamma & \text{if } y_i < -\gamma \\ 0 & \text{if } -\gamma \le y_i \le \gamma \\ y_i - \gamma & \text{if } y_i > \gamma \end{cases}$$

which naturally gives sparse outputs. In other words, $x = \mathrm{prox}_{\gamma\|\cdot\|_1}(y)$ lies on $\mathcal{M}_I$ (see (3.3)) where I is the complementary of support of x.

Actually, one readily checks that $\|\cdot\|_1$ is partly smooth at x relative to the structure manifold $\mathcal{M}_I$. In particular, the restriction of $\|\cdot\|_1$ to the manifold $\mathcal{M}_I$ is locally smooth at x; the ℓ_1 norm thus admits a Riemannian gradient and Hessian at point x:

$$\mathrm{grad}\,\|\cdot\|_1(x) = \mathrm{sgn}(x) \qquad \text{and} \qquad \mathrm{Hess}\,\|\cdot\|_1(x) = 0,$$

where $\mathrm{sgn}(x) \in \{-1, 0, 1\}$ denotes the sign of x, null when $x = 0$. □

Example 3.4 (nuclear norm). Following the notation of Example 3.2, we consider the nuclear norm (3.2). where Σ denotes the diagonal term of the singular value decomposition of x. This function is convex, and thus prox-regular at every point with $r = 0$. Its proximity operator admits a closed form: for matrix y $(= U\Sigma V^\top)$,

$$\mathrm{prox}_{\gamma\|\cdot\|_*}(y) = U(\Sigma - \gamma)_+ V^\top,$$

where the coefficient (i, j) of $(\Sigma - \gamma)_+$ is defined as $\max(\Sigma_{ij} - \gamma, 0)$. Thus, $x = \mathrm{prox}_{\gamma\|\cdot\|_*}(y)$ has low rank, by construction. Said otherwise, x lies on $\mathcal{M}_r$ (see (3.4)) where $r = \mathrm{rank}(\Sigma - \gamma)_+$.

Similarly, one readily checks that $\|\cdot\|_*$ is partly smooth at x relative to the structure manifold $\mathcal{M}_r$. In particular, the restriction of $\|\cdot\|_*$ to the manifold $\mathcal{M}_r$ is locally smooth; the nuclear norm thus admits a Riemannian gradient and Hessian at point x: denoting $\eta = UMV^\top + U_p V^\top + UV_p^\top \in T_x\mathcal{M}_r$ a tangent vector,

$$\begin{aligned} \mathrm{grad}\,\|\cdot\|_*(x) &= UV^\top \\ \mathrm{Hess}\,\|\cdot\|_*(x)[\eta] &= U\left[\tilde{F} \circ (M - M^\top)\right] V^\top + U_p \Sigma^{-1} V^\top + U\Sigma^{-1} V_p^T, \end{aligned}$$

where $\circ$ denotes the Hadamard product and $\tilde{F} \in \mathbb{R}^{\bar{r}\times\bar{r}}$ is such that $\tilde{F}_{ij} = 1/(\Sigma_{jj} + \Sigma_{ii})$ if $\Sigma_{jj} \neq \Sigma_{ii}$, and $\tilde{F}_{ij} = 0$ otherwise. We provide a self-contained derivation of these derivatives in Appendix B.2. □

3.3 COLLECTING STRUCTURE WITH THE PROXIMAL GRADIENT

In this section, we show that the proximal gradient operator smoothly locates structure in nonsmooth nonconvex settings. Building on this result, we then formulate minimal assumptions on points for which the proximal gradient detects structure.

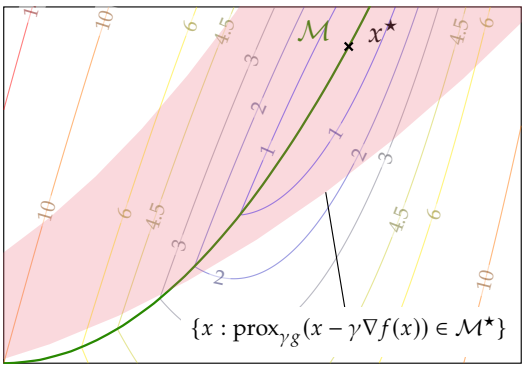


Figure 3.2: Illustration of Theorem 3.1 on the additive function $F(x) = 10(x_1 - 1)^2 + 5|x_1^2 - x_1|$. The minimizer is $x^\star = (1, 1)$, the structure manifold is $\mathcal{M} = \{x \in \mathbb{R}^2 : x_1^2 = x_2\}$. The red area shows the points mapped to $\mathcal{M}$ by the proximal gradient operator.

3.3.1 *Smoothness and localization of the proximal gradient operator*

The results of this section are built on g being a partly smooth and prox-regular function. Under this assumption, we show in the next theorem that the proximal gradient smoothly locates active manifolds: if some input $\bar{y}$ is mapped onto $\mathcal{M}$, then the proximal gradient is $\mathcal{M}$-valued and $\mathcal{C}^1$ around $\bar{y}$; see Figure 3.2.

Theorem 3.1 (Proximal gradient points smoothly locate manifolds). *Let f be a $\mathcal{C}^2$ function on $\mathbb{R}^n$ and g a lower semi-continuous function on $\mathbb{R}^n$. Suppose that g is both r-prox-regular at $\bar{x}$ and partly-smooth relative to $\mathcal{M}$ at $\bar{x}$.*

Take $\gamma, \bar{\gamma}$ such that $0 < \gamma < \bar{\gamma} \leq 1/r$ and $\bar{x} = \mathrm{prox}_{\bar{\gamma} g}(\bar{y} - \bar{\gamma} \nabla f(\bar{y}))$. If

i) $\frac{1}{\gamma}(\bar{y} - \bar{x}) - \nabla f(\bar{y}) \in \mathrm{ri}\, \partial g(\bar{x})$ *(the relative interior of the subdifferential at $\bar{x}$);*

ii) either a) γ is sufficiently close to $\bar{\gamma}$, or b) $\bar{y}$ is sufficiently close to $\bar{x}$;

then, the proximal gradient $y \mapsto \mathrm{prox}_{\gamma g}(y - \gamma \nabla f(y))$ is $\mathcal{C}^1$ and $\mathcal{M}$-valued near $\bar{y}$.

This result is based on the sensitivity analysis of partly smooth functions (Lewis, 2002, Sec. 5). The proof extends and refines the rationale of Daniilidis et al. (2006, Th. 28) on the proximal operator, recalled in this thesis as Proposition 2.3. Theorem 3.1 thus extends existing results on three aspects: *i)* it describes the proximal gradient operator rather than the proximal operator, *ii)* it allows for a full stepsize range of $(0, 1/r)$ in the proximal gradient, and *iii)* it describes the identification property near any structured point, rather than only minimizers.

Before moving to the proof, we note that the assumptions of Theorem 3.1 are rather tight: prox-regularity and partial smoothness are minimal assumptions in this setting. The following example shows on an example that, when assumption *i)* fails, the proximal gradient mapping looses the property of mapping a *full* neighborhood of $\bar{y}$ to $\mathcal{M}$ along with its smoothness.

Remark 3.1 (Proximal gradient at non-qualified points.). Assumption *i)* is necessary to ensure the identification of $\mathcal{M}$ on a full neighborhood. This assumption fails when $\frac{1}{\gamma}(\bar{y} - \bar{x}) - \nabla f(\bar{y})$ lies on the relative boundary of $\partial g(\bar{x})$. In this case, the proximal gradient may still identify depending on additional quantities, such as its initialization, stepsize, *etc*. Fadili et al. (2018) show that enlarged

identification properties still holds when *i)* fails; Bareilles and Iutzeler (2020) illustrate this situation. We also note that the distance from $\frac{1}{\gamma}(\bar{y} - \bar{x}) - \nabla f(\bar{y})$ to the relative boundary of $\partial g(\bar{x})$ can be computed after the prox computation without significant additional cost for the ℓ_1 and nuclear norms.

We illustrate failure of *i)* on a simple example. Take $f(x) = \frac{1}{2}(x-1)^2$ and $g(x) = |x|$ for any $x \in \mathbb{R}$, and $\gamma \in (0,1)$. The unique minimizer lies at the origin and $f + g$ is partly smooth there relative to the manifold $\mathcal{M} = \{0\}$. However, the minimizer is not an r-structured critical point as it is not qualified: $0 \notin \operatorname{ri} \partial(f+g)(0) = \operatorname{ri}[-2;0] = (-2;0)$. The proximal gradient operator of $f + g$ writes

$$\operatorname{prox}_{\gamma g}(y - \gamma \nabla f(y)) = \begin{cases} (1-\gamma)y + 2\gamma & \text{if } y \leq \frac{-2\gamma}{1-\gamma} \\ 0 & \text{if } \frac{-2\gamma}{1-\gamma} \leq y \leq 0 \\ (1-\gamma)x & \text{if } 0 \leq y \end{cases},$$

there is no neighborhood of 0 on which the operator is smooth and $\mathcal{M}$-valued.△

Proof (of Theorem 3.1). Adopting the same reasoning as in Lewis (2002, Sec. 5) and Daniilidis et al. (2006, Sec. 4.1), we consider the function

$$\begin{aligned} \rho : \mathbb{R}^n \times \mathbb{R}^n &\to \mathbb{R} \\ (x, y) &\mapsto g(x) + \tfrac{1}{2\gamma}\|x - y + \gamma \nabla f(y)\|^2, \end{aligned}$$

and denote by $\rho_y = \rho(\cdot, y)$. Computing the proximal gradient $\operatorname{prox}_{\gamma g}(y - \gamma \nabla f(y))$ can then be seen as minimizing the parameterized function ρ_y.

Step 1. As a first step, we study the minimizers of ρ_y restricted to $\mathcal{M}$, for y near $\bar{y}$. We consider the parametric manifold optimization problem, for y near $\bar{y}$:

$$\min_{x \in \mathcal{M}} \rho_y(x). \tag{$P_{\mathcal{M}}(y)$}$$

Since g is $\mathcal{C}^2$-partly-smooth relative to $\mathcal{M}$ and f is $\mathcal{C}^2(\mathbb{R}^n)$, ρ_y is twice continuously differentiable on $\mathcal{M}$. Moreover, the r-prox-regularity gives easily (see Lemma A.7) that $\rho_{\bar{y}}$ is lower-bounded by $(\frac{1}{\gamma} - r)\|\cdot - \bar{x}\|^2/2$ on a neighborhood of $\bar{x}$ in $\mathbb{R}^n$ and, a fortiori, in $\mathcal{M}$. Thus $\bar{x}$ is a strong local minimizer. The Riemannian sufficient optimality conditions Lemma A.1 imply

$$\operatorname{grad} \rho_{\bar{y}}(\bar{x}) = 0 \qquad \operatorname{Hess} \rho_{\bar{y}}(\bar{x}) \geq \left(\frac{1}{\gamma} - r\right) I \succ 0,$$

which are the conditions to apply the implicit functions theorem, as follows.

We consider the equation $\Phi(x, y) = 0$, for x, y near $\bar{x}, \bar{y}$, where $\Phi : \mathcal{M} \times \mathbb{R}^n \to T\mathcal{B}$ is defined as $\Phi(x, y) = \operatorname{grad} \rho_y(x)$. This function is continuously differentiable on a neighborhood of $(\bar{x}, \bar{y})$, and its differential relative to $\bar{x}$ at that point, $\operatorname{Hess} \rho_{\bar{y}}(\bar{x})$, is invertible. The implicit function theorem thus grants the existence of neighborhoods $\mathcal{N}_{\bar{x}}, \mathcal{N}_{\bar{y}}$ of $\bar{x}$, $\bar{y}$ in $\mathcal{M}$, $\mathbb{R}^n$, and a continuously differentiable function $\hat{x} : \mathcal{N}_{\bar{y}} \to \mathcal{N}_{\bar{x}}$ such that, for any y in $\mathcal{N}_{\bar{y}}$, $\Phi(\hat{x}(y), y) = \operatorname{grad} \rho_y(\hat{x}(y)) = 0$. Actually, $\hat{x}(y)$ is a strong minimizer of ρ_y on $\mathcal{M}$ for y close enough to $\bar{y}$. Indeed, the mapping $\hat{x}$ is continuous on $\mathcal{N}_{\bar{y}}$, so that $y \mapsto \operatorname{Hess} \rho_y(\hat{x}(y))$ is also continuous there and the property $\operatorname{Hess} \rho_{\bar{y}}(\hat{x}(\bar{y})) \succ 0$ extends locally around $\bar{y}$.

Step 2. As a second step, we turn to show that the minimizer $\hat{x}(y)$ of ρ_y on $\mathcal{M}$ is actually a strong critical point of ρ_y in $\mathbb{R}^n$ (Lewis, 2002, Def. 5.3), and thus the

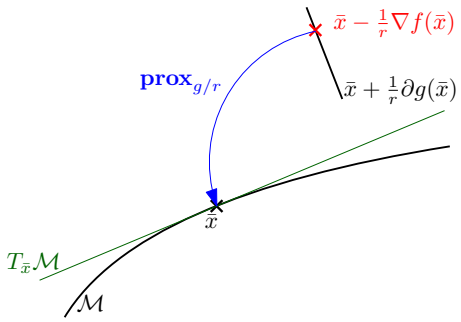


Figure 3.3: Illustration of a r-structured critical point. Point i) is illustrated by the blue arrow, and point ii) implies that the red cross is in the interior of the black segment. Partial smoothness appears in the fact that the black segment is perpendicular to the tangent plane of $\mathcal{M}$ at $\bar{x}$.

proximal gradient of point y. More precisely, we claim that, for y near $\bar{y}$ and $x = \hat{x}(y)$, there holds $0 \in \operatorname{ri} \partial \rho_y(x)$, that is

$$\frac{1}{\gamma}(y - x) - \nabla f(y) \in \operatorname{ri} \partial g(x).$$

This property holds at $(\bar{x}, \bar{y})$ by assumption. By contradiction, assume there exist sequences of points (y_r) with limit $\bar{y}$, $(x_r) = (\hat{x}(y_r))$ with limit $\bar{x} = \hat{x}(\bar{y})$ and (h_r) of unit norm $\|h_r\| = 1$ such that for all r, h_r separates 0 from $\partial \rho_{y_r}(x_r)$:

$$\inf_{h \in \partial \rho_{y_r}(x_r)} \langle h_r, h \rangle \geq 0.$$

Since (h_r) is bounded, a converging subsequence can be extracted from it, let $\bar{h}$ denote its limit. At the cost of renaming iterates, we assume that $\lim_{r\to\infty} h_r = \bar{h}$. The above property still holds at the limit $r \to \infty$. Indeed, let $\bar{u} \in \partial \rho_{\bar{y}}(\bar{x})$. Since g is partly smooth, the mapping $(x, y) \in \mathcal{N}_{\bar{x}} \times \mathcal{N}_{\bar{y}} \mapsto \partial \rho_y(x) = \partial g(x) + \frac{1}{\gamma}(x - y)$ is continuous. Therefore, there exists a sequence (u_r) such that $u_r \in \partial \rho_{y_r}(x_r)$ and $\lim_{r\to\infty} u_r = \bar{u}$. We have for all r: $\langle u_r, h_r \rangle \geq 0$, which yields at the limit $\langle \bar{u}, \bar{h} \rangle \geq 0$. Thus $\bar{h}$ separates 0 from $\partial \rho_{\bar{y}}(\bar{x})$, which contradicts our assumption. Conclusion. We thus have a continuously differentiable function $\hat{x}$ defined on a neighborhood of $\bar{y}$ such that i) $\hat{x}(\bar{y}) = \bar{x}$, ii) $\hat{x}(y)$ is a strong minimizer of ρ_y on $\mathcal{M}$, iii) $0 \in \operatorname{ri} \partial \rho_y(\hat{x}(y))$.

This last point tells us that $(y - \hat{x}(y))/\gamma - \nabla f(y) \in \partial g(\hat{x}(y))$. The characterization of proximity by the optimality condition (Lemma 2.2) gives that $\hat{x}(y) = \operatorname{prox}_{\gamma g}(y - \gamma \nabla f(y))$ for y close enough to $\bar{y}$. □

3.3.2 *Assumptions for structure identification*

Theorem 3.1 captures the localization properties of the proximal gradient operator. It also enables us to precisely define a condition under which a point can be localized. We formalize it in the definition of *r-structured critical points*, an illustration of which is depicted on Fig. 3.3.

Definition 3.1 . A point $\bar{x}$ of a C^2 submanifold $\mathcal{M}$ is *r-structured critical* for (f, g) if we have:

i) proximal gradient stability: $\bar{x} = \operatorname{prox}_{g/r}(\bar{x} - 1/r \nabla f(\bar{x}))$;

ii) qualification condition: $0 \in \operatorname{ri}(\nabla f + \partial g)(\bar{x})$;

iii) prox-regularity: g is r-prox-regular at $\bar{x}$;

iv) partial smoothness: g is partly-smooth at $\bar{x}$ with respect to $\mathcal{M}$.

While points ii), iii), iv) are standard in the literature (see e.g., (Daniilidis et al., 2006)), point i) is not always explicited (an exception is for the notion of identifiability in Drusvyatskiy and Lewis (2014)). It is directly verified when g is convex (for any $r > 0$), but this is not the case when g is nonconvex, as shown in Example 3.5.

Without point i) of our assumption, the following results would still hold by replacing "*for any* $\gamma \in (0, 1/r)$" by "*for* γ *small enough*", we did not take this option since we wanted to de-correlate the local identification from the stepsize choice.

Example 3.5 (Lack of proximal gradient stability). The following example shows that in the nonconvex setting, ii) and iii) do not necessarily imply i). Take f null and g as follows, then the proximity operator of g at 0 writes:

$$g(x) = \begin{cases} x^2/2 & \text{if } |x| \leq 1 \\ 1 - 3x/2 & \text{if } x \geq 1 \\ 1 + 3x/2 & \text{if } x \leq -1 \end{cases}, \qquad \operatorname{prox}_{\gamma g}(0) = \begin{cases} 0 & \text{if } \gamma \in (0, 8/9) \\ \{-3\gamma/2, 0, 3\gamma/2\} & \text{if } \gamma = 8/9 \\ \{-3\gamma/2, 3\gamma/2\} & \text{if } \gamma > 8/9. \end{cases}$$

The function g is 1-prox-regular at 0, there holds $0 \in \operatorname{ri} \partial g(0) = \{0\}$, and yet 0 is not a fixed point of the proximal operator with stepsizes close to 1. □

3.4 GENERAL PROXIMAL ALGORITHM WITH RIEMANNIAN ACCELERATION

As already mentioned, the output of a proximity operator often comes with the knowledge of the manifold on which it lives. In this section, we leverage this property to an algorithmic advantage by temporarily reducing our working space to the identified structure. "Smooth" structures, involving smooth submanifolds and smooth restrictions on it, emerge locally and open the way to Newton acceleration.

Let us start by specifying the blanket assumptions on the problem ($\mathcal{P}$). These assumptions are mostly common except for the third point, which directly comes from our idea of using the proximal operator both for the optimization itself and as an oracle for the current structure of the iterates.

Assumption 3.1 . The functions f and g are proper and

i) f is $\mathcal{C}^2(\mathbb{R}^n)$ with an L-Lipschitz continuous gradient;

ii) g is lower semi-continuous;

iii) $\operatorname{prox}_{\gamma g}$ is non-empty on $\mathbb{R}^n$ for any $\gamma > 0$;

iv) $F(x) = f(x) + g(x)$ is bounded below.

In this setup, we propose a general algorithm (Algorithm 3.1) which consists in, first, performing a proximal gradient step $x_k \in \operatorname{prox}_{\gamma g}(y_{k-1} - \gamma \nabla f(y_{k-1}))$ that provides both the current point x_k and the manifold $\mathcal{M}_k$ where it lies, and, second, carrying out a Riemannian optimization update $\operatorname{ManAcc}_{\mathcal{M}_k}$ on the current manifold. This algorithm is general in the sense that we do not precise for now what is the Riemannian step ManAcc.

In Section 3.4.1, we show that Algorithm 3.1 retains the global convergence properties of the proximal gradient algorithm. In Section 3.4.2, we study how

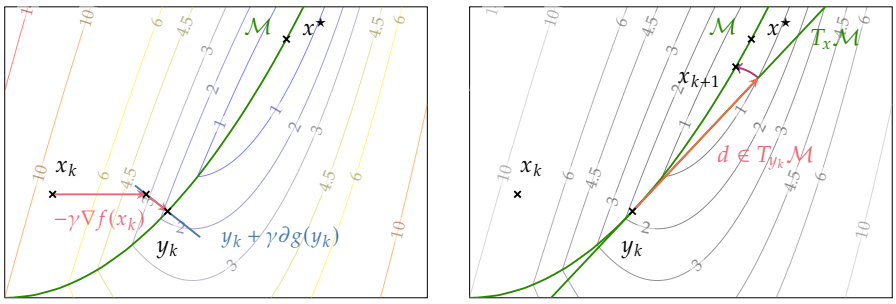


Figure 3.4: Illustration of Algorithm 3.1. Left pane: proximal gradient step from x_k mapped to the optimal manifold $\mathcal{M}^\star$, and the area of points mapped to $\mathcal{M}^\star$ by the proximal gradient. Right pane: ManAcc step, decomposed as a (Riemannian Newton) step $d \in T_x\mathcal{M}^\star$ in the tangent space and its retraction x_{k+1} onto $\mathcal{M}^\star$.

Riemannian methods with local superlinear convergence (such as Riemannian Newton's method) propagate their rate to Algorithm 3.1. We will investigate later in Section 3.5 the Riemannian Newton acceleration falling into this scheme.

Algorithm 3.1: General structure exploiting algorithm

Require: Pick x_0 arbitrary, $\gamma \in (0, 1/L)$.

1: **repeat**
2: Compute $x_k \in \operatorname{prox}_{\gamma g}(y_{k-1} - \gamma\nabla f(y_{k-1}))$ and get $\mathcal{M}_k \ni x_k$
3: Update $y_k = \mathrm{ManAcc}_{\mathcal{M}_k}(x_k)$ on the current manifold
4: **until** stopping criterion

3.4.1 *Global convergence*

The following result shows that Algorithm 3.1 converges to a critical value of F, and all accumulation points of its iterates are critical points. For this to hold, we only need the mild assumption that the manifold update does not increase the functional value. This offers a broad choice of methods since this kind of descent is easily obtained by line search as discussed in Section 3.5.1.

Theorem 3.2 (Global convergence). *Let Assumption 3.1 hold and take $\gamma \in (0, 1/L)$. Suppose that the manifold update $\mathrm{ManAcc}_{\mathcal{M}}$ provides descent, that is for any x in $\mathcal{M}$*

$$F(\mathrm{ManAcc}_{\mathcal{M}}(x)) \leq F(x).$$

Then, Algorithm 3.1 generates non-increasing functional values ($F(x_{k+1}) \leq F(y_k) \leq F(x_k)$ for all k) and all limit points of (x_k) and (y_k) are critical points of F, that share the same functional value.

Proof. It is well-known that the proximal gradient update provides a descent (see the result and reference in Appendix A.2). Choosing y_k such that $F(y_k) \leq F(x_k)$ (by assumption on the manifold update) yields:

$$F(x_{k+1}) \overset{\text{Lemma A.5}}{\leq} F(y_k) - \frac{1-\gamma L}{2\gamma}\|x_{k+1} - y_k\|^2 \leq F(x_k) - \frac{1-\gamma L}{2\gamma}\|x_{k+1} - y_k\|^2. \tag{3.5}$$

The sequence $(F(x_k))$ is thus non-increasing and lower-bounded, therefore it converges. Besides, any accumulation point of (x_k) is a critical point of F. Indeed, summing equation (3.5) for $k = 1, \dots, n$ yields:

$$\frac{1-\gamma L}{2\gamma} \sum_{k=1}^{n} \|x_{k+1} - y_k\|^2 \leq F(x_1) - F(x_{n+1}) \leq F(x_1) - \inf F < +\infty.$$

Therefore the general term of the above series $\|x_{k+1} - y_k\|^2$ converges to 0, which implies, by Lemma A.6, that the distance from $\partial F(x_k)$ to 0 converges to 0. The outer-semi continuity property of the limiting subdifferential allows to conclude that every accumulation point of (x_k) is a critical point of F. Finally, all limit points share the same functional value as F is lower semi-continuous. □

3.4.2 *Local identification and superlinear convergence*

Using the structure identification result of the proximal gradient Theorem 3.1, we can guarantee that our method, Algorithm 3.1, benefits from superlinear convergence provided that the considered Riemannian method converges (locally) superlinearly around a limit point.

Theorem 3.3 (Local convergence). *Let Assumption 3.1 hold and take $\gamma \in (0, 1/L)$, where L is the Lipschitz constant for ∇f. Assume that Algorithm 3.1 generates a sequence (y_k) which admits at least one limit point $\bar{x}$ such that:*

i) $\bar{x} \in \mathcal{M}$ is a r-structured critical point for (f, g) with $r < 1/\gamma$;

ii) $\mathrm{ManAcc}_{\mathcal{M}}$ has superlinear convergence rate of order $1 + \theta \in (1, 2]$ near $\bar{x}$ in $\mathcal{M}$: for some $q > 0$ and $x \in \mathcal{M}$ near $\bar{x}$,

$$\mathrm{dist}^{geo}_{\mathcal{M}}(\mathrm{ManAcc}_{\mathcal{M}}(x), \bar{x}) \leq q \, \mathrm{dist}^{geo}_{\mathcal{M}}(x, \bar{x})^{1+\theta}.$$

Then, after some finite time:

a) the full sequence (x_k) lies on $\mathcal{M}$;

b) x_k converges to $\bar{x}$ superlinearly with the same order as ManAcc*:*

$$\mathrm{dist}^{geo}_{\mathcal{M}}(x_{k+1}, \bar{x}) \leq c \ \mathrm{dist}^{geo}_{\mathcal{M}}(x_k, \bar{x})^{1+\theta} \qquad \text{for some } c > 0. \tag{3.6}$$

Proof. Let us note $\mathsf{T}(y) = \mathrm{prox}_{\gamma g}(y - \gamma \nabla f(y))$ for $y \in \mathbb{R}^n$. The part i) of the assumptions enables us to show the existence of some neighborhood of $\bar{x}$ on which the proximal gradient operation is $\mathcal{M}$-valued and Lipschitz continuous. More precisely, Theorem 3.1 implies that there exists $\delta_1 > 0$ and $C > 0$ such that,

$$\mathsf{T}(y) \in \mathcal{M} \quad \text{and} \quad \|\mathsf{T}(y) - \mathsf{T}(\bar{x})\| \leq C\|y - \bar{x}\| \qquad \text{for all } y \text{ in } \mathcal{B}(\bar{x}, \delta_1).$$

Now, if y belongs to $\mathcal{M}$, we get that there exists $\varepsilon_1 > 0$ such that for any y in $\mathcal{B}_{\mathcal{M}}(\bar{x}, \varepsilon_1)$, $\mathsf{T}(y) \in \mathcal{M}$; but in addition, the Euclidean Lipschitz continuity can be translated into a Riemannian one (see Lemma A.4) since for some $\delta > 0$,

$$\begin{aligned}(1-\delta)\,\mathrm{dist}^{\mathrm{geo}}_{\mathcal{M}}(\mathsf{T}(y), \bar{x}) = (1-\delta)\,\mathrm{dist}^{\mathrm{geo}}_{\mathcal{M}}(\mathsf{T}(y), \mathsf{T}(\bar{x})) &\leq \|\mathsf{T}(y) - \mathsf{T}(\bar{x})\| \\ &\leq C\|y - \bar{x}\| \leq C(1+\delta)\,\mathrm{dist}^{\mathrm{geo}}_{\mathcal{M}}(y, \bar{x}) \end{aligned} \tag{3.7}$$

Hence, there is $q_1 > 0$ such that for any y in $\mathcal{B}_{\mathcal{M}}(\bar{x}, \varepsilon_1)$

$$\operatorname{dist}^{\mathrm{geo}}_{\mathcal{M}}(\mathsf{T}(y), \bar{x}) = \operatorname{dist}^{\mathrm{geo}}_{\mathcal{M}}(\mathsf{T}(y), \mathsf{T}(\bar{x})) \leq q_1 \ \operatorname{dist}^{\mathrm{geo}}_{\mathcal{M}}(y, \bar{x}). \tag{3.8}$$

Then, the part ii) of the assumptions gives us the existence of ε_2, $q_2 > 0$ and $\theta \in (0, 1)$ such that, for any x in $\mathcal{B}_{\mathcal{M}}(\bar{x}, \varepsilon_2)$,

$$\operatorname{dist}^{\mathrm{geo}}_{\mathcal{M}}(\mathrm{ManAcc}_{\mathcal{M}}(x), \bar{x}) \leq q_2 \ \operatorname{dist}^{\mathrm{geo}}_{\mathcal{M}}(x, \bar{x})^{1+\theta}. \tag{3.9}$$

Let us now take any $x \in \mathcal{B}_{\mathcal{M}}(\bar{x}, \varepsilon)$ where $\varepsilon = \min(\varepsilon_1, \varepsilon_2, (\varepsilon_1/q_2)^{\frac{1}{1+\theta}}, (q_2 q_1)^{-\frac{1}{\theta}})$: (*i*) Since $x \in \mathcal{B}_{\mathcal{M}}(\bar{x}, \varepsilon_2)$, the manifold update (3.9) yields

$$\operatorname{dist}^{\mathrm{geo}}_{\mathcal{M}}(\mathrm{ManAcc}_{\mathcal{M}}(x), \bar{x}) \leq q_2 \ \operatorname{dist}^{\mathrm{geo}}_{\mathcal{M}}(x, \bar{x})^{1+\theta} \leq q_2 \ \varepsilon^{1+\theta} \leq \varepsilon_1.$$

(*ii*) As $\mathrm{ManAcc}_{\mathcal{M}}(x)$ lies in $\mathcal{B}_{\mathcal{M}}(\bar{x}, \varepsilon_1)$, the proximal gradient update (3.8) applied to $y = \mathrm{ManAcc}_{\mathcal{M}}(x)$ gives

$$\begin{aligned} \operatorname{dist}^{\mathrm{geo}}_{\mathcal{M}}(\mathsf{T}(\mathrm{ManAcc}_{\mathcal{M}}(x)), \bar{x}) &\leq q_1 \operatorname{dist}^{\mathrm{geo}}_{\mathcal{M}}(\mathrm{ManAcc}_{\mathcal{M}}(x), \bar{x}) \\ &\leq q_1 q_2 \ \operatorname{dist}^{\mathrm{geo}}_{\mathcal{M}}(x, \bar{x})^{1+\theta} \leq q_1 q_2 \ \varepsilon^{\theta} \operatorname{dist}^{\mathrm{geo}}_{\mathcal{M}}(x, \bar{x}). \end{aligned} \tag{3.10}$$

Since $q_2 q_1 \varepsilon^{\theta} \leq 1$ by construction, this gives

$$\operatorname{dist}^{\mathrm{geo}}_{\mathcal{M}}(\mathsf{T}(\mathrm{ManAcc}_{\mathcal{M}}(x)), \bar{x}) \leq \operatorname{dist}^{\mathrm{geo}}_{\mathcal{M}}(x, \bar{x}) \qquad \text{for any } x \in \mathcal{B}_{\mathcal{M}}(\bar{x}, \varepsilon). \tag{3.11}$$

We have thus proved the existence of a neighborhood $\mathcal{B}_{\mathcal{M}}(\bar{x}, \varepsilon)$ of $\bar{x}$ in $\mathcal{M}$ which is stable for an iteration of Algorithm 3.1 and over which one iteration has a superlinear improvement of order $1 + \theta$ (by (3.10)).

Finally, since $\bar{x}$ is a limit point of (y_k), there exists $K < \infty$ such that $y_K \in \mathcal{B}(\bar{x}, (1-\delta)\varepsilon/C)$. Besides, (3.7) tells us that $\operatorname{dist}^{\mathrm{geo}}_{\mathcal{M}}(\mathsf{T}(y_K), \bar{x}) \leq \varepsilon$ and thus x_k and y_k belong to $\mathcal{B}_{\mathcal{M}}(\bar{x}, \varepsilon)$ for all $k > K$ by (3.11). We conclude that $x_{k+1} = \mathsf{T}(y_k) \in \mathcal{M}$ for all $k \geq K$, and, using (3.10), that we have (3.6) with $c = q_1 q_2$, for all $k > K$.□

3.5 RIEMANNIAN NEWTON ACCELERATION, IN PRACTICE

In this section, we investigate the possibilities of manifold acceleration within Algorithm 3.1. We show in Sections 3.5.2 and 3.5.3 how to use Riemannian (truncated) Newton accelerations within our framework and derive quadratic (superlinear) convergence guarantees. A technical difficulty to ensure global convergence when interlacing proximal gradient updates with Riemannian Newton accelerations is to guarantee some functional decrease. Thus, we first study in Section 3.5.1 the use of line search for $\mathrm{ManAcc}_{\mathcal{M}}$ in our context.

3.5.1 *Ensuring functional descent while preserving local rates: line search*

We use in the following convergence proofs three properties of $\mathrm{ManAcc}_{\mathcal{M}}$: it should produce an update that lives on $\mathcal{M}$, enjoy a superlinear local convergence rate, and not degrade function value. For this last point, we consider a simple line search and we prove that, under mild assumptions, it helps to find a point which decreases function value, and retains the favorable local properties of Newton's method. Surprisingly, this result does not appear in the standard references on Riemannian optimization. We provide here the necessary developments inspired from the classical monograph by Dennis Jr and Schnabel (1996).

Standing at point $x \in \mathcal{M}$ with a proposed direction $\eta \in T_x\mathcal{M}$, a stepsize $\alpha > 0$ is *acceptable* if it satisfies the following *Armijo* condition

$$F(\mathrm{R}_x(\alpha\eta)) \leq F(x) + m_1\alpha\langle \operatorname{grad} F(x), \eta\rangle, \qquad \text{for } 0 < m_1 < 1/2. \tag{3.12}$$

The line search employs a second-order retraction R_x, e.g., the exponential map, a projection retraction (Absil and Malick, 2012), or any other second-order retraction (Boumal, 2022).[1] The conditions under which stepsizes satisfying the Armijo rule exist are discussed in Dennis Jr and Schnabel (1996, Sec 6.3), the following lemma can then be derived.

Lemma 3.4 . *Let Assumption 3.1 hold and consider a manifold $\mathcal{M}$ equipped with a retraction R and a pair $(x, \eta) \in T\mathcal{B}$. If F is differentiable on $\mathcal{M}$ at x, $\langle \operatorname{grad} F(x), \eta\rangle < 0$, and $m_1 < 1$, then there exists $\hat{\alpha} > 0$ such that any stepsize $\alpha \in (0, \hat{\alpha})$ is acceptable by the Armijo rule (3.12).*

Proof. We adapt a part of the proof of Dennis Jr and Schnabel (1996, Th. 6.3.2) for the Armijo rule and the Riemannian setting. Since $m_1 < 1/2$, for any α sufficiently small there holds

$$F \circ \mathrm{R}_x(\alpha\eta) \leq F \circ \mathrm{R}_x(0) + m_1 \,\mathrm{D}\,(F \circ \mathrm{R}_x)\,(0)[\alpha\eta] = F(x) + m_1\alpha\langle \operatorname{grad} F(x), \eta\rangle.$$

Since F is bounded below, there exists a smallest $\hat{\alpha}$ such that $F(\mathrm{R}_x(\hat{\alpha}\eta)) = F(x) + m_1\hat{\alpha}\langle \operatorname{grad} F(x), \eta\rangle$. Thus all stepsizes in $(0, \hat{\alpha})$ are acceptable by (3.12). □

In addition, a line search performed near a minimizer with a Newton direction should accept the unit stepsize, so that *the* full Newton step may be taken. This is the case when the Riemannian Hessian around this minimizer is positive definite as stated by the next lemma, which is a direct corollary of Theorem B.1.

Lemma 3.5 . *Let Assumption 3.1 hold and consider a manifold $\mathcal{M}$ equipped with a retraction R, a point $x^\star \in \mathcal{M}$ and a pair $(x, \eta) \in T\mathcal{B}$. Assume that F is twice differentiable on $\mathcal{M}$ near $x^\star$ and that $x^\star$ is a strong local minimizer on $\mathcal{M}$, that is $\operatorname{Hess} f(x^\star)$ is positive definite. If the direction η brings a superlinear improvement towards $x^\star$, that is $\operatorname{dist}^{geo}_{\mathcal{M}}(\mathrm{R}_x(\eta), x^\star) = o(\operatorname{dist}^{geo}_{\mathcal{M}}(x, x^\star))$ as $x \to x^\star$, and $0 < m_1 < 1/2$, then η is acceptable by the Armijo rule (3.12) with unit stepsize $\alpha = 1$.*

In the following, we will consider a *backtracking* line search for finding an acceptable stepsize α: the unit stepsize is first tried, and then the search space is reduced geometrically. In practice, we use exactly Dennis Jr and Schnabel (1996, Alg. A6.3.1), which features polynomial interpolation of $f \circ \mathrm{R}_x$ in the search space.

3.5.2 *Riemannian Newton & quadratic convergence*

We construct a manifold update based on the Riemannian Newton method (Absil et al., 2009a, Chap. 6), which is the simplest method with a local quadratic convergence. It consists in finding $d \in T_x\mathcal{M}$ that minimizes the second order model (2.3) of F at point $x \in \mathcal{M}$, or equivalently that solves Newton equation (see Boumal (2022, Sec. 6.2)):

$$\operatorname{grad} F(x) + \operatorname{Hess} F(x)[d] = 0 \tag{Newton eq.}$$

1 Indeed, in many applications of Riemannian optimization, computing geodesics and the exponential map can be costly and then retractions provide an efficient alternative. For this reason, we consider here second-order retractions (Absil et al., 2009a; Boumal, 2022).

Algorithm 3.2: ManAcc-Newton

Require: Manifold $\mathcal{M}$, point $x \in \mathcal{M}$
1: Find d in $T_x\mathcal{M}$ that solves (Newton eq.)
2: Find α satisfying the Armijo condition (3.12) with direction d
3: **return** $y = R_x(\alpha d)$

Theorem 3.6 . *Let Assumption 3.1 hold and take $\gamma \in (0, 1/L)$. Consider the sequence of iterates (x_k) generated by Algorithm 3.1 equipped with the Riemannian Newton manifold update (Algorithm 3.2). If* $\operatorname{Hess} F(x_k)$ *is positive definite at each step, then all limit points of (x_k) are critical points of F and share the same functional value.*

Furthermore, assume that the sequence (y_k) admits a limit point $x^\star$ such that

i) $x^\star \in \mathcal{M}$ is a r-structured critical point for (f, g) with $r < 1/\gamma$;

ii) $\operatorname{Hess}_{\mathcal{M}} F(x^\star) \succ 0$ *and* $\operatorname{Hess}_{\mathcal{M}} F$ *is locally Lipschitz around $x^\star$.*

Then, after some finite time,

a) the sequence (x_k) lies on $\mathcal{M}$;

b) x_k converges to $x^\star$ quadratically: for large k, there exists $c > 0$ such that

$$\operatorname{dist}^{geo}_{\mathcal{M}}(x_{k+1}, x^\star) \leq c \ \operatorname{dist}^{geo}_{\mathcal{M}}(x_k, x^\star)^2.$$

Proof. As the Riemannian Hessian is assumed to be positive definite, Newton's direction is a descent direction:

$$\langle \operatorname{grad} F(x_k), d_k \rangle = -\langle \operatorname{grad} F(x_k), \operatorname{Hess} F(x_k)^{-1} \operatorname{grad} F(x_k) \rangle < 0.$$

The Riemannian Newton manifold step is therefore well-defined, and the line search terminates by Lemma 3.4, so that the manifold update is well-defined and provides descent ($F(y_k) \leq F(x_k)$). Theorem 3.2 thus ensures that all limit points of (x_k) are critical points of F and share the same functional value.

Now we apply the local convergence of Riemannian Newton (Absil et al., 2009a, Th. 6.3.2): assumption ii) ensures that the Riemannian Newton direction d computed in step 1 of Algorithm 3.2 provides a quadratic improvement on a neighborhood of $x^\star$ on $\mathcal{M}$. Moreover, the line search returns the unit-stepsize after some finite time: $\alpha = 1$ is tried first, and is acceptable for directions providing superlinear improvement by Lemma 3.5. Thus the whole Riemannian Newton update provides quadratic improvement after some finite time. Using this and assumption i), Theorem 3.3 applies and yields the results. □

This theorem states that alternating proximal gradient steps and Riemannian Newton steps converges quadratically to structured points under virtually the same assumptions required by the Euclidean Newton method. The two standard issues of Newton's method therefore hold in our setting: at each iteration, a linear system has to be solved to produce the Newton direction; and this direction does not always provide descent (without positive definiteness of the Hessian). We show in the next section that truncated versions overcome these issues in our framework.

3.5.3 *Riemannian Truncated Newton & superlinear convergence*

We consider a manifold update based on a truncated Newton procedure (Dembo and Steihaug, 1983). (Riemannian) Truncated Newton consists in solving

(Newton eq.) partially by using a (Riemannian) conjugate gradient procedure so that whenever the resolution of (Newton eq.) is stopped, the resulting direction provides descent on the function. The quality of the truncated Newton direction is controlled by a parameter $\eta \in [0, 1)$ which bounds the ratio of residual and gradient norms:

$$\| \operatorname{grad} F(x) + \operatorname{Hess} F(x)[d] \| \leq \eta \| \operatorname{grad} F(x) \|. \qquad \text{(Inexact Newton eq.)}$$

Algorithm 3.3: ManAcc-Newton-CG

Require: Manifold $\mathcal{M}$, point $x \in \mathcal{M}$, convergence defining parameter $\theta \in (0, 1]$
1: Let $\eta = \| \operatorname{grad} F(x) \|^{\theta}$
2: Find d that solves (Inexact Newton eq.)
3: Find α satisfying the Armijo condition (3.12) with direction d
4: **return** $y = R_x(\alpha d)$

Theorem 3.7 . *Let Assumption 3.1 hold and take $\gamma \in (0, 1/L)$. Consider the sequence of iterates (x_k) generated by Algorithm 3.1 equipped with the Riemannian Truncated Newton manifold update (Algorithm 3.3). Then all limit points of (x_k) are critical points of F and share the same function value.*

Furthermore, assume that sequence (y_k) admits a limit point $x^\star$ such that

i) $x^\star \in \mathcal{M}$ is a r-structured critical point for (f, g) with $r < 1/\gamma$;

ii) $\operatorname{Hess}_{\mathcal{M}} F(x^\star) \succ 0$ and $\operatorname{Hess}_{\mathcal{M}} F$ is locally Lipschitz around $x^\star$.

iii) we take $\eta_k = \mathcal{O}(\| \operatorname{grad} F(x_k) \|^{\theta})$, for some $\theta \in (0, 1]$.

Then, for k large enough, the full sequence (x_k) lies on $\mathcal{M}$, and x_k converges to $x^\star$ superlinearly with order $1 + \theta$: for large k, there exist $c > 0$,

$$\operatorname{dist}^{geo}_{\mathcal{M}}(x_{k+1}, x^\star) \leq c \ \operatorname{dist}^{geo}_{\mathcal{M}}(x_k, x^\star)^{1+\theta}.$$

Proof. Applying the analysis of Dembo and Steihaug (1983, Lemma A.2) to the approximate resolution of (Inexact Newton eq.) on the euclidean space $T_x\mathcal{M}$ yields:

$$\langle \operatorname{grad} F(x), d \rangle \leq -\min(1, \| \operatorname{Hess} F(x) \|^{-1}) \| \operatorname{grad} F(x) \|^2.$$

Therefore, the direction provided by (Inexact Newton eq.) is a descent direction, and by Lemma 3.4 the line search terminates: the updates are well-defined and provide descent. Thus, as in the proof of Theorem 3.6 we get that every accumulation point of the iterate sequence is a critical point for F. We can apply now the local convergence of the Riemannian truncated Newton method (Absil et al., 2009a, Th. 8.2.1): assumptions ii) and iii) ensure that the direction d computed in step 1 of Algorithm 3.3 provides a local superlinear improvement towards $x^\star$. The end of the proof is the same as the one of the proof of Theorem 3.6. □

3.6 NUMERICAL ILLUSTRATIONS

In this section, we illustrate the effect of Newton acceleration. We consider Algorithm 3.1 equipped with either the Newton update of Algorithm 3.2, denoted 'Alt. Newton' or the truncated Newton update of Algorithm 3.3, denoted 'Alt. Truncated Newton'. These methods are compared to the Proximal Gradient and the Accelerated Proximal Gradient, which serve as baseline. The

algorithms and problems are implemented in Julia (Bezanson et al., 2017); experiments may be reproduced using the code available at https://github.com/GillesBareilles/NewtonRiemannAccel-ProxGrad.

We report the numerical results in figures showing a) the suboptimality $F(x_k) - F(x^\star)$ of the current iterate x_k versus time, and b) the dimension of the current manifold $\mathcal{M}_k \ni x_k$ versus iteration. We also report a table comparing the algorithms at the first iteration that makes suboptimality lower than tolerances 10^{-3} and 10^{-9} for various measures summarized in the following table:

$F(x_k) - F(x^\star)$	Suboptimality at current iteration.
#prox. grad. steps	Number of proximal gradient steps, each involve computing $\nabla f(\cdot)$ and $\operatorname{prox}_{\gamma g}(\cdot)$ once.
#ManAcc steps	Number of Riemannian steps, each involve computing grad $F(\cdot)$ once and Hess $F(\cdot)[\cdot]$ multiple times (one per Conjugate Gradient iteration).
#Hess $F(\cdot)[\cdot]$	Number of Riemannian Hessian-vector products, approximates the effort spent in manifold updates since algorithm started.
$\#f$	Number of calls to $f(x)$, one per iteration + some for the line search + some for the backtracking estimation of the Lipschitz constant.
$\#g$	Number of calls to $g(x)$, one per iteration + some for the line search.

The proximal gradient updates, present in all methods, include a backtracking procedure that maintains an estimate of the Lipschitz constant of ∇f, so that the proximal gradient step length is taken as the inverse of that estimate. The Conjugate Gradient used to solve (Newton eq.) and (Inexact Newton eq.) follows Boumal (2022, Alg. 6.2); it is stopped when the (in)exactness criterion is met, or after 50 iterations for the logistic problem and 150 for the trace-norm one, or when the inner direction d makes the ratio $\langle \operatorname{Hess} F(x_k)[d], d\rangle / \|d\|^2$ small.[2] The manifold updates are completed by a backtracking line search started from unit stepsize, a direct implementation of Dennis Jr and Schnabel (1996, Alg. 6.3.1).

3.6.1 *Two-dimensional nonsmooth example*

We consider the piecewise quadratic problem of Lewis and Wylie (2019):

$$\min_{x\in\mathbb{R}^2} 2x_1^2 + x_2^2 + |x_1^2 - x_2|.$$

The objective function is partly-smooth relative to the parabola $\{x : x_2 = x_1^2\}$, for which an expression for the tangent space, the orthogonal projection on tangent space, a second-order retraction and conversion from Euclidean gradients and Hessian-vector products to Riemannian ones are readily available.

We detail here the oracles of $f(x) \triangleq 2x_1^2 + x_2^2$ and $g(x) \triangleq |x_1^2 - x_2|$:

2 Each CG iteration requires one Hessian-vector product, avoiding to form the Hessian matrix. A test on this ratio is used to detect a direction of quasi-negative curvature for the (Riemannian) Hessian, which is a stopping criterion of the Conjugate Gradient. In our implementation, we require this quantity to be smaller than 10^{-15} for the Newton method. For the truncated version, we reduce the threshold when getting close to the solution: initialized at 1, the threshold is decreased by a factor 10 each time the unit-step is accepted by the line search.

- *proximity operator:* For $\gamma < 1/2$, there holds

$$\operatorname{prox}_{\gamma g}(x) = \begin{cases} (\frac{x_1}{1+2\gamma}, x_2 + \gamma) & \text{if } x_2 \leq \frac{x_1^2}{(1+2\gamma)^2} - \gamma \\ (\frac{x_1}{1+4\gamma t-2\gamma}, x_2 + 2\gamma t - \gamma) & \text{if } \frac{x_1^2}{(1+2\gamma)^2} - \gamma \leq x_2 \leq \frac{x_1^2}{(1-2\gamma)^2} + \gamma \\ (\frac{x_1}{1-2\gamma}, x_2 - \gamma) & \text{if } \frac{x_1^2}{(1-2\gamma)^2} + \gamma \leq x_2 \end{cases}$$

  where t solves $x_2^2 + (-2\gamma t + \gamma - x_2)(1 + 4\gamma t - 2\gamma)^2 = 0$.

- *Riemannian gradient and Hessian:* Since g is identically null on $\mathcal{M}$, for any point $(x, \eta) \in T\mathcal{B}$, $\operatorname{grad} g(x) = 0$ and $\operatorname{Hess} g(x)[\eta] = 0$. Moreover, Euclidean gradient and Hessian-vector product are converted to Riemannian ones using equations (2.1) and (2.2):

$$\operatorname{grad} f(x) = \operatorname{proj}_x(\nabla f(x))$$
$$\operatorname{Hess} f(x)[\eta] = \operatorname{proj}_x\left(\nabla^2 f(x)[\eta] - \begin{pmatrix} 2\eta_1 \\ 0 \end{pmatrix} \left\langle \nabla f(x), \begin{pmatrix} 2x_1 \\ -1 \end{pmatrix} \right\rangle \frac{1}{1+4x_1^2}\right),$$

  and the orthogonal projection onto $T_x\mathcal{M}$ writes

$$\operatorname{proj}_x(d) = d - \left\langle d, \begin{pmatrix} 2x_1 \\ -1 \end{pmatrix} \right\rangle \frac{1}{1+4x_1^2} \begin{pmatrix} 2x_1 \\ -1 \end{pmatrix}.$$

We run the proximal gradient, its accelerated counterpart, and Algorithm 3.1 with the Newton update Algorithm 3.2. For our illustrative purposes, the proximal gradient steps of all algorithms have a constant stepsize $\gamma = 0.05$, all algorithms are started from point $(2, 3)$.

OBSERVATIONS. The iterates are displayed in Fig. 3.5. The Proximal Gradient iterates reach the parabola in finite time, and then converge linearly on the parabola while the Accelerated Proximal Gradient iterates "overshoot" the optimal manifold (see Bareilles and Iutzeler (2020)). The iterates of the Alt. Newton method stay on the parabola and the quadratic convergence behavior appears clearly since two Newton updates bring suboptimality below 10^{-3}, and one additional step gets it below 10^{-12}.

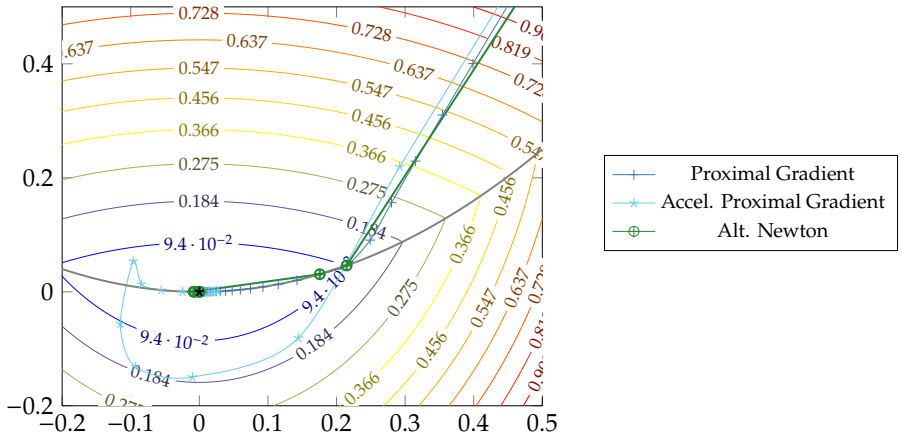
3.6.2 ℓ_1-regularized logistic problem

We now turn to the ℓ_1-regularized logistic problem:

$$\min_{x \in \mathbb{R}^n} \frac{1}{m} \sum_{i=1}^{m} \log(1 + \exp(-y_i \langle A_i, x \rangle)) + \lambda \|x\|_1,$$

where $A \in \mathbb{R}^{m \times n}$, $y \in \{-1, 1\}^m$, and $\lambda > 0$. The nonsmooth part $g(x) = \lambda \|x\|_1$ is described in Section 3.2.

We consider an instance where $n = 4000$, $m = 400$, $\lambda = 10^{-2}$ and the final manifold has dimension 249. The coefficients of A are drawn independently following a normal law. From a sparse random vector s, y_i is set to 1 with probability $1/(1 + \exp(-\langle A_i, s \rangle))$, and -1 otherwise. All algorithms start from



Algorithm	tol	F-F*	proxgrad	gradF	HessF	f	g
Prox. Gradient	$1 \cdot 10^{-3}$	$7.74 \cdot 10^{-4}$	29	–	–	30	30
Prox. Gradient	$1 \cdot 10^{-9}$	$7.59 \cdot 10^{-10}$	60	–	–	61	61
Accel. Prox. Gradient	$1 \cdot 10^{-3}$	$9.63 \cdot 10^{-4}$	16	–	–	17	17
Accel. Prox. Gradient	$1 \cdot 10^{-9}$	$5.18 \cdot 10^{-10}$	63	–	–	64	64
Alt. Newton	$1 \cdot 10^{-3}$	$1.49 \cdot 10^{-4}$	2	2	10	7	7
Alt. Newton	$1 \cdot 10^{-9}$	$8.75 \cdot 10^{-13}$	3	3	15	10	10

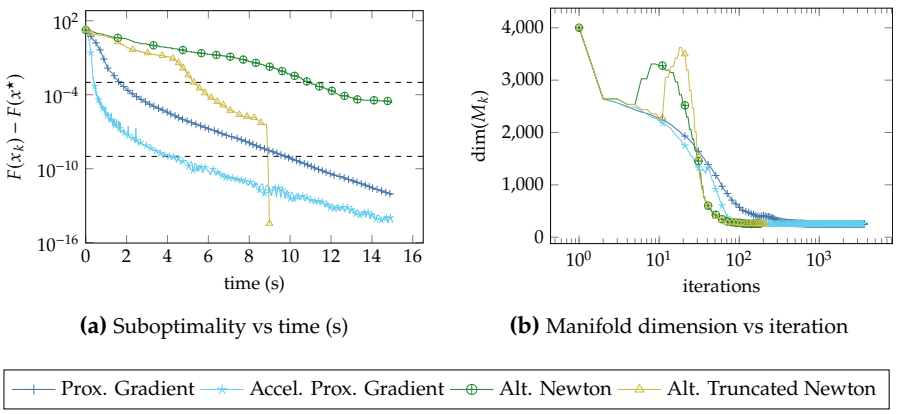
Figure 3.5: Nonsmooth example

the same point which is the output of 35 iterations of the accelerated proximal gradient randomly initiated.

OBSERVATIONS. The experiments are presented in Fig. 3.6.[3] The optimal manifold is identified around iteration 200 for all methods except for Proximal Gradient, which needs 1000 iterations. The two baselines Proximal Gradient and its accelerated version show linear convergence, with a better rate for the non accelerated version once the final manifold is reached. Alt. Truncated Newton shows superlinear acceleration, while Alt. Newton fails to converge in the given time budget.

As iterations grow, the (Accelerated) Proximal Gradient identifies manifolds of decreasing dimension in a roughly monotonical way. Alt. Truncated Newton behaves differently: after identifying monotonically manifolds of dimension lower than 2000, the dimension of the current manifold jumps to about 3000 for about 10 iterations, to finally reach quickly the final manifold. We believe that this partial loss of identified structure is caused by iterates getting close to a point having one non-null but very small coordinate. There, the second-order Taylor extension is valid on a small set however it may lead to a Newton step that lies outside that set, thus driving the iterate away. The same behavior occurs for Alt. Newton. This difficulty can be related to the well-known problem of constraint activation in nonlinear programming. Despite this, Algorithm 3.1 retains a good rate overall.

3 In Figs. 3.6a and 3.6b, the marks help distinguish curves. In particular they do not indicate that the algorithm has performed one (or any number of) iteration.



(a) Suboptimality vs time (s)

(b) Manifold dimension vs iteration

Algorithm	tol	F-F*	proxgrad	gradF	HessF	f	g
Prox. Gradient	$1 \cdot 10^{-3}$	$9.96 \cdot 10^{-4}$	357	–	–	779	358
Prox. Gradient	$1 \cdot 10^{-9}$	$9.97 \cdot 10^{-10}$	2,306	–	–	4,677	2,307
Accel. Prox. Gradient	$1 \cdot 10^{-3}$	$9.26 \cdot 10^{-4}$	90	–	–	246	91
Accel. Prox. Gradient	$1 \cdot 10^{-9}$	$9.9 \cdot 10^{-10}$	953	–	–	1,972	954
Alt. Newton	$1 \cdot 10^{-3}$	$9.76 \cdot 10^{-4}$	62	61	6,303	556	427
Alt. Newton	$1 \cdot 10^{-9}$	–	–	–	–	–	–
Alt. Truncated Newton	$1 \cdot 10^{-3}$	$9.56 \cdot 10^{-4}$	51	50	2,616	437	321
Alt. Truncated Newton	$1 \cdot 10^{-9}$	$3.77 \cdot 10^{-15}$	105	105	5,091	742	572

Figure 3.6: Logistic-ℓ_1 problem

3.6.3 *Trace-norm regularized problem*

We consider the following matrix regression problem:

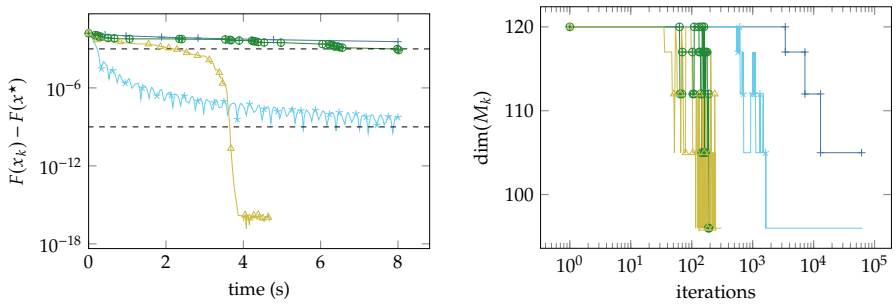
$$\min_{x \in \mathbb{R}^{n_1 \times n_2}} \frac{1}{2} \sum_{i=1}^{m} \left(\langle A_i, x \rangle - y_i \right)^2 + \lambda \|x\|_*, \tag{3.13}$$

where $A_i \in \mathbb{R}^{n_1 \times n_2}$ for $i = 1, \ldots, m$, $y \in \mathbb{R}^m$ and λ denotes a positive scalar. The nonsmooth part $g(x) = \lambda \|x\|_*$ is described in Section 3.2.

We consider an instance of (3.13) where $n_1 = 10$, $n_2 = 12$, $m = 60$, $\lambda = 10^{-2}$ and the final manifold is that of matrices of rank 6. The coefficients of the A_i's are drawn independently from a normal law. From a sparse random vector s, y_i is taken as $\langle A_i, s \rangle + \xi_i$, where ξ_i follows a centered normal law with variance 0.01^2. All algorithms start from the same point which is the output of 10^3 iterations of the accelerated proximal gradient randomly initiated.

OBSERVATIONS. The experiments are presented in Fig. 3.7.[4] We see on Fig. 3.7a that the Proximal Gradient algorithm and its accelerated version converge sublinearly. We conjecture that this slow rate is due to the lack of strong convexity of the objective problem. Alt. Truncated Newton converges superlinearly, and shows the interest of the Newtonian acceleration. Figure 3.7b shows that the Proximal Gradient does not reach the final optimal manifold within the budget of iterations; similarly for the Newton method, within the budget of time.

4 In Figs. 3.7a and 3.7b, the marks help distinguish curves. In particular they do not indicate that the algorithm has performed one (or any number of) iteration.



(a) Suboptimality vs time (s) **(b)** Manifold dimension vs iteration

Algorithm	tol	F-F*	proxgrad	gradF	HessF	f	g
Prox. Gradient	$1 \cdot 10^{-3}$	–	–	–	–	–	–
Prox. Gradient	$1 \cdot 10^{-9}$	–	–	–	–	–	–
Accel. Prox. Gradient	$1 \cdot 10^{-3}$	$9.99 \cdot 10^{-4}$	1,489	–	–	3,073	1,490
Accel. Prox. Gradient	$1 \cdot 10^{-9}$	$9.86 \cdot 10^{-10}$	43,283	–	–	86,661	43,284
Alt. Newton	$1 \cdot 10^{-3}$	$9.83 \cdot 10^{-4}$	93	93	28,063	873	687
Alt. Newton	$1 \cdot 10^{-9}$	–	–	–	–	–	–
Alt. Truncated Newton	$1 \cdot 10^{-3}$	$9.7 \cdot 10^{-4}$	76	76	16,342	738	568
Alt. Truncated Newton	$1 \cdot 10^{-9}$	$2.27 \cdot 10^{-11}$	128	128	27,786	1,101	879

Figure 3.7: Trace-norm problem

3.6.4 *Robustness of Riemannian Newton accelerations*

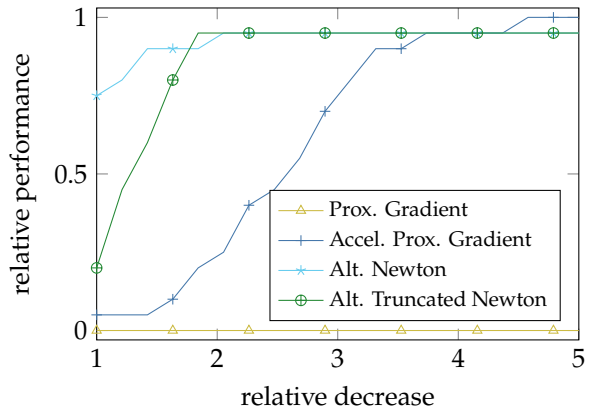


Figure 3.8: Performance profile for the time to decrease suboptimality below 10^{-9}

We illustrate the robustness of the Newton acceleration on several instances of the trace-norm problem (3.13). More precisely, we compare the 4 algorithms on 20 random instances of the tracenorm problem, in terms of wallclock time required to reach a suboptimality of 10^{-9}. We then provide in Fig. 3.8 a performance profile (i.e., the ordinate of a curve at abscissa $t \geq 1$ indicates the proportion of problems for which the corresponding algorithm was able to satisfy the criterion within t times the best algorithm time for each problem; see Dolan and Moré (2002)).

We observe the following on Fig. 3.8. The ordinate at origin of a curve gives the proportion of problems for which the corresponding algorithm performed best: methods with Newton acceleration are the most efficient in 95%(= 75% + 20%) of the instances. Furthermore, they require about 2.5× less time to converge in half of the instances. Note also that the proximal gradient is completely outperformed by the others algorithms since it takes 5× more time than the best algorithm, for all instances.

4

LOCAL NEWTON METHOD FOR NONSMOOTH COMPOSITE MINIMIZATION

This chapter incorporates material from Bareilles et al. (2022a).

4.1 INTRODUCTION

In this chapter, we consider nonsmooth optimization problems of the form

$$\min_{x \in \mathbb{R}^n} F(x) \triangleq g(c(x)), \tag{4.1}$$

where the inner mapping $c : \mathbb{R} \to \mathbb{R}^m$ is smooth and the outer function $g : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ is nonsmooth and may be nonconvex, but admits an explicit proximity operator. In Chapter 3, we considered the additive composite model, we now turn to the composition model. This model is more difficult to handle in general, and encompasses the additive model as a particular case. Here, we illustrate our developments on two classes of functions: the pointwise maximum of m smooth real-valued functions c_i

$$F(x) = \max_{i=1,\dots,m} (c_i(x)) \tag{4.2}$$

and the maximum eigenvalue of a parametrized symmetric real matrix c

$$F(x) = \lambda_{\max}(c(x)). \tag{4.3}$$

In these two examples and many others, subgradients of F can be computed and thus the composite function can be minimized using standard nonsmooth optimization algorithms (e.g.,subgradient methods, gradient sampling (Burke et al., 2020), nonsmooth BFGS (Lewis and Overton, 2013),[1] or black box bundle methods (Hiriart-Urruty and Lemaréchal, 1993)). Nevertheless, these methods do not exploit the fact that F is a composition of a smooth mapping c, which can hinder their performance. In contrast, the so-called prox-linear methods leverage this composite expression by introducing an extension of the proximity operator where the nonlinear mapping c is iteratively replaced by a first-order Taylor approximation (Lewis and Wright, 2016). These methods benefit from theoretical convergence guarantees, and nicely generalize to Taylor-like approximations (Drusvyatskiy et al., 2021; Bolte et al., 2020). However these methods are not always directly implementable because the prox-linear step may be hard to compute, as in (4.3). In a similar spirit, a variant of bundle methods also exploits the composite expression of the function, by leveraging the subgradients of the outer nonsmooth function and the derivatives of the

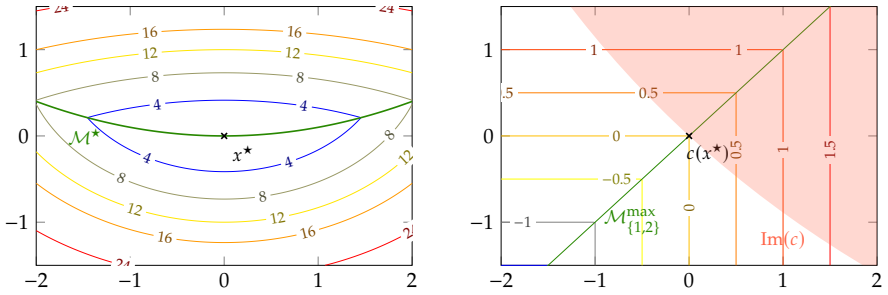
1 Contrary to the other mentioned algorithms, nonsmooth BFGS has no theoretical convergence guarantees when applied to nonsmooth functions.

inner mapping; see Sagastizábal (2013) for general composite functions (4.1) and Helmberg and Rendl (2000) for (4.3).

In this chapter, we propose an optimization algorithm for solving (4.1) exploiting that the nonsmooth objective function $F = g \circ c$ writes as a composition between a smooth mapping c and a simple nonsmooth function g, which displays some smooth substructure, as discussed below.

4.1.1 *Smooth substructure, identification, and existing algorithms*

For many composite functions, including (4.2) and (4.3), the nondifferentiability points locally organize into *smooth manifolds over which F evolves smoothly*. We illustrate in Figure 4.1 such a smooth substructure for a maximum of two functions.



(a) Level curves of F **(b)** Level curves of g and image of c

Figure 4.1: Smooth substructure on a simple example ($n = m = 2$). The figures show the level curves of $g(y) = \max(y_1, y_2)$ (on the right, in the intermediate space) and of $F = g \circ c$, with two quadratic functions $c_1(x)$, $c_2(x)$ (on the left, in the input space). The manifolds of non-differentiability are in green; the image of c is the red area.

The smooth substructure of F can help in solving (4.1). Indeed, if the optimal solution $x^\star$ belongs to a manifold $\mathcal{M}^\star$ that is known beforehand, then minimizing the nonsmooth function F over $\mathbb{R}^n$ boils down to minimizing the smooth restriction $F|_{\mathcal{M}^\star}$ over this smooth *optimal manifold* $\mathcal{M}^\star$. This would enable to solve (4.1) by smooth constrained optimization algorithms, such as Sequential Quadratic Programming (SQP) methods (see e.g., Nocedal and Wright (2006); Bonnans et al. (2006)). The main difficulty in practice is that *we do not know $\mathcal{M}^\star$ in advance*.

Thus, the algorithms exploiting this smooth substructure require two ingredients:

i) a mechanism to identify the optimal manifold;

ii) an efficient method to minimize F restricted to this manifold.

For general convex functions, the algorithm of Mifflin and Sagastizábal (2005) mixes a proximal bundle iteration (as a heuristic for identification) and a so-called $\mathcal{U}$-Newton iteration (which interprets as an SQP step; see Miller and Malick (2005, Sec. 5)). The obtained superlinear rate hinges on the identification of the optimal manifold.

For max-of-smooth functions (4.2), Womersley and Fletcher (1986) pioneered the idea of seeking the optimal manifold and using it to make second-order steps. Their identification heuristic uses the indices of the maximal function

along a descent direction. Recently, Lewis and Wylie (2019); Han and Lewis (2023) investigate a related setting and propose bundle-like algorithms incorporating high-order information that converge (super)linearly on max-of-smooth functions when the optimal manifold is known.

For the maximum eigenvalue of a parametrized matrix (4.3), a specific version of the $\mathcal{U}$-Newton method discussed above is studied by Noll and Apkarian (2005). Again, the identification mechanism is a heuristic determining the multiplicity of the maximal eigenvalue and the optimization step is an SQP iteration.

None of these methods guarantee identification of the optimal manifold: they either assume that the optimal manifold is known in advance, or rely on heuristics for identification. Here, we aim at further harnessing the smooth substructure of $F = g \circ c$ to have *guaranteed local identification* of the optimal manifold and then *guaranteed quadratic convergence* when using SQP iterations.

4.1.2 *Contributions and outline*

We propose a local second-order algorithm for solving the nonsmooth composite problem (4.1) that identifies the optimal manifold of non-differentiability. The two main ingredients of our algorithm are the following:

i) we use the explicit proximal operator of g with chosen stepsizes to provide a guaranteed identification procedure;

ii) for a candidate manifold $\mathcal{M}$, we make an SQP iteration minimizing a smooth extension of $F|_{\mathcal{M}}$ subject to the constraint of belonging to $\mathcal{M}$.

Proximal-based operators identify the manifolds of minimizers under some natural geometrical assumptions: we looked at the proximal and proximal gradient operators in Proposition 2.3 and Theorem 3.1; see also Lee (2023); Lewis and Wright (2016). Here, we only have access to the proximity operator of g. In order to exploit the structure it provides, we face the double challenge of, first, identifying the smooth structure around a point which is not a minimizer for g, and, second, deducing the corresponding structure of $F = g \circ c$. Thus, our main technical contribution is to establish that $\operatorname{prox}_{\gamma g}$ maps a point y close to $c(x^\star)$ to $c(\mathcal{M}^\star)$. The step γ should be carefully chosen, in particular larger than the distance of y to $c(\mathcal{M}^\star)$.

We combine this new identification result with standard SQP-steps to propose a local algorithm for minimizing the composite function F. We pay a special attention to prevent the quadratic convergence of SQP from jeopardizing identification: we prove that, for a well-chosen stepsize policy, the method identifies the optimal structure and locally converges quadratically. We illustrate numerically these properties on problems of the form (4.2) and (4.3).

OUTLINE OF THIS CHAPTER. First, in Section 4.2, we illustrate the technical tools to describe the manifold identification brought by proximity operators. Furthermore, we lay out two technical properties needed for proximal identification in the composite setting. In Section 4.3, we show our main result consisting in a description of a stepsize range for which the proximity operator of g identifies the optimal manifold locally around a minimizer. In Section 4.4 we detail the proposed method combining SQP-steps and proximal identification steps. Finally, we present in Section 4.5 numerical illustrations of our method and of the identification result. Some results concerning our two examples have been deferred to Appendix C.

4.2 SETTING AND ASSUMPTIONS

Let us start by representing schematically the type of functions we consider:

$$\mathbb{R}^n \xrightarrow[\text{smooth mapping}]{c} \operatorname{Im}(c) \subset \mathbb{R}^m \xrightarrow[\text{nonsmooth function}]{g} \mathbb{R} \cup \{+\infty\}.$$

Throughout the chapter, we denote by x points in the *input space* $\mathbb{R}^n$ and by y points in the *intermediate space* $\mathbb{R}^m$.

In all the results presented in this chapter, we make the following assumption that describes the minimal global properties on g and c to conduct our reasoning.

Assumption 4.1 . The mapping $c : \mathbb{R}^n \to \mathbb{R}^m$ is $\mathcal{C}^2$, the function $g : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ is proper and lower semi-continuous.

In the remainder of this section, we illustrate on our running examples the proximity operator in Section 4.2.1 and the structure manifolds in Section 4.2.2. Then, in Section 4.2.3 we introduce two assumptions required for identification and show they hold on our examples.

4.2.1 *Proximity operator: examples*

Example 4.1 (Maximum). The subdifferential of $g(y) = \max(y_1, \dots, y_m)$ is

$$\partial \max(y) = \operatorname{Conv}\left\{e_i : y_i = \max(y)\right\},$$

where e_i is the i-th element of the Cartesian basis of $\mathbb{R}^m$. This function is convex, thus globally prox-regular and prox-bounded everywhere (with parameters 0). Its proximity operator is given (coordinate-wise) by

$$\left[\operatorname{prox}_{\gamma \max}(y)\right]_i = \begin{cases} s & \text{if } y_i > s \\ y_i & \text{else} \end{cases}$$

where s is the unique real number such that $\sum_{\{i:y_i>s\}}(y_i - s) = \gamma$. □

Example 4.2 (Maximum eigenvalue). Denote the eigenvalue decomposition of a point $y \in \mathbb{S}_m$ as $y = E \operatorname{Diag}(\lambda) E^\top$, where $\lambda \in \mathbb{R}^m$ is a vector with decreasing entries and $E \in \mathbb{R}^{m \times m}$ an orthogonal matrix. The subdifferential of the maximum eigenvalue at y writes (Lewis, 2002, Ex. 3.6)

$$\partial \lambda_{\max}(y) = \{E_{1:r} Z E_{1:r}^\top, Z \in \mathbb{S}_r, Z \succeq 0, \operatorname{trace} Z = 1\}$$

where r is the multiplicity of the maximum eigenvalue of y. This function is convex, thus prox-regular and prox-bounded (with parameters 0). Its proximity operator can be expressed using the one of the max function as

$$\operatorname{prox}_{\gamma \lambda_{\max}}(y) = E \operatorname{Diag}(\operatorname{prox}_{\gamma \max}(\lambda)) E^\top.$$ □

4.2.2 *Structure manifolds: examples*

To highlight the relation between a structure manifold and the corresponding function g, we use the notation $\mathcal{M}^g$ for the structure manifold related to g.

Example 4.3 . The structure manifolds of max are

$$\mathcal{M}_I^{\max} = \{y \in \mathbb{R}^m : y_i = \max(y) \text{ for } i \in I\},$$

where $I \subset \{1, \dots, m\}$. A smooth manifold-defining map for $\mathcal{M}_I^{\max}$ is $h : \mathbb{R}^m \to \mathbb{R}^{|I|-1}$ such that $h(y)_l = y_{i_l} - y_{i_{|I|}}$, where $|I|$ denotes the size of I and i_l the l-th element of I (with some ordering). As required, this map is surjective. At any point $y \in \mathbb{R}^m$, the maximum is partly smooth relative to $\mathcal{M}_I^{\max}$, where $I = \{i : y_i = \max(y)\}$. □

Example 4.4 . The structure manifolds of $\lambda_{\max}$ in $\mathbb{S}_m$ consist of all matrices having a largest eigenvalue with fixed multiplicity r:

$$\mathcal{M}_r^{\lambda_{\max}} = \{y \in \mathbb{S}_m : \lambda_1(y) = \cdots = \lambda_r(y)\}.$$

A manifold-defining map of $\mathcal{M}_r^{\lambda_{\max}}$ is described in Shapiro and Fan (1995) and $\lambda_{\max}$ is partly smooth relative to $\mathcal{M}_r^{\lambda_{\max}}$ at any point $y \in \mathcal{M}_r^{\lambda_{\max}}$. □

In view of the expression of the proximity operators in our examples, their output naturally lie on the structure manifolds described above. More precisely, $\operatorname{prox}_{\gamma \max}(y)$ belongs to the structure manifold $\mathcal{M}_I^{\max}$, where I collects the indices of the k largest entries of y and k grows as γ increases. Similarly, $\operatorname{prox}_{\gamma\lambda_{\max}}(y)$ belongs to the structure manifold $\mathcal{M}_r^{\lambda_{\max}}$, where r increases as γ does. This observation is at the core of the ability of proximal operators to identify neighboring structure manifolds.

4.2.3 *Structure Identification*

The proximity operator identifies structure locally around critical points: all points near a minimizer are mapped to its structure manifold, and this structure is revealed during the computation of the operator.[2]

As discussed in Proposition 2.3.

In the composite setting we consider, the proximity operator of F cannot be explicitly computed. However, $\operatorname{prox}_{\gamma g}$ is available and can provide some structure in the intermediate space $\mathbb{R}^m$ that we would like to exploit. To do so, we introduce two properties (satisfied by two running examples), that will allow us to retrieve the structural information in the intermediate space near points that are not minimizers of g.

NORMAL ASCENT PROPERTY. The first property holds at point $\bar{y} \in \mathcal{M}^g$ if the nonsmooth function g strictly increases on all directions on which it is nonsmooth.

Property 4.1 (Normal ascent). A function g satisfies the *normal ascent* property at point $\bar{y}$ if 0 lies in the relative interior of the projection of $\partial g(\bar{y})$ on the normal space at $\bar{y}$, that is:

$$0 \in \operatorname{ri} \operatorname{proj}_{N_{\bar{y}}\mathcal{M}^g} \partial g(\bar{y}).$$

Notice that this normal ascent property is weaker than the usual property $0 \in \operatorname{ri} \partial F(\bar{y})$ appearing in stability results near nonsmooth critical points, such

2 Computing exactly the *structure* of the output point of the operator, as can be done for the prox, is opposed to merely observing the structure of the output after its computation. This last option is not desirable in our opinion as it entails delicate numerical questions such as testing equality between reals for the maximum, or computing the multiplicity of the maximal eigenvalue of a matrix.

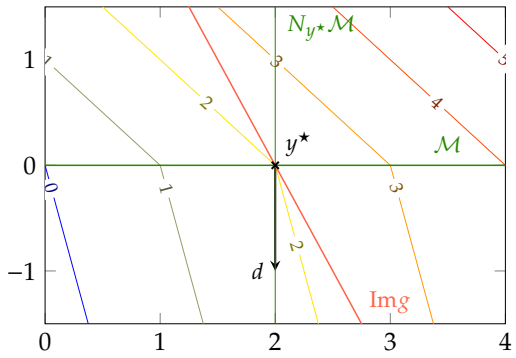


Figure 4.2: Illustration of the level-curves of function g in Example 4.6, along with the image of c and the tangent and normal spaces to $\mathcal{M}^g$ at the minimizer.

as Theorem 3.1 or Proposition 2.3. Indeed, we study here the stability of the proximity operator of g near points that are *not* minimizers of g. This property has a natural interpretation when g is regular and Lipschitz: it requires that the function (strictly) increases in all directions normal to $\mathcal{M}^g$, leaving tangent directions free. We discuss this interpretation precisely in the next remark.

Remark 4.1 (Positive directional derivative). In a "nice" setting where g is Lipschitz continuous and regular at $\bar{y}$, Property 4.1 implies that the directional derivative of g along any normal direction $d \in N_{\bar{y}}\mathcal{M}^g$ is positive. Indeed, in that case one-sided directional derivatives are well-defined (Rockafellar and Wets, 1998, p. 358, Th. 9.16), and the derivative along direction w equals $\max_{v\in\partial g(x)}\langle v, w\rangle$. Along a normal direction $d \in N_y\mathcal{M}^g$, by partial smoothness the directional derivative writes $\max_{v_n\in \mathrm{proj}_{N_y\mathcal{M}^g}(\partial g(x))}\langle v_n, d\rangle$. Property 4.1 ensures the existence of $\alpha > 0$ such that $\alpha d \in \mathrm{proj}_{N_y\mathcal{M}^g}(\partial g(x))$, making the derivative positive. △

Let us briefly discuss that, even if Property 4.1 may look strong, in practice it is not. For a given nonsmooth function F which can be decomposed as $F = g \circ c$, Property 4.1 may not hold for g at $c(x^\star)$ for a minimizer $x^\star$. Nevertheless, the property often holds for a different decomposition $F = \tilde{g} \circ \tilde{c}$. We give two examples where changing the decomposition of F ensures that Property 4.1 holds at minimizers.

Example 4.5 (Normal ascent for regularized-type problem). Consider the minimization of $F(x) = f(x) + r(x)$, where $f(x) = \frac{3}{2}x$ and $r(x) = |x| - \frac{1}{2}x$, whose minimizer is $x^\star = 0$. This writes as a composite problem by setting $c(x) = (f(x), x)$ and $g(y) = y_1 + r(y_2)$. We note first that Property 4.1 does not hold for g at $c(x^\star)$, the structure manifold of g at $c(x^\star)$ being $\mathcal{M}^g = \mathbb{R}\times\{0\}$. However the function also writes $F(x) = \tilde{f}(x) + \tilde{r}(x)$, with $\tilde{f}(x) = x$ and $\tilde{r}(x) = |x|$. Letting similarly $\tilde{c}(x) = (\tilde{f}(x), x)$ and $\tilde{g}(y) = y_1 + \tilde{r}(y_2)$, we now get that Property 4.1 holds for $\tilde{g}$ at $\tilde{c}(x^\star)$. □

Example 4.6 (Normal ascent property for composite problems). Consider minimizing $F = g \circ c$, with

$$g(y) = \begin{cases} y_1 + y_2 & \text{if } y_1 > 0 \\ y_1 + 0.25\, y_2 & \text{else} \end{cases}, \qquad c(x) = \begin{pmatrix} 2 - x \\ 2x \end{pmatrix}.$$

The minimizer is $x^\star = 0$, since g is strictly increasing at all $y \in \mathrm{Im}(c)$ near $y^\star = c(x^\star)$; see Fig. 4.2. However the normal ascent property does not hold at $x^\star$: g is decreasing at $y^\star$ along the normal direction $(0; -1)$.

The composite function boils down to $F(x) = 2 + \max(x, -0.5x) = \tilde{g} \circ \tilde{c}(x)$, where $\tilde{g}(y) = 2 + \max(y)$ and $\tilde{c}(x) = (x, -0.5x)$. With this decomposition, $\tilde{g}$ does satisfy the normal ascent property at $x^\star$. □

CURVE PROPERTY. The second property is more technical and controls the velocity of a curve on the manifold $\mathcal{M}^g$.

Property 4.2 (Curve property). A function g partly smooth at $\bar{y}$ relative to $\mathcal{M}^g$ satisfies the *curve property* at $\bar{y}$ when there exists a neighborhood $\mathcal{N}_{\bar{y}}$ of $\bar{y}$ and $T > 0$ such that any smooth application $e : \mathcal{N}_{\bar{y}} \times [0, T] \to \mathcal{M}^g$ such that $e(y, 0) = \mathrm{proj}_{\mathcal{M}^g}(y)$, $\frac{d}{dt} e(y, t)|_{t=0} = -\operatorname{grad} g(\mathrm{proj}_{\mathcal{M}^g}(y))$ satisfies

$$\| \mathrm{proj}_{N_{e(y,t)}\mathcal{M}^g}(e(y, t) - y)\| \leq \mathrm{dist}_{\mathcal{M}^g}(y) + \tilde{L}\, t^2 \quad \text{for all } y \in \mathcal{N}_{\bar{y}}, t \in [0, T],$$

where $\mathrm{dist}_{\mathcal{M}^g}(y) \triangleq \|y - \mathrm{proj}_{\mathcal{M}^g}(y)\|$ is the distance between $\mathcal{M}^g$ and y and $\operatorname{grad} g(p) \in T_p\mathcal{M}^g$ denotes the Riemannian gradient of g obtained as $\operatorname{grad} g(p) = \mathrm{proj}_{T_p\mathcal{M}^g}(\partial g(p))$.

The idea behind this property is to ensure that the differential of the projection of the (time dependent) normal space is (uniformly) negligible at time 0. Note that for affine spaces, we trivially have $\| \mathrm{proj}_{N_{e(y,t)}\mathcal{M}^g}(y - e(y, t))\| = \mathrm{dist}_{\mathcal{M}^g}(y)$ for all t near 0: the normal spaces are equal at all points of the manifold.

These two properties are satisfied at any structured point for the two non-smooth functions max and $\lambda_{\max}$ of our running examples as detailed in the following lemma. The proofs for the two functions are rather direct but require precise technical descriptions; we defer them to Appendix C.

Lemma 4.3 . *Consider either:*

- $g = \max$, $\bar{y} \in \mathbb{R}^m$, *and the structure manifold* $\mathcal{M}_I^{\max}$ *(of Example 4.3);*
- $g = \lambda_{\max}$, $\bar{y} \in \mathbb{S}_m$, *and the structure manifold* $\mathcal{M}_r^{\lambda_{\max}}$ *(of Example 4.4).*

Then, Properties 4.1 and 4.2 hold at $\bar{y}$.

Finally, the structure provided by $\mathrm{prox}_{\gamma g}$ lies in the intermediate space $\mathbb{R}^m$, while the optimization variable lives in $\mathbb{R}^n$. In order to transfer the structure information to the input space, we will also require the smooth map $c : \mathbb{R}^n \to \mathbb{R}^m$ to be *transversal* to $\mathcal{M}^g \subset \mathbb{R}^m$ at some point $\bar{x} \in \mathbb{R}^n$, which holds when $\mathcal{M}^g$ is a manifold around $c(\bar{x})$ and the following (equivalent) conditions hold:

$$\operatorname{Ker} \mathrm{D}\, c(\bar{x})^* \cap N_{c(\bar{x})}\mathcal{M}^g = \{0\} \quad \text{or} \quad T_{c(\bar{x})}\mathcal{M}^g + \operatorname{Im} \mathrm{D}\, c(\bar{x}) = \mathbb{R}^m. \tag{4.4}$$

In that case, the set $c^{-1}(\mathcal{M}^g)$ is a submanifold of $\mathbb{R}^n$ (Lee, 2003, Th. 6.30), whose normal space has the same dimension as the one of $\mathcal{M}^g$. Furthermore, we have (Lee, 2003, Ex. 6-10)

$$N_x c^{-1}(\mathcal{M}^g) = \mathrm{D}\, c(x)^* N_{c(x)}\mathcal{M}^g \quad \text{and} \quad T_x c^{-1}(\mathcal{M}^g) = \mathrm{D}\, c(x)^{-1} T_{c(x)}\mathcal{M}^g. \tag{4.5}$$

4.3 COLLECTING STRUCTURE WITH THE PROXIMITY OPERATOR

We show in this section how to exactly detect the optimal structure manifold of the composite function $F = g \circ c$ around a point $\bar{x}$ using the proximity operator of g.

In our nonconvex and nonsmooth setting, we seek only structured points which satisfy certain assumptions summarized in our definition of a *qualified point*.

We adapt the notion of qualified point, already present in Chapter 3, to the setting of this chapter.

Definition 4.1 (Qualified points). A point $\bar{x} \in \mathbb{R}^n$ is *qualified* relative to a decomposition (g, c) of F and manifold $\mathcal{M}^g$ if

1. g is prox-bounded and prox-regular at $c(\bar{x})$;
2. g is partly smooth at $c(\bar{x})$ relative to $\mathcal{M}^g$;
3. c is transversal to $\mathcal{M}^g$ at $\bar{x}$;
4. g satisfies Properties 4.1 and 4.2 at point $c(\bar{x})$.

Three of these assumptions constrain only the nonsmooth function g and are easily verifiable in practice. Only the transversality condition limits the range of acceptable smooth mappings; see e.g., Lewis (2002, Sec. 4). For such *qualified* points, we get two useful properties: first, F is partly smooth at $\bar{x}$ relative to the manifold $\mathcal{M}$, locally defined as $\mathcal{M} \triangleq c^{-1}(\mathcal{M}^g) \ni \bar{x}$ by the chain rule of Lewis (2002, Th. 4.2), and second, the operator $\operatorname{prox}_{\gamma g}$ is single-valued, locally Lipschitz, and defined by its optimality condition near $c(\bar{x})$.

4.3.1 *Main result:* $\operatorname{prox}_{\gamma g} \circ c$ *as a structure detector*

We show in the following theorem that if x is near a qualified point of F with structure $\mathcal{M}$, then $\operatorname{prox}_{\gamma g}(c(x))$ will output a point on $\mathcal{M}^g = c(\mathcal{M})$, the structure manifold of g corresponding to $\mathcal{M}$ (in the intermediate space). Our theorem provides precise conditions on x and γ that guarantee this structure identification and forms the main theoretical contribution of the chapter. We illustrate this behavior in Figures 4.3 and 4.4.

The position of this result with respect to the literature is discussed right after in Remark 4.2, and the proof is given in the following Section 4.3.2, in a succession of technical lemmas. We stress that we give guarantees on the *structure* to which the point $\operatorname{prox}_{\gamma g}(c(x))$ belongs, rather than on the point itself.

Theorem 4.4 (Prox for structure detection). *Consider a function $F = g \circ c$ and a point $\bar{x}$. Assume that $\bar{x}$ is* qualified *relative to a manifold $\mathcal{M}^g \subset \mathbb{R}^m$. Then, there exists a neighborhood $\mathcal{N}_{\bar{x}}$ of $\bar{x}$ and a constant Γ such that, for all $x \in \mathcal{N}_{\bar{x}}$,*

$$\operatorname{prox}_{\gamma g}(c(x)) \in \mathcal{M}^g \text{ for all } \gamma \in [\varphi(\operatorname{dist}_{\mathcal{M}}(x)), \Gamma],$$

where $\operatorname{dist}_{\mathcal{M}}(x)$ denotes the distance from x to the manifold $\mathcal{M}$ and φ is defined as

$$\varphi(t) = \frac{c_{ri}}{2\tilde{L}}\left(1 - \sqrt{1 - \frac{4\tilde{L}c_{map}t}{c_{ri}^2}}\right) = \frac{c_{map}}{c_{ri}}t + \frac{\tilde{L}c_{map}^2}{c_{ri}^3}t^2 + o(t^2),$$

with c_{ri}, c_{map}, and $\tilde{L}$ (of Property 4.2) positive constants.

In particular, there exists $L > 0$, $\varepsilon > 0$ such that

$$\|x - x^\star\| \leq \varepsilon \text{ and } L\|x - x^\star\| \leq \gamma \leq \Gamma \implies \operatorname{prox}_{\gamma g}(c(x)) \in \mathcal{M}^g.$$

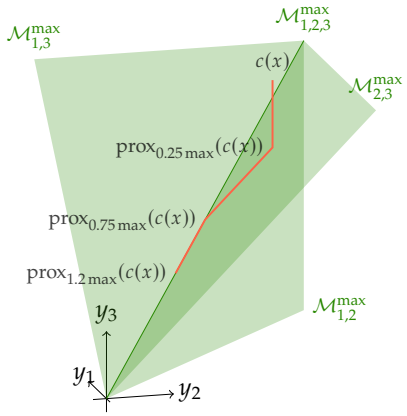


Figure 4.3: Illustration of the main result in the intermediate space, on the function of Fig. 4.4. The structure manifolds of $\max : \mathbb{R}^3 \to \mathbb{R}$ are displayed as the three half-planes and the line in green. The red line illustrates the curve $\gamma \mapsto \text{prox}_{\gamma \max}(c(x))$. When $\gamma < 0.25$, the curve does not lie on any structure manifold. For $\gamma \in [0.25, 0.75)$, the curve lies on the optimal manifold $\mathcal{M}^{\max}_{2,3}$. For $\gamma \geq 0.75$, the curve lies on $\mathcal{M}^{\max}_{1,2,3}$.

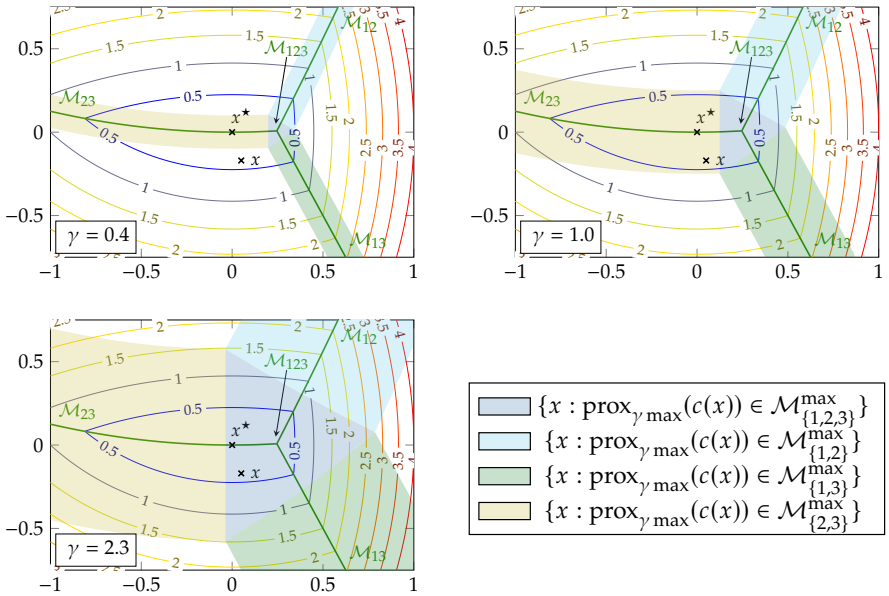


Figure 4.4: Illustration of the main result on a maximum of three quadratic functions, with $\bar{x} \in \mathcal{M}^{\max}_{\{1,2\}}$ and a point $\tilde{x}$ near $\bar{x}$. The three figures show the areas where $\text{prox}_{\gamma g} \circ c$ detects manifolds for three stepsizes: $\gamma = 0.4$ (upper left), $\gamma = 1$ (upper right) and $\gamma = 2.3$ (lower left). We see on the upper left fig. that $\text{prox}_{\gamma g} \circ c$ detects no structure from $\bar{x}$ because γ is too small, and in contrast, on the lower fig., that it wrongly detects too much structure ($\mathcal{M}^{\max}_{\{1,2,3\}}$) because γ is too large. On the upper right fig., the optimal manifold is detected with γ chosen in the right interval.

Note that Property 4.2 is only used to compute explicitly an interval of γ guaranteed to provide the correct structure; the existence of that interval holds independently.

Remark 4.2 (Relation with existing results). The difference between Theorem 4.4 and existing results lies in two aspects. First, the identification properties of the proximal operator (Daniilidis et al., 2006, Th. 28, recalled as Proposition 2.3), the proximal-gradient operator Theorem 3.1, or even approximate prox-gradient operators (Lee, 2023) give structure information directly in the input space (even in abstract algorithmic frameworks (Hare and Lewis, 2004, Th. 4) or (Lewis and Zhang, 2013, Th. 4.10)). In the composite case, the proximity operator reveals structure in the intermediate space only, and extra work is required to bring it back to the input space.

Second, most existing results investigate identification properties near minimizers, and not just arbitrary points (two notable exceptions give results near arbitrary structured points: (Lewis and Zhang, 2013) for an abstract algorithmic framework, and Theorem 3.1 for the proximal gradient). Here, we evaluate $\mathrm{prox}_{\gamma g}$ near $c(\bar{x})$, a point without any specific properties (even if $\bar{x}$ is a local minimizer). This is why we need Property 4.1 to guarantee identification in the intermediate space, and bring the structure information to the input space. △

Remark 4.3 (About prox-linear methods). Prox-linear methods are known to identify structure on composite problems (Lewis and Wright, 2016). Specifically, Lewis and Wright (2016, Th. 4.11) establishes that, after some finite time, an intermediate quantity belongs to the structure manifold of $c(x^\star)$. It is then mentioned that this information could be used to take efficient second-order steps to minimize F along the identified manifold. Whether this can be done generically is unclear to us: checking that this quantity, obtained from the subproblem solution, belongs to a structure manifold is delicate. Though it is reasonable if the subproblem is solved with a suitable active-set method, it becomes delicate if only an approximation of the subproblem solution is available, using e.g., interior point methods. In that case, the quantity will be somewhat close to the structure manifold, and one would have to resort to ε-based tests. △

Remark 4.4 (Theorem 4.4 as a structure identification tool). In contrast with the identification of prox-linear methods, Theorem 4.4 provides a simple result for the detection of structure manifolds near any point $x \in \mathbb{R}^n$. We also underline that the bounds on the range of γ that provide correct identification are surprisingly simple: the upper bound is constant and the lower bound is essentially a linear function of the distance to the manifold. These simple and explicit bounds allow us to build a simple algorithm in the forthcoming Section 4.4. △

4.3.2 *Proof of Theorem 4.4*

The main difficulty of the proof is to build a suitable identification result for the nonsmooth function g. Theorem 4.4 (identification for $g \circ c$) would then follow by taking into account the action of the smooth map c.

To derive an identification result on g, we have to give conditions on y and γ so that $p = \mathrm{prox}_{\gamma g}(y)$ lies on the considered manifold $\mathcal{M}^g$. Since g is prox-regular

and prox-bounded at point $c(\bar{x})$, Theorem 2.1 allows us to characterize this relation by its first-order optimality condition:

$$y \in p + \gamma \partial g(p).$$

Whenever $p \in \mathcal{M}^g$ (which is what we want to show), this inclusion decomposes along $T_p\mathcal{M}^g$ and $N_p\mathcal{M}^g$ as:

$$\operatorname{proj}_{T_p\mathcal{M}^g}(y-p) = \gamma \operatorname{grad} g(p) \tag{4.6}$$

$$\operatorname{proj}_{N_p\mathcal{M}^g}(y-p) \in \gamma \operatorname{proj}_{N_p\mathcal{M}^g} \partial g(p). \tag{4.7}$$

Thus we will show that for suitable (y, γ), there is a unique p that satisfies these two equations. We do so by considering the smooth tangent component Eq. (4.6) first and then the nonsmooth normal component Eq. (4.7) as follows:

- We first show in Lemma 4.5 that for y near $\bar{y}$ and γ small, there exists a unique point $p = e(y, \gamma)$ on $\mathcal{M}^g$ that satisfies Eq. (4.6), which depends smoothly on γ and y. This result is obtained by applying the implicit function theorem.

- Then, we prove in Lemma 4.6 that $e(y, \gamma)$ also satisfies the second inclusion Eq. (4.7) if γ belongs to the interval $[\varphi^g(\operatorname{dist}_{\mathcal{M}^g}(y)), \Gamma^g]$. This result is a consequence of the application of some variational analysis tools.

Putting these two results together, we obtain the existence and uniqueness of a point $p = e(y, \gamma) \in \mathcal{M}^g$ verifying both Eq. (4.6) and Eq. (4.7) for all y near $\bar{y}$ and $\gamma \in [\varphi^g(\operatorname{dist}_{\mathcal{M}^g}(y)), \Gamma^g]$. By the first-order optimality condition presented above, this point is necessarily $\operatorname{prox}_{\gamma g}(y)$.

Finally, this identification result in the intermediate space on g is transferred back to the input space using transversality.

Part 1: tangent optimality

We first show that, for y near $\bar{y}$ and γ small, there is a unique point p *on the manifold* $\mathcal{M}^g$ that satisfies the tangent component of this optimality condition:

$$\operatorname{proj}_{T_p\mathcal{M}^g}(y-p) = \gamma \operatorname{grad} g(p), \tag{4.8}$$

where $\operatorname{grad} g(p) \triangleq \operatorname{proj}_{T_p\mathcal{M}^g} \partial(g(p))$ is unique by the sharpness property of partial smoothness, and matches the Riemannian gradient of g on $\mathcal{M}^g$ (see Section 2.3). Such points p are given by a smooth manifold-valued application $e(y, \gamma)$, the existence of which is guaranteed by the following lemma.

Lemma 4.5 . *Consider a function $g : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$, a point $\bar{y} \in \mathbb{R}^m$, and a manifold $\mathcal{M}^g$ with g partly smooth at $\bar{y}$ relative to $\mathcal{M}^g$. Then, there exists a smooth curve $e : \mathcal{N}_{\bar{y}} \times \mathcal{N}_0 \to \mathcal{M}$ defined on a neighborhood of $(\bar{y}, 0)$ in $\mathbb{R}^m \times \mathbb{R}_+$ such that*

- *for all $y \in \mathcal{N}_{\bar{y}}$, $e(y, 0) = \operatorname{proj}_{\mathcal{M}^g}(y)$ and $\frac{d}{d\gamma} e(y, \gamma)|_{\gamma=0} = -\operatorname{grad} g(\operatorname{proj}_{\mathcal{M}^g}(y))$;*

- *for all $y \in \mathcal{N}_{\bar{y}}$, $\gamma \in \mathcal{N}_0$, Eq. (4.8) is satisfied for $p = e(y, \gamma)$.*

Proof. We define the mapping $\Phi : \mathbb{R}^m \times \mathbb{R} \times \mathcal{M}^g \to \cup_{x \in \mathcal{M}^g} T_x\mathcal{M}^g$ as

$$\Phi(y, \gamma, p) = \gamma \operatorname{grad} g(p) - \operatorname{proj}_{T_p\mathcal{M}^g}(y-p)$$

and consider the equation $\Phi(y, \gamma, p) = 0$ near the point $(\bar{y}, 0, \bar{y})$. Using the smoothness of g on $\mathcal{M}^g$ given by partial smoothness, we have that this mapping is continuously differentiable on a neighborhood of $(\bar{y}, 0, \bar{y})$. We see that its differential with respect to p is $\mathrm{D}_p\, \Phi(\bar{y}, 0, \bar{y}) = I$. Indeed, for $\eta \in T_p \mathcal{M}^g$,

$$\mathrm{D}_p\, \Phi(y, \gamma, p)[\eta] = \gamma\, \mathrm{Hess}\, g(p)[\eta] + \eta - \mathrm{D}_{p'} \left(p' \mapsto \mathrm{proj}_{T_{p'} \mathcal{M}^g}(y - p) \right)(p)[\eta].$$

At point $(\bar{y}, 0, \bar{y})$, the first term vanishes, and the third term writes

$$\mathrm{D}_{p'} \left(p' \mapsto \mathrm{proj}_{T_{p'} \mathcal{M}^g}(0) \right)(\bar{y})[\eta]$$

and vanishes as well as the differential of the null function $p' \mapsto \mathrm{proj}_{T_{p'} \mathcal{M}^g}(0)$. Thus $\mathrm{D}_p\, \Phi(\bar{y}, 0, \bar{y}) = I$ is invertible. The implicit functions theorem thus grants the existence of neighborhoods $\mathcal{N}^1_{\bar{y}}$, $\mathcal{N}^2_0$, $\mathcal{N}^3_{\bar{y}}$ of $\bar{y}$, 0, $\bar{y}$ in $\mathbb{R}^m$, $\mathbb{R}$, $\mathcal{M}^g$ and a continuously differentiable function $c : \mathcal{N}^1_{\bar{y}} \times \mathcal{N}^2_0 \to \mathcal{N}^3_{\bar{y}}$ such that, for any $(y, \gamma) \in \mathcal{N}^1_{\bar{y}} \times \mathcal{N}^2_0$, Equation (4.8) is satisfied with $p = e(y, \gamma)$. For $y \in \mathcal{N}^1_{\bar{y}}$, $e(y, 0)$ satisfies $y - e(y, 0) \in N_{e(y,0)} \mathcal{M}^g$, which is the first-order optimality condition of $e(y, 0) = \mathrm{proj}_{\mathcal{M}^g}(y)$. Possibly reducing $\mathcal{N}^1_{\bar{y}}$ so that, for all $y \in \mathcal{N}^1_{\bar{y}}$ $\mathrm{proj}_{\mathcal{M}^g}(y)$ is well-defined and unique, the previous optimality condition is equivalent to $e(y, 0) = \mathrm{proj}_{\mathcal{M}^g}(y)$. Besides, differentiating $\Phi(y, \gamma, e(y, \gamma)) = 0$ relative to γ at $\gamma = 0$ yields

$$\begin{aligned} \mathrm{D}_\gamma\, e(y, 0) &= -[\mathrm{D}_p\, \Phi(y, 0, \mathrm{proj}_{\mathcal{M}^g}(y))]^{-1}\, \mathrm{D}_\gamma\, \Phi(y, 0, \mathrm{proj}_{\mathcal{M}^g}(y)) \\ &= -\,\mathrm{grad}\, g(\mathrm{proj}_{\mathcal{M}^g}(y)), \end{aligned}$$

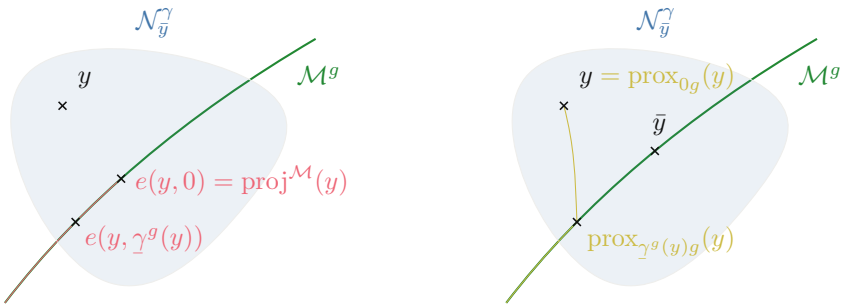
which concludes the proof. □

Part 2: normal optimality

The previous lemma shows that for every (y, γ) one can find a point $e(y, \gamma)$ *on the manifold* $\mathcal{M}^g$ that solves the tangent part of the optimality condition (4.8). The next lemma determines the values of y and γ for which the whole optimality condition

$$y \in e(y, \gamma) + \gamma\, \mathrm{ri}\, \partial g(e(y, \gamma)) \tag{4.9}$$

holds, as illustrated in Figure 4.5a.



(a) the curve $\gamma \mapsto e(y, \gamma)$ on $\mathcal{M}$.

(b) the curve $\gamma \mapsto \mathrm{prox}_{\gamma g}(y)$ on $\mathcal{M}$ for $\gamma \geq \varphi(\mathrm{dist}_{\mathcal{M}^g}(y))$.

Figure 4.5: Illustration of Lemma 4.6 and its consequences.

Lemma 4.6 . *Consider a function g, a point $\bar{y} \in \mathbb{R}^m$ and a manifold $\mathcal{M}^g$ such that g is partly smooth at $\bar{y}$ relative to $\mathcal{M}^g$ and that g satisfies Property 4.1 at $\bar{y}$. Let e denote a smooth $\mathcal{M}$-valued application defined on a neighborhood of $(\bar{y}, 0)$ provided by Lemma 4.5. Then, there exists $C > 0$ such that:*

1. *for all $\gamma \in [0, C]$, $e(\bar{y}, \gamma)$ verifies (4.9) with $y = \bar{y}$,*

2. *for all $\gamma \in [0, C]$, there exists a neighborhood $\mathcal{N}_{\bar{y}}^{\gamma}$ of $\bar{y}$ such that, for all $y \in \mathcal{N}_{\bar{y}}^{\gamma}$, $e(y, \gamma)$ verifies (4.9),*

Further assume that g satisfies Property 4.2 at $\bar{y}$ with constant $\tilde{L}$, then

3. *there exist $\Gamma^g > 0$ and a neighborhood $\mathcal{N}_{\bar{y}}$ of $\bar{y}$ such that for all $y \in \mathcal{N}_{\bar{y}}$*

$$e(y, \gamma) \text{ verifies (4.9) for all } \gamma \in [\varphi^g(\operatorname{dist}_{\mathcal{M}^g}(y)), \Gamma^g],$$

where $c_{ri} \geq 0$ and $\varphi^g(t) = \frac{c_{ri}}{2\tilde{L}}\left(1 - \sqrt{1 - \frac{4\tilde{L}t}{c_{ri}^2}}\right) = \frac{1}{c_{ri}}t + \frac{\tilde{L}}{c_{ri}^3}t^2 + o(t^2)$.

The proof consists in finding the points y, γ such that $0 \in \operatorname{ri} \Psi(y, \gamma)$, where the mapping $\Psi : \mathbb{R}^m \times \mathbb{R} \to \cup_{x \in \mathcal{M}^g} N_x \mathcal{M}^g$ is defined as

$$\Psi(y, \gamma) = \operatorname{proj}_{N_{e(y,\gamma)}\mathcal{M}^g}\left(\frac{1}{\gamma}\left(e(y, \gamma) - y\right) + \partial g\left(e(y, \gamma)\right)\right).$$

Items *i)* and *ii)* are shown by extending the property $0 \in \Psi(\bar{y}, 0)$ to a neighborhood of $(\bar{y}, 0)$, using the inner-semicontinuity properties of Ψ. We then derive explicit bounds on the interval of steps such that $0 \in \operatorname{ri} \Psi(y, \gamma)$: for a fixed $y \in \mathcal{N}_{\bar{y}}$, when γ decreases past some value, say $\underline{\gamma}(y)$, the condition $0 \in \operatorname{ri} \Psi(y, \gamma)$ no longer holds. Precisely at $\underline{\gamma}(y)$, 0 lies on the (relative) boundary of $\Psi(y, \underline{\gamma}(y))$: denoting $\operatorname{rbd} S \triangleq S \setminus \operatorname{ri} S$ the relative boundary of set S,

$$0 \in \operatorname{rbd} \operatorname{proj}_{N_{e(y,\underline{\gamma}(y))}\mathcal{M}^g}\left(\frac{1}{\underline{\gamma}(y)}\left(e(y, \underline{\gamma}(y)) - y\right) + \partial g\left(e(y, \underline{\gamma}(y))\right)\right).$$

Denoting $\partial^N g(p) \triangleq \operatorname{proj}_{N_p \mathcal{M}^g}(\partial g(p))$ the projection of the subdifferential on the normal space of its structure manifold and taking norms yields:

$$\begin{aligned}\|\operatorname{proj}_{N_{e(y,\underline{\gamma}(y))}\mathcal{M}^g}(y - e(y, \underline{\gamma}(y)))\| &\geq \underline{\gamma}(y) \inf_{v_n \in \operatorname{rbd} \partial^N g(e(y,\underline{\gamma}(y)))} \|v_n\| \\ &\geq \underline{\gamma}(y) \underbrace{\inf_{p \in \mathcal{N}_{\bar{y}}} \inf_{v_n \in \operatorname{rbd} \partial^N g(p)} \|v_n\|}_{\triangleq c_{\mathrm{ri}}}.\end{aligned}$$

By partial smoothness, ∂g is continuous on $\mathcal{M}^g$ at $\bar{y}$, and thus in particular inner-semicontinuous. The inclusion $0 \in \operatorname{ri} \operatorname{proj}_{N_{\bar{y}}\mathcal{M}^g} \partial g(\bar{y})$ therefore holds on a neighborhood of $\bar{y}$ on $\mathcal{M}^g$ (Daniilidis et al., 2006, Lemma 20), thus making the constant c_{ri} positive. We note that this kind of quantity also appears as the *modulus of identifiability*, proposed recently by Lewis and Tian (2022, Def. 2.3), where it has the same property: its positivity enables the identification of the associated structure manifold.

Using Property 4.2, the left-hand side is upper bounded by a simpler expression:

$$\tilde{L}\underline{\gamma}(y)^2 + \operatorname{dist}_{\mathcal{M}^g}(y) \geq c_{\mathrm{ri}}\underline{\gamma}(y), \quad \text{that is} \quad \underline{\gamma}(y) \leq \frac{c_{\mathrm{ri}}}{2\tilde{L}}\left(1 - \sqrt{1 - \frac{4\tilde{L}\operatorname{dist}_{\mathcal{M}^g}(y)}{c_{\mathrm{ri}}^2}}\right),$$

which provides the expression for φ^g used in the lemma.

Proof. Item i) We first consider $\Psi_{\bar{y}}(\cdot) = \Psi(\bar{y}, \cdot)$. Since $\bar{y} \in \mathcal{M}^g$, Lemma 4.5 tells us that $e(\bar{y}, \gamma) = \bar{y} - \gamma \operatorname{grad} g(\bar{y}) + o(\gamma)$, and thus

$$\Psi_{\bar{y}}(0) = \operatorname{proj}_{N_{\bar{y}}\mathcal{M}^g}\left(-\operatorname{grad} g(\bar{y}) + \partial g(\bar{y})\right) = \operatorname{proj}_{N_{\bar{y}}\mathcal{M}^g}(\partial g(\bar{y}))$$

where we used that $\operatorname{grad} g(\bar{y}) \in T_{\bar{y}}\mathcal{M}^g$ is orthogonal to $N_{\bar{y}}\mathcal{M}^g$.

Property 4.1 provides that $0 \in \operatorname{ri} \Psi_{\bar{y}}(0)$. We now turn to showing that there exists C' such that, for all $\gamma \in [0, C']$, $0 \in \operatorname{ri} \Psi_{\bar{y}}(\gamma)$.

By contradiction, assume there exist a sequence $\gamma_k \to 0$ such that $0 \notin \operatorname{ri} \Psi_{\bar{y}}(\gamma_k)$. This means that there exists a sequence of unit norm vectors (s_k) such that for all k,

$$\langle s_k, z\rangle \leq 0 \text{ for all } z \in \Psi_{\bar{y}}(\gamma_k). \tag{4.10}$$

As a bounded sequence, s_k admits at least one limit point, say $\bar{s}$. Take $\bar{z} \in \Psi_{\bar{y}}(0)$. The continuity of ∂g (by partial smoothness, item iv), of $\gamma \mapsto (e(\bar{y}, \gamma) - \bar{y})/\gamma$ (by smoothness of e), and of $\gamma \mapsto \operatorname{proj}_{N_{e(\bar{y},\gamma)}\mathcal{M}^g}$ (by smoothness of $\mathcal{M}^g$) yield the continuity of $\Psi_{\bar{y}}$ as a set-valued map. This mapping is thus inner-semicontinuous (Rockafellar and Wets, 1998, Def. 5.4), so there exists a sequence $z_k \in \Psi_{\bar{y}}(\gamma_k)$ such that z_k converges to $\bar{z}$. Taking the correct subsequence and renaming iterates, we can write $s_k \to \bar{s}$ and $z_k \to \bar{z}$. Equation (4.10) provides $\langle s_k, z_k\rangle \leq 0$ for all k, which gives at the limit $\langle \bar{s}, \bar{z}\rangle \leq 0$. This actually holds for all $\bar{z} \in \Psi_{\bar{y}}(0)$: $\bar{s}$ separates 0 and $\Psi(0)$, which contradicts $0 \in \operatorname{ri} \Psi_{\bar{y}}(0)$.

Finally, let us take the constant C such that $[0, C]$ is included in $[0, C']$ and the neighborhood of 0 provided by Lemma 4.5. Then, for any $\gamma \in [0, C]$, adding the two orthogonal inclusions $0 \in \operatorname{ri} \Psi_{\bar{y}}(\gamma)$ and $0 = \Phi(y, \gamma, c(y, \gamma))$, we obtain that $e(\bar{y}, \gamma)$ verifies (4.9) with $y = \bar{y}$.

Item ii) Let $\gamma \in [0, C]$. We turn to show the existence of a neighborhood $\mathcal{N}_{\bar{y}}^{\gamma}$ of $\bar{y}$ such that, for all $y \in \mathcal{N}_{\bar{y}}^{\gamma}$, $e(y, \gamma)$ verifies (4.9). By contradiction, assume that there exists a sequence (y_k) that converges to $\bar{y}$ such that (4.9) fails for (y_k, γ). Since the tangent component of (4.9) does hold, necessarily $0 \notin \operatorname{ri} \Psi(y_k, \gamma)$. However, the mapping $y \mapsto \Psi(y, \gamma)$ is inner-semicontinuous (from the same arguments as in the proof of *item i)* and there holds $0 \in \operatorname{ri} \Psi(\bar{y}, \gamma)$. A reasoning similar to that of *item i)* reveals the contradiction.

Item iii) Define $\mathcal{N}_{\bar{y}}$ a neighborhood of $\bar{y}$ and Γ^g a positive constant such that Property 4.2 applies over $\mathcal{N}_{\bar{y}} \times [0, \Gamma^g]$, $\mathcal{N}_{\bar{y}}$ is contained in $\cup_{\gamma \in [0,C]} \mathcal{N}_{\bar{y}}^{\gamma} \cap \mathcal{N}_{\bar{y}}^{C}$ and $0 \in \operatorname{ri} \Psi(y, \gamma)$ holds for all $(y, \gamma) \in \mathcal{N}_{\bar{y}} \times [0, \Gamma^g]$. The second condition can be met on a nontrivial neighborhood of $(\bar{y}, 0)$: it holds at that point, and Ψ is inner-semicontinuous ($e(y, \gamma)$ lies on $\mathcal{M}^g$ and ∂g is inner-semicontinuous by partial smoothness of g).

Let $y \in \mathbb{N}_{\bar{y}}$ and $\gamma \in [\varphi^g(\operatorname{dist}_{\mathcal{M}^g}(y)), \Gamma^g]$. We show that $0 \in \operatorname{ri} \Psi(y, \gamma)$, that is

$$\operatorname{proj}_{N_{e(y,\gamma)}\mathcal{M}^g}(y - e(y, \gamma)) \in \gamma \operatorname{ri} \partial^N g\left(e(y, \gamma)\right).$$

Combining this with the orthogonal inclusion $0 = \Phi(y, \gamma, e(y, \gamma))$ yields the claim.

The inequality $\varphi^g(\mathrm{dist}_{\mathcal{M}^g}(y)) \leq \gamma$ implies $\tilde{L}\gamma^2 + \mathrm{dist}_{\mathcal{M}}(y) \leq \gamma c_{\mathrm{ri}}$. We have successively by definition of $\mathbb{N}_{\bar{y}}$ and the above bound that

$$\begin{aligned}\|\operatorname{proj}_{N_{e(y,\gamma)}\mathcal{M}^g}(y - e(y,\gamma))\| \leq \mathrm{dist}_{\mathcal{M}}(y) + \tilde{L}\gamma^2 &\leq \gamma c_{\mathrm{ri}} \\ &\leq \gamma \inf\{\|n\|, n \in \operatorname{rbd} \partial^N g(e(y,\gamma))\}.\end{aligned}$$

This means that $\operatorname{proj}_{N_{e(y,\gamma)}\mathcal{M}^g}(y - e(y,\gamma))$ belongs to the ball of center 0 and radius $\gamma \inf\{\|n\|, n \in \operatorname{rbd} \partial^N(g(e(y,\gamma)))\}$ in $N_{e(y,\gamma)}\mathcal{M}^g$. Besides, this ball is included in $\gamma\partial^N(g(e(y,\gamma))$ since $0 \in \partial^N g(e(y,\gamma)$ by definition of $\mathbb{N}_{\bar{y}}$. Therefore, $0 \in \operatorname{ri} \Psi(y,\gamma)$ for all $y \in \mathbb{N}_{\bar{y}}$ and $\gamma \in [\varphi^g(\mathrm{dist}_{\mathcal{M}^g}(y)), \Gamma^g]$. □

Part 3: From the intermediate space to the input space

To conclude the proof of Theorem 4.4, we will first identify the curve $e(y,\gamma)$ to $\operatorname{prox}_{\gamma g}(y)$ and thus prove that it belongs to the sought manifold, as illustrated in Fig. 4.5b. Then, this intermediate identification result is brought back to the input space using transversality.

Proof (of Theorem 4.4). The standing assumptions allow to call Lemma 4.6 at point $c(\bar{x})$ with manifold $\mathcal{M}^g$. This yields the neighborhood $\mathcal{N}_{c(\bar{x})}$, constants Γ^g and C, a function φ^g, and a smooth mapping $e : \mathcal{N}_{c(\bar{x})} \times [0, C] \to \mathcal{M}^g$ such that, for $y \in \mathcal{N}_{c(\bar{x})}$ and $\gamma \in [\varphi^g(\mathrm{dist}_{\mathcal{M}^g}(y)), \Gamma^g]$, $e(y,\gamma)$ verifies the optimality condition (4.9) of $e(y,\gamma) = \operatorname{prox}_{\gamma g}(y)$. Besides, since g is prox-regular and prox-bounded at point $c(\bar{x})$, these properties also hold on a neighborhood of that point. Under these conditions, Theorem 2.1 allows to recover the equality $e(y,\gamma) = \operatorname{prox}_{\gamma g}(y)$. Take $\mathcal{N}_{\bar{x}} = c^{-1}(\mathcal{N}_{c(\bar{x})})$, a neighborhood of $\bar{x}$ as the preimage of a neighborhood of $c(\bar{x})$ by the continuous c. For all $x \in \mathcal{N}_{\bar{x}}$,

$$\operatorname{prox}_{\gamma g}(c(x)) \in \mathcal{M}^g \text{ for all } \gamma \in [\varphi^g(\mathrm{dist}_{\mathcal{M}^g}(c(x))), \Gamma^g].$$

We turn to show that, for some constant $c_{\mathrm{map}} > 0$, there holds $\mathrm{dist}_{\mathcal{M}^g}(c(x)) \leq c_{\mathrm{map}} \mathrm{dist}_{\mathcal{M}}(x)$ for all $x \in \mathcal{N}_{\bar{x}}$. Let $x \in \mathcal{N}_{\bar{x}}$ and $x^{\mathcal{M}} = \operatorname{proj}_{\mathcal{M}}(x)$, so that $\mathrm{dist}_{\mathcal{M}}(x) = \|x^{\mathcal{M}} - x\|$. Using successively that $c(x^{\mathcal{M}}) \in \mathcal{M}^g$ and smoothness of c, there holds for x near $\bar{x}$

$$\begin{aligned}\mathrm{dist}_{\mathcal{M}^g}(c(x)) &\leq \|c(x) - c(x^{\mathcal{M}})\| \\ &\leq \|\operatorname{Jac}_c(x^{\mathcal{M}}) \cdot (x - x^{\mathcal{M}})\| + \mathcal{O}(\|x - x^{\mathcal{M}}\|^2) \\ &\leq \left(\sup_{v_n \in N_{x^{\mathcal{M}}}\mathcal{M}, \|v_n\|=1} \|\operatorname{Jac}_c(x^{\mathcal{M}}) \cdot v_n\| \right) \|x - x^{\mathcal{M}}\| + \mathcal{O}(\|x - x^{\mathcal{M}}\|^2) \\ &\leq \underbrace{\left(\sup_{x' \in \mathcal{N}_{\bar{x}}} \sup_{v_n \in N_{x'}\mathcal{M}, \|v_n\|=1} \|\operatorname{Jac}_c(x') \cdot v_n\| \right)}_{C''} \|x - x^{\mathcal{M}}\| + \mathcal{O}(\|x - x^{\mathcal{M}}\|^2).\end{aligned}$$

We show by contradiction that the constant C'' is positive. If $C'' = 0$, there exists $v_n \in N_{\bar{x}}\mathcal{M}$ of unit norm such that $\mathrm{D}\, c(\bar{x}) v_n = 0$. By Eq. (4.5), we have $v_n = \mathrm{D}\, c(\bar{x})^* \hat{v}_n$ for some $\hat{v}_n \in N_{c(\bar{x})}\mathcal{M}^g$, so that $\mathrm{D}\, c(\bar{x})\, \mathrm{D}\, c(\bar{x})^* d = 0$. Pre-multiplying by $\hat{v}_n^*$ yields $\|\mathrm{D}\, c(\bar{x})^* \hat{v}_n\|^2 = 0$: there holds $\hat{v}_n \in \operatorname{Ker}(\mathrm{D}\, c(\bar{x})^*) \cap N_{c(\bar{x})}\mathcal{M}^g$. The

transversality condition Eq. (4.4) implies $\hat{v}_n = 0$, and in turn $v_n = 0$, which contradicts the fact that this vector has unit length.

Therefore, for all $x \in \mathcal{N}_{\bar{x}}$ and a constant $c_{\text{map}} > C''$, there holds $\operatorname{dist}_{\mathcal{M}^g}(c(x)) \leq c_{\text{map}} \operatorname{dist}_{\mathcal{M}}(x)$. Monotony of φ^g implies that $\varphi^g(\operatorname{dist}_{\mathcal{M}^g}(c(x))) \leq \varphi^g(c_{\text{map}} \operatorname{dist}_{\mathcal{M}}(x))$, which yields the claimed bounds with

$$\varphi(t) = \frac{c_{\text{ri}}}{2\tilde{L}}\left(1 - \sqrt{1 - \frac{4\tilde{L}c_{\text{map}}t}{c_{\text{ri}}^2}}\right) \quad \text{and} \quad \Gamma = \Gamma^g.$$

Finally, we show the existence of positive constants ε, L such that

$$\|x - \bar{x}\| \leq \varepsilon \text{ and } L\|x - \bar{x}\| \leq \gamma \leq \Gamma \Longrightarrow \operatorname{prox}_{\gamma g}(c(x)) \in \mathcal{M}^g.$$

Since $\bar{x} \in \mathcal{M}$, $\operatorname{dist}_{\mathcal{M}}(\cdot) \leq \|\cdot - \bar{x}\|$. By monotony and smoothness of φ, there exists $L > 0$ such that $\varphi(\operatorname{dist}_{\mathcal{M}^\star}(\cdot)) \leq L\|\cdot - x^\star\|$ over $\mathcal{B}(x^\star, \varepsilon)$. Reducing ε if necessary so that $L\varepsilon < \Gamma$ yields the result. □

4.4 A LOCAL NEWTON ALGORITHM FOR NONSMOOTH COMPOSITE MINIMIZATION

In this section, we use the results of Section 4.3 to propose an optimization method that locally identifies the structure of a minimizer and converges quadratically to this point.

Recall the basic idea: if the optimal manifold $\mathcal{M}^\star$ corresponding to a minimizer $x^\star$ is known, the *nonsmooth* optimization problem turns into a *smooth constrained* optimization problem. In turn, this problem can be solved using algorithms from smooth constrained optimization such as Sequential Quadratic Programming.

Using this idea and the structure identification mechanism developed in the previous section, we propose a method which: i) uses the proximity operator of g to gather structure in the intermediate space, ii) brings back this structure to the input space, and iii) optimizes smoothly along the identified manifold. The resulting algorithm is precisely described in Section 4.4.1 and then analyzed in Section 4.4.2.

4.4.1 *Description of the algorithm*

We proceed to describe the three steps exposed above. The full algorithm is depicted in Algorithm 4.1.

Gathering structure. We showed in Theorem 4.4 that near a qualified point in $\mathbb{R}^n$, the operator $\operatorname{prox}_{\gamma g}(c(\cdot))$ provides the optimal structure $\mathcal{M}^{g\star}$ (in the intermediate space $\mathbb{R}^m$) for an explicit range of steps. We thus define from the current iterate $x_k \in \mathbb{R}^n$ and stepsize γ_k the working manifold $\mathcal{M}_k^g$ (in the intermediate space) as the structure of $\operatorname{prox}_{\gamma_k g}(c(x_k))$. One technical point is to guarantee that, after some time, $\gamma_k \in [L\|x_k - x^\star\|, \Gamma]$ so that the optimal manifold is identified; this is done by decreasing γ_k linearly at each iteration.

From the intermediate to the input space. We now have a structure manifold $\mathcal{M}_k^g$ in the intermediate space, and can define $\tilde{g}_k$, a smooth extension of g on $\mathcal{M}_k^g$ to $\mathbb{R}^m$. Using a local equation h_k^g of $\mathcal{M}_k^g$, we define the smooth map $h_k = h_k^g \circ c : \mathbb{R}^n \to \mathbb{R}^{p_k}$, which locally defines $\mathcal{M}_k = c^{-1}(\mathcal{M}_k^g)$. Similarly, a smooth extension of F on $\mathcal{M}_k$ is defined by $\tilde{F}_k = \tilde{g}_k \circ c$.

Optimizing in the input space. We can now take steps to minimize the smooth extension $\tilde{F}_k$ on the smooth set $\mathcal{M}_k$ characterized by $h_k(x) = 0$:

$$\min_{x \in \mathbb{R}^n} \tilde{F}_k(x) \quad \text{s.t.} \quad h_k(x) = 0.$$

We turn to an elementary version of the traditional second-order Sequential Quadratic Programming methodology; see e.g., Bonnans et al. (2006, Chap. 14). At iteration k, the SQP direction $d_k^{\mathrm{SQP}}(x_k)$ at point x_k is defined as the solution of the following quadratic problem:

$$\begin{aligned} d_k^{\mathrm{SQP}}(x_k) = \operatorname*{arg\,min}_{d \in \mathbb{R}^n} \quad & \langle \nabla \tilde{F}_k(x_k), d \rangle + \frac{1}{2} \langle \nabla^2_{xx} L_k(x_k, \lambda_k(x_k)) d, d \rangle \\ \text{s.t.} \quad & h_k(x_k) + \mathrm{D}\, h_k(x_k) d = 0 \end{aligned} \tag{4.11}$$

where $\nabla^2_{xx} L_k$ denotes the Hessian of the Lagrangian $L_k(x, \lambda) = \tilde{F}_k(x) + \langle \lambda, h_k(x) \rangle$, and the multiplier $\lambda_k(x_k)$ defined from the following least-squares problem:

$$\lambda_k(x_k) = \operatorname*{arg\,min}_{\lambda \in \mathbb{R}^{p_k}} \left\| \nabla \tilde{F}_k(x_k) + \sum_{i=1}^{p_k} \lambda_i \nabla h_{k,i}(x_k) \right\|^2. \tag{4.12}$$

Finally, we check that $x_k + d_k^{\mathrm{SQP}}(x_k)$ provides a functional decrease in order to avoid degrading the iterate when the current structure is suboptimal. If the test is not verified, x_k is not updated and γ_k is decreased until a satisfying structure is detected.

Algorithm 4.1: General structure exploiting algorithm

Require: Pick x_0 near a minimizer, γ_0 large enough.

repeat
  $\gamma_k = \frac{\gamma_{k-1}}{2}$
  Compute $\mathrm{prox}_{\gamma_k g}(c(x_k))$ and obtain $\mathcal{M}_k^g$ locally defined by h_k^g
  $h_k = h_k^g \circ c$ (local equation of $\mathcal{M}_k$), $\tilde{F}_k = \tilde{g}_k \circ c$ (smooth extension)
  Compute $d_k^{\mathrm{SQP}}(x_k)$ by solving (4.11)
  if $F(x_k + d_k^{\mathrm{SQP}}(x_k)) \leq F(x_k)$ **then**
    $x_{k+1} = x_k + d_k^{\mathrm{SQP}}(x_k)$
  else
    $x_{k+1} = x_k$
  end if
until stopping criterion

Remark 4.5 (Complexity of one iteration). The main computational cost of one iteration of Algorithm 4.1 consists in the resolution of the quadratic program (4.11). Its plain resolution incurs a $\mathcal{O}(n^3)$ complexity. However, efficient approaches *reduce* this problem to a quadratic program on the subspace $\operatorname{Ker} \mathrm{D}\, h_k(x_k)$, which has dimension $\dim(\mathcal{M}_k)$. We refer to (Bonnans et al., 2006, Chap. 14) for an in-depth exposition of these techniques. The cost of an iteration is thus $\mathcal{O}(\dim(\mathcal{M}_k)^3)$. In situations where minimizers are highly structured (i.e., $\dim(\mathcal{M}^\star) \ll n$) this complexity may be comparable with the $\mathcal{O}(n^2)$ iteration complexity of classical nonsmooth optimization algorithms, such as nonsmooth BFGS (Lewis and Overton, 2013). △

4.4.2 *Convergence of Algorithm 4.1*

We proceed to give the result guaranteeing identification and local quadratic convergence of Algorithm 4.1.

In order to benefit from the quadratic rate of SQP, the elements of (4.11) should have the minimal regularity typically required by smooth constrained Newton methods (see e.g., Bonnans et al. (2006, Th. 14.5)); we thus make the following assumption.

Assumption 4.2 (Regularity of functions). The smooth extension and the manifold defining map are $\mathcal{C}^2$ with Lipschitz second derivatives, and the Jacobian of the constraints is full rank near the solution.

In order to focus on the algorithmic originality of the method, we slightly simplify the situation and make the two following algorithmic assumptions.

Assumption 4.3 (Nonconvex stability). The iterates of Algorithm 4.1 remain in the connex component of the sublevel set $\{x : F(x) \leq F(x_0)\}$ that contains $x^\star$.

This assumption ensures that an update that does not increase functional value remains in the neighborhood of the minimizer $x^\star$. It is naturally satisfied when F is convex, or when $x^\star$ is a global minimizer of F and x_0 is close enough to $x^\star$.

Assumption 4.4 (No Maratos effect). The iterates of Algorithm 4.1 are such that a step d that makes $x + d$ quadratically closer to x does not increase function value: $F(x + d) \leq F(x)$.

In smooth constrained optimization, getting closer (even at quadratic rate) to a minimizer does not imply decrease of objective value and constraint violation (measured by a merit function). This so-called Maratos effect (see e.g., Bonnans et al. (2006)) is one of the main difficulties in globalizing SQP schemes, which is out of the scope of the current chapter. We thus assume this effect does not affect our algorithm in theory, and use in practice one of the successful refinements, as discussed in Section 4.5.2.

We are now ready for the main convergence result of Algorithm 4.1, which establish that, after some finite time, the iterates identify exactly the optimal manifold and converge to the minimizer at a quadratic rate.

Theorem 4.7 (Exact identification and quadratic convergence). *Consider a function $F = g \circ c$ and $x^\star$ a strong minimizer,[3] qualified relative to the optimal manifold $\mathcal{M}^\star$. Assume that the smooth extension $\tilde{F}$ of F relative to $\mathcal{M}^\star$ and the corresponding manifold defining map h satisfy Assumption 4.2.*

If x_0 and $F(x_0)$ are close enough to $x^\star$ and $F(x^\star)$, γ_0 is large enough and the simplifying algorithmic Assumptions 4.3 and 4.4 hold, then there exists $C > 0$ such that the iterates $(x_k, \mathcal{M}_k)$ generated by Algorithm 4.1 verify:

$$\mathcal{M}_k = \mathcal{M}^\star \quad \text{and} \quad \|x_{k+1} - x^\star\| \leq C\|x_k - x^\star\|^2 \quad \text{for all } k \text{ large enough.}$$

The proof of this result consists in two steps. We first show the existence of a neighborhood of initialization on which the proximity operator will eventually identify the optimal manifold, once the stepsize has been sufficiently decreased. From this point onward, we prove that the SQP step provides a quadratic improvement and that the stepsize policy makes the manifold identification stable.

3 There exists $\eta > 0$, $\varepsilon > 0$ such that $F(x) \geq F(x^\star) + \eta\|x - x^\star\|^2$ for all $x \in \mathcal{B}(x^\star, \varepsilon)$.

Proof. Local identification of the optimal structure. By Theorem 4.4, there exists a ball centered around $x^\star$ of radius $\varepsilon_1 > 0$ and two positive constants L, Γ such that, for all $x \in \mathcal{B}(x^\star, \varepsilon_1)$ and $\gamma \in [L\|x - x^\star\|, \Gamma]$, $\operatorname{prox}_{\gamma g}(c(x))$ belongs to the optimal manifold $\mathcal{M}^{g\star} = c(\mathcal{M}^\star)$.

Local quadratic convergence of SQP on the optimal structure. Let us assume that the optimal manifold has been identified. The least square multiplier λ is defined by the optimality condition of (4.12):

$$\lambda(x) = -[\operatorname{Jac}_h(x) \operatorname{Jac}_h(x)^\top]^{-1} \operatorname{Jac}_h(x) \nabla \tilde{F}(x).$$

and since h is smooth and its Jacobian is full-rank near $x^\star$, λ is a Lipschitz continuous function near $x^\star$.

Since $x^\star$ is a strong minimizer of F, the Hessian of the Lagrangian restricted to the tangent space is positive definite. Indeed, since $x^\star$ is a strong minimizer of F on $\mathcal{M}^\star$, the Riemannian Hessian relative to the optimal manifold is positive definite. With the choice of multiplier (4.12), the Riemannian Hessian is exactly the Hessian of the Lagrangian restricted to the tangent space to $\mathcal{M}^\star$ at $x^\star$ (see Boumal (2022, Sec. 7.7)), which is thus itself positive definite.

Thus, using the local quadratic convergence of SQP (Bonnans et al., 2006, Th. 14.5), we get that there exists a ball centered around $x^\star$ of radius $\varepsilon_2 > 0$ such that the SQP step computed at a point x in that neighborhood relative to the optimal manifold provides a quadratic improvement towards $x^\star$. Reducing ε_2 if necessary, we can in addition have that the convergence is at least linear with rate $1/2$.

Initialization, identification, and quadratic convergence. Let $\varepsilon = \min(\varepsilon_1, \varepsilon_2, \Gamma/(2L))$. We will now show that initializing with $x_0 \in \{x : F(x) \leq F(x^\star) + \eta\varepsilon^2\}$ and $\gamma_0 \geq \Gamma$ provides the claimed behavior.

First, the functional decrease test of the algorithm and Assumption 4.4 guarantee that all iterates satisfy $F(x_k) \leq F(x_0)$. Using that $x^\star$ is a strong minimizer, we get that $\eta\|x_k - x^\star\|^2 \leq F(x_k) - F(x^\star) \leq F(x_0) - F(x^\star) \leq \eta\varepsilon^2$, and thus that the iterates remain in $\mathcal{B}(x^\star, \varepsilon)$.

Second, as $L\|x - x^\star\| \leq \Gamma/2$ for all $x \in \mathcal{B}(x^\star, \varepsilon)$ by construction, the fact that $\gamma_0 > \Gamma$ and (γ_k) decreases with geometric rate $1/2$ implies that there exists K such that $L\|x_K - x^\star\| \leq \gamma_K \leq \Gamma$.

Now, assume that at iteration $k \geq K$, $L\|x_k - x^\star\| \leq \gamma_k \leq \Gamma$. Since $x_k \in \mathcal{B}(x^\star, \varepsilon_1)$, we have from above that $\mathcal{M}^\star$ is identified. Thus, the SQP step is performed relative to the optimal manifold and $x_k + d_k^{\mathrm{SQP}}(x_k)$ brings a linear improvement of factor $1/2$ at least. Assumption 4.3 ensures that $F(x_k + d_k^{\mathrm{SQP}}(x_k)) \leq F(x_k)$ so that $x_{k+1} = x_k + d_k^{\mathrm{SQP}}(x_k)$ and thus

$$L\|x_{k+1} - x^\star\| \leq \frac{L}{2}\|x_k - x^\star\| \leq \frac{\gamma_k}{2} = \gamma_{k+1}.$$

This shows that $L\|x_{k+1} - x^\star\| \leq \gamma_{k+1} \leq \Gamma$, which completes the induction. We get that $\gamma_k \in [L\|x_k - x^\star\|, \Gamma]$ for all $k \geq K$. Finally, we have that for all $k \geq K$, $\mathcal{M}_k = \mathcal{M}^\star$ and x_{k+1} is quadratically closer to $x^\star$ than x_k. □

Remark 4.6 (Generalizations). Theorem 4.7 actually holds for any decrease factor of γ_k in $(0, 1)$ with the presented SQP update, or actually any superlinearly convergent update (e.g., a quasi-Newton type update). The above result is also readily adapted to an update that converges merely linearly, as long as its rate of convergence is faster than that of γ_k. This opens the possibility of using SQP

methods that rely only on first-order information (see e.g., Bolte and Pauwels (2016)). △

4.5 NUMERICAL ILLUSTRATIONS

In this section, we provide numerical illustrations for our results. Our goal here is twofold:

1. to illustrate the identification of the optimal manifold by the proximity operator near a minimizer as provided by Theorem 4.4;
2. to demonstrate the applicability of Algorithm 4.1 and observe the quadratic rates predicted by Theorem 4.7 on our running examples.

4.5.1 *Test problems*

We first consider the minimization of a maximum of smooth functions (4.2):

$$\min_{x\in\mathbb{R}^n} \max_{i=1,\dots,m} (c_i(x)).$$

We take the celebrated `MaxQuad` instance, where $n = 10$, $m = 5$ and each c_i is quadratic convex, making the whole function F convex (Bonnans et al., 2006, p. 153). In this instance, the optimal manifold is $\mathcal{M}_I^{\max}$ with $I = \{2, 3, 4, 5\}$.

Second, we consider the minimization of the maximum eigenvalue of an affine mapping (4.3):

$$\min_{x\in\mathbb{R}^n} \lambda_{\max}\left(A_0 + \sum_{i=1}^{n} x_i A_i\right).$$

We take $n = 25$ and we generate randomly $n + 1$ symmetric matrices of size 50. In this instance, the multiplicity of the maximum eigenvalue at the minimizer is $r = 3$.

We note that the "maximum" structure of a point, that is the partial smoothness manifold with smallest dimension, is $\mathcal{M}_r^{\lambda_{\max}}$ with $r = 6$. In particular, the multiplicity r cannot reach the matrix size $m = 50$. Indeed, the codimension of $\mathcal{M}_r^{\lambda_{\max}}$, that is the dimension of its normal spaces, should be lower than that of $\mathbb{R}^n$: $r(r+1)/2 - 1 \leq 25$, that is $r \leq 6$ (see the discussion in Shapiro and Fan (1995, pp. 555-556, Eq. 2.5)).

4.5.2 *Numerical setup*

All the algorithms are implemented in Julia (Bezanson et al., 2017); experiments may be reproduced using the code available online[4].

ALGORITHM. For the initialization of Algorithm 4.1, we set γ_0 as the smallest γ such that $\operatorname{prox}_{\gamma g}(c(x_0))$ has the most structure (e.g., if $g = \max$, we increase γ until the output of the proximity operator sets all coordinates to the same value). We solve the quadratic subproblem (4.11) providing the SQP step by the reduced system approach presented in Bonnans et al. (2006, p. 133). Tangent vectors are expressed in an orthonormal basis of the nullspace of the

4 See https://github.com/GillesBareilles/LocalCompositeNewton.jl for Algorithm 4.1 and https://github.com/GillesBareilles/NonSmoothSolvers.jl for the baselines.

Jacobian of the constraints at the current iterate. At iterate x_k, a second-order correction step $d^{\text{corr}}[x_k]$ is added to the SQP step $d^{\text{SQP}}[x_k]$. It is obtained as $d^{\text{corr}}[x_k] = \arg\min_{d \in \mathbb{R}^n} \{\|h(x_k) + \text{Jac}_h(x_k)\ d\|, \text{ s.t. } d \in \text{Im}\,\text{Jac}_h(x_k)^\top\}$. The full-step is thus $x_k + d^{\text{SQP}}[x_k] + d^{\text{corr}}[x_k]$.

BASELINES. For the two nonsmooth problems, we compare with the nonsmooth BFGS algorithm of (Lewis and Overton, 2013) (nsBFGS) and the gradient sampling algorithm (Burke et al., 2020). The nsBFGS method is not covered by any theoretical guarantees; it is known to perform relatively well in practice, often displaying a linear rate of convergence. In contrast, the Gradient Sampling algorithm generates with probability one a sequence of iterates for which all cluster points are Clarke stationary for F (Burke et al., 2020, Th. 3.1).[5] It is known however to have an iteration cost significantly higher than that of nonsmooth BFGS.

On a technical note, the iterations of nsBFGS stop as soon as the algorithm "breaks down in theory", and the quadratic subproblem of the gradient sampling iteration is solved by the method presented in Wolfe (1976).

Other methods could be considered relevant baselines.We mentioned the existence of methods for general composite functions, such as the composite bundle (Sagastizábal, 2013). In Section 1.3.2, we also reviewed schemes with potential superlinear rate specific to the maximum of smooth functions (Womersley and Fletcher, 1986), and the maximum eigenvalue minimization (Noll and Apkarian, 2005; Helmberg et al., 2014). We do not include these methods in our comparison since they are difficult to implement efficiently, contrary to the two methods proposed. We will review the $\mathcal{VU}$-algorithm (Mifflin and Sagastizábal, 2005) in Chapter 5.

ORACLES. Traditional methods for nonsmooth optimization, and notably bundle methods, require a first-order oracle:

$$x \mapsto (F(x), v) \quad \text{where } v \in \partial F(x)$$

while Gradient Sampling and nsBFGS require additionally to know if F is differentiable at point x. Algorithm 4.1 requires rather different information oracles:

$$\begin{aligned} x &\mapsto F(x) \\ x &\mapsto \mathcal{M}^g \ni \text{prox}_{\gamma g}(c(x)) \\ \mathcal{M}, x &\mapsto h(x), \text{Jac}_h(x), \nabla\tilde{F}(x), \nabla^2 L(x, \lambda). \end{aligned}$$

The second part of the oracle provides the candidate structure at point x. The last part of the oracle, which requires a point *and a candidate structure*, provides the second-order information of F required by the SQP step.

4.5.3 *Experiments*

Figure 4.6 reports the suboptimality of the considered methods in terms of CPU time and each marker corresponds to one iteration. All algorithms are initialized at a point x_0 obtained by running nsBFGS for several iterations.

5 This holds when F is locally Lipschitz over $\mathbb{R}^n$ and lower bounded, the algorithm iterates indefinitely and the sampling radius decreases to 0.

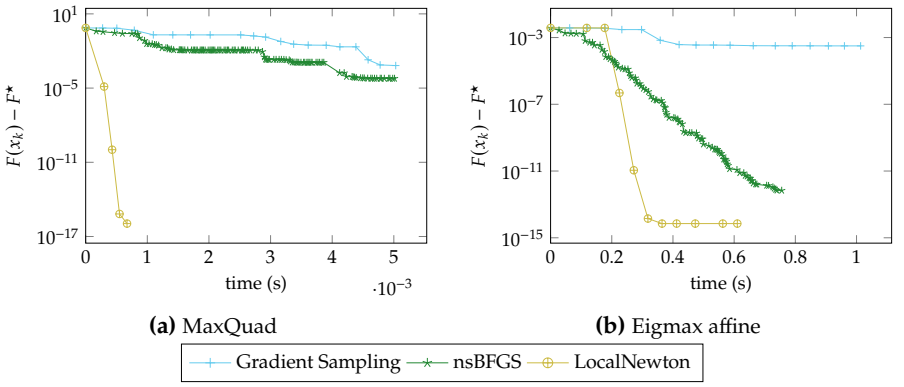


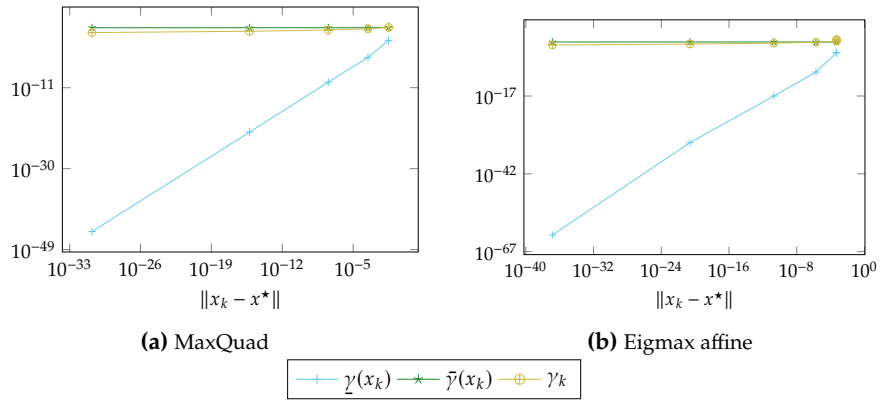
Figure 4.6: Suboptimality vs time (s)

Our algorithm compares favorably to nsBFGS and Gradient Sampling: it converges in a handful of iterations and less time. Note that this happens even though the iteration cost of our algorithm is higher than that of the other methods. Indeed, the oracles of our method are more complex and a quadratic problem needs to be solved, while the iteration cost of nsBFGS and Gradient Sampling is dominated by the computation of function values and subgradients at each trials of the line search.

In terms of identification, our method finds the correct manifold at the first iteration for MaxQuad, and at the third iteration for Eigmax. From that point, the iterates of Algorithm 4.1 reach machine precision in 3 iterations. This illustrates the quadratic convergence, and supports the idea that, for nondifferentiable problems as well, it is worth computing higher-order information to get fast local methods.

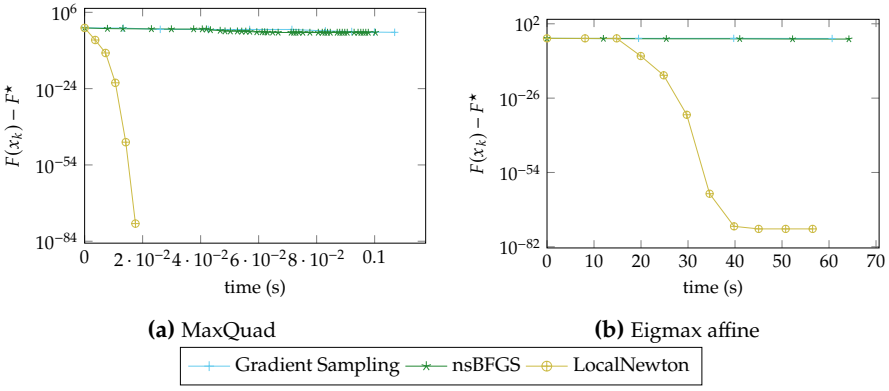
Figures 4.7 and 4.8 show the behavior of the same algorithms on the same problems with a higher numerical precision than Figure 4.6.[6] This allows to observe the identification of the algorithm and the quality of the bounds of Theorem 4.4. For each iterate x_k of Algorithm 4.1, we report the current step γ_k along with the minimal and maximal steps $\underline{\gamma}(x_k), \bar{\gamma}(x_k)$ such that $\mathrm{prox}_{\gamma g}(c(x_k))$ belongs to the optimal manifold. A first remark is that, as predicted by Theorem 4.7, the pair x_k, γ_k satisfies the identification condition $\gamma_k \in [L\|x_k - x^\star\|, \Gamma]$ after a few iterations. We also observe that $\bar{\gamma}(x_k)$ is near constant and that $\underline{\gamma}(x_k)$ converges to zero linearly with $\|x_k - x^\star\|$, as predicted by our result. Finally, we note that even though the initial point is away from the minimizer ($\|x_0 - x^\star\| \approx 10^{-2}$) and arbitrary, thus likely a differentiable point, the initialization of γ_0 ensures a quick identification.

6 Indeed, the flexibility of the Julia language allows to use the same implementation with the high precision `BigFloat` type, which precision is $1.73 \cdot 10^{-72}$, or the usual `Float64` type, which precision is $2.22 \cdot 10^{-16}$.



(a) MaxQuad **(b)** Eigmax affine

Figure 4.7: Stepsize γ_k vs iteration



(a) MaxQuad **(b)** Eigmax affine

Figure 4.8: Suboptimality vs time (s)

5

TOWARDS A GLOBAL NEWTON METHOD FOR NONSMOOTH COMPOSITE MINIMIZATION

5.1 INTRODUCTION

In this chapter, we consider the minimization of composite nonsmooth functions

$$\min_{x \in \mathbb{R}^n} F(x) = g \circ c(x), \tag{5.1}$$

where c is a smooth mappnig, and g is a nonsmooth real-valued function. Recall that in Chapter 4, we provided an algorithm, Algorithm 4.1 that, when started near a minimizer, identifies the smooth substructure and converges locally fast to it. We consider here the next step: we aim at providing a variant of the algorithm with same guarantees *when started at arbitrary points*.

A direct application of Algorithm 4.1 from arbitrary points is not satisfactory, since the algorithm can get trapped in nonminimizing points. Indeed, the structure detection tool introduced in Chapter 4 is not discriminating enough, as it detects the smooth substructure of any nearby nonsmooth point, including points which admit normal descent directions. We will come back on this behavior in details. This is a sufficient property near minimizers, but away from minimizers, the structure detection tool does not provide enough information anymore.

A possible fix would be a "two-phase" algorithm: first, finding a point near a minimizer with, e.g., nonsmooth BFGS or a bundle method, and then running the local Algorithm 4.1. The issue with such a scheme lies in choosing when to switch between algorithms. Furthermore, if the phase 1 method is very slow on a particular instance, a neighborhood of the minimizer will not be reached in reasonable time, making the fast local convergence unreachable in practice. This kind of difficulties is exactly what we wish to avoid in this thesis, by proposing methods that identify and exploit structure adaptively, with guaranteed identification.

5.1.1 *Towards globalization: tools and issues*

We review here three elements for globalization, as well as the two main difficulties we face when applying them.

NEWTON-LIKE STEPS. Building on Chapter 4, we can use Newton-type Sequential Quadratic Programming (SQP) steps to obtain the fast local convergence. At point x with candidate manifold $\mathcal{M}$, the SQP direction $d_{\mathcal{M}}^{\mathrm{SQP}}(x)$ is defined as

$$d_{\mathcal{M}}^{\mathrm{SQP}}(x) = \underset{d\in\mathbb{R}^n}{\arg\min} \quad \langle \nabla \tilde{F}_k(x_k), d\rangle + \frac{1}{2}\langle M_k d, d\rangle \qquad \text{s.t.} \quad h_k(x_k) + \mathrm{D}\, h_k(x_k)d = 0 \tag{5.2}$$

where h defines manifold $\mathcal{M}$, and $\tilde{F}$ is a smooth extension of F relative to $\mathcal{M}$. The main difference with Chapter 4 is that M_k may be the hessian of the Lagrangian $\nabla^2_{xx} L_k(x_k, \lambda_k(x_k))$, or any positive definite matrix. This flexibility allows to deal with non positive-definite Hessians, and encompasses quasi-Newton strategies such as BFGS or truncated Newton. This is a first element of globalization.

LINESEARCH PROCEDURE. In order to obtain a functional decrease at each iteration, we propose to use a linesearch procedure: at point x_k with direction $d(= d^{\mathrm{SQP}}(x))$, find some steplength $\alpha > 0$ that satisfies the following Armijo condition:

$$F(x + \alpha d) \leq F(x) + \alpha m F'(x; d), \tag{5.3}$$

where $m \in (0, \frac{1}{2})$ and $F'(x; d) = \max_{v\in\partial F(x)} \langle v, d\rangle$. This is the second element of globalization.

MAIN DIFFICULTY #1. One delicate point is to guarantee that the linesearch does not jeopardize the final quadratic rate brought by the SQP steps. To do so, we need to make sure that the unit stepsize is acceptable in a neighborhood of the minimizer. We will see on a simple example that this may not be possible, even arbitrarily close to minimizers. This issue was already present in Algorithm 4.1 as the algorithm incorporates a functional descent test. It was handled with Assumption 4.4 for the analysis.

This difficulty also appeared in Chapter 3; see Theorem B.1.

IDENTIFICATION. The last element of globalization is the identification procedure. We expect the globalized identification procedure to incorporate the behavior of the local algorithm. In addition, around arbitrary points, we only want to select manifolds that locally do not admit normal descent directions; see Section 5.4.1. In addition, the SQP step should be a descent direction for the objective function to make the linesearch possible; mathematically, we need $F'(x; d_{\mathcal{M}}^{\mathrm{SQP}}) < 0$.

Such manifolds are called "identifiable" by Davis et al. (2021).

MAIN DIFFICULTY #2. At this point, we can point out a difficulty with the local identification scheme of Algorithm 4.1. The structure detection tool $\mathrm{prox}_{\gamma g} \circ c$ doesn't provide enough information: it essentially selects manifolds based on function value information, and is thus oblivious to first-order information. In particular, it may select manifolds that admit descent directions in their normal space. On Fig. 5.1, $\mathrm{prox}_{\gamma g} \circ c$ detects structure $\mathcal{M}_2$ on a neighborhood of x_2. A direct application of Algorithm 4.1 could result in the convergence to point x_2. This point is indeed a minimizer of F *along* the manifold, but not for the full function F: there exists descent directions in the normal space to $\mathcal{M}_2$ at x_2.

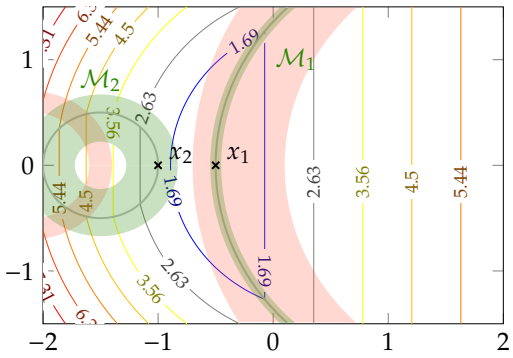


Figure 5.1: Attraction areas of $\text{prox}_{\gamma F}$ (red), and of $\text{prox}_{\gamma g} \circ c$ (green) on a maximum of three smooth functions. The nonsmooth function admits two substructure manifolds $\mathcal{M}_1$ and $\mathcal{M}_2$; points x_1 and x_2 are the minimizers of the restriction of F on these manifolds. Only x_1 is optimal for F. As expected, both operators detect the correct structure in a neighborhood of x_1, and $\text{prox}_{\gamma F}$ is not stable near x_2. However, $\text{prox}_{\gamma g} \circ c$ does detect the structure of x_2 on a neighborhood of this point, thus potentially trapping algorithms in that non-optimal substructure.

5.1.2 *Motivation, approach and algorithm*

We now lay out our approach. We first draw inspiration from the field of smooth constrained programming, and then present our approach and the preliminary results obtained.

A DETOUR IN NONLINEAR OPTIMIZATION. The composite problem (5.1) includes the particular class of ℓ_1 merit functions of nonlinear programming (Bonnans et al., 2006). These functions appear when transforming the minimization of a (smooth) function under (smooth) equality and inequality constraints into the unconstrained minimization of a crafted nonsmooth function. In that setting, finding the smooth substructure of a minimizer corresponds to finding exactly which inequality constraints hold with equality at the minimizer. The topic of globalizing an "active set" approach has been extensively studied, with several approaches including linesearch or filter methods (Nocedal and Wright, 2006). We set aside filter methods, which rely on the "multiobjective" nature of nonlinear programming, and focus here on linesearch methods. The theoretical analyses of linesearch active set methods (see e.g., Spellucci (1998)) show that, for nonlinear programming, *(1)* one can build a linesearch SQP scheme that generate iterates which limit points satisfy KKT conditions and *(2)* if a qualification condition holds on one of the limit points, then the methods identifies exactly the minimizer structure and converges to it quadratically.

This is an instance of problems with chosen nonsmoothness, discussed in Chapter 1.

OBJECTIVES AND CONTRIBUTIONS. We aim to generalize the above classic results in smooth constrained optimization to the setting of nonsmooth composite optimization. For a correctly specified algorithm, we wish to show that:

(i) all limit points of the iterate sequence are critical points,

(ii) if one limit point is qualified, then its structure is eventually identified and the iterates converge to it quadratically.

We propose Algorithm 5.1, an "ideal" algorithm that alternates a structure identification step and a linesearch on efficient Newton-like steps. In this

Algorithm 5.1: Ideal global algorithm for structured composite optimization

repeat
  Obtain $\mathcal{M}_k$, h_k by an identification procedure at point x_k ▷ Structure detection
  Select a smooth extension $\tilde{F}_k$ and a matrix M_k
  Compute $d_k^{\mathrm{SQP}}(x_k)$ by solving (5.2) ▷ Structure exploitation
  Find α_k by Armijo linesearch on F (5.3)
  $x_{k+1} = x_k + \alpha_k d_k^{\mathrm{SQP}}(x_k)$.
until stopping criterion

prospective chapter, we leave out the identification step to heuristics (related to objective *(i)*, see Section 5.3.3), and show first results towards showing the feasibility of such a linesearch SQP method (related to objective *(ii)*). Specifically, we give the following results:

- With a second order correction, the SQP step provides descent near structured minimizers which structure is correctly guessed;
- We show that near a qualified minimizer, the full SQP direction provides descent;

OUTLINE OF THIS CHAPTER. In Section 5.2, we introduce some technical tools and results on nonsmooth analysis and SQP steps. In Section 5.3, we provide theoretical guarantees on the behavior of the method after identification of the optimal manifold: well-posedness of the linesearch and eventual admissibility of the unit stepsize that ensures a fast local rate. In Section 5.4, we introduce a tractable optimality condition for structured nonsmooth problems, and make a step towards an identification procedure by proposing a way to detect smooth substructures with only normal ascent directions. We briefly discuss the standing questions in Section 5.5, including the standing question of combining linesearch steps with an identification procedure (objective *(i)*). Finally, we propose in Section 5.6 an identification heuristic and numerical illustrations of a proposed algorithm.

5.2 A CLOSER LOOK INTO NONSMOOTHNESS AND SQP STEPS

In this section, we discuss technical properties that will be useful in the following developments. First, we introduce two geometrical properties on nonsmooth functions, and second, we discuss the structure of SQP steps.

5.2.1 *Two properties on nonsmooth functions*

We consider in this section, and more generally in this chapter, a nonsmooth *real-valued* function $F : \mathbb{R}^n \to \mathbb{R}$.

We introduce two properties, taken in Davis et al. (2021), that will be central in studying the descent of SQP steps near minimizers.

Property 5.1 ((b) regular). We say that F is *(b)-regular* along $\mathcal{M}$ at $\bar{x}$ if, for any $\delta > 0$, there exists a neighborhood $\mathcal{N}_{\bar{x}}$ of $\bar{x}$ such that

$$|F(y) - F(x) - \langle v, y - x\rangle| \leq \delta\sqrt{1 + \|v\|^2}\|x - y\|$$

hold for all $x \in \mathcal{N}_{\bar{x}}$, $y \in \mathcal{M} \cap \mathcal{N}_{\bar{x}}$ and $v \in \partial F(x)$.

Broadly speaking, this property allows to control, near substructure manifolds, the growth of the function by the distance to the manifold and elements of the subdifferential of F on $\mathcal{M}$.

Remark 5.1 (Validity of (b)-regularity). There are several ways to establish that a function is (b)-regular.

First, Davis et al. (2021, Th. 2.6.2) provides a chain rule for composite functions $F = g \circ c$: if g is (b)-regular at $\bar{y}$, $\mathcal{M}^g$ is a $\mathcal{C}^p$-smooth manifold, the restriction of g to $\mathcal{M}^g$ is $\mathcal{C}^p$ smooth, and transversality (recall Eq. (4.4)) holds at $\bar{y}$, then F is (b)-regular at $c^{-1}(\bar{y})$ relative to $c^{-1}(\mathcal{M})$. One readily checks this property on the nonsmooth functions max and $\lambda_{\max}$, which were the main examples of Chapter 4.

More generally, for any real-valued continuous definable function, and any $p > 0$, there exists a partition of $\mathbb{R}^n$ into finitely many $\mathcal{C}^p$-smooth manifolds such that F is (b)-regular at any $\bar{x} \in \mathcal{M}$ along $\mathcal{M}$, for any manifold $\mathcal{M}$ (Davis et al., 2021, Th. 2.7.4). △

Property 5.2 (proximal aiming). We say that F satisfies the *proximal aiming* property a $\bar{x}$ when there exist a constant $c > 0$ and a neighborhood $\mathcal{N}_{\bar{x}}$ of $\bar{x}$ such that

$$\langle v, x - \operatorname{proj}_{\mathcal{M}}(x) \rangle \geq c \operatorname{dist}_{\mathcal{M}}(x), \tag{5.4}$$

for all $x \in \mathcal{N}_{\bar{x}}$ and $v \in \partial F(x)$.

This property captures the fact that subgradients form an acute angle with the opposite of the direction to the nonsmoothness manifold of $\bar{x}$.

Remark 5.2 (Validity of proximal aiming). By Davis et al. (2021, Cor. 2.1.5), the proximal aiming property holds when there exists a locally smooth manifold $\mathcal{M} \ni \bar{x}$ over which F is locally smooth, F is locally Lipschitz, (b)-regular along $\mathcal{M}$ at $\bar{x}$, subdifferentially regular at $\bar{x}$, and there holds:

$$\inf\{\|v\|, v \in \partial F(x), x \in \mathcal{N}_{\bar{x}} \setminus \mathcal{M}\} > 0,$$

where $\mathcal{N}_{\bar{x}}$ denote a neighborhood of $\bar{x}$. Note the proximity between the above assumption, the modulus of identifiability of Lewis and Tian (2022, Def. 2.3), and the positivity of $c_{\text{ri}} (= \inf_{p \in \mathcal{N}_{\bar{y}}} \inf_{v_n \in \operatorname{rbd} \partial^N g(p)} \|v_n\|)$ that appears in the proof of Theorem 4.4. It interprets as all normal directions being ascent for F. △

5.2.2 *Anatomy of SQP-type steps*

We introduce tools that will be useful in the analysis of the local convergence of SQP steps. These facts are standard in the SQP literature; see Bonnans et al. (2006, Chap. 14).

TRANSLATED MANIFOLD. We will need to consider points close to a target subspace $\mathcal{M}$ but not on it. At such a point x, we introduce the translated manifold $\mathcal{M}_x \triangleq h^{-1}(\{h(x)\})$, where $h : \mathbb{R}^n \to \mathbb{R}^p$ is a manifold-defining map for $\mathcal{M}$. Therefore, x lies on $\mathcal{M}_x$ and we have tangent and normal spaces there. *See Section 2.3.*

We will require an additional property for our manifold defining maps, which we lay out below.

Property 5.3 . Consider a function F, partly smooth at point $\bar{x}$ relative to a manifold $\mathcal{M}$. We say that the manifold defining map h of $\mathcal{M}$ *agrees with F at first-order* if there exists a neighborhood $\mathcal{N}_{\bar{x}}$ of $\bar{x}$ such that, for any $x \in \mathcal{N}_{\bar{x}}$,

$$\operatorname{Par}(\partial F(x)) \subset N_x \mathcal{M}_x.$$

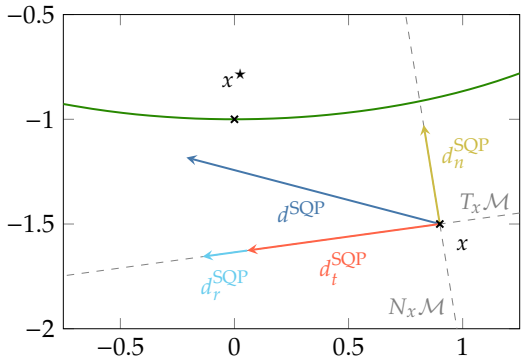


Figure 5.2: Illustration of the different components of an SQP step described in Lemma 5.4.

Here, $\mathrm{Par}(A)$ denotes the linear space spanned by A, defined as $\mathrm{Par}(A) \triangleq \mathrm{Aff}(A) - a$, for $a \in A$.

In practice, one checks that usual manifold defining maps for the maximum of smooth functions or the maximum eigenvalue meet this condition.

In this setting, we introduce the following objects:

- $Z^-(x) \in \mathbb{R}^{p\times n}$: a basis of the tangent space $T_x\mathcal{M}_x$;
- $\mathrm{grad}_{\mathcal{M}} F(x) \triangleq Z^{-\top}(x)v \in T_x\mathcal{M}_x$ for $v \in \partial F(x)$: the *reduced gradient* of the smooth function $\tilde{F}$;
- $\mathrm{Hess}_{\mathcal{M}} F(x) \triangleq Z^{-\top}(x)MZ^-(x)$: the *reduced Hessian*, that is the restriction to $T_x\mathcal{M}_x$ of the linear operator M in the step computation Eq. (5.2).

Here, $Z^{-\top}$ denotes the transpose of Z^-.

In the following, we drop the dependence in x when no confusion is possible.

SQP STEP. The SQP step splits in three components, which we illustrate on Fig. 5.2. This decomposition essentially comes form the developments in Bonnans et al. (2006, Chap. 14).

Lemma 5.4 (SQP step decomposition). *The SQP step d^{SQP} Eq. (5.2) for minimizing a smooth function $\tilde{F}$ on the manifold $\mathcal{M}$, computed at point x, writes:*

$$d^{\mathrm{SQP}}(x) = d_n^{\mathrm{SQP}}(x) + d_t^{\mathrm{SQP}}(x) + d_r^{\mathrm{SQP}}(x),$$

where

- $d_n^{\mathrm{SQP}}(x) = \arg\min_{d\in N_x\mathcal{M}_x} \|h(x) + \mathrm{D}\,h(x)\cdot d\|$ *is a Newton-Raphson step, that solely aims at increasing the feasibility relative to $\mathcal{M}$.*
- $d_t^{\mathrm{SQP}}(x) = -Z^-(x)[\mathrm{Hess}_{\mathcal{M}} F(x)]^{-1}\,\mathrm{grad}_{\mathcal{M}} F(x) \in T_x\mathcal{M}_x$ *interprets, when x is feasible but not optimal, as the Riemannian Newton step prior its retraction on the manifold; see Absil et al. (2009b) for details.*
- $d_r^{\mathrm{SQP}}(x) = -Z^-(x)[\mathrm{Hess}_{\mathcal{M}} F(x)]^{-1}Z^{-\top}(x)Md_n^{\mathrm{SQP}}(x) \in T_x\mathcal{M}_x$ *is a residual term, required to obtain a quadratic convergence rate.*

Proof. This decomposition is a simple rewriting in our notation of Bonnans et al. (2006, Eqs. 14.33, 14.34). □

5.3 AFTER IDENTIFICATION: VALIDITY OF LINESEARCH ON SQP STEPS

In this section, we study the SQP step relative to F, assuming that the optimal manifold has been correctly identified. We first show in Section 5.3.1 that the SQP is a descent direction for F near a minimizer. In Section 5.3.2, we show that taking a full-length plain SQP step may increase function value, and prove that adding a second-order correction resolves this issue. Finally, we discuss in Section 5.3.3 a condition under which this correction is dispensable, allowing to alleviate its cost in practice.

These results parallel closely those from Non Linear Programming; they have been largely inspired by Bonnans et al. (2006, Chap. 17).

5.3.1 *Descent of SQP steps*

We first show that an SQP step near a minimizer $\bar{x}$ and relative to this point is a descent direction.

Lemma 5.5 (descent of SQP step near identifiable minimizers). *Consider a function F and a point $\bar{x}$ such that $0 \in \partial F(\bar{x})$, F is partly smooth at $\bar{x}$ relative to $\mathcal{M}$, assume that F meets the proximal aiming property at $x^\star$ (property 5.2), and that the manifold defining map for h agrees with F at $\bar{x}$ (property 5.3).*

Then there exists a neighborhood $\mathcal{N}_{\bar{x}}$ such that, for $x \in \mathcal{N}_{\bar{x}}$,

$$\begin{aligned} F'(x; d^{\mathrm{SQP}}) &= -\langle \mathrm{Hess}_{\mathcal{M}} F(x) d_t^{\mathrm{SQP}}, d_t^{\mathrm{SQP}} \rangle + \max_{v \in \partial F(x)} \langle v, d_n^{\mathrm{SQP}} \rangle + o(\mathrm{dist}_{\mathcal{M}}(x)) \quad (5.5) \\ &\leq -\langle \mathrm{Hess}_{\mathcal{M}} F(x) d_t^{\mathrm{SQP}}, d_t^{\mathrm{SQP}} \rangle - c\, \mathrm{dist}_{\mathcal{M}}(x) + o(\mathrm{dist}_{\mathcal{M}}(x)), \end{aligned}$$

where the steps are computed at point x.

Proof. The directional derivative is obtained from the subdifferential as follows:

$$F'(x; d^{\mathrm{SQP}}) = \max_{v \in \partial F(x)} \langle v, d^{\mathrm{SQP}} \rangle.$$

Using property 5.3 and injecting the structure of the SQP step described in Lemma 5.4, we get

$$\begin{aligned} F'(x; d^{\mathrm{SQP}}) &= \max_{v \in \partial F(x)} \left\langle v, d_n^{\mathrm{SQP}} + d_t^{\mathrm{SQP}} + d_r^{\mathrm{SQP}} \right\rangle \\ &= \max_{v \in \partial F(x)} \langle v, d_n^{\mathrm{SQP}} \rangle + \langle \mathrm{grad}_{\mathcal{M}} F(x), d_t^{\mathrm{SQP}} \rangle + \langle \mathrm{grad}_{\mathcal{M}} F(x), d_r^{\mathrm{SQP}} \rangle. \end{aligned}$$

Combining $d_r^{\mathrm{SQP}} = \mathcal{O}(\mathrm{dist}_{\mathcal{M}}(x))$ and $\mathrm{grad}_{\mathcal{M}} F(x) = o(1)$ yields $\langle \mathrm{grad}_{\mathcal{M}} F(x), d_r^{\mathrm{SQP}} \rangle = o(\mathrm{dist}_{\mathcal{M}}(x))$. Therefore,

$$F'(x; d^{\mathrm{SQP}}) = \max_{v \in \partial F(x)} \langle v, d_n^{\mathrm{SQP}} \rangle + \langle \mathrm{grad}_{\mathcal{M}} F(x), d_t^{\mathrm{SQP}} \rangle + o(\mathrm{dist}_{\mathcal{M}}(x)).$$

Since $d_t^{\mathrm{SQP}} = -Z^- [\mathrm{Hess}_{\mathcal{M}} F(x)]^{-1} \mathrm{grad}_{\mathcal{M}} F(x)$, we get

$$\langle \mathrm{grad}_{\mathcal{M}} F(x), d_t^{\mathrm{SQP}} \rangle = -\langle \mathrm{Hess}_{\mathcal{M}} F(x) d_t^{\mathrm{SQP}}, d_t^{\mathrm{SQP}} \rangle,$$

which yields the first part of the result. The second part follows from the proximal aiming (Eq. (5.4)), and the fact that $d_n^{\mathrm{SQP}} = \mathrm{proj}_{\mathcal{M}}(x) - x + o(\mathrm{dist}_{\mathcal{M}}(x))$ (Eq. (5.8)). □

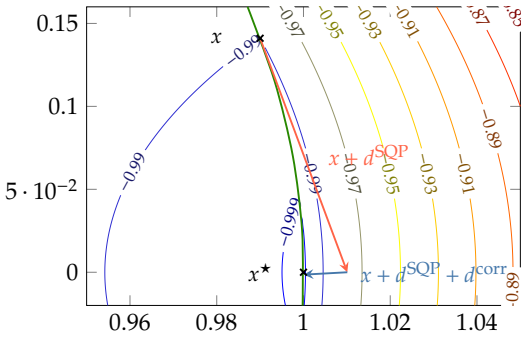


Figure 5.3: The Maratos effect: an SQP step d^{SQP} from point x increases function value; see Example 5.1. Adding a second-order correction step d^{corr} Eq. (5.6) reduces function value.

5.3.2 *Eventual admissibility of unit step size*

Iterating SQP steps in a neighborhood of a minimizer generates points that converge quadratically to the minimizer (Bonnans et al., 2006, Th. 14.5). To globalize the method, we use a linesearch procedure: we follow the SQP step only by some factor that ensures a sufficient decrease of the function value that satisfies Armijo's equation Eq. (5.3). At this point, it is not clear that the globalized scheme still converges locally quadratically. Indeed, following the SQP step with a unit stepsize may cause a functional increase, even when the step brings the next iterate quadratically closer to the minimizer.

We review an example of this phenomenon, called *Maratos effect* in the following example. We then propose a way to fix it.

Example 5.1 (Functional ascent of SQP step). We illustrate the "Maratos effect" on a simple function, and illustrate it on Fig. 5.3. This example is directly adapted from non linear programming (Bonnans et al., 2006, 17.6). Consider

$$F(x) = -x_1 + \max\left(1.6\|x\|^2 - 1.6, 0.4\|x\|^2 - 0.4\right).$$

The plain SQP step from point $x = (\cos\theta, \sin\theta)$ admits a closed form expression

$$d^{\mathrm{SQP}}(x) = (\sin^2\theta, -\sin\theta\cos\theta),$$

so that $F(x) = -\cos\theta$ and $F(x + \alpha d^{\mathrm{SQP}}(x)) = -\cos\theta + 0.4\alpha\sin^2\theta$. In particular, $F(x + d^{\mathrm{SQP}}(x)) > F(x)$ for all θ: the full SQP step increases functional value for x arbitrarily close to $x^\star$. □

We adopt one classical possibility to fix this issue, the so-called second-order correction step (Bonnans et al., 2006, pp. 310-314). At point x with SQP step d^{SQP}, one computes d^{corr} as a Newton-Raphson step aimed at improving feasibility of $x + d^{\mathrm{SQP}}$:

$$d^{\mathrm{corr}} = \operatorname*{arg\,min}_{d \in N_x\mathcal{M}_x} \|h(x + d^{\mathrm{SQP}}(x)) + \mathrm{D}\,h(x)\cdot d\|. \tag{5.6}$$

This step is illustrated on Fig. 5.3.

Adding a second-order correction to the SQP step result ensures satisfaction of an Armijo rule, as shown in the following theorem.

Theorem 5.6 (Eventual admissibility of unit step). *Consider a function F and point $x^\star$ such that F is partly smooth at $x^\star$ relative to some manifold $\mathcal{M}^\star$, $x^\star$ is a strong minimizer:*

$$0 \in \operatorname{ri} \partial F(x^\star) \quad \textit{and} \quad \operatorname{Hess} F(x^\star) \succ 0,$$

and assume that F meets the proximal aiming property at $x^\star$ (property 5.2) and is (b)*-regular at $x^\star$ along $\mathcal{M}$ (property 5.1), and that the manifold defining map for h agrees with F at $\bar{x}$ (property 5.3). Finally, assume that matrix M is taken such that, when x goes to $x^\star$, there holds:*

$$\left\langle \operatorname{Hess}_{\mathcal{M}} F(x) d_t^{\mathrm{SQP}}, d_t^{\mathrm{SQP}} \right\rangle \geq \left\langle \operatorname{proj}_{T_{x^\star}\mathcal{M}^\star} \operatorname{Hess} F(x^\star) \operatorname{proj}_{T_{x^\star}\mathcal{M}^\star} d_t^{\mathrm{SQP}}, d_t^{\mathrm{SQP}} \right\rangle + o(\|d^{\mathrm{SQP}}(x)\|^2) + o(\operatorname{dist}_{\mathcal{M}}(x)) \quad (5.7)$$

Then there exists a neighborhood $\mathcal{N}_{x^\star}$ of $x^\star$ such that, if $x \in \mathcal{N}_{x^\star}$ and $m \in (0, 1/2)$,

$$F(x + d^{\mathrm{SQP}}(x) + d^{\mathrm{corr}}(x)) \leq F(x) + mF'(x; d^{\mathrm{SQP}}(x)),$$

where $d^{\mathrm{SQP}}(x)$ denotes the SQP step (5.2) and $d^{\mathrm{corr}}(x)$ the second-order correction (5.6), both performed relative to $\mathcal{M}^\star$.

The proof consists in two steps: first building a local description of F at points x and $x + d^{\mathrm{SQP}} + d^{\mathrm{corr}}$ with precision $o(\operatorname{dist}_{\mathcal{M}}(x)) + o(\|d^{\mathrm{SQP}}\|^2)$, then using it to show the sufficient decrease of the corrected SQP update with steplength 1. Figure 5.4 depicts the situation.

The effect of adding the second-order correction appears in Eq. (5.13): if x_+ fails to incorporate $d^{\mathrm{corr}}(x)$, then $\langle v_{x_+}, x_+^{\mathcal{M}} - x_+ \rangle = \mathcal{O}(\operatorname{dist}_{\mathcal{M}}(x + d^{\mathrm{SQP}}))$. This term cannot be controlled, in particular when x is near $\mathcal{M}^\star$ and the manifold has non-null curvature as in Fig. 5.3.

We first require a simple result: the tangent space $T_x\mathcal{M}_x$ evolve slowly as x moves away from $\mathcal{M}$. The following proofs are based on three points: some point x, $x^{\mathcal{M}}$ its projection on $\mathcal{M}$ and the local minimum $x^\star$. We consider the situation when x tends to $x^\star$ and use the o and $\mathcal{O}$ notation.

Lemma 5.7 (Tangent spaces along normal directions). *Consider a point x, contained in a manifold $\mathcal{M}$ locally defined by a $\mathcal{C}^2$ mapping h, and a smooth vector $d(x)$ that vanishes as x goes to $\bar{x}$. Then,*

$$\operatorname{proj}_{T_x\mathcal{M}_x}(d(x)) = \operatorname{proj}_{T_{x^{\mathcal{M}}}\mathcal{M}}(d(x)) + o(\operatorname{dist}_{\mathcal{M}}(x)).$$

where $x^{\mathcal{M}}$ denotes the orthogonal projection of x on $\mathcal{M}$.

Proof. The projections of a fixed vector d on the tangent spaces write

$$\operatorname{proj}_{T_x\mathcal{M}_x}(d) = d - \left[\mathrm{D}\, h(x)^\top\right]^\dagger d$$
$$\operatorname{proj}_{T_{x^{\mathcal{M}}}\mathcal{M}}(d) = d - \left[\mathrm{D}\, h(x^{\mathcal{M}})^\top\right]^\dagger d$$

where the symbol † denotes the Moore Penrose pseudo-inverse. Since h is $\mathcal{C}^2$, the two projections are related by a Taylor development, which first order term is: Since the above functions are $\mathcal{C}^1$, there holds, on a neighborhood $\mathcal{N}_x$ of x,

$$\|\operatorname{proj}_{T_x\mathcal{M}_x}(d) - \operatorname{proj}_{T_{x^{\mathcal{M}}}\mathcal{M}_x}(d)\| \leq \left\| \left[\mathrm{D}\, h(x)^\top\right]^\dagger - \left[\mathrm{D}\, h(x^{\mathcal{M}})^\top\right]^\dagger \right\| \|d\|$$

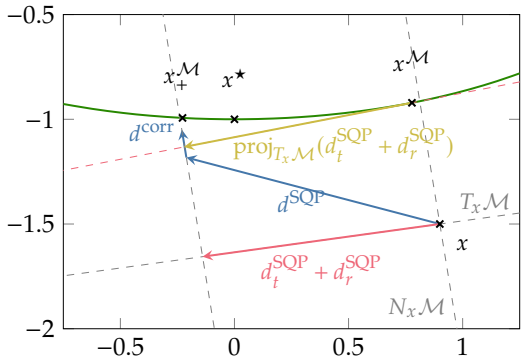


Figure 5.4: Illustration of the points and vectors that appear in the proof of Theorem 5.6.

$$\leq \sup_{u \in \mathcal{N}_x} \left\| \mathrm{D}\left(x \mapsto \left[\mathrm{D}\, h(x)^\top\right]^\dagger\right)(u) \right\| \underbrace{\|x - x^{\mathcal{M}}\|}_{=\mathrm{dist}_{\mathcal{M}}(x)} \|d\|$$

Applying the above inequality with a vector $d(x) = o(1)$ yields the result. □

We now proceed with the proof of Theorem 5.6.

Proof (of Theorem 5.6). Consider a point x near $x^\star$, and let $x_+ = x + d^{\mathrm{SQP}}(x) + d^{\mathrm{corr}}(x)$. We denote $x^{\mathcal{M}} = \mathrm{proj}_{\mathcal{M}}(x)$ and $x_+^{\mathcal{M}} = \mathrm{proj}_{\mathcal{M}}(x_+)$.

As a preliminary step, we collect some useful estimates on the SQP and correction steps, from the proof of Bonnans et al. (2006, Th. 17.7):

$$d_n^{\mathrm{SQP}} = x^{\mathcal{M}} - x + o(\mathrm{dist}_{\mathcal{M}}(x)) \tag{5.8}$$

$$\mathrm{dist}_{\mathcal{M}}(x) \leq C\|h(x)\| \tag{5.9}$$

$$\|h(x + d^{\mathrm{SQP}} + d^{\mathrm{corr}})\| = o(\|d^{\mathrm{SQP}}\|^2). \tag{5.10}$$

STEP 1. We first connect $F(x)$ and $F(x_+)$ via the intermediate points $x^{\mathcal{M}}$ and $x_+^{\mathcal{M}}$, with precision $o(\|\,\mathrm{dist}_{\mathcal{M}}(x)\|) + o(\|d^{\mathrm{SQP}}(x)\|^2)$.

First, the *(b)*-regularity property 5.1 provides, for any $v_x \in \partial F(x)$,

$$\begin{aligned} F(x) &\geq F(x^{\mathcal{M}}) + \langle v_x, x - x^{\mathcal{M}} \rangle + o(\mathrm{dist}_{\mathcal{M}}(x)) \\ &= F(x^{\mathcal{M}}) - \langle v_x, d_n^{\mathrm{SQP}} \rangle + o(\mathrm{dist}_{\mathcal{M}}(x)), \end{aligned} \tag{5.11}$$

where we used the estimate (5.8).

Then, the second-order Taylor model of F on $\mathcal{M}$ Eq. (2.3) gives, for any second-order retraction R,

$$F(x_+^{\mathcal{M}}) = F(x^{\mathcal{M}}) + \langle \mathrm{grad}\, F(x^{\mathcal{M}}), \eta \rangle + \frac{1}{2}\langle \mathrm{Hess}\, F(x^{\mathcal{M}})\eta, \eta \rangle + o(\|\eta\|^2), \tag{5.12}$$

where $\eta \in T_{x^{\mathcal{M}}}\mathcal{M}$ is defined implicitly by $x_+^{\mathcal{M}} = \mathrm{R}_{x^{\mathcal{M}}}(\eta)$.

Finally, the *(b)*-regularity property 5.1 provides, for $v_{x_+} \in \partial F(x_+)$,

$$F(x_+^{\mathcal{M}}) \geq F(x_+) + \langle v_{x_+}, x_+^{\mathcal{M}} - x_+ \rangle + o(\mathrm{dist}_{\mathcal{M}}(x_+))$$

Applying Cauchy-Schwarz' inequality and using local boundedness of ∂F near $\bar{x}$, we deduce $\langle v_{x_+}, x_+^{\mathcal{M}} - x_+ \rangle = \mathcal{O}(\mathrm{dist}_{\mathcal{M}}(x_+))$. Applying the estimates detailed in Eqs. (5.9) and (5.10) yields $\mathrm{dist}_{\mathcal{M}}(x_+) = \mathcal{O}(h(x_+)) = o(\|d^{\mathrm{SQP}}\|^2)$. Therefore,

$$F(x_+^{\mathcal{M}}) \geq F(x_+) + o(\|d^{\mathrm{SQP}}\|^2). \tag{5.13}$$

Summing up Eqs. (5.11)–(5.13) yields, for any $v_x \in \partial F(x)$,

$$F(x) \geq F(x_+) - \langle v_x, d_n^{\mathrm{SQP}} \rangle - \langle \mathrm{grad}\, F(x^{\mathcal{M}}), \eta \rangle - \frac{1}{2} \langle \mathrm{Hess}\, F(x^{\mathcal{M}})\eta, \eta \rangle + o(\mathrm{dist}_{\mathcal{M}}(x)) + o(\|d^{\mathrm{SQP}}\|^2) + o(\|\eta\|^2). \tag{5.14}$$

We now turn to explicit quantity η by expressing it in terms of d^{SQP}. We choose to use in (5.12) the orthographic retraction, defined in Proposition A.2, as it locally admits an explicit inverse:

$$\begin{aligned}\eta &= \mathrm{proj}_{T_{x^{\mathcal{M}}}\mathcal{M}}(x_+^{\mathcal{M}} - x^{\mathcal{M}}) \\ &= \mathrm{proj}_{T_{x^{\mathcal{M}}}\mathcal{M}}(x_+^{\mathcal{M}} - x_+) + \mathrm{proj}_{T_{x^{\mathcal{M}}}\mathcal{M}}(x_+ - x) + \mathrm{proj}_{T_{x^{\mathcal{M}}}\mathcal{M}}(x - x^{\mathcal{M}}),\end{aligned}$$

where the first term is bounded by $\mathrm{dist}_{\mathcal{M}}(x_+) = o(\|d^{\mathrm{SQP}}\|^2)$ and the third is null as $x^{\mathcal{M}} - x$ is normal to $\mathcal{M}$ at $x^{\mathcal{M}}$. Since $x_+ = x + d^{\mathrm{SQP}} + d^{\mathrm{corr}}$,

$$\eta = \mathrm{proj}_{T_{x^{\mathcal{M}}}\mathcal{M}}(d^{\mathrm{SQP}} + d^{\mathrm{corr}}) + o(\|d^{\mathrm{SQP}}\|^2).$$

Applying first Lemma 5.7, since $d^{\mathrm{SQP}} + d^{\mathrm{corr}} = o(1)$, and then the SQP step decomposition introduced in Lemma 5.4 with the fact that $d_n^{\mathrm{SQP}}, d^{\mathrm{corr}} \in N_x\mathcal{M}_x$ yields

$$\begin{aligned}\eta &= \mathrm{proj}_{T_x\mathcal{M}_x}(d^{\mathrm{SQP}} + d^{\mathrm{corr}}) + o(\mathrm{dist}_{\mathcal{M}}(x)) + o(\|d^{\mathrm{SQP}}\|^2) \\ &= d_t^{\mathrm{SQP}} + d_r^{\mathrm{SQP}} + o(\mathrm{dist}_{\mathcal{M}}(x)) + o(\|d^{\mathrm{SQP}}\|^2).\end{aligned}$$

We can now explicit the η terms of Eq. (5.14):

$$\begin{aligned}\langle \mathrm{grad}\, F(x^{\mathcal{M}}), \eta \rangle &= \langle \mathrm{grad}\, F(x^{\mathcal{M}}), d_t^{\mathrm{SQP}} \rangle + o(\mathrm{dist}_{\mathcal{M}}(x)) + o(\|d^{\mathrm{SQP}}(x)\|^2) \\ \langle \mathrm{Hess}\, F(x^{\mathcal{M}})\eta, \eta \rangle &= \langle \mathrm{Hess}\, F(x^{\mathcal{M}})d_t^{\mathrm{SQP}}, d_t^{\mathrm{SQP}} \rangle + o(\mathrm{dist}_{\mathcal{M}}(x)) + o(\|d^{\mathrm{SQP}}(x)\|^2),\end{aligned}$$

where we used that $d_r^{\mathrm{SQP}}(x) = \mathcal{O}(\mathrm{dist}_{\mathcal{M}}(x))$, $\mathrm{grad}\, F(x^{\mathcal{M}}) = o(1)$, and $d_t^{\mathrm{SQP}}(x) = o(1)$. Notice that $\mathrm{grad}\, F(x^{\mathcal{M}}) = \mathrm{grad}_{\mathcal{M}} F(x) + \mathcal{O}(\mathrm{dist}_{\mathcal{M}}(x))$, so that, with Lemma 5.4,

$$\langle \mathrm{grad}\, F(x^{\mathcal{M}}), d_t^{\mathrm{SQP}} \rangle = -\langle \mathrm{Hess}_{\mathcal{M}} F(x) d_t^{\mathrm{SQP}}, d_t^{\mathrm{SQP}} \rangle + o(\mathrm{dist}_{\mathcal{M}}(x)).$$

Besides, using that the Riemannian hessian of F is continuous on $\mathcal{M}$, we get

$$\langle \mathrm{Hess}\, F(x^{\mathcal{M}})d_t^{\mathrm{SQP}}, d_t^{\mathrm{SQP}} \rangle = \langle \mathrm{Hess}\, F(x^\star)d_t^{\mathrm{SQP}}, d_t^{\mathrm{SQP}} \rangle + o(\|d^{\mathrm{SQP}}\|^2),$$

where, by a slight abuse of notation, we omit the projection on $T_{x^\star}\mathcal{M}^\star$ surrounding $\mathrm{Hess}\, F(x^\star)$. Finally, we reach the following estimate:

$$F(x) \geq F(x_+) - \langle v_x, d_n^{\mathrm{SQP}} \rangle + \langle \mathrm{Hess}_{\mathcal{M}} F(x) d_t^{\mathrm{SQP}}, d_t^{\mathrm{SQP}} \rangle - \frac{1}{2} \left\langle \mathrm{Hess}\, F(x^\star) d_t^{\mathrm{SQP}}, d_t^{\mathrm{SQP}} \right\rangle + o(\mathrm{dist}_{\mathcal{M}}(x)) + o(\|d^{\mathrm{SQP}}\|^2). \tag{5.15}$$

STEP 2. We now show that $F(x + d^{\mathrm{SQP}} + d^{\mathrm{corr}}) \leq F(x) + mF'(x, d^{\mathrm{SQP}})$.

Combining (5.15) with $v_x \in \arg\max_{x \in \partial F(x)} \langle v, d_n^{\mathrm{SQP}} \rangle$ and the estimate of the directional derivative near a minimizer (5.5) yields

$$F(x_+) - F(x) - mF'(x; d^{\mathrm{SQP}}) \leq (1-m) \max_{x \in \partial F(x)} \langle v, d_n^{\mathrm{SQP}} \rangle$$
$$+ \left\langle \left((m-1) \operatorname{Hess}_{\mathcal{M}} F(x) + \frac{1}{2} \operatorname{Hess} F(x^\star) \right) d_t^{\mathrm{SQP}}, d_t^{\mathrm{SQP}} \right\rangle + o(\operatorname{dist}_{\mathcal{M}}(x)) + o(\|d^{\mathrm{SQP}}\|^2).$$

Using assumption (5.7), we reach:

$$F(x_+) - F(x) - mF'(x; d^{\mathrm{SQP}}) \leq (1-m) \max_{x \in \partial F(x)} \langle v, d_n^{\mathrm{SQP}} \rangle$$
$$+ \left(m - \frac{1}{2} \right) \left\langle \operatorname{Hess} F(x^\star) d_t^{\mathrm{SQP}}, d_t^{\mathrm{SQP}} \right\rangle + o(\operatorname{dist}_{\mathcal{M}}(x)) + o(\|d^{\mathrm{SQP}}\|^2)$$

Since $\operatorname{Hess} F(x^\star)$ is positive definite on $T_{x^\star}\mathcal{M}$ and the max term is negative by the proximal aiming property (5.4), the result holds as soon as $m \in (0, \frac{1}{2})$. □

5.3.3 *When to correct?*

The second-order correction term ensures functional decrease with the unit stepsize near minimizers. It is however somewhat costly: it requires to evaluate the smooth map (and potentially its eigenvalue decomposition) at a new point. A second look at the proof of Theorem 5.6 shows that, when

$$\|d_n^{\mathrm{SQP}}\| \geq C \|d_t^{\mathrm{SQP}} + d_r^{\mathrm{SQP}}\|$$

for some constant $C > 0$, the plain SQP step provides functional descent.

Note that this fits exactly Example 5.1: the points are exactly feasible, so that the restoration step is null. Again, this situation mirrors that of nonlinear programming; see (Bonnans et al., 2006, Prop. 17.8).

Theorem 5.8 (Avoiding second-order corrections). *Consider a function F and point $x^\star$ such that F is partly smooth at $x^\star$ relative to some manifold $\mathcal{M}^\star$, $x^\star$ is a strong minimizer:*

$$0 \in \operatorname{ri} \partial F(x^\star) \quad \textit{and} \quad \operatorname{Hess} F(x^\star) \succ 0,$$

and assume that F meets the proximal aiming property at $x^\star$ (property 5.2) and is (b)*-regular at $x^\star$ along $\mathcal{M}$ (property 5.1). Let C denote any positive constant. Then for all iterations such that*

$$\|d_n^{\mathrm{SQP}}(x_k)\| \geq C \|d_t^{\mathrm{SQP}}(x_k) + d_r^{\mathrm{SQP}}(x_k)\|,$$

the Armijo rule holds with unit stepsize:

$$F(x + d^{\mathrm{SQP}}(x)) \leq F(x) + mF'(x; d^{\mathrm{SQP}}(x)).$$

Proof. The proof follows closely the above developments, we omit it. □

5.4 VALIDITY OF LOCAL STRUCTURE

5.4.1 *A structure-based implementable stopping criterion*

Let $\bar{x}$ be a minimizer of F. Fermat's rule gives necessary optimality conditions:

$$0 \in \partial F(\bar{x}). \tag{5.16}$$

When F is partly smooth at $\bar{x}$ relative to some manifold $\mathcal{M}$, this implies:

$$\begin{aligned} \bar{x} &\in \mathcal{M} \\ \operatorname{grad} F(\bar{x}) &= 0 \\ 0 &\in \operatorname{proj}_{N_{\bar{x}}\mathcal{M}} \partial F(\bar{x}) \end{aligned}$$

The first condition is ensured by partial smoothness; the other two follow from projecting Fermat's rule (5.16) on the tangent and normal spaces. Recall that, the projection of the subdifferential on the tangent space is a unique vector: the Riemannian gradient of the restriction of F to $\mathcal{M}$, denoted $\operatorname{grad} F(\bar{x})$.

A practical convergence criterion is a quantity that continuously goes to zero as the point which it qualifies goes to a minimizer. The two last items above are problematic for points near $\bar{x}$ but not necessarily on $\mathcal{M}$: the Riemannian gradient of F is not defined at $x \in \mathbb{R}^n \setminus \mathcal{M}$, and the subdifferential dimensionality changes; recall e.g., Fig. 1.1.

SUBDIFFERENTIAL SMOOTH EXTENSION. We propose to deal with this difficulty by using information on the smooth substructure. We introduce here a smooth extension of ∂F relative to some structure manifold $\mathcal{M}$, that is, a continuous set-valued map that matches ∂F on $\mathcal{M}$.

Definition 5.1 (Subdifferential smooth extension). Consider a function F and a point $\bar{x}$ at which F is partly smooth relative to manifold $\mathcal{M}$. The set-valued application $\partial^{\mathcal{M}} F : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is a *smooth extension of the subdifferential* relative to $\mathcal{M}$ at $\bar{x}$ if

- $\partial^{\mathcal{M}} F(x) = \partial F(x)$ for all x near $\bar{x}$ on $\mathcal{M}$, and
- $\partial^{\mathcal{M}} F$ is continuous on a neighborhood of $\bar{x}$ in $\mathbb{R}^n$.

These two conditions seem strong, but it is straightforward to build valid smooth extensions of subdifferentials for the maximum of smooth functions, the maximum eigenvalue of a parameterized function, or ℓ_1-regularized functions. We propose such extensions in Examples 5.2–5.4.

Remark 5.3 (Relations with the literature). In a similar fashion, Liu et al. (2019) study enlargement of subdifferentials that incorporate information on the nonsmooth function along normal directions. The studied enlargements are smooth set-valued maps that depend on a parameter ε, such that the maps converge to the usual subdifferential as ε goes to zero. This is the main difference with the extensions discussed here, which instead depend on a structure manifold $\mathcal{M}$. These structure-aware enlargements are intended to replace in bundle methods the usual ε-subdifferential, as an attempt to make the structure identification of the $\mathcal{VU}$-bundle algorithm more stable in practice. △

APPROXIMATE OPTIMALITY CONDITION. We thus consider the following approximate optimality conditions for a pair $(x, \mathcal{M})$:

$$\begin{cases} \|h(x)\| & \leq \varepsilon \\ \operatorname{dist}\left(0, \partial^{\mathcal{M}} F(x)\right) & \leq \varepsilon \end{cases} \tag{5.17}$$

where h defines $\mathcal{M}$ locally. A point x that verifies Eq. (5.17) is such that there exists a point $\bar{x} \in \mathcal{M}$ at distance $\mathcal{O}(\varepsilon)$ of x such that $\operatorname{dist}(0, \partial F(\bar{x})) \leq \mathcal{O}(\varepsilon)$. This last condition implies that $\bar{x}$ is almost optimal: $F'(\bar{x}, d) \geq -\mathcal{O}(\varepsilon)\|d\|$.

EXAMPLES. We now provide valid smooth extensions of the subdifferential on three functions discussed in this thesis.

Example 5.2 (Maximum of smooth functions). We consider the maximum of smooth functions $F(x) = \max \circ c$. At a point x with smooth substructure $\mathcal{M}_I^{\max}$, upon satisfaction of the transversality condition Eq. (4.4), the subdifferential writes:

$$\partial F(x) = \operatorname{Conv}\left\{\nabla c_i(x) : i \in I\right\}.$$

A natural smooth extension, defined for *any* $x \in \mathbb{R}^n$, is:

$$\partial^{\mathcal{M}_I^{\max}} F(x) = \operatorname{Conv}\left\{\nabla c_i(x) : i \in I\right\}. \qquad \square$$

Example 5.3 (Maximum eigenvalue). We consider the maximum eigenvalue of a smoothly parameterized matrix. At a point x with smooth substructure $\mathcal{M}_r^{\lambda_{\max}}$, upon satisfaction of the transversality condition Eq. (4.4), the subdifferential writes:

$$\partial F(x) = \{\mathrm{D}\, c^*(x) \cdot E(x) Z E(x)^\top, \text{ for } Z \in \mathbb{S}_r, \operatorname{tr}(Z) = 1\},$$

See Shapiro and Fan (1995) for the smooth basis construction.

where $E(x) \in \mathbb{R}^{m \times r}$ denotes a *smooth* orthonormal basis of the eigenspace corresponding to the r largest eigenspaces. Therefore, a natural smooth extension, defined for *any* $x \in \mathbb{R}^n$, is:

$$\partial^{\mathcal{M}_r^{\lambda_{\max}}} F(x) = \{\mathrm{D}\, c^*(x) \cdot E(x) Z E(x)^\top, \text{ for } Z \in \mathbb{S}_r, \operatorname{tr}(Z) = 1\}. \qquad \square$$

Example 5.4 (ℓ_1-regularization). Finally, we consider an ℓ_1 regularized function $F(x) = f(x) + \lambda\|x\|_1$. At point x with structure $\mathcal{M}_I^{\ell_1} = \{x \in \mathbb{R}^n : x_i = 0, i \in I\}$, the subdifferential writes $\partial F(x) = \nabla f(x) + \lambda \partial \|\cdot\|_1(x)$, where

$$[\partial \|\cdot\|_1(x)]_i = \partial |\cdot|(x_i) = \begin{cases} -1 & \text{if } x_i < 0 \\ [-1, 1] & \text{if } x_i = 0 \cdot \\ 1 & \text{if } x_i > 0 \end{cases}$$

A natural smooth extension, defined at *any* $x \in \mathbb{R}^n$, is $\partial^{\mathcal{M}_I^{\ell_1}} F(x) = \nabla f(x) + \lambda \partial^{\mathcal{M}_I^{\ell_1}} \|\cdot\|_1(x)$, where the smooth extension $\partial^{\mathcal{M}_I^{\ell_1}} \|\cdot\|_1(x)$ is defined coordinate-wise as follows:

$$\left[\partial^{\mathcal{M}_I^{\ell_1}} \|\cdot\|_1(x)\right]_i = \begin{cases} [-1, 1] & \text{if } i \in I \\ \partial |\cdot|(x_i) & \text{else} \end{cases}. \qquad \square$$

Remark 5.4 . We report the subdifferential enlargements proposed by Liu et al. (2019) on Examples 5.2 and 5.4.

- When $F(x) = \max \circ c$, the proposed subdifferential enlargement is (see 5.4)

$$\delta_\varepsilon F(x) = \operatorname{Conv}\{\nabla c_i(x) : i \in I_\varepsilon(x)\}, \quad \text{where} \quad I_\varepsilon(x) = \{i : x_i \geq f(x) - \varepsilon\}.$$

- When $F(x) = f(x) + \lambda\|x\|_1$, the proposed subdifferential enlargement is (see 6.2.2)

$$\partial_\varepsilon F(x) = \nabla f(x) + \lambda \operatorname{Conv}\{s \in \mathbb{R}^n : s_i = \pm 1, \langle s, x\rangle \geq \|x\|_1 - \varepsilon\}. \qquad \triangle$$

5.4.2 *Towards structure screening: detecting nearby normal descent direction*

In this subsection, we look more closely to qualified points. We discuss the geometrical meaning of this qualification condition in the structured setting and show that qualification can be detected locally. This will be relevant in the manifold selection of the forthcoming Algorithm 5.2.

NEAR QUALIFIED POINTS. Most guarantees in this thesis are given for *qualified* points, that is point that satisfy a strengthened version of Fermat's rule:

$$0 \in \operatorname{ri} \partial F(\bar{x}).$$

When F is partly smooth at $\bar{x}$ relative to some manifold $\mathcal{M}$, this equation formalizes the fact that the function strictly increases in all directions normal to $\mathcal{M}$ at $\bar{x}$: $F'(\bar{x}; d) > 0$ for all $d \in N_x\mathcal{M}$.

This property actually holds at all points near $\bar{x}$ on $\mathcal{M}$, as proved by Daniilidis et al. (2006, Lemma 20): they show that there exists a neighborhood $\mathcal{N}_{\bar{x}}$ of $\bar{x}$ such that, for all $x \in \mathcal{N}_{\bar{x}} \cap \mathcal{M}$,

$$\operatorname{proj}_{\partial F(x)}(0) \in \operatorname{ri} \partial F(x). \tag{5.18}$$

This interprets as $F'(x; d) > 0$, for all $d \in N_x\mathcal{M}$ and $x \in \mathcal{N}_{\bar{x}} \cap \mathcal{M}$.

With a smooth extension of the subdifferential relative to $\mathcal{M}$, we can now extend this property to a neighborhood of $\bar{x}$ *on* $\mathbb{R}^n$, which we formalize in the next lemma. This gives an implementable criterion to detect structure manifolds which admit only ascent normal directions, from points that are *near*, but *not on*, the structure manifold. This provides a useful tool for the structure identification process, discussed in Section 5.6.1. Note that the following lemma considers a point $\bar{x}$ with structure $\mathcal{M}$ that admits only normal ascent directions. In particular, $\bar{x}$ is not necessarily a minimizer of F *on* the manifold.

Lemma 5.9 (Persistence of normal ascent). *Consider a function F, a point $\bar{x}$ and a manifold $\mathcal{M}$ such that F is prox-regular at $\bar{x}$, and F is partly smooth at $\bar{x}$ relative to $\mathcal{M}$. If*

$$\operatorname{proj}_{\partial F(\bar{x})}(0) \in \operatorname{ri} \partial F(\bar{x})$$

Then, for all x near $\bar{x}$ on $\mathbb{R}^n$, there holds, equivalently,

$$\begin{aligned} 0 &\in \operatorname{ri} \operatorname{proj}_{N_x\mathcal{M}_x}(\partial^{\mathcal{M}} F(x)) \\ \operatorname{proj}_{\partial^{\mathcal{M}} F(x)}(0) &\in \operatorname{ri} \partial^{\mathcal{M}} F(x) \end{aligned} \tag{5.19}$$

This condition captures for F, in the input space $\mathbb{R}^n$, the same geometrical property as property 4.1 did for g, in the intermediate space $\mathbb{R}^m$.

Proof. We follow closely the proof of Daniilidis et al. (2006, Lemma 20). With a basis of $N_x\mathcal{M}_x$ that depends smoothly on x, we first construct a continuous function $x \mapsto \psi_x$ such that $\psi_x : N_x\mathcal{M}_x \to \mathbb{R}^p$ is a linear bijection. Now, the multifunction $G : \mathbb{R}^n \rightrightarrows \mathbb{R}^p$ defined by $G(x) = \psi_x(\partial^{\mathcal{M}} F(x) - \mathrm{proj}_{\partial^{\mathcal{M}} F(x)}(0))$ is continuous around $\bar{x}$ by continuity of ψ_x and $\partial^{\mathcal{M}} F(x)$. Besides, it satisfies $\mathrm{proj}_{\partial^{\mathcal{M}} F(x)}(0) \in \mathrm{ri}\, \partial^{\mathcal{M}} F(x) \Leftrightarrow 0 \in \mathrm{int}\, G(x)$. The argument of Daniilidis et al. (2006, Lemma 20) applies to this setting and yields the conclusion. □

REMOTE DETECTION OF QUALIFIED POINTS AND SCREENING. We now discuss how to compute, on our examples, the distance from zero to the smooth subdifferential extension (5.17), and how this can at the same time provide whether Eq. (5.19) holds or not.

Example 5.5 (Maximum of smooth functions). Building on Example 5.2, the distance from zero to the subdifferential is obtained by solving

$$\min_{\alpha \in \Delta_I} \left\| \sum_{i \in I} \alpha_i \nabla c_i(x) \right\|,$$

where $\Delta_I = \{\alpha \in \mathbb{R}^{|I|} : \alpha_i \geq 0 \text{ for all } i \in I, \sum_{i \in I} \alpha_i = 1\}$ denotes the simplex with coordinates I.

Under transversality at x, condition (5.19) holds when the minimizer $\alpha^\star$ belongs to the interior of the simplex, that is when $\alpha_i^\star > 0$ for all $i \in I$. □

Example 5.6 (Maximum eigenvalue). Building on Example 5.3, the distance from zero to the subdifferential is obtained by solving

$$\min_{Z \in \mathrm{Sp}_r} \left\| \mathrm{D}\, c(x)^* \cdot EZE^\top \right\|$$

where $E \in \mathbb{R}^{m \times r}$ is an orthonormal basis of the eigenspace associated with r largest eigenvalues of $c(x)$, and $\mathrm{Sp}_r = \{Z \in \mathbb{S}_r, Z \succeq 0, \mathrm{trace}(Z) = 1\}$ denotes the spectraplex of dimension r.

Under transversality at x, condition (5.19) holds when the minimizer $Z^\star$ belongs to the interior of the spectraplex, that is when all its eigenvalues are positive. □

Example 5.7 (ℓ_1-regularization). Building on Example 5.4, the distance from zero to the subdifferential is obtained by solving

$$\min_{v \in \mathbb{R}^n} \|\nabla f(x) + \lambda v\|_2^2 = \sum_{i \in I} \left(\max(|\nabla f_i(x)| - \lambda), 0\right)^2 + \sum_{i \notin I} \left(\nabla f_i(x) + \lambda\, \mathrm{sgn}(x_i)\right)^2,$$

where $\mathrm{sgn}(t)$ denotes the sign of t, and $\nabla f_i(x)$ denotes the i-th coordinate of $\nabla f(x)$. We have assumed that $\mathcal{M}_I^{\ell_1}$ contains all null coordinates of x.

This time, condition (5.19) holds when the minimizer $v^\star$ satisfies $v_i \in (-1, 1)$ for $i \in I$. □

Remark 5.5 (Relations with the literature). The above mentioned problems allow to quantify the optimality of a pair $(x, \mathcal{M})$ and to detect the existence of normal ascent directions. We point out below similarities with three structure detection schemes proposed in the literature.

For bundle methods, Mifflin and Sagastizábal (2005) detects structure by observing the solution of a particular quadratic program constrained on the simplex (denoted γ-QP). In particular, it is the activity of this solution, that is

indices of the non-null coordinates, that allows to estimate the local smooth substructure.

The spectral bundle method of Helmberg et al. (2014) detects structure by considering the minimizer of a quadratic program constrained on the spectraplex. In particular, their guess for the local smooth substructure is based on the rank of the minimizing matrix.

Recently, Bertrand et al. (2022) proposed a working set approach to minimizer functions regularized with separable functions. The working set contains, as $\mathcal{M}_I^{\ell_1}$, a subset of the coordinates. They are ranked and chosen using a score function that quantifies a distance to optimality akin to (5.18); see Bertrand et al. (2022, Eq. 5). △

5.5 WHAT IS STILL MISSING

In the previous sections, we have gathered partial guarantees on ingredients of Algorithm 5.1. We quickly outline here the remaining questions and the approach we tried to follow.

The above developments show that a linesearch SQP method may retain the fast quadratic rate near minimizers, assuming identification of the optimal manifold.

The main difficulty consists in combining the identification procedure with the linesearch. We list below some of the questions we faced, and the approach we adopted, again strongly inspired by developments of nonlinear programming (Bonnans et al., 2006, Chap. 17).

- At each iteration, we would expect the SQP step to be a descent direction for F, that is: $F'(x, d^{\mathrm{SQP}}) < 0$. This is a necessary condition to execute a linesearch on F. This condition could be satisfied either by the manifold selection scheme, or by a careful adjustment of the SQP step, following Spellucci (1998, Eq. 14).

- A second difficulty is to show that the linesearch stepsize cannot be arbitrarily small.[1] For example, this happens on the one dimensional function $\varphi(x) = \max(-x, 0.5x, 2x - 3)$ if one considers the SQP step from point $t < 0$ arbitrarily close to zero, relative to the smooth substructure $\mathcal{M} = \{2\}$. Here, the trouble comes from the identification of the (non-optimal) structure $\mathcal{M} = \{2\}$ rather than the closer (optimal) $\mathcal{M} = \{0\}$.

- With the two above elements, we would expect to be able to extract properties on the limit points of the sequence of iterates. Indeed, summing the Armijo rule provided in Eq. (5.3) yields for lower bounded objectives:

$$\sum_{k \geq 0} \alpha_k F'(x_k; d_k^{\mathrm{SQP}}) \leq F(x_0) - \min F.$$

  If the linesearch step is bounded away from zero, we get $\lim_k F'(x_k; d_k^{\mathrm{SQP}}) = 0$. Is this informative enough to obtain that any limit point $\bar{x}$ satisfies $0 \in \partial F(\bar{x})$? Lemma 5.5 seems to imply that, near minimizers and relative to the optimal manifold, the above limit would imply that $\bar{x}$ lies on the manifold and that it is critical for F restricted to the manifold ($d_t^{\mathrm{SQP}} = 0$). Does this extend away from minimizers?

1 One would also expect that the linesearch terminates after a finite number of trials. These two aspects are closely related.

These questions deal with the behavior of nonsmooth functions away from minimizers. We believe that providing grounded answers requires to go beyond the local description provided by e.g., partial smoothness, by using global information on nonsmoothness. In this respect, the recent line of work developed by Davis et al. (2021) seems promising: they provide a *global* description of the nonsmoothness manifolds in the form of a Whitney stratification of the domain of the function, with expressive local geometrical properties such as properties 5.1 and 5.2. More work and insight will be necessary to provide precise answers to these questions.

5.6 NUMERICAL ILLUSTRATIONS

In this section, we propose an algorithm that integrates the globalization guarantees, and ideas presented in this chapter. We detail here the identification heuristic we rely on, and then we illustrate the behavior of this global algorithm on several problems, including those of Chapter 4.

5.6.1 *Identification heuristic and proposed algorithm*

We propose an identification scheme that combines the ideas of the local Newton method Algorithm 4.1, and the detection of normal ascent directions, discussed in Section 5.4.2.

Building on Chapter 4, we use the proximity operator as a structure detector: at point x and for step γ, we consider the manifold $\mathcal{M}$ relative to which g is partly smooth at point $\operatorname{prox}_{\gamma g}(x)$. We use here the prox as a *filter* for relevant information. We remark that varying γ and collecting $\mathcal{M} \ni \operatorname{prox}_{\gamma g}(x)$ provides only a subset of the possible substructure manifolds. Note also that for γ near 0, the dimension of the structure manifold $\mathcal{M}$ increases as γ reduces, to generally reach $\mathcal{M} = \mathbb{R}^n$ for γ small enough.

We propose to select the candidate manifold in this subset of manifolds, picking the one with the smaller dimension, that meets the normal ascent condition (5.19). Let us illustrate this heuristic on a simple example.

Example 5.8 (Manifold selection for $\max \circ c$). Here is an instance of the proposed identification heuristic. At some point, say we have

$$c(x) \approx [100, 3, 23, 21, -12].$$

Then the proximal detection tool suggests the following manifolds, ranked by increasing dimension:

$$I = \{1,2,3,4,5\}, \quad I = \{2,3,4,5\}, \quad I = \{2,4,5\}, \quad I = \{2,5\}, \quad I = \{5\}$$

Removing the manifolds $\mathcal{M}_I^{\max}$ whose feasibility is too big leaves only a subset of these manifolds. It remains to evaluate Eq. (5.19) successively to find the manifold with smallest dimension that meets this normal ascent condition. □

Note that this heuristic requires only one evaluation of c and Jac_c, for all these computations.

PROPOSED ALGORITHM. The proposed algorithm is summarized in Algorithm 5.2. It includes the described heuristic and the stopping criterion presented in Section 5.4.1.

Algorithm 5.2: Global Newton algorithm (heuristic)

Require: Set $x_0 \in \mathbb{R}^n$, $\mathcal{M}_0 = \mathbb{R}^n$, $\gamma_{\text{init}} > 0$, $m \in (0, 1/2)$, $\varepsilon > 0$. For example, take $m = 10^{-3}$ $\gamma_{\text{init}} = 10^2$ and $\varepsilon = 10^{-13}$.

repeat
  Form $\gamma \mapsto \mathcal{M}^\gamma \ni \text{prox}_{\gamma g} \circ c(x_k)$ ▷ Structure detection
  $\gamma_{k+1} = \gamma_{\text{init}}$
  while $\mathcal{M}^{\gamma_{k+1}}$ does not satisfy the normal ascent property Eq. (5.19), and is not the last manifold **do**
    Reduce γ_{k+1} to the greater value changing $\mathcal{M}^{\gamma_{k+1}}$
  end while
  Set $\mathcal{M}_{k+1} = \mathcal{M}^{\gamma_{k+1}}$ ($\mathbb{R}^n$ if no satisfactory manifold was found)
  Select a smooth extension $\tilde{F}_k$ and a matrix M_k
  Compute $d_k^{\text{SQP}}(x_k)$ by solving (5.2) ▷ Structure exploitation
  Obtain x_{k+1} from the linesearch procedure Algorithm 5.3.
until $\|h_k(x_k)\| \leq \varepsilon$ and $\text{dist}\left(0, \partial^{\mathcal{M}_k} F(x_k)\right) \leq \varepsilon$

We provide here some implementation details. The backtracking linesearch selects the higher step of the form $\alpha = (1/2)^i$, for some integer i. We compute a truncated SQP step at each iteration, following Bonnans et al. (2006, Alg. TSQP, p. 305). This is done by taking M as the exact Hessian of the Lagrangian, and solving problem (5.2) exactly, or until a negative curvature direction is found. We handle the Maratos effect by adding the second-order correction, defined in Eq. (5.6), only at iterations where it is necessary. Specifically, we use the approach described in Chauvier et al. (2003, p. 20), summarized in Algorithm 5.3. The algorithm stops when the pair $(x_k, \mathcal{M}_k)$ meets the optimality conditions Eq. (5.17).

5.6.2 *Numerical experiments*

TEST PROBLEMS. We consider the same problems and similar instances as in the numerical experiments of Section 4.5, which we detail below.

First, we take the celebrated `MaxQuad` instance, of the form $F(x) = \max_{i=1,\dots,m}(c_i(x))$, where $n = 10$, $m = 5$ and each c_i is quadratic convex, making the whole function F convex (Bonnans et al., 2006, p. 153). For this instance, the optimal manifold is $\mathcal{M}_I^{\max}$ with $I = \{2, 3, 4, 5\}$.

Second, we consider the problems `F3d-Uv`, which also write as maximum of quadratics with $n = 3$, and $m = 4$; see Mifflin and Sagastizábal (2005, Sec. 6).[2] The mappings are defined as

$$c(x) = \begin{pmatrix} \frac{1}{2}\left(x_1^2 + x_2^2 + \gamma x_3^2\right) - x_2 - x_3 \\ x_1^2 - 3x_1 \\ x_2 \\ x_3 \end{pmatrix} - \beta^v,$$

2 We report slightly corrected mappings – see c_4 and β_4^2, so that the minimizers and optimal function value of the problem match the values announced in Table 1.

Name	n	$\mathcal{M}^\star$	$\dim \mathcal{M}^\star$	$F(x^\star)$	$x^\star$	starting point
`MaxQuad`	10	$\mathcal{M}^{\max}_{\{2,3,4,5\}}$	7	$-^\dagger$	$-^\dagger$	$\mathbb{1}$
`F3d-U0`	3	$\mathcal{M}^{\max}_{\{1,2,3,4\}}$	0	0	$(1,0,0)$	$x^\star + (100, 33, -100)$
`F3d-U1`	3	$\mathcal{M}^{\max}_{\{1,3,4\}}$	1	0	$(0,0,0)$	$x^\star + (100, 33, -100)$
`F3d-U2`	3	$\mathcal{M}^{\max}_{\{1,3\}}$	2	0	$(0,0,10)$	$x^\star + (100, 33, -100)$
`F3d-U3`	3	$\mathcal{M}^{\max}_{\{1\}}$	3	0	$(0,1,10)$	$x^\star + (100, 33, -100)$
`eigmax`	25	$\mathcal{M}^{\lambda_{\max}}_{3}$	20	–	–	$x^\star + 0.1\mathbb{1}$

Table 5.1: Description of the experiment problems. $\mathbb{1}$ denotes the vector of ones. For the approximate optimal point and functional value ($-^\dagger$), see (Bonnans et al., 2006, p. 153).

where β^ν, for $\nu \in \{0, \dots, 3\}$ denotes the $(\nu+1)$th column of

$$\beta = \begin{pmatrix} 0.5 & 0 & -5 & -5.5 \\ -2 & 10 & 10 & 10 \\ 0 & 0 & 0 & 11 \\ 0 & 0 & 20 & 20 \end{pmatrix}$$

Finally, we consider maximal eigenvalue problems of the form

$$\min_{x \in \mathbb{R}^n} \lambda_{\max}\left(A_0 + \sum_{i=1}^{n} x_i A_i\right).$$

We take $n = 25$ and we generate randomly $n+1$ symmetric matrices of size 50. For this instance, the multiplicity of the maximum eigenvalue at the minimizer is $r = 3$.

All problems data are described in Table 5.1. We report the results in Figures 5.5–5.10.

ALGORITHMS & IMPLEMENTATION DETAILS. As in Chapter 4, we solve the problems with the nonsmooth BFGS (Lewis and Overton, 2013), the Gradient Sampling method (Burke et al., 2020), and Algorithm 5.2. In addition, we also solve the problems with the $\mathcal{VU}$-algorithm of Mifflin and Sagastizábal (2005). As suggested in the paper, the algorithm does not make use of second-order information in the $\mathcal{U}$-Newton steps, but rather approximates it from first-order information with a quasi-Newton BFGS scheme. We stop these algorithms after a fixed number of iterations, or when their termination criteria are met, when they have one. All the algorithms are implemented in Julia (Bezanson et al., 2017); experiments may be reproduced using the code available online.[3]

ITERATION COSTS PER ALGORITHM. At this stage, we summarize the iteration costs of each algorithm. The nonsmooth BFGS and the $\mathcal{VU}$-bundle method require a changing number of function values and subgradients per iteration. Indeed, the former method employs a linesearch, which may require several trials before finding a suitable point, and the latter refines a cutting plane model of the function, which requires several oracle calls before the model is precise enough to take a serious step. In contrast, the Gradient Sampling requires a

3 See `https://codeberg.org/GillesBareilles/GlobalCompositeNewton.jl` for Algorithm 5.2 and `https://github.com/GillesBareilles/NonSmoothSolvers.jl` for the baselines.

Algorithm 5.3: Efficient linesearch with second-order correction.

if $F(x_k + d_k^{\mathrm{SQP}}) \leq F(x_k) + mF'(x_k, d_k^{\mathrm{SQP}})$ **then**
  $x_{k+1} = x_k + d^{\mathrm{SQP}}$
else
  if $\|d_n^{\mathrm{SQP}}\| \leq C\|d_t^{\mathrm{SQP}}\|$ **then**
    Compute the second–oder correction d^{corr}
    if $F(x_k + d^{\mathrm{SQP}} + d^{\mathrm{corr}}) \leq F(x + d^{\mathrm{SQP}})$ **then**
      Do an arc-search along $\alpha \mapsto x_k + \alpha d^{\mathrm{SQP}} + \alpha^2 d^{\mathrm{corr}}$
    else
      Do a line-search along $\alpha \mapsto x_k + \alpha d^{\mathrm{SQP}}$
    end if
  else
    Do a line-search along $\alpha \mapsto x_k + \alpha d^{\mathrm{SQP}}$
  end if
end if

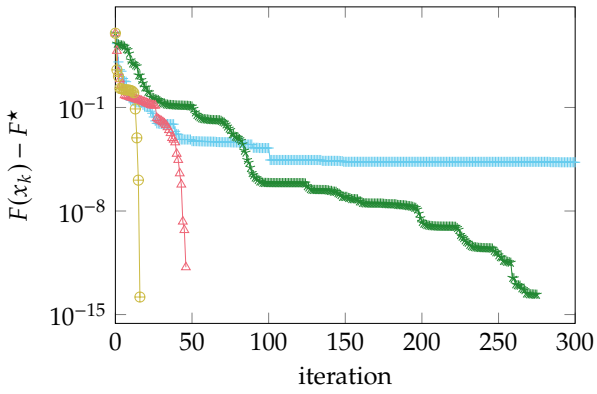
constant number of black-box oracle calls per iteration, set to twice the dimension of the optimization space $\mathbb{R}^n$ in the experiments. Algorithm 5.2 requires information of a different nature: the value of $c(x)$ and its (partial) eigenvalue decomposition, the Jacobian $\mathrm{Jac}_h(x)$, and the hessian of the Lagrangian for the smooth reduced problem obtained from identification. Besides, the linesearch requires evaluating F at several points. Comparison with respect to iterations is therefore not completely fair.

But the running time of the algorithms is not entirely satisfactory as a progress measure: obtaining implementations of comparable qualities requires different implementation effort, depending on the algorithms complexities.

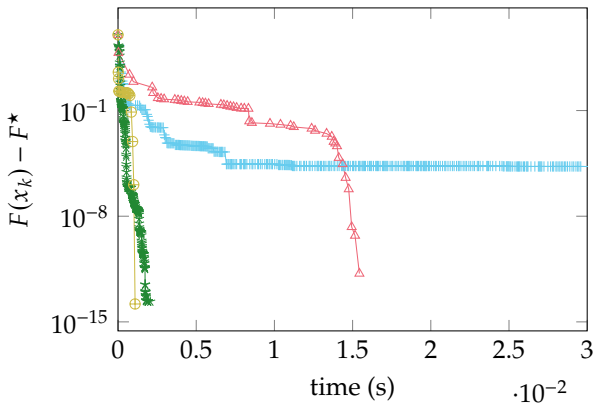
ILLUSTRATIONS. We report on Fig. 5.5a to 5.10a the evolution of the suboptimality versus iterations, and on Fig. 5.5b to 5.10b the evolution of suboptimality versus time. On the `MaxQuad` problem, Algorithm 5.2 obtains the correct manifold at iteration 13. Three iterations later, the iterates $(x_k, \mathcal{M}_k)$ satisfy the optimality condition (5.17) with $\varepsilon = 2 \cdot 10^{-13}$. The linesearch selects the unit stepsize during these iterations, as hinted by the quadratic convergence rate displayed in Fig. 5.5. On the maximum eigenvalue problem, the proposed algorithm obtains the right manifold at iteration 8, and the unit stepsize at iteration 13 until convergence. Similar remarks hold for Algorithm 5.2 on the `F3d-U`ν instances.

The serious steps of the $\mathcal{VU}$-algorithm converge as well very fast on all instances, even with only approximate second-order information. We also note that the algorithm succeeds in detecting the optimal manifold dimension in each instance, except for the `F3d-U0` problem.

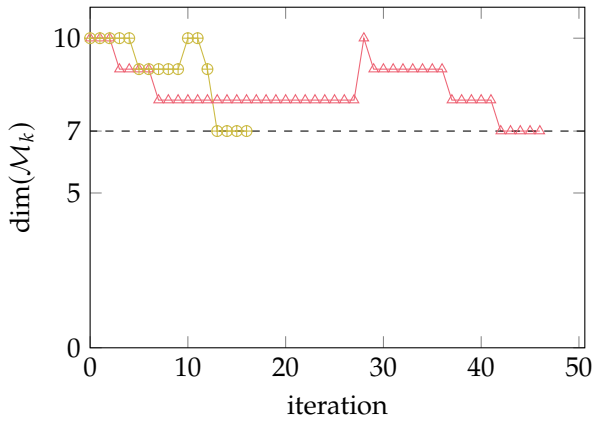
We report in Figs. 5.5d and 5.10d the optimality condition Eq. (5.17) and the normal ascent test Eq. (5.19) at the last point of each algorithm, relative to the optimal manifold for both problems. Only the $\mathcal{VU}$-algorithm and Algorithm 5.2 get a really high precision, and our implementation of the $\mathcal{VU}$-algorithm appears to favor feasibility over optimality along structure manifolds. We also note that, across all problems, the final point of each algorithm meets the normal ascent condition Eq. (5.19), although some points may be away from the minimizer.



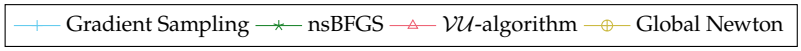
(a) Suboptimality vs iterations



(b) Suboptimality vs time (s)



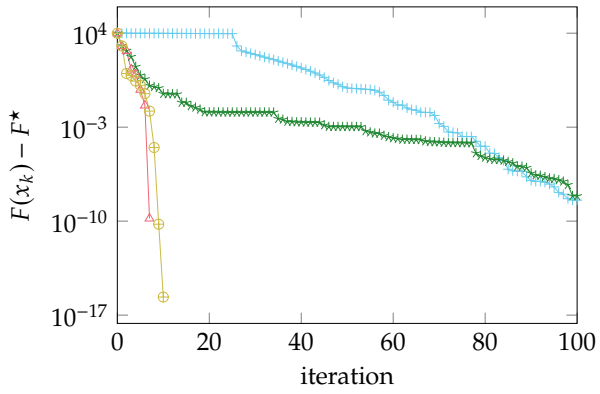
(c) Manifold dimension vs iterations



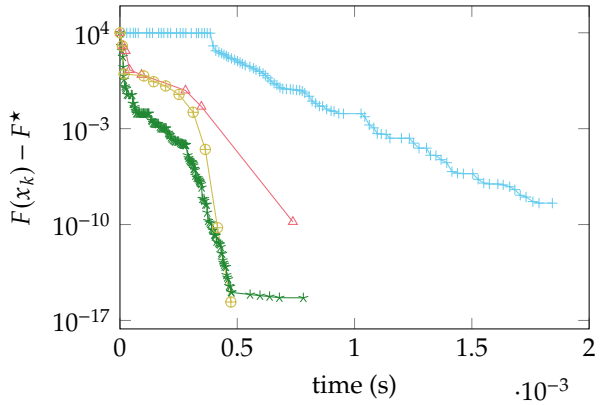
Algorithm	$\|h(x)\|$	$\operatorname{dist}(\partial^{\mathcal{M}} F(\bar{x}), 0)$	$\operatorname{proj}_{\partial F(\bar{x})}(0) \in \operatorname{ri} \partial F(\bar{x})$
Gradient Sampling	$1.08 \cdot 10^{-3}$	$2.2 \cdot 10^{-4}$	true
nsBFGS	$3.29 \cdot 10^{-11}$	$9.31 \cdot 10^{-9}$	true
$\mathcal{VU}$-algorithm	$5.98 \cdot 10^{-16}$	$7.67 \cdot 10^{-6}$	true
GlobalNewton	$2.95 \cdot 10^{-14}$	$2.06 \cdot 10^{-14}$	true

(d) Optimality conditions (5.17) and (5.19) at the final point of the algorithms, relative to the optimal manifold $\mathcal{M}^{\max}_{\{2,3,4,5\}}$.

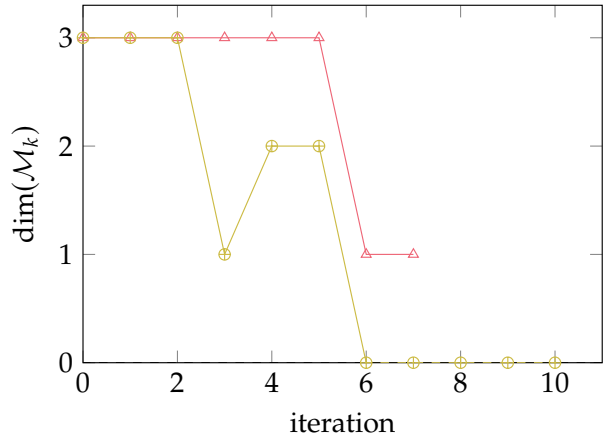
Figure 5.5: `MaxQuad` problem



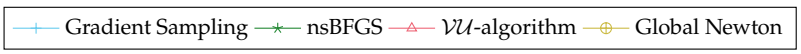
(a) Suboptimality vs iterations



(b) Suboptimality vs time (s)



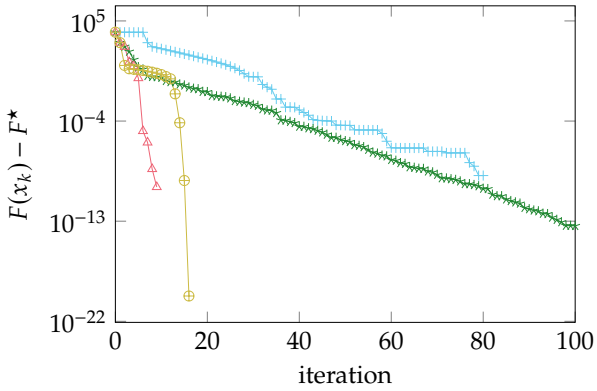
(c) Manifold dimension vs iterations



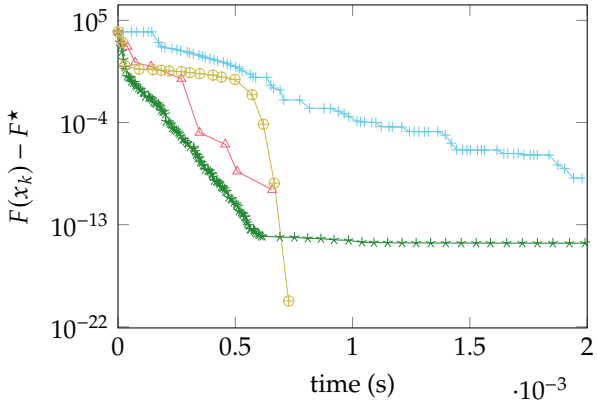
Algorithm	$\|h(x)\|$	$\operatorname{dist}(\partial^{\mathcal{M}} F(\bar{x}), 0)$	$\operatorname{proj}_{\partial F(\bar{x})}(0) \in \operatorname{ri} \partial F(\bar{x})$
Gradient Sampling	$4.43 \cdot 10^{-9}$	$1.3 \cdot 10^{-16}$	true
nsBFGS	$1.89 \cdot 10^{-15}$	$1.39 \cdot 10^{-16}$	true
$\mathcal{VU}$-algorithm	$2.1 \cdot 10^{-10}$	$2.11 \cdot 10^{-16}$	true
GlobalNewton	$2.49 \cdot 10^{-16}$	$8.33 \cdot 10^{-17}$	true

(d) Optimality conditions (5.17) and (5.19) at the final point of the algorithms, relative to the optimal manifold $\mathcal{M}^{\max}_{\{1,2,3,4\}}$. Note that the optimal manifold has dimension 0, it reduces to $\{x^\star\}$ near $x^\star$. This explains the fast rate of the $\mathcal{VU}$-algorithm, independent of the poor quality of the $\mathcal{U}$ space and Newton step on this example.

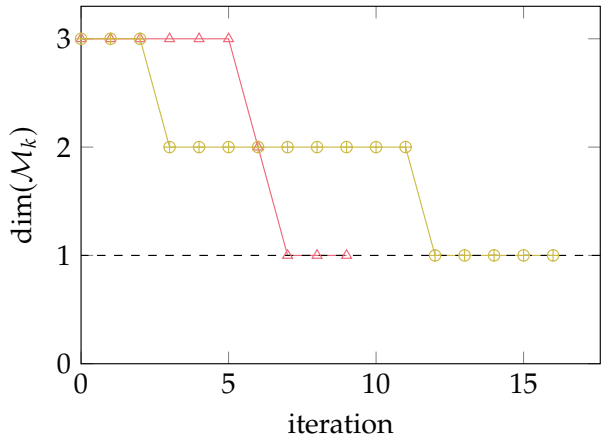
Figure 5.6: `F3d-U0` problem



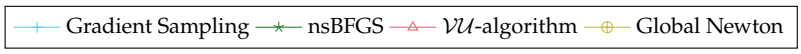
(a) Suboptimality vs iterations



(b) Suboptimality vs time (s)



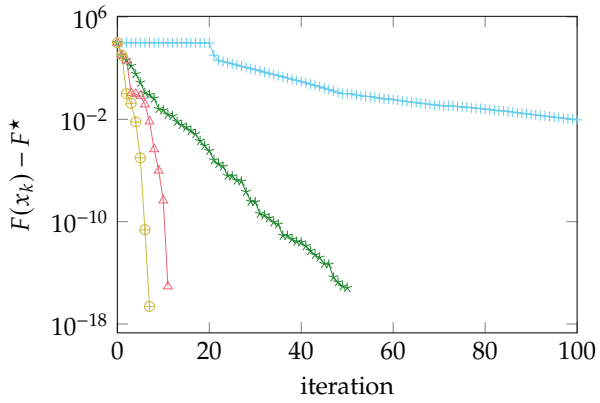
(c) Manifold dimension vs iterations



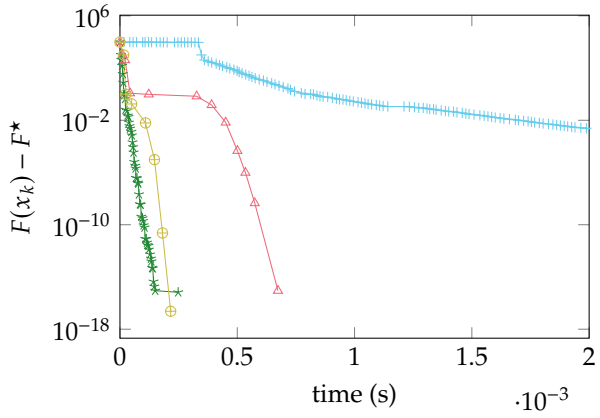
Algorithm	$\|h(x)\|$	$\operatorname{dist}(\partial^{\mathcal{M}} F(\bar{x}), 0)$	$\operatorname{proj}_{\partial F(\bar{x})}(0) \in \operatorname{ri} \partial F(\bar{x})$
Gradient Sampling	$2.26 \cdot 10^{-9}$	$5.4 \cdot 10^{-7}$	true
nsBFGS	$1.05 \cdot 10^{-15}$	$3.64 \cdot 10^{-8}$	true
$\mathcal{VU}$-algorithm	$2.44 \cdot 10^{-14}$	$8.84 \cdot 10^{-6}$	true
GlobalNewton	$2.83 \cdot 10^{-20}$	$7.67 \cdot 10^{-15}$	true

(d) Optimality conditions (5.17) and (5.19) at the final point of the algorithms, relative to the optimal manifold $\mathcal{M}^{\max}_{\{1,3,4\}}$.

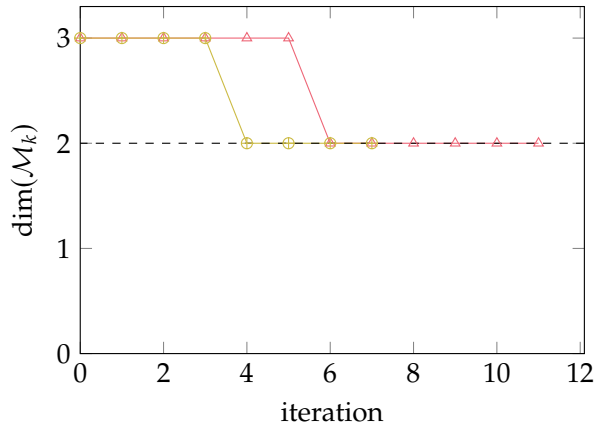
Figure 5.7: `F3d-U1` problem



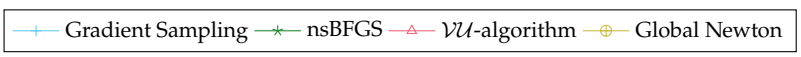
(a) Suboptimality vs iterations



(b) Suboptimality vs time (s)



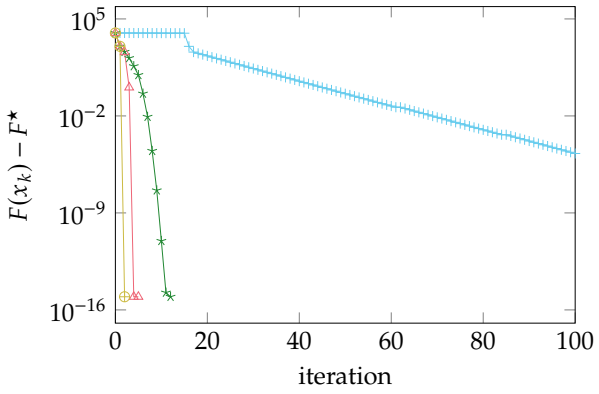
(c) Manifold dimension vs iterations



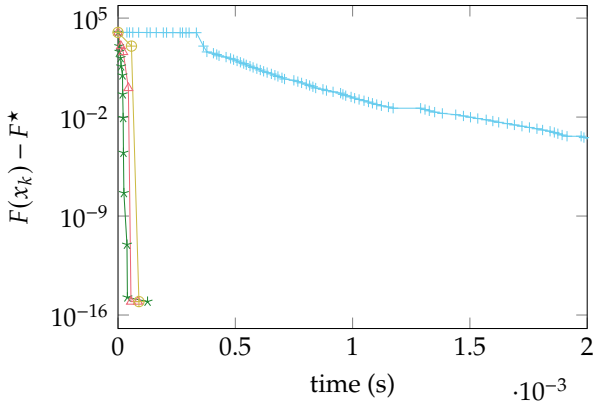
Algorithm	$\|h(x)\|$	$\operatorname{dist}(\partial^{\mathcal{M}} F(\bar{x}), 0)$	$\operatorname{proj}_{\partial F(\bar{x})}(0) \in \operatorname{ri} \partial F(\bar{x})$
Gradient Sampling	$5.67 \cdot 10^{-8}$	$1.98 \cdot 10^{-6}$	true
nsBFGS	$5.01 \cdot 10^{-16}$	$5.42 \cdot 10^{-9}$	true
$\mathcal{VU}$-algorithm	$5.43 \cdot 10^{-16}$	$2.33 \cdot 10^{-10}$	true
GlobalNewton	$2.36 \cdot 10^{-17}$	$6.58 \cdot 10^{-17}$	true

(d) Optimality conditions (5.17) and (5.19) at the final point of the algorithms, relative to the optimal manifold $\mathcal{M}^{\max}_{\{1,3\}}$.

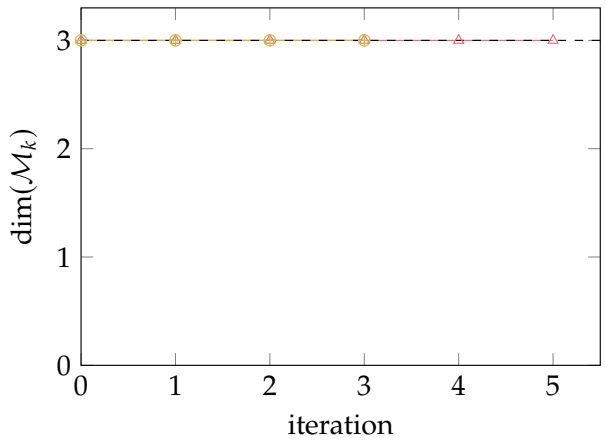
Figure 5.8: `F3d-U2` problem



(a) Suboptimality vs iterations



(b) Suboptimality vs time (s)

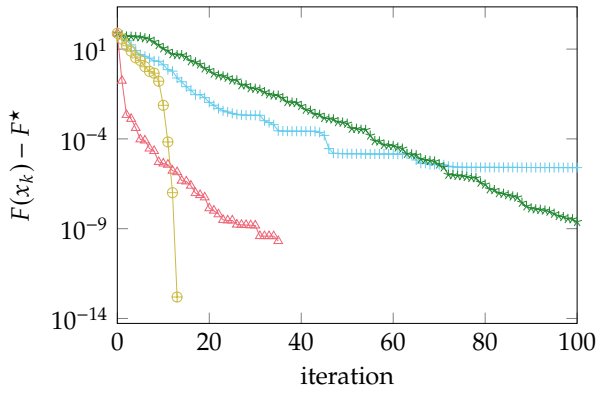


(c) Manifold dimension vs iterations

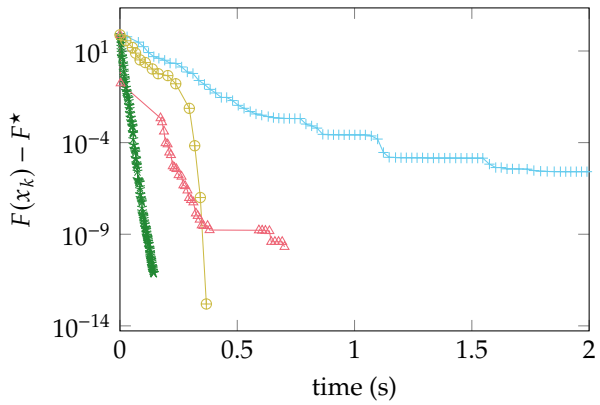
Algorithm	$\|h(x)\|$	$\operatorname{dist}(\partial^{\mathcal{M}} F(\bar{x}), 0)$	$\operatorname{proj}_{\partial F(\bar{x})}(0) \in \operatorname{ri} \partial F(\bar{x})$
Gradient Sampling	–	$8.26 \cdot 10^{-7}$	–
nsBFGS	–	$5.02 \cdot 10^{-11}$	–
$\mathcal{VU}$-algorithm	–	$4.45 \cdot 10^{-16}$	–
GlobalNewton	–	$1.5 \cdot 10^{-20}$	–

(d) Optimality conditions (5.17) and (5.19) at the final point of the algorithms, relative to the optimal manifold $\mathcal{M}^{\max}_{\{1\}}$. The function is differentiable near $x^\star$, therefore $\operatorname{dist}(\partial^{\mathcal{M}} F(\bar{x}), 0)$ boils down to $\|\nabla F(\bar{x})\|$. The other two criteria are not relevant here, as they quantify optimality of $\bar{x}$ along nonsmooth directions.

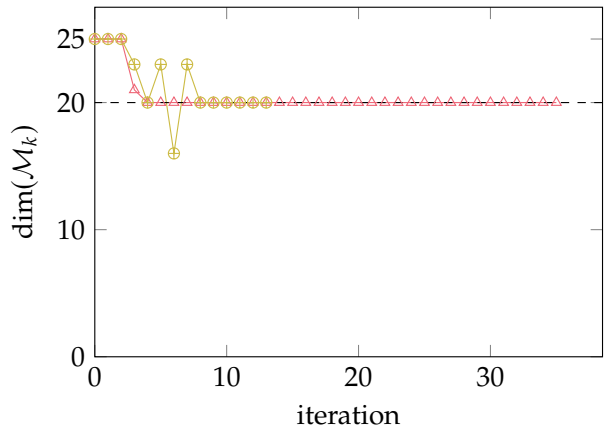
Figure 5.9: `F3d-U3` problem



(a) Suboptimality vs iterations



(b) Suboptimality vs time (s)



(c) Manifold dimension vs iterations

Gradient Sampling — nsBFGS — $\mathcal{VU}$-algorithm — Global Newton

Algorithm	$\|h(x)\|$	$\operatorname{dist}(\partial^{\mathcal{M}} F(\bar{x}), 0)$	$\operatorname{proj}_{\partial F(\bar{x})}(0) \in \operatorname{ri} \partial F(\bar{x})$
Gradient Sampling	$6.87 \cdot 10^{-6}$	$2.24 \cdot 10^{-5}$	true
nsBFGS	$5.52 \cdot 10^{-12}$	$4.5 \cdot 10^{-6}$	true
$\mathcal{VU}$-algorithm	$1.09 \cdot 10^{-13}$	$2.66 \cdot 10^{-5}$	true
GlobalNewton	$3.63 \cdot 10^{-13}$	$1.63 \cdot 10^{-13}$	true

(d) Optimality conditions (5.17) and (5.19) at the final point of the algorithms, relative to the optimal manifold $\mathcal{M}_r^{\lambda^{\max}}$, with $r = 3$.

Figure 5.10: Eigmax Problem

6

CONCLUSION & PERSPECTIVES

This thesis considers the question of detecting and leveraging the smooth substructure of nonsmooth functions that appear in some machine learning, signal processing, and control applications. After a chapter of introduction (Chapter 1), and a chapter of recalls (Chapter 2), the contributions of this thesis were presented in Chapters 3–5. We briefly review them here, to better lay down a number of on-going works, further questions, and perspectives.

In Chapter 3, we considered the minimization of additive nonsmooth functions. For such nonconvex functions, we first proved that the proximal gradient operator detects smooth substructure: its sends neighborhoods of nonsmooth minimizers to the optimal submanifold (Theorem 3.1). Then, we proposed an algorithm that alternates a proximal-gradient step with an efficient Riemannian step on the detected manifold (Algorithm 3.1). We proved that this algorithm converges to critical points and that, if one of the minimizers meets a natural geometrical condition, the algorithm eventually detects the smooth substructure and converges at a quadratic speed (Theorems 3.2 and 3.3). We also observed this fast convergence numerically on classical structured regression problems.

In Chapter 4, we studied the local structure of functions that write as a composition of a nonsmooth function with a smooth mapping. When the proximity operator of the nonsmooth function is explicitly available, we showed that it can be used to detect the smooth substructure of minimizers of the full function (Theorem 4.4). We used this information to propose a local Newton method that minimizes the objective by exploiting the detected structure (Algorithm 4.1). This method is guaranteed to identify the structure of the minimizer and to converge quadratically (Theorem 4.7), which we illustrated on two popular nonsmooth problems.

In Chapter 5, we also considered the minimization of composite functions, but from a global perspective. We proposed preliminary results aimed at designing an algorithm that converges from any starting point, and with the same local guarantees as the Newton algorithm of Chapter 4. There, we touched more closely than before the difficulties of globalization and of the combinatorial aspect brought by identification. We showed that linesearch SQP method, with a correction term, still retains the fast local Newton rate (Theorem 5.6). We also proposed an optimality criterion for algorithms that explicitly use the smooth substructure of nonsmooth functions. While a theoretically sound combination of linesearch and identification procedure has yet to be reached, we illustrated numerically the behavior of a heuristic global Newton algorithm (Algorithm 5.2) on two nonsmooth problems.

Nonsmooth optimization is an exciting field; we are glad to have obtained some results and proposed promising algorithms. During this PhD, we discovered the beauties and difficulties of nonsmoothness; in particular, we have faced many technical challenges. Some of them were quite subtle, and probably

explain why interest in structure exploitation slowed down after founding contributions in the 2000s, including Mifflin and Sagastizábal (2005); Lewis (2002), central in this work. We are glad to note a renewed interest in such questions, with e.g., Han and Lewis (2023); Davis and Drusvyatskiy (2019); Lee (2023). Facing these challenges allowed us to propose several refined technical results, such as the description of identification of the proximal gradient *away from minimizers*, or the precise description of the proximity operator behavior *relative to its stepsize*. On the other hand, the question of globalizing the local Newton method for composite minimization, discussed in Chapter 5, proved to be particularly delicate for us. The recent preprint Davis et al. (2021) was released timely, providing a rich theoretical framework based on Whitney stratification of definable functions. This approach sheds an interesting light on the globalization question, allowing the preliminary developments of Chapter 5.

The key in Chapter 5, and more generally in this thesis, was to combine a theoretical study of the nonsmoothness, a review of the recent literature, and thorough numerical experimentation. Comparing theoretical ideas with their behavior in practice was, for us, a very effective way to build and refine intuition on the behavior of methods or procedures. From that point, seeking out fine geometrical reasons for these intuitions brought forth most results of this thesis.

Along the way, we encountered many interesting questions, ranging from intriguing technical facts to appealing applications. While some questions remained at the stage of remarks, others led to promising developments. I list below some of these questions: first, the questions regarding applications to specific problems, and second, those regarding extensions of the results presented in this thesis.

ADDITIVE SETTINGS WITH SEVERAL NONSMOOTH FUNCTIONS. Chapter 3 considered additive functions with two elements. One could consider the wider setting of functions that write as $f + g + h(L\cdot)$ with one smooth and two nonsmooth functions with an explicit prox. This function can be written as a composition of a smooth map and a nonsmooth function with an explicit prox, that appear in signal processing e.g., 3d mesh denoising (Repetti et al., 2015; Thouvenin et al., 2022). The composition-based methods of Chapters 4 and 5 thus apply. How would these methods compare to the splitting methods such Condat-VU (Condat, 2013; Vũ, 2013) or three-operator splitting (Davis and Yin, 2017)? Can these two approaches be combined to obtain a globally convergent and locally fast algorithm akin to the Riemannian acceleration of Chapter 3? I would like to explore these questions in a medium run.

MINIMIZATION OF SMOOTH FUNCTIONS OVER PROJECTION-EXPLICIT SETS. The additive setting of Chapter 3 captures the minimization of a smooth function over a set which admits an explicit projection; it suffices to take g as the indicator function of the set. In this wide setting, let us underline the minimization of quadratics over the simplex, which forms the bundle subproblem (Bonnans et al., 2006, Eq. 10.9), or on the spectraplex, which appears in the spectral bundle (Helmberg et al., 2014, Sec. 5.2) (and in Chapter 5). The interest of the algorithm of Chapter 3, in this case a Newton acceleration of the *projected* gradient, is the identification of the minimizer smooth substructures, which corresponds to the active constraints. This is indeed the relevant information for the mentioned problems: finding the active linear functions for the bundle subproblem, or the rank of the matrix that minimizes over the spectraplex. These problems are respectively dealt with by a specific QP solver on the simplex (Kiwiel, 1986),

or heuristically in Helmberg et al. (2014). An interesting direction would thus be to adapt the algorithm of Chapter 3 to this setting, and investigate their applicability in the above settings. Notably, this requires adapting linesearches to extended-value functions.

We now turn to perspectives to improve the algorithms and results present in this thesis.

IMPROVE THE GLOBAL EFFICIENCY OF NEWTON ACCELERATION OF PROXGRAD. The Newton acceleration of the proximal gradient converges fast around minimizers, as discussed in Chapter 3. However, the observed convergence rate away from minimizers is still similar to that of the proximal gradient. Several extensions are possible.

First, both building blocks can be refined: the truncation strategy used in Chapter 3 was a first attempt to improve on the plain Newton method but other Newton accelerations could be considered (e.g., trust-region (Absil et al., 2007), cubic regularization (Agarwal et al., 2021)), as well as other proximal algorithms (e.g., prox-Newton (Lee et al., 2014), fast proximal gradient (Beck and Teboulle, 2009)).

Second, computing a proximal gradient can be costly in some settings e.g., with nuclear-norm regularization of large problems, since it amounts to computing an SVD. For such problems, the Burrer-Monteiro (Waldspurger, 2021) approach represents the low-rank iterates in factorized form. Could this approach be adapted for the Riemannian acceleration of the proximal gradient? One could consider an algorithm which performs at each step either a proximal gradient step or a Riemannian step, and try to limit the number of prox steps. How to decide weather it is beneficial to perform an identification or a Riemannian step? How often should an identification step, less efficient than a Riemannian step, be performed? Going step further, can the proximal gradient be altogether replaced by a computationally lighter identification procedure? Ideas from Chapter 5 could help.

IDENTIFICATION NEAR NONQUALIFIED MINIMIZERS. In practical settings the minimizers can be *non*-qualified, which means that both $0 \in \partial F(\bar{x})$ and $0 \notin \operatorname{ri} \partial F(\bar{x})$ hold, at least numerically. This setting was studied for "mirror-stratifiable" functions of Fadili et al. (2018). It leads to a "sandwhich" identification behavior: the algorithm provably converges to one of a subset of structure manifolds, and in practice, it often identifies the manifold with largest dimension. What is the behavior of the algorithms introduced in this thesis on non qualified problems? Do they retain some identification and local convergence properties? More generally, how can this partial substructure be exploited?

GENERALITY OF PROPERTIES NORMAL ASCENT AND CURVE. The developments of Chapter 4 required the introduction of Properties 4.1 and 4.2, which hold on our examples. How constraining are these properties? Regarding the normal ascent property, as discussed on an example, it seems a linear tilt along a normal direction could ensure it holds. Can this remark be formalized and integrated to the theory properly? Besides, the curve property was proved to hold at any nonsmooth points of the maximum and maximum eigenvalue functions. For what function does this property fails? Is there a simple criterion on the nonsmooth function that ensures this property?

BIBLIOGRAPHY

P-A Absil, Christopher G Baker, and Kyle A Gallivan. Trust-region methods on riemannian manifolds. *Foundations of Computational Mathematics*, 7(3):303–330, 2007.

P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2009a.

P. A. Absil, Jochen Trumpf, Robert Mahony, and Ben Andrews. All roads lead to Newton: Feasible second-order methods for equality-constrained optimization. Technical report, August 2009b.

Pierre-Antoine Absil and Jérôme Malick. Projection-like retractions on matrix manifolds. *SIAM J. Optim.*, 22:135–158, 2012.

Naman Agarwal, Nicolas Boumal, Brian Bullins, and Coralia Cartis. Adaptive regularization with cubics on manifolds. *Mathematical Programming*, 188(1):85–134, July 2021. ISSN 0025-5610, 1436-4646. doi: 10.1007/s10107-020-01505-1.

Pierre Apkarian, Dominikus Noll, Jean-Baptiste Thevenet, and Hoang Duong Tuan. A Spectral Quadratic-SDP Method with Applications to Fixed-Order H2 and H∞ Synthesis. *European Journal of Control*, 10(6):527–538, January 2004. ISSN 0947-3580. doi: 10.3166/ejc.10.527-538.

Aleksandr Y. Aravkin, Robert Baraldi, and Dominique Orban. A proximal quasi-newton trust-region method for nonsmooth regularized optimization. *SIAM Journal on Optimization*, 32(2):900–929, 2022.

Hedy Attouch and Jérôme Bolte. On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. *Mathematical Programming*, 116 (1):5–16, 2009.

Francis Bach, Rodolphe Jenatton, Julien Mairal, Guillaume Obozinski, et al. Optimization with sparsity-inducing penalties. *Foundations and Trends® in Machine Learning*, 4 (1):1–106, 2012.

Francis R. Bach. Consistency of Trace Norm Minimization. *The Journal of Machine Learning Research*, 9:1019–1048, June 2008. ISSN 1532-4435.

Gilles Bareilles and Franck Iutzeler. On the interplay between acceleration and identification for the proximal gradient algorithm. *Computational Optimization and Applications*, 77(2):351–378, 2020.

Gilles Bareilles, Franck Iutzeler, and Jérôme Malick. Harnessing structure in composite nonsmooth minimization. (arXiv:2206.15053), June 2022a. doi: 10.48550/arXiv.2206.15053.

Gilles Bareilles, Franck Iutzeler, and Jérôme Malick. Newton acceleration on manifolds identified by proximal gradient methods. *Mathematical Programming*, August 2022b. ISSN 1436-4646. doi: 10.1007/s10107-022-01873-w.

Heinz H. Bauschke and Patrick L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. CMS Books in Mathematics. Springer International Publishing, Cham, 2017. ISBN 978-3-319-48310-8 978-3-319-48311-5. doi: 10.1007/978-3-319-48311-5.

Amir Beck. *First-order methods in optimization*, volume 25. SIAM, 2017.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

Stephen Becker, Jalal Fadili, and Peter Ochs. On quasi-newton forward-backward splitting: Proximal calculus and convergence. *SIAM Journal on Optimization*, 29(4): 2445–2481, 2019.

A. Ben-Tal, L. El Ghaoui, and A. Nemirovski. *Robust optimization*. Princeton University Press, 2009.

Quentin Bertrand, Quentin Klopfenstein, Pierre-Antoine Bannier, Gauthier Gidel, and Mathurin Massias. Beyond L1: Faster and Better Sparse Models with skglm, April 2022.

Jeff Bezanson, Alan Edelman, Stefan Karpinski, and Viral B Shah. Julia: A fresh approach to numerical computing. *SIAM review*, 59(1):65–98, 2017.

Jérôme Bolte and Edouard Pauwels. Majorization-Minimization Procedures and Convergence of SQP Methods for Semi-Algebraic and Tame Programs. *Mathematics of Operations Research*, 41(2):442–465, May 2016. ISSN 0364-765X, 1526-5471. doi: 10.1287/moor.2015.0735.

Jérôme Bolte, Trong Phong Nguyen, Juan Peypouquet, and Bruce W Suter. From error bounds to the complexity of first-order descent methods for convex functions. *Mathematical Programming*, pages 1–37, 2015.

Jérôme Bolte, Zheng Chen, and Edouard Pauwels. The multiproximal linearization method for convex composite problems. *Mathematical Programming*, 182(1):1–36, July 2020. ISSN 1436-4646. doi: 10.1007/s10107-019-01382-3.

Joseph-Frédéric Bonnans, Jean Charles Gilbert, Claude Lemaréchal, and Claudia A Sagastizábal. *Numerical optimization: theoretical and practical aspects*. Springer Science & Business Media, 2006.

Nicolas Boumal. An introduction to optimization on smooth manifolds. To appear with Cambridge University Press, Apr 2022. URL http://www.nicolasboumal.net/book.

O. Briant, C. Lemaréchal, Ph. Meurdesoif, S. Michel, N. Perrot, and F. Vanderbeck. Comparison of bundle and classical column generation. *Mathematical Programming*, 113(2):299–344, 2008.

James V Burke and Jorge J Moré. On the identification of active constraints. *SIAM Journal on Numerical Analysis*, 25(5):1197–1211, 1988.

James V Burke, Frank E Curtis, Adrian S Lewis, Michael L Overton, and Lucas EA Simões. Gradient sampling methods for nonsmooth optimization. In *Numerical Nonsmooth Optimization*, pages 201–225. Springer, 2020.

Emmanuel Candès and Benjamin Recht. Simple bounds for recovering low-complexity models. *Mathematical Programming*, 141(1):577–589, October 2013. ISSN 1436-4646. doi: 10.1007/s10107-012-0540-0.

Emmanuel J. Candès, Justin K. Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, 59(8):1207–1223, 2006. ISSN 1097-0312. doi: 10.1002/cpa.20124.

Laurent Chauvier, Antonio Fuduli, and Charles Jean Gilbert. A truncated sqp algorithm for solving nonconvex equality constrained optimization problems. In *High Performance Algorithms and Software for Nonlinear Optimization*, pages 149–176. Springer, 2003.

Laurent Condat. A primal–dual splitting method for convex optimization involving lipschitzian, proximable and linear composite terms. *Journal of Optimization Theory and Applications*, 158(2):460–479, 2013.

Rafael Correa and Claude Lemaréchal. Convergence of some algorithms for convex minimization. *Mathematical Programming*, 62(1):261–275, February 1993. ISSN 1436-4646. doi: 10.1007/BF01585170.

Aris Daniilidis, Warren Hare, and Jérôme Malick. Geometrical interpretation of the predictor-corrector type algorithms in structured optimization problems. *Optimization*, 55(5-6):481–503, 2006.

Aris Daniilidis, Claudia Sagastizábal, and Mikhail Solodov. Identifying structure of nonsmooth convex functions by the bundle technique. *SIAM Journal on Optimization*, 20(2):820–840, 2009. doi: 10.1137/080729864.

Damek Davis and Dmitriy Drusvyatskiy. Active strict saddles in nonsmooth optimization. page 43, 2019.

Damek Davis and Wotao Yin. A three-operator splitting scheme and its optimization applications. *Set-valued and variational analysis*, 25(4):829–858, 2017.

Damek Davis, Dmitriy Drusvyatskiy, and Liwei Jiang. Subgradient methods near active manifolds: Saddle point avoidance, local convergence, and asymptotic normality. *arXiv:2108.11832 [cs, math]*, August 2021.

Ron S Dembo and Trond Steihaug. Truncated-newtono algorithms for large-scale unconstrained optimization. *Mathematical Programming*, 26(2):190–212, 1983.

John E Dennis Jr and Robert B Schnabel. *Numerical methods for unconstrained optimization and nonlinear equations*. SIAM, 1996.

Elizabeth D. Dolan and Jorge J. Moré. Benchmarking optimization software with performance profiles. *Mathematical Programming*, 91(2):201–213, January 2002. ISSN 1436-4646. doi: 10.1007/s101070100263.

D. Drusvyatskiy and C. Paquette. Efficiency of minimizing compositions of convex functions and smooth maps. *Mathematical Programming*, 178(1-2):503–558, November 2019. ISSN 0025-5610, 1436-4646. doi: 10.1007/s10107-018-1311-3.

D. Drusvyatskiy, A. D. Ioffe, and A. S. Lewis. Nonsmooth optimization using Taylor-like models: Error bounds, convergence, and termination criteria. *Mathematical Programming*, 185(1-2):357–383, January 2021. ISSN 0025-5610, 1436-4646. doi: 10.1007/s10107-019-01432-w.

Dmitriy Drusvyatskiy and Adrian S Lewis. Optimality, identifiability, and sensitivity. *Mathematical Programming*, 147(1-2):467–498, 2014.

Jalal Fadili, Jerome Malick, and Gabriel Peyré. Sensitivity analysis for mirror-stratifiable convex functions. *SIAM Journal on Optimization*, 28(4):2975–3000, 2018.

M. Fazel, H. Hindi, and S.P. Boyd. A rank minimization heuristic with application to minimum order system approximation. In *Proceedings of the 2001 American Control Conference. (Cat. No.01CH37148)*, volume 6, pages 4734–4739 vol.6, June 2001. doi: 10.1109/ACC.2001.945730.

Osman. Güler. On the Convergence of the Proximal Point Algorithm for Convex Minimization. *SIAM Journal on Control and Optimization*, 29(2):403–419, March 1991. ISSN 0363-0129. doi: 10.1137/0329022.

X. Y. Han and Adrian S. Lewis. Survey Descent: A Multipoint Generalization of Gradient Descent for Nonsmooth Optimization. *SIAM Journal on Optimization*, pages 36–62, March 2023. ISSN 1052-6234. doi: 10.1137/21M1468450.

Warren Hare and Adrian S Lewis. Identifying active constraints via partial smoothness and prox-regularity. *Journal of Convex Analysis*, 11(2):251–266, 2004.

Warren Hare and Claudia Sagastizábal. Computing proximal points of nonconvex functions. *Mathematical Programming*, 116(1-2):221–258, 2009.

Grace Hechme-Doukopoulos, Sandrine Brignol-Charousset, Jérôme Malick, and Claude Lemaréchal. The short-term electricity production management problem at edf. *Optima Newsletter*, 84:2–6, 2010.

C. Helmberg and F. Rendl. A Spectral Bundle Method for Semidefinite Programming. *SIAM Journal on Optimization*, 10(3):673–696, January 2000. ISSN 1052-6234. doi: 10.1137/S1052623497328987.

C. Helmberg, M.L. Overton, and F. Rendl. The spectral bundle method with second-order information. *Optimization Methods and Software*, 29(4):855–876, July 2014. ISSN 1055-6788, 1029-4937. doi: 10.1080/10556788.2013.858155.

J.-B. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms*. Springer Verlag, Heidelberg, 1993. Two volumes.

Franck Iutzeler and Jérôme Malick. Nonsmoothness in Machine Learning: Specific Structure, Proximal Identification, and Applications. *Set-Valued and Variational Analysis*, 28(4):661–678, December 2020. ISSN 1877-0533, 1877-0541. doi: 10.1007/s11228-020-00561-1.

Krzysztof C. Kiwiel. A Method for Solving Certain Quadratic Programming Problems Arising in Nonsmooth Optimization. *IMA Journal of Numerical Analysis*, 6(2): 137–152, 1986. ISSN 0272-4979, 1464-3642. doi: 10.1093/imanum/6.2.137.

Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix Factorization Techniques for Recommender Systems. *Computer*, 42(8):30–37, August 2009. ISSN 1558-0814. doi: 10.1109/MC.2009.263.

D. Kuhn, P.M. Esfahani, V. Anh Nguyen, and S. Shafieezadeh-Abadeh. Wasserstein distributionally robust optimization: Theory and applications in machine learning. In *Operations Research & Management Science in the Age of Analytics*. INFORMS, 2019.

Ching-pei Lee. Accelerating inexact successive quadratic approximation for regularized optimization through manifold identification. *Mathematical Programming*, January 2023. ISSN 1436-4646. doi: 10.1007/s10107-022-01916-2.

Jason D Lee, Yuekai Sun, and Michael A Saunders. Proximal newton-type methods for minimizing composite functions. *SIAM Journal on Optimization*, 24(3):1420–1443, 2014.

John M. Lee. *Introduction to Smooth Manifolds*. Graduate Texts in Mathematics. Springer-Verlag, New York, 2003. ISBN 978-0-387-21752-9. doi: 10.1007/978-0-387-21752-9.

Claude Lemaréchal, François Oustry, and Claudia Sagastizábal. The u-lagrangian of a convex function. *Transactions of the American mathematical Society*, 352(2):711–729, 2000.

Claude Lemaréchal, Adam Ouorou, and Georgios Petrou. A bundle-type algorithm for routing in telecommunication data networks. *Computational Optimization and Applications*, 44(3):385, December 2007. ISSN 1573-2894. doi: 10.1007/s10589-007-9160-7.

A. S. Lewis and S. J. Wright. A proximal method for composite minimization. *Mathematical Programming*, 158(1):501–546, July 2016. ISSN 1436-4646. doi: 10.1007/s10107-015-0943-9.

A. S. Lewis and S. Zhang. Partial Smoothness, Tilt Stability, and Generalized Hessians. *SIAM Journal on Optimization*, 23(1):74–94, January 2013. ISSN 1052-6234, 1095-7189. doi: 10.1137/110852103.

A. S. Lewis, Jingwei Liang, and Tonghua Tian. Partial Smoothness and Constant Rank. *SIAM Journal on Optimization*, 32(1):276–291, March 2022. ISSN 1052-6234. doi: 10.1137/19M1237909.

Adrian Lewis and Tonghua Tian. Identifiability, the kl property in metric spaces, and subgradient curves. *arXiv preprint arXiv:2205.02868*, 2022.

Adrian Lewis and Calvin Wylie. A simple Newton method for local nonsmooth optimization. *arXiv:1907.11742 [cs, math]*, July 2019.

Adrian S Lewis. Active sets, nonsmoothness, and sensitivity. *SIAM Journal on Optimization*, 13(3):702–725, 2002.

Adrian S. Lewis and Michael L. Overton. Nonsmooth optimization via quasi-Newton methods. *Mathematical Programming*, 141(1):135–163, October 2013. ISSN 1436-4646. doi: 10.1007/s10107-012-0514-2.

Jingwei Liang, Jalal Fadili, and Gabriel Peyré. Local linear convergence of forward–backward under partial smoothness. In *Advances in Neural Information Processing Systems*, pages 1970–1978, 2014.

Jingwei Liang, Jalal Fadili, and Gabriel Peyré. Activity identification and local linear convergence of forward–backward-type methods. *SIAM Journal on Optimization*, 27(1):408–437, 2017a.

Jingwei Liang, Jalal Fadili, and Gabriel Peyré. Local Convergence Properties of Douglas–Rachford and Alternating Direction Method of Multipliers. *Journal of Optimization Theory and Applications*, 172(3):874–913, March 2017b. ISSN 1573-2878. doi: 10.1007/s10957-017-1061-z.

Shuai Liu, Claudia Sagastizábal, and Mikhail Solodov. Subdifferential Enlargements and Continuity Properties of the V U $$\mathcal { {VU}}$$ -Decomposition in Convex Optimization. In Seyedehsomayeh Hosseini, Boris S. Mordukhovich, and André Uschmajew, editors, *Nonsmooth Optimization and Its Applications*, volume 170, pages 55–87. Springer International Publishing, Cham, 2019. ISBN 978-3-030-11369-8 978-3-030-11370-4. doi: 10.1007/978-3-030-11370-4_4.

Mathurin Massias, Joseph Salmon, and Alexandre Gramfort. Celer: a fast solver for the lasso with dual extrapolation. In *International Conference on Machine Learning*, pages 3321–3330, 2018.

Robert Mifflin and Claudia Sagastizábal. A $\mathcal{VU}$-algorithm for convex minimization. *Mathematical programming*, 104(2-3):583–608, 2005.

Scott A Miller and Jérôme Malick. Newton methods for nonsmooth convex minimization: connections among-lagrangian, riemannian newton and sqp methods. *Mathematical programming*, 104(2-3):609–633, 2005.

Jean-Jacques Moreau. Proximité et dualité dans un espace Hilbertien. *Bull. Soc. Math. France*, 93(2):273–299, 1965.

Eugene Ndiaye, Olivier Fercoq, Alexandre Gramfort, and Joseph Salmon. Gap safe screening rules for sparsity enforcing penalties. *The Journal of Machine Learning Research*, 18(1):4671–4703, 2017.

Yurii Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.

Yurii E Nesterov. A method for solving the convex programming problem with convergence rate $O(1/k^2)$. In *Dokl. Akad. Nauk SSSR*, volume 269, pages 543–547, 1983.

Jorge Nocedal and Stephen Wright. *Numerical optimization*. Springer Science & Business Media, 2006.

Dominikus Noll and Pierre Apkarian. Spectral bundle methods for non-convex maximum eigenvalue functions: Second-order methods. *Mathematical Programming*, 104(2-3):729–747, November 2005. ISSN 0025-5610, 1436-4646. doi: 10.1007/s10107-005-0635-y.

Julie Nutini, Mark Schmidt, and Warren Hare. "Active-set complexity" of proximal gradient: How long does it take to find the sparsity pattern? *Optimization Letters*, 13(4):645–655, June 2019. ISSN 1862-4480. doi: 10.1007/s11590-018-1325-z.

François Oustry. The $\mathcal{U}$-Lagrangian of the Maximum Eigenvalue Function. *SIAM Journal on Optimization*, 9(2):526–549, January 1999. ISSN 1052-6234. doi: 10.1137/S1052623496311776.

René Poliquin and R Rockafellar. Prox-regular functions in variational analysis. *Transactions of the American Mathematical Society*, 348(5):1805–1838, 1996.

Audrey Repetti, Emilie Chouzenoux, and Jean-Christophe Pesquet. A random block-coordinate primal-dual proximal algorithm with application to 3d mesh denoising. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3561–3565. IEEE, 2015.

R. Tyrrell Rockafellar and Roger J.-B. Wets. *Variational Analysis*. Grundlehren Der Mathematischen Wissenschaften. Springer-Verlag, Berlin Heidelberg, 1998. ISBN 978-3-540-62772-2. doi: 10.1007/978-3-642-02431-3.

Claudia Sagastizábal. Nonsmooth Optimization: Thinking Outside of the Black Box. *SIAG/OPT Views-and-News*, 22(2):11, 2011. doi: 10.1.1.681.8726.

Claudia Sagastizábal. Composite proximal bundle method. *Mathematical Programming*, 140(1):189–233, August 2013. ISSN 1436-4646. doi: 10.1007/s10107-012-0600-5.

Otmar Scherzer, Grasmair Markus, Harald Grossauer, Markus Haltmeier, and Frank Lenzen. *Variational Methods in Imaging*, volume 167 of *Applied Mathematical Sciences*.

Springer New York, New York, NY, 2009. ISBN 978-0-387-30931-6 978-0-387-69277-7. doi: 10.1007/978-0-387-69277-7.

Alexander Shapiro. On a Class of Nonsmooth Composite Functions. *Mathematics of Operations Research*, 28(4):677–692, November 2003. ISSN 0364-765X, 1526-5471. doi: 10.1287/moor.28.4.677.20512.

Alexander Shapiro and Michael K. H. Fan. On Eigenvalue Optimization. *SIAM Journal on Optimization*, 5(3):552–569, August 1995. ISSN 1052-6234, 1095-7189. doi: 10.1137/0805028.

P. Spellucci. An SQP method for general nonlinear programs using only equality constrained subproblems. *Mathematical Programming*, 82(3):413–448, August 1998. ISSN 0025-5610, 1436-4646. doi: 10.1007/BF01580078.

Pierre-Antoine Thouvenin, Arwa Dabbech, Ming Jiang, Abdullah Abdulaziz, Jean-Philippe Thiran, Adrian Jackson, and Yves Wiaux. Parallel faceted imaging in radio interferometry via proximal splitting (Faceted HyperSARA): II. Code and real data proof of concept, September 2022.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

Samuel Vaiter, Gabriel Peyré, and Jalal Fadili. Low complexity regularization of linear inverse problems. In *Sampling Theory, a Renaissance*, pages 103–153. Springer, 2015.

B.-C. Vũ. A splitting algorithm for dual monotone inclusions involving cocoercive operators. *Advances in Computational Mathematics*, 38(3):667–681, 2013.

Irène Waldspurger. Lecture notes on non-convex algorithms for low-rank matrix recovery, May 2021.

Philip Wolfe. Finding the nearest point in A polytope. *Mathematical Programming*, 11(1): 128–149, December 1976. ISSN 0025-5610, 1436-4646. doi: 10.1007/BF01580381.

R. S. Womersley and R. Fletcher. An algorithm for composite nonsmooth optimization problems. *Journal of Optimization Theory and Applications*, 48(3):493–523, March 1986. ISSN 0022-3239, 1573-2878. doi: 10.1007/BF00940574.

Stephen J Wright. Identifiable surfaces in constrained optimization. *SIAM Journal on Control and Optimization*, 31(4):1063–1079, 1993.

APPENDIX

A

ELEMENTARY RESULTS ON THE PROXIMAL GRADIENT AND RIEMANNIAN OPTIMIZATION

In this appendix, we provide elementary, though usefull, technical results and tools that are useful in our developments. These results may be seen as folklore knowledge, and appear more or less explicitly in the literature.

A.1 TECHNICAL RESULTS ON RIEMANNIAN MANIFOLDS

We first state Riemannian sufficient optimality conditions (Lemma A.1), then build a specific retraction (Proposition A.2), and finally show that Euclidean and Riemannian distances are locally equivalent (Lemma A.4).

A.1.1 *Sufficient optimality conditions*

Using Taylor extension on smooth curves (see e.g., Boumal (2022, Chap 4.2, 6.1)), we recover in the Riemannian setting sufficient optimality conditions for strong minimizers.

Used to prove Theorem 3.1.

Lemma A.1 (Sufficient optimality conditions). *Consider a manifold $\mathcal{M}$ embedded in $\mathbb{R}^n$, a point $\bar{x}$ of $\mathcal{M}$ and a function f defined on a neighborhood of $\bar{x}$ in $\mathcal{M}$ for which $\bar{x}$ is a strong minimizer i.e., for all x near $\bar{x}$ in M,*

$$f(x) \geq f(\bar{x}) + \frac{c}{2}\|x - \bar{x}\|^2.$$

Then, there holds

$$\operatorname{grad} f(\bar{x}) = 0 \quad \textit{and} \quad \operatorname{Hess} f(\bar{x}) \succeq cI.$$

Proof. Let $\eta \in T_{\bar{x}}\mathcal{M}$ and denote γ a geodesic going through $\bar{x}$ with velocity η at $t = 0$. The quadratic growth assumption can be applied at $x = \gamma(t)$, which allows to write

$$\frac{1}{t}(f \circ \gamma(t) - f \circ \gamma(0)) \geq \frac{c}{2}\left\|\frac{\gamma(t) - \gamma(0)}{\sqrt{t}}\right\|^2.$$

Taking the limit $t \to 0$ yields $\langle \operatorname{grad} f(\bar{x}), \eta \rangle \geq 0$. The same reasoning holds with $x = \gamma(-t)$ and yields the converse inequality, so that $\langle \operatorname{grad} f(\bar{x}), \eta \rangle = 0$.

Besides, summing the quadratic growth conditions applied at $\gamma(t)$ and $\gamma(-t)$ provides

$$\frac{1}{t^2}(f \circ \gamma(t) - 2f \circ \gamma(0) + f \circ \gamma(-t)) \geq \frac{c}{2}\left\|\frac{\gamma(t) - \gamma(0)}{t}\right\|^2 + \frac{c}{2}\left\|\frac{\gamma(-t) - \gamma(0)}{t}\right\|^2.$$

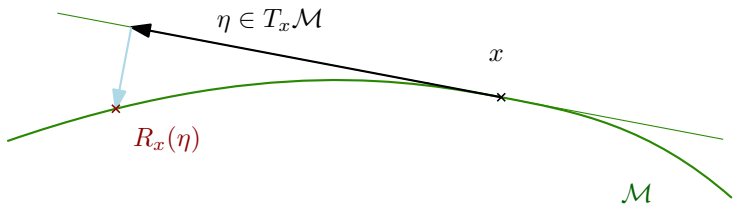


Figure A.1: Illustration of the Orthographic retraction (Proposition A.2).

Taking the limit as $t \to 0$ yields $\langle \mathrm{Hess}\, f(\bar{y})[\eta], \eta \rangle \geq c\|\eta\|^2$. As η is picked arbitrarily in $T_{\bar{x}}\mathcal{M}$, the results are obtained. □

A.1.2 *Orthographic retraction*

This term echoes the orthographic map projection of Earth.

We introduce a specific retraction and show it is a second-order retraction (Absil and Malick, 2012). This retraction was introduced in Miller and Malick (2005, Th. 2.2), before the notions of (second-order) retraction became established; it is illustrated in Fig. A.1.

Used to prove Lemma A.4 and Theorem 5.6

Proposition A.2 . *Consider a p-dimensional $\mathcal{C}^k$-submanifold $\mathcal{M}$ of $\mathbb{R}^n$ around a point $\bar{x} \in \mathcal{M}$. The mapping $\mathrm{R} : T\mathcal{B} \to \mathcal{M}$, defined for $(x, \eta) \in T\mathcal{B}$ near $(\bar{x}, 0)$ by $\mathrm{proj}_x(\mathrm{R}(x, \eta)) = \eta$ defines a second-order retraction near $(\bar{x}, 0)$. The point-wise retraction, defined as $\mathrm{R}_x = \mathrm{R}(x, \cdot)$, is locally invertible with inverse $\mathrm{R}_x^{-1} = \mathrm{proj}_x$.*

Proof. Let $\Psi : \mathbb{R}^n \to \mathbb{R}^{n-p}$ denote a $\mathcal{C}^k$ function defining $\mathcal{M}$ around $\bar{x}$: for all x close enough to $\bar{x}$, there holds $x \in \mathcal{M} \Leftrightarrow \Psi(x) = 0$, and $\mathrm{D}\Psi(x)$ is surjective. Consider the equation $\Phi(x, \eta_t, \eta_n) = 0$ around $(\bar{x}, 0, 0)$, with

$$\begin{aligned} \Phi : \{x, \eta_t, \eta_n : x \in \mathcal{M}, \eta_t \in T_x\mathcal{M}, \eta_n \in N_x\mathcal{M}\} &\to \mathbb{R} \\ x, \eta_t, \eta_n &\mapsto \Psi(x + \eta_t + \eta_n). \end{aligned}$$

The partial differential $\mathrm{D}_{\eta_n} \Phi(\bar{x}, 0, 0)$ is, for $\xi_n \in N_{\bar{x}}\mathcal{M}$,

$$\mathrm{D}_{\eta_n} \Phi(\bar{x}, 0, 0)[\xi_n] = \mathrm{D}\Psi(\bar{x})[\xi_n].$$

Since $\bar{x} \in \mathcal{M}$, $\mathrm{D}_{\eta_n} \Phi(\bar{x}, 0, 0)$ is surjective from $N_{\bar{x}}\mathcal{M}$ to $\mathbb{R}^{n-p}$ so its a bijection. The implicit function theorem provides the existence of neighborhoods $\mathcal{N}_{\bar{x}}^1 \subset \mathcal{M}$, $\mathcal{N}_0^2 \subset \cup_{x \in \mathcal{M}} T_x\mathcal{M}$, $\mathcal{N}_0^3 \subset \cup_{x \in \mathcal{M}} N_x\mathcal{M}$ and a unique $\mathcal{C}^k$ function $\eta_n : \mathcal{N}_{\bar{x}}^1 \times \mathcal{N}_0^2 \to \mathcal{N}_0^3$ such that, for all $x \in \mathcal{N}_{\bar{x}}^1$, $\eta_t \in \mathcal{N}_0^2$ and $\eta_n \in \mathcal{N}_0^3$, $\eta_n(\bar{x}, 0) = 0$ and

$$\Phi(x, \eta_t, \eta_n(x, \eta_t)) = 0 \Leftrightarrow x + \eta_t + \eta_n(x, \eta_t) \in \mathcal{M}.$$

It also provides an expression for the partial derivative of η_n at $(x, 0)$ along η_t: for $\xi_t \in T_x\mathcal{M}$,

$$\mathrm{D}_{\eta_t} \eta_n(x, 0)[\xi_t] = -\left[\mathrm{D}_{\eta_n} \Phi(x, 0, 0)\right]^{-1} \mathrm{D}_{\eta_t} \Phi(x, 0, 0)[\xi_t].$$

As noted before, $\mathrm{D}_{\eta_n} \Phi(x, 0, 0)$ is bijective since $x \in \mathcal{M}$. Besides, $\mathrm{D}_{\eta_t} \Phi(x, 0, 0) = \mathrm{D}\Phi(x)[\xi_t] = 0$ since $T_x\mathcal{M}$ identifies as the kernel of $\mathrm{D}\Phi(x)$. Thus $\mathrm{D}_{\eta_t} \eta_n(x, 0) = 0$.

Now, define a map $\mathrm{R} : \mathcal{N}_{\bar{x}}^1 \times \mathcal{N}_0^2 \to \mathcal{M}$ by $\mathrm{R}(x, \eta_t) = x + \eta_t + \eta_n(x, \eta_t)$. This map has degree of smoothness $\mathcal{C}^k$ since η_n is $\mathcal{C}^k$, satisfies $\mathrm{R}(x, 0) = x$ since $\eta_n(x, 0) = 0$ and satisfies $\mathrm{D}_{\eta_t} \eta_n(x, 0) = I + \mathrm{D}_{\eta_t}(x, 0) = I$. Thus R defines a retraction on a neighborhood of $(\bar{x}, 0)$.

We turn to show the second-order property of R. Consider the smooth curve c defined as $c(t) = \mathrm{R}(x, t\eta)$ for some $x \in \mathcal{N}_{\bar{x}}^1$, $\eta_t \in T_x\mathcal{M} \cap \mathcal{N}_0^2$. It's first derivative writes

$$c'(t) = \eta + \mathrm{D}_{\eta_t}\, \eta_n(x, t\eta)[\eta] = \eta.$$

The acceleration of the curve c is obtained by computing the derivative of $c'(\cdot)$ in the ambient space and then projecting onto $T_x\mathcal{M}$. Thus $c''(t) = 0$ and in particular, $c''(0) = 0$ which makes R a second-order retraction. □

A.1.3 *Euclidean spaces and manifolds, back and forth*

We show here that, locally, the Euclidean and Riemannian distances are equivalent.

This result is used to prove Theorem B.1 and Lemma A.4.

Lemma A.3 . *Consider a point $\bar{x}$ of a Riemannian manifold $\mathcal{M}$, equipped with a retraction* R *such that $\mathrm{R}_{\bar{x}}$ is $\mathcal{C}^2$. For any $\varepsilon > 0$, there exists a neighborhood $\mathcal{N}_{\bar{x}}$ of $\bar{x}$ in $\mathcal{M}$ such that*

$$(1-\varepsilon)\,\mathrm{dist}_{\mathcal{M}}(x, \bar{x}) \leq \|\,\mathrm{R}_{\bar{x}}^{-1}(x)\| \leq (1+\varepsilon)\,\mathrm{dist}_{\mathcal{M}}(x, \bar{x}) \qquad \textit{for all } x \in \mathcal{N}_{\bar{x}}.$$

where $\mathrm{R}_{\bar{x}}^{-1} : \mathcal{M} \to T_{\bar{x}}\mathcal{M}$ is the smooth inverse of $\mathrm{R}_{\bar{x}}$ defined locally around $\bar{x}$.

Proof. The retraction at $\bar{x}$ can be inverted locally around 0. Indeed, as $\mathrm{D}\,\mathrm{R}_{\bar{x}}(0_{T_{\bar{x}}\mathcal{M}}) = I$ is invertible and $\mathrm{R}_{\bar{x}}$ is $\mathcal{C}^2$, the implicit function theorem provides the existence of a $\mathcal{C}^2$ inverse function $\mathrm{R}_{\bar{x}}^{-1} : \mathcal{M} \to T_{\bar{x}}\mathcal{M}$ defined locally around $\bar{x}$. Furthermore, one shows by differentiating the relation $\mathrm{R}_{\bar{x}} \circ \mathrm{R}_{\bar{x}}^{-1}$ that the differential of $\mathrm{R}_{\bar{x}}^{-1}$ at $\bar{x}$ is the identity.

We consider the function $f : \mathcal{M} \to \mathbb{R}$ defined by $f(x) = \|\log_{\bar{x}}(x)\| - \|\,\mathrm{R}_{\bar{x}}^{-1}(x)\|$. Clearly $f(\bar{x}) = 0$, and $\mathrm{D}\,f(\bar{x}) = 0$ as the differentials of both $\mathrm{R}_{\bar{x}}^{-1}$ and logarithm at $\bar{x}$ are the identity. In local coordinates $\hat{x} = \log_{\bar{x}} x$ around $\bar{x}$, f is represented by the function $\hat{f} = f \circ \exp_{\bar{x}} : T_{\bar{x}}\mathcal{M} \to \mathbb{R}$. As $\hat{f}(\hat{\bar{x}}) = 0$, $\mathrm{D}\,\hat{f}(\hat{\bar{x}}) = 0$ and $\hat{f}$ is $\mathcal{C}^2$, there exists some $C > 0$ such that

$$-C\|\hat{x} - \hat{\bar{x}}\|^2 \leq \hat{f}(\hat{x}) \leq C\|\hat{x} - \hat{\bar{x}}\|^2 \qquad \text{in a neighborhood } \hat{\mathcal{N}} \text{ of } \hat{\bar{x}},.$$

For any $\varepsilon > 0$, by taking a small enough neighborhood $\hat{\mathcal{N}}' \subset \hat{\mathcal{N}}$, there holds

$$-\varepsilon\|\hat{x} - \hat{\bar{x}}\| \leq \hat{f}(\hat{x}) \leq \varepsilon\|\hat{x} - \hat{\bar{x}}\|.$$

Thus for all x in $\mathcal{N}_{\bar{x}} = \mathrm{R}_{\bar{x}}(\hat{\mathcal{N}}')$,

$$-\varepsilon\|\log_{\bar{x}}(x)\| \leq \|\log_{\bar{x}}(x)\| - \|\,\mathrm{R}_{\bar{x}}^{-1}(x)\| \leq \varepsilon\|\log_{\bar{x}}(x)\|,$$

as $\hat{x} = \log_{\bar{x}}(x)$, $\hat{\bar{x}} = 0$. We conclude with $\mathrm{dist}_{\mathcal{M}}(x, \bar{x}) = \|\hat{x} - \hat{\bar{x}}\| = \|\log_{\bar{x}}(x)\|$. □

Used in to prove Theorem 3.3.

Lemma A.4 . *Consider a point $\bar{x}$ of a Riemannian manifold $\mathcal{M}$. For any $\varepsilon > 0$, there exists a neighborhood $\mathcal{N}_{\bar{x}}$ of $\bar{x}$ in $\mathcal{M}$ such that, for all $x \in \mathcal{N}_{\bar{x}}$,*

$$(1-\varepsilon)\,\mathrm{dist}_{\mathcal{M}}(x, \bar{x}) \leq \|x - \bar{x}\| \leq (1+\varepsilon)\,\mathrm{dist}_{\mathcal{M}}(x, \bar{x}),$$

where $\|x - \bar{x}\|$ is the Euclidean distance in the ambient space.

Proof. Let $\bar{x}$, x denote two close points on $\mathcal{M}$. Consider the tangential retraction introduced in Proposition A.2. As a retraction, it satisfies:

$$\mathrm{R}_{\bar{x}}(\eta) = \mathrm{R}_{\bar{x}}(0) + \mathrm{D}\,\mathrm{R}_{\bar{x}}(0)[\eta] + \mathcal{O}(\|\eta\|^2) = \bar{x} + \mathcal{O}(\|\eta\|^2).$$

Taking $x = \mathrm{R}_{\bar{x}}(\eta)$ allows to write $x = \bar{x} + \mathcal{O}(\|\mathrm{R}_{\bar{x}}^{-1}(x)\|^2)$, so that for any small $\varepsilon_1 > 0$, there exists a small enough neighborhood $\mathcal{N}_{\bar{x}}^1 \subset \mathcal{N}_{\bar{x}}$ of $\bar{x}$ in $\mathcal{M}$ such that

$$(1-\varepsilon_1)\|\mathrm{R}_{\bar{x}}^{-1}(x)\| \leq \|x - \bar{x}\| \leq (1+\varepsilon_1)\|\mathrm{R}_{\bar{x}}^{-1}(x)\|.$$

By Lemma A.3, for $\varepsilon_2 > 0$ small enough, there exists a neighborhood $\mathcal{N}_{\bar{x}}^2 \subset \mathcal{N}_{\bar{x}}$ of $\bar{x}$ such that,

$$(1-\varepsilon_2)\operatorname{dist}_{\mathcal{M}}(x,\bar{x}) \leq \|\mathrm{R}_{\bar{x}}^{-1}(x)\| \leq (1+\varepsilon_2)\operatorname{dist}_{\mathcal{M}}(x,\bar{x}). \qquad \square$$

With ε_1, ε_2 such that $1-\varepsilon = (1-\varepsilon_1)(1-\varepsilon_2)$, we combine the two estimates to conclude.

A.2 TECHNICAL RESULTS ON THE PROXIMAL GRADIENT.

Out of completeness, we present below basic results about the proximal gradient along with their proofs. We refer to Beck (2017, Chap. 10) for a general reference on the topic.

This first lemma ensures functional descent of the proximal gradient, as soon as the stepsize γ is lower than $1/L$.

This result is used to prove Theorem 3.2.

Lemma A.5 (Functional descent). *Let Assumption 3.1 hold. For any $y \in \mathbb{R}^n$, $\gamma \in \mathbb{R}_+$, and $x \in \operatorname{prox}_{\gamma g}(y - \gamma\nabla f(y))$, we have*

$$F(x) + \frac{1-\gamma L}{2\gamma}\|x-y\|^2 \leq F(y).$$

Proof. By definition,

$$\begin{aligned} x &\in \operatorname*{arg\,min}_{u\in\mathbb{R}^n} \left(g(u) + \frac{1}{2\gamma}\|u - (y - \gamma\nabla f(y))\|^2 \right) \\ &= \operatorname*{arg\,min}_{u\in\mathbb{R}^n} \underbrace{\left(f(y) + \langle \nabla f(y), u - y\rangle + g(u) + \frac{1}{2\gamma}\|u-y\|^2 \right)}_{\triangleq s_y(u)} \end{aligned}$$

and the optimality of x implies that $s_y(x) \leq s_y(y)$, i.e.,

$$f(y) + \langle \nabla f(y), x - y\rangle + g(x) + \frac{1}{2\gamma}\|x-y\|^2 \leq f(y) + g(y).$$

Finally, the L-Lipschitz continuity the gradient of f implies that $f(x) \leq f(y) + \langle \nabla f(y), x-y\rangle + L/2\|y-x\|^2$ (Bauschke and Combettes, 2017, Th. 18.15). Combined with the previous equation, this yields the result. $\square$

This second lemma links the output of the proximal gradient operator and the subdifferential of F, see Bolte et al. (2015, Prop. 13).

This result is used to prove Theorem 3.2.

Lemma A.6 (Bound on distance to subdifferential). *Let Assumption 3.1 hold. For any $y \in \mathbb{R}^n$, $\gamma \in \mathbb{R}_+$, and $x = \operatorname{prox}_{\gamma g}(y - \gamma\nabla f(y))$, we have*

$$\operatorname{dist}(0, \partial F(x)) \leq \frac{L\gamma + 1}{\gamma}\|y - x\|.$$

Proof. The first order optimality condition defining x are

$$0 \in \partial g(x) + \frac{1}{\gamma}(x - (y - \gamma \nabla f(y))) \Leftrightarrow 0 \in \partial g(x) + \nabla f(x) + \frac{x-y}{\gamma} + \nabla f(y) - \nabla f(x),$$

which can be rewritten as $\frac{y-x}{\gamma} + \nabla f(x) - \nabla f(y) \in \partial F(x)$. Using the L-Lipschitz continuity of ∇f yields the following bound:

$$\operatorname{dist}(0, \partial F(x)) \leq \left\| \frac{y-x}{\gamma} + \nabla f(x) - \nabla f(y) \right\| \leq \left(\frac{1}{\gamma} + L \right) \|y - x\|. \qquad \square$$

Finally, we show below that any critical point of the proximal-gradient minimization problem is actually a strong local minimizer, provided that g is prox-regular there. This result appears more or less explicitly in some articles, including Daniilidis et al. (2006).

This result is used to prove Theorem 3.1.

Lemma A.7 . *Let f and g denote two functions and $\bar{x}$, $\bar{y}$ two points such that f is differentiable at $\bar{y}$ and g is r-prox-regular at $\bar{x}$ for subgradient $\frac{1}{\gamma}(\bar{y} - \bar{x}) - \nabla f(\bar{y}) \in \partial g(\bar{x})$ with $\gamma \in (0, 1/r)$. Then, the function $\rho_{\bar{y}} : x \mapsto g(x) + \frac{1}{2\gamma}\|\bar{y} - \gamma \nabla f(\bar{y}) - x\|^2$ satisfies*

$$\rho_{\bar{y}}(x) \geq \rho_{\bar{y}}(\bar{x}) + \frac{1}{2}\left(\frac{1}{\gamma} - r \right) \|x - \bar{x}\|^2, \qquad \textit{for all } x \textit{ near } \bar{x}.$$

Proof. Prox-regularity of g at $\bar{x}$ with subgradient $\frac{1}{\gamma}(\bar{y} - \gamma \nabla f(\bar{y}) - \bar{x}) \in \partial g(\bar{x})$ writes

$$g(x) \geq g(\bar{x}) + \frac{1}{\gamma}\langle \bar{y} - \gamma \nabla f(\bar{y}) - \bar{x}, x - \bar{x} \rangle - \frac{r}{2}\|x - \bar{x}\|^2.$$

The identity $2\langle b - a, c - a \rangle = \|b - a\|^2 + \|c - a\|^2 - \|b - c\|^2$ applied to the previous scalar product yields:

$$g(x) \geq g(\bar{x}) + \frac{1}{2\gamma}\|\bar{y} - \gamma \nabla f(\bar{y}) - \bar{x}\|^2 + \frac{1}{2\gamma}\|x - \bar{x}\|^2 - \frac{1}{2\gamma}\|\bar{y} - \gamma \nabla f(\bar{y}) - x\|^2 - \frac{r}{2}\|x - \bar{x}\|^2,$$

which rewrites

$$\underbrace{g(x) + \frac{1}{2\gamma}\|\bar{y} - \gamma \nabla f(\bar{y}) - x\|^2}_{=\rho_{\bar{y}}(x)} \geq \underbrace{g(\bar{x}) + \frac{1}{2\gamma}\|\bar{y} - \gamma \nabla f(\bar{y}) - \bar{x}\|^2}_{=\rho_{\bar{y}}(\bar{x})} + \frac{1}{2}\left(\frac{1}{\gamma} - r \right) \|x - \bar{x}\|^2,$$

which is the claimed inequality. $\square$

B

TECHNICAL RESULTS ON RIEMANNIAN OPTIMIZATION

In this appendix, we provide some results on Riemannian optimization that are important in our developments, and that we have not been able to locate in the existing literature.

B.1 ACCEPTATION OF THE UNIT STEPSIZE BY RIEMANNIAN LINE SEARCH ALGORITHMS

We provide here a technical result used in the proofs of Section 3.5. Theorem B.1 ensures that a step providing superlinear improvement towards a strong minimizer locally decreases function value. Such a step is thus directly acceptable by an Armijo line search. The result and proof adapt Bonnans et al. (2006, Th. 4.16) to the Riemannian setting.

Theorem B.1 (Soundness of the Riemannian line search). *Consider a manifold $\mathcal{M}$ equipped with a retraction R and a twice differentiable function $f : \mathcal{M} \to \mathbb{R}$ that admits a strong local minimizer $x^\star$, that is, a point such that $\mathrm{Hess}\, f(x^\star)$ is positive definite. If x is close to $x^\star$, η brings a superlinear improvement towards $x^\star$, that is $\mathrm{dist}_{\mathcal{M}}(\mathrm{R}_x(\eta), x^\star) = o(\mathrm{dist}_{\mathcal{M}}(x, x^\star))$ as $x \to x^\star$, and $0 < m_1 < 1/2$, then η is acceptable by the Armijo rule (3.12) with unit stepsize $\alpha = 1$.*

Proof. Let $x, \eta \in T\mathcal{B}$ denote a pair such that x is close to $x^\star$ and $\mathrm{dist}_{\mathcal{M}}(\mathrm{R}_x(\eta), x^\star) = o(\mathrm{dist}_{\mathcal{M}}(x, x^\star))$. For convenience, let $x_+ = \mathrm{R}_x(\eta)$ denote the next point.

Following Absil et al. (2009a) (see e.g. the proof of Th. 6.3.2), we work in local coordinates around $x^\star$, representing any point $x \in \mathcal{M}$ by $\widehat{x} = \log_{x^\star}(x)$ and any tangent vector $\eta \in T_x\mathcal{M}$ by $\widehat{\eta}_x = \mathrm{D}\log_{x^\star}(x)[\eta]$. The function f is represented by $\widehat{f} = f \circ \exp_{x^\star} : T_{x^\star}\mathcal{M} \to \mathbb{R}$. Defining the coordinates via the logarithm grants the useful property that the Riemannian distance of any two points $x, y \in \mathcal{M}$ matches the euclidean distance between their representatives: $\mathrm{dist}_{\mathcal{M}}(x, y) = \|\widehat{x} - \widehat{y}\|$. Besides, there holds

$$\mathrm{D}\, f(x)[\eta] = \mathrm{D}\, \widehat{f}(\widehat{x})[\widehat{\eta}] \quad \text{and} \quad \mathrm{Hess}\, f(x)[\eta, \eta] = \mathrm{D}\, \widehat{f}(\widehat{x})[\widehat{\eta}, \widehat{\eta}]. \tag{B.1}$$

Indeed, $\mathrm{D}\, f(x)[\eta] = (f \circ \gamma)'(0)$ and $\mathrm{Hess}\, f(x)[\eta, \eta] = (f \circ \gamma)''(0)$, where γ denotes the geodesic curve defined by $\widehat{\gamma}(t) = \widehat{x} + t\widehat{\eta}$. Using $f \circ \gamma = \widehat{f} \circ \widehat{\gamma}$, one obtains the result.

Step 1. We derive an approximation of $\mathrm{D}\, f(x)[\eta] = \langle \mathrm{grad}\, f(x), \eta \rangle$ in terms of $\mathrm{D}^2\, \widehat{f}(\widehat{x^\star})[\widehat{x} - \widehat{x^\star}]^2$. To do so, we go through the intermediate quantity $\mathrm{D}\, \widehat{f}(\widehat{x})[\widehat{x_+} - \widehat{x}]$, and handle precisely the $o(\cdot)$ terms. By smoothness of $\widehat{f}$ and since $\mathrm{D}\, \widehat{f}(\widehat{x^\star}) = 0$, Taylor's formula for $\mathrm{D}\, \widehat{f}$ writes

$$\mathrm{D}\, \widehat{f}(\widehat{x})[\widehat{x_+} - \widehat{x}] = \mathrm{D}^2\, \widehat{f}(\widehat{x^\star})[\widehat{x_+} - \widehat{x}, \widehat{x} - \widehat{x^\star}] + o(\|\widehat{x} - \widehat{x^\star}\|^2)$$

$$= -\mathrm{D}^2 \widehat{f}(\widehat{x^\star})[\widehat{x} - \widehat{x^\star}]^2 + \mathrm{D}^2 \widehat{f}(\widehat{x^\star})[\widehat{x_+} - \widehat{x^\star}, \widehat{x} - \widehat{x^\star}] + o(\|\widehat{x} - \widehat{x^\star}\|^2)$$
$$= -\mathrm{D}^2 \widehat{f}(\widehat{x^\star})[\widehat{x} - \widehat{x^\star}]^2 + o(\|\widehat{x} - \widehat{x^\star}\|^2),$$

where, in the last step, we used that $\|\widehat{x_+} - \widehat{x^\star}\| = o(\|\widehat{x} - \widehat{x^\star}\|)$ to get that $\|\mathrm{D}^2 \widehat{f}(\widehat{x^\star})[\widehat{x_+} - \widehat{x^\star}, \widehat{x} - \widehat{x^\star}]\| = \|\mathrm{D}^2 \widehat{f}(\widehat{x^\star})\| \|\widehat{x_+} - \widehat{x^\star}\| \|\widehat{x} - \widehat{x^\star}\| = o(\|\widehat{x} - \widehat{x^\star}\|^2)$. We now turn to show that $\mathrm{D}\widehat{f}(\widehat{x})[\widehat{x_+} - \widehat{x}]$ behaves as $\mathrm{D} f(x)[\eta]$ up to $o(\|\widehat{x_+} - \widehat{x}\|^2)$. Since $\mathrm{D} f(x)[\eta] = \mathrm{D}\widehat{f}(\widehat{x})[\widehat{\eta}]$ by (B.1), there holds:

$$\|\mathrm{D} f(x)[\eta] - \mathrm{D}\widehat{f}(\widehat{x})[\widehat{x_+} - \widehat{x}]\| = \|\mathrm{D}\widehat{f}(\widehat{x})[\widehat{\eta} - (\widehat{x_+} - \widehat{x})]\| \leq \|\mathrm{D}\widehat{f}(\widehat{x})\| \|\widehat{\eta} - (\widehat{x_+} - \widehat{x})\|.$$

As f is twice differentiable and exp is $\mathcal{C}^\infty$, $\widehat{f}$ is twice differentiable as well. In particular its derivative is locally Lipschitz continuous, so that for $\widehat{x}$ near $\widehat{x^\star}$, we obtain a first estimate:

$$\|\mathrm{D}\widehat{f}(\widehat{x})\| = \|\mathrm{D}\widehat{f}(\widehat{x}) - \mathrm{D}\widehat{f}(\widehat{x^\star})\| = \mathcal{O}(\|\widehat{x} - \widehat{x^\star}\|).$$

Besides, the following estimate holds $\|\widehat{\eta} - (\widehat{x_+} - \widehat{x})\| = o(\|\widehat{x} - \widehat{x^\star}\|)$. Indeed, as the function $\log_{x^\star} \circ \mathrm{R}_x : T_x\mathcal{M} \to T_{x^\star}\mathcal{M}$ is differentiable, there holds for $\eta \in T_x\mathcal{M}$ small,

$$\log_{x^\star}(\mathrm{R}_x(\eta)) = \log_{x^\star}(\mathrm{R}_x(0)) + \mathrm{D}\log_{x^\star}(\mathrm{R}_x(0))[\mathrm{D}\,\mathrm{R}_x(0)[\eta]] + o(\|\eta\|),$$

which simplifies to $\widehat{x_+} = \widehat{x} + \widehat{\eta} + o(\|\eta\|)$. Lemma A.3 allows to write $\|\eta\| = \|\mathrm{R}_x^{-1}(x_+)\| = \mathcal{O}(\mathrm{dist}_{\mathcal{M}}(x, x_+))$. Using the triangular inequality and the assumption that $\mathrm{dist}_{\mathcal{M}}(x_+, x^\star) = o(\mathrm{dist}_{\mathcal{M}}(x, x^*))$ we get

$$\mathrm{dist}_{\mathcal{M}}(x, x_+) \leq \mathrm{dist}_{\mathcal{M}}(x, x^\star) + \mathrm{dist}_{\mathcal{M}}(x^\star, x_+) = \mathcal{O}(\mathrm{dist}_{\mathcal{M}}(x, x^*)) = \mathcal{O}(\|\widehat{x} - \widehat{x^\star}\|),$$

so that the second estimate holds.

Combining the two above estimates allows to conclude that

$$\|\mathrm{D} f(x)[\eta] - \mathrm{D}\widehat{f}(\widehat{x})[\widehat{x_+} - \widehat{x}]\| = o(\|\widehat{x} - \widehat{x^\star}\|^2),$$

so that overall,

$$\mathrm{D} f(x)[\eta] = \mathrm{D}\widehat{f}(\widehat{x})[\widehat{x_+} - \widehat{x}] + o(\|\widehat{x} - \widehat{x^\star}\|^2) = -\mathrm{D}^2 \widehat{f}(\widehat{x^\star})[\widehat{x} - \widehat{x^\star}]^2 + o(\|\widehat{x} - \widehat{x^\star}\|^2).$$

Using that $\|\widehat{x} - \widehat{x^\star}\| = \mathrm{dist}_{\mathcal{M}}(x, x^\star)$ and $\mathrm{D}^2 \widehat{f}(\widehat{x^\star}) = \mathrm{Hess}\, f(x^\star)$ (B.1), we obtain

$$\mathrm{D} f(x)[\eta] = -\mathrm{Hess}\, f(x^\star)[\widehat{x} - \widehat{x^\star}]^2 + o(\mathrm{dist}_{\mathcal{M}}(x, x^\star)^2). \tag{B.2}$$

Step 2. The function f admits a second-order development around $x^\star$: applying Eq. (2.3) with the exponential map $\exp_{x^\star}$ as a second-order retraction yields

$$f(x) = f(x^\star) + \frac{1}{2}\mathrm{Hess}\, f(x^\star)[\widehat{x} - \widehat{x^\star}]^2 + o(\mathrm{dist}_{\mathcal{M}}(x, x^\star)^2), \tag{B.3}$$

where we used that $\mathrm{dist}_{\mathcal{M}}(x, x^\star) = \|\log_{x^\star}(x) - \log_{x^\star}(x^\star)\|$. Denote $0 < l \leq L$ the lower and upper eigenvalues of $\mathrm{Hess}\, f(x^\star)$. The combination (B.3) + m_1(B.2) writes

$$f(x) + m_1 \mathrm{D} f(x)[\eta] = f(x^\star) + (\frac{1}{2} - m_1)\mathrm{Hess}\, f(x^\star)[\widehat{x} - \widehat{x^\star}]^2 + o(\mathrm{dist}_{\mathcal{M}}(x, x^\star)^2)$$

$$\geq f(x^\star) + (\frac{1}{2} - m_1)l \operatorname{dist}_{\mathcal{M}}(x, x^\star)^2 + o(\operatorname{dist}_{\mathcal{M}}(x, x^\star)^2),$$

Let $\varepsilon > 0$ such that $\frac{1}{2}L\varepsilon^2 < (\frac{1}{2} - m_1)l$. As $\operatorname{dist}_{\mathcal{M}}(x_+, x^\star) = o(\operatorname{dist}_{\mathcal{M}}(x, x^\star))$, for x close enough to $x^\star$ there holds $\operatorname{dist}_{\mathcal{M}}(x_+, x^\star) \leq \varepsilon \operatorname{dist}_{\mathcal{M}}(x, x^\star)$. Combining this with the second-order development of f at x_+, there holds:

$$\begin{aligned} f(x_+) &= f(x^\star) + \frac{1}{2}\operatorname{Hess} f(x^\star)[\widehat{x_+} - \widehat{x^\star}]^2 + o(\operatorname{dist}_{\mathcal{M}}(x_+, x^\star)^2) \\ &\leq f(x^\star) + \frac{1}{2}L \operatorname{dist}_{\mathcal{M}}(x_+, x^\star)^2 + o(\operatorname{dist}_{\mathcal{M}}(x_+, x^\star)^2) \\ &\leq f(x^\star) + \frac{1}{2}L\varepsilon^2 \operatorname{dist}_{\mathcal{M}}(x, x^\star)^2 + o(\operatorname{dist}_{\mathcal{M}}(x, x^\star)^2). \end{aligned}$$

Subtracting the two estimates yields

$$f(x_+) - (f(x) + m_1 \operatorname{D} f(x)[\eta]) \leq \left(\frac{1}{2}L\varepsilon^2 - (\frac{1}{2} - m_1)l\right) \operatorname{dist}_{\mathcal{M}}(x, x^\star)^2 + o(\operatorname{dist}_{\mathcal{M}}(x, x^\star)^2),$$

which ensures that the Armijo condition is satisfied. □

B.2 RIEMANNIAN DERIVATIVES OF THE NUCLEAR NORM

We compute in Lemma B.2 the derivatives of the matrices involved in the singular value decomposition. These results may be seen as part of folklore, but, up to our knowledge, there are not explicitly written in the literature. Based on these derivatives, we compute in Proposition B.3 the Riemannian gradient and Hessian of the trace-norm function.

Lemma B.2 . *Consider the manifold of fixed rank matrices $\mathcal{M}_r$, a pair $(x, \eta) \in T\mathcal{B}$ and a smooth curve $c : I \to \mathcal{M}_r$ such that $c(0) = x$, $c'(0) = \eta$. Besides, let $U(t)$, $\Sigma(t)$, $V(t)$ denote smooth curves of $St(m, r)$, $\mathbb{R}^{r\times r}$, $St(n, r)$ such that $\gamma(t) = U(t)\Sigma(t)V(t)^\top$. The derivatives of the decomposition factors at $t = 0$ write*

$$\begin{aligned} U' &= U\left(F \circ \left[U^\top \eta V \Sigma + \Sigma V^\top \eta^\top U\right]\right) + (I_m - UU^\top)\eta V \Sigma^{-1} \\ V' &= V\left(F \circ \left[\Sigma U^\top \eta V + V^\top \eta^\top U \Sigma\right]\right) + (I_n - VV^\top)\eta^\top U \Sigma^{-1} \\ \Sigma' &= I_k \circ \left[U^\top \eta V\right], \end{aligned}$$

where I_k is the identity of $\mathbb{R}^{k\times k}$, $\circ$ denotes the Hadamard product and $F \in \mathbb{R}^{r\times r}$ is such that $F_{ij} = 1/(\Sigma_{jj}^2 - \Sigma_{ii}^2)$ if $\Sigma_{jj} \neq \Sigma_{ii}$, and $F_{ij} = 0$ otherwise. Equivalently, when the tangent vector is represented as $\eta = UMV^\top + U_pV^\top + UV_p^\top$, the above expressions simplify to

$$\begin{aligned} U' &= U\left(F \circ \left[M\Sigma + \Sigma M^\top\right]\right) + U_p\Sigma^{-1} \\ V' &= V\left(F \circ \left[\Sigma M + M^\top \Sigma\right]\right) + V_p\Sigma^{-1} \\ \Sigma' &= I_k \circ M, \end{aligned}$$

Proof. We consider the curve γ and all components and derivatives at $t = 0$, therefore we don't mention evaluation time. Differentiating $\gamma = U\Sigma V^\top$ yields

$$\eta = U'\Sigma V^\top + U\Sigma' V^\top + U\Sigma V'^\top \tag{B.4}$$

As a tangent vector to the Stiefel manifold at point U, U' can be expressed as Absil et al. (2009a, Ex. 3.5.2)

$$U' = U\Omega_U + U_\perp B_U, \tag{B.5}$$

where $\Omega_U \in \mathbb{R}^{r\times r}$ is a skew-symmetric matrix, $B_U \in \mathbb{R}^{m-r\times m-r}$, and $U_\perp$ is any matrix such that $U^\top U_\perp = 0$ and $U_\perp^\top U_\perp = I_{m-r}$. Similarly, $V' = V\Omega_V + V_\perp B_V$, where $\Omega_V \in \mathbb{R}^{r\times r}$ is skew-symmetric, $B_V \in \mathbb{R}^{n-r\times n-r}$, and $V_\perp$ is any matrix such that $V^\top V_\perp = 0$ and $V_\perp^\top V_\perp = I_{n-r}$.

Computing $U^\top \times$ (B.4) $\times V$ yields

$$U^\top \eta V = \Omega_U \Sigma + \Sigma' + \Sigma\Omega_V^\top.$$

Looking at the diagonal elements of this equation yields the derivative of the diagonal component of η. This is done by taking the Hadamard product of both sides of previous equation with the identity matrix of $\mathbb{R}^{r\times r}$, and writes

$$\Sigma' = I_r \circ \left[U^\top \eta V\right].$$

The off-diagonal elements of this equation write

$$\bar{I}_r \circ \left[U^\top \eta V\right] = \Omega_U \Sigma + \Sigma\Omega_V^\top, \tag{B.6}$$

where $\bar{I}_r$ has zeros on the diagonal and ones elsewhere. Adding (B.6)Σ and Σ(B.6)$^\top$ yields

$$\bar{I}_r \circ \left[U^\top \eta V\Sigma + \Sigma V^\top \eta^\top U\right] = \Omega_U \Sigma^2 - \Sigma^2 \Omega_U,$$

which decouples coefficient-wise. At coefficient (ij), with $i \neq j$,

$$\left[U^\top \eta V\Sigma + \Sigma V^\top \eta^\top U\right]_{ij} = [\Omega_U]_{ij}\left(\Sigma_{jj}^2 - \Sigma_{ii}^2\right),$$

hence $\Omega_U = F \circ \left[U^\top \eta V\Sigma + \Sigma V^\top \eta^\top U\right]$, where $F \in \mathbb{R}^{m-r\times r}$ has zeros on the diagonal and for $i \neq j$, $F_{ij} = 1/(\Sigma_{jj}^2 - \Sigma_{ii}^2)$ if $\Sigma_{jj}^2 \neq \Sigma_{ii}^2$, 0 otherwise. Besides, left-multiplying (B.4) by $U_\perp^\top$ yields $U_\perp^\top \eta = U_\perp^\top U'\Sigma V^\top$, which rewrites, using the decomposition (B.5) of U', as $U_\perp^\top \eta = B_U \Sigma V^\top$. Hence $B_U = U_\perp^\top \eta V \Sigma^{-1}$ and we get the complete expression for U' by assembling the expressions of Ω_U and B_U with the decomposition (B.5). The term $U_\perp^\top U_\perp$ is eliminated using that $U^\top U + U_\perp^\top U_\perp = I_m$.

Let's follow the same steps to get expressions for V'. Adding Σ(B.6) and (B.6)$^\top\Sigma$ yields

$$\bar{I}_r \circ \left[\Sigma U^\top \eta V + V^\top \eta^\top U\Sigma\right] = \Omega_V \Sigma^2 - \Sigma^2 \Omega_V,$$

from which we get $\Omega_V = F \circ \left[\Sigma U^\top \eta V + V^\top \eta^\top U\Sigma\right]$. Besides, right-multiplying (B.4) by $V_\perp$ yields $\eta V_\perp = U\Sigma V'^\top V_\perp$, which rewrites using the decomposition $V' = V\Omega_V + V_\perp B_V$ as $\eta V_\perp = U\Sigma B_V^\top$. Hence $B_V = V_\perp^\top \eta^\top U\Sigma^{-1}$, and we get the claimed formula by eliminating the $V_\perp$ terms with $V^\top V + V_\perp^\top V_\perp = I_n$. The simplified expressions are obtained using that $U^\top U = I_m$, $U^\top U_p = 0$, $V^\top V = I_n$ and $V^\top V_p = 0$. □

We are now ready to give the expression of the Riemannian gradient and Hessian of the nuclear norm.

Proposition B.3 . *The nuclear norm $g = \|\cdot\|_*$ restricted to $\mathcal{M}_r$ is $\mathcal{C}^2$ and admits a smooth second-order development of the form (2.3) near any point $x = U\Sigma V^\top \in \mathcal{M}_r$. Denoting $\eta = UMV^\top + U_pV^\top + UV_p^\top \in T_x\mathcal{M}_r$ a tangent vector, there holds:*

$$\begin{aligned}\operatorname{grad} g(x) &= UV^\top \\ \operatorname{Hess} g(x)[\eta] &= U\left[\tilde{F} \circ (M - M^\top)\right]V^\top + U_p\Sigma^{-1}V^\top + U\Sigma^{-1}V_p^T,\end{aligned}$$

where $\circ$ denotes the Hadamard product and $\tilde{F} \in \mathbb{R}^{r\times r}$ is such that $\tilde{F}_{ij} = 1/(\Sigma_{jj} + \Sigma_{ii})$ if $\Sigma_{jj} \neq \Sigma_{ii}$, and $\tilde{F}_{ij} = 0$ otherwise.

Proof. Let $c : I \to \mathcal{M}_r$ denote a smooth curve over $\mathcal{M}_r$ such that $\gamma(0) = x$ and $\gamma'(0) = \eta$, and consider $\varphi = \|c(\cdot)\|_* : I \to \mathbb{R}$. Writing the decomposition $c(t) = U(t)\Sigma(t)V(t)^\top$, for $U(t)$, $\Sigma(t)$, $V(t)$ smooth curves of $St(m,r)$, $\mathbb{R}^{r\times r}$, $St(n,r)$ allows to write $\varphi(t) = \operatorname{Tr}(\Sigma(t))$. Applying Lemma B.2 yields

$$\varphi'(0) = \operatorname{Tr}(\Sigma'(0)) = \operatorname{Tr}(U^\top\eta V) = \operatorname{Tr}(\eta VU^\top) = \langle \eta, UV^\top\rangle,$$

so that $\operatorname{grad} g(x) = UV^\top \in T_X\mathcal{M}$.

In order to obtain the Riemannian Hessian, let $\bar{Z} : I \to \mathbb{R}^n$ denote a smooth extension of $\operatorname{grad} g(c(\cdot))$, defined by $\bar{Z}(t) = U(t)V(t)^\top$. The Riemannian Hessian is then obtained as $\operatorname{Hess} g(x)[\eta] = \operatorname{proj}_x \bar{Z}'(0)$. The derivative of $\bar{Z}$ at 0 is simply $\bar{Z}'(0) = U'V^\top + UV'^\top$ and thus writes, applying Lemma B.2

$$\begin{aligned}\bar{Z}'(0) = U\left(F \circ \left[M\Sigma + \Sigma M^\top\right]\right)V^\top + U_p\Sigma^{-1}V^\top \\ + U\left(F \circ \left[\Sigma M + M^\top\Sigma\right]\right)^\top V^\top + U\Sigma^{-1}V_p^\top\end{aligned}$$

This expression simplifies to the statement by using the fact that F is antisymmetric and applying the identity $(A \circ B)^\top = A^\top \circ B^\top$. □

C

THE MAXIMUM AND MAXIMUM EIGENVALUE SATISFY THE NORMAL ASCENT AND CURVE PROPERTIES

In Chapter 3, we built a proximal identification scheme for additive functions that satisfy some properties. We show here that the maximum and the maximum eigenvalue indeed meet the normal ascent Property 4.1 and curve properties Property 4.2.

We begin with a lemma that simplifies verification of Property 4.2.

Lemma C.1 . *Consider a function g, partly smooth at a point $\bar{y}$ relative to a manifold $\mathcal{M}^g$, and a smooth application $e : \mathcal{N}_{\bar{y}} \times [0,T] \to \mathcal{M}^g$ defined for a neighborhood $\mathcal{N}_{\bar{y}}$ of $\bar{y}$ and $T > 0$ such that $e(y,0) = \mathrm{proj}_{\mathcal{M}^g}(y)$, $\frac{d}{dt}e(y,t)|_{t=0} = -\operatorname{grad} g(\mathrm{proj}_{\mathcal{M}^g}(y))$. If $\mathrm{D}\left(t \mapsto \mathrm{proj}_{N_{e(y,t)}\mathcal{M}^g}(\mathrm{proj}_{\mathcal{M}}(y) - y)\right) = 0$ for all $y \in \mathcal{N}_{\bar{y}}$, then g satisfies Property 4.2 at point $\bar{y}$.*

Proof. We denote $\theta(y,t) = \mathrm{proj}_{N_{e(y,t)}\mathcal{M}^g}(e(y,t) - y)$. First,

$$\frac{d}{dt}\theta(y,t)|_{t=0} = \mathrm{D}\left(t \mapsto \mathrm{proj}_{N_{e(y,t)}\mathcal{M}^g}(\mathrm{proj}_{\mathcal{M}^g}(y) - y)\right) + \mathrm{proj}_{N_{\mathrm{proj}_{\mathcal{M}}(y)}\mathcal{M}^g}\left(\mathrm{D}(t \mapsto (e(y,t) - y))(0)\right),$$

where the first term is null by assumption and the second is also null since it is the normal projection of the tangent vector $\operatorname{grad} g(\mathrm{proj}_{\mathcal{M}^g}(y))$. Thus, $\frac{d}{dt}\theta(y,t)|_{t=0} = 0$. Using this fact and smoothness of θ, Taylor's theorem with Lagrange remainder yields, for all $y \in \mathcal{N}_{\bar{y}}$, the existence of $\bar{t} \in [0,T]$ such that, for all $t \in [0,T]$,

$$\theta(y,t) = \theta(y,0) + \frac{t^2}{2}\frac{d^2}{dt^2}\theta(y,\bar{t}).$$

Therefore, for all $y \in \mathcal{N}_{\bar{y}}$ and $t \in [0,T]$,

$$\|\theta(y,t)\| \leq \|\theta(y,0)\| + \frac{t^2}{2}\sup_{\bar{t}\in[0,T]}\frac{d^2}{dt^2}\theta(y,\bar{t}) \leq \|\theta(y,0)\| + t^2\tilde{L},$$

where $\tilde{L} = \sup_{y\in\mathcal{N}_{\bar{y}}}\sup_{\bar{t}\in[0,T]}\frac{d^2}{dt^2}\theta(y,\bar{t})$. □

We can now proceed with the proof of Lemma 4.3, divided into two parts corresponding to the two cases of the result. The case $g = \max$ comes easily, due to the polyhedral nature of the function.

Lemma C.2 . *Consider $g = \max$, a point $\bar{y} \in \mathbb{R}^m$ and the corresponding structure manifold $\mathcal{M}_I^{\max}$ (of Example 4.3). Then Properties 4.1 and 4.2 hold at $\bar{y}$.*

Proof. Normal ascent Take $y \in \mathcal{M}_I^{\max}$ for some active indices $I \subset \{1,\ldots,m\}$. A normal direction $d \in N_y\mathcal{M}_I^{\max}$ is such that $d_i = 0$ for $i \notin I$ and $\sum_{i\in I} d_i = 0$. Thus

$\max(y + td) = y_i + td_i$ with $i = \arg\max_i d_i$, and $\mathrm{D}\max(y)[d] = \lim_{t \searrow 0}(\max(y + td) - \max(y))/t = d_i > 0$ for all $d \neq 0$.

Curve assumption Since the structure manifold of max are affine subspaces, the normal spaces are equal at all points of the manifold. Therefore the derivative of the projection at a parametrized point is null and Lemma C.1 provides the result. □

The case $g = \lambda_{\max}$ is not difficult *per se*, but requires a precise description of the geometry of the maximum eigenvalue function and its structure manifolds; we refer to Shapiro and Fan (1995); Oustry (1999) for the derivation of these tools.

Lemma C.3 . *Consider $g = \lambda_{\max}$, a point $\bar{y} \in \mathbb{S}_m$ and the corresponding structure manifold $\mathcal{M}_r^{\lambda_{\max}}$ (of Example 4.4). Then Properties 4.1 and 4.2 hold at $\bar{y}$.*

Proof. Normal ascent Take $y \in \mathcal{M}_r^{\lambda_{\max}}$, let $U \in \mathbb{R}^{m \times r}$ denote a basis of the first eigenspace of matrix y and $d \in N_y\mathcal{M}_r^{\lambda_{\max}}$. The normal space at $y \in \mathcal{M}_r^{\lambda_{\max}}$ writes (Oustry, 1999, Th. 4.3, Cor. 4.8)

$$N_y\mathcal{M}_r^{\lambda_{\max}} = \{U(y)ZU(y)^\top, Z \in \mathbb{S}_r, \mathrm{trace}(Z) = 0\}.$$

Therefore, $d = UZU^\top$ for some $Z \in \mathbb{S}_r$ such that $\mathrm{trace}(Z) = 0$. Let $s = U(I/r + \alpha Z)U^\top$ where $\alpha > 0$ is small enough so that s is positive definite. Since s has also unit trace, it is a subgradient of $\lambda_{\max}$ at y (Oustry, 1999, Th. 4.1). Thus $\lambda'_{\max}(y; d) = \sup_{v \in \partial\lambda_{\max}(y)} \langle v, d\rangle \geq \langle s, d\rangle = \langle I/r + \alpha Z, Z\rangle = \alpha\|Z\|^2$. Hence $\lambda'_{\max}(y; d) > 0$ for any $d \in N_y\mathcal{M}_r^{\lambda_{\max}} \setminus \{0\}$.

Curve assumption Let $\bar{y} \in \mathcal{M}_r^{\lambda_{\max}}$. For any $y \in \mathbb{S}_m$, we denote by $P(y)$ the orthogonal projection on the eigenspace corresponding to the r largest eigenvalues of y (counting multiplicities). This operator is smooth. We can define a mapping $U : \mathbb{S}_m \to \mathbb{R}^{m \times r}$ such that: $U(y)^\top U(y) = I_r$, $P(y) = U(y)U(y)^\top$, U is smooth near our reference point $\bar{y}$ and its derivative at $\bar{y}$ satisfies $\mathrm{D}U(\bar{y})^\top U(\bar{y}) = 0$. The mapping U defines a smooth orthonormal basis of the eigenspace corresponding to the r largest eigenvalues (Shapiro and Fan, 1995, p. 557). Finally, for a point $y' \in \mathcal{M}_r^{\lambda_{\max}}$, the projection of $d \in \mathbb{S}_m$ on $N_{y'}\mathcal{M}_r^{\lambda_{\max}}$ writes

$$\mathrm{proj}_{N_{y'}\mathcal{M}_r^{\lambda_{\max}}}(d) = U(y')\left\{U(y')^\top dU(y') - \frac{1}{r}\mathrm{trace}(U(y')^\top dU(y'))I_r\right\}U(y')^\top.$$

Now, fix y near $\bar{y}$, consider the eigenbasis U with reference point $e(y, 0) = \mathrm{proj}_{\mathcal{M}_r^{\lambda_{\max}}}(y)$. Following Lemma C.1, let $v : t \mapsto \mathrm{proj}_{N_{e(y,t)}\mathcal{M}_r^{\lambda_{\max}}}(d)$ with $d = \mathrm{proj}_{\mathcal{M}_r^{\lambda_{\max}}}(y) - y$. We can now give an explicit expression of $v(t)$ and show that $\frac{d}{dt}v(0)$ is null. Denoting $U(t) = U(e(y, t))$, we have

$$v(t) = U(t)\underbrace{\left\{U(t)^\top dU(t) - \frac{1}{r}\mathrm{trace}(U(t)^\top dU(t))I_r\right\}}_{\triangleq \chi(t)}U(t)^\top.$$

First, as d is a normal vector to $\mathcal{M}_r^{\lambda_{\max}}$ at point $\mathrm{proj}_{\mathcal{M}_r^{\lambda_{\max}}}(y)$, there exists $Z \in \mathbb{S}_r$ such that $d = U(0)ZU(0)^\top$. Using that $\mathrm{D}U(0)^\top U(0) = 0$ yields

$$\mathrm{D}U(0)^\top dU(0) = \mathrm{D}U(0)^\top U(0)ZU(0)^\top U(0) = 0.$$

Then, one readily checks that $U(0)\,\mathrm{D}\,\chi(0)U(0) = 0$.

We turn to the term $\mathrm{D}\,U(0)\chi(0)U(0)^\top$. A quick computation from the eigen decomposition of y shows that d writes $U(0)ZU(0)^\top$, where Z is actually diagonal. Therefore, $\chi(0) = Z - (1/r)\,\mathrm{trace}(Z)I_r$ is a diagonal matrix, so that

$$\mathrm{D}\,U(0)\chi(0)U(0)^\top = \sum_{i=1}^{r} \chi(0)_{ii}\,\mathrm{D}\,U_i(0)U_i(0)^\top.$$

Following Shapiro and Fan (1995), the differential of $t \mapsto U(e(y,t))$ at $t = 0$ writes

$$\mathrm{D}\,U_i(0) = \sum_{k=r+1}^{m} \frac{1}{\lambda_1 - \lambda_k} U_k(0)U_k(0)^\top \eta U_i(0),$$

with $\eta = \mathrm{grad}\,\lambda_{\max}(\mathrm{proj}_{\mathcal{M}_r^{\lambda_{\max}}}(y))$. Using that $\lambda_{\max}(y) = (1/r)\sum_{i=1}^{r} U_i(y)^\top y U_i(y)$, we compute the Riemannian gradient: $\mathrm{grad}\,\lambda_{\max}(y) = (1/r)\sum_{i=1}^{r} U_i(y)^\top U_i(y)$ (see Boumal (2022, Sec. 7.7)). By orthogonality of the smooth basis of eigenvectors, the terms $U_k(0)^\top U_i(0)$ vanish for all $i \in \{1,\dots,r\}$ and $k \in \{r+1,\dots,m\}$. We get that $\mathrm{D}\,U(0)\chi(0)U(0)^\top = 0$, and thus that $\mathrm{D}\,\nu(0) = 0$. Thus, Lemma C.1 applies and yields the result. □

LIST OF FIGURES

LIST OF ALGORITHMS

COLOPHON

This manuscript was typeset with LaTeX 2$_\varepsilon$ using Hermann Zapf's Palatino type face (the actual Type 1 PostScript fonts used were URW Palladio L and FPL). The monospaced text (hyperlinks, etc.) was typeset in *Bera Mono*, originally developed by Bitstream, Inc. as "Bitstream Vera" (with Type 1 PostScript fonts by Malte Rosenau and Ulrich Dirr).

The typographic style of this dissertation was inspired by the authoritative genius of Bringhurst's *Elements of Typographic Style*, ported to LaTeX by André Miede, the original designer of the `classicthesis` template. Any unsightly deviations from these works should be attributed solely to the author's (not always successful) efforts to conform to the awkward A4 paper size.

Structured nonsmooth optimization: proximal identification, fast local convergence, and applications

Grenoble, December 2, 2022

Gilles Bareilles