Nonsmooth optimization: conditioning, convergence and semi-algebraic models

Adrian S. Lewis

Abstract. Variational analysis has come of age. Long an elegant theoretical toolkit for variational mathematics and nonsmooth optimization, it now increasingly underpins the study of algorithms, and a rich interplay with semi-algebraic geometry illuminates its generic applicability. As an example, alternating projections – a rudimentary but enduring algorithm for exploring the intersection of two arbitrary closed sets – concisely illustrates several far-reaching and interdependent variational ideas. A transversality measure, intuitively an angle and generically nonzero, controls several key properties: the method's linear convergence rate, *a posteriori* error bounds, sensitivity to data perturbations, and robustness relative to problem description. These linked ideas emerge in a wide variety of computational problems. Optimization in particular is rich in examples that depend, around critical points, on "active" manifolds of nearby approximately critical points. Such manifolds, central to classical theoretical and computational optimization, exist generically in the semi-algebraic case. We discuss examples from eigenvalue optimization and stable polynomials in control systems, and a prox-linear algorithm for large-scale composite optimization applications such as machine learning.

Mathematics Subject Classification (2010). Primary 90C31, 49K40, 65K10; Secondary 90C30, 14P10, 93D20.

Keywords. variational analysis, nonsmooth optimization, inverse function, alternating projections, metric regularity, semi-algebraic, convergence rate, condition number, normal cone, transversality, quasi-Newton, eigenvalue optimization, identifiable manifold.

1. Introduction: the Banach fixed point theorem

Our topic — sensitivity and iterative algorithms for numerical inversion and optimization — has deep roots in the Banach fixed point theorem, so we begin our quick introduction there. Given a Euclidean space $\mathbf{E}$ (a finite-dimensional real inner product space), we seek to invert a map $F\colon \mathbf{E} \to \mathbf{E}$. In other words, given a data vector $y \in \mathbf{E}$, we seek a solution vector $x \in \mathbf{E}$ satisfying $F(x) = y$. We analyze this problem around a particular solution $\bar{x} \in \mathbf{E}$ for data $\bar{y} = F(\bar{x})$. A good exposition on the idea of inversion, close in spirit to our approach here, is the monograph of Dontchev and Rockafellar [26].

Given a constant ρ such that the map $I - \rho F$ (where I is the identity) has Lipschitz modulus $\tau = \operatorname{lip}(I - \rho F)(\bar{x}) < 1$ (meaning that the map is locally a strict contraction), Banach's 1922 argument [2] shows that the *Picard iteration*

$$x_{k+1} = x_k - \rho(F(x_k) - y) \quad (\text{for } k = 0, 1, 2, \ldots)$$

Proceedings of the International Congress of Mathematicians, Seoul, 2014

converges linearly to a solution $\hat{x}$ of the equation $F(x) = y$, for any data vector y near $\bar{y}$, when initiated near $\bar{x}$. Furthermore, by starting sufficiently near $\bar{x}$, we can ensure an upper bound on the *linear rate* arbitrarily close to τ: in other words, for any constant $\bar{\tau} > \tau$ we know $\bar{\tau}^{-k}|x_k - \hat{x}| \to 0$. This construction shows that the inverse map F^{-1} agrees (graphically) around the point $(\bar{y}, \bar{x})$ with a single-valued function having Lipschitz modulus $\frac{\rho}{1-\tau}$. We thus see the sensitivity of solutions to data perturbations, and error bounds on the distance from approximate solutions x to the true solution in terms of the *a posteriori* error $F(x) - y$.

Similar classical arguments show robustness in the problem description: for linear maps $A\colon \mathbf{E} \to \mathbf{E}$ with norm less than the bound $\frac{\rho}{1-\tau}$, the perturbed map $F + A$ retains (locally) a Lipschitz inverse. Less classically, as we shall see (though related to the Eckart-Young theorem [40]), the bound $\frac{\rho}{1-\tau}$ is optimal.

Consider the even simpler case when the map F is linear, self-adjoint, and positive semidefinite, with maximum and minimum eigenvalues Λ and λ respectively. In the generic case when F is actually positive definite, we could choose $\rho = \frac{1}{\Lambda}$, and then $\tau = 1 - \frac{\lambda}{\Lambda}$. The Picard iteration becomes simply the method of steepest descent for the convex quadratic function $\frac{1}{2}\langle x, Fx\rangle - \langle y, x\rangle$, with constant step size. The key constant $\frac{1}{1-\tau}$ controlling the algorithm's convergence rate, sensitivity, error bounds and robustness, is just $\frac{\Lambda}{\lambda}$, the condition number of F. This constant is also closely associated with the linear convergence rate of other algorithms, such as steepest descent with exact line search, and the method of conjugate gradients.

A broad paradigm, originating with Demmel [22], relates the computational difficulty of a problem instance (here indicated by convergence rate) with the distance to the nearest "ill-posed" instance (in this case one where Lipschitz invertibility breaks down). An extensive theory of Renegar (see [65]), analogous to the theory above for convex quadratic minimization, concerns feasibility and optimization problems with constraints of the form $y \in Fx + K$, for linear maps F and convex cones K: in that case, the algorithms in question are interior-point methods [60].

Over the next couple of sections, we illustrate and study these ideas more broadly. In each case, we consider a computational problem involving inversion or optimization (which amounts to inverting a gradient-type mapping), and study a "regularity" modulus at a particular solution. We observe how that modulus controls error bounds, sensitivity analysis, robustness in the problem description, and the local linear convergence rate of simple iterative algorithms.

2. Variational analysis and alternating projections

We next consider the problem of *set intersection*: given two nonempty closed sets X and Y in the Euclidean space $\mathbf{E}$, we simply seek a point $z \in X \cap Y$. Like our first example, this problem involves a kind of inversion: we seek a point z such that $(0, 0)$ lies in the set $(X - z) \times (Y - z)$, a set we can view as a function of z.

We denote the distance from a point $y \in \mathbf{E}$ to X by $d_X(y)$, and the set of nearest points (or *projection*) by $P_X(y)$. We consider the method of *alternating projections*, which simply repeats the iteration

$$x_{k+1} \in P_X(y_k), \quad y_{k+1} \in P_Y(x_{k+1}).$$

For convex sets, this method has a long history dating back at least to a 1933 work of von Neumann [76], with a well understood convergence theory: a good survey is [3]. While typically slow, its simplicity lends it enduring appeal, even for nonconvex sets. Robust control theory, for example, abounds in low-rank matrix equations, and projecting a matrix M onto the (nonconvex) set of matrices of rank no larger than r is easy: we simply zero out all but the r largest singular values in the singular value decomposition of M (an approach tried in [38], for example). Furthermore, for our current purposes, the method of alternating projections perfectly illustrates many core ideas of variational analysis, as well as our broad thesis.

Central to our discussion is the notion of transversality. If x is a nearest point in the set X to a point $y \in \mathbf{E}$, then any nonnegative multiple of the vector $y - x$ is called a *proximal normal* to X at x: such vectors comprise a cone $N_X^p(x)$. We say that X and Y intersect *transversally* at a point $\bar{z}$ in their intersection when there exists an angle $\theta > 0$ such that the angle between any proximal normal to X and proximal normal to Y, both at points near $\bar{z}$, is always less than $\pi - \theta$. The supremum of such θ is the *transversality angle*. When X and Y are smooth manifolds, transversality generalizes the classical notion [47]. We then have the following special case of a result from [28].

Theorem 2.1 (Convergence of alternating projections). *Initiated near any transversal intersection point for two closed sets, the method of alternating projections converges linearly to a point in the intersection. If the transversality angle is $\bar{\theta} > 0$, then we can ensure an upper bound on the convergence rate arbitrarily close to $\cos^2(\frac{\bar{\theta}}{2})$ by initiating sufficiently near the intersection point.*

Notable in this result (unlike all previous analysis, such as [52]) is the absence of any assumptions on the two intersecting sets, such as convexity or smoothness. Central to the proof is the Ekeland variational principle [35].

Modern variational analysis grew out of attempts to expand the broad success of convex analysis — an area for which Rockafellar's seminal monograph [67] remains canonical — and to unify it with classical smooth analysis. Classical analysis relies crucially on limiting constructions: for example, the definition of transversally intersecting smooth manifolds (a special case of our property) involves their tangent spaces. The more general property described above also has a limiting flavor, and we can express it more succinctly using a limiting construction. This construction originated in Clarke's 1973 thesis [16, 17], in a convexified form, and a couple of years later, in the raw form we describe here (including implications for transversality) in work reported in Mordukhovich's paper [57] along with contemporaneous joint studies with Kruger ranging from [59] to [45]. It is fundamental to variational analysis: the expository monographs [8, 18, 58, 68] each provide excellent surveys and historical discussion, [7] is a gentler introduction, and [43, p. 112] recounts some early history. The monograph [26] is particularly attuned to our approach here.

A limit of proximal normals to the set X at a sequence of points approaching a point $x \in X$ is simply called a *normal* at x. Such vectors comprise a closed cone $N_X(x)$, possibly nonconvex, called the *normal cone*. With this notation, transversality at the point $\bar{z}$ is simply the property

$$N_X(\bar{z}) \cap -N_Y(\bar{z}) = \{0\},$$

and the transversality angle is the minimal angle between pairs of vectors in the cones $N_X(\bar{z})$ and $-N_Y(\bar{z})$.

The idea of a normal vector to a closed set $X \subset \mathbf{E}$ is a special case of the idea of a "subgradient" of a lower semicontinuous extended-real-valued function on $\mathbf{E}$. For simplicity of exposition, in this essay we confine ourselves to properties of normals, but many of the results that we present extend to subgradients.

The terminology of "normals" we use here is consistent with classical usage for smooth manifolds and convex sets, a fact fruitfully seen in a broader context. A set $X \subset \mathbf{E}$ is nonempty, closed and convex if and only if its projection mapping P_X is everywhere single-valued. More generally [64], X is *prox-regular* at a point $x \in X$ when P_X is everywhere single-valued nearby. In that case, the limiting construction above is superfluous: all normals are proximal, so the cones $N_X(x)$ and $N_X^p(x)$ coincide (and are closed and convex). Prox-regularity applies more broadly than convexity, to smooth manifolds, for example. A set $\mathcal{M} \subset \mathbf{E}$ is a $\mathcal{C}^{(2)}$ *manifold* around a point $\bar{x} \in \mathcal{M}$ if it can be described locally as $F^{-1}(0)$, where the map $F\colon \mathbf{E} \to \mathbf{F}$ is twice continuously differentiable, with surjective derivative at $\bar{x}$. In that case, classical analysis shows $\mathcal{M}$ is prox-regular at $\bar{x}$.

For convex sets X and Y, transversality fails at a common point exactly when there exists a separating hyperplane through that point; a small translation of one set then destroys the intersection. The following result [45], a local generalization of the separating hyperplane theorem, hints at the power of transversality.

Theorem 2.2 (Extremal principle). *On any neighborhood of a point where two closed sets intersect transversally, all small translations of the sets must intersect.*

This principle is a unifying theme in the exposition [58], for example. One proof proceeds constructively, using alternating projections [52].

Another consequence of transversality is the existence of an error bound, discussed in [41, p. 548], estimating the distance to the intersection of the two sets in terms of the distances to each separately. Notice, in the product space $\mathbf{E} \times \mathbf{E}$, we have the relationship $d_{X\times Y}(z,z) = \sqrt{d_X^2(z) + d_Y^2(z)}$ for any point $z \in \mathbf{E}$.

Theorem 2.3 (Error bound). *If closed sets X and Y intersect transversally at a point $\bar{z}$, then there exists a constant $\rho > 0$ such that all points z near $\bar{z}$ satisfy*

$$d_{X\cap Y}(z) \leq \rho d_{X\times Y}(z,z).$$

Intuitively, when the transversality angle is small, we expect to need a large constant ρ in the error bound above. We can make this precise through a single result, discussed in [52], subsuming the preceding two.

Theorem 2.4 (Sensitivity). *Sets X and Y intersect transversally at a point $\bar{z}$ if and only if there exists a constant $\rho > 0$ such that all points z near $\bar{z}$ and all small translations X' of X and Y' of Y satisfy*

$$d_{X'\cap Y'}(z) \leq \rho d_{X'\times Y'}(z,z).$$

The infimum of such ρ is $(1-\cos\bar{\theta})^{-\frac{1}{2}}$, where $\bar{\theta}$ is the transversality angle.

We see a pattern of ideas analogous to those for inversion via the Picard iteration: an algorithm whose linear convergence rate is governed by a sensitivity modulus. To pursue the analogy intuitively a little further, when the transversality angle $\bar{\theta}$ is small, we expect a small change in the problem description to destroy transversality. To illustrate, suppose $\bar{z} = 0$, and at that point choose unit normals u and v to the sets X and Y respectively with an angle

of $\pi - \bar{\theta}$ between them. Now consider the orthogonal map R on the space $\mathbf{E}$ rotating the u-v plane through an angle $\bar{\theta}$ and leaving its orthogonal complement invariant, and so that $Ru = -v$. The sets RX and Y are no longer transversal at zero, since Ru is normal to RX. A natural way to measure the change in the problem description is to view the original problem as $(I, I)z \in X \times Y$, and the perturbed problem as $(R^{-1}, I)z \in X \times Y$, the size of the change being the norm $\|I - R^{-1}\| = 2 \sin \frac{\bar{\theta}}{2}$. In Section 4, where we consider the broader pattern, we see that in fact a somewhat smaller change will destroy transversality. However, rather than pursue the analogy further now, we first consider whether transversality is a realistic assumption in concrete settings.

3. Generic transversality of semi-algebraic sets

Like most areas of analysis, the reach of general variational analysis is limited by pathological examples. In our present context, for example, consider the set intersection problem in $\mathbf{R}^3 = \mathbf{R}^2 \times \mathbf{R}$, for the two sets $X = \mathbf{R}^2 \times \{0\}$ and

$$Y \;=\; \big\{(w, r) : r \geq f(w)\big\},$$

where the function f is a famous 1935 example of Whitney [77] that is continuously differentiable and has an arc of critical points with values ranging from -1 to 1. Thus for every number s in the interval $[-1, 1]$ there exists a critical point w with $f(w) = s$: hence the vector $(0, -1)$ is normal to Y at the point (w, s), so clearly the intersection of the translated set $X + (x, s)$ (for any point $x \in \mathbf{R}^2$) and the set Y is not transversal at the point (w, s). We have arrived at an example of two closed sets for which, after translations, the failure of transversality is not uncommon.

On the other hand, in concrete computational settings we do not expect to encounter Whitney's example. To be more precise, we take as an illustrative model of "concrete" computation the world of *semi-algebraic* sets. We view the Euclidean space $\mathbf{E}$ as isomorphic to the space $\mathbf{R}^n$ (for some dimension n), and consider finite unions of sets, each defined by finitely-many polynomial inequalities. This world, and its generalizations in models of "tame" geometry first promoted by Grothendieck [39], strike happy compromises between broad generality and good behavior. Concise and clear surveys appear in [19, 20, 74].

On the one hand, semi-algebraic sets comprise a rich class: in particular, they may be neither convex nor smooth. They are, furthermore, often easy to recognize without recourse to the basic definition, due to the Tarski-Seidenberg Theorem: the projection of a semi-algebraic set onto a subspace is semi-algebraic. Applying this principle repeatedly shows that sets like the cone of real positive semidefinite symmetric matrices and sets of matrices of bounded rank are semi-algebraic.

On the other hand, semi-algebraic sets cannot be too pathological (or "wild", in Grothendieck's terminology). For example, although nonsmooth in general, they stratify into finite unions of analytic manifolds, so have a natural notion of *dimension*, namely the largest dimension of any manifold in a stratification. Another important example for us concerns the term "generic". In this essay, we call a property that depends on a data vector y in a Euclidean space $\mathbf{E}$ *generic* when it holds except for y in a set $Z \subset \mathbf{E}$ of measure zero. Unlike the general case, if Z is semi-algebraic, then the following properties are equivalent:

- Z has measure zero.

- Z has dimension strictly less than that of $\mathbf{E}$.
- the complement of Z is dense.
- the complement of Z is topologically generic.

We call semi-algebraic sets Z with these properties *negligible*.

No semi-algebraic analog can exist of the example we constructed from Whitney's function. Specifically, we have the following result [28], a special case of a powerful generalization we discuss later.

Theorem 3.1 (Generic transversality). *Suppose X and Y are semi-algebraic subsets of $\mathbf{E}$. Then for all vectors z outside a negligible semi-algebraic subset of $\mathbf{E}$, transversality holds at every point in the intersection of the sets $X - z$ and Y.*

Practical variational problems are often highly structured, involving sparse data, for example. Nonetheless, this result is reassuring: it suggests that, for concrete intersection problems with sets subject to unstructured perturbations, transversality is a reasonable assumption.

4. Measuring invertibility: metric regularity

Our sketch hints at an intriguing web of ideas concerning computational inversion:

- *Sensitivity* of solutions to data perturbation
- *Linear error bounds* for trial solutions in terms of measured error
- *Robustness* in problem description
- Local *linear convergence* of simple solution algorithms.

A single *modulus* (a condition number or angle in our examples) quantifies all four properties. We call problem instances *well-posed* when the modulus is finite, and, within broad problem classes, this property is *generic*. As we now describe, these interdependent ideas are very pervasive indeed.

To capture the abstract idea of computational inversion, we consider two Euclidean spaces $\mathbf{E}$ and $\mathbf{F}$ and a set-valued mapping Φ on $\mathbf{E}$ whose images are subsets of $\mathbf{F}$: we write $\Phi\colon \mathbf{E} \rightrightarrows \mathbf{F}$. Given a data vector $\bar{y} \in \mathbf{F}$, our problem is to find a solution $x \in \mathbf{E}$ to the generalized equation $\bar{y} \in \Phi(x)$. This model subsumes, of course, the example of a classical equation, when Φ is single-valued and smooth, but it is much more versatile than its abstract simplicity might suggest, modeling inequalities rather than just equations, for instance.

To illustrate the power of the approach, among many further examples, we keep in mind two in particular. The first we have seen already. Given two sets X and Y in the space $\mathbf{E}$, if we consider the mapping

$$\Phi\colon \mathbf{E} \rightrightarrows \mathbf{E}^2 \text{ defined by } \Phi(z) = (X - z) \times (Y - z), \tag{4.1}$$

then the problem $0 \in \Phi(z)$ is just set intersection.

For the second example, we return to the normal cone $N_X(x)$ to a nonempty closed set X in $\mathbf{E}$, but now thought of as a mapping $N_X\colon \mathbf{E} \rightrightarrows \mathbf{E}$ (defining $N_X(x) = \emptyset$ for $x \notin \mathbf{E}$). Solutions $x \in \mathbf{E}$ of the generalized equation $\bar{y} \in N_X(x)$ are *critical points* for the linear

optimization problem $\sup_X \langle \bar{y}, \cdot \rangle$. This terminology is in keeping with the classical notion when X is a smooth manifold, while for convex X, critical points are just maximizers. For simplicity, this essay concentrates on linear rather than general optimization. However, that restriction involves little loss of generality: for example, minimizing a function $f\colon \mathbf{E} \to \mathbf{R}$ is equivalent to a linear optimization problem over the *epigraph* of f:

$$\inf\{\tau : (x,\tau) \in \operatorname{epi} f\}, \quad \text{where} \quad \operatorname{epi} f \;=\; \big\{(x,\tau) \in \mathbf{E} \times \mathbf{R} : \tau \geq f(x)\big\}.$$

The fundamental idea, unifying the kinds of error bounds and sensitivity analysis we have illustrated so far, is *metric regularity* of the mapping Φ at a point $\bar{x} \in \mathbf{E}$ for a data vector $\bar{y} \in \Phi(\bar{x})$: the existence of a constant $\rho > 0$ such that

$$d_{\Phi^{-1}(y)}(x) \leq \rho d_{\Phi(x)}(y) \;\text{ for all } (x,y) \text{ near } (\bar{x},\bar{y}). \tag{4.2}$$

We call $\bar{y}$ a *critical value* if Φ is not metrically regular for $\bar{y}$ at some point in $\mathbf{E}$.

Inequality (4.2) is a locally uniform linear bound on the error between a trial solution x and the true solution set $\Phi^{-1}(y)$ for data y, in terms of the measured error from y to the trial image $\Phi(x)$. It captures both error bounds (where $y = \bar{y}$) and sensitivity analysis (where y varies). In highlighting metric regularity, we are implicitly supposing inversion to be computationally hard: the set $\Phi(x)$ is more tractable than the set $\Phi^{-1}(y)$.

The mapping Φ is *closed* when its graph

$$\operatorname{gph} \Phi \;=\; \big\{(x,y) \in \mathbf{E} \times \mathbf{F} : y \in \Phi(x)\big\}$$

is closed. It is *semi-algebraic* when its graph is semi-algebraic, and then its *graphical dimension* is the dimension of its graph. Around any point $(\bar{x},\bar{y}) \in \operatorname{gph} \Phi$ we define three constants:

- The *modulus* is the infimum of the constants $\rho > 0$ such that the metric regularity inequality (4.2) holds.
- The *radius* is the infimum of the norms of linear maps $G\colon \mathbf{E} \to \mathbf{F}$ such that the mapping $\Phi + G$ is not metrically regular at $\bar{x}$ for $\bar{y} + G\bar{x}$.
- The *angle* is the transversality angle for the sets $\operatorname{gph} \Phi$ and $\mathbf{E} \times \{\bar{y}\}$ at the point $(\bar{x},\bar{y})$.

These quantities are strongly reminiscent of the linked ideas opening this section. The first constant quantifies error bounds and sensitivity. The second concerns how robust the problem is under linear perturbations: within that class, it measures the distance to the nearest ill-posed (metrically irregular) instance. By Theorem 2.1, the third quantity controls the local linear convergence rate of at least one simple conceptual algorithm for finding a solution x near $\bar{x}$ to the generalized equation $\bar{y} \in \Phi(x)$: alternating projections on the sets $\operatorname{gph} \Phi$ and $\mathbf{E} \times \{\bar{y}\}$.

In this essay we concentrate on what we loosely call "simple" algorithms, relying only on basic evaluations and properties of the mapping Φ. By contrast, Newton-type schemes use, or assume and approximate, tangential ("higher-order") properties of $\operatorname{gph} \Phi$. For an extensive discussion relating metric regularity and the convergence of Newton-type methods, see [26]. The conceptual algorithm above belongs to the class of *proximal point* methods, which minimize functions f using the iteration

$$x_{k+1} \;\in\; \operatorname{argmin}\big\{f(x) + |x - x_k|^2\big\}.$$

In this case, $f(x) = d^2_{\Phi(x)}(\bar{y})$.

Between the three diverse quantities we have introduced, we have the following extraordinarily simple and general relationship.

Theorem 4.1 (Metric regularity). *At any point in the graph of any closed set-valued mapping we have*

$$\text{radius} \;=\; \frac{1}{\text{modulus}} \;=\; \tan(\text{angle}).$$

The first equality is [25, Theorem 1.5], while the second is a version of the "coderivative criterion" for metric regularity (whose history is discussed in [68, p. 418]). For reasons of space, we omit a dual "derivative criterion", expressible using tangents in the place of normals: see [24] for a discussion.

For instance, consider our motivating example, the intersection problem for two sets X and Y discussed in Section 2. Equation (4.1) describes the corresponding mapping. Theorem 2.4 (Sensitivity) and a calculation [52] shows that the modulus is $(1 - \cos\bar{\theta})^{-\frac{1}{2}}$, where $\bar{\theta}$ is the transversality angle for X and Y at the intersection point. The radius is therefore $\sqrt{2}\sin\frac{\bar{\theta}}{2}$. The proximal point method above is that for minimizing the function $d^2_X + d^2_Y$.

We consider our earlier example of normal cone operators shortly. First, however, we return to the classical single-valued case. When the mapping Φ is linear, standard linear algebra shows that metric regularity is equivalent to surjectivity, and the Eckart-Young Theorem identifies the radius as just the smallest singular value of Φ. More generally, for continuously differentiable Φ, the Lusternik-Graves Theorem amounts to the fact that the modulus of Φ at any point $\bar{x}$ agrees with that of its linear approximation there (see [26]), and hence equals the reciprocal of the smallest singular value of the derivative of Φ at $\bar{x}$.

This classical case also guides us on the question of whether metric regularity is a generic property. In this case, the set of critical values $C \subset \mathbf{F}$ is just the image under the mapping Φ of the set in $\mathbf{E}$ where the derivative of Φ is not surjective. The example of Whitney that we discussed earlier shows that C may be large (in that case an interval in $\mathbf{R}$) even for continuously differentiable Φ. However, assuming Φ is sufficiently smooth, Sard's theorem [69] guarantees that C has measure zero. In this sense, metric regularity is typical.

To address this question for set-valued mappings, we consider the semi-algebraic world, in which we have the following striking result of Ioffe [42].

Theorem 4.2 (Semi-algebraic Sard). *The set of critical values of any semi-algebraic set-valued mapping is semi-algebraic and negligible.*

In computational practice, generic results like this one may often be of limited consequence, since generalized equations often involve highly structured data. Nonetheless, like its special case, Theorem 3.1 (Generic transversality), the result provides a reassuring baseline: for concrete generalized equations with unstructured data, metric regularity is a reasonable assumption.

5. Interlude: nonsmooth optimization via quasi-Newton methods

Metric regularity spans a broad range of inversion and optimization problems. Its suggestive links to convergence rates tempt us to study linearly convergent algorithms, whenever we encounter them, through the lens of metric regularity. An important recurrent theme in the

work of the late Paul Tseng, for example, was the use of error bounds in linear convergence results [56].

An intriguing case is the popular BFGS method [61] (named for its inventors, Broyden, Fletcher, Goldfarb and Shanno) for minimizing a function $f\colon \mathbf{R}^n \to \mathbf{R}$. The BFGS algorithm is a quasi-Newton method, so called by association with the Newton iteration for minimizing a $\mathcal{C}^{(2)}$ function f:

$$x_{k+1} = x_k - \nabla^2 f(x_k)^{-1} \nabla f(x_k).$$

The BFGS method replaces the inverse Hessian by an approximation H_k in the space of n-by-n symmetric matrices $\mathbf{S}^n$, and involves a step length $\alpha_k > 0$:

$$x_{k+1} = x_k - \alpha_k H_k \nabla f(x_k).$$

We then choose H_{k+1} to be the minimizer over the positive-definite matrices of the strictly convex function

$$H \;\mapsto\; \text{trace}(H_k^{-1} H) - \ln\det H \tag{5.1}$$

(see [36]), subject to a linear constraint called the *secant condition*:

$$H\big(\nabla f(x_{k+1}) - \nabla f(x_k)\big) \;=\; x_{k+1} - x_k.$$

The secant condition forces H_{k+1} to behave like the true inverse Hessian in the direction of the last step taken, while the objective (5.1) keeps H_{k+1} close to H_k, since its unconstrained minimizer is at H_k. A simple formula [61] expresses H_{k+1} explicitly as a rank-two perturbation of H_k.

The step length α_k is chosen by a line search on the univariate function

$$\alpha \;\mapsto\; h(\alpha) \;=\; f\big(x_k - \alpha H_k \nabla f(x_k)\big),$$

aiming to satisfy two conditions (called the Armijo and Wolfe conditions):

$$h(\alpha) < h(0) + c_1 h'(0)\alpha \;\text{ and }\; h'(\alpha) > c_2 h'(0).$$

The constants $c_1 < c_2$ in the interval $(0, 1)$ are fixed at the outset. The Armijo condition requires the decrease in the value of h to be a reasonable fraction of its instantaneous decrease at zero, while the Wolfe condition prohibits steps that are too small, by requiring a reasonable reduction in the rate of decrease in h, and ensures the existence of a positive-definite H_{k+1} satisfying the secant condition. A simple bisection scheme finds a suitable step α_k by maintaining the endpoints of a search interval such that the Armijo condition holds on the left and fails on the right, checking both conditions at the midpoint, and then halving the interval accordingly. For a thorough description, see [53].

The BFGS algorithm has been a method of choice for smooth minimization for several decades. It is robust and fast, typically converging superlinearly to a local minimizer. Given its motivation — approximating a Hessian — it seems astonishing that the algorithm also serves as an excellent general-purpose method for *nonsmooth* nonconvex minimization. In principle the algorithm might encounter a point x_k at which the function f is not differentiable, and thereby break down, but with generic initialization, no such breakdowns seem to occur.

A systematic study [53] investigated this phenomenon, and we return to an example from that study later in this work. In general, the BFGS method, when applied to minimize a

semi-algebraic Lipschitz function and with generic initialization, always seems to generate a sequence of function values (including all those computed in the line search) that converges to a stationary value — the value of the function at a point near which convex combinations of gradients are arbitrarily small. Furthermore, for nonsmooth stationary values, that convergence is *linear*! Our current context demands the obvious question: does some condition number or modulus of metric regularity govern the convergence rate of the BFGS method on nonsmooth problems? We seem far from any understanding of this question.

6. Strong regularity and second-order properties

One way to strengthen the metric regularity property is especially important for sensitivity analysis and numerical methods. A set-valued mapping $\Phi\colon \mathbf{E} \rightrightarrows \mathbf{F}$ is clearly metrically regular at a point $\bar{x}$ for a value $\bar{y}$ when the graph of the inverse mapping Φ^{-1} coincides locally with the graph of a single-valued Lipschitz map $G\colon \mathbf{F} \to \mathbf{E}$ around the point $(\bar{y}, \bar{x})$. In that case, we call Φ *strongly metrically regular* (terminology deriving from Robinson [66]); the regularity modulus coincides with the Lipschitz modulus $\operatorname{lip} G(\bar{y})$. We call $\bar{y}$ a *weakly critical value* if there exists a point in $\mathbf{E}$ at which Φ is not strongly metrically regular for $\bar{y}$.

We began, in Section 1, with an example of strong metric regularity: a single-valued map $\Phi\colon \mathbf{E} \to \mathbf{E}$ such that $I - \rho\Phi$ is, locally, a strict contraction. A related example derives from the setting of the classical inverse function theorem: a continuously differentiable map $\Phi\colon \mathbf{E} \to \mathbf{E}$ is strongly metrically regular at points where the derivative of Φ is invertible.

Less classically, suppose the set $\mathcal{M} \subset \mathbf{E}$ is a $\mathcal{C}^{(2)}$ manifold around a point $\bar{x} \in \mathcal{M}$, and consider the mapping Φ defined in terms of the normal cone $N_{\mathcal{M}}$ by

$$\Phi(x) = \begin{cases} x + N_{\mathcal{M}}(x) & (x \in \mathcal{M}) \\ \emptyset & (x \notin \mathcal{M}). \end{cases}$$

Strong metric regularity holds at $\bar{x}$ for $\bar{x}$, because around the point $(\bar{x}, \bar{x})$, the inverse mapping Φ^{-1} agrees graphically with the projection operator $P_{\mathcal{M}}$, which is single-valued and Lipschitz.

The previous section included a conceptual algorithm for solving metrically regular generalized equations, whose convergence rate is controlled by the modulus. Assuming, instead, strong metric regularity (of the mapping Φ at the point $\bar{x}$, for the value 0, say), Pennanen [62] linked the modulus to the convergence rates of algorithms closer to computational practice (in "multiplier methods"). For example, for any constant c larger than twice the modulus, there exists a neighborhood U of $\bar{x}$ such that, with initial point $x_0 \in U$, the proximal-point-type iteration

$$x_k - x_{k+1} \in c\Phi(x_{k+1}), \quad \text{with } x_{k+1} \in U \tag{6.1}$$

always generates sequences converging linearly to a solution of the generalized equation $0 \in \Phi(x)$. (The bound on the rate behaves, as $c \to \infty$, like $\frac{\sqrt{5}}{c}$ times the modulus.) In keeping with our focus on simple algorithms, we pass by the strong connections between strong metric regularity and Newton methods (most importantly in numerical optimization via sequential quadratic programming). That line of investigation, first pursued in [44], is discussed at length in the monograph [26].

Unlike critical values, weakly critical values may be common, even for semi-algebraic mappings. For example, mapping every point to the whole range space $\mathbf{F}$ results in every value being weakly critical. As the next result [32] makes clear, however, this behavior can only result from a large graph.

Theorem 6.1 (Strong semi-algebraic Sard). *If a set-valued mapping $\Phi\colon \mathbf{E} \rightrightarrows \mathbf{F}$ is semi-algebraic and has graphical dimension no larger than $\dim \mathbf{F}$, then its set of weakly critical values is semi-algebraic and negligible.*

This result applies to single-valued semi-algebraic maps $\Phi\colon \mathbf{E} \to \mathbf{E}$ in particular. More interesting for optimizers, however, is the following corollary [27, 30].

Theorem 6.2 (Normal cone mapping). *For any closed semi-algebraic set $X \subset \mathbf{E}$, the normal cone mapping $N_X\colon \mathbf{E} \rightrightarrows \mathbf{E}$ has graphical dimension equal to $\dim \mathbf{E}$, and hence its set of weakly critical values is semi-algebraic and negligible.*

This result suggest that, for concrete linear optimization problems over a set $X \subset \mathbf{E}$ with unstructured objective $\langle \bar{y}, \cdot \rangle$ and solution $\bar{x}$, strong metric regularity of the normal cone mapping N_X is a reasonable assumption. As the next result [31] reveals, this type of property is closely related to second-order conditions in optimization, classically guaranteeing quadratic growth via a Hessian condition. Following our pared-down approach, we focus on linear optimization, but, as before, we could consider a more general problem of the form $\inf_X f$ as seeking a point (x, τ) in the set epi $f \cap (X \times \mathbf{R})$ to maximize the linear function $-\tau$.

Theorem 6.3 (Strong regularity and quadratic growth). *Given a closed set X and a vector $\bar{y} \in \mathbf{E}$, suppose that the point $\bar{x} \in X$ is a local maximizer of the linear function $\langle \bar{y}, \cdot \rangle$ over X. Consider the following three properties:*

- *The normal cone mapping N_X is strongly metrically regular at $\bar{x}$ for $\bar{y}$.*
- *For some scalar $\kappa > 0$ and neighborhood U of $\bar{x}$, "uniform quadratic growth" holds: for all vectors y near $\bar{y}$, there exists a point $x \in X \cap U$ so*

$$\langle y, x' \rangle \leq \langle y, x \rangle - \kappa |x' - x|^2 \ \text{ for all } \ x' \in X \cap U. \tag{6.2}$$

- *The "negative definite" condition holds:*

$$(z, w) \in N_{\operatorname{gph} N_X}(\bar{x}, \bar{y}) \ \text{ and } \ w \neq 0 \implies \langle z, w \rangle < 0.$$

In general, the first condition implies the second two, and so, if X is semi-algebraic, then for all $\bar{y}$ outside a negligible semi-algebraic set, all three conditions hold. If, on the other hand, X is prox-regular at $\bar{x}$, then all three conditions are equivalent.

This result has multiple roots, and deserves some comments. Bonnans and Shapiro include a careful study of uniform quadratic growth in their monograph [6]. Geometrically, condition (6.2) describes a ball with surface containing the point x, center on the ray $x - \mathbf{R}_+ y$, and containing the set X around x. It is natural to include the final special case of a prox-regular set, because the first condition alone turns out to imply a local form of prox-regularity [31]. Assuming prox-regularity, a fourth equivalent notion is *tilt stability* [63].

The link with second-order conditions is not surprising, because the regularity modulus of the normal cone mapping is related via Theorem 4.1 (Metric regularity) to the transversality angle at the intersection point $(\bar{x}, \bar{y})$ for the sets $\operatorname{gph} N_X$ and $\mathbf{E} \times \{\bar{y}\}$, which in turn is just the minimal angle between the subspace $\{0\} \times \mathbf{F}$ and the cone

$$N_{\operatorname{gph} N_X}(\bar{x}, \bar{y})$$

appearing in the third condition. Mordukhovich [58] uses exactly this iterated normal cone construction to define his *generalized Hessian*.

For a semi-algebraic set X, the result above, while interesting, dramatically understates the good behavior of N_X^{-1}: it will typically be single-valued and not just Lipschitz but *analytic*. We explore far-reaching consequences next.

7. Identifiability and the active set philosophy

Given a closed set $X \subset \mathbf{E}$ and a data vector $\bar{y} \in \mathbf{E}$, consider once again the linear optimization problem

$$\sup_X \langle \bar{y}, \cdot \rangle. \tag{7.1}$$

Recall that a point $x \in X$ is *critical* when $\bar{y} \in N_X(x)$.

A wide variety of iterative methods for the linear optimization problem generate *asymptotically critical* sequences (x_k) in X for $\bar{y}$, meaning that some sequence of normals $y_k \in N_X(x_k)$ converges to $\bar{y}$ (implying in particular that any limit point of (x_k) is critical for the problem (7.1)). We aim to profit from this behavior by simplifying the possibly complicated underlying set X. We first illustrate with two examples: alternating projections, and proximal point methods.

Suppose we seek a critical point for our linear optimization problem by applying the proximal point method (6.1) to the mapping defined on $\mathbf{E}$ by $x \mapsto \Phi(x) = N_X(x) - \bar{y}$. Assume the normal cone mapping N_X is strongly metrically regular at $\bar{x}$ for $\bar{y}$. We arrive at the following relationship between iterates, in a neighborhood of a solution $\bar{x}$:

$$\frac{1}{c}(x_k - x_{k+1}) + \bar{y} \in N_X(x_{k+1}).$$

This uniquely defines a sequence in a neighborhood of any fixed solution that converges, providing the constant c is large enough. Since the left-hand side must therefore converge to $\bar{y}$, the sequence of iterates (x_k) is asymptotically critical.

As another example, given two closed sets X and Y in $\mathbf{E}$, we could rewrite the set intersection problem as the linear optimization problem $\sup\{-\tau : (x, y, \tau) \in S\}$, where $S \subset \mathbf{E}^2 \times \mathbf{R}$ is the set defined by the constraint $\tau \geq \frac{1}{2}|x - y|^2$. A quick calculation shows that the method of alternating projections generates two sequences of points, $x_k \in X$ and $y_k \in Y$, satisfying

$$(0, x_k - x_{k+1}, -1) \in N_S(s_k), \quad \text{where } s_k = \Big(x_{k+1}, y_k, \frac{1}{2}|x_{k+1} - y_k|^2\Big).$$

Under reasonable conditions — those of Theorem 2.1 (Convergence of alternating projections), for example — we know that both sequences converge to a point $z \in X \cap Y$. Hence the sequence (s_k) is asymptotically critical for the problem.

We now introduce a simple but powerful variational idea for the set X. We call a subset $\mathcal{M} \subset X$ *identifiable* at a point $\bar{x} \in X$ for the vector $\bar{y} \in \mathbf{E}$ if every asymptotically critical sequence for $\bar{y}$ converging to $\bar{x}$ must eventually lie in $\mathcal{M}$. If $\mathcal{M}$ is also a $\mathcal{C}^{(2)}$ manifold at $\bar{x}$, then we simply call it an *identifiable manifold* at $\bar{x}$ for $\bar{y}$. Such a subset hence balances two competing demands: as a subset of the typically nonsmooth set X, it must be small enough to be a smooth manifold, and yet large enough to capture the tail of every asymptotically critical sequence.

The existence of an identifiable manifold seems, at first sight, a demanding condition. The next result [29], on the other hand, shows that such a manifold uniquely captures important sensitivity information about how critical points for the linear optimization problem (7.1) vary under data perturbation. Furthermore, its existence forces the critical point $\bar{x} \in X$ for the vector $\bar{y} \in \mathbf{E}$ to be *nondegenerate*: $\bar{y}$ must lie not just in the normal cone $N_X(\bar{x})$, but in its relative interior — its interior relative to its span. It also forces a local form of prox-regularity [29], rather as in the Section 6, so for transparency we simply assume prox-regularity.

Theorem 7.1 (Identifiability, uniqueness, and sensitivity). *Suppose the set $X \subset \mathbf{E}$ is prox-regular at the point $\bar{x} \in X$. If X has an identifiable manifold $\mathcal{M}$ at $\bar{x}$ for the vector $\bar{y} \in \mathbf{E}$, then that manifold is locally unique. Indeed, for any sufficiently small neighborhood U of $\bar{y}$, the manifold $\mathcal{M}$ coincides locally around $\bar{x}$ with the set $N_X^{-1}(U)$. Furthermore, $\bar{x}$ must then be a nondegenerate critical point for $\bar{y}$.*

To illustrate, consider the case when the set X is a polyhedron. If $\bar{x}$ is a nondegenerate critical point for the problem $\sup_X \langle \bar{y}, \cdot \rangle$, then the set of maximizers (or in other words the face of X exposed by the vector $\bar{y}$) is an identifiable manifold at $\bar{x}$ for $\bar{y}$. We discuss more varied examples in the following sections.

A set X may easily have no identifiable manifold at the critical point $\bar{x}$ in question, even when X is closed, convex and semi-algebraic, and $\bar{x}$ is nondegenerate. An example is the set

$$X \;=\; \big\{(u, v, w) \in \mathbf{R}^3 : w^2 \geq u^2 + v^4,\; w \geq 0\big\},$$

at the point $\bar{x} = (0, 0, 0)$ for the vector $\bar{y} = (0, 0, -1)$. However, as we shall see shortly, at least for semi-algebraic examples such as this one, such behavior is unusual. Furthermore, as the next result [29] makes clear, the existence of an identifiable manifold has broad and powerful consequences for optimization.

Theorem 7.2 (Identifiability, active sets, and partial smoothness). *Suppose the set $X \subset \mathbf{E}$ is prox-regular at the point $\bar{x} \in X$, and has an identifiable manifold $\mathcal{M}$ there for the vector $\bar{y} \in \mathbf{E}$. Then the following properties hold:*

- **Smooth reduction:** *The graphs of the normal cone mappings N_X and $N_{\mathcal{M}}$ coincide around the point $(\bar{x}, \bar{y})$.*
- **Sharpness:** *The normal cone $N_X(\bar{x})$ spans the normal space $N_{\mathcal{M}}(\bar{x})$.*
- **Active set philosophy:** *For any small neighborhood V of $\bar{x}$, if the vector $y \in \mathbf{E}$ is near $\bar{y}$, then the two optimization problems of maximizing the linear function $\langle y, \cdot \rangle$ over the sets $X \cap V$ and $\mathcal{M} \cap V$ are equivalent.*
- **Second-order conditions:** *The rate of quadratic growth*

$$\liminf_{x \to \bar{x}} \frac{\langle \bar{y}, \bar{x} - x \rangle}{|\bar{x} - x|^2}$$

is independent of whether the limit is taken over $x \in X$ or $x \in \mathcal{M}$.

Given the multiple flavors of this result, some commentary is useful. Perhaps most striking is the "active set" result, which reduces the original optimization problem over the potentially nonsmooth and high-dimensional set X to the restricted optimization problem over the smooth and potentially lower-dimensional subset $\mathcal{M}$. Exactly this phenomenon drives the elimination of inequality constraints inherent in classical active set methods for optimization [61], and also the big reduction in dimension crucial to "sparse optimization" in contemporary machine learning and compressed sensing applications. In a huge recent literature, a particularly pertinent example is [80].

Underlying the active set assertion is the "smooth reduction" result that the mappings N_X and $N_{\mathcal{M}}$ graphically coincide, locally. Since $\mathcal{M}$ is a smooth manifold, its normal cone mapping is easy to understand through classical analysis. In particular, second-order properties like the negative definite condition in Theorem 6.3 (Strong regularity and quadratic growth), which may in general appear formidably abstract, now become purely classical [55]. For example, the $\liminf$ in the final second-order condition above, when computed over $\mathcal{M}$, simply involves a Hessian computation for the function $\langle \bar{y}, \cdot \rangle$ restricted to the manifold $\mathcal{M}$.

The "sharpness" property at the point $\bar{x}$ is geometric in essence: we call the set X *sharp* (or "V-shaped") there around the manifold $\mathcal{M}$. In [50], extending Wright's notion of an "identifiable surface" for active set methods in convex optimization [79], the set X is called *partly smooth* at the point $\bar{x}$ relative to the $\mathcal{C}^{(2)}$ manifold $\mathcal{M}$ when this sharpness property holds, the normal cone mapping N_X is continuous at $\bar{x}$ when restricted to $\mathcal{M}$, and *Clarke regularity* holds on $\mathcal{M}$. This latter property concerns *tangent* directions $z \in \mathbf{E}$ at any point $x \in X$ (limits of directions to nearby points in X): it requires $\langle y, z \rangle \leq 0$ for all normals $y \in N_X(x)$.

Partial smoothness is closely related to identifiability. In general, consider a point $\bar{x}$ in a set X and a proximal normal $\bar{y} \in N_X^p(\bar{x})$. On the one hand, suppose that the critical point $\bar{x}$ is nondegenerate for $\bar{y}$, and partial smoothness holds relative to a $\mathcal{C}^{(2)}$ manifold $\mathcal{M}$. In particular, X must then be Clarke regular at $\bar{x}$. However, if we strengthen this assumption slightly, from Clarke to prox-regularity, then $\mathcal{M}$ must be an identifiable manifold. On the other hand, suppose conversely that $\mathcal{M}$ is an identifiable manifold. As we have seen, $\bar{x}$ must then be nondegenerate, and furthermore a local version of partial smoothness must hold [29].

The existence of an identifiable manifold, as Theorem 7.2 makes clear, is a powerful property. Remarkably, according to the following result [32], for semi-algebraic optimization this property holds generically.

Theorem 7.3 (Generic identifiability). *Given any closed semi-algebraic set $X \subset \mathbf{E}$, there exists an integer K such that, for all vectors $y \in \mathbf{E}$ outside some negligible semi-algebraic subset of $\mathbf{E}$, the following properties hold. The linear optimization problem*

$$\sup_X \langle y, \cdot \rangle.$$

has no more than K local maximizers. At each local maximizer $x \in X$, the normal cone mapping is strongly metrically regular for y; there exists an identifiable manifold $\mathcal{M}$ at x for y, and the normal cone mappings N_X and $N_{\mathcal{M}}$ coincide around the point (x, y). Furthermore, x is a nondegenerate critical point, and X is sharp around $\mathcal{M}$ there: in other words, y lies in the interior of the normal cone $N_X(x)$ relative to its span, which is just the

normal space $N_{\mathcal{M}}(x)$. In addition, the following quadratic growth condition holds:

$$\liminf_{\substack{x' \to x \\ x' \in X}} \frac{\langle y, x - x' \rangle}{|x - x'|^2} > 0.$$

The key ingredients of the proof have appeared through our discussion. Generic strong metric regularity follows from Theorem 6.2, and in that case, the inverse image of a small neighborhood of y under the normal cone mapping N_X (or in other words the set of nearby approximately critical points) will generically comprise an identifiable manifold. The consequences then flow from Theorems 6.3, 7.1, and 7.2. For convex sets X, this result appeared in [5].

In this result, we can view the existence of an identifiable manifold in conjunction with nondegeneracy and the quadratic growth condition as comprising the natural "second-order sufficient conditions" for our optimization problem. Classically, the generic validity of such conditions has a long history, dating back to [70]. Here we have taken a fresh, abstract approach, assuming nothing about the structure of the problem beyond its concrete (semi-algebraic) nature.

8. Optimization over stable polynomials

We have argued that ideas of identifiable manifolds and active set methods in optimization merge seamlessly. Less standard, but an elegant computational illustration of the appearance of an identifiable manifold, is a problem of Blondel [4]. The original question (with generous prizes of Belgian chocolate) highlighted the difficulty of simultaneous plant stabilization in continuous-time control.

The crucial idea of stability in dynamical systems and control theory involves *stable* and *strictly stable* polynomials $p(z)$ (for the complex variable $z \in \mathbf{C}$): polynomials with all zeroes in the closed or open left half-planes respectively. Blondel's problem seeks stable polynomials p, q, r with real coefficients and satisfying

$$r(z) = (z^2 - 2\delta z + 1)p(z) + (z^2 - 1)q(z),$$

for a real parameter $\delta \in [0.9, 1)$. If $\delta = 1$, then $r(1) = 0$, so no solution exists.

An optimization approach to this problem in [10], for any fixed parameter value δ, varies a cubic polynomial p and scalar q to minimize numerically a real variable α under the condition that the two polynomials $z \mapsto p(z + \alpha)$ and $z \mapsto r(z + \alpha)$ are both stable. The numerical results in [10] strongly suggest that when $\delta \in [0.9, 0.96]$, the minimum value $\bar{\alpha}$ is negative, as required for stability. If, furthermore, δ is close to, and no larger than, the value $\bar{\delta} = \frac{1}{2}\sqrt{2 + \sqrt{2}} \approx 0.924$, then the optimal polynomials $\bar{p}$ and $\bar{r}$ are not only stable but have a persistent structure: $\bar{p}$ is strictly stable, and $\bar{r}$ is a multiple of the polynomial $z \mapsto (z - \bar{\alpha})^5$. This structure defines a manifold $\mathcal{M}$ in the space of variables (α, p, q, r), which, once divined numerically, leads to a solution to Blondel's problem in closed form for such δ. Not surprisingly, $\mathcal{M}$ is the identifiable manifold for our optimization problem.

Underlying this striking appearance of an identifiable manifold is a remarkable property of stable polynomials. To understand this property, we first identify monic polynomials p of degree n with vectors $\tilde{p}$ in the space $\mathbf{C}^n$ (with the usual inner product), via the correspondence $p(z) = z^n + \sum_{j<n} \tilde{p}_j z^j$, and thereby consider them as constituting a Euclidean

space. Within that space, we then consider the set of stable polynomials Δ_n. The basic variational geometry of this nonconvex set is challenging. Around any polynomial with a multiple imaginary zero, Δ_n is nonsmooth, and indeed, with a suitable interpretation, nonlipschitz. Notice, for example, that monic polynomials $p(z)$ near the polynomial z^n have zeroes whose dependence on the coefficient vector $\tilde{p}$ is nonlipschitz.

On the other hand, despite these structural challenges, the set of monic stable polynomials is certainly semi-algebraic. Theorem 7.3 (Generic identifiability) therefore implies the *generic* existence of an identifiable manifold around solutions of linear optimization problems over stable polynomials. However, the following beautiful result of Burke and Overton [13] holds not just generically, but *always*.

Theorem 8.1. *The set of monic stable polynomials of degree n is Clarke regular everywhere.*

The techniques of [13] (which treat regions more general than the left half-plane) show more. For any monic stable polynomial p, the normal cone $N_{\Delta_n}(p)$ depends on the "pattern" of imaginary zeroes of p (which we specify simply by listing the multiplicities of those zeroes as we move down the imaginary axis). Using the language of partial smoothness from Section 7, we arrive at the following result.

Theorem 8.2 (Partial smoothness of the stable polynomials). *Around any polynomial in the set of monic stable polynomials Δ_n, the subset of polynomials with the same pattern of imaginary zeroes constitute a manifold, with respect to which Δ_n is partly smooth.*

It is exactly this property that underlies the identifiable manifold in [10] for Blondel's problem. The set of stable n-by-n matrices — those whose eigenvalues all lie in the left half plane — enjoys parallel properties around any stable *nonderogatory* matrix (one whose eigenvalues all have geometric multiplicity one) [11, 12]. One explanation [51] is to note that the characteristic polynomial map from the space of matrices to monic polynomials has surjective derivative at any nonderogatory matrix, enabling a standard calculus rule.

9. An eigenvalue optimization example

We can be confident of the generic existence of an identifiable manifold for a semi-algebraic optimization problem, by Theorem 7.3 (Generic identifiability), under no assumptions whatsoever about the problem's presentation. Optimization algorithms sometimes reveal clues about the identifiable manifold as they proceed. For the polynomial stabilization example in Section 8, numerical results from a simple general-purpose nonsmooth optimization method point to the identifiable manifold, helped along by our understanding of the potential structure for such manifolds. The BFGS method that we discussed in Section 5 naturally accumulates identifiable manifold information as it nears an optimal solution.

Suppose the BFGS method for minimizing a function $f\colon \mathbf{R}^n \to \mathbf{R}$ converges to a local minimizer $\bar{x} \in \mathbf{R}^n$ at which there exists an identifiable manifold $\mathcal{M}$. By this, we mean that the set $\big\{\big(x, f(x)\big) : x \in \mathcal{M}\big\}$ is an identifiable manifold of the epigraph $\operatorname{epi} f$, at the point $\big(\bar{x}, f(\bar{x})\big)$, for the vector $(0, -1)$. We deduce a dichotomy: on the one hand, the restriction of f to the manifold $\mathcal{M}$ is smooth, and on the other hand, the sharpness condition in Theorem 7.2 (Identifiability, active sets, and partial smoothness) shows that the gradient ∇f jumps as we move orthogonally across $\mathcal{M}$. The inverse Hessian approximation H_k should reflect this dichotomy: a basis of eigenvectors spanning an approximation to the tangent space to

$\mathcal{M}$ at $\bar{x}$ corresponds to a well-scaled set of eigenvalues, whereas the eigenvectors spanning the orthogonal complement correspond to eigenvalues converging to zero. In numerical experiments on semi-algebraic Lipschitz functions, we indeed see exactly this behavior [53].

The following example from [1] is illuminating:

$$\inf\Big\{\prod_{i=1}^{q}\lambda_i(A\circ X) : X\in\mathbf{S}^p_+,\ X_{ii}=1 \text{ for all } i\Big\}.$$

Here, $\mathbf{S}^p_+$ denotes the cone of p-by-p positive semidefinite matrices, $A\in\mathbf{S}^p$ is a given data matrix, $\circ$ denotes the componentwise (Hadamard) matrix product, and λ_i denotes the ith largest eigenvalue (counted by multiplicity). It is not hard to frame this optimization problem as the unconstrained minimization of a suitable function f, expressed in terms of the nonsmooth nonconvex function $\prod_{i=1}^{q}\lambda_i$ on the space $\mathbf{S}^p$: see [53] for the modeling details.

The results from multiple runs of the BFGS method on an example with $p=20$ and $q=10$ are typical for eigenvalue optimization [53]. Generic symmetric matrices have no multiple eigenvalues, but optimal solutions of semidefinite programs (see [72]) and more general eigenvalue optimization problems usually do, precisely due to their identifiable manifolds. That is the case here: as observed in Section 5, the BFGS trial function values consistently converge linearly, and at termination, the nine eigenvalues $\lambda_6,\lambda_7,\lambda_8,\ldots,\lambda_{14}$ of the matrix $A\circ X$ are coalescing.

Given a permutation-invariant function $h\colon\mathbf{R}^p\to\mathbf{R}$, that function of the vector $\lambda(Z)\in\mathbf{R}^p$ with components the eigenvalues of a matrix variable $Z\in\mathbf{S}^p$ inherits many properties from h. One important example is convexity [48], a generalization of von Neumann's characterization of unitarily invariant matrix norms [75], but the list of such properties is extensive [49]. In particular [21], given a matrix $\bar{Z}\in\mathbf{S}^p$, if h has an identifiable manifold $\mathcal{M}$ at the point $\lambda(\bar{Z})$, then at $\bar{Z}$ the composite function $h\big(\lambda(\cdot)\big)$ has an identifiable manifold $\{Z\in\mathbf{S}^n : \lambda(Z)\in\mathcal{M}\}$.

The permutation-invariant function $h\colon\mathbf{R}^p\to\mathbf{R}$ here is $h(x)=\prod_{i=1}^{q}[x]_i$, where the map $x\mapsto[x]$ rearranges the components of the vector $x\in\mathbf{R}^p$ into nonincreasing order. For this function h we can easily check in the example that the set

$$\big\{x\in\mathbf{R}^{20} : [x]_5>[x]_6=[x]_7=\cdots=[x]_{14}>[x]_{15}\big\}$$

is an identifiable manifold. Hence

$$\big\{Z\in\mathbf{S}^{20} : \lambda_5(Z)>\lambda_6(Z)=\lambda_7(Z)=\cdots=\lambda_{14}(Z)>\lambda_{15}(Z)\big\}$$

is an identifiable manifold for our objective function $\prod_{i=1}^{10}\lambda_i$. Classical matrix analysis [46, p. 141] shows that this manifold of symmetric matrices with an eigenvalue of multiplicity nine has codimension $\frac{1}{2}9(9+1)-1=44$.

Examining the BFGS output [53] and counting the number of eigenvalues of the inverse Hessian approximations H_k that converge to zero reveals the answer 44 — exactly the codimension of the identifiable manifold at the optimal solution. To confirm, around the final iterate we can plot the behavior of the objective function along the eigenvectors of H_k. Sure enough, along those eigenvectors corresponding to the vanishing eigenvalues, the objective function is V-shaped; along other eigenvectors, it is smooth. To summarize, with no *a priori* input about the underlying structure of the problem, and no *a posteriori* interpretation, the BFGS method nonetheless accurately approximates the geometry of the identifiable manifold.

10. Identifiability and a prox-linear algorithm

We have argued that, independent of the presentation of an optimization problem, an identifiable manifold is typically there to be found, and is a powerful tool once known. However, the manner in which a particular algorithm profits from such knowledge will likely depend on the explicit structure of the underlying problem. The classical example is the active set methodology for optimization under inequality constraints, which considers equality-constrained subproblems based on an estimate of the "active set" of constraints — those that are tight at optimality. We end this survey with a discussion of a practical algorithm [54], designed for large-scale applications in areas such as machine learning, and well-suited to the application of identifiability.

Given two Euclidean spaces $\mathbf{E}$ and $\mathbf{F}$ and a closed set $Y \subset \mathbf{F}$, we consider optimization problems of the following form:

$$\inf_{x \in \mathbf{E}} \big\{ f(x) : g(x) \in Y \big\}, \tag{10.1}$$

where the functions $f \colon \mathbf{E} \to \mathbf{R}$ and $g \colon \mathbf{E} \to \mathbf{F}$ are $\mathcal{C}^{(2)}$ smooth. Crucially, we suppose that the set Y is, in some sense, simple. We define "simple" operationally: we assume that we can solve relatively easily *prox-linear subproblems* of the form

$$\inf_{d \in \mathbf{E}} \big\{ \tilde{f}(d) + \mu |d|^2 : \tilde{g}(d) \in Y \big\}, \tag{10.2}$$

for *affine* functions $\tilde{f} \colon \mathbf{E} \to \mathbf{R}$ and $\tilde{g} \colon \mathbf{E} \to \mathbf{F}$, and a *prox parameter* $\mu > 0$. In the algorithm we describe, $\tilde{f}$ and $\tilde{g}$ are the linear approximations to f and g at the current iterate x_k:

$$\tilde{f}(d) = f(x_k) + Df(x_k)d \ \text{ and } \ \tilde{g}(d) = g(x_k) + Dg(x_k)d.$$

Consider again the example of optimization under inequality constraints, when the set Y is just a positive orthant. The corresponding prox-linear subproblem reduces to projection onto a polyhedron, a relatively easy problem computationally. Simpler still is the l_1-constrained least squares problem

$$\inf_{x \in \mathbf{R}^n} \big\{ |Ax - b|^2 : |x|_1 \leq \tau \big\},$$

for given $\tau > 0$, used to find sparse approximate solutions to huge linear systems $Ax = b$ in popular procedures such as LASSO and LARS [15, 23, 34, 71]. Corresponding prox-linear subproblems at the point x have the form

$$\inf_{d \in \mathbf{R}^n} \big\{ 2\langle Ax - b, Ad \rangle + \mu |d|^2 : |x + d|_1 \leq \tau \big\}. \tag{10.3}$$

This problem reduces to projection onto the l_1-ball, for which very fast algorithms are available: simple $O(n \log n)$ methods appear in [33, 73], and [33] describes an approach in expected linear time. (The computational simplicity of the singly-constrained convex program (10.3) is not surprising: its Lagrangian is separable in the components d_i, so can be minimized in linear time.) The nuclear-norm-constrained least squares approach for low-rank matrix equations is similar [14]. This time the corresponding subproblem is projection onto the nuclear-norm-ball (consisting of matrices whose singular values sum to at most one), which again is relatively easy: we simply replace the vector of singular values appearing in the singular value decomposition by its projection onto the l_1-ball.

Returning to the general problem (10.1), we consider a local minimizer $\bar{x}$ at which the set Y is prox-regular and satisfies the following standard constraint qualification:

$$\text{span } N_Y\big(g(\bar{x})\big) \cap \text{Null}\big(Dg(\bar{x})^*\big) \;=\; \{0\}. \tag{10.4}$$

This condition implies that $\bar{x}$ must satisfy the natural *first-order optimality condition*: there exists a *Lagrange multiplier* $y \in \mathbf{F}$ (in fact unique) such that

$$y \in N_Y\big(g(\bar{x})\big) \;\text{ and }\; Df(\bar{x}) + Dg(\bar{x})^* y = 0. \tag{10.5}$$

Furthermore, for any point $x \in \mathbf{E}$ near $\bar{x}$, if the prox parameter μ is large enough, then the prox-linear subproblem (10.2) has a unique small local minimizer $d(x)$, and in fact $d(x) = O(|x - \bar{x}|)$.

The basic structure of the algorithm we describe is standard in optimization. The prox parameter μ controls the size of the trial step suggested by the prox-linear subproblem. When μ is large enough, we can correct the trial step to generate a reasonable fraction of the improvement predicted by linearization. If that proves impossible, we retrench, rejecting the trial step and increasing μ.

To be more precise, suppose the current iterate is $x \in \mathbf{E}$, and the current value of the prox parameter is $\mu > 0$. We first calculate the trial step $d = d(x)$, the appropriate local minimizer for the prox-linear subproblem (10.2), so in particular

$$g(x) + Dg(x)d \in Y \tag{10.6}$$

holds. We then calculate the new iterate $x^+ \in \mathbf{E}$ by trying to *correct* the trial point $x + d$, aiming at three conditions. First, the correction should be not too large relative to the step:

$$\big|x^+ - (x + d)\big| \leq \frac{1}{2}|d|.$$

Secondly, the new iterate should be feasible: $g(x^+) \in Y$. Thirdly, the actual decrease in the objective should be at least a reasonable fraction of that predicted by linearization:

$$\frac{f(x) - f(x^+)}{f(x) - f(x + d)} \geq \frac{1}{2}.$$

Assuming μ is sufficiently large, the constraint qualification (10.4) ensures that such a correction x^+ exists. If we find it, we accept it as our new current iterate and proceed; if not, we reject it, double μ, and try the whole process again. A standard argument shows a rudimentary convergence result: any limit point of the sequence of iterates must satisfy the first-order optimality condition.

The ideas behind this algorithm date back three decades [9, 37]. An implementable version in general must overcome two hurdles. The first — that the prox-linear subproblem may have several local minimizers — may arise, but only for nonconvex sets Y. The second concerns the correction mechanism, which we leave unspecified. When the map g is linear, or in particular just the identity, the algorithm is workable without correction. The algorithm we have described for the special case $\inf_Y f$, for closed convex Y and smooth f (which covers l_1-constrained least squares, for example), is closely related to the successful SPARSA code for compressed sensing [78]. Some kind of correction step is crucial when the map g is nonlinear, no matter how large the prox parameter μ. In particular, the linearized constraint

(10.6) does not guarantee the feasibility condition $g(x+d) \in Y$. Even when the trial step is feasible, we may want to enhance it using second-order information, leading us back to the idea of identifiability.

The basic prox-linear algorithm that we have described is versatile: it is often simple to implement and applicable to large-scale problems. In general, however, its convergence is slow. For example, consider unconstrained minimization of a strictly convex quadratic: in this simple case, $f(x) = \langle x, Ax\rangle$ for a positive-definite self-adjoint map $A\colon \mathbf{E} \to \mathbf{E}$, the map g is just the identity, and the set Y is just $\mathbf{E}$. The prox-linear algorithm then becomes the method of steepest descent for f with a fixed step size, an algorithm that, as we observed in the introduction, converges linearly but slowly when the map A is ill-conditioned. If our algorithm can readily access second-order information, we might hope to accelerate convergence.

So far we have supposed that the set Y is simple enough to render the prox-linear subproblems relatively easy. Now assume we know more, namely the structure of the set's identifiable manifolds. For example, the identifiable manifolds of the l_1-ball in $\mathbf{R}^n$ are simply its interior along with the sets of vectors x with norm $|x|_1 = 1$ and constant sign pattern $(\operatorname{sgn} x_i)$, where $\operatorname{sgn}\gamma = \gamma/|\gamma|$, or zero if $\gamma = 0$.

This structural information about the set Y allows us to impose a *second-order optimality condition* of the kind guaranteed generically by Theorem 7.3 (Generic identifiability). Specifically, consider any point $\bar{x} \in \mathbf{E}$ satisfying feasibility ($g(\bar{x}) \in Y$), the constraint qualification (10.4), and the first-order optimality condition (10.5), and now suppose furthermore that Y has an identifiable manifold $\mathcal{M}$ at the point $g(\bar{x})$ for the normal vector y and that the objective f grows quadratically on the manifold $g^{-1}(\mathcal{M})$ around $\bar{x}$. This latter condition is classical, amounting to the requirement that the Hessian of the Lagrangian function $f+\langle y, g\rangle$ at $\bar{x}$ be positive definite on the tangent space to $\mathcal{M}$ at $\bar{x}$.

From the second-order optimality condition we deduce powerful consequences. First, around the critical point $\bar{x}$, the objective f must in fact grow at least quadratically not just on the identifiable manifold $\mathcal{M}$ but on the whole set Y. Secondly, initiated nearby, the prox-linear algorithm must converge to $\bar{x}$. Thirdly, the sequence of trial iterates $g(x_k)+Dg(x_k)d_k$ in Y generated by the prox-linear subproblems is asymptotically critical for the Lagrange multiplier y, and hence eventually lies in $\mathcal{M}$. The algorithm thus *identifies* $\mathcal{M}$, in principle allowing an eventual reduction of the original optimization problem to the classical equality-constrained problem $\inf\{f(x) : g(x) \in \mathcal{M}\}$, and thereby opening up the possibility of second-order methods and accelerated convergence, as in [80] for the LASSO problem.

11. Afterthoughts and acknowledgements

Variational analysis and nonsmooth optimization deserve a wide audience. A flourishing toolkit of elegant theory for several decades, the discipline's more computational impact is only now coming into focus. In its full generality (skirted here) the field can seem at first formidably technical. However, as this essay has tried to emphasize, the core ideas — the normal cone and metric regularity, for example — are intuitive and powerful in both theory and algorithms. Semi-algebraic variational analysis makes for an illuminating concrete testing ground for the theory. The reach of variational analysis in applications ranges from its historical roots in optimal control and the calculus of variations, through more recent domains such as eigenvalue optimization and robust control, and on to burgeoning areas like

compressed sensing and machine learning. The field is thriving.

The material in this essay strongly reflects what I have tried to learn from the many co-authors and mentors with whom I have been lucky enough to work. Among them, I would especially like to mention Jon Borwein (who taught me variational analysis), Jim Burke and Michael Overton (my enthusiastic companions watching theory made manifest on a computer screen), Jim Renegar (an inspiring source of encouragement), Jérôme Bolte and Aris Daniilidis (with whom I first explored the semi-algebraic world), and most recently Dima Drusvyatskiy and Alex Ioffe. Thanks too to Asen Dontchev, Mike Todd, and Steve Wright for their broad support, and their helpful suggestions on this manuscript.

The author is grateful to the Dipartimento di Ingegneria Informatica Automatica e Gestionale at the Università di Roma La Sapienza for its hospitality during the writing of this paper.

References

[1] K. Anstreicher and J. Lee, *A masked spectral bound for maximum-entropy sampling*, In A. Bucchianico, A. Läuter, and H. P. Wynn, editors, MODA 7 – Advances in Model-Oriented Design and Analysis, Springer, Berlin, 2004, pp. 1–10.

[2] S. Banach, *Sur les opérations dans les ensembles abstraits et leur application aux équations intégrales*, Fundamenta Mathematicae **3** (1922), 133–181.

[3] H.H. Bauschke and J.M. Borwein, *On projection algorithms for solving convex feasibility problems*, SIAM Review **38** (1996), 367–426.

[4] V. Blondel, *Simultaneous Stabilization of Linear Systems*, Springer, Berlin, 1994.

[5] J. Bôlte, A. Daniilidis, and A. S. Lewis, *Generic optimality conditions for semi-algebraic convex programs*, Mathematics of Operations Research **36** (2011), 55–70.

[6] J. F. Bonnans and A. Shapiro, *Perturbation Analysis of Optimization Problems*, Springer, New York, 2000.

[7] J. M. Borwein and A. S. Lewis, *Convex Analysis and Nonlinear Optimization*, Springer, New York, second edition, 2006.

[8] J. M. Borwein and Q. J. Zhu, *Techniques of Variational Analysis*, Springer, New York, 2005.

[9] J. V. Burke, *Descent methods for composite nondifferentiable optimization problems*, Mathematical Programming **33** (1985), 260–279.

[10] J. V. Burke, D. Henrion, A. S. Lewis, and M. L. Overton, *Stabilization via nonsmooth, nonconvex optimization*, IEEE Transactions on Automatic Control **51** (2006), 1760–1769.

[11] J. V. Burke, A. S. Lewis, and M. L. Overton, *Optimal stability and eigenvalue multiplicity*, Foundations of Computational Mathematics **1** (2001), 205–225.

[12] J. V. Burke and M. L. Overton, *Variational analysis of non-Lipschitz spectral functions*, Mathematical Programming **90** (2001), 317–352.

[13] ———, *Variational analysis of the abscissa mapping for polynomials*, SIAM Journal on Control and Optimization **39** (2001), 1651–1676.

[14] J.-F. Cai, E. Candès, and Z. Shen, *A singular value thresholding algorithm for matrix completion*, SIAM Journal on Optimization **20** (2010), 1956–1982.

[15] E. J. Candès and T. Tao, *Near-optimal signal recovery from random projections: universal encoding strategies*, IEEE Transactions on Information Theory **52** (2007), 5406–5425.

[16] F.H. Clarke, *Necessary Conditions for Nonsmooth Problems in Optimal Control and the Calculus of Variations*, PhD thesis, University of Washington, Seattle, 1973.

[17] ———, *Generalized gradients and applications*, Transactions of the American Mathematical Society **205** (1975), 247–262.

[18] F. H. Clarke, Yu. S. Ledyaev, R. J. Stern, and P. R. Wolenski, *Nonsmooth Analysis and Control Theory*, Springer-Verlag, New York, 1998.

[19] M. Coste, *An Introduction to O-minimal Geometry*, RAAG Notes, 81 pages, Institut de Recherche Mathématiques de Rennes, 1999.

[20] ———, *An Introduction to Semialgebraic Geometry*, RAAG Notes, 78 pages, Institut de Recherche Mathématiques de Rennes, 2000.

[21] A. Daniilidis, D. Drusvyatskiy, and A. S. Lewis, *Orthogonal invariance and identifiability*, SIAM Journal on Optimization, 2014. To appear.

[22] J.W. Demmel, *On condition numbers and the distance to the nearest ill-posed problem*, Numerische Mathematik **51** (1987), 251–289.

[23] D. Donoho, *Compressed sensing*, IEEE Transactions on Information Theory **52** (2006), 1289–1306.

[24] A. L. Dontchev and H. Frankowska, *On derivative criteria for metric regularity*, In Computational and Analytical Mathematics, Springer Proceedings in Mathematics and Statistics, Springer, New York, 2013, pp. 365–374.

[25] A. L. Dontchev, A. S. Lewis, and R. T. Rockafellar, *The radius of metric regularity*, Transactions of the American Mathematical Society **355** (2003), 493–517.

[26] A. L Dontchev and R. T. Rockafellar, *Implicit Functions and Solution Mappings: a View from Variational Analysis*, Springer, New York, 2009.

[27] D. Drusvyatskiy, A. D. Ioffe, and A. S. Lewis, *The dimension of semi-algebraic subdifferential graphs*, Nonlinear Analysis **75** (2012), 1231–1245.

[28] ———, *Alternating projections and coupling slope*, 2014. Preprint. arXiv:1401.7569.

[29] D. Drusvyatskiy and A. S. Lewis, *Optimality, identifiability, and sensitivity*, Mathematical Programming, 2013. DOI 10.1007/s10107-013-0730-4.

[30] _____, *Semi-algebraic functions have small subdifferentials*, Mathematical Programming, Series B **140** (2013), 5–29.

[31] _____, *Tilt stability, uniform quadratic growth, and strong metric regularity of the subdifferential*, SIAM Journal on Optimization **23** (2013), 256–267.

[32] _____, *Strong regularity of semi-algebraic mappings*, 2014. Preprint.

[33] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra, *Efficient projections onto the l_1-ball for learning in high dimensions*, In Proceedings of the 25th International Conference on Machine Learning, Helsinki, 2008.

[34] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, *Least angle regression*, Annals of Statistics **32**(2) (2004), 407–499.

[35] I. Ekeland, *On the variational principle*, Journal of Mathematical Analysis and Applications **47** (1974), 324–353.

[36] R. Fletcher, *A new variational result for quasi-Newton formulae*, SIAM Journal on Optimization **1** (1991), 18–21.

[37] R. Fletcher and E. Sainz de la Maza, *Nonlinear programming and nonsmooth optimization by successive linear programming*, Mathematical Programming **43** (1989), 235–256.

[38] K. M. Grigoriadis and R. E. Skelton, *Low-order control design for LMI problems using alternating projection methods*, Automatica **32** (1996), 1117–1125.

[39] A. Grothendieck, *Esquisse d'un programme*, In L. Schneps and P. Lochak, editors, Geometric Galois Actions, volume 1. Cambridge University Press, Cambridge, U.K., 1997. London Mathematical Society Lecture Note Series 242.

[40] R. A. Horn and C. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, U.K., 1985.

[41] A. D. Ioffe, *Metric regularity and subdifferential calculus*, Russian Mathematical Surveys **55** (2000), 501–558.

[42] _____, *A Sard theorem for tame set-valued mappings*, Journal of Mathematical Analysis and Applications **335** (2007), 882–901.

[43] _____, *On the theory of subdifferentials*, Advances in Nonlinear Analysis **1** (2012), 47–120.

[44] N. H. Josephy, *Newton's method for generalized equations and the PIES energy model*, PhD thesis, Dept. of Industrial Engineering, University of Wisconsin-Madison, 1979.

[45] A. Y. Kruger and B. S. Mordukhovich, *Extremal points and the Euler equation in nonsmooth analysis*, Doklady Akademia Nauk BSSR (Belorussian Academy of Sciences) **24** (1980), 684–687.

[46] P. D. Lax, *Linear Algebra*, Wiley, New York, 1997.

[47] J. M. Lee, *Introduction to Smooth Manifolds*, Springer, New York, 2003.

[48] A. S. Lewis, *Convex analysis on the Hermitian matrices*, SIAM Journal on Optimization **6** (1996), 164–177.

[49] ———, *Nonsmooth analysis of eigenvalues*, Mathematical Programming **84** (1999), 1–24.

[50] ———, *Active sets, nonsmoothness and sensitivity*, SIAM Journal on Optimization **13** (2003), 702–725.

[51] ———, *Eigenvalues and nonsmooth optimization*, In L.M. Pardo, A. Pinkus, E. Suli, and M.J. Todd, editors, Foundations of Computational Mathematics, Santander 2005, Cambridge University Press, Cambridge, U.K., 2005, pp. 208–229.

[52] A. S. Lewis, D. R. Luke, and J. Malick, *Local linear convergence for alternating and averaged nonconvex projections*, Foundations of Computational Mathematics **3** (2009), 485–513.

[53] A. S. Lewis and M. L. Overton, *Nonsmooth optimization via quasi-Newton methods*, Mathematical Programming **141** (2013), 135–163.

[54] A. S. Lewis and S. J. Wright, *A proximal method for composite minimization*, 2008. Preprint. arXiv:0812.0423v1.

[55] A. S. Lewis and S. Zhang, *Partial smoothness, tilt stability, and generalized Hessians*, SIAM Journal on Optimization **23** (2013), 74–94.

[56] Z.-Q. Luo and P. Tseng, *On the linear convergence of descent methods for convex essentially smooth minimization*, SIAM Journal on Control and Optimization **30** (1992), 408–425.

[57] B. S. Mordukhovich, *Maximum principle in the problem of time optimal response with nonsmooth constraints*, Journal of Applied Mathematics and Mechanics **40** (1976), 960–969.

[58] ———, *Variational Analysis and Generalized Differentiation, I: Basic Theory; II: Applications*, Springer, New York, 2006.

[59] B. S. Mordukhovich and A. Y. Kruger, *Necessary optimality conditions in the problem of terminal control with nonfunctional constraints. (Russian)*, Doklady Akademia Nauk BSSR (Belorussian Academy of Sciences) **20** (1976), 1064–1067.

[60] Y. E. Nesterov and A. S. Nemirovskii, *Interior-Point Polynomial Algorithms in Convex Programming*, SIAM, Philadelphia, 1994.

[61] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer, New York, second edition, 2006.

[62] T. Pennanen, *Local convergence of the proximal point algorithm and multiplier methods without monotonicity*, Mathematics of Operations Research **27** (2002), 170–191.

[63] R. A. Poliquin and R. T. Rockafellar, *Tilt stability of a local minimum*, SIAM Journal on Optimization **8** (1998), 287–299.

[64] R. A. Poliquin, R. T. Rockafellar, and L. Thibault, *Local differentiability of distance functions*, Transactions of the American Mathematical Society **352** (2000), 5231–5249.

[65] J. Renegar, *Condition numbers, the barrier method, and the conjugate gradient method*, SIAM Journal on Optimization **6** (1996), 879–912.

[66] S. M. Robinson, *Strongly regular generalized equations*, Mathematics of Operations Research **5** (1980), 43–62.

[67] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, N.J., 1970.

[68] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Springer, Berlin, 1998.

[69] A. Sard, *The measure of the critical values of differentiable maps*, Bulletin of the American Mathematical Society **48** (1942), 883–890.

[70] J. E. Spingarn and R. T. Rockafellar, *The generic nature of optimality conditions in nonlinear programming*, Mathematics of Operations Research **4** (1979), 425–430.

[71] R. Tibshirani, *Regression shrinkage and selection via the LASSO*, Journal of the Royal Statistical Society B **58** (1996), 267–288.

[72] M. J. Todd, *Semidefinite optimization*, Acta Numerica **10** (2001), 515–560.

[73] E. van den Berg and M. P. Friedlander, *Probing the Pareto frontier for basis pursuit solutions*, SIAM Journal on Scientific Computing **31** (2008), 890–912.

[74] L. van den Dries and C. Miller, *Geometric categories and o-minimal structures*, Duke Mathematics Journal **84** (1996), 497–540.

[75] J. von Neumann, *Some matrix inequalities and metrization of matric-space*, Tomsk University Review **1** (1937), 286–300. In: Collected Works, Pergamon, Oxford, 1962, Volume IV, 205-218.

[76] _____, *Functional Operators*, volume II. Princeton University Press, Princeton, NJ, 1950. Reprint of notes distributed in 1933.

[77] H. Whitney, *A function not constant on a connected set of critical points*, Duke Mathematics Journal **1** (1935), 514–517.

[78] S. J. Wright, R. D. Nowak, and M. A. T. Figueiredo, *Sparse reconstruction by separable approximation*, IEEE Transactions on Signal Processing **57** (2009), 2479–2493.

[79] S. J. Wright, *Identifiable surfaces in constrained optimization*, SIAM Journal on Control and Optimization **31** (1993), 1063–1079.

[80] _____, *Accelerated block-coordinate relaxation for regularized optimization*, SIAM Journal on Optimization **22** (2012), 159–186.

School of Operations Research and Information Engineering, Cornell University, Ithaca NY 14853, USA

E-mail: adrian.lewis@cornell.edu