Local structure and algorithms in nonsmooth optimization

Adrian S. Lewis

(joint work with James V. Burke and Michael L. Overton)

The Belgian chocolate problem. Illustrating the difficulty of control design problems, Blondel [1] proposed the following problem in 1994:

> Given a real δ, find stable real polynomials p and q such that the polynomial $r(s) = (s^2 - 2\delta s + 1)p(s) + (s^2 - 1)q(s)$ is also stable.

(We call a polynomial p *stable* if its *abscissa* $\alpha(p) = \max\{\mathrm{Re}\, s : p(s) = 0\}$ is nonpositive.) Clearly the problem is unsolvable if $\delta = 1$, since then $r(1) = 0$; more delicate results (summarized in [7]) show it remains unsolvable for $\delta < 1$ close to 1. Blondel offered a prize of 1kg of Belgian chocolate for the case $\delta = 0.9$, a problem solved via randomized search in [7].

To illustrate the theme of this talk, we first outline (based on joint work with D. Henrion) a more systematic, optimization approach to the chocolate problem. We fix the degrees of the polynomials p and q (say 3, for example), without loss of generality suppose p is monic, and consider the resulting problem

$$(CP) \qquad \min\{\alpha(pqr) : p, q \text{ cubic},\ p \text{ monic}\}.$$

A feasible solution with negative objective value would solve Blondel's problem.

A simple nonsmooth algorithm. For nonsmooth optimization problems like (CP), it is convenient to have on hand a simply-implementable, intuitive, robust algorithm for minimizing a nonsmooth function f. We present such a method in [3]. To motivate it, suppose for simplicity f (unlike the abscissa α) is Lipschitz.

Fundamental for good behavior in nonsmooth optimization is the *regularity* of the function f at points x, which means we can write the directional derivative as

$$f'(x; d) = \limsup_{y \to x} \nabla f(x)^T d, \text{ for all } d$$

(noting the almost everywhere differentiability of f on its domain in $\mathbf{R}^n$). Both convex and smooth functions are regular. Assuming regularity, we can check that the steepest descent direction at x is

$$-\lim_{\epsilon \downarrow 0} \mathrm{argmin}\big\{\|d\| : d \in \mathrm{conv}\{\nabla f(y) : y \in x + \epsilon B\}\big\},$$

where B denotes the unit ball. The *Gradient Sampling* algorithm of [3] approximates this direction by a random vector

$$G_\epsilon^m(x) = -\mathrm{argmin}\big\{\|d\| : d \in \mathrm{conv}\{\nabla f(Y_i) : i = 1, 2, \dots, m\}\big\},$$

for some fixed radius ϵ, fixed $m > n$, and independent, uniformly distributed, random points $Y_i \in x + \epsilon B$. (In practice, we add the point x.) The algorithm then performs a simple line search along this direction, and repeats.

The performance of Gradient Sampling. The Gradient Sampling algorithm is intuitive, and straightforward to implement when function and gradient evaluations are cheap. Experiments on a wide variety of examples are very promising [3]. Rigorous justifications include the almost sure convergence of the search direction $G^m_\epsilon(x)$ to a "robust" steepest descent direction as the sample size m grows [2], and convergence results for the algorithm under a variety of underlying assumptions and implementation regimes (for reducing the radius ϵ, for example) [3]. Among these results, however, the following fact is particularly suggestive of the "smoothing" effect of the algorithm.

Theorem 1. *The expectation of the search direction $G^m_\epsilon(x)$ depends continuously on the point x.*

We sketch a proof suggested by S. Henderson. First, we sample the points Y_i corresponding to the current point x, as above. Next, we construct random points Y_i' corresponding to a perturbed point x', but "coupled" with the points Y_i as follows. If $Y_i \in x' + \epsilon B$, then we set $Y_i' = Y_i$; otherwise we choose Y_i' uniformly distributed on the set $(x' + \epsilon B) \setminus (x + \epsilon B)$. The resulting random points Y_i' are mutually independent, and uniformly distributed on the ball $x' + \epsilon B$, as required. Since the set $(x + \epsilon B) \setminus (x' + \epsilon B)$ has measure $O(\|x - x'\|)$, the sets $\{Y_i\}$ and $\{Y_i'\}$ (and hence the vectors $G^m_\epsilon(x)$ and $G^m_\epsilon(x')$) are identical with probability $1 - O(\|x - x'\|)$. On the other hand, even if this latter event does not occur, since f is Lipschitz, the vector $G^m_\epsilon(x) - G^m_\epsilon(x')$ is uniformly bounded. In summary, the expectation of this latter vector must be $O(\|x - x'\|)$.

Solving the chocolate problem. The Gradient Sampling algorithm suggests numerically that the solution of the problem (CP) for any value of δ near 0.9 has a distinctive structure: the polynomial q is a constant, and the polynomial r has a negative real zero of order five. Armed with this observation, a simple hand calculation reveals a unique feasible solution of this form under the assumption $\delta < \frac{1}{2}\sqrt{2 + \sqrt{2}} \approx 0.924$, in particular solving Blondel's problem.

A nice exercise in nonsmooth calculus verifies our numerical observation that the above solution is indeed a local minimizer for the problem (CP), at least when we further restrict the polynomial q to be constant. The requisite nonsmooth chain rule we need relies heavily on the following striking result [4].

Theorem 2. *The abscissa α is regular throughout the set of degree-k polynomials.*

Structural persistence in nonsmooth optimization. The persistent solution structure for the chocolate problem (CP) as the parameter δ varies illustrates another important feature of concrete nonsmooth optimization problems, akin to active set phenomena in nonlinear programming. For classical nonlinear programs, the second-order sufficient conditions have several important consequences:

- (i) the current point is a strict local minimizer;
- (ii) as we perturb the problem's parameters, this minimizer varies smoothly on an "active" manifold;
- (iii) we can calculate perturbed minimizers via smooth systems of equations.

Properties (ii) and (iii) do not rely fundamentally on second-order theory, and indeed they also hold for a broad class of nonsmooth functions introduced in [6].

For simplicity once again, we restrict attention to Lipschitz functions f. We call f *partly smooth* relative to the *active manifold* $\mathcal{M}$ if f is regular throughout $\mathcal{M}$ and the directional derivative $f'(x;d)$ is continuous as x varies on $\mathcal{M}$, with

$$f'(x;-d) > -f'(x;d) \text{ whenever } 0 \neq d \perp \mathcal{M} \text{ at } x.$$

This last condition enforces a "vee-shape" on the graph of f around a "ridge" corresponding to $\mathcal{M}$. Partial smoothness holds, for example, for the function $x \mapsto \max\{x_i\}$, the Euclidean norm, and the maximum eigenvalue of a symmetric matrix, and the property is typically preserved under smooth composition, generating a wealth of applications. Furthermore, critical points of partly smooth functions typically satisfy the sensitivity properties (ii) and (iii) above.

The structural persistence we first observed numerically in the chocolate problem (CP) is explained by the following refinement of Theorem 2. We associate with any polynomial p a *list of multiplicities* of those zeroes of p with real part equal to the abscissa, listed in order of decreasing imaginary part.

Theorem 3. *The abscissa α is partly smooth relative to any manifold of polynomials having a fixed list of multiplicities.*

By contrast with the sensitivity properties (ii) and (iii) above, convenient checks for property (i) (strict local minimality) do typically involve second-order analysis. For partly smooth functions f, the extra assumption we need is *prox-regularity* [5]. This property requires, locally, that the nearest-point projection onto the epigraph $\{(x,r) : r \geq f(x)\}$ should be unique (as typically holds if f is the pointwise maximum of some smooth functions, for example). The question of the prox-regularity of the abscissa α remains open. The essential ingredient is the following question, with which we end.

Question 1. *Does every degree-k polynomial $p(s)$ near the polynomial s^k have a unique nearest stable polynomial?*

References

[1] V.D. Blondel *Simultaneous stabilization of linear systems*, Springer, Berlin (1994).

[2] J.V. Burke, A.S. Lewis and M.L. Overton *Approximating subdifferentials by random sampling of gradients*, Math. of O.R. **27** (2002), 567–584.

[3] J.V. Burke, A.S. Lewis and M.L. Overton *A robust gradient sampling algorithm for nonsmooth, nonconvex optimization*, SIAM J. Opt, to appear.

[4] J.V. Burke and M.L. Overton *Variational analysis of the abscissa mapping for polynomials*, SIAM J. Control Opt. **39** (2001), 1651–1676.

[5] W.L. Hare and A.S. Lewis, *Identifying active sets via partial smoothness and prox-regularity*, J. Convex Analysis (2004), to appear.

[6] A.S. Lewis, *Active sets, nonsmoothness and sensitivity*, SIAM J. Opt. **13** (2003), 702–725.

[7] V.V. Patel, G. Deodhare and T. Viswanath *Some applications of randomized algorithms for control system design*, Automotica **28** (2002), 2085–2092.