Sparse recovery in Derivative-Free Optimization

Afonso S. Bandeira
Program in Applied and Computational Mathematics
Princeton University
ajsb@math.princeton.edu

First of all, it is an honor to receive the INFORMS Optimization Society best student paper award. Before going into the content of the paper, I want to give a bit of context in which I did this research. This paper was essentially my master thesis work in the University of Coimbra, in Portugal, back in 2009-2010 (In fact, my master thesis has the same title [Ban10]). Katya Scheinberg and Luís Nunes Vicente proposed me to work on the connection between the, then recent, developments in sparse recovery and Compressed Sensing (subject I was, and still am, quite interested in) and Derivative-Free Optimization, a subject I was not familiar with at the time but which I very quickly learned to enjoy. During this period I was extremely fortunate to have had the opportunity to work with both Luis and Katya and, together, we wrote the awarded paper, "Computation of sparse low degree interpolating polynomials and their application to derivative-free optimization" [BSV12]. This work has spurred both parallel [BSV11] and future research [BSV13] that I will also briefly describe below.

The framework is unconstrained optimization: one wants to minimize a (sufficiently smooth) function $f\colon \mathbb{R}^n \to \mathbb{R}$ over $\mathbb{R}^n$. In many applications function evaluations are particularly expensive and one has no access to function derivatives (an important example is when the goal is to do parameter optimization and each evaluation requires an expensive simulation). These applications motivate the interest in optimizing f without using its derivatives and using as few function evaluations as possible, this is known as Derivative-Free Optimization. An excellent introduction to the topic is given in the book [CSV09].

One type of algorithms used in DFO are the, so called, model-based trust-region methods. Essentially, they operate by iteratively picking a small region, known as the trust-region, $B \subset \mathbb{R}^n$ (say a ball) and build (via approximation on samples of f) a model $m\colon B \to \mathbb{R}$ of f in B. The idea is for $m(x)$ to be easier to optimize in B while being a reliable model of f, and for minimizer of $m(x)$ to be an estimate for the minimizer of f in B. Depending on the location of the minimizer and its value on f the trust-region is updated and the procedure repeated.

A very popular class of models used is the quadratic polynomials, as these are reasonably simple while being able capturing curvature of the function (unlike linear models). For the sake of simplicity, let us suppose for the moment that f is itself a quadratic and we want to find the model $m = f$. Constructing $m(x)$ from p function evaluations of f corresponds to solving a linear system. Let $\phi_1, \ldots, \phi_N$ be a basis for the quadratic polynomials of n variables (meaning $N = \frac{(n+3)n+2}{2}$). We can write $m(x) = \sum_{i=1}^N \alpha_i \phi_i(x)$ and focus on estimating the $\{\alpha_i\}_{i=1}^N$. In fact, a function evaluation of f at y_j gives a linear constraint on α,

$$\sum_{i=1}^N \alpha_i \phi_i(y_j) = f(y_j).$$

A sample set Y of p points corresponds to p linear constraints on α which we represent by an interpolation matrix $M(\phi, Y)$

$$M(\phi, Y)\alpha = f(Y). \tag{1}$$

In order for (1) to be a determined system one needs $p \geq N = \frac{(n+3)n+2}{2}$ function evaluations in each iterations, which is very often too expensive.

Indeed, without any extra information on the structure of f, this is the best one can do. Fortunately, most functions that arise from applications have special structures. For example, in the parameter estimation problem, it is rather unlikely that every pair of parameters is interacting (in a relevant way). Pairs of parameters not interacting should correspond to zero joint derivatives which suggests sparsity of the function's Hessian. This motivated us to pursue techniques that exploited the Hessian sparsity in order to construct reliable models with far fewer samples, which is precisely the subject of the paper [BSV12].

Provided we choose a basis $\{\phi\}_{i=1}^N$ such that Hessian sparsity translates into sparsity in α, the Hessian sparsity of f corresponds to sparsity of the solution of the linear system (1). Around a decade ago, the seminal work of Candes, Donoho, and others [CRT06a], [CRT06b], [CT05], [CT06], [Don06], spurred a vast body of exciting research in the area of sparse recovery (also known as Compressed Sensing) which provides very good understanding of when one is able to recover a sparse vector α from a underdetermined linear system. The main contribution of this paper is leveraging, and adapting, these results to estimate α (which gives a model $m(x)$), via (1), using significantly less than $N = \mathcal{O}(n^2)$ function evaluations.

In a nutshell, the theory of Compressed Sensing tells us that, if $M(\phi, Y)$ satisfies a certain property (known as the Restricted Isometry Property (RIP) [Can08]), the sparse vector α can be recovered by ℓ_1 minimization (essentially, it is the vector with minimum ℓ_1 norm which still satisfies the linear constraints). Matrices satisfying the Restricted

Isometry Property are notably difficult to build and computationally hard to certify [BDMS13] (I have spent some time thinking about this myself [BFMW13]) but random constructions are known to yield RIP matrices for a number of rows (corresponding to samples) p on the order of $p = \mathcal{O}(k \log N)$, where k is the sparsity of the vector, and N the ambient dimension (in our case $N = \mathcal{O}(n^2)$). This means that, as long as $M(\phi, Y)$ is RIP, the number of samples needed is no longer on the order of vector dimension but instead, on the order of the sparsity of α (with a small logarithmic loss). Moreover, ℓ_1 minimization can be formulated as a linear program thus enjoying many efficient algorithms.

Classically, the results in Compressed Sensing guaranteeing the RIP property for random matrices mostly concern matrices with independent entries. In our setting, however, we are constrained to a very structured interpolation matrix $M(\phi, Y)$. Knowing how difficult constructing good deterministic RIP matrices seems to be, we opted to "inject randomness" in the matrix by taking the sample set Y to be random (while the basis $\{\phi\}_{i=1}^N$ is fixed and deterministic). In fact, provided that the basis $\{\phi\}_{i=1}^N$ satisfies certain properties, a sufficiently large random sample set Y gives an interpolation matrix $M(\phi, Y)$ which is known [Rau10] to be RIP with high probability. In our paper [BSV12], we are able to build a basis $\{\phi\}_{i=1}^N$ both inheriting sparsity from Hessian sparsity and satisfying the properties needed to yield RIP interpolation matrices. This, together with a particular choice of trust-region and sampling measure for Y allowed us to show that, with high probability, ℓ_1 minimization succeeds in recovering α from the linear measurements (1) with as few as

$$p = \mathcal{O}(n \log^4(n))$$

samples, provided that the Hessian of f has $\mathcal{O}(n)$ non-zero entries. Note that this number of samples (corresponding to function evaluations) is considerably less than the $\mathcal{O}(n^2)$ samples that would be needed in the classical case.

In general, f is not a quadratic polynomial. However, as long as f is sufficiently smooth, one can show that the procedure sketched above gives, with the same number of samples, a model m that approximates f in B essentially as well as its second-order Taylor approximation (these are known as fully-quadratic models). The idea is to replace f with its second-order Taylor approximation in the arguments above. In that case, each sample of f can be regarded as a noisy sample of the quadratic approximation. Fortunately, the guarantees given in the theory of sparse recovery often come with robustness to noise and, in this case, we can leverage such results to ensure the recovery of a fully-quadratic model of f, with high probability.

The sparsity assumption used is on the Hessian of the function, however the coefficients α also describe the gradient and constant term which may not be sparse. This means that there are some entries of the vector α that are not believed to be sparse. Motivated by this fact, we investigated the problem of sparse recovery for partially sparse vectors [BSV11]. We showed that, not very surprisingly, one should do ℓ_1 minimization only on the entries that are believed to be sparse.

Using the machinery described above, we developed a model-based trust-region method based on minimum ℓ_1 model construction. In our experiments, this method was able to compete with state of the art Derivative-Free methods, such as NEWUOA [Pow06], [Pow03] on the standard problem data base CUTEr [GOT03].

A natural question raised by this work regard the convergence of methods based on this type of random models. Recall that recovery is only ensured with high probability at each iteration and so it will likely fail on some iterations. This question was the target of further research [BSV13] where we showed that, under somewhat general conditions, the convergence guarantees for model-based trust-region methods can be adapted to handle this uncertainty in the model construction step. Essentially, we showed that, as long as the probability of constructing a good model on each iteration is over one half, these methods still converge.

REFERENCES

[Ban10] A. S. Bandeira, *Computation of sparse low degree interpolating polynomials and their application to derivative-free optimization*, Master's thesis, Dept. of Mathematics, Univ. of Coimbra, 2010.

[BDMS13] A.S. Bandeira, E. Dobriban, D.G. Mixon, and W.F. Sawin, *Certifying the restricted isometry property is hard*, IEEE Trans. Inform. Theory **59** (2013), no. 6, 3448–3450.

[BFMW13] A. S. Bandeira, M. Fickus, D. G. Mixon, and P. Wong, *The road to deterministic matrices with the restricted isometry property*, Journal of Fourier Analysis and Applications **19** (2013), no. 6, 1123–1149.

[BSV11] A. S. Bandeira, K. Scheinberg, and L. N. Vicente, *On partial sparse recovery*, Tech. report, CMUC, Department of Mathematics, University of Coimbra, Portugal, 2011.

[BSV12] ———, *Computation of sparse low degree interpolating polynomials and their application to derivative-free optimization*, Mathematical Programming **134** (2012), 223–257.

[BSV13] ———, *Convergence of a trust region method based on probabilistic models*, Available online at arXiv:1304.2808 [math.OC] (2013).

[Can08] E. J. Candès, *The restricted isometry property and its implications for compressed sensing*, C. R. Acad. Sci. Paris, Ser. I **346** (2008), 589–592.

[CRT06a] E. J. Candès, J. Romberg, and T. Tao, *Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information*, IEEE Trans. Inform. Theory **52** (2006), 489–509.

[CRT06b] ———, *Stable signal recovery from incomplete and inaccurate measurements*, Comm. Pure Appl. Math. **59** (2006), 1207–1223.

[CSV09] A. R. Conn, K. Scheinberg, and L. N. Vicente, *Introduction to derivative-free optimization*, MPS-SIAM Series on Optimization, SIAM, Philadelphia, 2009.

[CT05] E. J. Candès and T. Tao, *Decoding by linear programming*, IEEE Trans. Inform. Theory **51** (2005), 4203–4215.

[CT06] ———, *Near optimal signal recovery from random projections: universal encoding strategies?*, IEEE Trans. Inform. Theory **52** (2006), 5406–5425.

[Don06] D. L. Donoho, *Compressed sensing*, IEEE Trans. Inform. Theory **52** (2006), 1289–1306.

[GOT03] N. I. M. Gould, D. Orban, and P. L. Toint, *Cuter and sifdec: A constrained and unconstrained testing environment, revisited*, ACM Trans. Math. Softw. **29** (2003), no. 4, 373–394.

[Pow03] M. J. D. Powell, *On trust region methods for unconstrained minimization without derivatives*, Mathematical Programming **97** (2003), no. 3, 605–623.

[Pow06] ———, *The newuoa software for unconstrained optimization without derivatives*, Large-Scale Nonlinear Optimization (G. Pillo and M. Roma, eds.), Nonconvex Optimization and Its Applications, vol. 83, Springer US, 2006, pp. 255–297.

[Rau10] H. Rauhut, *Compressive sensing and structured random matrices*, Theoretical Foundations and Numerical Methods for Sparse Recovery (M. Fornasier, ed.), Radon Series Comp. Appl. Math., vol. 9, deGruyter, 2010, pp. 1–92.