Intrinsic methods for optimization problems

Department of Mathematical Sciences
University of Copenhagen
Universitetsparken 5
DK-2100 Copenhagen, Denmark
Email: topsoe@math.ku.dk

***Abstract*— General optimization techniques typically depend on analytical tools involving differentiation. However, for many problems, special *intrinsic* tools lead in a more natural way to insight. For information theory, instances of this phenomenon includes problems on the determination of capacity, the search for universal codes or maximum entropy distributions and the calculation of information projections. Problems from other areas may also be treated by the intrinsic methods here presented.**

I. Introduction

Recently, the charming monograph [2]: "Optimization: Insights and Applications" by Brinkhuis and Tikhomirov appeared. The authors discuss at length and in depth what appears as trivial routine to many a high-school student, the *theorem of Fermat* on the efficiency of the "$f'(x) = 0$-method" when seeking extrema of a real function. True, this method (in its modern formulation and with availability of a fully developed differential calculus) is a wonderful and essential result. And so are refinements as those associated with the names of Lagrange, Karush, Kuhn and Tucker for problems of optimization involving constraints with (or without) the presense of convexity assumptions. Noone can argue meaningfully against that. And the recorded success of the approach is overwhelming and spans all fields of science.

Though this is so, one should be open to the possibility that when you deal with a more specialized problem, there may be a chance that intrinsic techniques specific to the particular subject area may well be applicable and even provide more insight. Let us go further and propose the thesis that *when a specific problem of optimization is "canonical", i.e. works with "just the right concepts" and reflects "just the right questions", then intrinsic tools are the way forward to insight*.

As a trivial example, seek the minimum of $x^2 - 6x + 1$. Differentiate – and you are done. But is it not better to use the intrinsic rules of algebra and simply rewrite the expression as $(x-3)^2-8$? Of course it is. And we see things more clearly, e.g. indeed, we do find a *minimum* this way.

The attitude expressed is illustrated well by problems of information theory. Not that surprising. Did not Shannon [15] teach us "just the right concepts" and helped us ask "just the right questions"?

Some words about the scope of the contribution are in order. Rather than presenting results in a logical meticulous manner, it is the purpose to demonstrate that intrinsic methods is the way forward in many situations where analytical methods is most commonly applied. This is achieved – so is the intention at least – by going through one model example (MaxEnt calculations), by using that example as motivation for an axiomatic approach and finally by indicating further examples (including one from geometry) which can be handled by the theory indicated. We shall not dwelve on technical assumptions or attempt completeness. More details will be published in two papers under preparation, one for the "Journal of Global Optimization", another for the electronic journal "Entropy". Also, manuscripts on my homepage may be helpful.

II. MaxEnt

Let X denote the *preparation* consisting of all probability distributions on a discrete alphabet, A, for which the mean *energy*, $\overline{E}$, is prescribed. Using $\langle x, \cdot \rangle$ for mean value w.r.t. a probability distribution x and E for the energy function, we may write the preparation as $X = \{\langle x, E\rangle = \overline{E}\}$. The problem is to determine the *MaxEnt-distribution*, the distribution in the preparation with maximal entropy. Denoting by $H_{\max}$ the maximum value in question, we thus seek $x \in X$ with $H(x) = H_{\max}$. The significance of this problem (and its natural extensions) which occur in many disguises is recognized since long as witnessed in [11], for example.

If you introduce Lagrange multipliers, one to express that x is a normalized measure, another to express the constraint, you are soon led to the quantities that govern the solution. Easy and fast. But it pays to go more slowly about it. First, let us remind ourselves what Shannon taught us: entropy may be conceived as minimal average code length. Idealizing in a well known way, this fact may be expressed by the introduction of the set Y of *codes* (or rather, *idealized code length functions*), functions $i \mapsto y_i$ on A that satisfy *Kraft's equality*,

$$\sum_{i \in A} e^{-y_i} = 1\,. \tag{1}$$

Defining the function Φ on $X \times Y$ – referred to as the *complexity function*, but interpreted as *average code length* – by

$$\Phi(x, y) = \langle x, y\rangle\,, \tag{2}$$

we realize that entropy may be obtained from Φ by the formula

$$H(x) = \inf \Phi_x\,, \tag{3}$$

the infimum (minimum in fact) of the *marginal function*, $\Phi_x : y \mapsto \Phi(x, y)$. Comparing Φ and H we are led to consider the

redundancy

$$D(x,y) = \Phi(x,y) - H(x) = \sum_{i\in A} x_i \ln \frac{x_i}{\exp(-y_i)}\,. \qquad (4)$$

The three basic quantities, complexity, entropy and redundancy are connected by the *linking identity*

$$\Phi(x,y) = H(x) + D(x,y)\,. \qquad (5)$$

Furthermore, we realize that the *fundamental inequality*

$$D(x,y) \geq 0 \qquad (6)$$

holds with equality if and only if y is *adapted to* x, written $y = \hat{x}$, i.e. $y_i = \ln \frac{1}{x_i}$ for every $i \in A$.

The proper intrinsic handling of the MaxEnt problem then rests on two facts, one general and the other more specific. For the general observation we introduce the notion of a *robust code*. This is an element $y \in Y$ for which there exists a finite constant, h, the *level of robustness*, such that $\Phi(x,y) = h$ for all $x \in X$.

Lemma 2.1 (robustness lemma): Assume that $x^* \in X$ and that $y^* = \widehat{x^*}$ is robust. Then x^* is the MaxEnt distribution and $H_{\max} = h$, the level of robustness.

Proof: This follows since $H(x^*) = \Phi(x^*, y^*) = h$ and since, for $x \in X \setminus \{x^*\}$, $H(x) < H(x) + D(x, y^*) = \Phi(x, y^*) = h$. ∎

The specific fact we need is the simple observation that every code y which is of the form $y = \alpha + \beta E$ with α and β two constants is robust. In view of Kraft's equality, these codes really only contain one parameter, say β. One may then attempt to adjust β so that the distribution x with $\hat{x} = y$ is a member of the preparation X. If you succeed, you are done – by the robustness lemma, you have found the MaxEnt distribution.

When you compare the analytical approach with the approach above via robustness, you realize that you are – of course – led to the same expressions and will then obtain the same result by the two methods. The latter approach is just as fast, more elementary and tells a good deal more about the structure. This will be even more evident if you conceive Φ as an *objective function* for a two-person zero-sum game with one player, a *maximizer*, choosing a strategy from X and the other, a *minimizer*, choosing a strategy from Y. For instance, from the game theoretical approach you derive the *Pythagorean* as well as the *reverse Pythagorean* inequalities: With assumptions as in Lemma 2.1,

$$H(x) + D(x, y^*) \leq H_{\max} \text{ for all } x \in X\,, \qquad (7)$$

$$H_{\max} + D(x^*, y) \leq R(y) \text{ for all } y \in Y\,. \qquad (8)$$

Here, $R(y) = \sup \Phi^y$, the supremum of the marginal function Φ^y, conceived as a *risk* for the minimizer in applying the strategy y. The Pythagorean inequality (and identity) goes back to Čencov [3] and to Csiszár [4]. Details about the game indicated may be found in [18] and in [9].

III. An axiomatic approach

Given are two *strategy sets*, X and Y, a map $x \mapsto \hat{x}$ between them, the *response*, and a triple of functions (Φ, H, D), the *information triple*, with Φ and D defined on $X \times Y$ and H defined on X. Terminology is much as in Sectiuon II, thus the functions are termed, respectively, *complexity, entropy* and *redundancy* and when $y = \hat{x}$, we say that y is *adapted to x* or that x *matches* y.

Though we do not really need special assumptions for the first results on robustness, we may as well single out right from the start the two axioms which to our mind point to the most useful information triples:

Axiom 1 (linking, fundamental inequality): The relations (5) and (6) hold and, for every $(x,y) \in X \times Y$, $D(x,y) = 0 \Leftrightarrow y = \hat{x}$.

Axiom 2 (marginal affinity):[1] The strategy set X is convex and Φ is affine in its first variable, i.e. the marginal function Φ^y is affine for every $y \in Y$.

It is convenient to introduce the set $MOL(X)$ of *molecular measures over* X. These are probability distributions with finite support over X. We often denote such distributions by the letter α and then we denote by $b(\alpha)$ the corresponding *barycentre*: $b(\alpha) = \sum_{x\in X} \alpha_x x$.

From the axioms one can derive important identities and inequalities related to convexity- and concavity properties. The most important one among these is the *compensation identity* [2] which states that, for each $\alpha \in MOL(X)$ and each $y \in Y$,

$$\sum_{x\in X} \alpha_x D(x,y) = \sum_{x\in X} \alpha_x D(x, \widehat{b(\alpha)}) + D(b(\alpha), y)\,. \qquad (9)$$

A simple proof is obtained by copying the argument used in the proof of Theorem 6.1 of [19].

The two-person zero-sum game γ_ϕ is the game with Φ as objective function, *Player I* (a maximizer) choosing strategies from X and *Player II* (a minimizer) choosing strategies from Y. The *Player-I value* at x, $H(x) = \inf \Phi_x$, is the *entropy of* x, and the *Player-II value at* $y \in Y$, $R(y) = \sup \Phi^y$ the *risk at* y. The *global values* of the game are the MaxEnt and the MinRisk values: $H_{\max} = \sup_{x\in X} H(x)$ and $R_{\min} = \inf_{y\in Y} R(y)$. Notions of optimal strategies are defined in the natural way as strategies (x^*, y^*) with $H(x^*) = H_{\max}$ and $R(y^*) = R_{\min}$. The game is in *(game theoretical) equilibrium* if $H_{\max} = R_{\min}$ and this number is finite. If this is the case and x^* and y^* are optimal strategies – an ideal situation from a game theoretical point of view – we express this by writing $\gamma_\phi \in GTE(x^*, y^*)$.

Parallel to the setting in Section II we introduce the notion of *robustness* and one may prove a general version of the robustness lemma, including (7) (with equality) and (8). More general, but equally simple to prove is the following result:

[1] for some purposes, concavity instead of affinity will do, but this mainly concerns non-constructive existence theorems which we shall not enter into here.

[2] Authors terminology, I guess not standard. In quantum information theory known as *Donalds identity*, cf. [6]. Possibly, it first appeared in [17].

Lemma 3.1: Let y^* be a Player-II strategy and assume that the marginal function Φ^{y^*} assumes a finite maximal value at a point x^* which matches y^*. Then $\gamma_\phi \in GTE(x^*, y^*)$ and (7) and (8) hold.

The game theoretical approach becomes much richer when several *preparations*, say belonging to a *preparation family* $\mathcal{X}$ of non-empty subsets of X are involved. The *subgames* $\gamma_\phi(X_0)$ with $X_0 \in \mathcal{X}$ are defined in the natural way by restricting Φ to $X_0 \times Y$.

The *exponential family* $\mathcal{E}_{II} = \mathcal{E}_{II}(\mathcal{X})$ associated with the family of games $\gamma_\phi(X_0)$, $X_0 \in \mathcal{X}$ is defined as the set of Player-II strategies which are robust for all games in the family. The corresponding exponential family in the Player-I domain is denoted $\mathcal{E}_I$ and consists of all Player-I strategies which match a strategy in $\mathcal{E}_{II}$. Sometimes it is convenient to work with $\mathcal{E}_{II}$, sometimes with $\mathcal{E}_I$. Let us use $\mathcal{E}$ to stand for either of the families – which one will follow from the context.

We call $(\mathcal{E}, \mathcal{X})$ *complete* if $\mathcal{E}$ is the disjoint union of the intersections $\mathcal{E} \cap X_0$ with X_0 ranging over $\mathcal{X}$[3]. The robustness lemma in a setting with completeness is the following one, considered a simple but important example of the paradigm of intrinticity:

Theorem 3.1: Assume that $(\mathcal{E}, \mathcal{X})$ is a complete pair in the sense just explained. Then, for a preparation $X_0 \in \mathcal{X}$, $X_0 \in GTE(x^*, y^*)$ with y^* adapted to that strategy x^* which is the unique point of intersection of $\mathcal{E}$ and X_0. Furthermore, (7) (with equality) and (8) hold.

Other useful notions and results apply in relation to the above concepts (e.g. regarding maximal preparations). This is presently being investigated (and confrontation with the huge literature on exponential families in statistics sought).

Regarding axioms, topological considerations may be added in order to allow one to prove existence theorems in cases when simple optimal strategies do not exist. These results, however, are non-constructive and rely on relatively advanced methods and are, therefore, considered to be outside the scope of the present contribution. Results in the direction indicated – with implications such as separation- and duality theorems – will be published separately. Ideas go back to [18], see also more recent expressions of the ideas in [5].

IV. Generation of information triples

The success of the intrinsic methods pointed out so far for the identification of optimal strategies rests on the possibilities for the creation of interesting information triples satisfying Axioms 1 and 2.

We shall point to three methods: *integration of atomic triples*, *relativization* and *randomization*. When below we talk about information triples, it is understood that they satisfy Axioms 1 and 2.

[3]more general settings (sacrificing disjointness) are needed to cover situations with possible *loss of entropy*, a phenomenon first observed in [10], see also [18] and [9]

A. Atomic Triples, Integration

We consider an interval $I \subseteq]-\infty, \infty]$ and seek information triples (ϕ, h, d) with $X = Y = I$ and with the identity as response. It turns out that, suppressing regularity conditions, the only possibilities are to generate such triples in a way closely modelled after an approach by Bregman, cf. [1]. The construction is based on a strictly concave function h and this is used to define the complexity and the redundancy (or *divergence*) functions by the formulas:

$$\phi(x, y) = h(y) + (x - y)h'(y)\,, \tag{10}$$

$$d(x, y) = h(y) - h(x) + (x - y)h'(y)\,. \tag{11}$$

Triples constructed in this way are considered the simplest ones and are said to be *atomic*. The example with

$$h(x) = x \ln \frac{1}{x}\,, \tag{12}$$

say on $I = [0, 1]$, is of central importance for information theory. An example with a geometric flavour is given by $h(x) = -(x - y_0)^2$ on $]-\infty, \infty[$.

By a natural process of integration, one may generate more complex information triples from atomic triples. Thus from (12) we may generate the triple (Φ, H, D) of Section II which is of course a key example for information theory. But the freedom of choice of h opens up for a wide range of further possibilities. Within the field of non-extensive statistical physics, popular variants of Shannon (-Boltzmann-Gibbs) quantities can be discussed and then treated by intrinsic methods, cf. [20], [13].

Similarly, from the other example above we may generate an information triple on Hilbert space given by

$$\Phi(x, y) = \|x - y\|^2 - \|x - y_0\|^2\,, \tag{13}$$

$$H(x) = -\|x - y_0\|^2\,, \tag{14}$$

$$D(x, y) = \|x - y\|^2\,. \tag{15}$$

B. Relativization

Consider an information triple (Φ, H, D) and let $y_0 \in Y$. Then a new information triple may be obtained by looking at the functions "relative" to y_0, considered as a *prior*. In this way we obtain a new information triple $(\tilde{\Phi}, \tilde{H}, \tilde{D})$ given by

$$\tilde{\Phi}(x, y) = D(x, y) - D(x, y_0) \tag{16}$$

$$\tilde{H}(x) = -D(x, y_0) \tag{17}$$

$$\tilde{D}(x, y) = D(x, y)\,. \tag{18}$$

For this triple, $\tilde{\Phi}(x, y) = D(x, y) - D(x, y_0)$ is considered as the negative of an *updating gain* and we realize that, in case the original triple is the one from Section II, we are led to consider the familiar *minimum information discrimination principle* which originated with Kullback, [12]. See also [8] where the game theoretical point of view is a guiding principle.

It is of interest to note that the triple $(\tilde{\Phi}, \tilde{H}, \tilde{D})$ only requires that the function D is given beforehand and that, modulo regularity conditions, the necessary and sufficient conditions that (16)-(18) defines an information triple is that D satisfies the compensation identity.

C. randomization

A given information triple with strategy sets X and Y and response $x \mapsto \hat{x}$ generate in a natural way a new information triple for which the strategy sets are $MOL(X)$, Y and the response is $\alpha \mapsto b(\alpha)$. Two natural constructions may be considered. We only mention what appears as the most useful construction. It is given by the formulas

$$\tilde{\Phi}_0(\alpha, y) = \sum_{x \in X} \alpha_x D(x, y)\,, \tag{19}$$

$$\tilde{H}_0(\alpha) = \sum_{x \in X} \alpha_x D(x, \widehat{b(\alpha)})\,, \tag{20}$$

$$\tilde{D}_0(\alpha, y) = D(b(\alpha), y)\,. \tag{21}$$

Without going into details let us mention that this construction, when specialized to the information theoretical quantities lies behind the *redundance-capacity theorem*, cf. [7] and [14]. Further, the usual *Kuhn-Tucker conditions* of importance in this connection really come under the heading of what we refer to as intrinsic methods. In fact they can be derived easily, either directly from the compensation identity or from Lemma 3.1.

Another comment is worth while making: When we specialize to the triple given by (13)-(15), we are led to consider a problem from location theory, the *Sylvester problem*: "*It is required to find the least circle which shall contain a given system of points in the plane*", cf. [16] – in fact this is the full text of [16]!

V. Conclusions

By axiomatizing basic properties which are well known from information theory, we demonstrated that a theory with a wide range of possibilities emerges. Especially the handling of optimization problems is facilitated, both within and outside information theory. The focus has been on the possibility to use simple intrinsic methods adapted to the optimization problems at hand. The thesis has been defended that for natural problems as those which arise from the theory initiated, it is possible to device such methods, rather than turning to an automatic application of standard analytical tools based on the differential calculus.

References

[1] L. M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Comput. Math. and Math. Phys.*, 7:200–217, 1967. Translated from Russian.

[2] J. Brinkhuis and V. Tikhomirov. *Optimization: Insights and Applications*. Princeton University Press, 2005.

[3] N. N. Čencov. A nonsymmetric distance between probability distributions, entropy and thePythagorean theorem. *Math. Zametki*, 4:323–332, 1968. (in Russian).

[4] I. Csiszár. I-divergence geometry of probability distributions and minimization problems. *Ann. Probab.*, 3:146–158, 1975.

[5] I. Csiszár. Generalized projections for non-negative functions. *Acta Math. Hungar.*, 68:161–185, 1995.

[6] M. J. Donald. On the relative entropy. *Commun. Math. Phys.*, 105:13–34, 1985.

[7] R. G. Gallager. *Source coding with side information and universal coding*. unpublished manuscript, 1977.

[8] P. D. Grünwald and A. P. Dawid. Game theory, maximum entropy, minimum discrepancy, and robust Bayesian decision theory. *Annals of Mathematical Statistics*, 32(4):1367–1433, 2004.

[9] Peter Harremoës and Flemming Topsøe. Maximum entropy fundamentals. *Entropy*, 3(3):191–226, Sept. 2001.

[10] R. S. Ingarden and K. Urbanik. Quantum informational thermodynamics. *Acta Physica Polonica*, 21:281–304, 1962.

[11] J. N. Kapur. *Maximum Entropy Models in Science and Engineering*. Wiley, New York, 1993. first edition 1989.

[12] S. Kullback. *Information Theory and Statistics*. Wiley, New York, 1959.

[13] J. Naudts. Generalized exponential families and associated entropy functions. arXiv:cond-mat/0803.0104, 2008.

[14] B. Ya. Ryabko. Comments on "a source matching approach to finding minimax codes". *IEEE Trans. Inform. Theory*, 27:780–781, 1981. Including also the ensuing Editor's Note.

[15] C. E. Shannon. A mathematical theory of communication. *Bell Syst. Tech. J.*, 27:379–423 and 623–656, 1948.

[16] J. J. Sylvester. A question in the geometry of situation. *Quarterly Journal of Pure and Applied Mathematics*, 1:79, 1857.

[17] F. Topsøe. An information theoretical identity and a problem involving capacity. *Studia Scientiarum Mathematicarum Hungarica*, 2:291–292, 1967.

[18] F. Topsøe. Information theoretical optimization techniques. *Kybernetika*, 15(1):8 – 27, 1979.

[19] F. Topsøe. Basic concepts, identities and inequalities – the toolkit of informationtheory. *Entropy*, 3:162–190, 2001. http://www.unibas.ch/mdpi/entropy/ [ONLINE].

[20] F. Topsøe. Exponential Families and MaxEnt Calculations for Entropy Measures of Statistical Physics. In Qurati Rapisarda Tsallis Abe, Hermann, editor, *Complexity, Metastability, and Non-Extensivity, CTNEXT07*, volume 965 of *AIP Conference Proceedings*, pages 104–113, 2007.