Towards Optimal Techniques for Solving Global Optimization Problems: Symmetry-Based Approach

Christodoulos A. Floudas[1] and Vladik Kreinovich[2]

[1] Department of Chemical Engineering, Princeton University, Princeton, NJ 08544, USA floudas@titan.princeton.edu
[2] Department of Computer Science, University of Texas at El Paso, El Paso, TX 79968, USA vladik@utep.edu

1 Introduction

1.1 Global Optimization – an Important Practical Problem

In many practical situations, we have several possible actions, and we must choose the best action. For example, we must find the best design of an object, or the best control of a plant. The set of possible actions is usually characterized by parameters $x = (x_1, \ldots, x_n)$, and the result of different actions (controls) is characterized by an *objective function* $f(x)$.

In some cases, the objective function describes losses or expenses; in such cases, the problem of finding the best action (design, or control) can be described as the problem of *global minimization*, i.e., the problem of finding the values x for which the function $f(x)$ attains the smallest possible value.

In other cases, the objective function describes gain; in such cases, the problem of finding the best action can be described as the problem of *global maximization*, i.e., the problem of finding the values x for which the function $f(x)$ attains the largest possible value.

Global minimization and global maximization are particular cases of *global optimization*.

Similar problems arise in *data processing*, when we have a model characterized by several parameters x_i, and we need to find the values of these parameters which provide the best fit for the data, i.e., for which the discrepancy $f(x)$ between the data and the model is the smallest possible.

Actual and potential real-world applications of global optimization are overviewed, e.g., in [Pin96].

1.2 Global Optimization is a Difficult Computational Problem

In general, the problem of finding the exact values x that minimize a given objective function $f(x)$ is computationally difficult (NP-hard); see, e.g., [Vav91].

Crudely speaking, NP-hardness means that (provided that P$\neq$NP) it is not possible to have an algorithm that solves *all* optimization problems in reasonable time. In other words, no matter how good is an algorithm for solving global optimization optimization problems, there will always be cases in which better results are possible.

1.3 Variety of Global Optimization Techniques

Since we cannot hope for a single algorithm for global optimization, new algorithms are constantly designed, and the existing algorithms are constantly modified. As a result, we have a wide variety of different global optimization techniques and methods; see, e.g., [HP95].

There exist classes of objective functions for which efficient algorithms for global optimization are possible. It is therefore natural to try to reduce general hard-to-solve global optimization problems to problems from such classes.

One class for which global optimization is easier-to-solve is the class of quadratic objective functions. Namely, it is known that a global optimum of an objective function $f(x)$ is attained at a point x at which all the partial derivatives of this function are equal to 0. For a quadratic function $f(x)$, we can thus find the desired optimum by solving a system of linear equations $\frac{\partial f}{\partial x_i} = 0$. It is therefore natural to find a minimum of $f(x)$ by approximating a function $f(x)$ with a linear or quadratic expression – i.e., in effect, by consider gradient descent-type techniques and/or their second-order analogues.

Another important class is the class of convex functions – for which there are efficient algorithms for finding the global minimum. Not surprisingly, there are numerous effective global optimization techniques that reduce the general global optimization problems to convex ones; see, e.g., [Flo00, TS02].

In many real-life situations, the objective function is complex, and it is difficult to approximate it by a quadratic and/or by a convex objective function on its entire domain. In such situations, it is reasonable to subdivide the original domain into smaller subdomains and approximate $f(x)$ by different functions on different subdomains; see, e.g., [KK96].

There also exist numerous heuristic and semi-heuristic techniques which emulate the way optimization is done in nature: e.g., genetic algorithms simulate the biological evolution which, in general, leads to the birth and survival individuals and species which are best fit for a given environment; see, e.g., [Mih96].

1.4 Problem: Which Techniques is the Best?

We have already mentioned that there is a wide variety of different global optimization techniques. Because of this variety, every time we have a new optimization problem, we must select the best technique for solving this problem.

This selection problem is made even more complex by the fact that most techniques for solving global optimization problems have *parameters* that need to be adjusted to the problem or to the class of problems. For example, in gradient methods, we can select different step sizes.

When we have a *single* parameter (or few parameters) to choose, it is possible to empirically try many values and come up with an (almost) optimal value. Thus, in such situations, we can come up with optimal version of the corresponding technique.

In other approaches, e.g., in methods like convex underestimators (described in detail in the next section), instead of selecting the value of single *number*-valued parameter, we have select the auxiliary *function*. It is not practically possible to test all possible functions, so it is not easy to come up with an optimal version of the corresponding technique.

1.5 What We Do in This Chapter

In this chapter, we consider the problem of selecting the best auxiliary function within a given global optimization technique. Specifically, we show that in many such selection situations, natural symmetry requirements enable us either to analytically solve the problem of finding the optimal auxiliary function, or at least reduce this problem to the easier-to-solve problem of finding a few parameters.

In particular, for convex understimators, we show that we can thus explain both the αBB method [AAF98, ADA98, Flo00, MF94] and its modifications recently proposed in [AF04, AF06].

2 Case Study: Selecting Convex Underestimators

2.1 Why Convex Underestimators?

It is well known that convex functions are computationally easier to minimize than non-convex ones; see, e.g., [Flo00]. This relative easiness is not only an empirical fact, it also has a theoretical justification; see, e.g., [KK05, Vav91].

Because of this relative easiness, one of the approaches to minimization of a non-convex function $f(x) = f(x_1, \ldots, x_n)$ (under certain constraints) over a box $[x^L, x^U] = [x_1^L, x_1^U] \times \ldots \times [x_n^L, x_n^U]$ is to first minimize its convex "underestimator", i.e., a convex function $L(x) \le f(x)$.

- Since the new function $L(x)$ is convex, it is easy to minimize;

- since $L(x)$ is an underestimator, i.e., $L(x) \leq f(x)$, the minimum of $L(x)$ is a lower bound for the minimum of $f(x)$.

By selecting $L(x)$ as close to $f(x)$ as possible, we can get estimates for $\min f(x)$ which are as close to the actual minimum as possible.

The quality of approximation improves when the boxes become smaller. So, to get more accurate bounds on $\min f(x)$, we can:

- bisect the box $[x^L, x^U]$ into sub-boxes,
- use the above technique to estimate $\min f$ over each sub-box, and
- return the smallest of these estimates as the lower bound for $\min f$ over the entire box $[x^L, x^U]$.

2.2 Example: αBB Techniques

A known efficient approach to designing a convex underestimator is the αBB global optimization algorithm [AAF98, ADA98, Flo00, MF94], in which we select an underestimator $L(x) = f(x) + \Phi(x)$, where

$$\Phi(x) = -\sum_{i=1}^{n} \alpha_i \cdot (x_i - x_i^L) \cdot (x_i^U - x_i). \tag{1}$$

Here, the parameters α_i are selected in such a way that the resulting function $L(x)$ is convex and still not too far away from the original objective function $f(x)$.

2.3 Natural Generalization of αBB Techniques

In many optimization problems, αBB techniques are very efficient, but in some non-convex optimization problems, it is desirable to improve their performance. One way to do that is to provide a more general class of methods, with more parameters to tune.

In the αBB techniques, for each coordinate x_i, we have a single parameter α_i affecting this coordinate. Changing α_i is equivalent to a linear re-scaling of x_i. Indeed, if we change the unit for measuring x_i to a new unit which is λ_i times smaller, then all the numerical values become λ_i times larger: $x_i \to y_i = g_i(x_i)$, where $g_i(x_i) = \lambda_i \cdot x_i$. In principle, we can have two different re-scalings:

- $x_i \to y_i = g_i(x_i) = \lambda_i \cdot x_i$ on the interval $[x_i^L, x_i]$, and
- $x_i \to z_i = h_i(x_i) = \mu_i \cdot x_i$ on the interval $[x_i, x_i^U]$.

If we substitute the new values $y_i = g_i(x_i)$ and $z_i = h_i(x_i)$ into the formula (1), then we get the following expression

$$\Phi(x) = -\sum_{i=1}^{n} \alpha_i \cdot (g_i(x_i) - g_i(x_i^L)) \cdot (h_i(x_i^U) - h_i(x_i)). \tag{2}$$

For the above linear re-scalings, we get

$$\widetilde{\Phi}(x) = -\sum_{i=1}^{n} \widetilde{\alpha}_i \cdot (x_i - x_i^L) \cdot (x_i^U - x_i),$$

where $\widetilde{\alpha}_i = \alpha_i \cdot \lambda_i \cdot \mu_i$.

From this viewpoint, a natural generalization is to replace *linear* re-scalings $g_i(x_i)$ and $h_i(x_i)$ with *non-linear* ones, i.e., to consider convex underestimators of the type $L(x) = f(x) + \Phi(x)$, where $\Phi(x)$ is described by the formula (2) with non-linear functions $g_i(x_i)$ and $h_i(x_i)$. Now, instead of selecting a number α_i for each coordinate i, we have an additional freedom of choosing arbitrary non-linear functions $g_i(x_i)$ and $h_i(x_i)$. Which are the best choices?

2.4 Empirical Fact: Exponential Functions $g_i(x_i)$ and $h_i(x_i)$ Are the Best

In [AF04, AF06], several different non-linear functions have been tried, and it turned out that among the tested functions, the best results were achieved for the exponential functions $g_i(x_i) = \exp(\gamma_i \cdot x_i)$ and $h_i(x_i) = -\exp(-\gamma_i \cdot x_i)$. For these functions, the expression (2) can be somewhat simplified: indeed,

$$\alpha_i \cdot (g_i(x_i) - g_i(x_i^L)) \cdot (h_i(x_i^U) - h_i(x_i)) = \alpha_i \cdot (e^{\gamma_i \cdot x_i} - e^{\gamma_i \cdot x_i^L}) \cdot (-e^{-\gamma_i \cdot x_i^U} + e^{-\gamma_i \cdot x_i}) =$$

$$\widetilde{\alpha}_i \cdot (1 - e^{\gamma_i \cdot (x_i - x_i^L)}) \cdot (1 - e^{\gamma_i \cdot (x_i^U - x_i)}),$$

where $\widetilde{\alpha}_i \stackrel{\text{def}}{=} \alpha_i \cdot e^{\gamma_i \cdot (x_i^U - x_i^L)}$.

2.5 Questions

Two related questions naturally arise:

- first, a *practical* question: an empirical choice is made by using only finitely many functions; is this choice indeed the best – or there are other, even better functions $g_i(x_i)$ and $h_i(x_i)$, which we did not discover because we did not try them?
- second, a *theoretical* question: how can we explain the above empricial fact?

2.6 Natural Idea of Symmetry: Intuitive Motivations for Shift-Invariance

The starting (0) point for measuring each coordinate x_i is often a matter of arbitrary choice; e.g.:

- Fahrenheit and Celsius scales use different starting points for measuring temperature,

- different calendars use different starting points as Year 0,

etc.

If a selection of the functions $g_i(x_i)$ and $h_i(x_i)$ is "optimal" (in some intuitive sense), then the results of using these optimal functions should not change if we simply change the starting point for measuring x_i – i.e., replace each value x_i with a new value $x_i + s$, where s is the shift in the starting point. Indeed, otherwise, if the "quality" of the resulting convex underestimators changes with shift, we could apply a shift and get better functions $g_i(x_i)$ and $h_i(x_i)$ – which contradicts to our assumption that the selection of $g_i(x_i)$ and $h_i(x_i)$ is already optimal.

So, the "optimal" choices $g_i(x_i)$ and $g_i(x_i)$ can be determined from the requirement that each component $\alpha_i \cdot (g_i(x_i) - g_i(x_i^L)) \cdot (h_i(x_i^U) - h_i(x_i))$ in the sum (2) be invariant under the corresponding shift. Let us describe this requirement in precise terms.

Definition 1. *A pair of smooth functions $(g(x), h(x))$ from real numbers to real numbers is* shift-invariant *if for every s and α, there exists $\widetilde{\alpha}(\alpha, s)$ such that for every x^L, x, and x^U, we have*

$$\alpha \cdot (g(x) - g(x^L)) \cdot (h(x^U) - h(x)) =$$

$$\widetilde{\alpha}(\alpha, s) \cdot (g(x+s) - g(x^L + s)) \cdot (h(x^U + s) - h(x+s)) \tag{3}$$

Comment. Smoothness is needed because smooth functions are easier to optimize, and we therefore want our techniques to preserve smoothness.

2.7 Consequences of Shift-Invariance

At first glance, shift invariance is a reasonable but weak property. It turns out, however, that this seemingly weak property actually almost uniquely determines the optimal selection of exponential functions:

Proposition 1. *If a pair of functions $(g(x), h(x))$ is shift-invariant, then this pair is either exponential or linear, i.e., each of the functions $g(x)$ and $h(x)$ has the form $g(x) = A + C \cdot \exp(\gamma \cdot x)$ or $g(x) = A + k \cdot x$.*

Comments.

- For reader's convenience, all the proofs are placed in a separate (last) section.
- One can easily see that adding a constant to each of the functions $g(x)$ and $h(x)$ does not change the expression (2), so we can safely assume that each of these functions has the form $g(x) = \exp(\gamma \cdot x)$ and $h(x) = x$.

2.8 Additional Symmetry $x \to -x$ and the Final Result

In addition to shift, another natural symmetry is changing the sign: e.g., for electric charge, the fact that electrons are negatively charged is just a matter of definition; we can as well consider them positively charged. If we require that the expression (2) remain invariant if we change the sign, i.e., replace x by $-x$, then we get the relation between $g(x)$ and $h(x)$: $h(x) = -g(-x)$. So, if a pair $(g(x), h(x)$ is shift-invariant and sign-invariant, then:

- either $g(x) = \exp(\gamma \cdot x)$ and $h(x) = -\exp(-\gamma \cdot x)$,
- or $g(x) = h(x) = x$.

In other words, *the optimal generalized αBB scheme is either the original αBB, or the scheme with exponential functions* described in [AF04, AF06]. Thus, we have answers to both above questions:

- yes, the exponential functions are indeed optimal, and
- yes, we have a theoretical explanation of why they are optimal – because they are the only pair of functions which satisfies the condition of symmetry (shift-invariance and sign-invariance) that optimal pairs should satisfy.

2.9 Auxiliary Result: Scale Invariance

In addition to changing the starting point for x, we can also (as we have mentioned) change a unit for measuring x, i.e., consider *scaling* transformations $x \to \lambda \cdot x$. Shall we require scale-invariance as well? In other words, shall we require that the expression (2) be invariant not only w.r.t. shifts but w.r.t scalings as well?

We already know that there are only two shift-invariant solutions: exponential and linear functions. Out of these two solutions, only the linear solution – corresponding to αBB – is scale-invariant. Thus, if we also require scale-invariance, we restrict ourselves only to αBB techniques – and miss on (often better) exponential generalizations.

Since we cannot require *both* shift- and scale-invariance, a natural next question is: what if we only require scale invariance?

Definition 2. *A pair of smooth functions $(g(x), h(x))$ from real numbers to real numbers is* scale-invariant *if for every λ and α, there exists $\widetilde{\alpha}(\alpha, \lambda)$ such that for every x^L, x, and x^U, we have*

$$\alpha \cdot (g(x) - g(x^L)) \cdot (h(x^U) - h(x)) =$$

$$\widetilde{\alpha}(\alpha, \lambda) \cdot (g(\lambda \cdot x) - g(\lambda \cdot x^L)) \cdot (h(\lambda \cdot x^U) - h(\lambda \cdot x)) \tag{4}$$

Proposition 2. *If a pair of functions $(g(x), h(x))$ is scale-invariant, then this pair is either exponential or linear, i.e., each of the functions $g(x)$ and $h(x)$ has the form $g(x) = A \cdot x^\gamma$ or $g(x) = A + k \cdot \ln(x)$.*

From the theoretical viewpoint, these functions may look as good as the exponential functions coming from shift invariance, and in practice, they do not work so well.

The problem with these solutions is that, as we have mentioned, we want to preserve smoothness. Both linear and exponential functions which come from shift-invariance are infinitely differentiable for all x and hence, adding the corresponding term $\Phi(x)$ will not decrease the smoothness level of the objective function.

In contrast, in general, the functions $g(x) = x^\gamma$ which come from scale invariance are not infinitely differentiable at $x = 0$. They are differentiable only for integer values γ. So, if we use scale invariance to select a convex underestimator, we end up with a new parameter γ which only attains integer-valued values and is, thus, less flexible than the continuous-valued parameters coming from scale-invariance.

2.10 Auxiliary Shift-Invariance Results

Instead of an expression (2), we can consider an even more general expression

$$\Phi(x) = -\sum_{i=1}^{n} \alpha_i \cdot a_i(a, x^L) \cdot b_i(x_i, x_i^U). \tag{5}$$

Whet can we conclude from shift-invariance in this more general case?

Definition 3. *A pair of smooth functions $(a(x, x^L), b(x, x^U))$ from real numbers to real numbers is* shift-invariant *if for every s and α, there exists $\widetilde{\alpha}(\alpha, s)$ such that for every x^L, x, and x^U, we have*

$$\alpha \cdot a(x, x^L) \cdot b(x, x^U) =$$

$$\widetilde{\alpha}(\alpha, s) \cdot a(x + s, x^L + s) \cdot b(x + s, x^U + s). \tag{6}$$

Proposition 3. *If a pair of functions $(a(x, x^L), b(x, x^U))$ is shift-invariant, then*

$$a(x, x^L) \cdot b(x, x^U) = A(x - x^L) \cdot B(x^U - x) \cdot e^{\gamma \cdot x^L}$$

for some functions $A(x)$ and $B(x)$ and for some real number γ.

Comment. If we additionally require that the expression $a(x, x^L) \cdot b(x, x^U)$ be invariant under $x \to -x$, then we conclude that $B(x) = A(x)$.

Another shift-invariance result comes from the following observation. Both αBB expression $-(x - x^L) \cdot (x^U - x)$ and the generalized expression

$$-(1 - e^{\gamma \cdot (x - x^L)}) \cdot (1 - e^{\gamma \cdot (x^U - x)})$$

have the form $a(x - x^L) \cdot a(x^U - x)$ with $a(0) = 0$. The differences $x - x^L$ and $x^U - x$ come from the fact that we want these expressions to be shift-invariant.

The product form makes sense, since we want the product to be 0 on each border $x = x^L$ and $x = x^U$ of the corresponding interval $[x^L, x^U]$.

On the other hand, it is well known that optimizing a product is more difficult than optimizing a sum; since we will be minimizing the expression $f(x) + \Phi(x)$, it is therefore desirable to be able to reformulate it in terms of the easier-to-minimize sum, e.g., as $b(x - x^L) + b(x^U - x) + c(x^U - x^L)$ for some functions b and c (for minimization purposes, c does not depend on x and is thus a constant). It is worth mentioning that both the αBB expression and its exponential generalization allow such representation:

- from the known equality $a \cdot b = \frac{1}{2}((a+b)^2 - a^2 - b^2)$, we conclude that

$$-(x - x^L) \cdot (x^U - x) = \frac{1}{2} \cdot (x - x^L)^2 + \frac{1}{2} \cdot (x^U - x)^2 - \frac{1}{2} \cdot (x^U - x^L)^2;$$

- for the exponential function, simply multiplying the two sums leads to the desired expression:

$$-(1 - e^{\gamma \cdot (x - x^L)}) \cdot (1 - e^{\gamma \cdot (x^U - x)}) = -1 + e^{\gamma \cdot (x - x^L)} + e^{\gamma \cdot (x^U - x)} - e^{\gamma \cdot (x^U - x^L)}.$$

Interestingly, the above two expressions are the only one which have this easiness-to-compute property:

Definition 4. *We say that a smooth function $a(x)$ from real numbers to real numbers describes an* easy-to-compute *underestimator if $a(0) = 0$, $a'(0) \neq 0$, and there exist smooth functions $b(x)$ and $c(x)$ such that for every x, x^L, and x^U, we have*

$$a(x - x^L) \cdot a(x^U - x) = b(x - x^L) + b(x^U - x) + c(x^U - x^L). \tag{7}$$

Comment. The condition $a'(0) \neq 0$ comes from the fact that otherwise, for small $\Delta x \stackrel{\text{def}}{=} x - x^L$ and $x^U - x$, each value $a(x - x^L)$ will be quadratic in $x - x^L$, the resulting product will be fourth order, and we will not be able to compensate for quadratic non-convex terms in the original objective function $f(x)$ – which defeats the purpose of using $f(x) + \Phi(x)$ as a *convex* underestimator.

Proposition 4. *The only functions which describe easy-to-compute underestimators are $a(x) = k \cdot x$ and $a(x) = k \cdot (1 - e^{\gamma \cdot x})$.*

Comment. This is already a second shift-invariance related results which selects linear and exponential functions as "the best" in some reasonable sense. In the following section, we show that this is not an accident: namely, we will prove that *any* "natural" shift-invariant optimality cruetrion on the set of all possible underestimator methods selects either a linear or an exponential function.

3 Selecting Convex Underestimatiors: From Informally "Optimal" to Formally Optimal Selections

3.1 In the Previous Section, We Used Informal "Optimality"

In the above text, we argued that if a selection is optimal (in some reasonable sense), than it is natural to expect that this selection should be shift-invariant. We used this argument to justify the empirical selection of convex underestimators.

In this section, we will go one step further, and explain that the empirical selection is indeed optimal – in the precise mathematical sense of this word.

3.2 What Are We Selecting?

In effect, we are selecting the functions $g(x)$ and $h(x)$. However, as we have mentioned earlier, what we are really interested in is the corresponding family of functions

$$\Phi(x) = -\alpha \cdot (g(x) - g(x^L)) \cdot (h(x^U) - h(x)).$$

The difference is that (as we have mentioned) we can change one (or both) of the functions $g(x)$ and $h(x)$ and still end up with the same class of functions. For example, if we replace the original function $g(x)$ with a new function $\widetilde{g}(x) = A \cdot g(x) + B$, then we end up with the same class of functions $\Phi(x)$. With this in mind, let us introduce the following definition.

Definition 5. *By a* family*, we mean the family of functions*

$$F = \{-\alpha \cdot (g(x) - g(x^L)) \cdot (h(x^U) - h(x))\}, \tag{8}$$

where $g(x)$ and $h(x)$ are fixed, and α goes over all real numbers.

Denotation. We will denote a family generated by functions $g(x)$ and $h(x)$ by $F(g, h)$.

In these terms, the question is how to select, out of all possible families, the family which is optimal in some reasonable sense, i.e., which is optimal in the sense of some optimality criterion.

3.3 What is an Optimality Criterion?

When we say that some *optimality criterion* is given, we mean that, given two different families F and F', we can decide whether the first or the second one is better, or whether these families are equivalent w.r.t. the given criterion. In mathematical terms, this means that we have a *pre-ordering relation* $\preceq$ on the set of all possible families.

3.4 We Want to Solve an Ambitious Problem: Enumerate All Families that are Optimal Relative to Some Natural Criteria

One way to approach the problem of choosing the "best" family F is to select *one* optimality criterion, and to find a family that is the best with respect to this criterion. The main drawback of this approach is that there can be different optimality criteria, and they can lead to different optimal solutions. It is, therefore, desirable not only to describe a family that is optimal relative to some criterion, but to describe *all* families that can be optimal relative to different natural criteria[3]. In this section, we are planning to implement exactly this more ambitious task.

3.5 Examples of Optimality Criteria

Pre-ordering is the general formulation of optimization problems in general, not only of the problem of choosing a family F. In general optimization theory, in which we are comparing arbitrary *alternatives* a', a'', ..., from a given set A, the most frequent case of such a pre-ordering is when a *numerical criterion* is used, i.e., when a function $J : A \to R$ is given for which $a' \preceq a''$ iff $J(a') \leq J(a'')$.

Several natural numerical criteria can be proposed for choosing a function J. For example, we can take, as a criterion, the *average* number of iterations that lead to determining all global minima with a given relative accuracy (average in the sense of some natural probability measure on the set of all problems).

Alternatively, we can fix a class of problems, and take the largest number of iterations for problems of this class as the desired (numerical) optimality criterion.

Many other criteria of this type can be (and have actually been) proposed. For such "worst-case" optimality criteria, it often happens that there are several different alternatives that perform equally well in the worst case, but whose performance differ drastically in the average cases. In this case, it makes sense, among all the alternatives with the optimal worst-case behavior, to choose the one for which the average behavior is the best possible. This very natural idea leads to the optimality criterion that is not described by one numerical optimality criterion $J(a)$: in this case, we need *two* functions: $J_1(a)$ describes the worst-case behavior, $J_2(a)$ describes the average-case behavior, and $a \preceq b$ iff either $J_1(a) < J_1(b)$, or $J_1(a) = J_1(b)$ and $J_2(a) \leq J_2(b)$.

We could further specify the described optimality criterion and end up with *one* natural criterion. However, as we have already mentioned, the goal

[3] In this phrase, the word "natural" is used informally. We basically want to say that from the purely mathematical viewpoint, there can be weird ("unnatural") optimality criteria. In our text, we will only consider criteria that satisfy some requirements that we would, from the common sense viewpoint, consider reasonable and natural.

of this chapter is not to find *one* family that is optimal relative to some criterion, but to describe *all* families that are optimal relative to some natural optimality criteria. In view of this goal, in the following text, we will not specify the criterion, but, vice versa, we will describe a very general class of *natural* optimality criteria.

So, let us formulate what "natural" means.

3.6 What Optimality Criteria are Natural?

We have already mentioned that the value x often represents the value of some measured quantity, and that the numerical value of a measured quantity changes if we select a new starting point. It is natural to require that the relative quality of two families does not depend on the choice of the starting point.

How does replacing a starting point change the family F? If we replace a starting point by a new one that is smaller by a constant s, then the quantity that was initially described by a value x will be described by a new value $x+s$. Correspondingly, x^L is replaced by x^L+s, and x^U by x^U+s. Thus, after this shift T_s, the original family (8) turns into the new family

$$T_s(F) \stackrel{\text{def}}{=} \{-\alpha \cdot (g(x+s) - g(x^L+s)) \cdot (h(x^U+s) - h(x+s))\}. \qquad (9)$$

In these terms, the above requirement is that if F is better than F', then the "shifted" F (i.e., the family $T_s(F)$) should be better than the "shifted" F' (i.e., than $T_s(F')$).

There is one more reasonable requirement for a criterion, that is related with the following idea: If the criterion does not select a single optimal family, i.e., if it considers several different families equally good, then we can always use some other criterion to help select between these "equally good" ones, thus designing a two-step criterion. If this new criterion still does not select a unique family, we can continue this process until we arrive at a combination multi-step criterion for which there is only one optimal family. Therefore, we can always assume that our criterion is *final* in this sense.

Definition 6. *By an* optimality criterion, *we mean a* pre-ordering *(i.e., a transitive reflexive relation) $\preceq$ on the set A of all possible families. An optimality criterion $\preceq$ is called:*

- *shift-invariant if for all F, F', and s, $F \preceq F'$ implies $T_s(F) \preceq T_s(F')$.*
- *final if there exists one and only one family F that is preferable to all the others, i.e., for which $F' \preceq F$ for all $F' \neq F$.*

Proposition 5.

- *If a family F is optimal w.r.t. some shift-invariant final optimality criterion, then this family F is generated by linear or exponential functions $g(x)$ and $h(x)$.*

- *For every two exponential or linear functions $g(x)$ and $h(x)$, there exists a shift-invariant final optimality criterion for which the only optimal family is $F(g,h)$.*

Comments.

- In other words, if the optimality criterion satisfies the above-described natural properties, then *the optimal convex underestimator is generated by linear or exponential functions.*
- If, in addition to shift-invariance, we also require sign-invariance, then we conclude that either both functions $g(x)$ and $h(x)$ are linear (as in αBB), or both are exponential (as in the empirically best generalization of αBB).

4 Other Cases when a Symmetry-Based Approach Leads to Optimal Techniques for Solving Global Optimization Problems

Similar symmetry-based ideas have been applied to produce an optimal auxiliary function in other aspects of global optimization. Let us overview the main results obtained by following this direction.

4.1 Optimal Bisection

As we have mentioned, applying the optimization technique to the original function (or its convex underestimator) on the original box $[x^L, x^U]$ is not always the best strategy. One way to improve the optimization algorithm is to subdivide (e.g., bisect) the box into several sub-boxes and apply optimization techniques to these sub-boxes. Some of these sub-boxes must be further subdivided, etc. Two natural questions arise:

- which box should we select for bisection?
- which variable shall we use to bisect the selected box?

To answer both questions, several heuristic techniques have been proposed, and there has been an extensive empirical comparative analysis of these techniques. It turns out that for both questions, the symmetry-based approach enables us to theoretically justify the empirical selection:

- Until recently, for subdivision, a box B was selected for which the computed lower bound $\underline{f}(B)$ was the smallest possible. Recently (see, e.g, [CG98, CGC00]), it was shown that the optimization algorithms converge much faster if we select, instead, a box B with the largest possible value of the ratio

$$I_0 = \frac{\widetilde{f} - \underline{f}(B)}{\overline{f}(B) - \underline{f}(B)},$$

where $\widetilde{f}$ is a current upper bound on the actual global minimum. In [KC01], we give a symmetry-based theoretical justification for this empirical criterion. Namely, we condider all possible indictaor functions $I(\underline{f}(B), \overline{f}(B), \widetilde{f})$, and we show that:

- first, that the empirically best criterion I_0 is the only one that is *invariant* w.r.t. some reasonable symmetries – namely, shift and scaling; and
- second, that this criterion is *optimal* in some (symmetry-related) reasonable sense.

- We can bisect a given box in n different ways, depending on which of n sides we decided to halve. So, the natural question appears: which side should we cut? i.e., where to bisect a given box? Historically the first idea was to cut the *longest* side (for which $x_i^U - x_i^L \to \max$). It was shown (in [Rat92, Rat94]) that much better results are achieved if we choose a side i for which $|d_i|(x_i^U - x_i^L) \to \max$, where d_i is the known approximation for the partial derivative $\dfrac{\partial f}{\partial x_i}$. In [KK98], we consider arbitrary selection criteria, i.e., functions

$$S(f, d_1, \ldots, d_n, x_1^L, x_1^U, \ldots, x_n^L, x_n^U),$$

which map available information into an index $S \in \{1, 2, \ldots, n\}$, and we show that the empirically best box-splitting strategy is the only scale-invariant one – and is, thus, optimal under any scale-invariant final optimality criterion.

4.2 Optimal Selection of Penalty (Barrier) Functions

A similar approach can be used for reducing constraint optimization to non-constrained one. A well-known Lagrange multiplier method minimizes a function $f(x)$ under a constraint $g(x) = 0$ by reducing it to the unconstrained problem of optimizing a new objective function $f(x) + \lambda \cdot g(x)$. One of the known approaches to solving a similar problem with a constraint $g(x) \geq 0$ is the *penalty* (*barrier*) method in which we reduce the original problem to the un-constrained problem of optimizing a new objective function $f(x) + \lambda \cdot g(x) + \mu \cdot P(g(x))$, for an appropriate (non-linear) penalty function $P(y)$. Traditionally, the most widely used penalty functions are $P(y) = y \cdot \ln(y)$ and $P(y) = y^2$.

In [NK97], we show that the only y-scale-invariant families $\{\lambda \cdot y + \mu \cdot P(y)\}$ are families corresponding to $P(y) = y \cdot \ln(y)$ and $P(y) = y^\alpha$ for some real number α. Thus, under any scale-invariant optimality criterion, the optimal penalty function must indeed take one of these forms.

This example also shows that we can go beyond theoretical justification of empirically best heuristic, towards finding new optimal heuristics: indeed, for penalty functions, instead of single-parameter families $\{\lambda \cdot y + \lambda \cdot P(y)\}$, we can consider multiple-parameter families

$$\{\lambda \cdot y + \mu_1 \cdot P_1(y) + \ldots + \mu_m \cdot P_m(y)\}$$

for several functions $P_1(y), \ldots, P_m(y)$. In this case, the optimal functions have also been theoretically found: they are of the type

$$P_i(y) = L^{\alpha_i} \cdot (\ln(y))^{p_i}$$

for real (or complex) values α_i and non-negative integer values of p_i.

4.3 Other Examples

Similar symmetry-based techniques provide an explanation of several other empirically optimal techniques:

- sometimes, it is beneficial to (slightly) enlarge the original (non-degenerate) box $[x^L, x^U]$ and thus improve the performance of the algorithm; the empirically efficient "epsilon-inflation" technique [Rum80, Rum92]

  $$[x_i^L, x_i^U] \to [(1+\varepsilon)x_i^L - \varepsilon \cdot x_i^U, (1+\varepsilon)x_i^U - \varepsilon \cdot x_i^L]$$

  was proven to be the only shift- and scale-invariant technique and thus, the only one optimal under an arbitrary shift-invariant and scale-invariant optimality criterion [KSM97] (see also [Rum98]);
- by using shift-invariance, we explain why the probability proportional to $\exp(-\gamma \cdot f(x))$ is optimal in simulated annealing [NK97],
- by using scale- and shift-invariance, we explain why exponential and power re-scalings of the objective function are optimal in genetic algorithms [NK97];
- by using appropriate symmetries, we also explain, in [ISK02], the empirically optimal selection of probabilities in swarm ("ant") optimization (see, e.g., [KES01]).

5 Proofs

5.1 Proof of Proposition 1

For $\alpha = 1$, the condition (3) takes the form

$$(g(x) - g(x^L)) \cdot (h(x^U) - h(x)) =$$

$$C(s) \cdot (g(x+s) - g(x^L + s)) \cdot (h(x^U + s) - h(x+s)), \tag{10}$$

where we denoted $C(s) \stackrel{\text{def}}{=} \widetilde{\alpha}(1, s)$. To simplify this equation, let us separate the variables:

- let us move all terms containing x^L to the left-hand side – by dividing both sides by $(g(x+s) - g(x^L + s))$, and

- let us move all terms containing x^U to the right-hand side – by dividing both sides by $(h(x^U) - h(x))$.

As a result, we arrive at the following equation:

$$\frac{g(x) - g(x^L)}{g(x+s) - g(x^L+s)} = C(s) \cdot \frac{h(x^U+s) - h(x+s)}{h(x^U) - h(x)}. \tag{11}$$

Let us denote the left-hand side of this equation by A. By definition, the value A depends on x, s, and x^L. Since A is equal to the right-hand side, and the right-hand side does not depend on x^L, the expression A cannot depend on x^L, so $A = A(x,s)$, i.e.,

$$\frac{g(x) - g(x^L)}{g(x+s) - g(x^L+s)} = A(x,s). \tag{12}$$

Multiplying both sides by the denominator, we conclude that

$$g(x) - g(x^L) = A(x,s) \cdot (g(x+s) - g(x^L+s)). \tag{13}$$

Differentiating both sides by x^L, we conclude that

$$-g'(x^L) = -A(x,s) \cdot g'(x^L+s), \tag{14}$$

i.e., equivalently,

$$\frac{g'(x^L)}{g'(x^L+s)} = A(x,s). \tag{15}$$

In this equation, the left-hand side does not depend on x, so the right-hand does not depend on x either, i.e., $A(x,s) = A(s)$. Thus, the equation (13) takes the form

$$a(s) \cdot (g(x) - g(x^L)) = (g(x+s) - g(x^L+s)), \tag{16}$$

where we denoted $a(s) \stackrel{\text{def}}{=} 1/A(s)$.

The function $g(x)$ is smooth, hence the function $a(s)$ is smooth too – as the ratio of two smooth functions. Differentiating both sides of the equation (16) with respect to s and taking $s = 0$, we get

$$a \cdot (g(x) - g(x^L)) = (g'(x) - g'(x^L)), \tag{17}$$

where $a \stackrel{\text{def}}{=} a'(0)$.

To simplify this equation, let us separate the variables, i.e., let us move all the term depending on x to the right-hand side and all the terms depending on x^L to the left-hand side. As a result, we arrive at the following:

$$g'(x^L) - a \cdot g(x^L) = g'(x) - a \cdot g(x). \tag{18}$$

The right-hand side is a function of x only, but since it is equal to the left-hand side – which does not depend on x at all – it is simply a constant. If we denote this constant by b, we get the following equation:

$$g'(x) - a \cdot g(x) = b, \tag{19}$$

i.e.,

$$\frac{dg}{dx} = a \cdot g + b \tag{20}$$

and

$$\frac{dg}{a \cdot g + b} = dx. \tag{21}$$

When $a = 0$, integrating both sides of this equation, we get $\frac{1}{b} \cdot g(x) = x + C$, i.e., $g(x) = b \cdot x + b \cdot C$. When $a \neq 0$, then for $\widetilde{g}(x) \stackrel{\text{def}}{=} g(x) + \frac{b}{a}$, we get

$$\frac{d\widetilde{g}}{a \cdot \widetilde{g}} = dx. \tag{22}$$

hence $\frac{1}{a} \cdot \ln(\widetilde{g}(x)) = x + C$ thence $\ln(\widetilde{g}(x)) = a \cdot x + a \cdot C$, so $\widetilde{g}(x) = C \cdot \exp(a \cdot x)$ and $g(x) = \widetilde{g}(x) - \frac{b}{a} = C \cdot \exp(a \cdot x) + C_1$ for some constants C, a, and C_1. The proposition is proven.

5.2 Proof of Proposition 2

By introducing new variables $X = \ln(x)$, $X^L = \ln(x^L)$, and $X^U = \ln(x^U)$ – so that $x = \exp(X)$, $x^L = \exp(X^L)$, and $x^U = \exp(X^U)$, and by introducing new functions $G(X) = g(\exp(x))$ and $H(X) = h(\exp(x))$, one can easily check that if the pair $(g(x), h(x))$ is scale-invariant, then the new pair $(G(X), H(X))$ is shift-invariant.

We already know, from Proposition 1, how shift-invariant pairs look like: we have either $G(X) = A + C \cdot \exp(\gamma \cdot X)$ or $G(X) = A + k \cdot X$. From the definition of $G(X)$, we conclude that $g(x) = G(\ln(x))$; thus, we have either $g(x) = A + C \cdot \exp(\gamma \cdot \ln(x)) = A + C \cdot x^\gamma$ or $g(x) = A + k \cdot \ln(x)$. The proposition is proven.

5.3 Proof of Proposition 3

For $\alpha = 1$, the shift invariance requirement (6) takes the form

$$C(s) \cdot a(x + s, x^L + s) \cdot b(x + s, x^U + s) = a(x, x^L) \cdot b(x, x^U), \tag{23}$$

where $C(s) \stackrel{\text{def}}{=} \widetilde{\alpha}(1, s)$. Let us separate the variables by dividing both sides of this equation by $a(x, x^L)$ and $b(x, x^U)$; we then get

$$C(s) \cdot \frac{a(x+s, x^L+s)}{a(x, x^L)} = \frac{b(x+s, x^U+s)}{b(x, x^U)}. \tag{24}$$

The left-hand side ℓ of this equality depends only on x, x^L, and s. Since it is equal to the right-hand side, which does not depend on x^L at all, we can conclude that ℓ only depends on x and s:

$$C(s) \cdot \frac{a(x+s, x^L+s)}{a(x, x^L)} = \ell(x, s), \tag{25}$$

i.e., equivalently,

$$\frac{a(x+s, x^L+s)}{a(x, x^L)} = \widetilde{\ell}(x, s), \tag{26}$$

where $\widetilde{\ell}(x, s) \stackrel{\text{def}}{=} \dfrac{\ell(x, s)}{C(s)}$. For convenience (and without losing generality), we can describe $\widetilde{\ell}$ as depending on x and $x+s$:

$$\frac{a(x+s, x^L+s)}{a(x, x^L)} = N(x, x+s), \tag{27}$$

where $N(x, a) \stackrel{\text{def}}{=} \widetilde{\ell}(x, a-x)$.

We can perform the transition from x to $x+s$ in one step, as above, or we can first go to $x+(-x)=0$, and then to $0+(x+s)=x+s$. We then have

$$N(x, x+s) = \frac{a(x+s, x^L+s)}{a(x, x^L)} =$$

$$\frac{a(0+(x+s), (x^L-x)+(x+s))}{a(0, x^L-x)} \cdot \frac{a(x+(-x), x^L-x)}{a(x, x^L)} = \tag{28}$$

$$N(0, x+s) \cdot N(x, 0),$$

i.e.,

$$N(x, x+s) = N(0, x+s) \cdot N(x, 0). \tag{29}$$

For $s=0$, the equation (27) leads to $N(x, x)=1$, hence from (29), we conclude that $N(0, x) \cdot N(x, 0) = 1$ thence $N(x, 0) = \dfrac{1}{N(0, x)}$; thus, (29) takes the form

$$N(x, x+s) = \frac{n(x+s)}{n(x)}, \tag{30}$$

where $n(x) \stackrel{\text{def}}{=} N(0, x)$. Substituting (30) into the formula (27), we conclude that

$$\frac{a(x+s, x^L+s)}{n(x+s)} = \frac{a(x, x^L)}{n(x)}. \tag{31}$$

In particular, for $s=-x^L$, we conclude that

$$\frac{a(x,x^L)}{n(x)} = \frac{a(x-X^L,0)}{n(x-x^L)}, \tag{32}$$

i.e.,

$$a(x,x^L) = A_0(x-x^L)\cdot n(x), \tag{33}$$

where $A_0(z) \stackrel{\text{def}}{=} \dfrac{a(z,0)}{n(z)}$. Similarly, $b(x,x^U) = B_0(x^U-x)\cdot m(x)$ for some functions $B(z)$ and $m(x)$. Hence,

$$a(x,x^L)\cdot b(x,x^U) = A_0(x-x^L)\cdot B_0(x^U-x)\cdot p(x), \tag{34}$$

where $p(x) \stackrel{\text{def}}{=} m(x)\cdot n(x)$.

In this expression, the terms $A_0(x-x^L)$ and $B_0(x^U-x)$ are shift-invariant, so shift-invariance (23) of the product (34) means that $C(s)\cdot p(x+s) = p(x)$ for all x and s, i.e., that

$$p(x+s) = c(s)\cdot p(x), \tag{35}$$

where $c(s) \stackrel{\text{def}}{=} 1/C(s)$. Since the functions a and b are smooth, the functions p and c are smooth as well. Differentiating both sides of (35) w.r.t. s and substituting $s=0$, we conclude that $p'(x) = \gamma\cdot p(x)$, where $\gamma \stackrel{\text{def}}{=} c'(0)$, hence $\dfrac{dp}{dx} = \gamma\cdot p$, $\dfrac{dp}{p} = \gamma\cdot dx$, and $\ln(p(x)) = \gamma\cdot x + C_1$; thus, $p(x) = C_2\cdot\exp(\gamma\cdot x)$.

Since $\exp(\gamma\cdot x) = \exp(\gamma\cdot(x-x^L))\cdot\exp(\gamma\cdot x^L)$, the equation (34) takes the desired form

$$a(x,x^L)\cdot b(x,x^U) = A(x-x^L)\cdot B_0(x^U-x)\cdot e^{\gamma\cdot x^L}, \tag{36}$$

where $A(z) \stackrel{\text{def}}{=} A_0(z)\cdot C_2\cdot\exp(\gamma\cdot z)$. The proposition is proven.

5.4 Proof of Proposition 4

For convenience, let us introduce new variables $X \stackrel{\text{def}}{=} x-x^L$ and $Y \stackrel{\text{def}}{=} x^U-x$. In terms of these variables, $x^U-x^L = X+Y$, and thus, the desired formula (7) takes the form

$$a(X)\cdot a(Y) = b(X)+b(Y)+c(X+Y). \tag{37}$$

Differentiating both sides of this equality w.r.t. Y, we conclude that

$$a(X)\cdot a'(Y) = b'(Y)+c'(X+Y). \tag{38}$$

Differentiating once again, this time w.r.t. X, we conclude that

$$a'(X)\cdot a'(Y) = c''(X+Y). \tag{39}$$

In particular, for $Y = 0$, we get

$$a'(X) \cdot a'(0) = c''(X). \tag{40}$$

Substituting this expression for $c''(X)$ into the formula (39), we conclude that

$$a'(X) \cdot a'(Y) = a'(X+Y) \cdot a'(0). \tag{41}$$

Dividing both sides by $a'(0)$, we get

$$\frac{a'(X)}{a'(0)} \cdot \frac{a'(Y)}{a'(0)} = \frac{a'(X+Y)}{a'(0)}, \tag{42}$$

i.e.,

$$A(X+Y) = A(X) \cdot A(Y), \tag{43}$$

where $A(X) \stackrel{\text{def}}{=} \dfrac{a'(X)}{a'(0)}$. Differentiating both sides of (43) by Y and substituting $Y = 0$, we conclude that $A'(X) = \gamma \cdot A(X)$, where $\gamma \stackrel{\text{def}}{=} A'(0)$. Similarly to the proof of Proposition 3, we get $A(X) = C_1 \cdot \exp(\gamma \cdot X)$ for some constant C_1. Therefore, $a'(X) = a'(0) \cdot A(X) = C_2 \cdot \exp(\gamma \cdot X)$, where $C_2 \stackrel{\text{def}}{=} a'(0) \cdot C_1$. Thus:

- If $\gamma = 0$, we get $a'(X) = C_2$, hence $a(X) = C_2 \cdot X + C_3$ for some constant C_3. From the condition $a(0) = 0$, we conclude that $C_3 = 0$.
- If $\gamma \neq 0$, then $a(X) = C_3 \cdot \exp(\gamma \cdot X) + C_4$, where $C_3 \stackrel{\text{def}}{=} \dfrac{C_2}{\gamma}$. Here too, from the condition that $a(0) = 0$, we conclude that $a(X) = C_4 \cdot (1 - \exp(\gamma \cdot X))$.

The proposition is proven.

5.5 Proof of Proposition 5

We have already shown, in the proof of Proposition 1, that:

- for linear or exponential functions, the corresponding family is shift-invariant, and
- vice versa, that if a family is shift-invariant, then it has the form $F(g,h)$ for some linear or exponential functions $g(x)$ and $h(x)$.

1°. To prove the first part of Proposition 5, we thus need to show that for every shift-invariant and final optimality criterion, the corresponding optimal family F_{opt} is shift-invariant, i.e., that $T_s(F_{\text{opt}}) = F_{\text{opt}}$ for all s. Then, the result will follow from Proposition 1.

Indeed, the transformation T_s is invertible, its inverse transformation is a shift by $-s$: $T_s^{-1} = T_{-s}$. Now, from the optimality of F_{opt}, we conclude that for every $F' \in A$, $T_s^{-1}(IF') \preceq F_{\text{opt}}$. From the invariance of the optimality

criterion, we can now conclude that $F' \preceq T_s(F_{\text{opt}})$. This is true for all $F' \in A$ and therefore, the family $T(F_{\text{opt}})$ is optimal.

But since the criterion is final, there is only one optimal indicator function; hence, $T_s(F_{\text{opt}}) = F_{\text{opt}}$. So, the optimal family is indeed invariant and hence, due to Proposition 1, it coincides with $F(g,h)$ for some linear or exponential functions $g(x)$ and $h(x)$. The first part is proven.

2°. Let us now prove the second part of Proposition 5. Let $g(x)$ and $h(x)$ be fixed linear or exponential functions, and let $F_0 = F(g,h)$ be the corresponding family. We will then define the optimality criterion as follows: $F \preceq F'$ iff F' is equal to this F_0.

Since the family F_0 is shift-invariant, thus the defined optimality criterion is also shift-invariant. It is also clearly final.

The family F_0 is clearly optimal w.r.t. this shift-invariant and final optimality criterion. The proposition is proven.

Acknowledgments

This chapter was supported in part by NASA under cooperative agreement NCC5-209, NSF grants EAR-0225670 and DMS-0532645, Army Research Lab grant DATM-05-02-C-0046, Star Award from the University of Texas System, Texas Department of Transportation grant No. 0-5453, and the National Institute of Health grant R01 GM52032. The authors are greatly thankful to the editors and to the anonymous referees for valuable comments.

References

[AAF98] Adjiman, C.S., Androulakis, I., Floudas, C.A.: A global optimization method, αBB, for general twice-differentiable constrained NLP II. Implementation and computational results. Computers and Chemical Engineering, **22**, 1159–1179 (1998)

[ADA98] Adjiman, C.S., Dallwig, S., Androulakis, I., Floudas, C.A.: A global optimization method, αBB, for general twice-differentiable constrained NLP I. Theoretical aspects. Computers and Chemical Engineering, **22**, 1137–1158 (1998)

[AF04] Akrotirianakis, I.G., Floudas, C.A.: Computational experience with a new class of convex underestimators: box-constrained NLP problems. Journal of Global Optimization, **29**, 249–264 (2004)

[AF06] Akrotirianakis, I.G., Floudas, C.A. A new class of improved convex underestimators for twice continuously differentiable constrained NLPs. Journal of Global Optimization, to appear.

[CG98] Casado, L.G., García, I.: New load balancing criterion for parallel interval global optimization algorithm, In: Proc. of the 16th IASTED International Conference, Garmisch-Partenkirchen, Germany, February 1998, 321–323 (1998)

[CGC00] Casado, L.G., García, I., Csendes, T.: A new multisection technique in interval methods for global optimization. Computing, **65**, 263–269 (2000)
[Flo00] Floudas, C.A.: Deterministic Global Optimization: Theory, Methods, and Applications. Kluwer, Dordrecht (2000)
[HP95] Horst, R., Pardalos P.M. (eds.): Handbook of Global Optimization, Kluwer, Dordrecht (1995)
[ISK02] Iourinski, D., Starks, S.A., Kreinovich, V., Smith, S.F.: Swarm intelligence: theoretical proof that empirical techniques are optimal. In: Proceedings of the 2002 World Automation Congress WAC'2002, Orlando, Florida, June 9–13, 107–112 (2002)
[KK96] Kearfott, R.B.: Rigorous Global Search: Continuous Problems. Kluwer, Dordrecht (1996)
[KK98] Kearfott, R.B., Kreinovich, V.: Where to Bisect a Box? A Theoretical Explanation of the Experimental Results. In: Alefeld, G., Trejo, R.A. (eds.), Interval Computations and its Applications to Reasoning Under Uncertainty, Knowledge Representation, and Control Theory. Proceedings of MEXICON'98, Workshop on Interval Computations, 4th World Congress on Expert Systems, México City, México (1998)
[KK05] Kearfott, R.B., Kreinovich, V.: Beyond Convex? Global Optimization Is Feasible Only for Convex Objective Functions: A Theorem. Journal of Global Optimization, **33**, 617–624 (2005)
[KES01] Kennedy, J., Eberhart, R., Shi, Y.: Swarm Intelligence. Morgan Kaufmann, San Francisco, California (2001)
[KC01] Kreinovich, V., Csendes, T.: Theoretical Justification of a Heuristic Subbox Selection Criterion. Central European Journal of Operations Research CEJOR, **9**, 255–265 (2001)
[KSM97] Kreinovich, V., Starks, S.A., Mayer, G.: On a Theoretical Justification of The Choice of Epsilon-Inflation in PASCAL-XSC. Reliable Computing, **3**, 437–452 (1997)
[MF94] Maranas, C.D., Floudas, C.A.: Global minimal potential energy conformations for small molecules. Journal of Global Optimization, **4**, 135–170 (1994)
[Mih96] Michalewicz, Z.: Genetic Algorithms + Data Structures = Evolution Programs. Springer, Berlin (1996)
[NK97] Nguyen, H.T., Kreinovich, V. Applications of continuous mathematics to computer science. Kluwer, Dordrecht (1997)
[Pin96] Pinter, J.D.: Global Optimization in Action. Kluwer, Dordrecht (1996)
[Rat92] Ratz, D., Automatische Ergebnisverifikation bei globalen Optimierungsproblemen. Ph.D. dissertation, Universität Karlsruhe (1992)
[Rat94] Ratz, D.: Box-Splitting Strategies for the Interval Gauss–Seidel Step in a Global Optimization Method. Computing, **53**, 337–354 (1994)
[Rum80] Rump, S.M.: Kleine Fehlerschranken bei Matrixproblemen, Ph.D. dissertation, Universität Karlsruhe (1980)
[Rum92] Rump, S.M.: On the solution of interval linear systems. Computing, **47**, 337–353 (1992)
[Rum98] Rump, S.M.: A Note on Epsilon-Inflation. Reliable Computing, **4**(4), 371–375 (1998)
[TS02] Tawarmalani, M., Sahinidis, N.V.: Convexification and Global Optimization in Continuous and Mixed-Integer Nonlinear Programming: Theory, Algorithms, Software, and Applications. Kluwer, Dordrecht (2002)

[Vav91] Vavasis, S.A.: Nonlinear Optimization: Complexity Issues. Oxford University Press, New York (1991)