© International Journal of Combinatorial Optimization Problems and Informatics, Vol. 7, No. 3, Sep-Dec 2016, pp. 87-93. ISSN: 2007-1558.

Structure of Filled Functions: Why Gaussian and Cauchy Templates Are Most Efficient

Vyacheslav Kalashnikov[1,2,3], Vladik Kreinovich[4], José Guadalupe Flores-Muñiz[5], Nataliya Kalashnykova[5]

[1]*Tecnológico de Monterrey (ITESM), Campus Monterrey, Mexico*
[2]*Central Economics & Mathematics Institute (CEMI), Moscow, Russia*
[3]*Sumy State University (SumDU), Sumy, Ukraine*
[4]*University of Texas at El Paso (UTEP), El Paso, TX, USA*
[5]*Universidad Autónoma de Nuevo León (UANL), San Nicolás de los Garza, NL, Mexico,*

kalash@itesm.mx, vladik@utep.edu, jose_guadalupe64@hotmail.com, nkalash2009@gmail.com

Abstract. One of the main problems of optimization algorithms is that they often end up in a local optimum. It is, therefore, necessary to make sure that the algorithm gets out of the local optimum and eventually reaches the global optimum. One of the promising ways guiding one from the local optimum is prompted by the filled function method. It turns out that empirically, the best smoothing functions to use in this method are the Gaussian and Cauchy functions. In this paper, we provide a possible theoretical explanation of this empirical effect.

Keywords: Optimization Algorithms, Filled Function Method, Gaussian and Cauchy Functions.

1. Introduction

Many optimization techniques end up in a local optimum. Therefore, to solve a global optimization problem, it is necessary to move out of the local minimum, so that eventually, we will end up in a global minimum (or at least in a better local minimum).

One of the most efficient techniques for avoiding a local optimum is the *filled function* method originally proposed by Renpu in [1]; *see* also papers by Kalashnikov et al. [2], and by Wu et al. [3] – [4] .

In this method, once we reach a local optimum x^*, then, instead of optimizing the original objective function $f(x)$, we optimize an auxiliary expression

$$K\left(\frac{x-x^*}{\sigma}\right)\cdot F\left(f(x), f\left(x^*\right), x\right)+G\left(f(x), f\left(x^*\right), x\right), \quad (1)$$

for appropriate functions $K(x), F\left(f, f^*, x\right)$, and $G\left(f, f^*, x\right)$, and for an appropriate value σ. Once we find the optimum of this auxiliary expression, we use it as a new first approximation to find the optimum of the original objective function $f(x)$.

An interesting empirical fact is as follows. Although several different functions $K\left(x-x^*\right)$ have been proposed, it turns out that the most computationally efficient functions are the Gaussian functions

$$K(x)=exp\left(-\|x\|^2\right), \quad (2)$$

Received Jan 29, 2016 / Accepted Aug 21, 2016

and the Cauchy function

$$K(x) = \frac{1}{1+\|x\|^2}; \quad (3)$$

see, e.g., [2] – [3].

A natural question arises: Are these functions indeed the most efficient, or they are simply the most efficient among a few functions that have been tried, and there are other, more efficient functions $K(x)$?

In this paper, we formulate the above question as a precise mathematical problem, and we show that in this formulation, the Gaussian and the Cauchy functions are indeed the most efficient ones. This result provides a possible theoretical explanation for the above empirical fact. This result also shows that – at least in this formulation – the Gaussian and the Cauchy functions $K(x)$ are indeed the best. This will hopefully make users more confident in (these versions of) the function filling method.

The rest of the paper is organized as follows. Section 2 specifies the considered problem, while Section 3 describes the desired properties of the smoothing functions. In Section 4, we show that any infinitely divisible probability distribution generates smoothing functions boasting the required properties. Finally, in Section 5, additional requirements to the smoothing functions are discussed, and the empirical fact that the Gaussian and Cauchy distribution are most efficient for constructing the filled functions is explained. Conclusions, acknowledgments, and a list of references complete the paper.

2. Problem Specification

One of the known ways to eliminate local optima is to apply a *weighted smoothing* (see, e.g., [5]), i.e., to replace the original objective function $f(x)$ with a "smoothed" one:

$$f^*(x) = \int_{-\infty}^{\infty} K\left(\frac{x-x'}{\sigma}\right) \cdot f(x')\,dx', \quad (4)$$

for some smooth nonnegative weighting function $K(x)$ and for an appropriate value of σ.

Remark 1. The weighting function is usually selected in such a way that $K(-x) = K(x)$, and $\int_{-\infty}^{\infty} K(x)\,dx < +\infty$.

- The first condition comes from the fact that a priori, we have no reason to prefer different orientation of the coordinate system.
- The second requirement arises from the natural expectation that for a constant function $f(x) = const$, smoothing leads to a finite constant. □

Remark 2. Selecting the value of the parameter σ is really important. Indeed,

- When $\sigma \to 0$ the smoothed function tends to the original function $f(x)$.
- When $\sigma \to \infty$, the smoothed function simply becomes a constant equal to $\int_{-\infty}^{\infty} f(x)dx$. □

Because of that, taking into account Remark 2, we need to guarantee that:

- When the value of σ is too small, that is, when the smoothing only covers very tiny neighborhood of each point *x*, the smoothed function $f^*(x)$ is so close to the original function $f(x)$ that one still observes *all* the local optima.
- On the other hand, if the parameter σ has a too large value, the smoothed function $f^*(x)$ differs very much from the original objective function $f(x)$ so that the optimum of the smoothed function may have nothing in common with the optimum of the original objective function $f(x)$.

So, for the smoothing function to work, it is important to select an appropriate value of σ, which can be found by several iterations of a try-and-error procedure. This procedure may be roughly described as follows:

***Procedure 1* (*to find a good approximation of* σ).**

- If we have smoothed the objective function "too much", i.e., if we have selected a too large value of σ, then we need to "unsmooth" it, that it, to take a smaller value of σ.
- Vice versa, if we haven't smoothed the objective function to a proper grade, that is, of we have selected σ to be too small, then we need to smooth it a bit more, i.e., to pick a larger value of σ.

2.1. Computationally effective filled functions

However, the most important characteristic of the selected fill function is its computational effectiveness. For the process to be computationally effective, it is desirable to be able to use the previous smoothing result.

Suppose that we first applied smoothing with some value σ and it turns out that we need to apply smoothing with a larger value $\sigma' > \sigma$. In principle, we can then start again with the original objective function $f(x)$ and apply the smoothing with the new parameter.

However, once we have smoothed the function too much, it is difficult to unsmooth it. Therefore, a usual approach is that we first try some small smoothing. If the resulting smoothed function $f^*(x)$ still leads to a similar local maximum, we smooth it some more, etc.

For a small σ, to find each value $f^*(x)$ of the smoothed function, we only need to consider values of $f(x')$ in a small neighborhood of *x*. The larger σ, the larger this neighborhood, the more values $f(x')$ we need to take into account and thus, the more computations we require.

Thus, once we have a smoothed function $f^*(x)$ corresponding to some value of σ, and we need to compute a smoothed function $f^{**}(x)$ corresponding to a larger value $\sigma' > \sigma$, it is more computationally efficient not to start "from scratch" and apply smoothing with parameter σ' to the original objective function $f(x)$, but rather apply a small additional smoothing to the smoothed function $f^*(x)$.

3. Resulting Requirements on the Smoothing Function

Before we formulate the resulting requirements on the smoothing function $K(x)$, from now on, we will omit the lower and upper limits in the integrals involved, thus always meaning that we integrate over the full domains of the participating functions.

From our reasoning outlined above, we deduce the requirement that for every σ' and σ, there must exist an appropriate value Δ_σ such that applying smoothing with parameter Δ_σ to the (already) smoothed function

$$f^*(x) = \int K\left(\frac{x - x'}{\sigma}\right) \cdot f(x')\,dx', \tag{5}$$

will lead us to the desired function

$$f^{**}(x) = \int K\left(\frac{x - x'}{\sigma'}\right) \cdot f(x')\,dx'. \tag{6}$$

In other words, we need to make sure that for every objective function $f(x)$, we have

$$f^*(x) = \int K\left(\frac{x-x'}{\sigma'}\right) \cdot f(x')\,dx' = \int K\left(\frac{x-x'}{\Delta_\sigma}\right) \cdot f^*(x')\,dx'. \quad (7)$$

3.1. Analysis of the requirement

By substituting the expression

$$f^*(x') = \int K\left(\frac{x'-x''}{\sigma}\right) \cdot f(x'')\,dx'', \quad (8)$$

into the above formula, we conclude that

$$\int K\left(\frac{x-x'}{\sigma'}\right) \cdot f(x')\,dx' = \int K\left(\frac{x-x'}{\Delta_\sigma}\right) \cdot \left(\int K\left(\frac{x'-x''}{\sigma}\right) \cdot f(x'')\,dx''\right) dx'. \quad (9)$$

Furthermore, changing the order of integration and renaming the variables in the right-hand side (namely, swapping x' and x'') brings us to the relationship

$$\int K\left(\frac{x-x'}{\sigma'}\right) \cdot f(x')\,dx' = \int K'(x-x') \cdot f(x')\,dx', \quad (10)$$

where we denoted

$$K'(x-x') := \int K\left(\frac{x-x''}{\Delta_\sigma}\right) \cdot K\left(\frac{x''-x'}{\sigma}\right) dx''. \quad (11)$$

Eq. (11) must hold for every objective function $f(x)$. Thus, the corresponding functions $K\left(\frac{x-x'}{\sigma'}\right)$ and $K'(x-x')$ must coincide:

$$K\left(\frac{x-x'}{\sigma'}\right) = \int K\left(\frac{x-x''}{\Delta_\sigma}\right) \cdot K\left(\frac{x''-x'}{\sigma}\right) dx''. \quad (12)$$

4. Infinitely Divisible Probability Distributions

The requirement of Eq. (12) can be reduced to the analysis of infinitely divisible probability distributions. The function $K(x)$ is nonnegative, and its integral $\int K(x)\,dx$ is finite. Thus, after dividing $K(x)$ by the value of this integral, we come to a probability density function

$$\rho_X(x) = \frac{K(x)}{\int K(y)\,dy}, \quad (13)$$

boasting the necessary properties: $\rho_X(x) \geq 0$ and $\int \rho_X(x)\,dx = 1$. Here, X denotes the random variable having the probability density defined by Eq. (13).

Now, for any real number $\sigma > 0$, the probability density function (pdf) for $Y = \sigma \cdot X$ can be found as follows. By definition, the original pdf can be characterized by the equality

$$P\{x \leq X \leq x + dx\} = \rho_X(x)\,dx. \quad (14)$$

(Here, by *P*, we denote probability of the event in question). Therefore, the pdf for *Y* must satisfy

$$P\{y \leq Y \leq y + dy\} = \rho_Y(y)\,dy. \quad (15)$$

Now since $Y = \sigma \cdot X$ we have

$$P\{y \leq Y \leq y + dy\} = P\{y \leq \sigma \cdot X \leq y + dy\} = P\left\{\frac{y}{\sigma} \leq X \leq \frac{y}{\sigma} + \frac{dy}{\sigma}\right\}. \quad (16)$$

Hence, Eq. (16) together with the definition of $\rho_X(x)$, imply

$$P\{y \le Y \le y + dy\} = \rho_X\left(\frac{y}{\sigma}\right) \cdot \frac{dy}{\sigma}, \tag{17}$$

thus yielding the key relationship between the two density functions:

$$\rho_Y(y) = \rho_X\left(\frac{y}{\sigma}\right) \cdot \frac{1}{\sigma}. \tag{18}$$

Now, based upon [6] – [9], we introduce the following concept.

Definition 1. A distribution is called *infinitely divisible* if for any two values $\sigma_1 > 0$ and $\sigma_2 > 0$, and for every two independent random variables X_1 and X_2 distributed according to this distribution, their linear combination $Y = Y_1 + Y_2$, where $Y_i = \sigma_i \cdot X_i, i = 1,2,$ also has the representation $Y = \sigma_Y \cdot X$ for some $\sigma_Y > 0$ and some random variable *X* boasting the same distribution as X_1 and X_2.

As an example of an infinitely divisible distribution, one can consider a Gaussian (normal) distribution with mean 0 and standard deviation 1. In this case, $\sigma_Y = \sqrt{\sigma_1^2 + \sigma_2^2}$.

Since the variables $Y_1 = \sigma_1 \cdot X_1$ and $Y_2 = \sigma_2 \cdot X_2$ are independent, their joint probability density function is equal to the product of the corresponding probability densities:

$$\rho_{Y_1,Y_2}(y_1, y_2) = \rho_{Y_1}(y_1) \cdot \rho_{Y_2}(y_2). \tag{19}$$

In order to deduce the probability density $\rho_Y(y)$ of $Y = Y_1 + Y_2$, we have to "sum up" the values $\rho_{Y_1,Y_2}(y_1, y_2)$ corresponding to all the pairs (y_1, y_2) giving $y_1 + y_2 = y$, that is, for which $y_1 = y - y_2$. In other words, we take the integral

$$\rho_Y(y) = \int_{-\infty}^{\infty} \rho_{Y_1}(y - y_2)\, \rho_{Y_2}(y_2)\, dy_2. \tag{20}$$

Now since $Y_i = \sigma_i \cdot X_i, i = 1,2,$ and $Y = \sigma_Y \cdot X$, we have

$$\rho_{Y_i}(y_i) = \rho_{X_i}\left(\frac{y_i}{\sigma_i}\right) \cdot \frac{1}{\sigma_i}, i = 1,2, \text{ and } \rho_Y(y) = \rho_X\left(\frac{y}{\sigma_Y}\right) \cdot \frac{1}{\sigma_Y}. \tag{21}$$

Therefore, Eq. (20) reduces to the identity

$$\frac{1}{\sigma_Y} \cdot \rho_X\left(\frac{y}{\sigma_Y}\right) = \frac{1}{\sigma_1} \cdot \frac{1}{\sigma_2} \cdot \int_{-\infty}^{\infty} \rho_X\left(\frac{y - y_2}{\sigma_1}\right) \cdot \rho_X\left(\frac{y_2}{\sigma_2}\right) dy_2. \tag{22}$$

Recall that here, $\rho_X(x) = const \cdot K(x)$, so the above formula Eq. (22) finally takes the form

$$K\left(\frac{y}{\sigma_Y}\right) \propto \int_{-\infty}^{\infty} K\left(\frac{y - y_2}{\sigma_1}\right) \cdot K\left(\frac{y_2}{\sigma_2}\right) dy_2, \tag{23}$$

For $y = x - x'$, $y_2 = x'' - x'$, $\sigma_1 = \Delta_\sigma$, $\sigma_2 = \sigma$, and $\sigma_Y = \sigma'$, Eq. (23) clearly coincides with our requirement for the smoothing function $K(x)$ described above by Eq. (12). In other words, we have just established an interesting theoretical result:

Proposition 1. *Any infinitely divisible probability distribution generates a smoothing function* $K(x)$ *satisfying requirement described by Eq.* (*12*). □

5. Additional Requirements and Computational Hints

Other requirements necessary for the construction of filled functions are:

(i) *Symmetry*. The condition is equivalent to $\rho_X(-x) = \rho_X(x)$, so we only need to consider symmetric probability distributions.

(ii) *Computability*. The smoothing function $K(x)$ must be easy to compute.

Indeed, for the resulting computational process to be efficient, we need to guarantee that the function $K(x)$ is easy to compute. Ideally, $K(x)$ should boast an explicit formula whose computation consists in performing a small number of arithmetic operations and applying easy-to-calculate elementary functions.

It is well known that out of all symmetric infinitely divisible distributions, only two distributions have such an explicit expression: the Gaussian distribution corresponding to $K(x) \propto exp(-x^2)$, and the Cauchy distribution for which $K(x) \propto \frac{1}{1+x^2}$; see, again, [6] – [9]. This explains why these two smoothing cores are used to construct filled functions.

5.1. Approximating the integral with a sum

The above arguments explain why instead of optimizing the original function $f(x)$, we should optimize its smoothed version

$$f^*(x) = \int K\left(\frac{x-x'}{\sigma}\right) \cdot f(x')\,dx'. \quad \textbf{(24)}$$

In most practical cases, the only way to compute an integral is to approximate it by the weighted sum of the values of the corresponding functions calculated at different points. The simplest possible case is when we consider one or two points; then, we get a linear combination of two values of $f(x)$ with weights proportional to $K\left(\frac{x-x'}{\sigma}\right)$. But this is exactly what the filled function does in order to find a way from a current local optimum to another one.

The latter means we indeed get a theoretical explanation of the empirical fact – that the Gaussian and Cauchy smoothing functions $K(x) \propto exp(-x^2)$ and $K(x) \propto \frac{1}{1+x^2}$, respectively, prove to be the most efficient in the filled function method.

6. Concluding Remarks

It is well known that in non-convex optimization the filled function techniques are very popular in order to make steps from one local optimum to another. In this paper, a methodology is proposed to explain the empirical fact that only Gaussian and Cauchy distributions are mostly used when generating the filled functions.

The explanation is made by analyzing the main requirements to the smoothing functions serving as integral cores when developing the smoothed functions that replace the original objective functions locally in the filled function method.

Then, it is demonstrated that any infinitely divisible probability distribution generates smoothing functions boasting the required properties. Finally, additional requirements to the smoothing functions are discussed, and the empirical fact that the Gaussian and Cauchy distribution are most efficient for constructing the filled functions is explained. Indeed, the latter distributions are the only ones among all symmetric distributions boasting explicit formulas for their computation.

Acknowledgments

This work was supported by the SEP-CONACYT grant CB-2013-01-221676 from the Mexican Consejo Nacional de Ciencia y Tecnología (CONACYT). It was also partly supported by the US National Science Foundation grants HRD-0734825 and HRD-1242122 (Cyber-ShARE Center of Excellence), as well as by DUE-0926721, and by an award "UTEP and Prudential Actuarial Science Academy and Pipeline Initiative" from Prudential Foundation.

This work was performed when José Guadalupe Flores Muñiz visited the University of Texas at El Paso (UTEP) during his research stay August 04 – December 31, 2016.

The authors would also like to thank two anonymous reviewers whose comments have helped improving the presentation and the paper's structure.

References

1. Renpu, G.E.: A filled function method for finding a global minimizer of a function of several variables. Mathematical Programming, 46(1), 57-67 (1988).
2. Kalashnikov, V.V., Herrera Maldonado, R.C., Camacho-Vallejo, J.-F., Kalashnykova, N.I.: A heuristic algorithm solving bilevel toll optimization problems. The International Journal of Logistics Management, 27(1), 31-51 (2016).
3. Wu, Z.Y., Bai, F.S., Yang, Y.J., Mammadov, M.: A new auxiliary function method for general constrained global optimization. Optimization, 62(2), 193-210 (2013).
4. Wu, Z.Y., Mammadov, M., Bai, F.S., and Yang, Y.J.: A filled function method for nonlinear equations. Applied Mathematics and Computation, 189(2), 1196-1204 (2007).
5. Addis, B., Locatelli, M., and Schoen, F.: Local optima smoothing for global optimization. Optimization Methods and Software, 20(4-5), 417-437 (2005).
6. Johnson, N.L., Kotz, S., and Balakrishnan, N.: Continuous Univariate Distributions, Vol. 2, Wiley, New York (1995).
7. Klenke, A.: Probability Theory: A Comprehensive Course. Springer, Berlin-Heidelberg-New York (2014).
8. Sato, K.-I.: Lévy Processes and Infinitely Divisible Distributions. Cambridge University Press, Cambridge, UK (1999).
9. Steutel, F.W., and Van Harn, K.: Infinite Divisibility of Probability Distributions on the Real Line. Marcel Dekker, New York (2003).