Why Gaussian and Cauchy Functions Are Efficient in Filled Function Method: A Possible Explanation

José Guadalupe Flores Muñiz[1], Vyacheslav V. Kalashnikov[2,3], Nataliya Kalashnykova[1,4], and Vladik Kreinovich[5]
[1]Dept. Phys. and Math., Universidad Autónoma de Nuevo León
San Nicolás de los Garza, México, jose_guadalupe64@hotmail.com
nkalash2009@gmail.com
[2]Department of Systems and Industrial Engineering
Tecnológico de Monterrey ITESM, Campus Monterrey, Mexico,
kalash@itesm.mx
[3]Department of Experimental Economics, Central Economics and Math. Inst. (CEMI), Moscow, Russia
[4]Department of Computer Science, Sumy State University, Ukraine
[5]Department of Computer Science, University of Texas at El Paso
El Paso, Texas 79968, USA, vladik@utep.edu

1. Outline

- One of the main problems of optimization algorithms is that they end up in a local optimum.
- It is necessary to get out of the local optimum and eventually reach the global optimum.
- One of the promising methods to leave the local optimum is the filled function method.
- Empirically, the best smoothing functions in this method are the Gaussian and the Cauchy functions.
- In this talk, we provide a possible theoretical explanation for this empirical result.

2. Formulation of the Problem

- Many optimization techniques end up in a local optimum.
- So, to solve a global optimization problem, it is necessary to move out of the local minimum.
- Eventually, we should 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 Renpu's *filled function* method.
- In this method, once we reach a local optimum x^*, we optimize an auxiliary expression

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

for some $K, F, G,$ and σ.

3. Formulation of the Problem (cont-d)

- We use its optimum as a new first approximation to find the optimum of $f(x)$.
- Several different functions $K(x - x^*)$ have been proposed.
- It turns out that the most computationally efficient functions are the Gaussian and Cauchy functions

$$K(x) = \exp(-\|x\|^2), \quad K(x) = \frac{1}{1 + \|x\|^2}.$$

- Are these function indeed the most efficient?
- Or they are simply the most efficient among a few functions that have been tried?

4. What We Plan to Do

- In this talk, we formulate the above question as a precise mathematical problem.
- 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 results also shows that 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.

5. Need for Smoothing

- One of the known ways to eliminate local optima is to apply a *weighted smoothing*.
- In this method, we replace the original objective function $f(x)$ with a "smoothed" one

$$f^*(x) \stackrel{\text{def}}{=} \int K\left(\frac{x-x'}{\sigma}\right) \cdot f(x')\,dx', \text{ for some } K(x) \text{ and } \sigma.$$

- The weighting function is usually selected in such a way that $K(-x) = K(x)$ and $\int K(x)\,dx < +\infty$.
- The first condition comes from the fact that we have no reason to prefer different orientations of coordinates.
- The second condition is that for $f(x) = \text{const}$, smoothing should leads to a finite constant.

6. Need to Select an Appropriate Value σ

- When σ is too small, the smoothing only covers a very small neighborhood of each point x.
- The smoothed function $f^*(x)$ is close to the original objective function $f(x)$.
- So, we will still observe all the local optima.
- On the other hand, if σ is too large, the smoothed function $f^*(x)$ is too different from $f(x)$.
- So the optimum of the smoothed function may have nothing to do with the optimum of $f(x)$.
- So, for the smoothing method to work, it is important to select an appropriate value of σ.

7. We May Need Several Iterations to Find an Appropriate σ

- Our first estimate for σ may not be the best.
- If we have smoothed the function too much, then we need to "un-smooth" it, i.e., to select a smaller σ.
- If we have not smoothed the function enough, then we need to smooth it more, i.e., to select a larger σ.

8. Computationally Efficient Smoothing: Analysis

- Once we have smoothed the function too much, it is difficult to un-smooth 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 small σ:
  - to find each value $f^*(x)$ of the smoothed function,
  - we only need to consider values of $f(x')$ in a small vicinity of x.
- The larger σ, the larger this vicinity, so:
  - the more values $f(x')$ we need to take into account,
  - and thus the more computations we need.

9. Computationally Efficient Smoothing: Conclusion

- Let's assume that we have a smoothed function $f^*(x)$ corresponding to some value of σ.
- We need to compute a smoothed function $f^{**}(x)$ corresponding to a larger value $\sigma' > \sigma$.
- It is thus more computationally efficient *not* to apply smoothing with σ' to the original $f(x)$.
- Instead, we should apply a small additional smoothing to the smoothed function $f^*(x)$.

10. Resulting Requirement on the Smoothing Function $K(x)$

- For every σ' and σ, there should be an appropriate value $\Delta\sigma$.
- Then, after we get
$$f^*(x) = \int K\left(\frac{x - x'}{\sigma}\right) \cdot f(x')\,dx'.$$
- A smoothing with $\Delta\sigma$ should lead to the desired function
$$f^{**}(x) = \int K\left(\frac{x - x'}{\sigma'}\right) \cdot f(x')\,dx'.$$
- In other words, we need to make sure that for every objective function $f(x)$, we have
$$\int K\left(\frac{x - x'}{\sigma'}\right) \cdot f(x')\,dx' = \int K\left(\frac{x - x'}{\Delta\sigma}\right) \cdot f^*(x')\,dx'.$$

11. Analyzing The Above Requirement

- The above requirement leads to:

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

- The function $K(x)$ is non-negative, and its integral $\int K(x)\,dx$ is finite.
- Thus, after dividing $K(x)$ by the value of this integral, we get a probability density function (pdf):

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

- For this pdf:

$$\rho\left(\frac{x-x'}{\sigma'}\right) = \int \rho\left(\frac{x-x''}{\Delta\sigma}\right) \cdot \rho\left(\frac{x''-x'}{\sigma}\right)\,dx''.$$

12. Analysis (cont-d)

$$\rho\left(\frac{x-x'}{\sigma'}\right)=\int \rho\left(\frac{x-x''}{\Delta\sigma}\right)\cdot\rho\left(\frac{x''-x'}{\sigma}\right)\,dx''.$$

- Let X denote the random variable with the probability density function $\rho_X(x)$.
- The, the LHS is pdf of $\sigma'\cdot X$.
- The RHS is a pdf of the sum of two independent random variables $\sim\sigma\cdot X$ and $\sim\Delta\sigma\cdot X$.
- The requirement that the sum is similarly distributed means that $\rho(x)$ is *infinitely divisible*.
- Among symmetric infinitely divisible distributions, only Gaussian and Cauchy have analytical expressions:

$$\rho(x)\sim\exp(-x^2);\quad \rho(x)\sim\frac{1}{1+x^2}.$$

- All others requires complex algorithms to compute.

13. This Leads to the Desired Explanation (Almost)

- Computational efficiency implies that:
  - the smoothing $K(x)$
  - is proportional to the cdf of an infinitely divisible distribution.
- Among symmetric infinitely divisible distributions, only Gaussian and Cauchy have analytical expressions:

$$\rho(x) \sim \exp(-x^2); \quad \rho(x) \sim \frac{1}{1+x^2}.$$

- All others requires complex algorithms to compute.
- Thus, the most computationally efficient smoothing functions are the Gaussian and the Cauchy ones.

14. Final Step In Our Explanation: We Need to Approximate the Integral With a Sum

- The above arguments explain that:
  - instead of optimizing the original function $f(x)$,
  - we should optimize its smoothed version
$$\int K\left(\frac{x-x'}{\sigma}\right)\cdot f(x')\,dx'.$$
- 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 at different points.
- The simplest case is when we consider one or two points.
- Then, we get a linear combination of two values $f(x)$ with weights proportional to $K\left(\frac{x-x'}{\sigma}\right)$.

15. Conclusion

- We get a linear combination of two values $f(x)$ with weights proportional to $K\left(\frac{x-x'}{\sigma}\right)$.
- This is exactly what the filled function method does.
- Thus, we indeed get an explanation of the empirical fact – that:
  - the functions $K(x) \sim \exp(-x^2)$ and $K(x) \sim \frac{1}{1+x^2}$
  - are the most efficient in the filled function method.

16. Acknowledgments

- This work was supported by a grant from Mexico 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) and
    * DUE-0926721,
  - and by an award from Prudential Foundation.
- This work was performed when José Guadalupe Flores Muñiz visited the University of Texas at El Paso.