Efficiency of Gaussian and Cauchy functions in the Filled Function method

José Guadalupe Flores Muñiz,
Vyacheslav V. Kalashnikov,
Nataliya Kalashnykova,
and Vladik Kreinovich

27 de octubre de 2016

Outline I

- 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.

Formulation of the Problem I

- In the Renpu's **filled function** 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.

- We use its optimum as a new first approximation to find the optimum of $f(x)$.

Formulation of the Problem II

- Several different functions $K(x - x^*)$ have been proposed, but 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?

Need for Smoothing I

- 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

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

for some $K(x)$ and σ.

Need for Smoothing II

- 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.

Need to Select an Appropriate Value σ I

- 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.

We May Need Several Iterations to Find an Appropriate σ I

- 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.

Computationally Efficient Smoothing: Analysis I

- 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.

Computationally Efficient Smoothing: Conclusion I

- 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 $f^*(x)$.

Resulting Requirement on the Smoothing Function $K(x)$ I

- 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

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

Analyzing The Above Requirement I

- 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):

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

Analyzing The Above Requirement II

- For this pdf:

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

- Let X denote the random variable with the probability density function $\rho_X(x)$.
- Then, 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*.

This Leads to the Desired Explanation (Almost) I

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

$$\rho(x) \sim \exp(-x^2);$$

$$\rho(x) \sim \frac{1}{1+x^2}.$$

All others requires complex algorithms to compute.

- Thus, the most computationally efficient smoothing functions are the Gaussian and the Cauchy ones.

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

- 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.

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

- 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 $K\left(\dfrac{x-x'}{\sigma}\right)$.
- For example, the function:

$$Q_{p,t^*}(t) := -e^{-\|t-t^*\|^2} g_{\frac{2}{5}u(t^*)}(u(t)) - \rho s_{\frac{2}{5}u(t^*)}(u(t)),$$

where $u_b(v)$ and $s_b(v)$ are cubic splines, is used in [3].

Conclusion I

- 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.

Acknowledgments I

- 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.

Referencias I

[1] B. Addis, M. Locatelli, and F. Schoen, "Local optima smoothing for global optimization", *Optimization Methods and Software*, 2005, Vol. 20, No. 4–5, pp. 417–437.

[2] N. L. Johnson, S. Kotz, and N. Balakrishnan, *Continuous Univariate Distributions*, Vol. 2, Wiley, New York, 1995.

[3] V. V. Kalashnikov, R. C. Herrera Maldonado, and J.-F. Camacho-Vallejo, "A heuristic algorithm solving bilevel toll optimization problem", *The International Journal of Logistics Management*, 2016, Vol. 27, No. 1, pp. 31–51.

[4] A. Klenke, *Probability Theory: A Comprehensive Course*, Springer, Berlin, Hiedelberg, New York, 2014.

Referencias II

[5] G. E. Renpu, "A filled function method for finding a global minimizer of a function of several variables", *Mathematical Programming*, 1988, Vol. 46, No. 1, pp. 57–67.

[6] K.-I. Sato, *Lévy Processes and Infinitely Divisible Distributions*, Cambridge University Press, Cambridge, UK, 1999.

[7] F. W. Steutel and K. Van Harn, *Infinite Divisibility of Probability Distributions on the Real Line*, Marcel Dekker, New York, 2003.

[8] Z. Y. Wu, F. S. Bai, Y. J. Yang, and M. Mammadov, "A new auxiliary function method for general constrained global optimization", *Optimization*, 2013, Vol. 62, No. 2, pp. 193–210.

Referencias III

[9] Z. Y. Wu, M. Mammadov, F. S. Bai, and Y. J. Yang, "A filled function method for nonlinear equations", *Applied Mathematics and Computation*, 2007, Vol. 189, No. 2, pp. 1196–1204.