 doi:10.20944/preprints202206.0361.v1

Article

A Short Study on Minima Distribution

Loc Nguyen

Loc Nguyen's Academic Network, Vietnam
ng_phloc@yahoo.com

Abstract: Global optimization is an imperative development of local optimization because there are many problems in artificial intelligence and machine learning requires highly acute solutions over entire domain. There are many methods to resolve the global optimization, which can be classified into three groups such as analytic methods (purely mathematical methods), probabilistic methods, and heuristic methods. Especially, heuristic methods like particle swarm optimization and ant bee colony attract researchers because their effective and practical techniques which are easy to be implemented by computer programming languages. However, these heuristic methods are lacking in theoretical mathematical fundamental. Fortunately, minima distribution establishes a strict mathematical relationship between optimized target function and its global minima. In this research, I try to study minima distribution and apply it into explaining convergence and convergence speed of optimization algorithms. Especially, weak conditions of convergence and monotonicity within minima distribution are drawn so as to be appropriate to practical optimization methods.

Keywords: global optimization; minima distribution; particle swarm optimization; PSO

1. Introduction

Given target function $f(x)$ concerned in a compact set $X \subset R^n$ where R^n is the n-dimension real field, the optimization problem is how to find out its minimizer x^* as follows:

$$x^* = \underset{x \in X}{\operatorname{argmin}} f(x)$$

Let $f^* = f(x^*)$ and let X^* is the set of all x^*, the optimization problem is started with local optimization methods such as gradient descent and Newton-Raphson given condition that f(x) is convex on entire X. As a progressive development, global optimization concerns arbitrary $f(x)$, which can be classified into three groups such as analytic methods, probabilistic methods, and heuristic methods. Some purely mathematical methods assert optimized solutions but there is a lack of information about convergence process of $f(x)$ to f^* in optimization algorithms. Fortunately, Luo (Luo, 2019) established the strict relationship between the target function $f(x)$ and its minima f^* when Luo defined the minima distribution $m^{(k)}(x)$ as follows (Luo, 2019, p. 5):

$$m^{(k)}(x) = \frac{\tau^k(x)}{\int_X \tau^k(t)\mathrm{d}t}$$

Function $\tau(x)$ is defined as $\tau(x) = \rho(f(x))$ where $\rho(y)$ is positive and monotonically decreasing for y such that $y = f(x)$ where x belongs to X. Luo proved that the integral of the association of $f(x)$ and $m^{(k)}(x)$ which can be understood that mean of $f(x)$ within minima distribution will approach f^*.

$$\lim_{k \to \infty} \int_X f(x) m^{(k)}(x)\mathrm{d}x = f^*$$

The equation above specifies the convergence property of optimization algorithms. Luo also proposed the monotonicity property as the following ordered sequences for all real k and real $\Delta k > 0$.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 June 2022 doi:10.20944/preprints202206.0361.v1

$$\int_X f(x)m^{(k)}(x)\mathrm{d}x \geq \int_X f(x)m^{(k+\Delta k)}(x)\mathrm{d}x \geq f^*$$

Based on monotonicity property, monotonic shrinkage is derived as follows:

$$X \supseteq D_f^{(k)} \supseteq D_f^{(k+\Delta k)} \supseteq X^*$$

Where $D_f^{(k)}$ is called significant set of $m^{(k)}(x)$.

$$D_f^{(k)} = \left\{x \in X : f(x) \leq \int_X f(t)m^{(k)}(t)\mathrm{d}t\right\}$$

Luo's convergence conditions are strict because $\tau(x)$ is defined based $\rho(f(x))$ and in turn, $\rho(f(x))$ is monotonically decreasing. In some practical cases, convergence property can be achieved without concerning $\rho(f(x))$. Therefore, in this research, I draw weaker conditions for convergence property and monotonicity property. Section 2 draws weak conditions of convergence and monotonicity whereas section 3 continues to mention convergence speed. Section 4 describes an experiment on convergence speed associated with particle swarm optimization (PSO) algorithm. Section 5 is the conclusions.

2. Weak conditions of convergence and monotonicity

Two most important properties of minima distribution are convergence and monotonicity. Stability and shrinkage are derived from the two properties. As a summary, Luo (Luo, 2019, p. 5) proposed the following condition for satisfying convergence and monotonicity of minima distribution: "Positive function $\tau(x)$ is defined based on function ρ such that $\tau(x) = \rho(f(x))$ where ρ is monotonically decreasing and ρ is positive, for instance, $\rho(y) > 0$ for all x belonging to the domain X such that $y = f(x)$". However, it is not necessary to strictly define $\tau(x)$ based on ρ and $f(x)$ for convergence. Following the proofs of theorem 1 by Luo (Luo, 2019, p. 7), a *weak convergence condition* is drawn with two following requirements as follows:

- Function $\tau(x)$ is positive for all x belonging to domain X. This requirement of positiveness can be understood that $\tau(x)$ is only necessary to be nonzero because $\tau(x)$ is positive if $-\tau(x)$ is negative and nonzero, and vice versa.
- Given any x which is not a minimizer of $f(x)$, there always exists an open set having nonzero Lebesgue measure such that $\tau(t) > \tau(x)$ for all t belonging to this open set. This requirement asserts that $\lim_{k\to\infty} m^{(k)}(x) = 0, \forall x \notin X^*$ given positive $\tau(x)$.

The convergence property with this condition is easily proved by similar proofs of theorem 1 (vi, vii) and theorem 2 (Luo, 2019, pp. 7-8). Of course, if $\tau(x)$ is defined as $\tau(x) = \rho(f(x))$ where ρ is monotonically decreasing and ρ is positive, the weak convergence condition is satisfied. This condition is useful in practical cases that $\tau(x)$ is defined as a non-constant, positive, and differentiable function such that $\tau(x) < f(x), \forall x \in X$ or $\tau(x)$ has no actual minimizer with note that some infima may not be actual minima, for example, $\tau(x) = \exp(-x)$, $\tau(x) = \exp(-x^2)$, $\tau(x) = \exp(x)$, etc. Therefore, if positive function $\tau(x)$ is continuous, the weak convergence condition only requires that any x which is not a minimizer of $f(x)$ is not a minimizer of $\tau(x)$ too.

Within the weak convergence condition, theorem 3 for nonnegativity (Luo, 2019, p. 8) is still obtained because $\tau(x)$ is positive.

$$\int_X f(x)m^{(k)}(x)\mathrm{d}x \geq f^*$$

Monotonicity can be achieved if the first order derivative of $\int_X f(x)m^{(k)}(x)\mathrm{d}x$ with regard to k is greater than or equal to 0. Suppose $m^{(k)}(x)$ is differentiable, we have:

$$\frac{\mathrm{d}}{\mathrm{d}k}\int_X f(x)m^{(k)}(x)\mathrm{d}x = \int_X f(x)\frac{\mathrm{d}m^{(k)}(x)}{\mathrm{d}k}\mathrm{d}x$$

Suppose positive function $\tau^k(x)$ is differentiable with regard to k and its derivatives with regard to k is formulated as follows:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 June 2022 doi:10.20944/preprints202206.0361.v1

$$\frac{\mathrm{d}\tau^k(x)}{\mathrm{d}k} = \tau^k(x)g(x)$$

In other words, $g(x)$ is specified as follows:

$$g(x) = \frac{\mathrm{d}\tau^k(x)/\mathrm{d}k}{\tau^k(x)} \tag{1.1}$$

Following the proof of theorem 1 (v) by Luo (Luo, 2019, p. 7), we have:

$$\frac{\mathrm{d}m^{(k)}(x)}{\mathrm{d}k} = m^{(k)}(x)\left(g(x) - E^{(k)}(g)\right) \tag{1.2}$$

Where:

$$E^{(k)}(g) = \int_X g(x)m^{(k)}(x)\mathrm{d}x$$

Therefore, we obtain (Luo, 2019, p. 9):

$$\frac{\mathrm{d}}{\mathrm{d}k}\int_X f(x)m^{(k)}(x)\mathrm{d}x = E^{(k)}(fg) - \left(E^{(k)}(f)\right)\left(E^{(k)}(g)\right) \tag{1.3}$$

Where:

$$E^{(k)}(fg) = \int_X f(x)g(x)m^{(k)}(x)\mathrm{d}x$$

According to Gurland's inequality (Luo, 2019, p. 6), if $f(x)$ and $g(x)$ are inversely proportional, for instance, $f(x)$ is nonincreasing and $g(x)$ is nondecreasing in every open set, and vice versa then, $E^{(k)}(yg) \le (E^{(k)}(y))(E^{(k)}(g))$. As a result, a *weak monotonicity condition* is drawn that function $\tau(x)$ is defined such that $f(x)$ and $g(x)$ are inversely proportional where $g(x)$ is specified by equation 1.1. Note that the weak monotonicity condition follows the weak convergence condition.

Because the target function $f(x)$ is too complicated to determine whether $f(x)$ and $g(x)$ are inversely proportional, we concern the case that $g(x)$ can be specified as $g(x) = h(y)$ where $y = f(x)$. In other words, if g is function of $f(x)$, we have:

$$\frac{\mathrm{d}}{\mathrm{d}k}\int_X f(x)m^{(k)}(x)\mathrm{d}x = E^{(k)}(yh) - \left(E^{(k)}(y)\right)\left(E^{(k)}(h)\right) \tag{1.4}$$

Where:

$$E^{(k)}(y) = \int_X ym^{(k)}(x)\mathrm{d}x\,, E^{(k)}(h) = \int_X h(y)m^{(k)}(x)\mathrm{d}x\,, E^{(k)}(yh) = \int_X yh(y)m^{(k)}(x)\mathrm{d}x$$

Given y and $h(y)$ where y is increasing, the weak monotonicity condition is simplified that function $\tau(x)$ following the weak convergence condition is defined such that $h(y)$ is nonincreasing function with regard to y.

As usual, $\tau^k(x)$ is the k^{th}-power function of $\tau(x)$ as $\tau^k(x) = (\tau(x))^k$. Let k denote knowledge amount of any optimization algorithms, now $\tau^k(x)$ is generalized as $\tau^k(x) = w_k(\tau(x))$ where $w_k(\tau)$ is called knowledge function for τ which has two properties as follows:

$$w_k(\tau_1\tau_2) = w_k(\tau_1)w_k(\tau_2)$$
$$\lim_{k\to\infty} w_k\big(\tau(x)\big) = 0 \text{ if } \tau(x) < 1$$

Note, the knowledge function $w_k(\tau)$ is function of k, which is determined based on the value $\tau(x)$. The first property implies that $w_k(\tau)$ is a homomorphism with regard to multiplication. In trivial cases, we have $w_k(\tau) = (\tau(x))^k$, $w_k(\tau) = \big(\tau(x)\big)^{\exp(k)}$, etc. With the two properties of $w_k(\tau)$, it is easy to assert the weak convergence condition by the proofs of theorem 1 (vi, vii) by Luo (Luo, 2019, p. 7). Based on two weak conditions of convergence and monotonicity, convergence speed is proposed in the next section.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 June 2022 doi:10.20944/preprints202206.0361.v1

3. Convergence speed

For all real k and real $\Delta k > 0$, two successive integrals within the weak convergence condition are determined as follows:

$$\int_X f(x)m^{(k)}(x)\mathrm{d}x \geq f^*, \int_X f(x)m^{(k+\Delta k)}(x)\mathrm{d}x \geq f^*$$

Convergence speed is defined as the absolute differential of two successive integrals as follows:

$$c_\tau = \left| \lim_{\Delta k \to 0} \frac{\int_X f(x)m^{(k)}(x)\mathrm{d}x - \int_X f(x)m^{(k+\Delta k)}(x)\mathrm{d}x}{\Delta k} \right|$$

It is easy to recognize that the convergence speed is the absolute value of the first order derivative of $\int_X f(x)m^{(k)}(x)\mathrm{d}x$ specified by equation 1.3.

$$c_\tau = \left| \frac{\mathrm{d}}{\mathrm{d}k} \int_X f(x)m^{(k)}(x)\mathrm{d}x \right| = \left| E^{(k)}(fg) - \left(E^{(k)}(f)\right)\left(E^{(k)}(g)\right) \right| \tag{2.1}$$

Note, $g(x)$ is defined by equation 1.1. The magnitude of c_τ is proportional to the difference between $f(x)$ and $g(x)$ which is represented by the following derivative:

$$\frac{\mathrm{d}}{\mathrm{d}x}\big(f(x) - g(x)\big) = f'(x) - g'(x)$$

Here we can ignore the target function $f(x)$ because it is not parameter for optimization algorithms. Therefore, let Q_τ be the metric that measures the convergence speed, which is defined as follows:

$$Q_\tau = |g'(x)| \tag{2.2}$$

Actually, Q_τ is absolutely slope of function g. The steeper Q_τ is, the larger c_τ is, which in turn, the faster convergence speed. For example, if the definition of function τ is associated with normal distribution given mean 0 and variance 1, it becomes:

$$\tau_N(x) = \exp\left(-\frac{x^2}{2}\right)$$

Therefore,

$$\frac{\mathrm{d}\tau_N^k(x)}{\mathrm{d}k} = \left(-\frac{x^2}{2}\right)\exp\left(-\frac{x^2}{2}\right)$$

$$Q_{\tau_N} = \left|\frac{\mathrm{d}g_N(x)}{\mathrm{d}x}\right| = \left|\frac{\mathrm{d}}{\mathrm{d}x}\left(-\frac{x^2}{2}\right)\right| = |x|$$

If the definition of function τ is associated with Gumbel distribution for extreme value given location parameter 0 and scale parameter 1, it becomes:

$$\tau_G(x) = \exp\Big(-\big(x + \exp(-x)\big)\Big)$$

Therefore,

$$\frac{\mathrm{d}\tau_G^k(x)}{\mathrm{d}k} = \Big(-\big(x + \exp(-x)\big)\Big)\exp\Big(-\big(x + \exp(-x)\big)\Big)$$

$$Q_{\tau_G} = \left|\frac{\mathrm{d}g_G(x)}{\mathrm{d}x}\right| = \left|\frac{\mathrm{d}}{\mathrm{d}x}\Big(-\big(x + \exp(-x)\big)\Big)\right| = |\exp(-x) - 1|$$

Obviously, convergence speed of $\tau_G(x)$ is faster than $\tau_N(x)$ because $g_G(x)$ is steeper than $g_N(x)$. Given normal distribution with mean μ and variance σ^2 then, $\tau_N(x)$ is redefined as follows:

$$\tau_N(x) = \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

$$Q_{\tau_N} = \frac{|x-\mu|}{\sigma^2}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 June 2022 doi:10.20944/preprints202206.0361.v1

Variance σ^2 reflects the slope of $\tau_N(x)$. Therefore, the smaller variance σ^2 is, the steeper function $\tau_N(y)$ is. The next section describes an experiment with varying variance σ^2.

4. Experiment with PSO algorithm

In this research, convergence speed of minima distribution is concerned, which in turn derives the conclusion that function $g(x)$ extracted from the derivative of $\tau^k(x)$ should be steeper in order to improve the convergence speed. In particle swarm optimization (PSO) algorithm, movement of particles obeys normal distribution. Therefore, this section describes an experiment by varying variance of such normal distribution with note that $\tau(x)$ here is associated with the PSO normal distribution. The iterative PSO algorithm (Poli, Kennedy, & Blackwell, 2007) was developed by James Kennedy and Russell C. Eberhart updates velocities and positions of particles at each iteration as follows:

$$v_i = v_i + U(0, \phi_1) \otimes (p_i - x_i) + U(0, \phi_2) \otimes (p_g - x_i)$$

$$x_i = x_i + v_i$$

Note, x_i and p_i be current position and best position of particle i whereas p_g be the global best position over entire swarm of particles. Functions $U(0, \phi_1)$ and $U(0, \phi_2)$ generate random vectors whose elements are random numbers in the ranges $[0, \phi_1]$ and $[0, \phi_2]$, respectively, where ϕ_1 and ϕ_2 are parameters. The operator $\otimes$ denotes component-wise multiplication of two points (Poli, Kennedy, & Blackwell, 2007, p. 3). Kennedy and Eberhart (Poli, Kennedy, & Blackwell, 2007, p. 13), (Pan, Hu, Eberhart, & Chen, 2008, p. 3), (al-Rifaie & Blackwell, 2012, p. 51) asserted that, given p_i and p_g, each x_i follows normal distribution with mean $(p_i+p_g)/2$ and variance $(p_i-p_g)^2$. In this experiment, I put random movements into updating positions of particles, following normal distribution. Moreover, I modify this variance as $\sigma^2 = r^2(p_i-p_g)^2$ with $r = 1, 2, 3, 4, 5$ to monitor the convergence speed of PSO. Note, r is called convergence rate. The target function is (Sharma, Chhamunya, Gupta, Sharma, & Bansal, 2015, p. 24):

$$f(\boldsymbol{x} = (x_1, x_2)^T) = -\cos(x_1)\cos(x_2)\exp(-(x_1 - \pi)^2 - (x_2 - \pi)^2)$$

Note, super script "^{T}" denotes vector transposition operator. Given true minimizer $\boldsymbol{x}^* = (3.1416, 3.1416)^T$ with true minimum $f^* = -1$, Table 1 shows the experimental results with minimum biases and iteration counts over convergence rates.

Table 1. Experimental results with increased convergence rates.

	Minimum	Minimum bias	Minimizer	Iteration count
r=1	–0.9996	0.0004	$(3.1275, 3.1327)^T$	18
r=2	–0.9960	0.0040	$(3.1911, 3.1270)^T$	16
r=3	–0.9982	0.0018	$(3.1231, 3.1127)^T$	14
r=4	–0.9992	0.0008	$(3.1545, 3.1607)^T$	15
r=5	–0.9999	0.0001	$(3.1356, 3.1465)^T$	23

From Table 1, convergence speed is increased from r=1 to r=4 because the iteration count is decreased. Especially, at $r = 5$, convergence speed is not increased as expectation but the best converged value (–0.9999) is obtained with smallest bias (0.0001). Therefore, it is no doubt that convergence speed is improved by increasing the variance which makes function $g(x)$ steeper.

5. Conclusion

Some optimization algorithms like PSO take advantages of distribution of x instead of taking advantages of $f(x)$. In other words, they define implicitly $\tau(x)$ as function of x instead of function of $f(x)$ like $\tau(x) = \rho(f(x))$. Therefore, it is reasonable to assert their convergence by the weak convergence condition. The convergence speed also depends on acuteness of the knowledge function $w_k(\tau)$, besides the slope Q_τ of $\tau(x)$. It is inferred that $w_k(\tau)$ implies the knowledge amount of given optimization algorithm after each iteration.

 doi:10.20944/preprints202206.0361.v1

For PSO, heuristic movement of particles after each iteration reflects how fast the power function $w_k(\tau)$ approaches when k approaches positive infinity. In the future trend, I will research minima distribution with Bayesian optimization because Bayesian optimization takes full advantages of prior information which is the knowledge amount associated with the knowledge function.

References

al-Rifaie, M. M., & Blackwell, T. (2012). Bare Bones Particle Swarms with Jumps. In M. Dorigo, M. Birattari, C. Blum, A. L. Christensen, A. P. Engelbrecht, R. Groß, & T. Stützle (Ed.), *International Conference on Swarm Intelligence. Lecture Notes in Computer Science 7461*, pp. 49-60. Brussels: Springer Berlin. doi:10.1007/978-3-642-32650-9_5

Luo, X. (2019, May 24). Minima distribution for global optimization. *arXiv preprint*. doi:10.48550/arXiv.1812.03457

Pan, F., Hu, X., Eberhart, R., & Chen, Y. (2008, September 21). An Analysis of Bare Bones Particle Swarm. *IEEE Swarm Intelligence Symposium 2008 (SIS 2008)* (pp. 1-5). St. Louis, MO, US: IEEE. doi:10.1109/SIS.2008.4668301

Poli, R., Kennedy, J., & Blackwell, T. (2007, June). Particle swarm optimization. (M. Dorigo, Ed.) *Swarm Intelligence, 1*(1), 33-57. doi:10.1007/s11721-007-0002-0

Sharma, K., Chhamunya, V., Gupta, P. C., Sharma, H., & Bansal, J. C. (2015, September). Fitness based Particle Swarm Optimization. (A. K. Verma, P. K. Kapur, & U. Kumar, Eds.) *International Journal of System Assurance Engineering and Management, 6*(3), 319-329. doi:10.1007/s13198-015-0372-4