Particle Swarm optimization Algorithm and Convergence Analysis

Mohamed Keteb - Francesca Crucinio

Research Internship at King's College

7 janvier 2025

Sommaire

Introduction

Goal : How to minimize a function which is not convex, a such methodology or algorithm is needed in machine learning where the objectives (loss function) are not convex (Neural networks).
Given a compact set $K \subset \mathbb{R}^n$ and $f \in \mathcal{C}(K, \mathbb{R})$,

$$x^\star \in \arg\min_{x \in K} f(x) \tag{$\mathcal{P}$}$$

General statement on PSO

The PSO methodology is a dynamic approach to optimization that can be viewed, as the name suggests, as a swarm of particles interacting with each other. The equations that describes PSO are as follows,
$\forall i \in \{1, \ldots, N\}, \forall t \in \{0, \ldots, T-1\}$,

$$v_{t+1}^i = w v_t^i + c_1 r_1 (p_t^i - x_t^i) + c_2 r_2 (g_t - x_t^i) \tag{1}$$

$$x_{t+1}^i = x_t^i + v_{t+1}^i \tag{2}$$

$$p_{t+1}^i := \arg \min_{u \in \{x_{0:t+1}^i\}} f(u) = \arg \min_{u \in \{x_{t+1}^i, p_t^i\}} f(u) \tag{3}$$

$$g_{t+1} := \arg \min_{u \in \{x_{0:t+1}^{1:N}\}} f(u) = \arg \min_{u \in \{x_{t+1}^{1:N}, g_t\}} f(u) = \arg \min_{u \in \{p_{t+1}^{1:N}, g_t\}} f(u) \tag{4}$$

c_1 individual sensibility and c_2 social sensibility, r_1 and r_2 are uniformly distributed in the interval $(0, 1)$ it is also common to use a velocity clamping setting to prevent partcile from going outside the seach space.

Implementing PSO

We will be using the following test functions that are highly non-convex,

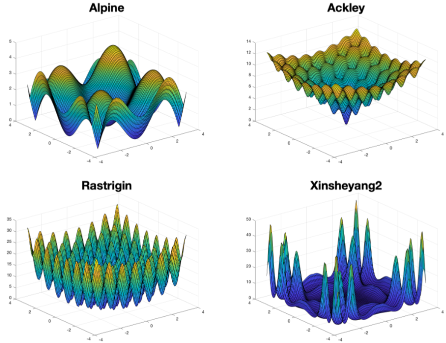
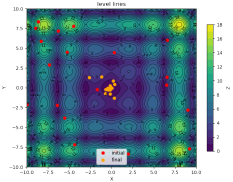
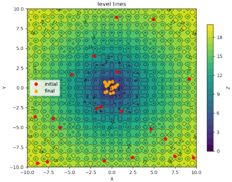


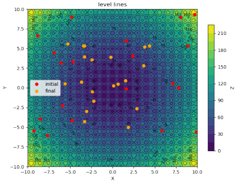
Figure: Non convex test functions in 2 dimensions

Implementing PSO

For $K = [-10, 10]^2$ we have the following results,

[Alpine] 

[Ackley] 

[Rastrigin] 

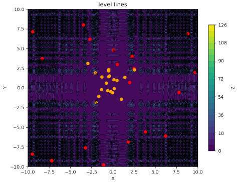
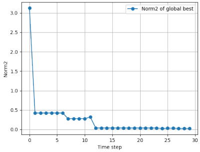
[Xinsheyang2] 

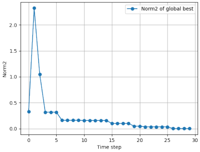
Figure: Evolution of 20 particles : initially in red, and in yellow after 30 iterations.

We can check the convergence imperially,

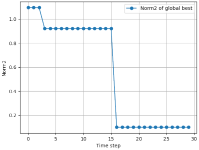
[Alpine]



[Ackley]



[Rastrigin]



[Xinsheyang2]

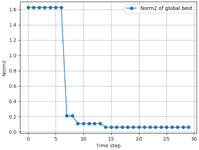


Figure: L^2 norm of the global best across time (30 iterations).

Markov chain modelization of PSO

To begin with we can rewrite the equations to have a compact equation as follows,$\forall i \in \{1, \ldots, N\}$,

$$x_{t+1}^i = x_t^i + w(x_t^i - x_{t-1}^i) + c_1 r_1 (p_t^i - x_t^i) + c_2 r_2 (g_t - x_t^i) \tag{5}$$

The first part has been to notice that, if we define $\xi_t^i := (x_t^i, x_{t-1}^i, p_t^i, g_t) \in K^4$ with a state stape denoted by $\mathcal{S}$ then if we aggregate all the particle,

$$\zeta_t = (\xi_t^1, \ldots, \xi_t^N) = (x_t^{1:N}, x_{t-1}^{1:N}, p_t^{1:N}, g_t^{1:N}) \in K^{4N} \tag{6}$$

with a state spase denoted by Θ is a Homogeneous Markov Chain.

A definition of convergence

A first insight to convergence (Proposition 3 in the report)
The sequence $(F(\zeta_t))$ is a almost surely convergent but we do not have any information about the convergence of the swarm sequence, A first approach has been studied in Qian and Li 2018.
(Definition 7 in the report) $\forall \epsilon > 0$, we say that the $(g_t)_{t \geq 0}$ is ϵ-convergent if and only if,

$$\mathbb{P}(g_t \in B_\epsilon) \xrightarrow[t \to \infty]{} 1 \tag{7}$$

With $B_\epsilon := \{x \in K, f(x^\star) \leq f(x) \leq f(x^\star) + \epsilon\}$.
If ϵ-convergence holds for all $\epsilon > 0$ then we have convergence.

An interesting quantity for convergence

An interesting quantity has been introduced, Given a subset $B \subset K$, $\forall t \geq 0, \forall N \geq 2$,

$$\Phi_B(\zeta_t) = \sum_{i=1}^{N} (\mathbf{1}_B(x_t^i) + \mathbf{1}_B(p_t^i) + \mathbf{1}_B(g_t))$$

The link between this quantity and the convergence is given by the following results, Given $\epsilon > 0, t \geq 0$,

$$(\Phi_{B_\epsilon}(\zeta_t) \neq 0) = (g_t \in B_\epsilon) \tag{8}$$

We can also easily prove that, Given $\epsilon > 0$,

$$\forall s > t \geq 0, \ (\Phi_{B_\epsilon}(\zeta_t) \neq 0) \subset (\Phi_{B_\epsilon}(\zeta_s) \neq 0) \tag{9}$$

A first result

In terms of convergence, it shows that if a certain time the global best is optimal, then it remains optimal B_ϵ is kind of an absorbent state of the global best process, (Theorem 1 in the report)

We can then state the first trivial result, Given $\omega \in \Omega$ that defines a trajectory $(\zeta_t(\omega))_{t\geq 0}$ and $\epsilon > 0$, if there exists $\tau \geq 0$ such that $\Phi_{B_\epsilon}(\zeta_\tau(\omega)) \neq 0$ then

$$\forall t > \tau,\ g_t(\omega) \in B_\epsilon \tag{10}$$

Leveraging the support

We will be interesting in the feasible region which is the support of the conditional measure in this set up, namely at time t we would like to know what are the feasible region by the swarm at time $t+1$, recall that,

$$\mathbb{P}(dx_{t+1}^i \mid \zeta_t) = [x_t^i + w(x_t^i - x_{t-1}^i)] \oplus c_1(p_t^i - x_t^i) \cdot U_1 \oplus c_2(g_t - x_t^i) \cdot U_2$$

$\forall \zeta \in \theta$,

$$\mathcal{R}_{t+1}(\zeta) := \bigcup_{i=1}^{N} Supp(\mathbb{P}(dx_{t+1}^i \mid \zeta_t = \zeta)) = \bigcup_{i=1}^{N} \prod_{j=1}^{n} Supp(\mathbb{P}(dx_{t+1}^i(j) \mid \zeta_t = \zeta)) \tag{11}$$

So the support of the measure $\mathbb{P}(dx_{t+1}^i(j) \mid \zeta_t = \zeta)$ could be determined as the support of the sum of two random variables uniformly distributed.

A result that uses the feasible region

Given $\epsilon > 0$ and $\zeta \in \Theta$, if there exists $t > 0$ such that $x^\star \in \mathcal{R}_{t+1}(\zeta)$ then there exists $\rho_\epsilon > 0$,

$$\mathbb{P}(\Phi_{B_\epsilon}(\zeta_{t+1}) \neq 0 \,|\, \zeta_t = \zeta) \geq \rho_\epsilon > 0 \tag{12}$$

We can algo give the expression of ρ_ϵ,

$$\rho_\epsilon = (\min\{\frac{1}{8c_1c_2m^2}(\frac{\delta}{3})^2, \frac{c_1 \wedge c_2}{4c_1c_2m}\frac{\delta}{3}, \frac{1}{8c_1c_2m^2}(\frac{\delta}{2})^2, \frac{\delta}{2c_1m}, \frac{\delta}{2c_2m}, 1\})^n$$

with $m := \max_{1 \leq i \leq n}\{|x_i^\star - \delta|, |x_i^\star + \delta|\}$ **Curse of dimensionality**

A result that is sterile in application but rich in instructional value

(Theorem 3 in the report) Given $\epsilon > 0$, if there exists an increasing mapping $\alpha : \mathbb{N} \longrightarrow \mathbb{N}$ such as,

$$\forall \zeta \in \Theta, \forall t \geq 0, \; \mathbb{P}(\Phi_{B_\epsilon}(\zeta_{\alpha(t)+1}) \neq 0 \,|\, \zeta_{\alpha(t)} = \zeta) \geq \rho_\epsilon > 0 \tag{13}$$

Remarks on the proof

I got inspired by this proof to state the following result that doesn't use any increasing map but another stronger assumption, (Theorem 4 in the report)

Given $\epsilon > 0$, if for all $\zeta \in \Delta$ there exists $t \geq 1$ such that $x^\star \in \mathcal{R}_{t+1}(\zeta)$ then we have ϵ-convergence.

Is the feasible region a good direction ?

There is a results that shows that the feasible region is not negligible in the analysis, in fact,

Given the Markov chain, $(\zeta_t)_{t\geq 0}$ that describes PSO, and $w \in \Omega$ that gives a trajectory $(\zeta_t(w))_{t\geq 0}$,if $x^\star \notin \overline{\bigcup_{t\geq 0} \mathcal{R}_{t+1}(\zeta_t(w))}$ then there exists $\epsilon > 0$ such as,

$$\forall t \geq 0,\ g_t(w) \notin B_\epsilon$$

A second point of view on convergence

This approach defines an optimal set for a particle and derives the optimal set of the swarm as a union, Let us denote Γ the optimal set of the swarm (process (ζ_t)),

$$\zeta_t = (\xi_t^1, \ldots, \xi_t^N) \in \Gamma \iff \exists i \in \{1, \ldots, N\},\ \xi_t^i \in \mathcal{M}$$

with the optimal set $\mathcal{M}$ defined as follows,

$$\mathcal{M} := (K \times K \times \{x^\star\} \times \{x^\star\}) \cap \mathcal{S}$$

We can easily show that Γ is an absorbing state ($\zeta \in \Gamma$, $\mathbb{P}(\zeta_{t+1} \in \Gamma \,|\, \zeta_t = \zeta) = 1$)

A definition of convergence base on Γ

This definition has been proposed in the article (Gang Xu 2017), The Markov chain, $(\zeta_t)_{t\geq 0}$ that describes PSO is convergent if and only if

$$\mathbb{P}(\zeta_t \in \Gamma) \xrightarrow[t\to\infty]{} 1$$

An inductive proof in the report show that by using the fact that $x^i_{t+1} \,|\, \zeta_t$ has a density we can show that, $\forall t \geq 0$, $\mathbb{P}(\zeta_t \in \Gamma) = 0$

- Empirically PSO could be a good method to use for non-convex function but it could suffer from the curse of dimensionality
- The feasible region is a good way to tackle convergence and our last result shift the problem to the initialization and then to have some information on the area that contains the minimum
- The second approach of convergence is way too strong
- We can add some assumption on the functions Ackley seems to be a quasi-convex (conjecture)
- We planned with Francesca Crucinio to work on a proximal variant of PSO

.