PARTICLE SWARM OPTIMIZATION ALGORITHM AND CONVERGENCE ANALYSIS

Mohamed Keteb
ENSAE, IP Paris
mohamed.keteb@ensae.fr

Francesca Romana Crucinio
King's College London
francesca_romana.crucinio@kcl.ac.uk

ABSTRACT

optimization algorithms are essential components of machine learning. A well-known algorithm in this field is the gradient descent algorithm, which is used to minimize a differentiable loss function. However, it requires the objective function to be differentiable, which is not always the case. In this paper, we focus on a meta-heuristic algorithm called Particle Swarm optimization (PSO), introduced by Kennedy and Eberhart in the paper [1]. PSO is widely used because it only requires the objective function to be continuous, not differentiable, not even convex. Even if the convergence of an algorithm is evident experimentally, it is crucial to provide a formal proof of convergence. The convergence analysis of PSO is still not fully developed. In this paper, we present an analysis of the convergence of PSO using Markov chain.

Keywords Simulation, optimization, Interacting Particles, Markov Chain, Convergence

KING'S College LONDON

Contents

1 Introduction

Optimization is a crucial field that aims to find the global minimum or, equivalently, the global maximum of a function with certain regularity properties. The well-known gradient descent algorithm, along with its stochastic variant, has significantly advanced the performance of machine learning. Thanks to its efficiency, gradient descent is widely used in deep learning to tune neural network parameters by minimizing a loss function. Despite the irregularities often present in these minimization problems, stochastic gradient descent (SGD) achieves remarkable results. In optimization, we commonly assume properties such as continuity, differentiability, convexity (or strong convexity), and smoothness. However, in practice, objective functions are often barely differentiable, which presents challenges for optimization methods like gradient descent.

Another interesting way to address the lack of regularity in objective functions in practice is by using particle-based algorithms. These algorithms are used in optimization and for sampling from a target probability distribution, which is useful for tasks like data generation or calculating the normalizing constant of posterior distributions in the Bayesian setting. Particle-based algorithms fall under a broader category based on metaheuristics, which are designed to find near-optimal solutions to complex optimization problems. These methods are particularly effective when traditional approaches struggle due to non-linearity, high dimensionality, or the overall complexity of the solution space.

Among particle-based approaches, Particle Swarm optimization (PSO) is one of these techniques, first introduced by Kennedy and Eberhart in [1]. Initially inspired by the social behavior of bird flocking and fish schooling, PSO models the search for an optimal solution as a collective effort by particles moving through the problem's search space. Each particle represents a potential solution and is guided by both its own experience and the experiences of neighboring particles. This dynamic enables a balance between exploration (searching new areas) and exploitation (refining known good areas), which helps efficiently locate near-optimal solutions.

This report delves into the theoretical analysis of PSO, which was investigated during the internship. The goal is to contribute to a deeper understanding of how PSO works and to identify new directions for exploration to guarantee convergence.The structure of the report is as follows: In the first part, we present the fundamental dynamic equations of PSO and explain the metaheuristic behind it. We then implement PSO to observe its behavior on irregular functions (i.e., those that are continuous but non-convex). In the second part, we model PSO as a Markov chain to leverage its properties for convergence analysis. Finally, we explore various modes of convergence, identifying one that appears promising for further investigation and another that seems unsuitable for this purpose.

In this report, we focus on the problem of minimization. It is important to note that, without loss of generality, the theoretical framework and results presented here are equally applicable to the maximization problem. Given a compact space $K \subset \mathbb{R}^n$, with n the dimension of the search space, and $f \in \mathscr{C}(K, \mathbb{R})$ a continuous function. We want to solve the following minimization problem,

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

Note that f is not necessarily convex and not necessarily differentiable, The lack of regularity makes the task more complicated, we cannot use for instance the gradient descent which usually used for strongly convex and smooth functions. Consequently, Particle Swarm optimization algorithm introduces a novel heuristic designed to approximate the global minimum in this general context.

2 Particle Swarm optimization Algorithm (PSO)

2.1 Dynamic Equation of PSO

The heuristic we will be explaining in this part comes from the original paper of PSO [1, Section 2]. 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. Analogous to the behavior of flocks of birds, PSO is inspired by how a group of birds might search for food to maximize their utility. The birds interact with each other until the entire flock converges to the area that maximizes their utility. Birds rely on their own experiences and follow the group to some extent, thus they also consider the overall position of the group. These interactions represent a trade-off that the group must manage to converge as quickly as possible to the optimal area.

The dynamic system is composed of N particles, x_t^i is the position in $\mathbb{R}^n$ of the i-th particle at time t. Likewise v_t^i is the velocity of the i-th particle at time t. The following system gives the dynamic of the N particles across time, $\forall i \in \{1, \dots, N\}, \forall t \in \{0, \dots, T-1\}$,

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

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

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

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

with, w the inertia weight which is a user-supplied constant $(0 \leq w \leq 1)$, c_1, c_2 are two user-supplied constants $(0 \leq c_1, c_2 \leq 2)$ and r_1, r_2 are two random variables drawn form $\mathscr{U}([0,1])$, the uniform law in $[0,1]$, at each step of the velocity update.

The personal best of the particle i at time t, p_t^i is the best minimizer found so far by the particle i. The global best g_t at time t is the best minimizer within the swarm at time t.

The two equations (1) and (2) that drive the particles can be simplified in one dynamic equation in $\mathbb{R}^n$, as follows,

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

2.2 Behavior of Particle and Velocity Clamping

In the paper [2, Page 4], we can find a description of the two behaviors of a particle in PSO. In fact, each particle in the swarm is driven by two behaviors that we can notice in the velocity equation (1). The first behavior is represented by the cognitive component $c_1 r_1 (p_t^i - x_t^i)$ of the velocity. This component drives the local search of the particle in the search space (here K). This term mimics the fact that the particle relies on its own experience to approach the solution. The second behavior of a particle is the ability to follow the group, the social or exploitation term $c_2 r_2 (g_t - x_t^i)$ try to mimic this behavior that unable to have a global search in the search space.

If we set $c_1 = 0,\ c_2 \neq 0$, then individual exploration is eliminated, and each particle has no memory. Intuitively, if a particle becomes trapped in a local minimum, the entire group will converge to this local minimum. On the other hand, if we set $c_1 \neq 0,\ c_2 = 0$, we have no social interaction, and each particle relies solely on its own research, allowing for more exploration of the space. However, it is often more efficient to have $c_1, c_2 \neq 0$ because it combines both behaviors. By finding the right combination of constants, the algorithm can yield a better minimizer.

Often, optimization problems are constrained, and particles must adhere to these constraints, here the constrain is the compact K. If c_1, c_2 are too large, depending on the problem, the velocity will also become too large. As we can observe in the second dynamic equation, (2), The next position of the particle may not necessarily satisfy the constraints. That is why the velocity at the time $t+1$ in the equation (1) depends on the previous velocity which is controlled by the inertia coefficient w. In the first instance, w needs to be sufficiently close to 1 to ensure that the velocity is not too small, which could lead to issues of exploration. On the other hand, we need to set a maximum velocity v_{max} for the particles, which depends on the shape of the constraints, see [2, Page 4].

In order to prevent the particles from moving excessively beyond the search space, we employ a technique known as velocity clamping, which limits the maximum velocity of each particle.

Definition 1. *see [2, Page 4] If our search space K is a compact in $\mathbb{R}$, we can define the maximum velocity as follows, if $K = [x_{inf}, x_{sup}]$*

$$v_{max} = k \cdot \frac{x_{sup} - x_{inf}}{2}$$

With $k \in (0,1)$ represents a user-supplied velocity clamping factor. If a particle i at time t has a velocity v_t^i larger than v_{max}, then we set it to v_{max}. If v_t^i is lower than $-v_{max}$, we set it to $-v_{max}$.

We can extend this definition to dimensions beyond 2, in fact if the compact K is a product of line segments, we can therefore define the maximum velocity vector as the vector with the clamping velocities of the previous definition on each line segment,$\forall k \in (0,1)$,

$$K = \prod_{i=1}^{n} [a_i, b_i],\ v_{max} = (v_{max,i})_{1 \leq i \leq n} \text{ with } v_{max,i} = k \cdot \frac{b_i - a_i}{2} \tag{6}$$

2.3 Experimental results of the PSO algorithm

2.3.1 PSO Algorithm

The table 1, presents the constants used in the experiments conducted in this study.

Table 1: PSO Constants

Constants	`c_1`	`c_2`	`w`	`k`	`N`	`T`
Set value	1.9	1	0.9	0.5	20	30

Algorithm 1 PSO Algorithm

1: **Input :** c_1, c_2, w, T, k and N,
2: **Initialisation :**
Draw x_0^i from $\mathscr{U}(K)$ for $i = 1 \dots N$
$v_0 = (0)_{1 \leq i \leq N}$ and compute v_{max}
$p_0 = (x_0^i)_{1 \leq i \leq N}$, $g_0 = arg\min_{u \in \{x_0^{1:N}\}} f(u)$
3: **for** $t = 1$ **to** T **do**
4: Generate r_1, r_2 from $\mathscr{U}[0,1]$
5: $v_t^i = wv_{t-1}^i + c_1 r_1 (p_{t-1}^i - x_{t-1}^i) + c_2 r_2 (g_{t-1} - x_{t-1}^i), i = 1, \dots, N$
6: For $i = 1, \dots, N$ verify if v_t^i satisfies the velocity clamping condition
7: $x_t^i = x_{t-1}^i + v_t^i, i = 1, \dots, N$
8: For $i = 1, \dots, N$ verify if x_t^i is in the search space K otherwise set the components to the boundaries
9: $p_t^i = arg\min_{u \in \{p_{t-1}^{1:N}, x_t^{1:N}\}} f(u)$
10: $g_t = arg\min_{u \in \{g_{t-1}, x_t^{1:N}\}} f(u)$
11: **end for**
12: **Return** g_T

Note that in the PSO algorithm from line 6 we use the velocity clamping strategy to prevent the particles form going outside the search space K. Therefore we compare component wise the velocity of each particle with the vector v_{max}. By taking in the experiments K as a product of line segments, we use the generalized definition of v_{max} discussed in equation (6).

Note that in the line 8 we have to verify if the new particles are outside the search space K, here again for simplicity by taking K as a product of line space then it is sufficient to verify if each component of x_t^i is in the proper line segment otherwise we set the component to the nearest boundary.

The estimator provided by the PSO algorithm for the true solution $x^\star$ is the last global best g_T, therefore we can compute the MSE of this estimate,

$$MSE(N,T) = \mathbb{E}[\|g_T - x^\star\|^2],$$

we can estimate this expectation by the classical Monte Carlo estimator, let R denote the number of PSO runs,

$$\hat{MSE}(N,T) = \frac{1}{R}\sum_{k=1}^{R} \|g_{T,k} - x^\star\|^2 \tag{7}$$

We run PSO on non-convex test functions available in the paper [3, Page 13],$\forall x \in \mathbb{R}^n$,

- Alpine : This non-convex differentiable function has a global minimum at $x_{min} = (0,\dots,0)$.

$$f(x) = \sum_{i=1}^{n} |x_i \sin(x_i) + 0.1x_i|$$

- Ackley : This function is continuous, non-differentiable and non-convex and has its global minimum at $x_{min} = (0,\dots,0)$.

$$f(x) = -20\exp(-0.2\sqrt{\frac{1}{n}\|x\|_2^2}) - \exp(\frac{1}{n}\sum_{i=1}^{n}\cos(2\pi x_i)) + 20 + \exp(1)$$

- Rastrigin : The Rastrigin function is continuous, differentiable and convex, has lots of local minima and a global minimum at $x_{min} = (0,\dots,0)$.

$$f(x) = 10n + \sum_{i=1}^{n} x_i^2 - 10\cos(2\pi x_i)$$

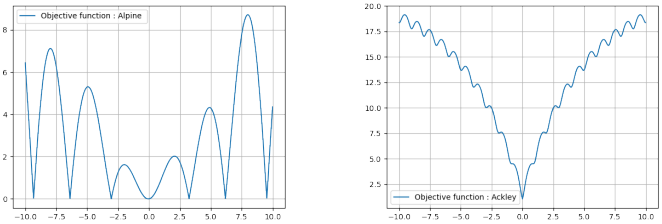
- Xinsheyang2 : This function is continuous but not differentiable and non-convex with a global minimum at $x_{min} = (0,\dots,0)$.

$$f(x) = \sum_{i=1}^{n} |x_i| \exp(-\sum_{i=1}^{n} \sin(x_i^2))$$

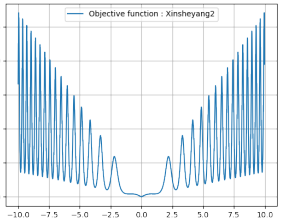
2.3.2 Experiments in Dimension 1

We run PSO on the compact set $K = [-10, 10]$ for the four test functions Figure 1,

The experiment in Figure 2 is that for each function, we run PSO and we only keep the random first state in orange and the last state of the dynamic in green, we also represent the evolution of the absolute value of the global best across the PSO run.

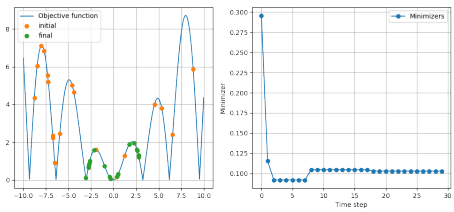


(a) Alpine (b) Ackley (c) Rastrigin

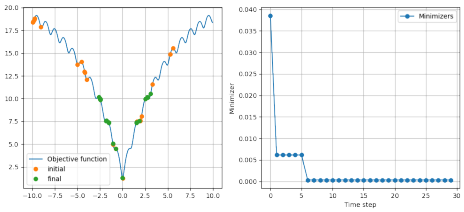


(d) Xinsheyang2

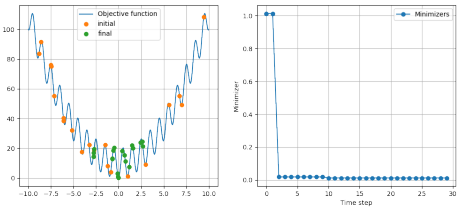
Figure 1: Non convex test functions in 1 dimension



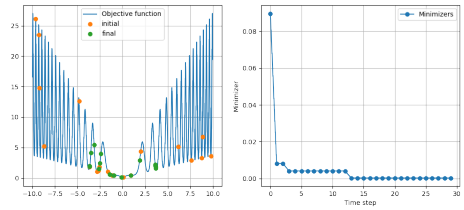
(a) Alpine



(b) Ackley



(c) Rastrigin



(d) Xinsheyang2

Figure 2: Dynamic of particles and evolution of the global best across time.

The result of the experiment in Figure 2 shows a typical issue of optimization algorithm using particles. In fact we can notice that for each function that some green particles are trapped in local minima. The main issue is to avoid the local minima by letting r_1, r_2 be stochastic. However, it can generally be observed that the orange points, representing the initial random states, tend to get stuck

in local minima as shown in Figure 2(d), while the green points successfully move closer to the global minimum.

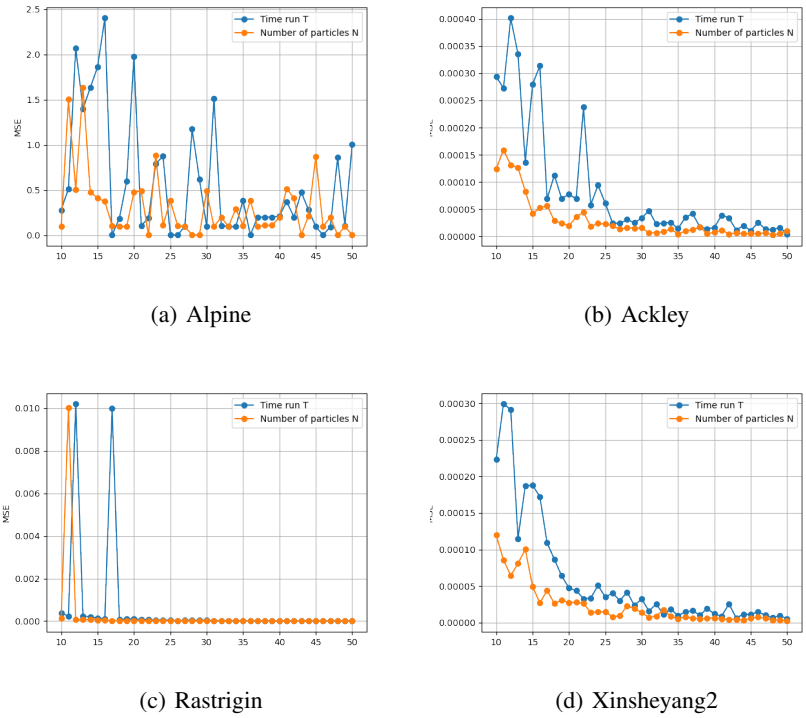
It can be observed that, for each test function, the absolute value of the global best decreases over time, with the exception in Figure 2(a). This scenario can occur when the global best solution approaches a local minimum and subsequently jumps to another, more distant local minimum.

We aim to evaluate the mean squared error (MSE) by repeating the experiment described above. The MSE is a function of two parameters, N and T and the empirical MSE is, $\forall N, T, R > 0$,

$$\hat{MSE}(N,T) = \frac{1}{R}\sum_{k=1}^{R}(g_{T,k})^2,\ R = 100 \tag{8}$$

For a given test function in Figure 3, the blue line represents the scenario where N is fixed at 20 particles, and the $\hat{MSE}$ is plotted as T varies from 10 to 50. Conversely, the orange line represents the case where T is fixed at 30 iterations, and the $\hat{MSE}$ is plotted as N varies from 10 to 50.

The Figure 3 provides insight into the convergence behavior of the algorithm across different test functions. We observe that, in some cases, such as in Figure 3(b), 3(c) and 3(d) the MSE approaches zero quite rapidly. However, for Figure 3(a), the convergence remains less clear, which may be attributed to the different landscape of the Alpine function.



(a) Alpine (b) Ackley

(c) Rastrigin (d) Xinsheyang2

Figure 3: Component wise empirical convergence of $\hat{MSE}(N,T)$

2.3.3 Experiments in Dimension 2

In this part we run PSO in a 2D compact search space $K = [-10, 10]^2 \subset \mathbb{R}^2$ with constants in Table 1 and for the four test functions in Figure 4 taken from [3, Page 14].

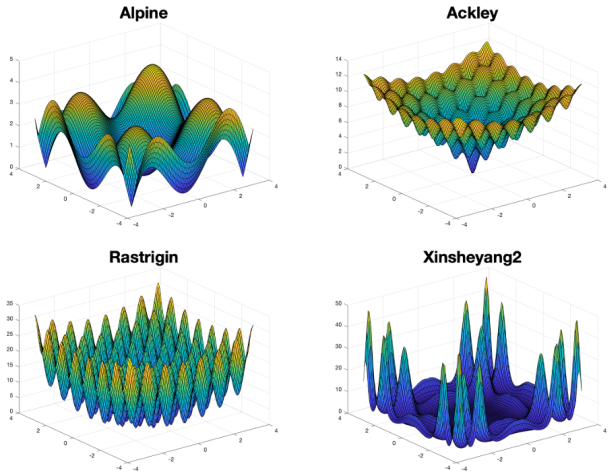
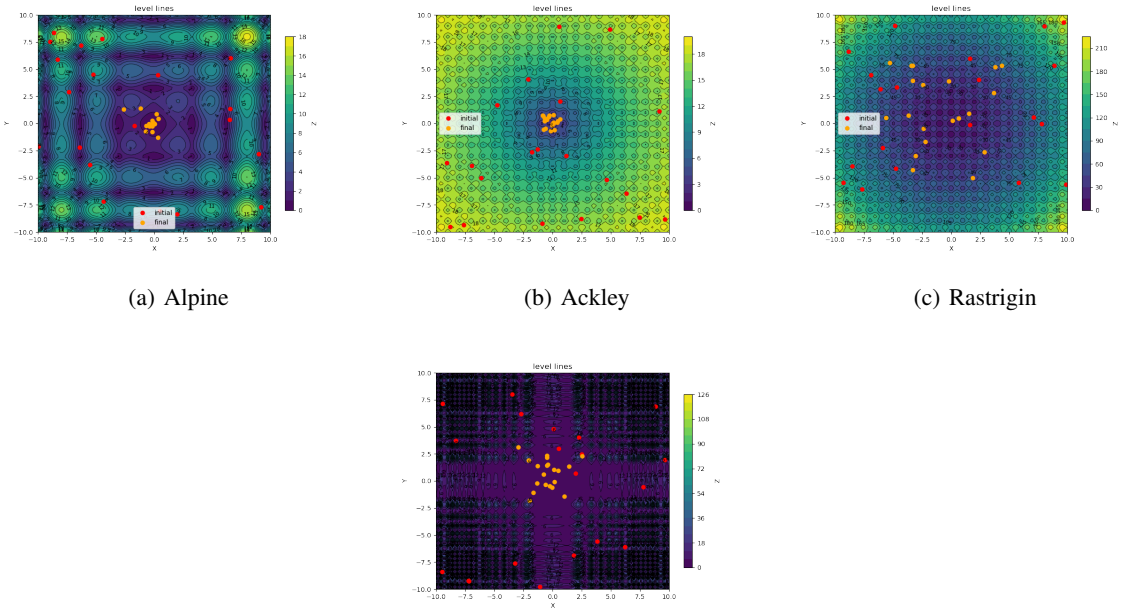


Figure 4: Non convex test functions in 2 dimensions

The Figure 7 presents the results of a PSO run. The red particles represent the random initial states, drawn from a uniform distribution over K, while the orange points correspond to the final states after 30 iterations (refer to Table 1).

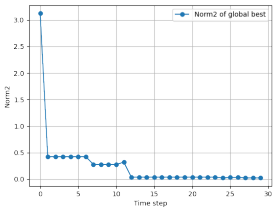


(a) Alpine (b) Ackley (c) Rastrigin

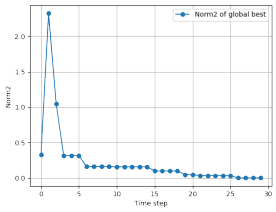
(d) Xinsheyang2

Figure 5: Evolution of 20 particles : initially in red, and in yellow after 30 iterations.

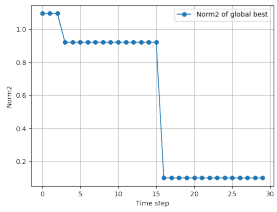
In Figure 5 it is observed that the final states are satisfying, as shown in Figures 5(a) and 5(b), where the yellow particles (final states) are concentrated in the dark zone with a lower gradient. In the other cases 5(c), 5(d) , the particles are more spread but still close to the global minimum $x^\star = (0,0)$.



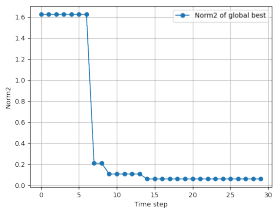
(a) Alpine



(b) Ackley



(c) Rastrigin



(d) Xinsheyang2

Figure 6: L^2 norm of the global best across time (30 iterations).

To better understand the convergence of the particles, we evaluate whether the global best approaches the global minimum in terms of the L^2 norm. As observed in Figure 6, after a few iterations, the L^2 norm of the global best rapidly decreases to zero over time.

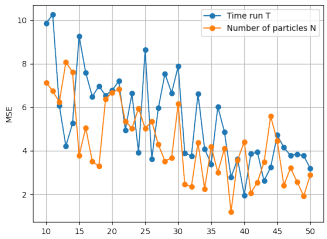
We can now adopt a weaker approach to convergence by assessing the empirical risk, which corresponds to convergence in probability by Markov's inequality, $\forall N, T, R > 0$,

$$\hat{MSE}(N,T) = \frac{1}{R}\sum_{k=1}^{R}(g_{T,k}[1])^2 + (g_{T,k}[2])^2 \tag{9}$$

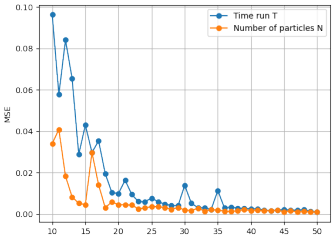
In Figure 7, for a given function, the blue line represents the scenario where N is fixed at 20 particles, and the $\hat{MSE}$ is plotted as T varies from 10 to 50. Conversely, the orange line represents the case where T is fixed at 30 iterations, and the $\hat{MSE}$ is plotted as N varies from 10 to 50.

It is observed that the $\hat{MSE}$ in Figures 7(b) and 7(d) approaches 0 more quickly than the $\hat{MSE}$ in Figure 7(c). Meanwhile, the $\hat{MSE}$ in Figure 7(a) decreases but does not appear to converge to 0 for this time range.

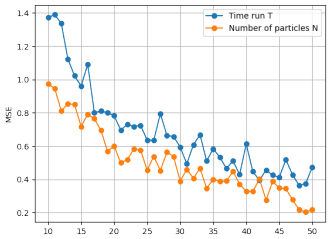
The key takeaway from these experiments is that the PSO estimate of the global minimum appears to be consistent with respect to both N and T. However, this convergence is noticeably influenced by the shape of the objective function f.



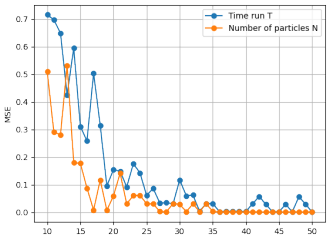
(a) Alpine



(b) Ackley



(c) Rastrigin



(d) Xinsheyang2

Figure 7: $\hat{MSE}(N,T)$ for the four test functions

3 Markov Chain Modeling of PSO

In this section, we develop a Markov chain model to describe the dynamics of PSO. First, we define Markov chains by introducing the transition kernel, followed by an explanation of the concept of homogeneity of Markov chains. The construction proceeds in two steps : first, we consider a random process representing a single particle like in [4]. In the second step, we aggregate all the individual processes and properly prove that the resulting process is a homogeneous Markov chain by giving more details than in [4].

3.1 Introduction to Markov Chains

we need to introduce the core object of the Markov chain theory which is the transition kernel.

Definition 2. *see [5, Page 6, Definition 1.2.1]*
Let $(X,\mathscr{X})$ and $(Y,\mathscr{Y})$ be two measurable spaces. A kernel P on $X\times\mathscr{Y}$ is a mapping $P : X\times\mathscr{Y}\to[0,\infty]$ satisfying the following conditions.

1. *$\forall x\in X$, the mapping $P(x,\cdot) : A\mapsto P(x,A)$ is a measure on $\mathscr{Y}$*
2. *$\forall A\in\mathscr{Y}$, the mapping $P(\cdot,A) : x\mapsto P(x,A)$ is a measurable function from $(X,\mathscr{X})$ to $([0,\infty],\mathscr{B}([0,\infty])$*

If $\forall x\in X, P(x,Y)=1$ then P is a Markov kernel.

Before presenting the model for the dynamic of Particle Swarm optimization (PSO) using a Markov chain, we will first provide a proper definition of a Markov chain with random variables,

Definition 3. *see [5, Page 4, Definition 1.1.1]*
Given a filtered probability space $(\Omega, \mathscr{F}, (\mathscr{F}_n)_{n\geq 0}, \mathbb{P})$ and a measurable space $(X, \mathscr{X})$, a $\mathscr{X}$-valued random process $(X_n)_{n\geq 0}$ adapted to $(\mathscr{F}_n)_{n\geq 0}$ is a Markov chain if and only if

$$\forall n \in \mathbb{N}, \forall A \in \mathscr{X},\ \mathbb{P}(X_{n+1} \in A \,|\, \mathscr{F}_n) = \mathbb{P}(X_{n+1} \in A \,|\, X_n)\ a.s. \tag{10}$$

Note that this definition is an equality between two transition kernels, object that we define later.

The equality (10) holds almost surely between two Markov kernels, specifically the Markov kernel of X_{n+1} given $X_n, \ldots, X_0$ and the Markov kernel of X_{n+1} given X_n. One can also adopt the following notation to mean the equality between the two Markov kernels of the equality (10),

$$X_{n+1} \,|\, (X_n, \ldots, X_0) = X_{n+1} \,|\, X_n$$

The underlying intuition behind Markov chains is that the next state depends solely on the current state, such that prior information is irrelevant.

If the transition kernel of X_{n+1} given $X_n, \ldots, X_0$ is equal to a constant Markov kernel P that is independent of n, then the Markov chain is said to be homogeneous.

Definition 4. *Given a filtered probability space $(\Omega, \mathscr{F}, (\mathscr{F}_n)_{n\geq 0}, \mathbb{P})$ and a measurable space $(X, \mathscr{X})$, a $\mathscr{X}$-valued Markov chain $(X_n)_{n\geq 0}$ adapted to $(\mathscr{F}_n)_{n\geq 0}$ is a homogeneous if and only if there exists a transition kernel P such that,*

$$\forall A \in \mathscr{X}, \forall n \geq 0,\ \mathbb{P}(X_{n+1} \in A \,|\, X_n) = P(X_n, A)\ \mathbb{P} - a.s. \tag{11}$$

3.2 The Swarm as a Homogeneous Markov Chain

The first random process we can define contains all the information about a particle, as follows,

Definition 5. *Let K be a compact set and $N \geq 1$, the number of particles, $\forall i \in \{1, \ldots, N\}, \forall t \geq 1$,*

$$\xi_t^i := (x_t^i, x_{t-1}^i, p_t^i, g_t) \in K^4 \tag{12}$$

and $\xi_0^i := (x_0^i, x_0^i, p_0^i, g_0)$,
Let $\mathscr{S} \subset K^4$ be the state space, the random process $(\xi_t^i)_{t\geq 0}$ takes its value in $(\mathscr{S}, \mathscr{B}(\mathscr{S}))$.

We can notice in the dynamic equation (5) that the next position of a particle is computed with the present and the former position, the present personal best and the global best. Having noticed that, one can then define a vector such as the its next state depends only on the present state,

We define a random process that aggregates all individual random processes in order to model the swarm

Definition 6. *The swarm random process is define as follows,* $\forall t \geq 1, \forall N \geq 1$,

$$\zeta_t = (\xi_t^1, \ldots, \xi_t^N) = (x_t^{1:N}, x_{t-1}^{1:N}, p_t^{1:N}, g_t^{1:N}) \in K^{4N} \tag{13}$$

and $\zeta_0 := (x_0^{1:N}, x_0^{1:N}, p_0^{1:N}, g_0^{1:N})$, *Let* $\Theta := \mathscr{S}^N$ *be the state space, the random process* $(\zeta_t)_{t\geq 0}$ *takes its value in* $(\Theta, \mathscr{B}(\Theta))$.

The next proposition states that the aggregated random process satisfies the Markov property,

Proposition 1. *The swarm random process* $(\zeta_t)_{t\geq 0}$ *is a Markov chain for the canonical filtration,* $\mathscr{F}_t := \sigma(\zeta_0, \ldots, \zeta_t)$

Proof. we want to show that,$\forall t > 0$,

$$\zeta_{t+1} \,|\, (\zeta_t, \ldots, \zeta_0) = \zeta_{t+1} \,|\, \zeta_t$$

For that we can understand what determines the behavior of each particle of the swarm at time $t+1$, $\forall i \in \{1, \ldots, N\}$,

$$\xi_{t+1}^i = (x_{t+1}^i, x_t^i, p_{t+1}^i, g_{t+1})$$

We compute with the synamic equation (5) the first component of ξ_{t+1}^i with the vector ξ_t^i which is contained in the swarm ζ_t. The second component x_t^i is also provided by ξ_t^i without any computation. p_{t+1}^i can take only two values and we compute it as follows :

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

It shows that only ξ_t^i is needed and the last component is the global best at time $t+1$ which is obtained as follows :

$$g_{t+1} = arg \min_{u \in \{g_t, (x_{t+1}^j)_{1\leq j\leq N}\}} f(u)$$

The global best g_{t+1} to be computed needs the previous global best which is in the previous swarm and the new positions of all the particles that we compute only with ζ_t. Therefore, $(\zeta_t)_{t\geq 1}$ is a Markov chain. ■

After having shown that the swarm process is a Markov chain, one can wonder if this Markov chain is homogeneous.

Proposition 2. *The Markov chain $(\zeta_t)_{t\geq 1}$ adapted to $\mathscr{F}_t = \sigma(\zeta_1,\dots,\zeta_t)$ is a homogeneous Markov chain.*

To simplify the notation, we use the symbol $\oplus$ to denote the law obtained by the addition of independent random variables, so the density of the resulting law is the convolution of the individual densities.

Proof. We want to show that the transition kernel does not depend on time.$\forall t > 0$

$$\zeta_{t+1} \,|\, \zeta_t = \zeta_2 \,|\, \zeta_1$$

The main arguments that makes this Markov chain homogeneous is the fact that the constants w, c_1 and r_2 in the dynamic equation (5) are time independent. Then this dynamic equation gives the density of x^i_{t+1} (14) does not depend on time and then the density of the whole swarm does not depend on time because the densities of the global best and the personal bests at time $t+1$ depends on the density of x^i_{t+1} which is time independent, $\forall i \in \{1,\dots,N\}, U_1, U_2 \sim \mathscr{U}[0,1]$,

$$\mathbb{P}(dx^i_{t+1} \,|\, \zeta_t) = [x^i_t + w(x^i_t - x^i_{t-1})] \oplus c_1(p^i_t - x^i_t)\cdot U_1 \oplus c_2(g_t - x^i_t)\cdot U_2 \tag{14}$$

■

Notice that we do not have $\zeta_{t+1} \,|\, \zeta_t = \zeta_1 \,|\, \zeta_0$ because the definition of ζ_0 is such as, $p^i_0 = x^i_0$ and then the transition law (14) for $t = 0$ is, $\forall U_2 \sim \mathscr{U}[0,1]$,

$$\mathbb{P}(dx^i_1 \,|\, \zeta_0) = x^i_t \oplus c_2(g_0 - x^i_0)\cdot U_2 \tag{15}$$

which defers from the transition law of $\zeta_{t+1} \,|\, \zeta_t$. The process is homogeneous from $t \geq 1$.

4 Different Convergence Analysis

In this section, we aim to detail and discuss three types perspectives on the convergence analysis of PSO. This part is divided into three subsections. In the first, we discuss the convergence of $(f(g_t))_{t\geq 0}$ which arises form the study in [6]. In the second, we analyze a type of convergence closely related to convergence in probability well studied in [7]. Finally, we examine a last type of convergence proposed in [4] which seems to be too strong and question its relevance.

4.1 Introductory Insights into Convergence

In this section, we aim to naively study the behavior of the global best sequence and the mapped sequence $(f(g_t))_{t\geq 0}$.

let $F : K^{4\times N} \longrightarrow \mathbb{R}$, be the mapping such that, $\forall x = [x_{i,j}]_{\substack{1\leq i\leq 4\\ 1\leq j\leq N}} \in K^{4\times N}$, $F(x) = f(x_{4,1})$. Therefore F maps the swarm process to $f(g_t)$, $\forall t \geq 0$, $F(\zeta_t) = f(g_t)$.

Proposition 3. *Given the Markov chain $(\zeta_t)_{t\geq 0}$ that describes PSO, there exists a random variable $\hat{\zeta} \in K^{4N}$ such as,*

$$F(\zeta_t) \xrightarrow[t\to\infty]{a.s} L := F(\hat{\zeta}) \tag{16}$$

Note that L is a random variable,

Proof. As the objective function f is continuous on K, F is continuous on K^{4N}. Thanks to Tychonoff's theorem K^{4N} is a compact set, then by continuity $F(K^{4N})$ is compact.

$\forall \omega \in \Omega$, $(F(\zeta_t(\omega)))_{t\geq 0} \in F(K^{4N})^{\mathbb{N}}$ is a bounded sequence and it is a non-increasing sequence by construction, in fact the global bests are such as, $\forall t \geq 0$,

$$F(\zeta_t(w)) = f(g_t(w)) \leq f(g_{t+1}(w)) = F(\zeta_{t+1}(w))$$

so according to The Monotone convergence theorem, it converges to a limit $L(w) \in \mathbb{R}$ such as,

$$F(\zeta_t(w)) \xrightarrow[t\to\infty]{} L(w)$$

and since $F(K^{4N})$ is a compact so it is closed then, $L \in F(K^{4N})$, there exists a random variable $\hat{\zeta}$ such as, $L = F(\hat{\zeta})$. ■

This result shows that all the trajectories of the random process $(f(g_t))_{t\geq 0}$ will converge, however the limit is not necessarily the same for all the trajectories. Moreover, we do not have any a priori information about the limit L. This result do not give any information about the asymptotic behavior of the global best process $(g_t)_{t\geq 0}$.

Regarding the behavior of $(\zeta_t)_{t\geq 0}$, we cannot claim that (16) implies the convergence of $(\zeta_t)_{t\geq 0}$ to $\hat{\zeta}$ almost surely by an argument of correspondence, as claimed in [6, Page 70, Proposition 5.6].

4.2 First Approach to Convergence

In the previous section, we have proven that the random process of the fitness values, $(f(g_t))_{t\geq 0}$, converges. However, it remains unclear whether the sequence of global bests, $(g_t)_{t\geq 0}$, constitutes a minimizing sequence or even a converging process. In this section, we aim to determine whether the sequence of global bests $(g_t)_{t\geq 0}$ converges in the sense defined below, this convergence was previously studied in [7].

To avoid repetition, in the following results and definitions, the random process $(\zeta_t)_{t\geq 0}$ will consistently represent the swarm process describing PSO, as defined in (13).

First, we define the specific notion of convergence under consideration,

Definition 7. *$\forall \varepsilon > 0$, we say that the $(g_t)_{t\geq 0}$ is ε-convergent if and only if,*

$$\mathbb{P}(g_t \in B_\varepsilon) \xrightarrow[t\to\infty]{} 1 \tag{17}$$

With $B_\varepsilon := \{x \in K, f(x^\star) \leq f(x) \leq f(x^\star) + \varepsilon\}$.

If ε-convergence holds for all $\varepsilon > 0$ then we have convergence.

From the outset, we have not imposed any assumptions on the uniqueness of the solution to the minimization problem, which may, in general, have multiple solutions. Therefore, the definition of convergence we consider is flexible, as the event of interest is $g_t \in B_\varepsilon$. Here, B_ε represents the pre-image of $B(f(x^\star), \varepsilon)$ by f which is continuous, so B_ε encompasses a union of regions surrounding all possible solutions. In this framework, g_t is permitted to approach any solution, accounting for the potential non-uniqueness of $x^\star$.

The convergence under consideration is weaker than the convergence in probability of $(g_t)_{t\geq 0}$ to $x^\star$. Let us denote by $f^{-1}[A]$ the pre-image of a subset $A \subset \mathbb{R}$. As f is continuous, $\forall \varepsilon > 0, \exists \delta > 0$, $B(x^\star, \delta) \subset f^{-1}[B(f(x^\star), \varepsilon)]$. If we suppose that $(g_t)_{t\geq 0}$ converges to $x^\star$ in probability, then $\mathbb{P}(g_t \in B(x^\star, \delta)) \xrightarrow[t\to\infty]{} 1$. Thus, inequality $\mathbb{P}(g_t \in B(x^\star, \delta)) \leq \mathbb{P}(g_t \in B_\varepsilon)$ completes the proof.

4.2.1 A preliminary simple result.

In this section, we aim to present an initial, straightforward result, and introduce a useful function that will serve as a foundation for a more advanced analysis afterwards,

Definition 8. *Given a subset $B \subset K$, $\forall t \geq 0, \forall N \geq 2$,*

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

and let $\Delta := \{\zeta \in \Theta,\ \Phi_{B_\varepsilon}(\zeta) = 0\}$

The random process of interest is $(\Phi_{B_\varepsilon}(\zeta_t))_{t\geq 0}$ for $\varepsilon > 0$., This process is a counting process, though it is unclear whether it qualifies as a Markov chain. What we can state with certainty is that this process is a function of the Markov chain $(\zeta_t)_{t\geq 0}$. A natural question arises: is a function of a Markov chain itself a Markov chain? If the mapping is injective, the answer is affirmative. However, in this case, the mapping is not injective. Fortunately, determining whether this process is a Markov chain is not essential for the forthcoming analysis.

We now state some results involving the function Φ, which establish a connection with ε-convergence,

Lemma 1. *see [7, Page 4051, Lemma 3] Given $\varepsilon > 0, t \geq 0$,*

$$(\Phi_{B_\varepsilon}(\zeta_t) \neq 0) = (g_t \in B_\varepsilon) \tag{18}$$

Proof. $\forall \varepsilon > 0, t \geq 0, \forall \omega \in (g_t \in B_\varepsilon)$ one has,

$$\begin{aligned}\Phi_{B_\varepsilon}(\zeta_t(\omega)) &= \sum_{i=1}^{N} (\mathbf{1}_{B_\varepsilon}(x_t^i(\omega)) + \mathbf{1}_{B_\varepsilon}(p_t^i(\omega)) + \mathbf{1}_{B_\varepsilon}(g_t(\omega))) \\ &= \sum_{i=1}^{N} \mathbf{1}_{B_\varepsilon}(x_t^i(\omega)) + \mathbf{1}_{B_\varepsilon}(p_t^i(\omega)) + N \geq N > 0\end{aligned}$$

Conversely, $\forall \omega \in (g_t \notin B_\varepsilon)$ one has,

$$f(x^\star) + \varepsilon < f(g_t(\omega)) \leq f(p_t^i(\omega)) \leq f(x_t^i(\omega))$$

therefore, $\{p_t^i(\omega), x_t^i(\omega)\} \notin B_\varepsilon$ and we have shown,

$$\Phi_{B_\varepsilon}(\zeta_t(\omega)) = 0$$

■

From **Lemma 1** we can derive simply a property of monotonicity as follows,

Lemma 2. *see [7, Page 4015, Lemma 4] Given* $\varepsilon > 0$,

$$\forall s > t \geq 0,\ (\Phi_{B_\varepsilon}(\zeta_t) \neq 0) \subset (\Phi_{B_\varepsilon}(\zeta_s) \neq 0) \tag{19}$$

Proof. According to **Lemma 1**, it is equivalent to show that,

$$\forall s > t \geq 0,\ (g_t \in B_\varepsilon) \subset (g_s \in B_\varepsilon)$$

Given $s > t \geq 0$, and $\omega \in (g_t \in B_\varepsilon)$ one has,

$$f(x^\star) \leq f(g_t(\omega)) \leq f(x^\star) + \varepsilon$$

and by construction of the PSO algorithm one has,

$$f(g_s(\omega)) \leq f(g_t(\omega))$$

By combining the two inequality one has,

$$f(x^\star) \leq f(g_s(\omega)) \leq f(g_t(\omega)) \leq f(x^\star) + \varepsilon$$

Therefore we have shown, $w \in (g_s \in B_\varepsilon)$.

■

Lemma 2 establishes that the sequence of events $(g_t \in B_\varepsilon)_{t \geq 0}$ is non-decreasing, implying that the set B_ε is absorbing for the random process $(g_t)_{t \geq 0}$. Similarly, this lemma indicates that the pre-image $\Phi_{B_\varepsilon}^{-1}[\mathbb{R}^\star]$ is an absorbing state for the Markov chain $(\zeta_t)_{t \geq 0}$.

Using **Lemma 2**, we can now state a first result regarding the trajectories of $(\zeta_t)_{t \geq 0}$ which are the maps $t \mapsto \zeta_t(\omega)$ for $\omega \in \Omega$, representing realizations of the swarm process. These trajectories correspond to what is observed when running the algorithm.

Theorem 1. *see [7, Page 4052, Theorem 1], Given $\omega \in \Omega$ that defines a trajectory $(\zeta_t(\omega))_{t\geq 0}$ and $\varepsilon > 0$, if there exists $\tau \geq 0$ such that $\Phi_{B_\varepsilon}(\zeta_\tau(\omega)) \neq 0$ then*

$$\forall t > \tau,\ g_t(\omega) \in B_\varepsilon \tag{20}$$

The proof of this theorem is very straight forward as it is a direct application of the **Lemma 2**.

This theorem is unsatisfactory because it only states that for a given realization of PSO and a fixed,$\varepsilon > 0$, small, if, by chance, the fitness of the global best at some time $t > 0$ is ε distant from $f(x^\star)$ then the global best remains in the ε-optimal region B_ε.. This result is obvious given the construction of the algorithm. What we would like to establish is a statement of the form : for any $\varepsilon > 0$ there almost surely exists a time t $(0 < t < +\infty)$ such that $g_t \in B_\varepsilon$.

Such a result would provide a strong guarantee that, for any fixed error $\varepsilon > 0$, the trajectories of the PSO algorithm will almost surely approach the global minimum as closely as desired. However, it is important to note that such a result is not equivalent to the convergence under consideration and does not necessarily imply it.

In the next section, we aim to obtain better results by leveraging the fact that we are working with a Markovian process.

4.2.2 Core results and advanced analysis

In this section, we aim to provide a more advanced analysis by investigating the behavior of the swarm. Specifically, we focus on understanding the transition probabilities of each particle in order to define the regions that each particle can potentially reach.

Let us denote bu $Supp(\mu)$ the support of any measure μ, and recall that for all $i \in \{1,\dots,N\}$,

$$\mathbb{P}(dx^i_{t+1} \,|\, \zeta_t) = [x^i_t + w(x^i_t - x^i_{t-1})] \oplus c_1(p^i_t - x^i_t) \cdot U_1 \oplus c_2(g_t - x^i_t) \cdot U_2$$

Definition 9. *Given a state $\zeta \in \Theta$ and $t > 0$. The feasible states at time $t+1$ that can be reached by the particle i of the swarm is :*

$$\mathscr{R}_{t+1}(\zeta) := \bigcup_{i=1}^{N} Supp(\mathbb{P}(dx^i_{t+1} \,|\, \zeta_t = \zeta)) \subset K$$

Note that $x^i_t + w(x^i_t - x^i_{t-1}), c_1(p^i_t - x^i_t)$ and $c_2(g_t - x^i_t)$ as ζ_t is fixed to ζ.

The probability measure $\mathbb{P}(dx^i_{t+1} \,|\, \zeta_t = \zeta)$ is defined on $K \subset \mathbb{R}^n$. We can decompose the support of this measure into the product of the supports of its marginal probabilities. Let $x^i_{t+1}(j)$ denote the j-th component of the vector x^i_{t+1}.The feasible region is then, for all $\zeta \in \Theta$,

$$\mathscr{R}_{t+1}(\zeta) = \bigcup_{i=1}^{N} \prod_{j=1}^{n} Supp(\mathbb{P}(dx_{t+1}^{i}(j) \,|\, \zeta_t = \zeta)) \tag{21}$$

At this point, we can elucidate the support of the marginal distributions to better understand the region that can be explored by the swarm. Since the marginal distributions are obtained by summing two independent uniform random variables (see Equation (14)), we require the following lemma,

Lemma 3. *see [7, Page 4049, Lemma 2]*
Assume that Y and Z are two independent random variables, and uniformly distributed in the range $[0,1]$*, then the probability density of the random variable* $X = aY + bZ + c$ $(a \neq 0, b \neq 0, a \neq b)$ *is*

$$f_X(x) = \begin{cases} \frac{x-\alpha}{|a||b|} & \text{if } \alpha \leq x \leq \beta \\ \frac{|a| \wedge |b|}{|a||b|} & \text{if } \beta \leq x \leq \gamma \\ \frac{\tau - x}{|a||b|} & \text{if } \gamma \leq x \leq \tau \\ 0 & \text{Otherwise} \end{cases}$$

With $\alpha = \min\{0,a\} + \min\{0,b\} + c$, $\beta = \min\{\min\{0,a\} + \max\{0,b\}, \max\{0,a\} + \min\{0,b\}\} + c$, $\gamma = \max\{\min\{0,a\} + \max\{0,b\}, \max\{0,a\} + \min\{0,b\}\}$ *and* $\tau = \max\{0,a\} + \max\{0,b\} + c$

From **Lemma 3**, we observe that the density of X is positive on the interval $[\alpha, \tau]$,implying that $Supp(\mathbb{P}_X) = [\alpha, \tau]$,

Now we can apply this result to have the support of the marginals, $\forall j \in \{1, \ldots, n\}, \forall t > 0$,

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

By applying **Lemma 3** with,

$$\begin{aligned} a &= c_2(g_t(j) - x_t^i(j)) \\ b &= c_1(p_t^i(j) - x_t^i(j)) \\ c &= x_t^i(j) + w(x_t^i(j) - x_{t-1}^i(j)) \end{aligned}$$

and $Y = r_2$, $Z = r_1$, we have, for all $\zeta \in \Theta$,

$$Supp(\mathbb{P}(x_{t+1}^{i}(j) \,|\, \zeta_t = \zeta)) = [\alpha_t^i(j), \tau_t^i(j)]$$

We can then write the feasible region as an union of hypercube,$\forall \zeta \in \Theta$,

$$\mathscr{R}_{t+1}(\zeta) = \bigcup_{i=1}^{N} \prod_{j=1}^{n} [\alpha_t^i(j), \tau_t^i(j)] \tag{22}$$

Note that for $t = 0$,

$$x_1^i(j) = x_0^i(j) + c_2 r_2 (g_0(j) - x_0^i(j))$$

Then the conditional variable $x_1^i(j)|\zeta_0 = \zeta \sim \mathscr{U}[x_0^i(j), x_0^i(j) + r_2(g_0(j) - x_0^i(j))]$ Then the feasible region can be written simply as follows,

$$\forall \zeta \in \Theta,\ \mathscr{R}_1(\zeta) = \bigcup_{i=1}^{N} \prod_{j=1}^{n} [x_0^i(j), x_0^i(j) + c_2(g_0(j) - x_0^i(j))] \tag{23}$$

At this point, we can state a result that establishes the connection between the feasible region and our problem,

Proposition 4. *see [7, Page 4051, Lemma 5]*
Given $\varepsilon > 0$ and $\zeta \in \Theta$, if there exists $t > 0$ such that $x^\star \in \mathscr{R}_{t+1}(\zeta)$ then there exists $\rho_\varepsilon > 0$,

$$\mathbb{P}(\Phi_{B_\varepsilon}(\zeta_{t+1}) \neq 0 \,|\, \zeta_t = \zeta) \geq \rho_\varepsilon > 0 \tag{24}$$

In the **Proposition 4** the expression of ρ_ε is $\rho_\varepsilon := \tilde{\rho_\varepsilon}^n$ with,

$$\tilde{\rho_\varepsilon} = \min\{\frac{1}{8c_1c_2m^2}(\frac{\delta}{3})^2, \frac{c_1 \wedge c_2}{4c_1c_2m}\frac{\delta}{3}, \frac{1}{8c_1c_2m^2}(\frac{\delta}{2})^2, \frac{\delta}{2c_1m}, \frac{\delta}{2c_2m}, 1\} \in (0,1)$$

Note that the lower bound ρ_ε is dependent on the constants δ ans m. The constant δ comes from the continuity of f, in fact by definition $\forall \varepsilon > 0$, there exists $\delta > 0$ such that,

$$\prod_{i=1}^{n} [x_i^\star - \delta, x_i^\star + \delta] \subset f^{-1}[B(f(x^\star), \varepsilon)]$$

And m is defined as $m := \max_{1 \leq i \leq n}\{|x_i^\star - \delta|, |x_i^\star + \delta|\}$.

From this point, we will present the main results concerning the convergence mode defined in Equation(17). To begin with, we will state a technical result from probability theory to provide rigorous proofs, followed by a theorem derived from [7], which assumes having the lower bound of Proposition 4 for a subsequence. Next, we will present two additional results inspired by this proof, using the lower bound from Proposition 4. Finally, we will conclude this section with a result from the article [7] which underlines the importance of the feasible region within this convergence mode.

Let us state a technical theorem from probability theory,

Theorem 2. ***(Jirina's theorem or the disintegration theorem)*** *Let X and Y be two random variables taking values in $(E, \mathscr{E})$ and $(F, \mathscr{F})$ respectively then there exists a transition kernel $\nu_{X|Y}$ on $E \times \mathscr{F}$ such that,*

$$\forall A \in \mathscr{E}, \forall B \in \mathscr{F},\ \mathbb{P}(X \in A, Y \in B) = \int_A \nu_{Y|X}(x, B)\, \mathbb{P}_X(dx)$$

and for $x \in E$, we call $\nu_{Y|X}(x, \cdot)$ the conditional law of the random variable Y given $X = x$ that we can also write $\mathbb{P}_{Y|X=x}$

This technical theorem provides a foundation for understanding conditional variables when conditioning on an event that may have probability zero, which will be applied in the following proof.

We can now state a result whose proof contains valuable ideas that will be useful later on,

Theorem 3. *see [7, Page 4052, Theorem 2] Given $\varepsilon > 0$, if there exists an increasing mapping $\alpha : \mathbb{N} \longrightarrow \mathbb{N}$ such as,*

$$\forall \zeta \in \Theta, \forall t \geq 0,\ \mathbb{P}(\Phi_{B_\varepsilon}(\zeta_{\alpha(t)+1}) \neq 0 \,|\, \zeta_{\alpha(t)} = \zeta) \geq \rho_\varepsilon > 0 \tag{25}$$

Then we have ε-convergence.

Note that this theorem assume having the lower bound of **Proposition 4** for a subsequence of the random process $(\zeta_t)_{t\geq 1}$.

Proof. Given $\varepsilon > 0$,

From **Lemma 2** one has, $\forall t > 0,\ (\Phi_{B_\varepsilon}(\zeta_t) = 0) \subset \cdots \subset (\Phi_{B_\varepsilon}(\zeta_1) = 0)$ then

$$(\Phi_{B_\varepsilon}(\zeta_1) = 0) \cap \cdots \cap (\Phi_{B_\varepsilon}(\zeta_t) = 0) = (\Phi_{B_\varepsilon}(\zeta_t) = 0) \tag{26}$$

Now let's show the convergence of $(g_{\alpha(t)})_{t\geq 0}$ in the sense already defined, $\forall t \geq 0$,

$$\mathbb{P}(g_{\alpha(t+1)} \in B_\varepsilon) = \mathbb{P}(\Phi_{B_\varepsilon}(\zeta_{\alpha(t+1)}) \neq 0) = 1 - \mathbb{P}(\Phi_{B_\varepsilon}(\zeta_{\alpha(t+1)}) = 0) \tag{27}$$

Let us expand the probability $\mathbb{P}(\Phi_{B_\varepsilon}(\zeta_{\alpha(t+1)}) = 0)$,

$$\begin{aligned}
\mathbb{P}(\Phi_{B_\varepsilon}(\zeta_{\alpha(t+1)}) = 0) &= \mathbb{P}(\Phi_{B_\varepsilon}(\zeta_1) = 0, \ldots, \Phi_{B_\varepsilon}(\zeta_{\alpha(t+1)}) = 0) \\
&\quad \text{We use the property in (26)} \\
&= \mathbb{P}(\Phi_{B_\varepsilon}(\zeta_1) = 0) \prod_{k=1}^{\alpha(t+1)-1} \mathbb{P}(\Phi_{B_\varepsilon}(\zeta_{k+1}) = 0 \,|\, \Phi_{B_\varepsilon}(\zeta_k) = 0, \ldots, \Phi_{B_\varepsilon}(\zeta_1) = 0) \\
&\quad \text{We expand the intersection by using conditioning} \\
&= \mathbb{P}(\Phi_{B_\varepsilon}(\zeta_1) = 0) \prod_{k=1}^{\alpha(t+1)-1} \mathbb{P}(\Phi_{B_\varepsilon}(\zeta_{k+1}) = 0 \,|\, \Phi_{B_\varepsilon}(\zeta_k) = 0) \\
&\quad \text{We use the property in (26) to simplify the conditioning} \\
&\leq \prod_{k=1}^{t} \mathbb{P}(\Phi_{B_\varepsilon}(\zeta_{\alpha(k+1)}) = 0 \,|\, \Phi_{B_\varepsilon}(\zeta_{\alpha(k+1)-1}) = 0)
\end{aligned} \tag{28}$$

To have the upper bound (28) We upper bound the probability in front of the product by 1, retaining only the terms corresponding to the subsequence α in the product, while bounding the other terms by 1.

Let's simplify the product (28), by writing the conditional probabilities as kernels and by applying the homogeneity of $(\zeta_t)_{t\geq 1}$. Since $\Phi_{B_\varepsilon}(\zeta_{\alpha(k+1)-1}) = 0$ there exists $\zeta \in \Delta := \{\zeta \in \Theta,\ \Phi_{B_\varepsilon}(\zeta) = 0\}$ such that $\zeta_{\alpha(k+1)-1} = \zeta$. $\forall k \in \{1, \ldots, t\}$,

$$\begin{aligned}
\mathbb{P}(\Phi_{B_\varepsilon}(\zeta_{\alpha(k+1)}) = 0 \,|\, \Phi_{B_\varepsilon}(\zeta_{\alpha(k+1)-1}) = 0) &= \nu_{\Phi_{B_\varepsilon}(\zeta_{\alpha(k+1)})\,|\,\Phi_{B_\varepsilon}(\zeta_{\alpha(k+1)-1})}(0,\{0\}) \\
&\quad \text{By definition of the conditional law} \\
&= \nu_{\Phi_{B_\varepsilon}(\zeta_{\alpha(k+1)})\,|\,\zeta_{\alpha(k+1)-1}}(\zeta,\{0\}) \\
&\quad \text{As previously justified above} \\
&= \mathbb{P}(\Phi_{B_\varepsilon}(\zeta_{\alpha(k+1)}) = 0 \,|\, \zeta_{\alpha(k+1)-1} = \zeta) \\
&\quad \text{By definition of the conditional law} \\
&= \mathbb{P}(\Phi_{B_\varepsilon}(\zeta_{\alpha(k+1)+1}) = 0 \,|\, \zeta_{\alpha(k+1)} = \zeta) \\
&\quad \text{By homogeneity of the Markov chain } (\zeta_t)_{t\geq 1} \\
&\leq 1-\rho_\varepsilon \quad \text{By assumption see (25).}
\end{aligned}$$

then the inequality (28) becomes, $\forall t \geq 0$,

$$0 \leq \mathbb{P}(\Phi_{B_\varepsilon}(\zeta_{\alpha(t+1)}) = 0) \leq (1-\rho_\varepsilon)^t$$

By plugging the upper bound (28) we can lower bound the expression (27) as follows,

$$1 \geq \mathbb{P}(g_{\alpha(t+1)} \in B_\varepsilon) \geq 1-(1-\rho_\varepsilon)^t \xrightarrow[t\to\infty]{} 1$$

The sequence $(\mathbb{P}(g_t \in B_\varepsilon))_{t\geq 0}$ is bounded and non-decreasing, and thus converges to 1, as it is already a cluster point, as shown above.

■

In this theorem, we do not use the lower bound from **Proposition 4** as a result, but rather assume the existence of a subsequence that satisfies such a lower bound. Moreover, the existence of such a subsequence has not been proven in the article [7]. Following the approach of the previous proof, it is not difficult to generalize this theorem to use **Proposition 4** without assuming the resulting lower bound.

Theorem 4. *Given $\varepsilon > 0$, if for all $\zeta \in \Delta$ there exists $t \geq 1$ such that $x^\star \in \mathscr{R}_{t+1}(\zeta)$ then we have ε-convergence.*

Proof. Following the same path as the previous proof we have, $\forall t \geq 1$,

$$0 \leq \mathbb{P}(\Phi_{B_\varepsilon}(\zeta_t) = 0) = \mathbb{P}(\Phi_{B_\varepsilon}(\zeta_1) = 0) \prod_{k=1}^{t-1} \mathbb{P}(\Phi_{B_\varepsilon}(\zeta_{k+1}) = 0 \,|\, \Phi_{B_\varepsilon}(\zeta_k) = 0) \tag{29}$$

For all $k \in \{1,\dots,t-1\}$, the conditional probabilities can be rewritten thanks to the existence of a $\zeta \in \Delta$ as follows,

$$\begin{aligned}
\mathbb{P}(\Phi_{B_\varepsilon}(\zeta_{k+1}) = 0 \,|\, \Phi_{B_\varepsilon}(\zeta_k) = 0) &= \nu_{\Phi_{B_\varepsilon}(\zeta_{k+1})\,|\,\Phi_{B_\varepsilon}(\zeta_k)}(0,\{0\}) \\
&= \nu_{\Phi_{B_\varepsilon}(\zeta_{k+1})\,|\,\zeta_k}(\zeta,\{0\}) \\
&= \mathbb{P}(\Phi_{B_\varepsilon}(\zeta_{k+1}) = 0 \,|\, \zeta_k = \zeta)
\end{aligned}$$

By assumption there exists $t_0 \geq 1$ such that $x^\star \in \mathscr{R}_{t_0+1}(\zeta)$ then by **Proposition 4** there exists $\rho_\varepsilon > 0$ such that,

$$\mathbb{P}(\Phi_{B_\varepsilon}(\zeta_{t_0+1}) = 0 \,|\, \zeta_{t_0} = \zeta) \leq 1 - \rho_\varepsilon$$

and by homogeneity we have, $\forall k \in \{1, \ldots, t-1\}$,

$$\mathbb{P}(\Phi_{B_\varepsilon}(\zeta_{k+1}) = 0 \,|\, \zeta_k = \zeta) = \mathbb{P}(\Phi_{B_\varepsilon}(\zeta_{t_0+1}) = 0 \,|\, \zeta_{t_0} = \zeta) \leq 1 - \rho_\varepsilon$$

Then one can upper bound the equality in (29) as follows,

$$0 \leq \mathbb{P}(\Phi_{B_\varepsilon}(\zeta_t) = 0) \leq \mathbb{P}(\Phi_{B_\varepsilon}(\zeta_1) = 0)(1 - \rho_\varepsilon)^{t-1} \xrightarrow[t \to \infty]{} 0$$

and

$$\mathbb{P}(g_t \in B_\varepsilon) = \mathbb{P}(\Phi_{B_\varepsilon}(\zeta_t) \neq 0) \xrightarrow[t \to \infty]{} 1$$

which completes the proof.

■

Theorem 4 indicates that if we can ensure that the feasible region contains $x^\star$ at some point, then convergence will be achieved. However, in [7], the existence of such a moment is not addressed. At this point, we face a new challenge which is finding a way to ensure that the feasible region will contain $x^\star$ at some point, possibly by tuning the constants c_1 and c_2 this direction remains somewhat unclear and requires further clarification.

To motivate the consideration of the feasible region in the convergence analysis, we can state a theorem that claim if $x^\star$ is not contained in the closure of all successive feasible regions, then the global best cannot approach $x^\star$ as closely as desired.

Theorem 5. *see [7, Page 4053, Theorem 3] 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 $x^\star \notin \overline{\bigcup_{t\geq 0} \mathscr{R}_{t+1}(\zeta_t(w))}$ then there exists $\varepsilon > 0$ such as,*

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

Proof. Let $\mathscr{C} := \overline{\bigcup_{t\geq 0} \mathscr{R}_{t+1}(\zeta_t(w))}$, as $\mathscr{C}$ is a compact set and f is continuous, we define $f_{min} := \min_{\mathscr{C}} f$. By assumption $x^\star \notin \mathscr{C}$ which guarantee that $f_{min} - f(x^\star) > 0$. Then $\forall \varepsilon \in (0, f_{min} - f(x^\star))$, one can notice that, $B_\varepsilon \cap \mathscr{C} = \emptyset$.

In fact, if $B_\varepsilon \cap \mathscr{C} \neq \emptyset$, then there exists, $u \in B_\varepsilon \cap \mathscr{C}$, such as

$$f(x^\star) \leq f(u) \leq f(x^\star) + \varepsilon < f_{min}$$

However, as $u \in \mathscr{C}$, we have $f(u) \geq f_{min}$. It shows that the intersection is an empty set. $\forall t \geq 0$,

$$\mathbf{1}_{g_t(w) \in B_\varepsilon} = \mathbf{1}_{\bigcup_{k=0}^{t} \bigcup_{i=1}^{N} (x_k^i(w) \in B_\varepsilon)} \leq \sum_{k=0}^{t} \sum_{i=1}^{N} \mathbf{1}_{x_k^i(w) \in B_\varepsilon}$$

However, for all $t \geq 0$ and $k \in \{1, \ldots, t\}$, $x_k^i(w) \in \mathscr{C} \subset B_\varepsilon^c$ then $\mathbf{1}_{g_t(w) \in B_\varepsilon} = 0$. ■

4.3 Alternative Perspective on Convergence

In some paper we can find another approach to the convergence which is different from the ε-convergence. Let's define the optimal region of the particle swarm in order to define a convergence of the PSO algorithm, see [6] and [8].

First of all we define the set of the optimal states of the particles represented by the random process $(\xi_t^i)_{t\geq 1}$. Then we define the sates such as the swarm is optimal.

Definition 10. *see [6]*

Let $\mathscr{A}$ denote the set of the global minima of f, the set of optimal states of the particle i is :

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

with $x^\star \in \mathscr{A}$,

which means when the personal best and the global best reach the same theoretical solution of the minimizing problem $(\mathscr{P})$ the particle is optimal. Now we can define the optimal swarm state $\Gamma \subset \Theta$ which is the set of the optimal states of the random process $(\zeta_t)_{t\geq 1}$.

Definition 11. *see [6] Given the Markov chain, $(\zeta_t)_{t\geq 0}$ that describes PSO, we define Γ as:*

$$\Gamma := \{(x_1,\dots,x_N) \in \Theta \mid \exists i \in \{1,\dots,N\}, x_i \in \mathscr{M}\}$$

Which means that the state of the swarm is optimal if any particle is optimal.

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

The sets $\mathscr{M}$ and Γ verify a property of closeness or absorbance which is define as follows,

Definition 12. *Let P on $X \times \mathscr{X}$ be a Markov kernel, a set $B \in \mathscr{X}$ is called closed or absorbing if and only if*

$$\forall x \in B,\ P(x,B) = 1$$

Namely, we cannot reach another state outside of B if the chain is already in B,

Now one can properly state the following two propositions,

Proposition 5. *see [6] $\mathscr{M}$ is an absorbing set for the random process $(\xi_t^i)_{t\geq 0}$.*

Proof. Given $i \in \{1,\dots,N\}$ and $\xi \in \mathscr{M}$, let's show that $\mathbb{P}(\xi_{t+1}^i \in \mathscr{M} \mid \xi_t^i = \xi) = 1$

If $\xi_t^i = \xi \in \mathscr{M}$, then by definition $p_t^i = g_t = x^\star$, The next personal best is define as follows,

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

it shows that $p_{t+1} := x^\star$, likewise for the next global best we have ,

$$g_{t+1} := \arg \min_{u \in \{x_{t+1}^{1:N}, g_t\}} f(u) = x^\star$$

therefore $\xi_{t+1}^i \in \mathscr{M}$ which completes the proof. ■

We have the same result for the set Γ,

Proposition 6. *see [6] Γ is an absorbing set of the random process $(\zeta_t)_{t\geq 0}$.*

The proof is pretty straight forward as it derives from the previous proof,

Proof. Given $\zeta \in \Gamma$, let's show that $\mathbb{P}(\zeta_{t+1} \in \Gamma \,|\, \zeta_t = \zeta) = 1$

If $\zeta_t = \zeta \in \Gamma$, then by definition there exists $i \in \{1,\dots,N\}$ such that $\xi_t^i \in \mathscr{M}$. The same proof of the closeness of $\mathscr{M}$ shows that $\xi_{t+1}^i \in \mathscr{M}$. By definition it means that $\zeta_{t+1} \in \Gamma$ which completes the proof. ■

At this point we can define the convergence under consideration studied in [6].

Definition 13. *The Markov chain, $(\zeta_t)_{t\geq 0}$ that describes PSO is convergent if and only if*

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

We can intuitively understand that this convergence is more powerful than the ε-convergence . In fact asking for $x^\star \in \Gamma$ happens is equivalent to ask that at a certain time t the position of a particle is exactly equal to $x^\star$, however x_t^i has a conditional densities which is then non-atomic and can apply that this event is a null probability event. This idea is clarified by the following proposition,

Proposition 7. *Given the Markov chain $(\zeta_t)_{t\geq 0}$ that describes PSO then*

$$\forall t \geq 0,\ \mathbb{P}(\zeta_t \in \Gamma) = 0 \tag{30}$$

Proof. We proof the proposition by induction,

Initialization:

Notice that for $i \in \{1,\dots,N\}$, $x_0^i \sim \mathscr{U}(K)$, $\mathbb{P}(x_0^i = x^\star) = 0$. then by the union bound inequality we have,

$$0 \leq \mathbb{P}(\zeta_0 \in \Gamma) = \mathbb{P}(\bigcup_{i=1}^{N}(x_0^i = x^\star)) \leq \sum_{i=1}^{N} \mathbb{P}(x_0^i = x^\star) = 0$$

Inductive Step:

Assume that for a fixed $t > 0$, we have $\mathbb{P}(\zeta_t \in \Gamma) = 0$ then,

$$\begin{aligned}
\mathbb{P}(\zeta_{t+1} \in \Gamma) &= \mathbb{P}(\zeta_{t+1} \in \Gamma \,|\, \zeta_t \notin \Gamma) \\
&\quad \zeta_t \notin \Gamma \text{ is an event almost sure by the inductive hypothesis} \\
&= \mathbb{P}(\bigcup_{i=1}^{N} x_{t+1}^i = x^\star \,|\, \zeta_t \notin \Gamma) \\
&\quad \text{Since the state } t \text{ is not optimal to be optimal at } t+1 \text{ one positions has to reach } x^\star \\
&\leq \sum_{i=1}^{N} \mathbb{P}(x_{t+1}^i = x^\star \,|\, \zeta_t \notin \Gamma) \qquad (31) \\
&\quad \text{By union bound}
\end{aligned}$$

Let's rewrite the conditional probability by using Jirina theorem (see **Theorem 2**),

$$\begin{aligned}
\mathbb{P}(x_{t+1}^i = x^\star \,|\, \zeta_t \notin \Gamma) &= \mathbb{P}(x_{t+1}^i = x^\star \cap \zeta_t \in \Gamma^c) \\
&\quad \zeta_t \notin \Gamma \text{ is an event almost sure} \\
&= \int_{\Gamma^c} \nu_{x_{t+1}^i|\zeta_t}(x, \{x^\star\}) \mathbb{P}_{\zeta_t}(dx) \\
&\quad \text{see } \textbf{Theorem 2}
\end{aligned}$$

Since $x_{t+1}^i|\zeta_t$ has a density with respect to the Lebesgue measure, (see **Lemma 3** and (14)) we have, $\nu_{x_{t+1}^i|\zeta_t}(x, \{x^\star\}) = 0$ then, $\forall i \in \{1, \ldots, N\}$,

$$\mathbb{P}(x_{t+1}^i = x^\star \,|\, \zeta_t \notin \Gamma) = 0$$

Finally the inequality (31) shows that, $\mathbb{P}(\zeta_{t+1} \in \Gamma) = 0$ ■

Proposition 7 shows as a conclusion that $\mathbb{P}(\zeta_t \in \Gamma) \xrightarrow[t\to\infty]{} 0 \neq 1$, unless a mistake in the analysis, this mode of convergence is not worth studying.

5 Discussion and conclusion

The implementation of PSO highlights the common issue of particle-based optimization algorithms becoming stuck in local minima. Additionally, it reveals that the performance of PSO can vary significantly depending on the shape of the objective function. From basic convex optimization theory, it is well-known that optimization algorithms generally perform well on smooth, gently varying functions. The empirical analysis of the empirical risk showed that this risk tends to converge to zero as T (number of iterations) and N (number of particles) increase.

We have presented three approaches to convergence. The first, a naive and very simple approach, provides almost no information about the behavior of $(\zeta_t)_{t\geq 0}$ and $(g_t)_{t\geq 0}$. This approach only tells us that the fitness of the global best, $(f(g_t))_{t\geq 0}$ converges, but it gives no insight into whether the limit is the global minimum or not. The second, more advanced approach leverages the fact that PSO can be modeled as a Markov chain. A careful examination of the proof of **Theorem 3** led us to formulate a new theorem using the lower bound from **Proposition 4**. However, this new approach assumes the existence of a time t such that the next feasible region contains $x^\star$, the global optimum. The concept of a feasible region introduces a new direction for investigation, as emphasized by **Theorem 5**. The last mode of convergence was addressed thanks to the previous work, which shows that the definition of convergence under consideration is too strong. In fact, this definition fails to account for the absolute continuity of the conditional variable with respect to the Lebesgue measure. As a result, we have shown that this alternative perspective on convergence is inappropriate.

In this work, we integrated insights from various studies on convergence, observing that in some papers, convergence was analyzed without fully considering the role of transition probabilities. A key challenge was thoroughly understanding the proofs of convergence and adapting them, especially when the underlying arguments were not immediately clear. Our focus on transition probabilities provided an additional perspective on the final definition of convergence, highlighting aspects that may have been overlooked in previous approaches.

An interesting new approach is to consider a possible variant of PSO, as there are already many, see [9]. A general idea to simplify an optimization problem is regularization, which involves introducing a penalty term, such as in ridge or lasso regression. One possible regularization that can be applied to PSO is the use of the Proximity Map. The proximity map is interesting because it relates to the fixed-point problem, see [10, Page 131]. Indeed, by iterating a minimization algorithm using the proximity map, we achieve convergence to the fixed points of the proximal operator, which are the minimizers of the objective function. The idea to explore would be to update the global bests using the proximity map.

References

[1] James Kennedy and Russell Eberhart. Particle swarm optimization. In *Proceedings of ICNN'95-international conference on neural networks*, volume 4, pages 1942–1948. IEEE, 1995.

[2] James Blondin. Particle swarm optimization: A tutorial. *Academia*, pages 1–5, 2009.

[3] Claudia Totzeck and Marie-Therese Wolfram. Consensus-based global optimization with personal best. *arXiv preprint arXiv:2005.07084*, 2020.

[4] Feng PAN et al. Analysis of standard particle swarm optimization algorithm based on markov chain. *Science Direct*, 2013.

[5] Randal Douc, Eric Moulines, Pierre Priouret, and Philippe Soulier. *Markov Chains*. Springer Series in Operations Research and Financial Engineering. Springer Nature Switzerland AG, Cham, 1 edition, 2018.

[6] Guosong Yu Gang Xu. On convergence analysis of particle swarm optimization algorithm. *Journal of Computational and Applied Mathematics*, pages 65–73, 2017.

[7] Weiyi Qian and Ming Li. Convergence analysis of standard particle swarm optimization algorithm and its improvement. *Soft Computing*, 22(13):4047–4070, 2018.

[8] Xiao-Ting LI Feng PAN. Analysis of standard particle swarm optimization algorithm based on markov chain. *Procedia Computer Science*, 18:1216–1225, 2013.

[9] Tareq M Shami, Ayman A El-Saleh, Mohammed Alswaitti, Qasem Al-Tashi, Mhd Amen Summakieh, and Seyedali Mirjalili. Particle swarm optimization: A comprehensive survey. *IEEE Access*, 10:10031–10061, 2022.

[10] Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends® in Optimization*, 1(3):127–239, 2014.

Mohamed Keteb

ENSAE 2nd year
Applied internship
2023 - 2024
Supervisor : Francesca Romana Crucinio

On convergence analysis of Particle Swarm Optimization with Markov Chain

—

Summary Note

Context of the internship and subject

The applied internship took place in the Department of Mathematics at King's College, under the supervision of Dr. Francesca Romana Crucinio. The focus of the internship was on optimization techniques, which are crucial to the success of modern machine learning methods. optimization naturally arises in the context of machine learning, where it is essential for minimizing loss functions that are often neither convex nor regular. Techniques such as gradient descent are commonly used; however, newer approaches like stochastic gradient descent (SGD) have shown promising results in complex, non-convex cases. Despite this, SGD does not guarantee convergence in all scenarios.

Ce stage est donc axé sur le développement de nouvelles techniques d'optimisation pour des fonctions objectif dans un cadre très général, tout en fournissant des analyses de convergence qui garantissent un progrès vers un minimum global. Dans ce stage, nous nous concentrons spécifiquement sur l'analyse de convergence d'un algorithme existant appartenant à une classe alternative de méthodes d'optimisation connue sous le nom d'algorithmes d'optimisation à base de particules.

Particle-based algorithms approach optimization by evolving particles over time, thereby allowing for the exploration of the search space to approach a global minimum. The strength of this class of algorithms lies in the fact that the assumptions about the objective functions are as unconstraining as possible, limited solely to continuity.

The algorithm under consideration in this internship in Particle Swarm optimization (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. Analogous to the behavior of flocks of birds, PSO is inspired by how a group of birds might search for food to maximize their utility. The birds interact with each other until the entire flock converges to the area that maximizes their utility. Birds rely on their own experiences and follow the group to some extent, thus they also consider the overall position of the group. These interactions represent a trade-off that the group must manage to converge as quickly as possible to the optimal area.

On convergence of Particle Swarm Optimization

First, we implemented the Particle Swarm optimization (PSO) algorithm to highlight key challenges inherent to particle-based optimization methods. Specifically, particles can become trapped in local minima, limiting the algorithm's effectiveness. The results demonstrate that the algorithm's performance is highly dependent on the shape and characteristics of the objective function. The empirical analysis of the empirical risk showed that this risk tends to converge to zero as T (number of iterations) and N (number of particles) increase.

In the first part of our analysis, we represent PSO as a Markov chain, demonstrating that the Markov chain modeling the swarm is homogeneous. After establishing this analytical framework, we define the notion of convergence. As a preliminary step, we show, somewhat naively, that the sequence of global best positions $(g_t)_{\geq 0}$ the best position identified by any particle up to time t has a fitness value $f(g_t)$ that converges over time. However, this initial result does not provide information on whether this limit represents a global minimum or any details on the convergence properties of $(g_t)_{t\geq 0}$ itself.

Next, we define two potential modes of convergence for the sequence of the global best. The first is a weaker form of convergence in probability, while the second is a stronger convergence criterion. To analyze these, we introduce supporting lemmas and preliminary results, leveraging the Markov chain framework by examining the transition law to gain insights into the regions that particles are able to explore. The key result in this part shows that if the global minimum $x^\star$ lies within the feasible region of particles, specifically, within the support of the transition probability, then the PSO algorithm converges in a defined sense. This result indicates that when the global minimum is accessible to the particles at some time t, convergence of the algorithm can be assured.

In the final part, we explore a stronger convergence mode, defined by defining optimal sets of the random process that describes PSO. We define convergence as the probability that the Markov chain enters these optimal sets, approaching 1 as t goes to infinity. However, our previous analysis reveals that the transition probability of the Markov chain has a density with respect to the Lebesgue measure, suggesting that this mode of convergence will not be well-suited to this context.

In this work, we integrated insights from various studies on convergence, observing that in some papers, convergence was analyzed without fully considering the role of transition probabilities. A key challenge was thoroughly understanding the proofs of convergence and adapting them, especially when the underlying arguments were not immediately clear. Our focus on transition probabilities provided an additional perspective on the final definition of convergence, highlighting aspects that may have been overlooked in previous approaches.

Mohamed Keteb

ENSAE 2ème année
Stage d'application
2023 - 2024
Maître de stage : Francesca Romana Crucinio

Analyse par chaîne de Markov de l'algorithme Particle Swarm Optimization

—

Note de synthèse

Contexte du stage et du sujet

Le stage appliqué s'est déroulé au Département de Mathématiques du King's College, sous la supervision de la Dr Francesca Romana Crucinio. L'objectif du stage portait sur les techniques d'optimisation, qui sont essentielles pour le succès des méthodes modernes d'apprentissage automatique. L'optimisation apparaît naturellement dans le contexte de l'apprentissage automatique, où elle est cruciale pour minimiser les fonctions de perte, qui sont souvent ni convexes ni régulières. Des techniques comme la descente de gradient sont couramment utilisées ; cependant, des approches plus récentes, telles que la descente de gradient stochastique (SGD) qui est plus performante lorsque la fonction objectif n'est plus convexe, bien que la convergence vers la solution optimale ne soit pas garantie dans ce cas.

Ce stage se situe dans le domaine de recherche qui vise à développer de nouvelles techniques d'optimisation pour des fonctions objectif dans un cadre très général, tout en fournissant des analyses de convergence qui garantissent un progrès vers un minimum global. Dans ce stage, nous nous concentrons spécifiquement sur l'analyse de convergence d'un algorithme existant appartenant à une classe alternative de méthodes d'optimisation connue sous le nom d'algorithmes d'optimisation à base de particules.

Les algorithmes à base de particules abordent l'optimisation en faisant évoluer des particules au fil du temps, permettant ainsi d'explorer l'espace de recherche pour approcher un minimum global. La force de cette classe d'algorithmes réside dans le fait que les hypothèses sur les fonctions objectives sont les moins contraignantes possible, se limitant uniquement à la continuité.

L'algorithme considéré dans ce stage est l'algorithme Particle Swarm Optimization (PSO). La méthodologie PSO est une approche dynamique de l'optimisation qui peut être vue, comme le suggère le nom, comme un essaim de particules interagissant les unes avec les autres. À l'instar du comportement des bandes d'oiseaux, PSO s'inspire de la manière dont un groupe d'oiseaux pourrait chercher de la nourriture pour maximiser son utilité. Les oiseaux interagissent entre eux jusqu'à ce que l'ensemble du groupe converge vers la zone qui maximise leur utilité. Les oiseaux s'appuient sur leurs propres expériences et suivent le groupe dans une certaine mesure, tenant ainsi compte de

la position globale du groupe. Ces interactions représentent un compromis que le groupe doit gérer pour converger aussi rapidement que possible vers la zone optimale.

L'analyse de convergence de PSO

Tout d'abord, nous avons implémenté PSO pour mettre en évidence les principaux défis inhérents aux méthodes d'optimisation à base de particules. En particulier, les particules peuvent se retrouver piégées dans des minima locaux, ce qui limite l'efficacité de l'algorithme. Les résultats démontrent que la performance de l'algorithme dépend fortement de la forme et des caractéristiques de la fonction objectif. L'analyse empirique du risque empirique a montré que ce risque tend à converger vers zéro à mesure que T (le nombre d'itérations) et N (le nombre de particules) augmentent.

Dans la première partie de notre analyse, nous modélisons PSO comme une chaîne de Markov homogène. Après avoir établi ce cadre analytique, nous définissons la notion de convergence. En tant qu'étape préliminaire, nous montrons, de manière quelque naïve, que la séquence des meilleures positions globales $(g_t)_{t\geq 0}$, la meilleure position identifiée par l'ensemble des particules jusqu'à l'instant t, a une valeur de fitness $f(g_t)$ qui converge. Cependant, ce résultat initial ne fournit pas d'informations sur le fait que cette limite représente un minimum global ou sur les propriétés de convergence du processus $(g_t)_{t\geq 0}$.

Ensuite, nous définissons deux modes de convergence potentiels pour la suite des meilleures positions globales. Le premier est une forme plus faible de convergence en probabilité, tandis que le second constitue un critère de convergence plus fort. Pour analyser ces modes, nous introduisons des lemmes et des résultats préliminaires. La modélisation par chaîne de Markov appelle à examiner la loi de transition pour obtenir des informations sur les régions que les particules sont capables d'explorer. Le résultat clé dans cette partie montre que si le minimum global $x^\star$ se trouve dans la région atteignable par les particules, c'est à dire le support de la probabilité de transition, alors l'algorithme PSO converge dans un sens bien défini.

Dans la dernière partie, nous explorons un mode de convergence plus fort, défini par des états optimaux du processus aléatoire qui décrit PSO. Nous définissons la convergence comme la probabilité que la chaîne de Markov entre dans un l'état optimal, approchant 1 lorsque t tend vers l'infini. Cependant, notre analyse précédente révèle que la probabilité de transition de la chaîne de Markov a une densité par rapport à la mesure de Lebesgue, ce qui suggère que ce mode de convergence ne sera pas bien adapté à ce contexte.

Dans ce travail, nous avons intégré des perspectives provenant de diverses études sur la convergence, observant que dans certains articles, la convergence était analysée sans prendre pleinement en compte le rôle des probabilités de transition. Un défi majeur consistait à bien comprendre les preuves de convergence et à les adapter, en particulier lorsque les arguments sous-jacents n'étaient pas immédiatement clairs. Notre attention sur les probabilités de transition a fourni une perspective supplémentaire sur la dernière définition de la convergence, mettant en lumière des aspects qui ont pu être négligés dans les approches précédentes.