

Research Article

Dynamics and Stability of the Sampling Distribution of Particle Swarm Optimisers via Moment Analysis

Riccardo Poli

Department of Computing and Electronic Systems, University of Essex, Wivenhoe Park, Colchester CO4 3SQ, UK

Correspondence should be addressed to Riccardo Poli, rpoli@essex.ac.uk

Received 23 July 2007; Accepted 3 December 2007

Recommended by T. Blackwell

For stochastic optimisation algorithms, knowing the probability distribution with which an algorithm allocates new samples in the search space is very important, since this explains how the algorithm really works and is a prerequisite to being able to match algorithms to problems. This is the only way to beat the limitations highlighted by the no-free lunch theory. Yet, the sampling distribution for velocity-based particle swarm optimisers has remained a mystery for the whole of the first decade of PSO research. In this paper, a method is presented that allows one to exactly determine all the characteristics of a PSO's sampling distribution and explain how it changes over time during stagnation (i.e., while particles are in search for a better personal best) for a large class of PSO's.

Copyright © 2008 Riccardo Poli. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

1. INTRODUCTION

Let us consider the basic form of PSO with inertia weight, which is controlled by the following two equations:

$$\begin{aligned}v_{t+1}^i &= wv_t^i + \phi_1 \otimes (y^i - x_t^i) + \phi_2 \otimes (\hat{y} - x_t^i), \\x_{t+1}^i &= x_t^i + v_{t+1}^i,\end{aligned}\quad (1)$$

where ϕ_1 and ϕ_2 represent vectors of random numbers uniformly distributed in $[0, c_1]$ and $[0, c_2]$, respectively, (c_1 and c_2 are nonnegative constants), \otimes is component wise multiplication, y^i is the position of the i th particle's personal best, \hat{y} is the position of the particle's neighbourhood best, and w is a constant in $[0, 1]$. Following Kennedy's graphical examinations of the trajectories of individual particles and their responses to variations in key parameters [1], traditionally, this model (and the related PSO with constriction [2], which is, provably, mathematically equivalent) has been studied theoretically under strong simplifying assumptions such as isolated single individuals, search stagnation (i.e., no improved solutions are found) and *absence of randomness* [2–12]. While these assumptions make it possible to make theoretical progress, they also make the results of the analyses difficult to carry over to the real PSO (where stochasticity is present).

Only very few attempts to understand the behaviour of the PSO in the *presence of stochasticity* have been made. For example, Clerc [13] analysed the distribution of velocities of one particle controlled by the standard PSO update rule with inertia and stochastic forces and was able to show that a particle's new velocity is the sum of three components: a forward force, a backward force, and noise. Kadiramanathan et al. [14] were able to study the stability of particles in the presence of stochasticity by using Lyapunov stability analysis.

These are good and important first steps in the direction of modelling the real PSO. However, they only marginally scratch the surface in relation to one of the most important open problems in PSO research: the characterisation of the PSO's sampling distribution and its changes over time. Why is this question so fundamental? For two reasons. Firstly, this scientific knowledge gives us much greater understanding of an algorithm's behaviour. Secondly, on a more practical level, the only way to beat the limitations implied by NFL [15] is to match algorithms to problems. Naturally, at least in part, this can be achieved by a trial-and-error approach involving testing different algorithms and parameter settings on a particular problem or class of problems until a suitable match is found. A more analytic approach to this problem, however, requires two things: (a) knowing the search distribution of an algorithm and (b) checking whether this is well matched to

the features of a problem's fitness landscape. Unfortunately, the sampling distribution of the PSO has been unavailable during the whole first decade (and beyond) of PSO research.

In recent work [16], we introduced a novel method, which allowed us to exactly determine the mean and standard deviation of the sampling distribution of the canonical PSO as well as their changes over any number of generations. The only assumption we made was stagnation, that is, we studied the sampling distribution produced by particles in search for a better personal best. In this paper, we generalise the technique and show how it can be used to determine any number of moments of the sampling distribution of PSO's.

The paper is organised as follows. In Section 2, we provide a summary of the results presented in [16]. In Section 3, we generalised the method to moments of any order. In Section 4, we apply the technique to compute the skewness and kurtosis of the canonical PSO's sampling distribution. In Section 5, we further extend the generality of our technique so as to cover a variety of PSO's, not just the one in (1). This allows us to compare, on a theoretical basis, different forms of PSO. In Section 6, we then use the moments of the PSO's sampling distribution to approximately reconstruct the distribution itself, making it possible, for the first time, to understand the search strategy implemented by the canonical PSO, while searching for new improvements. We discuss the results and conclude in Section 7.

2. DYNAMICS OF FIRST- AND SECOND ORDER MOMENTS OF THE PSO'S SAMPLING DISTRIBUTION

2.1. Derivation of dynamic equations for the moments

During stagnation, each particle behaves independently. Also, each dimension is treated independently. So we can analyse each particle's behaviour in isolation. Therefore, we can drop the superscript i in (1), and rewrite them as

$$x_{t+1} = x_t(1 + w) - x_t(\phi_1 + \phi_2) - wx_{t-1} + \phi_1 y + \phi_2 \hat{y}, \quad (2)$$

where we used the relation $v_t = x_t - x_{t-1}$. Note that on the r.h.s. of this equation, while w , y , and \hat{y} are constants, x_t , ϕ_1 , ϕ_2 , and x_{t-1} are stochastic variables. As a consequence, x_{t+1} on the l.h.s. of the equation is a stochastic variable too (in fact, it is a stochastic function of the stochastic variables appearing in the r.h.s.).

In [16], we applied the expectation operator to both sides of the equation obtaining

$$E[x_{t+1}] = E[x_t](1 + w) - E[x_t](E[\phi_1] + E[\phi_2]) - wE[x_{t-1}] + E[\phi_1]y + E[\phi_2]\hat{y}, \quad (3)$$

where we performed the substitution $E[x_t\phi_i] = E[x_t]E[\phi_i]$ because of the statistical independence between ϕ_i and x_t . Assuming that $c_1 = c_2 = c$, that is, the ϕ_1 and ϕ_2 are both uniformly distributed in $[0, c]$, we have $E[\phi_i] = c/2$, and so

$$E[x_{t+1}] = E[x_t](\rho - c) - wE[x_{t-1}] + c\frac{y + \hat{y}}{2}, \quad (4)$$

where we renamed $(1 + w) = \rho$.

If we square both sides of (2), we obtain

$$x_{t+1}^2 = (x_t\rho - x_t\phi_1 - x_t\phi_2 - wx_{t-1} + \phi_1 y + \phi_2 \hat{y})^2. \quad (5)$$

Expanding the r.h.s., we obtain an equation expressing the stochastic variable x_{t+1}^2 as a function of other stochastic variables, that is, x_t^2 , $x_{t-1}x_t$, ϕ_1^2 , and so forth. By applying the expectation operator to both sides of this equation, one obtains

$$\begin{aligned} E[x_{t+1}^2] &= E[x_t^2](\rho^2 - 4\mu\rho + 2\nu + 2\mu^2) \\ &\quad + E[x_{t-1}x_t](-2w\rho + 4w\mu) + E[x_{t-1}^2](w^2) \\ &\quad + E[x_t](2\mu y\rho + 2\mu\hat{y}\rho - 2\nu y - 2\mu^2\hat{y} - 2\mu^2 y - 2\nu\hat{y}) \\ &\quad + E[x_{t-1}](-2\mu y w - 2\mu\hat{y} w) + \nu y^2 + 2\mu^2 y\hat{y} + \nu\hat{y}^2, \end{aligned} \quad (6)$$

where we set $\mu = E[\phi_i] = c/2$ and $\nu = E[\phi_i^2] = c^2/3$, for brevity.

If, instead, we multiply both sides of (2) by x_t and apply the expectation operator, we obtain

$$E[x_{t+1}x_t] = E[x_t^2](\rho - c) - wE[x_t x_{t-1}] + E[x_t]c\frac{y + \hat{y}}{2}. \quad (7)$$

Equations (4), (6), and (7) form a system of coupled second-order difference equations using which, given appropriate initial conditions, one can compute $E[x_t]$, $E[x_t^2]$ and $E[x_t x_{t-1}]$ for any t .

2.2. Analysis of the dynamics

Equations (4), (6), and (7) can also be written in matrix notation as the extended first order system

$$\mathbf{z}(t+1) = M\mathbf{z}(t) + \mathbf{b}, \quad (8)$$

where

$$\begin{aligned} \mathbf{z}(t) &= \left(E[x_t] \quad E[x_{t-1}] \quad E[x_t^2] \quad E[x_{t-1}^2] \quad E[x_t x_{t-1}] \right)^T, \\ M &= \begin{pmatrix} \rho - c & -w & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ A1 & A2 & A3 & w^2 & 2w(2\mu - \rho) \\ 0 & 0 & 1 & 0 & 0 \\ c\rho & 0 & \rho - c & 0 & -w \end{pmatrix}, \\ \mathbf{b} &= \begin{pmatrix} c\rho \\ 0 \\ \nu y^2 + 2\mu^2 y\hat{y} + \nu\hat{y}^2 \\ 0 \\ 0 \end{pmatrix}, \end{aligned} \quad (9)$$

where $A1 = 4\rho(\mu\rho - \nu - \mu^2)$, $A2 = -4\mu w\rho$, $A3 = \rho^2 - 4\mu\rho + 2\nu + 2\mu^2$.

It is then trivial to verify under what conditions $E[x_t]$, $E[x_t^2]$, and $E[x_t x_{t-1}]$ will converge to stable fixed points. We need to have that all eigenvalues of M must be within the unit

circle, that is, $\Lambda_m = \max_i |\lambda_i| < 1$. When this happens, we will say that the PSO is *order-2 stable*.

Naturally, when $\Lambda_m < 1$, in principle, we could derive (either symbolically or numerically) the fixed point for the system, which we will denote as \mathbf{z}^* . This would be simply given by

$$\mathbf{z}^* = (I - M)^{-1} \mathbf{b}. \quad (10)$$

When the system is order-2 stable, by the simple change of variables $\tilde{\mathbf{z}}(t) = \mathbf{z}(t) - \mathbf{z}^*$, we can represent the dynamics of the system via the following linear homogeneous equation:

$$\tilde{\mathbf{z}}(t+1) = M\tilde{\mathbf{z}}(t), \quad (11)$$

which can trivially be integrated to obtain the following explicit solution for the dynamics of the first two moments:

$$\tilde{\mathbf{z}}(t) = M^t \tilde{\mathbf{z}}(0). \quad (12)$$

3. HIGHER-ORDER MOMENTS

In [16], we obtained recursions which describe the dynamics of first- and second-order moments of the sampling distribution of a standard PSO during stagnation, as summarised in the previous section. One may then wonder whether it would be possible to follow a similar approach to study the dynamics of higher-order moments.

3.1. Derivation of dynamic equations for higher-order moments

The fundamental question is “what quantities would we have to deal with if we took higher powers of both sides of (2) as we did to derive (5)?” Generally, the r.h.s. would be a sum of terms of the form

$$a_0 x_t^{a_1} x_{t-1}^{a_2} w^{a_3} \phi_1^{a_4} \phi_2^{a_5} y^{a_6} \hat{y}^{a_7}, \quad (13)$$

where a_k are suitable constants. Naturally, taking powers of (2) and then multiplying both sides by some power of x_t would also lead to equations involving terms such as those in (13). That is, for any choice of $b_1 \in \mathbb{N}$ and $b_2 \in \mathbb{N}$,

$$x_{t+1}^{b_1} x_t^{b_2} = \sum_i a_0 x_t^{a_1} x_{t-1}^{a_2} w^{a_3} \phi_1^{a_4} \phi_2^{a_5} y^{a_6} \hat{y}^{a_7}, \quad (14)$$

where a_{k_i} are suitable constants. If we then take expectations for both sides, we obtain

$$E[x_{t+1}^{b_1} x_t^{b_2}] = \sum_i a_0 w^{a_3} y^{a_6} \hat{y}^{a_7} E[\phi_1^{a_4}] E[\phi_2^{a_5}] E[x_t^{a_1} x_{t-1}^{a_2}], \quad (15)$$

where we used the independence of ϕ_1 , ϕ_2 , $x_t x_{t-1}$, and, of course, their powers. If ϕ_1 and ϕ_2 are uniformly distributed in the same range, namely, $[0, c]$, it is easy to verify that

$$E[\phi_j^n] = \frac{c^n}{n+1}. \quad (16)$$

So

$$E[x_{t+1}^{b_1} x_t^{b_2}] = \sum_i \omega_i E[x_t^{a_1} x_{t-1}^{a_2}], \quad (17)$$

where

$$\omega_i = \left(\frac{a_0 w^{a_3} c^{a_4+a_5} y^{a_6} \hat{y}^{a_7}}{(1+a_4)(1+a_5)} \right). \quad (18)$$

It is important to note here that because (2) is linear in x_t and x_{t-1} , all the terms on the r.h.s. of (14) and (17) respect the relation $a_1 + a_2 \leq b_1 + b_2$. This implies that *it is possible to construct recursions for moments of arbitrary order*.

3.2. Complexity of the calculations

If one wanted to push the analysis developed in Section 2 up to order 3, one would need to instantiate (17) for $E[x_{t+1}^3]$, $E[x_{t+1}^2 x_t]$, $E[x_{t+1} x_t^2]$ and add the resulting equations to (4), (6), and (7). If one wanted to go to order 4, an additional set of four equations (for $E[x_{t+1}^4]$, $E[x_{t+1}^3 x_t]$, $E[x_{t+1}^2 x_t^2]$, and $E[x_{t+1} x_t^3]$) would be needed, bringing the total to 10.

More generally, in order to compute statistics of order n , one needs to construct and iterate

$$Q(n) = \frac{n \times (n+1)}{2} \quad (19)$$

second-order difference equations. Since, after expansion, the r.h.s. of (2) contains 7 atomic terms of the form in (13), the r.h.s. of (17) contains 7^{b_1} such terms (Note that the exponent b_2 does not influence the number of terms because the recursion for $E[x_{t+1}^{b_1} x_t^{b_2}]$ is obtained as follows: (a) we compute $x_{t+1}^{b_1}$, which is given by an expression containing 7^{b_1} terms; (b) we multiply each term by $x_t^{b_2}$, which changes the exponents a_i but does not alter the number of terms; (c) we apply the expectation operator, which again does not modify the number of terms.) So the total number of terms one needs to compute to construct the equations for order- n statistics is 7 (for the order 1 equations) plus $(7^2 + 7)$ (for the order 2 equations) plus $(7^3 + 7^2 + 7)$ (for the order 3 equations), and so forth. This gives us a total of

$$T(n) = \sum_{i=1}^n (n-i+1) \times 7^i \quad (20)$$

terms. Note that $T(n)$ grows exponentially approximately as 1.36×7^n . So although the number of equations one needs to deal with grows quadratically, the computational effort required to instantiate them is exponential. For instance, $T(1) = 7$, $T(2) = 63$, $T(3) = 462$, $T(4) = 3262$, and $T(5) = 22869$. The growth in number of terms can be reduced if one makes explicit use of ρ (i.e., by adding the factor ρ^{a_8} in (13)). Then $T(n) = O(6^n)$. Either way, manually deriving equations for moments of order 3 is already very laborious. The process, however, is clearly mechanisable. This can be done using computer algebra systems, or by explicitly representing and manipulating the ω_i 's in each equation (this is what we did). As a result of mechanisation, computing the equations for up to order 6 or 7 is feasible with an ordinary personal computer.

Some of the ω_i 's in (17) present the same pattern of exponents for w , c , y , and \hat{y} , so terms can be collected leading to more compact equations (e.g., compare (5) and (6)). Also,

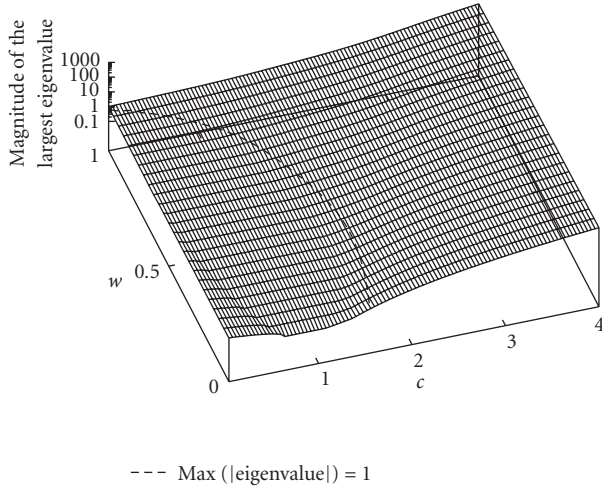


FIGURE 1: Magnitude of the largest eigenvalue of M_4 as a function of the parameters w and c . The curved line on the surface encloses the order-4 stable region.

given their size, one will normally want to study (e.g., integrate) (17) numerically. In this case, w , c , γ , and \hat{y} are all numerical parameters. So the ω_i 's become constants and, after collecting terms, each equation contains at most $Q(n)$ terms, which, as we know, is quadratic in the order n . As a result, although the complexity of the construction of the motion equations for the moments is exponential in the order of the moments, their numerical integration is only of order $O(n^4)$.

Naturally, the system of $Q(n)$ second-order difference equations necessary to predict the dynamics of moments of order 1 to n can be turned into a system of order 1 of the form in (8), via the choice

$$z(t) = \begin{pmatrix} E[x_t] \\ E[x_{t-1}] \\ E[x_t^2] \\ E[x_t x_{t-1}] \\ E[x_{t-1}^2] \\ E[x_t^3] \\ E[x_t^2 x_{t-1}] \\ E[x_t x_{t-1}^2] \\ E[x_{t-1}^3] \\ \vdots \\ E[x_{t-1}^n] \end{pmatrix}. \quad (21)$$

This effectively means adding artificial update equations of the form $E[x_t^k] = E[x_{t-1}^k]$ for $k = 1, \dots, n$, bringing the total to $Q'(n) = Q(n) + n$. The transition matrix for the system is, therefore, of size $Q'(n) \times Q'(n)$. We will denote this with M_n . For example, for $n = 4$, which would allow one to study the mean, variance, skewness, and kurtosis of the sampling distribution as a function of t , M_4 is merely a 14×14 matrix.

Interestingly, $Q'(n)$ grows so slowly that one can perform an eigenvalue analysis for any M_n that one is able to compute.

That is, the expensive part of the process is the construction of M_n . Once this is done, iterating the system, establishing its stability, or finding its fixed points is a trivial matter.

In the next section, we provide results for statistics of order 3 and 4, that is, $n = 4$, a value of n for which computing M_n takes only a few seconds. However, before we do this, we need to consider the initial conditions for the system.

3.3. Initial conditions

In order to provide initial conditions for the moments' dynamics, we need to compute $E[x_0^k]$ and $E[x_1^k x_0^l]$ for generic $k > 0$ and $l \geq 0$.

If we are at the very first iteration of the PSO algorithm, the initial conditions for the dynamics of the moments are related to the ranges used to initialise positions and velocities in the PSO. Under the assumption that a particle's initial position, x_0 , is chosen uniformly at random in a symmetric range $[-\Omega, \Omega]$, we have

$$E[x_0^k] = \begin{cases} 0 & \text{if } k \text{ is odd,} \\ \frac{\Omega^k}{k+1} & \text{otherwise.} \end{cases} \quad (22)$$

In order to compute $E[x_1^k x_0^l]$, we need to consider the equation

$$x_1 = x_0 + wv_0 - x_0(\phi_1 + \phi_2) + \phi_1 y + \phi_2 \hat{y}, \quad (23)$$

where a particle's initial velocity, v_0 , is a stochastic variable uniformly distributed the range $[-\Omega_v, \Omega_v]$ (often $\Omega_v = \Omega$). By taking the k th power of both sides of the equation, multiplying by x_0^l , and taking expectations, as we did to construct (17), one obtains the desired expressions for $E[x_1^k x_0^l]$. Like for (17), these expressions contain a number of terms that grows exponentially with n . However, this process, too, can be trivially mechanised.

If the PSO is not at its first iteration, and a new improvement has just been found to a particle's personal best or neighbourhood best, then x_0 is not a stochastic variable, but a constant determined by the position where the particle found the last improvement.

4. SKEWNESS AND KURTOSIS OF THE PSO'S SAMPLING DISTRIBUTION

We constructed the recursions for moments of up to order 4 as described in the previous section for the canonical PSO. Naturally, for higher-order moments, we can do exactly the same type of eigenvalue analysis we did in [16] for the mean and standard deviation of the sampling distribution. For example, Figure 1 shows the magnitude of the largest eigenvalue, Λ_m , of M_4 as a function of the parameters w and c . The system is order-4 stable, that is, mean, variance, skewness, and kurtosis have a stable fixed point, whenever $\Lambda_m < 1$, that is, within the curved region enclosed by the contour shown in the figure.

For easier comparison, we show the lines where $\Lambda_m = 1$ for M_1 , M_2 , M_3 , and M_4 in Figure 2 (ordered from top to

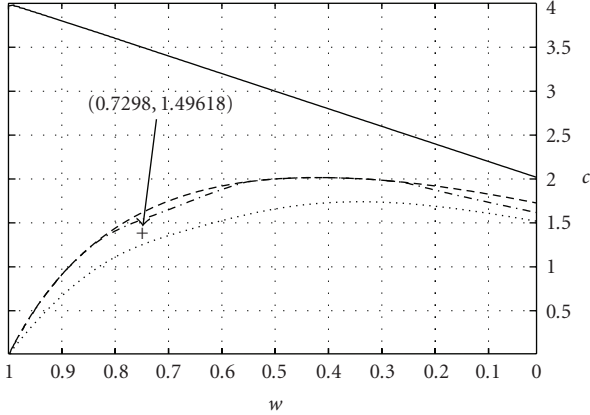
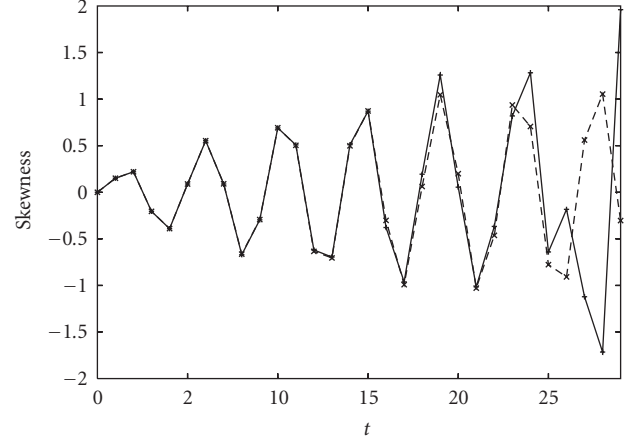


FIGURE 2: Plot of the regions of order-1, -2, -3, and -4 stability for the canonical PSO.

bottom, resp.). The regions of order-1, -2, -3, and -4 stability are nested. Note how the $\Lambda_m = 1$ lines for M_2 and M_3 coincide for many values of w . Note also that the standard setting, $c = 1.49618$ and $w = 0.7298$, lays within the narrow region of order-3 stability. This implies that while mean, variance, and skewness of the standard PSO tend to a fixed point, kurtosis is unstable and will tend to grow indefinitely. Interestingly, a growth in the kurtosis of samples was observed by Kennedy [17], although this was effectively computed under the assumption that the sampling distribution is time independent. So the values of x_t recorded in a run at t grows were treated as different samples from the same distribution, while we now know this is incorrect.

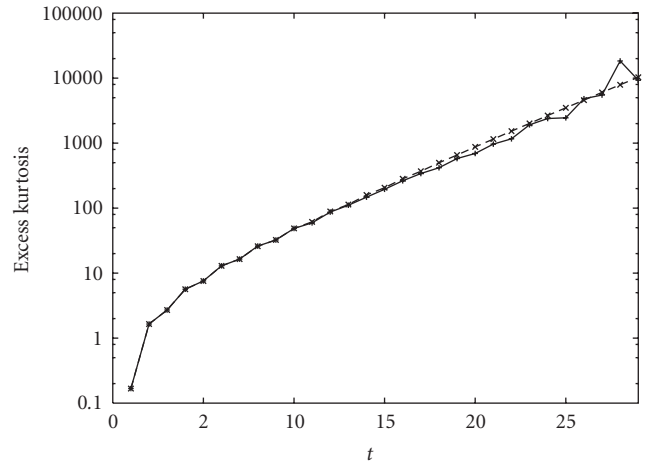
That the predictions of the model are exact is also confirmed by the comparison of the dynamics of predicted and recorded higher-order moments. Figure 3(a) shows a comparison between the skewness $E[(x_t - \mu_t)^3]/\sigma_t^3$ computed using our model and the average positions of the particle recorded in one billion (1 000 000 000) real runs in the first 30 iterations for the case $c = 1.49618$, $w = 0.7298$, $y = 0$, $\hat{y} = 1$, and $\Omega = 5$. As one can see, there is a very good match between the model’s predictions and the stagnation behaviour of particles in real runs. Only after about 20–25 generations, the sampling errors start accumulating enough to show significant differences. We know that these differences are due to sampling errors because before performing 1 000 000 000 runs, we attempted to use much smaller sets of runs. With fewer runs, we observed that errors were visible well before generation 25. For example, with 1 000 000 runs, there is a good match up to around iteration 15.

As shown in Figure 3(b), the model also predicts very well the behaviour of the (excess) kurtosis $E[(x_t - \mu_t)^4]/\sigma_t^2 - 3$ of the sampling distribution. (Following standard practice, in this paper whenever we use the term “kurtosis,” we will refer to the excess kurtosis $E[(x_t - \mu_t)^4]/\sigma_t^2 - 3$. The excess kurtosis of the normal distribution to 0.) Note that for $c = 1.49618$ and $w = 0.7298$, the system is order-3 stable, and so, although the oscillations of the skewness shown in Figure 3(a) appear to grow bigger and bigger, suggesting instability, this



—+— Skewness real runs
-*- Predicted skewness

(a)



—+— Kurtosis real runs
-*- Predicted kurtosis

(b)

FIGURE 3: Comparison between predicted and experimental skewness and (excess) kurtosis of the PSO sampling distribution for $c = 1.49618$, $w = 0.7298$, $y = 0$, $\hat{y} = 1$, and $\Omega = 5$. Kurtosis grows exponentially, and so it is plotted on a logarithmic scale. The first point is not plotted because the excess kurtosis was negative (-1.2).

is actually only a transient effect, as shown in Figure 4 where we integrate the equations over 200 generations instead of 30.

5. COMPARISON BETWEEN PSO’S

In the previous sections, we studied the canonical PSO with the restriction that the acceleration coefficients, c_1 and c_2 , were identical: $c_1 = c_2 = c$. One may wonder, however, whether allowing such coefficients to differ would produce qualitatively very different dynamics. For example, what would happen if we set one of the c_i ’s to zero as in a purely cognitive or purely social PSO model? This effectively would reduce to one the sources of random influences on a parti-

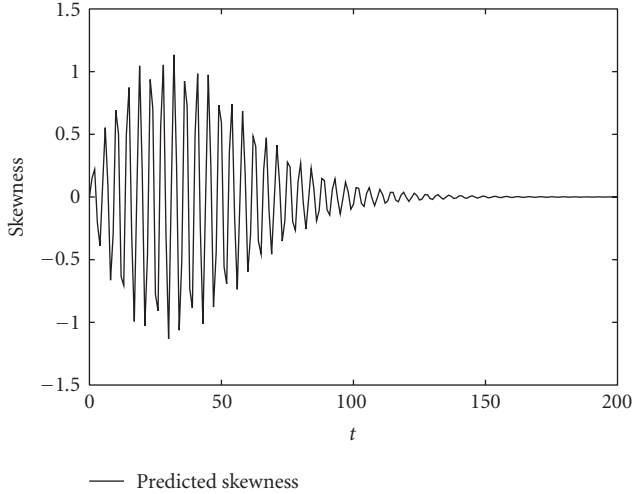


FIGURE 4: Predicted dynamics of the skewness of the PSO sampling distribution over 200 generations for $c = 1.49618$, $w = 0.7298$, $y = 0$, $\hat{y} = 1$, and $\Omega = 5$.

cle's dynamics. Conversely, one might wonder what would happen if we increased the number of such sources of influence, as is done, for example, in the fully informed particle swarm (FIPS) [18, 19].

5.1. Moments of generalised PSO's

To answer these (and other) important questions on the sampling distribution of different PSO models, we adopt a FIPS-like general class of PSO's described by the following difference equation:

$$x_{t+1} = x_t + wv_t + \sum_{i=1}^m \phi_i(\hat{y}_i - x_t), \quad (24)$$

where the ϕ_i 's are stochastic variables uniformly distributed in the range $[0, c_i]$, c_i are constants, and the \hat{y}_i 's are the personal best positions of neighbours of the particle (the particle itself may be included in its own neighbourhood). Naturally, this equation can be converted into the following:

$$x_{t+1} = x_t(1 + w) - wx_{t-1} - \sum_{i=1}^m \phi_i x_t + \sum_i \phi_i \hat{y}_i, \quad (25)$$

which is a generalisation of (2). All of the steps we performed in Section 3 can be repeated for (25). These lead to recursions of the form in (17) with the only difference that the coefficients ω_i take the more general form

$$\omega_i = \left(\frac{a_0 w^{a_{w_i}} c_1^{a_{c_1}} \cdots c_m^{a_{c_m}} \hat{y}_1^{a_{y_1}} \cdots \hat{y}_m^{a_{y_m}}}{\prod_{j=1}^m (1 + a_{c_j})} \right), \quad (26)$$

where a_0 , a_{w_i} , a_{c_1}, \dots, a_{c_m} , a_{y_1}, \dots, a_{y_m} are appropriate constants.

Because (25) contains $3 + 2 \times m$ terms, the complexity of the expansion now grows as $O((3 + 2 \times m)^n)$, where n is

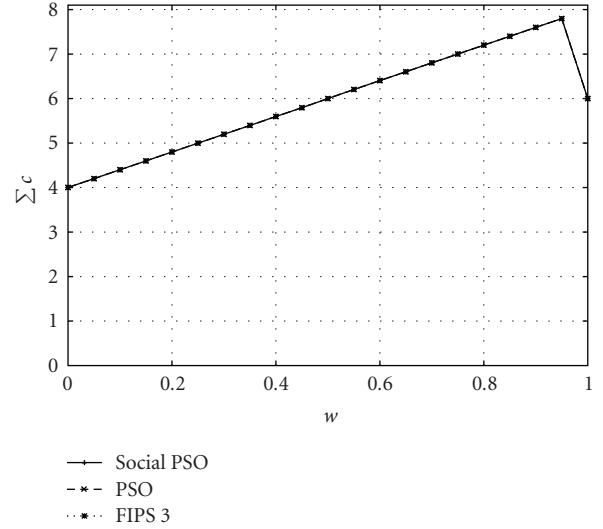


FIGURE 5: Lines below which the mean of the sampling distributions for a social PSO, a canonical PSO, and FIPS with a neighbourhood of three individuals have a fixed point (order-1 stability). The three lines coincide.

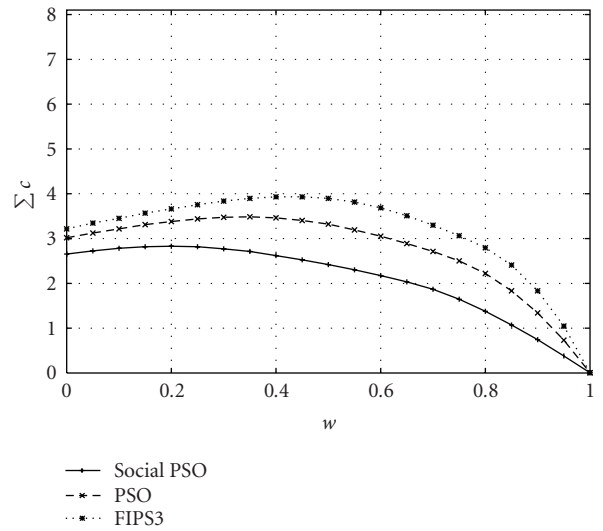


FIGURE 6: Lines below which the variance of the sampling distributions for a social PSO (bottom), a canonical PSO (middle), and FIPS with a neighbourhood of three individuals (top) have a fixed point (order-2 stability).

the order of the moments we are interested in. So the larger m , the heavier the computation load required to compute M_n . Once the transition matrices M_n are computed, however, they are exactly of the same size for all PSO models within the class defined by (24). Initial conditions can be found following the approach described in Section 3. Calculations are expensive but can be mechanised. We did this for the examples described below.

5.2. Three PSO's we want to compare

An extensive comparison of different PSO's is beyond the scope of this paper. However, as an example of the kind of

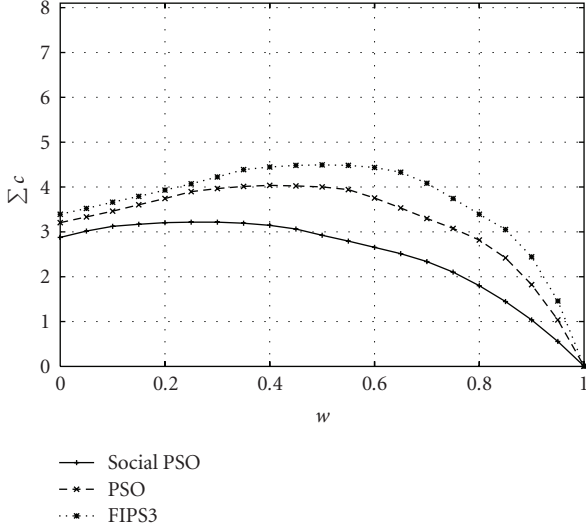


FIGURE 7: Lines below which a social PSO, a canonical PSO, and FIPS3 are order-3 stable (from bottom to top, resp.).

comparisons one can make using our approach, we considered the PSO's in (24) with $N = 3$. Within this class of PSO's, we considered three variants as follows:

- a purely social variant of PSO, which we will call *social PSO* for brevity, where $c_1 > 0$ and $c_2 = c_3 = 0$ (due to symmetries, the behaviour of a purely cognitive PSO where $c_2 > 0$ and $c_1 = c_3 = 0$ is effectively identical to that of this social PSO);
- the *canonical PSO* we have studied so far in the paper, which is obtained by setting $c_1 = c_2 > 0$ and $c_3 = 0$;
- the simplest version of FIPS with a neighbourhood of three individuals, for example, obtained using an *lbest* topology and an interaction radius of 1, where $c_1 = c_2 = c_3 > 0$. We will call this version *FIPS3*.

In order to perform a fair comparison of the stability properties of these PSO variants, we study them in conditions where the sum of the amplitudes of the random components, ϕ_i , is identical across models. That is, we set $c = \sum_i c_i$, and compare models with the same c value. Again, we analyse eigenvalues.

5.3. Results of the comparison

Figures 5–8 show the lines in the (w, c) plane where the magnitude of the largest eigenvalue of M_n , Λ_m is 1 for $n = 1, 2, 3, 4$ and for the three PSO variants mentioned above. Each figure represents a different moment. In the plot for the n th moment, the region below each line represents the set of all (w, c) pairs for which the n th moment of the distribution of the PSO corresponding to that line is stable. Let us analyse these figures in detail.

Firstly, we should note that the regions of order-1 stability for the three models are identical. This is because the dy-

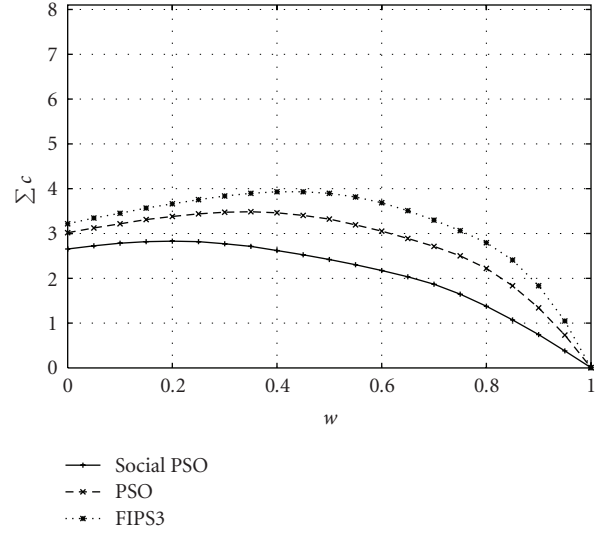


FIGURE 8: Lines below which the kurtosis of the sampling distributions for a social PSO (bottom), a canonical PSO (middle), and FIPS3 (top) have a fixed point (order-4 stability).

namics of the mean of the three models is governed by equations of the same form, namely,

$$E[x_{t+1}] = E[x_t] \left(1 + w - \frac{c}{2}\right) - wE[x_{t-1}] + \text{constant}, \quad (27)$$

where the constant term may differ in different PSO variants. (This is irrelevant for the stability of the system, since stability is determined by the homogeneous part of the equation.) Note also that the rightmost point in each plot is an artifact due to the fact that, at $w = 1$, $\Lambda_m = 1$ for M_1 irrespective of the value of c .

The regions where the variance is stable for the three models, instead, are different, with FIPS3 having the largest region of order-2 stability, followed by the canonical PSO, and finally, by the social PSO. Exactly the same happens with skewness (Figure 7) and kurtosis (Figure 8), with the order-3 stability region largely coinciding with the order-2 ones also for FIPS3 and the social PSO.

These results are counter intuitive. One would expect that the more sources of randomness, the ϕ_i 's, there are, the more a PSO should be unstable. However, the exact opposite happens. The social PSO, where the only influence is ϕ_1 , is the least stable of all models, while FIPS3, which has three sources of randomness, is the most stable. What are the reasons for this behaviour?

We can understand this by rewriting (25) as follows:

$$x_{t+1} = x_t(1 + w) - wx_{t-1} - x_t\Phi_m + \Psi_m, \quad (28)$$

where $\Phi_m = \sum_{i=1}^m \phi_i$ and $\Psi_m = \sum_i \phi_i \hat{y}_i$. Both Φ_m and Ψ_m are the sum of independent and uniformly distributed variables: the variables ϕ_i in the case of Φ_m and the variables $\hat{y}_i \phi_i$ in the case of Ψ_m .

We know that $\sum_i c_i = c$. To simplify our treatment, let us further assume that the ϕ_i 's are i.i.d., that is, that all c_i are identical, and so $c_i = c/m$. We can then apply the central limit

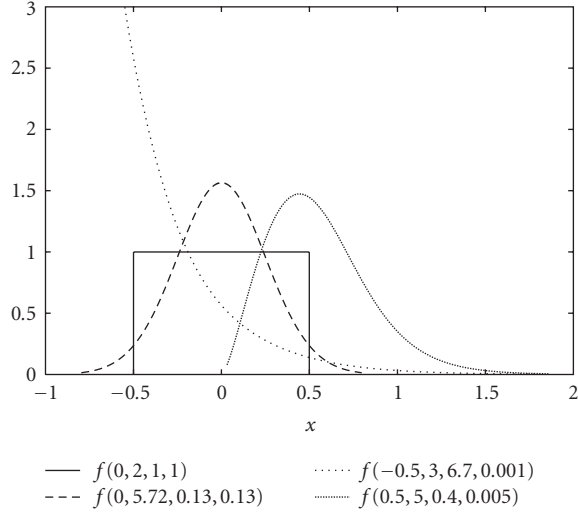


FIGURE 9: Examples of GLD density functions.

theorem to Φ_m . This predicts that for sufficiently large m , the distribution of Φ_m is approximately Gaussian with mean $\sum_i c_i/2 = c/2$ and variance $\sum_i (c_i^2/3 - (c_i/2)^2) = \sum_i c_i^2/12 = c^2/(12m)$. So the larger m , the smaller the variance of Φ_m and the stochasticity of (28).

In the case of the stochastic variable Ψ_m , the quantities $\phi_i \hat{y}_i$ are not identically distributed even if all c_i are identical. This is because, in principle, each \hat{y}_i may be different. This prevents the use of the standard central limit theorem. We can, however, apply Lyapunov's central limit theorem to Ψ_m . The conditions for its application are as follows:

- (1) the variables $\phi_i \hat{y}_i$ must have finite mean, which is the case since $\mu_i = E[\phi_i \hat{y}_i] = c_i \hat{y}_i/2 = c \hat{y}_i/(2m)$,
- (2) the variables $\phi_i \hat{y}_i$ must have finite variance, which, again, is the case since $\sigma_i^2 = E[(\phi_i \hat{y}_i - \mu_i)^2] = (c_i \hat{y}_i)^2/12 = (c \hat{y}_i/m)^2/12$,
- (3) the variables $\phi_i \hat{y}_i$ must have finite third central moment, which is satisfied since $r_i^3 = E[(\phi_i \hat{y}_i - \mu_i)^3] = 0$, and finally,
- (4) the Lyapunov condition, $\lim_{m \rightarrow \infty} ((\sum_{i=1}^m r_i^3)^{1/3} / (\sum_{i=1}^m \sigma_i^2)^{1/2}) = 0$, must be satisfied, which, again, is the case since all $r_i^3 = 0$.

Then for sufficiently large m , also the distribution of Ψ_m is approximately Gaussian with mean $\sum_i c \hat{y}_i/(2m) = (c/2) \times (\sum_i \hat{y}_i/m)$ and variance $(c^2/12m) \times (\sum_i \hat{y}_i^2/m)$. Note that $\sum_i \hat{y}_i/m$ and $\sum_i \hat{y}_i^2/m$ are the mean \hat{y}_i and the mean \hat{y}_i^2 , respectively. So these are finite quantities if, as is normally the case, all \hat{y}_i are finite. (PSO search is normally confined to a prefixed, finite region of \mathbb{R}^N , and so all \hat{y}_i must be finite.) So like for Φ_m , the larger m , the smaller the variance of Ψ_m , and consequently, the less the stochasticity of (28).

Effectively, the larger m , the more Φ_m and Ψ_m become deterministic and approach constant values. This explains why adding more and more sources of randomness—while keeping c constant—produces progressively more and more stable PSO's.

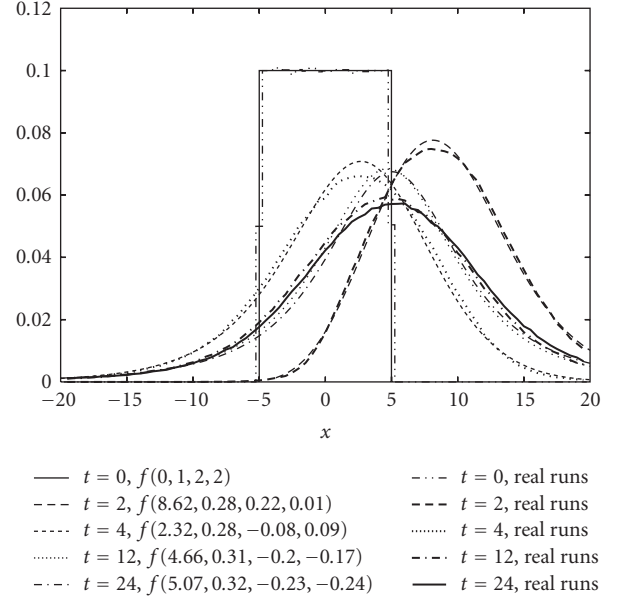


FIGURE 10: Estimates of the sampling distribution of a canonical PSO with parameters $c = 1.49618$, $w = 0.7298$, $\gamma = 0$, $\hat{\gamma} = 10$, and $\Omega = 5$ during stagnation, reconstructed via GLD best fitting versus histograms over 1 000 000 real runs. Snapshots at times $t = 0, 2, 4, 12$, and 24 are shown. For each theoretical sampling distribution, we report the parameters of the corresponding GLD.

6. THE DENSITY FUNCTION OF THE CANONICAL PSO'S SAMPLING DISTRIBUTION

The technique described in this paper, in principle, would allow one to determine all the moments of the sampling distribution of the PSO at all times. The question then is “could we derive the PSO sampling distribution itself?” The answer is of course in the positive since knowing all the moments of a distribution implies knowing its moment generating function. This, in turn, allows one to obtain the density function of the distribution via inverse Laplace transform.

In practice, however, it is impossible to compute all the moments of the PSO sampling distribution. This is for two reasons. Firstly, there are infinitely many such moments. Secondly, as we have seen in the previous sections, the cost of computing moments is exponential in the order of the moments. The next question is then “to what extent can we still reconstruct the PSO's density function from a finite number of moments?” This is an instance of the well-known *truncated moment problem*, a difficult, inverse ill-posed problem for which many approaches have been proposed. Here, we consider only one such approach.

A particularly simple idea is to consider a family of density functions $f(x; \lambda_1, \lambda_2, \dots)$ with parameters λ_1, λ_2 , and so forth, with sufficient expressive power to represent distributions with widely different shapes, with more or less asymmetries, with tails of different characteristics, and so on. Then one can use optimisation techniques to identify the parameters of the distribution f which minimise the difference between the moments of f and the moments of the PSO's

sampling distribution. This is called the *moment-matching method*. Once the parameters λ_1 , λ_2 , and so forth, are identified, f can be used as an approximation of the true PSO sampling distribution. This approach to reconstructing probability distributions from moments was proposed [20] (see also [21, 22]), where a *generalised lambda distribution* (GLD) was used. We adopt this same approach here.

GLD is a four-parameter distribution defined via its quantile function as follows:

$$Q(u) = \lambda_1 + \frac{1}{\lambda_2} \left(\frac{u^{\lambda_3} - 1}{\lambda_3} + \frac{1 - (1 - u)^{\lambda_4}}{\lambda_4} \right), \quad (29)$$

where $u \in [0, 1]$. Its density function is given by

$$f(x; \lambda_1, \lambda_2, \lambda_3, \lambda_4) = \left(\frac{dQ(u)}{du} \right)^{-1} = \frac{\lambda_2}{u^{(\lambda_3-1)} + (1-u)^{(\lambda_4-1)}}, \quad (30)$$

where $u = Q^{-1}(x)$.

The GLD is enormously flexible in terms of the shape of the distribution. For example, as shown in Figure 9, the uniform, Gaussian, exponential, and Gamma distributions are all special cases of GLD. Effectively, λ_1 determines the location of the distribution, λ_2 determines its scale, while λ_3 and λ_4 determine other shape characteristics. In particular, only if $\lambda_3 = \lambda_4$, the distribution is symmetric.

Because GLD has 4 parameters, all we need is four moments—the mean, variance, skewness, and kurtosis—in the PSO’s sampling distribution in order to identify such parameters with the moment-matching method described above.

As an illustration, we apply this technique to reconstruct the sampling distribution during stagnation of a canonical PSO with parameters $c = 1.49618$, $w = 0.7298$, $\gamma = 0$, $\hat{y} = 10$, and $\Omega = 5$. With these parameters, $\lim_{t \rightarrow \infty} E[x_t] = 5$. In Figure 10, we show snapshots at times $t = 0, 2, 4, 12$, and 24 of the theoretical sampling distribution together with estimates of the distribution based on 1 000 000 actual runs. In all cases, the match between the moments of the GLD and those of the PSO sampling distribution was exact (within experimental errors). Also, there is considerable agreement between the theoretical lines and histograms obtained in real runs. Note how widely the mean of the density function oscillates around 5 in the first few generations. Also note the asymmetry in the distributions due to the oscillations of the skewness.

7. DISCUSSION AND CONCLUSIONS

Several theoretical analyses of the dynamics of particle swarms have been offered in the literature over the last decade. These have been very illuminating. However, virtually, all have relied on substantial simplifications and on the assumption that the particles are deterministic. Naturally, these simplifications make it impossible to derive an exact characterisation of the sampling distribution of the PSO.

In previous work [16], by using surprisingly simple techniques, we started by exactly determining, perhaps, the most

important characteristic of a PSO’s sampling distribution, its variance, and we have been able to explain how it changes over any number of generations. The only assumption we made is stagnation. Here, we extended this technique to the study of higher-order statistics. In particular, we analysed, in detail, the skewness and kurtosis of the distribution. Because of the complexity of the calculations involved, this required mechanising the derivation of the recursions for these moments.

We applied the analysis to the PSO with inertia weight, but the analysis is also valid for the PSO with constriction, because of the well-known equivalence of these two models (via a simple parameter mapping). We also generalised our model so as to include FIPS. This made it possible to explicitly compare the stability of different forms of PSO, leading to a deeper understanding of their properties. In particular, we showed that while FIPS and standard forms of PSO present exactly the same order-1 stability, in FIPS, higher-order moments are more stable than in the other PSO’s, and we were able to explain why this is the case, using two forms of central limit theorem (Section 5). Elsewhere [23], we also applied the same type of analysis to derive the stability regions of a simpler form of velocity-based PSO.

Finally, with all these tools in hand, we went in search for the “holy grail”—the actual PSO sampling density function. We treated the problem as an ill-posed inverse problem, which we regularised thanks to the use (and best fit) of a family of distributions—the generalised lambda distribution (GLD). All empirical evidence we have suggests that this distribution approximates very closely the sampling behaviour of PSO’s (naturally, with parameters that are functions of time, i.e., $\lambda_i = \lambda_i(t)$). We will attempt to characterise the PSO’s sampling distribution in more depth in future research.

Whether or not GLD is the exact PSO sampling distribution or just a very good approximation, if one could determine (again, either exactly or approximately) how the $\lambda_i(t)$ ’s are affected by the parameters c , w , γ , and \hat{y} and by the initial conditions x_0 and v_0 , it would then be possible to accurately simulate the behaviour of the PSO by sampling from $f(x; \lambda_1(t), \lambda_2(t), \lambda_3(t), \lambda_4(t))$. This is easily done since GLD deviates can trivially be produced by picking u uniformly at random in the interval $[0, 1]$ and applying (29), that is, $Q(U[0, 1])$ is generalised lambda distributed. We will study this more sophisticated form of bare-bones PSO [17] in future research.

Bare-bone PSO’s have some resemblance with estimation of distribution algorithms (EDA’s). EDA’s are powerful population-based searchers where the variation operations traditionally implemented via crossover and mutation in evolutionary algorithms are replaced by the process of sampling from a distribution (a review of EDA’s is available in [24]). The distribution is modified generation after generation, by using information obtained from the more fit individuals in the population. The objective of these changes is to increase the probability of generating individuals with high fitness. As one can clearly see from this description, a bare-bone PSO performs a very similar form of search to an EDA. However, there is a difference between the two. While

in a typical EDA the whole new population is generated from a single distribution, in a bare-bone PSO, each individual has a personal sampling distribution from which only one sample is drawn at each iteration. (The sampling distributions of different particles may share some characteristics, since the sampling distribution of a particle is determined by both personal best and neighbourhood best, and the neighbourhood best can easily be the same for different particles.) Nonetheless, one could still interpret a bare-bone PSO as an unusual type of EDE in which multiple distributions are used and updated based on fitness during the search.

ACKNOWLEDGMENTS

The author would like to acknowledge the support by EP-SRC (Grant no. GR/T11234/01, “Extended Particle Swarms (XPS)”) and the help and comments from Maurice Clerc, David Broomhead, Susanna Wau Men Au-Yeung, and Roberto Renó. The anonymous reviewers and the editors are also thanked for their useful feedback.

REFERENCES

- [1] J. Kennedy, “The behavior of particles,” in *Proceedings of the 7th International Conference on Evolutionary Programming (EP '98)*, V. W. Porto, N. Saravanan, D. Waagen, and A. E. Eiben, Eds., pp. 581–589, Springer, San Diego, Calif, USA, March 1998.
- [2] M. Clerc and J. Kennedy, “The particle swarm—explosion, stability, and convergence in a multidimensional complex space,” *IEEE Transactions on Evolutionary Computation*, vol. 6, no. 1, pp. 58–73, 2002.
- [3] E. Ozcan and C. K. Mohan, “Analysis of a simple particle swarm optimization system,” in *Intelligent Engineering Systems Through Artificial Neural Networks*, vol. 8, pp. 253–258, ASME Press, St. Louis, Mo, USA, 1998.
- [4] E. Ozcan and C. K. Mohan, “Particle swarm optimization: surfing the waves,” in *Proceedings of the IEEE Congress on Evolutionary Computation (CEC '99)*, vol. 3, pp. 1939–1944, IEEE Service Center, Washington, DC, USA, July 1999.
- [5] F. van den Bergh, *An analysis of particle swarm optimizers*, Ph.D. thesis, Department of Computer Science, University of Pretoria, Pretoria, South Africa, 2002.
- [6] A. P. Engelbrecht, *Fundamentals of Computational Swarm Intelligence*, John Wiley & Sons, New York, NY, USA, 2005.
- [7] K. Yasuda, A. Ide, and N. Iwasaki, “Adaptive particle swarm optimization,” in *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, vol. 2, pp. 1554–1559, Washington, DC, USA, October 2003.
- [8] T. M. Blackwell, “Particle swarms and population diversity,” *Soft Computing*, vol. 9, no. 11, pp. 793–802, 2005.
- [9] B. Brandstatter and U. Baumgartner, “Particle swarm optimization—mass-spring system analogon,” *IEEE Transactions on Magnetics*, vol. 38, no. 2, pp. 997–1000, 2002.
- [10] I. C. Trelea, “The particle swarm optimization algorithm: convergence analysis and parameter selection,” *Information Processing Letters*, vol. 85, no. 6, pp. 317–325, 2003.
- [11] E. F. Campana, G. Fasano, and A. Pinto, “Dynamic system analysis and initial particles position in particle swarm optimization,” in *Proceedings of the IEEE Swarm Intelligence Symposium (SIS '06)*, Indianapolis, Ind, USA, May 2006.
- [12] E. F. Campana, G. Fasano, D. Peri, and A. Pinto, “Particle swarm optimization: efficient globally convergent modifications,” in *Proceedings of the 3rd European Conference on Computational Mechanics, Solids, Structures and Coupled Problems in Engineering (ECCM '06)*, C. A. Mota Soares, J. A. C. Martins, H. C. Rodrigues, and J. A. C. Ambrosio, Eds., Lisbon, Portugal, June 2006.
- [13] M. Clerc, “Stagnation analysis in particle swarm optimisation or what happens when nothing happens,” Tech. Rep. CSM-460, Department of Computer Science, University of Essex, Colchester, UK, August 2006.
- [14] V. Kadiramanathan, K. Selvarajah, and P. J. Fleming, “Stability analysis of the particle dynamics in particle swarm optimizer,” *IEEE Transactions on Evolutionary Computation*, vol. 10, no. 3, pp. 245–255, 2006.
- [15] D. H. Wolpert and W. G. Macready, “No free lunch theorems for search,” *IEEE Transactions on Evolutionary Computation*, vol. 1, no. 1, pp. 67–82, 1997.
- [16] R. Poli and D. Broomhead, “Exact analysis of the sampling distribution for the canonical particle swarm optimiser and its convergence during stagnation,” in *Proceedings of the 9th Genetic and Evolutionary Computation Conference (GECCO '07)*, pp. 134–141, ACM Press, London, UK, July 2007.
- [17] J. Kennedy, “Bare bones particle swarms,” in *Proceedings of the IEEE Swarm Intelligence Symposium (SIS '03)*, pp. 80–87, Indianapolis, Ind, USA, April 2003.
- [18] J. Kennedy and R. Mendes, “Population structure and particle swarm performance,” in *Proceedings of the IEEE Congress on Evolutionary Computation (CEC '02)*, vol. 2, pp. 1671–1676, Honolulu, Hawaii, USA, May 2002.
- [19] R. Mendes, *Population topologies and their influence in particle swarm performance*, Ph.D. thesis, Departamento de Informatica, Escola de Engenharia, Universidade do Minho, Braga, Portugal, 2004.
- [20] J. S. Ramberg, P. R. Tadikamalla, E. J. Dudewicz, and E. F. Mykytka, “A probability distribution and its uses in fitting data,” *Technometrics*, vol. 21, no. 2, pp. 201–214, 1979.
- [21] A. Lakhany and H. Mausser, “Estimating the parameters of the generalized lambda distribution,” *Algo Research Quarterly*, vol. 3, no. 3, pp. 47–58, 2000.
- [22] S. W. M. AuYeung, “Finding probability distributions from moments,” M.S. thesis, Imperial College, London, UK, 2003.
- [23] R. Poli, D. Bratton, T. M. Blackwell, and J. Kennedy, “Theoretical derivation, analysis and empirical evaluation of a simpler particle swarm optimiser,” in *Proceedings of the IEEE Congress on Evolutionary Computation (CEC '07)*, IEEE Press, Singapore, September 2007.
- [24] P. Larrañaga and J. A. Lozano, *Estimation of Distribution Algorithms: A New Tool for Evolutionary Computation*, Kluwer Academic Publishers, Boston, Mass, USA, 2002.