Desperately Searching for Something

Clive E. Bowman[1] and Peter Grindrod[1]

[1]Mathematical Institute, University of Oxford, Oxford OX2 6GG UK

September 13, 2022

Abstract

There is a growing interest in *novelty search*: that is, in sampling a *parameter space* to search for radical or unexpected behaviour(s), occurring as a consequence of parameter choice, being input to some downstream complex system, process, or service that will not yield to analysis, without imposing any specific pre-ordained objective function, or fitness function to be optimised. We mean "parameter" in the widest sense, including system learnables, non-autonomous forcing, sequencing and all inputs.

Depending upon the nature of the underlying parameter space of interest one may adopt a rather wide range of search algorithms. We do consider that this search activity has *meta-objectives*, though: one is of achieving diversity (efficiently reaching out across the space in some way); and one is of achieving some minimum density (not leaving out large unexplored holes). These are in tension. In general, the computational costs of both of these qualities become restrictive as the dimension of the parameter spaces increase; and consequently their balance is harder to maintain. We may also wish for a substantial random element of search to provide some *luck* in discovery and to avoid any naive preset sampling patterns.

We consider archive-based methods within a range of spaces: finite discrete spaces, where the problem is straightforward (provided we are patient with the random element); Euclidean spaces, of increasing dimension, that become very lonely places; and infinite dimensional spaces. Our aim is to discuss a raft of distinctive search concepts, that respond to identified challenges, and rely on a rather diverse range of mathematical ideas. This arms practitioners with a range of highly practical methods.

However applications requiring novelty search arise, one should avoid rushing to code-up a standard evolving search algorithm and instead give some thought to the nature and requirements of the search: there is a range of effective options available. We give some considered advice.

> *You have your way. I have my way. As for the right way,*
> *the correct way, and the only way, it does not exist.*
> Thus Spake Zarathustra, Friedrich Nietzsche 1883.

1 Introduction

For over a decade now [1, 2], the idea that we could benefit from novelty searches, rather than simply maximising pre-set objectives has been advanced. This idea does not decry the true advances in various methods of optimisation, whether via "hill walking" multivariate gradient methods (when one or more derivates are available) or via non-derivative optimisation (from Nelder-Mead to simulated annealing, genetic algorithms, and particle swarms). The thesis is that all such methods optimise a foreseen set of objectives, usually through the setting of objective or fitness functions. The alternative, that of abandoning objectives and searching spaces for novel (even surprising) solutions, applies not just to scientific and technological fields but also to many other endeavours that must make decisions or prioritizations against pre-set performance or preference measures (the *curse of consensus*).

More recently it has been argued that, even without setting any technical objectives to be optimised, the aim of novelty search itself does indeed assert some *meta objectives*: those of fully searching a parameter domain (with possibly unknown constraints) for novel or radical mechanisms while not missing possibilities through an incomplete search, leaving unexplored holes [3, 4].

Novelty search as an exploration algorithm was recently considered within European Union's Horizon 2020 program [5, 6], where it was asserted that novelty search asymptotically behaves like a uniform random search process within the *parameter space*. Yet archive-based search methods, including methods with *memory*, should really do better (in some ways), and fulfil the above meta objectives. There is no requirement that new samples should be *independent* perturbations of elements drawn from the existing archive, but rather could be co-dependent on many such (recently added) elements.

We are encouraged and excited by the idea of novelty search as a repost to conventional, *consensual*, thinking. Any pre-set objective function may induce the law of unintended consequences (the are examples a-plenty in [1]) and may rule out phenomena that do not conform to the prevalent group-think. The imposed consensus is really the large-scale problem, while the technical optimisation challenge is the small-scale problem which garners almost all of the attention. Objective optimisation is thus low risk and may well fail to make any giant leaps of progress. It is largely aimed at the incremental grinding out of advantages within the present paradigm as opposed to disruptive, game changing, progress.

The term *novelty detection* is often used to mean anomaly detection [7, 8], a form of single class supervised discrimination. One creates a model of normality by using observations (hopefully free from the presence of anomalies), and then, deploying a metric derived from that model and allowing for variability, one identifies incoming anomalous observations that are extreme or unexpected. In applications the observations are generated (given to us) from some particular source(s) over which we have little control. So, though related, novelty detection is distinct from novelty search: the two should not be conflated. However, if we are successful in novelty search, then we might evaluate the resulting samples in a number of informal ways, perhaps according to a list of non-exchangeable performance summaries. In that case a novelty/anomaly detection algorithm might be extremely valuable and could even feed-back to bias the subsequent search. In this paper we will set aside the anomaly detection challenge, for the most part, yet discus the possible role of behaviour recognition [6] in section 3.8.

More recent work [9] adopts an approach, inspired by [1], for efficient exploration within reinforcement learning. This leverages a low-dimensional encoding of the environment learned with a combination of model-based and model-free objectives. Just as introduced below, and suggested in [1], this method uses the local sparsity of regions containing a new sample as a surrogate for novelty (measured by the distance to the new sample's nearest neighbour in the existing sample archive, or the average distance over the K nearest neighbours). In [10] the authors provide a review and suggest some open problems within deep reinforcement learning (DRL) applications and several promising directions for future research: these include some that are relevant to the methods set out, and the conclusions drawn, here. In particular, they say, *"The difficulty of DRL naturally increases with the growing of the state-action space. For example, real-world robots often have high-dimensional sensory inputs like images or high-frequency radar signals, and have a large number of degrees of freedom... In general, when facing large state-action space, exploration can be inefficient,... the exploration issues in large action space are [subject to a] lack of study currently... how to efficiently explore in a large action space is still an open question."*

The core problem with novelty search is a wholly practical one. One that is too often obscured by the details of particular applications. One should always ask how many dimensions are there? Are any of the parameter dimensions unbounded?

We might describe a possible parameter space by representing a wide set of parameters/inputs (to some system). Yet inevitably a search of such domains must involve sampling. So how best to sample, especially when spaces become very large?

Within many applications the parameter space may be high-dimensional. The curse of dimensionality is challenging and our intuition of two and three dimensional toy problems may be wholly misleading.

In this paper we wish to contribute to two major technical issues pertaining to novelty searches. Firstly, we will look at procedures for sampling domains that might engender wide coverage (diversity), while at the same time creating a relatively uniform sampling density. It seems to us surprising that many rather useful and illuminating concepts from some diverse areas of mathematics have not so far been drawn into the methodologies of novelty search. We will discuss some of these in responding to the sampling, and sampling bias, aspirations that must underpin the inherent the tension between sample packing (density), and effective and fast reach (diversity). As the dimension of the sample space increases there is a need to have some common and rigorous foundations and some comparative concepts: the alternative is a free for all.

Secondly, in the spirit of novelty, we wish to break out of finite dimensional (parameter) searches, and consider novelty search within infinite dimensional spaces. This seems to be largely absent from the literature whereas the equivalent classical optimisation (variational) problems are familiar to all.

We summarise our findings in the last discussion at the end of this paper.

2 Preliminaries

Throughout we will use $\mathbb{R}^n$ to denote the n dimensional real Euclidean space, and $\mathbb{R}^+$ to denote the set of non-negative real numbers.

We adopt the framework introduced in [3]. Let (X,δ) denote a metric space, defined over a non-empty set of points, X, and equipped with a metric $\delta : X \times X \to \mathbb{R}^+$. For any point $a \in X$ and distance $\varepsilon > 0$, we let $N_\varepsilon(a)$ denote the ε-neighbourhood of a: $N_\varepsilon(a) = \{x \in X | \delta(x,a) < \varepsilon\}$. X will denote the parameter space within which we wish to search for parameters that give rise to extreme behaviours when they are input to some complex systems (that itself does not yield to analysis).

For any subset $A \subset X$, the ε-neighbourhood of A is the union of the ε-neighbourhoods of all points in A:

$$A_\varepsilon = \cup_{a\in A} N_\varepsilon(a) = \{x \in X | \min_{a\in A} \delta(x,a) < \varepsilon\}.$$

Definition

i An ε-cover is a subset $A \subset X$ such that $X = A_\varepsilon$.

ii An ε-packing is a subset $A \subset X$ such that $\delta(a,b) > 2\varepsilon$ for all $a,b \in A$.

iii An ε-net is a subset $A \subset X$ that is an ε-cover of X and an $\varepsilon/2$-packing.

Intuitively an ε-net is a set of points that is well spread out, yet is an ε-cover for X.

An ε-net is not a unique entity: there may be many such nets over the same space. Any ε-net is often merely an aspiration, possessing both coverage (reach) and packing (density) attributes.

We will consider a range of methods that can grow a net as an archive, by generating and possibly accepting suitable candidates: a new candidate $x \in X$ will be added into A provided it is sufficiently far away from points already in A (typically being greater than $\varepsilon > 0$, fixed). It is trivial to generalise this test to take an average of the distances from x to its $K \geq 1$ nearest neighbours, already in A; though of course that approach could violate any packing requirement, so we leave this for now. The idea of using the K-near neighbours to reflect local sparseness of A within U goes back to the initial concept of novelty search [1, 2, 3]. We see no issue in adopting this K-nearest neighbour principle with any of the candidate generation methods discussed below.

We begin by considering two quite distinct examples.

In coding theory, a Hamming space is usually defined where $X = \{0,1\}^n$, the set of all 2^n binary strings of length n, equipped with the Hamming metric, an integer in $\{0,..,,n\}$, counting the number of differences between any two such strings.

In Figure 1 we show an ε-net for the $n = 16$ Hamming space and $\varepsilon = 4$. Here $|A| = 478$. In fact in this case an ε-net consists of an $\varepsilon/2$-packing made up of neighbourhoods about each point in A, each containing exactly 1+16+120=137 points in U (all at distances ≤ 2). Such a packing, A, may thus contain at most $2^{16}/137$= 478.4' points. The net is equivalent to an error-correcting code, correcting at most two errors; with efficiency $\log_2 478/16$ (55.6%).

In this case the net A was generated recursively by growing A as an *archive*, one element at a time. At each step we randomly draw a parent, an existing element of the evolving archive A, and randomly mutate it to form a new *candidate* which is accepted if it lies outside of the existing ε-neighbourhood of A. So each successive candidate element is added providing it is sufficiently *novel*, compared to existing elements. Obviously, when A is necessarily finite, the last few elements of A become harder and harder to find.

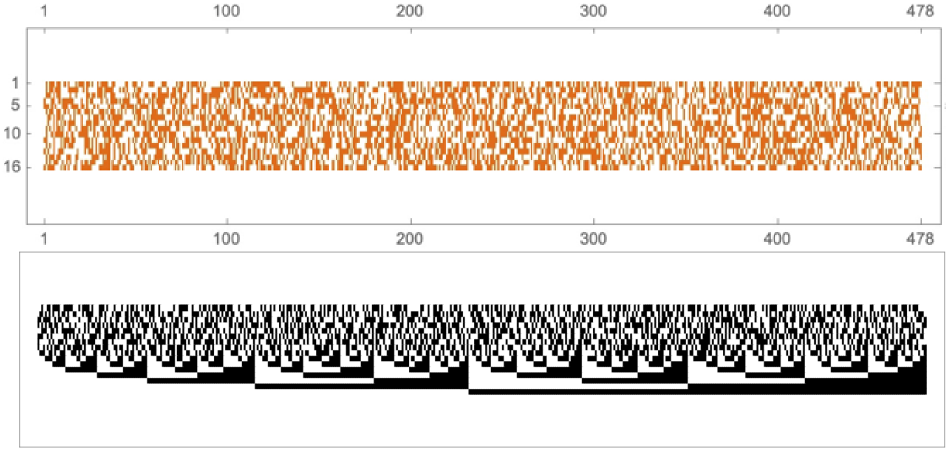

Figure 1: An ε-net for the $n = 16$ Hamming space with $\varepsilon = 4$, and 478 elements: each lies with a minimum pairwise distance of 4 from all others. We show (above, orange) each element vertically as the columns and the arrange the elements horizontally in the order in which they were added to the net. We show (below, black) the same elements this time ordered horizontally by their binary value.

Consider next a rather different example of an unbounded Euclidean space, where again we successively grow a net as an *archive*, A, of elements within $X = \mathbb{R}^2$. To do so we draw an existing element, to be a parent, from the present archive, A, at random, and then *mutate* it by incrementing each of its components independently with probability of 1/2, by the possible addition of a random perturbation drawn uniformly and independently from [-1,1]. Such a candidate element is accepted provided it is a distance of greater that $\varepsilon = 0.5$ from any existing elements within the archive. The result is shown in Figure 2. As observed previously in [3], the net-growth process gets slower and slower as one is more and more likely to draw the parent element from the interior of the net rather than lying close to its periphery, and thus being more likely to generate an external, *novel*, candidate.

The idea of constructing an ε-net, or an even approximation to one, in order to cover, and thus search over, a large metric space is key to the concept of seeking *novelty* rather than maximising any pre-set objectives [1, 2]. The general point is that by pre-setting (any) objectives one simultaneously commits to only obtaining foreseeable progress, within a pre-set frame or pre-set paradigm; while one risks entrenching any present consensus thinking about what might be possible; and one also risks getting any optimisation process stuck within the basins of attraction for local, rather than a global, maxima. The latter objection is a practical question of optimisation, and often besets derivative-free optimisation: it is a challenge to that discipline. Yet the former objections are questioning the underpinning (and often implicit) assumptions made in seeking solutions to any well-defined optimisation problems.

The largest such assumption is that we are even able to state an objective, without throwing at least some babies out with the bathwater. Stanley and Lehman [1, 2] argue coherently that this is damaging in many areas of science and in wider (objective) decision-making.

Recently [3, 4] suggested that, in fact, the desire to search spaces via the quest for novelty is itself a *meta objective*, which questions how have we decided which space to search, and why? In any case the search for novelty inevitably involves a tension between the twin objectives of achieving growing coverage together with an efficient packing.

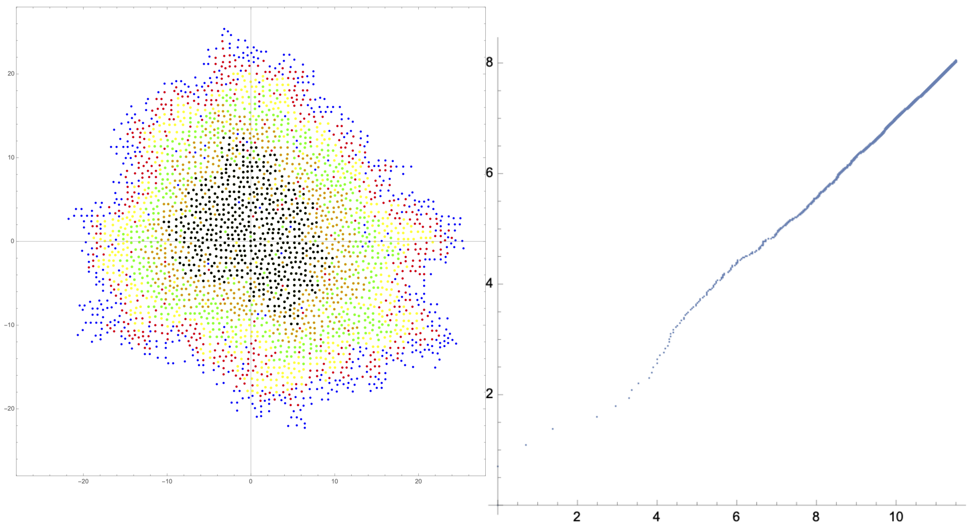

Figure 2: Left: a growing ε-net for $\mathbb{R}^2$ and $\varepsilon = 1/2$, with 500, 1000, 1500,...,3000 successive points. Right:$\log_e$ of the points added versus $\log_e$ of the candidates tested.

In [4] the author concludes, "*Still, selecting random individuals [as parents] from inside an archive is inefficient. As the archive grows, the ratio of the leading edge to its overall volume will reduce, making it harder and harder to sample from areas that are likely to produce "novel" solutions. This problem gets worse as dimensionality is increased, of course.*" This suggests that we bias parent selection to favour the leading edge, the periphery, of the evolving archive.

3 Novel methods for novelty searches

3.1 Recency bias in parent selection

Given the suggestion in [4], that parent-selection be biased towards the evolving archive periphery, we require an inexpensive way to achieve this: that is, avoiding the testing and assaying of many possible parents. Yet the information necessary to do this is already at hand (and unexploited, thus far): it is the order in which the the past elements have been added to the archive.

Rather than selecting a parent independently from the whole evolving archive, we will bias to select uniformly from amongst relatively recent additions, which by definition must be mostly in the periphery. In Figure 3 we apply this to the construction of a net in $\mathbb{R}^2$, introduced in the previous section. We see that although this is much more efficient in terms of adding elements to the whole archive, when it becomes too biased it tends to leave gaps within the interior of the net, as well as shifting the whole via a kind of group random walk, and induces a ragged periphery.

This method makes absolutely no difference to the performance of the example of net construction for the finite Hamming space, given in the last section, since all distances are bounded there: everywhere is thus relatively close to everywhere else. The inefficiency arose from (randomly) trying to find the last few elements of the net. In fact, in that case, following an initial automated random search, an exhaustive search of the remaining possibilities appears to be much more efficient than grinding on.

3.2 Additional random rotations

Consider a problem where the potential parameter space is rotationally invariant, such at $X = \mathbb{R}^n$. Instead of generating candidates solely via random perturbations of existing elements, here we include an additional random rotation drawn uniformly from $[0, 2\pi)$.

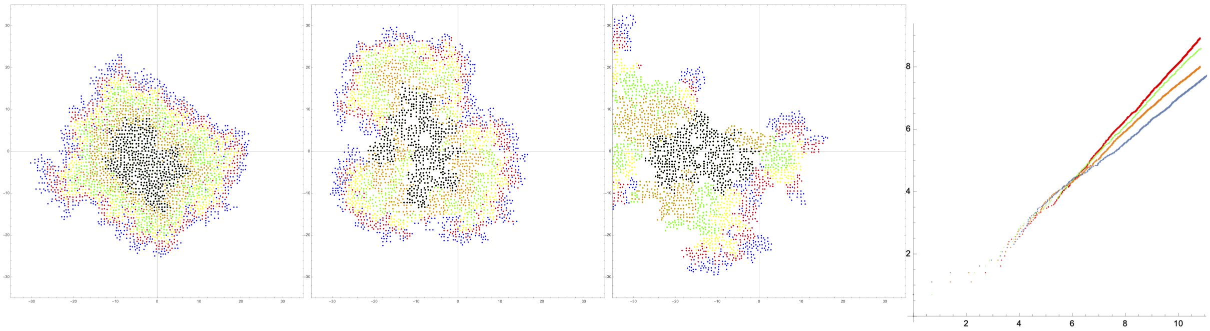

Figure 3: Growing ε-nets for $\mathbb{R}^2$ and $\varepsilon = 1/2$, with 500, 1000, 1500,...,3000 successive points. From the left: selecting parents from the most recent 50%, 20% and 10% of the archive. Right: $\log_e$ of the points added versus $\log_e$ of the candidates tested, selecting parents from the most recent 100% (blue), 50% (orange), 20% (green), and 10% (red) of the archive.

To illustrate consider the case shown in Figure 3 where we select from only the most recent 20% or 50% of elements added to the archive. Now, in addition, we impose a random rotation and we obtain the situation shown in Figure 4. This fixes the random walk and ragged edge problems and yet is more efficient than the comparable no-rotation case in terms of having candidates selected.

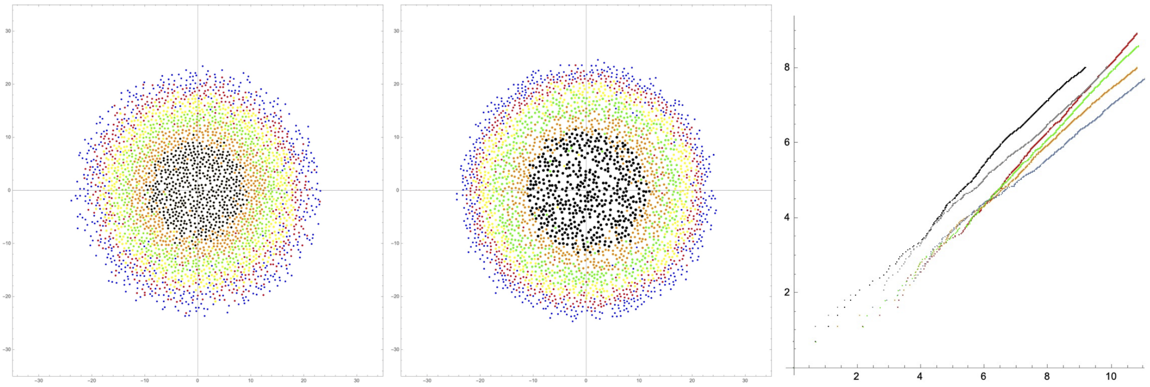

Figure 4: From the left: Growing ε-nets for $\mathbb{R}^2$ and $\varepsilon = 1/2$, with 500, 1000, 1500,...,3000 successive coloured points, by selecting and perturbing parents drawn from the most recent 50% and the most recent 20% of the archive together with an additional random rotations. Right: $\log_e$ of the points added versus $\log_e$ of the candidates tested: for sampling the most recent 50% (gray) and 20% (black), compared to the results from Figure 3 (selecting parents from the most recent 100% (blue), 50% (orange), 20% (green), and 10% (red) of the archive, with no additional rotations).

We conclude that additional random rotations may be very useful provided that we have some *a priori* reason that the growing net should maintain some symmetry (thus avoiding a ragged periphery) and also should suppress a group random walk.

3.3 Dimensional considerations

Bernhard Schölkopf once said, "A high-dimensional space is a lonely place" [11]. How efficient are the packings achieved by our evolving nets spreading out in $\mathbb{R}^n$?

By varying the dimension n we observe that the fraction of candidates accepted (being a distance greater than ε from any previous elements of the archive) increases. Yet in this section we will show that the packing efficiency decreases. These phenomena are both consequences of the *loneliness* of higher dimensions: with increasing n it is easier to generated new candidates elements that are isolated, and yet much harder to pack these efficiently.

The volume a ball in $\mathbb{R}^n$ of radius R is

$$V(n,R) = \frac{\pi^{n/2}R^n}{\Gamma(\frac{n}{2}+1)}.$$

Optimal packing densities, ρ_n, for hyper-spheres in n dimensions come into play here. These volume fractions are known [12] for $n = 1,..,8$.

n	ρ_n
2	$\frac{\pi}{2\sqrt{3}} = 0.9069$
3	$\frac{\pi}{3\sqrt{2}} = 0.7405$
4	$\frac{\pi^2}{16} = 0.6169$
5	$\frac{\pi^2}{15\sqrt{2}} = 0.4653$
6	$\frac{\pi^3}{48\sqrt{3}} = 0.3729$
7	$\frac{\pi^3}{105} = 0.2953$
8	$\frac{\pi^4}{384} = 0.2537$

In our previous examples, within sections 2, 3.1 and 3.2, we chose the centroids of balls to be at a distance greater than $\varepsilon = 0.5$ from one another. Hence we had packed balls of radius of $\varepsilon/2 = 0.25$ about each centroid. Suppose we have m such n-dimensional balls, each of radius $\varepsilon/2 = 1/4$, totally contained within a large n-dimensional ball of radius 8 centred at the origin (the centroids of all such $\varepsilon/2$ balls being less that 1.25 from the origin). Them the archived density is approxiately

$$\rho_a = m\frac{V(n,1/4)}{V(n,3/2)} = m\frac{(1/4)^n}{8^n} = \frac{m}{32^n}.$$

In Figure 5 we show an example for $n = 2$. This achieves an approximate packing density of $\rho_a = 0.43$, compared to an optimal packing density of 0.74.

For $n = 3$ a similar calculation yields $\rho_a = 0.21$, compared to an optimal packing density of 0.61.

For $n = 6$ a similar calculation yields $\rho_a = 0.005$, compared to an optimal packing density of 0.37.

Clearly searching a high dimensional spaces is expensive, and randomised methods, succesively adding individual elements to growing archives, are relatively poor at packing. Any opportunity for dimensionality reduction (of defining parameter space) should be accepted gratefully.

3.4 Lattice searches

Alternatively, we might accept this deficiency as a challenge, and design novel random methods which can form an achieve that is far closer to the optimal packing densities in dimension $n > 3$.

For example we take a naive rectilinear lattice of elements, within $\mathbb{R}^6$, with centroids at $0.5z$ for suitable values of $z \in \mathbb{Z}^6$, we find $\rho_a = 0.039$, compared to an optimal packing density of 0.37: almost an order of magnitude better than the random individual element approach, above. Similarly such a rectilinear grid in $\mathbb{R}^3$ dimensions achieves $\rho_a = 0.35$.

In fact in [12] the lattice arrangements that obtain the optimal packings are discussed. So one could achieve those densities provided one is prepared to accept such a regular lattice underpinning one's search purposes (up to $n = 8$ at least). We refer the reader to [13] which is a most useful reference in this regard.

Of course any lattice searches may suffer from systematic biases, since a certain structure is imposed: one may be *unlucky* without incorporating a more random element of search.

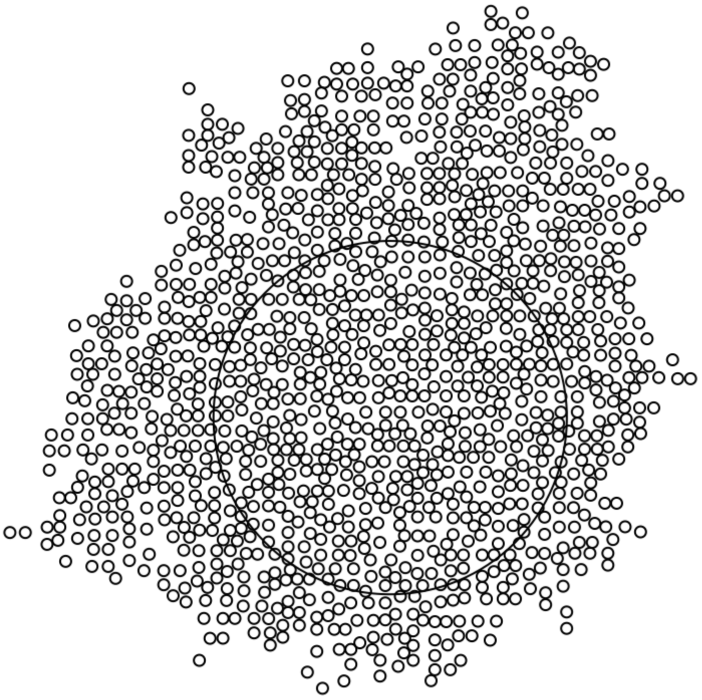

Figure 5: In $n = 2$ dimensions, 437 spheres of radius $\varepsilon/2 = 1/4$ are located inside a larger circular domain of radius 8. We estimate a packing density of ρ_a =0.43, compared to the optimal density of 0.74.

3.5 Halton points

In this section we consider an alternative method of selecting a sample to cover a space in a non-uniform way (as opposed to a uniform lattice or sequential random candidates).

Halton point sets, and the related Hammersley point sets, are two well-known, *low discrepancy* sequences, and have been used for quasi-Monte Carlo integration, ray tracing and other applications [14]. A deterministic formula generates a uniformly distributed and stochastic-looking sampling pattern at low computational cost, over the unit cube in $\mathbb{R}^n$. For each of n dimensions we chose distinct primes, $\{p_1, p_2, ..., p_n\}$, and construct as many points, K, as we require, as follows.

Briefly, for any $k = 0, 1, ...K$, and any prime p, we write $k = q_0 + a_1 p + a_2 p^2 + ... + a_r p^r$, where each a_i in an integer in $\{0, 1, ..., p-1\}$. We define the function

$$\Phi_p(k) = \frac{a_0}{p} + \frac{a_1}{p^2} + ... + \frac{a_r}{p^{r+1}}.$$

Then the kth Halton point is given by

$$(\Phi_{p_1}(k), \Phi_{p_1}(k), ..., \Phi_{p_n}(k))^T \in \mathbb{R}^n \quad k = 0, 1, ...K.$$

In practice, for our purposes, we may ignore the $k = 0$ point at the origin. The advantage with Halton points over the closely related Hammersley points is that if we should require more samples (thus increasing K) then that will not change the values of the previously calculated samples [14, 15]. An archive with this property is called *hierarchical*.

We give an example in $\mathbb{R}^2$ and in $\mathbb{R}^3$ Figure 6.

In $\mathbb{R}^6$ we generated a set of $4^6 = 4,096$ points within $[0,1]^6$ with a minimum separation of 0.099' (using primes 2, 3, 5, 7, 11, and 13). This leads to a packing density of 0.00031, which is wildly inefficient, nevertheless those sample points are well spread out through the 6-cube. For any point in the cube, the nearest element of the set lies a median distance of 0.189 away, with the 5th and 95th percentiles given by (0.121,0.247). This compares to a *corner to corner* distance in the 6-cube of $\sqrt{6}$=2.449.

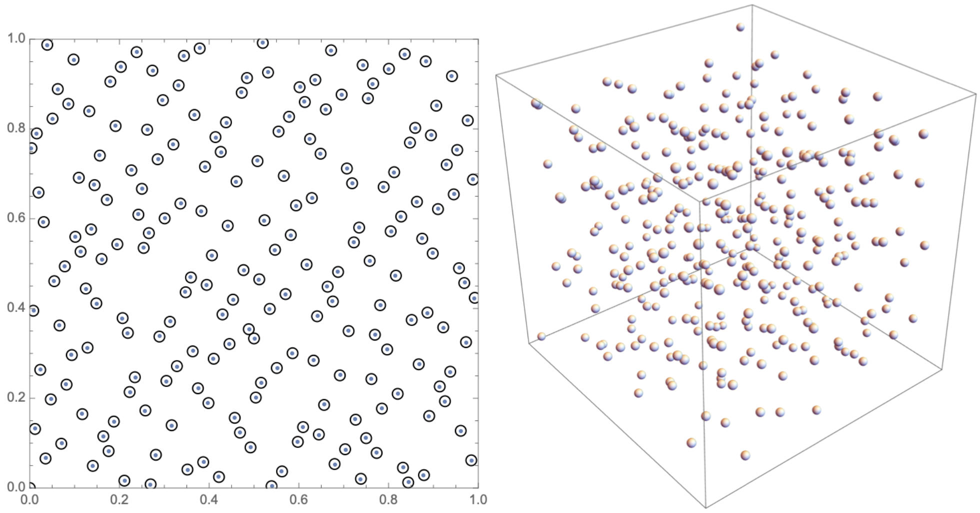

Figure 6: Left: A set of 200 Halton points in $\mathbb{R}^2$, generated with $p_1 = 2$ and $p_2 = 3$, with the largest balls, of radii= 0.01184, centred at each point, such that no two balls intersect. Right: A set of $7^3 = 343$ Halton points $\mathbb{R}^3$ generated with $p_1 = 2$, $p_2 = 3$ and $p_3 = 5$.

For a $7^6 = 117,649$ point set, generated in the same way, the median distance from any point is reduced to 0.104, with the 5th and 95th percentiles given by (0.067,0.133).

Halton points might be accepted wholesale, up to a certain number K, covering a cube in n dimensions, or we might consider them successively and reject any candidate that is too close to existing candidates. In that way the process will terminate once the n-cube is sufficiently packed.

3.6 Lévy flights and foraging

We have already discussed how the process of successive parent selection and subsequent mutation really produces a number of random walks, when we follow a sequential chain of successive parents and offspring. We have seen that the parent selection may benefit from a recency bias, as in section 3.1. Here we examine mutation process, incrementally perturbing parents to produce candidate offspring. So far we have spoken quite generally about the mutation. Here we show that there are two extreme options (and a continuum between): that of having perturbations with a finite variance, resulting in a Brownian random walk; and that of having perturbations with an infinite variance, resulting in a Lévy flight.

Brownian motion may not be the best way to sample a space. Classical Fickian diffusion has particle walks with displacement increments selected independently from a distribution which has a finite variance. The resulting Gaussian profile for the probability distribution for the (un-constrained) particle positions follows from the Central Limit Theorem. Lévy flights draw independent increments from a distribution with a *fat tail*, meaning they have no finite variance. A distribution with inverse quadratic decay is most popular in applications, such as a standard Cauchy distribution: $1/\left(\pi(1+x^2)\right)$. Consequently, compared to Brownian motion, there are disproportionately many more larger increments.

For a few decades the hypothesis of Lévy foraging has been advanced: the idea that conscious animals do not move around and search randomly, blindly; instead that they obey Lévy flights [16, 17]. There is some debate about the observations and optimality of such foraging within ecology, although the sense in which such behaviour is mathematically optimal, at least within two dimensional space, is now clarified [18]. The potential effectiveness of high-dimensional Lévy searches has not yet been thoroughly examined. Theory predicts that three-dimensional Lévy searching can be advantageous when foragers are effectively blind and need to come into contact with a target to establish its presence [16]. Accordingly searches based on Lévy flights may be far more successful in achieving the diversity (dispersion) meta objective. We illustrate this approach, and motivate the use of the Cauchy distribution as follows.

Let $X = L^p(\mathbb{R})$ for $p \geq 1$. Let $f \in L^1(\mathbb{R})$ denote a real nonnegative function with unit mass, and r be a positive real parameter. For time $t > 0$ we define the "similarity" function

$$F_t(x) = \frac{1}{t^r} f\left(\frac{x}{t^r}\right),$$

and the corresponding family mappings S_t, evolving any initial distribution, $u \in X$, at time $t = 0$, via the convolutions

$$S_t.u = F_t * u, \quad u \in X.$$

Then for any well-defined, memoryless, time-dependent evolution we must have $S_{t_a+t_b}.u = S_{t_b}.S_{t_a}.u$, for all $t_a, t_b > 0$. This requires that S satisfies the semigroup property: $S_{t_a+t_b} = S_{t_a}.S_{t_b}$ for $t_a,\ t_b\ >\ 0$. Taking Fourier transforms we see that this last is true if and only $\hat{F}_t(k) = e^{tg(k)}$, for some even function $g(k)$ vanishing at $k = 0$, and tending to $-\infty$ for k large. On the other hand, given the definition of F_t, a direct calculation of the transform shows that $\hat{F}_t(k) = \hat{f}(t^r|k|)$.

Putting these together we must have $g(k) = -\mu.|k|^{1/r}$, for some constant $\mu > 0$ (without loss of generality we take $\mu = 1$) so that, $\hat{f}(t^r k) = e^{-t|k|^{1/r}}$.

If $r = 1/2$ we recover the usual Fickian process: $F_t(x)$ is the fundamental solution of the diffusion equation. The semigroup S_t is the Gauss-Weierstrasse semigroup.

If $r = 1$ we obtain $F_t(x) = 1/\left(\pi t(1 + (x/t)^2)\right)$, the standard Cauchy distribution. The semigroup S_t is called the Poisson or Cauchy semigroup.

These two examples are very well known [19, 20] with the general case $1/2 \leq r \leq 1$ treated in [19]: in fact, for r outside of this range f is not non-negative so we may rule out such cases.

Variations on Lévy motions and the corresponding fractional partial differential equations have been very widely proposed for use within pricing models in finance applications, extending the Black-Scholes method [21]. Almost all of the resulting analysis seems to resort to numerical (sampling) methods [22]. The semigroup property ensures that there is no memory dependence of solutions so the adoption of this class of dispersive process does not invite infinite opportunities for arbitrage.

The space searched grows like t^r, so a Lévy flight may well be a good choice for diversity (reach): but the overall density of such an archive will be decreasing as more distant samples are added.

In Figure 7 we show 1000 point archives developed with a Brownian process and a Lévy flight precess in $\mathbb{R}^2$, elements in both archives are coloured by the order in which they were accepted. In both cases, at each step, we considered only the most recent 50% of the archive elements as parents and accepted candidates provided they were a minimum distance of $\varepsilon = 0.5$ from any existing archive member. The radical nature of the Lévy flight method is apparent (not all points are shown). This lack of packing is likely to become more of an issue as the dimension of the underlying parameter space increases.

In many applications Lévy flight method are developed with controlled truncations of the Cauchy distribution. Note that these archives are formed from the superposition of many walks/flights, with Markov chains formed from successive (parent, successful candidate) pairs.

3.7 Novelty search within infinite dimensions

Novelty search within infinite dimensional function spaces is largely absent from the literature, due to obvious difficulties. Nevertheless it might be desirable.

The equivalent classical optimisation problem is dealt with by the calculus of variations, where we seek a function, from within a chosen space, that can optimise some functional that is defined in terms of a suitable real valued integral. Typically we will be optimising a functional within a subspace of a Hilbert space, such as $L^2(\Omega)$, the square integral functions defined over a domain, Ω, with sufficient smoothness, and satisfying certain boundary or decay conditions. Then the calculus of variations yields an equation to be satisfied by functions that achieve an optima. Subspaces X of $L^2(\Omega)$ are usually defined by demanding some additional smoothness through the existence of a set number of generalised derivatives also lying inside $L^2(\Omega)$, resulting in X being a suitable Sobolev space [24].

What of searches within such infinite dimensional spaces?

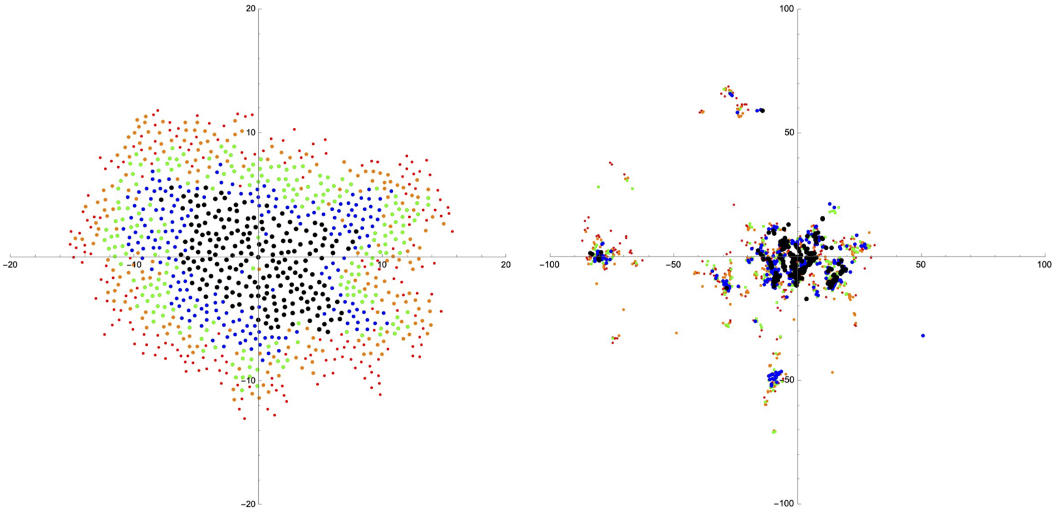

Figure 7: 1000 point archives developed with a Brownian process (left) and a Lévy flight process (right), with a minimum distance of $\varepsilon = 0.5$ from any existing archive member, and considering only the most recent 50% of the archive as parents. Points are coloured by their ordering: black points $\{1,200\}$, blue $\{201,400\}$, green $\{401,600\}$, orange $\{601,800\}$, red $\{801,1000\}$. Note the scale of the right hand plot (where a few far away points are not shown).

Bases for spaces depend of the nature of Ω and the boundary or decay conditions we impose, but could include eigenfunctions of a self-adjoint operator, so that they are orthogonal. Fourier series represent such an orthonormal expansion for a function (trigonometrical functions being eigenfunctions of the Laplacian), and the (power law) decay of the Fourier coefficients at infinity determine the smoothness of the function that is so represented.

Alternatively, different types of wavelets provide suitable orthonormal bases for L^2, where each basis function is a dilation and a translation of a common function, called the generating wavelet. There are many wavelet bases. A Daubechies wavelet [23] is both continuous and compactly supported. The possibility of compact support is thrilling to initiates. It is also continuously differentiable to some order. The wavelet coefficients in an expansion must decay as the dilations become smaller and smaller in order to ensure that the function of interest is smooth (and, thus, that the series expansion converges). Compactly supported wavelets generally avoid the Gibbs effects, that Fourier expansions have in approximating discontinuities.

Let us assume that we are searching within an infinite dimensional space, X, elements of which are represented by an orthonormal basis with associated expansion coefficients satisfying a decay constraint (ensuring sufficient smoothness). Then we may perturb such a function by perturbing the basis expansion coefficients provided they do not violate the decay constraint. The norm on the space is equivalent to the (possibly weighted) sum of squares of the expansion coefficients.

An archive growing by parent selection incorporating recency bias (as in section 3.1) and mutation (via random controlled perturbations to the basis expansion) may work well. But notions of packing may be absent. Recency bias in the most extreme case, when we select only the most recent element added to the archive as the parent, is equivalent to making a random walk, a single Markov chain. In general many such random walks all superimposed represent a diffusion process. Of course here we are thinking about random walks and diffusion within infinite dimensions. That should not be confused with an infinite dimensional diffusion dynamic for a distribution function over a finite dimensional domain.

This appears to be a specialist and yet rather neglected field of novelty search.

Of course, in terms of numerical computations we would only ever hold a finite number of basis coefficients; or alternatively we might hold values of the functions of interest at finite set of points in Ω, or a finite (truncated) part of any other expansion (in terms of finite elements, for example). So we rely on the theory to ensure that nothing untoward will undermine such an approach.

As an example, in Figure 8 we adopt a Fourier cosine series expansion for functions of one variable defined on $\Omega = [0,1]$ with zero Neumann boundary conditions. We also impose a power law decay condition on the Fourier coefficients to ensure that the search space, X, consists of differentiable functions. We show every 10th element of a growing archive with parent selection biased to the most recent 20% of archive elements. This is similar to the example in section 3.1. This approach is good in practice but it is poor in theory due to the necessary truncation of the basis expansion. If one is content with the truncation, and the consequent over imposition of smoothness, then this may well be useful though.

A notion of packing is impossible in an infinity of dimensions, the best we can have is an assurance that elements of the archive are separated by a minimum distance of $\varepsilon = 0.5$ from one another, without being able to pack any bounded sub-domain.

Applications of an infinite diemsional parameter space would include situations where one *forces* a complex system with an externally derived time-dependent function, resulting in a non-autonomous system (such as time dependent effects resulting from future climate variations, or future human behaviour).

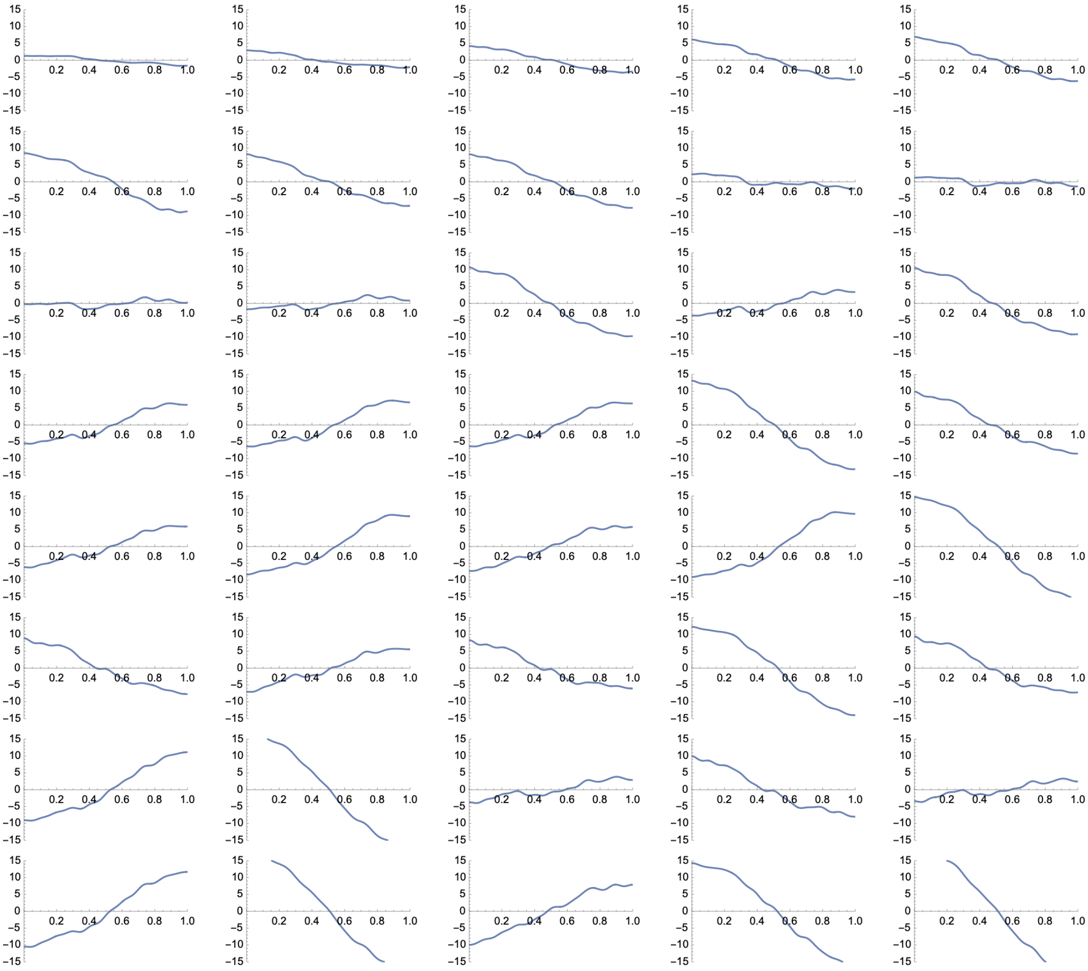

Figure 8: An archive of 160 successive functions (we show every 4th element added), defined on $\Omega = [0,1]$ satisfying zero Neumann boundary conditions, with each being a minimal separation of $\varepsilon = 0.5$ in the $L^2([0,1])$ norm from all previous elements.

3.8 Behaviour recognition controlling novelty search

So far we have looked at a range of methods that can generate candidates to be added to an archive. In general we have a parameter metric space (X,δ), and a growing archive of points, $A \subset X$. A new candidate $x \in X$ is added into A provided it is sufficiently far away from points in A; that is, we have

$$\min\{\delta(x,a)|a \in A\} > \varepsilon,$$

for some $\varepsilon > 0$, fixed. We imposed this condition on the candidates generated via allowable perturbations of parents, drawn from (usually recent) existing element of A, whether via Brownian process, Lévy fight, or any other mechanism.

It is straightforward to generalise this test condition to taking an average distance over the $K \geq 1$ nearest neighbours:

$$\frac{1}{K}\sum_{k=1}^{K} \delta(x,a_k(x)) > \varepsilon,$$

where the $a_k(x) \in A$ are the K nearest neighbours of x within A. In fact, the idea of using K-near neighbours to reflect local sparseness of A goes back to the initial concepts of novelty search [1, 2, 3]. We see no issue in adopting this principle to test any of the candidate generation methods discussed above.

This may be taken a step further by introducing a mapping $\phi : X \rightarrow B$, where (B,d) is another metric space, called *Behaviour space*. Let $A_B \subset B$ be the image of an archive $A \subset X$ under the mapping ϕ.

Then we may impose the above sparseness requirement in B, as follows. A new candidate $x \in X$ is added into A provided $\phi(x)$ is sufficiently far away from points in A_B; that is, we have

$$\frac{1}{K}\sum_{k=1}^{K} d(\phi(x),b_k(x)) > \varepsilon,$$

where the $b_k(x) \in A_B$ are the K nearest neighbours of $\phi(x)$ within A_B.

Hence A within X can be grown by novelty sampling and growing A_B within B. In this way we can avoid over sampling regions of X where behaviour, ϕ, is relatively constant, and instead bias the search towards regions where ϕ is relatively variable or even volatile. In practice B might be defined in terms of a range of observable, non-exchangeable (not co-dependent) performance measures.

In *Behaviour Recognition Novelty Search* [6] the authors do away with the archive, instead deploying a generational approach with a generational reference set within X. The importance though lies in the flexibility of choice of ϕ and B. This objective has been shown to be sufficient to outperform fitness-based optimisation in problems with deceptive optima. Within this framing it is also possible to combine the novelty objective with a fitness multi-objective.

4 Discussion

One needs to consider the nature of each and every novelty search problem on its own merits. In particular, the chosen search process needs to reflect whether the parameter (search) space is discrete and finite or countable; or else is it is infinite of either low or high but finite dimension, and is bounded or unbounded; or else it is of infinite dimension.

One may also have different priorities in distinct search situations. The dimension, n, of a parameter space should not be swept under the carpet. It should be embraced. As n increases density (packing) becomes excessively costly and also limits how far one can search. Even searching an n-cube the necessary sample would grow exponentially with n to maintain density. Yet in the early days ”novelty” was defined in terms of a sparsity-surrogate measure (distance to nearest archived sample, or the the average distance over the K nearest such samples). In many applications n might be large, though, and, by definition, we can have no objective or fitness function, that is to be maximised, and can limit (constrain) or direct the search.

One may decide that “reach” (diversity) within the archive is more important than packing. But then one should choose mutations (perturbations) which may be large, and avoid the trap of finite variance (the default, over-use, of Gaussians and so on).

- If the parameter space, X, is finite or else is bounded within $\mathbb{R}^n$ (n finite), then random archive packing is clearly possible but this may take some computational effort to squeeze out the last few elements. Alternatively rectilinear lattice-based methods are very simple, yet may contain possible systematic biases. We recommend using pseudo-lattice methods, which avoid rectilinear alignments yet also limit clustering, such as the use of Halton points as candidates for the archive generation in the n-cube in $\mathbb{R}^n$. This is highly efficient and can achieve a relatively uniform density (and thus packing). As n becomes larger this is certainly preferable.
- If the parameter space, X, is unbounded within $\mathbb{R}^n$ (n finite), where packing (the existence of unexplored voids) is very important, then any exhaustive is search impossible. Archive methods using mutations (applied to parents) with finite variance are equivalent to Brownian motion (diffusion) in $\mathbb{R}^n$. We also recommend using a recency-bias in parent selection so that the outer extremes of the archive are grown efficiently (avoiding the selection of parents within the interior of spaces already sampled). As the dimension n becomes large though, diffusion may be far too slow.
- If the parameter space, X, is unbounded in $\mathbb{R}^n$ (n finite), where packing is far less important than achieving a wide reach (a large sampling variation), then diffusion (mutations with finite variance whence the Central Limit Theorem applies) is far too inefficient. In this case we recommend that the mutations applied to parents should be such that they have an infinite variance in $\mathbb{R}^n$. The resulting Lévy flights in n-dimensions has occasional very large jumps. This is explorative (and known to be effective for small n within foraging). The behaviour for n larger will need to be analysed. We also recommend using a recency bias in parent selection so that the outer extremes of the archive are grown efficiently.
- If the parameter space, X, is an infinite dimensional parameter space, usually a function space, whether unbounded or bounded, then any packing is simply impossible (even for the unit ball: one may always add new elements equidistant form all existing members of an archive). In this case we recommend choosing a suitable basis (compactly supported wavelets or other bases) depending on the space and the nature of the elements, and the only relevant aim can be reach exploration. Therefore *a priori* truncations of the basis should be avoided, unless one is happy imposing some smoothness assumption. Perhaps one might allow some systematic increase in the length of the chosen basis expansion for candidates, so that more recent elements of the archive exhibit relatively more fine-scale structures. This last, combined with a recency bias and perhaps a Lévy flight process, may prove most effective.

Funding and Licence

This research was not funded by any external body. No conflicts of interests are known. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript (AAM) version arising from this submission.

Acknowledgments

We are pleased to acknowledge the advice given by Des Higham, Tom Melham, and Steve Roberts.

References

[1] Lehman, J. and Stanley, K.O. (2011), Abandoning Objectives: evolution through the search for novelty alone, Evolutionary Computation journal, (19):2, pages 189-223, Cambridge, MA: MIT Press.

[2] Stanley, K.O. and Lehman, J. (2015), Why Greatness Cannot Be Planned, The Myth of the Objective, Springer.

[3] Wiegand, R.P. (2020), The objective of simple novelty search. In Proceedings from the 2020 Florida Artificial Intelligence Research Symposium Conference.

[4] Wiegand, R.P. (2021), Population-Based Novelty Searches Can Converge. The International FLAIRS Conference Proceedings, 34. https://doi.org/10.32473/flairs.v34i1.128753.

[5] Doncieux, S., Laflaquière, A., Coninx, A. (2019), Novelty search: a Theoretical Perspective. GECCO '19: Genetic and Evolutionary Computation Conference, Jul 2019, Prague Czech Republic, France. pp.99-106, 10.1145/3321707.3321752 . hal-02561846

[6] Salehi, A., Coninx, A., Doncieux, S. (2021), BR-NS: an Archive-less Approach to Novelty Search, arXiv - CS - Artificial Intelligence (IF), Pub Date : 2021-04-08, DOI: arxiv-2104.03936

[7] Pimentel, M.A.F., Clifton, D.A, Clifton, L. and Tarassenko, L. (2014), A review of novelty detection. Signal Process. 99 (June, 2014), 215–249, https://doi.org/10.1016/j.sigpro.2013.12.026.

[8] Smith, M., Reece, S., Roberts, S. et al. (2014), Maritime abnormality detection using Gaussian processes. Knowl Inf Syst 38, 717–741.

[9] Tao, R.Y., François-Lavet, V. and Pineau, J. (2022), Novelty Search in Representational Space for Sample Efficient Exploration, arXiv https://arxiv.org/abs/2009.13579.

[10] Yang, T., Tang, H., Bai, C., Liu ,J., Hao, J., Meng, Z., Liu, P. and Z Wang, Z. (2022), Exploration in Deep Reinforcement Learning: A Comprehensive Survey, arXiv https://arxiv.org/abs/2109.06668.

[11] B. Schölkopf (2014), Twitter account, https://twitter.com/bschoelkopf/status/503554842829549568?lang=en.

[12] Weisstein, E.W., Hypersphere Packing, From MathWorld – A Wolfram Web Resource, https://mathworld.wolfram.com/HyperspherePacking.html.

[13] Conway, J.H. and Sloane, N.J.A. (1995), What are all the best sphere packings in low dimensions?, Discrete & computational geometry 13.3-4 (1995): 383-404. http://eudml.org/doc/131369.

[14] Wong, T.-T. Luk, W.-S. & Heng, P.-A. (1997), Sampling with Hammersley and Halton Points. Journal of Graphics Tools. 2. 10.1080/10867651.1997.10487471.

[15] Niederreiter, H. (1992), Random number generation and quasi-Monte Carlo methods. CBMS-NSF regional conference series in applied mathematics : 63, SIAM, Philadelphia.

[16] Reynolds A.M. (2015). Extending Lévy search theory from one to higher dimensions: Lévy walking favours the blind. Proceedings. Mathematical, physical, and engineering sciences, 471(2179), 20150123, https://doi.org/10.1098/rspa.2015.0123.

[17] Viswanathan, G.M., Buldyrev, S.V., Havlin, S., da Luz, M.G.E., Raposo, E.P., Stanley, H.E. (1999), Optimizing the success of random searches. Nature 401, 911–914. (doi:10.1038/44831).

[18] Guinard, B. and Korman, A. (2021), Intermittent inverse-square Lévy walks are optimal for finding targets of all sizes, Sci. Adv. 2021; 7 : eabe8211, https://www.science.org/doi/epdf/10.1126/sciadv.abe8211.

[19] Feller, W., (1971), An Introduction to Probability Theory and its Applications, Vol II, Wiley New York, 1971.

[20] McBride, A.C. (1987), Semigroups of linear operators: an introduction, Pitman Research Notes in mathemtics 156, Longman, 1987.

[21] Matacz, A., (2000), Financial modelling and option theory with the truncated Levy process, (available arXiv.cond-mat/9710197 v1 20 Octm 1997), International Journal of Theoretical and Applied Finance, Vol. 3, No. 1, P. 143, 2000.

[22] Cartea, A., del-Castillo-Negrete, D., (2006), Fractional diffusion models of options prices in markets with jumps, BWPEF0604, Birkbeck working papers in economics and finance, Birkbeck College, London.

[23] Daubechies, I. (1988), Orthonormal bases of compactly supported wavelets, Comm. Pure Appl. Math. 41 : 7 , pp. 909–996.

[24] Folland, G.B., (1995), Introduction to Partial Differential Equations. Second Edition (Mathematical Notes, 102), Princeton University Press; Revised edition (October 15, 1995).