Adaptive parameterized improving hit-and-run for global optimization

Wei Wang, Archis Ghate and Zelda B. Zabinsky*

Industrial and Systems Engineering, University of Washington, Seattle, WA 98195-2650, USA

(Received 30 April 2008; final version received 19 May 2009)

We build on improving hit-and-run’s (IHR) prior success as a Monte Carlo random search algorithm for global optimization by generalizing the algorithm’s sampling distribution. Specifically, in place of the uniform step-size distribution in IHR, we employ a family of parameterized step-size distributions to sample candidate points. The IHR step-size distribution is a special instance within this family. This parameterization is motivated by recent results on efficient decentralized search in the so-called Small World problems. To improve the performance of the algorithm, we adaptively tune the parameter based on the success rate of obtaining improving points. We present analytical and numerical results on simple spherical programmes to illustrate the key ideas of the relationship between the parametrization and algorithm performance. These results are then extended to global optimization problems with Lipschitz continuous objective functions. Our preliminary numerical results demonstrate the potential benefit of considering parameterized versions of IHR.

Keywords: global optimization; improving hit-and-run; adaptive stochastic search

1. Introduction

Markov chain-based stochastic search algorithms are commonly used for solving global optimization problems of the form

\[(P) \quad \min_{x \in S} f(x),\]

where \(x \in \mathbb{R}^n\), \(S \subset \mathbb{R}^n\) is convex, compact and of full dimension, and \(f\) is a Lipschitz continuous real-valued function over \(S\). These methods work informally as follows. At the current iterate \(x\), they employ a candidate generator Markov chain to generate a candidate point \(y\) and then either accept or reject \(y\) using a probabilistic acceptance filter.

A good candidate generator is critical for the efficiency of a Markov chain method for solving continuous optimization problems [20,21]. Existing generators typically include two components – a direction and a step-size along this direction, and hence can be broadly classified into (i) fixed step-size along a uniformly random direction [7,8,19]; (ii) uniformly distributed step-size
on a variable length interval that can shrink or expand along a uniformly random or coordinate direction [9,21]; (iii) uniformly random step along the intersection of a uniformly random direction with the feasible region [5,24,25,35]; and (iv) uniformly random step along the intersection of an analytically chosen optimal non-uniform direction with the feasible region [16,17]. Although researchers acknowledge that there is a potential benefit in stochastically biasing the step-size in a non-uniform and adaptive manner along a chosen direction [21], there are very few analytical results that precisely quantify and pursue this notion.

On the other hand, several acceptance filters have been extensively analysed in the literature. Perhaps the simplest of these is improving hit-and-run (IHR) introduced by Zabinsky et al. [35] that employs the well-known hit-and-run (HR) [6,27] candidate generator and accepts a candidate point only if it improves the function value at the current iterate. IHR may be viewed as a special case of simulated annealing (SA) [18], which was originally proposed for discrete optimization and has also been extended to continuous problems [19–21,24,28,33]. SA uses a Metropolis [22] filter parameterized by the so-called temperature, which is gradually reduced to a very small non-negative value so that candidate points that do not improve the function value of the current iterate are accepted with diminishing probabilities. Early research efforts to design variants of the SA acceptance filter were mainly computational (see Szu and Hartley [29] for fast SA (FSA) that uses a Cauchy filter, Tsallis [30] for generalized SA (GSA) with a filter that uses two parameters). These were followed by analytical results about limiting behaviour of similar extensions [4,31]. Locatelli [20] posed an open question as to which restrictions on SA can be relaxed to preserve convergence in probability. This ensued renewed interest in designing non-adaptive as well as adaptive variants of the SA acceptance filter [2,3,12,14,15,19,21,26,32,33].

This paper is a preliminary attempt at closing the gap between the scarcity of results on adaptive, parameterized, stochastic step-sizes in candidate generators and the wide range of existing work on acceptance filters that incorporate these features. Since we wish to isolate the impact of step-size distributions and exploit them to improve algorithm performance, we employ one of the simplest direction choices – uniformly distributed on the surface of a hypersphere centred at the current iterate as in the HR generator [27], and one of the simplest acceptance filters – accept a candidate only if it is improving as in IHR [35]. We thus design an extension of IHR where the step-size distribution is chosen from a family of single-parameter distributions that generalizes the uniformly distributed step-size in the original version of HR [27]. This parameter is denoted by ‘$a$’ and specifically, modifies the distribution of step-size $\lambda$. Given a direction choice, $\lambda$ is sampled from a probability density function that is roughly proportional to $(1/|\lambda|^a)$. This family is borrowed from [13], where a parameterized step-size was used in a decentralized search algorithm for solving the so-called Small World problem. Although the Small World problem is not directly an optimization problem, in [11] and [13], it is viewed as a Markov chain parameterized by $a$, where the expected hitting time characterizes the effectiveness of passing a message from one person to another randomly chosen person. Ghate and Smith [13] showed that the expected hitting time for the Small World problem is minimized when $a = 1$, and that $a = 1$ is the unique choice of $a$ that is scale invariant among all non-negative values. They suggest a potential application of this Markov chain family parameterized by $a$ to continuous global optimization. We explore the impact of $a = 1$ and other values of $a$ on the performance of IHR, as opposed to $a = 0$ which corresponds to the uniform step-size choice in the original version. Note that a large positive value of $a$ biases the step-size towards the current iterate imparting a ‘local’ character to the search, a value of zero corresponds to a uniform distribution and a negative value achieves a more ‘global’ search. In an algorithmic implementation, an important question in each iteration is how to choose an appropriate value of $a$. The above delicate tradeoff suggests that an adaptive choice of this parameter may help the overall search process and is in fact the central idea in adaptive parameterized IHR (APIHR).
The adaptive choice of the value of the parameter $a$ in APIHR is motivated by the observation that in many of the commonly employed generators the probability that the candidate has a better function value than the current iterate falls sharply as the value of the current iterate gets closer to the optimum (see the analysis of IHR [35] for spherical programs). In each iteration of APIHR, the idea is to choose a value of parameter ‘$a$’ that maximizes the probability of improvement.

As in the complexity analysis of IHR presented in [35], we first focus on spherical programs – optimization problems with spherical feasible regions and concentric spherical level sets around the centre of the feasible region. For spherical programs, the probability of improvement can be expressed exactly as a function of $a$ (see Theorem 3.1). Our analytical and empirical results show that $a = 0$ achieves higher improvement probabilities than all $a < 0$ (see Theorem 3.4 and Figure 4), which allows us to concentrate on non-negative $a$ values. Furthermore, while there is no strict dominance relationship among $a \geq 0$, we focus on the values $a = 0$ and $1$ for APIHR because $a = 0$ corresponds to the original IHR and $a = 1$ is the best parameter for the Small World problem. Also the values $a = 0$ and 1 have special characteristics with regard to the probability of improvement, as discussed later and illustrated in Figure 5. For these two values on spherical programs, we find $a = 1$ performs better than $a = 0$ when the radius $r$ of the current level set is below a certain threshold value, and the probability of improvement is higher for $a = 0$ only when $r$ is very large. Thus, APIHR on a spherical program employs a ‘threshold optimal policy’. It sets $a = 0$ initially, and switches to $a = 1$ when the radius of the current level set drops below the threshold. This threshold quickly approaches the radius of the feasible region as the problem dimension is increased, implying that $a = 1$ appears more and more effective throughout the progress of the algorithm for higher dimensional spherical programs.

The probability of improvement cannot be calculated exactly when the objective function is Lipschitz continuous even when the feasible region is spherical owing to the irregular shape of the level sets. Therefore, we exploit the notion of a worst case function in [26,34] whose level sets are spherical and contained within the level sets of the original Lipschitz function. Consequently, lower bounds on the probabilities of improvement can be derived and compared within an APIHR implementation. Our numerical results demonstrate that an APIHR implementation performs better than the original IHR with $a = 0$ on a set of test problems.

The rest of the paper is organized as follows. In the next section we introduce a parameterized version of the IHR algorithm. In Section 3, we develop APHIR and demonstrate its applications to both spherical and Lipschitz programs. Numerical results are also presented. Section 4 concludes with final remarks.

2. Parameterized IHR

For the optimization problem $(P)$, we define

$$y_* = \min_{x \in S} f(x), \quad x_* = \arg \min_{x \in S} f(x), \quad y^* = \max_{x \in S} f(x).$$

We restate the IHR algorithm based on [35].

Step 0. Initialize $X_0 \in S$, $Y_0 = f(X_0)$ and $j = 0$.

Step 1. Generate a direction $D_j$ uniformly distributed on the surface of the unit hyper-sphere centred at $X_j$. 
**Step 2.** Generate a candidate point \( X_j + \lambda_j D_j \) at step-size \( \lambda_j \) that is uniformly distributed over the interval \( L_j = \{ \lambda \in \mathbb{R} : X_j + \lambda D_j \in S \} \) that defines the set of feasible step-sizes from the current iterate \( X_j \) along direction \( D_j \).

**Step 3.** Accept the candidate point only if it is improving, i.e. set
\[
X_{j+1} = \begin{cases} 
X_j + \lambda_j D_j & \text{if } f(X_j + \lambda_j D_j) < Y_j \\
X_j & \text{otherwise}
\end{cases}
\]
and set \( Y_{j+1} = f(X_{j+1}) \).

**Step 4.** Stop if a stopping criterion is met; otherwise set \( j \leftarrow j + 1 \) and go to Step 1.

Let \( a \) be a real number. We now recall a family of parameterized step-size distributions with densities proportional to \( 1/|\lambda|^a \) from Ghate and Smith [13]. Note when \( a > 0, 1/|\lambda|^a \rightarrow \infty \) as \( |\lambda| \rightarrow 0 \). For this technical reason, we let \( \epsilon > 0 \) be any real number negligibly small when compared with the diameter of the feasible region, for example, the machine precision. Suppose the iterate \( X_j = x_j \in S \) and the direction \( D_j = \xi_j \) for some \( j \), and let \( B(x_j, \epsilon) \) denote the \( n \)-dimensional ball of radius \( \epsilon \) centred at the current iterate position \( x_j \). The parameterized step-size \( \lambda_j \) is then generated from a probability density function that is proportional to \( 1/|\lambda|^a \) over \( \Lambda_j = \{ \lambda \in \mathbb{R} : x_j + \lambda \xi_j \in S \setminus B(x_j, \epsilon) \} \) and is uniform over \( \Gamma_j = \{ \lambda \in \mathbb{R} : x_j + \lambda \xi_j \in S \cap B(x_j, \epsilon) \} \). Note that \( \Gamma_j \cap \Lambda_j = \emptyset \) and \( \Gamma_j \cup \Lambda_j = L_j = \{ \lambda \in \mathbb{R} : x_j + \lambda \xi_j \in S \} \). For \( a \leq 0 \), we no longer have the above singularity issue, and \( \lambda_j \) is simply sampled from \( \Lambda_j = L_j \), and the region \( \Gamma_j = \emptyset \).

Given the iterate position \( x_j \) and direction \( \xi_j \), we henceforth use \( g^a_j(\lambda) \) to denote the conditional probability density function for the parameterized step-size \( \lambda_j \). Mathematically,
\[
g^a_j(\lambda) = \begin{cases} 
\frac{1/|\lambda|^a}{C_a(x_j, \xi_j)} & \text{for } \lambda \in \Lambda_j, \\
1/C_a(x_j, \xi_j) & \text{for } \lambda \in \Gamma_j,
\end{cases}
\]
where \( C_a(x_j, \xi_j) \) is a normalization constant that ensures that the density function integrates to one over \( L_j \). More specifically,
\[
C_a(x_j, \xi_j) = \int_{\lambda \in \Lambda_j} \frac{1}{|\lambda|^a} \, d\lambda + \int_{\lambda \in \Gamma_j} \, d\lambda.
\]

Parameterized IHR is the same algorithm as IHR except the step-sizes in Step 2 are generated according to the density function \( g^a_j \). We call this modified step in parameterized IHR Step 2’ stated as follows:

**Step 2’.** Generate a candidate point \( X_j + \lambda_j D_j \) at step-size \( \lambda_j \) distributed according to the density function \( g^a_j \) over the interval \( L_j = \{ \lambda \in \mathbb{R} : X_j + \lambda D_j \in S \} \) that defines the set of feasible step-sizes from the current iterate \( X_j \) along direction \( D_j \).

The performance of parameterized IHR intuitively should depend on the value of \( a \) employed in Step 2’. We investigate this with numerical results that demonstrate the disadvantages of using a single value of \( a \) fixed a priori. We use three values of \( a \), \(-1, 0, 1\), and four test problems from [1]: (1) Ackley problem (10-dimensional), (2) Rastrigin problem (10-dimensional), (3) Levy and Montalvo problem (10-dimensional) and (4) Bohachevsky problem (2-dimensional) (see Appendix 1, for detailed descriptions of these problems). We ran 100 independent trials each with several iterations (30,000 for the first three whereas 250 for the fourth since it is 2-dimensional) for each of these four functions. For a reasonable visual comparison, the function
value axes of our plots in Figure 1 were normalized to the interval \([0, 1]\). We plot the average of \((f(X_j) - y_\ast)/(f(X_0) - y_\ast)\) versus \(j\) over 100 trials.

For all the four problems, the plots show that \(a = 0\) always performs better than \(a = -1\). For the Ackley problem, \(a = 1\) takes the lead but gets surpassed by \(a = 0\) and \(-1\) in the middle. It is also shown that \(a = 0\) and \(-1\) outperform \(a = 1\) for the Bohachevsky problem, whereas \(a = 1\) does best for the Rastrigin problem. On the other hand, on the Levy and Montalvo problem \(a = 0\) and \(-1\) perform better initially and \(a = 1\) takes over gradually. All of the above observations suggest that parameterized IHR performance can be enhanced by using an adaptively chosen value for parameter \(a\) rather than fixing it a priori. This is implemented in APIHR, as discussed in the next section.

3. Adaptive parameterized Improving Hit-and-Run (APIHR)

The idea in APIHR is to choose, in each iteration, a value \(a\) that maximizes the probability of improvement. We show that, on spherical programs, the values of \(a\) can be reduced to considering simply \(a = 0\) and \(1\) without undermining the benefits of a high probability of improvement. Moreover, these two particular choices have their own significance: \(a = 0\) corresponds to the IHR algorithm in [35], and \(a = 1\) possesses unique efficiency properties shown in [13] for decentralized search. We use \(P_a(f(X_{j+1}) < f(X_j)|X_j = x)\), or \(P_a\) for brevity, to denote the conditional
probability of improvement associated with each $a$ value and modify Step 2′ in parameterized IHR to the following Step 2′ in APIHR:

Step 2′. If $P_0 > P_1$, generate step-size $\lambda_j$ distributed according to $\theta_j^0$, else generate step-size $\lambda_j$ distributed according to $g_j^1$ over the interval $L_j = \{\lambda \in \mathbb{R} : X_j + \lambda D_j \in S\}$ that defines the set of feasible step-sizes from the current iterate $X_j$ along direction $D_j$.

The key in an APIHR implementation is the calculation of the conditional probabilities $P_a$. We discuss how this can be accomplished for spherical and Lipschitz problems in the remainder of this paper. Numerical results are also presented.

### 3.1 Spherical programs

The optimization problem ($P$) is called a spherical program if the objective function $f$ can be written as

$$f(x) = h(r),$$

where $r = \|x - x_s\|_2$ and $h$ is a strictly monotonically increasing function of $r$ [35]. Here $\cdot \|_2$ is the Euclidean norm on $\mathbb{R}^n$. Also, the feasible region is an $n$-dimensional ball centred at $x_s$ with radius $q$, i.e. $S = B(x_s, q)$. We assume without loss of generality (since APIHR generates a sequence of iterates with non-increasing function values) that the initial point $X_0$ is on the boundary of $S$ and $Y_0 = y^* = h(q)$. Note that the level sets of a spherical program are nested spheres centred at the optimum $x_s$, and this made the analysis of IHR in [35] tractable. In this section, we present the central ideas and derivations involved in analysing the behaviour of our parameterized candidate generators on spherical programs. These will prove helpful in designing algorithms for solving Lipschitz continuous problems in the next section.

We are primarily interested in the conditional probability of improvement $P_a$ for the special case of the spherical program, which can be written as

$$P_a(f(X_{j+1}) < f(X_j)|X_j = x) = P_a(h(\|X_{j+1} - x_s\|_2) < h(\|x - x_s\|_2)|X_j = x),$$

$$= P_a(\|x + \lambda_j D_j - x_s\|_2 < \|x - x_s\|_2|X_j = x).$$

In Theorem 3.1, we provide the more general probability $P_a(\|x + \lambda_j D_j - x_s\|_2 < s|X_j = x)$, for $0 < s \leq \|x - x_s\|_2$, denoted $P_a(s, r_s)$ with $r_s = \|x - x_s\|_2$. Owing to rotational symmetry in spherical programs, it is more convenient to use a spherical coordinate system where the position $x$ of an iterate is defined by $r_s$ and direction $\xi$ by the angle $\theta_\xi$ that it makes with a reference direction connecting $x$ to $x_s$. In this notation, $C_a(x, \xi)$ in formula (2) is replaced by $\tilde{C}_a(r_s, \theta_\xi)$.

**Theorem 3.1** For any spherical program and any real $a$, given the current point $x \in S$ with $r = \|x - x_s\|_2$, the next candidate point sampled as in parameterized IHR has probability $P_a(s, r)$ for $0 < s \leq r \leq q$ given by

$$P_a(s, r) = \frac{2}{B(n - 1/2, 1/2)} \int_{\theta_0}^{\theta_+} \int_{\rho_-}^{\rho_+} \frac{\sin^{n-2}(\theta)}{\tilde{C}_a(r, \theta) \rho^a} d\rho d\theta,$$

(3)

where $\theta_0 = \arcsin s/r$, $\rho_- = r \cos \theta - \sqrt{s^2 - r^2 \sin^2 \theta}$, $\rho_+ = r \cos \theta + \sqrt{s^2 - r^2 \sin^2 \theta}$ (see Figure 2), $B(n - 1/2, 1/2)$ is the beta function and $\tilde{C}_a(r, \theta)$ is calculated in Appendix 2. Simplified expressions for $P_a(s, r)$ for four cases: $a < 0$, $a = 0$, $a > 0$ and $a \neq 1$, and $a = 1$, are given in Equation (A2–A5) respectively.
Figure 2. Illustration of probability of improvement calculations.

Proof See Appendix 3.

Corollary 3.2 When \( a \leq 0 \), the conditional probability of improvement for \( 0 < r \leq q \) is

\[
P_a(r, r) = \frac{2}{B(n - 1/2, 1/2)} \int_0^{\pi/2} \frac{(2r \cos \theta)^{1-a}}{l_+^{1-a} + l_-^{1-a}} \sin^{n-2}(\theta) \, d\theta,
\]

(4)

where \( l_\pm = \sqrt{q^2 - r^2 \sin^2 \theta} \pm r \cos \theta \) (see Figure 2 and Appendix 2).

Specifically for \( a = 0 \) we have a simplified expression

\[
P_0(r, r) = \frac{2}{B(n - 1/2, 1/2)} \int_0^{\pi/2} \frac{r \cos \theta}{\sqrt{q^2 - r^2 \sin^2 \theta}} \sin^{n-2}(\theta) \, d\theta.
\]

(5)

Proof Equation (4) follows from Theorem 3.1 immediately by letting \( s = r, \theta_0 = \pi/2 \) and substituting the corresponding \( \tilde{C}_a(r, \theta) \). Equation (5) follows from Equation (4) with \( a = 0 \), see also [35].

Corollary 3.3 When \( x \) is at least distance \( \epsilon \) from the boundary of \( S \), that is \( 0 < r \leq q - \epsilon \), the conditional probability of improvement for \( a > 0, a \neq 1 \) is bounded by

\[
P_a(r, r) \geq \frac{2}{B(n - 1/2, 1/2)} \int_0^{\arccos \epsilon/2r} \frac{(2r \cos \theta)^{1-a} - \epsilon^{1-a}}{l_+^{1-a} + l_-^{1-a} - 2\epsilon^{1-a} + 2\epsilon(1-a)} \sin^{n-2}(\theta) \, d\theta,
\]

(6)

and the conditional probability of improvement for \( a = 1 \) is bounded by

\[
P_1(r, r) \geq \frac{\int_0^{\arccos \epsilon/2r} \log(2r \cos \theta / \epsilon) \sin^{n-2}(\theta) \, d\theta}{B(n - 1/2, 1/2)(\log(\sqrt{q^2 - r^2 / \epsilon}) + \epsilon)}.
\]

(7)

Proof Due to the treatment of the \( \epsilon \) ball, for each of the two cases \( (a > 0, a \neq 1 \) and \( a = 1 \)), we have chosen to state a lower bound that excludes the (negligible) probability of sampling the
candidate point from the improving portion of the $\epsilon$ ball around the current iterate. In other words, the desired probability is bounded below by the probability of sampling from $B(x_*, r) \setminus B(x, \epsilon)$ (Figure 3). Notice that $\theta_0 = \arccos \frac{\epsilon}{2r}$, in addition the limits of integration on $\rho$ are $\rho_- = \epsilon$ and $\rho_+ = 2r \cos \theta$ for $\theta \in [0, \theta_0]$. The lower bound then follows immediately from Theorem 3.1.

We now devote our analysis to the value of parameter $a$ that maximizes the probability of improvement on each iteration of a spherical program for use in APIHR. In particular, the following theorem reveals important dominance relationships among $a \leq 0$ in terms of probability of improvement.

**Theorem 3.4** For any spherical program, $a = 0$ yields a higher probability of improvement than all $a \leq -1$ for all $0 < r \leq q$. In addition, $a = 0$ yields a higher probability of improvement than all $a \in (-1, 0)$ for $0 < r \leq q/3$.

**Proof** See Appendix 4.

The first part of Theorem 3.4 rules out the necessity to select any $a \leq -1$. Unfortunately we are unable to show strict dominance of $a = 0$ over $a \in (-1, 0)$ at every iteration due to the complicated expression of $P_a(r, r)$, but the second portion of Theorem 3.4 provides a partial analytical result that $a = 0$ dominates $a \in (-1, 0)$ for relatively small $r$ within the range $0 < r \leq q/3$. Numerically, as is shown in Figure 4a, $a = 0$ dominates all $a \in (-1, 0)$ over the entire range of $r$, and the blowup plot for large $r$ in Figure 4b numerically supports the dominance beyond $0 < r \leq q/3$. These analytical and empirical results support the conclusion that we only need to consider non-negative $a$ values, which is however not surprising – in our spherical program (unimodal) context, a negative $a$-value that can potentially guide the search to jump out of a local optimum is not needed.

We now turn our attention to $a \geq 0$. Since expressions (6) and (7) are too complicated to be analysed directly, we plot them using MATLAB in Figure 4c, with a blowup in Figure 4d. We select $a = 0, 0.5, 1, 1.5$ as representatives for different parameter values. We find unlike the case when $a \leq 0$, no individual $a$ dominates other values for the entire range of $r$. More specifically, for very large $r$ values (with iterate close to the boundary of the feasible region), $a = 0$ yields a higher probability of improvement, while as $r$ drops below a certain level large $a$ seems more promising.
To more efficiently investigate the dependence of improving probability on $a$, we plot $P_a$ as a continuous function of $a$, as shown in Figure 5. We choose two very large $r$ near the boundary of feasible region and two other intermediate values for $r$ because Figure 4c indicates that the superiority of $a$ depends on $r$, the distance of the current iterate to the centre of the spherical feasible region. For each selected $r$, a curve of $P_a$ over the range $a \in [-3, 4]$ is generated. The dash-dotted lines ($a < 0$) reinforce our earlier result that $a = 0$ dominates $a < 0$ for all $r$-values. In the range $0 \leq a \leq 1$, the dotted lines are either monotonically decreasing or increasing, depending on whether or not $r$ is close to the boundary. More interestingly, regardless of the value $r$ takes, the probability of improvement quickly ‘converges’ to $1/2$ shortly after $a$ exceeds 1; in other words, increasing $a$-value beyond 1 yields little or no extra amount of probability of improvement. This observation leads us to view the values of $a = 0$ and 1 as limiting values for the probability of improvement. Thus we focus our analysis on $a = 0$ and 1 for the rest of the paper.

In Figure 4c and d, we observe that for large values of $r$, $a = 0$ yields the higher probability of improvement, which unfortunately drops sharply to 0 as $r$ decreases. On the other hand, the probability of improvement for $a = 1$ drops at a much slower rate and dominates $a = 0$ for $r \geq r^*$. In other words, $r^*$ defines a ‘threshold optimal policy’ for spherical programs: for $r > r^*$ choose $a = 0$ and for $r \leq r^*$ choose $a = 1$. Moreover, if we fix $q$ and $\epsilon$, then as the dimension $n$ increases the intersection point, $r^*$ approaches $q$, the radius of the feasible region, as shown in Figure 6 for $q = 1000$, $\epsilon = 0.001$. This suggests that as $n$ gets large, the probability of improvement for $a = 1$ dominates that of $a = 0$ for almost all values of $r$. 

Figure 4. Probabilities of improvement for $n = 20$, $q = 1000$ and $\epsilon = 0.001$. 
Figure 5. Probability of improvement as a continuous function of $a$.

We provide numerical results on two spherical programs with objective functions

$$f_1(x) = \|x\|_2,$$

$$f_2(x) = \log(\|x\|_2 + 1) + 2\|x\|_2 + 1.\$$

For both test functions, we choose parameters $n = 20$, $q = 1000$, $\epsilon = 0.001$ as in our plots in Figures 4 and 5. We numerically estimate the threshold value $r^* = 969$ by equating the right-hand sides of Equations (5) and (7). We ran 100 trials of 10,000 iterations each for three cases: parameterized IHR for $a = 0$, $a = 1$ and APIHR. The average normalized function values are

Figure 6. Impact of dimension $n$ on the threshold radius $r^*$. 
plotted versus iterations in Figure 7. The plots show that APIHR reaches the optimal value after about 7000 iterations, whereas parameterized IHR with $a=0$ does not reach near the optimum value even after 10,000 iterations. However, at the beginning of the run, $a=0$ outperforms both $a=1$ and APIHR for many more iterations than required to pass the $r^*$ threshold. We believe this is due to IHR ($a=0$) making large improvements even though the probability of improvement is higher for $a=1$. Moreover, as expected, APIHR almost exactly follows parameterized IHR with $a=1$ since as noted earlier, $a=1$ yields the higher probability of improvement for almost all iterations (or in other words, for almost all values of $r$).

### 3.2 Extension to Lipschitz problems

In the previous section, our analysis focused on spherical programs where the probability of improvement for both $a=0$ and $1$ can be derived exactly. Now we extend the theory to problems with Lipschitz continuous objective functions. We continue to assume that the feasible region is spherical with radius $q$ and we allow the optimum $x_*$ to be located anywhere in $S$. Let $L$ be the Lipschitz constant of the objective function $f$ in problem (P). That is,

$$|f(x_1) - f(x_2)| \leq L\|x_1 - x_2\|_2 \quad \forall x_1, x_2 \in S.$$ 

For $f(x)$ the level sets are not necessarily spherical and thus we cannot obtain exact expressions of probability of improvement. However, based on the concept of a worst case function introduced in [34] and used in [26], we can derive lower bounds on these probabilities of improvement. Specifically, for problem (P), given the $j$th iterate $X_j$, the worst case function $w(z)$ is defined as

$$w(z) = \min\{f(X_j), f(x_*) + L\|z - x_*\|_2\}, \quad \forall z \in \mathbb{R}^n.$$ 

The key motivation behind this definition comes from three properties of the worst case function as noted in Lemma 3.5.

**Lemma 3.5** Given iterate $X_j$, let $S_{w(X_j)}$ and $S_{f(X_j)}$ be the corresponding level sets of $w(z)$ and $f(z)$ respectively, i.e.

$$S_{w(X_j)} = \{z \in S : w(z) < w(X_j)\}, \quad S_{f(X_j)} = \{z \in S : f(z) < f(X_j)\}$$
then we have the following:

1. \( w(X_j) = f(X_j) \),
2. \( \forall z \in S_{w(X_j)}, \ w(z) = f(x_a) + L\|z - x_a\|_2 \),
3. \( S_{w(X_j)} \subseteq S_{f(X_j)} \).

**Proof** By Lipschitz continuity, \( f(X_j) \leq f(x_a) + L\|X_j - x_a\|_2 \), thus

\[
w(X_j) = \min\{f(X_j), f(x_a) + L\|X_j - x_a\|_2\} = f(X_j).
\]

This proves Equation (1). From Equation (1) and definition of \( S_{w(X_j)} \) we know that \( \forall z \in S_{w(X_j)}, w(z) = w(X_j) > w(z) = \min\{f(X_j), f(x_a) + L\|z - x_a\|_2\}, \) thus \( w(z) = f(x_a) + L\|z - x_a\|_2 \). This proves Equation (2). From Equation (2) and the Lipschitz condition we get \( \forall z \in S_{w(X_j)}, f(X_j) = w(X_j) > w(z) = f(x_a) + L\|z - x_a\|_2 \geq f(z) \), which asserts \( z \in S_{f(X_j)} \) and hence \( S_{w(X_j)} \subseteq S_{f(X_j)} \). This proves Equation (3).

Fact (3) in Lemma 3.5 implies that the probability of sampling the candidate point from \( S_{f(X_j)} \), i.e., the probability of improvement, is bounded below by the probability of sampling the candidate from \( S_{w(X_j)} \). Moreover, the shape of \( S_{w(X_j)} \) is ‘nice’, in fact an open ball intersected with \( S \), as shown here

\[
S_{w(X_j)} = \{z \in S : w(z) < w(X_j)\} = \{z \in S : w(z) < f(X_j)\} \text{ from fact (1) above}
\]

\[
= \{z \in S : \min\{f(X_j), f(x_a) + L\|z - x_a\|_2\} < f(X_j)\} \text{ by } w(z) \text{ definition}
\]

\[
= \{z \in S : f(x_a) + L\|z - x_a\|_2 < f(X_j)\} \text{ since } \min\{a, b\} < a \iff b < a
\]

\[
= \left\{z \in S : \|z - x_a\|_2 < \frac{f(X_j) - f(x_a)}{L}\right\}.
\]

Therefore, we characterize set \( S_{w(X_j)} \) as \( B(x_a, f(X_j) - f(x_a)/L) \cap S \). The shape of \( S_{w(X_j)} \) in general depends (among other quantities) on the location of \( x_a \). We consider two cases below.

**Case 1** Global minimum at the centre of \( S \)

Suppose \( x_a \) is at the centre of \( S \), \( X_j = x \), let \( r_x = \|x - x_a\|_2 \) and \( s_x = f(x) - f(x_a)/L \). Then by Lipschitz continuity \( s_x \leq r_x \) and \( S_{w(X_j)} = B(x_a, s_x) \). Therefore,

\[
P_a(X_{j+1} \in S_{w(X_j)}|X_j = x) = P_a(X_{j+1} \in B(x_a, s_x)|X_j = x) = P_a(s_x, r_x).
\]

This observation leads to the following corollary from expressions (A3) and (A5) in Appendix 3.

**Corollary 3.6** For Lipschitz continuous optimization problems with a global minimum at the centre of a spherical feasible region \( S \), the probability of improvement is bounded below by

\[
P_0(X_{j+1} \in S_{w(X_j)}|X_j = x) = \frac{2}{B(n - 1/2, 1/2)} \int_0^{\Theta_0} \frac{\sqrt{s_x^2 - r_x^2 \sin^2 \Theta}}{\sqrt{q^2 - r_x^2 \sin^2 \Theta}} \sin^{n-2}(\Theta) \, d\Theta,
\]

(8)
Optimization Methods & Software

\[ P_1(X_{j+1} \in S_w(x_j) | X_j = x) = \frac{\int_0^{\theta_0} \left( \log r_x \cos \theta + \sqrt{s_x^2 - r_x^2 \sin^2 \theta} \right) \frac{\sin n - 2(\theta)}{r_x \cos \theta - \sqrt{s_x^2 - r_x^2 \sin^2 \theta}} d\theta}{B(n - 1/2, 1/2)(\log(\sqrt{q^2 - r_x^2}/\epsilon) + \epsilon)}, \quad (9) \]

where \( \theta_0 = \arcsin \frac{s_x}{r_x}, \quad r_x = \| x - x^* \|_2 \) and \( s_x = f(x) - f(x^*)/L \).

To gain more insight into the relationship between Equations (8) and (9), it is helpful to generate a plot of these lower bounds on probabilities of improvement (similar to Figure 4). We notice Equations (8) and (9) depend on not only \( r_x \) but also \( s_x \); however, when \( f(x) \) depends on \( x \) only through \( \| x \|_2 \), in other words, \( f(x) = h(r_x) \) where \( h \) is Lipschitz continuous, then by definition \( s_x \) becomes a function of \( r_x \) and hence both Equations (8) and (9) depend only on \( r_x \). Thus the plot of lower bounds on probabilities of improvement against \( r_x \) can be generated (Figure 8a and c).

To illustrate, we apply the above methodology to test problems 5 and 6 adapted from [10] (see Appendix 1 for detailed descriptions of these problems). Figure 8a and c shows the lower bounds on probabilities of improvement corresponding to \( a = 0 \) and 1, and Figure 8b and d shows the individual performance of parameterized IHR employing each value of \( a \). Figure 8 reveals a similar pattern as in the spherical program context: for large values of \( r_x \), \( a = 0 \) has a larger lower bound on probabilities of improvement and parameterized IHR with \( a = 0 \) makes faster improvement in terms of objective function value; however, as \( r_x \) gets below a certain value and closer to zero, \( a = 1 \) takes over.

![Figure 8](image-url)
Case 2  Position of global minimum unknown

In the previous case with \( x^* \) at the centre of \( S \), we showed that \( S_{w(X_j)} \) is a sphere entirely contained in \( S \). In the more realistic case where the location of \( x^* \) is anywhere in \( S \) and in particular is unknown, \( S_{w(X_j)} = B(x^*, s) \cap S \) may be a portion of a sphere. In this case we cannot compute the exact probability of sampling from \( S_{w(X_j)} \). We thus provide lower bounds on this probability for \( a = 0 \) and 1 in the following theorem and use them in our APIHR implementation. Note that as the location of \( x^* \) is unknown in this case, we wish to calculate lower bounds that do not depend on location of \( x^* \).

Let \( s_x = f(x) - f(x^*) \) as before and now we distinguish between \( r_x = \|x - x^*\|_2 \) and \( r_o = \|x - O\|_2 \), where \( O \) is the centre of \( S \). Note that since \( x^* \) is not necessarily at \( O \), \( r_x \) and \( r_o \) may be different. By Lipschitz continuity, we have \( s_x \leq r_x \) and note that \( r_x \leq 2q \).

We define a lower bound \( l_x \) on \( s_x/r_x \) as

\[
\frac{s_x}{r_x} \geq \frac{s_x}{2q} = \frac{f(x) - f(x^*)}{2qL} := l_x.
\]

When the current iterate is \( x \in S \), from an implementation view point, it is necessary to derive lower bounds on the probability of improvement that do not depend on \( r_x = \|x - x^*\|_2 \) and in particular the location of \( x^* \) since it is unknown but may depend on \( r_o = \|x - O\|_2 \). We accomplish this in the following theorem.

**Theorem 3.7**  For Lipschitz continuous optimization problems with location of the global minimum unknown, the probability of sampling from \( S_{w(X_j)} \) (which is the lower bound of probability of improvement) is further bounded below by

\[
P_0(X_{j+1} \in S_{w(X_j)} | X_j = x) \geq \frac{2 \int_0^{2\arcsin l_x/2} l_x \left( 1 - \cos \theta + \sqrt{l_x^2 - \sin^2 \theta} \right) \sin^{n-2}(\theta) \, d\theta}{B(n - 1/2, 1/2)}
\]

and

\[
P_1(X_{j+1} \in S_{w(X_j)} | X_j = x) \geq \frac{2 \int_0^{2\arcsin l_x/2} \left( \log 1/\cos \theta - \sqrt{l_x^2 - \sin^2 \theta} \right) \sin^{n-2}(\theta) / \hat{C}(r_o) \, d\theta}{B(n - 1/2, 1/2)},
\]

where \( l_x = f(x) - f(x^*)/2qL \) and

\[
\hat{C}(r_o) = \begin{cases} 
2 \log \frac{\sqrt{q^2 - r_o^2}}{\epsilon} + 2\epsilon & 0 < r_o \leq q - \epsilon, \\
\log \frac{q + r_o}{\epsilon} + \epsilon + q & q - \epsilon < r_o \leq q.
\end{cases}
\]

**Proof**  See Appendix 5.

3.3 Numerical results

Based on Theorem 3.7, we implement APIHR on test functions 5–12 (see Appendix 1 for detailed descriptions of these problems) and compare with parameterized IHR with \( a = 0 \) and 1. The implementation of APIHR compares the lower bounds in Theorem 3.7 at each iteration, and chooses the \( a \)-value with the higher lower bound. The average normalized function values obtained
Figure 9. Numerical results for eight Lipschitz continuous test functions.
Table 1. Numerical results on test problems 5–12.

<table>
<thead>
<tr>
<th>Test functions</th>
<th>Dim $d$</th>
<th>Total iterations</th>
<th>Optimal value</th>
<th>Mean best value found by</th>
<th>$p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>APIHR $a=0$ $a=1$ $a=0$ $a=1$</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>3000</td>
<td>$-2.6352$</td>
<td>$-2.6298$ $-2.3888$ $-2.5888$ $2.9E^{-005}$ $3.2E^{-001}$</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>3000</td>
<td>$1.1585$</td>
<td>$1.4712$ $1.5232$ $1.4842$ $3.8E^{-005}$ $8.5E^{-001}$</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>10</td>
<td>3000</td>
<td>$-0.3335$</td>
<td>$0.3497$ $0.5662$ $0.3397$ $5.3E^{-008}$ $4.0E^{-001}$</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>10</td>
<td>3000</td>
<td>$2.8140$</td>
<td>$4.7374$ $5.4903$ $4.8374$ $3.8E^{-005}$ $4.0E^{-001}$</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>10</td>
<td>3000</td>
<td>$-3.5000$</td>
<td>$-3.3833$ $-3.2907$ $-3.3809$ $3.9E^{-002}$ $9.6E^{-001}$</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>3000</td>
<td>$-3.5000$</td>
<td>$-3.0178$ $-2.6538$ $-3.0135$ $9.2E^{-031}$ $8.8E^{-001}$</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>5</td>
<td>1000</td>
<td>$-2.1951$</td>
<td>$-2.1265$ $-2.0887$ $-2.1177$ $3.1E^{-007}$ $2.3E^{-001}$</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>3</td>
<td>1000</td>
<td>$-1.0000$</td>
<td>$-0.9928$ $-0.9895$ $-0.9863$ $1.2E^{-001}$ $1.6E^{-001}$</td>
<td></td>
</tr>
</tbody>
</table>

over 100 independent trials of APIHR, $a=0$ and 1 are plotted versus iterations for these eight problems in Figure 9. In Table 1, we list relevant information and summarize the key numerical results for each test function: the first four columns exhibit respectively the problem number, dimension, total number of iterations as well as optimal value of each test function. Columns 5–7 show the mean best function value over the 100 independent runs. With the last iteration data of each trial, we statistically test whether the mean performance of APIHR is better than that of either $a=0$ or 1 individually, the $p$-values of the two sample $t$-tests are reported in last two columns of Table 1.

Figure 9 shows for most of the test problems initially $a=0$ does better but sooner (problem 5, 7–9 and 11) or later (problem 6) $a=1$ and APIHR catch up and outperform $a=0$. For the shifted sinusoidal problem (Problem 10), $a=1$ and APIHR dominate $a=0$ for the entire run. For problem 12, $a=0$ beats $a=1$ through the whole course and APIHR switches from $a=1$ to 0 in the middle. We notice from the last two columns of Table 1 that APIHR significantly outperforms $a=0$ (with $p$-value less than 0.05) on all test problems except 6 and 12, but none of the tests show that APIHR is significantly better than $a=1$. In fact, for problem 6, none of the parameter settings yield satisfying performance, while for problem 12, within 1000 iterations each of the three parameter settings can guide the search close enough to the global minimum, which thus dims the difference between them. We believe that this is due to the conservative lower bounds on probability of improvement that we used, so that APIHR almost always used $a=1$. Although our bounds on probability of improvement indicate $a=1$ is usually the best choice in APIHR, other performance measures such as expected amount of improvement may select $a=0$ or even negative $a$-values in multi-modal problems.

For problems 6–8, 10 and 11, the numerical results do not find the global optimum for these challenging functions. We intend to embed APIHR in a SA context with a cooling schedule. Even with no cooling schedule, APIHR demonstrates the benefit of tuning the step-size parameter.

4. Conclusions

We introduced an algorithm called APIHR, where the parameter $a$ influences the step-size distribution in the candidate generator. We switch the parameter $a$ between 0 and 1 based on the probability of improvement at each iteration. We derive exact expressions for probabilities of improvement for spherical programs and numerically evaluate the integrals to indicate a threshold policy for $a$. For the more general Lipschitz program, we derive lower bounds on improving probabilities and use them to adaptively select the parameter value. Numerical results on Lipschitz test problems demonstrate the overall improvement in performance achieved by APIHR over IHR with $a=0$. 
References


Appendix 1. Test problems

(1) **Ackley problem** [1]

\[
\min_x f(x) = -20 \exp \left( -0.02 \frac{\|x\|_2}{\sqrt{n}} \right) - \exp \left( \frac{1}{n} \sum_{i=1}^{n} \cos(2\pi x_i) \right) + 20 + e
\]

subject to \(-30 \leq x_i \leq 30, i \in \{1, 2, \ldots, n\}\). The global minimum \(x^*_a\) is located at the origin with \(f(x^*_a) = 0\). Our test was performed for \(n = 10\).

(2) **Rastrigin problem** [1]

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

subject to \(-5.12 \leq x_i \leq 5.12, i \in \{1, 2, \ldots, n\}\). The global minimum \(x^*_a\) is located at the origin with \(f(x^*_a) = 0\). Our test was performed for \(n = 10\).

(3) **Levy and Montalvo problem** [1]

\[
\min f(x) = 0.1 \left[ \sin^2(3\pi x_1) + \sum_{i=1}^{n-1} (x_i - 1)^2 (1 + \sin^2(3\pi x_{i+1})) + (x_n - 1)^2 (1 + \sin^2(2\pi x_n)) \right],
\]

subject to \(-5 \leq x_i \leq 5, i \in \{1, 2, \ldots, n\}\). The global minimum \(x^*_a\) is located at \(x^*_a = (1, 1, \ldots, 1)\) with \(f(x^*_a) = 0\). Our test was performed for \(n = 10\).

(4) **Bohachevsky problem** [1]

\[
\min f(x) = x_1^2 + 2x_2^2 - 0.3 \cos(3\pi x_1) - 0.4 \cos(4\pi x_2) + 0.7,
\]

subject to \(-50 \leq x_1, x_2 \leq 50\). The global minimum \(x^*_a\) is located at the origin with \(f(x^*_a) = 0\).

(5) Adapted from **Famularo et al.** [10]

\[
\min_{\|x\| \leq 2\pi} f(x) = -\frac{1}{500} \left( \frac{4}{\pi} \left( \|x\|_2 - \frac{3}{10} \right) - 4 \right)^6 + \frac{3}{100} \left( \frac{4}{\pi} \left( \|x\|_2 - \frac{3}{10} \right) - 4 \right)^4
\]

\[ -\frac{27}{500} \left( \frac{4}{\pi} \left( \|x\|_2 - \frac{3}{10} \right) - 4 \right)^2 + \frac{3}{2}.\]

The Lipschitz constant is \(L = 12.442132\) and the global minimum is located at the origin with \(f_* = -2.6352\). Our test was performed for \(n = 2\).
(6) Adapted from Famularo et al. [10]
\[
\min_{\|x\| \leq 5} f(x) = \sum_{i=1}^{5} \frac{1}{5} \sin((i + 1)\|x\|_2 - 1) + 2.
\]

The Lipschitz constant is \( L = 3.843648 \) and the global minimum is located at the origin with \( f(x_*) = 1.1585 \). Our test was performed for \( n = 2 \).

(7) Adapted from Famularo et al. [10]
\[
\min_{\|x\| \leq 4} f(x) = \exp(-\cos(-4\|x\|_2 - 7)) + \frac{1}{250(-4\|x\|_2 - 7)^2} - 1.
\]

The Lipschitz constant is \( L = 6.387862 \) and the global minimum is located at the origin with \( f(x_*) = -0.3335 \). Our test was performed for \( n = 10 \).

(8) Adapted from Famularo et al. [10]
\[
\min_{\|x\|_2 \leq 8.1} f(x) = -\cos(3(8.1 - \|x\|_2)) |(8.1 - \|x\|_2) \sin(8.1 - \|x\|_2)| + 8.1.
\]

The Lipschitz constant is \( L = 23.625414 \) and the global minimum is located at the origin with \( f(x_*) = 2.814 \). Our test was performed for \( n = 10 \).

(9) \textit{Sinusoidal problem} [26]
\[
\min_{\|x\|_2 \leq 90} f_5(x) = -2.5 \prod_{i=1}^{n} \sin(x_i + 90) - \prod_{i=1}^{n} \sin(5(x_i + 90)).
\]

The Lipschitz constant is \( L = 23.72 \) and the global minimum is located at the origin with \( f(x_*) = -3.5 \). Our test was performed for \( n = 10 \).

(10) \textit{Shifted sinusoidal problem} [26]
\[
\min_{\|x\|_2 \leq 90} f_6(x) = -2.5 \prod_{i=1}^{n} \sin(x_i + 60) - \prod_{i=1}^{n} \sin(5(x_i + 60)).
\]

This is a shifted version of problem 9. The Lipschitz constant is \( L = 23.72 \) and the global minimum is located at the origin \( x_* = (30, 30, \ldots, 30) \) with \( f(x_*) = -3.5 \). Our test was performed for \( n = 10 \).

(11) Adapted from Mladineo [23]
\[
\min_{\|x\|_2 \leq 1} f(x) = -\frac{1}{n} \sum_{i=1}^{n} x_i^2 + \prod_{i=1}^{n} \cos(10 \ln(i + 1)x_i) - 1.
\]

The Lipschitz constant depends on the dimension \( n \). Our test was performed for \( n = 5 \) and the Lipschitz constant \( L = 10 \) and the global minimum is located at the origin with \( f(x_*) = -2.1951 \).

(12) Adapted from Mladineo [23]
\[
\min_{\|x\|_2 \leq 4} f(x) = -\sin(x_1) \sin(x_1x_2) \sin(x_1x_2x_3).
\]

The Lipschitz constant is \( L = 19.39 \). There are multiple global minimal with \( f(x_*) = -1 \). Our test was performed for \( n = 3 \).
Appendix 2. Calculation of $\tilde{C}_a(r, \theta)$

In this appendix, we drop the subscript $j$ indicating iteration for ease of notation. Consider any point $x \in S$ with $r = \|x - x^*\|_2$ and angle $\theta$ corresponding to direction $D$. Let $\Lambda = \{ \lambda \in \mathbb{R} : x + \lambda D \in S \setminus B(x, \epsilon) \}$, $\Gamma = \{ \lambda \in \mathbb{R} : x + \lambda D \in S \cap B(x, \epsilon) \}$ and $L = \Gamma \cup \Lambda = \{ \lambda \in \mathbb{R} : x + \lambda D \in S \}$. As in Equation (2),

$$\tilde{C}_a(r, \theta) = \int_{\lambda \in \Lambda} \frac{1}{|\lambda|^a} d\lambda + \int_{\lambda \in \Gamma} d\lambda.$$  \hspace{1cm} (A1)

Also we define $l_-$ and $l_+$ as

$$l_+ = \sqrt{q^2 - r^2 \sin^2 \theta + r \cos \theta}, \quad l_- = \sqrt{q^2 - r^2 \sin^2 \theta - r \cos \theta}.$$

A.1 $a \leq 0$

For $a \leq 0$ we have $\epsilon = 0$ since there is no singularity issue. Therefore

$$\tilde{C}_a(r, \theta) = \int_{\lambda \in L} d\lambda = \int_{-l_-}^{l_+} |\lambda|^{-a} d\lambda = \frac{l_+^{1-a} + l_-^{1-a}}{1-a}.$$  \hspace{1cm} (A2)

Specifically when $a = 0$,

$$\tilde{C}_a(r, \theta) = 2\sqrt{q^2 - r^2 \sin^2 \theta}.$$  \hspace{1cm} (A3)

A.2 $a > 0, a \neq 1$

For $a > 0$, the step-size distribution is uniform within the $\epsilon$ ball and $1/|\lambda|^a$ elsewhere; therefore, $\tilde{C}_a(r, \theta)$ depends on the location of $x$, specifically on $r$. We treat $a = 1$ separately in Section A.3 because the term $1/|\lambda|^a$ integrates to a rational polynomial for $a > 0$ and $a \neq 1$. We consider two cases, when $x$ is further than $\epsilon$ from the boundary (case 1) and when $x$ is within $\epsilon$ of the boundary (case 2).

Case 1 When $0 < r \leq q - \epsilon$,

$$\tilde{C}_a(r, \theta) = \int_{-l_-}^{-\epsilon} \frac{1}{|\lambda|^a} d\lambda + \int_{-\epsilon}^{\epsilon} d\lambda + \int_{\epsilon}^{l_+} \frac{1}{|\lambda|^a} d\lambda = 2\epsilon + \frac{l_+^{1-a} + l_-^{1-a} - 2\epsilon^{1-a}}{1-a}.$$  \hspace{1cm} (A4)

Case 2 When $q - \epsilon < r \leq q$,

$$\tilde{C}_a(r, \theta) = \int_{-l_-}^{\epsilon} d\lambda + \int_{\epsilon}^{l_+} \frac{1}{|\lambda|^a} d\lambda = \epsilon + l_+ + \frac{(2q \cos \theta)^{1-a} - \epsilon^{1-a}}{1-a}.$$  \hspace{1cm} (A5)

A.3 $a = 1$

When $a = 1$, the term $1/|\lambda|$ integrates to a log expression rather than the rational counterpart for $a > 0, a \neq 1$. We again get two cases depending on the distance of $x$ from the boundary.
Case 1 When $0 < r \leq q - \epsilon$,
\[
\tilde{C}_a(r, \theta) = \int_{-\epsilon}^{-\epsilon} \frac{1}{\lambda} \, d\lambda + \int_{-\epsilon}^{\epsilon} \frac{1}{\lambda} \, d\lambda + \int_{\epsilon}^{l_+} \frac{1}{\lambda} \, d\lambda = 2\epsilon + 2 \log \frac{\sqrt{q^2 - r^2}}{\epsilon}.
\]

Case 2 When $q - \epsilon < r \leq q$,
\[
\tilde{C}_a(r, \theta) = \int_{-\epsilon}^{\epsilon} \frac{1}{\lambda} \, d\lambda + \int_{\epsilon}^{l_+} \frac{1}{\lambda} \, d\lambda = \epsilon + l_+ + \log \frac{l_+}{\epsilon}.
\]

Appendix 3. Proof of Theorem 3.1

Let $0 < s \leq r \leq q$ where $r = \|x - x_u\|_2$. Then we have [13,35]
\[
P_a(s, r) = \frac{2\Gamma(1 + n/2)}{\pi^{n/2}} \int_{\theta=0}^{\theta_0} \int_{\phi_1=0}^{\pi} \cdots \int_{\phi_{n-2}=0}^{\pi} \int_{\rho_-}^{\rho_+} \frac{d^n \rho}{C_a(r, \theta) \rho^{n+a-1}},
\]
where $d^n \rho$ is the hyperspherical volume element of the form
\[
d^n \rho = \rho^{n-1} \sin^{n-2}(\theta) \prod_{k=1}^{n-3} \sin^{n-2-k}(\phi_k) \, d\rho d\theta d\phi_1 \cdots d\phi_{n-2}.
\]
For $s \leq r \leq q$, we have \(\theta_0 = \arcsin s/r\) and \(\rho_- = r \cos \theta - \sqrt{s^2 - r^2 \sin^2 \theta}\), \(\rho_+ = r \cos \theta + \sqrt{s^2 - r^2 \sin^2 \theta}\) (Figure 2). So we get
\[
P_a(s, r) = A \int_{\theta=0}^{\theta_0} \int_{\rho_-}^{\rho_+} \frac{\sin^{n-2}(\theta)}{\tilde{C}_a(r, \theta) \rho^a} \, d\rho d\theta,
\]
where
\[
A = \left( \frac{2\Gamma(1 + n/2)}{\pi^{n/2}} \int_{\phi_1=0}^{\pi} \cdots \int_{\phi_{n-2}=0}^{\pi} \prod_{k=1}^{n-3} \sin^{n-2-k}(\phi_k) \, d\phi_1 \cdots d\phi_{n-2} \right).
\]
Note that the surface area element for an $n$-dimensional unit ball is given by
\[
d\omega = \sin^{n-2}(\theta) \prod_{k=1}^{n-3} \sin^{n-2-k}(\phi_k) \, d\phi_1 \cdots d\phi_{n-2} d\theta
\]
and its surface area is $n\pi^{n/2}/\Gamma(1 + n/2)$, so we have
\[
A = \frac{2\Gamma(1 + n/2) \int_{\theta=0}^{\pi/2} \int_{\phi_1=0}^{\pi} \cdots \int_{\phi_{n-2}=0}^{\pi} \sin^{n-2}(\theta) \prod_{k=1}^{n-3} \sin^{n-2-k}(\phi_k) \, d\phi_1 \cdots d\phi_{n-2} d\theta}{n\pi^{n/2} \int_{0}^{\pi/2} \sin^{n-2}(\theta) \, d\theta}.
\]
\[
= \frac{\Gamma(1 + n/2)}{n\pi^{n/2} \int_{0}^{\pi/2} \sin^{n-2}(\theta) \, d\theta} \left( 2 \int_{\theta=0}^{\pi/2} \int_{\phi_1=0}^{\pi} \cdots \int_{\phi_{n-2}=0}^{\pi} d\omega \right),
\]
\[
= \frac{\Gamma(1 + n/2)}{n\pi^{n/2} \int_{0}^{\pi/2} \sin^{n-2}(\theta) \, d\theta} \left( \frac{n\pi^{n/2}}{\Gamma(1 + n/2)} \right) = \frac{1}{B(\frac{1}{2}, \frac{1}{2})},
\]
where $B(n - 1/2, 1/2)$ is the beta function. Therefore we get

$$P_a(s, r) = \frac{2}{B(n - 1/2, 1/2)} \int_{\theta=0}^{\theta_0} \int_{\rho=0}^{\rho_+} \frac{\sin^{n-2}(\theta)}{C_a(r, \theta) \rho^a} \, d\rho \, d\theta.$$  

We can get more detailed expressions by substituting $\tilde{C}_a(r, \theta)$ for the four cases: (i) $a < 0$, (ii) $a = 0$, (iii) $a > 0$ but $a \neq 1$ and (iv) $a = 1$, and integrating with respect to $\rho$, where, for cases (iii) and (iv) we assume $x$ is further than $\epsilon$ from the boundary, i.e. $0 < r \leq q - \epsilon$, whereas we allow $0 < r \leq q$ for cases (i) and (ii). We obtain

$$P_{a<0}(s, r) = \frac{2}{B(n - 1/2, 1/2)} \int_{\theta=0}^{\theta_0} \frac{\rho_+^{1-a} - \rho_-^{1-a}}{l_+^{1-a} + l_-^{1-a}} \sin^{n-2}(\theta) \, d\theta,$$  \hspace{1cm} (A2)

$$P_0(s, r) = \frac{2}{B(n - 1/2, 1/2)} \int_{\theta=0}^{\theta_0} \frac{\sqrt{s^2 - r^2 \sin^2 \theta}}{\sqrt{q^2 - r^2 \sin^2 \theta}} \sin^{n-2}(\theta) \, d\theta,$$  \hspace{1cm} (A3)

$$P_{a>0,a\neq1}(s, r) = \frac{2}{B(n - 1/2, 1/2)} \int_{\theta=0}^{\theta_0} \frac{\rho_+^{1-a} - \rho_-^{1-a}}{l_+^{1-a} + l_-^{1-a} - 2\epsilon^{1-a} + 2\epsilon(1-a)} \sin^{n-2}(\theta) \, d\theta,$$  \hspace{1cm} (A4)

$$P_1(s, r) = \frac{\int_{\theta=0}^{\theta_0} (\log r \cos \theta + \sqrt{s^2 - r^2 \sin^2 \theta} \, r \cos \theta - \sqrt{q^2 - r^2 \sin^2 \theta} \, r \cos \theta \sin^{n-2}(\theta) \sin^{n-2}(\theta) \, d\theta}{B(n - 1/2, 1/2)(\log(\sqrt{q^2 - r^2/\epsilon}) + \epsilon)},$$  \hspace{1cm} (A5)

where

$$\rho_{\pm} = r \cos \theta \pm \sqrt{s^2 - r^2 \sin^2 \theta} \quad \text{and} \quad l_{\pm} = \sqrt{q^2 - r^2 \sin^2 \theta \pm r \cos \theta},$$

are introduced earlier in Theorem 3.1 and Appendix 2.

### Appendix 4. Proof of Theorem 3.4

Using Equations (4) and (5) in Corollary 3.2, we seek conditions under which the following inequality holds,

$$P_0(r, r) - P_a(r, r) = \frac{2}{B(n - 1/2, 1/2)} \int_0^{\pi/2} \left( 2r \cos \theta - \frac{(2r \cos \theta)^{1-a}}{l_+^{1-a} + l_-^{1-a}} \right) \sin^{n-2}(\theta) \, d\theta,$$

$$= \frac{2}{B(n - 1/2, 1/2)} \int_0^{\pi/2} 2r \cos \theta \left( \frac{1}{l_+^{1-a} + l_-^{1-a}} - \frac{(2r \cos \theta)^{-a}}{l_+^{1-a} + l_-^{1-a}} \right) \sin^{n-2}(\theta) \, d\theta \times \sin^{n-2}(\theta) \, d\theta \geq 0.$$

A sufficient condition for the above to hold true is

$$\frac{1}{l_+^{1-a} + l_-^{1-a}} - \frac{(2r \cos \theta)^{-a}}{l_+^{1-a} + l_-^{1-a}} \geq 0 \quad \text{for all } \theta \in [0, \pi/2],$$

which is equivalent to

$$l_+^{1-a} + l_-^{1-a} \geq (l_+ \times l_-)(2r \cos \theta)^{-a} \quad \text{for all } \theta \in [0, \pi/2].$$  \hspace{1cm} (A6)
Note $l_+ = l_- + 2r \cos \theta$, thus Equation (A6) is further equivalent to
\[
(l_- + 2r \cos \theta)^{1-a} + l_-^{1-a} \geq (2l_- + 2r \cos \theta)(2r \cos \theta)^{-a} = 2l_- (2r \cos \theta)^{-a} + (2r \cos \theta)^{1-a}. \tag{A7}
\]
Bernoulli’s inequality states that when $x > -1$,
\[
(1 + x)^h \geq 1 + hx \quad \text{for all } h \geq 1,
\]
and because $1 - a \geq 1$, we expand the left-hand side of Equation (A7) as
\[
(l_- + 2r \cos \theta)^{1-a} + l_-^{1-a} = \left(1 + \frac{l_-}{2r \cos \theta}\right)^{1-a} (2r \cos \theta)^{1-a} + l_-^{1-a}
\geq \left(1 + (1 - a)\frac{l_-}{2r \cos \theta}\right)(2r \cos \theta)^{1-a} + l_-^{1-a}
= (2r \cos \theta)^{1-a} + (1 - a)l_- (2r \cos \theta)^{-a} + l_-^{1-a}.
\]
Thus Equation (A7) would hold if
\[
(2r \cos \theta)^{1-a} + (1 - a)l_- (2r \cos \theta)^{-a} + l_-^{1-a} \geq 2l_- (2r \cos \theta)^{-a} + (2r \cos \theta)^{1-a},
\]
which can be simplified to
\[
l_-^{1-a} \geq (1 + a)(2r \cos \theta)^{-a} \quad \text{for all } \theta \in [0, \pi/2]. \tag{A8}
\]
When $l_+ = l_-$, $2r \cos \theta = 0$ hence Equation (A8) is trivially true.
When $l_+ > l_-$, Equation (A8) can be written as
\[
\left(\frac{l_-}{2r \cos \theta}\right)^{-a} = \left(\frac{l_-}{l_+ - l_-}\right)^{-a} \geq 1 + a. \tag{A9}
\]
Two sufficient conditions can be immediately derived to establish Equation (A9):
(1) $a \leq -1$, which makes $(l_-/l_+ - l_-)^{-a} \geq 0 \geq 1 + a$.
(2) $l_-/l_+ - l_- \geq 1$, or $l_- \geq l_+/2$, which makes $(l_-/l_+ - l_-)^{-a} \geq 1 > 1 + a$.

The first condition proves the first statement in Theorem 3.4. For the second condition, substituting $l_\pm = \sqrt{q^2 - r^2 \sin^2 \theta} \pm r \cos \theta$ into $l_- \geq l_+/2$, we get
\[
\sqrt{q^2 - r^2 \sin^2 \theta} \geq 3r \cos \theta. \tag{A10}
\]
Because Equation (A10) must hold for all $\theta \in [0, \pi/2]$, it is also true for $\theta = 0$, thus we get $r \leq q/3$ and the second part of the theorem is proved.

Appendix 5. Proof of Theorem 3.7

We begin with a preliminary lemma.

**Lemma A.1** Given the current iterate $X_j = x$ and the location of the optimum $x_*$ in $S$, consider two balls $B(x, r_x)$ and $B(x_*, s_x)$ with $r_x = \|x - x_*\|_2$ and $s_x = f(x) - f(x_*)/L$, and denote...
their two intersections by A and B. Then at least one of the points of intersection is in the feasible region; moreover, at least one of arc $Ax_*$ and arc $Bx_*$ is in the feasible region.

Proof Notice the three triangles $xOAx$, $xOx_*$ and $xOB$ share a common side $xO$, in addition $xA = xx_* = xB$ and $\angle OxA < \angle Oxx_* < \angle OxB$, so by the law of cosine we have $OA < Ox_* < OB$. Since $x_* \in S$, $Ox_* < q$, thus $OA < q$, i.e. $A \in S$.

Moreover, let $\hat{x}$ be any point lying on arc $Ax_*$, similarly we observe that triangles $xO\hat{x}$, $xOx_*$ share a common side $xO$, in addition $xx_* = xx_*$ but $\angle Ox\hat{x} < \angle Ox_*< q$ and hence $\hat{x} \in S$. Because $\hat{x}$ is arbitrarily chosen from arc $Ax_*$, we have arc $Ax_* \in S$ (Figure A1). ■

Following Lemma A.1 we have arc $Ax_* \in S$ and hence by convexity sector $xAx_* \in S$. To obtain a lower bound on the probability of improvement from $x$, we calculate the probability of sampling from the intersection of sector $xAx_*$ and $B(x_*, sx)$. Note that $||A - x_*||_2 = sx$ and $\theta_0 = 2 \arcsin (sx/2rx_*)$. Consider a sampled direction with angle $\theta$ to the reference line $xx_*$. For $\theta \in [0, \theta_0]$ the lower limit of $\rho$ is $\rho_- = rx_* \cos \theta - \sqrt{s_x^2 - r_x^2 \sin^2 \theta}$. Let $\phi$ be the angle between this sampled direction and line $xO$. Note that $\phi = \angle OxA + \theta_0 - \theta$ and $\phi$ depends on the location of $x_*$. From Appendix 2 the normalization constant corresponding to $a = 0$ is $\tilde{C}_a(r_o, \phi) = 2\sqrt{q^2 - r_o^2 \sin^2 \phi}$.

Using Theorem 3.1 and substituting $\tilde{C}_a(r_o, \phi)$, $P_0(X_{j+1} \in S_w(x_j) | X_j = x)$ is bounded below as

$$
\geq \frac{2}{B(n - 1/2, 1/2)} \int_0^{\theta_0} \int_{\rho_-}^{r_{x_*}} \frac{\sin^{n-2}(\theta)}{2\sqrt{q^2 - r_o^2 \sin^2 \phi}} \, d\rho \, d\theta
= \frac{1}{B(n - 1/2, 1/2)} \int_0^{\theta_0} \frac{r_{x_*} - \rho_-}{\sqrt{q^2 - r_o^2 \sin^2 \phi}} \sin^{n-2}(\theta) \, d\theta
$$

Figure A1. Illustration of the proof of Lemma A.1
\[ 1 \frac{r_x - r_x \cos \theta + \sqrt{s_x^2 - r_x^2 \sin^2 \theta}}{\sqrt{q^2 - r_o^2 \sin^2 \phi}} \sin^{n-2}(\theta) \, d\theta \]
\[ = 1 \frac{r_x \left(1 - \cos \theta + \sqrt{(s_x^2/r_x^2) - \sin^2 \theta} \right)}{q} \sin^{n-2}(\theta) \, d\theta \]
\[ \geq 1 \frac{s_x \left(1 - \cos \theta + \sqrt{(s_x^2/r_x^2) - \sin^2 \theta} \right)}{q} \sin^{n-2}(\theta) \, d\theta \]

since \( s_x \leq r_x \). Note \( s_x/r_x \geq s_x/2q = f(x) - f(x_*)/2QL = l_x \), so we get

\[ P_0(X_{j+1} \in S_w(X_j)|X_j = x) \geq 1 \frac{1}{B(n - 1/2, 1/2)} \times \int_0^{2\arcsin l_x/2} \frac{2l_x \left(1 - \cos \theta + \sqrt{l_x^2 - \sin^2 \theta} \right)}{\sin^{n-2}(\theta) \, d\theta} \]

which does not depend on the location of \( x_* \).

When \( a = 1 \), we similarly have that \( P_1(X_{j+1} \in S_w(X_j)|X_j = x) \) is bounded below as

\[ \geq 2 \frac{\log 1}{B(n - 1/2, 1/2)} \int_0^{\theta_0} \frac{\sin^{n-2}(\theta)}{C_a(r_o, \phi)} \, d\theta \]
\[ = 2 \frac{\log 1}{B(n - 1/2, 1/2)} \int_0^{\theta_0} \left( \frac{r_x}{r_x \cos \theta - \sqrt{s_x^2 - r_x^2 \sin^2 \theta}} \right) \sin^{n-2}(\theta) \, d\theta \]
\[ = 2 \frac{\log 1}{B(n - 1/2, 1/2)} \int_0^{\theta_0} \left( \frac{1}{\cos \theta - \sqrt{l_x^2 - \sin^2 \theta}} \right) \sin^{n-2}(\theta) \, d\theta \]
\[ \geq 2 \frac{\log 1}{B(n - 1/2, 1/2)} \int_0^{2\arcsin l_x/2} \left( \frac{1}{\cos \theta - \sqrt{l_x^2 - \sin^2 \theta}} \right) \sin^{n-2}(\theta) \, d\theta. \]

Since \( \phi \) depends on \( x_* \), we need to get an upper bound say \( \hat{C}(r_o) \) on \( C_a(r_o, \phi) \) that is independent of \( \phi \). This is accomplished by considering two cases depending on the position of \( x \) (specifically \( r_o \)) as in Appendix 2.

**Case 1** When \( 0 < r_o \leq q - \epsilon \),

\[ \hat{C}_a(r_o, \phi) = 2 \log \frac{\sqrt{q^2 - r_o^2}}{\epsilon} + 2\epsilon \equiv \hat{C}(r_o). \]
Case 2  When $q - \epsilon < r_o \leq q$,

$$\tilde{C}_a(r_o, \phi) = \log \frac{l_+}{\epsilon} + \epsilon + l_-,$$

where

$$l_+ = \sqrt{q^2 - r_o^2 \sin^2 \phi + r_o \cos \phi}, \quad l_- = \sqrt{q^2 - r_o^2 \sin^2 \phi - r_o \cos \phi}.$$

This is bounded above by

$$\tilde{C}_a(r_o, \phi) \leq \log \frac{q + r_o}{\epsilon} + \epsilon + q \equiv \hat{C}(r_o).$$

So we get

$$P_1(X_{j+1} \in S_w(X_j)|X_j = x) \geq \frac{2}{B(n - 1/2, 1/2)} \int_{0}^{2 \arcsin l_+/2} \left( \frac{1}{\cos \theta - \sqrt{l_+^2 - \sin^2 \theta}} \right) \sin^{n-2}(\theta) \frac{\hat{C}(r_o)}{\hat{C}(r_o)} d\theta,$$

which does not depend on the location of $x$, where

$$\hat{C}(r_o) = \begin{cases} 
2 \log \frac{q - r_o^2}{\epsilon} + 2\epsilon & 0 < r_o \leq q - \epsilon, \\
\log \frac{q + r_o}{\epsilon} + \epsilon + q & q - \epsilon < r_o \leq q.
\end{cases}$$