Journal of Global Optimization **20:** 33–47, 2001.
© 2001 *Kluwer Academic Publishers. Printed in the Netherlands.*

Implementing pure adaptive search for global optimization using Markov chain sampling

DANIEL J. REAUME[1], H. EDWIN ROMEIJN[2] and ROBERT L. SMITH[3,⋆]
[1]*Research and Development Center, General Motors, Warren, Michigan 48090 (e-mail: daniel_reaume@gmr.com);* [2]*Department of Industrial and Systems Engineering, University of Florida, 303 Weil Hall, P.O. Box 116595, Gainesville, Florida 32611-6595 (e-mail: romeijn@ise.ufl.edu);* [3]*Department of Industrial and Operations Engineering, The University of Michigan, Ann Arbor, Michigan 48109-2117 (e-mail: rlsmith@umich.edu)*

(Received 7 March 2000; accepted in revised form 14 November 2000)

Abstract. The Pure Adaptive Search (PAS) algorithm for global optimization yields a sequence of points, each of which is uniformly distributed in the level set corresponding to its predecessor. This algorithm has the highly desirable property of solving a large class of global optimization problems using a number of iterations that increases at most linearly in the dimension of the problem. Unfortunately, PAS has remained of mostly theoretical interest due to the difficulty of generating, in each iteration, a point uniformly distributed in the improving feasible region. In this article, we derive a coupling equivalence between generating an *approximately* uniformly distributed point using Markov chain sampling, and generating an *exactly* uniformly distributed point with a certain probability. This result is used to characterize the complexity of a PAS-implementation as a function of (a) the number of iterations required by PAS to achieve a certain solution quality guarantee, and (b) the complexity of the sampling algorithm used. As an application, we use this equivalence to show that PAS, using the so-called Random ball walk Markov chain sampling method for generating nearly uniform points in a convex region, can be used to solve most convex programming problems in polynomial time.

Key words: Global optimization, Markov chain sampling, Coupling, Complexity.

1. Introduction

Consider global optimization problems of the following form:

$$\min_{x\in S} f(x) \tag{P}$$

⋆ The work of this author was supported in part by (i) the National Science Foundation under NSF Grant DMI-9820744, (ii) the Department of Defense Research & Engineering (DDR& E) Multidisciplinary University Research Initiative (MURI) on “Low Energy Electronics Design for Mobile Platforms” and managed by the Army Research Office (ARO) under grant ARO DAAH04-96-1-0377; and (iii) by the Department of Defense Augmentation Awards for Science and Engineering Research Training (ASSERT) on “Optimization Algorithms for Low Power Mobile Platforms” and managed by the Army Research Office under grant ARO DAAG55-98-1-0155. The work of this author was also supported in part by a visiting professorship at the Delft University of Technology during the period January 1 through April 30, 1999.

where the objective function f is continuous over the feasible region $S \subset \mathbb{R}^d$, which is a compact, convex body, i.e., S is compact, convex, and its interior is nonempty.

The Pure Adaptive Search (PAS) algorithm was first developed by Patel, Smith and Zabinsky [14] for convex programming problems, and later extended to a large class of global optimization algorithms by Zabinsky and Smith [16]. It proceeds by generating a sequence of points in the feasible region S, with the property that each point is uniformly distributed in the level set corresponding to its predecessor. Under mild conditions, it was shown that the expected number of iterations required to obtain a solution with a given error increases at most *linearly* in the dimension of the problem. Bulger and Wood [3] extended this result under even milder conditions as a byproduct of studying a class of algorithms called Hesitant Adaptive Search.

What has prevented the practical use of PAS is that the problem of generating a uniformly distributed point in an arbitrary, or even convex, set is an extremely difficult one. However, progress has been made during the last decade using Markov chain sampling techniques. In particular, it is now possible to generate, in polynomial time, a point in a convex set that is *approximately* uniformly distributed. The first link between these methods and PAS has been made by Gademann [5], who developed a polynomial time implementation of PAS for linear programming problems.

In this paper we first derive a coupling equivalence between generating an approximately uniformly distributed point with certainty, and an exactly uniformly distributed point with a certain probability. Next we show that, using PAS, we can find a point that is approximately optimal with given probability in a linear number of iterations. These two results are then used to prove the main result of this paper: the computational complexity of finding a solution to (P) with prespecified probability and error is equal to the number of iterations required by PAS to achieve a certain solution quality guarantee (which, for certain global optimization problems, can be shown to be of the order of the dimension d of the problem) times the computational complexity of generating a uniformly distributed point in a full-dimensional compact set of dimension d with prespecified probability and error using a Markov chain sampling technique. This result implies that the question of finding the complexity of (approximately) solving an optimization problem reduces to the question of finding the complexity of (approximate) random sampling in certain classes of regions. As an application, we use the result to show that there exists a polynomial time implementation of the Pure Adaptive Search algorithm for most convex programming problems.

2. Approximate and Exact Sampling

Let $X = (X_n;\ n = 0, 1, 2, \dots)$ be a Markov chain defined on the measurable space $(S, \mathcal{B})$, where $\mathcal{B}$ is the restriction of the Borel σ-field on $\mathbb{R}^d$ to S. If the Markov chain has a limiting distribution π, then we can use this Markov chain to sample approximately from the distribution π: starting from some arbitrary point $x_0 \in S$, we simulate the Markov chain for a large number of iterations. The last point thus generated is approximately distributed according to π. Our goal in this section is to convert this result into one regarding the generation of points distributed *exactly* according to the limiting distribution π of the Markov chain.

This goal can be achieved, if we, instead of considering *deterministic* stopping times, consider *random* stopping times. The following example illustrates our motivation. Suppose we start with a point $X_0 = x_0$ in an d-dimensional hypercube C and randomly select, with equal probability, one of the coordinate directions e_i ($i = 1, \dots, d$). We then generate X_1 uniformly over the line segment in C passing through x_0 and parallel to e_i. Repeating this process, we obtain a Markov chain X over C with uniform limiting distribution (see Berbee et al. [2] and Bélisle, Romeijn and Smith [1]). For any finite deterministic n, X_n cannot be uniformly distributed over C since there is a positive probability of choosing the same coordinate direction in the first n iterations. On the other hand, if we have selected all d coordinate directions by the T-th iteration, then X_T is exactly uniformly distributed over C. Note that we do not have a contradiction since T is a random time, not a deterministic one.

To study such random times at which a Markov chain attains exactly a particular distribution, we turn to results from coupling theory. A *coupling* of two Markov chains X and X' is a random vector $(\widehat{X}, \widehat{X}')$ such that the distributions of X and X' are the marginal distributions of $(\widehat{X}, \widehat{X}')$. A *coupling time* T corresponding to a coupling $(\widehat{X}, \widehat{X}')$ of two Markov chains X and X' is a random variable T such that $\widehat{X}_n = \widehat{X}'_n$ for all $n \geqslant T$ (see Lindvall [9]).

Now let us assume that the Markov chain is *Harris recurrent with respect to* π (or π*-recurrent*), which means that, for all sets $B \in \mathcal{B}$ with positive measure $\pi(B) > 0$,

$$\Pr(X_n \in B \text{ for some } n \geqslant 1 | X_0 = x) = 1 \quad \text{for all } x \in S$$

and *aperiodic*, i.e., there exists no k-tuple $(A_1, \dots, A_k)$ of $k > 1$ disjoint sets in $\mathcal{B}$ such that (for all n)

$$\begin{aligned} \Pr(X_{n+1} \in A_{j+1} | X_n = x) = 1 \quad & x \in A_j,\ j = 1, \dots, k-1 \\ \Pr(X_{n+1} \in A_1 | X_n = x) = 1 \quad & x \in A_k \end{aligned}$$

(see Orey [13]). It will be convenient to denote the shifted sequence $\{X_k\}_{k=n}^{\infty}$ formed by discarding the first n iterates of X by $\theta_n X$.

A generalization of Goldstein's theorem (see Lindvall [9]) relates coupling times to the *total variation* distance between the distribution of the iterates of a Markov chain and the limiting distribution of the chain.

DEFINITION 2.1. *If π_1 and π_2 are probability measures over a measurable space $(E, \mathcal{F})$, then the* total variation distance *between π_1 and π_2 is equal to*

$$\|\pi_1 - \pi_2\| \equiv 2 \sup_{F \in \mathcal{F}} (\pi_1(F) - \pi_2(F)).$$

THEOREM 2.2. *For Markov chains $X = \{X_n\}_{n=0}^{\infty}$ and $X' = \{X'_n\}_{n=0}^{\infty}$, the following are equivalent:*

(i) There exists a coupling of X and X' with coupling time T such that

$$\| \Pr(\theta_n X \in \cdot) - \Pr(\theta_n X' \in \cdot)\| = 2 \Pr(T > n)$$

and

$$\lim_{n \to \infty} \Pr(T > n) = 0.$$

(ii) The total variation distance between the distributions of $\theta_n X$ and $\theta_n X'$ converges to zero as $n \to \infty$.

Proof.

(i) ⇒ (ii):

This is immediate by Definition 2.1.

(ii) ⇒ (i):

Goldstein's theorem and Lindvall [9] (sections III.14 and III.15) state that (ii) implies the existence of a coupling and corresponding coupling time with the property that

$$\| \Pr(\theta_n X \in \cdot) - \Pr(\theta_n X' \in \cdot)\| \leqslant 2 \Pr(T > n)$$

and

$$\lim_{n \to \infty} \Pr(T > n) = 0.$$

Lindvall [9] then strengthens this result to the existence of another coupling with the property

$$\| \Pr(\theta_n X \in \cdot) - \Pr(\theta_n X' \in \cdot)\| = 2 \Pr(T > n)$$

and

$$\lim_{n \to \infty} \Pr(T > n) = 0$$

which proves the desired result. □

To be able to apply Theorem 2.2, we convert total variation distances between the distributions of iterates of Markov chains to total variation distances between the distributions of the shifted chains with the following result.

THEOREM 2.3. *Let X and X' be two Markov chains over a measurable space $(S, \mathscr{B})$ having the same transition kernel, but possibly differing in their initial distribution. Fix $\delta > 0$ and suppose that the total variation distance between the distributions of the random variables X_n and X'_n is bounded from above by δ, for some $n \geqslant 0$. Then the total variation distance between the distributions of the stochastic processes $\theta_n X$ and $\theta_n X'$ is bounded from above by 2δ.*

Proof. Consider any set $B \in \mathscr{B}^\infty$. Let μ and λ denote the distributions of X_n and X'_n, respectively. Let $\widehat{B}$ denote the projection of B onto S equal to the set of n-th elements of each sequence in B.

$$\begin{aligned}
&\Pr(\theta_n X \in B) - \Pr(\theta_n X' \in B)\\
&\quad = \int_S \Pr(\theta_n X \in B | X_n = x)\, d\mu - \int_S \Pr(\theta_n X' \in B | X'_n = x)\, d\lambda\\
&\quad = \int_{\widehat{B}} \Pr(\theta_n X \in B | X_n = x)\, d\mu - \int_{\widehat{B}} \Pr(\theta_n X' \in B | X'_n = x)\, d\lambda\\
&\quad = \int_{\widehat{B}} \Pr(\theta_n X \in B | X_n = x)\, d\mu - \int_{\widehat{B}} \Pr(\theta_n X \in B | X_n = x)\, d\lambda
\end{aligned}$$

where the last equality is due to the chains sharing the same Markov kernel. Now denote the signed measure $\mu - \lambda$ by ν. By the Hahn decomposition Theorem (see Halmos [7]), there exists a partition of $\widehat{B}$ into sets $\widehat{B_1}$ and $\widehat{B_2}$ which are positive and negative with respect to ν. In other words, ν is a non-negative measure on $\widehat{B_1}$ while $-\nu$ is a non-negative measure on $\widehat{B_2}$. Hence, we have that

$$\begin{aligned}
&\Pr(\theta_n X \in B) - \Pr(\theta_n X' \in B)\\
&\quad = \int_{\widehat{B}} \Pr(\theta_n X \in B | X_n = x)\, d\nu\\
&\quad = \int_{\widehat{B_1}} \Pr(\theta_n X \in B | X_n = x)\, d\nu + \int_{\widehat{B_2}} \Pr(\theta_n X \in B | X_n = x)\, d\nu\\
&\quad = \int_{\widehat{B_1}} \Pr(\theta_n X \in B | X_n = x)\, d\nu - \int_{\widehat{B_2}} \Pr(\theta_n X \in B | X_n = x)\, d(-\nu)
\end{aligned}$$

Thus,

$$\begin{aligned}
&|\Pr(\theta_n X \in B) - \Pr(\theta_n X' \in B)|\\
&\quad = \left| \int_{\widehat{B_1}} \Pr(\theta_n X \in B | X_n = x)\, d\nu - \int_{\widehat{B_2}} \Pr(\theta_n X \in B | X_n = x)\, d(-\nu) \right|\\
&\quad \leqslant \int_{\widehat{B_1}} \Pr(\theta_n X \in B | X_n = x)\, d\nu + \int_{\widehat{B_2}} \Pr(\theta_n X \in B | X_n = x)\, d(-\nu)\\
&\quad = \nu(\widehat{B_1}) + (-\nu)(\widehat{B_2})\\
&\quad = \mu(\widehat{B_1}) - \lambda(\widehat{B_1}) + \lambda(\widehat{B_2}) - \mu(\widehat{B_2}).
\end{aligned}$$

But then, since

$$\| \Pr(X_n \in \cdot) - \Pr(X'_n \in \cdot)\| \leqslant \delta$$

we have from the definition of total variation distance that

$$\mu(\widehat{B}_1) - \lambda(\widehat{B}_1) \leqslant \tfrac{1}{2}\delta$$

and

$$\lambda(\widehat{B}_2) - \mu(\widehat{B}_2) \leqslant \tfrac{1}{2}\delta.$$

Hence

$$| \Pr(\theta_n X \in B) - \Pr(\theta_n X' \in B)| \leqslant \delta$$

and thus, again by definition of total variation distance, we have that

$$\| \Pr(\theta_n X \in \cdot) - \Pr(\theta_n X' \in \cdot)\| \leqslant 2\delta. \qquad \square$$

The following theorem is an application of Theorem 2.2:

THEOREM 2.4. *Consider an aperiodic Markov chain X that is Harris recurrent with respect to its stationary distribution π. In addition, let X' be a Markov chain having the same transition kernel, but with a possibly different initial distribution. Then there exists a coupling of X and X' with coupling time T such that, for all $n = 0, 1, 2, \ldots,$*

$$\| \Pr(\theta_n X \in \cdot) - \Pr(\theta_n X' \in \cdot)\| = 2 \Pr(T > n).$$

Proof. Since X and X' are aperiodic and Harris recurrent, the total variation distance between the distributions of X_n and X'_n converges to zero (as $n \to \infty$; see Orey [13]). Theorem 2.3 then tells us that the total variation distance between $\theta_n X$ and $\theta_n X'$ converges to zero as well (as $n \to \infty$). Theorem 2.2 now yields the desired result. □

We may now present the main results from this section. Theorem 2.5 shows that a point generated from a distribution that is within δ of some distribution π is statistically indistinguishable from a point that is, with probability at least $1-\delta$, *exactly* generated from the distribution π. Corollary 2.6 then uses this fact to conclude that, as soon as a Markov chain has been run long enough to generate iteration points having distribution close to its limiting distribution π in total variation, these points are essentially indistinguishable from points generated *exactly* from the distribution π.

THEOREM 2.5. *Suppose a Harris chain X over a measurable space $(S, \mathcal{B})$ is such that its first element, X_0, has distribution within δ in total variation distance from its limiting distribution π. Then, with probability at least $1 - \delta$, X_0 is distributed* exactly *according to π.*

Proof. Let X' be a Harris chain having the same transition kernel as X but whose initial distribution is its limiting distribution. By hypothesis, the total variation distance between the distribution of X_0 and the limiting distribution for X is less than δ. Now let T be a random variable such that X_T is the first iterate of X distributed exactly according to π. Applying Theorem 2.3 to X, the total variation distance between the distributions of $\theta_0 X = X$ and $\theta_0 X' = X'$ is at most 2δ. Since X is a Harris chain, we may apply Theorem 2.4 to X and X'. There therefore exists a coupling of X and X' such that the first coupling time T satisfies

$$\begin{aligned}\Pr(T = 0) &= 1 - \Pr(T > 0)\\ &= 1 - \tfrac{1}{2}\| \Pr(\theta_0 X \in \cdot) - \Pr(\theta_0 X' \in \cdot)\|\\ &\geqslant 1 - \tfrac{1}{2} \cdot 2\delta\\ &= 1 - \delta\end{aligned}$$

But by definition of a first coupling time, X_T is the first iterate of X distributed exactly according to the distribution of the corresponding iterate of X'. Since X' had initial limiting distribution, this completes the proof. □

COROLLARY 2.6. *Suppose a Markov chain X over a measurable space $(S, \mathscr{B})$ is such that its n-th element, X_n, has distribution within δ in total variation distance from its limiting distribution π. Then, with probability of at least $1 - \delta$, X_n is distributed exactly according to π.*

3. Pure Adaptive Search

We now return to the global optimization problem

$$\min_{x \in S} f(x). \tag{P}$$

Applied to (P), the PAS algorithm proceeds as follows:

Pure Adaptive Search (PAS)

Step 0. Set $k = 0$ and $y_0 = \infty$.

Step 1. Generate a point x_{k+1} uniformly distributed in $\{x \in S : f(x) < y_k\}$.

Step 2. Set $y_{k+1} = f(x_{k+1})$, increment k and return to step 1.

Without loss of generality, we can assume that the range of f on S is equal to $[0, 1]$, so that the optimal value of (P) is equal to 0. Moreover, we can define the error of a feasible solution x to (P) to be $f(x)$. Now define the number of iterations required for the convergence of PAS to a solution having an error of ε with probability $1 - \alpha$ as

$$K_{\alpha,\varepsilon} = \min\{k : \Pr(Y_k \leqslant \varepsilon) \geqslant 1 - \alpha\}$$

and let the random variable N_ε denote the number of iterations needed by PAS to obtain a solution to (P) with error at most ε.

Clearly, the problem instance (P) is characterized by the pair (S, f). A *class* $\mathcal{P}$ of global optimization problems is then a set $\mathcal{P}$ of pairs (S, f), each corresponding to a particular instance of (P). In the remainder, let $\mathcal{C}$ denote the class of convex programming problems, i.e., problems where S and f are both convex. Moreover, let $\mathcal{C}^1$ denote the class of convex programming problems having a unique optimum. For this latter class, Patel, Smith, and Zabinsky [14] derived an upper bound on the number of iterations required for the convergence of PAS to a solution having at most a given error. Their bound was improved by Schmeiser and Wang [15], yielding the following result.

THEOREM 3.1. *Consider a convex programming problem from the class* $\mathcal{C}^1$. *Then*

$$K_{\alpha,\varepsilon} \leqslant K_{\alpha,\varepsilon}^{(\mathcal{C}^1)} \equiv (d+1) \cdot \ln\left(\frac{1}{\alpha\varepsilon}\right).$$

Zabinsky and Smith [16] derived a similar result, which was improved upon by Bulger and Wood [3], for the class $\mathcal{G}_L$ of global optimization problems, where the objective function f is Lipschitz continuous.

THEOREM 3.2. *Consider a global optimization problem from the class* $\mathcal{G}_L$. *Let the diameter of the feasible region* S *be* Δ, *and the Lipschitz constant of the objective function* f *be* κ. *Then*

$$\mathcal{E}(N_\varepsilon) \leqslant 1 + d \cdot \ln\left(\frac{\kappa\Delta}{\varepsilon}\right)$$

where the random variable N_ε *denotes the number of iterations needed by PAS to obtain a solution with error at most* ε.

Note that the first result gives a bound on the number of iterations needed to obtain a point with at most a given error *with a certain probability*, while the second result gives a bound on the *expected* number of iterations needed to obtain a point with at most a given error. The following result shows a relationship between the two.

THEOREM 3.3. *Consider the problem* (P). *Let*

$$K_{\alpha,\varepsilon} = \min\{k : \Pr(Y_k \leqslant \varepsilon) \geqslant 1-\alpha\}. \tag{1}$$

Then

$$K_{\alpha,\varepsilon} < \frac{\mathcal{E}(N_\varepsilon)}{\alpha}.$$

Proof. Clearly,

$$\Pr(Y_k \leqslant \varepsilon) = \Pr(N_\varepsilon \leqslant k).$$

Thus,

$$K_{\alpha,\varepsilon} = \min\{k : \Pr(N_\varepsilon \leqslant k) \geqslant 1 - \alpha\}$$

which implies that both

$$\Pr(N_\varepsilon \leqslant K_{\alpha,\varepsilon}) \geqslant 1 - \alpha$$

and

$$\Pr(N_\varepsilon \leqslant K_{\alpha,\varepsilon} - 1) < 1 - \alpha$$

so that

$$\Pr(N_\varepsilon \geqslant K_{\alpha,\varepsilon}) > \alpha$$

But then

$$\begin{aligned} \mathcal{E}(N_\varepsilon) &= \sum_{k=1}^{\infty} \Pr(N_\varepsilon \geqslant k) \\ &\geqslant \sum_{k=1}^{K_{\alpha,\varepsilon}} \Pr(N_\varepsilon \geqslant k) \\ &\geqslant \sum_{k=1}^{K_{\alpha,\varepsilon}} \Pr(N_\varepsilon \geqslant K_{\alpha,\varepsilon}) \\ &= K_{\alpha,\varepsilon} \cdot \Pr(N_\varepsilon \geqslant K_{\alpha,\varepsilon}) \\ &> \alpha K_{\alpha,\varepsilon} \end{aligned}$$

yielding the desired result. □

For global optimization problems, this now yields

COROLLARY 3.4. *Consider a global optimization problem from the class $\mathcal{G}_L$. Let the diameter of the feasible region S be Δ, and the Lipschitz constant of the objective function f be κ. Then*

$$K_{\alpha,\varepsilon} \leqslant K_{\alpha,\varepsilon}^{(\mathcal{G}_L)} \equiv \frac{1}{\alpha} \cdot \left(1 + d \cdot \ln\left(\frac{\kappa\Delta}{\varepsilon}\right)\right).$$

4. Complexity of Pure Adaptive Search for Global Optimization

4.1. ORACLES AND COMPLEXITY

An often used way of measuring the complexity of algorithms is using the concept of *oracle calls*. An oracle is a ‘black box’ that performs tasks of a predefined type, such as checking membership of a point in a set or evaluating an objective function. This informational approach to complexity lends itself aptly to the discussion of optimization questions since objective and constraint functions may be arbitrarily difficult to compute.

If the maximum number of oracle calls required by an algorithm to solve a certain class of problems is bounded by a positive constant times some function γ depending on a number of problem characteristics, then the algorithm is said to have *complexity of order* $O(\gamma)$ with respect to the oracles employed. If γ is a polynomial function in its parameters, then the algorithm is said to be *polynomial*, or the algorithm is said to enjoy *polynomial complexity*.

It is possible that for all but a handful of pathological problem instances an algorithm may perform far better than the worst case complexity measure suggests. However, since there is no way to easily identify such troublesome problems a priori, much less the distribution of the frequency of occurrence of problems of varying difficulty, we must consider the worst case scenario to fairly and theoretically evaluate performance.

For our purposes, we consider three types of oracles. A *membership oracle* for a set S takes as input a point x and returns a yes/no answer as to whether or not x is contained in S. An *evaluation oracle* for a function f takes as input a point x in the domain of f and returns the value of $f(x)$. A *separation oracle* for a convex set S returns, when given a point x, either the assertion that $x \in S$ or a hyperplane h such that S is completely contained in the halfspace defined by h that does not contain x. For more information on oracles and complexity, we recommend Grötschel, Lovász and Schrijver [6].

4.2. COMPLEXITY OF PAS

In section 2 we have defined the concept of a Markov chain sampling algorithm. We can define the complexity of such a sampling algorithm, on a given measurable space $(S, \mathcal{B})$ and for a given limiting distribution π, as the number of oracle calls necessary to sample a point whose distribution is within some prespecified distance to π. When we consider a *class* of Markov chain sampling algorithms, we can define the complexity of this class of sampling algorithms as the maximum of the complexities over each element of the class.

Suppose that we have a class of Markov chain sampling algorithms, say $\mathcal{M}(\mathcal{R})$, for generating uniformly distributed points in each of the sets in any class of (finite-dimensional) sets $\mathcal{R}$. Moreover, let the complexity of this class of algorithms be given by $\gamma(d, \delta, \mathcal{R})$, i.e., for each d-dimensional set in $\mathcal{R}$, the number of oracle calls necessary to sample a point whose distribution is within δ in total variation distance of the uniform distribution on that set is at most equal to $\gamma(d, \delta, \mathcal{R})$.

Now let $\mathcal{G}$ denote a class of global optimization problems, and let

$$\mathcal{S}_{\mathcal{G}} = \{\{x \in S : f(x) < y\} : y \in (0, 1],\ (S, f) \in \mathcal{G}\}$$

i.e., $\mathcal{S}_{\mathcal{G}}$ is the set of all sets that can occur as a level set of a problem in the class $\mathcal{G}$. The corresponding class of Markov chain sampling algorithms is then $\mathcal{M}(\mathcal{S}_{\mathcal{G}})$. The following theorem contains the main result of this section.

THEOREM 4.1. *Consider global optimization problems of the form (P). Then the PAS algorithm applied to a problem from the class $\mathcal{G}$ provides, with probability at least $1 - \alpha$, a solution with an error of at most ε, using a number of oracle calls that is bounded from above by*

$$K \cdot \gamma(d, \delta, \mathcal{S}_{\mathcal{G}})$$

if

$$K \geqslant K^{(\mathcal{G})}_{\alpha/2,\varepsilon}$$

and

$$\delta \leqslant \frac{\alpha}{2K}$$

where $K^{(\mathcal{G})}_{\alpha/2,\varepsilon}$ is an upperbound on $K_{\alpha/2,\varepsilon}$ (as defined in equation (1)) for all global optimization problems (P) from the class $\mathcal{G}$.

Proof. First note that, by Corollary 2.6, the class of Markov chain sampling algorithms $\mathcal{M}(\mathcal{S}_{\mathcal{G}})$ yields points that are, with prespecified probability, exactly uniformly distributed. Now assume that we run the PAS algorithm using the Markov chain sampling algorithm for K iterations. Failure of this algorithm can have two causes. Firstly, in each of the iterations of the PAS algorithm, the Markov chain sampler could fail to generate an exactly uniformly distributed point in the level set under consideration. Secondly, even with exact uniform samples in each iteration, the PAS algorithm itself may fail to deliver a solution with error at most ε. Now let E denote the event that the PAS algorithm using the Markov chain sampler does not yield a solution with error at most ε, and let A denote the event that the Markov chain sampler yields an exactly uniformly distributed point in each of the iterations of PAS. Furthermore, let A^c denote the complement of event A. Then,

$$P(E) = P(E|A) \cdot P(A) + P(E|A^c) \cdot P(A^c) \leqslant P(E|A) + P(A^c).$$

The probability of error is at most α if we ensure that

$$P(E|A) \leqslant \tfrac{1}{2}\alpha \tag{2}$$

and

$$P(A^c) \leqslant \tfrac{1}{2}\alpha. \tag{3}$$

Inequality (2) follows easily if we choose $K \geqslant K_{\alpha/2,\varepsilon}$. Now consider event A. By the conditions in the theorem, and Corollary 2.6, we can sample an exactly uniformly distributed point in each iteration with at least some prespecified probability $1-\delta'$. Thus,

$$\begin{aligned} P(A) &\geqslant (1-\delta')^K \\ &\geqslant 1 - K\delta' \end{aligned}$$

(and $P(A^c) \leqslant K\delta'$). Thus, inequality (3) follows if we choose

$$\delta' \leqslant \frac{\alpha/2}{K}. \qquad \square$$

For the classes of Lipschitz continuous and convex programming problems introduced in section 3, we then obtain the following results.

COROLLARY 4.2. *Consider global optimization problems of the form (P). Then the PAS algorithm applied to a problem from the class $\mathcal{G}_L$ provides, with probability at least $1-\alpha$, a solution with an error of at most ε, using a number of oracle calls that is bounded from above by*

$$\frac{1}{\alpha} \cdot \left(1 + d \cdot \ln\left(\frac{\kappa\Delta}{\varepsilon}\right)\right) \cdot \gamma(d, \delta, \mathcal{S}_{\mathcal{G}_L})$$

with

$$\delta \leqslant \frac{\alpha}{\frac{2}{\alpha} \cdot \left(1 + d \cdot \ln\left(\frac{\kappa\Delta}{\varepsilon}\right)\right)}.$$

COROLLARY 4.3. *Consider convex optimization problems of the form (P). Then the PAS algorithm applied to a problem from the class $\mathcal{C}^1$ provides, with probability at least $1-\alpha$, a solution with an error of at most ε, using a number of oracle calls that is bounded from above by*

$$(d+1) \cdot \ln\left(\frac{1}{\alpha\varepsilon}\right) \cdot \gamma(d, \delta, \mathcal{S}_{\mathcal{C}^1})$$

with

$$\delta \leqslant \frac{\alpha}{2(d+1) \cdot \ln\left(\frac{1}{\alpha\varepsilon}\right)}.$$

5. Complexity of Pure Adaptive Search for Convex Programming

In this section we will make the result in Corollary 4.3 more concrete by considering sampling methods that can be used to sample uniformly distributed points in a large class of convex bodies.

5.1. POLYNOMIAL TIME MARKOV CHAIN SAMPLERS

Markov chain samplers have received a great deal of attention during the last decade, since it was shown that some of these enjoy polynomial time complexity, not only for sampling approximately uniformly distributed points, but also, for example, for estimating the volume of a convex body. We will review the literature on this topic here, and use this to obtain a polynomial time implementation of PAS for most convex programming problems.

5.1.1. *Rounding Convex Bodies*

In order to ensure that it is possible to find even a single point in the convex set S in finite time, it is reasonable to assume that this set is *well-guaranteed*. This means that we know the centers and radii of two spheres, one contained inside S and the other containing S. Note that, without the restriction imposed by the outer ball, S could be located anywhere in $\mathbb{R}^d$, and without the existence of the inner ball, S could be arbitrarily small. The *asphericity* or 'sandwiching ratio' of S denotes the ratio between the inscribed and circumscribed balls guaranteeing S, and is a measure of the 'roundness' of S.

As we will see below, the complexity of sampling from the convex body S depends strongly on the asphericity of the set. It is therefore desirable to transform S in such a way that the asphericity is reduced, preferably to be a polynomial function of the dimension d of S. (The transformed set is then often called *well-rounded*.) Hereby we restrict ourselves to affine transformations, so that an (approximately) uniformly distributed point in the transformed set, say $A(S)$, is (approximately) uniformly distributed in the original set S as well. The best known general result is that, for any convex set S, the Löwner-John ellipsoid (see Grötschel, Lovász and Schrijver [6]) defines an affine transformation A^{LJ}, such that $B \subseteq A^{\mathrm{LJ}}(S) \subseteq d\,B$ where B is the unit sphere centered at the origin. Unfortunately, no efficient algorithm for finding the transformation A^{LJ} exists. With the ellipsoid algorithm, however, it is possible to find an affine transformation A^{E} such that $B \subseteq A^{\mathrm{E}}(S) \subseteq d\sqrt{d}\,B$ using a number of oracle calls that is polynomial in the dimension d (see Grötschel, Lovász and Schrijver [6]).

5.1.2. *Random Ball Walk*

The walk is a simple Markov chain first presented by Lovász and Simonovits [11] and examined in greater depth in [12]. When applied to a convex body S, it has uniform limiting distribution and proceeds as follows:

Random ball walk

Step 0. Let $n = 0$ and let $x_0 \in S$.

Step 1. Increment n and generate a candidate point z_n uniformly over $B(x_n, \rho)$, the d-dimensional ball of radius ρ, centered at x_n.

Step 2. If $x_n \in S$, then let $x_n = z_n$, otherwise let $x_n = x_{n-1}$.

Step 3. Return to step 1.

Kannan, Lovász and Simonovits [8] show that, with appropriate choice of ρ, the Random ball walk, to generate points with distribution within δ of uniform in total variation distance in a convex body, has polynomial complexity in the dimension d, the asphericity R, and the precision δ.

THEOREM 5.1. *The number of iterations $\gamma_{\mathrm{BW}}(d, \delta, \mathcal{S}_{\mathcal{C}})$ needed by the Random ball walk, with appropriately chosen parameter ρ, to generate a point with distribution within δ in total variation distance from the uniform distribution over a set $S \in \mathcal{S}_{\mathcal{C}}$, where $\mathcal{S}_{\mathcal{C}}$ is the set of all finite-dimensional convex bodies, is polynomial in d and δ.*

Proof. As noted above, an affine transformation $A(S)$ of S having asphericity $R = d^c$ for some constant c can be found in polynomial time. By Kannan, Lovász and Simonovits [8], a point in $A(S)$ with distribution within δ in total variation of the uniform distribution can then be found using a polynomial number of oracle calls, which proves the desired result. □

5.2. COMPLEXITY OF PAS

Combining the results of the previous section with Corollary 4.3 yields the following complexity result for PAS applied to most convex programming problems.

THEOREM 5.2. *Consider convex programming problems in $\mathcal{C}^1$. Then the PAS algorithm, using the Random ball walk sampler, provides, with probability at least $1-\alpha$, a solution with an error in value of at most ε, using a number of oracle calls that is polynomial in d, α and ε.*

Proof. By Theorem 5.1, and using Corollary 2.6 a point that is, with suitably chosen minimal probability, uniformly distributed in S can be generated using a polynomial number of oracle calls, using the Random ball walk samping method. Since the same result holds for subsequent iterations of PAS, an application of Corollary 4.3 now yields that PAS requires a polynomial number of oracle calls to solve a convex programming problem in $\mathcal{C}^1$ with probability at least $1-\alpha$ and with error at most ε. □

6. Concluding Remarks

We have shown that Pure Adaptive Search can be implemented using Markov chain samplers for generating its iterates. Since we are assured that the number of iterates of Pure Adaptive Search grows at most linearly in the dimension of the problem for a large class of global optimization problems, the overall efficiency of the procedure rests on the efficiency of the Markov chain sampler used for obtaining each iterate.

Markov chain samplers are currently an extremely active area of research which has already achieved remarkable results (see e.g. Diaconis and Freedman [4] for an up-to-date survey, as well as the recent paper by Lovász [10]). Although we currently are guaranteed polynomial performance only for convex regions, the promise of polynomial samplers for more general non-convex regions offers the potential for polynomial procedures for truly global optimization problems.

References

1. Bélisle, C.J.P., Romeijn, H.E. and Smith, R.L. (1993), Hit-and-Run algorithms for generating multivariate distributions, *Mathematics of Operations Research* 18(2): 255–266.
2. Berbee, H.C.P., Boender, C.G.E., Rinnooy Kan, A.H.G., Scheffer, C.L., Smith, R.L. and Telgen, J. (1987), Hit-and-Run algorithms for the identification of nonredundant linear inequalities. *Mathematical Programming* 37: 184–207.
3. Bulger D.W. and Wood, G.R. (1998), Hesitant adaptive search for global optimisation. *Mathematical Programming* 81(1): 89–102.
4. Diaconis, P. and Freedman, D. (1998), Iterated random functions. Technical Report no. 511, Department of Statistics, University of California, Berkeley, California.
5. Gademann, A.J.R.M. (1993), *Linear Optimization in Random Polynomial Time*. PhD thesis, Department of Applied Mathematics, University of Twente, Enschede, The Netherlands.
6. Grötschel, M., Lovász, L. and Schrijver, A. (1993), *Geometric Algorithms and Combinatorial Optimization*. Springer-Verlag.
7. Halmos, P.R. (1950), *Measure Theory*. Van Nostrand, New York.
8. Kannan, R., Lovász, L. and Simonovits, M. (1997), Random walks and an $O^*(n^5)$ volume algorithm for convex bodies. Technical Report 1092, Department of Computer Science, Yale University, New Haven, Connecticut.
9. Lindvall, T. (1992), *Lectures on the Coupling Method*. Wiley.
10. Lovász, L. (1999), Hit-and-Run mixes fast. *Mathematical Programming*, 86(3): 443–461.
11. Lovász, L. and Simonovits, M. (1992), On the randomized complexity of volume and diameter. *IEEE Transactions*, pages 482–491.
12. Lovász, L. and Simonovits, M. (1993), Random walks in a convex body and an improved volume algorithm. *Random Structures and Algorithms*, 4(4): 359–412.
13. Orey, S. (1971), *Limit Theorems for Markov Chain Transition Probabilities*. Van Nostrand, New York.
14. Patel, N.R., Smith, R.L. and Zabinsky, Z.B. (1988), Pure adaptive search in Monte Carlo optimization. *Mathematical Programming* 43: 317–328.
15. Schmeiser, B.W. and Wang, J. (1995), On the performance of Pure Adaptive Search. In C. Alexopoulos, K. Kang, W.R. Lilegdon and D. Goldsman (eds), *Proceedings of the 1995 Winter Simulation Conference*, pages 353–356.
16. Zabinsky Z.B. and Smith, R.L. (1992), Pure adaptive search in global optimization. *Mathematical Programming* 53(3): 323–338.