

# HOPPING WITH AN ADAPTIVE HOP PROBABILITY DISTRIBUTION

Englander, A. C.<sup>+</sup>, Englander, J. A.<sup>#</sup>, and Carter, M.J.<sup>\*</sup>

Monotonic Basin Hopping (MBH) is a stochastic global search technique that may be used to design complex interplanetary trajectories. Previous research on MBH empirically demonstrated that a bi-polar Pareto distribution is an effective way to generate search “hops.” However no analytical foundation exists to explain this performance. In this work we provide the beginning of that analytical foundation and also introduce a new variant of MBH that uses an adaptive probability distribution to generate the hops. The technique is demonstrated on historical interplanetary trajectory design problems.

## INTRODUCTION

Monotonic Basin Hopping (MBH), a form of global random search that is often accompanied by a small-neighborhood gradient search, has been used to design and optimize interplanetary space-probe trajectories for many years.<sup>1,2,3,4</sup> The history of MBH goes back to the late 1990s when it was used in molecular modeling.<sup>5,6,7</sup> The random increments driving the random search, here referred to as  $\Delta x_p$ , are drawn from a probability distribution that we refer to as  $\text{pdf}(\Delta x_p)$  or simply  $p$ . Historically,  $p$  was Uniform or Gaussian, scaled according to the domain of the objective function  $f$ . In 2014, two of the current authors (Englander and Englander) demonstrated, empirically, that making  $p$  Bi-polar Pareto (here referred to as Iso-directional Pareto) provides significant benefits in terms of speeding up MBH time-to-convergence-to-solution for a broad family of  $f$  relevant to the design and optimization of interplanetary space-probe trajectories.<sup>8</sup> In our view, a satisfying analytical explanation was not provided. However, the empirical evidence was sufficiently compelling that the one of authors (J.E.) adopted the Iso-directional Pareto  $p$  as the default distribution for generating  $\Delta x_p$  in the NASA Goddard Space Flight Center’s Evolutionary Mission Trajectory Generator (EMTG) software.

This paper provides a summary of the analytical explanation that was missing from the 2014 paper, and we use that analysis to develop an adaptive  $p[t]$  that generates  $\Delta x_p[t]$  to further improve MBH time-to-convergence for  $f$  relevant to trajectory design and optimization. In addition, in the Appendix, we provide the widely-thought-to-be-missing proof for the asymptotic convergence of MBH, in order to justify our focus on speeding up, and improving the accuracy of, MBH convergence without discussion of whether convergence to the global minimum will ever occur.

We demonstrate the performance of this modified MBH on the re-resign and optimization of the historical Pioneer 11 mission, as modeled by the PyKep trajectory modeling toolbox provided open-source by the European Space Agency’s Advanced Concepts Team (ACT).<sup>9</sup>

---

<sup>+</sup>Department of Electrical and Computer Engineering, University of New Hampshire, Kingsbury Hall, Durham, NH 03824, email: [ace@englander-associates.net](mailto:ace@englander-associates.net)

<sup>#</sup> Navigation and Mission Design Branch, NASA Goddard Space Flight Center, Greenbelt, MD 20771

<sup>\*</sup> Department of Electrical and Computer Engineering, University of New Hampshire, Kingsbury Hall W319, Durham, NH 03824

When surveying the MBH literature it should be kept in mind that MBH is sometimes used as a generator, based on directed random search, of “landing sites” around which small-neighborhood gradient search is performed. When limited-range or limited-time local gradient search is incorporated into MBH, the optimizer does not get “stuck” in local minima because, in such cases, the MBH algorithm makes lists of what are referred to as “putative” minima, and compares them at the end of the optimization process to find the deepest (global) minimum. Much of the earliest MBH literature describes it being used in that manner.

Some current practitioners of MBH also use local gradient search to evaluate  $f$  within small neighborhoods around each new “hop” when:  $f$  is assumed to be locally smooth but not flat (otherwise the gradient search fails and wastes time); and/or the problem dimensionality is large; and/or there is reason to suspect that the global minimum may be located very near a constraint, discontinuity, or singularity; and the only risk to using a gradient search solver is its numerical instabilities which can be managed heuristically, and leads to, worst case, a waste of computing time. Much more effective in practice, when possible, is the use of local gradient search specifically based on small-neighborhood analytical derivatives. The use of analytical derivatives in MBH was pioneered by Ellison *et al.* (2018).<sup>10</sup>

No form of gradient search is used in the present work.

In 1977, Baba, Shoman, and Sawargi, working in the field of Operations Research, (OR) provided an asymptotic convergence proof for stochastic global optimization that is applicable to MBH.<sup>11</sup> They also provided necessary conditions for asymptotic convergence. Adaptation of their asymptotic convergence proof to MBH is contained in Appendix I. In 1981, Solis and Wets, also working the field of OR, re-stated and provided additional examples for the asymptotic convergence proof for stochastic global optimization provided earlier by Baba *et al.*<sup>12</sup>. Neither the 1977 paper by Baba, Shoman, and Sawargi, or the 1981 paper by Solis and Wets limited the finite number of dimensions in any way. The asymptotic convergence proof for stochastic global optimization by Baba *et al.* is summarized in the 2003 lecture slides by Aghassi.<sup>13</sup>

Less than twenty years after the 1981 paper by Solis and Wets, various forms of MBH emerged as effective ways to solve global non-convex optimization problems in molecular-modeling, and in space-probe trajectory design. In the excitement over being able to use MBH to solve such difficult non-convex optimization problems, the focus towards empirical results and away from analysis. The analysis of MBH convergence times, especially faster than asymptotic convergence times and how to estimate, predict, and accelerate them, remained un-addressed until the present work.

Wales and Doye provided one of the earliest papers on MBH.<sup>5</sup> They used a form of MBH, which they called “basin-hopping or Monte Carlo-minimization”, to find the lowest energy structures of Lennard-Jones clusters (a macro-molecular system) containing up to 110 atoms. Their success with that applied problem was a landmark achievement in global non-convex optimization. They described their method as having been designed to exploit features they believed were present in the energy landscapes of Lennard-Jones clusters, namely global non-convexity accompanied by local convexity in very small neighborhoods. To them, that was a prescription for MBH incorporating local gradient search into global random search even though Baba *et al.* had formally shown that local search was already inherent in random search if particular conditions are met. Thus, Wales and Doye began a tradition of MBH incorporating local gradient search into global random search. The analytical framework summarized in this paper, along with the empirical results presented here, show that for many  $f$  local gradient search is unnecessary, provided that  $p$  is “well suited” to  $f$  as described in this paper. Wales and Doye provided analytical support for the energy landscapes of Lennard-Jones clusters being globally non-convex but locally convex, but no analysis of the convergence of MBH.

In 2000, Leary discussed a form of MBH in his work on global optimization on funneling landscapes.<sup>6</sup> His application was “Molecular conformation problems arising in computational chemistry require the global minimization of a non-convex potential energy function representing the interactions of, for example, the component atoms in a molecular system.” Leary went on to say, “Typically, the number of local minima on the potential energy surface grows exponentially with system size, and often becomes enormous even for relatively modestly sized systems.”. He noted that MBH-like methods were well suited to molecular conformation problems because they found a small set of very deep local minima from which the global minimum could be chosen.

Locatelli and Schoen also applied MBH to Lennard-Jones clusters.<sup>7</sup> In their Abstract they describe a method “incorporating an already existing and quite successful method, monotonic basin hopping”. Locatelli and Schoen regarded basin hopping as a method for finding a set of deep local minima which they referred to as set of “putative” global minima in the domain being searched. They reported a method that “improves by two orders of magnitude the speed and the percentage of success in finding the global optima of clusters of 75, 98 and 102 atoms”. The global minimization of the energy landscape of Lennard-Jones clusters for these quantities of atoms ere, at the time, widely known as being the most challenging that had so far been solved. Locatelli and Schoen refer to their method as “a stochastic method for intelligently sampling the space of feasible solutions’ and emphasize that whereas the problem they solved solving was globally non-convex, it can be assumed to be locally convex, and that therefore one should incorporate some limited amount of gradient-search in the small neighborhood around each stochastically chosen sample.

Vasile, Minisci, and Locatelli provided detailed a comparison of global optimization methods that are applicable to space-probe trajectory design, including MBH.<sup>1</sup> Their paper was not a major contribution to MBH as much as it was a contribution to the analysis of planetary “transfers” modeled as Lambert’s problem. However, they used that analysis of the physics of planetary “transfers” to show that the problems were globally non-convex and, in small neighborhoods, locally convex. On that basis, they argued methods for optimizing such problems would benefit from a combination of global search and small-neighborhood gradient search.

In 2011, Yam, Lorenzo, and Izzo published the first solution for the semi-autonomous design and global optimization of low-thrust (ion) propelled space-probe missions in which the selection of targets for “fly-by” gravity assists are predetermined by humans, and the transfer trajectories are optimized by machine.<sup>2</sup> Such missions are called “fixed missions”. Yam, Lorenzo, and Izzo made significant and important contributions to the method and applications of MBH, including ways of measuring, documenting, and displaying the performance of empirical results. However, they did not provide any analysis of the convergence of MBH or address how to speed-up up the convergence of MBH.

Englander and Conway, based on Englander’s years-earlier Ph.D. thesis, published the first solution for the fully autonomous (machine-generated) design and global optimization of low-thrust (ion) propelled space-probe missions with variable missions.<sup>4</sup> (Variable missions are missions in which the selection of flyby targets for gravity assists, as well as the transfer trajectories, are machine-optimized.) Earlier, Englander, Conway, and Williams, published the first solution to fully autonomous design and global optimization of a variable mission sequence, using chemical propulsion.<sup>3</sup> Among Englander’s primary tools and methods was MBH. The landmark contribution of Englander’s earlier Ph.D. thesis was showing not only that fully autonomous design and global optimization of low-thrust space-probe missions with variable mission sequence could be accomplished, and but also a hybrid method (combining Genetic Algorithms with MBH).

In 2014, Englander and Englander published the first empirical evidence that, for MBH operating on real and high-dimensional space-probe trajectory optimization problems as well as

synthetic low-dimensional optimization problems, convergence times are significantly impacted by the probability distribution from which the random search increments are drawn<sup>8</sup>. Furthermore, for a wide variety of  $f$ , particular probability distributions speed-up MBH convergence much more than others. While the empirical results described in that paper were sufficiently compelling to lead that paper’s first author to use them to improve the MBH-based trajectory optimization software used by NASA Goddard Space Flight Center, that paper’s second author was dissatisfied by the lack of analytical support for those empirical results. The then-missing analytical support is provided in the present work.

In 2019, Izzo and the Advanced Concepts Team at the European Space Agency, published a set of Python modules named PyKep for solving Kepler’s models as solutions to Lambert’s problem.<sup>9</sup> A few years earlier, they had published models of historical space probe missions that could be designed and optimized by casting their objective functions in a form generated by previous versions of PyKep. PyKep and the published European Space Agency models of historical space probe missions, have been used in the present work to bench-mark test applications based the analytically supported framework described in this paper.

## OVERVIEW OF MBH

This paper describes an analytical framework for, and an application bench-mark test of, of a form of MBH that does not use any form of local gradient search; instead it uses an adaptive probability distribution from which the increments that drive the random search are drawn. Throughout this paper,  $t$  refers to MBH time-steps and should not be confused with flight time.

It is not the intent of this paper to argue the extent to which, or under what conditions, local gradient search is not needed in MBH. Rather, local gradient search was omitted here because it obscured the central theme of the analytical framework and might have made the significance of the empirical results less apparent.

This section of the paper is general with respect to any positive integer number of dimensions including 1-D, so, here, variables that would be vectors in dimensions higher than 1-D are not written in bold font.

The MBH algorithm used here is described after introducing some essential concepts:

Objective function  $f$  is a single value function defined over a domain  $X$ . Here,  $f$  is total thrust, the minimization of which is equivalent to minimizing total fuel consumption. The task of the optimizer is to find the optimal solution  $(x^*, f^*) = (\text{argmin}_{\text{global}}(f), \text{min}_{\text{global}}(f))$ , where  $\text{argmin}_{\text{global}}(f)$  is the point in  $X$  at which  $f$  is globally minimized.  $f$  may be non-convex in various challenging respects, and high-dimensional. In the case of the optimization of the Pioneer 11 trajectory,  $f$  is non-convex over a 3-dimensional  $X$ , and contains quasi-periodicities and a challenging singularity located very close to the optimal solution.

Objective function  $f$  is assumed to have a lower, but not necessarily an upper, bound.  $f$  may or may not have analytical or numerically estimable gradients. In any case, in the present work, gradients are not used. Whether or not some kinds of proxies for gradients are indirectly used is a matter of interpretation and argument.

$X$  is bounded but may be disconnected and may contain clusters of sub-spaces in which  $f$  cannot be evaluated (i.e. in which  $f$  is not “feasible”). In the application described in this paper,  $f$  is not “feasible” at the singularity that corresponds to the 180-degree transfer from Earth to Jupiter. (The problem is challenging, in part, because that singularity sits extremely close to the optimal solution.)  $X$  may have any finite number of dimensions, In the current application problem,  $X$  has

three dimensions: date of launch from Earth; date of arrival at Jupiter; date of arrival at Saturn. Those three dimensions indirectly “drive” several other dimensions of control variables, but only those three dimensions of decision variables are directly optimized.

$\min_{\text{global}}(f)$  is assumed to exist and be unique, although there may be arbitrarily many  $\min_{\text{local}}(f)$  that are arbitrarily located, and nearly as deep, relative to  $\min_{\text{global}}(f)$ . Multiple  $\min_{\text{local}}(f)$  nearly as deep as  $\min_{\text{global}}(f)$  may slow the time-to-converge, but they do not impact the guarantee of asymptotic convergence.

The stochastic optimization process searches for  $(x^*, f^*) = (\arg\min_{\text{global}}(f), \min_{\text{global}}(f))$  in a manner specified by the MBH algorithm, provided below in this section.

The search is driven by randomly generated search increments.  $p(\Delta x_p)$  is the pdf of the distribution from which the random search increments  $\Delta x_p$  are drawn.  $q(\Delta x_q)$  is the distribution of increments  $\Delta x_p$  that move the MBH towards  $(x^*, f^*)$ , written as  $q(\Delta x_q) = p(\Delta x_p \mid \Delta x_p \text{ results in a further minimization of } f)$ . Without having yet specified the MBH algorithm, it is already clear that  $q(\Delta x_q) = p(\Delta x_p \mid \Delta x_p \text{ results in a further minimization of } f)$ , implies  $q(\Delta x_q) = p(\Delta x_p \mid \Delta x_p \sim q(\Delta x_q))$ , which shows the benefit of achieving “similarity” between  $p$  and  $q$ .

Distribution  $q$  is characteristic of  $f$ . This is established by measure  $g[d]$  in the following manner:

$f$  is the objective function defined over domain  $X$ .  $p$  is the distribution of all hops  $\Delta x$ , which from here on are referred to as  $\Delta x_p \sim p$ .  $q$  is the distribution of “accepted” hops in the sense that  $x[t]$  is allowed to move, rather than be forced to remain-in-place, only when  $f[x[t] + \Delta x_p] < f[x[t-1]]$  in which case, by definition,  $\Delta x_p \sim q$  because  $q = p(\Delta x_p \mid \Delta x_p \text{ is “accepted”})$ .

$$g[d] = 1/\text{measure}(X) \int I(u) du, \forall u, \forall x \text{ in } X, \quad (1)$$

$$\text{where } I(u) = 1 \text{ if } f[v] < f[d] \forall v < u; 0 \text{ otherwise,} \quad (2)$$

and  $d$  is the depth of a slice through  $f$  across  $X$ .  $I$  is an indicator function, and  $u$  and  $v$  are “dummy variables” within the integral.

$g[d]$  applies to any  $f$  defined over any finite number of dimensions, because  $g$  nonetheless remains a functional and  $d$  remains a scalar parameter.

$g[d]$  can be interpreted as a measure that is proportional to the probability, at depth  $d$ , of a random probe finding a yet smaller value of  $f$  at depths deeper than  $d$ .  $g[d]$  is a decreasing function, where  $\max(g[d]) = 1$  when  $d = \max_{\text{global}}(f)$ , and  $\min(g[d]) = 0$  when  $d = \min_{\text{global}}(f)$ . That implies that as  $d \rightarrow$  smaller (implying descending deeper into  $f$ ), smaller values of  $f$  become increasingly rare and are less likely to be found. By geometric probability, smaller values of  $f$  being less likely to be found implies that smaller values of  $f$  are found less often. Thus  $g[d]$  is not only determined by  $f$ ;  $g[d]$  determines frequency of the event that  $x[t]$  advances, meaning  $\Delta x_p$  is “accepted”, meaning  $\Delta x_p \sim q$ .

Because  $g[d]$  is characteristic of  $f$ , and frequency of the event  $\Delta x_p \sim q$  is determined by  $g[d]$ ,  $q$  is characteristic of  $f$ . Thereby,  $g[d]$  establishes that distribution  $q$  is characteristic of  $f$ . Later,  $g[d]$  will be used to establish that  $q[t]$  is a time-varying distribution.

$q$  is not necessarily unique to  $f$ . Indeed, there may be equivalent classes  $\Theta = \{f_i\}$  such that all  $f_i \in \Theta$  share a very “similar” set of  $\{q_i\}$ . That depends on the similarity of the  $f_i$ , which is consistent with the analytic support constructed above the empirical results provided in Englander and Englander where some  $p$  were very effective on a wide range of  $f$ . The similarity of the  $\{f_i\}$  can be measured by suitable norms on their respective Fourier Transform Magnitudes or Wavelet Transform Magnitudes, and the similarity between any two  $(q_j, q_k)$  in  $\{q_i\}$  can be measured using Kullback-Leibler divergence discussed below.

Distribution  $q$  can be deterministically constructed by a deterministic tree search of  $f$  provided that  $f$  is known and a very large amount of computing power and time is available just for constructing and analyzing  $q$ . In practice,  $q$  is estimated by a Monte Carlo method in near real-time with a single on-going MBH.

Distribution  $q$  directly varies as to the proximity of  $f[x[t]]$  to  $f^*$ , and therefore with respect to  $t$ . Because,  $q$  is time-varying it is written as  $q[t]$ .  $g[d]$ , defined above, shows that  $q$  is time-varying by showing that, as each lower value of  $f$  is found, the frequency with which a yet-lower value of  $f$  is found typically decreases (the amount of time required to find a yet-lower value of  $f$  typically increases).

Because  $q[t]$  is time-varying, as  $t$  increases,  $q[t]$  develops a narrower mid-section, longer tails, and higher excess kurtosis. (The kurtosis of a probability density function is its fourth central moment divided by the square of its second moment.) That can be analytically proven, but that proof is outside the scope of this paper. Instead, Figure 8 illustrates  $\Delta x_p[t] \sim p[t]$  and  $\Delta x_q[t] \sim q[t]$ , with  $p[t]$  adapted to  $q[t]$  as explained below, and Figure 9 illustrates the histograms of  $q[t]$  and  $p[t]$ , respectively. In Figure 9, the narrow mid-section, long tails, and high excess kurtosis of  $q[t]$  and adapted  $p[t]$  are apparent. Whereas the kurtosis of a Gaussian pdf is 3.0 (mesokurtic), the kurtosis of  $q[t]$  is much larger (leptokurtic). In hindsight, this explains the empirical results reported in Englander and Englander, wherein the reported “best” performing pdf was narrow bodied, long tailed, and had high excess kurtosis – namely the so-called Bipolar Pareto distribution. This also explains why, in that paper, the performance of MBHs driven by Uniform and Gaussian distributions were relatively disappointing; neither of those distributions has excess kurtosis.

Distribution  $p$  can be chosen prior to the start of an MBH, as it was in Englander and Englander, or it can be adapted to  $q[t]$  during a single on-going MBH, as it is in the present work. Because  $q[t]$  is time-varying, adapting  $p[t]$  to  $q[t]$  in near real-time is more effective than using a  $p$  that is fixed. That is because an adaptive  $p[t]$  can be made more similar to  $q[t]$  than a fixed  $p$ . Therefore, an adaptive  $p[t]$  can be made to be more “better-suited” to  $f$ , the search can be made more “efficient”, and the convergence will therefore be sped-up. The goal is to make  $p[t]$  as “similar” to  $q[t]$  as possible, in near-real-time, without allowing the scale parameter of  $p[t]$  to decay to zero (in very rare pathological cases outside the scope of this paper, if the scale parameter of  $p[t]$  decays to zero, the MBH may converge to a local minimum).

Search “efficiency” will be given a formal definition later in this section, after the MBH algorithm is specified. Until then, an intuitive sense of “efficiency”, the ratio of useful output (“accepted” hops) to total input (all hops), is adequate.

The “similarity” between  $p$  (or  $p[t]$ ) and  $q[t]$  is also given a formal definition later in this section, where it is shown that “similarity” between  $p$  and  $q[t]$  is measurable using the Kullback-Leibler divergence  $D_{K-L}(p, q)$ . Here, an intuitive sense of “the similarity” of two probability distributions (e.g. in terms of the similarity of their shapes, or the numerical values of their respective mean, variance, skew, kurtosis, etc.) is adequate.

Having established essential concept and terminology above, the specification of the MBH algorithm (without adaptive  $p[t]$ ) follows:

1. At every time-step, given  $x[t]$  defined by the previous time-step or resulting from the random initialization of  $x[0]$ , evaluate  $f[x[t]]$ . For reasons that will become clear, except in the case of  $x[0]$ ,  $f[x[t]]$  can always be evaluated. Further, the random selection of  $x[0]$  can always be repeated and/or perturbed until  $f[x[0]]$  can be evaluated
2. Then, within the same time-step, draw  $\Delta x \sim p$  and generate  $\xi[t] = (x[t] + \Delta x[t])$
3. Evaluate  $f[\xi[t]]$  if  $f$  can be evaluated at  $\xi[t]$
4. Then, within the same time-step, if  $f[\xi[t]]$  can be evaluated, compare  $f[\xi[t]]$  to  $f[x[t]]$
5. In that same time-step, if  $f[\xi[t]]$  can be evaluated and  $f[\xi[t]] < f[x[t]]$ , then  $f[x[t]] \triangleq f[\xi[t]]$  and  $x[t] \triangleq \xi[t]$ ; else  $f[x[t]] \triangleq f[x[t]]$  and  $x[t] \triangleq x[t]$ , meaning that  $f[x[t]]$  and  $x[t]$  remain unchanged. Here,  $\triangleq$  denotes “is set equal to”.
6. Thus, in MBH, at every time-step,  $x[t]$  only moves if that move brings  $f[x[t]]$  closer to  $f^*$ . Otherwise  $x[t]$  is forced to wait-in-place

MBH can be extended to any finite number of dimensions using Cartesian or polar (meaning (direction, radius)) coordinates, in which case, the random increments are vectors  $\Delta \mathbf{x}[t]$  and the incumbent and candidate positions are vectors  $\mathbf{x}[t]$  and  $\xi[t]$ , respectively.

When describing the current application in other sections of this paper,  $\Delta \mathbf{x}[t]$ ,  $\xi[t]$  and  $\Delta \mathbf{x}[t]$  will signify vectors, and  $\Delta x_n[t]$ ,  $\xi_n[t]$  and  $x_n[t]$  will signify elements within those vectors indexed by dimensions  $n = 1, 2, 3$ .

It is common in the MBH literature to refer to the search increments as “hops”.

Whereas MBH was widely believed to lack an asymptotic convergence proof, one of the present authors (A.E.) found and adapted one from the historical literature of Operations research (Baba, *et al.*).<sup>10</sup> The adaptation and application to MBH of that asymptotic convergence is provided in the **APPENDIX**. That convergence proof enables the focus of the present work to focus be, instead, on the analysis, prediction, and speed-up of practical (much faster-than-asymptotic) MBH convergence times.

A contraction mapping is not the basis of the asymptotic convergence proof for MBH used too analytically support the method used in this paper, however the contraction mapping

$$\rho(f[x[t]], f[x^*]) \leq \rho(f[x[t-1]], f[x^*]) \quad (3)$$

applies to MBH. In the present work, this contraction mapping is used to analytically explain the decreasing availability of  $f[x] < f[x[t]]$  as  $x[t] \rightarrow x^*$ , which, by any asymptotic convergence proof applicable to MBH, means that as  $t$  grows large,  $x[t] \rightarrow x^*$ . By that decreasing availability of  $f[x] < f[x[t]]$  as  $t$  grows large, the present work establishes that statements that are directly supportable when stated in terms of  $x[t] \rightarrow x$  are indirectly supportable when stated in terms of the passage of time.

Given the definition of MBH (numbered steps 1 through 6 above), the “efficiency” of an MBH searching for  $f^*$ , and, the “suitability” of  $p$  to  $f$ , are defined as follows:

Recall that  $\Delta x[t]$  such that  $\xi[t] = (x[t] + \Delta x[t])$ ,  $x[t]$  remains in  $X$  or is relocated so that it remains in  $X$ ,  $f[\xi[t]]$  can be evaluated, and  $f[\xi[t]] < f[x[t]]$  so  $f[x[t]]$  is set to  $f[x[t]]$  and  $x[t]$  is set to  $\xi[t]$ . The present work refers to such  $\Delta x$  as “accepted”  $\Delta x$ . Otherwise  $\Delta x$  referred to as “rejected”  $\Delta x$ .

Let  $I[t] = 1$  if  $\Delta x[t]$  is “accepted”; otherwise  $I[t] = 0$ . (4)

The “efficiency”  $\eta$  of an MBH in its search for  $f^*$  on a particular  $f$ , during the time interval from  $t = \tau_a$  to  $t = \tau_b$  inclusive is:

$$\eta[\tau_a : \tau_b] = (1/(\tau_b - \tau_a)) \sum_{j=\tau_a}^{\tau_b} I(j) \quad (5)$$

where  $[\tau_a : \tau_b]$  means from  $t = \tau_a$  to  $t = \tau_b$ , inclusive, for  $\tau_b > \tau_a$ . Clearly, the more often  $\Delta x[t]$  is accepted, the more frequently  $f[x[t]] < f[x[t-1]]$ , and the more “efficient” the MBH search for  $f^*$  is when it is operating on given a  $f$  during time interval  $[\tau_a : \tau_b]$ .

Note that, “efficiency”  $\eta$  is defined in terms of  $\Delta x$ . In the present work, the analytic support work makes the simplifying assumption that analyzing the pdf  $p$ , from which  $\Delta x$  are drawn, is an adequate way to describe  $\Delta x$  and thereby analyze search “efficiency”. When  $f$  contains particular complicated hyper-geometries, such as multi-dimensional periodicities and singularities, some properties of the statistical dynamics of the  $\Delta x[t]$  may matter. Within the current analytical framework, statistical dynamics of the  $\Delta x[t]$  are of secondary importance compared to pdfs, and so are left for future investigation. Nonetheless, “efficiency” is shown to be the result of the  $\Delta x$  being “accepted” frequently, which (in large part) implies that the  $p$  from which the  $\Delta x$  are drawn, is, in that sense, “well-suited” to  $f$ .

The importance of  $p$ , and its impact on MBH convergence times, was overlooked throughout the decades-long history of the literature and applications of MBH. Englander and Englander provided the first paper about the role of  $p$  (not referring to  $p$  by name), but that paper was motivated by intuition and the investigation and supported only by empirical evidence. The analytical framework presented here is the first to address search “efficiency” in terms of  $p$  being “well-suited to  $f$ ”, meaning, as will now be shown,  $p$  being measurably similar  $q$ .

From here on, “accepted”  $\Delta x$  is referred to as  $\Delta x_q$ , and the super-set (super-sequences) all  $\Delta x$ , whether they are “accepted” or “rejected”, is referred to as  $\Delta x_p$ . By the definition of  $\Delta x_q$  being a  $\Delta x_q$  that was “accepted”,  $q(\Delta x_q) = p(\Delta x_p \mid f[x[t] + \Delta x_p] < f[x[t]])$ , meaning  $q(\Delta x_q) = p(\Delta x_p \mid \Delta x_p \text{ is “accepted”})$ , meaning  $q(\Delta x_q) = p(\Delta x_p \mid \Delta x_p \sim q(\Delta x_q))$ .

Distribution  $p$  is said to be “well-suited” to a particular  $f$  if most of the  $\Delta x_p$  are “accepted” when the MBH is operating on that  $f$  and, therefore,  $\eta[\tau_a : \tau_b]$  is equal to or greater than  $1/2$  in most intervals  $[\tau_a : \tau_b]$ . The more “well-suited” is to a particular  $f$ , the more “efficient” an MBH operating on that  $f$  using that  $p$  will be, and the closer  $\eta[\tau_a : \tau_b]$  will be to 1.

From  $q(\Delta x_q) = p(\Delta x_p \mid \Delta x_p \sim q(\Delta x_q))$  it is clear that  $p$  is “well-suited” to a particular  $f$  when  $p$  is “similar” to  $q(\Delta x_q)$  in the sense of  $p$  frequently generates  $\Delta x_p \sim q(\Delta x_q)$ .

A formal measure for the similarity between  $p$  and  $q$  is the Kullback-Leibler divergence  $D_{K-L}(p, q)$ . The Kullback-Leibler divergence  $D_{K-L}(p, q)$ , sometimes written as  $D_{KL}(p \parallel q)$ , is a measure in statistics that quantifies how “close” a binned probability  $p_i$  in a distribution  $p$  is to a model binned probability  $q_i$  in a distribution  $q$ , where the distributions are probability mass functions with equal width and aligned bins, and  $p_i$  and  $q_i$  are the respective normalized frequencies of occurrence in  $i^{\text{th}}$  bins  $p_i$  and  $q_i$ .<sup>13</sup>

$$D_{K-L}(p, q) = \sum_i p_i (\log_b \frac{p_i}{q_i}), \text{ where } b \text{ is the base of the logarithm} \quad (6)$$

In information theory  $b = 2$  and the units of  $D_{K-L}(p, q)$  are “bits”. In the present work,  $b = e$  (the base of the “natural” logarithm) and the units of  $D_{K-L}(p, q)$  are “nats”.

The Kullback-Leibler divergence has been shown to be related to likelihood theory.<sup>15</sup> However, here it is adequate to say, intuitively, that  $D_{K-L}(p,q)$  “captures” and compares properties of distributions that would otherwise require the extraction of a large number of moments. In that sense,  $D_{K-L}(p,q)$  is not only effective, but also parsimonious and computationally simple as long as the probability mass functions are properly normalized, and their bins are properly constructed and aligned. By the definition of entropy used in information theory,  $D_{K-L}(p,q)$  is the relative entropy between  $p$  and  $q$ .

It is clear that as  $p$  becomes more “similar” to  $q$ ,  $D_{K-L}(p,q) \rightarrow 0$  because  $\sum_i p_i \rightarrow 1$ ,  $\frac{p_i}{q_i} \rightarrow 1 \forall i$ , and  $\log_b \frac{p_i}{q_i} \rightarrow 0$ .

The use of the Kullback-Leibler divergence  $D_{K-L}(p,q)$  to measure the “similarity” between  $p$  and  $q$ , and thereby the search efficient of MBH operating of  $f$  by using  $p$ , has an interesting historical background. The Kullback-Leibler divergence was developed, and is used, in information theory. There it is interpreted as the “inefficiency” of a coder coding messages built from an alphabet distributed as to probability mass function  $p$  when, in fact, the messages were built from an alphabet distributed as to probability mass function  $q$ . That “inefficiency” due to the dis-similarity between  $p$  and  $q$  specifies the number of additional “bits” of information that must be included in the coded message in order to increase the likelihood of it being decoded correctly. As  $p \rightarrow q$ ,  $D_{K-L}(p,q) \rightarrow 0$  and the coder is more “efficient”. It is therefore all-the more interesting that  $D_{K-L}(p,q)$  can be used to provide analytic support for the notion of “efficiency” in MBH.

The “similarity” of  $p$  to  $q$  can be achieved either by adapting  $p[t]$  to a near real-time estimate of  $q[t]$ , or by empirically making a fixed  $p$  that is determined to be “well-suited” to  $f$ . The latter is possible (to a limited degree) because making  $p$  “well-suited” to  $f$  has an obvious beneficial impact on convergence MBH times, so even a trial-and-error approach is rewarding. That was the method used in Englander and Englander. Methods for adapting  $p[t]$  to estimates of  $q[t]$ , in real real-time within a single on-going MBH, are provided in the **APPLICATION** section of this paper.

That  $q(\Delta x_q)$  is a time-varying pdf  $q[t]$  that typically has a monotonically non-increasing scale parameter was established above using  $g[d]$  (Equations 1 and 2). But even without  $g[d]$ , it is intuitive that as  $x[t] \rightarrow x^*$ ,  $\|\Delta x_q\|$  needs to decrease in order for  $\xi[t]$  not to “overshoot”  $x^*$ , or else  $\xi[t] = (x[t] + \Delta x[t])$  would not be accepted and become  $x[t]$ . It is  $\xi[t] = (x[t] + \Delta x[t])$  being “accepted” that makes  $\Delta x[t] \sim q(\Delta x_q)$ . Therefore  $q[t]$  must have a decreasing scale factor as  $x[t] \rightarrow x^*$ . That dependency on  $x[t]$  becomes a dependency on  $t$  by either the asymptotic convergence proof by Baba, *et al.*, or by the contraction mapping described in above. Thus,  $q[t]$  must have a decreasing scale factor as  $t$  grows large.

In practice,  $q[t]$  is estimated using a Monte Carlo method, which requires that one assume that an estimable  $q[t]$  exists and has, at least initially, an assumed functional form. Here, that initial functional form is assumed to be approximately a Laplace distribution. The basis for this assumption is  $g[d]$ , which shows that  $\|\Delta x_q\|$  is typically decreasing at a rate determined by  $d$  as  $t$  grows large.

Because it was established above that the similarity between  $p[t]$  and  $q[t]$  is measurable by  $D_{K-L}(p,q)$ , and the central theme of this paper is that adapting  $p[t]$  to estimated  $q[t]$  speeds-up convergence, it is therefore desirable to show, analytically, that  $D_{K-L}(p,q)$  is predictive of MBH

convergence for a given  $f$ . That analysis is beyond the scope of this paper. Instead, the comments in the next two paragraphs are offered.

Improved search “efficiency” and accelerated MBH convergence times thereby, can be measured and analyzed by First Passage Time Densities (FPTDs), which are pdfs of First Passage Times (FPTs), where an FPT is defined as the first arrival at  $(x^*, f^*)$  of a search probe, i.e. hopper, starting from a randomly initialized starting location initial point  $(x[0], f[0])$ . In addition, in the case of challenging  $f$ , improved “efficiency” and the acceleration of convergence times thereby, enables the solution to be found reliably, whereas otherwise the solution may not be found at all within practical amounts of time. For most  $f$ ,  $D_{K-L}(p, q)$  is highly predictive of MBH FPTDs and therefore MBH convergence times.

FPTDs display near-optimal search “efficiency” when  $\mu/\sigma$  is nearly maximized, where  $\mu$  is the mean and  $\sigma^2$  is the variance. Because of the typically exponential decay in  $g[d]$ , implying typically exponential FPTs to  $(x^*, f^*)$ , MBH FPTDs can be modeled as Gamma distributions  $\Gamma(\alpha, \beta)$ . When MBH FPTDs are modelled as a Gamma distribution, the shape parameter  $\alpha$  and the inverse scale parameter  $\beta$  can be used characterize improvements in the speed and variability of MBH convergence-times.

$D_{K-L}(p, q)$  and FPTDs are used for analytical purposes only. In practice, the estimation of  $q[t]$ , and the adaptation of  $p[t]$  to  $q[t]$ , so that  $p[t]$  is “driven” to become “similar” to  $q[t]$ , are implemented using signal processing methods that support rapid estimation of statistical parameters within a single on-going MBH. The similarity between  $p[t]$  and  $q[t]$  is visually apparent in Figures 8 and 9. Further, the apparent similarity between  $p[t]$  and  $q[t]$ , correlates with the empirical performance data in Table 1.

The extent to which, for some  $f$ , there are properties of the  $\Delta x_p$  that affect MBH convergence but are not detectable in  $q[t]$ , is a subject of on-going investigation. Such properties of the  $\Delta x_p$  may be detectable in time-varying autocorrelations or Wavelet Transforms but be “washed out” in the process of constructing or estimating  $q[t]$ . The hypothesis of the current work is that if such properties of the  $\Delta x_p$  exist, they are far less significant than properties of the  $\Delta x_p$  that are detectable in  $q[t]$ , and that, in any, case they are largely accounted for in the process of adapting  $p[t]$  to  $q[t]$ . The empirical data shown in Table 1 suggests that, for the application described in this paper, effectiveness of adapting  $p[t]$  to  $q[t]$  is already quite high.

To summarize: In large part, search “efficiency” is primarily determined by how “well-suited”  $p$  is to  $f$ , which is a function of how “similar” to  $p$  is to  $q[t]$ . The similarity between  $p$  and  $q[t]$ , as measured by  $D_{K-L}(p, q)$  being small, correlates with the shape of MBH FPTDs displaying near-optimal search “efficiency”.

Nothing said in this section or in the **APPENDIX** is limited to any dimensionality, along as the number of dimensions is a positive integer.

It was empirically established in Englander and Englander that a fixed  $p$  can be designed so as to make the hop distribution reasonably “well-suited” to a family of  $f$ , thereby improving MBH performance for that family of objective functions.<sup>8</sup> The analytical framework explains that a given objective function  $f$  has a “preferred” hop distribution  $q[t]$ , leading to the hypothesis that equivalence classes  $\{f_i\}$  having similar  $p_i[t]$  exist, which would support the empirical finding that a fixed  $p$  can be designed so as to make the hop distribution reasonably “well-suited” to a family of  $f$ .

However, because  $q[t]$  is time-varying, adapting  $p[t]$  to  $q[t]$ , in near real-time within a single on-going MBH, is more effective than a well-designed fixed  $p$ . The adaptation of  $p[t]$  to  $q[t]$  is described in the next section.

### MBH WITH ADAPTIVE HOP DISTRIBUTION

Implementations of the framework involve the Monte Carlo estimation of the distribution  $q[t]$  that is characteristic of a given objective function  $f$ , and an understanding of how, if  $q[t]$  could be used to drive the MBH, the result would be very efficient, fast, accurate searches. However, MBH hops are draws from distribution  $p$ . Therefore, efficient, fast, accurate searches can only be achieved by choosing or adapting  $p$  so that it is highly similar to  $q$ . It has been explained above that the Kullback-Leibler divergence  $D_{K-L}(p, q)$  measures the similarity between distributions  $p$  and  $q$ . It then follows that the K-L distance between  $q$  and  $p$  is an indicator of search efficiency and speed. But  $D_{K-L}(p, q)$  is not a real-time process, so, while it is an essential analytical tool, it cannot be used adapt  $p[t]$  to  $q[t]$  in near real-time, within a single on-going MBH.

There are various ways to adapt  $p[t]$  to  $q[t]$  in near real-time, within a single on-going MBH, but all of them involve the near real-time estimation of the scale parameter of  $q[t]$ , and the near real-time adaptation of the scale parameter of  $p[t]$  to it. The method for doing that used in the implantation used for this paper is specified in the **PSEUDO-CODE** section. There,  $p[\lambda[t]]$  is the currently adapted  $p$  with currently adapted scale parameter  $\lambda[t]$ . It can be seen that:

$$\lambda[t] = ((\max(f[\mathbf{x}[:, t]])/(\max(f[\mathbf{x}[:, t-1]])) + \varepsilon) \quad (7)$$

where  $:$  is across the multiple hoppers and  $\varepsilon$  is a small fixed positive number (e.g. 0.1) that prevents  $\lambda[t]$  from decaying to 0.

$\lambda[t]$  is the current estimate of the normalized scale parameter of  $q[t]$ , varying between being initially 1 when  $\max(f[\mathbf{x}[:, t]])$  is large, eventually approaching 0 as  $\max(f[\mathbf{x}[:, t]])$  approaches 0. Multiple simultaneous hoppers are used only to improve the estimate of  $\lambda[t]$ .  $\max(f[\mathbf{x}[:, t]])$  is used to assure that the scale parameter of  $q[t]$  is not estimated to be near-zero until the  $m^{\text{th}}$  hopper having the largest value  $f[\mathbf{x}[m, t]]$  is evaluating  $f[\mathbf{x}[m, t]]$  as being near-zero.

That algorithm for  $\lambda[t]$  works in the Pioneer 11 problem because  $\max(f[\mathbf{x}[:, t]])$  can be expected to eventually approach 0. In more general settings in which  $f[\mathbf{x}[:, t]]$  may not decay to near-zero, the range  $0 \leq \lambda[t] \leq 1$  needs to be established by other normalizations. One effective method is to use a lightly low-pass filtered estimate of  $\text{var}(f[\mathbf{x}[:, t]])/\text{var}(f[\mathbf{x}[:, 0]])$  where  $\text{var}(\cdot)$  is the variance across  $(\cdot)$ . Another effective method is to use a lightly low-pass filtered estimate  $\text{average}(\text{abs}(f[\mathbf{x}[:, t]]))/\text{average}(\text{abs}(f[\mathbf{x}[:, 0]]))$ . What matters is to estimate of the normalized scale parameter of  $q[t]$  in the form  $0 \leq \lambda[t] \leq 1$ . There are many ways to achieve that.

### APPLICATION TO PIONEER 11

Because a specifically 3-D application is being discussed, vectors are now written in bold font.

The application problem addressed in this paper is the redesign and optimization of historical Pioneer 11, using the PyKep 2-body model.

Pioneer 11 was launched by NASA in April of 1973. Its scientific mission was to study the asteroid belt, as well as solar wind and cosmic rays, around Jupiter and Saturn. Pioneer 11 was the first probe to encounter Saturn, and the second to fly through the asteroid belt and perform a flyby of Jupiter. Pioneer 11 then achieved escape velocity that enabled it to leave the Solar System. Due

to power constraints and the vast distance to the probe, the last routine contact with the spacecraft was on September 30, 1995. The last transmission of scientifically useful data from Pioneer 11 was received on November 24, 1995.\* Here, Pioneer 11's trajectory refers to the path from Earth to Jupiter, and then Jupiter to Saturn.

Pioneer 11 was equipped with a very small tanks for propellant, as can be seen from the replica of it hanging in the Smithsonian Institute Air and Space Museum and in diagrams of it that are available on-line. Therefore, the optimal trajectory is known *a priori* to be one that used very little propellant, which implies a very small total change in velocity,  $\Delta v$ , in meters per second.

The historical mission used 17 pounds of propellant in May of 1974 to re-target it so that it would fly to Saturn after it's flyby of Jupiter.\* The solution provided by the adaptive MBH implementation described in this paper uses near-zero total  $\Delta v$ , implying the use of very little fuel.

The precision of the model used in the current paper is not fine enough to determine whether the adaptive MBH solution uses more or less propellant than the historic mission. However, the significance of the bench-marking data in this paper, tabulated in TABLE 1, is that the problem's objective function has challenging hyper-geometric properties that test the analytical framework upon which this paper is built. Those hyper-geometric challenges of the problem arise from: quasi-periodicities due to the relative positions of the probe, Earth, Jupiter, and Saturn, throughout the mission; singularities introduced by the transcription's use of Lagrange's solution to Lambert's problem, which is undefined at transfer angles of 180 degrees; and penalty functions that the ACT transcription uses to bound those singularities. These are hyper-geometric properties are visible in the small neighborhoods around the solution's  $f^*$  in the heatmaps shown in Figures 2 through 7.

Among the practical benefits of using PyKep is that it is server-based, written in C++, and exposed to Python via an API. The implementation described in this paper was coded in Python and only needed to query the server-based ACT model relatively infrequently (a maximum of 500 times per MBH, for each of 16 simultaneous hoppers, totaling a maximum of 8,000 queries). This enabled an implementation of a 3-dimensional model that has extremely fine granularity (IEEE Floating Point precision) in  $f$  and  $X$ , and to perform each MBH in a matter of seconds on a conventional laptop.

The inputs into PyKep are three decision variables, each of which represents a date in PyKep coordinates. Each of the three decision variables is bounded so as to stay within a range centered around the corresponding dates of the historical Pioneer 11 mission. The calendar coordinates are: Launch from Earth on April 6, 1973, 02:11:00 UTC; fly-by of Jupiter on December 3, 1974; and fly-by Saturn on September 1, 1979.\*

The range on  $x_1$  is ( -9861.25, -9494.50 ); the range on  $x_2$  is ( 0.25, 1000.00 ); and the range on  $x_3$  is ( 0.25, 2000.0 ). The lower bound on  $x_1$  is set by PyKep. It is a negative number because PyKep's date convention uses January 1, 2000 00:00:01 UTC as 0.0. The resolution in each of  $x_1$ ,  $x_2$ , and  $x_3$ , is IEEE Floating Point precision.

Using the method using 16 hoppers that communicate as specified in the **PSEUDO-CODE** section, the range of  $f^* = 0.0003$  through 0.0285, with 10-trial average of 0.0067, was found within fewer than 500 MBH time-steps. A plot of the resulting trajectory for  $f^* = 0.0011$  is provided in Figure 1.

This paper does not evaluate the performance of the adaptive implantation by comparing it to the PyKep  $f^*$  for the historical Pioneer 11 mission. Instead, the PyKep model of Pioneer 11, is used as it was intended by its developers: as a generator of a challenging  $f$ , for the comparison of different optimization methodologies – in this paper, a non-adaptive MBH driven by a non-adaptive

\*[https://en.wikipedia.org/wiki/Pioneer\\_11](https://en.wikipedia.org/wiki/Pioneer_11)

leptokurtic (narrow-bodied, long-tailed)  $p$ ; a non-adaptive Gaussian  $p$ ; and the adaptive  $p$  which is the center-piece of this paper.

When only one hopper is used and  $p$  is a non-adaptive Gaussian distribution  $p$ ,  $p$  resembles the distribution  $p = \mathcal{N}(x[t - 1], \mathbf{I})$  in the example given by Baba *et al.* in the **APPENDIX**. When that is done for 10 MBH trials and statistics are taken, the 500-time-step  $f^*$  ranges from 2.5302 to 242.3353, with an average of 90.8899. The 10,000-time-step  $f^*$  ranges from 0.7494 to 15.2358, with an average of 6.7002.

When only one hopper is used and  $p$  is a non-adaptive leptokurtic (narrow-bodied, long-tailed) pdf that resembles the Bipolar Pareto distribution described in Englander and Englander, the 10-trial, 500-time-step  $f^*$  ranges from 0.5991 to 17.5510, with an average of 4.7793, and the 10,000-time-step estimated  $f^*$  ranges from 0.0242 to 0.8234, with an average of 0.2245.

In the analytical framework, typical convergence times and variability in convergence times, for an MBH operating on a given  $f$  using a given  $p$ , are measured using First Passage Time Density (FPTD) functions. That requires performing thousands of MBHs on the same  $f$ , and an *a priori* knowledge of true  $f^*$ . Therefore, while that is a valuable tool for constructing the analytical framework, it is usually impractical – if not impossible – in practice. Instead, empirical data is used to make the performance comparisons.

The table below summarizes the comparative performance of the adaptive, non-adaptive leptokurtic, and non-adaptive Gaussian implementations.

| p                                      | Number of<br>hoppers | Number of<br>hops | Number of<br>PyKep<br>queries | min( $f^*$ ) | max( $f^*$ ) | average( $f^*$ ) | std.<br>dev( $f^*$ ) |
|----------------------------------------|----------------------|-------------------|-------------------------------|--------------|--------------|------------------|----------------------|
| Adaptive                               | 16                   | 500               | 8,000                         | 0.0003       | 0.0250       | 0.0067           | 0.0082               |
| Fixed<br>Laplace-like<br>(leptokurtic) | 1                    | 500               | 500                           | 0.5991       | 17.5510      | 4.3793           | 5.0067               |
|                                        | 1                    | 10,000            | 10,000                        | 0.0242       | 0.8234       | 0.2245           | 0.2541               |
| Fixed<br>Gaussian                      | 1                    | 500               | 500                           | 2.5302       | 242.3553     | 90.8899          | 83.0177              |
|                                        | 1                    | 10,000            | 10,000                        | 0.7494       | 15.2358      | 6.7022           | 5.0392               |

**TABLE 1: A comparison of  $f^*$  for various  $p$  using the PyKep model for Pioneer 11**

## CONVERGENCE OF MBH

The adaptation, to MBH, of an asymptotic convergence proof for global optimization using directed random search, by Baba *et al.* (1977), is provided in the **APPENDIX**.

That MBH convergence can be sped-up dramatically, why, and how, is the central theme of the analytic framework described in this paper.

Using the analytical framework, one can show that the K-L distance between  $q$  and  $p$  is an indicator of search efficiency and speed, and that that FPTDs can be used to characterize MBH search efficiency and speed, and that therefore K-L distance between  $q$  and  $p$  can be used to characterize and even predict MBH FPTDs. The search efficiency and speed,  $D_{K-L}(p, q)$  and the FPTDs all tell the same story from different perspectives.

Further, as illustrated in Figures 8 and 9, as the adaptation of  $p[t]$  to  $q[t]$  drives  $p[t]$  to become similar to  $q[t]$ . Intuitively matching Figures 8 and 9 with Table 1 implies that adaptation of  $p[t]$  to  $q[t]$ , and the resulting similarity of  $p[t]$  and  $q[t]$ , is the source of the MBH performance improvement. While a presentation of more thorough empirical evidence that attributes the MBH performance improvements to the similarity between  $p[t]$  and  $q[t]$  would require many more plots than can be shown here, that attribution follows directly from the analytical framework.

## SUMMARY

This paper has summarized an analytical framework for explain why an MBH adaptive hop length distribution  $p[t]$  dramatically speeds discovery of the optimal solution. An application of the framework to the PyKep 2-body model of the historical Pioneer 11 mission provides empirical support for that framework.

Preliminary indications are that adapting hop distribution  $p[t]$  in a manner that drives it to become similar to a given objective function's preferred distribution hop  $q[t]$  results in an efficient, fast, and accurate search. Consistent with Englander and Englander, well-chosen hop distributions  $p$  can also improve the efficiency, speed, and accuracy of MBH, convergence, but not as much as adaptive hop length distributions, for reasons explained by the summarized analytical framework.

**PSEUDO-CODE**


---

**Algorithm 1:** Stochastic Monotonic Basin Hopping with adaption of  $p[t]$  to  $q[t]$ 


---

```

if t = 0:
    # t is the MBH time-step
    for m in range (1, Num_hoppers):
        Randomly initialize of  $\mathbf{x}[m,0]$  and evaluate  $f[\mathbf{x}[m,0]]$ 
        # Num_hoppers = 16;  $\mathbf{x}[m, t]$  is an N-dimensional vector;  $f[\mathbf{x}[m, t]]$  is a scalar
    for t in range(1, Terminus):
        for m in range (1, Num_hoppers):
            evaluate  $f[\mathbf{x}[m, t-1]]$ 
            for n in range (1, Num_dimensions):
                draw  $\Delta x_{p_{n,m}} \sim p[\lambda[t-1]]$  and generate  $\xi[m, t]_n = (\mathbf{x}[m, t] + \Delta x_p[m, t])_n$ .
                #  $\xi[m, t]_n$  and  $(\mathbf{x}[m, t] + \Delta x_p[m, t])_n$  are the  $n^{\text{th}}$  dimensional component of
                vectors  $\xi[m, t]$  and  $(\mathbf{x}[m, t] + \Delta x_p[m, t])$ , respectively
                #  $\Delta x_{p_{n,m}} = \lambda[t-1] \cdot k_n \cdot \prod_1^3 \delta$  where  $k_n$  is a scale factor applied to each of
                the three dimensions as to their respective ranges, and  $\delta$  is an
                independent draw from  $\mathcal{N}(0,1)$ .  $\xi[m, t]$  is an N-dimensional vector
                comprised of elements  $\xi[m, t]_n$  bounded so that  $\xi[m, t]_n \subset X$ 
            evaluate  $f[\xi[m, t]]$  #  $f[\xi[t]]$  is a scalar
            if  $f[\xi[m, t]] < f[\mathbf{x}[m, t-1]]$ :
                set  $f[\mathbf{x}[m, t]] = f[\xi[m, t]]$  and  $\mathbf{x}[m, t] = \xi[m, t]$ 
            if  $f[\xi[m, t]] \geq f[\mathbf{x}[m, t-1]]$ :
                set  $f[\mathbf{x}[m, t]] = f[\mathbf{x}[m, t-1]]$  and  $\mathbf{x}[m, t] = \mathbf{x}[m, t-1]$ 
        hopperbest = argminm( $f[\mathbf{x}[:, t]]$ ) # : denotes across the M hoppers
        for all m <> hopperbest:
             $\mathbf{x}[m, t] = \mathbf{x}[\text{hopper}_{\text{best}}, t]$ 
             $f[\mathbf{x}[m, t]] = f[\mathbf{x}[\text{hopper}_{\text{best}}, t]]$ 
             $\lambda[t] = ((\max(f[\mathbf{x}[:, t]])) / (\max(f[\mathbf{x}[:, t-1]]))) + \varepsilon$ 
        # : denotes across the m hoppers and  $\varepsilon$  is a small positive constant that prevents
        the denominator from going to zero if  $\max(f[\mathbf{x}[:, t-1]])$  ever goes to zero
        t = t+1

```

---

## FIGURES

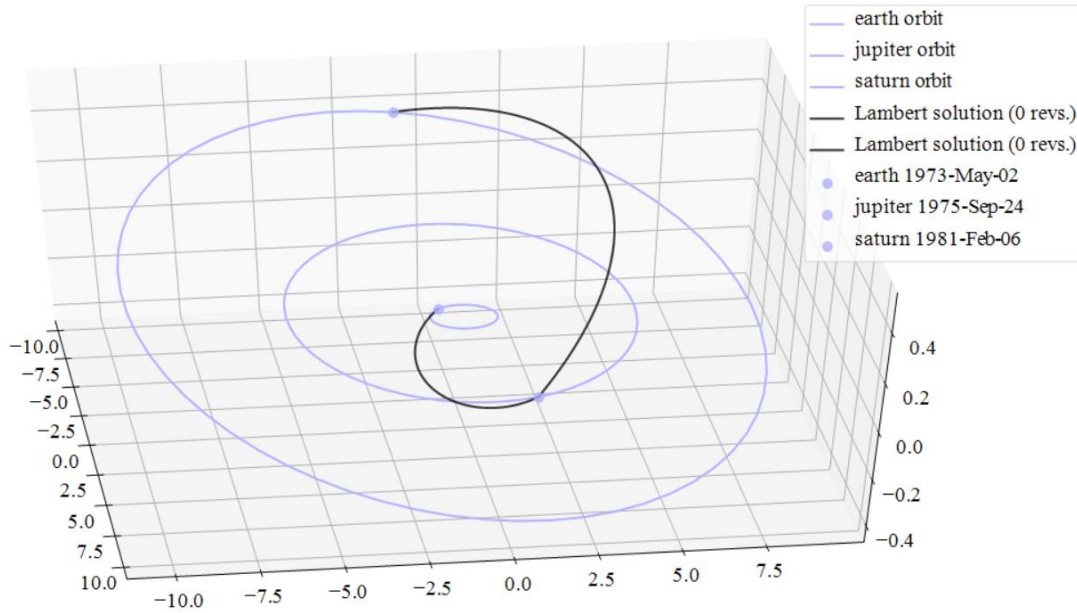


**Figure 1: The  $f^* = 0.003$  Pioneer 11 trajectory, found by the adaptive MBH, rendered in the ESA ACT's PyKep toolkit (the z-axis is magnified 5x)**

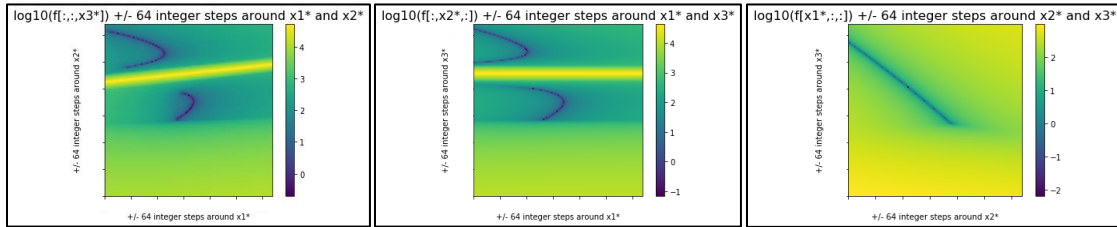


**Figure 2, 3, 4: Heatmaps of  $\text{Log}_{10}(f[x_1, x_2, x_3])$  in the +/- 64 neighborhood 3-D around  $x_n^*$**

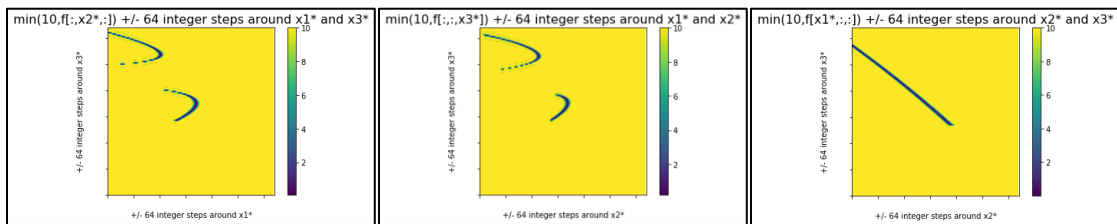


**Figure 5, 6, 7: Heatmaps of  $\min(10, f[x_1, x_2, x_3])$  in the +/- 64 neighborhood in 3-D around  $x_n^*$**

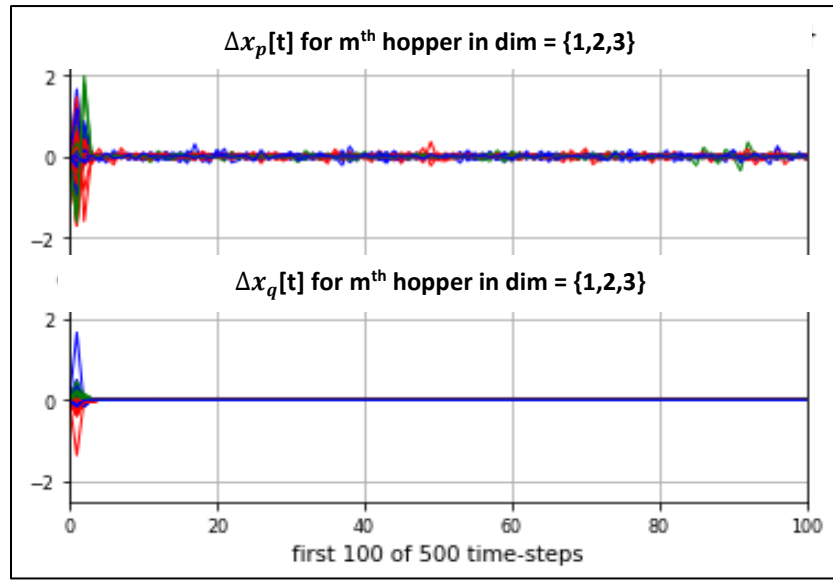


Figure 8:  $\Delta x_{q_n}[t] \sim q[t]$ , and  $\Delta x_{p_n}[t] \sim p[t]$  adapted to  $\Delta x_{q_n}[t] \sim q[t]$   
(Colors red, green, and blue correspond to the three dimensions of  $X$ .)

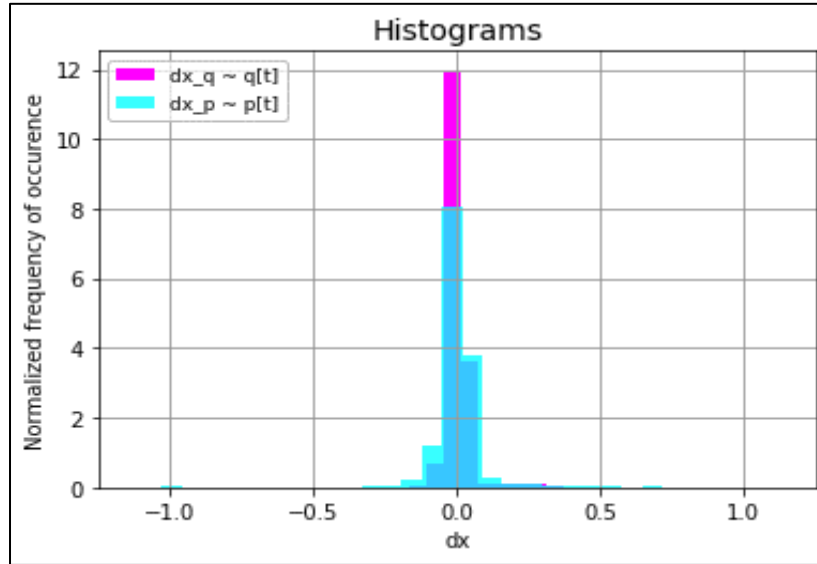


Figure 9: Corresponding histograms of  $\Delta x_{q_n}[t]$ , and  $\Delta x_{p_n}[t] \sim p[t]$  adapted to  $\Delta x_{q_n}[t] \sim q[t]$

## REFERENCES

- <sup>1</sup>Vasile, M., Minisci, E. and Locatelli, M., (2010), Analysis of Some Global Optimization Algorithms for Space Trajectory Design. *Journal of Spacecraft and Rockets*, Volume 47, No. 2, pages 334 - 444
- <sup>2</sup>Yam, C.H., Lorenzo, D. D. and Izzo, D., (2011), Low-thrust trajectory design as a constrained global optimization problem. *Proceedings of the Institution of Mechanical Engineers, Part G: Journal of Aerospace Engineering*, Volume 225, No. 11, pages 1243 - 1251
- <sup>3</sup>Englander J.A., Conway B.A., Williams, T., Automated Mission Planning via Evolutionary Algorithms. *Journal of Guidance, Control, and Dynamics*. 2012;35 (6):1878-1887
- <sup>4</sup>Englander J.A., Conway B.A., An Automated Solution of the Low-Thrust Interplanetary Trajectory Problem. *Journal of Guidance, Control, and Dynamics*. 2017; 40(1):15-27. doi:10.2514/1.G002124
- <sup>5</sup>Wales, D.J. and Doye, J.P.K., Global optimization by basin-hopping and the lowest energy structures of Lennard-Jones clusters containing up to 110 Atoms. *Journal of Physical Chemistry. A*, 101, 5111-5116 (1997).
- <sup>6</sup>Leary, R. H., (2000), Global Optimization on Funneling Landscapes. *Journal of Global Optimization*, Volume 18, No. 4, pages 367 - 383
- <sup>7</sup>Locatelli, M., and Schoen, F., (2003), Efficient Algorithms for Large Scale Optimization: Lennard-Jones Clusters. *Computational Optimization and Applications* 26 (2003): 173-190.
- <sup>8</sup>Englander, J. A. and Englander, A. C., (2014), Tuning Monotonic Basin Hopping: Improving the Efficiency of Stochastic Search as Applied to Low-Thrust Trajectory Optimization, *Proceedings of the 24th International Symposium on Space Flight Dynamics*, Laurel, MD
- <sup>9</sup>Izzo, D., (2019, February 22). esa/pykep: (Version v2.3). Zenodo. <http://doi.org/10.5281/zenodo.2575462>
- <sup>10</sup>Ellison, D. H., Conway, B. A., Englander, J. A. and Ozimek, M. T., (2018), Analytic Gradient Computation for Bounded-Impulse Trajectory Models Using Two-Sided Shooting. *Journal of Guidance, Control, and Dynamics*, Volume 41, No. 7, pages 1449 - 1462
- <sup>11</sup>Baba, N., Showman, T., and Sawaragi, Y., (1977), A Modified Convergence Theorem for Random Optimization Algorithm, *Information Science*, No. 13, (1977)
- <sup>12</sup>Solis, F. J., and Wets, R. J. B (1981), Minimization by Random Search Techniques, *Mathematics of Operations Research*, Vol. 6: 19 – 30 (1981)
- <sup>13</sup>Aghassi, M., *Minimization by Random Search Techniques* by Solis and Wets, and Introduction to Sampling Methods, published slides from a lecture presented in 2003
- <sup>14</sup>Cover, T.M., and Thomas, J.A., *Elements of information theory*. Wiley, New York, 1991
- <sup>15</sup>Jonathon Shlens, Google Research, 2014, arXiv:14042000v1 [cs.IT] 8 Apr 2014

## APPENDIX: ADAPTATION, TO MBH, OF THE ASYMPTOTIC CONVERGENCE PROOF BY BABA, ET AL.

Based on Baba, Shoman, and Sawaragi, restated and used by Solis and Wets, and included in Aghassi, Minimization by Random Search Techniques by Solis and Wets, and Introduction to Sampling Methods, published slides from a lecture presented in 2003: <sup>11, 12, 13</sup>

### Conceptual Algorithm:

Here, all notation has been revised to align with the notation used in the analytical frame work presented above, including replacing  $x^0$  with  $x[0]$ ,  $\xi^k$  with  $\xi[t]$ ,  $x^k$  with  $x[t]$ ,  $S$  with  $X$ , and distribution  $\mu^k$  with  $p[t]$ . In the present work,  $t$  is the discretized index of successive MBH operations, the same as  $k$  is in the asymptotic convergence proofs by Baba, Shoman, and Sawaragi, and Solis and Wets.  $X$  replaces  $S$ , not only to align the notation in the body of this paper analytic support provided above, but because MBH is typical thought of as operating on an  $f$  having a domain that is hyper-cube rather than a hyper-sphere. No changes in meaning or loss of generality resulted from these changes in notation. The proof is general with respect to any positive integer number of dimensions including 1-D, so here, so all variables that are not matrices are represented as scalars.

Find  $x[0] \in X$ . Set  $t := 0$

1. Generate  $\xi[t] \in \mathbb{R}^n$  (random) from distribution  $p[t]$ . More precisely, in order to adapt this asymptotic convergence proof to MBH, generate  $\xi[t] = (x[t-1] + \Delta x_p[t])$ ,  $\Delta x_p[t]$  random,  $\Delta x_p[t] \sim p[t]$ ,  $\xi[t] \in X$ ,  $f[\xi[t]] \in \mathbb{R}^n$
2. Set  $x[t] = D(x[t-1], \xi[t])$ . In order to adapt this asymptotic convergence proof to MBH, define  $D$  as choose  $\xi[t]$  if  $f[\xi[t]] < f[x[t-1]]$ , otherwise choose  $x[t-1]$
3. Go to step 1

This provides both local search and global search.

Local search  $\Rightarrow \text{supp}(p[t])$  is bounded and  $v(X \cap \text{supp}(p[t])) < v(X)$ , where  $\text{supp}$  denotes support and  $v$  is a measure in the measure-theoretic sense

Global search  $\Rightarrow \text{supp}(p[t])$  is such that  $v(X \cap \text{supp}(p[t])) < v(X)$

### Sufficient conditions for Convergence:

1.  $D$  s.t.  $\{f[x[t]]\}$  is non-increasing. This is satisfied by the MBH algorithm.
  - a.  $f[D(x, \xi)] \leq f[x[t-1]]$
  - b.  $\xi[t] \in X \Rightarrow f[D(x, \xi)] \leq \min\{f[\xi[t]], f[x[t-1]]\}$
2. Zero probability of repeatedly missing any positive volume subset of  $X$ 
  - a. Stated formally,  $\forall A \subseteq X$ , where  $A$  is some sub-space (or the entirety) of  $X$ , s.t.  $v(A) > 0$ ,  $\prod_{t=0}^{\infty} (1 - p[t](A)) = 0$
  - b. In other words, the sampling strategy based on  $p[t]$  cannot consistently ignore any part of  $X$  with positive volume

Example given by Baba *et al.*:

$$D = \begin{cases} \xi[t], \xi[t] \in X, f[\xi[t]] < f[x[t-1]] \\ x[t-1], \text{otherwise} \end{cases}$$

$$p = \mathcal{N}(x[t-1], \mathbf{I})$$

Sufficient Condition 1,  $D$  s.t.  $\{f[x[t]]\}_{t=0}^{\infty}$  is non-increasing, is satisfied by construction.

Sufficient Condition 2, Zero probability of repeatedly missing any positive volume subset of  $X$ , is satisfied because  $X$  is contained in the support of  $\mathcal{N}(x[t-1], \mathbf{I})$

This is MBH when  $p$  is Gaussian and the domain of  $f$  is the unit variance hyper-sphere  $S$ .

The asymptotic convergence (occurring at some unspecified time prior to  $t = \text{infinity}$ ) is proven.

Interestingly, the central theme (and novelty) of the present work shows, analytically and by empirical evidence, that  $p = \mathcal{N}(x[t-1], \mathbf{I})$  results in sub-optimal convergence times that are much slower than convergence times that are achieved when  $p$  is “well-suited” to  $f$ , meaning when  $p$  is similar to  $q$ , as defined earlier. Empirical support for the sub-optimality of Gaussian is evident in Table 1.