RANDOM SEARCH UNDER ADDITIVE NOISE

Luc Devroye
School of Computer Science
McGill University
Montreal, Canada H3A 2K6

and

Adam Krzyżak
Department of Computer Science
Concordia University
Montreal, Canada H3G 1M8

§1. Sid's contributions to noisy optimization

From the early days in his career, Sid Yakowitz showed interest in noisy function optimization. He realized the universality of random search as an optimization paradigm, and was particularly interested in the minimization of functions Q without making assumptions on the form of Q. Especially the noisy optimization problem appealed to him, as exact computations of Q come often at a tremendous cost, while rough or noisy evaluations are computationally cheaper. His early contributions were with Fisher (Fisher and Yakowitz, 1976; Yakowitz and Fisher, 1973). The present paper builds on these fundamental papers and provides further results along the same lines. It is also intended to situate Sid's contributions in the growing random search literature.

Always motivated by the balance between accurate estimation or optimization and efficient computations, Sid then turned to so-called bandit problems, in which noisy optimization must be performed within a given total computational effort (Yakowitz and Lowe, 1991).

The computational aspects of optimization brought him closer to learning and his work there included studies of game playing strategies (Yakowitz, 1989; Yakowitz and Kollier, 1992), epidemiology (Yakowitz, 1992; Yakowitz, Hayes and Gani, 1992) and communication theory (Yakowitz and Vesterdahl, 1993). Sid formulated machine learning invariably as a noisy optimization problem, both over finite and infinite sample spaces: Yakowitz (1992), Yakowitz and Lugosi (1990), and Yakowitz, Jayawardena and Li (1992) summarize his main views and results in this respect.

Another thread he liked to follow was stochastic approximation, and in particular the Kiefer-Wolfowitz method (1952) for the local optimization in the presence of noise. In a couple of technical reports in 1989 and in his 1993 SIAM paper, Sid presented globally convergent extensions of this method by combining ideas of random search and stochastic approximation.

We have learned from his insights and shared his passion for nonparametric estimation, machine learning and algorithmic statistics. Thank you, Sid.

§2. Formulation of search problem

We wish to locate the global minimum of a real-valued function Q on some search domain $\mathcal{X}$, a subset of R^d. As we pose it, this problem may have no solution. First of all, the function Q may not have a minimum on $\mathcal{X}$ (consider $\mathcal{X} = (0,1)$ and $Q(x) = x$), and if a minimum exists, it may not be unique (consider the real line and $Q(x) = \sin(x)$), and if it exists and is unique, it may be nearly impossible to find it exactly, although we can hope to approximate it in some sense. But is even that possible? Take for example the function on R^d defined by $||x||^2$ everywhere except on a finite set A on which the function takes the value -1. Without a priori information about the location of the points of A, it is impossible to locate any point in A, and thus to find a global minimum. To get around this, we take a probabilistic view. Assume that we can probe our space with the help of a probability distribution μ such as the uniform density on $[0,1]^d$ or the standard normal distribution on R^d. If X is a random variable with probability distribution μ, we can define the global minimum by the essential infimum:

$$q_\mu = \operatorname{ess\,inf}\ Q(X) .$$

This means that $\mathbb{P}(Q(X) < q_\mu) = 0$ and $\mathbb{P}(Q(X) < q_\mu + \epsilon) > 0$ for all $\epsilon > 0$. The value of q_μ depends heavily on μ. It is the smallest possible value that we can hope to reach in a search process if the search is carried out at successive independent points $X_1, \ldots, X_n, \ldots$ with common probability distribution μ. For example, if μ is the uniform distribution on R^d, then q_μ is the (Lebesgue) essential infimum of Q in the unit cube. To push the formalism to an extreme, we could say that a couple (Q, μ) defines a search problem if

(i) Q is a Borel measurable function on R^d.
(ii) μ is a probability measure on the Borel sets of R^d.
(iii) $q_\mu > -\infty$.

Formally, a search algorithm is a sequence of mappings g_{n+1} from $\mathcal{X}^n$ to $\mathcal{X}$, $n \geq 0$, where $X_{n+1} = g_{n+1}(X_1, \ldots, X_n)$ is a place at which $Q(X_{n+1})$ is computed or evaluated. The objective is have $\min(Q(X_1), \ldots, Q(X_n))$ tend to q_μ, and if possible, to assure that the rate of this convergence is fast. In random search methods, the mapping g_{n+1} is replaced by a distribution on $\mathcal{X}$ that is a function of $X_1, \ldots, X_n$, and X_{n+1} is a random variable drawn from that distribution. The objective remains the same. Noisy optimization problems will be formally defined further on in the paper.

§3. Random search: a brief overview

Random search methods are powerful optimization techniques. They include pure random search, adaptive random search, simulated annealing, genetic algorithms, neural networks, evolution strategies, nonparametric estimation methods, bandit problems, simulation optimization, clustering methods, probabilistic automata and random restart. The ability of random search methods to locate the global extremum made them indispensable tool in many areas of science and engineering. The explosive growth of random search is partially documented in books such as those by Aarts and Korst (1989), Ackley (1987), Ermoliev and Wets (1988), Goldberg (1989), Holland (1992), Pintér (1996), Schwefel (1977, 1981), Törn and Žilinskas (1989), Van Laarhoven and Aarts (1987), Wasan (1969) and Zhigljavsky (1991).

Random search algorithms are usually easy and inexpensive to implement. Since they either ignore the past or use a small collection of points from iteration to iteration they are easily parallelizable. Convergence of most random search procedures is not affected by the cost function, in particular its smoothness or multimodality. In a minimax sense, random search is more powerful than deterministic search: this means it is nearly the best method in the worst possible situation (discontinuities, high dimensionality, multimodality) but possibly the worst method in the best situation (smoothness, continuity, unimodality) (Jarvis, 1975). The simplest random search method the pure random search can be used to select a starting point for more sophisticated random search techniques and also can act as a benchmark against which the performance of other search algorithms are measured. Also, random search is much less sensitive than deterministic search to function evaluations perturbed by the additive noise and that motivates the present paper.

In ordinary random search, we denote by X_n^* the best estimate of the (global) minimum after n iterations, and by X_n a random probe point. In pure random search, $X_1, \ldots, X_n$ are i.i.d. with a given fixed probability measure over the parameter space $\mathcal{X}$. The simple ordinary random search algorithm is given below:

$$X_{n+1}^* = \begin{cases} X_{n+1} & \text{if } Q(X_{n+1}) < Q(X_n^*) \\ X_n^* & \text{otherwise.} \end{cases}$$

In local random search on a discrete space, X_n usually is a random neighbor of X_n^*, where the definition of a neighborhood depends upon the application. In local random search in a Euclidean space, one might set

$$X_{n+1} = X_n^* + W_n \ ,$$

where W_n is a random perturbation usually centered at zero. The fundamental properties of pure random search (Brooks, 1958) are well documented. Let $F(u) \stackrel{\text{def}}{=} \mathbb{P}\{Q(X_1) \le u\}$ be the distribution function of $Q(X_1)$. Then $F(Q(X_n^*))$ is approximately distributed as E/n, where E is an exponential random variable. This follows from the fact that if F is nonatomic,

$$\mathbb{P}\{F(Q(X_n^*)) > t/n\} = (1 - t/n)^n \to e^{-t} \ , \ t > 0 \ .$$

Note first of all the distribution-free character of this statement: its universality is both appealing and limiting. We note in passing here that many papers have been written about how one could decide to stop random search at a certain point.

To focus the search somewhat, random covering methods may be considered. For example, Lipschitz functions may be dealt in the following manner (Shubert, 1972): at the trial points X_i, we know Q and can thus derive piecewise linear bounds on Q. The next trial point X_{n+1} is given by

$$X_{n+1} = \arg\min_{x \in \mathcal{X}} \max_{i \le n} \{Q(X_i) - C\|x - X_i\|\}$$

where C is the Lipschitz constant. This is a beautiful approach, whose implementation for large d seems very hard. For noisy problems, or when the dimension is large, a random version of this was proposed in Devroye (1978). If $\mathcal{X}$ is compact, X_{n+1}^* is taken uniformly in $\mathcal{X}$ minus the union of the n balls centered at the X_i's $(1 \le i \le n)$ with radius $(Q(X_i) - Q(X_n^*))/C$. If C is unknown, replace it in the formula for the radius by C_n and let $C_n \to \infty$ such that $C_n^d/n \to 0$ and $(C_{n+1}/C_n)^d = 1 + o(1/n)$ (example: $C_n = \exp((\log n)^p)$ for $p \in (0,1)$). Then $Q(X_n) \to \min Q$ almost surely.

Global random search is a phrase used to denote many methods. Some of these methods proceed in a local manner, yet find a global minimum. Assume for example that we set

$$X_{n+1} = X_n^* + \sigma_n N_{n+1} \ ,$$

where $N_1, N_2, \ldots$ are i.i.d. normal random vectors, and $\sigma_n \to 0$ is a given deterministic sequence. The new probe point is not far from the old best point, as if one is trying to mimic local descent algorithms.

However, over a compact set, global convergence takes place whenever $\sigma_n \sqrt{\log n} \to \infty$. This is merely due to the fact that $N_1, N_2, \ldots, N_n$ form a cloud that becomes dense in the expanding sphere of radius $\sqrt{2 \log n}$. Hence, we will never get stuck in a local minimum. The convergence result does not put any restrictions on Q. The above result, while theoretically pleasing, is of modest value in practice as σ_n must be adapted to the problem at hand. A key paper in this respect is by Matyas (1965), who suggests making σ_n adaptive and setting

$$X_{n+1} = X_n^* + \sigma_n N_{n+1} + D_{n+1} ,$$

where D_{n+1} is a preferred direction that is made adaptive as well. A rule of thumb, that may be found in several publications (see Devroye, 1972, and more recently, Bäck, Hoffmeister and Schwefel, 1991), is that σ_n should increase after a successful step, and decrease after a failure, and that the parameters should be adjusted to keep the probability of success around $1/5$. Schumer and Steiglitz (1968) and others investigate the optimality of similar strategies for local hill-climbing. Alternately, σ_n may be found by a one-dimensional search along the direction given by N_{n+1} (Bremermann, 1968; Gaviano, 1975).

In simulated annealing, one works with random probes as in random search, but instead of letting X_{n+1}^* be the best of X_{n+1} (the probe point) and X_n^* (the old best point), a randomized decision is introduced, that may be reformulated as follows (after Hajek and Sasaki, 1989):

$$X_{n+1}^* = \begin{cases} X_{n+1} & \text{if } Q(X_{n+1}) - Q(X_n^*) \le t_n E_n \\ X_n^* & \text{otherwise.} \end{cases}$$

where t_n is a positive constant depending for now on n only and $E_1, E_2, \ldots$ is an i.i.d. sequence of positive random variables. The best point thus walks around the space at random. If t_n, the temperature, is zero, we obtain ordinary random search. If $t_n = \infty$, $X_1^*, X_2^*, \ldots$ is a random walk over the parameter space. If $t_n > 0$ and E_n is exponentially distributed, then we obtain the Metropolis Markov chain or the Metropolis algorithm (Metropolis et al, 1953; Kirkpatrick, Gelatt and Vecchi, 1983; Meerkov, 1972; Cerny, 1985; Hajek and Sasaki, 1989). Yet another version of simulated annealing has emerged, called the heat bath Markov chain (Geman and Hwang, 1986; Aluffi-Pentini et al, 1985), which proceeds by setting

$$X_{n+1}^* = \begin{cases} X_{n+1} & \text{if } Q(X_{n+1}) + t_n Y_n \le Q(X_n^*) + t_n Z_n \\ X_n^* & \text{otherwise,} \end{cases}$$

where now $Y_1, Z_1, Y_2, Z_2, \ldots$ are i.i.d. random variables and t_n is the temperature parameter. If the Y_i's are distributed as the extreme-value distribution (with distribution function $\exp(e^{-x})$) then we obtain the original version of the heat bath Markov chain. Note that each Y_i is then distributed as $\log\log(1/U)$ where U is uniform $[0,1]$, so that computer simulation is not hampered.

The two schemes are not dramatically different. The heat bath Markov chain as we presented it here has the feature that function evaluations are intentionally corrupted by noise. This clearly reduces the information content and must slow down the algorithm. Most random search algorithms take random steps but do not add noise to measurements; in simulated annealing, one deliberately destroys valuable information. It should be possible to formulate an algorithm that does not corrupt expensive function evaluations with noise (by storing them) and outperforms the simulated annealing algorithm in some sense. One should be careful though and only compare algorithms that occupy equal amounts of storage for the program and the data.

We now turn to the choice of t_n. In view of the representation given above, it is clear that $\mathrm{E}\{Q(X_n^*) - \min Q\}$ is bounded from below by a constant times t_n as t_n is the threshold we allow in steps away from the minimum. Hence the need to make t_n small. This need clashes with the condition of convergence (typically, t_n must be at least $c/\log n$ for some constant $c > 0$). The condition of convergence depends upon the setting (the space $\mathcal{X}$ and the definition of X_{n+1} given X_n^*). We briefly deal with the specific case of finite-domain simulated annealing below. In continuous spaces, progress has been made by Vanderbilt and Louie (1984), Dekkers and Aarts (1991), Bohachevsky, Johnson and Stein (1986), Gelfand and Mitter (1991), and Haario and Saksman (1991). Other key references on simulated annealing include Aarts and Korst (1989), Van Laarhoven and Aarts (1987), Anily and Federgruen (1987), Gidas (1985), Hajek (1988), and Johnson, Aragon, McGeoch and Schevon (1989).

Further work seems required on an information-theoretic proof of the inadmissibility of simulated annealing and on a unified treatment of multistart and simulated annealing, where multistart is a random search procedure in which one starts at a randomly selected place at given times or whenever one is stuck in a local minimum.

On a finite connected graph, simulated annealing proceeds by picking a trial point uniformly at random from its neighbors. Assume the graph is regular, i.e., each node has an equal number of neighbors. If we keep the temperature $t > 0$ fixed, then there is a limiting distribution for X_n^*, called the Gibbs distribution or Maxwell-Boltzmann distribution: for the Metropolis algorithm, the asymptotic probability of node i is proportional to $e^{-Q(i)/t}$. Interestingly, this is independent of the structure of the graph. If we now let $t_n \to 0$ then with probability tending to one, X_n^* belongs to the collection of local minima. With probability tending to one, X_n^* belongs to the set of global minima if additionally, $\sum_n e^{-\Delta/t_n} = \infty$ (for example, $t_n = c/\log(n+1)$ for $c \geq \Delta$ will do). Here Δ is the maximum of all depths of strictly local minima (Hajek, 1988). The only condition on the graph is that all connected components of $\{x : Q(x) \leq c\}$ are strongly connected for any c. The slow convergence of t_n puts a severe lower bound on the convergence rate of simulated annealing.

Let us consider optimization on a compact of $\mathbf{R}^d$, and let Q be bounded there. If we let $X_{n+1} - X_n^*$ have a fixed density f that is bounded from below by a constant times the indicator of the unit ball, then X_n^* in the Metropolis algorithm converges to the global minimum in probability if $t_n \downarrow 0$, yet $t_n \log n \to \infty$. Bohachevsky, Johnson and Stein (1986) adjust t_n during the search to make the probability of accepting a trial point hover near a constant. Nevertheless, if t_n is taken as above, the rate of convergence to the minimum is bounded from below by $1/\log n$, which is much slower than the polynomial rate we would have if Q were multimodal but Lipschitz.

Several ideas deserve more attention as they lead to potentially efficient algorithms. These are listed here in arbitrary order. In 1975, Jarvis introduced <u>competing searches</u> such as competing local random searches. If N is the number of such searches, a trial (or time unit) is spent on the i-th search with probability p_i, where p_i is adapted as time evolves; a possible formula is to replace p_i by $\alpha p_i + (1-\alpha)(c/Q(X_i))^b$, where $\alpha \in (0,1)$ is a weight, c and b are constants, and X_i is the trial point for the i-th competing search. More energy is spent on promising searches.

This idea was pushed further by several researchers in one form or another. Several groups realized that when two searches converge to the same local minimum, many function evaluations could be wasted. Hence the need for <u>on-line clustering</u>, the detection of points that belong somehow to the same local valley of the function. See Becker and Lago (1970), Törn (1974, 1976), de Biase and Frontini (1978), Boender et al (1982), and Rinnooy Kan and Timmer (1984, 1987).

The picture is now becoming clearer—it pays to keep track of several base points, i.e., to increase the storage. In Price's <u>controlled random search</u> for example (Price, 1983), one has a cloud of points of size about $25d$, where d is the dimension of the space. A random simplex is drawn from these points, and the worst point of this simplex is replaced by a trial point, if this trial point is better. The trial point is picked at random inside the simplex.

Independently, the German school developed the Evolutionsstrategie (Rechenberg, 1973; Schwefel, 1981). Here a population of base points gives rise to a population of trial points. Of the group of trial points, we keep the best N, and repeat the process.

Bilbro and Snyder (1991) propose tree annealing: all trial points are stored in tree format, with randomly picked leaves spawning two children. The leaf probabilities are determined as products of edge probabilities on the path to the root, and the tree represents the classical k-d tree partition of the space. Their approach is at the same time computationally efficient and fast.

To deal with high-dimensional spaces, the coordinate projection method of Zakharov (1969) and Hartman (1973) deserves some attention. Picture the space as being partitioned by a $N \times \cdots \times N$ regular grid. With each marginal interval of each coordinate we associate a weight proportional to the likelihood that the global minimum is in that interval. A cell is grabbed at random in the grid according to these (product) probabilities, and the marginal weights are updated. While this method is not fool-proof, it attempts at least to organize global search effort in some logical way.

Consider a population of points, called a generation. By selecting good points, modifying or mutating good points, and combining two or more good points, one may generate a new generation, which, hopefully, is an improvement over the parent generation. Iterating this process leads to the evolutionary search method (Bremermann, 1962, 1968; Rechenberg, 1973; Schwefel, 1977; Jarvis, 1975) and a body of methods called genetic algorithms (Holland, 1975). Mutations may be visualized as little perturbations by noise vectors in a continuous space. However, if $\mathcal{X}$ is the space $\{0,1\}^d$, then mutations become bit flips, and combinations of points are obtained by merging bit strings in some way. The term cross-over is often used. In optimization on graphs, mutations correspond to picking a random neighbor. The selection of good points may be extinctive or preserving, elitist or non-elitist. It may be proportional or based on ranks. As well, it may be adaptive and allow for immigration (new individuals). In some cases, parents never die and live in all subsequent generations. The population size may be stable or explosive. Intricate algorithms include parameters of the algorithm itself as part of the genetic structure. Convergence is driven by mutation and can be proved under conditions not unlike those of standard random search. Evolution strategies aim to mimic true biological evolution. In this respect, the early work of Bremermann (1962) makes for fascinating reading. Ackley's thesis (1987) provides some practical implementations. In a continuous space, the method of generations as designed by Ermakov and Zhigljavsky (1983) lets the population size change over time. To form a new generation, parents are picked with probability proportional to

$$\frac{Q^k(X_i)}{\sum_j Q^k(X_j)} ,$$

and random perturbation vectors are added to each individual, where k is to be specified. The latter are distributed as $\sigma_n Z_n$, where the Z_n's are i.i.d. and σ_n is a time-dependent scale factor. This tends to maximize Q if we let k tend to infinity at a certain rate. For more recent references, see Goldberg (1989), Schwefel (1995) or Banzhaf, Nordin and Keller (1998).

§4. Noisy optimization by random search: a brief survey

Here is a rather general optimization problem: for each point $x \in \mathcal{X}$, we can observe a random process $Y_1, \ldots, Y_n, \ldots$ with $Y_n \to Q(x)$ almost surely, where Q is the function to be minimized. We refer to this as the noisy optimization model. For example, at x, we can observe independent copies of $Q(x) + \xi$, where ξ is measurement noise satisfying $\mathbf{E}\xi = 0$ and $\mathbf{E}|\xi| < \infty$. Averaging these observations naturally leads to a sequence Y_n with the given convergence property. In simulation optimization, Y_n may represent a simulation run for a system parametrized by x. It is necessary to take n large for accuracy, but taking n too large would be wasteful for optimization. Beautiful compromises are awaiting the analyst. Finally, in some cases, $Q(x)$ is known to be the expected value or an integral, as in $Q(x) = \int_A q(x,t)\,dt$ or $Q(x) = \mathbf{E}\{q(x,T)\}$ where A is a fixed set and T is a given random variable. In both cases, Y_n may represent a certain Monte Carlo estimate of $Q(x)$, which may be made as accurate as desired by taking n large enough.

By additive noise, we mean that each $Q(x)$ is corrupted by an independent realization of a random variable Z, so that we can only observe $Q(x) + Z$. The first question to ask is whether ordinary random search is still convergent. Formally, if $Z_1, Z_2, \ldots$ are independent realizations of Z, the algorithm generates trials $X_1, X_2, \ldots$, and at X_i observes $Y_i = Q(X_i) + Z_i$. Then X_n^* is defined as the trial point among $X_1, \ldots, X_n$ with the lowest value Y_i. Assume that with probability at least $\alpha > 0$, X_n is sampled according to a fixed distribution with support on $\mathcal{X}$. Even though the decisions are arbitrary, as in simulated annealing, and even though there is no converging temperature factor, the above algorithm may be convergent in some cases, i.e., $Q(X_n^*) \to \inf Q$ in probability. For stable noise, i.e., noise with distribution function G satisfying

$$\lim_{x \downarrow -\infty} \frac{G(x-\epsilon)}{G(x)} = 0, \text{ all } \epsilon > 0 ,$$

such as normally distributed noise, or indeed, any noise with tails that decrease faster to zero than exponential, then we have convergence in the given sense. The reader should not confuse our notion of stability which is taken from the order statistics literature (Geffroy, 1958) with that of the stable distribution. Stable noise is interesting because an i.i.d. sequence $\eta_1, \ldots, \eta_n$ drawn from G, satisfies $\min(\eta_1, \ldots, \eta_n) - a_n \to 0$ in probability for some sequence a_n. See for example Rubinstein and Weissman (1979). Additional results are presented in this paper.

In noisy optimization in general, it is possible to observe a sample drawn from distribution F_x at each x, with F_x possibly different for each x. The mean of F_x is $Q(x)$. If there are just two x's, and the probe points selected by us are $X_1, \ldots, X_n$, where each of the X_i's is one of the x's, then the purpose in <u>bandit problems</u> is to minimize

$$A_n = \frac{1}{n} \sum_{i=1}^{n} Q(X_i)$$

in some sense (by, e.g., keeping $\mathbf{E}\{A_n\}$ small). This minimization is with respect to the sequential choice of the X_i's. Obviously, we would like all X_i's to be exactly at the best x, but that is impossible since some sampling of the non-optimal value or values x is necessary. Similarly, we may sometimes wish to minimize

$$B_n = \sum_{i=1}^{n} 1_{[X_i \neq x^*]}$$

where x^* is the global minimum of Q. This is relevant whenever we want to optimize a system on the fly, such as an operational control system or a game-playing program. Strategies have been developed based upon certain parametric assumptions on the F_x's or in a purely nonparametric setting. A distinction is also made between finite horizon and infinite horizon solutions. With a finite number of bandits, if at least one F_x is nondegenerate, then for any algorithm, we must have $\mathbf{E}B_n \geq c \log n$ for some constant $c > 0$ on some optimization problem (Robbins, 1952; Lai and Robbins, 1985).

In the case of bounded noise, Yakowitz and Lowe (1991) devised a play-the-leader strategy in which the trial point X_n is the best point seen thus far (based on averages) unless $n = \lfloor ae^k + b \rfloor$ for some integer k (a and b are fixed positive numbers), at which times X_n is picked at random from all possible choices. This guarantees $\mathbf{E}B_n = O(\log n)$. Thus, the optimum is missed at most $\log n$ times out of n.

Another useful strategy for parametric families F_x was proposed by Lai and Robbins (1985). Here confidence intervals are constructed for all $Q(x)$, $x \in \mathcal{X}$. The x with the smallest lower confidence interval endpoint is sampled. Exact lower bounds were derived by them for this situation. For two normal distributions with means $\mu_1 < \mu_2$ and variances σ_1^2 and σ_2^2, Holland (1973) showed that $\mathbf{E}B_n \geq (2\sigma_1^2/(\mu_2 - \mu_1) + o(1)) \log n$.

Yakowitz and Lugosi (1989) illustrate how one may optimize an evaluation function on-line in the Japanese game of gomoku. Here each F_x represents a Bernoulli distribution and $Q(x)$ is nothing but the probability of winning against a random opponent with parameters x.

In a noisy situation when $\mathcal{X}$ is uncountable, we may minimize Q if we are given infinite storage. More formally, let $X_1, X_2, \ldots$ be trial points, with the only restriction being that at each n, with probability at least α_n, X_n is sampled from a distribution whose support is the whole space $\mathcal{X}$ (such as the normal density, or the uniform density on a compact). The support of a random variable X is the smallest closed set S such that $\mathbf{P}\{X \in S\} = 1$. We also make sure that at least λ_n observations are available for each X_i at time n. If the noise is additive, we may consider the λ_n^2 pairings for all the observations at each of X_i and X_j, recording all values of $W(i,j)$, the number of wins of X_i over X_j, $1 \leq i \leq j \leq n$, where a win occurs when for a pair of observations (Y, Y'), $Y < Y'$. For each X_i, let $Z_i = \min_{j \neq i} W(i,j)$, and define X_n^* as the trial point with maximal Z_i value. If $\lambda_n / \log n \to \infty$, and $\sum \alpha_n = \infty$, then $Q(X_n^*) \to \operatorname{ess\,inf} Q(X)$ almost surely (Devroye, 1977; Fisher and Yakowitz, 1973). Interestingly, there are no conditions whatever on the noise distribution. With averaging instead of a statistic based on ranks, a tail condition on the noise would have been necessary. Details and proofs are provided in this paper. For non-additive noise,

$$\sup_x \mathbf{E} \left\{ e^{t|Y|} | X = x \right\} < \infty$$

for all $0 < t \leq t_0$ (where Y is drawn from F_x) suffices for example when X_n^* is obtained by minimizing the λ_n-averages at the trial points.

Gurin (1966) was the first to explore the idea of averages of repeated measurements. Assume again the α_n condition on the selection of trial points and let $\widehat{Q}$ denote the average of λ_n observations. Then, if $\epsilon_n \geq 0$, Gurin proceeds by setting

$$X_{n+1}^* = \begin{cases} X_{n+1} & \text{if } \widehat{Q}(X_{n+1}) < \widehat{Q}(X_n^*) - \epsilon_n \\ X_n^* & \text{otherwise.} \end{cases}$$

This is contrary to all principles of simulated annealing, as we are gingerly accepting new best points by virtue of the threshold ϵ_n. Devroye (1976) has obtained some sufficient conditions for the strong convergence of $Q(X_n^*) \to \operatorname{ess\,inf} Q(X)$. One set includes $\epsilon_n \equiv 0$, $\sup_x \mathbf{V}\{Y | X = x\} < \infty$, and $\sum 1/\sqrt{\lambda_n} =$

∞ (a very strong condition indeed). If $\epsilon_n > 0$ and for each x, $|Y - Q(x)|$ is stochastically smaller than Z where $\mathbb{E}e^{tZ} < \infty$ for some $t > 0$, then $\epsilon_n \to 0$ and $\lambda_n \epsilon_n^2 / \log n \to 0$ are sufficient as well. In the latter case, the conditions insure that with probability one, we make a finite number of incorrect decisions. Other references along the same lines include Marti (1982), Pintér (1984), Karmanov (1974), Solis and Wets (1978), Koronacki (1976) and Tarasenko (1977).

§5. Optimization and nonparametric estimation

To extract the maximum amount of information from past observations, we might store these observations and construct a nonparametric estimate of the regression function $Q(x) = \mathbb{E}\{Y|X = x\}$, where Y is an observation from F_x. Assume that we have n pairs (X_i, Y_i), $1 \le i \le n$, where a diverging number of X_i's are drawn from a global distribution, and the Y_i's are corresponding noisy observations. Estimate $Q(x)$ by $\widehat{Q}(x)$, which may be obtained by averaging those Y_i's whose X_i is among the k nearest neighbors of x. It should be obvious that if $\|\widehat{Q} - Q\|_\infty \to 0$ almost surely, then $Q(X_n^*) \to \operatorname{ess\,inf} Q(X)$ almost surely if $X_n^* = \arg\min_i \widehat{Q}(X_i)$. To this end, it suffices for example that $k/n \to 0$, $k/\log n \to \infty$, that the noise be uniformly bounded, and that $\mathcal{X}$ be compact. Such nonparametric estimates may also be used to identify local minima.

§6. Noisy optimization: formulation of the problem

We consider a search problem (Q, μ) on a subset B of $\mathbb{R}^d$, where μ is the probability distribution of a generic random variable X that has support on B. Typically, μ is the uniform distribution on B. For every x, it is possible to obtain an i.i.d. sequence $Y_1, \ldots, Y_n, \ldots$ distributed as $Q(x) + \eta$, where η is a random variable ("the noise") with a fixed but unknown distribution. We can, if we wish, demand to see as little or as much of the Y_n sequence as we wish. With this formulation, it is still possible to define a random search procedure such that $Q(X_n^*) \to q_\mu \stackrel{\text{def}}{=} \operatorname{ess\,inf} Q(X)$ almost surely for all search problems (Q, μ) and all distributions of η. Note that we do not even assume that η has a mean. Throughout this paper, F is the distribution function of $Q(X) - q_\mu$. The purpose of this paper is to draw attention to such universally convergent random search algorithms that do not place any conditions on μ and F, just as Sid Yakowitz showed us in 1973 (Yakowitz and Fisher, 1973).

§7. Pure random search

In this section, we analyze the behavior of unaltered pure random search under additive noise. The probe points $X_1, \ldots$ form an i.i.d. sequence drawn from a distribution with probability distribution μ. At each probe point X_n, we observe $Y_n = Q(X_n) + \eta_n$, where the η_i's are i.i.d. random variables distributed as η. Then we define

$$X_n^* = X_i \text{ if } 1 \le i \le n \text{ and } Y_i = \min_{1 \le j \le n} Y_j.$$

This is nothing but the pure random search algorithm, employed as if we were unaware of the presence of any noise. Our study of this algorithm will reveal how noise-sensitive or robust pure random search really is. Not unexpectedly, the behavior of the algorithm depends upon the nature of the noise distribution. The noise will be called **stable** if for all $\epsilon > 0$,

$$\lim_{x \downarrow -\infty} \frac{G(x-\epsilon)}{G(x)} = 0,$$

where G is the distribution function of η, and $0/0$ is considered as zero. This will be called <u>Gnedenko's condition</u> (see Lemma 1). A sufficient condition for stability is that G has a density g and

$$\lim_{x \downarrow -\infty} \frac{g(x)}{G(x)} = \infty$$

.

EXAMPLES. If G does not have a left tail (i.e., $G(x_0) = 0$ for some $x_0 > -\infty$), then the noise is stable. Normal noise is also stable, but double exponential noise is not. In fact, the exponential distribution is on the borderline between stability and non-stability. Distributions with a diverging hazard rate as we travel from the origin out to $-\infty$ are stable. Thus, stable noise distributions have small left tails. In fact, $\int_{-\infty}^0 |x|^k G(dx) < \infty$ for all k. $\square$

The reason why stable noise will turn out to be manageable, is that $\min(\eta_1, \ldots, \eta_n)$ is basically known to fall into an interval of arbitrary small positive length around some deterministic value a_n with probability tending to one for some sequence $\{a_n\}$. It could thus happen that $a_n \to -\infty$ as $n \to \infty$, yet this is not a problem. This was also observed by Rubinstein and Weissman (1979). In the next section, we obtain a necessary and sufficient condition for the weak convergence of $Q(X_n^*)$ for the pure random search algorithm.

THEOREM 1. *If η is stable, then $Q(X_n^*) \to q_\mu$ in probability. Conversely, if η is not stable, then $Q(X_n^*)$ does not tend to q_μ in probability for any search problem for which for all ϵ small enough, $F(2\epsilon) - F(\epsilon) > 0$.*

We picked the name "stable noise" because the minimum η_n^* of $\eta_1, \ldots, \eta_n$ is stable in the sense used in the literature on order statistics, that is, there exists a sequence a_n such that $\eta_n^* - a_n \to 0$ in probability. We will prove the minimal properties needed further on in this section. The equivalence property A of Lemma 1 is due to Gnedenko (1943), while parts B and C are inherent in the fundamental paper of Geffroy (1958).

LEMMA 1.

A. *G is the distribution function of stable noise if and only if $\eta_n^* - a_n \to 0$ in probability for some sequence a_n.*

B. *If $\eta_n^* - a_n \to 0$ in probability for some sequence a_n, then $nG(a_n - \epsilon) \to 0$ and $nG(a_n + \epsilon) \to \infty$ as $n \to \infty$ for all $\epsilon > 0$. Also, if $b_n = G^{\mathrm{inv}}(1/n)$, then $nG(b_n - \epsilon) \to 0$ and $nG(b_n + \epsilon) \to \infty$ as $n \to \infty$ for all $\epsilon > 0$. Note: $G^{\mathrm{inv}}(u) = \inf\{t : G(t) \geq u\}$.*

C. *If the noise distribution is not stable, then there exist positive constants $a < b$, a sequence $\{a_n\}$, a subsequence n_i and an $\epsilon > 0$ such that $n_i G(a_{n_i} - \epsilon) \geq a$ and $n_i G(a_{n_i}) \leq b$ for all i.*

PROOF. We begin with property B. Note that by assumption, $(1-G(a_n-\epsilon))^n \to 1$, and thus $nG(a_n-\epsilon) \to 0$. Also, $(1-G(a_n+\epsilon))^n \to 0$ implies $nG(a_n+\epsilon) \to \infty$. Observe that $nG(b_n) \le 1 \le nG(b_n+u)$ for any $u > 0$. This shows that eventually, $a_n + \epsilon \ge b_n \ge a_n - \epsilon$. Thus, $nG(b_n + 2\epsilon) \ge nG(a_n + \epsilon) \to \infty$ and $nG(b_n - 2\epsilon) \le nG(a_n - \epsilon) \to 0$.

Let us turn to A. We first show that B implies Gnedenko's condition. We can assume without loss of generality that a_n is monotone decreasing since a_n can be replaced by $G^{\text{inv}}(1/n)$ in view of property B. For every $u < a_1$, we find n such that $a_n > u \ge a_{n+1}$. Thus, $G(a_n+\epsilon) \ge G(u+\epsilon) \ge G(a_{n+1}+\epsilon)$ and $G(a_n-\epsilon) \ge G(u-\epsilon) \ge G(a_{n+1}-\epsilon)$. Thus,

$$\frac{G(a_{n+1}+\epsilon)}{G(a_n-\epsilon)} \le \frac{G(u+\epsilon)}{G(u-\epsilon)} \le \frac{G(a_n+\epsilon)}{G(a_{n-1}-\epsilon)}.$$

The case of bounded a_n is trivial, so assume $a_n \to -\infty$. Now let $u \to -\infty$ (and thus $n \to \infty$), and deduce that $G(u+\epsilon)/G(u-\epsilon) \to \infty$. Since ϵ is arbitrary, we obtain Gnedenko's condition.

Next, part A follows if we can show that Gnedenko's condition implies the existence of the sequence a_n; proving the existence of a_n is equivalent to proving the existence of a_n such that $nG(a_n+\epsilon) \to \infty$ and $nG(a_n-\epsilon) \to 0$ for all $\epsilon > 0$. Let us take $a_n = G^{\text{inv}}(1/n)$. From the definition of G^{inv}, we note that for any $u > 0$, $v \in (0,1)$, $G(G^{\text{inv}}(v)) \le v \le G(G^{\text{inv}}(v)+u)$. Thus, by the Gnedenko condition, for $\epsilon > 0$,

$$nG(a_n+\epsilon) \ge \frac{G(a_n+\epsilon)}{G(a_n+\epsilon/2)} \to \infty \ .$$

Similarly,

$$nG(a_n-\epsilon) \le \frac{G(a_n-\epsilon)}{G(a_n)} \to 0.$$

This concludes the proof of part A.

For part C, we see that necessarily $a_n \to \infty$. We define $a_n = G^{\text{inv}}(1/n)$. By assumption, there exists an $\epsilon > 0$, a sequence $x_k \uparrow \infty$, and an $a > 0$ such that $G(x_k - 2\epsilon)/G(x_k+\epsilon) \ge a$ for all k. Next, by definition of a_n, we note that for infinitely many indices n, we have $x_k - \epsilon \le a_n \le x_k$. These define the subsequence n_i that we will use. Observe that $nG(a_n) \le 1$ for all n, while for all n with $x_k - \epsilon \le a_n \le x_k$,

$$nG(a_n-\epsilon) \ge \frac{G(a_n-\epsilon)}{G(a_n+\epsilon)} \ge \frac{G(x_k-2\epsilon)}{G(x_k+\epsilon)} \ge a > 0.\square$$

PROOF OF THEOREM 1. Let F be the distribution function of $Q(X_1) - q_\mu$, and let G be the distribution function of η. We first show that stable noise is sufficient for convergence. For brevity, we denote q_μ by q. Furthermore, $\epsilon > 0$ is an arbitrary constant, and $a_n = G^{\text{inv}}(1/n)$. Observe that the event $[Q(X_n^*) \le q+3\epsilon]$ is implied by $A_n \cap B_n$, where A_n is the event that for some $i \le n$, $Q(X_i) \le q+\epsilon$ and simultaneously, $\eta_i \le a_n + \epsilon$; and B_n is the event that for all $i \le n$ we either have $Q(X_i) \le q + 3\epsilon$ or $\eta_i > a_n - \epsilon$. The convergence follows if we can show that $\mathbf{P}\{A_n\} \to 1$ and $\mathbf{P}\{B_n\} \to 1$ as $n \to \infty$.

$$\begin{aligned}\mathbf{P}\{A_n^c\} &= \mathbf{P}\{\cap_{i=1}^n \{[Q(X_i) > q+\epsilon] \cup [Q(X_i) \le q+\epsilon][\eta_i > a_n+\epsilon]\}\} \\ &= (1 - F(\epsilon) + F(\epsilon)(1-G(a_n+\epsilon)))^n \\ &\le \exp(-nF(\epsilon)G(a_n+\epsilon)).\end{aligned}$$

This tends to zero by property B of Lemma 1. Next,

$$\begin{aligned}\mathbf{P}\{B_n^c\} &= \mathbf{P}\left\{\cap_{i=1}^n \{[Q(X_i) \le q+3\epsilon] \cup [Q(X_i > q+3\epsilon][\eta_i \le a_n - \epsilon]\}\right\} \\ &= (F(3\epsilon) + (1-F(3\epsilon))G(a_n-\epsilon))^n \\ &\le \exp(-n(1-F(3\epsilon))(1-G(a_n-\epsilon))) \\ &\sim \exp(-(1-F(3\epsilon))n)\end{aligned}$$

where we used property B of Lemma 1 again. This concludes the proof of the sufficiency.

The necessity is obtained as follows. Since G is not stable, we can find positive constants $a < b$, a sequence $\{a_n\}$, a subsequence n_i and an $\epsilon > 0$ such that $n_i G(a_{n_i} - 2\epsilon) \ge a$ and $n_i G(a_{n_i}) \le b$ for all i (Lemma 1, property C). Let n be in this subsequence n_i, and let $(N, M, n-N-M)$ be the multinomial random vector with the number of $Q(X_i) - q$ values ($i \le n$) in $[0,\epsilon]$, $(\epsilon, 2\epsilon]$ and $(2\epsilon, \infty]$, respectively. We first condition on this vector. Clearly, if for some $Q(X_i) - q$ in the second interval we have $\eta_i \le a_n - 2\epsilon$, while for all $Q(X_i) - q$ in the first interval, we have $\eta_i > a_n$, then $Q(X_n^*) > q + \epsilon$. Thus, the conditional probability of this event is

$$\begin{aligned}&\mathbf{P}\left\{Q(X_n^*) > q_\mu + \epsilon\right\} \ge (1-G(a_n))^N (1-(1-G(a_n-2\epsilon))^M) \\ &\ge (1-b/n)^N (1-(1-a/n)^M).\end{aligned}$$

To un-condition, we use the multinomial moment generating function

$$\mathbf{E}\left\{s^N t^M v^{n-N-M}\right\} = (sp_1 + tp_2 + vp_3)^n \ ,$$

where $s, t, v \ge 0$ and p_1, p_2, p_3 are the parameters of the multinomial distribution. This yields

$$\begin{aligned}\mathbf{P}\left\{Q(X_n^*) - q_\mu > \epsilon\right\} &\ge \mathbf{E}\left\{(1-b/n)^N (1-(1-a/n)^M)\right\} \\ &= (1-F(\epsilon)b/n)^n - (1-F(\epsilon)b/n - (F(2\epsilon)-F(\epsilon))a/n)^n \\ &\sim \exp(-bF(\epsilon))(1-\exp(-a(F(2\epsilon)-F(\epsilon)))) > 0,\end{aligned}$$

provided that $F(2\epsilon) - F(\epsilon) > 0$. This can be guaranteed, since we can make ϵ smaller without compromising the validity of the lower bound. This concludes the proof of Theorem 1. $\square$

§8. Strong convergence and strong stability

There is a strong convergence analog of Theorem 1. We say that G is **strongly stable** noise if the minimal order statistic of $\eta_1, \ldots, \eta_n$ is strongly stable, i.e., there exists a sequence of numbers a_n such that $\eta_n^* - a_n \to 0$ almost surely as $n \to \infty$.

THEOREM 2. *If η is strongly stable, then $Q(X_n^*) \to q_\mu$ almost surely.*

PROOF. Since strong stability implies stability, we recall from Lemma 1 that we can assume without loss of generality that $a_n = G^{\mathrm{inv}}(1/n)$. For $\delta > 0$, let $a_{\delta n} = G^{\mathrm{inv}}(1/(\delta n))$. Observe that in any case $a_n - a_{\delta n} \to 0$ as $n \to \infty$. Let $\epsilon > 0$ be arbitrary, and let S_n be the set of indices between 1 and n for which $Q(X_i) \le q_\mu + \epsilon$. Let $|S_n|$ denote the cardinality of this set. As $|S_n|$ is binomial $(n, F(\epsilon))$, it is easy to see that $|S_n| \ge nF(\epsilon)/2$ except possibly finitely often with probability one. Define

$$\eta_n^* = \min_{i \le n} \eta_i \ , \ \eta_n^{\#} = \min_{i \in S_n} \eta_i .$$

Since $|S_n| \to \infty$ almost surely, we have $\eta_n^* - a_n \to 0$ and $\eta_n^{\#} - a_{|S_n|} \to 0$ almost surely. Define $c = F(\epsilon)/2$. Consider the following inclusion of events (assuming for convenience that cn is integer-valued):

$$
\begin{aligned}
&[Q(X_n^*) > q_\mu + 4\epsilon \text{ i.o. }] \\
&\quad \subseteq [|S_n| < cn \text{ i.o. }] \cup [\eta_n^* \leq a_n - \epsilon \text{ i.o. }] \cup \\
&\qquad [\eta_n^{\#} \geq a_{|S_n|} + \epsilon \text{ i.o. }] \cup [a_n \leq a_{cn} - \epsilon \text{ i.o. }].
\end{aligned}
$$

The event on the right hand side has zero probability. Hence so does the event on the left hand side. $\square$

It is more difficult to provide characterizations of strongly stable noises, although several sufficient and a few necessary conditions are known. For an in-depth treatment, we refer to Geffroy (1958). It suffices to summarize a few key results here. The following condition due to Geffroy is sufficient:

$$\lim_{x \downarrow -\infty} \frac{G(x+\epsilon)}{G(x)\log(1/G(x))} = \infty \ .$$

This function comes close to being necessary. Indeed, if G is strongly stable, and $G(x)/G(x+\epsilon)$ is monotone in the left tail beyond some point, then Geffroy's condition must necessarily hold. If G has a density g, then another sufficient condition is that

$$\lim_{x \downarrow -\infty} \frac{g(x)}{G(x)\log\log(1/G(x))} = \infty \ .$$

It is easy to verify now that noise with density $c\exp(-|x|^\alpha)$ is strongly stable for $a > 1$ and is not stable when $a \leq 1$. The borderline is once again close to the double exponential distribution. To more clearly identify the threshold cases consider the noise distribution function given by $G(x) = \exp(-|x|\log|x|h(|x|)), x \leq -1$, where $h(x)$ is a positive function. It can be shown that for constant $h(x) \equiv H > 0$, the noise is stable but not strongly stable. However, if $h(|x|) \uparrow \infty$ as $x \to -\infty$, then G is strongly stable (Geffroy, 1958).

§9. Mixed random search

Assume next that we are not using pure random search, in the hope of assuring consistency for more noise distributions, or speeding up the method of the previous section. A certain minimum amount of global search is needed in any case. So, we will consider the following prototype model of an algorithm: X_n has distribution given by $\alpha_n\mu + (1-\alpha_n)\mu_n$, where $\{\alpha_n\}$ is a sequence of probabilities, and μ_n is an arbitrary probability distribution that may depend upon the past (all X_i with $i < n$, and all observations made up to time n). We call this mixed random search, since with probability α_n, the trial point is generated according to the pure random search distribution μ. In the noiseless case, $\sum_{n=1}^\infty \alpha_n = \infty$ is necessary and sufficient for strong convergence of $Q(X_n^*)$ to q_μ for all search problems and all ways of choosing μ_n. One is tempted to think that under stable noise and the same condition on α_n, we still have at least weak convergence. Unfortunately, this is not so. What has gone wrong is that it is possible that too few probe points have small Q-values, and that the smallest η_i corresponding to these probe values is not small enough to "beat" the smallest η_i among the other probe values. In the next theorem, we establish that convergence under stable noise conditions can only be guaranteed when a positive fraction of the search effort is spent on global search, e.g. when $\inf_n \alpha_n > 0$. Otherwise, we can still have convergence of $Q(X_n^*)$ to q_μ, but we lose the guarantee, as there are several possible counterexamples.

Theorem 3. *If* $\liminf_{n\to\infty} \frac{1}{n}\sum_{i=1}^n \alpha_i \ge a$ *for some* $a > 0$*, then under stable noise conditions,* $Q(X_n^*) \to q_\mu$ *in probability, and under strong stable noise conditions,* $Q(X_n^*) \to q_\mu$ *almost surely. Conversely, if* $\liminf_{n\to\infty} \frac{1}{n}\sum_{i=1}^n \alpha_i = 0$*, then there exists a search problem, a stable noise distribution* G*, and a way of choosing the sequence* μ_n *such that* $Q(X_n^*)$ *does not converge weakly to* q_μ*.*

Proof. We only prove the first part with strong stability. We mimic the proof of Theorem 2 with the following modification: let S_n be the set of indices between 1 and n for which $Q(X_i) \le q_\mu + \epsilon$ and X_i is generated according to the $\alpha_i\mu$ portion of the mixture distribution. Note that $|S_n|$ is distributed as $\sum_{i=1}^n A_i$, where the A_i's are i.i.d. Bernoulli random variables with parameter $\alpha_i F(\epsilon)$. Note that $|S_n| / \sum_{i=1}^n \alpha_i F(\epsilon) \to 1$ almost surely, so that $|S_n| \ge naF(\epsilon)/2$ except possibly finitely often with probability one. Then apply the event inclusion of Theorem 2 with $c = aF(\epsilon)/2$. The weak convergence is obtained in a similar fashion from the inclusion

$$[Q(X_n^*) > q_\mu + 4\epsilon] \subseteq [|S_n| < cn] \cup [\eta_n^* \le a_n - \epsilon] \cup [\eta_n^\# \ge a_{|S_n|} + \epsilon] \cup [a_n \le a_{cn} - \epsilon],$$

where we use the notation of Theorem 2. All events on the right-hand-side have probability tending to zero with n. □

§10. Strategies for general additive noise

From the previous sections, we conclude that under some circumstances, noise can be tolerated in pure random search. However, it is very difficult to verify whether the noise at hand is indeed stable; and the rate of convergence takes a terrible beating with some stable noise distributions. There are algorithms whose convergence is guaranteed under all noise distributions, and whose rate of convergence depends mainly upon the search distribution F, and not on the noise distribution! Such niceties do not come free: a slower rate of convergence results even when the algorithm operates under no noise; and in one of the two strategies discussed further on, the storage requirements grow unbounded with time.

How can we proceed? If we stick to the idea of trying to obtain improvements of X_n^* by comparing observations drawn at X_{n+1} with observations drawn at X_n^*, then we should be very careful not to accept X_{n+1} as the new X_{n+1}^* unless we are reasonably sure that $Q(X_{n+1}) < Q(X_n^*)$. Thus, several noise-perturbed observations are needed at each point, and some additional protection is needed in terms of thresholds that give X_n^* the edge in a comparison. This is only natural, since X_n^* embodies all the information gathered so far, and we should not throw it away lightly. To make the rate of convergence less dependent upon the noise distribution, we should consider comparisons between observations that are based on the relative ranks of these only. This kind of solution was first proposed in Devroye (1977).

Yakowitz and Fisher (1973) proposed another strategy, in which no information is ever thrown away. We store for each X_n all the observations ever made at it. At time $n+1$, draw more observations at all the previous probe points and at a new probe point X_{n+1}, and choose X_{n+1}^* from the entire pool of probe points. From an information theoretic point of view, this is a clever though costly policy. The decision which probe point to take should be based upon ranks, once again. Yakowitz and Fisher (1973) employ the empirical distribution functions of the observations at each probe point. Devroye (1977) in a related approach advocates the use of the Wilcoxon-Mann-Whitney rank statistic and modifications of it. Note that the fact that no information is ever discarded makes these algorithms nonsequential in nature; this is good for parallel implementations, but notoriously bad when rates of convergence are considered.

Consider first the nonsequential strategy in its crudest form: define X_n^* as the "best" among the first n probe points $X_1, \ldots, X_n$, which in turn are i.i.d. random vectors with probability distribution μ. Let $\lambda_n \uparrow \infty$ be a sequence of integers to be chosen by the user. We make sure that at each time n, the function is sampled λ_n times at each n. If the previous observations are not thrown away, then this means that $\lambda_n + (n-1)(\lambda_n - \lambda_{n-1})$ new observations are necessary, λ_n at X_n, and $\lambda_n - \lambda_{n-1}$ at each of X_i, $1 \le i < n$. The observations are stored in a giant array $Y_{ij}, 1 \le i \le n, 1 \le j \le \lambda_n$. As an index of the "goodness" of X_i, we could consider the average

$$Q_{ni} = \frac{1}{\lambda_n} \sum_{j=1}^{\lambda_n} Y_{ij}.$$

The best point is the one with the smallest average. Clearly, this strategy cannot be universal since for good performance, it is necessary that the law of large numbers applies, and thus that $\mathbf{E}|\eta| < \infty$, where η is the noise random variable. However, in view of its simplicity and importance, we will return to this solution in a further subsection.

If we order all the components $(Y_{i1}, \ldots, Y_{i\lambda_n})$ of the i-th vector so as to obtain $(Y_{i(1)} < \cdots < Y_{i(\lambda_n)})$, then other measures of the goodness may include the medians M_{ni} of these components, or "quick and dirty" methods such as Gastwirth's statistic (Gastwirth, 1966)

$$G_{ni} = \frac{1}{10}(3Y_{i(\lambda_n/3)} + 4Y_{i(\lambda_n/2)} + 3Y_{i(2\lambda_n/3)}).$$

We might define X_n^* as that point among $X_1, \ldots, X_n$ with the smallest value of G_{ni}.

We could also introduce the notion of pairwise competitions between the X_i: for example, we say that X_i wins against X_k if $G_{ni} < G_{nk}$. The X_i with the most wins is selected to be X_n^*. This view leads to precisely that X_i with the smallest value of G_{ni}. However, pairwise competitions can go further, as we now illustrate. Yakowitz and Fisher (1973) thought it useful to work with the empirical distribution functions $F_{ni}(x) = \frac{1}{\lambda_n} \sum_{j=1}^{\lambda_n} 1_{[Y_{ij} \le x]}$. Our approach may be considered as a tournament between $X_1, \ldots, X_n$, in which $\binom{n}{2}$ matches are played, one per pair (X_i, X_j). Let

$$D_{ij} = \sup_x (F_{ni}(x) - F_{nj}(x)).$$

We say that X_i wins its match with X_j when $D_{ij} \ge D_{ji}$. We define X_n^* as that member of $\{X_1, \ldots, X_n\}$ with the most wins. In case of ties, we take the member with the smallest index. We call this the tournament method. Rather than using the Kolmogorov-Smirnov based statistic suggested above, one might also consider other rank statistics based on medians of observations or upon the generalized Wilcoxon statistic (Wilcoxon, 1945; Mann and Whitney, 1947), where only comparisons between Y_{ij}'s are used.

The number of function evaluations used up to iteration n is $n\lambda_n$. In addition, in some cases, some sorting may be necessary, and in nearly all the cases, the entire Y_{ij} array needs to be kept in storage. Also, there are $\binom{n}{2}$ matches to determine the wins, leading to a complexity, in the n-th iteration alone of about $n^2\lambda_n$. This seems to be extremely wasteful. We discuss some time-saving modifications further on. We provide a typical result here (Theorem 4) for the tournament method.

THEOREM 4. *Let X_n^* be chosen by the tournament method. If $\lambda_n / \log n \to \infty$, then $Q(X_n^*) \to q_\mu$ almost surely as $n \to \infty$.*

PROOF. Fix $\epsilon > 0$ and let G be the noise distribution, and F_n is the empirical distribution function obtained from λ_n observations taken at $X_i = x$. Note that $F_n(y+Q(x))$ can be considered as an estimate of $G(y)$. In fact, by the Glivenko-Cantelli lemma, we know that $\sup_y |F_n(y+Q(x)) - G(y)| \to 0$ almost surely. But much more is true. By an inequality due to Dvoretzky, Kiefer and Wolfowitz (1956), in a final form derived by Massart (1990), we have for all $\delta > 0$,

$$\mathrm{P}\left\{\sup_y |F_n(y+Q(x)) - G(y)| \geq \delta\right\} \leq 2e^{-2\lambda_n \delta^2}.$$

For $\epsilon > 0$, we define the positive constant

$$\delta = \sup_x (G(x+\epsilon) - G(x)),$$

and let A_n be the event that for all $i \leq n$, $\sup_y |F_{ni}(y+Q(X_i)) - G(y)| < \delta/4$. Clearly, then,

$$\mathrm{P}\left\{A_n^c\right\} \leq 2ne^{-\lambda_n \delta^2/8}.$$

An important event for us is T_ϵ, the event that all matches (X_i, X_j) $(i \neq j, i, j, \leq n)$ with $Q(X_i) \leq Q(X_j) - \epsilon$ have a fair outcome, that is, X_i wins against X_j. Let us compute a bound on the probability of T_ϵ^c: We observe that $T_\epsilon^c \subseteq A_n^c$ so that $\mathrm{P}\{T_\epsilon^c\} \leq \delta/4 < 0 + 2(\delta/4)$. To see this, fix (i,j) with $Q(X_i) \leq Q(X_j) - \epsilon$. Then $D_{ij} \leq D_{ji}$ (X_i loses) if

$$\sup_x (F_{ni}(x) - F_{nj}(x)) \leq \sup_x (F_{nj}(x) - F_{ni}(x))$$

but this in turn implies that

$$\delta - 2(\delta/4) < 0 + 2(\delta/4)$$

which is impossible. Thus, every such match has a fair outcome.

Consider next the tournament, and partition the X_i's in four groups according to whether $Q(X_i)$ belongs to one of these intervals: $[q_\mu, q_\mu + \epsilon]$, $(q_\mu + \epsilon, q_\mu + 2\epsilon]$, $(q_\mu + 2\epsilon, q_\mu + 3\epsilon]$, $(q_\mu + 3\epsilon, \infty)$. The cardinalities of the groups are N_i, $1 \leq i \leq 4$. If T_ϵ holds, then any member of group 1 wins its match against any member of groups 3 and 4, for at least $N_3 + N_4$ wins. Any member of group 4 can at most win against other members of group 4 or all members of group 3, for at most $N_3 + N_4 - 1$ wins. Thus, the tournament winner must come from groups 1, 2 or 3, unless there is no X_i in any of these groups. Thus,

$$Q(X_n^*) \leq q_\mu + 3\epsilon \ .$$

We showed the following:

$$\mathrm{P}\{Q(X_n^*) > q_\mu + 3\epsilon\} \leq \mathrm{P}\{T_\epsilon^c \cup [N_1 + N_2 + N_3 = 0]\} \leq 2ne^{-2\lambda_n \delta^2} + (1 - F(3\epsilon))^n.$$

This is summable in n for every $\epsilon > 0$, so that we can conclude $Q(X_n^*) \to q_\mu$ almost surely by the Borel-Cantelli lemma. □

It is a simple exercise to modify Theorem 4 to include <u>mixing schemes</u>, i.e., schemes of sampling in which X_n has probability measure $\alpha_n \mu + (1-\alpha_n)\mu_n$, where $\alpha_n \in [0,1]$ and μ_n is an arbitrary probability measure. We note that $Q(X_n^*)$ still converges to q_μ almost surely if we merely add the standard mixing condition

$$\sum_{n=1}^{\infty} \alpha_n = \infty.$$

Following the proof of Theorem 4, we notice indeed that for arbitrary N,

$$\bigcup_{n\geq N} [Q(X_n^*) > q_\mu + 3\epsilon] \subseteq \bigcup_{n\geq N} A_n^c \cup \left[\bigcap_{i=1}^{N} [Q(X_i) > q_\mu + 3\epsilon]\right].$$

Hence,

$$\mathrm{P}\left\{\sup_{n\geq N} Q(X_n^*) > q_\mu + 3\epsilon\right\} \leq \sum_{n\geq N} 2ne^{-\lambda_n \delta^2/8} + \exp\left(-F(3\epsilon)\sum_{i=1}^{N}\alpha_i\right).$$

Here we used an argument as in the proof of Theorem 3 in the section on mixed global random search. The bound tends to 0 with N, and thus $Q(X_n^*) \to q_\mu$ almost surely if $\lambda_n/\log n \to \infty$.

Consider the simplest sequential strategy, comparable to a new player entering at each iteration in the tournament, and playing against the best player seen thus far. Assume that the X_i are i.i.d. with probability measure μ, and that at iteration n, we obtain samples Y_{nj} and Y_{n-1,j^*}, $1 \leq j \leq \lambda_n$, at X_n and X_{n-1}^* respectively. For comparing performances, we use suitably modified statistics such as

$$Z_n = \sup_y (F_n(y) - F_{n-1}^*(y)),$$

where F_n is the empirical distribution function for the Y_{nj} sample, and F_{n-1}^* is the empirical distribution function for the other sample. Define X_n^* according to the rule

$$X_n^* = \begin{cases} X_n & \text{if } Z_n < \epsilon_n \\ X_{n-1}^* & \text{otherwise,} \end{cases}$$

where $\epsilon_n \geq 0$ is a threshold designed to give some advantage to X_{n-1}^*, since the information contained in X_{n-1}^* is too valuable to throw away without some form of protection. We may introduce mixing as long as we can guarantee that for any Borel set A, $\mathrm{P}\{X_n \in A\} \geq \alpha_n \mu(A)$, and α_n is the usual mixing coefficient. This allows us to combine global and local search and to concentrate global search efforts in promising regions of the search domain.

Theorem 5 (devroye, 1978). *Assume that the above sequential strategy is used with mixing, and that comparisons are based upon Z_n. Assume furthermore that $\lim_{n\to\infty}\epsilon_n = 0$, $\sum_{n=1}^{\infty}\alpha_n = \infty$. If $\lambda_n\epsilon_n^2/2 - \log(1/\alpha_n) \to \infty$, then $Q(X_n^*) \to q_\mu$ in probability as $n \to \infty$. If*

$$\sum_{n=1}^{\infty} \exp(-\lambda_n \epsilon_n^2/2) < \infty ,$$

then $Q(X_n^*) \to q_\mu$ with probability one as $n \to \infty$.

Proof. We will use fact that if $Q(X_n^*) \to q_\mu$ in probability, and

$$\sum_{n=1}^{\infty} \mathrm{P}\left\{Q(X_{n+1}^*) > Q(X_n^*)\right\} < \infty ,$$

then $Q(X_n^*) \to q_\mu$ almost surely. This follows easily from the inequality

$$\mathrm{P}\left\{\bigcup_{n\geq N} [Q(X_n^*) > q_\mu + \epsilon]\right\} \leq \mathrm{P}\left\{Q(X_N^*) > q_\mu + \epsilon\right\} + \sum_{n=N}^{\infty} \mathrm{P}\left\{Q(X_{n+1}^*) > Q(X_n^*)\right\}.$$

Now,

$$\begin{aligned}
\mathbf{P}\left\{Q(X^*_{n+1}) > Q(X^*_n)\right\} &\leq \mathbf{P}\left\{Q(X_{n+1}) > Q(X^*_n); Z_n > \epsilon_n\right\} \\
&\leq \mathbf{P}\left\{\sup_y |F_n(y) - G(y - Q(X_n)| > \frac{\epsilon_n}{2}\right\} \\
&\quad + \mathbf{P}\left\{\sup_y |F^*_{n-1}(y) - G(y - Q(X^*_{n-1})| > \frac{\epsilon_n}{2}\right\} \\
&\leq 4\exp(-\lambda_n \epsilon_n^2/2),
\end{aligned}$$

where we used the Dvoretzky-Kiefer-Wolfowitz inequality. Thus, the summability condition is satisfied. To obtain the weak convergence, we argue as follows. Define

$$\rho_n = \inf_{x,x':Q(x')\leq Q(x)-\epsilon} \mathbf{P}\left\{X^*_{n+1} = X_{n+1} | X_{n+1} = x', X^*_n = x\right\}$$

and

$$p_n = \sup_{x,x':Q(x')\geq Q(x)} \mathbf{P}\left\{X^*_{n+1} = X_{n+1} | X_{n+1} = x', X^*_n = x\right\} \ .$$

Then

$$\begin{aligned}
\xi_{n+1} &\stackrel{\text{def}}{=} \mathbf{P}\left\{Q(X^*_{n+1}) > q_\mu + 2\epsilon\right\} \\
&\leq \mathbf{P}\left\{Q(X^*_n) > q_\mu + 2\epsilon; [Q(X_{n+1}) \leq q_\mu + \epsilon; X^*_{n+1} = X_{n+1}]^c\right\} \\
&\quad + \mathbf{P}\left\{Q(X^*_n) \leq q_\mu + 2\epsilon; Q(X_{n+1}) > q_\mu + 2\epsilon; X^*_{n+1} = X_{n+1}\right\} \\
&\leq \xi_n(1 - \alpha_n F(\epsilon)\rho_n) + p_n \ .
\end{aligned}$$

A bit of analysis shows that $\xi_n \to 0$ when $\sum_{n=1}^\infty \alpha_n \rho_n = \infty$ and either $\sum_n p_n < \infty$ or $p_n/(\alpha_n q_n) \to 0$. But we have already seen that

$$p_n \leq 4\exp(-\lambda_n \epsilon_n^2/2).$$

Define $\delta = \sup_x (G(x+\epsilon) - G(x))$. Then, for $\epsilon_n \leq \delta/2$,

$$q_n \geq 1 - 4\exp(-\lambda_n \delta^2/2).$$

Thus, when $\lambda_n \to \infty$, $\epsilon_n \to 0$, and either $\sum_n \alpha_n = \infty$ or $\alpha_n^{-1}\exp(-\lambda_n\epsilon_n^2/2) \to 0$, we have weak convergence of $Q(X^*_n)$ to q_μ. This concludes the proof of Theorem 5. □

Choosing ϵ_n and λ_n is an arbitrary process. Can we do without? For example, can we boldly choose $\epsilon_n = 0$ and still have guaranteed convergence for all search problems and all noises? The answer is yes, provided that λ_n increases faster than quadratically in n. This result may come as a bit of a surprise, since we based the proof of Theorem 5 on the observation that the event $[Q(X^*_n) > Q(X^*_{n-1})]$ occurs finitely often with probability one. We will now allow this event to happen infinitely often with any positive probability, but by increasing λ_n quickly enough, the total sum of the "bad" moves $\sum_n (Q(X^*_n) - Q(X^*_{n-1}))_+$ is finite almost surely.

Theorem 6. *Assume that the sequential strategy is used with mixing, and that comparisons are based upon Z_n. Assume furthermore that $\epsilon_n \equiv 0$, $\sum_{n=1}^\infty \alpha_n = \infty$, and $\sum_{n=1}^\infty 1/\sqrt{\lambda_n} < \infty$. Then $Q(X^*_n) \to q_\mu$ with probability one as $n \to \infty$.*

PROOF. Let $\epsilon > 0$ be arbitrary. Observe that

$$\bigcup_{n\geq N} [Q(X_n^*) > q_\mu + 2\epsilon] \subseteq [Q(X_N^*) > q_\mu + \epsilon] \cup \left[\sum_{n\geq N} (Q(X_{n+1}^*) - Q(X_n^*))_+ > \epsilon \right] .$$

Thus, strong convergence follows from weak convergence if we can prove that

$$\sum_{n=1}^{\infty} (Q(X_{n+1}^*) - Q(X_n^*))_+ < \infty \text{ almost surely} .$$

This follows if

$$\sum_{n=1}^{\infty} \mathbf{1}_{\left[Q(X_{n+1}^*) > Q(X_n^*)+1\right]} < \infty \text{ almost surely}$$

and

$$\sum_{n=1}^{\infty} \min\left(1, (Q(X_{n+1}^*) - Q(X_n^*))_+\right) < \infty \text{ almost surely} .$$

By the Beppo-Levi theorem, the former condition is implied by

$$\sum_{n=1}^{\infty} \mathbf{P}\left\{Q(X_{n+1}^*) > Q(X_n^*) + 1\right\} < \infty.$$

By the Beppo-Levi theorem, the latter condition is implied by

$$\sum_{n=1}^{\infty} \mathbf{E}\left\{\min\left(1, (Q(X_{n+1}^*) - Q(X_n^*))_+\right)\right\} < \infty .$$

Define

$$p_n(s) = \sup_{x,y: Q(y)-Q(x)=s} \mathbf{P}\left\{X_{n+1}^* = X_{n+1} | X_{n+1} = y, X_n^* = x\right\} .$$

We recall that

$$p_n(s) \leq 4 \exp(-\lambda_n \delta(s)^2/2),$$

where

$$\delta(s) = \sup_x (G(x + s/2) - G(x)) .$$

We note that if L is the distance between the third and first quartiles of G, then $\delta(s) \geq s/(2s+4L)$. This is easily seen by partitioning the two quartile interval of length L into $\lceil 2L/s \rceil \leq 1 + 2L/s$ intervals of length $s/2$ or less. The maximum probability content of one of these intervals is at least $1/2(1 + 2L/s)$. From the proof of Theorem 5,

$$\begin{aligned}
&\mathbf{E}\left\{\min\left(1, (Q(X_{n+1}^*) - Q(X_n^*))_+\right)\right\} \\
&\quad \leq \sup_{1\geq s>0} s p_n(s) \\
&\quad \leq \sup_{1\geq s>0} 4s \exp(-\lambda_n \delta(s)^2/2) \\
&\quad \leq \sup_{1\geq s>0} 4s \exp(-\lambda_n s^2/2(2s+4L)^2) \\
&\quad = O\left(1/\sqrt{\lambda_n}\right) .
\end{aligned}$$

This is summable in n, as required. Also,

$$\mathbb{P}\left\{Q(X_{n+1}^*) > Q(X_n^*) + 1\right\} \le p_n(1)$$

and this is summable in n. To establish the weak convergence of $Q(X_n^*)$, we begin with the following inclusion of events, in which m is a positive integer:

$$[Q(X_{n+m}^*) > q_\mu + 2\epsilon] \subseteq \left[\bigcap_{k=n+1}^{n+m} A_k\right] \cup \left[\sum_{k=n}^{n+m} (Q(X_{n+1}^*) - Q(X_n^*))_+ > \epsilon\right],$$

where

$$A_k \stackrel{\text{def}}{=} [Q(X_k) > q_\mu + \epsilon/2] \cup [Q(X_k) \le q_\mu + \epsilon/2]\left[\sup_x |F_k(x) - G(x + Q(X_k))| > \delta(\epsilon)/2\right]$$
$$\cup [Q(X_k) \le q_\mu + \epsilon/2]\left[\sup_x |F_{k-1}^*(x) - G(x + Q(X_{k-1}^*))| > \delta(\epsilon)/2\right].$$

Here we used the notation of the proof of Theorem 5. By estimates obtained there, we note that

$$\begin{aligned}\mathbb{P}\left\{\bigcap_{k=n+1}^{n+m} A_k\right\} &\le \prod_{k=n+1}^{n+m}\left(1 - \alpha_k F(\epsilon/2)\left(1 - 4e^{-\lambda_k \delta(\epsilon)^2/2}\right)\right)\\ &\le \prod_{k=n+1}^{n+m}(1 - \alpha_k F(\epsilon/2)/2)\\ &\le \exp\left(-\sum_{k=n+1}^{n+m}\alpha_k F(\epsilon/2)/2\right),\end{aligned}$$

provided that n is large enough. For any fixed n, we can choose m so large that this upper bound is smaller than a given small constant. Thus, $\mathbb{P}\left\{Q(X_{n+m}^*) > q_\mu + 2\epsilon\right\}$ is smaller than a given small constant if we first choose n large enough, and then choose m appropriately. This concludes the proof of Theorem 6 when Z_n is used. $\square$

Both the sequential and nonsequential strategies can be applied to the case in which we compare points on the basis of Q_{ni}, the average of λ_n observations made at X_i. This is in fact nothing more than the situation we will encounter when we wish to minimize a regression function. Indeed, taking averages would only make sense when the mean exists. Assume thus that we have the regression model

$$Q(x) = \mathbb{E}\left\{g(x, \eta)\right\},$$

where g is a real-valued function, and η is some random variable. For fixed x, we cannot observe realizations of η, but rather an i.i.d. sample $Y_1, Y_2, \ldots$, where $Y_i = g(x, \eta_i)$, and the η_i's are i.i.d. In the additive noise case, we have the special form $g(x, \eta) = Q(x) + \eta$, where η is a zero mean random variable. We first consider the nonsequential model, in which the probe points are i.i.d. with probability measure μ. The convergence is established in Theorem 7. Clearly, the choice of λ_n (and thus our cost) increases with the size of the tail or tails of η. In the best scenario, λ_n should increase faster than $\log n$.

Theorem 7. *Let X_n^* be chosen on the basis of the smallest value of Q_{ni}. Assume that η is such that $\mathbb{E}|\eta| < \infty$, $\mathbb{E}\eta = 0$. Then $Q(X_n^*) \to q_\mu$ in probability as $n \to \infty$ when condition A holds: $\mathbb{P}\{|\eta| > n\} = o(1/n^{t+1})$ for some $t > 0$ (a sufficient condition for this is $\mathbb{E}|\eta|^{1+t} < \infty$), and $\liminf \lambda_n / n^{1/t} > 0$. We have strong convergence if B or C hold: (condition B) $\mathbb{E}e^{t\eta} < \infty$ for all t in an open neighborhood of the origin, and $\lim_{n\to\infty} \lambda_n / \log n = \infty$; (condition C) $\mathbb{P}\{|\eta| > n\} = o(1/n^{t+1})$ for some $t > 0$, $\liminf \lambda_n/n^{1/t} > 0$, and $\sum_n \lambda_n^{-t} < \infty$. Finally, for any noise with zero moment, there exists a sequence $\{\lambda_n\}$ such that $Q(X_n^*) \to q_\mu$ almost surely as $n \to \infty$.*

Proof. Let A_n be the event that for all $i \le n$, $|Q_{ni} - Q(X_i)| \le \epsilon/2$. We note that

$$\mathbb{P}\{A_n^c\} \le n\mathbb{P}\left\{\left|\frac{1}{\lambda_n}\sum_{i=1}^{\lambda_n} \eta_i\right| > \epsilon/2\right\}.$$

Arguing as in the proof of Theorem 4, we have

$$\mathbb{P}\left\{Q(X_n^*) > q_\mu + 2\epsilon\right\} \le \mathbb{P}\{A_n^c\} + (1 - F(\epsilon))^n.$$

This implies weak convergence of $Q(X_n^*)$ to q_μ if $\mathbb{P}\{A_n^c\} \to 0$. From the work of Baum and Katz (1985) (see Petrov (1975, pp. 283-286)), we retain that $\mathbb{P}\{A_n^c\} = o(n/\lambda_n^t)$ for all $\epsilon > 0$ if $\mathbb{P}\{|\eta| \ge n\} = o(n^{-t-1})$, where $t \ge 0$. If $\mathbb{E}|\eta|^{1+t} < \infty$ for some $t > 0$, then by Theorem 28 of Petrov (1975), $\mathbb{P}\{|\eta| > n\} = o(1/n^{t+1})$, so condition A is satisfied. Finally, if $\mathbb{E}e^{t\eta} \le \infty$ for all t in an open neighborhood of the origin, then $\mathbb{P}\{A_n^c\} = O(n\rho^{-\lambda_n})$ for some constant $\rho \in (0,1)$ (see e.g. Petrov (1975, pp. 54-55)). This proves the weak convergence portion of the theorem.

The strong convergence under condition B follows without trouble from the above bounds and the Borel-Cantelli lemma. For condition C, we need the fact that if $\mathbb{P}\left\{|n^{-1}\sum_{i=1}^n \eta_i| \ge \epsilon\right\} = o(n^{-u})$ for every $\epsilon > 0$, where $u > 0$ is a constant, then we have

$$\mathbb{P}\left\{\sup_{n\ge N}\left|n^{-1}\sum_{i=1}^n \eta_i\right| \ge \epsilon\right\} = o(N^{-u})$$

(see Petrov, 1975, p. 284). Now observe that under condition C,

$$\begin{aligned}
\mathbb{P}\left\{\sup_{n\ge N} Q(X_n^*) > q_\mu + 2\epsilon\right\} &\le \mathbb{P}\left\{\sup_{n\ge N}\sup_{i\le n}|Q_{ni} - Q(X_i)| > \epsilon/2\right\} + (1-F(\epsilon))^N \\
&\le \sum_{i=1}^\infty \mathbb{P}\left\{\sup_{n\ge\max(i,N)} |Q_{ni} - Q(X_i)| > \epsilon/2\right\} + (1-F(\epsilon))^N \\
&= \sum_{i=1}^{N-1} \mathbb{P}\left\{\sup_{n\ge N}|Q_{ni} - Q(X_i)| > \epsilon/2\right\} + (1-F(\epsilon))^N \\
&\quad + \sum_{i=N}^\infty \mathbb{P}\left\{\sup_{n\ge i}|Q_{ni} - Q(X_i)| > \epsilon/2\right\} + (1-F(\epsilon))^N \\
&= o(N\lambda_N^{-t}) + \sum_{i=N}^\infty c\lambda_i^{-t} + (1-F(\epsilon))^N
\end{aligned}$$

for some constant $c > 0$. This tends to 0 with N.

The last part of the theorem follows from the weak law of large numbers. Indeed, there exists a function $\omega(u)$ with $\omega(u) \to 0$ as $u \to \infty$ such that $\mathbb{P}\left\{|n^{-1}\sum_{i=1}^{n}\eta_i| > \epsilon\right\} = \omega(n)$. Thus, $\mathbb{P}\left\{A_n^c\right\} \leq n\omega(\lambda_n)$. Clearly, this is countable in n if we choose λ_n so large that $\omega(\lambda_n) \leq 2^{-n}$. Therefore, the last part of the theorem follows by the Borel-Cantelli lemma. □

The remarks regarding rates of convergence made following Theorem 4 apply here as well. What is new though is that we have lost the universality, since we have to impose conditions on the noise. If we apply the algorithm with some predetermined choice of λ_n, we have no guarantee whatsoever that the algorithm will be convergent. And even if we knew the noise distribution, it may not be possible to avoid divergence for any manner of choosing λ_n.

§11. Universal convergence

In the search for universal optimization methods, we conclude with the following observation. Let Q be a function on the positive integers with finite infimum. Assume that for each $x \in Z_+ = \{1, 2, 3, \ldots\}$, there exists an infinite sequence of i.i.d. random variables $Y(x,1), Y(x,2), \ldots$, called the observations. We have $Q(x) = \mathbb{E}\{Y(x,1)\}$. A search algorithm is a sequence of functions

$$f_{n+1}(x,k; X_1, K_1, X_2, K_2, \ldots, X_n, K_n)\ , n \geq 0,$$

where as a function of (x,k), f_{n+1} describes a distribution. The sequence $(X_1, K_1, \ldots, X_n, K_n)$ is called the history. For each n, starting with $n = 0$, we generate a pair (X_n, K_n) from the distribution given by $f_n(x,k)$. This pair allows us to look at $Y(X_n, K_n)$. Thus, after n iterations, we have accessed at most n observations. A search algorithm needs g_n, a function of $X_i, K_i, Y(X_i, K_i), 1 \leq i \leq n$, that maps to Z_+ to determine which integer is taken as the best estimate X_n^* of the minimum: $X_n^* = g_n(X_1, K_1, Y(X_1, K_1), \ldots, X_1, K_1, Y(X_1, K_1)$. A search algorithm is a sequence of mappings (f_n, g_n). A search algorithm is universally convergent if for all functions Q with $\inf_x Q(x) > -\infty$, and all distributions of $Y(x,1), x \in Z_+$, $Q(X_n^*) \to \inf_x Q(x)$ in probability. We do not know if a universally convergent search algorithm exists. The difficulty of the question follows from the following observation. At time n, we have explored at most n integers and looked at at most n observations. Assume that we have n observations at each of the first n integers (consider this as a present of $n^2 - n$ additional observations). Let us average these observations, and define X_n^* as the integer with the smallest average. While at each integer, the law of large numbers holds, it is not true that the averages converge at the same rate to their means, and this procedure may actually see $Q(X_n^*)$ diverge to infinity in some probabilistic sense.

§12. References

E. Aarts and J. Korst, *Simulated Annealing and Boltzmann Machines*, John Wiley, New York, 1989.

D. H. Ackley, *A Connectionist Machine for Genetic Hillclimbing*, Kluwer Academic Publishers, Boston, 1987.

F. Aluffi-Pentini, V. Parisi, and F. Zirilli, "Global optimization and stochastic differential equations," *Journal of Optimization Theory and Applications*, vol. 47, pp. 1–16 , 1985.

S. Anily and A. Federgruen, "Simulated annealing methods with general acceptance probabilities," *Journal of Applied Probability*, vol. 24, pp. 657–667, 1987.

W. Banzhaf, P. Nordin, and R. E. Keller, *Genetic Programming : An Introduction : On the Automatic Evolution of Computer Programs and Its Applications*, Morgan Kaufman, San Mateo, CA, 1998.

L. Baum and M. Katz, “Convergence rates in the law of large numbers,” *Transactions of the American Mathematical Society*, vol. 121, pp. 108–123, 1965.

R. W. Becker and G. V. Lago, “A global optimization algorithm,” in: *Proceedings of the 8th Annual Allerton Conference on Circuit and System Theory*, pp. 3–12, 1970.

G. A. Bekey and M. T. Ung, “A comparative evaluation of two global search algorithms,” *IEEE Transactions on Systems, Man and Cybernetics*, vol. SMC-4, pp. 112–116, 1974.

G. L. Bilbro and W. E. Snyder, “Optimization of functions with many minima,” *IEEE Transactions on Systems, Man and Cybernetics*, vol. SMC-21, pp. 840–849, 1991.

C. G. E. Boender, A. H. G. Rinnooy Kan, L. Stougie, and G. T. Timmer, “A stochastic method for global optimization,” *Mathematical Programming*, vol. 22, pp. 125–140, 1982.

I. O. Bohachevsky, M. E. Johnson, and M. L. Stein, “Generalized simulated annealing for function optimization,” *Technometrics*, vol. 28, pp. 209–217, 1986.

H. J. Bremermann, “Optimization through evolution and recombination,” in: *Self-Organizing Systems*, (edited by M. C. Yovits, G. T. Jacobi and G. D. Goldstein), pp. 93–106, Spartan Books, Washington, D.C, 1962.

H. J. Bremermann, “Numerical optimization procedures derived from biological evolution processes,” in: *Cybernetic Problems in Bionics*, (edited by H. L. Oestreicher and D. R. Moore), pp. 597–616, Gordon and Breach Science Publishers, New York, 1968.

S. H. Brooks, “A discussion of random methods for seeking maxima,” *Operations Research*, vol. 6, pp. 244–251, 1958.

S. H. Brooks, “A comparison of maximum-seeking methods,” *Operations Research*, vol. 7, pp. 430–457, 1959.

T. Bäck, F. Hoffmeister, and H.-P. Schwefel, “A survey of evolution strategies,” in: *Proceedings of the Fourth International Conference on Genetic Algorithms*, (edited by R. K. Belew and L. B. Booker), pp. 2–9, Morgan Kaufman Publishers, San Mateo, CA, 1991.

V. Cerny, “Thermodynamical approach to the traveling salesman problem: an efficient simulation algorithm,” *Journal of Optimization Theory and Applications*, vol. 45, pp. 41–51, 1985.

A. Dekkers and E. Aarts, “Global optimization and simulated annealing,” *Mathematical Programming*, vol. 50, pp. 367–393, 1991.

L. Devroye, “The compound random search algorithm,” in: *Proceedings of the International Symposium on Systems Engineering and Analysis, Purdue University*, vol. 2, pp. 195–110, 1972.

L. Devroye, “On the convergence of statistical search,” *IEEE Transactions on Systems, Man and Cybernetics*, vol. SMC-6, pp. 46–56, 1976.

L. Devroye, "On random search with a learning memory," in: *Proceedings of the IEEE Conference on Cybernetics and Society, Washington*, pp. 704–711, 1976.

L. Devroye, "An expanding automaton for use in stochastic optimization," *Journal of Cybernetics and Information Science*, vol. 1, pp. 82–94, 1977.

L. Devroye, "The uniform convergence of nearest neighbor regression function estimators and their application in optimization," *IEEE Transactions on Information Theory*, vol. IT-24, pp. 142–151, 1978.

L. Devroye, "Rank statistics in multimodal stochastic optimization," Technical Report, School of Computer Science, McGill University, 1978.

L. Devroye, "Progressive global random search of continuous functions," *Mathematical Programming*, vol. 15, pp. 330–342, 1978.

L. Devroye, "Global random search in stochastic optimization problems," in: *Proceedings of Optimization Days 1979, Montreal*, 1979.

L. de Biase and F. Frontini, "A stochastic method for global optimization: its structure and numerical performance," in: *Towards Global Optimization 2*, (edited by L. C. W. Dixon and G. P. Szegö), pp. 85–102, North Holland, Amsterdam, 1978.

A. Dvoretzky, J. C. Kiefer, and J. Wolfowitz, "Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator," *Annals of Mathematical Statistics*, vol. 27, pp. 642–669, 1956.

S. M. Ermakov and A. A. Zhiglyavskii, "On random search for a global extremum," *Theory of Probability and its Applications*, vol. 28, pp. 136–141, 1983.

Yu. Ermoliev and R. Wets, "Stochastic programming, and introduction," in: *Numerical Techniques of Stochastic Optimization*, (edited by R. J.-B. Wets and Yu. M. Ermoliev), pp. 1–32, Springer-Verlag, New York, 1988.

L. Fisher and S. J. Yakowitz, "Uniform convergence of the potential function algorithm," *SIAM Journal on Control and Optimization*, vol. 14, pp. 95–103, 1976.

J. L. Gastwirth, "On robust procedures," *Journal of the American Statistical Association*, vol. 61, pp. 929–948, 1966.

M. Gaviano, "Some general results on the convergence of random search algorithms in minimization problems," in: *Towards Global Optimization*, (edited by L. C. W. Dixon and G. P. Szegö), pp. 149–157, North Holland, New York, 1975.

J. Geffroy, "Contributions à la théorie des valeurs extrêmes," *Publications de l'Institut de Statistique des Universités de Paris*, vol. 7, pp. 37–185, 1958.

S. B. Gelfand and S. K. Mitter, "Weak convergence of Markov chain sampling methods and annealing algorithms to diffusions," *Journal of Optimization Theory and Applications*, vol. 68, pp. 483–498, 1991.

S. Geman and C.-R. Hwang, "Diffusions for global optimization," *SIAM Journal on Control and Optimization*, vol. 24, pp. 1031–1043, 1986.

B. Gidas, "Global optimization via the Langevin equation," in: *Proceedings of the 24th IEEE Conference on Decision and Control, Fort Lauderdale*, pp. 774–778, 1985.

A B. V. Gnedenko, *Sur la distribution du terme maximum d'une série aléatoire*, *Annals of Mathematics*, vol. 44, pp. 423–453, 1943.

D. E. Goldberg, *Genetic Algorithms in Search Optimization and Machine Learning*, Addison-Wesley, Reading, Mass, 1989.

L. S. Gurin, "Random search in the presence of noise," *Engineering Cybernetics*, vol. 4, pp. 252–260, 1966.

L. S. Gurin and L. A. Rastrigin, "Convergence of the random search method in the presence of noise," *Automation and Remote Control*, vol. 26, pp. 1505–1511, 1965.

H. Haario and E. Saksman, "Simulated annealing process in general state space," *Advances in Applied Probability*, vol. 23, pp. 866–893, 1991.

B. Hajek, "Cooling schedules for optimal annealing," *Mathematics of Operations Research*, vol. 13, pp. 311–329, 1988.

B. Hajek and G. Sasaki, "Simulated annealing—to cool or not," *Systems and Control Letters*, vol. 12, pp. 443–447, 1989.

J. H. Holland, "Genetic algorithms and the optimal allocation of trials," *SIAM Journal on Computing*, vol. 2, pp. 88–105, 1973.

J. H. Holland, *Adaptation in Natural and Artificial Systems : An Introductory Analysis With Applications to Biology, Control, and Artificial Intelligence*, MIT Press, Cambridge, Mass, 1992.

R. A. Jarvis, "Adaptive global search by the process of competitive evolution," *IEEE Transactions on Systems, Man and Cybernetics*, vol. SMC-5, pp. 297–311, 1975.

D. S. Johnson, C. R. Aragon, L. A. McGeogh, and C. Schevon, "Optimization by simulated annealing: an experimental evaluation; part I, graph partitioning," *Operations Research*, vol. 37, pp. 865–892, 1989.

A. H. G. Rinnooy Kan and G. T. Timmer, "Stochastic methods for global optimization," *American Journal of Mathematical and Management Sciences*, vol. 4, pp. 7–40, 1984.

V. G. Karmanov, "Convergence estimates for iterative minimization methods," *USSR Computational Mathematics and Mathematical Physics*, vol. 14(1), pp. 1–13, 1974.

J. Kiefer and J. Wolfowitz, "Stochastic estimation of the maximum of a regression function," *Annals of Mathematical Statistics*, vol. 23, pp. 462–466, 1952.

S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, "Optimization by simulated annealing," *Science*, vol. 220, pp. 671–680, 1983.

J. Koronacki, "Convergence of random-search algorithms," *Automatic Control and Computer Sciences*, vol. 10(4), pp. 39–45, 1976.

H. L. Kushner, "Asymptotic global behavior for stochastic approximation via diffusion with slowly decreasing noise effects: global minimization via Monte Carlo," *SIAM Journal on Applied Mathematics*, vol. 47, pp. 169–185, 1987.

T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Advances in Applied Mathematics*, vol. 6, pp. 4–22, 1985.

H. B. Mann and D. R. Whitney, "On a test of whether one or two random variables is stochastically larger than the other," *Annals of Mathematical Statistics*, vol. 18, pp. 50–60, 1947.

K. Marti, "Minimizing noisy objective functions by random search methods," *Zeitschrift für Angewandte Mathematik und Mechanik*, vol. 62, pp. T377–T380, 1982.

K. Marti, "Stochastic optimization in structural design," *Zeitschrift für Angewandte Mathematik und Mechanik*, vol. 72, pp. T452–T464, 1992.

P. Massart, "The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality," *Annals of Probability*, vol. 18, pp. 1269–1283, 1990.

J. Matyas, "Random optimization," *Automation and Remote Control*, vol. 26, pp. 244–251, 1965.

S. M. Meerkov, "Deceleration in the search for the global extremum of a function," *Automation and Remote Control*, vol. 33, pp. 2029–2037, 1972.

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equation of state calculation by fast computing machines," *Journal of Chemical Physics*, vol. 21, pp. 1087–1092, 1953.

J. B. Mockus, *Bayesian Approach to Global Optimization*, Kluwer Academic Publishers, Dordrecht, Netherlands, 1989.

R. Männer and H.-P. Schwefel, "Parallel Problem Solving from Nature," vol. 496, Lecture Notes in Computer Science, Springer-Verlag, Berlin, 1991.

V. V. Petrov, *Sums of Independent Random Variables*, Springer-Verlag, Berlin, 1975.

M. A. Pinsky, *Lecture Notes on Random Evolution*, World Scientific Publishing Company, Singapore, 1991.

J. Pintér, "Convergence properties of stochastic optimization procedures," *Mathematische Operationsforschung und Statistik, Series Optimization*, vol. 15, pp. 405–427, 1984.

J. Pintér, *Global Optimization in Action*, Kluwer Academic Publishers, Dordrecht, 1996.

W. L. Price, "Global optimization by controlled random search," *Journal of Optimization Theory and Applications*, vol. 40, pp. 333–348, 1983.

I. Rechenberg, *Evolutionsstrategie—Optimierung technischer Systeme nach Prinzipien der biologischen Evolution*, Frommann-Holzboog, Stuttgart, 1973.

A. H. G. Rinnooy Kan and G. T. Timmer, “Stochastic global optimization methods part II: multi level methods,” *Mathematical Programming*, vol. 39, pp. 57–78, 1987.

A. H. G. Rinnooy Kan and G. T. Timmer, “Stochastic global optimization methods part I: clustering methods,” *Mathematical Programming*, vol. 39, pp. 27–56, 1987.

H. Robbins, “Some aspects of the sequential design of experiments,” *Bulletin of the American Mathematical Society*, vol. 58, pp. 527–535, 1952.

R. Y. Rubinstein and I. Weissman, “The Monte Carlo method for global optimization,” *Cahiers du Centre d’Etude de Recherche Operationelle*, vol. 21, pp. 143–149, 1979.

M. A. Schumer and K. Steiglitz, “Adaptive step size random search,” *IEEE Transactions on Automatic Control*, vol. AC-13, pp. 270–276, 1968.

H.-P. Schwefel, *Modellen mittels der Evolutionsstrategie*, Birkhäuser Verlag, Basel, 1977.

H.-P. Schwefel, *Numerical Optimization of Computer Models*, John Wiley, Chichester, 1981.

H.-P. Schwefel, *Evolution and Optimum Seeking*, Wiley, New York, 1995.

C. Sechen, *VLSI Placement and Global Routing using Simulated Annealing*, Kluwer Academic Publishers, 1988.

G. R. Shorack and J. A. Wellner, *Empirical Processes with Applications to Statistics*, John Wiley, New York, 1986.

B. O. Shubert, “A sequential method seeking the global maximum of a function,” *SIAM Journal on Numerical Analysis*, vol. 9, pp. 379–388, 1972.

F. J. Solis and R. B. Wets, “Minimization by random search techniques,” *Mathematics of Operations Research*, vol. 1, pp. 19–30, 1981.

G. S. Tarasenko, “Convergence of adaptive algorithms of random search,” *Cybernetics*, vol. 13, pp. 725–728, 1977.

A. Törn, *Global Optimization as a Combination of Global and Local Search*, Skriftserie Utgiven av Handelshogskolan vid Abo Akademi, Abo, Finland, 1974.

A. Törn, “Probabilistic global optimization, a cluster analysis approach,” in: *Proceedings of the EURO II Conference, Stockholm, Sweden*, pp. 521–527, North Holland, Amsterdam, 1976.

A. Törn and A. Žilinskas, *Global Optimization*, Lecture Notes in Computer Science, vol. 350, Springer-Verlag, Berlin, 1989.

K. Uosaki, H. Imamura, M. Tasaka, and H. Sugiyama, “A heuristic method for maxima searching in case of multimodal surfaces,” *Technology Reports of Osaka University*, vol. 20, pp. 337–344, 1970.

D. Vanderbilt and S. G. Louie, “A Monte Carlo simulated annealing approach to optimization over continuous variables,” *Journal of Computational Physics*, vol. 56, pp. 259–271, 1984.

P. J. M. Van Laarhoven and E. H. L. Aarts, *Simulated Annealing: Theory and Applications*, D. Reidel, Dordrecht, 1987.

M. T. Wasan, *Stochastic Approximation*, Cambridge University Press, New York, 1969.

F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, vol. 1, pp. 80–83, 1945.

S. Yakowitz, "Automatic learning: theorems for concurrent simulation and optimization," in: *1992 Winter Simulation Conference Proceedings*, (edited by J. J. Swain, D. Goldsman, R. C. Crain and J. R. Wilson), pp. 487–493, ACM, Baltimore, MD, 1992.

S. J. Yakowitz, "A statistical foundation for machine learning, with application to go-moku," *Computers and Mathematics with Applications*, vol. 17, pp. 1095–1102, 1989.

S. J. Yakowitz, "A globally-convergent stochastic approximation," Technical Report, Systems and Industrial Engineering Department, University of Arizona, Tucson, AZ, 1989.

S. J. Yakowitz, "On stochastic approximation and its generalizations," Technical Report, Systems and Industrial Engineering Department, University of Arizona, Tucson, AZ, 1989.

S. J. Yakowitz, "A decision model and methodology for the AIDS epidemic," *Applied Mathematics and Computation*, vol. 55, pp. 149–172, 1992.

S. J. Yakowitz, "Global stochastic approximation," *SIAM Journal on Control and Optimization*, vol. 31, pp. 30–40, 1993.

S. J. Yakowitz and L. Fisher, "On sequential search for the maximum of an unknown function," *Journal of Mathematical Analysis and Applications*, vol. 41, pp. 234–259, 1973.

S. J. Yakowitz, R. Hayes, and J. Gani, "Automatic learning for dynamic Markov fields, with applications to epidemiology," *Operations Research*, vol. 40, pp. 867–876, 1992.

S. J. Yakowitz, T. Jayawardena, and S. Li, "Theory for automatic learning under Markov-dependent noise, with applications," *IEEE Transactions on Automatic Control*, vol. AC-37, pp. 1316–1324, 1992.

S. J. Yakowitz and M. Kollier, "Machine learning for blackjack counting strategies," *Journal of Forecasting and Statistical Planning*, vol. 33, pp. 295–309, 1992.

S. J. Yakowitz and W. Lowe, "Nonparametric bandit methods," *Annals of Operations Research*, vol. 28, pp. 297–312, 1991.

S. J. Yakowitz and E. Lugosi, "Random search in the presence of noise, with application to machine learning," *SIAM Journal on Scientific and Statistical Computing*, vol. 11, pp. 702–712, 1990.

S. J. Yakowitz and A. Vesterdahl, "Contribution to automatic learning with application to self-tuning communication channel," Technical Report, Systems and Industrial Engineering Department, University of Arizona, 1993.

A. A. Zhigljavsky, *Theory of Global Random Search*, Kluwer Academic Publishers, Hingham, MA, 1991.