Reflections on Bandit Problems and Selection Methods in Uncertain Environments

Günter Rudolph
Universität Dortmund
Fachbereich Informatik, LS XI
D–44221 Dortmund, Germany
Rudolph@LS11.Informatik.Uni-Dortmund.de

Abstract

The behavior of selection methods used in evolutionary algorithms that operate in uncertain environments is investigated in the framework of parametric two–armed bandit problems. Asymptotically optimal selection strategies are based on the sequential probability ratio test which is proved to perform up to four times better than analogous strategies based on the optimal fixed size sample test. A variant of local binary tournament selection in a spatially structured population is shown to behave like a sequential test provided that the population size is optimally adjusted.

1 Introduction

John Holland (1975) used the so–called two–armed bandit problem as a formal model to derive general guidelines for the design of genetic algorithms that operate in uncertain environments. These uncertainties may appear in various situations: Assume that we have to control some physical system but do not completely know its law of motion. All we know is that there is a finite (say) number of alternatives — each of them being true with some probability. Consequently, the performance of each control strategy is a random variable and we are seeking for a control strategy with maximum expected performance. In case of static function optimization uncertainties arise whenever the objective function value is stochastically perturbed. If these perturbations are additive with zero mean it is desirable to choose that admissible point whose associated random objective function value has maximum mean.

An obvious (but not necessarily optimal) method to identify the random variable with maximum mean is based on the following fact: Let $X_1, \ldots, X_n$ be independent and identically distributed random variables with finite mean μ and variance σ^2. Then the mean and variance of the average $\bar{X}_n$ are $\mathsf{E}[\,\bar{X}_n\,] = \mu$ and $\mathsf{V}[\,\bar{X}_n\,] = \sigma^2/n$. Thus, the more samples are drawn the smaller is the uncertainty about the true mean. If we would draw infinitely many samples we would get the true value (zero variance). In practice, however, only a finite number of samples can be drawn so that this procedure is not immune from errors. Assume that at most $N < \infty$ trials can be drawn in total. The question is: How many trials should be allocated to each random variable to gather information about their (unknown) means before selecting the random variable with suspected maximum mean? Since this decision may be wrong with some probability (which decreases as the number of trials increase) there must exist a strategy that minimizes the sum of expected information costs and decision losses.

A general formal model to analyze the situation above is a version of the two–armed bandit problem (or k–armed in general). Suppose we are faced with a gambling machine equipped with two handles (i.e., the two–armed bandit). Each time we put a coin into the slot we may pull one arm which leads to a random payoff according to some probability distribution. Let the random payoff be represented by the two independent random variables X and Y. Our goal is to maximize the expected payoff under the constraint that our budget is restricted to $N < \infty$ coins. Thus, we need a strategy that quickly and reliably identifies the random variable with higher (or smaller) mean in order to allocate the remaining trials to the suspected superior arm. Unfortunately, quickness and reliability are conflicting objectives so that some compromise must be sought for. This multi–criteria optimization problem can be mapped into a single–criteria problem if we accept that information costs and decision losses can be measured in the same scale. If so, the sum of information costs and decision losses represents the overall loss to be minimized.

In the general problem formulation given above we are confronted with a nonparametric bandit problem. Although this situation is the more realistic one in practice, we shall consider parametric bandit problems here. It is assumed that we know the probability (density) functions f_0 and f_1 of both random variables, but we do not know which density is associated with which random variable. Thus, initially we have

$$\mathsf{P}\{\,(f_x = f_0, f_y = f_1)\,\} = \mathsf{P}\{\,(f_x = f_1, f_y = f_0)\,\} = \frac{1}{2}\,.$$

Chernoff (1959) remarks that optimal strategies for such bandit problems are difficult to characterize whereas asymptotically optimal results ought to be easily available. His approach was as follows: Without loss of generality let f_0 be the density with largest mean. Choose one arm arbitrarily, X say, and pull it repeatedly. The realizations of the random variables serve as input to a statistical procedure to test the hypothesis $H_0 : f_x = f_0$ versus hypothesis $H_1 : f_x = f_1$. If the statistical test suggest rejecting hypothesis H_0 then the remaining trials will be allocated to arm Y, otherwise we stay at arm X for the remaining trials. In fact, building on Chernoff's results Kiefer and Sacks (1963) proved that this two–staged strategy is asymptotically optimal, provided that we are using the 'optimal' statistical test.

The quality of a statistical test depends on three parameters: The sample size n and the two error probabilities $\alpha = \mathsf{P}\{\,\text{reject } H_0 \,|\, H_0 \text{ true}\}$ and $\beta = \mathsf{P}\{\,\text{accept } H_0 \,|\, H_0 \text{ false}\}$. Since the minimization of all three quantities yields conflicting goals, two parameters must be fixed while the third one is minimized. Under the scenario to test a simple hypothesis against another simple hypothesis optimal tests are known. Neyman and Pearson (1933) have developed a method to construct statistical tests such that the error probability β is minimized for given α and sample size n. Another type of optimality is associated with the sequential probability ratio test (SPRT) developed by Abraham Wald (1945). This test is of sequential nature and it minimizes the expected number of samples for given α and β regardless which hypothesis is actually true (the optimality is shown in Wald and Wolfowitz 1948).

Since the SPRT will be the method of choice here, it will be briefly presented in section 2 before we sketch the analysis for a single player in section 3. In section 4 the scenario is changed in that we use the SPRT in case of paired observations and compare its performance with fixed sample size tests of Holland (1975) and of Neyman/Pearson type. Section 5 is devoted to the situation in which the feedback of the gambling machine is not the realized outcome of both arms but merely which arm was better. Note that this type of ranking information is processed by tournament selection rules in evolutionary algorithms. This is the theme of section 6 where evolutionary selection methods are interpreted as an ingredient of a sequential statistical test. First, we investigate a population of two individuals whose fitness is represented by the outcome of the random payoff received from the two–armed bandit. As might have been expected, the performance is tremendously bad because of the missing individual memory. Then it is shown in case of a spatially structured population with a local tournament selection rule how a population can be used to store information. Moreover, the optimal population size (depending on N and the distribution of the random variables) is calculated numerically.

2 The Sequential Probability Ratio Test in a Nutshell

The sequential probability ratio test (SPRT) can be used to test a simple hypothesis versus another simple hypothesis. Let f_0 and f_1 be two probability density functions (or their discrete analogue) that are completely known. Suppose that the probability density function (p.d.f.) f_x of random variable X is either f_0 or f_1. Then the SPRT to test

$$H_0 : f_x = f_0 \quad \text{versus} \quad H_1 : f_x = f_1$$

is defined as follows: Choose constants $0 < A < 1 < B < \infty$ and define

$$\lambda_n = \prod_{i=1}^{n} \frac{f_1(X_i)}{f_0(X_i)}.$$

Let $T = \min\{n \in \mathbb{N} : \lambda_n \notin (A,B)\}$ be the random time at which the sequence λ_n leaves the open interval (A,B) for the first time. Stop sampling at time T and reject H_0 if $\lambda_T \geq B$, and accept H_0 if $\lambda_T \leq A$. This test errs with the probabilities

$\alpha = \mathsf{P}\{\,\lambda_T \geq B \,|\, H_0 \text{ true}\} = \mathsf{P}\{\,\text{reject } H_0 \,|\, H_0 \text{ true}\}$ and

$\beta = \mathsf{P}\{\,\lambda_T \leq A \,|\, H_1 \text{ true}\} = \mathsf{P}\{\,\text{reject } H_1 \,|\, H_1 \text{ true}\}$.

Remark: For the sake of brevity, probabilities and expectations conditioned by the event that H_i is true will be indexed by the subscript i. For example, $\beta = \mathsf{P}_1\{\,\lambda_T \leq A\,\}$.

The error probabilities are connected with the constants A and B via the fundamental inequalities

$$\begin{aligned} \beta &\leq A\,(1-\alpha) \text{ and} && (1)\\ \alpha &\leq (1-\beta)/B && (2) \end{aligned}$$

that fail to be equalities only because (in general) λ_T does not hit the boundaries A and B exactly. If we agree that the excess over the boundary can be neglected we may set $A = \beta/(1-\alpha)$, $B = \alpha/(1-\beta)$, and the inequalities for the expected stopping times below would turn to equalities as well:

$$\mathsf{E}_0[\,T\,] \geq \frac{1}{\kappa_0}\left(\alpha \log\frac{1-\beta}{\alpha} + (1-\alpha)\log\frac{\beta}{1-\alpha}\right) \quad (3)$$

$$\mathsf{E}_1[\,T\,] \geq \frac{1}{\kappa_1}\left(\beta \log\frac{\beta}{1-\alpha} + (1-\beta)\log\frac{1-\beta}{\alpha}\right) \quad (4)$$

where

$$\kappa_i = \int_{-\infty}^{\infty} \log\left(\frac{f_1(x)}{f_0(x)}\right) f_i(x)\,dx = \mathsf{E}_i\left[\log\frac{f_1(X)}{f_0(X)}\right]$$

for $i = 0, 1$ with $\kappa_0 < 0 < \kappa_1$. The constants κ_i can be interpreted as the Kullback–Leibler information numbers $I_0 = -\kappa_0 > 0$ and $I_1 = \kappa_1 > 0$ which serve as a measure of the hardness of distinguishing between both p.d.f.s if hypothesis H_i is true: The smaller is I_i, the more difficult is the problem. These constants will play an important role in subsequent sections.

3 Decisions Based on Observations from a Single Arm

Let X and Y denote the random variables associated with the random rewards of the two–armed bandit. If we draw random variable X its p.d.f. may be either f_0 or f_1 with the same probability. Without loss of generality let $\mu_0 > \mu_1$ with $\mu_i = \int x\, f_i(x)\,dx$.

Use the SPRT to test $H_0 : f_x = f_0$ versus $H_1 : f_x = f_1$. After the test has stopped at random time T it has drawn T random variables with p.d.f. f_x. If the test accepts the hypothesis H_0 the remaining $N{-}T$ random variables are drawn with p.d.f. f_x, otherwise with p.d.f. f_y. Thus, the random reward is

$$\begin{aligned} R \;=\; \sum_{i=1}^{T} X_i \quad &+\quad 1_{\{H_0 \text{ accepted}\}} \sum_{i=1}^{N-T} X_i \\ &+\quad 1_{\{H_0 \text{ rejected}\}} \sum_{i=1}^{N-T} Y_i \end{aligned}$$

provided that $T \le N$. Consequently, the test will be stopped at step $\min\{T, N\}$.

In the sequel we shall make use of the fact that the expectation of a sum of T independent and identically distributed random variables X_i, where T is a stopping time depending on the outcomes of X_i until T, is equal to the product $\mathsf{E}[\,T\,] \cdot \mathsf{E}[\,X_1\,]$ (see e.g. Gut 1988, p. 22).

Let H_0 be true. Then the expected reward is

$$\begin{aligned} \mathsf{E}_0[\,R\,] &= n_0\,\mu_0 + (1-\alpha)\,(N - n_0)\,\mu_0 + \alpha\,(N - n_0)\,\mu_1 \\ &= n_0\,\mu_0 + (N - n_0)\,[\,(1-\alpha)\,\mu_0 + \alpha\,\mu_1] \end{aligned}$$

where $\mathsf{E}_0[\,X\,] = \mu_0$, $\mathsf{E}_0[\,Y\,] = \mu_1$, $n_0 = \mathsf{E}_0[\,T\,]$ and $\mathsf{E}_0[\,1_{\{H_0\text{rejected}\}}\,] = \mathsf{P}_0\{\,H_0 \text{ rejected}\,\} = \alpha$. If H_1 is true the expected reward is

$$\begin{aligned} \mathsf{E}_1[\,R\,] &= n_1\,\mu_1 + \beta\,(N - n_1)\,\mu_1 + (1-\beta)\,(N - n_1)\,\mu_0 \\ &= n_1\,\mu_1 + (N - n_1)\,[\,(1-\beta)\,\mu_0 + \beta\,\mu_1] \end{aligned}$$

where $\mathsf{E}_1[\,X\,] = \mu_1$, $\mathsf{E}_1[\,Y\,] = \mu_0$, $n_1 = \mathsf{E}_1[\,T\,]$ and $\mathsf{E}_1[\,1_{\{H_0\text{accepted}\}}\,] = \mathsf{P}_1\{\,H_0 \text{ accepted}\,\} = \beta$. Since each hypothesis is equally likely the expected reward is given by

$$\begin{aligned} \mathsf{E}[\,R\,] &= \mathsf{P}\{\,H_0 \text{ true}\,\}\,\mathsf{E}_0[\,R\,] + \mathsf{P}\{\,H_1 \text{ true}\,\}\,\mathsf{E}_1[\,R\,] \\ &= N\,\mu_0 - \delta\,\left[\frac{\alpha+\beta}{2}\,N - \frac{\alpha}{2}\,n_0 + \frac{1-\beta}{2}\,n_1\right] \end{aligned}$$

where $\delta = \mu_0 - \mu_1 > 0$. Since $N\,\mu_0$ is the maximum reward under perfect information the expected loss of this method is

$$\mathsf{E}[\,L\,] = \delta\,\left[\frac{\alpha+\beta}{2}\,N - \frac{\alpha}{2}\,n_0 + \frac{1-\beta}{2}\,n_1\right]. \tag{5}$$

Under the assumption that the excess over the boundaries A or B can be neglected we may take inequalities (3) and (4) like equalities that can be inserted into equation (5). Partial differentiation with respect to α and β leads to the necessary optimality conditions. An analytic solution of the resulting system of nonlinear equations seems hopeless, but it is clear that the optimal error probabilities α^* and β^* will nonlinearly depend on the constants κ_0, κ_1, and N. If the p.d.f.s f_0 and f_1 are normal densities the constants κ_i are

$$\begin{aligned} \kappa_0 &= -\frac{1}{2}\left[\frac{(\mu_0-\mu_1)^2}{\sigma_1^2} + \frac{\sigma_0^2}{\sigma_1^2} - \log\frac{\sigma_0^2}{\sigma_1^2} - 1\right], \\ \kappa_1 &= \frac{1}{2}\left[\frac{(\mu_0-\mu_1)^2}{\sigma_0^2} + \frac{\sigma_1^2}{\sigma_0^2} - \log\frac{\sigma_1^2}{\sigma_0^2} - 1\right]. \end{aligned}$$

Since $x - 1 \ge \log x$ the equations above reveal that different variances facilitate the discrimination between the densities.

The scenario considered so far is tailored to a single player[1]. In evolutionary algorithms, however, a population of players/individuals is gambling in parallel. In the simplest case the population will consist of two individuals. Therefore it is reasonable to begin the analysis with strategies which must make decisions based on paired observations.

4 Decisions Based on Paired Observations

Suppose that there are two identical[2] gambling machines and two players each having $N/2$ coins. The players aim at maximizing the joint expected payoff. Their strategy is that the first player pulls arm X while the other player pulls arm Y at the other bandit until they agree which arm is better. Then both will choose the suspected better arm at each bandit for the remaining trials. Thus, both players are observing the sequence of pairs (X_i, Y_i) until they make a decision

[1] For small N the same scenario was analyzed in Macready and Wolpert (1996), who derived optimal strategies for $N = 1$ and $N = 2$ in case of Gaussian random variables.

[2] Two–armed bandits are identical if the have the same assignment of random variables to arms.

about the suspected better arm. The situation can be equivalently described as follows:

Let $D = X - Y$ so that $\mathsf{E}[\,D\,] = \mathsf{E}[\,X\,] - \mathsf{E}[\,Y\,]$ and $\mathsf{V}[\,D\,] = \mathsf{V}[\,X\,] + \mathsf{V}[\,Y\,]$. Thus, no matter which p.d.f. is associated with which random variable, the variance of D is $\eta^2 = \sigma_0^2 + \sigma_1^2$. The expectation of D, however, may be either $\theta_0 = \mu_0 - \mu_1 > 0$ or $\theta_1 = \mu_1 - \mu_0 = -\theta_0$. If the pairs (μ_i, σ_i) completely specify the distributions of the random variables, it is equivalent to saying that the p.d.f. of random variable D is

$$\tilde{f}_i(x) = \int_{-\infty}^{\infty} f_i(z)\, f_{1-i}(z-x)\, dz$$

where i is either 0 or 1 with the same probability. As a consequence, the players' decision to prefer a specific arm may be based on testing hypothesis $H_0 : f_D = \tilde{f}_0$ versus $H_1 : f_D = \tilde{f}_1$. If H_0 is rejected they will choose arm Y at both bandits, otherwise arm X. Provided that they use an optimal statistical test and that N is sufficiently large, their strategy should be asymptotically optimal.

4.1 Optimal Fixed Size Sample Test

After n paired samples the observation costs are $n\,\delta$ no matter which hypothesis is true. To determine the decision loss two subcases must be considered: If H_0 is true the test errs with probability α and $N - 2\,n$ trials are allocated to the wrong arm. Thus, the decision loss is $\alpha\,(N - 2\,n)\,\delta$. Similarly, if H_1 is true the decision loss is $\beta\,(N - 2\,n)\,\delta$. Since each event is equally likely the entire expected loss is

$$\mathsf{E}[\,L\,] = \delta \left[n + (N - 2\,n)\,\frac{\alpha + \beta}{2} \right] . \tag{6}$$

Suppose that X and Y are continuous random variables with $f_x \neq f_y$ and support $\mathbb{R}$. Then $\tilde{f}_0 \neq \tilde{f}_1$ and both p.d.f.s have support $\mathbb{R}$ as well. We like to test the hypothesis $H_0 : f_D = \tilde{f}_0$ against $H_1 : f_D = \tilde{f}_1$. For given n and $\alpha > 0$ the optimal Neyman–Pearson test runs as follows: Let

$$\lambda_n(D_1, \ldots, D_n) = \prod_{i=1}^{n} \frac{\tilde{f}_1(D_i)}{\tilde{f}_0(D_i)} \tag{7}$$

and reject H_0 if $\lambda_n(D_1, \ldots, D_n) > \xi$, otherwise accept H_0, where ξ is the solution of the equation

$$\mathsf{P}\{\, \lambda_n \leq \xi \mid H_0 \text{ true} \} = 1 - \alpha \,.$$

This test minimizes the error probability β for given $\alpha > 0$ and sample size $n < \infty$. Moreover, it is guaranteed that $\beta < 1 - \alpha$.

Suppose that f_0 and f_1 are p.d.f.s of normally distributed random variables with distribution $\mathcal{N}(\mu_i, \sigma_i^2)$, so that $\tilde{f}_0$ and $\tilde{f}_1$ are the densities of normal random variables with known means $\theta_0 > \theta_1$ and common known variance $\eta^2 > 0$. Then (7) becomes

$$\lambda_n(D_1, \ldots, D_n) = \exp\left(-\frac{\theta_0 - \theta_1}{\eta^2} \sum_{i=1}^{n} D_i + n\,\frac{\theta_0^2 - \theta_1^2}{2\,\eta^2} \right) .$$

Notice that $S_n = \sum_{i=1}^{n} D_i \sim \mathcal{N}(n\,\theta_0, n\,\eta^2)$ if hypothesis H_0 is true, and $S_n \sim \mathcal{N}(n\,\theta_1, n\,\eta^2)$ otherwise. To determine ξ we have to solve the equation

$$\begin{aligned} \mathsf{P}_0\{\, \lambda_n \leq \xi \,\} &= \mathsf{P}_0 \left\{ S_n \geq \frac{n\,(\theta_0^2 - \theta_1^2) - 2\,\eta^2 \log \xi}{2\,(\theta_0 - \theta_1)} \right\} \\ &= \Phi\left(\frac{n\,(\theta_0 - \theta_1)^2 + 2\,\eta^2 \log \xi}{2\,(\theta_0 - \theta_1)\,\eta\,\sqrt{n}} \right) \\ &= \Phi\left(\frac{\log \xi}{\gamma_n} + \frac{\gamma_n}{2} \right) \\ &\overset{!}{=} 1 - \alpha \,. \end{aligned}$$

where $\gamma_n = (\theta_0 - \theta_1)\,\sqrt{n}/\eta$ and $\Phi(\cdot)$ denotes the distribution function of the standard normal random variable. It follows that

$$\frac{\log \xi}{\gamma_n} + \frac{\gamma_n}{2} \overset{!}{=} u_\alpha := \Phi^{-1}(1 - \alpha)$$

and finally

$$\xi = \exp\left(\gamma_n\, u_\alpha - \frac{\gamma_n^2}{2} \right) . \tag{8}$$

The test rejects hypothesis H_0 if $\lambda_n(D_1, \ldots, D_n) > \xi$ which is equivalent to the condition

$$\bar{D}_n \leq \theta_0 - u_\alpha\, \eta/\sqrt{n} \tag{9}$$

where $\bar{D}_n = S_n/n$ is the average of $D_1, \ldots, D_n$. Taking into account equation (8) the error probability of the second kind can be easily obtained via

$$\begin{aligned} \beta \;=\; \mathsf{P}_1\{\, \lambda_n \leq \xi \,\} &= \mathsf{P}_1\{\, S_n \geq n\,\theta_0 - u_\alpha\,\eta\,\sqrt{n} \,\} \\ &= \Phi(u_\alpha - \gamma_n) \,. \end{aligned} \tag{10}$$

Let $\mathcal{L}(n, \alpha)$ denote the expected loss for given $n \in \mathbb{N}$ and $\alpha \in (0, 1)$. Owing to equations (6) and (10) we obtain

$$\mathcal{L}(n, \alpha) = \delta \left[n + (N - 2\,n)\,\frac{\alpha + \Phi(u_\alpha - \gamma_n)}{2} \right] .$$

Thus, the task to determine the minimal expected loss requires the solution of a mixed–integer optimization problem. Partial differentiation with respect to α leads to the necessary condition

$$\frac{\partial}{\partial\,\alpha}\, \Phi(\Phi^{-1}(1 - \alpha) - \gamma_n) \overset{!}{=} -1 \,.$$

Let $\Phi'(x) = \varphi(x)$. Since

$$\begin{aligned} \frac{\partial}{\partial\,\alpha}\, \Phi(\Phi^{-1}(1 - \alpha) - \gamma_n) &= -\frac{\varphi(\Phi^{-1}(1 - \alpha) - \gamma_n)}{\varphi(\Phi^{-1}(1 - \alpha))} \\ &= -\exp\left(\frac{2\,\gamma_n\, u_\alpha - \gamma_n^2}{2} \right) \end{aligned}$$

the necessary condition reduces to $u_\alpha = \gamma_n/2$ which may be equivalently expressed as

$$\begin{aligned} \gamma_n &= 2\,u_\alpha\,, && (11)\\ \alpha &= 1-\Phi(\gamma_n/2)\,, && (12)\\ n &= 4\,\Phi^{-2}(1-\alpha)/\gamma_1^2\,. && (13) \end{aligned}$$

Insertion of equation (11) into the condition that the test rejects H_0 equation (9) becomes

$$\bar{D}_n \le \theta_0 - u_\alpha\,\eta/\sqrt{n} = (\theta_0+\theta_1)/2\,.$$

Since $\theta_0 = -\theta_1$ the optimal fixed size sample test rejects H_0 if $\bar{D}_n \le 0$, or if $\bar{X}_n \le \bar{Y}_n$ since $\bar{D}_n = \bar{X}_n - \bar{Y}_n$. Notice that this is exactly the test proposed in Holland (1975), pp. 83–85.

It remains to determine the optimal sample size n^*. Insertion of equation (11) into the equation (10) yields $\alpha = \beta$. Taking into account equation (12) and noting that $\gamma_1^2/2 = \kappa = (\theta_0-\theta_1)^2/(2\,\eta^2)$ the expected loss can be rewritten to $\mathcal{L}(n) = \delta\,[\,n + (N-2\,n)\,\Phi(-c\,\sqrt{n})\,]$ where $c^2 = \kappa/2$. Differentiation with respect to n leads to the necessary condition

$$\frac{2\,\Phi(c\,n^{1/2})-1}{c\,n^{1/2}\,\varphi(c\,n^{1/2})} \overset{!}{=} \frac{N}{2\,n} - 1\,. \qquad (14)$$

Suppose that $c\,N$ is small. Consequently, $c\,n^{1/2}$ is even much smaller and a series expansion of the l.h.s. of the equation above yields $2 + 2\,n\,c^2/3 = N/(2\,n) - 1$ and finally

$$n^* = \frac{6\,N}{(18^2 + 48\,c^2\,N)^{1/2} + 18} \longrightarrow \frac{N}{6}$$

as $c \to 0$, or equivalently, as $\kappa \to 0$. Since n^* increases as c decreases the optimally adjusted test never uses more than $N/6$ paired trials until the final decision is being made. But notice that $\alpha^* = \beta^* \to 1/2$ as $n^* \to N/6$.

To investigate the general dependence of n^* from c (or κ) and N a normalization of the constants and variables is useful. Let $\tilde{n} = c^2\,n$ and $K = N\kappa$ where $c^2 = \kappa/2$. The optimality criterion (14) changes to

$$4\,\tilde{n}\,\left[\frac{2\,\Phi(\tilde{n}^{1/2})-1}{\tilde{n}^{1/2}\,\varphi(\tilde{n}^{1/2})} + 1\right] \overset{!}{=} K \qquad (15)$$

Evidently, the root $\tilde{n}^*$ of equation (15) only depends on the constant K. For large K the root is approximately located at

$$\tilde{n}^* \approx 2\,\log\left(\frac{K}{4\,\log K}\right) \qquad (16)$$

and for small K at

$$\tilde{n}^* \approx \frac{3\,K}{(18^2 + 24\,K)^{1/2} + 18}\,. \qquad (17)$$

The approximations given in equations (16) and (17) are quite accurate for $K > 10^3$ resp. $K < 1$. Nevertheless, a comparison with the results for the SPRT given in the next subsection will be based on the exact values.

4.2 Optimal Sequential Test

If we use the sequential probability ratio test in lieu of the optimal test with fixed sample size, the expected loss can be derived as follows: If hypothesis H_0 is true the test stops on average after n_0 steps. Thus, the observation costs are $n_0\,\delta$. Since the test errs with probability α the decision loss is $\alpha\,(N-n_0)\,\delta$. Similarly, if H_1 is true the observation costs are $n_1\,\delta$ and the decision loss is $\beta\,(N-n_1)\,\delta$. Since each event is equally likely the entire expected loss is

$$\mathcal{L}(\alpha,\beta) = \delta\left[\frac{n_0+n_1}{2} + \frac{\alpha+\beta}{2}\,N - \alpha\,n_0 - \beta\,n_1\right]\,.$$

Since $\kappa := \kappa_1 = -\kappa_0 = (\theta_0-\theta_1)^2/(2\,\eta^2)$ and each hypothesis is equally likely nothing can be gained from choosing different error probabilities. Thus, we may set $\alpha = \beta$ and the expected loss becomes

$$\mathcal{L}(\alpha) = \delta\,[\,n + \alpha\,(N-2\,n)\,]$$

where $n = n_0 = n_1$. Assuming that the excess over the boundaries A or B is neglectable, we may take inequalities (3) and (4) like equalities so that the expected loss is given by

$$\mathcal{L}(\alpha) = \delta\,\left[\alpha\,N + \frac{(1-2\,\alpha)^2}{\kappa}\,\log\left(\frac{1}{\alpha}-1\right)\right]\,. \qquad (18)$$

Differentiation with respect to α leads to the necessary condition

$$N\kappa \overset{!}{=} \frac{(1-2\,\alpha)^2}{\alpha\,(1-\alpha)} + 4\,(1-2\,\alpha)\,\log\left(\frac{1}{\alpha}-1\right)\,. \qquad (19)$$

If $\kappa \to 0$ for fixed N the optimal error probability α^* converges monotonically increasing to $1/2$. Now assume that $N \gg \kappa^{-1}$. In this case α must be small and we may choose $\alpha = 1/(N\kappa - 4\,\log(N\kappa) + 3)$ to obtain an asymptotical solution of the optimality equation (19). This approximation is quite accurate for $K = N\kappa \ge 10^2$.

The appropriate quantity to compare the SPRT with the fixed size sample test (FSST) is the ratio of expected losses in dependence from K. Figure 1 reveals that the FSST and SPRT perform equally well (or bad) for small K, but as soon as K increases the FSST is beaten by the SPRT. It can be shown that the ratio converges monotonically increasing to 4 as $K \to \infty$. As a consequence, the claim (Holland 1975, p. 84) that the paired sampling strategy with the optimal FSST leads (asymptotically) to the minimal expected loss is untenable.

5 Decisions Based on Ranks

In the precedent section the players used statistical tests basing on the actual realizations of the random variables. Here, we shall investigate the situation if the feedback from the gambling machine reduces to

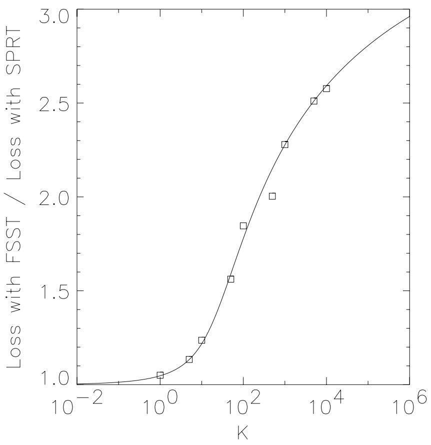

Figure 1: Ratio of minimal losses between fixed size sample test (FSST) and sequential probability ratio test (SPRT). The squares indicate the results from simulations with the SPRT.

the indication which of the two arms was better. The payoffs, however, are still the realizations of the random variables which is not observable by the players until the Nth trial has been done. Without any doubt, the task to maximize the total reward must be more difficult than the task considered previously. Notice that this scenario reflects the situation of using a (binary) tournament selection method in evolutionary algorithms. The question is: What are the additional costs of the loss of information?

Let X and Y be the random variables that are not directly observable by the players. When the players pull arm X at the first and arm Y at the second bandit, each bandit returns a 0 or 1, where a 1 indicates which bandit has realized the smaller payoff. Since the ranking of the outcomes of two continuous random variables is unique, it suffices to observe the sequence of zeros and ones associated with the first bandit. Thus, the players observe the random indicator variable $Z = 1_{\{X<Y\}}$ of the event $\{X < Y\}$.

Let $p = \mathsf{P}\{\, Z = 1 \,\}$, $p_0 = \mathsf{P}\{\, Z = 1 \,|\, \mu_x < \mu_y \,\} > 1/2$ and suppose that the players know the value of p_0. As a consequence, they may use a statistical procedure to test the hypothesis $H_0 : p = p_0$ against hypothesis $H_1 : p = p_1 = 1 - p_0$. If H_0 is rejected they will choose arm X at both bandits for the remaining trials, otherwise arm Y. The SPRT for this task runs as follows:

Since $\mathsf{P}_0\{\, Z = k \,\} = p_0^k\,(1-p_0)^{1-k}$ and $\mathsf{P}_1\{\, Z = k \,\} = (1-p_0)^k\, p_0^{1-k}$ the likelihood ratio function is

$$\lambda_n = \prod_{i=1}^{n} \frac{(1-p_0)^{Z_i}\, p_0^{1-Z_i}}{p_0^{Z_i}\,(1-p_0)^{1-Z_i}} = \left(\frac{p_0}{1-p_0}\right)^{n-2\,S_n}$$

where $S_n = \sum_{i=1}^{n} Z_i$. As in the previous section, nothing can be gained from choosing different error probabilities. Thus, we may set $\alpha = \beta$ and $1/A = B = \alpha/(1-\alpha)$. The expected stopping time is

$$n = \mathsf{E}_0[\,T\,] = \mathsf{E}_1[\,T\,] = \frac{1-2\,\alpha}{\kappa}\,\log\frac{1-\alpha}{\alpha}$$

where

$$\kappa = (2\,p_0 - 1)\,\log\frac{p_0}{1-p_0}\,. \qquad (20)$$

Notice that the expected loss and the optimality criterion are identical to equations (18) and (19), respectively. Only the values for the Kullback–Leibler information numbers are different. Suppose that X and Y are normal random variables as in the previous section. In this case we obtain

$$p_0 = \Phi\left(\frac{\mu_0 - \mu_1}{(\sigma_0^2 + \sigma_1^2)^{1/2}}\right) = \Phi\left(\sqrt{\frac{\tilde\kappa}{2}}\right) > \frac{1}{2}$$

where $\tilde\kappa$ denotes the Kullback–Leibler information number of the previous section. A closer look at equation (20) reveals that $\kappa = \kappa(\tilde\kappa)$ is a function of $\tilde\kappa$ and that

$$\frac{\tilde\kappa}{\kappa(\tilde\kappa)} \to \begin{cases} 4 & \text{as } \tilde\kappa \to \infty, \\ \pi/2 & \text{as } \tilde\kappa \to 0. \end{cases}$$

Figure 2 shows the general behavior of the ratio above. Thus, the additional information costs for a prescribed error probability α consists of 57 % up to 300 % more paired observations.

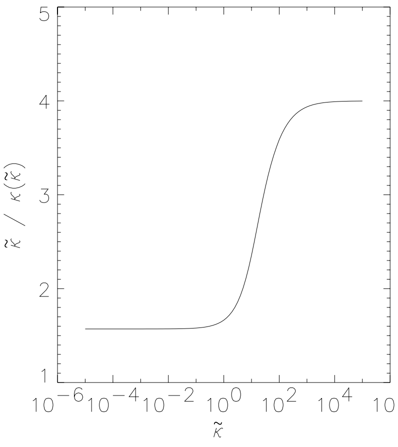

Figure 2: Comparison of Kullback–Leibler information numbers.

6 Evolutionary Decisions Based on Ranking Information

Now we are in the position to compare the quality of selection methods used in evolutionary algorithms with the optimal methods by interpreting the selection procedures as an ingredient of a statistical test. The scenario is as follows:

An individual is of type x if it pulls arm X, and of type y if it pulls arm Y. Random variation operators like mutation or crossover are not considered at the moment. The population is initialized by distributing the types x and y uniformly among the individuals. Therefore, the population size ν must be an even number. In each generation each individual pulls the arm associated with its type and the outcome of the random variable is the fitness of the individual. Then some selection method is used to form the population of the next generation. The population is said to be converged if all individuals are of the same type. The objective is to maximize the cumulated fitness values within $N < \infty$ pulls in total. Thus, the maximum number of trials must be a multiple of the population size.

The simplest case is given for a population of two individuals. Initially, we have one individual of type x and one individual of type y. As soon as the selection methods produces an identical population at random time T, the population is converged and the remaining trials are allocated to the same arm. Since the selection methods' behavior only depend on the outcome of the random variables and not on the type of the individuals, we can assume w.l.o.g. that $\mu_x > \mu_y$. Until the population is converged the expected loss is $\delta\, n$ where $\delta = \mu_x - \mu_y$ and $n = \mathsf{E}[\,T\,]$. If the population is converged to type y the selection procedure has committed an error with probability α and $N - 2\,n$ trials are allocated to the inferior arm. Thus, the expected loss is $\mathcal{L} = \delta\,[\,n + \alpha\,(N - 2\,n)]$.

In order to determine the values for α and n, we shall use a finite Markov chain model. For $\nu = 2$ individuals the Markov chain can attain the three states $S \in \{0, 1, 2\}$, where S denotes the number of individuals of type x. Thus, the Markov chain will be absorbed by the states 0 and 2 (identical populations) and it starts at state 1. Let q be the probability to transition from state 1 to state 0, p the probability to transition from state 1 to 2, and $r = 1 - p - q$ the probability to stay at state 1. Then the expected time until absorption is $\mathsf{E}[\,T\,] = 1/(p+q)$ while the absorption probabilities are $\mathsf{P}\{\,S_T = 0 \,|\, S_0 = 1\} = q/(p+q)$ and $\mathsf{P}\{\,S_T = 2 \,|\, S_0 = 1\} = p/(p+q)$. Evidently, the error probability is $\alpha = \mathsf{P}\{\,S_T = 0 \,|\, S_0 = 1\}$.

In the sequel we shall consider tournament selection rules. Let $\varepsilon = \mathsf{P}\{\,X > Y\,\} > 1/2$. If we employ binary tournament selection *without* replacement the selection procedure picks out both individuals and chooses the individual with higher fitness. This process is repeated once to obtain a complete new population. Therefore we obtain $p = \varepsilon$, $r = 0$, and $q = 1 - \varepsilon$. As a result, the expected loss is $\mathcal{L} = \delta\,[\,(1-\varepsilon)\,N + 2\,\varepsilon - 1\,]$.

In case of binary tournament selection *with* replacement the selection procedure draws two individuals at random, so that it is possible to draw the same individual twice. We obtain $p = \varepsilon^2$, $r = 2\,\varepsilon\,(1-\varepsilon)$, $q = (1-\varepsilon)^2$, and finally

$$\mathcal{L} = \delta\,\frac{(1-\varepsilon)^2\,(1 - 2\,\varepsilon + 2\,\varepsilon^2)^3\,N + 2\,\varepsilon - 1}{(1 - 2\,\varepsilon + 2\,\varepsilon^2)^2}$$

which is less than the loss of binary tournament selection *without* replacement for $N \geq 6$ and $\varepsilon \in (1/2, 1)$, but tremendously worse than the optimal methods.

The reason for the unsatisfactory behavior stems from the fact, that the transition probabilities (p, q, r) and hence the error probability α as well as the expected stopping time n depends only on ε and not on the maximum sample size N. We could achieve such a dependence if the maximum sample size N (and ε) become control parameters of the selection procedure. In case of selection methods based on ranking information, the optimal procedure would be the SPRT considered in the previous section.

But there is another way to improve the situation. Notice that the SPRT may be interpreted as follows: The individuals pull different arms and store the number of competitions won. As soon as one individual has won a certain prescribed number of competitions (depending on N) it is considered superior to the other individual which will not have a chance to reproduce again. In evolutionary algorithms the functionality of a memory is not provided by the individuals but by the population itself. Thus, we have to increase the population size ν. To keep the analysis simple, we shall consider a population arranged in a spatial structure. Suppose that the individuals are placed at the nodes of a degenerated undirected graph (i.e., a bi–directional list). Each individual can compete only with those individuals which are its direct neighbors. With probability $1/2$ it will compete with its left or right neighbor. Initially, the first $\nu/2$ individuals from left to right are of type x and the rest of type y. The Markov chain describing the evolution of the population has $\nu{+}1$ states $S \in \{0, \ldots, \nu\}$. Again, state S denotes the number of x's in the population, so that $S_0 = \nu/2 = m$. The Markov chain changes its state if an individual of type x is replaced by an individual of type y or vice versa. Due to our special initialization of the population such an event only happens once during each generation. Thus, the probability that the number of x's will be incremented is p, the probability of a decrement is q and the probability to keep the status quo is r. Here, $p = \varepsilon/2$, $r = 1/2$, and $q = (1-\varepsilon)/2$. Since $p > q$ this Markov chain is nothing more than a random walk with drift. The absorption probability to state 0 (i.e., the error probability) is

$$\alpha_m = \mathsf{P}\{\,S_T = 0 \,|\, S_0 = m\} = \frac{1}{a^m + 1}$$

($a = p/q$) while the expected absorption time is

$$n(m) = \mathsf{E}[\, T \mid S_0 = m \,] = \frac{m}{p-q} \cdot \frac{a^m - 1}{a^m + 1} \, .$$

The total number of pulls at the inferior arms depends on the random number of times V_i, that the Markov chain is in state $i = 1, \ldots, \nu - 1$. The expectation of V_i is

$$\mathsf{E}[\, V_i \mid S_0 = m \,] = \begin{cases} \dfrac{a^i - 1}{(a^m + 1)\,(p - q)} & \text{, if } i = 1, \ldots, m \\[2ex] \dfrac{a^m - a^{i-m}}{(a^m + 1)\,(p - q)} & \text{, otherwise.} \end{cases}$$

At position i there are $(\nu - i)$ pulls at the inferior arm. This happens $\mathsf{E}[\, V_i \mid S_0 = m \,]$ times on average. Thus, the average total number of pulls at the inferior arm until absorption is

$$W_m = \sum_{i=1}^{\nu-1} \mathsf{E}[\, V_i \mid S_0 = m \,]\,(\nu - i) =$$

$$\frac{m}{2\,(p-q)} \left[\frac{a+1}{a-1}\,\frac{a^m - 1}{a^m + 1} + m\,\frac{a^m - 3}{a^m + 1} \right]$$

so that the expected loss is

$$\mathcal{L}(m) = \delta \, [\, W_m + \alpha_m \, (N - 2\, m\, n(m)) \,]$$

Since $\nu = 2\,m$ we can control the error probability α_m and all other quantities by the choice of the population size. It is clear that the optimal population size ν^* will depend on the difference $p - q$, the ratio $a = p/q$ *and* the maximum sample number N.

Figure 3 shows a comparison of the different selection strategies by the ratio of the loss of each method with loss of the optimal SPRT. Since the maximum sample size was kept fixed at $N = 10^6$ the Kullback–Leibler information number $\tilde{\kappa}$ of section 4 ranges between 10^{-10} and 10^0. Two points deserve special attention: First, the SPRT based on ranking information leads to a better strategy than using the FSST based on the p.d.f.s of the differences of both random variables. Second, the optimal population size is remarkably small ($\nu^* = 84$) even for the difficult problems with $\tilde{\kappa} = 10^{-8}$. One might conjecture that the performance of the local tournament methods improves under random initialization. The analysis of this case, however, is considerably more complicated than the initialization considered here.

7 Conclusions

Holland's claim that the fixed size sample test leads to an asymptotically optimal strategy was falsified. A strategy using the sequential probability ratio test performs up to four times better. Local tournament selection methods in evolutionary algorithms may be interpreted as an ingredient of a sequential statistical test. Their performance is acceptable well provided that the population size is optimally adjusted.

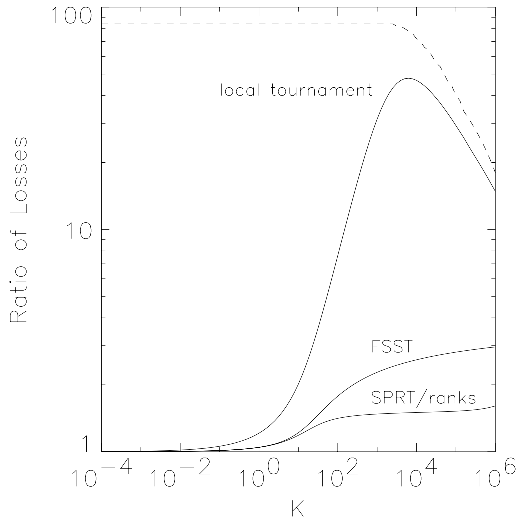

Figure 3: Ratios of losses of the fixed size sample test (FSST), sequential probability ratio test (SPRT) based on ranking information, and local binary tournament selection with the optimal SPRT. The dashed line indicates the optimal population size.

Acknowledgments

The work presented is a result of the *Collaborative Research Center SFB 531* sponsored by the Deutsche Forschungsgesellschaft (DFG).

References

Chernoff, H. (1959). Sequential design of experiments. *Ann. Math. Stat. 30*, 755–770.

Gut, A. (1988). *Stopped Random Walks.* New York: Springer.

Holland, J. H. (1975). *Adaptation in natural and artificial systems.* Ann Arbor: The University of Michigan Press.

Kiefer, J. and J. Sacks (1963). Asymptotically optimal sequential inference and design. *Ann. Math. Stat. 34*, 705–750.

Macready, W. G. and D. H. Wolpert (1996). On 2–armed gaussian bandits and optimization. Technical Report SFI–TR–96–03–009, Santa Fe Institute, Santa Fe (NM).

Neyman, J. and E. S. Pearson (1933). On the problem of the most efficient tests of statistical hypotheses. *Philos. Trans. Roy. Soc. London, Ser. A 231*, 289–337.

Wald, A. (1945). Sequential tests of statistical hypotheses. *Ann. Math. Stat. 16*, 117–186.

Wald, A. and J. Wolfowitz (1948). Optimum character of the sequential probability ratio test. *Ann. Math. Stat. 19*, 326–339.