On the Use of Evolution Strategies for Optimising Certain Positive Definite Quadratic Forms

Dirk V. Arnold
Faculty of Computer Science
Dalhousie University
Halifax, Nova Scotia
Canada B3H 1W5
dirk@cs.dal.ca

ABSTRACT

This paper studies the performance of multi-recombinative evolution strategies using isotropically distributed mutations on a class of convex quadratic objective functions that is characterised by the presence of only two different eigenvalues of their Hessian. A simplified model of the strategy's behaviour is developed. Using it, expressions that approximately describe the stationary state that is attained when the mutation strength is adapted are derived. The performance achieved when using cumulative step length adaptation is compared with that obtained when using optimally adapted step lengths.

Categories and Subject Descriptors

G.1.6 [**Optimization**]: Unconstrained Optimization; I.2.8 [**Problem Solving, Control Methods, and Search**]; I.2.6 [**Learning**]: Parameter Learning

General Terms

Algorithms, Performance, Theory

Keywords

Evolution strategy, cumulative step length adaptation, positive definite quadratic form

1. INTRODUCTION

Analyses of the performance of evolutionary algorithms (EAs) on selected objective functions serve the purposes of highlighting differences as well as strengths and weaknesses of strategy variants, of deriving recommendations with regard to the setting of strategy parameters, and of gaining insights that may be of use when developing adaptation strategies. In the realm of continuous (i.e., real-valued) evolutionary optimisation, such analyses have predominantly focused on classes of objective functions that are both amenable to mathematical analysis and suitable for revealing important aspects of the algorithms being studied. Examples include the sphere and corridor models, the ridge function class, and various noisy and time varying variants thereof [1, 4, 12].

The sphere and parabolic ridge models are at opposite ends of the spectrum of test functions in that while the condition number (i.e., the ratio of largest to smallest eigenvalue) of the Hessian matrix of the former is one (and thus minimal), that of the latter is infinite. As real-world optimisation problems exhibit various degrees of ill-conditioning, it is desirable to extend analyses of the behaviour of EAs to problems with condition numbers between those two extremes. A class of test functions that offers the opportunity to consider varying condition numbers while retaining mathematical tractability is the class of positive definite quadratic forms (PDQFs). A number of PDQFs have been used extensively in empirical investigations of the behaviour of EAs [7, 13]. However, there have been few attempts to study properties of EAs optimising PDQFs other than the sphere model analytically.

Among the exceptions is an investigation of the steady state of evolution strategies optimising general ellipsoidal objective functions disturbed by noise [5] as well as two important recent papers by Jägersküpper [9, 10] that study the performance of the $(1+1)$-ES on PDQFs of bounded bandwidth (i.e., with condition number in $\mathcal{O}(1)$) as well as on a particular class of ill-conditioned PDQFs. Jägersküpper derives the following results:

- For a PDQF of bounded bandwidth in $\mathbb{R}^N$, the number of steps needed to reduce the approximation error to a 2^{-b}-fraction ($b \geq 1$ polynomial in N) is $\Omega(bN)$ both in expectation and with overwhelming probability, independently of how the mutation strength is adapted. If the 1/5th rule is used (and the mutation strength is initialised appropriately), then the number of steps is $\mathcal{O}(bN)$ with overwhelming probability.
- For PDQFs with only two different eigenvalues of their Hessian, both occurring in equal proportions, and with condition number ξ polynomially bounded in N such that $1/\xi \to 0$ as $N \to \infty$, if the mutation strength is initialised appropriately, then the number of steps needed to reduce the initial approximation error to a 2^{-b} fraction ($b \geq 1$ polynomial in N) is $\Theta(b\xi N)$ with overwhelming probability.

Loosely speaking, the first of the two results states that for quadratic functions "sufficiently close" to the sphere model,

GECCO'07, July 7–11, 2007, London, England, United Kingdom.

the $(1+1)$-ES can at best achieve linear convergence, with the speed of convergence inversely proportional to the dimensionality N of the search space. If the mutation strength is adapted using the 1/5th rule, the performance differs from optimal performance by no more than a constant factor. The behaviour is thus asymptotically the same as on the sphere model [8]. The second result states that on the particular class of ill-conditioned PDQFs considered, for large condition numbers the speed of convergence decreases proportionally with the degree of ill-conditioning.

Jägersküpper's results provide a valuable, rigorously derived understanding of important aspects of search properties of the $(1+1)$-ES on both mildly and certain severely ill-conditioned PDQFs. However, they are qualitative in that the asymptotic notation hides constants. For example, while a PDQF with condition number 10 is "close to the sphere" in an asymptotic sense, to a user who applies an EA to a 40-dimensional optimisation problem with that condition number, it is significant what fraction of the performance on the sphere model is obtained. It is also of interest to study population-based rather than point-based strategy variants, and to compare different step length adaptation mechanisms with regard to what fraction of the optimal performance they are able to achieve.

This paper takes a step in direction of an understanding of population-based EAs in ill-conditioned environments by studying the behaviour of an adaptive, multi-recombinative evolution strategy on PDQFs of the form

$$f(\mathbf{x}) = \xi \sum_{i=1}^{N\vartheta} x_i^2 + \sum_{i=N\vartheta+1}^{N} x_i^2 \tag{1}$$

where $\mathbf{x} = \langle x_1, \ldots, x_N \rangle \in \mathbb{R}^N$ and where $\vartheta \in [0, 1]$ is such that $N\vartheta$ is integer. Rather than employing asymptotic notation and attempting to derive results rigorously, a simplified dynamic model of the optimisation process is studied. The simplifications assume that N and ϑ are such that both $N\vartheta$ and $N(1-\vartheta)$ are large. Computer experiments are used to evaluate the accuracy of the approximations. The approach makes it possible to determine (approximately) optimal parameter settings, and to reveal information about constants that are hidden in the asymptotic notation employed in the rigorous approach.

For symmetry reasons, it can be assumed without loss of generality that $\xi \geq 1$. Clearly, the parameter ξ is the condition number of the Hessian matrix of the function. It is important to note that while in the formulation in Eq. (1) the coordinate basis coincides with the principal axes of the Hessian and the function is thus separable, this does not constitute a limitation as both mutation and recombination operators of the strategy considered here are isotropic. The function could be subjected to an arbitrary rotation (and thus made non-separable) without influencing the results.

Several test functions that are frequently used for (empirically) evaluating EAs occur as special cases of Eq. (1). For $\vartheta \in \{0, 1\}$ or for $\xi = 1$, the sphere model is obtained. For $\vartheta = (N-1)/N$, Eq. (1) becomes the cigar function; for $\vartheta = 1/N$, it is the discus (or tablet) function. Neither of the latter cases is included in the discussion here due to the assumption that both $N\vartheta$ and $N(1-\vartheta)$ are large. The special case that $\vartheta = 0.5$ is referred to by Hansen and Ostermeier [7] as the two-axes function and is the ill-conditioned case considered by Jägersküpper [9, 10].

The remainder of the paper is organised as follows. Section 2 gives a brief description of the $(\mu/\mu, \lambda)$-CSA-ES. In Section 3, the symmetries of the class of PDQFs considered are exploited to describe the behaviour of the strategy using a small number of state variables. Stochastic evolution equations are derived that describe the dynamics of those variables for single time steps. Section 4 introduces several simplifications that make it possible to obtain an analytical solution to the evolution equations. Computer experiments are used to evaluate the accuracy of the predictions. Section 5 is devoted to the analysis of the step length adaptation mechanism. Section 6 concludes with a brief discussion of the results and their significance.

2. STRATEGY

The $(\mu/\mu, \lambda)$-CSA-ES is an evolution strategy for the optimisation of functions $f : \mathbb{R}^N \to \mathbb{R}$ that uses the cumulative step length adaptation mechanism proposed by Ostermeier et al. [11] for the control of its mutation strength. In every time step the strategy computes the centroid of the population of candidate solutions as a search point $\mathbf{x} \in \mathbb{R}^N$ that mutations are applied to. A vector $\mathbf{s} \in \mathbb{R}^N$ that is referred to as the search path is used to accumulate information about the directions of the most recently taken steps. An iteration of the $(\mu/\mu, \lambda)$-CSA-ES updates the search point along with the search path and the mutation strength of the strategy in five steps:

1. Generate λ offspring candidate solutions $\mathbf{y}^{(i)} = \mathbf{x} + \sigma\mathbf{z}^{(i)}$, $i = 1, \ldots, \lambda$, where mutation strength $\sigma > 0$ determines the step length and the $\mathbf{z}^{(i)}$ are mutation vectors consisting of N independent, standard normally distributed components.

2. Determine the objective function values $f(\mathbf{y}^{(i)})$ of the offspring candidate solutions and compute the average

$$\mathbf{z}^{(\mathrm{avg})} = \frac{1}{\mu} \sum_{k=1}^{\mu} \mathbf{z}^{(k;\lambda)} \tag{2}$$

of the μ best of the $\mathbf{z}^{(i)}$. The index $k;\lambda$ refers to the kth best of the λ offspring candidate solutions. Vector $\mathbf{z}^{(\mathrm{avg})}$ is referred to as the progress vector.

3. Update the search point according to

$$\mathbf{x} \leftarrow \mathbf{x} + \sigma\mathbf{z}^{(\mathrm{avg})}. \tag{3}$$

4. Update the search path according to

$$\mathbf{s} \leftarrow (1-c)\mathbf{s} + \sqrt{\mu c(2-c)}\mathbf{z}^{(\mathrm{avg})} \tag{4}$$

where the cumulation parameter c is set to $1/\sqrt{N}$.

5. Update the mutation strength according to

$$\sigma \leftarrow \sigma \exp\left(\frac{\|\mathbf{s}\|^2 - N}{2DN}\right) \tag{5}$$

where damping parameter D is set to $\sqrt{N}$.

See [6] for a more comprehensive discussion of evolution strategies and their parameters.

The behaviour of the $(\mu/\mu, \lambda)$-CSA-ES on the class of PDQFs defined in Eq. (1) is illustrated in Fig. 1. While a single run of the strategy is shown, the same qualitative

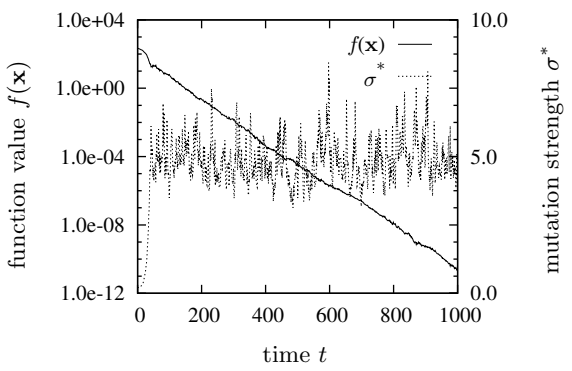

Figure 1: Objective function value $f(\mathbf{x})$ of the search point and normalised mutation strength σ^* plotted against time t for a typical run of the $(3/3, 10)$-CSA-ES ($N = 40$, $\xi = 10$, $\vartheta = 0.5$).

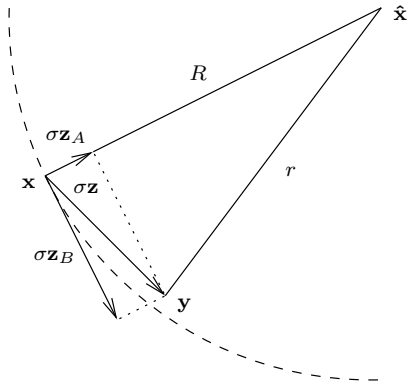

Figure 2: *Decomposition of a vector $\mathbf{z}$ into components $\mathbf{z}_A$ and $\mathbf{z}_B$. Vector $\mathbf{z}_A$ is parallel to $\hat{\mathbf{x}} - \mathbf{x}$, vector $\mathbf{z}_B$ is in the $(N-1)$-dimensional hyperplane perpendicular to that.*

behaviour can be observed in other runs as well as for other settings of the parameters. Initially, the strategy is in a transitory period (that lasts for about 40 time steps in the figure). The length of that period increases with increasing values of N and ξ, and during it, the behaviour of the strategy depends on how $\mathbf{x}$ and σ are initialised. Afterwards, the strategy enters a phase that is stationary in that the average relative progress becomes constant (i.e., the logarithm of the objective function value of the search point decreases linearly). The steeper the slope of a regression line fitted to the graph of the logarithmic function values, the faster the progress toward the optimal solution. That slope is

$$\Delta = \mathrm{E}\left[-\log\left(\frac{f(\mathbf{x}^{(t+1)})}{f(\mathbf{x}^{(t)})}\right)\right] \tag{6}$$

where superscripts indicate time, and is referred to as the quality gain of the strategy. It is the purpose of the following sections to compute an approximation to that quality gain.

3. DYNAMIC SYSTEM

The following discussion builds extensively on techniques previously used in connection with the sphere model [1, 4, 12]. The analysis of the behaviour of evolution strategies on the sphere model relies on a decomposition of vectors that is illustrated in Fig. 2. Consider points $\mathbf{x}$ and $\mathbf{y} = \mathbf{x} + \sigma\mathbf{z}$ at distances R and r from the centre $\hat{\mathbf{x}}$ of the sphere, respectively. Let $z_A = \mathbf{z} \cdot (\hat{\mathbf{x}} - \mathbf{x})/R$ denote the signed length of the component of $\mathbf{z}$ in direction of the centre of the sphere. Decompose vector $\mathbf{z}$ into vectors $\mathbf{z}_A = z_A(\hat{\mathbf{x}} - \mathbf{x})/R$ and $\mathbf{z}_B = \mathbf{z} - \mathbf{z}_A$. It follows using elementary geometry that

$$\begin{aligned} r^2 &= (R - \sigma z_A)^2 + \sigma^2 \|\mathbf{z}_B\|^2 \\ &= R^2 - 2R\sigma z_A + \sigma^2 \|\mathbf{z}\|^2. \end{aligned} \tag{7}$$

Moreover, if $\mathbf{z}$ is a mutation vector then z_A is standard normally distributed and $\|\mathbf{z}\|^2$ is χ^2_N-distributed.

The PDQF defined in Eq. (1) is the weighted sum of two independent, spherically symmetric functions. Let $\mathbf{x} \in \mathbb{R}^N$ be the search point of the strategy and write

$$R_1^2 = \sum_{i=1}^{N\vartheta} x_i^2 \quad \text{and} \quad R_2^2 = \sum_{i=N\vartheta+1}^{N} x_i^2$$

for the squared distances from the centres of the two spheres, yielding

$$f(\mathbf{x}) = \xi R_1^2 + R_2^2.$$

Let $\mathbf{z}_1 = \langle z_1, \ldots, z_{N\vartheta} \rangle \in \mathbb{R}^{N\vartheta}$ consist of the first $N\vartheta$ components of vector $\mathbf{z} \in \mathbb{R}^N$ and let $\mathbf{z}_2 = \langle z_{N\vartheta+1}, \ldots, z_N \rangle \in \mathbb{R}^{N(1-\vartheta)}$ consist of the remaining $N(1-\vartheta)$ components. Then for $\mathbf{y} = \mathbf{x} + \sigma\mathbf{z}$ the respective squared distances

$$r_1^2 = \sum_{i=1}^{N\vartheta} y_i^2 \quad \text{and} \quad r_2^2 = \sum_{i=N\vartheta+1}^{N} y_i^2$$

are in immediate analogy to Eq. (7)

$$r_1^2 = R_1^2 - 2R_1\sigma z_{A1} + \sigma^2 \|\mathbf{z}_1\|^2 \tag{8}$$

$$r_2^2 = R_2^2 - 2R_2\sigma z_{A2} + \sigma^2 \|\mathbf{z}_2\|^2 \tag{9}$$

where z_{A1} and z_{A2} are standard normally distributed, $\|\mathbf{z}_1\|^2$ is $\chi^2_{N\vartheta}$-distributed, and $\|\mathbf{z}_2\|^2$ is $\chi^2_{N(1-\vartheta)}$-distributed. Furthermore, using Eqs. (8) and (9), the objective function value of $\mathbf{y}$ is

$$\begin{aligned} f(\mathbf{y}) &= \xi r_1^2 + r_2^2 \\ &= f(\mathbf{x}) - 2R_1\xi\sigma z_{A1} - 2R_2\sigma z_{A2} + \xi\sigma^2\|\mathbf{z}_1\|^2 + \sigma^2\|\mathbf{z}_2\|^2. \end{aligned}$$

Selection picks those μ of the λ mutation vectors $\mathbf{z}$ that yield the smallest values of $f(\mathbf{x} + \sigma\mathbf{z})$. Recombination according to Eq. (2) averages the selected mutation vectors. With the progress vector $\mathbf{z}^{(\mathrm{avg})}$ taking the place of $\mathbf{z}$, Eqs. (8) and (9) describe the evolution of the state of the strategy (without step length adaptation) from one time step to the next. It thus remains to compute $z_{A1}^{(\mathrm{avg})}$, $z_{A2}^{(\mathrm{avg})}$, $\|\mathbf{z}_1^{(\mathrm{avg})}\|^2$, and $\|\mathbf{z}_2^{(\mathrm{avg})}\|^2$.

4. STATIC PERFORMANCE

The stochastic dynamic system described in the previous section does not allow for an exact analytical solution. For the sphere model, an approximate solution can be obtained by making several simplifications that in essence rely on the assumption that the search space dimensionality N is very high [4, 12]. For the PDQF in Eq. (1) it is assumed that both

$N\vartheta$ and $N(1-\vartheta)$ are large, making it possible to use the simplifications for the sphere model for both of the spherically symmetric functions that form the objective.

For the range of mutation strengths that afford positive quality gain, the variance of the normally distributed terms in Eqs. (8) and (9) for growing N increasingly outweighs that of the χ^2-distributed ones. For high search space dimensionality, it is thus possible to replace the χ^2-distributed random variables with their mean values, yielding

$$r_1^2 = R_1^2 - 2R_1\sigma z_{A1} + N\vartheta\sigma^2$$
$$r_2^2 = R_2^2 - 2R_2\sigma z_{A2} + N(1-\vartheta)\sigma^2$$

as simplified evolution equations. Similarly, the objective function value of offspring candidate solution $\mathbf{y} = \mathbf{x} + \sigma\mathbf{z}$ is

$$f(\mathbf{y}) = f(\mathbf{x}) - 2R_1\xi\sigma z_{A1} - 2R_2\sigma z_{A2} + N\vartheta\xi\sigma^2 + N(1-\vartheta)\sigma^2.$$

The first, fourth, and fifth terms on the right hand side are identical for all offspring; the second and third terms are normally distributed. Therefore, as the task is minimisation, the offspring selected to survive are those with the largest values of $R_1\xi z_{A1} + R_2 z_{A2}$. The signed lengths z_{A1} and z_{A2} of the components of the mutation vectors that point in direction of the optimum are concomitants of the order statistics $f(\mathbf{y}^{(k;\lambda)})$ upon which selection is based. While ideally both are large and the values of both impact selection, they are not selected independently. The contribution to the objective function value of one of the two spheres acts as noise impacting selection of the signed length of the component of the mutation vector toward the optimum of the other.

According to Eq. (2), the signed lengths in direction of the optimum of the progress vector are the averages of the respective signed lengths of the selected mutation vectors. The noise-to-signal ratios that impact selection are $R_2/(R_1\xi)$ and $R_1\xi/R_2$, respectively. Using results on expected values of concomitants of order statistics derived in [1, 2] it follows that

$$\mathrm{E}\left[z_{A1}^{(\mathrm{avg})}\right] = \frac{c_{\mu/\mu,\lambda}}{\sqrt{1 + (R_2/(R_1\xi))^2}} \tag{10}$$

and

$$\mathrm{E}\left[z_{A2}^{(\mathrm{avg})}\right] = \frac{c_{\mu/\mu,\lambda}}{\sqrt{1 + (R_1\xi/R_2)^2}} \tag{11}$$

where $c_{\mu/\mu,\lambda}$ is the $(\mu/\mu,\lambda)$-progress coefficient defined in [4] and describes the effect of selection in the absence of noise. Moreover, in analogy to results derived for the sphere model in that same reference,

$$\mathrm{E}\left[\frac{\|\mathbf{z}_1^{(\mathrm{avg})}\|^2}{N\vartheta}\right] = \mathrm{E}\left[\frac{\|\mathbf{z}_2^{(\mathrm{avg})}\|^2}{N(1-\vartheta)}\right] = \frac{1}{\mu} \tag{12}$$

where the factor μ in the denominator reflects the fact that averaging uncorrelated random vectors results in a vector with a length that is reduced compared to the lengths of the vectors being averaged. Eqs. (10), (11), and (12) are exact in the limit $N \to \infty$ if $\vartheta \in (0,1)$ is fixed. They will be seen below to yield approximations provided that both $N\vartheta$ and $N(1-\vartheta)$ are sufficiently large.

Using Eqs. (10), (11), and (12) in Eqs. (8) and (9) makes it possible to obtain the expected behaviour of the search point in a single time step. With normalised mutation strength $\sigma^* = \sigma N\vartheta/R_1$ and location parameter

$$\zeta = \frac{R_2}{R_1\xi}$$

that determines the relative positions of the search point on the two spheres it follows that

$$\mathrm{E}\left[{R_1^{(t+1)}}^2\right] = {R_1^{(t)}}^2\left[1 - \frac{2}{N\vartheta}\left(\frac{\sigma^* c_{\mu/\mu,\lambda}}{\sqrt{1+\zeta^2}} - \frac{{\sigma^*}^2}{2\mu}\right)\right] \tag{13}$$

and

$$\mathrm{E}\left[{R_2^{(t+1)}}^2\right] = {R_2^{(t)}}^2\left[1 - \frac{2}{N\vartheta}\left(\frac{\sigma^* c_{\mu/\mu,\lambda}}{\xi\sqrt{1+\zeta^2}} - \frac{{\sigma^*}^2(1-\vartheta)}{2\mu\vartheta\xi^2\zeta^2}\right)\right] \tag{14}$$

where superscripts indicate time.

If the step length is adapted successfully, then σ and R_1 decrease in equal proportions and consequently, the normalised mutation strength σ^* fluctuates around a stationary average value (compare Fig. 1). The relative magnitude of the variations decreases with increasing N. The same holds for location parameter ζ. Assuming that the normalised mutation strength is constant, an approximation to the steady state value of ζ can be obtained from Eqs. (13) and (14). The terms in the square brackets in those equations must be (approximately) equal as otherwise ζ would have a bias toward either larger or smaller values. Introducing standardised mutation strength $\bar{\sigma} = \sigma^*/(\mu c_{\mu/\mu,\lambda})$, elementary transformations yield stationarity condition

$$2\frac{\xi - 1}{\xi}\zeta^2 = \bar{\sigma}\sqrt{1+\zeta^2}\left(\zeta^2 - \frac{1-\vartheta}{\vartheta\xi^2}\right) \tag{15}$$

that can be solved numerically for the location parameter. Fig. 3 illustrates how ζ depends on the normalised mutation strength σ^*. The lines in the figure have been obtained from Eq. (15). The dots represent values measured in runs of the evolution strategy with constant σ^*.[1] It can be seen that the quality of the approximation is quite good for a wide range of condition numbers ξ. The most significant deviations occur for $\vartheta = 0.25$, where for $N = 40$ the sphere with the higher curvature is only 10-dimensional and variations have a significant impact on measured values. Agreement of the values measured for $N = 400$ with theoretical predictions is generally good. It can also be seen that while for larger values of ξ the location parameter is relatively independent of both ϑ and ξ, it (and with it the trajectory of the search point) depends strongly on the value of the normalised mutation strength.

It remains to compute the quality gain achieved by the strategy. It can be seen from Eqs. (13) and (14) that the magnitude of the changes of R_1^2 and R_2^2 in a single time step is inversely proportional to N. (This has been shown rigorously for the case of the $(1+1)$-ES by Jägersküpper [10].) The quality gain of the strategy, too, is thus inversely proportional to the search space dimensionality. For large values of N, the logarithm in Eq. (6) can thus be expanded into a Taylor series with terms beyond the linear one dropped,

[1] Keeping σ^* constant of course requires information typically not available to optimisation algorithms. The experiments only serve the purpose of evaluating the quality of the approximation derived and do not constitute a viable optimisation strategy.

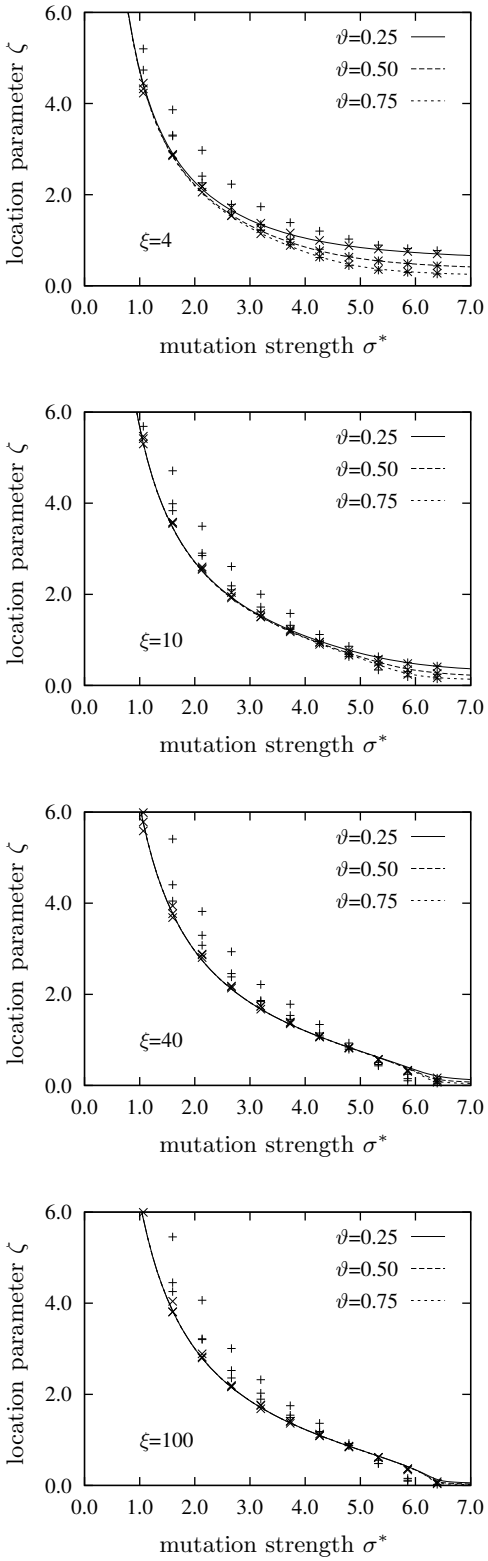

Figure 3: Location parameter ζ of the $(\mu/\mu,\lambda)$-ES plotted against normalised mutation strength σ^* for $\mu = 3$, $\lambda = 10$, $\xi \in \{4, 10, 40, 100\}$, and $\vartheta \in \{0.25, 0.50, 0.75\}$. The dots mark measurements made in runs of the strategy in search spaces with $N = 40$ (+) and $N = 400$ (×).

yielding

$$\begin{aligned}\Delta &= \mathrm{E}\left[1 - \frac{f(\mathbf{x}^{(t+1)})}{f(\mathbf{x}^{(t)})}\right] \\ &= 1 - \frac{\mathrm{E}\left[{R_1^{(t+1)}}^2\right]\xi + \mathrm{E}\left[{R_2^{(t+1)}}^2\right]}{{R_1^{(t)}}^2\xi + {R_2^{(t)}}^2}\end{aligned}$$

where the expected values in the second line can be computed using Eqs. (13) and (14). Taking into account that in the steady state the values in the parentheses in those equations are equal, the squared distances R_1^2 and R_2^2 cancel out. Introducing normalised quality gain $\Delta^* = \Delta N\vartheta/2$ along with standardised quantity $\bar{\Delta} = \Delta^*/(\mu c_{\mu/\mu,\lambda}^2)$ yields

$$\bar{\Delta} = \frac{\bar{\sigma}}{\sqrt{1+\zeta^2}} - \frac{\bar{\sigma}^2}{2}. \tag{16}$$

Fig. 4 illustrates for several values of $\xi > 1$ how the normalised quality gain of the strategy depends on the normalised mutation strength. The lines have been obtained from Eq. (16) with the location parameter computed numerically using Eq. (15). The dots mark measurements made in runs of the strategy with constant σ^*. It can be seen that the quality of the approximation is good for small values of the normalised mutation strength, but that relatively large values of N are required to observe good agreement of measurements and predictions for larger values of σ^*. The inaccuracies for small N stem from replacing χ^2-distributed random variables with their means, from the linearisation of the logarithm in the definition of the quality gain, and from variations in the value of the location parameter.

The results thus obtained allow computing the maximum mutation strength that affords non-negative quality gain. Demanding that $\bar{\Delta} = 0$ in Eq. (16) and using Eq. (15) to solve for the location parameter yields $\zeta^2 = (1-\vartheta)/(\vartheta\xi)$. The resulting standardised mutation strength is

$$\bar{\sigma}_{\max} = \frac{2}{\sqrt{1+(1-\vartheta)/(\vartheta\xi)}}$$

and corresponds to the positive zero of the quality gain in Fig. 4. As $0 < \bar{\sigma}_{\max} < 2$ for any setting of ϑ and ξ, useful values of the normalised mutation strength σ^* are always smaller than $2\mu c_{\mu/\mu,\lambda}$.

Finally, Eqs. (15) and (16) can be used to compute optimal settings of the mutation strength and the corresponding quality gain. For $\xi = 1$, the optimal standardised mutation strength and quality gain are $\bar{\sigma}_{\mathrm{opt}} = \sqrt{\vartheta}$ and $\bar{\Delta}_{\mathrm{opt}} = \vartheta/2$ in agreement with results for the sphere model [4] (notice the differing normalisations). For $\xi > 1$, computing the derivative of Eq. (16) and using Eq. (15) to eliminate the standardised mutation strength results in

$$\begin{aligned}&2\vartheta^2\xi^3\zeta^6 + 2\vartheta(1-\vartheta)\xi(1-2\xi)\zeta^4 \\ &\qquad + 2\vartheta(1-\vartheta)\xi(2-\xi)\zeta^2 - 2(1-\vartheta)^2 = 0.\end{aligned} \tag{17}$$

Solving the equation numerically for ζ and using Eqs. (15) and (16) to compute the standardised mutation strength and quality gain yields the dashed lines in Fig. 6. The asymptotic behaviour for large condition numbers can be determined analytically. For large ξ, solving Eq. (17) yields $\zeta = ((1-\vartheta)/(\vartheta\xi))^{0.25}$. The corresponding standardised mutation strength is $\bar{\sigma}_{\mathrm{opt}} = 2$, resulting in standardised quality gain $\bar{\Delta}_{\mathrm{opt}} = 2/\xi$. The inverse proportionality of the optimal

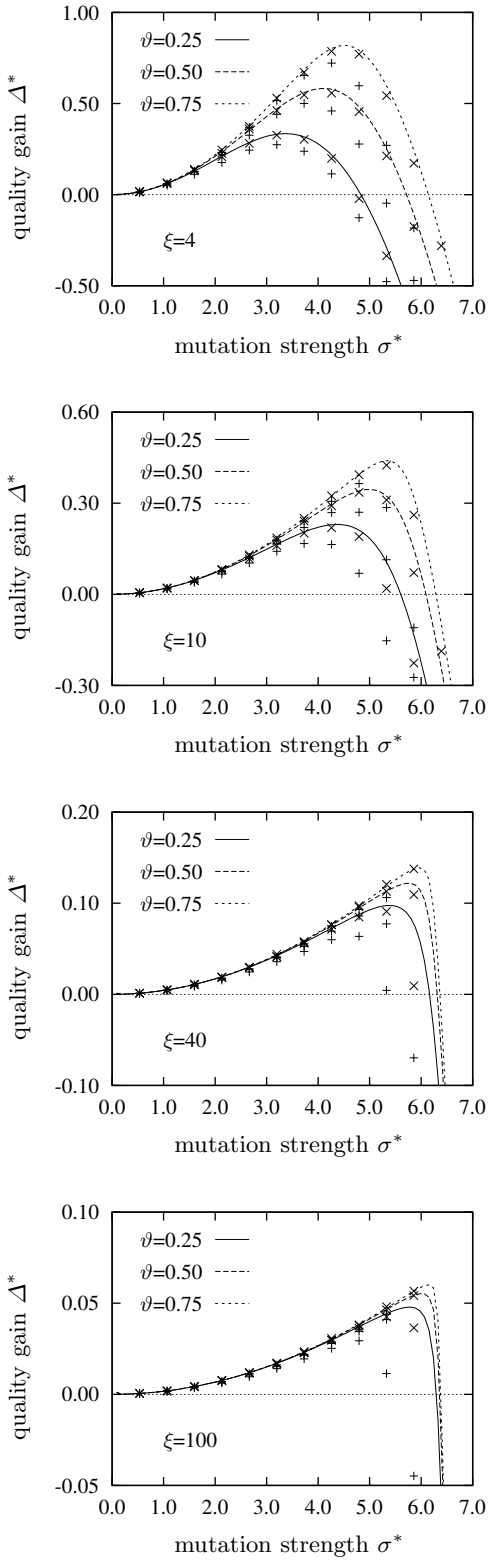

Figure 4: Normalised quality gain Δ^* of the $(\mu/\mu, \lambda)$-ES plotted against normalised mutation strength σ^* for $\mu = 3$, $\lambda = 10$, $\xi \in \{4, 10, 40, 100\}$, and $\vartheta \in \{0.25, 0.50, 0.75\}$. The dots mark measurements made in runs of the strategy in search spaces with $N = 40$ (+) and $N = 400$ (×).

quality gain to the condition number for large values of ξ parallels the corresponding result for the $(1 + 1)$-ES derived by Jägersküpper [10].

5. DYNAMIC PERFORMANCE

In the previous section, it was assumed that σ^* is constant. As can be seen in Fig. 1, successful step length adaptation leads to the normalised mutation strength fluctuating around a stationary average value. This section computes an approximation to that average value assuming that the mutation strength is adapted using cumulative step length adaptation as described in Section 2.

Compared to Section 4, considering cumulative step length adaptation requires introducing several additional state variables. The performance of the $(\mu/\mu, \lambda)$-CSA-ES on the sphere model is analysed in [1, 3]. For that model, the additional state variables are the signed length s_A of the component of the search path that points in direction of the optimum, the squared length $\|\mathbf{s}\|^2$ of the search path, and the normalised mutation strength σ^*. The approach to the analysis is the same as that employed in Section 4: derive equations that describe the expected behaviour of the state variables in a single time step, make simplifications for large N, and determine steady state values of the variables by assuming that their expected values do not change.

In [1, 3], steady state values

$$s_A = \sqrt{\frac{\mu(2-c)}{c}} c_{\mu/\mu,\lambda} \left(\frac{1}{\sqrt{1+\delta^2}} - \frac{\sigma^*}{\mu c_{\mu/\mu,\lambda}} \right) \tag{18}$$

where δ denotes the noise-to-signal ratio, and

$$\|\mathbf{s}\|^2 = N + \frac{2(1-c)}{\sqrt{c(2-c)}} \sqrt{\mu} s_A \mathrm{E}\left[z_A^{(\mathrm{avg})} \right] \tag{19}$$

are derived for the noisy sphere model. For the PDQF in Eq. (1), the signed lengths of the components of the search path toward the optimum need to be considered separately for the two spheres that form the objective. Requiring stationarity of the expected value yields in direct analogy to Eq. (18)

$$s_{A1} = \sqrt{\frac{\mu(2-c)}{c}} c_{\mu/\mu,\lambda} \left(\frac{1}{\sqrt{1+\zeta^2}} - \frac{\sigma^*}{\mu c_{\mu/\mu,\lambda}} \right). \tag{20}$$

Notice that the noise-to-signal ratio is ζ as seen in Section 4 and is a result of contributions to the objective function value from the second sphere that interfere with the selection of good components of the mutation vector for the first. An analogous calculation for the second sphere yields

$$s_{A2} = \sqrt{\frac{\mu(2-c)}{c}} c_{\mu/\mu,\lambda} \left(\frac{\zeta}{\sqrt{1+\zeta^2}} - \frac{\sigma^*(1-\vartheta)}{\mu c_{\mu/\mu,\lambda} \vartheta \xi \zeta} \right) \tag{21}$$

where the extra factors in the second summand result from the fact that the normalisation of the mutation strength introduced in Section 4 uses quantities from the first sphere.

Replicating the steps made in the derivation of Eq. (19) yields for the PDQF from Eq. (1)

$$\|\mathbf{s}\|^2 = N + \frac{2(1-c)}{\sqrt{c(2-c)}} \sqrt{\mu} \left(s_{A1} \mathrm{E}\left[z_{A1}^{(\mathrm{avg})} \right] + s_{A2} \mathrm{E}\left[z_{A2}^{(\mathrm{avg})} \right] \right).$$

Using Eqs. (10), (11), (20), and (21) and rearranging terms

yields

$$\|\mathbf{s}\|^2 = N + \frac{2(1-c)}{c}\mu c_{\mu/\mu,\lambda}^2 \cdot \left[1 - \frac{\bar{\sigma}}{\sqrt{1+\zeta^2}}\left(1 + \frac{1-\vartheta}{\vartheta\xi}\right)\right] \quad (22)$$

as an approximation for the steady state squared length of the search path.

It remains to compute the (average) normalised mutation strength that using cumulative step length adaptation results in. Rewriting Eq. (5) in terms of the normalised mutation strength, assuming that σ^* is unchanged by the update rule and cancelling it out, and squaring both sides of the equation yields

$$\frac{{R_1^{(t+1)}}^2}{{R_1^{(t)}}^2} = \exp\left(\frac{\|\mathbf{s}\|^2 - N}{DN}\right). \quad (23)$$

According to Eq. (22) with the settings for c and D from Section 2, the argument to the exponential function is inversely proportional to N. For high search space dimensionality, the exponential can thus be expanded into a Taylor series with terms beyond the linear one dropped. Replacing the left hand side of Eq. (23) with its expected value and using Eq. (13) it follows that

$$1 - \frac{2\mu c_{\mu/\mu,\lambda}^2}{N\vartheta}\left(\frac{\bar{\sigma}}{\sqrt{1+\zeta^2}} - \frac{\bar{\sigma}^2}{2}\right) = 1 + \frac{2(1-c)}{cDN}\mu c_{\mu/\mu,\lambda}^2\left[1 - \frac{\bar{\sigma}}{\sqrt{1+\zeta^2}}\left(1 + \frac{1-\vartheta}{\vartheta\xi}\right)\right].$$

Simplifying and recognising that $(1-c)$ tends to one as N increases yields

$$\xi\left(\bar{\sigma}^2 - 2\vartheta\right)\sqrt{1+\zeta^2} = 2\bar{\sigma}(1-\vartheta)(\xi - 1) \quad (24)$$

as a condition that the steady state mutation strength generated by cumulative step length adaptation can be obtained from. Notice that for $\xi = 1$ the previously obtained result for the sphere model is recovered [1, 3].

Fig. 5 illustrates how the normalised mutation strength σ^* and the resulting normalised quality gain Δ^* of the $(\mu/\mu, \lambda)$-CSA-ES depend on the condition number ξ of the Hessian. The lines have been obtained by using Eqs. (15) and (24) to compute the normalised mutation strength and the location parameter. The normalised quality gain has subsequently been obtained from Eq. (16). The dots mark measurements made in runs of the $(\mu/\mu, \lambda)$-CSA-ES. It can be seen that the accuracy of the approximation increases with increasing search space dimensionality. While deviations for $N = 40$ are in the double digit range, those for $N = 400$ are generally below 10%. Not shown, larger values of μ and λ typically require larger values of N for the same degree of accuracy.

Finally, Fig. 6 compares the mutation strength and quality gain obtained when using cumulative step length adaptation with the optimal values derived in Section 4. Due to the use of standardised quantities, the curves are independent of μ and λ. It can be seen that for $\xi = 1$, as previously found for the sphere model [1, 3], the mutation strength generated is larger than optimal by a factor of $\sqrt{2}$, and that the resulting quality gain is below optimal by a factor of $2(\sqrt{2} - 1)$. For large values of the condition number the situation is reversed

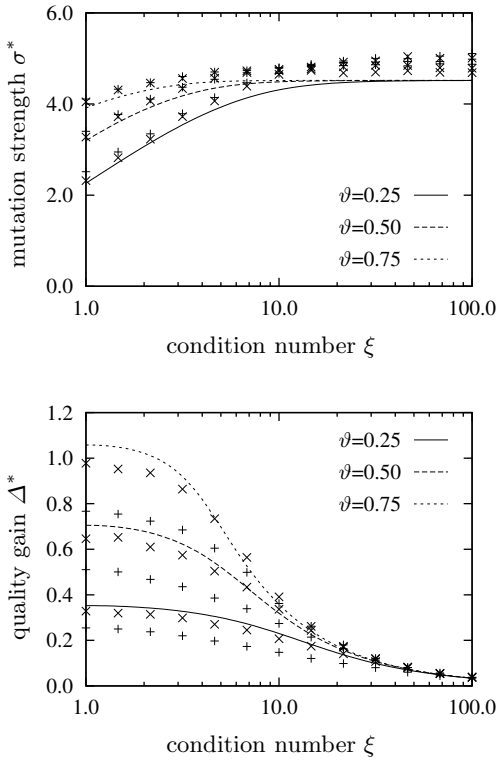

Figure 5: Average normalised mutation strength σ^* and normalised quality gain Δ^* of the $(\mu/\mu, \lambda)$-CSA-ES plotted against condition number ξ for $\mu = 3$, $\lambda = 10$, and $\vartheta \in \{0.25, 0.50, 0.75\}$. The dots mark measurements made in runs of the strategy with $N = 40$ (+) and $N = 400$ (×).

in that the adapted mutation strength is below the optimal one. The standardised mutation strength approaches $\sqrt{2}$ while asymptotically, $\bar{\sigma} = 2$ is optimal. The resulting standardised quality gain asymptotically equals $\bar{\Delta} = 1/\xi$ and is thus half of the asymptotically optimal quality gain derived in Section 4.

6. DISCUSSION AND FUTURE WORK

The results derived in this paper extend those by Jägersküpper by considering a more advanced evolution strategy and a generalisation of the ill-conditioned objective function in [10]. Most importantly, they add detail by providing an approximation to the quality gain for arbitrary degrees of ill-conditioning and not using asymptotic notation that hides constants, albeit at the cost of a loss of mathematical rigour.

In [10] Jägersküpper contends that "Gaussian mutations adapted by the 1/5th rule make the optimization process stabilize such that the trajectory of the evolving search point takes course very close to the gentlest descent of the ellipsoidal fitness landscape, i.e., in the region of (almost maximum curvature)". The results arrived at here suggest that for ill-conditioned functions the mutation strength employed in fact has a significant impact on the trajectory of the search point. Fig. 3 illustrates that the deviation from the "gentlest descent" trajectory increases with increasing normalised mutation strength.

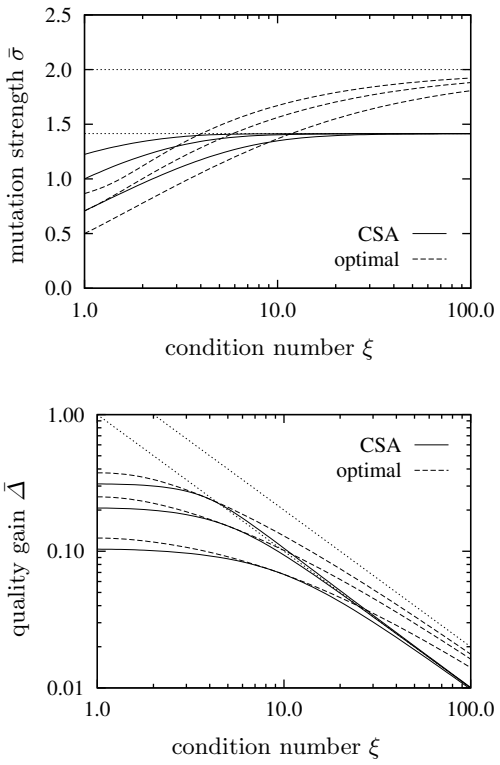

Figure 6: Standardised mutation strength $\bar{\sigma}$ and standardised quality gain $\bar{\Delta}$ of the $(\mu/\mu, \lambda)$-CSA-ES plotted against the condition number ξ of the Hessian. The lines correspond to, from bottom to top, $\vartheta = 0.25$, 0.50, and 0.75. Shown are both optimal values (dashed lines) and values generated by cumulative step length adaptation (solid lines). The dotted lines indicate the limit behaviour for large ξ.

The result for the standardised quality gain derived here is independent of the population size parameters μ and λ. As a result of the standardisation and properties of the $(\mu/\mu, \lambda)$-progress coefficient [4], the (not standardised) quality gain is thus roughly proportional to the population size if both μ and λ are increased in equal proportions. However, it is important to keep in mind that as on the sphere model, the accuracy of the approximations decreases with increasing population size parameters, and that for finite N the speed-up is significantly sublinear unless μ and λ are small.

A further interesting insight gained from Fig. 6 is that both the optimal quality gain and the quality gain achieved with cumulative step length adaptation decrease very slowly with increasing condition number for values of ξ in the vicinity of one (the derivative $\partial\Delta/\partial\xi|_{\xi=1}$ equals zero). The gap between optimal performance and the performance of the $(\mu/\mu, \lambda)$-CSA-ES in fact narrows with increasing condition number before it starts widening. This may be of significance in connection with covariance matrix adaptation strategies such as the CMA-ES [7]. Such strategies strive to learn a mutation covariance matrix that transforms ill-conditioned objective functions locally into functions with low condition numbers. As adaptation is never perfect, condition numbers observed in practice are typically larger than one. Fig. 6 suggests that the impact of the imperfect adaptation of the covariance matrix is rather minor, and that cumulative step length adaptation (which is employed in the CMA-ES) is rather well suited for this case.

Finally, numerous ways of extending the results presented in this paper are conceivable. It is desirable to obtain a better understanding of the behaviour of EAs for PDQFs other than the class considered here. The cigar and discus functions are two natural candidates to be considered. It is also of interest to investigate other mutation strength adaptation mechanisms, and to compare their performance with that of cumulative step length adaptation.

ACKNOWLEDGEMENTS

This research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

7. REFERENCES

[1] D. V. Arnold. *Noisy Optimization with Evolution Strategies*. Kluwer Academic Publishers, 2002.

[2] D. V. Arnold and H.-G. Beyer. Local performance of the $(\mu/\mu_I, \lambda)$-ES in a noisy environment. In W. N. Martin et al., editors, *Foundations of Genetic Algorithms 6*, pages 127–141. Morgan Kaufmann, 2001.

[3] D. V. Arnold and H.-G. Beyer. Performance analysis of evolutionary optimization with cumulative step length adaptation. *IEEE Transactions on Automatic Control*, 49(4):617–622, 2004.

[4] H.-G. Beyer. *The Theory of Evolution Strategies*. Springer Verlag, 2001.

[5] H.-G. Beyer and D. V. Arnold. The steady state behavior of $(\mu/\mu_I, \lambda)$-ES on ellipsoidal fitness models disturbed by noise. In E. Cantú-Paz et al., editors, *Genetic and Evolutionary Computation — GECCO 2003*, pages 525–536. Springer Verlag, 2003.

[6] H.-G. Beyer and H.-P. Schwefel. Evolution strategies — A comprehensive introduction. *Natural Computing*, 1(1):3–52, 2002.

[7] N. Hansen and A. Ostermeier. Completely derandomized self-adaptation in evolution strategies. *Evolutionary Computation*, 9(2):159–195, 2001.

[8] J. Jägersküpper. Analysis of a simple evolutionary algorithm for minimization in Euclidean spaces. In *Automata, Languages, and Programming — ICALP 2003*, pages 1068–1079. Springer Verlag, 2003.

[9] J. Jägersküpper. Rigorous runtime analysis of the $(1 + 1)$ ES: 1/5-rule and ellipsoidal fitness landscapes. In A. H. Wright et al., editors, *Foundations of Genetic Algorithms 8*, pages 260–281. Springer Verlag, 2005.

[10] J. Jägersküpper. How the $(1 + 1)$ ES using isotropic mutations minimizes positive definite quadratic forms. *Theoretical Computer Science*, 361(1):38–56, 2006.

[11] A. Ostermeier, A. Gawelczyk, and N. Hansen. Step-size adaptation based on non-local use of selection information. In Y. Davidor et al., editors, *Parallel Problem Solving from Nature — PPSN III*, pages 189–198. Springer Verlag, 1994.

[12] I. Rechenberg. *Evolutionsstrategie '94*. Friedrich Frommann Verlag, 1994.

[13] H.-P. Schwefel. *Evolution and Optimum Seeking*. Wiley, 1995.