Theoretical Computer Science 832 (2020) 42–67

Contents lists available at ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

Quality gain analysis of the weighted recombination evolution strategy on general convex quadratic functions ☆

Youhei Akimoto [a,*], Anne Auger [b], Nikolaus Hansen [b]

[a] Faculty of Engineering, Information and Systems, University of Tsukuba, Japan
[b] Inria, RandOpt Team, CMAP, Ecole Polytechnique, France

ARTICLE INFO

Article history:
Received 15 November 2017
Received in revised form 9 April 2018
Accepted 10 May 2018
Available online 26 June 2018

Keywords:
Evolution strategy
Weighted recombination
Quality gain analysis
Optimal step-size
General convex quadratic function

ABSTRACT

Quality gain is the expected relative improvement of the function value in a single step of a search algorithm. Quality gain analysis reveals the dependencies of the quality gain on the parameters of a search algorithm, based on which one can derive the optimal values for the parameters. In this paper, we investigate evolution strategies with weighted recombination on general convex quadratic functions. We derive a bound for the quality gain and two limit expressions of the quality gain. From the limit expressions, we derive the optimal recombination weights and the optimal step-size, and find that the optimal recombination weights are independent of the Hessian of the objective function. Moreover, the dependencies of the optimal parameters on the dimension and the population size are revealed. Differently from previous works where the population size is implicitly assumed to be smaller than the dimension, our results cover the population size proportional to or greater than the dimension. Numerical simulation shows that the asymptotically optimal step-size well approximates the empirically optimal step-size for a finite dimensional convex quadratic function.

1. Introduction

Background Evolution Strategies (ES) are randomized search algorithms to minimize a black-box function $f : \mathbb{R}^N \to \mathbb{R}$ in continuous domain, where neither the gradient nor the Hessian matrix of the objective function is available. The most advanced and commonly used category of evolution strategies is covariance matrix adaptation evolution strategy (CMA-ES) [2,3], which is recognized as the state-of-the-art black box continuous optimizer. It generates multiple candidate solutions from a multivariate normal distribution. They are evaluated on the objective function. The distribution parameters such as the mean vector and the covariance matrix are updated by using the candidate solutions and their ranking information, where the objective function values are not directly used. Due to its population-based and comparison-based nature, the algorithm is invariant to any strictly increasing transformation of the objective function in addition to the invariance to scaling, translation and rotation of the search space [4]. These invariance properties guarantee that the algorithm shows exactly the same behavior on a function f and on its transformation $g \circ f \circ T$, where $g : \mathbb{R} \to \mathbb{R}$ is a strictly increasing function and $T : \mathbb{R}^N \to \mathbb{R}^N$ is a combination of scaling, translation and rotation defined as $T : \boldsymbol{x} \mapsto a \cdot \mathbf{U}(\boldsymbol{x} - \boldsymbol{b})$ with a positive

☆ This is the extension of our extended abstract presented at FOGA'2017 [1].

* Corresponding author.
E-mail addresses: akimoto@cs.tsukuba.ac.jp (Y. Akimoto), anne.auger@inria.fr (A. Auger), nikolaus.hansen@inria.fr (N. Hansen).

https://doi.org/10.1016/j.tcs.2018.05.015
0304-3975/

real $a > 0$, an N dimensional vector $\boldsymbol{b} \in \mathbb{R}^N$, and an N dimensional orthogonal matrix $\mathbf{U}$. These invariance properties are the essence of the success of CMA-ES.

The performance evaluation of evolutionary algorithms is often based on empirical studies such as benchmarking on a test function suite [5,6] and well-considered performance assessment [7,8]. It is easier to check the performance of an algorithm on a specific problem in simulation than to analyze it mathematically. The invariance properties of an algorithm then generalize the empirical result to a class of infinitely many functions defined by the invariance relation. On the other hand, theoretical studies often require simplification of algorithms and assumptions on the objective function, because of the difficulty of the analysis of advanced algorithms due to their comparison-based and population-based nature and the complex adaptation mechanisms. Nevertheless, theoretical studies lead us to a better understanding of algorithms and reveals the dependency of the performance on the interval parameter settings. For example, the recombination weights in CMA-ES are selected based on the mathematical analysis of an evolution strategy [9].[1] The theoretical result of the optimal step-size on the sphere function is used to design a box constraint handling technique [10] and to design a termination criterion for a restart strategy [11]. A recent variant of CMA-ES [12] exploits the theoretical result of the optimal rate of convergence of the step-size to estimate the condition number of the product of the covariance matrix and the Hessian matrix of the objective function.

Quality gain analysis *Quality gain* and *progress rate* analysis [13–15] measure the expected progress of the mean vector in one step. On one side, differently from convergence analysis (e.g., [16]), analyses based on these quantities do not guarantee the convergence and often take a limit to derive an explicit formula. Moreover, the step-size adaptation and the covariance matrix adaptation are not taken into account. On the other side, one can derive quantitative explicit estimates of these quantities, which is not the case in convergence analysis. The quantitative explicit formulas are particularly useful to know the dependency of the expected progress on the parameters of the algorithm such as the population size, number of parents, and recombination weights, which we may not recognize from empirical studies of algorithms. The above mentioned recombination weights in CMA-ES are derived from the quality gain analysis of evolution strategies [9].

Although quality gain analysis is not meant to guarantee the convergence of the algorithm since it analyzes only a single step expected improvement, the progress rate is linked to the convergence rate of algorithms. It is directly related to the convergence rate of an "artificial" algorithm where the step-size is proportional to the distance to the optimum on the sphere function (see e.g., [17]). Moreover, the convergence rate of this artificial algorithm gives a bound on the convergence rate of algorithms that implement a proper step-size adaptation. For $(1 + \lambda)$ or $(1, \lambda)$ ESs the bound holds on any function with a unique global optimum; that is, any step-size adaptive $(1 \dot{+} \lambda)$-ES optimizing any function f with a unique global optimum can not achieve a convergence rate faster than the convergence rate of the artificial algorithm on the sphere function where the step-size is the distance to the optimum times the optimal constant [18–20].[2] For algorithms implementing recombination, this bound still holds on spherical functions [19,20].

Related work In this paper, we investigate ESs with weighted recombination on a general convex quadratic function. ESs with weighted recombination samples multiple candidate solutions at one time and compute the weighted average of the candidate solutions to update the distribution mean vector. Weighted recombination ESs are among the most important categories of ESs since the standard CMA-ES and most of the recent variants of CMA-ES [21–23] employ weighted recombination.

The first analysis of weighted recombination ESs were done in [9], where the quality gain has been derived on the infinite dimensional sphere function $f : x \mapsto \|x\|^2$. The optimal step-size and the optimal recombination weights are derived. Reference [24] studied a variant of weighted recombination ESs called $(\mu/\mu_I, \lambda)$-ES, where $(\mu/\mu_I, \lambda)$ stands for intermediate recombination, where the recombination weights are equal for the best μ candidate solutions and zero for the other $\lambda - \mu$ candidate solutions. The analysis has been performed on the quadratic functions with the Hessian $\mathbf{A} = \frac{1}{2}\operatorname{diag}(\alpha, \ldots, \alpha, 1, \ldots, 1)$, where the number $\lfloor N\theta \rfloor$ of diagonal elements that are $\alpha > 1$ is controlled by the ratio θ of short axes. Reference [25] studied the $(1 + 1)$-ES with the one-fifth success rule on the same function and showed the convergence rate of $\Theta(1/(\alpha N))$. Reference [26] studied ES with weighted recombination on the same function. Their results, progress rate and quality gain, depend on the so-called localization parameter, the steady-state value of which is then analyzed to obtain the steady-state quality gain. References [27,28] studied the progress rate and the quality gain of $(\mu/\mu_I, \lambda)$-ES on the general convex quadratic model.

The quality gain analysis and the progress rate analysis in the above listed references rely on a geometric intuition of the algorithm in the infinite dimensional search space and on various approximations. On the other hand, the rigorous derivation of the progress rate (or convergence rate of the algorithm with step-size proportional to the distance to the optimum) on the sphere function provided for instance in [17,20,29,30] only holds on spherical functions and provides solely a limit without a bound between the finite dimensional convergence rate and its asymptotic limit. The result of this

[1] The weights of CMA-ES were set before the publication [9] because the theoretical result of optimal weights on the sphere was known before the publication.

[2] More precisely, $(1 \dot{+} \lambda)$-ES optimizing any function f (that may have more than one global optimum) can not converge towards a given optimum x^* faster in the search space than the artificial algorithm with step-size proportional to the distance to x^*.

paper is different in that we consider the general weighted recombination on the general convex quadratic objective and cover finite dimensional cases as well as the limit $N \to \infty$.

Contributions We study the weighted recombination ES on a general convex quadratic function $f(x) = \frac{1}{2}(x - x^*)^{\mathrm{T}}\mathbf{A}(x - x^*)$ on the finite N dimensional search space. We investigate the quality gain ϕ, that is, the expectation of the relative function value decrease. We decompose ϕ as the product of two functions: g, a function that depends only on the mean vector of the sampling distribution and the Hessian $\mathbf{A}$, and $\bar{\phi}$, the so-called *normalized quality gain* that depends essentially on all the algorithm parameters such as the recombination weights and the step-size. We approximate $\bar{\phi}$ by an analytically tractable function φ. We call φ the *asymptotic normalized quality gain*. The main contributions are summarized as follows.

First, we derive the error bound between $\bar{\phi}$ and φ for finite dimension N. To the best of our knowledge, this is the first work that performs the quality gain analysis for finite N and provides an error bound. The asymptotic normalized quality gain and the bounds in this paper are improved over the previous work [1]. Thanks to the explicit error bound derived in the paper, we can treat the population size λ increasing with N and provide (for instance) a rigorous sufficient condition on the dependency between λ and N such that the per-iteration quality gain scales with $O(\lambda/N)$ for algorithms with intermediate recombination [15].

Second, we show that the error bound between $\bar{\phi}$ and φ converges to zero as the learning rate c_m for the mean vector update tends to infinity. We derive the optimal step-size and the optimal recombination weights for φ, revealing the dependencies of these optimal parameters on λ and N. In contrast, the previous works of quality gain analysis mentioned above take the limit $N \to \infty$ while λ is fix, hence assuming $\lambda \ll N$. Therefore, they do not reveal the dependencies of $\bar{\phi}$ and the optimal parameters on λ when $\lambda \not\ll N$. We validate in experiments that the optimal step-size derived for $c_m \to \infty$ provides a reasonable estimate of the optimal step-size even for $c_m = 1$.

Third, we prove that φ converges toward $\bar{\phi}_\infty$ as $N \to \infty$ under the condition $\lim_{N\to\infty} \mathrm{Tr}(\mathbf{A}^2)/\mathrm{Tr}(\mathbf{A})^2 = 0$, where $\bar{\phi}_\infty$ is the limit of $\bar{\phi}$ on the sphere function for $N \to \infty$ derived in [9]. The condition $\lim_{N\to\infty} \mathrm{Tr}(\mathbf{A}^2)/\mathrm{Tr}(\mathbf{A})^2$ holds, for example, for positive definite $\mathbf{A}$ with bounded eigenvalues. It also holds for some positive semi-definite $\mathbf{A}$ and for some positive definite $\mathbf{A}$ with unbounded eigenvalues, for example with eigenvalues in $[1, \sqrt{N}]$. The result implies that the optimal recombination weights are independent of $\mathbf{A}$, whereas the optimal step-size heavily depends on $\mathbf{A}$ and the distribution mean. This part of the contribution is a generalization of the previous foundation in [27,28], but the proof methodology is rather different. Furthermore, the error bound between $\bar{\phi}$ and φ derived in this paper allows us to further investigate how fast φ converges toward $\bar{\phi}_\infty$ as $N \to \infty$, depending on the eigenvalue distribution of $\mathbf{A}$.

Organization This paper is organized as follows. In Section 2, we formally define the evolution strategy with weighted recombination. The quality gain analysis on the infinite dimensional sphere function is revisited. In Section 3, we derive the quality gain bound for a finite dimensional convex quadratic function. In Section 4, important consequences of the quality gain bound are discussed. In Section 5, we conclude our paper. Properties of the normal order statistics that are important to understand our results are summarized in Appendix A and the detailed proofs of lemmas are provided in Appendix B.

Notation We apply the following mathematical notations throughout the paper. For integers $n, m \in \mathbb{N}$ such that $n \leq m$, we denote the set of integers between n and m (including n and m) by $[\![n, m]\!]$. Binomial coefficients are denoted as $\binom{m}{n} = \frac{m!}{(m-n)!n!}$. For real numbers $a, b \in \mathbb{R}$ such that $a \leq b$, the open and the closed intervals are denoted as (a, b) and $[a, b]$, respectively. For an N-dimensional real vector $x \in \mathbb{R}^N$, let $[x]_i$ denote the i-th coordinate of x. A sequence of length n is denoted as $(x_i)_{i=1}^n = (x_1, \cdots, x_n)$, or just as (x_i), and an infinite sequence is denoted as $(x_i)_{i=1}^\infty$. For $x \in \mathbb{R}$, the absolute value of x is denoted by $|x|$. For $x \in \mathbb{R}^N$, the Euclidean norm is denoted by $\|x\| = \left(\sum_{i=1}^N [x]_i^2\right)^{\frac{1}{2}}$. Let $\mathbb{1}_{\texttt{condition}}$ be the indicator function which is 1 if `condition` is true and 0 otherwise. Let Φ be the cumulative density function (c.d.f.) deduced by the (one-dimensional) standard normal distribution $\mathcal{N}(0, 1)$. Let $\mathcal{N}_{i:\lambda}$ be the i-th smallest random variable among λ independently and standard normally distributed random variables, i.e., $\mathcal{N}_{1:\lambda} \leq \cdots \leq \mathcal{N}_{\lambda:\lambda}$. The expectation of a random variable (or vector) X is denoted as $\mathbb{E}[X]$. The conditional expectation of X given Y is denoted as $\mathbb{E}[X \mid Y]$. For a function f of n random variables $(X_i)_{i=1}^n$, the conditional expectation of $F = f(X_1, \ldots, X_n)$ given X_k for some $k \in [\![1, n]\!]$ is denoted as $\mathbb{E}_k[F] = \mathbb{E}[F \mid X_k]$. Similarly, the conditional expectation of F given X_k and X_l for different $k, l \in [\![1, n]\!]$ is denoted as $\mathbb{E}_{k,l}[F] = \mathbb{E}[F \mid X_k, X_l]$.

2. Formulation

2.1. Evolution strategy with weighted recombination

We consider an evolution strategy with weighted recombination. At each iteration $t \geq 0$, it draws λ independent random vectors $Z_1, \ldots, Z_\lambda$ from the N-dimensional standard normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$, where $\mathbf{0} \in \mathbb{R}^N$ is the zero vector and $\mathbf{I}$ is the identity matrix of dimension N. The candidate solutions $X_1, \ldots, X_\lambda \sim \mathcal{N}(\boldsymbol{m}^{(t)}, (\sigma^{(t)})^2\mathbf{I})$ are computed as $X_i = \boldsymbol{m}^{(t)} + \sigma^{(t)} Z_i$, where $\boldsymbol{m}^{(t)} \in \mathbb{R}^N$ is the mean vector and $\sigma^{(t)} > 0$ is the standard deviation, also called the step-size or the mutation strength. The candidate solutions are evaluated on a given objective function $f : \mathbb{R}^N \to \mathbb{R}$. Without loss of generality (w.l.o.g.), we assume f to be minimized. Let $i : \lambda$ be the index of the i-th best candidate solution among

Algorithm 1 Single step of the weighted recombination ES solving f.

procedure ES($\boldsymbol{m}, \sigma, (w_k)_{k=1}^{\lambda}, c_m$)
  for $i = 1, \ldots, \lambda$ **do** ▷ Generate and evaluate λ candidate solutions
    $Z_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
    $X_i = \boldsymbol{m} + \sigma Z_i$
    Evaluate $f(X_i)$
  end for
  $W(i; (X_k)_{k=1}^{\lambda}) = \sum_{k=1+l_i}^{u_i} w_k/(u_i - l_i)$ ▷ Compute the weights with (2)
  $\boldsymbol{m} \leftarrow \boldsymbol{m} + c_m \sigma \sum_{i=1}^{\lambda} W(i; (X_k)_{k=1}^{\lambda}) Z_i$ ▷ Update the mean with (3)
  return $\boldsymbol{m}$
end procedure

$X_1, \ldots, X_\lambda$, i.e., $f(X_{1:\lambda}) \le \cdots \le f(X_{\lambda:\lambda})$, and $w_1 \ge \cdots \ge w_\lambda$ be the real-valued recombination weights. W.l.o.g., we assume $\sum_{i=1}^{\lambda} |w_i| = 1$. Let $\mu_w = 1/\sum_{i=1}^{\lambda} w_i^2$ denote the so-called effective variance selection mass. The mean vector is updated according to

$$\boldsymbol{m}^{(t+1)} = \boldsymbol{m}^{(t)} + c_m \sum_{i=1}^{\lambda} w_i (X_{i:\lambda} - \boldsymbol{m}^{(t)}) \ , \tag{1}$$

where $c_m > 0$ is the learning rate of the mean vector update.

In this paper we reformulate (1) to investigate the algorithm with mathematical rigor. Hereunder, we write the candidate solutions, $X_1, \ldots, X_\lambda$, and the corresponding random vectors, $Z_1, \ldots, Z_\lambda$, as sequences $(X_i)_{i=1}^{\lambda}$ and $(Z_i)_{i=1}^{\lambda}$ for short. First, we introduce the weight function

$$W(i; (X_k)_{k=1}^{\lambda}) := \sum_{k=1+l_i}^{u_i} \frac{w_k}{u_i - l_i} \ , \qquad \text{where} \qquad \begin{cases} l_i = \sum_{j=1}^{\lambda} \mathbb{1}_{f(X_j) < f(X_i)} \\ u_i = \sum_{j=1}^{\lambda} \mathbb{1}_{f(X_j) \le f(X_i)} \end{cases} \ , \tag{2}$$

i.e., l_i and u_i are the numbers of strictly and weakly better candidate solutions than X_i, respectively. The weight value for X_i is the arithmetic average of the weights w_k for the tie candidate solutions. In other words, all the tie candidate solutions have the same weight values. If there is no tie, the weight value for the i-th best candidate solution $X_{i:\lambda}$ is simply w_i. In the following, we drop the subscripts and the superscripts for sequences unless they are unclear from the context and write simply as $(X_k) = (X_k)_{k=1}^{\lambda}$. With the weight function, we rewrite the mean vector update (1) as

$$\boldsymbol{m}^{(t+1)} = \boldsymbol{m}^{(t)} + c_m \sum_{i=1}^{\lambda} W(i; (X_k))(X_i - \boldsymbol{m}^{(t)}) \ . \tag{3}$$

The above update (3) is equivalent with the original update (1) if there is no tie among λ candidate solutions. If the objective function is a convex quadratic function, there will be no tie with probability one. Therefore, they are equivalent with probability one. Algorithm 1 summarizes the single step of the algorithm, where we rewrite (3) by using $X_i - \boldsymbol{m}^{(t)} = \sigma^{(t)} Z_i$.

The above formulation is motivated twofold. One is to well define the update even when there is tie. In our formulation, tie candidate solutions receive the equal recombination weights. The other is a technical reason. In (1) the already sorted candidate solutions $X_{i:\lambda}$ are all correlated and they are not anymore normally distributed. However, they are assumed to be normally distributed in the previous work [9,27,28]. To ensure that such an approximation leads to the asymptotically true quality gain limit, a mathematically involved analysis has to be done. See [17,29,30] for details. In (3), the weight function explicitly includes the ranking computation and X_i are still independent and normally distributed. This allows us to derive the quality gain on a convex quadratic function rigorously.

2.2. Quality gain analysis on the spherical function

The quality gain is defined as the expectation of the relative decrease of the function value. Formally, it is the conditional expectation of the relative decrease of the function value conditioned on the mean vector $\boldsymbol{m}^{(t)} = \boldsymbol{m}$ and the step-size $\sigma^{(t)} = \sigma$, defined as follows.

Definition 1. The *quality gain* of Algorithm 1 given $\boldsymbol{m}^{(t)} = \boldsymbol{m}$ and $\sigma^{(t)} = \sigma$ is

$$\phi(\boldsymbol{m}, \sigma) = \frac{\mathbb{E}[f(\boldsymbol{m}) - f(\mathrm{ES}(\boldsymbol{m}, \sigma, (w_k)_{k=1}^{\lambda}, c_m))]}{f(\boldsymbol{m}) - f(x^*)} \ , \tag{4}$$

where $x^* \in \mathbb{R}^N$ is (one of) the global minimum point of f. Note that the quality gain depends also on $(w_k)_{k=1}^{\lambda}$, c_m, and the dimension N.

Results Algorithm 1 solving a spherical function $f(x) = \|x\|^2$ is analyzed in [9]. For this purpose, the *normalized step-size* and the *normalized quality gain* are introduced as

$$\bar{\sigma} = \sigma \frac{c_m N}{\|\boldsymbol{m}\|} \qquad \text{and} \qquad \bar{\phi}(\boldsymbol{m}, \bar{\sigma}) = \frac{N}{2}\phi\left(\boldsymbol{m}, \sigma = \frac{\bar{\sigma}\,\|\boldsymbol{m}\|}{c_m N}\right) , \tag{5}$$

respectively. This normalization of the step-size suggests that σ is proportional to $\|\boldsymbol{m}\|$ and inverse proportional to c_m and to N. The normalized step-size $\bar{\sigma}$ is proportional to the ratio between the actual step-size and the distance between the current mean and the optimal solution. This reflects the scale invariance of the algorithm on the sphere function, that is, the single step response is solely determined by the normalized step-size. The dimension N in the numerator implies that the step-size σ needs to be inversely proportional to N. The normalized quality gain $\bar{\phi}$ is simply the quality gain given $\bar{\sigma}$ scaled by $N/2$. The scaling by N reflects that the convergence speed can not exceed $O(1/N)$ for any comparison based algorithm [31]. By taking $N \to \infty$, the normalized quality gain converges pointwise (w.r.t. $\bar{\sigma}$) to

$$\begin{aligned}\bar{\phi}_\infty(\bar{\sigma}, (w_k)) &:= \lim_{N\to\infty} \bar{\phi}(\boldsymbol{m}, \bar{\sigma}) = -\bar{\sigma}\sum_{i=1}^{\lambda} w_i \mathbb{E}[\mathcal{N}_{i:\lambda}] - \frac{\bar{\sigma}^2}{2\mu_w} \\ &= \frac{\mu_w}{2}\left(\sum_{i=1}^{\lambda} w_i \mathbb{E}[\mathcal{N}_{i:\lambda}]\right)^2 \left(1 - \left(\frac{\bar{\sigma}}{\bar{\sigma}^*((w_k))} - 1\right)^2\right) ,\end{aligned} \tag{6}$$

where $\bar{\sigma}^*((w_k))$ denotes the normalized step-size $\bar{\sigma}$ optimizing $\bar{\phi}_\infty$ given (w_k) and is given by

$$\bar{\sigma}^*((w_k)) = -\mu_w \sum_{i=1}^{\lambda} w_i \mathbb{E}[\mathcal{N}_{i:\lambda}] \ . \tag{7}$$

A formal proof of this result is presented in Theorem 2 of [29] relying on the uniform integrability of some random variable proved in [30].

Consider the optimal recombination weights that maximize $\bar{\phi}_\infty$ in (6). The optimal recombination weights are given independently of $\bar{\sigma}$ by

$$w_k^* = -\frac{\mathbb{E}[\mathcal{N}_{k:\lambda}]}{\sum_{i=1}^{\lambda} |\mathbb{E}[\mathcal{N}_{i:\lambda}]|} \tag{8}$$

and $\bar{\phi}_\infty$ is written as

$$\bar{\phi}_\infty(\bar{\sigma}, (w_k^*)) = \frac{\sum_{i=1}^{\lambda} \mathbb{E}[\mathcal{N}_{i:\lambda}]^2}{2}\left(1 - \left(\frac{\bar{\sigma}}{\sum_{i=1}^{\lambda} |\mathbb{E}[\mathcal{N}_{i:\lambda}]|} - 1\right)^2\right) \ . \tag{9}$$

Note that $\bar{\sigma}^*((w_k^*)) = \sum_{i=1}^{\lambda} |\mathbb{E}[\mathcal{N}_{i:\lambda}]|$. Given $\bar{\sigma}^*$ and (w_k^*), we achieve the maximal value of $\bar{\phi}_\infty$ that is $\bar{\phi}_\infty(\bar{\sigma}^*((w_k^*)), (w_k^*)) = \sum_{i=1}^{\lambda} \mathbb{E}[\mathcal{N}_{i:\lambda}]^2/2$.

Remarks The optimal normalized step-size (7) and the normalized quality gain (6) given $\bar{\sigma}^*$ depend on (w_k). Particularly, they are proportional to μ_w. For instance, under the optimal weights (8), we have $-\sum_{i=1}^{\lambda} w_i \mathbb{E}[\mathcal{N}_{i:\lambda}] = \sum_{i=1}^{\lambda} \mathbb{E}[\mathcal{N}_{i:\lambda}]^2 / \sum_{i=1}^{\lambda} |\mathbb{E}[\mathcal{N}_{i:\lambda}]| \approx (\pi/2)^{1/2}$ for a sufficiently large λ.[3] Then, from (6) and (7) we know $\bar{\sigma} \propto \mu_w$ and $\bar{\phi}_\infty \propto \mu_w$. Moreover, using the relation $\mu_w = \left(\sum_{i=1}^{\lambda} |\mathbb{E}[\mathcal{N}_{i:\lambda}]|\right)^2 / \sum_{i=1}^{\lambda} \mathbb{E}[\mathcal{N}_{i:\lambda}]^2 \approx (2/\pi)\lambda$, we can reword it as that the optimal step-size and the normalized quality gain given $\bar{\sigma}^*$ are proportional to λ. Fig. 1 shows how $\bar{\phi}_\infty/\lambda$ scales with λ when the optimal step-size $\sigma = \bar{\sigma}^*((w_k))\,\|\boldsymbol{m}\|/(c_m N)$ is set. This shows that the normalized quality gain, and hence the optimal normalized step-size, are proportional to λ for standard weight schemes. When the optimal weights are used, $\bar{\phi}_\infty/\lambda$ goes up to 0.5 as λ increases. On the other hand, nonnegative weights can not achieve the value of $\bar{\phi}_\infty/\lambda$ above 0.25. The CMA type weights are designed to approximate the optimal nonnegative weights, where the first half are proportional to the optimal setting and the last half are zero. The truncation weights result in a smaller normalized quality gain. It is shown in [15] that the truncation weights achieve $\bar{\phi}_\infty \in O(\mu \log(\lambda/\mu))$.

The normalized quality gain limit $\bar{\phi}_\infty$ depends only on the normalized step-size $\bar{\sigma}$ and the weights (w_k). Since the normalized step-size does not change if we multiply c_m by some factor and divide σ by the same factor, c_m does not have any impact on $\bar{\phi}_\infty$, hence on the quality gain ϕ. This is unintuitive and is not true in a finite dimensional space. The step-size σ realizes the standard deviation of the sampling distribution and it has an impact on the ranking of the

[3] We used the facts $\lim_{\lambda\to\infty} \sum_{i=1}^{\lambda} \mathbb{E}[\mathcal{N}_{i:\lambda}]^2/\lambda = 1$ and $\lim_{\lambda\to\infty} \sum_{i=1}^{\lambda} |\mathbb{E}[\mathcal{N}_{i:\lambda}]|/\lambda = (2/\pi)^{1/2}$. See Appendix A for details.

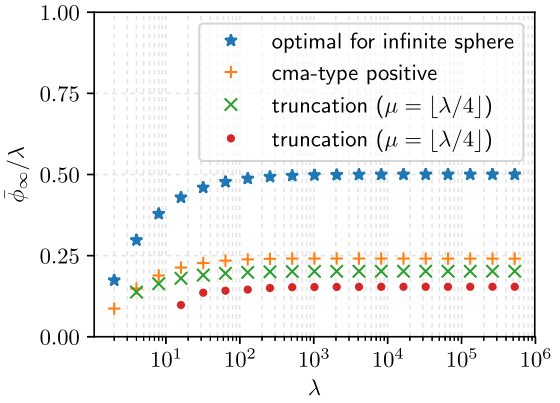


Fig. 1. The normalized quality gain limit $\bar{\phi}_\infty(\bar{\sigma}^*((w_k)),(w_k))$ divided by λ. Four different weight schemes are employed: the optimal weights ($w_k \propto -\mathbb{E}[\mathcal{N}_{k:\lambda}]$), the weights used in the CMA-ES ($w_k \propto \max(\ln(\frac{\lambda+1}{2}) - \ln(k), 0)$), and the truncation weights ($w_k = 1/\mu$ for $k = 1, \ldots, \mu$ and $w_k = 0$ for $k = \mu+1, \ldots, \lambda$) with $\mu = \lfloor \lambda/4 \rfloor$ and $\mu = \lfloor \lambda/10 \rfloor$. All the weights are scaled so that $\sum_{k=1}^{\lambda} |w_k| = 1$. The value of $\mathbb{E}[\mathcal{N}_{i:\lambda}]$ is approximated by the Blom's formula (see Appendix A).

candidate solutions. On the other hand, the product σc_m is the step-size of the $\boldsymbol{m}$-update that depends on the ranking of the candidate solutions. The normalized quality gain limit provided above tells us that the ranking of the candidate solutions is independent of $\bar{\sigma}$ in the infinite dimensional space. We will discuss this further in Section 4.

The quality gain is to measure the improvement in one iteration. If we generate and evaluate λ candidate solutions every iteration, the quality gain per evaluation (f-call) is $1/\lambda$ times smaller, i.e., the quality gain per evaluation is $1/N$, rather than λ/N. It implies that the number of iterations to achieve the same amount of the quality gain is inversely proportional to λ. This is the best we can hope for when the algorithm is implemented on a parallel computer. However, since the above result is obtained in the limit $N \to \infty$ while λ is fixed, it is implicitly assumed that $\lambda \ll N$. The optimal down scaling of the number of iterations indeed only holds for $\lambda \ll N$. In practice, the quality gain per iteration tends to level out as λ increases. We will revisit this point in Section 4 and see how the optimal values for $\bar{\sigma}$ and $\bar{\phi}$ depend on N and λ when both are finite.

3. Quality gain analysis on general quadratic functions

In this section we investigate the normalized quality gain of Algorithm 1 minimizing a quadratic function with its Hessian $\nabla\nabla f(x) = \mathbf{A}$ assumed to be nonnegative definite and symmetric, i.e.,

$$f(x) = \frac{1}{2}(x - x^*)^{\mathrm{T}}\mathbf{A}(x - x^*) \ , \tag{10}$$

where $x^* \in \mathbb{R}^N$ is the global optimal solution.[4] W.l.o.g., we assume $\mathrm{Tr}(\mathbf{A}) = 1$.[5] For the sake of notation simplicity we denote the directional vector of the gradient of f at $\boldsymbol{m}$ by $\mathbf{e} = \frac{\nabla f(\boldsymbol{m})}{\|\nabla f(\boldsymbol{m})\|} = \frac{\mathbf{A}(\boldsymbol{m}-x^*)}{\|\mathbf{A}(\boldsymbol{m}-x^*)\|}$. To make the dependency of $\mathbf{e}$ on $\boldsymbol{m}$ clear, we sometimes write it as $\mathbf{e}_{\boldsymbol{m}}$.

3.1. Normalized quality gain and normalized step-size

We introduce the normalized step-size and the normalized quality gain. First of all, if the objective function is homogeneous around the optimal solution x^*, the optimal step-size must be a homogeneous function of degree 1 with respect to $\boldsymbol{m} - x^*$. This is formally stated in the following proposition. The proof is found in Appendix B.1.

Proposition 2. *Let $f : \mathbb{R}^N \to \mathbb{R}$ be a homogeneous function of degree n, i.e., $f(\alpha \cdot x) = \alpha^n f(x)$ for a fixed integer $n > 0$ for any $\alpha > 0$ and any $x \in \mathbb{R}^N$. Consider Algorithm 1 minimizing a function $f^* : x \mapsto f(x - x^*)$. Then, the quality gain is scale-invariant, i.e., $\phi(x^* + (\boldsymbol{m} - x^*), \sigma) = \phi(x^* + \alpha(\boldsymbol{m} - x^*), \alpha\sigma)$ for any $\alpha > 0$. Moreover, the optimal step-size $\sigma^* = \mathrm{argmax}_{\sigma \geq 0} \phi(\boldsymbol{m}, \sigma)$, if it is well-defined, is a function of $\boldsymbol{m} - x^*$. For the sake of simplicity we write the optimal step-size as a map $\sigma^* : \boldsymbol{m} - x^* \mapsto \sigma^*(\boldsymbol{m} - x^*)$. It is a homogeneous function of degree 1, i.e., $\sigma^*(\alpha \cdot (\boldsymbol{m} - x^*)) = \alpha\sigma^*(\boldsymbol{m} - x^*)$ for any $\alpha > 0$.*

[4] We use the following terminology in this paper. A nonnegative definite matrix $\mathbf{A}$ is a matrix having only nonnegative eigenvalues, i.e., $x^{\mathrm{T}}\mathbf{A}x \geq 0$ for all $x \in \mathbb{R}^N$. A nonnegative definite matrix $\mathbf{A}$ is called positive definite if $x^{\mathrm{T}}\mathbf{A}x > 0$ for all $x \in \mathbb{R}^N \setminus \{\mathbf{0}\}$, otherwise it is called positive semi-definite. If $\mathbf{A}$ is positive semi-definite, the optimum x^* is not unique.

[5] None of the algorithmic components and the quality measures used in the paper are affected by multiplying a positive constant to $\mathbf{A}$, or equivalently to f. To consider a general $\mathbf{A}$, simply replace $\mathbf{A}$ with $\mathbf{A}/\mathrm{Tr}(\mathbf{A})$ in the following of the paper.

Note that the quadratic function is homogeneous of degree 2, and the function $\boldsymbol{m} \mapsto \|\nabla f(\boldsymbol{m})\| = \|\mathbf{A}(\boldsymbol{m} - x^*)\|$ is homogeneous of degree 1 around x^*. The latter is our candidate for the optimal step-size. We define the normalized step-size, the scale-invariant step-size, and the normalized quality gain for a quadratic function as follows.

Definition 3. For a convex quadratic function (10), the *normalized step-size* $\bar{\sigma}$ and the *scale-invariant step-size* σ given $\bar{\sigma}$ are defined as $\bar{\sigma} = (\sigma c_m)/\|\nabla f(\boldsymbol{m})\|$ and $\sigma = (\bar{\sigma}/c_m)\|\nabla f(\boldsymbol{m})\|$.

Definition 4. Let $g : \mathbb{R}^N \to \mathbb{R}$ be the $\boldsymbol{m}$-dependent scaling factor of the normalized quality gain defined as $g(\boldsymbol{m}) = \|\nabla f(\boldsymbol{m})\|^2 / f(\boldsymbol{m})$. The *normalized quality gain* for a quadratic function is defined as $\bar{\phi}(\boldsymbol{m}, \bar{\sigma}) = \phi(\boldsymbol{m}, \sigma = \bar{\sigma}\|\nabla f(\boldsymbol{m})\|/c_m)/g(\boldsymbol{m})$.

Note that the normalized step-size and the normalized quality gain defined above coincide with (5) if $f(x) = \|x\|^2/(2N)$, where $\mathbf{A} = \mathbf{I}/N$, $\nabla f(\boldsymbol{m}) = \boldsymbol{m}/N$ and $g(\boldsymbol{m}) = 2/N$. Moreover, they are equivalent to Eq. (4.104) in [15] introduced to analyze the $(1+\lambda)$-ES and the $(1,\lambda)$-ES. The same normalized step-size has been used for $(\mu/\mu_I, \lambda)$-ES [27,28]. See Section 4.3.1 of [15] for the motivation of this normalization.

Non-isotropic Gaussian sampling Throughout the paper, we assume that the multivariate normal sampling distributions have an isotropic covariance matrix. We can generalize all the following results to an arbitrary positive definite symmetric covariance matrix $\mathbf{C}$ by considering a linear transformation of the search space. Indeed, let $f : x \mapsto \frac{1}{2}(x - x^*)^{\mathrm{T}}\mathbf{A}(x - x^*)$, and consider the coordinate transformation $x \mapsto y = \mathbf{C}^{-\frac{1}{2}}x$. In the latter coordinate system the function f can be written as $f(x) = \bar{f}(y) = \frac{1}{2}(y - \mathbf{C}^{-\frac{1}{2}}x^*)^{\mathrm{T}}(\mathbf{C}^{\frac{1}{2}}\mathbf{A}\mathbf{C}^{\frac{1}{2}})(y - \mathbf{C}^{-\frac{1}{2}}x^*)$. The multivariate normal distribution $\mathcal{N}(\boldsymbol{m}, \sigma^2\mathbf{C})$ is transformed into $\mathcal{N}(\mathbf{C}^{-\frac{1}{2}}\boldsymbol{m}, \sigma^2\mathbf{I})$ by the same transformation. Then, it is easy to prove that the quality gain on the function f given the parameter $(\boldsymbol{m}, \sigma, \mathbf{C})$ is equivalent to the quality gain on the function $\bar{f}$ given $(\mathbf{C}^{-\frac{1}{2}}\boldsymbol{m}, \sigma, \mathbf{I})$. The normalization factor $g(\boldsymbol{m})$ of the quality gain and the normalized step-size are then rewritten as

$$g(\boldsymbol{m}) = \frac{\left\|\mathbf{C}^{\frac{1}{2}}\mathbf{A}(\boldsymbol{m} - x^*)\right\|^2}{f(\boldsymbol{m})\,\mathrm{Tr}(\mathbf{C}^{\frac{1}{2}}\mathbf{A}\mathbf{C}^{\frac{1}{2}})}\ , \qquad \bar{\sigma} = \frac{\sigma c_m \,\mathrm{Tr}(\mathbf{C}^{\frac{1}{2}}\mathbf{A}\mathbf{C}^{\frac{1}{2}})}{\left\|\mathbf{C}^{\frac{1}{2}}\mathbf{A}(\boldsymbol{m} - x^*)\right\|}\ .$$

3.2. Conditional expectation of the weight function

The quadratic objective (10) can be written as

$$f(\boldsymbol{m} + \Delta) = f(\boldsymbol{m}) + \nabla f(\boldsymbol{m})^{\mathrm{T}}\Delta + \frac{1}{2}\Delta^{\mathrm{T}}\mathbf{A}\Delta\ . \tag{11}$$

The normalized quality gain on a convex quadratic function can be written as (using (11) with $\Delta = \boldsymbol{m}^{(t+1)} - \boldsymbol{m}^{(t)}$ and substituting (3))

$$\begin{aligned}\bar{\phi}(\boldsymbol{m}, \bar{\sigma}) = &-\bar{\sigma}\sum_{i=1}^{\lambda}\mathbb{E}\left[\mathbb{E}_i[W(i;(X_k))]\mathbf{e}^{\mathrm{T}}Z_i\right] \\ &-\frac{\bar{\sigma}^2}{2}\sum_{i=1}^{\lambda}\sum_{j=1}^{\lambda}\mathbb{E}\left[\mathbb{E}_{i,j}[W(i;(X_k))W(j;(X_k))]Z_i^{\mathrm{T}}\mathbf{A}Z_j\right]\ ,\end{aligned}$$

where $X_k = \boldsymbol{m}^{(t)} + \sigma^{(t)}Z_k$, and $\mathbb{E}_i$ and $\mathbb{E}_{i,j}$ are the conditional expectations given X_i and (X_i, X_j), respectively.

The following lemma provides the expression of the conditional expectation of the weight function, which allows us to derive the bound for the difference between $\bar{\phi}$ and φ. In the following, let

$$\begin{aligned}P_b(k; n, p) &= \binom{n}{k}p^k(1-p)^{n-k} \\ P_t(k, l; n, p, q) &= \binom{n}{l+k}\binom{l+k}{k}p^k q^l(1-(p+q))^{n-(k+l)}\end{aligned}$$

denote the probability mass functions of the binomial and trinomial distributions, respectively, where $0 \le k \le n$, $0 \le l \le n$, $k + l \le n$, $0 \le p \le 1$, $0 \le q \le 1$ and $p + q \le 1$. The proof of the lemma is provided in Appendix B.2.

Lemma 5. *Let $X \sim \mathcal{N}(\boldsymbol{m}, \sigma^2\boldsymbol{I})$ and $(X_i)_{i=1}^{\lambda}$ be λ i.i.d. copies of X. Let $F_f(t) = \Pr[f(X) < t]$ be the c.d.f. of the function value $f(X)$. Then, we have for any $i, j \in [\![1, \lambda]\!]$, $i \neq j$,*

$$\mathbb{E}_i[W(i;(X_k))] = u_1(F_f(f(X_i))) \ ,$$
$$\mathbb{E}_i[W(i;(X_k))^2] = u_2(F_f(f(X_i))) \ ,$$
$$\mathbb{E}_{i,j}[W(i;(X_k))W(j;(X_k))] = u_3(F_f(f(X_i)), F_f(f(X_j))) \ ,$$

where

$$u_1(p) = \sum_{k=1}^{\lambda} w_k P_b(k-1;\lambda-1,p) \ , \tag{12}$$

$$u_2(p) = \sum_{k=1}^{\lambda} w_k^2 P_b(k-1;\lambda-1,p) \ , \tag{13}$$

$$u_3(p,q) = \sum_{k=1}^{\lambda-1}\sum_{l=k+1}^{\lambda} w_k w_l P_t(k-1,l-k-1;\lambda-2,\min(p,q),|q-p|) \ . \tag{14}$$

Thanks to Lemma 5 and the fact that $(X_k)_{k=1}^{\lambda}$ are i.i.d., we can further rewrite the normalized quality gain as

$$\begin{aligned}\bar{\phi}(\boldsymbol{m},\bar{\sigma}) = &-\bar{\sigma}\lambda\mathbb{E}\big[u_1(F_f(f(X)))\mathbf{e}^{\mathrm{T}}Z\big] - \frac{\bar{\sigma}^2\lambda}{2}\mathbb{E}\big[u_2(F_f(f(X)))\big(Z^{\mathrm{T}}\mathbf{A}Z-1\big)\big] \\ &-\frac{\bar{\sigma}^2\lambda}{2}\mathbb{E}[u_2(F_f(f(X)))] - \frac{\bar{\sigma}^2(\lambda-1)\lambda}{2}\mathbb{E}\big[u_3(F_f(f(X)),F_f(f(\tilde{X})))Z^{\mathrm{T}}\mathbf{A}\tilde{Z}\big] \ .\end{aligned} \tag{15}$$

Here Z and $\tilde{Z}$ are independent and $\mathcal{N}(\mathbf{0},\mathbf{I})$-distributed, and $X = \boldsymbol{m} + \sigma Z$ and $\tilde{X} = \boldsymbol{m} + \sigma\tilde{Z}$, where $\sigma = \bar{\sigma}\,\|\nabla f(\boldsymbol{m})\|/c_m$ is the scale-invariant step-size. Note that X and $\tilde{X}$ are independent and $\mathcal{N}(\boldsymbol{m},\sigma^2\mathbf{I})$-distributed.

The following Lemma shows the Lipschitz continuity of u_1, u_2, and u_3. The proof is provided in Appendix B.3.

Lemma 6. *The functions u_1, u_2, and u_3 are ℓ_1-Lipschitz continuous, i.e., $|u_1(p_1)-u_1(p_2)| \le L_1|p_1-p_2|$, $|u_2(p_1)-u_2(p_2)| \le L_2|p_1-p_2|$, and $|u_3(p_1,q_1)-u_3(p_2,q_2)| \le L_3(|p_1-p_2|+|q_1-q_2|)$, with the Lipschitz constants*

$$L_1 = \sup_{0<p<1}\left|(\lambda-1)\sum_{k=1}^{\lambda-1}(w_{k+1}-w_k)P_b(k-1;\lambda-2,p)\right| \ ,$$

$$L_2 = \sup_{0<p<1}\left|(\lambda-1)\sum_{k=1}^{\lambda-1}(w_{k+1}^2-w_k^2)P_b(k-1;\lambda-2,p)\right| \ ,$$

$$\begin{aligned}L_3 = \max\Bigg[&\sup_{0<p<q<1}\left|\sum_{k=1}^{\lambda-2}\sum_{l=k+2}^{\lambda} w_l(w_{k+1}-w_k)P_t(k-1,l-k-2;\lambda-3,p,q-p)\right| \ , \\ &\sup_{0<p<q<1}\left|\sum_{k=1}^{\lambda-2}\sum_{l=k+2}^{\lambda} w_k(w_l-w_{l-1})P_t(k-1,l-k-2;\lambda-3,p,q-p)\right|\Bigg](\lambda-2) \ .\end{aligned}$$

Upper bounds for the above Lipschitz constants are discussed in Appendix B.4.

3.3. Theorem: normalized quality gain on convex quadratic functions

The following main theorem provides the error bound between $\bar{\phi}$ and φ.

Theorem 7. *Consider Algorithm 1 and let f be a convex quadratic objective function (10). Let the normalized step-size $\bar{\sigma}$ and the normalized quality gain $\bar{\phi}$ defined in Definition 3 and Definition 4, respectively. Let $\boldsymbol{e}_{\boldsymbol{m}} = \nabla f(\boldsymbol{m})/\|\nabla f(\boldsymbol{m})\|$ and $\alpha = \min\big(1,(\bar{\sigma}/c_m)\operatorname{Tr}(\boldsymbol{A}^2)^{1/2}\big)$. Define*

$$G(\alpha) = \min\left[1,\ \alpha\left(2+\frac{2^{\frac{1}{2}}(\ln(1/\alpha))^{\frac{1}{2}}}{\pi^{\frac{1}{2}}}+\frac{d_1(\boldsymbol{A})\ln(1/\alpha)}{(2\pi)^{\frac{1}{2}}\operatorname{Tr}(\boldsymbol{A}^2)^{\frac{1}{2}}}\right)\right] \tag{16}$$

and

$$\varphi(\bar{\sigma}, (w_k), \boldsymbol{e_m}, \boldsymbol{A}) = -\bar{\sigma} \sum_{i=1}^{\lambda} w_i \mathbb{E}[\mathcal{N}_{i:\lambda}] - \frac{\bar{\sigma}^2}{2} \sum_{i=1}^{\lambda} w_i^2 \left(1 - \boldsymbol{e_m}^{\mathrm{T}} \boldsymbol{A} \boldsymbol{e_m}\right) - \frac{\bar{\sigma}^2}{2} \boldsymbol{e_m}^{\mathrm{T}} \boldsymbol{A} \boldsymbol{e_m} \sum_{i=1}^{\lambda} \sum_{j=1}^{\lambda} w_i w_j \mathbb{E}[\mathcal{N}_{i:\lambda} \mathcal{N}_{j:\lambda}] \ , \tag{17}$$

and let L_1, L_2, L_3 be the Lipschitz constants of u_1, u_2 and u_3 defined in Lemma 5, respectively. Then,

$$\sup_{\boldsymbol{m} \in \mathbb{R}^N \setminus \{\boldsymbol{0}\}} \left| \bar{\phi}(\boldsymbol{m}, \bar{\sigma}) - \varphi(\bar{\sigma}, (w_k), \boldsymbol{e_m}, \boldsymbol{A}) \right| \leq \bar{\sigma} \lambda L_1 \left((2/\pi)^{\frac{1}{2}} G(\alpha) + (4\pi)^{-\frac{1}{2}} \alpha \right) + \bar{\sigma} c_m \lambda L_2 \left(2^{-\frac{1}{2}} G(\alpha) + (8\pi)^{-\frac{1}{2}} \alpha \right) \alpha + \bar{\sigma} c_m \lambda (\lambda - 1) L_3 \left((2/\pi)^{\frac{1}{2}} G(\alpha) + (2\pi^2)^{-\frac{1}{2}} \alpha \right) \alpha \ . \tag{18}$$

The above theorem claims that if the right-hand side (RHS) of (18) is sufficiently small, the normalized quality gain $\bar{\phi}$ is approximated by the asymptotic normalized quality gain φ defined in (17). Compared to $\bar{\phi}_\infty$ in (6) derived for the infinite dimensional sphere function, φ is different even when $\mathbf{A} \propto \mathbf{I}$. We investigate the properties of φ in Section 4.1. The situations where the RHS of (18) is sufficiently small are discussed in Section 4.2 and Section 4.3. We remark that Theorem 3.4 in [1] provides a bound for the difference between $\bar{\phi}$ and $\bar{\phi}_\infty$, instead of the difference between $\bar{\phi}$ and φ. Introducing φ allows us to consider a finite dimensional case and to derive a tighter bound.

3.4. Outline of the proof of the main theorem

In the following of the section and in Appendix B, let $Z_{\mathbf{e}} = \mathbf{e}^{\mathrm{T}} Z$, $Z_\perp = Z - Z_{\mathbf{e}} \mathbf{e}$, and $X = \boldsymbol{m} + \sigma Z$ for $Z \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Then, $Z_{\mathbf{e}} \sim \mathcal{N}(0, 1)$ and $Z_\perp \sim \mathcal{N}(\mathbf{0}, \mathbf{I} - \mathbf{e}\mathbf{e}^{\mathrm{T}})$ and they are independent. Define

$$H_N = \frac{f(\boldsymbol{m} + \sigma Z) - \mathbb{E}[f(\boldsymbol{m} + \sigma Z)]}{\sigma \, \|\nabla f(\boldsymbol{m})\|} \qquad \text{and} \qquad h(Z) = \frac{1}{2} \frac{\bar{\sigma}}{c_m} (Z^{\mathrm{T}} \mathbf{A} Z - 1) \ ,$$

where $\mathbb{E}[f(\boldsymbol{m} + \sigma Z)] = f(\boldsymbol{m}) + \sigma^2/2$. It is easy to see that $H_N = Z_{\mathbf{e}} + h(Z)$. Let F_f and F_N be the c.d.f. induced by $f(X)$ and H_N, respectively. Then, $F_f(f(X)) = F_N(H_N)$. Let $\tilde{Z}$ be the i.i.d. copy of Z and define $\tilde{Z}_{\mathbf{e}}$, $\tilde{Z}_\perp$, $\tilde{X}$, and $\tilde{H}_N$ analogously.

The first lemma allows us to approximate F_N (hence F_f) with the c.d.f. Φ of the standard normal distribution. The proof is based on the Lipschitz continuity of Φ and the tail bound of $h(Z)$ proved in Lemma 1 of [32]. The detail is provided in Appendix B.5.

Lemma 8. *Let α and $G(\alpha)$ be defined in Theorem 7. Then, $\sup_{t \in \mathbb{R}} |F_N(t) - \Phi(t)| \leq G(\alpha)$.*

The following three lemmas are used to bound each term on the RHS of (15). The proofs are straight-forward from the Lipschitz continuity of u_1, u_2 and u_3 and Lemma 8. The detailed proofs are found in Appendix B.6, Appendix B.7, and Appendix B.8, respectively.

Lemma 9. *Let L_1, α, and $G(\alpha)$ be the quantities appeared in Lemma 6 and Theorem 7. Then,*

$$\left| \mathbb{E}[u_1(F_f(f(X)))Z_{\boldsymbol{e}}] - \mathbb{E}[u_1(\Phi(Z_{\boldsymbol{e}}))Z_{\boldsymbol{e}}] \right| \leq L_1 \left((2/\pi)^{\frac{1}{2}} G(\alpha) + (4\pi)^{-\frac{1}{2}} \alpha \right) \ .$$

Lemma 10. *Let L_2, α, and $G(\alpha)$ be the quantities appeared in Lemma 6 and Theorem 7. Then,*

$$\left| \mathbb{E}[u_2(F_f(f(X)))(Z^{\mathrm{T}} \boldsymbol{A} Z - 1)] - \mathbb{E}[u_2(\Phi(Z_{\boldsymbol{e}}))(Z^{\mathrm{T}} \boldsymbol{A} Z - 1)] \right| \leq L_2 \left(2^{\frac{1}{2}} G(\alpha) + (2\pi)^{-\frac{1}{2}} \alpha \right) \mathrm{Tr}(\boldsymbol{A}^2)^{\frac{1}{2}} \ .$$

Lemma 11. *Let L_3, α, and $G(\alpha)$ be the quantities appeared in Lemma 6 and Theorem 7. Then,*

$$\left| \mathbb{E}[u_3(F_f(f(X)), F_f(f(\tilde{X})))Z^{\mathrm{T}} \boldsymbol{A} \tilde{Z}] - \mathbb{E}[u_3(\Phi(Z_{\boldsymbol{e}}), \Phi(\tilde{Z}_{\boldsymbol{e}}))Z^{\mathrm{T}} \boldsymbol{A} \tilde{Z}] \right| \leq L_3 \left((8/\pi)^{\frac{1}{2}} G(\alpha) + (2^{\frac{1}{2}}/\pi) \alpha \right) \mathrm{Tr}(\boldsymbol{A}^2)^{\frac{1}{2}} \ .$$

The asymptotic normalized quality gain φ in (17) is obtained by replacing the c.d.f. F_f of $f(X)$ in (15) with the c.d.f. Φ of the standard normal distribution. The above lemmas are used to bound the difference between $\bar{\phi}$ and φ. The following lemma provides the explicit form of each term of (17). The proof is straight-forward in light of Lemma 5. The detail can be found in Appendix B.9.

Lemma 12. *The functions u_1, u_2 and u_3 defined in Lemma 5 satisfy the following properties:*

$$\lambda \mathbb{E}[u_1(\Phi(Z_{\boldsymbol{e}}))Z_{\boldsymbol{e}}] = \sum_{i=1}^{\lambda} w_i \mathbb{E}[\mathcal{N}_{i:\lambda}] \ , \tag{19}$$

$$\lambda \mathbb{E}[u_2(\Phi(Z_{\boldsymbol{e}}))] = \sum_{i=1}^{\lambda} w_i^2 \ , \tag{20}$$

$$\lambda \mathbb{E}[u_2(\Phi(Z_{\boldsymbol{e}}))(Z^{\mathrm{T}}\boldsymbol{A}Z - 1)] = \boldsymbol{e}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{e} \sum_{i=1}^{\lambda} w_i^2 (\mathbb{E}[\mathcal{N}_{i:\lambda}^2] - 1) \ , \tag{21}$$

$$\lambda(\lambda-1)\mathbb{E}[u_3(\Phi(Z_{\boldsymbol{e}}), \Phi(\tilde{Z}_{\boldsymbol{e}}))Z^{\mathrm{T}}\boldsymbol{A}\tilde{Z}] = 2\boldsymbol{e}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{e} \sum_{k=1}^{\lambda-1} \sum_{l=k+1}^{\lambda} w_k w_l \mathbb{E}[\mathcal{N}_{k:\lambda}\mathcal{N}_{l:\lambda}] \ . \tag{22}$$

Now we finalize the proof of the main theorem. Using Lemma 12, we can rewrite (17) as

$$\begin{aligned}\varphi(\bar{\sigma}, (w_k)_{k=1}^{\lambda}, \mathbf{e}_{\boldsymbol{m}}, \mathbf{A}) = &-\bar{\sigma}\lambda \mathbb{E}[u_1(\Phi(Z_{\mathbf{e}}))Z_{\mathbf{e}}] - \frac{\bar{\sigma}^2}{2}\lambda \mathbb{E}[u_2(\Phi(Z_{\mathbf{e}}))] \\ &- \frac{\bar{\sigma}^2}{2}\lambda \mathbb{E}[u_2(\Phi(Z_{\mathbf{e}}))(Z^{\mathrm{T}}\mathbf{A}Z - 1)] - \frac{\bar{\sigma}^2}{2}\lambda(\lambda-1)\mathbb{E}[u_3(\Phi(Z_{\mathbf{e}}), \Phi(\tilde{Z}_{\mathbf{e}}))Z^{\mathrm{T}}\mathbf{A}\tilde{Z}] \ .\end{aligned}$$

From the equation (15) and the above expression of φ, we have

$$\begin{aligned}&\bar{\phi}(\boldsymbol{m}, \bar{\sigma}) - \varphi(\bar{\sigma}, (w_k)_{k=1}^{\lambda}, \mathbf{e}_{\boldsymbol{m}}, \mathbf{A}) \\ &= -\bar{\sigma}\lambda \mathbb{E}\left[(u_1(F_f(f(X))) - u_1(\Phi(Z_{\mathbf{e}})))Z_{\mathbf{e}}\right] \\ &\quad - \frac{\bar{\sigma}^2\lambda}{2}\mathbb{E}[(u_2(F_f(f(X))) - u_2(\Phi(Z_{\mathbf{e}})))] \\ &\quad - \frac{\bar{\sigma}^2\lambda}{2}\mathbb{E}\left[(u_2(F_f(f(X))) - u_2(\Phi(Z_{\mathbf{e}})))\left(Z^{\mathrm{T}}\mathbf{A}Z - 1\right)\right] \\ &\quad - \frac{\bar{\sigma}^2(\lambda-1)\lambda}{2}\mathbb{E}\left[(u_3(F_f(f(X)), F_f(f(\tilde{X}))) - u_3(\Phi(Z_{\mathbf{e}}), \Phi(\tilde{Z}_{\mathbf{e}})))Z^{\mathrm{T}}\mathbf{A}\tilde{Z}\right] \ .\end{aligned}$$

From the well-known fact (e.g., Theorem 2.1 of [33]) that for a random variable X with a continuous c.d.f. F_x the random variable $F_x(X)$ is uniformly distributed on $[0, 1]$, we can prove both $F_f(f(X))$ and $\Phi(Z_{\mathbf{e}})$ are uniformly distributed on $[0, 1]$. Therefore, we have $\mathbb{E}[u_2(F_f(f(X)))] = \mathbb{E}[u_2(\Phi(Z_{\mathbf{e}}))] = \mathbb{E}[u_2(\mathcal{U}[0, 1])]$, and the second term on the RHS of the above equality is zero. Applying the triangular inequality and Lemma 9, Lemma 10, and Lemma 11, we obtain (18). It completes the proof of Theorem 7.

4. Consequences

Theorem 7 tells that if the RHS of (18) is sufficiently small, the normalized quality gain $\bar{\phi}(\boldsymbol{m}, \bar{\sigma})$ is well approximated by $\varphi(\bar{\sigma}, (w_k)_{k=1}^{\lambda}, \mathbf{e}_{\boldsymbol{m}}, \mathbf{A})$ defined in (17). First we investigate the parameter values that are optimal for φ. Then, we consider the situations when the RHS of (18) is sufficiently small.

Let $\boldsymbol{n}_{(\lambda)}$ be the λ dimensional column vector whose i-th component is $\mathbb{E}[\mathcal{N}_{i:\lambda}]$ and $\mathbf{N}_{(\lambda)}$ be the λ dimensional symmetric matrix whose (i, j)-th elements are $\mathbb{E}[\mathcal{N}_{i:\lambda}\mathcal{N}_{j:\lambda}]$. Let $\boldsymbol{w}$ and $\bar{\boldsymbol{w}}$ be the λ dimensional column vector whose i-th element is w_i and $\bar{\sigma}w_i$, respectively. Now (17) can be written as

$$\varphi(\bar{\boldsymbol{w}}, \mathbf{e}, \mathbf{A}) = -\bar{\boldsymbol{w}}^{\mathrm{T}}\boldsymbol{n}_{(\lambda)} - \frac{1}{2}\left(1 - \mathbf{e}^{\mathrm{T}}\mathbf{A}\mathbf{e}\right)\bar{\boldsymbol{w}}^{\mathrm{T}}\bar{\boldsymbol{w}} - \frac{1}{2}(\mathbf{e}^{\mathrm{T}}\mathbf{A}\mathbf{e})\bar{\boldsymbol{w}}^{\mathrm{T}}\mathbf{N}_{(\lambda)}\bar{\boldsymbol{w}} \ . \tag{23}$$

In the following we use the following asymptotically true approximation for a sufficiently large λ (see (A.2) in Appendix A)

$$\frac{\bar{\boldsymbol{w}}^{\mathrm{T}}\mathbf{N}_{(\lambda)}\bar{\boldsymbol{w}}}{\lambda\|\bar{\boldsymbol{w}}\|^2} \approx \frac{(\bar{\boldsymbol{w}}^{\mathrm{T}}\boldsymbol{n}_{(\lambda)})^2}{\|\bar{\boldsymbol{w}}\|^2\left\|\boldsymbol{n}_{(\lambda)}\right\|^2} \approx \frac{(\bar{\boldsymbol{w}}^{\mathrm{T}}\boldsymbol{n}_{(\lambda)})^2}{\lambda\|\bar{\boldsymbol{w}}\|^2} \ . \tag{24}$$

By *"for a sufficiently large λ"*, we mean for a λ large enough to approximate the left hand side (LHS) of (24) by the right-most side (RMS). For a sufficiently large λ, (23) is approximated by

$$\varphi(\bar{\boldsymbol{w}}, \mathbf{e}, \mathbf{A}) \approx -\bar{\boldsymbol{w}}^{\mathrm{T}}\boldsymbol{n}_{(\lambda)} - \frac{1}{2}\left(1 - \mathbf{e}^{\mathrm{T}}\mathbf{A}\mathbf{e}\right)\bar{\boldsymbol{w}}^{\mathrm{T}}\bar{\boldsymbol{w}} - \frac{1}{2}(\mathbf{e}^{\mathrm{T}}\mathbf{A}\mathbf{e})(\bar{\boldsymbol{w}}^{\mathrm{T}}\boldsymbol{n}_{(\lambda)})^2 \ . \tag{25}$$

4.1. Asymptotic normalized quality gain and optimal parameters

As we mentioned in the previous section, φ in (17) is different from the normalized quality gain limit $\bar{\phi}_\infty$ in (6). Consider the sphere function $\mathbf{A} = \mathbf{I}/N$; then, since $\mathbf{e}^\mathrm{T}\mathbf{A}\mathbf{e} = 1/N$ for any $\mathbf{e}$ with $\|\mathbf{e}\| = 1$, we have

$$\varphi(\bar{\sigma}, (w_k), \mathbf{e}, \mathbf{A}) = \bar{\phi}_\infty(\bar{\sigma}, (w_k)_{k=1}^\lambda) + \frac{\bar{\sigma}^2}{2N}\sum_{i=1}^\lambda w_i^2 - \frac{\bar{\sigma}^2}{2N}\sum_{i=1}^\lambda\sum_{j=1}^\lambda w_i w_j \mathbb{E}[\mathcal{N}_{i:\lambda}\mathcal{N}_{j:\lambda}] \ .$$

Note that the second and the third terms on the RHS are proportional to $1/N$. By taking the limit for N to infinity, we have $\varphi = \bar{\phi}_\infty$. Therefore, the second and third terms describe how the finite dimensional cases are different from the infinite dimensional case. In particular, the last term prevents the quality gain from scaling up proportionally to λ when $\lambda \ll N$.

Optimal recombination weights The recombination weights optimal for φ are provided in the following proposition.

Proposition 13. *The asymptotic normalized quality gain φ (17) is optimized when $\bar{\boldsymbol{w}}$ is the solution to the following linear system of equations*

$$(\boldsymbol{I} + \boldsymbol{e}^\mathrm{T}\boldsymbol{A}\boldsymbol{e}(\boldsymbol{N}_{(\lambda)} - \boldsymbol{I}))\bar{\boldsymbol{w}} = -\boldsymbol{n}_{(\lambda)} \ , \tag{26}$$

where $\bar{\sigma}$ and w_i are uniquely determined using the condition $\sum_{i=1}^\lambda |w_i| = 1$. Then the optimal value of φ is $-\frac{1}{2}\boldsymbol{n}_{(\lambda)}^\mathrm{T}\bar{\boldsymbol{w}}^*$ where $\bar{\boldsymbol{w}}^*$ is the solution to the linear system (26).

Proof. We obtain (26) by taking the derivative of (23) with respect to $\bar{\boldsymbol{w}}$, and requiring $\partial\varphi(\bar{\boldsymbol{w}}, \mathbf{e}, \mathbf{A})/\partial[\bar{\boldsymbol{w}}]_i = 0$. This ends the proof. □

First, consider the limit for $N \to \infty$ while λ is fixed. As long as the largest eigenvalue $d_1(\mathbf{A})$ of $\mathbf{A}$ converges to zero as $N \to \infty$, i.e., $\lim_{N\to\infty} d_1(\mathbf{A}) = 0$, we have $\mathbf{e}^\mathrm{T}\mathbf{A}\mathbf{e} \to 0$ as $N \to 0$. Then, (26) reads $\bar{\boldsymbol{w}} = -\boldsymbol{n}_{(\lambda)}$. Therefore, we have the same optimal recombination weights as the ones derived for the infinite dimensional sphere function.

Next, consider a finite dimensional case. If λ is sufficiently large, the optimality condition (26) is approximated by

$$\left(\frac{1}{\lambda}\left(1 - \frac{1}{\mathbf{e}^\mathrm{T}\mathbf{A}\mathbf{e}}\right)\mathbf{I} + \frac{1}{\mathbf{e}^\mathrm{T}\mathbf{A}\mathbf{e}}\frac{\boldsymbol{n}_{(\lambda)}\boldsymbol{n}_{(\lambda)}^\mathrm{T}}{\|\boldsymbol{n}_{(\lambda)}\|^2}\right)\bar{\boldsymbol{w}} = -\frac{\boldsymbol{n}_{(\lambda)}}{\lambda} \ .$$

The solution to the above approximated condition is given by $\bar{\boldsymbol{w}} \propto -\boldsymbol{n}_{(\lambda)}$ independently of $\mathbf{A}$ and $\mathbf{e}$. It means, for a sufficiently large λ, the optimal recombination weights are approximated by the weights (8) optimal for the infinite dimensional sphere function.

Optimal normalized step-size The optimal $\bar{\sigma}$ under a given $(w_k)_{k=1}^\lambda$ is provided in the following proposition.

Proposition 14. *Given $\boldsymbol{w} = (w_1, \ldots, w_\lambda)$, the asymptotic normalized quality gain (17) is maximized when the normalized step-size $\bar{\sigma}$ is*

$$\bar{\sigma}^* = \frac{-\sum_{i=1}^\lambda w_i \mathbb{E}[\mathcal{N}_{i:\lambda}]}{\sum_{i=1}^\lambda w_i^2\left(1 - \boldsymbol{e}_{\boldsymbol{m}}^\mathrm{T}\boldsymbol{A}\boldsymbol{e}_{\boldsymbol{m}}\right) + \boldsymbol{e}_{\boldsymbol{m}}^\mathrm{T}\boldsymbol{A}\boldsymbol{e}_{\boldsymbol{m}}\sum_{i=1}^\lambda\sum_{j=1}^\lambda w_i w_j \mathbb{E}[\mathcal{N}_{i:\lambda}\mathcal{N}_{j:\lambda}]} \ , \tag{27}$$

then $\varphi(\bar{\sigma}^*, (w_k), \boldsymbol{e}, \boldsymbol{A}) = \frac{\bar{\sigma}^*}{2}\left(-\sum_{i=1}^\lambda w_i \mathbb{E}[\mathcal{N}_{i:\lambda}]\right)$.

Proof. It is a straight-forward consequence from differentiating (17) with respect to $\bar{\sigma}$ and solving $\partial\varphi/\partial\bar{\sigma} = 0$. □

For a sufficiently large λ (see (24)), one can rewrite and approximate (27) as

$$\begin{aligned}\bar{\sigma}^* &= \frac{-\boldsymbol{w}^\mathrm{T}\boldsymbol{n}_{(\lambda)}}{\left(1 - \mathbf{e}_{\boldsymbol{m}}^\mathrm{T}\mathbf{A}\mathbf{e}_{\boldsymbol{m}}\right)\|\boldsymbol{w}\|^2 + \mathbf{e}_{\boldsymbol{m}}^\mathrm{T}\mathbf{A}\mathbf{e}_{\boldsymbol{m}}\,\boldsymbol{w}^\mathrm{T}\mathbf{N}_{(\lambda)}\boldsymbol{w}} \\ &\approx \frac{(\mathbf{e}_{\boldsymbol{m}}^\mathrm{T}\mathbf{A}\mathbf{e}_{\boldsymbol{m}})^{-1}\mu_w(-\boldsymbol{w}^\mathrm{T}\boldsymbol{n}_{(\lambda)})}{(\mathbf{e}_{\boldsymbol{m}}^\mathrm{T}\mathbf{A}\mathbf{e}_{\boldsymbol{m}})^{-1} - 1 + \mu_w(-\boldsymbol{w}^\mathrm{T}\boldsymbol{n}_{(\lambda)})^2} \ .\end{aligned} \tag{28}$$

Note again that $-\boldsymbol{w}^\mathrm{T}\boldsymbol{n}_{(\lambda)} = -\sum_{i=1}^\lambda w_i \mathbb{E}[\mathcal{N}_{i:\lambda}] \in O(1)$ for the optimal weights, CMA-type non-negative weights, and truncation weights with fixed truncation ratio. To provide a better insight, consider the case of the sphere function ($\mathbf{A} = \mathbf{I}/N$). Then, the RMS of (28) reads

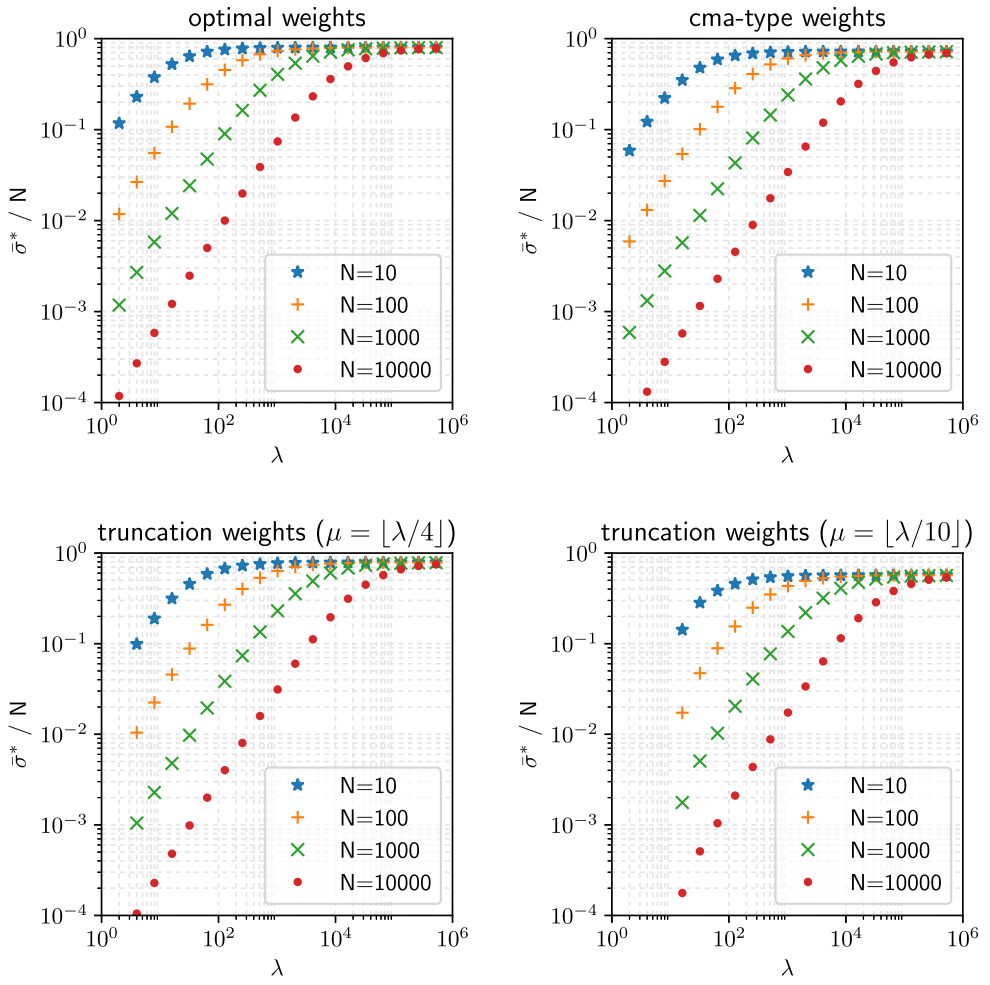


Fig. 2. Optimal normalized step-size on $N = 10, 100, 1000, 10000$ dimensional sphere function for different weight schemes and different population size λ.

$$\bar{\sigma}^* \approx \frac{N\mu_w(-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{n}_{(\lambda)})}{N - 1 + \mu_w(-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{n}_{(\lambda)})^2} \ . \tag{29}$$

Then, we find the following: (i) if $N \gg \mu_w$, then $\bar{\sigma}^* \approx \mu_w(-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{n}_{(\lambda)})$ and $\varphi \approx \mu_w(-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{n}_{(\lambda)})^2/2$; (ii) if $\mu_w \gg N$, then $\bar{\sigma}^* \approx N/(-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{n}_{(\lambda)})$ and $\varphi \approx N/2$. Fig. 2 visualizes the optimal normalized step-size (27) for various $\boldsymbol{w}$ on the sphere function. The optimal normalized step-size (27) scales linearly for $\lambda \leq N$ and it tends to level out for $\lambda > N$.

Geometric interpretation of the optimal situation On the infinite dimensional sphere function, we know that the optimal step-size puts the algorithm in the situation where $f(\boldsymbol{m})$ improves twice as much by $\boldsymbol{m}$ moving towards the optimum as it deteriorates by $\boldsymbol{m}$ moving randomly in the subspace orthogonal to the gradient direction [13]. On a finite dimensional convex quadratic function, we find the analogous result. From (15) and lemmas in Section 3.4, the first term of the asymptotic normalized quality gain (17), i.e. $-\bar{\sigma}\sum_{i=1}^{\lambda} w_i \mathbb{E}[\mathcal{N}_{i:\lambda}]$, is due to the movement of $\boldsymbol{m}$ in negative gradient direction, and the second and third terms are due to the random walk in the orthogonal subspaces.[6] The asymptotic normalized quality gain is maximized when the normalized step-size is set such that the first term is twice as large as the absolute value of the sum of the second and the third terms. That is, the amount of the decrease of $f(\boldsymbol{m})$ by $\boldsymbol{m}$ moving into the negative gradient direction is twice greater than the increase of $f(\boldsymbol{m})$ by $\boldsymbol{m}$ moving in its orthogonal subspace.

[6] More precisely, the second and the third terms come from the quadratic term in (11) that contains the information in the gradient direction as well. However, the above statement is true in the limit $N \to \infty$ as long as $\mathrm{Tr}(\mathbf{A}^2) \to 0$.

4.2. Infinitesimal step-size case

The RHS of (18), the error bound between $\bar{\phi}$ and φ, converges to zero when $\alpha \to 0$. One such situation is the limit of $\sigma / \|\boldsymbol{m}\| \to 0$ while $c_m \sigma$ can remain positive, i.e., in the *mutate small, but inherit large* situation, which is formally stated in the next corollary.

Corollary 15. *For any positive constant* $\mathrm{C} > 0$*,*

$$\lim_{\sigma/\|\boldsymbol{m}\| \to 0} \sup_{\bar{\sigma} \in (0, \mathrm{C}]} \sup_{\boldsymbol{m} \in \mathbb{R}^N \setminus \{\mathbf{0}\}} \frac{1}{\bar{\sigma}} \left| \bar{\phi}(\boldsymbol{m}, \bar{\sigma}) - \varphi(\bar{\sigma}, (w_k), \boldsymbol{e}_{\boldsymbol{m}}, \boldsymbol{A}) \right| = 0 \tag{30}$$

Proof. Note that the function $G(\alpha) \in \mathcal{O}(\alpha \ln(1/\alpha))$ as $\alpha \to 0$. Then, (18) reads

$$\begin{aligned} &\sup_{\boldsymbol{m} \in \mathbb{R}^N \setminus \{\mathbf{0}\}} \left| \bar{\phi}(\boldsymbol{m}, \bar{\sigma}) - \varphi(\bar{\sigma}, (w_k), \mathbf{e}_{\boldsymbol{m}}, \mathbf{A}) \right| \\ &\quad \in \bar{\sigma} \lambda \mathcal{O}(\alpha \ln(1/\alpha)) \left[L_1 + c_m \alpha L_2 + c_m \alpha (\lambda - 1) L_3 \right] . \end{aligned} \tag{31}$$

Note also that

$$\alpha = \frac{\bar{\sigma}}{c_m} \operatorname{Tr}(\mathbf{A}^2)^{\frac{1}{2}} = \frac{\sigma \operatorname{Tr}(\mathbf{A}^2)^{\frac{1}{2}}}{\|\nabla f(\boldsymbol{m})\|} \leq \frac{\sigma}{\|\boldsymbol{m}\|} \frac{\operatorname{Tr}(\mathbf{A}^2)^{\frac{1}{2}}}{d_N(\mathbf{A})}$$

and $\alpha c_m = \bar{\sigma} \operatorname{Tr}(\mathbf{A}^2)^{\frac{1}{2}} \leq C \operatorname{Tr}(\mathbf{A}^2)^{\frac{1}{2}}$. It implies that the RHS of (31) divided by $\bar{\sigma}$ is in $O(\alpha \ln(1/\alpha)) \subseteq o(\alpha^{1-\epsilon})$ for any $\epsilon > 0$ under the condition $\bar{\sigma} \leq \mathrm{C}$. Since $\alpha \to 0$ as $\sigma / \|\boldsymbol{m}\| \to 0$, (31) implies (30). □

If c_m is fixed, we have $\bar{\sigma} \to 0$ as $\sigma / \|\boldsymbol{m}\| \to 0$. Then, the asymptotic normalized quality gain (17) converges towards zero as the bound on the RHS of (18) goes to zero. The above corollary tells that the bound converges faster than $\bar{\sigma}$ does, while the asymptotic normalized quality gain decreases linearly in $\bar{\sigma}$. As a consequence, we find that the normalized quality gain approaches $-\bar{\sigma} \sum_{i=1}^{\lambda} w_i \mathbb{E}[\mathcal{N}_{i:\lambda}]$ as $\sigma / \|\boldsymbol{m}\| \to 0$.

Consider the case that $\bar{\sigma}$ is fixed, i.e., $c_m \sigma$ is fixed. Then, from the corollary we find that the normalized quality gain converges towards φ in (17) as $\sigma / \|\boldsymbol{m}\| \to 0$.

Taking $c_m \to \infty$ we obtain $\bar{\phi} \to \varphi$. Though resembling a numerical gradient estimation when $\sigma \to 0$, it is not quite practical to take a large c_m. Indeed we usually rather do the opposite. For noisy optimization, the idea of *rescaled mutations* (corresponding to a large σ and a small c_m) that was proposed by A. Ostermeier in 1993 (according to [13]) and analyzed in [34,35] is introduced to reduce the noise-to-signal ratio. If neither $c_m \gg 1$ nor $N \gg 1$, the RHS of (18) will not be small enough to approximate the normalized quality gain by φ in (17). Then the normalized step-size defined in (27) is not guaranteed to provide an approximation of the optimal normalized step-size. However, in practice, we observe that the normalized step-size defined in (27) provides a reasonable approximation of the optimal normalized step-size for $c_m \geq 1$. We will see it in Fig. 4.

4.3. Infinite dimensional case

The other situation when $\alpha \to 0$ occurs is the limit $N \to \infty$ under the condition $\lim_{N \to \infty} \operatorname{Tr}(\mathbf{A}^2) = 0$. In this case, since $\mathbf{e}^{\mathrm{T}} \mathbf{A} \mathbf{e} \leq \operatorname{Tr}(\mathbf{A}^2)^{\frac{1}{2}} \to 0$, the limit expression φ converges to $\bar{\phi}_\infty$, i.e., the same limit on the sphere function. It is stated in the following corollary, which is a generalization of the result obtained in [9] from the sphere function to a general convex quadratic function.

Corollary 16. *Let* $(\boldsymbol{A}_N)_{N=1}^{\infty}$ *be the sequence of nonnegative definite matrices satisfying* $\lim_{N \to \infty} \operatorname{Tr}(\boldsymbol{A}^2)^{\frac{1}{2}} = 0$*. Then,*

$$\lim_{N \to \infty} \sup_{\boldsymbol{m} \in \mathbb{R}^N \setminus \{\mathbf{0}\}} \left| \varphi(\bar{\sigma}, (w_k)_{k=1}^{\lambda}, \boldsymbol{e}, \boldsymbol{A}_N) - \bar{\phi}_\infty(\bar{\sigma}, (w_k)_{k=1}^{\lambda}) \right| = 0 \ , \tag{32}$$

where $\bar{\phi}_\infty(\bar{\sigma}, (w_k)_{k=1}^{\lambda}) = -\bar{\sigma} \sum_{i=1}^{\lambda} w_i \mathbb{E}[\mathcal{N}_{i:\lambda}] - \bar{\sigma}^2 / (2\mu_w)$ *as defined in* (6)*. Moreover,*

$$\lim_{N \to \infty} \sup_{\bar{\sigma} \in (0, \mathrm{C}]} \sup_{\boldsymbol{m} \in \mathbb{R}^N \setminus \{\mathbf{0}\}} \left| \bar{\phi}(\boldsymbol{m}, \bar{\sigma}) - \bar{\phi}_\infty(\bar{\sigma}, (w_k)_{k=1}^{\lambda}) \right| = 0 \ . \tag{33}$$

Corollary 16 shows that the normalized quality gain on a convex quadratic function converges towards the asymptotic normalized quality gain derived on the infinite dimensional sphere function as $\operatorname{Tr}(\mathbf{A}^2)^{\frac{1}{2}} \to 0$. It implies that the optimal values of the recombination weights and the normalized step-size are independent of the Hessian of the objective function,

and given by (8). It is a nice feature since we do not need to tune the weight values depending on the function. Since any twice continuously differentiable function is locally approximated by a quadratic function, the optimal weights derived here are expected to be locally optimal on any twice continuously differentiable function.

In the above corollary, the population size λ is a constant over the dimension N. However, in the default setting of the CMA-ES, the population size is $\lambda = 4 + \lfloor 3\ln(N) \rfloor$, meaning that the population size is unbounded. If the population size increases to infinity as $N \to \infty$, it is not guaranteed that the per-evaluation progress $\bar{\phi}/\lambda$ converges to $\bar{\phi}_\infty/\lambda$ as $N \to \infty$. The following proposition provides a sufficient condition on the recombination weights and the population size such that the per-evaluation progress $\bar{\phi}/\lambda$ converges to $\bar{\phi}_\infty/\lambda$ when λ increases as N increases.

Proposition 17. *Let $(\boldsymbol{A}_N)_{N=1}^\infty$ be the sequence of the Hessian matrix that satisfies $\lim_{N\to\infty} \mathrm{Tr}(\boldsymbol{A}_N^2)^{\frac{1}{2}} = 0$. Let $(\lambda_N)_{N=1}^\infty$ be the sequence of the population size and $(w_k^N)_{k=1}^{\lambda_N}$ be the sequence of the weights for the population size λ_N. Suppose for an arbitrarily small positive ϵ,*

$$\lambda_N^2 \in o\left(\frac{1}{d_1(\boldsymbol{A}_N)}\right) \quad \textit{and}$$
$$\max\left(\lambda_N, L_1^{\frac{1}{1-\epsilon}}\lambda_N^{\frac{2-\epsilon}{1-\epsilon}}, L_2^{\frac{1}{2-\epsilon}}\lambda_N^{\frac{3-\epsilon}{2-\epsilon}}, L_3^{\frac{1}{2-\epsilon}}\lambda_N^{\frac{4-\epsilon}{2-\epsilon}}\right) \in O\left(\frac{1}{\mathrm{Tr}(\boldsymbol{A}_N^2)^{\frac{1}{2}}}\right) .$$

Then,

$$\lim_{N\to\infty} \sup_{\bar{\sigma}\in(0,2\bar{\sigma}^*)} \sup_{\boldsymbol{m}\in\mathbb{R}^N\setminus\{\boldsymbol{0}\}} \frac{1}{\lambda_N} \left| \bar{\phi}(\boldsymbol{m}, \bar{\sigma}) - \bar{\phi}_\infty(\bar{\sigma}, (w_k^N)_{k=1}^{\lambda_N}) \right| = 0 , \tag{34}$$

where $\bar{\sigma}^*$ is the normalized step-size optimal for $\bar{\phi}_\infty(\bar{\sigma}, (w_k^N)_{k=1}^{\lambda_N})$, which is formulated in (7).

Proof. A sufficient condition for φ to converge to $\bar{\phi}_\infty$ for $\bar{\sigma} \in (0, 2\bar{\sigma}^*)$ is that the third term on the RHS of (17) converges to zero as $N \to \infty$. As we know from Appendix A that $\boldsymbol{w}^{\mathrm{T}}\mathbf{N}_{(\lambda)}\boldsymbol{w} \le d_1(\mathbf{N}_{(\lambda)})\|\boldsymbol{w}\|^2 \le \mathrm{Tr}(\mathbf{N}_{(\lambda)})\|\boldsymbol{w}\|^2 = \lambda/\mu_w$. On the other hand, $\mathbf{e}_{\boldsymbol{m}}^{\mathrm{T}}\mathbf{A}\mathbf{e}_{\boldsymbol{m}}$ is no greater than the greatest eigen value $d_1(\mathbf{A})$ of $\mathbf{A}$. The third term on the RHS of (17) is maximized when $\bar{\sigma} = \bar{\sigma}^* = \mu_w(-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{n}_{(\lambda)})$, where, $\mu_w(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{n}_{(\lambda)}) \le \sum_{i=1}^\lambda |\mathbb{E}[\mathcal{N}_{i:\lambda}]|$ and $\frac{1}{\lambda}\sum_{i=1}^\lambda |\mathbb{E}[\mathcal{N}_{i:\lambda}]| \to (2/\pi)^{\frac{1}{2}}$. From these arguments derives that the third term on the RHS of (17) converges to zero as $N \to \infty$ provided that $\lambda^2 d_1(\mathbf{A}) \to 0$ as $N \to 0$.

Next we consider the convergence of the bound (RHS of (31)). Remember that $\alpha = (\bar{\sigma}/c_m)\mathrm{Tr}(\mathbf{A}^2)^{\frac{1}{2}}$ and $G(\alpha) \in O(\alpha\ln(1/\alpha))$. For $\bar{\sigma} \in (0, 2\bar{\sigma}^*)$, we have $(\bar{\sigma}/c_m)\mathrm{Tr}(\mathbf{A}^2)^{\frac{1}{2}} \le 2(\bar{\sigma}^*/c_m)\mathrm{Tr}(\mathbf{A}^2)^{\frac{1}{2}}$. Since $\bar{\sigma}^* \in O(\lambda)$, we have $\bar{\sigma} \in O(\lambda)$ and $\alpha \in O(\lambda\,\mathrm{Tr}(\mathbf{A}^2)^{\frac{1}{2}})$. Then, the RHS of (31) divided by λ,

$$\begin{aligned} &O\left(\bar{\sigma}\alpha\ln(1/\alpha)[L_1 + L_2\alpha + L_3\lambda\alpha]\right) \\ &\quad\subseteq o\left(\bar{\sigma}\alpha^{1-\epsilon}[L_1 + L_2\alpha + L_3\lambda\alpha]\right) \\ &\quad\subseteq o\left(\lambda^{2-\epsilon}\,\mathrm{Tr}(\mathbf{A}^2)^{\frac{1-\epsilon}{2}}[L_1 + L_2\lambda\,\mathrm{Tr}(\mathbf{A}^2)^{\frac{1}{2}} + L_3\lambda^2\,\mathrm{Tr}(\mathbf{A}^2)^{\frac{1}{2}}]\right), \end{aligned}$$

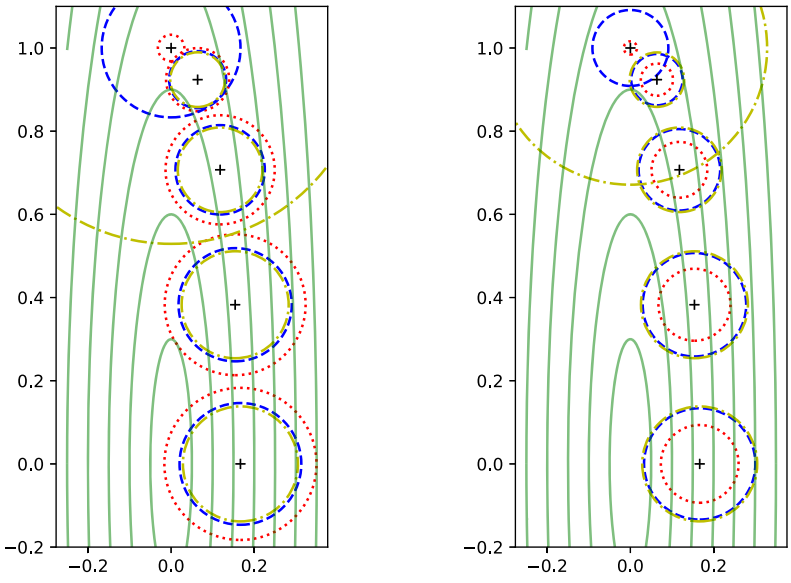
where the convergence of each term is supposed in the proposition. □

Consider the truncation weights with a fixed selection ratio $\lambda = \rho\mu$ for some $\rho > 1$ and the sequence of the Hessian matrices such that the condition number is bounded. From Appendix B.4, we have that $L_1 \in O(\lambda^{-1/2})$, $L_2 \in O(\lambda^{-3/2})$, and $L_3 \in O(\lambda^{-3/2})$. Moreover, we have $1/d_1(\mathbf{A}) \in O(N)$ and $1/\mathrm{Tr}(\mathbf{A}^2)^{\frac{1}{2}} \in O(N^{\frac{1}{2}})$. Then, the condition of Proposition 17 reduces to $\lambda \in o(N^{\frac{1}{3}})$. This condition is a rigorous (but probably not tight) bound for the scaling of λ such that the per-iteration convergence rate of a $(\mu/\mu, \lambda)$-ES with a fixed λ/μ on the sphere function scales like $O(\lambda/N)$ [15, Equation 6.140]. One can also deduce the condition for the optimal weights and the CMA-type positive weights (positive half of the optimal weights).

4.4. Effect of the eigenvalue distribution of the Hessian matrix

Corollary 16 implies that the optimal recombination weights are independent of the Hessian in the limit $N \to \infty$ as long as $\lim_{N\to\infty}\mathrm{Tr}(\mathbf{A}^2) = 0$. Moreover, Proposition 13 and Corollary 15 together imply that the same values approximate the optimal recombination weights for a sufficiently large λ in the limit of $\sigma/\|\boldsymbol{m}\| \to 0$. On the other hand, the step-size and the progress rate depend on the Hessian. In the following we discuss the effect of the Hessian followed by a simulation. To make the discussion more intuitive, we remove the condition $\mathrm{Tr}(\mathbf{A}) = 1$ and consider an arbitrary non-negative definite symmetric $\mathbf{A}$. All the statements above still hold by replacing $\mathbf{A}$ with $\mathbf{A}/\mathrm{Tr}(\mathbf{A})$.

Given $(w_k)_{k=1}^\lambda$, the optimal normalized step-size $\bar{\sigma}^*$ and the normalized quality gain $\bar{\phi}_\infty(\bar{\sigma}, (w_k))$ are independent of the Hessian $\mathbf{A}$ and the distribution mean $\boldsymbol{m}$. However, the step-size $\sigma = (\bar{\sigma}/c_m)\|\nabla f(\boldsymbol{m})\|$ and the quality gain



(a) Optimal weights (8) (b) Optimal positive weights

Fig. 3. The asymptotically optimal step-size $\sigma^* = \bar{\sigma}^* \|\nabla f(\boldsymbol{m})\|$ on $f(x) = x^{\mathrm{T}}\mathbf{A}x/2$ with $\mathbf{A} = \mathrm{diag}(1, 36)$. The circles with radius $\bar{\sigma}^* \|\nabla f(\boldsymbol{m})\|$ centered at $\boldsymbol{m} = 2\mathbf{A}^{-\frac{1}{2}}(\cos(\theta), \sin(\theta))$ with $\theta = \pi/2, 3\pi/8, \pi/4, \pi/8, 0$ are displayed, where the asymptotically optimal normalized step-size $\bar{\sigma}^*$ is computed using (27) with the optimal weights (8) (left) and with the optimal positive weights (right). Red dotted: $\lambda = 2$, Blue dashed: $\lambda = 10$, Yellow dot-dashed: $\lambda = 50$. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

$\phi(\boldsymbol{m}, \sigma) = g(\boldsymbol{m})\bar{\phi}_\infty(\bar{\sigma}, (w_k)_{k=1}^{\lambda})$ depend on them through $\|\nabla f(\boldsymbol{m})\|$ and $g(\boldsymbol{m}) = \|\nabla f(\boldsymbol{m})\|^2 / f(\boldsymbol{m})$. If $\boldsymbol{m}$ is on a contour ellipsoid ($f(\boldsymbol{m}) = 1$ for example), $g(\boldsymbol{m})$ increases as $\|\nabla f(\boldsymbol{m})\|$. In other words, the greater the optimal step-size is, the greater quality gain we achieve. These quantities are bounded as

$$\frac{d_N(\mathbf{A})}{\mathrm{Tr}(\mathbf{A})} \left\|\boldsymbol{m} - x^*\right\| \leq \|\nabla f(\boldsymbol{m})\| \leq \frac{d_1(\mathbf{A})}{\mathrm{Tr}(\mathbf{A})} \left\|\boldsymbol{m} - x^*\right\| \text{ and}$$
$$\frac{d_N(\mathbf{A})}{\mathrm{Tr}(\mathbf{A})} \leq \frac{g(\boldsymbol{m})}{2} \leq \frac{d_1(\mathbf{A})}{\mathrm{Tr}(\mathbf{A})} \ .$$

The lower and upper equalities for both of the above inequalities hold if and only if $\boldsymbol{m} - x^*$, or equivalently $\mathbf{e}_{\boldsymbol{m}}$, is parallel to the eigenspace corresponding to the smallest and largest eigenvalues of $\mathbf{A}$, respectively. Therefore, the optimal step-size and the quality gain can be different by the factor of at most $\mathrm{Cond}(\mathbf{A}) = d_1(\mathbf{A})/d_N(\mathbf{A})$. Fig. 3 visualizes example cases. The asymptotic optimal step-size heavily depends on the location of $\boldsymbol{m}$ if $\mathbf{A}$ is ill-conditioned. If we focus on the area around each circle, the function landscape looks like a parabolic ridge function. Note that a relatively large step-size displayed at $\boldsymbol{m} = (0, 1)$ for $\lambda > 2$ is because $\mathbf{e}_{\boldsymbol{m}}^{\mathrm{T}}\mathbf{A}\mathbf{e}_{\boldsymbol{m}} \ll 1$ in (27), resulting in $\bar{\sigma}^* \propto \mu_w$. The asymptotic normalized quality gain is derived for the limit $\sigma / \|\boldsymbol{m}\| \to 0$, and the update of the mean vector results in an approximation of the negative gradient direction. If the mean vector is exactly on the longest axis of the hyper-ellipsoid, the gradient points to the optimal solution and a large normalized step-size is desired. However, this never happens in practice, since the mean vector will not be exactly in such a situation with probability one. We also remark that the asymptotically optimal normalized step-size (27) is not monotonically increasing w.r.t. λ. Indeed, we see in Fig. 3a smaller step-sizes for greater λ values, whereas they are monotonic in Fig. 3b. The main difference is that the step-sizes with the optimal weights for $\lambda = 2$ can be greater than those with the optimal positive weights. Nonetheless, there is no guarantee that these figures reflect the *actually* optimal step-size precisely since displayed are the step-size optimal in the limit of c_m to infinity. Further investigation needs to be conducted.

Table 1 summarizes $d_N(\mathbf{A})/\mathrm{Tr}(\mathbf{A})$, $d_1(\mathbf{A})/\mathrm{Tr}(\mathbf{A})$ and $\mathrm{Tr}(\mathbf{A}^2)/\mathrm{Tr}(\mathbf{A})^2$ for different types of $\mathbf{A}$. The greater the first two quantities are, the greater the optimal step-size and hence the quality gain are. The smaller the last quantity is, the more reliable it is to approximate $\bar{\phi}$ with $\bar{\phi}_\infty$. If the condition number $\alpha = \mathrm{Cond}(\mathbf{A})$ is fixed, the worst case ($d_N(\mathbf{A})/\mathrm{Tr}(\mathbf{A})$) is maximized when the function has a discus type structure and is minimized when the function has a cigar type structure. The value of $d_N(\mathbf{A})/\mathrm{Tr}(\mathbf{A})$ will be close to $1/N$ as $N \to \infty$ for the discus type function, whereas it will be close to $1/(N\alpha)$ for the cigar. Therefore, the discus type function is as easy to solve as the sphere function if $N \gg \alpha$, while the cigar type function takes roughly $1/\alpha$ times more iterations to reach the same target function value. On the other hand, the inequality $\mathrm{Tr}(\mathbf{A}^2)/\mathrm{Tr}(\mathbf{A})^2 < 1/(N-1)$ holds independently of α on the cigar type function, while $\mathrm{Tr}(\mathbf{A}^2)/\mathrm{Tr}(\mathbf{A})^2$ depends heavily on α

Table 1
Different types of the eigenvalue distributions of $\mathbf{A}$. The second to fourth types (discus: $d_1(\mathbf{A}) = \alpha$ and $d_2(\mathbf{A}) = \cdots = d_N(\mathbf{A}) = 1$, ellipsoid: $d_i(\mathbf{A}) = \alpha^{\frac{i-1}{N-1}}$, cigar: $d_1(\mathbf{A}) = \cdots = d_{N-1}(\mathbf{A}) = \alpha$ and $d_N(\mathbf{A}) = 1$) have the condition number $\mathrm{Cond}(\mathbf{A}) = d_1(\mathbf{A})/d_N(\mathbf{A}) = \alpha$, while the last type has the condition number N.

Type	$\frac{d_N(\mathbf{A})}{\mathrm{Tr}(\mathbf{A})}$	$\frac{d_1(\mathbf{A})}{\mathrm{Tr}(\mathbf{A})}$	$\frac{\mathrm{Tr}(\mathbf{A}^2)}{\mathrm{Tr}(\mathbf{A})^2}$
Sphere	$\frac{1}{N}$	$\frac{1}{N}$	$\frac{1}{N}$
Discus	$\frac{1}{(N-1)+\alpha}$	$\frac{\alpha}{(N-1)+\alpha}$	$\frac{(N-1)+\alpha^2}{((N-1)+\alpha)^2}$
Ellipsoid	$\frac{\alpha^{\frac{1}{N-1}}-1}{\alpha^{\frac{N}{N-1}}-1}$	$\frac{\alpha^{\frac{N}{N-1}}-\alpha}{\alpha^{\frac{N}{N-1}}-1}$	$\frac{\left(\alpha^{\frac{2N}{N-1}}-1\right)/\left(\alpha^{\frac{2}{N-1}}-1\right)}{\left(\alpha^{\frac{N}{N-1}}-1\right)^2/\left(\alpha^{\frac{1}{N-1}}-1\right)^2}$
Cigar	$\frac{1}{(N-1)\alpha+1}$	$\frac{\alpha}{(N-1)\alpha+1}$	$\frac{(N-1)\alpha^2+1}{((N-1)\alpha+1)^2}$
$d_i(\mathbf{A}) = i$	$\frac{1}{N(N+1)/2}$	$\frac{1}{(N+1)/2}$	$\frac{\frac{1}{6}N(N+1)(2N+1)}{\left(N(N+1)/2\right)^2}$

on the discus type function. The fraction will not be sufficiently small and we can not approximate the normalized quality gain by $\bar{\phi}_\infty$ unless $\alpha \ll N$ holds.[7]

The condition $\lim_{N\to\infty} \mathrm{Tr}(\mathbf{A}^2)/\mathrm{Tr}(\mathbf{A})^2 = 0$ also hold for some positive semi-definite $\mathbf{A}$, where only $M < N$ eigenvalues of $\mathbf{A}$ are positive and the others are zero. That is, $d_1(\mathbf{A}) \geq \cdots \geq d_M(\mathbf{A}) > 0$ and $d_{M+1}(\mathbf{A}) = \cdots = d_N(\mathbf{A})$. In this case, the condition $\mathrm{Tr}(\mathbf{A}^2)/\mathrm{Tr}(\mathbf{A})^2 \to 0$ holds only if the dimension M of the effective search space tends to infinity as $N \to \infty$. The above inequalities are refined as follows. Let $\boldsymbol{m}^+$ and $\boldsymbol{m}^-$ be the decomposition of $\boldsymbol{m}$ such that $\boldsymbol{m}^-$ is the projection of $\boldsymbol{m}$ onto the hyper-plane through x^* spanned by the eigenvectors of $\mathbf{A}$ corresponding to the zero eigenvalue, and $\boldsymbol{m}^+ = \boldsymbol{m} - \boldsymbol{m}^-$. Then,

$$\frac{d_M(\mathbf{A})}{\mathrm{Tr}(\mathbf{A})} \leq \frac{g(\boldsymbol{m})}{2} \leq \frac{d_1(\mathbf{A})}{\mathrm{Tr}(\mathbf{A})}$$
$$\frac{d_M(\mathbf{A})}{\mathrm{Tr}(\mathbf{A})} \left\|\boldsymbol{m}^+\right\| \leq \frac{\|\nabla f(\boldsymbol{m})\|}{\mathrm{Tr}(\mathbf{A})} \leq \frac{d_1(\mathbf{A})}{\mathrm{Tr}(\mathbf{A})} \left\|\boldsymbol{m}^+\right\| \quad .$$

In this case, $g(\boldsymbol{m})$ can be $2/M$ if $d_1(\mathbf{A}) = \cdots = d_M(\mathbf{A}) > 0$ and $d_i(\mathbf{A}) = 0$ for $i \in [\![M+1, N]\!]$. The quality gain is then proportional to $2/M$, instead of $2/N$. That is, the evolution strategy with the optimal step-size solves the quadratic function with the effective rank M defined on the N dimensional search space as efficiently as it solves its projection onto the effective search space.

Comment on the algorithm dynamics The asymptotic quality gain depends on the distribution mean $\boldsymbol{m}$ through $g(\boldsymbol{m})$. In practice, we observe near worst case performance with $g(\boldsymbol{m}) \approx 2d_N(\mathbf{A})/\mathrm{Tr}(\mathbf{A})$, which implies that $\boldsymbol{m} - x^*$ is almost parallel to the eigenspace corresponding to the smallest eigenvalue $d_N(\mathbf{A})$ of the Hessian matrix. We provide an intuition to explain this behavior, which will be useful to understand the algorithm, even though the argument is not fully rigorous.

Consider Algorithm 1 with scale-invariant step-size (Definition 3). Lemma 9 implies that the order of the function values $f(X_i)$ coincide with the order of $[\mathcal{N}_i]_1 = \mathbf{e}^{\mathrm{T}}(X_i - \boldsymbol{m}^{(t)})/\sigma^{(t)}$, where $\mathbf{e} = \nabla f(\boldsymbol{m}^{(t)})/\left\|\nabla f(\boldsymbol{m}^{(t)})\right\|$. This is because if $Z \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, then $Z^{\mathrm{T}}\mathbf{A}Z/\mathrm{Tr}(\mathbf{A})$ in (11) almost surely converges to one by the strong law of large numbers as $N \to \infty$ under $\mathrm{Tr}(\mathbf{A}^2)/\mathrm{Tr}(\mathbf{A})^2 \to 0$. It means that the function value of a candidate solution is determined solely by the first component on the right-hand side of (11), that is, $\mathbf{e}^{\mathrm{T}}(X_i - \boldsymbol{m}^{(t)})/\sigma^{(t)}$. Since the ranking of the function value only depends on $\mathbf{e}^{\mathrm{T}}(X_i - \boldsymbol{m}^{(t)})/\sigma^{(t)}$, one may rewrite the update of the mean vector as

$$\boldsymbol{m}^{(t+1)} = \boldsymbol{m}^{(t)} + c_m\sigma^{(t)} \sum_{i=1}^{\lambda} w_i \mathcal{N}_{i:\lambda}(0,1) \cdot \mathbf{e} + c_m\sigma^{(t)}\mu_w^{-\frac{1}{2}}\mathcal{N}(\mathbf{0}, \mathbf{I} - \mathbf{e}\mathbf{e}^{\mathrm{T}}) \quad , \tag{35}$$

where $\mathcal{N}_{i:\lambda}(0,1)$ are the i-th order statistics from λ population of $\mathcal{N}(0,1)$, and $\mathcal{N}(\mathbf{0}, \mathbf{I} - \mathbf{e}\mathbf{e}^{\mathrm{T}})$ is the normally distributed random vector with mean vector $\mathbf{0}$ and the degenerated covariance matrix $\mathbf{I} - \mathbf{e}\mathbf{e}^{\mathrm{T}}$. It indicates that the mean vector moves along the gradient direction with the distribution $c_m\sigma^{(t)}\sum_{i=1}^{\lambda} w_i\mathcal{N}_{i:\lambda}(0,1)$, while it moves randomly in the subspace orthogonal to the gradient with the distribution $c_m\sigma^{(t)}\mu_w^{-\frac{1}{2}}\mathcal{N}(\mathbf{0}, \mathbf{I} - \mathbf{e}\mathbf{e}^{\mathrm{T}})$.

If the function is spherical, i.e. $\mathbf{A} \propto \mathbf{I}$, the mean vector does a symmetric, unbiased random walk on the surface of a hypersphere while the radius of the hypersphere gradually decreases due to the second term on (35). If the function is a

[7] However, the worst case scenario on the discus type function, $1/(\alpha + (N-1))$, describes an empirical observation [36] that the convergence speed of evolution strategy with isotropic distribution does not scale down with N for $N \ll \alpha$.

general convex quadratic function, $\mathbf{A} \propto \mathbf{I}$, the corresponding random walk on the surface of a hyperellipsoid becomes biased. Then, $\boldsymbol{m} - x^*$ tends to be parallel to the eigenspace corresponding to the smallest eigenvalue $d_N(\mathbf{A})$, which means that the quality gain is close to the worst case of $d_N(\mathbf{A})/\mathrm{Tr}(\mathbf{A})$. The reason may be explained as follows. The progress in one step is the largest in the short axis direction (parallel to the eigenvector corresponding to the largest eigenvalue of $\mathbf{A}$), and the smallest in the long axis direction (parallel to the eigenvector corresponding to the largest eigenvalue of $\mathbf{A}$). The short axis direction is quickly optimized and the situation gets close to the worst case, where it takes many iterations to escape from. Therefore, we observe the near worst situation in practice. Further theoretical investigation on the distribution of $\mathbf{e} = \nabla f(\boldsymbol{m}^{(t)}) / \left\| \nabla f(\boldsymbol{m}^{(t)}) \right\|$ should be done in the future work.

4.5. Experiments

To see the effect of the eigenvalue distribution of $\mathbf{A}$, we run the experiments. Four quadratic functions are considered: Sphere, Discus, Ellipsoid, Cigar functions of $N = 10$, 100, 1000 dimensions. The ES with the weights optimal for the infinite dimensional sphere, (8), and the optimal normalized step-size $\bar{\sigma}^*$ derived for $c_m \to \infty$, (27), times a constant factor is run for $T = 10000$ iterations. The empirical normalized quality gain is estimated as $(2/T)\sum_{t=T/2}^{T-1} \left[f(\boldsymbol{m}^{(t)}) - f(\boldsymbol{m}^{(t+1)}) \right] / \left[f(\boldsymbol{m}^{(t)}) g(\boldsymbol{m}^{(t)}) \right]$. The mean vector is initialized randomly by the normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$. Eleven independent runs are conduced for each setting. The results are compared with φ, which is supposed to approximate the empirical normalized quality gain for $c_m \gg 1$ and $N \gg 1$. Note that $\bar{\sigma}^*$ in (27) and φ in (17) depend on $\boldsymbol{m}$ through $\mathbf{e}^{\mathrm{T}}\mathbf{A}\mathbf{e}/\mathrm{Tr}(\mathbf{A})$. We replace $\mathbf{e}^{\mathrm{T}}\mathbf{A}\mathbf{e}/\mathrm{Tr}(\mathbf{A})$ with $d_N(\mathbf{A})/\mathrm{Tr}(\mathbf{A})$ based on the observation and the above discussion that the mean vector tends to be parallel to the eigenspace corresponding to the smallest eigenvalue of $\mathbf{A}$. Figs. 4 and 5 show the median (marker) and the 10%–90% interval (shaded area) of the empirical normalized quality gain for each c_m and the theoretically derived normalized quality gain formula discussed above. Note that the shaded area is almost invisible, implying that the number of runs and the number of iterations are sufficient to get accurate estimates.

We first focus on the results with $c_m = 1$ (the default setting). The empirical normalized quality gain gets closer to the normalized quality gain derived for the infinite dimensional quadratic function as N increases. The approach of the empirical normalized quality gain to the theory is the fastest for the sphere function ($\mathbf{A} = \mathbf{I}$). For convex quadratic functions with the same condition number of $\alpha = 10^6$, the speed of the convergence of the normalized quality gain to φ as $N \to \infty$ is the fastest for the cigar function, and the slowest for the discus function. This reflects the upper bound derived in Theorem 7 that depends on the ratio $\mathrm{Tr}(\mathbf{A}^2)/\mathrm{Tr}(\mathbf{A})^2$, whose value is summarized in Table 1. For the cigar function $\mathrm{Tr}(\mathbf{A}^2)/\mathrm{Tr}(\mathbf{A})^2$ is close to $1/(N-1)$, while for the discus function it is very close to 1 for $N \ll \alpha$ and we do not observe significant difference between results on different N.

A larger c_m led to a better empirical normalized quality gain for all cases, i.e., the empirical normalized quality gains became monotonically closer to the theoretical curve.[8] As c_m becomes greater while the normalized step-size is fixed, the ratio $\sigma / \|\boldsymbol{m}\|$ becomes smaller and tends to zero in the limit $c_m \to \infty$. As Corollary 15 implies, the normalized quality gain converges to φ in the limit $\sigma / \|\boldsymbol{m}\| \to 0$. Therefore, the results reflect the theory. Moreover, the theoretically optimal normalized step-size $\bar{\sigma}^*$ well approximates the empirically optimal normalized step-size $\bar{\sigma}$ that maximize the normalized quality gain for all cases when $c_m \geq 1$. As c_m becomes smaller, the empirically optimal normalized step-size $\bar{\sigma}$ becomes smaller compared to $\bar{\sigma}^*$. Note that the difference of the empirical normalized quality gain curves on the sphere function comes only from the randomness of the length of each step Z. If we replace Z with $(\mathbb{E}[\|Z\|]/\|Z\|)Z$ in the algorithm, the selection is independent of c_m values and is determined by the inner product of the step and the gradient of the objective function at the mean vector. Then, the effect of c_m goes away.

Comparing Fig. 4 and Fig. 5, the empirical curves are closer to the theoretical curves in Fig. 4. It reflects the fact that the bound between the normalized quality gain and the asymptotic normalized quality gain derived in Theorem 7 typically increases as λ increases. To approximate the theoretical curve, a larger c_m value is required when λ is greater. The peak of the empirical curves tend to be achieved at a smaller normalized step-size as λ or c_m becomes greater or smaller, respectively.

5. Conclusion

We perform the quality gain analysis of the weighted recombination evolution strategy (ES) on a convex quadratic function. Differently from the previous works, where the limit for the search space dimension N to infinity is considered, we derive the error bound between the so-called normalized quality gain and its limit expression for the finite dimension. We show that the bound converges to zero when (I) $N \to \infty$ as long as the Hessian $\mathbf{A}$ of the objective function satisfies $\mathrm{Tr}(\mathbf{A}^2)/\mathrm{Tr}(\mathbf{A})^2 \to 0$, or when (II) $\sigma / \|\boldsymbol{m}\| \to 0$. The limit expression of the normalized quality gain reveals that the optimal recombination weights are independent of the Hessian matrix in the limit (I). Moreover, if the effective variance selection

[8] Figure 4 in the previous work [1] shows non-monotonic change of empirical normalized quality gain over c_m, whereas in Figs. 4 and 5 of this paper shows a monotonic behavior. In Figure 4 in [1] $\bar{\sigma}^*$ is approximated with (7), whereas in the figures of this paper $\bar{\sigma}^*$ is computed with (27). The difference between these two quantities is less pronounced as N increases. The monotonic changes of the graphs are because $\bar{\sigma}^*$ in (27) approximates the optimal normalized step-size better than (7) on a finite dimensional quadratic function.

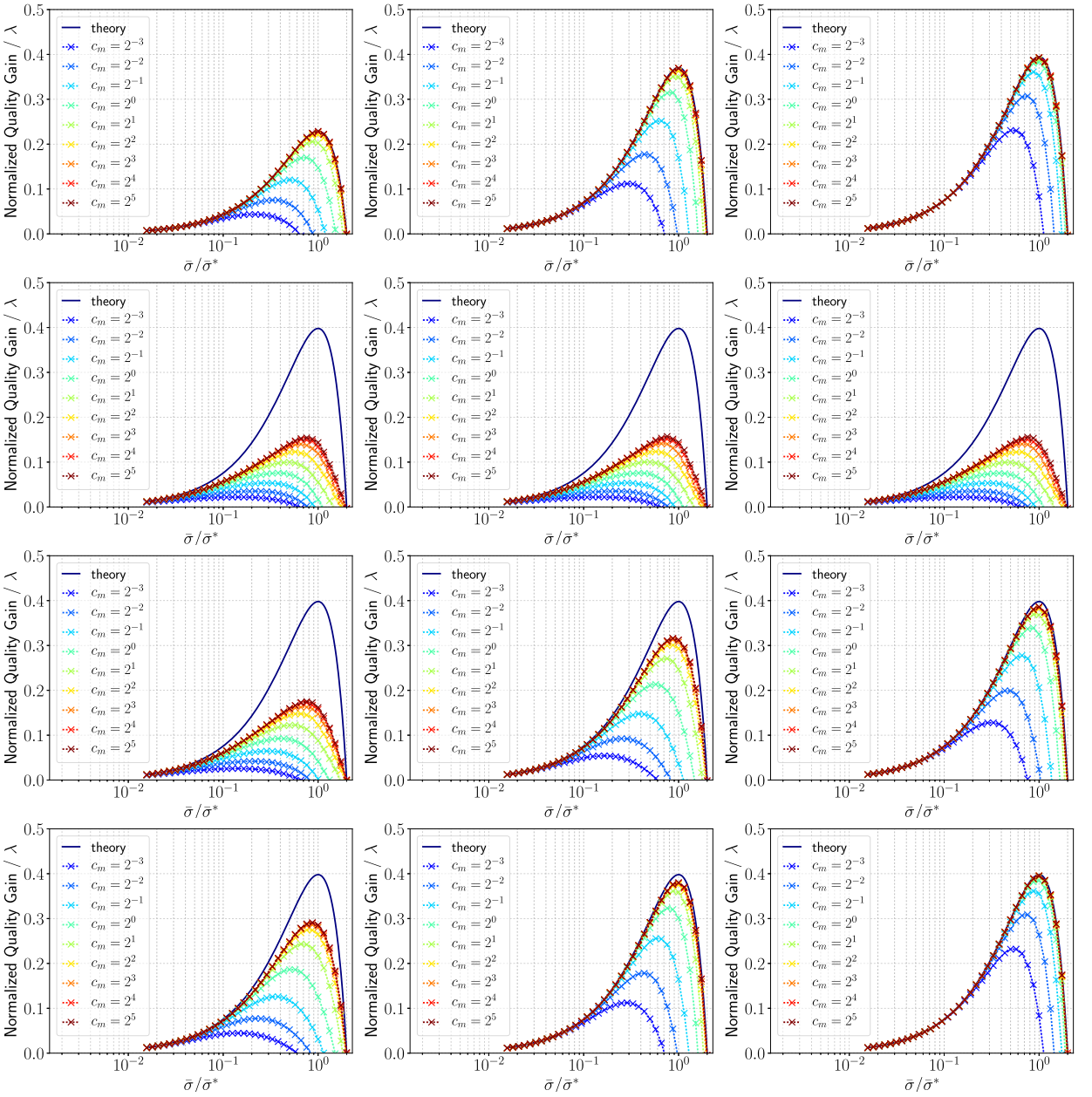


Fig. 4. Empirical normalized quality gain on four convex quadratic functions, Sphere, Discus, Ellipsoid and Cigar (from top to bottom) of dimension $N = 10$, 100 and 1000 (from left to right). The optimal weights (8) are used and $\lambda = 10$.

mass μ_w is sufficiently large, the optimal recombination weights for the limit (II) admits the same optimal recombination weights. The optimal normalized step-size for given recombination weights is derived. In the limit (I) the optimal normalized step-size is independent of $\mathbf{A}$, while the optimal step-size is proportional to the length of the gradient at the distribution mean. The limit (II) reveals the dependencies of the normalized step-size on N and μ_w.

The quality gain analysis provides a useful insight into the algorithmic behavior, even though it does not take into account the adaptation of the step-size. Knowing the optimal recombination weights (w_k^*) directly contributes to the optimal parameter setting. On the contrary, knowing the optimal normalized step-size $\bar{\sigma}^*$ does not lead to the optimal step-size control. This is because the optimal scale-invariant step-size σ^* in Definition 3 where $\bar{\sigma}$ is replaced with its optimal value $\bar{\sigma}^*$ is proportional to $\|\nabla f(\boldsymbol{m})\|$, which is unknown to the algorithm. The optimal step-size, however, is useful to evaluate step-size control mechanisms and to see how close to the optimal situation the step-size control mechanism is. Some theoretical insights into the adaptation mechanism of practical step-size adaptive methods is provided by the approached referred to as "dynamical system" approach by its authors. We refer to [27,28] for the recent development in the dynamical

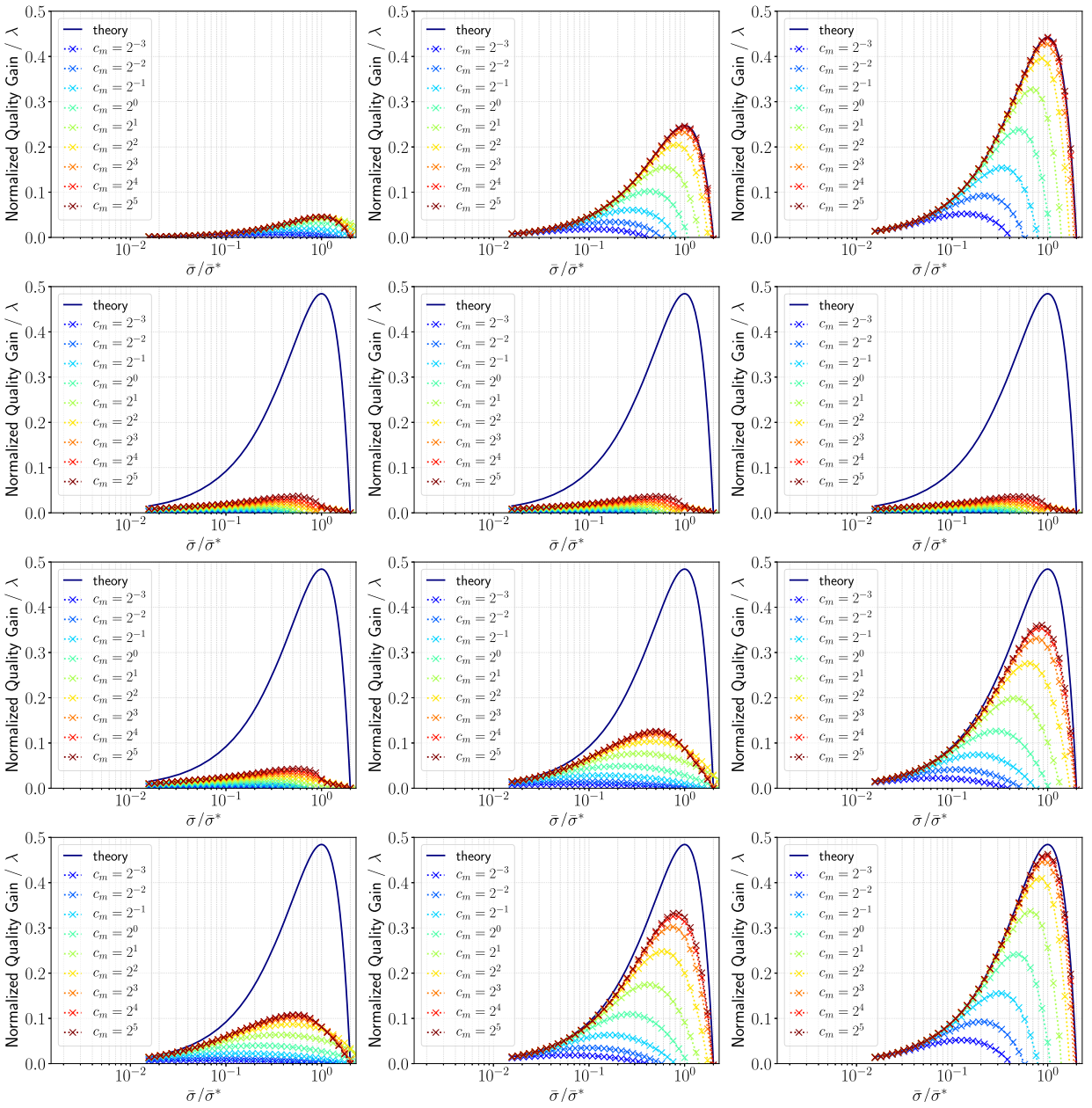


Fig. 5. Empirical normalized quality gain on four convex quadratic functions, Sphere, Discus, Ellipsoid and Cigar (from top to bottom) of dimension $N = 10$, 100 and 1000 (from left to right). The optimal weights (8) are used and $\lambda = 100$.

system approach. An important remaining question is: what is the optimal parameter update? Neither the quality gain analysis nor the dynamical system approach will answer this question. The optimal step-size on a quadratic function is revealed in this paper, however, it depends on the norm of the gradient, which is unknown to the real algorithm. A methodology to analyze the optimal update, rather than the optimal parameter value, hopefully including the covariance matrix update is desired in future work.

Acknowledgements

The authors thank *Dagstuhl seminar 17191: Theory of Randomized Optimization Heuristics* for providing the opportunity to present and discuss this work. This work is partially supported by JSPS KAKENHI Grant Number 15K16063.

Appendix A. Normal order statistics

Here we summarize some important properties of the moments of normal order statistics that are useful to understand the results in the paper.

The first moments of the normal order statistics have the properties: $\mathbb{E}[\mathcal{N}_{i:\lambda}] \leq \mathbb{E}[\mathcal{N}_{i+1:\lambda}]$, $\mathbb{E}[\mathcal{N}_{i:\lambda}] = -\mathbb{E}[\mathcal{N}_{\lambda+1-i:\lambda}]$, and $\sum_{i=1}^{\lambda} \mathbb{E}[\mathcal{N}_{i:\lambda}] = 0$. The second (product) moments of the normal order statistics have the following properties: $\sum_{j=1}^{\lambda} \mathbb{E}[\mathcal{N}_{i:\lambda}\mathcal{N}_{j:\lambda}] = 1$, $\sum_{i=1}^{\lambda} \mathbb{E}[\mathcal{N}_{i:\lambda}^2] = \sum_{i=1}^{\lambda}\sum_{j=1}^{\lambda} \mathbb{E}[\mathcal{N}_{i:\lambda}\mathcal{N}_{j:\lambda}] = \lambda$, and $\mathbb{E}[\mathcal{N}_{i:\lambda}\mathcal{N}_{j:\lambda}] = \mathbb{E}[\mathcal{N}_{j:\lambda}\mathcal{N}_{i:\lambda}] = \mathbb{E}[\mathcal{N}_{\lambda+1-i:\lambda}\mathcal{N}_{\lambda+1-j:\lambda}] = \mathbb{E}[\mathcal{N}_{\lambda+1-j:\lambda}\mathcal{N}_{\lambda+1-i:\lambda}]$.

Here we summarize useful inequalities about order statistics that are all listed in Section 35.1.6 of [37]. The positive dependency inequality tells that the order statistics are non-negatively correlated, $\mathrm{Cov}(\mathcal{N}_{i:\lambda}, \mathcal{N}_{j:\lambda}) = \mathbb{E}[\mathcal{N}_{i:\lambda}\mathcal{N}_{j:\lambda}] - \mathbb{E}[\mathcal{N}_{i:\lambda}]\mathbb{E}[\mathcal{N}_{j:\lambda}] \geq 0$. Together with $\sum_{j=1}^{\lambda} \mathrm{Cov}(\mathcal{N}_{i:\lambda}, \mathcal{N}_{j:\lambda}) = \sum_{j=1}^{\lambda} \mathbb{E}[\mathcal{N}_{i:\lambda}\mathcal{N}_{j:\lambda}] = 1$, we have $0 \leq \mathrm{Cov}(\mathcal{N}_{i:\lambda}, \mathcal{N}_{j:\lambda}) \leq 1$. It implies $\mathbb{E}[\mathcal{N}_{i:\lambda}]\mathbb{E}[\mathcal{N}_{j:\lambda}] \leq \mathbb{E}[\mathcal{N}_{i:\lambda}\mathcal{N}_{j:\lambda}] \leq \mathbb{E}[\mathcal{N}_{i:\lambda}]\mathbb{E}[\mathcal{N}_{j:\lambda}] + 1$.

Another important inequality is David inequality for normal distribution. It tells that $\Phi^{-1}\big(i/(\lambda+1)\big) \leq \mathbb{E}[\mathcal{N}_{i:\lambda}] \leq \min\left\{\Phi^{-1}\big(i/(\lambda+0.5)\big),\ \Phi^{-1}\big((i-0.5)/\lambda\big)\right\}$, where Φ is the c.d.f. of $\mathcal{N}(0,1)$. It proves an asymptotically tight approximation (Blom's approximation) $\mathbb{E}[\mathcal{N}_{i:\lambda}] \approx \Phi^{-1}\big(\frac{i-\alpha}{\lambda-2\alpha+1}\big)$ with $\alpha = 0.375$ for $i \leq \lceil \lambda/2 \rceil$. The following asymptotic equalities are also used (see Example 8.1.1 in [37])

$$\lim_{\lambda\to\infty} \frac{\mathbb{E}[\mathcal{N}_{\lambda:\lambda}] - \mathbb{E}[\mathcal{N}_{1:\lambda}]}{2(2\ln(\lambda))^{\frac{1}{2}}} = 1, \quad \lim_{\lambda\to\infty} \frac{1}{\lambda}\sum_{i=1}^{\lambda} |\mathbb{E}[\mathcal{N}_{i:\lambda}]| = \frac{2^{\frac{1}{2}}}{\pi^{\frac{1}{2}}}, \quad \lim_{\lambda\to\infty} \frac{1}{\lambda}\sum_{i=1}^{\lambda} \mathbb{E}[\mathcal{N}_{i:\lambda}]^2 = 1 \ . \tag{A.1}$$

Let $\boldsymbol{n}_{(\lambda)}$ be the λ dimensional column vector whose i-th component is $\mathbb{E}[\mathcal{N}_{i:\lambda}]$ and $\mathbf{N}_{(\lambda)}$ be the λ dimensional symmetric matrix whose (i,j)-th element is $\mathbb{E}[\mathcal{N}_{i:\lambda}\mathcal{N}_{j:\lambda}]$. The covariance matrix $\mathbf{N}_{(\lambda)} - \boldsymbol{n}_{(\lambda)}\boldsymbol{n}_{(\lambda)}^{\mathrm{T}}$ is by definition nonnegative definite. It implies the eigenvalues of $\mathbf{N}_{(\lambda)}$ are all nonnegative. Moreover, from the above mentioned fact derives that the sum of the eigenvalues is $\mathrm{Tr}(\mathbf{N}_{(\lambda)}) = \sum_{i=1}^{\lambda}\sum_{j=1}^{\lambda} \mathrm{Cov}(\mathcal{N}_{i:\lambda}, \mathcal{N}_{j:\lambda}) = \lambda$. Furthermore, the third asymptotic relation of (A.1) reads $\lim_{\lambda\to\infty} \mathrm{Tr}(\boldsymbol{n}_{(\lambda)}\boldsymbol{n}_{(\lambda)}^{\mathrm{T}})/\lambda = \lim_{\lambda\to\infty} \left\|\boldsymbol{n}_{(\lambda)}\right\|^2/\lambda = 1$. It implies, for any $\boldsymbol{x} \in \mathbb{R}^{\lambda} \setminus \{\mathbf{0}\}$, we have

$$\lim_{\lambda\to\infty} \frac{\boldsymbol{x}^{\mathrm{T}}\mathbf{N}_{(\lambda)}\boldsymbol{x}}{\lambda\left\|\boldsymbol{x}\right\|^2} = \lim_{\lambda\to\infty} \frac{(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{n}_{(\lambda)})^2}{\lambda\left\|\boldsymbol{x}\right\|^2} = \lim_{\lambda\to\infty} \frac{(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{n}_{(\lambda)})^2}{\left\|\boldsymbol{x}\right\|^2\left\|\boldsymbol{n}_{(\lambda)}\right\|^2} \ . \tag{A.2}$$

Appendix B. Proofs and derivations

B.1. Proof of Proposition 2

Proof. Let $\Delta = c_m \sum_{i=1}^{\lambda} W(i; (\boldsymbol{m} + \sigma Z_k)_{k=1}^{\lambda}) Z_i$, where $(Z_i)_{i=1}^{\lambda}$ are independent and N-variate standard normally distributed random vectors. Then,

$$\begin{aligned}
\phi(\boldsymbol{m}, \sigma) &= 1 - \mathbb{E}[f^*(\boldsymbol{m} + \sigma\Delta))]/f^*(\boldsymbol{m}) \\
&= 1 - \mathbb{E}[f(\boldsymbol{m} + \sigma\Delta - x^*)]/f(\boldsymbol{m} - x^*) \\
&= 1 - \alpha^{-n}\mathbb{E}[f(\alpha\cdot(\boldsymbol{m} + \sigma\Delta - x^*))/\alpha^{-n} f(\alpha\cdot(\boldsymbol{m} - x^*)) \\
&= 1 - \mathbb{E}[f(\alpha\cdot(\boldsymbol{m} + \sigma\Delta - x^*))]/f(\alpha\cdot(\boldsymbol{m} - x^*)) \\
&= 1 - \mathbb{E}[f^*(x^* + \alpha\cdot(\boldsymbol{m} - x^*) + \alpha\sigma\Delta)/f^*(x^* + \alpha\cdot(\boldsymbol{m} - x^*)) \\
&= \phi(x^* + \alpha(\boldsymbol{m} - x^*), \alpha\sigma) \ .
\end{aligned}$$

Note that $\phi(x^* + (\boldsymbol{m} - x^*), \sigma) = \phi(\boldsymbol{m}, \sigma)$. That is, the quality gain is scale invariant around $(x^*, 0)$. Moreover, the above equality implies that $\mathrm{argmax}_{\sigma}\, \phi(x^* + (\boldsymbol{m} - x^*), \sigma) = \mathrm{argmax}_{\sigma}\, \phi(x^* + \alpha(\boldsymbol{m} - x^*), \alpha\sigma)$, i.e., the optimal step-size at $x^* + \alpha(\boldsymbol{m} - x^*)$ is α times greater than the optimal step-size at $x^* + (\boldsymbol{m} - x^*)$. Therefore, the optimal step-size as a function of $\boldsymbol{m} - x^*$ is homogeneous of degree 1, i.e., $\sigma^*(\alpha\cdot(\boldsymbol{m} - x^*)) = \alpha\sigma^*(\boldsymbol{m} - x^*)$. □

B.2. Proof of Lemma 5

Proof. Since $(X_k)_{k=1}^{\lambda}$ are independent and normally distributed, the conditional probability of $\mathbb{1}_{f(X_k)<f(X_i)} = 1$ given X_i for any $k \neq i$ is $F_f(f(X_i))$. Then, the probability of $\sum_{k=1}^{\lambda} \mathbb{1}_{f(X_k)\leq f(X_i)}$ being a for $a \in [\![1, \lambda]\!]$ is given by $P_b(a-1; \lambda-1, p)$ with $p = F_f(f(X_i))$. Then, for any $\alpha \geq 0$,

$$\mathbb{E}_i[W(i; (X_k)_{k=1}^{\lambda})^{\alpha}] = \sum_{k=1}^{\lambda} w_k^{\alpha} P_b(k-1; \lambda-1, p) \ .$$

Similarly, the joint distribution of $\sum_{k=1}^{\lambda} \mathbb{1}_{f(X_k)\leq f(X_i)}$ and $\sum_{k=1}^{\lambda} \mathbb{1}_{f(X_k)\leq f(X_j)}$ is derived. Due to the symmetry between i and j, we can assume w.l.o.g. that $f(X_i) \leq f(X_j)$. Then, the joint probability of $\sum_{k=1}^{\lambda} \mathbb{1}_{f(X_k)\leq f(X_i)} = a$ and $\sum_{k=1}^{\lambda} \mathbb{1}_{f(X_k)\leq f(X_j)} = b$ for $a, b \in [\![1, \lambda]\!]$ is given by $P_t(a-1, b-a-1; \lambda-2, p, q-p)$ with $p = F_f(f(X_i))$ and $q = F_f(f(X_j))$ if $a < b$, and zero otherwise. Then,

$$\mathbb{E}_{i,j}[W(i; (X_k)_{k=1}^{\lambda}) W(j; (X_k)_{k=1}^{\lambda})] = \sum_{m=1}^{\lambda-1} \sum_{l=m+1}^{\lambda} w_m w_l P_t(m-1, l-m-1; \lambda-2, p, q-p) \ .$$

This ends the proof. □

B.3. Proof of Lemma 6

Proof. The derivative of u_1 is $\sum_{k=1}^{\lambda} w_k \binom{\lambda-1}{k-1} \frac{\mathrm{d}}{\mathrm{d}p}[p^{k-1}(1-p)^{\lambda-k}]$, where

$$\frac{\mathrm{d}}{\mathrm{d}p}[p^{k-1}(1-p)^{\lambda-k}] = (k-1)p^{k-2}(1-p)^{\lambda-k} - (\lambda-k)p^{k-1}(1-p)^{\lambda-k-1} \ .$$

Substituting the derivatives and rearranging the terms, we obtain

$$\frac{\mathrm{d}u_1(p)}{\mathrm{d}p} = (\lambda-1) \sum_{k=1}^{\lambda-1} (w_{k+1} - w_k) \binom{\lambda-2}{k-1} p^{k-1}(1-p)^{\lambda-k-1} \ .$$

The Lipschitz constant L_1 is the supremum of the absolute value of the derivative derived above. It completes the proof for the ℓ_1-Lipschitz continuity of u_1 and its Lipschitz constant. Since u_2 is equivalent to u_1 if w_i are replaced with w_i^2 in the definition of u_1, we have the ℓ_1-Lipschitz continuity of u_2 and its Lipschitz constant by replacing w_i with w_i^2 in the above argument.

The partial derivative of u_3 with respect to p is

$$\sum_{k=1}^{\lambda-1} \sum_{l=k+1}^{\lambda} w_k w_l \binom{\lambda-2}{l-2} \binom{l-2}{k-1} \frac{\partial}{\partial p}[\min(p,q)^{k-1} |q-p|^{l-k-1} (1-\min(p,q))^{\lambda-l}] \ ,$$

where

$$\frac{\partial}{\partial p} \min(p,q)^{k-1} |q-p|^{l-k-1} (1-\min(p,q))^{\lambda-l}$$
$$= \begin{cases} [(k-1)(q-p) - (l-k-1)p]p^{k-2}(q-p)^{l-k-2}(1-q)^{\lambda-l} & (p<q) \\ [(l-k-1)(1-p) - (\lambda-l)(p-q)]q^{k-1}(p-q)^{l-k-2}(1-p)^{\lambda-l-1} & (p>q) \ . \end{cases}$$

Substituting the derivatives and rearranging the terms, we obtain

$$\frac{1}{\lambda-2} \frac{\partial u_3(p,q)}{\partial p}$$
$$= \begin{cases} \sum_{k=1}^{\lambda-2} \sum_{l=k+2}^{\lambda} w_l (w_{k+1} - w_k) \binom{\lambda-3}{l-3} \binom{l-3}{k-1} p^{k-1}(q-p)^{l-k-2}(1-q)^{\lambda-l} & (p<q) \\ \sum_{k=1}^{\lambda-2} \sum_{l=k+2}^{\lambda} w_k (w_l - w_{l-1}) \binom{\lambda-3}{l-3} \binom{l-3}{k-1} q^{k-1}(p-q)^{l-k-2}(1-p)^{\lambda-l} & (p>q) \end{cases}$$

Since u_3 is differentiable with respect to p almost everywhere in $(0,1)$, it is Lipschitz continuous with respect to p. Its Lipschitz constant is $\sup_{q\in(0,1)} \sup_{p\in(0,q)\cup(q,1)} \left| \frac{\partial u_3(p,q)}{\partial p} \right|$. Due to the symmetry, $u_3(p,q)$ is ℓ_1-Lipschitz continuous on $[0,1]^2$ with the Lipschitz constant $L_3 = \sup_{q\in(0,1)} \sup_{p\in(0,q)\cup(q,1)} \left| \frac{\partial u_3(p,q)}{\partial p} \right|$. This completes the proof. □

B.4. Upper bounds of Lipschitz constants

For a general weight scheme, we have the following trivial upper bounds for the Lipschitz constants derived in Lemma 6,

$$L_1 \leq (\lambda-1) \max_{k\in[\![1,\lambda-1]\!]} \left| w_{k+1} - w_k \right| \ , \tag{B.1}$$

$$L_2 \leq (\lambda-1) \max_{k\in[\![1,\lambda-1]\!]} \left| w_{k+1}^2 - w_k^2 \right| \ , \tag{B.2}$$

$$L_3 \leq (\lambda-2) \max_{k\in[\![1,\lambda]\!]} \max_{l\in[\![1,k-2]\!]\cup[\![k+1,\lambda-1]\!]} |w_k| \cdot \left| w_{l+1} - w_l \right| \ . \tag{B.3}$$

These upper bounds are straight-forward from the facts $\sum_{k=1}^{\lambda-1} P_b(k-1;\lambda-2,p)=1$ and $\sum_{k=1}^{\lambda-2}\sum_{l=k+2}^{\lambda} P_t(k-1, l-k-2;\lambda-3,\min(p,q),|q-p|)=1$.

For the truncation weights with $3\le\mu\le\lambda-2$, we can obtain better bounds. The bounds of the factorial of $n\ge 1$ known by Robbins [38], namely,

$$(2\pi n)^{\frac{1}{2}}\left(\frac{n}{e}\right)^n \exp\left(\frac{1}{12n+1}\right) < n! < (2\pi n)^{\frac{1}{2}}\left(\frac{n}{e}\right)^n \exp\left(\frac{1}{12n}\right)$$

gives us an upper bound of $\binom{n}{k}$ for $0<k<n$

$$\binom{n}{k} < \left(\frac{n}{2\pi k(n-k)}\right)^{\frac{1}{2}}\left(\frac{n}{k}\right)^k\left(\frac{n}{n-k}\right)^{n-k} . \tag{B.4}$$

Here we used $\exp\left(\frac{1}{12n}-\frac{1}{12k+1}-\frac{1}{12(n-k)+1}\right)<1$. On the other hand, we have for $0<k<n$

$$\sup_{0\le p\le 1} p^k(1-p)^{n-k} = \left(\frac{k}{n}\right)^k\left(\frac{n-k}{n}\right)^{n-k} . \tag{B.5}$$

Since $w_{k+1}-w_k=-1/\mu$ for $k=\mu$ and $w_{k+1}-w_k=0$ for $k\neq\mu$, we have for $3\le\mu\le\lambda-2$,

$$\begin{aligned}
L_1 &= \sup_{0<p<1}\left|(\lambda-1)\frac{1}{\mu}\binom{\lambda-2}{\mu-1}p^{\mu-1}(1-p)^{\lambda-\mu-1}\right| \\
&= \frac{\lambda-1}{\mu}\binom{\lambda-2}{\mu-1}\left(\frac{\mu-1}{\lambda-2}\right)^{\mu-1}\left(\frac{\lambda-\mu-1}{\lambda-2}\right)^{\lambda-\mu-1} \\
&\le \frac{\lambda-1}{\mu}\left(\frac{\lambda-2}{2\pi(\mu-1)(\lambda-\mu-1)}\right)^{\frac{1}{2}} .
\end{aligned}$$

Analogously, since $w_{k+1}^2-w_k^2=-1/\mu^2$ for $k=\mu$ and $w_{k+1}^2-w_k^2=0$ for $k\neq\mu$, we obtain the bound of L_2: $L_2 < \left[(\lambda-1)/\mu^2\right]\cdot\left[(\lambda-2)/(2\pi(\mu-1)(\lambda-\mu-1))\right]^{1/2}$.

Moreover, since $w_k(w_l-w_{l-1})=-1/\mu^2$ for $l=\mu+1$ and $w_k(w_l-w_{l-1})=0$ otherwise, we have

$$\begin{aligned}
L_3 &= \sup_{q\in(0,1)}\sup_{p\in(q,1)}\sum_{k=1}^{\mu-1}\frac{\lambda-2}{\mu^2}\binom{\lambda-3}{\mu-2}\binom{\mu-2}{k-1}q^{k-1}(p-q)^{\mu-k-1}(1-p)^{\lambda-\mu-1} \\
&= \sup_{p\in(0,1)}\frac{(\lambda-2)}{\mu^2}\binom{\lambda-3}{\mu-2}(1-p)^{\lambda-\mu-1}\sup_{q\in(0,p)}\sum_{k=1}^{\mu-1}\binom{\mu-2}{k-1}q^{k-1}(p-q)^{\mu-k-1} \\
&= \frac{(\lambda-2)}{\mu^2}\binom{\lambda-3}{\mu-2}\sup_{p\in(0,1)}(1-p)^{\lambda-\mu-1}p^{\mu-2} \\
&= \frac{(\lambda-2)}{\mu^2}\binom{\lambda-3}{\mu-2}\left(\frac{\mu-2}{\lambda-3}\right)^{\mu-2}\left(\frac{\lambda-\mu-1}{\lambda-3}\right)^{\lambda-\mu-1} \\
&\le \frac{(\lambda-2)}{\mu^2}\left(\frac{\lambda-3}{2\pi(\mu-2)(\lambda-\mu-1)}\right)^{\frac{1}{2}} .
\end{aligned}$$

Here we used (B.4), (B.5), and the binomial relation $\sum_{k=1}^{\mu-1}\binom{\mu-2}{k-1}q^{k-1}(p-q)^{\mu-k-1}=p^{\mu-2}$.

B.5. Proof of Lemma 8

Proof. If $\alpha=1$, then $G(\alpha)=1$, and the inequality is trivial. Hence, we assume $\alpha<1$ in the following.

Remember that $H_N=Z_{\mathbf{e}}+h(Z)$. The absolute difference between $F_N(t)$ and $\Phi(t)$ is rewritten as follows

$$\begin{aligned}
&|F_N(t)-\Phi(t)| \\
&= |\Pr[H_N\le t]-\Pr[Z_{\mathbf{e}}\le t]| \\
&= |\Pr[Z_{\mathbf{e}}+h(Z)\le t]-\Pr[Z_{\mathbf{e}}\le t]| \\
&= \Pr[h(Z)\ge 0 \text{ and } t-h(Z)\le Z_{\mathbf{e}}\le t]+\Pr[h(Z)\le 0 \text{ and } t\le Z_{\mathbf{e}}\le t-h(Z)] .
\end{aligned}$$

With an arbitrary $\epsilon_+ > 0$, the first term on the RMS is upper bounded as

$$\begin{aligned}
&\Pr[h(Z) \geq 0 \text{ and } t - h(Z) \leq Z_{\mathbf{e}} \leq t] \\
&\leq \Pr[h(Z) \geq \epsilon_+] + \Pr[h(Z) < \epsilon_+ \text{ and } t - h(Z) \leq Z_{\mathbf{e}} \leq t] \\
&\leq \Pr[h(Z) \geq \epsilon_+] + \Pr[h(Z) < \epsilon_+ \text{ and } t - \epsilon_+ \leq Z_{\mathbf{e}} \leq t] \\
&\leq \Pr[h(Z) \geq \epsilon_+] + \Pr[t - \epsilon_+ \leq Z_{\mathbf{e}} \leq t] \\
&\leq \Pr[h(Z) \geq \epsilon_+] + (2\pi)^{-\frac{1}{2}} \epsilon_+ \ .
\end{aligned}$$

For the last inequality, we used that the density of the one-dimensional standard normal distribution is at most $(2\pi)^{-\frac{1}{2}}$ and $Z_{\mathbf{e}}$ is of the one-dimensional standard normal distribution. Analogously, we have for any $\epsilon_- > 0$

$$\begin{aligned}
\Pr[h(Z) \leq 0 \text{ and } t \leq Z_{\mathbf{e}} \leq t - h(Z)] &\leq \Pr[h(Z) \leq -\epsilon_-] + \Pr[t \leq Z_{\mathbf{e}} \leq t + \epsilon_-] \\
&\leq \Pr[h(Z) \leq -\epsilon_+] + (2\pi)^{-\frac{1}{2}} \epsilon_- \ .
\end{aligned}$$

Let $\tilde{h}(Z) = 2(c_m/\bar{\sigma})h(Z) = Z^{\mathrm{T}}\mathbf{A}Z - 1$, $\tilde{\epsilon}_+ = 2(c_m/\bar{\sigma})\epsilon_+$ and $\tilde{\epsilon}_- = 2(c_m/\bar{\sigma})\epsilon_-$. Then, $\Pr[h(Z) \geq \epsilon_+] = \Pr[\tilde{h}(Z) \geq \tilde{\epsilon}_+]$ and $\Pr[h(Z) \leq -\epsilon_-] = \Pr[\tilde{h}(Z) \leq -\tilde{\epsilon}_-]$. From Lemma 1 in [32] knows that for any $x \geq 0$

$$\begin{aligned}
\Pr\left[\tilde{h}(Z) \geq 2\operatorname{Tr}(\mathbf{A}^2)^{\frac{1}{2}} x^{\frac{1}{2}} + 2d_1(\mathbf{A})x\right] &\leq \exp(-x) \ , \\
\Pr\left[\tilde{h}(Z) \leq -2\operatorname{Tr}(\mathbf{A}^2)^{\frac{1}{2}} x^{\frac{1}{2}}\right] &\leq \exp(-x) \ .
\end{aligned} \tag{B.6}$$

Let $x = \ln(1/\alpha)$ and let ϵ_+ and ϵ_- such that

$$\begin{aligned}
\tilde{\epsilon}_+ &= 2\operatorname{Tr}(\mathbf{A}^2)^{\frac{1}{2}} x^{\frac{1}{2}} + 2d_1(\mathbf{A})x = 2\operatorname{Tr}(\mathbf{A}^2)^{\frac{1}{2}} \left((\ln(1/\alpha))^{\frac{1}{2}} + (d_1(\mathbf{A})/\operatorname{Tr}(\mathbf{A}^2)^{\frac{1}{2}}) \ln(1/\alpha)\right) \\
\tilde{\epsilon}_- &= 2\operatorname{Tr}(\mathbf{A}^2)^{\frac{1}{2}} x^{\frac{1}{2}} = 2\operatorname{Tr}(\mathbf{A}^2)^{\frac{1}{2}} (\ln(1/\alpha))^{\frac{1}{2}} \ .
\end{aligned}$$

Then, from (B.6) derives that

$$\begin{aligned}
&\Pr[h(Z) \geq \epsilon_+] + (2\pi)^{-\frac{1}{2}} \epsilon_+ \\
&= \Pr[\tilde{h}(Z) \geq \tilde{\epsilon}_+] + (\bar{\sigma}\tilde{\epsilon}_+)/(2(2\pi)^{\frac{1}{2}} c_m) \\
&\leq \alpha + (\bar{\sigma}\tilde{\epsilon}_+)/(2(2\pi)^{\frac{1}{2}} c_m) \\
&= \alpha + (2\pi)^{-\frac{1}{2}} (\bar{\sigma}/c_m) \operatorname{Tr}(\mathbf{A}^2)^{\frac{1}{2}} \left((\ln(1/\alpha))^{\frac{1}{2}} + \left(d_1(\mathbf{A})/\operatorname{Tr}(\mathbf{A}^2)^{\frac{1}{2}}\right) \ln(1/\alpha)\right) \\
&= \alpha\left(1 + (2\pi)^{-\frac{1}{2}} (\ln(1/\alpha))^{\frac{1}{2}} + (2\pi)^{-\frac{1}{2}} \left(d_1(\mathbf{A})/\operatorname{Tr}(\mathbf{A}^2)^{\frac{1}{2}}\right) \ln(1/\alpha)\right) \ .
\end{aligned}$$

Similarly, we have $\Pr[h(Z) \leq -\epsilon_-] + (2\pi)^{-\frac{1}{2}} \epsilon_- \leq \alpha\left(1 + (2\pi)^{-\frac{1}{2}} (\ln(1/\alpha))^{\frac{1}{2}}\right)$. Altogether, we obtain

$$|F_N(t) - \Phi(t)| \leq \alpha\left(2 + (2/\pi)^{\frac{1}{2}} (\ln(1/\alpha))^{\frac{1}{2}} + (2\pi)^{-\frac{1}{2}} \left(d_1(\mathbf{A})/\operatorname{Tr}(\mathbf{A}^2)^{\frac{1}{2}}\right) \ln(1/\alpha)\right) \ .$$

Since the RHS of the above inequality is independent of t, taking the supremum of both sides over $t \in \mathbb{R}$, we obtain the desired inequality. □

B.6. Proof of Lemma 9

Proof. First, note that $F_f(f(\boldsymbol{m} + \sigma Z)) = F_N(H_N)$ and $H_N = Z_{\mathbf{e}} + h(Z)$. Using Lemma 6, we have

$$\begin{aligned}
\left|\mathbb{E}[u_1(F_f(f(X)))Z_{\mathbf{e}}] - \mathbb{E}[u_1(\Phi(Z_{\mathbf{e}}))Z_{\mathbf{e}}]\right| &= |\mathbb{E}[u_1(F_N(H_N))Z_{\mathbf{e}}] - \mathbb{E}[u_1(\Phi(Z_{\mathbf{e}}))Z_{\mathbf{e}}]| \\
&\leq \mathbb{E}[|u_1(F_N(H_N)) - u_1(\Phi(Z_{\mathbf{e}}))| \cdot |Z_{\mathbf{e}}|] \leq L_1 \mathbb{E}[|F_N(H_N) - \Phi(Z_{\mathbf{e}})| \cdot |Z_{\mathbf{e}}|] \ .
\end{aligned}$$

Noting that Φ is Lipschitz continuous with the Lipschitz constant $(2\pi)^{-\frac{1}{2}}$, we have $|\Phi(H_N) - \Phi(Z_{\mathbf{e}})| \leq (2\pi)^{-\frac{1}{2}} |H_N - Z_{\mathbf{e}}| = (2\pi)^{-\frac{1}{2}} |h(Z)|$. On the other hand, Lemma 8 says that $|F_N(H_N) - \Phi(H_N)| \leq G(\alpha)$. From these inequalities we obtain

$$|F_N(H_N) - \Phi(Z_{\mathbf{e}})| = |F_N(H_N) - \Phi(H_N) + \Phi(H_N) - \Phi(Z_{\mathbf{e}})| \leq G(\alpha) + (2\pi)^{-\frac{1}{2}} |h(Z)| \ . \tag{B.7}$$

Using the inequality (B.7) and the Schwarz inequality and the identities $\mathbb{E}[|Z_{\mathbf{e}}|] = (2/\pi)^{\frac{1}{2}}$, $\mathbb{E}[Z_{\mathbf{e}}^2] = 1$, and

$$\mathbb{E}[|h(Z)|^2] = \left(\frac{1}{2}\frac{\bar{\sigma}}{c_m}\right)^2 \mathbb{E}\left[\left(Z^{\mathrm{T}}\mathbf{A}Z - 1\right)^2\right] = \left(\frac{1}{2}\frac{\bar{\sigma}}{c_m}\right)^2 (2\operatorname{Tr}(\mathbf{A}^2)) = \frac{\alpha^2}{2} \ , \tag{B.8}$$

we have

$$\begin{aligned}\mathbb{E}[|F_N(H_N) - \Phi(Z_{\mathbf{e}})| \cdot |Z_{\mathbf{e}}|] &\leq G(\alpha)\mathbb{E}[|Z_{\mathbf{e}}|] + (2\pi)^{-\frac{1}{2}}\mathbb{E}[|h(Z)| \cdot |Z_{\mathbf{e}}|] \\ &\leq G(\alpha)\mathbb{E}[|Z_{\mathbf{e}}|] + (2\pi)^{-\frac{1}{2}}\mathbb{E}[h(Z)^2]^{\frac{1}{2}}\mathbb{E}[Z_{\mathbf{e}}^2]^{\frac{1}{2}} \\ &= (2/\pi)^{\frac{1}{2}}G(\alpha) + (2\pi)^{-\frac{1}{2}}\mathbb{E}[h(Z)^2]^{\frac{1}{2}} \\ &= (2/\pi)^{\frac{1}{2}}G(\alpha) + (4\pi)^{-\frac{1}{2}}\alpha \ .\end{aligned}$$

Altogether, we obtain the inequality stated in the lemma. This completes the proof. □

B.7. Proof of Lemma 10

Proof. Analogously to the proof of Lemma 9, we have

$$\begin{aligned}&\left|\mathbb{E}[u_2(F_f(f(X)))(Z^{\mathrm{T}}\mathbf{A}Z - 1)] - \mathbb{E}[u_2(\Phi(Z_{\mathbf{e}}))(Z^{\mathrm{T}}\mathbf{A}Z - 1)]\right| \\ &\quad\leq L_2\mathbb{E}[|F_N(H_N) - \Phi(Z_{\mathbf{e}})| \cdot \left|Z^{\mathrm{T}}\mathbf{A}Z - 1\right|] \\ &\quad\leq L_2\mathbb{E}[(G(\alpha) + (2\pi)^{-\frac{1}{2}}\,|h(Z)|)\left|Z^{\mathrm{T}}\mathbf{A}Z - 1\right|] \\ &\quad= L_2(G(\alpha)\mathbb{E}[\left|Z^{\mathrm{T}}\mathbf{A}Z - 1\right|] + (2\pi)^{-\frac{1}{2}}\mathbb{E}[|h(Z)| \cdot \left|Z^{\mathrm{T}}\mathbf{A}Z - 1\right|]) \ .\end{aligned}$$

Applying the inequalities $\mathbb{E}[\left|Z^{\mathrm{T}}\mathbf{A}Z - 1\right|] \leq \mathbb{E}[(Z^{\mathrm{T}}\mathbf{A}Z - 1)^2]^{\frac{1}{2}} = (2\,\mathrm{Tr}(\mathbf{A}^2))^{\frac{1}{2}}$ and

$$\mathbb{E}[|h(Z)| \cdot \left|Z^{\mathrm{T}}\mathbf{A}Z - 1\right|] = \frac{1}{2}(\bar{\sigma}/c_m)\mathbb{E}[(Z^{\mathrm{T}}\mathbf{A}Z - 1)^2] = (\bar{\sigma}/c_m)\,\mathrm{Tr}(\mathbf{A}^2) = \alpha\,\mathrm{Tr}(\mathbf{A}^2)^{\frac{1}{2}} \ ,$$

we obtain the inequality stated in the lemma. This completes the proof. □

B.8. Proof of Lemma 11

Proof. Using Lemma 6, we have

$$\begin{aligned}&\left|\mathbb{E}[u_3(F_f(f(X)), F_f(f(\tilde{X})))Z^{\mathrm{T}}\mathbf{A}\tilde{Z}] - \mathbb{E}[u_3(\Phi(Z_{\mathbf{e}}), \Phi(\tilde{Z}_{\mathbf{e}}))Z^{\mathrm{T}}\mathbf{A}\tilde{Z}]\right| \\ &\quad\leq \mathbb{E}[\left|(u_3(F_N(H_N), F_N(\tilde{H}_N)) - u_3(\Phi(Z_{\mathbf{e}}), \Phi(\tilde{Z}_{\mathbf{e}}))\right| \cdot \left|Z^{\mathrm{T}}\mathbf{A}\tilde{Z}\right|] \\ &\quad\leq L_3\mathbb{E}[(|F_N(H_N) - \Phi(Z_{\mathbf{e}})| + \left|F_N(\tilde{H}_N) - \Phi(\tilde{Z}_{\mathbf{e}})\right|) \cdot \left|Z^{\mathrm{T}}\mathbf{A}\tilde{Z}\right|] \ .\end{aligned}$$

Then, using the equality $\mathbb{E}[\left|Z^{\mathrm{T}}\mathbf{A}\tilde{Z}\right| \mid Z] = (2/\pi)^{\frac{1}{2}}\,\|\mathbf{A}Z\|$ (since $\left|Z^{\mathrm{T}}\mathbf{A}\tilde{Z}\right|$ given Z is half-normally distributed), the symmetry of Z and $\tilde{Z}$, the Schwarz inequality, and the inequality (B.7), we have

$$\begin{aligned}&\mathbb{E}[(|F_N(H_N) - \Phi(Z_{\mathbf{e}})| + \left|F_N(\tilde{H}_N) - \Phi(\tilde{Z}_{\mathbf{e}})\right|) \cdot \left|Z^{\mathrm{T}}\mathbf{A}\tilde{Z}\right|] \\ &\quad\leq 2G(\alpha)\mathbb{E}[\left|Z^{\mathrm{T}}\mathbf{A}\tilde{Z}\right|] + 2(2\pi)^{-\frac{1}{2}}\mathbb{E}[|h(Z)| \cdot \left|Z^{\mathrm{T}}\mathbf{A}\tilde{Z}\right|] \ .\end{aligned}$$

On one hand, we have

$$\begin{aligned}\mathbb{E}[\left|Z^{\mathrm{T}}\mathbf{A}\tilde{Z}\right|] &= \mathbb{E}[\mathbb{E}[\left|Z^{\mathrm{T}}\mathbf{A}\tilde{Z}\right| \mid Z]] = (2/\pi)^{\frac{1}{2}}\mathbb{E}[\|\mathbf{A}Z\|] \\ &\leq (2/\pi)^{\frac{1}{2}}\mathbb{E}[\|\mathbf{A}Z\|^2]^{\frac{1}{2}} = (2/\pi)^{\frac{1}{2}}\,\mathrm{Tr}(\mathbf{A}^2)^{\frac{1}{2}} \ ,\end{aligned}$$

where we used $\mathbb{E}[\|\mathbf{A}Z\|^2] = \mathbb{E}[\mathrm{Tr}(\mathbf{A}ZZ^{\mathrm{T}}\mathbf{A})] = \mathrm{Tr}(\mathbf{A}\mathbb{E}[ZZ^{\mathrm{T}}]\mathbf{A}) = \mathrm{Tr}(\mathbf{A}^2)$. On the other hand, we have

$$\begin{aligned}\mathbb{E}[|h(Z)| \cdot \left|Z^{\mathrm{T}}\mathbf{A}\tilde{Z}\right|] &= \mathbb{E}[|h(Z)|\,\mathbb{E}[\left|Z^{\mathrm{T}}\mathbf{A}\tilde{Z}\right| \mid Z]] = (2/\pi)^{\frac{1}{2}}\mathbb{E}[|h(Z)| \cdot \|\mathbf{A}Z\|] \\ &\leq (2/\pi)^{\frac{1}{2}}\mathbb{E}[|h(Z)|^2]^{\frac{1}{2}}\mathbb{E}[\|\mathbf{A}Z\|^2]^{\frac{1}{2}} = \pi^{-\frac{1}{2}}\alpha\,\mathrm{Tr}(\mathbf{A}^2)^{\frac{1}{2}} \ ,\end{aligned}$$

where we used $\mathbb{E}[|h(Z)|^2] = \alpha^2/2$ derived in (B.8). Altogether, we obtain the inequality stated in the lemma. This completes the proof. □

B.9. Proof of Lemma 12

Proof. Let p be the probability density function of the one-dimensional standard normal distribution and $p_{i:\lambda}$ be the probability density function of $\mathcal{N}_{i:\lambda}$ and $p_{i,j:\lambda}$ be the joint probability density function of $\mathcal{N}_{i:\lambda}$ and $\mathcal{N}_{j:\lambda}$. It is well known that $p_{i:\lambda}(x) = \lambda\binom{\lambda-1}{i-1}\Phi(x)^{i-1}(1-\Phi(x))^{\lambda-i}p(x)$ and $p_{i,j:\lambda}(x,y) = \lambda(\lambda-1)\binom{\lambda-2}{j-2}\binom{j-1}{i-1}\Phi(x)^{i-1}(\Phi(y)-\Phi(x))^{(j-i-1)}(1-\Phi(x))^{\lambda-j}p(x)p(y)$ for $i<j$ and $x<y$, and $p_{i,j:\lambda}(x,y)=0$ for $i<j$ and $x\geq y$. Note also that $p_{i,j:\lambda}(x,y) = p_{j,i:\lambda}(y,x)$.

The functions u_1 and u_2 are then written using these p.d.f.s of the normal order statistics as $\lambda u_1(\Phi(x))p(x) = \sum_{k=1}^{\lambda} w_k p_{k:\lambda}(x)$ and $\lambda u_2(\Phi(x))p(x) = \sum_{k=1}^{\lambda} w_k^2 p_{k:\lambda}(x)$. From these identities, we obtain (19) and (20). The identity (21) is derived by using $\lambda\mathbb{E}[u_2(\Phi(Z_{\mathbf{e}}))(Z_{\mathbf{e}}^2-1)] = \sum_{i=1}^{\lambda} w_i^2(\mathbb{E}[\mathcal{N}_{i:\lambda}^2]-1)$ and $\mathbb{E}[u_2(\Phi(Z_{\mathbf{e}}))(Z^{\mathrm{T}}\mathbf{A}Z-1)] = \mathbb{E}[u_2(\Phi(Z_{\mathbf{e}}))(Z_{\mathbf{e}}^2-1)]\mathbf{e}^{\mathrm{T}}\mathbf{A}\mathbf{e}$, where the last equality is proved by using the expression $Z^{\mathrm{T}}\mathbf{A}Z = Z_{\mathbf{e}}^2\mathbf{e}^{\mathrm{T}}\mathbf{A}\mathbf{e} + Z_{\mathbf{e}}\mathbf{e}^{\mathrm{T}}\mathbf{A}Z_{\perp} + Z_{\perp}^{\mathrm{T}}\mathbf{A}Z_{\perp}$, the mutual independence between $Z_{\mathbf{e}}$ and $Z_{\perp}$, and $\mathbb{E}[Z_{\perp}]=0$ and $\mathbb{E}[Z_{\perp}^{\mathrm{T}}\mathbf{A}Z_{\perp}] = 1-\mathbf{e}^{\mathrm{T}}\mathbf{A}\mathbf{e}$.

Using $p_{i,j:\lambda}$, we can write

$$\lambda(\lambda-1)u_3(\Phi(x),\Phi(y))p(x)p(y) = \sum_{k=1}^{\lambda-1}\sum_{l=k+1}^{\lambda} w_k w_l \max(p_{k,l:\lambda}(x,y), p_{l,k:\lambda}(x,y)) \ .$$

The equality (22) is obtained by substituting the equality

$$\begin{aligned}
&\lambda(\lambda-1)\mathbb{E}[u_3(\Phi(Z_{\mathbf{e}}),\Phi(\tilde{Z}_{\mathbf{e}}))Z_{\mathbf{e}}\tilde{Z}_{\mathbf{e}}] \\
&= \sum_{k=1}^{\lambda-1}\sum_{l=k+1}^{\lambda} w_k w_l \iint Z_{\mathbf{e}}\tilde{Z}_{\mathbf{e}} \max(p_{k,l:\lambda}(x,y), p_{l,k:\lambda}(x,y))dxdy \\
&= \sum_{k=1}^{\lambda-1}\sum_{l=k+1}^{\lambda} w_k w_l \left(\iint_{x<y} xyp_{k,l:\lambda}(x,y)dxdy + \iint_{x\geq y} xyp_{l,k:\lambda}(x,y)dxdy \right) \\
&= \sum_{k=1}^{\lambda-1}\sum_{l=k+1}^{\lambda} w_k w_l \left(\iint xyp_{k,l:\lambda}(x,y)dxdy + \iint xyp_{l,k:\lambda}(x,y)dxdy \right) \\
&= 2\sum_{k=1}^{\lambda-1}\sum_{l=k+1}^{\lambda} w_k w_l \iint xyp_{k,l:\lambda}(x,y)dxdy \\
&= 2\sum_{k=1}^{\lambda-1}\sum_{l=k+1}^{\lambda} w_k w_l \mathbb{E}[\mathcal{N}_{k:\lambda}\mathcal{N}_{l:\lambda}]
\end{aligned}$$

into $\mathbb{E}[u_3(\Phi(Z_{\mathbf{e}}),\Phi(\tilde{Z}_{\mathbf{e}}))Z^{\mathrm{T}}\mathbf{A}\tilde{Z}] = \mathbb{E}[u_3(\Phi(Z_{\mathbf{e}}),\Phi(\tilde{Z}_{\mathbf{e}}))Z_{\mathbf{e}}\tilde{Z}_{\mathbf{e}}]\mathbf{e}^{\mathrm{T}}\mathbf{A}\mathbf{e}$. The last equality is obtained by using the expression $Z^{\mathrm{T}}\mathbf{A}\tilde{Z} = Z_{\mathbf{e}}\tilde{Z}_{\mathbf{e}}\mathbf{e}^{\mathrm{T}}\mathbf{A}\mathbf{e} + Z_{\mathbf{e}}\mathbf{e}^{\mathrm{T}}\mathbf{A}\tilde{Z}_{\perp} + \tilde{Z}_{\mathbf{e}}\mathbf{e}^{\mathrm{T}}\mathbf{A}Z_{\perp} + Z_{\perp}^{\mathrm{T}}\mathbf{A}\tilde{Z}_{\perp}$, the mutual independence between $Z_{\mathbf{e}}$, $\tilde{Z}_{\mathbf{e}}$, $Z_{\perp}$, and $\tilde{Z}_{\perp}$, and the equalities $\mathbb{E}[Z_{\perp}] = \mathbb{E}[\tilde{Z}_{\perp}] = 0$. □

References

[1] Y. Akimoto, A. Auger, N. Hansen, Quality gain analysis of the weighted recombination evolution strategy on general convex quadratic functions, in: Foundations of Genetic Algorithms, FOGA XIV, 2017, pp. 111–126.

[2] N. Hansen, S. Kern, Evaluating the cma evolution strategy on multimodal test functions, in: Parallel Problem Solving from Nature, PPSN VIII, 2004, pp. 282–291.

[3] N. Hansen, A. Auger, Principled design of continuous stochastic search: from theory to practice, in: Y. Borenstein, A. Moraglio (Eds.), Theory and Principled Methods for the Design of Metaheuristics, Springer, 2014.

[4] N. Hansen, Invariance, self-adaptation and correlated mutations in evolution strategies, in: M. Schoenauer, K. Deb, G. Rudolph, X. Yao, E. Lutton, J.J.M. Guervós, H.-P. Schwefel (Eds.), Parallel Problem Solving from Nature, PPSN VI, Springer, 2000, pp. 355–364.

[5] N. Hansen, A. Auger, R. Ros, S. Finck, P. Pošík, Comparing results of 31 algorithms from the black-box optimization benchmarking bbob-2009, in: Proceedings of Genetic and Evolutionary Computation Conference, 2010, pp. 1689–1696.

[6] L.M. Rios, N.V. Sahinidis, Derivative-free optimization: a review of algorithms and comparison of software implementations, J. Global Optim. 56 (3) (2013) 1247–1293.

[7] N. Hansen, A. Atamna, A. Auger, How to assess step-size adaptation mechanisms in randomised search, in: Parallel Problem Solving from Nature, PPSN XIII, Springer, 2014, pp. 60–69.

[8] O. Krause, T. Glasmachers, C. Igel, Qualitative and quantitative assessment of step size adaptation rules, in: Proceedings of the 14th ACM/SIGEVO Conference on Foundations of Genetic Algorithms, FOGA '17, ACM, New York, NY, USA, 2017, pp. 139–148.

[9] D.V. Arnold, Optimal weighted recombination, in: Foundations of Genetic Algorithms, FOGA VIII, Springer, 2005, pp. 215–237.

[10] N. Hansen, A.S.P. Niederberger, L. Guzzella, P. Koumoutsakos, A method for handling uncertainty in evolutionary optimization with an application to feedback control of combustion, IEEE Trans. Evol. Comput. 13 (1) (2009) 180–197.

[11] T. Yamaguchi, Y. Akimoto, Benchmarking the novel CMA-ES restart strategy using the search history on the bbob noiseless testbed, in: Proceedings of GECCO '17 Companion, 2017, pp. 1780–1787.
[12] Y. Akimoto, N. Hansen, Online model selection for restricted covariance matrix adaptation, in: Parallel Problem Solving from Nature, PPSN XIV, 2016, pp. 3–13.
[13] I. Rechenberg, Evolutionsstrategie '94, Frommann-Holzboog, Stuttgart-Bad Cannstatt, 1994.
[14] H.-G. Beyer, Towards a theory of 'evolution strategies': results for (1+, λ)-strategies on (nearly) arbitrary fitness functions, in: Parallel Problem Solving from Nature, PPSN III, 1994, pp. 58–67.
[15] H.-G. Beyer, The Theory of Evolution Strategies, Nat. Comput. Ser., Springer-Verlag, 2001.
[16] A. Auger, Convergence results for the (1, λ)-SA-ES using the theory of φ-irreducible markov chains, Theoret. Comput. Sci. 334 (1–3) (2005) 35–69.
[17] A. Auger, N. Hansen, Reconsidering the progress rate theory for evolution strategies in finite dimensions, in: Proceedings of Genetic and Evolutionary Computation Conference, GECCO '06, 2006, pp. 445–452.
[18] M. Jebalia, A. Auger, P. Liardet, Log-linear convergence and optimal bounds for the (1+ 1)-es, in: Evolution Artificielle, EA '07, 2008, pp. 207–218.
[19] M. Jebalia, A. Auger, Log-linear convergence of the scale-invariant $(\mu/\mu_w, \lambda)$-ES and optimal μ for intermediate recombination for large population sizes, in: Parallel Problem Solving from Nature, PPSN XI, 2010, pp. 52–62.
[20] A. Auger, Analysis of comparison-based stochastic continuous black-box optimization algorithms, Habilitation, Universitè Paris-Sud, 2015.
[21] R. Ros, N. Hansen, A simple modification in CMA-ES achieving linear time and space complexity, in: Parallel Problem Solving from Nature, PPSN X, 2008, pp. 296–305.
[22] I. Loshchilov, A computationally efficient limited memory CMA-ES for large scale optimization, in: Proceedings of Genetic and Evolutionary Computation Conference, GECCO '14, 2014, pp. 397–404.
[23] Y. Akimoto, N. Hansen, Projection-based restricted covariance matrix adaptation for high dimension, in: Proceedings of Genetic and Evolutionary Computation Conference, GECCO '16, 2016, pp. 197–204.
[24] D.V. Arnold, On the use of evolution strategies for optimising certain positive definite quadratic forms, in: Proceedings of Genetic and Evolutionary Computation Conference, GECCO '07, 2007, pp. 634–641.
[25] J. Jägersküpper, How the $(1 + 1)$ ES using isotropic mutations minimizes positive definite quadratic forms, Theoret. Comput. Sci. 361 (1) (2006) 38–56.
[26] S. Finck, H.-G. Beyer, Weighted recombination evolution strategy on a class of pdqf's, in: Foundations of Genetic Algorithms, FOGA X, 2009, pp. 1–12.
[27] H.-G. Beyer, A. Melkozerov, The dynamics of self-adaptive multirecombinant evolution strategies on the general ellipsoid model, IEEE Trans. Evol. Comput. 18 (5) (2014) 764–778.
[28] H.-G. Beyer, M. Hellwig, The dynamics of cumulative step size adaptation on the ellipsoid model, Evol. Comput. 24 (1) (2016) 25–57.
[29] A. Auger, D. Brockhoff, N. Hansen, Mirrored sampling in evolution strategies with weighted recombination, in: Proceedings of Genetic and Evolutionary Computation Conference, GECCO '11, 2011, pp. 861–868.
[30] M. Jebalia, A. Auger, Log-Linear Convergence of the Scale-Invariant $(\mu/\mu_w, \lambda)$-ES and Optimal μ for Intermediate Recombination for Large Population, Research Report RR-7275, INRIA, 2010.
[31] O. Teytaud, S. Gelly, General lower bounds for evolutionary algorithms, in: Parallel Problem Solving from Nature, PPSN IX, 2006, pp. 21–31.
[32] B. Laurent, P. Massart, Adaptive estimation of a quadratic functional by model selection, Ann. Statist. (2000) 1302–1338.
[33] L. Devroye, Non-Uniform Random Variate Generation, Springer, New York, 1986.
[34] H.-G. Beyer, Mutate large, but inherit small! On the analysis of rescaled mutations in $(\tilde{1}, \tilde{\lambda})$-ES with noisy fitness data, in: Parallel Problem Solving from Nature, PPSN V, 1998, pp. 109–118.
[35] D.V. Arnold, Weighted multirecombination evolution strategies, Theoret. Comput. Sci. 361 (2006) 18–37.
[36] N. Hansen, A. Ostermeier, Completely derandomized self-adaptation in evolution strategies, Evol. Comput. 9 (2) (2001) 159–195.
[37] A. DasGupta, Asymptotic Theory of Statistics and Probability, Springer Science & Business Media, 2008.
[38] H. Robbins, A remark on stirling's formula, Amer. Math. Monthly 62 (1) (1955) 26–29.