On spectral invariance of Randomized Hessian and Covariance Matrix Adaptation schemes

Sebastian U. Stich and Christian L. Müller

MOSAIC group, Institute of Theoretical Computer Science
and Swiss Institute of Bioinformatics,
ETH Zürich, CH 8092 Zürich, Switzerland
sstich@inf.ethz.ch, christian.mueller@inf.ethz.ch

Abstract. We evaluate the performance of several gradient-free variable-metric continuous optimization schemes on a specific set of quadratic functions. We revisit a randomized Hessian approximation scheme (D. Leventhal and A. S. Lewis. Randomized Hessian estimation and directional search, 2011), discuss its theoretical underpinnings, and introduce a novel, numerically stable implementation of the scheme (RH). For comparison we also consider closely related Covariance Matrix Adaptation (CMA) schemes. A key goal of this study is to elucidate the influence of the distribution of eigenvalues of quadratic functions on the convergence properties of the different variable-metric schemes. For this purpose we introduce a class of quadratic functions with parameterizable spectra. Our empirical study shows that (i) the performance of RH methods is less dependent on the spectral distribution than CMA schemes, (ii) that adaptive step size control is more efficient in the RH method than line search, and (iii) that the concept of the evolution path allows a paramount speed-up of CMA schemes on quadratic functions but does not alleviate the overall dependence on the eigenvalue spectrum. The present results may trigger research into the design of novel CMA update schemes with improved spectral invariance.

Keywords: gradient-free optimization, variable metric, Randomized Hessian, Covariance Matrix Adaptation, quadratic functions

1 Introduction

Randomized gradient-free (or black-box) optimization schemes are nowadays a ubiquitous tool for solving many practical problems in science and engineering where gradient or higher order information about the objective are difficult to compute or do not exist. Among the first proposed schemes that are still of considerable (theoretical) importance are adaptive step size random search (aSSRS) [1] and the (almost identical) well-known (1+1)-Evolution Strategy (ES) [2] in Evolutionary Computation (EC). To improve the poor performance of these schemes on ill-conditioned problems several fully adaptive schemes known as gradient-free variable-metric methods have been designed in the past 50 years.

All variable-metric schemes are iterative algorithms that share the idea of adapting a position vector and a quadratic form that defines a local metric between search points to best reflect the *local structure* of the underlying function. In gradient-free optimization two distinct classes of variable-metric methods are known: Randomized Hessian (RH) approximation schemes and Covariance Matrix Adaptation (CMA) schemes.

Randomized Hessian schemes closely follow their deterministic counterparts in nonlinear optimization. However, rather than using exact first- or second-order information they rely on approximations of gradients or Hessians found by finite differences or by estimators based on a finite collection of samples. Such approaches date back at least to 1970's [3]. In an excellent paper Marti [4] proposed several randomized Hessian update schemes taking the perspective of optimal control. Recently, Leventhal and Lewis [5] introduced a genuine RH algorithm with provable convergence guarantees which we further detail in Sec. 2.2.

Covariance Matrix Adaptation schemes follow the principle of sampling search points from the multivariate normal distribution and adapting mean and covariance according to different design principles. The first scheme of this kind, Gaussian Adaptation (GaA) [6], follows the principle of maximum entropy. A very popular modern algorithm is the Covariance Matrix Adaptation Evolution Strategy (CMA-ES) [7, 8]. One recent instantiation of this scheme comprises a derandomization of the sampling termed mirrored sampling [9] which we consider in Sec. 2.3.

One key strength of variable-metric methods is their invariance property to affine transformations. In addition (and more importantly in practice), they achieve the same convergence rate on all functions from the same coset modulo affine transformations, once the affine transformation has been learned in the course of optimization. How fast the different schemes learn affine transformations T is thus of fundamental importance. While theory suggests that the number of samples needed should be at least quadratic in the dimension, it is not yet fully understood how the efficiency of different variable-metric schemes depends on the *eigenvalue spectrum* of TT^T. In the EC community a small number of specific quadratic models have been proposed to probe this dependency. Key instances are the tablet, the discus, the two-axes, and the cigar function, as well as ellipsoidal functions with exponentially increasing eigenvalues [7, 8]. Rather than using these specific functions we here propose a novel set of quadratic functions that varies the shape of the spectral distribution (i) in an easy *parameteric* manner and (ii) with certain common constraints that ease a simpler interpretation of performance results. We demonstrate the excellent discriminative power of the set by numerical experiments with different variable-metric schemes.

The reminder of this paper is structured as follows. In Sec. 2 we outline a standard randomized optimization framework and revisit several representative RH and CMA schemes. In Sec. 3 we introduce the design of the quadratic function set. We also review the Rosenbrock function that serves as a test model with smoothly changing Hessian. In Sec. 4 we summarize the key results of the empirical study. We discuss these results and conclude the paper in Sec. 5.

2 Variable-metric gradient-free optimization schemes

We here present all optimization methods considered in this study. We first detail two non-adaptive randomized schemes, a specific (1+1)-ES and Random Pursuit (RP) [10], that serve as base algorithms. We then show how to couple these algorithms with Leventhal and Lewis's RH scheme. We then detail one instance of GaA [11, 12] and (1,4)-CMA-ES with mirrored sampling and sequential selection [9] with and without evolution path as representative CMA schemes.

genericSearch($\mathbf{x}_0, H_0, N, [\epsilon, \mu, \sigma_0, p]$)

for $k = 1$ *to* N **do**
  if *variable metric* **then**
    $H_k \leftarrow \text{updateHess}(H_{k-1}, \mathbf{x}_{k-1}, \epsilon)$
  else $H_k \leftarrow H_{k-1}$
  $\mathbf{u}_k \sim \mathcal{N}(0, H_k^{-1})$
  if *line search* **then**
    $\mathbf{x}_k \leftarrow \text{lineSearch}(\mathbf{x}_{k-1}, \mathbf{u}_k/\|\mathbf{u}_k\|, \mu)$
  else $(\mathbf{x}_k, \sigma_k) \leftarrow \text{aSS}(\mathbf{x}_{k-1}, \mathbf{u}_k, \sigma_{k-1}, p)$
return $\mathbf{x}_N$

updateHess($H, \mathbf{x}, \epsilon$)

$\mathbf{u} \sim S^{n-1}$
$\Delta_u \leftarrow \frac{f(\mathbf{x}+\epsilon\mathbf{u})-2f(\mathbf{x})+f(\mathbf{x}-\epsilon\mathbf{u})}{\epsilon^2} - \mathbf{u}^T H \mathbf{u}$
if $J := H + \Delta_u \cdot \mathbf{u}\mathbf{u}^T$ *psd* **then**
  $H_+ \leftarrow H + \Delta_u \cdot \mathbf{u}\mathbf{u}^T$
else
  $v \leftarrow \text{smallestEVec}(J)$
  $\Delta_v \leftarrow \frac{f(\mathbf{x}+\epsilon\mathbf{v})-2f(\mathbf{x})+f(\mathbf{x}-\epsilon\mathbf{v})}{\epsilon^2} - \mathbf{v}^T J \mathbf{v}$
  $H_+ \leftarrow \left(H + \Delta_v \cdot \mathbf{v}\mathbf{v}^T\right) + \Delta_u \cdot \mathbf{u}\mathbf{u}^T$
return H_+

aSS($\mathbf{x}, \mathbf{u}, \sigma, p$) *(adaptive step size)*

if $f(\mathbf{x} + \sigma\mathbf{u}) \leq f(\mathbf{x})$ **then**
  $\mathbf{x}_+ \leftarrow \mathbf{x} + \sigma\mathbf{u}$; $\sigma_+ \leftarrow \sigma \cdot \exp(1/3)$
else
  $\mathbf{x}_+ \leftarrow \mathbf{x}$; $\sigma_+ \leftarrow \sigma \cdot \exp\left(-\frac{p}{3(1-p)}\right)$
return $(\mathbf{x}_+, \sigma_+)$

lineSearch($\mathbf{x}, \mathbf{u}, \mu$)

let $\mathbf{x}^* := \mathbf{x} + \arg\min_\lambda f(\mathbf{x} + \lambda\mathbf{u}) \cdot \mathbf{u}$
if *relative accuracy* **then**
  find $\mathbf{x}_+ \in \left[(1-\mu)\mathbf{x} + \mu\mathbf{x}^*, \mathbf{x}^*\right]$
else
  find $\mathbf{x}_+ \in \left[\mathbf{x}^* - \mu\mathbf{u}, \mathbf{x}^* + \mu\mathbf{u}\right]$
return x_+

Fig. 1. Basic building blocks for variable-metric gradient-free optimization

2.1 Isotropic gradient-free optimization schemes

We consider two basic optimization schemes that iteratively generate a sequence of approximate solutions to the optimization problem $\min_{\mathbf{x}} f(\mathbf{x})$ for $f\colon \mathbb{R}^n \mapsto \mathbb{R}$. In each step a search direction is drawn $\mathbf{u} \sim \mathcal{N}(0, \mathbb{1}_n)$. The choice of the step size $\lambda \in \mathbb{R}$ is the key difference between the two schemes.
In Random Pursuit (RP), first proposed in [13] and analyzed in [10], the step size λ is determined by minimizing the objective function in direction $\mathbf{u}$, i.e. $\lambda \approx \arg\min_c f(\mathbf{x} + c\mathbf{u})$. For quadratic functions $f(\mathbf{x}) := \frac{1}{2}\mathbf{x}^T A\mathbf{x}$ with Hessian A, the expected one-step progress can be estimated as:

$$\mathbb{E}\left[f(\mathbf{x}_+) \mid \mathbf{x}\right] \leq \left(1 - \kappa(A^{-1})/n\right) f(\mathbf{x}), \tag{1}$$

where $\mathbf{x}$ is the current iterate, $\mathbf{x}_+ := \mathbf{x} + \lambda\mathbf{u}$ the next iterate, and $\kappa(A^{-1})$ denotes the condition number of A^{-1}. This statement can also be generalized to arbitrary smooth convex functions [10]. Stich et al. [10] showed that both relative and absolute errors in the line search do not hamper the convergence guarantees of RP. We use the built-in MATLAB routine `fminunc.m` with `optimset('TolX'=1E-4)`

as numerical gradient-free line search method with absolute tolerance μ.
In the (1+1)-ES the step size is dynamically controlled such as to approximately guarantee a certain probability p of finding an improving iterate. Depending on the underlying test function different optimality conditions can be formulated for the probability p. Schumer and Steiglitz [1] suggest the setting $p = 0.27$ which is considered throughout this work. We use immediate exponential step size control as explicitly formulated in the aSS sub-routine in Fig. 1. Jägersküpper [14] showed that, for quadratic functions, the dependence of the expected one-step progress of the (1+1)-ES on $\kappa(A)$ is almost identical to the one shown in Eq. (1).

2.2 Randomized Hessian approximation schemes

Assume that the random search direction $\mathbf{u}$ is not chosen from the standard normal distribution, but rather $\mathbf{u} \sim \mathcal{N}(0, H^{-1})$ for a positive definite matrix H. Then a standard analysis [5] shows that the factor of the one-step RP progress in Eq. (1) changes to $(1 - \kappa(HA^{-1})/n)$ for quadratic functions. A refined analysis by Stich et. al. [15] shows a dependence on $\mathrm{Tr}(AH^{-1})$. Hence, if a suitable matrix H with $AH^{-1} \approx \mathbb{1}_n$ can be found, the convergence of RP will be linear with the optimal rate $(1 - 1/n)$. Leventhal and Lewis [5] proposed the following iterative scheme to generate a sequence of Hessian estimates H that converge to A. In each step, a new iterate H_+ is generated as follows:

$$H_+ = H + \mathbf{u}^T (A - H)\, \mathbf{u} \cdot \mathbf{u}\mathbf{u}^T , \tag{2}$$

where $\mathbf{u} \sim S^{n-1}$ is a uniform random unit vector and $\mathbf{u}^T A\mathbf{u}$ is calculated by:

$$\mathbf{u}^T A\mathbf{u} = \left(f(\mathbf{x} + \epsilon\mathbf{u}) - 2f(\mathbf{x}) + f(\mathbf{x} - \epsilon\mathbf{u})\right)/\epsilon^2 , \tag{3}$$

for arbitrary $\epsilon > 0$. Whilst equality only holds for quadratic functions, for general twice differentiable functions the value can be approximated by choosing ϵ sufficiently small. It can be shown [5, Thm. 1] that

$$\mathbb{E}\left[\|H_+ - A\|_F\right] \leq \left(1 - 2/(n(n+2))\right) \|H - A\|_F , \tag{4}$$

holds, and the sequence $(H_k)_{k\geq 1}$ of estimates $H_k \to A$ a.s for $(k \to \infty)$.
Unfortunately, H_+ generated by Eq. (2) is not necessarily positive definite. We thus propose an additional correction step in our implementation of the update. If H_+ is not positive definite we perform a second deterministic update in the direction of the eigenvector that corresponds to the smallest (and the only negative) eigenvalue of H_+. By standard results from matrix perturbation theory, the resulting "twice updated" matrix will be positive (semi-)definite. The algorithm is also illustrated in Fig. 1. As we directly operate on the Cholesky decomposition of H, the condition on line 3 can be efficiently checked. For all quadratic functions we arbitrarily set $\epsilon = 1$, for the Rosenbrock function (see Section 3) we use $\epsilon = 1\text{E-}9$.
Combining the Hessian update with the different step size update schemes from the previous section, we arrive at two variable-metric gradient-free optimization schemes. We will refer to them as RH RP (Randomized Hessian Random Pursuit) and RH (1+1) (Randomized Hessian with aSSRS).

2.3 Covariance Matrix Adaptation schemes

CMA schemes are conceptually different from the presented RH scheme. New search points are sampled from a multivariate normal distribution whose parameter are updated in each iteration based on the information present in the evaluated samples. Many different adaptation schemes exist today. The covariance matrix can be adapted using different rank-1 [6, 7] or rank-k updates [8]. In addition, the CMA-ES scheme is augmented by an auxiliary variable called evolution path that takes into account the correlation of successive means taken over a finite horizon. This is similar in spirit to Rao-Blackwellization techniques in Marko Chain Monte Carlo methods [16] and Polyak's heavy ball method in first-order optimization [17]. We here select two specific instances of CMA schemes: (i) one that is as close as possible to the described RH scheme and (ii) one that is the fastest scheme for quadratic functions known today. The first scheme is a variant of GaA [12] in the (1+1) setting. In every iteration a single sample $\mathbf{x}_k \sim \mathcal{N}(\mathbf{m}_k, \sigma_k^2 C_k)$ is drawn. The aSS sub-routine is employed for step size adaptation. If $f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k)$ the mean $\mathbf{m}_{k+1} = \mathbf{x}_{k+1}$ and the covariance matrix is updated according to $C_k = (1-\alpha)C_k + \alpha(\mathbf{x}_{k+1} - \mathbf{m}_k)(\mathbf{x}_{k+1} - \mathbf{m}_k)^T$ with $\alpha = \log(n+1)/(n+1)^2$ [12]. This constitutes the simplest covariance update without evolution path. The second scheme considered here is the (1,4)-CMA-ES with mirrored sampling and sequential selection. Brockhoff and co-workers state that this scheme "is unbiased and appears to be faster, more robust, and as local as the (1+1)-CMA-ES" [9]. We also refer to [9] for a full description of this scheme and all parameter settings used. For the (1,4)-CMA-ES scheme has been retrieved from `http://coco.gforge.inria.fr/doku.php?id=bbob-2010-results`. We used the GaA code from `http://www.mosaic.ethz.ch/Downloads/GaA`.

3 Benchmark functions

For the presented variable-metric methods there is either theoretical or a large body of empirical evidence that, on quadratic functions, the sequence of estimated Hessians (or inverse Hessians, respectively) will converge after sufficiently many iterations. A reasonable assumption is that the difficulty of the approximation task is mainly determined by the distribution of the eigenvalues of the underlying Hessian. Thus far, this influence has been extensively studied for CMA schemes on specific quadratic model functions such as the tablet, the cigar, or ellipsoidal functions with exponentially increasing eigenvalues (see, e.g., [7, 8]). The exact dependency of variable-metric schemes on the spectral distribution remains, however, largely elusive because the spectral properties such as trace and condition are not constant across experiments. For RH schemes, we are not aware of any systematic empirical study. From the theory of RH schemes we know, however, that the expected progress for a fixed Hessian estimate H depends on $\kappa(AH^{-1})$ as well as on $\mathrm{Tr}(AH^{-1})$ where A denotes the Hessian [5, 15]. We thus propose a class of quadratic functions with different spectra under the constraint of equal trace and condition number L. The functions are constructed

Table 1. List of benchmark functions. All functions are quadratic except f_{Rosen}. For $f_{\text{Sigm}(a)}$ we use $a = 15, 8, 5, 2.8$ and for $f_{\text{Flat}(a)}$ we use $a = 6, 3.2, 2, 1.25$. The spectra of all the quadratic functions are depicted in Fig. 2b.

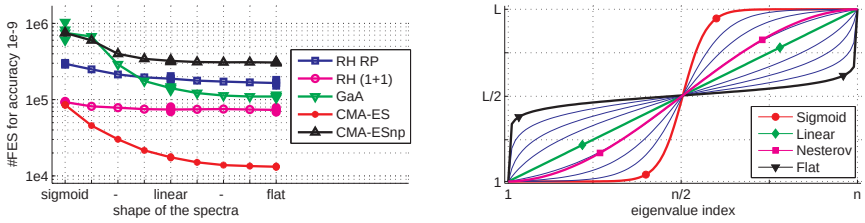
$$
\begin{aligned}
f_{\text{Sigm}(a)}(\mathbf{x}) &= \textstyle\sum_{i=1}^{n} \text{normalize}_i \left(\left(1 + e^{a - \frac{2a(t-1)}{n-1}} \right)^{-1} + \frac{1}{2} \right) (x_i - 1)^2 \\
f_{\text{Flat}(a)}(\mathbf{x}) &= \textstyle\sum_{i=1}^{n} \text{normalize}_i \left(-\log \left(\left(10^{-a} + \frac{(t-1)(1-2\cdot 10^{-a})}{n-1} \right)^{-1} - 1 \right) \right) (x_i - 1)^2 \\
f_{\text{Lin}}(\mathbf{x}) &= \textstyle\sum_{i=1}^{n} \text{normalize}_i \left(\frac{2t}{n+1} - 1 \right) (x_i - 1)^2 \\
f_{\text{Nes}}(\mathbf{x}) &= \textstyle\sum_{i=1}^{n} \text{normalize}_i \left(\sin \left(\frac{t\pi}{n+1} - \frac{\pi}{2} \right) \right) (x_i - 1)^2 \\
f_{\text{Rosen}}(\mathbf{x}) &= \textstyle\sum_{i=1}^{n-1} \left(100(x_{i+1} - x_i^2)^2 + (x_i - 1)^2 \right)
\end{aligned}
$$

$$
\text{normalize}_i(f(t)) = \frac{L-1}{2} \frac{f(i)}{|f(1)|} + \frac{L+1}{2}
$$

as follows: We choose the distribution of the Hessian eigenvalues according to three specific parametric functions outlined below. We then normalize the spectra such that the smallest eigenvalue equals 1, the largest equals L, and the trace equals $n(L+1)/2$ where n denotes the dimension. The function set is summarized in Tab. 1. For f_{Lin} the eigenvalues are linearly spaced. For f_{Sigm} the eigenvalues lie on the sigmoidal curve $(1 + e^{-t})^{-1}$ resulting in many eigenvalues being close to 1 or close to L with only a few intermediate eigenvalues. By distributing the eigenvalues proportional to the inverse of the sigmoid function $\log(1/t - 1)$ we find a family of suitable quadratic functions where most eigenvalues are concentrated around the mean $L/2$. The exact parameterizations are summarized in Tab. 1. The shape of the spectra is depicted in Fig. 2b. For large a we note that (i) $f_{\text{Sigm(a)}}$ becomes similar to the two-axes function [8] (half of the eigenvalues are 1, half of them are L) and (ii) $f_{\text{Flat(a)}}$ gets close to a cigar-like function (with one small eigenvalue and all others on the order of L). Another important feature of our parametric family is the fact that the sigmoidal function can closely approximate Nesterov's worst case function f_{Nes} [18] which has been used to show a lower complexity bound for first-order optimization (see again Fig. 2b for a sketch). Note that the present trace constraint prohibits the design of quadratic functions with exponentially distributed eigenvalues.
Finally, we also include the standard Rosenbrock function f_{Rosen} in the test set. The function serves as a test model with smoothly changing Hessian in order to study the valley-following abilities of the different variable-metric schemes.

4 Empirical study

We now highlight the key results of our empirical study. All algorithms and functions have been implemented in MATLAB and will be made publicly available at the authors' website. The (1,4)-CMA-ES with mirrored sampling and sequential selection (referred to as CMA-ES in the following) has been run both with and without evolution path (setting `CMA.ccum=1` in the referred MATLAB code). The latter variant is referred to as CMA-ESnp. For all performed experiments



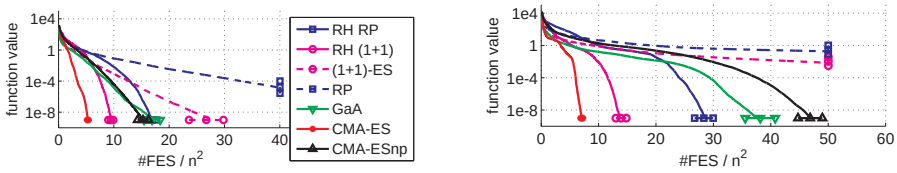
(a) Comparison for different spectra. (b) The spectra of the functions.

Fig. 2. (a): Relation between method performance and spectral distribution in $n = 50$ for $L = 1\text{E}6$. We recorded #FES needed to reach accuracy 1E-9 on all parametrized functions f_{Sigm}, f_{Flat} and f_{Lin}; the median of 51 runs is indicated by a marker. (b): Shape of the spectra of the quadratic benchmark functions. Thin blue lines show $f_{\text{Sigm(a)}}$ and $f_{\text{Flat(a)}}$ for intermediate a values.

the initial settings were $\mathbf{x}_0 = \mathbf{0}$, $H_0 = \mathbb{1}_n$ ($\mathbf{m}_0 = \mathbf{0}$, $C_0 = \mathbb{1}_n$, respectively). The initial step size of the algorithms with adaptive step size control was (empirically) set such that the target success probability $p = 0.27$ is met for $\mathbf{x}_0$. As performance measure we count the number of function evaluations (#FES) needed to reach a target function value below 1E-9.

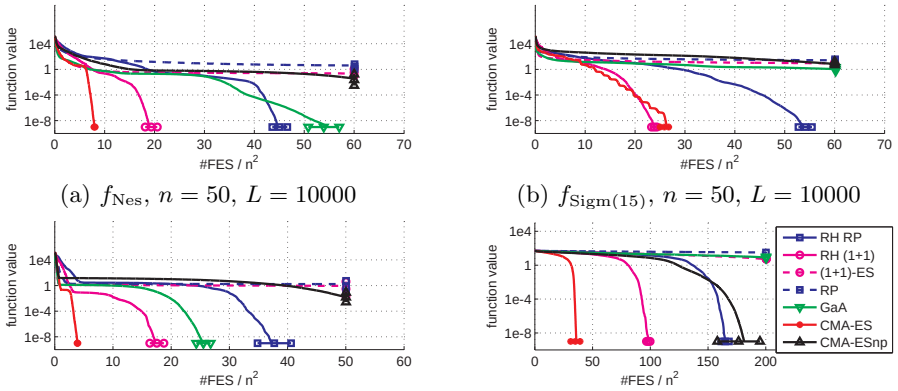
We first demonstrate the general influence of the spectral distribution on the performance of all introduced variable-metric schemes. Experimental set-up and results are summarized in Fig. 2. For all CMA schemes (CMA-ES, CMA-ESnp, and GaA) we see a strong monotone dependence of their performance on the spectral shape. The sigmoidal-shaped eigenspectrum presents the hardest problem, the flat spectrum the easiest. Both CMA-ES and GaA show the strongest run time dependence on the spectra with CMA-ES being the fastest algorithm on all functions and CMA-ESnp the slowest one. The performance of both RH schemes is much less dependent on the shape with RH (1+1) being almost invariant to the spectral distribution. We observe that RH (1+1) achieves the same performance on $f_{\text{Sigm}(15)}$ (the leftmost datum in Fig. 2a) as CMA-ES.

To study the influence of the condition number L on the qualitative convergence behavior of the different algorithms we present results on f_{Nes} as representative example in Figs. 3 and 4a for fixed dimension $n = 50$. The same



(a) f_{Nes}, $n = 50$, $L = 100$ (b) f_{Nes}, $n = 50$, $L = 1000$

Fig. 3. Evolution of function value vs. #FES on f_{Nes} for $L = 100$ (a) and $L = 1000$ (b). We recorded #FES needed to reach accuracy 1E-9. The median trajectory of 11 runs is depicted; mean and one standard deviation are indicated by markers.

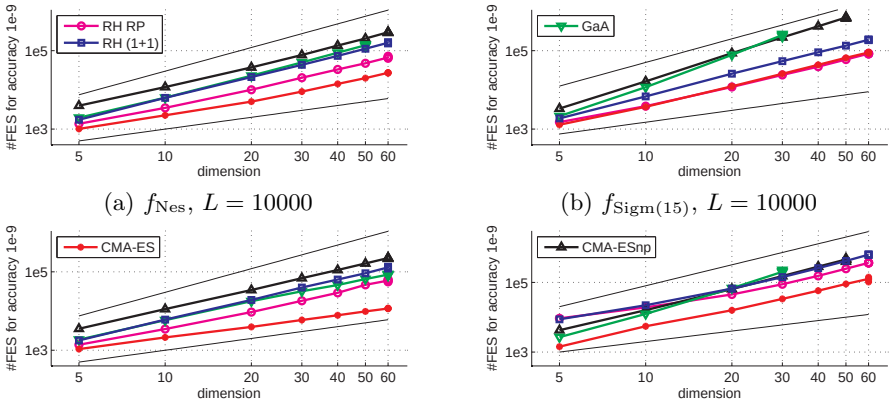


(a) f_{Nes}, $n = 50$, $L = 10000$ (b) $f_{\mathrm{Sigm}(15)}$, $n = 50$, $L = 10000$

(c) $f_{\mathrm{Flat}(6)}$, $n = 50$, $L = 10000$ (d) f_{Rosen}, $n = 50$

Fig. 4. Evolution of function value vs. #FES for different functions. We recorded #FES needed to reach accuracy 1E-9. The median trajectory of 11 runs is depicted; mean and one standard deviation are indicated by markers.

qualitative behavior has been observed for the other quadratic functions. We see that the non-adaptive schemes RP and (1+1)-ES are (as expected) not even competitive for $L = 100$. We will thus concentrate on variable-metric schemes in the further discussion. For $L \geq 1000$ we observe that the convergence behavior of all variable-metric methods can be divided into three phases, (i) an initial short tune-in phase with rapid progress, (ii) a learning (or adaptation) phase with marginal progress in function value (of length quadratic in n), and (iii) a convergence phase with strong function value decrease (of length linear in n). Moreover, we see that the slope of the trajectory *at the level of the target accuracy* is distinct for all schemes. This measured convergence rate reflects the efficiency of the adaptation process of the different schemes at this function value level. CMA-ES and RH (1+1) show the steepest descent, CMA-ESnp and GaA the flattest one. For $L = 10000$ CMA-ESnp and the non-adaptive methods do not reach the target accuracy within a FES budget of $60n^2$. These observations are generally confirmed on other quadratic functions having different spectral shapes with a few notable exceptions. We here exemplify the performance of the schemes on the two most extreme functions $f_{\mathrm{Sigm}(15)}$ and $f_{\mathrm{Flat}(6)}$ (with $L = 10000$ in $n = 50$) as well as on f_{Rosen} (as shown in Fig. 4). On $f_{\mathrm{Flat}(6)}$ (cf. Fig. 4c) the Hessian is well-approximated by all convergent schemes. The convergence rate in phase (iii) is best for CMA-ES followed by RH (1+1) and GaA. RH RP' convergence takes longer because per line search 5-10 FES are needed on average. CMA-ESnp is still in the adaptation phase within the displayed FES budget. On $f_{\mathrm{Sigm}(15)}$ (cf. Fig. 4b) we observe that CMA-ES' convergence rate is slower than the one of RH (1+1) above accuracy 1E-7 eventually converging at optimal rate below this level. This indicates that CMA-ES is still in adaptation phase even at low function value level. Both CMA-ESnp and GaA are still in the adaptation phase within the displayed FES budget. Inspection of the convergence trajectories also



(a) f_{Nes}, $L = 10000$ (b) $f_{\mathrm{Sigm}(15)}$, $L = 10000$

(c) $f_{\mathrm{Flat}(6)}$, $L = 10000$ (d) f_{Rosen}

Fig. 5. #FES to reach the target accuracy vs. dimension n in log-log scale. The median of 11 runs is depicted by a marker for all converged runs within the considered #FES budget. Thin lines indicate quadratic scaling (top) or linear scaling (bottom) .

reveals that the length of learning phase is responsible for CMA-ES' observed dependence on the spectral shape.

The experiments on f_{Rosen} confirm that all variable-metric schemes (except GaA) can efficiently learn a smoothly changing Hessian (without tune-in phase) confirming and extending known results for RH schemes [5] and CMA schemes [7]. We finally show the scaling behavior of the algorithms on selected functions in Fig. 5. All algorithms show the expected quadratic scaling with dimension (for $n \geq 20$) with two notable exceptions. While GaA and CMA-ESnp on $f_{\mathrm{Sigm}(15)}$ and GaA on f_{Rosen} exhibit scaling of higher order than quadratic, CMA-ES shows super-linear convergence on $f_{\mathrm{Flat}(6)}$. The latter result is in full agreement with the empirical tests of CMA-ES on the cigar function [7, 8].

5 Discussion and Conclusions

We have empirically tested the performance of several randomized gradient-free variable-metric optimization schemes on a novel set of quadratic functions whose spectral distribution ranges (for any fixed dimension n and condition number L) from a near-flat distribution to a sigmoidal shape under constant trace constraint. Using this benchmark set we have been able to show a clear *monotonic dependence* of the performance of CMA schemes on the shape of the spectrum. From the data we also conclude that the concept of the evolution path allows a paramount speed-up of CMA schemes but does not alleviate the dependence on the eigenvalue spectrum. The presented Randomized Hessian (RH) approximation schemes [5], on the other hand, have been shown to be less dependent or almost invariant to the specific distribution of eigenvalues. Our empirical results also indicate that coupling our novel, numerically stable implementation of the RH scheme with adaptive step size control is more efficient than a

scheme with approximate line search on all tested problems. We believe that the present results may trigger research into the design of novel CMA update schemes with improved spectral invariance. We also advocate the embedding of the proposed function set (most prominently the sigmoidal ones) in modern black-box optimization benchmark test suites. Investigating quadratic function sets under constant determinant and condition constraints (thus allowing exponentially distributed eigenvalues) will be subject of future research.

References

1. Schumer, M., Steiglitz, K.: Adaptive step size random search. Automatic Control, IEEE Transactions on **13**(3) (1968) 270–276
2. Rechenberg, I.: Evolutionsstrategie; Optimierung technischer Systeme nach Prinzipien der biologischen Evolution. Frommann-Holzboog (1973)
3. Betro, B., De Biase, L.: A Newton-like method for stochastic optimization. In: Towards Global Optimization. Volume 2. North-Holland (1978) 269–289
4. Marti, K.: Controlled random search procedures for global optimization. In: Stochastic Optimization. Volume 81 of Lecture Notes in Control and Information Sciences. Springer (1986) 457–474
5. Leventhal, D., Lewis, A.S.: Randomized Hessian estimation and directional search. Optimization **60**(3) (2011) 329–345
6. Kjellström, G., Taxen, L.: Stochastic Optimization in System Design. IEEE Trans. Circ. and Syst. **28**(7) (July 1981)
7. Hansen, N., Ostermeier, A.: Completely Derandomized Self-Adaption in Evolution Strategies. Evolutionary Computation **9**(2) (2001) 159–195
8. Hansen, N., Muller, S.D., Koumoutsakos, P.: Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (CMA-ES). Evol. Comput. **11**(1) (Spr 2003) 1–18
9. Brockhoff, D., Auger, A., Hansen, N., Arnold, D., Hohm, T.: Mirrored Sampling and Sequential Selection for Evolution Strategies. In: PPSN XI. Volume 6238 of LNCS. Springer (2010) 11–21
10. Stich, S.U., Müller, C.L., Gärtner, B.: Optimization of convex functions with Random Pursuit. http://arxiv.org/abs/1111.0194 (2011)
11. Müller, C.L., Sbalzarini, I.F.: Gaussian adaptation revisited - an entropic view on covariance matrix adaptation. In: EvoStar. LNCS, Springer (2010) 432–441
12. Müller, C.L., Sbalzarini, I.F.: Gaussian Adaptation as a unifying framework for continuous black-box optimization and adaptive Monte Carlo sampling. In: Evolutionary Computation (CEC), 2010 IEEE Congress on. (2010) 1–8
13. Mutseniyeks, V.A., Rastrigin, L.A.: Extremal control of continuous multiparameter systems by the method of random search. Eng.Cyb. **1** (1964) 82–90
14. Jägersküpper, J.: Rigorous runtime analysis of the (1+1) ES: 1/5-rule and ellipsoidal fitness landscapes. In: FOGA. Volume 3469 of LNCS. (2005) 356–361
15. Stich, S.U., Gärtner, B., Müller, C.L.: Variable Metric Random Pursuit. in preparation for Math. Prog. (2012)
16. Andrieu, C., Thoms, J.: A tutorial on adaptive MCMC. Statistics and Computing **18**(4) (December 2008) 343–373
17. Polyak, B.: Introduction to Optimization. Optimization Software - Inc, Publications Division, New York (1987)
18. Nesterov, Y.: Introductory Lectures on Convex Optimization. Kluwer, Boston (2004)