12th International Conference on Parallel Problem Solving From Nature - PPSN XII

On spectral invariance of Randomized Hessian and Covariance Matrix Adaptation schemes

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Sebastian U. Stich
stich@inf.ethz.ch

Christian L. Müller
christian.mueller@inf.ethz.ch

Motivation

Let $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T H\mathbf{x}$ quadratic to minimize. The simple random search scheme in [1] draws search directions $\mathbf{u}$ from *fixed* distribution $\mathcal{N}(\mathbf{0}, C)$ and generates improving iterates by a line search: $\mathbf{x}_+ = \mathbf{x} + \arg\min_t f(\mathbf{x}+t\mathbf{u})\cdot\mathbf{u}$. The progress depends on the spectra of H and can be estimated [2] as:

$$f(\mathbf{x}_+) \lessapprox \left(1 - \frac{\lambda_{\min}(HC)}{\mathrm{Tr}[HC]}\right) f(\mathbf{x})$$

Study the influence of the *eigenvalue spectra* of the Hessian H on the performance of *variable metric schemes*.

Contribution

We consider Covariance Matrix Adaptation schemes (CMA-ES [3], Gaussian Adaptation (GaA) [4]) and Randomized Hessian (RH) schemes from Leventhal and Lewis [5].

We provide a new, numerically stable implementation for RH and, in addition, combine the update with an adaptive step size strategy.

We design a class of quadratic functions with parametrizable spectra to study the influence of the spectra on the performance of variable metric schemes.

We empirically study 5 variable metric schemes on this function class and on Rosenbrock's function.

Conclusion

We observe a *monotonic dependence* of the performance of the studied variable metric schemes on the shape of the eigenvalue spectra.

The sigmoidal-shaped spectra of f_{Sigm} presents the hardest learning problem for all tested variable metric schemes.

Randomized Hessian schemes are less dependent on the spectra.

The use of an evolution path is crucial for CMA schemes to get superior performance.

Randomized Hessian update with adaptive step size has shown to be an effective variable metric scheme.

Algorithms

Randomized Hessian Update [2, 5]

- curvature estimate in a random direction $\mathbf{u} \sim S^{n-1}$
- rank-1 update of Hessian estimate H
- 2-4 additional function evaluations

$\Delta_u \leftarrow \frac{f(\mathbf{x}+\epsilon\mathbf{u})-2f(\mathbf{x})+f(\mathbf{x}-\epsilon\mathbf{u})}{\epsilon^2} - \mathbf{u}^T H\mathbf{u}$
if $J := H + \Delta_u \cdot \mathbf{u}\mathbf{u}^T$ *pd* **then**
  $H_+ \leftarrow H + \Delta_u \cdot \mathbf{u}\mathbf{u}^T$
else
  $v \leftarrow \mathrm{smallestEigenVector}(J)$
  $\Delta_v \leftarrow \frac{f(\mathbf{x}+\epsilon\mathbf{v})-2f(\mathbf{x})+f(\mathbf{x}-\epsilon\mathbf{v})}{\epsilon^2} - \mathbf{v}^T J\mathbf{v}$
  $H_+ \leftarrow \left(H + \Delta_v \cdot \mathbf{v}\mathbf{v}^T\right) + \Delta_u \cdot \mathbf{u}\mathbf{u}^T$

RH with line search (RH RP)

- *alternate* between Hessian estimate and search
- search direction $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, H^{-1})$
- line search to guarantee sufficient decrease

$$\mathbf{x}^* := \mathbf{x} + \arg\min_{t\in\mathbb{R}} f(\mathbf{x}+t\mathbf{u})\cdot\mathbf{u}$$
$$\mathbf{x}_+ \in [(1-\mu)\mathbf{x} + \mu\mathbf{x}^*, \mathbf{x}^*]$$

RH with adaptive step size control

- *alternate* between Hessian estimate and search
- search direction $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, H^{-1})$
- adaptive step size control (RH (1+1))

success:	$\mathbf{x}_+ = \mathbf{x} + \sigma\mathbf{u}$	$\sigma_+ = \sigma \cdot 1.40$
failure:	$\mathbf{x}_+ = \mathbf{x}$	$\sigma_+ = \sigma \cdot 0.88$

Gaussian Adaptation [4]

- covariance estimation based on maximum entropy principle, *rank-1 update along search direction*
- search direction $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, C)$
- adaptive step size control

CMA-Evolution Strategy [3]

- (1,4)-CMA-ES with *mirrored sampling* and sequential selection, default parameter setting
- rank-1 update based on covariance estimation and evolution path
- search direction $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, C)$

CMA-ES without Evolution Path

- (1,4)-CMA-ES with mirrored sampling and sequential selection, *no evolution path* (CMA-ESnp)
- rank-1 update based on covariance estimation

Convergence dependent on spectra

- CMA-ES and GaA show strongest dependence on spectra
- RH schemes are less dependent on spectra

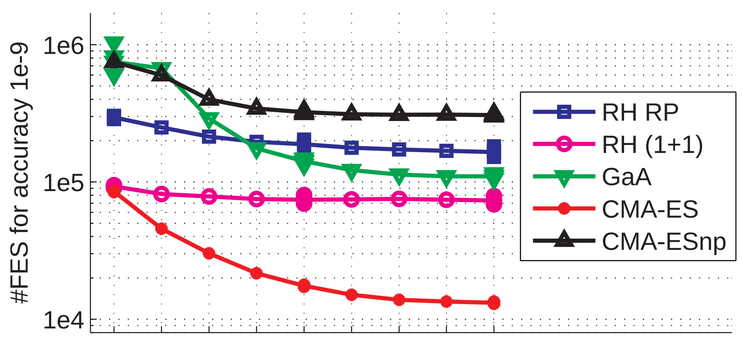

Figure: Relation between method performance and spectral distribution in $n = 50$ for $L = 1\mathrm{E}6$. We recorded #FES needed to reach accuracy 1E-9 on all parametrized functions f_{Sigm}, f_{Flat} and f_{Lin}; the median of 51 runs is indicated by a marker.

Test functions

- Quadratic functions $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T H\mathbf{x}$
- standard Rosenbrock function as non-convex test case

Simple design principles for the Hessian H:

- $H \in \mathbb{R}^n \times \mathbb{R}^n$ positive definite (*pd*)
- fixed condition number $\kappa(H) = L$
- fixed trace $\mathrm{Tr}[H] = \frac{n(L+1)}{2}$
- spectra $\lambda(H)$ easy to parametrize

Generic extension to known test functions (tablet, cigar).

$f_{\mathrm{Sigm}(15)}$	many eigenvalues close to 1 or to L
f_{Lin}	linearly spaced eigenvalues
$f_{\mathrm{Flat}(6)}$	most eigenvalues concentrated at the mean $L/2$
f_{Nes}	the same spectra as Nesterov's worst case function for first order methods
f_{Rosen}	smoothly changing Hessian

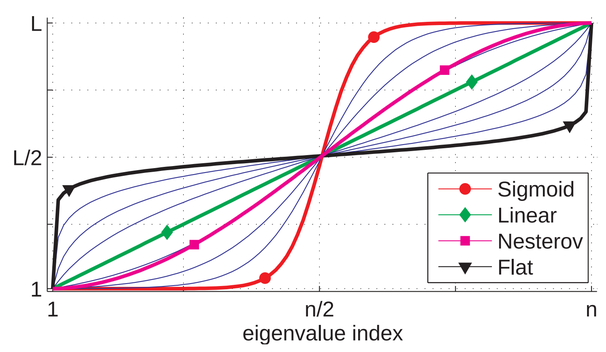

Figure: Shape of the spectra of the quadratic benchmark functions.

Scaling in dimension

- learning phase scales *quadratically* in the dimension n
- RH schemes *quadratic* throughout testbed
- GaA and CMA-ESnp on $f_{\mathrm{Sigm}(15)}$ *sub-quadratic*
- CMA-ES on $f_{\mathrm{Flat}(6)}$ *super-linear*

Trajectories

- convergence in three phases: (i) tune-in, (ii) learning, (iii) fast convergence
- convergence rate at *the level of the target accuracy* is best for CMA-ES and RH (1+1)

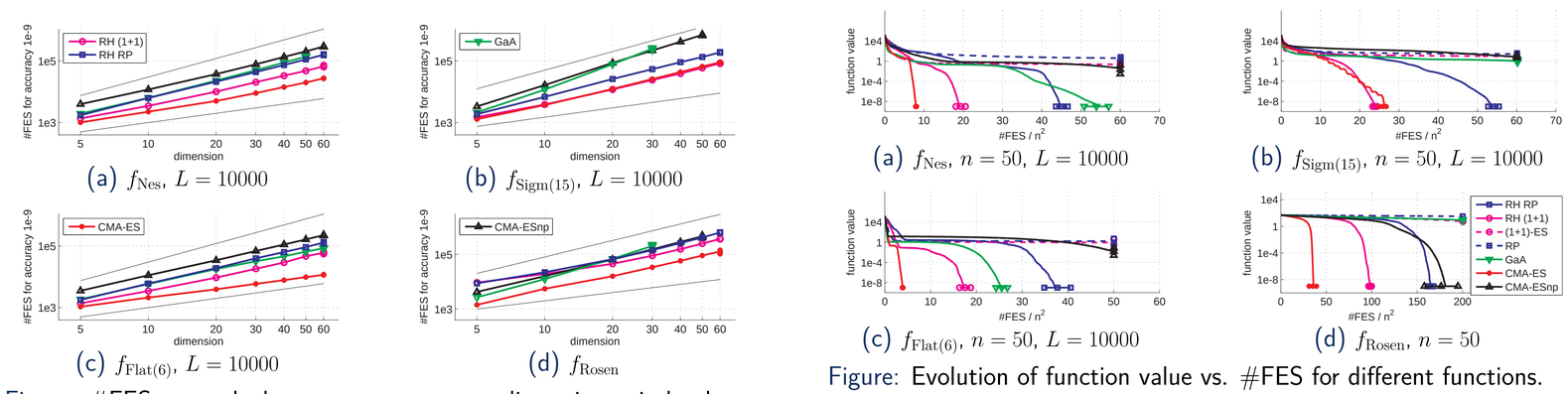

Figure: #FES to reach the target accuracy vs. dimension n in log-log scale. The median of 11 runs is depicted by a marker for all converged runs within the considered #FES budget. Thin lines indicate quadratic scaling (top) or linear scaling (bottom) .

Figure: 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.

Selected References

[1] Stich, S.U., Müller, C.L., Gärtner, B.:
Optimization of convex functions with Random Pursuit.
http://arxiv.org/abs/1111.0194 (2011)

[2] Stich, S.U., Gärtner, B., Müller, C.L.:
Variable Metric Random Pursuit.
in preparation for Math. Prog. (2012)

[3] 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

[4] 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 (2010) 1–8

[5] Leventhal, D., Lewis, A.S.:
Randomized Hessian estimation and directional search.
Optimization **60**(3) (2011) 329–345

Acknowledgments

We thank Dr. Bernd Gärtner and Dr. Ivo F. Sbalzarini for useful discussions.
Supported by the project CG Learning (FET-Open grant number: 255827).