MADALGO & CTIC Summer School 2011

Randomized Derivative-Free Optimization

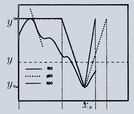
Sebastian U. Stich
sstich@ethz.ch
Institute of Theoretical Computer Science
Swiss Institute of Bioinformatics

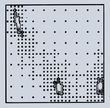
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Standard Approaches

Pure Random Search [1a]

- sample points uniformly in space
- exponential running time
- Pure adaptive random search (PAS) [1b]: sample sequence of points, all new points uniformly among those with better function value → not possible to do
- PAS analogous to randomized version of method of centers



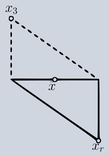


DIRECT [1c]

- iteratively evaluate function at grid
- in each iteration locally refine grid cells if potential to improve function value
- exponential running time

Nelder-Mead [1d]

- evaluate function at vertices of a simplex
- propose new search point: reflection of worst point through centroid
- can converge to non-stationary points
- Adaptation: simplex derivatives [1e]



Trust Region [1f]

- approximate objective function with models (often quadratic) on regions
- refine/expand regions on which the approximation is bad/good

[1a] Samuel H. Brooks. A discussion of random methods for seeking maxima. *Operations Research*, 6(2):244–251, 1958.

[1b] Zelda B. Zabinsky and Robert L. Smith. Pure adaptive search in global optimization. *Mathematical Programming*, 53:323–338, 1992.

[1c] Jones, Perttunen, and Stuckman. Lipschitzian optimization without the lipschitz constant. *Journal of Optimization Theory and Applications*, 79:157–181, 1993.

[1d] J. A. Nelder and R. Mead. A simplex method for function minimization. *The Computer Journal*, 7(4):308–313, 1965.

[1e] Andrew R. Conn, Katya Scheinberg, and Luis N. Vicente. *Introduction to derivative-free optimization*. 2009

[1f] Andrew R. Conn, Nicholas I. M. Gould, and Philippe L. Toint. *Trust-region method*. 2000

Evolution Strategies Explained

- $(1+1)$-ES described already in 1968 [2a]
- basis of more complex methods like CMA-ES
- convergence of (1+1)-ES linear on quadratic functions [2b]
- theoretical convergence results only known for very simple functions

Generic $(\mu+\lambda)$-Evolution Strategy (ES)

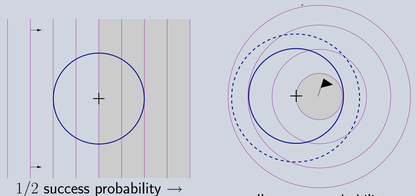
1: **for** $k=0$ to N **do**
2: $\mathcal{X}_k \leftarrow$(λ new points, based on $\mathcal{M}_{k-1}$)
3: $\textsc{EvaluateFunction}\ (\mathcal{X}_k)$
4: $\mathcal{M}_k \leftarrow \textsc{Select}\ \mu$ best points among $\mathcal{X}_k \cup \mathcal{M}_{k-1}$
5: **end for**
6: **return** $\textsc{BestPoint}\left(\bigcup_{k=0}^{N} \mathcal{M}_k\right)$

For (1+1)-ES: (with *step-size* σ_k)

$$x_{k+1} \sim \mathcal{N}(m_k, \sigma_k \cdot I_n) \qquad m_k = \textsc{Best}(x_k, m_{k-1})$$

Most ES work by only considering rank information of the current iterates. (Invariant under monotone transformations).

Step-size adaptation of ES [2c]



$1/2$ success probability → increase σ

small success probability → decrease σ

[2a] M. Schumer and K. Steiglitz. Adaptive step size random search. *Automatic Control, IEEE Transactions on*, 13(3):270 – 276, 1968.

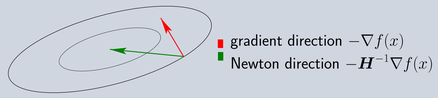
[2b] Jens Jägerskküpper. Rigorous runtime analysis of the (1+1) es: 1/5-rule and ellipsoidal fitness landscapes. *Foundations of Genetic Algorithms*, 3469:356–361. 2005.

[2c] A. Auger and N. Hansen. Tutorial: Evolution strategies and covariance matrix adaptation. available online.

The Right Metric

For f quadratic:

$$f(x) = f(x^*) + \frac{1}{2}(x-x^*)\boldsymbol{H}(x-x^*)$$
$$\boldsymbol{H} = \nabla^2 f(x^*)$$



- if $\boldsymbol{H} \approx I_n$ first order information is sufficient
- otherwise estimation of $\boldsymbol{H}^{-1}$ is necessary
- invariant under quadratic transformations
- estimating $\boldsymbol{H}$ only from zeroth-order information is difficult
- many update schemes are reasonable and used

Rank 1 update. Estimate

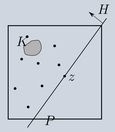
$$\boldsymbol{H}_{k+1}^{-1} \approx (1-\lambda_k)\boldsymbol{H}^{-1} + \lambda_k y_k y_k^T$$

y_k some vector (for example: $y_k \propto m_{k+1} - m_k$), λ_k weight

Other Methods

Hit and Run [7a]

- draw a random direction and on the segment contained in the convex body uniformly the next iterate
- originally proposed as sampling method to generate uniform samples from convex body [7b]
- fast mixing times (in special settings, e.g. log-convave functions) proven by Lovász [7c] and Dyer [7d]
- introduced as optimization algorithm by Bertsimas [7a]

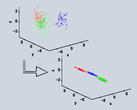


Random Cutting Plane [7e]

- based on cutting plane method [7f]
- RCP algorithm described first by Levin [7g]
- natural extension of 1D bisection method
- Not practical usable as for a generic convex set, computing the center of gravity is more difficult than solving the original optimization problem.
- Dabbene et al. [7e] approximate centroid by hit-and-run sampling

Random Conic Pursuit [7h]

- random direction (rank 1 matrix) is chosen at random and optimization over 2-dim subspace yields next iterate
- convergence on unconstrained SDP
- several adaptation heuristics proposed in the paper (adaptation of sampling distribution, boundaries) but no theoretical results



Picture [7j]

Simulated Annealing. [7k, 7l]

- inspiration and name from annealing in metallurgy
- new iterates proposed by sampling random points in a weighted neighborhood (acceptance rate changes according to *cooling schedule*)
- convergence results, but only low rates
- adaptation of Metropolis-Hastings algorithm

[7a] D. Bertsimas and S. Vempala. Solving convex programs by random walks. *J. ACM*, 51:540–556, July 2004.

[7b] R. L. Smith. Efficient Monte-Carlo procedures for generating points uniformly distributed over bounded regions. *Operation Research*, 32:1296i£¡-1308, 1984.

[7c] László Lovász. Hit-and-run mixes fast. *Mathematical Programming*, 86:443–461, 1999.

[7d] M. Dyer, A. Frieze, and R. Kannan. A random poly.-time alg. for approx. the volume of conv. bodies. *J. ACM*, 38:1–17, January 1991.

[7e] F. Dabbene, P. S. Shcherbakov, and B. T. Polyak. A randandomized cutting plane method with probabilistic geometric convergence. *Siam J. Optim.*, 20(6):3185–3207, 2010.

[7f] Jr. Kelley, J. E. The cutting-plane method for solving convex programs. *J. of the Soc. for Ind. and App. Mathematics*, 8(4):pp. 703–712, 1960.

[7g] A. Levin. On an algorithm for the minimization of convex functions. *Soviet Math. Dokl.*, 6:286–290, 1965.

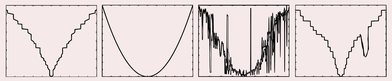
[7h] A. Kleiner, A. Rahimi, and M. Jordan. Random conic pursuit for SDP. *NIPS 23*:1135–1143. 2010.

[7j] E. P. Xing et al. Distance metric learning with application to clustering. In *NIPS'02*:505–512, 2002.

[7k] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Opt. by simulated annealing. *Science*, 220(4598):671–680, 1983.

[7l] V. Černý, Thermodynamical approach to the traveling salesman problem: An efficient simulation algorithm. *J. of Opt. Th. and App.*, 45:41–51, 1985.

Problem Statement



$$\arg\min_{x\in\mathbb{R}^n} f(x)$$

- access only zeroth-order information
- objective fuction given as black-box
- multi-modal functions (global minimizer)
- robust against (random) noise
- provable (fast) convergence?

Summary

For smooth functions, standard methods (gradient, Newton) will converge to local minimizer but are not robust against noise. Evolution strategies have been proven to effective in this setting. Highly developed methods (like CMA-ES) are at the moment (among) the best performing algorithms.

- convergence proof of random variants of standard methods relatively easy for convex functions
- convergence proofs for general ES still missing
- main issues: self-adaptation of strategy parameters (stepsize, Hessian estimation, lots of heuristics...)

Open Problems and Goals

Theoretical results to support experimental evidence.

- rigorous convergence results to compare with classical methods
- convergence/divergence results for noisy and multi-modal functions

Especially interesting would be to analyze particular heuristic used by many algorithms to self-estimate strategy parameters.

- step-size adaptation heuristics
- Hessian estimation (eg. by rank-μ updates)
- limited memory Hessian estimation possible?

Random Gradient [3a]

iteratively update

$$x_{k+1} = x_k - h_k g_\mu(x_k)$$

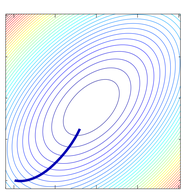
g directional derivative/finite difference in random direction:

$$g_0(x) = f'(x,u)\cdot u$$
$$g_\mu(x) = \frac{f(x+\mu u) - f(x)}{\mu}\cdot u$$

$u \sim \mathcal{N}(0, I_n)$

$h_k = h_k(L_1)$ stepsizes

μ accuracy



- First results by Polyak [3b], completely analyzed by Nesterov [3a] using Gaussian smoothing [3c]
- *random oracle* for derivative
- convergence for convex functions with L_1-Lipschitz continuous gradients
- **accelerated methods reach (provable) existing lower complexity bounds [3d] by factor $O(n)$**

[3a] Yu. Nesterov. Random Gradient-Free Minimization of Convex Functions. Technical report, ECORE, 2011.

[3b] B. Polyak. *Introduction to Optimization. Optimization Software - Inc.* Publications Division, New York, 1987.

[3c] A.Nemirovsky and D.Yudin. *Problem complexity and method efficiency in optimization.* John Wiley and Sons, New York, 1983.

[3d] Yu. Nesterov. *Introductory Lectures on Convex Optimization.* Kluwer, Boston, 2004.

Random Pursuit [4a]

iteratively update

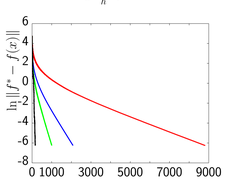
$$x_{k+1} = x_k - h_k g_\mu(x_k)$$

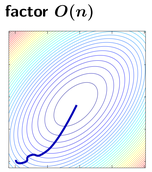
$u \sim \mathcal{N}(0, I_n)$

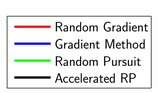
$h_k = \arg\min_h f(x_k + h\cdot u)$

line search in random direction

- Described by [4b], first analysis by Karmanov [4c], for approximative line search independently by [4a]
- convergence for smooth convex functions with bounded level sets and Lipschitz continuous gradients
- **reaches complexity bound of gradient method by factor $O(n)$**
- first approaches to variable metric by Leventhal [4d]







[4a] B. Gärtner, Ch. L. Müller, and S. U. Stich. Optimization of convex functions with random pursuit. in preparation.

[4b] F. J. Solis and R. J-B. Wets. Minimization by random search techniques. *Mathematical Operations Research*, 6:19–30, 1981.

[4c] V. G. Karmanov. On convergence of a random search method *(in Russian). Theory of Probability and its Applications*, 19(4):788–794, 1974.

[4d] D. Leventhal and A.S. Lewis. Randomized hessian estimation and directional search. *Optimization*, 60(3):329–345, 2011.

CMA-ES [5a]

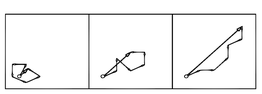
(μ,λ)-Evolution Strategy, iteratively update

m_k (mean) σ_k(stepsize) C_k (covariance)

propose new candidates

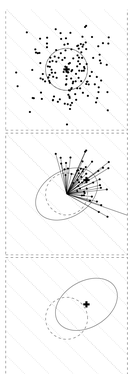
$$y_k^{(\cdot)} \sim \mathcal{N}(m_k, \sigma_k \cdot C_k)$$

- step-size adapt. by path lenght control
- covariance adapt. by rank μ updates



decrease σ increase σ

- variable metric (invariant under quadratic transformations)
- **Natural-gradient descent in parameter space [5b]**



[5a] N. Hansen and A. Ostermeier. Completely derandomized self-adaptation in evolution strategies. *Evolutionary Computation*, 9(2):159–195, 2001.

[5b] L. Arnold, A. Auger, N. Hansen, and Y. Ollivier. Information-Geometric Optimization Algorithms: A Unifying Picture via Invariance Principles.

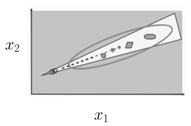
Gaussian Adaptation [6a, 6b]

$(1+1)$-Evolution Strategy iteratively update

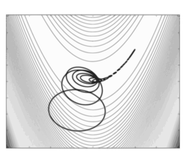
m_k (mean) σ_k(stepsize) C_k (covariance)

propose new candidates

$$y_k^{(\cdot)} \sim \mathcal{N}(m_k, \sigma_k \cdot C_k)$$



- covariance adaptation by rank 1 updates
- variable metric (invariant under quadratic transformations)
- described first by Kjellström, turned into effective optimizer by Müller [6c]
- very similar to $(1+1)$-CMA-ES without evolution path
- **approximation of level sets by ellipsoids (entropy maximization of search distribution with the constraint of fixed hitting probability)**



[6a] G. Kjellström. Network Optimization by Random Variation of Component Values. *Ericsson Technics*, 25(3):133–151, 1969.

[6b] G. Kjellströom and L. Taxen. Stochastic Optimization in System Design. *IEEE Trans. Circ. and Syst.*, 28(7), 1981.

[6c] Ch. L. Müller and I. F. Sbalzarini. Gaussian adaptation revisited - an entropic view on covariance matrix adaptation. In *EvoApplications (1)*:432–441, 2010.

Acknowledgments

My research is supported by the CG Learning project, which itself is supported by the FET (Future and Emerging Technologies) unit of the European Commission (EC) within the 7th Framework Programme of the EC under contract No. 255827.

Pictures from respectively cited papers unless indicated otherwise.