Randomized Derivative-Free Optimization

Standard Approaches

• Pure Random Search [1a]
  • Sample points uniformly at random
  • Small success probability

• Axial Random Search [%2d]
  • Sample sequence of points, all new points uniformly among those better than function value — not possible to do
  • PMK analogous to randomized version of method of centers

• DIRECT [1c]
  • Randomly evaluate function at grid
  • In each iteration locally refine grid until
  • Potential to improve function value
  • Exponential running time

• Nelder-Mead [1d]
  • Evaluate function at vertices of a simplex
  • John Wilker search: reflection of worst point through center
  • Converge to stationary points
  • Adaptation: simplex derivate [1e]

• Trust Region [1f]
  • Approximate function with models (when possible) at region
  • Refine/extend region on which the approximation is high grad

Evolution Strategies Explained

• (1+1)-ES described already in 1988 [3a]
• There is most complete method like CMA-ES
• Convergence of (1+1)-ES linear in quadratic functions [2b]
• Theoretical convergence results only known for very simple functions

• Genetic (1+1) Evolutionary Strategy (ES)
  1. for i = 0 to MAX
  2. u_k = x + r_k 
  3. evaluate fitness (f(x))
  4. i = i + 1
  5. end

For (1+1)-ES (with stepsize $m_k$):

• Small ES probability — decrease $m_k$ (1)
• small ES probability — increase $m_k$ (2)

For (1+1)-ES (with $m_k$)

• access only zeroth-order information
• Objective function given as black-box
• Multifunctional modes (global optimization)
• Robust against (random) noise
• Provably (fast) convergence

Problem Statement

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...)

Summary

• Convergence/divergence results for noisy and multi-modal functions
• Especially interesting would be to analyze particular heuristic used by many algorithms to self-optimize strategy parameters
• Step-size adaptation heuristics
• Hessian estimation (e.g. by rank-$r$ updates)
• Limited memory Hessian estimation possible?

The Right Metric

• For 1 quadratic
  $f(x) = f(y) + \frac{1}{2} \lambda (x-y)^2$
  - ES direction $\nabla f(x)$
  - ES gradient $\partial f(x)$
• If $\lambda = 1$, step size information is sufficient
• Otherwise estimation of $\lambda$ is necessary
• Invariant under quadratic transformations
• Estimation of $\lambda$ only from zeroth-order information is difficult
• Many update schemes are reasonable and used

Open Problems and Goals

• Convergence/divergence results to compare with classical methods
  • Convex/concave results for noisy and multi-modal functions
• Especially interesting would be to analyze particular heuristic used by many algorithms to self-optimizing strategy parameters
• Convergence/divergence results for noisy and multi-modal functions
• Step-size adaptation heuristics
• Hessian estimation (e.g. by rank-$r$ updates)
• Limited memory Hessian estimation possible?

CMA-ES [5a]

• Gaussian Adaptation [6a, 6b]

Randomized Pursuit [1b]

• Randomized (1+1) ES iterative update

Random Pursuit [3a]

• Iteratively update $x_{k+1} = x_k - h(x_k)$
  • y direction: direct assign $h(x_k)$
  • x direction: difference in random direction

• For first results by Poljak [3b], completely analyzed by Nesterov [3a], using Gaussian smoothing [3a]
  • Random walks for derivatives
•等待收敛 for convergence on convex functions
• Wait for Lipchitz continuous gradients
• Accelerated methods reach (provable) existing lower bounds [3d] by factor $O(\mu)$

• CMA-ES [5a]

Gaussian Adaptation [6a, 6b]

• For 1 quadratic
  $f(x) = f(y) + \frac{1}{2} \lambda (x-y)^2$
  - ES direction $\nabla f(x)$
  - ES gradient $\partial f(x)$
• If $\lambda = 1$, step size information is sufficient
• Otherwise estimation of $\lambda$ is necessary
• Invariant under quadratic transformations
• Estimation of $\lambda$ only from zeroth-order information is difficult
• Many update schemes are reasonable and used

Hit and Run [7a]

• Take a random direction and as the segment is contained in the convex body uniformly randomized with low variance
• Efficiently organized as sampling method to generate uniform samples from convex bodies

Random Conic Pursuit [7b]

• Converges on convex bodies by randomly proposed new candidates
  • Randomly proposed new candidates in weighted neighborhood (probabilistic moves based on acceptance criteria)
  • Convex/concave results to compare with classical methods

Simulated Annealing [7b, 7c]

• Inspiration and name from annealing in thermodynamics
• New iterates proposed by sampling random weight in a weighted neighborhood (probabilistic moves based on acceptance criteria)
• Convex/concave results to compare with classical methods
• Adaptive Metropolis-Hastings algorithm

Acknowledgments

My research is supported by the EC 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 sole indicated otherwise.