ISSUES IN STOCHASTIC SEARCH AND OPTIMIZATION

PerMIS 2004
NIST

James C. Spall
The Johns Hopkins University
Applied Physics Laboratory (JHU/APL)

james.spall@jhuapl.edu
### Issues in Stochastic Search and Optimization

**Johns Hopkins University, Applied Physics Laboratory (JHU/APL), 11100 Johns Hopkins Road, Laurel, MD, 20723**

**Proceedings of the 2004 Performance Metrics for Intelligent Systems Workshop (PerMIS ’04), Gaithersburg, MD on August 24-26 2004**
Performance Metrics and Optimization

• How are performance metrics used?
  – Sensitivity studies
  – System design
  – Decision aid for strategic planning
  – Adapting system over time
  – Detecting instability; avoiding unstable performance
  – Evaluating system reliability
  – Design of experiments
  – Mathematical modeling and parameter estimation
  – And on and on….

• Most of above involve optimization

• **Claim:** Impossible to have a performance metrics conference w/o **seriously** considering optimization!
Search and Optimization Algorithms as Part of Problem Solving

- There exist many deterministic and stochastic algorithms
- Algorithms are \textit{part} of the broader solution
- Need clear understanding of problem structure, constraints, data characteristics, political and social context, limits of algorithms, etc.
- “Imagine how much money could be saved if truly appropriate techniques were applied that go beyond simple linear programming.” (Z. Michalewicz and D. Fogel, 2000)
  - Deeper understanding required to provide truly appropriate solutions; COTS usually not enough!
- Many (most?) real-world implementations involve stochastic effects
Potpourri of Problems Using Stochastic Search and Optimization

• Minimize the costs of shipping from production facilities to warehouses
• Maximize the probability of detecting an incoming warhead (vs. decoy) in a missile defense system
• Place sensors in manner to maximize useful information
• Determine the times to administer a sequence of drugs for maximum therapeutic effect
• Find the best red-yellow-green signal timings in an urban traffic network
• Determine the best schedule for use of laboratory facilities to serve an organization’s overall interests
Two Fundamental Problems of Interest

• Let $\Theta$ be the domain of allowable values for a vector $\theta$
• $\theta$ represents a vector of “adjustables”
  – $\theta$ may be continuous or discrete (or both)
• Two fundamental problems of interest:

  **Problem 1.** Find the value(s) of a vector $\theta \in \Theta$ that minimize a scalar-valued loss function $L(\theta)$

  — or —

  **Problem 2.** Find the value(s) of $\theta \in \Theta$ that solve the equation $g(\theta) = 0$ for some vector-valued function $g(\theta)$

• Frequently (but not necessarily) $g(\theta) = \partial L(\theta)/\partial \theta$
Three Common Types of Loss Functions

Continuous

Discrete/
Continuous

Discrete
Stochastic Search and Optimization

• Focus here is on **stochastic** search and optimization:

  A. Random noise in input information (e.g., noisy measurements of \( L(\theta) \))

  — and/or —

  B. Injected randomness (Monte Carlo) in choice of algorithm iteration magnitude/direction

• Contrasts with deterministic methods
  – E.g., steepest descent, Newton-Raphson, etc.
  – Assume perfect information about \( L(\theta) \) (and its gradients)
  – Search magnitude/direction deterministic at each iteration

• Injected randomness (B) in search magnitude/direction can offer benefits in efficiency and robustness
  – E.g., Capabilities for global (vs. local) optimization
Some Popular Stochastic Search and Optimization Techniques

- Random search
- Stochastic approximation
  - Robbins-Monro and Kiefer-Wolfowitz
  - SPSA
  - NN backpropagation
  - Infinitesimal perturbation analysis
  - Recursive least squares
  - Many others
- Simulated annealing
- Genetic algorithms
- Evolutionary programs and strategies
- Reinforcement learning
- Markov chain Monte Carlo (MCMC)
- Etc.
Effects of Noise on Simple Optimization Problem
Example Search Path (2 variables): Steepest Descent with Noisy and Noise-Free Input
Example of Noisy Loss Measurements: Tracking Problem

• Consider tracking problem where controller and/or system depend on design parameters $\theta$
  – E.g.: Missile guidance, robot arm manipulation, attaining macroeconomic target values, etc.

• Aim is to pick $\theta$ to minimize mean-squared error (MSE):
  \[ L(\theta) = E\left(\|\text{actual output} - \text{desired output}\|^2\right) \]

• In general nonlinear and/or non-Gaussian systems, \textit{not possible} to compute $L(\theta)$

• Get \textit{observed} squared error $y(\theta) \equiv \| \cdot \|^2$ by running system

• Note that $y(\theta) = \| \cdot \|^2 = L(\theta) + \text{noise}$
  – Values of $y(\theta)$, not $L(\theta)$, used in optimization of $\theta$
Example of Noisy Loss Measurements: Simulation-Based Optimization

- Have credible Monte Carlo simulation of real system
- Parameters $\theta$ in simulation have physical meaning in system
  - E.g.: $\theta$ is machine locations in plant layout, timing settings in traffic control, resource allocation in military operations, etc.
- Run simulation to determine best $\theta$ for use in real system
- Want to minimize average measure of performance $L(\theta)$
  - Let $y(\theta)$ represent one simulation output ($y(\theta) = L(\theta) + \text{noise}$)
Some Key Properties in Implementation and Evaluation of Stochastic Algorithms

• Algorithm comparisons via number of evaluations of $L(\theta)$ or $g(\theta)$ (not iterations)
  – Function evaluations typically represent major cost

• Curse of dimensionality
  – E.g.: If $\text{dim}(\theta) = 10$, each element of $\theta$ can take on 10 values.
    Take 10,000 random samples: $\text{Prob}(\text{finding one of 500 best } \theta) = 0.0005$
  – Above example would be even much harder with only noisy function measurements

• Constraints

• Limits of numerical comparisons
  – Avoid broad claims based on numerical studies
  – Best to combine theory and numerical analysis
Global vs. Local Solutions

• Global methods *tend* to have following characteristics:
  – Inefficient, especially for high-dimensional $\theta$
  – Relatively difficult to use (e.g., require very careful selection of algorithm coefficients)
  – Shaky theoretical foundation for global convergence

• Much “hype” with many methods (genetic algorithm [GA] software advertisements):
  – “…can handle the most complex problems, including problems unsolvable by any other method.”
  – “…uses GAs to solve *any* optimization problem!”

• But there are *some* mathematically sound methods
  – E.g., *restricted settings* for GAs, simulated annealing, and SPSA
No Free Lunch Theorems

• Wolpert and Macready (1997) establish several “No Free Lunch” (NFL) Theorems for optimization

• NFL Theorems apply to settings where parameter set and set of loss function values are finite, discrete sets
  – Relevant for continuous $\theta$ problem when considering digital computer implementation
  – Results are valid for deterministic and stochastic settings

• Number of optimization problems—mappings from to set of loss values—is finite

• NFL Theorems state, in essence, that no one search algorithm is “best” for all problems
No Free Lunch Theorems—Basic Formulation

• Suppose that

\[ N_{\theta} = \text{number of values of } \theta \]
\[ N_L = \text{number of values of loss function} \]

• Then

\[ \left( N_L \right)^{N_{\theta}} = \text{number of loss functions} \]

• There is a finite (but possibly huge) number of loss functions

• Basic form of NFL considers average performance over all loss functions
Illustration of No Free Lunch Theorems (Example 1.7 in ISSO)

- Three values of $\theta$, two outcomes for noise free loss $L$
  - Eight possible mappings, hence eight optimization problems
- Mean loss across all problems is same regardless of $\theta$; entries 1 or 2 in table below represent two possible $L$ outcomes

<table>
<thead>
<tr>
<th>Map</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta_1$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>$\theta_2$</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>$\theta_3$</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>
No Free Lunch Theorems (cont’d)

• NFL Theorems state, in essence:

Averaging (uniformly) over all possible problems (loss functions L), all algorithms perform equally well

• In particular, if algorithm 1 performs better than algorithm 2 over some set of problems, then algorithm 2 performs better than algorithm 1 on another set of problems

Overall relative efficiency of two algorithms cannot be inferred from a few sample problems

• NFL theorems say nothing about specific algorithms on specific problems
Relative Convergence Rates of Deterministic and Stochastic Optimization

- Theoretical analysis based on convergence rates of iterates $\hat{\theta}_k$, where $k$ is iteration counter.
- Let $\theta^*$ represent optimal value of $\theta$.
- For deterministic optimization, a standard rate result is:
  \[ \|\hat{\theta}_k - \theta^*\| = O(c^k), \; 0 < c < 1 \]

- Corresponding rate with noisy measurements:
  \[ \|\hat{\theta}_k - \theta^*\| = O\left(\frac{1}{k^\lambda}\right), \; 0 < \lambda \leq \frac{1}{2} \]

- Stochastic rate inherently slower in theory and practice.
Concluding Remarks

- Stochastic search and optimization very widely used
  - Handles noise in function evaluations
  - Generally better for global optimization
  - Broader applicability to “non-nice” problems (robustness)
- Some challenges in practical problems
  - Noise dramatically affects convergence
  - Distinguishing global from local minima not generally easy
  - Curse of dimensionality
  - Choosing algorithm “tuning coefficients”
- Rarely sufficient to use theory for standard deterministic methods to characterize stochastic methods
- “No free lunch” theorems are barrier to exaggerated claims of power and efficiency of any specific algorithm
- Algorithms should be implemented in context: “Better a rough answer to the right question than an exact answer to the wrong one” (Lord Kelvin)
Selected References on Stochastic Optimization

Contact Info. and Related Web Sites

• *james.spall@jhuapl.edu*

• **www.jhuapl.edu/SPSA** (Web site on stochastic approximation algorithm)

• **www.jhuapl.edu/ISSO** (Web site on book *Introduction to Stochastic Search and Optimization*)