

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

Report Documentation Page				Form Approved OMB No. 0704-0188	
Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.					
1. REPORT DATE AUG 2004		2. REPORT TYPE		3. DATES COVERED 00-00-2004 to 00-00-2004	
4. TITLE AND SUBTITLE Issues in Stochastic Search and Optimization				5a. CONTRACT NUMBER	
				5b. GRANT NUMBER	
				5c. PROGRAM ELEMENT NUMBER	
6. AUTHOR(S)				5d. PROJECT NUMBER	
				5e. TASK NUMBER	
				5f. WORK UNIT NUMBER	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Johns Hopkins University, Applied Physics Laboratory (JHU/APL), 11100 Johns Hopkins Road, Laurel, MD, 20723				8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)				10. SPONSOR/MONITOR'S ACRONYM(S)	
				11. SPONSOR/MONITOR'S REPORT NUMBER(S)	
12. DISTRIBUTION/AVAILABILITY STATEMENT Approved for public release; distribution unlimited					
13. SUPPLEMENTARY NOTES Proceedings of the 2004 Performance Metrics for Intelligent Systems Workshop (PerMIS '04), Gaithersburg, MD on August 24-26 2004					
14. ABSTRACT					
15. SUBJECT TERMS					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT Same as Report (SAR)	18. NUMBER OF PAGES 22	19a. NAME OF RESPONSIBLE PERSON
a. REPORT unclassified	b. ABSTRACT unclassified	c. THIS PAGE unclassified			

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 **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 Θ be the domain of allowable values for a vector θ
- θ represents a vector of “adjustables”
 - θ may be continuous or discrete (or both)
- Two fundamental problems of interest:

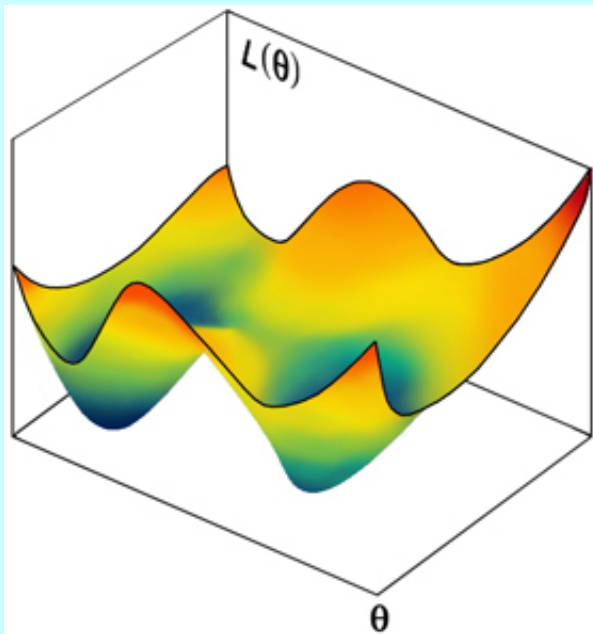
Problem 1. Find the value(s) of a vector $\theta \in \Theta$ that minimize a scalar-valued *loss function* $L(\theta)$

— or —

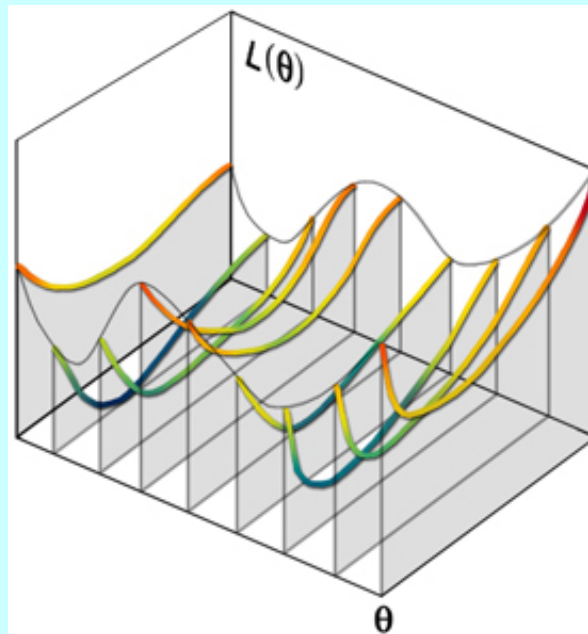
Problem 2. Find the value(s) of $\theta \in \Theta$ that solve the equation $\mathbf{g}(\theta) = \mathbf{0}$ for some vector-valued function $\mathbf{g}(\theta)$

- Frequently (but not necessarily) $\mathbf{g}(\theta) = \partial L(\theta) / \partial \theta$

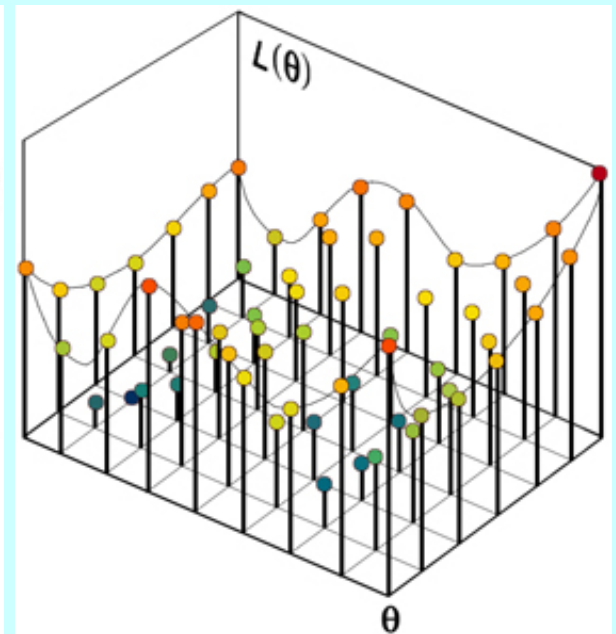
Three Common Types of Loss Functions



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 —

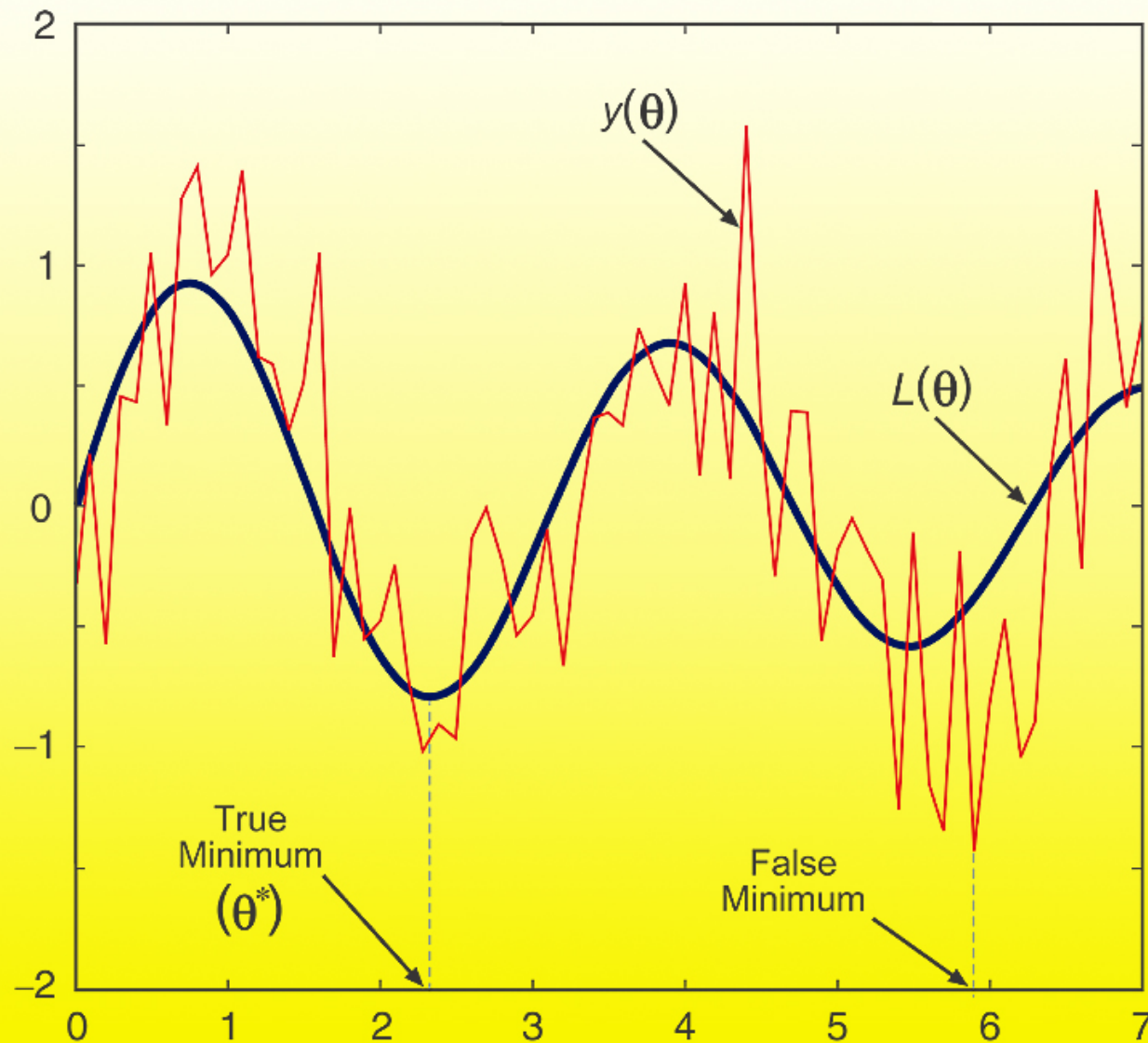
B. Injected randomness (Monte Carlo) in choice of algorithm iteration magnitude/direction

- 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

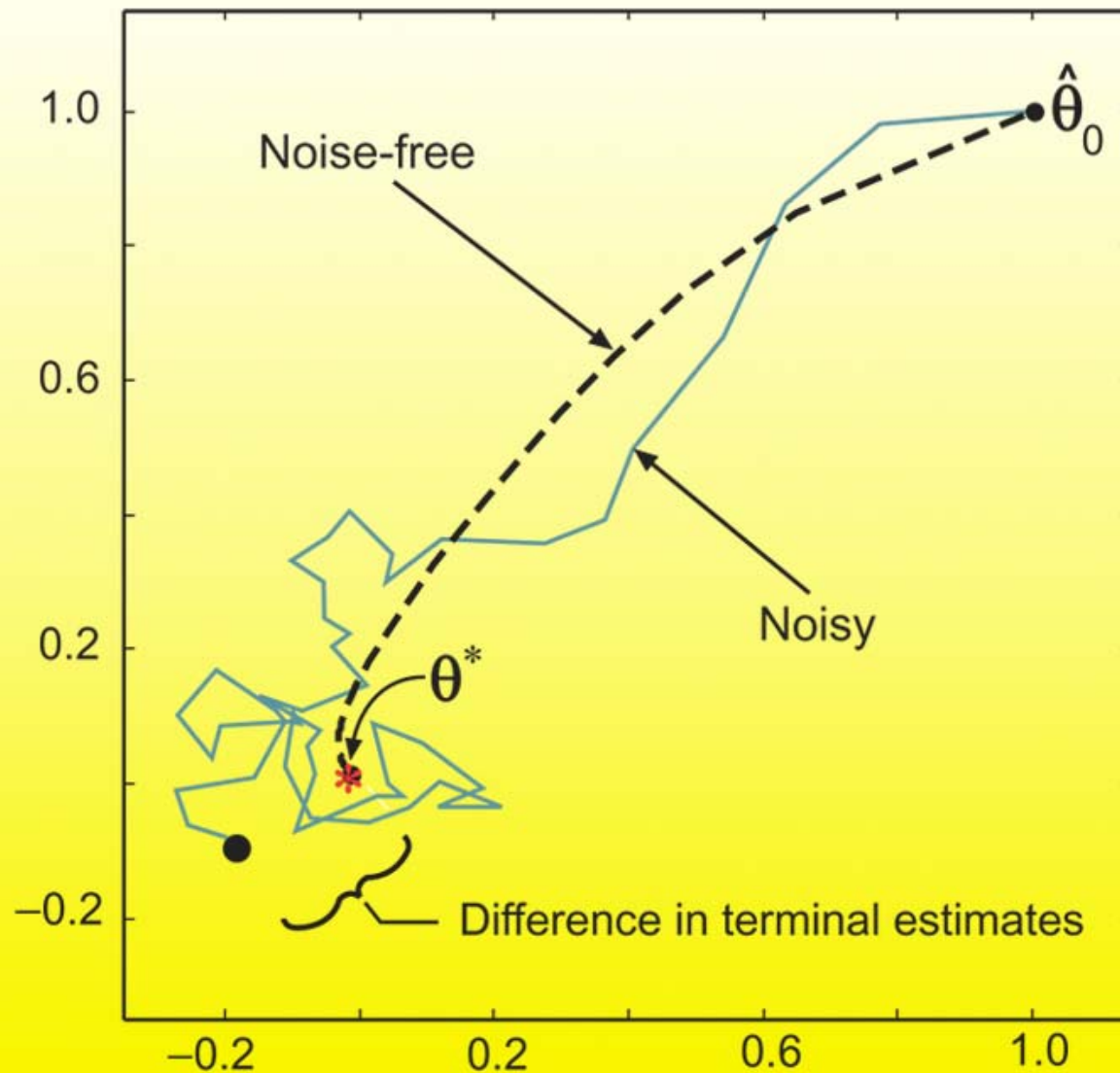
Some Popular Stochastic Search and Optimization Techniques

- 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

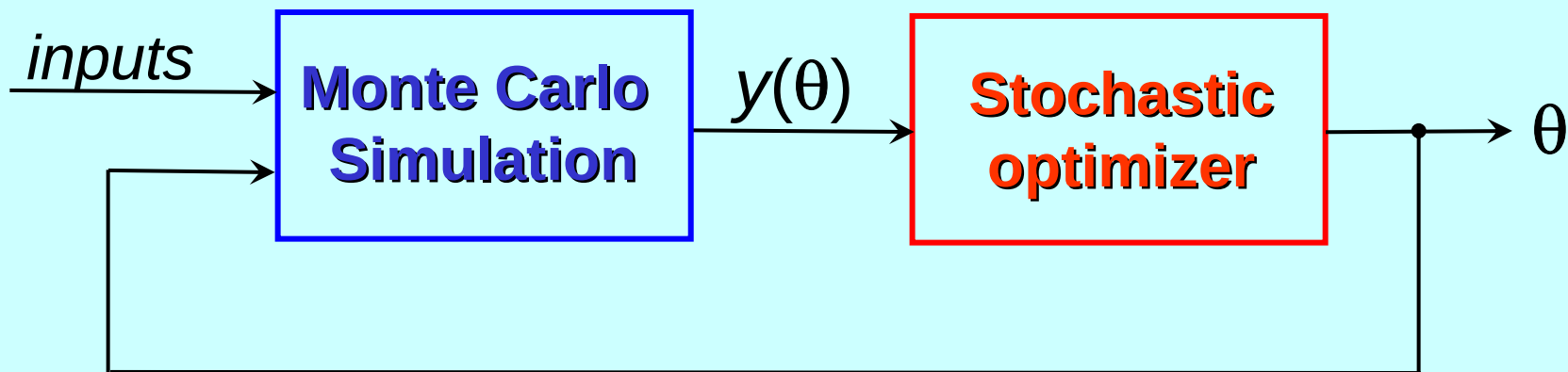


Example of Noisy Loss Measurements: Tracking Problem

- Consider tracking problem where controller and/or system depend on design parameters θ
 - E.g.: Missile guidance, robot arm manipulation, attaining macroeconomic target values, etc.
- Aim is to pick θ to minimize mean-squared error (MSE):
$$L(\theta) = E \left(\left\| \text{actual output} - \text{desired output} \right\|^2 \right)$$
- In general nonlinear and/or non-Gaussian systems, ***not possible*** to compute $L(\theta)$
- Get **observed** squared error $y(\theta) \equiv \left\| \cdot \right\|^2$ by running system
- Note that $y(\theta) = \left\| \cdot \right\|^2 = L(\theta) + \text{noise}$
 - Values of $y(\theta)$, not $L(\theta)$, used in optimization of θ

Example of Noisy Loss Measurements: Simulation-Based Optimization

- Have credible Monte Carlo simulation of real system
- Parameters θ in simulation have physical meaning in system
 - E.g.: θ is machine locations in plant layout, timing settings in traffic control, resource allocation in military operations, etc.
- Run simulation to determine best θ 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 $\dim(\theta) = 10$, each element of θ 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 θ
 - 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 Θ and set of loss function values are finite, discrete sets
 - Relevant for continuous θ problem when considering digital computer implementation
 - Results are valid for deterministic and stochastic settings
- Number of optimization problems—mappings from Θ 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_θ = number of values of θ

N_L = number of values of loss function

- Then

$(N_L)^{N_\theta}$ = 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 θ , two outcomes for noise free loss L
 - Eight possible mappings, hence eight optimization problems
- Mean loss across all problems is same regardless of θ ; entries 1 or 2 in table below represent two possible L outcomes

Map θ	1	2	3	4	5	6	7	8
θ_1	1	1	1	2	2	2	1	2
θ_2	1	1	2	1	1	2	2	2
θ_3	1	2	2	1	2	1	1	2

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 θ^* represent optimal value of θ
- For **deterministic** optimization, a standard rate result is:

$$\|\hat{\theta}_k - \theta^*\| = O(c^k), \quad 0 < c < 1$$

- Corresponding rate with **noisy measurements**

$$\|\hat{\theta}_k - \theta^*\| = O\left(\frac{1}{k^\lambda}\right), \quad 0 < \lambda \leq \frac{1}{2}$$

- 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

- Fogel, D. B. (2000), *Evolutionary Computation: Toward a New Philosophy of Machine Intelligence* (2nd ed.), IEEE Press, Piscataway, NJ.
- Fu, M. C. (2002), “Optimization for Simulation: Theory vs. Practice” (with discussion by S. Andradóttir, P. Glynn, and J. P. Kelly), *INFORMS Journal on Computing*, vol. 14, pp. 192–227.
- Goldberg, D. E. (1989), *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley, Reading, MA.
- Gosavi, A. (2003), *Simulation-Based Optimization: Parametric Optimization Techniques and Reinforcement Learning*, Kluwer, Boston.
- Holland, J. H. (1975), *Adaptation in Natural and Artificial Systems*, University of Michigan Press, Ann Arbor, MI.
- Kushner, H. J. and Yin, G. G. (2003), *Stochastic Approximation and Recursive Algorithms and Applications* (2nd ed.), Springer-Verlag, New York.
- Michalewicz, Z. and Fogel, D. B. (2000), *How to Solve It: Modern Heuristics*, Springer-Verlag, New York.
- Spall, J. C. (2003), *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control*, Wiley, Hoboken, NJ.
- Zhigljavsky, A. A. (1991), *Theory of Global Random Search*, Kluwer Academic, Boston.

Contact Info. and Related Web Sites

- [*james.spall@jhuapl.edu*](mailto:james.spall@jhuapl.edu)
- [*www.jhuapl.edu/SPSA*](http://www.jhuapl.edu/SPSA) (Web site on stochastic approximation algorithm)
- [*www.jhuapl.edu/ISSO*](http://www.jhuapl.edu/ISSO) (Web site on book *Introduction to Stochastic Search and Optimization*)