Model Reference Adaptive Search: A New Approach to Global Optimization

Steve Marcus

(joint with Jiaqiao Hu and Michael Fu)

October 11, 2005

A. JAMES CLARK SCHOOL OF ENGINEERING

The Institute for Systems Research

Outline

- Problem Setting
- Related Work and Motivation
- Model Reference Adaptive Search (MRAS)
- Properties of MRAS
  - Rigorous Global Convergence
  - Relationship to Cross-Entropy Method
- Numerical Examples
- Conclusions/Work in Progress/Future Work

Problem Setting

- Solution space $\chi \subseteq \Re^n$
  - continuous or discrete (combinatorial)
- Objective function $H(\cdot)$: $\chi \rightarrow \Re$
- Objective: find optimal $x^* \in \chi$ such that

$$x^* \in \arg\max_{x \in \chi} H(x)$$

  - Assumptions: existence, uniqueness (but possibly many local minima)

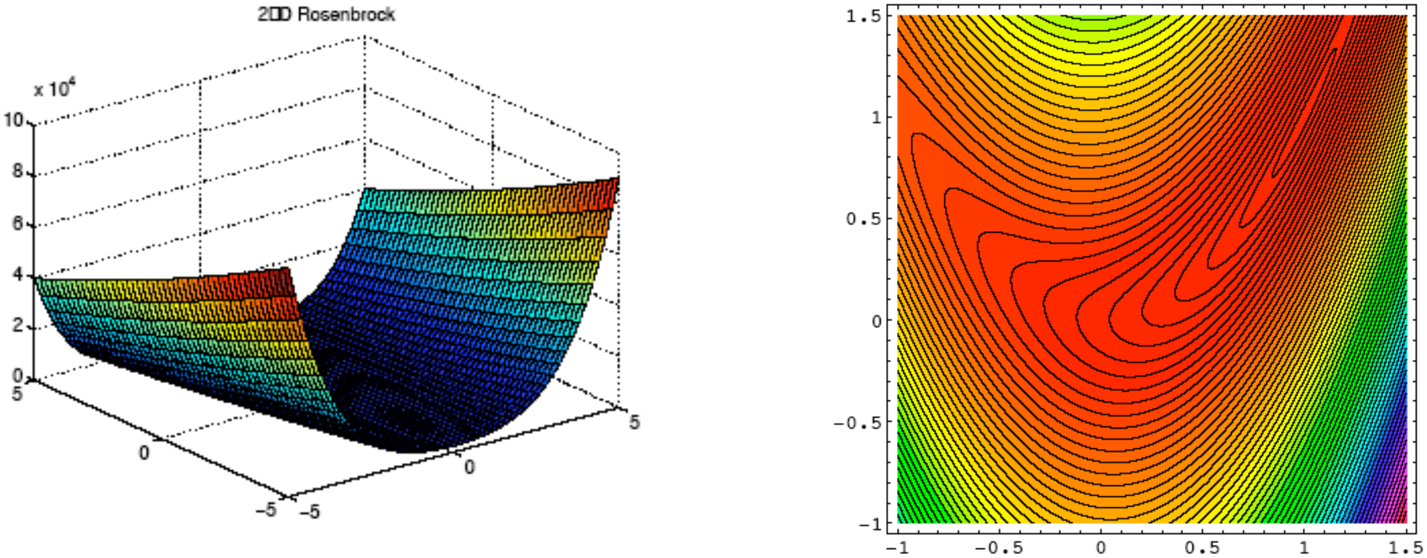
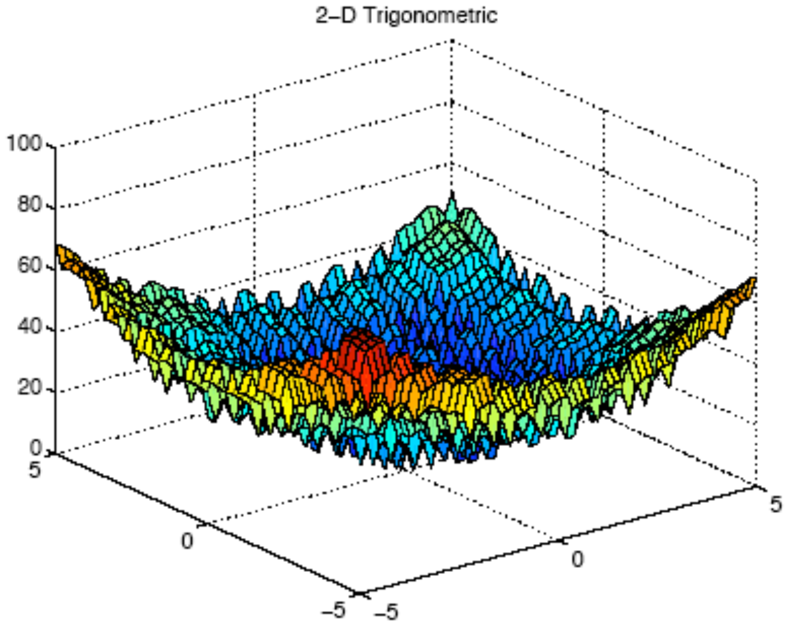


Figure 1: 2-D Rosenbrock function, where $-5 \le x_i \le 5,\ i = 1, 2$.



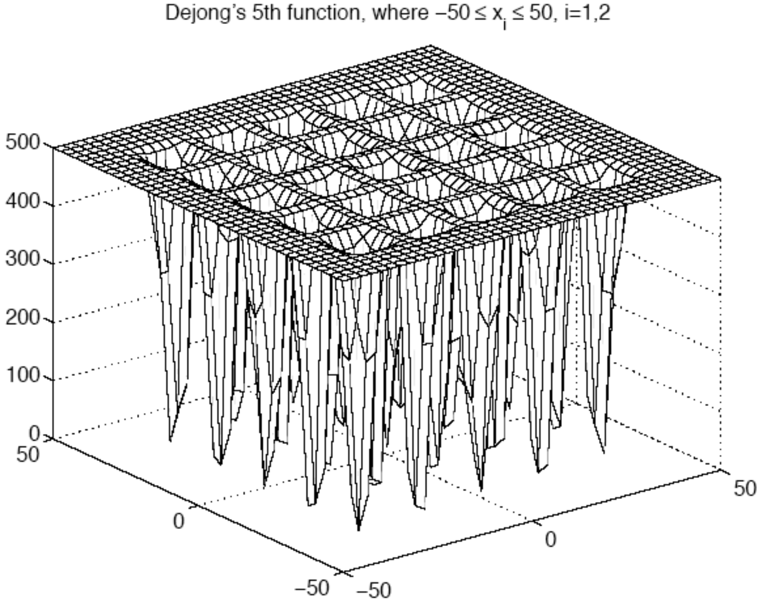


Figure 2: 2-D Trigonometric function, where $-5 \le x_i \le 5,\ i = 1, 2$.

Overview of Global Optimization Approaches

- **Instance-based approaches:** search for new solutions depends *directly* on previously generated solutions
  - **simulated annealing**
  - **genetic algorithms**
  - **tabu search**
  - **nested partitions**

Overview of Global Optimization Approaches (cont.)

- **Model-based search methods:** new solutions generated via an intermediate *probability distribution (model)* updated from previous generated solutions *(indirect* dependence).
  - **ant colony optimization**
  - **cross-entropy method (CE)**
  - **estimation of distribution algorithms (EDAs)**
  - At each iteration of the algorithm

    1. Generate population of candidate solutions (random samples) according to *probability distribution (model)* over the solution space

    2. Update parameters of model on basis of data in previous step in a way that will concentrate future search in regions containing high quality solutions

Brief Review of Genetic Algorithms (GAs)

- works with **population** of solutions
- update population (generate new generation):
  - **operators**, e.g., crossover, mutation,
    - often *probabilistic*
    - produces new candidates
  - **selection** (from old and new)

Estimation of Distribution Algorithms (EDAs)

- works with sequence of **probability distributions** over solution space (continuous pdf, discrete pmf)

Main Steps of typical procedure:

- initialization: starting distribution g_0
- until stopping rule satisfied, iterate the following:
  - generate population from current distribution
  - **evaluate** newly generated solutions
    and **select** some subset
    to **update** distribution

EDAs (continued)

similarities to GAs

- uses a population
- selection process
- randomized algorithm,
  but uses “model” (distribution) instead of operators

aka

- probabilistic model building genetic algorithms (PMBGAs)
- distribution estimation algorithms (DEAs)
- iterated density estimation algorithms (IDEAs)

Model-based Search Methods

- KEY QUESTION: how to update probability distributions?
  - Traditional EDAs use an explicit construction, can be difficult and computationally expensive, particularly for infinite solution spaces
  - Alternative: use parameterized family of distributions, and minimize distance to desired distributions (use projection)
    - Cross-Entropy (CE) method uses optimal importance sampling reference distribution
    - MRAS approach: general sequence of model reference distributions (don’t actually need to be computed)

MRAS

- Main characteristics

  - Given sequence of reference distributions $\{g_k(\cdot)\}$

  - works with a family of parameterized probability distributions $\{f(\cdot,\theta)\}$ over the solution space

  - fundamental steps at iteration k :

    * generate candidate solutions according to the current probability distribution $f(\cdot, \theta_k)$

    * calculate θ_{k+1} using data collected in previous step to bias future search toward promising regions, by minimizing distance between $f(\cdot,\theta)$ and $g_{k+1}(\cdot)$

  - Algorithm converges to optimal if $\{g_k(\cdot)\}$ does

MRAS: specific instantiation

- **Main idea:** Next distribution obtained by tilting previous

$$g_{k+1}(x) = \frac{H(x)g_k(x)}{E_{g_k}[H(X)]}, \ \forall x \in \chi.$$

Properties:

$$E_{g_{k+1}}[H(X)] \geq E_{g_k}[H(X)], \ \text{and}$$

$$\lim_{k \to \infty} E_{g_k}[H(X)] = H(x^*).$$

- Related to recursions found in EDAs, learning automata, “multiplicative weights”

- Other choices of $\{g_k(\cdot)\}$ result in other algorithms (e.g., cross-entropy)--this choice leads to global convergence

MRAS: specific instantiation

- **Obvious Difficulties**
  - requires enumerating all points in solution space
  - $g_k(x)$ may not be computationally tractable

- **Proposed Approach**
  - Monte Carlo (sampling) version
  - use parameterized distributions $\{f(\cdot,\theta)\}$
  - projection of $g_k(\cdot)$, which are *implicitly* generated

MRAS (deterministic version) Components

- positive continuous strictly increasing function $S(\cdot)$
- parameterized family of distributions $\{f(\cdot,\theta)\}$
- selection parameters ρ and non-decreasing $\{\gamma_k\}$, affecting distribution updates
  - ρ determines the proportion of solutions used
  - In iteration k, only solutions better than γ_k are in updating θ_{k+1}

MRAS Parameter Updating

- (1- ρ) quantiles w.r.t. $f(\cdot,\theta_k)$

$$\gamma_{k+1} = \sup_l \{ l : P_{\theta_k}(H(X) \geq l) \geq \rho \}$$

- update θ_{k+1} as

$$\theta_{k+1} = \arg\max_{\theta\in\Theta} \int_{x\in\chi} [S(H(x))]^k I\{H(x) > \gamma_{k+1}\} \ln f(x,\theta) dx$$

Lemma : θ_{k+1} minimizes the KL-distance between g_{k+1} and $f(\cdot,\theta)$, i.e.,

$$\theta_{k+1} = \arg\min_{\theta\in\Theta} D(g_{k+1} \mid f(\cdot,\theta)) := \arg\min_{\theta\in\Theta} E_{g_{k+1}}\left[\ln \frac{g_{k+1}(X)}{f(X,\theta)}\right] \text{, where}$$

$$g_{k+1}(x) = \frac{S(H(x)) I_{\{H(x)\geq\gamma_{k+1}\}} g_k(x)}{E_{g_k}[S(H(X)) I_{\{H(X)\geq\gamma_{k+1}\}}]}, \quad g_1(x) := \frac{I_{\{H(x)\geq\gamma_1\}}}{E_{\theta_0}[I_{\{H(X)\geq\gamma_1\}} \big/ f(X,\theta_0)]}$$

MRAS Basic Algorithm (deterministic version)

Initialization: Specify $\rho \in (0,1]$, $S(\cdot): \Re \rightarrow \Re^{+}$, $f(x, \theta_0) > 0 \ \forall x \in \chi$

- **Repeat** until a specified stopping rule is satisfied:
  - Calculate (1-ρ)-quantile

$$\gamma_{k+1} = \sup_{l} \{ l : P_{\theta_k}(H(X) \geq l) \geq \rho \}$$

  - Update parameter

$$\theta_{k+1} = \arg\max_{\theta \in \Theta} \int_{x \in \chi} [S(H(x))]^k I\{H(x) > \gamma_{k+1}\} \ln f(x, \theta) dx$$

Theory

- Restriction to natural exponential family (NEF)
  - covers broad class of distributions

  Examples: Gaussian, Poisson, binomial, geometric

- **Global convergence** can be established under some mild regularity conditions
  - multivariate Gaussian

$$\lim_{k\to\infty} \mu_k = x^*, \quad \lim_{k\to\infty} \Sigma_k = 0_{n\times n}$$

  - univariate independent components

$$\lim_{k\to\infty} E_{\theta_k}[X] = x^*.$$

Cross-Entropy (CE) Method

- pioneered by Rubinstein et al. (www.cemethod.org)
- originally for finding optimal
  parameterized importance sampling measure
- found that it could be applied to
  combinatorial optimization problems
- like EDAs and MRAS, updates distribution iteratively
- drawbacks: no proof of global convergence in general

MRAS Interpretation of CE Method

- (1- ρ) quantiles w.r.t. $f(\cdot,\theta_k)$

$$\gamma_{k+1} = \sup_l \{l : P_{\theta_k}(H(X) \geq l) \geq \rho\}$$

- update θ_{k+1} as

$$\theta_{k+1} = \arg\max_{\theta \in \Theta} E_{\theta_k}[\varphi(H(X)I_{\{H(X)\geq\gamma_{k+1}\}} \ln f(X,\theta)]$$

Lemma : θ_{k+1} minimizes the KL-distance between g_{k+1}^{ce} and $f(\cdot,\theta)$, where

$$g_{k+1}^{ce}(x) = \frac{\varphi(H(x))I_{\{H(x)\geq\gamma_{k+1}\}} f(x,\theta_k)}{E_{\theta_k}[\varphi(H(X))I_{\{H(X)\geq\gamma_{k+1}\}}]}$$

Relationship of MRAS with CE

- MRAS has general sequence of implicit reference models $\{g_k\}$, whereas CE uses the optimal importance sampling measure at each iteration
- MRAS provides general framework: CE can be interpreted by defining appropriate $\{g_k\}$, but the sequence depends on $\{f(\cdot,\theta_k)\}$ sequence
- Stronger theoretical convergence results for MRAS (global convergence for Monte Carlo version)
  - Uses the fact that $\{g_k\}$converge to optimal, which is not in general true for CE
- Computational comparison results reported later

MRAS$_1$ (Monte-Carlo version)

changes from deterministic version

- **finite number of samples, say N_k, at each iteration**
- **replace the true $(1-\rho)$-quantiles by sample quantiles**
- **replace the integrals (expected values) by sample averages**
- **ρ_k adaptively decreasing and N_k adaptively increasing**

Global convergence can be established

- multivariate normal case

$$\lim_{k\to\infty} \hat{\mu}_k = x^*, \text{ and } \lim_{k\to\infty} \hat{\Sigma}_k = 0_{n\times n} \quad \text{w.p.1.}$$

- independent univariate case

$$\lim_{k\to\infty} E_{\hat{\theta}_k}[X] = x^* \quad \text{w.p.1.}$$

Some Numerical Examples

- **Implementation issues**
  - parameter smoothing $\hat{\theta}_{k+1} \leftarrow \omega\,\hat{\theta}_{k+1} + (1-\omega)\,\hat{\theta}_k$
- **Numerical examples (minimization, not max)**
  - **Continuous optimization**
    - * Trigonometric function

$$1 + \sum_{i=1}^{20} 8\sin^2(7(x_i - 0.9)^2) + 6\sin^2(14(x_i - 0.9)^2) + (x_i - 0.9)^2$$

    - * Rosenbrock function

$$\sum_{i=1}^{19} 100(x_{i+1} - x_i^2)^2 + (x_i - 1)^2$$

    - * Powell singular function

$$\sum_{i=2}^{18} [(x_{i-1} + 10x_i)^2 + 5(x_{i+1} - x_{i+2})^2 + (x_i - 2x_{i+1})^4 + 10(x_{i-1} - x_{i+2})^4]$$

    - * Pinter, DeJong, Griewank functions
    - * Compared with CE (Kroese, Rubinstein, & Porotsky)

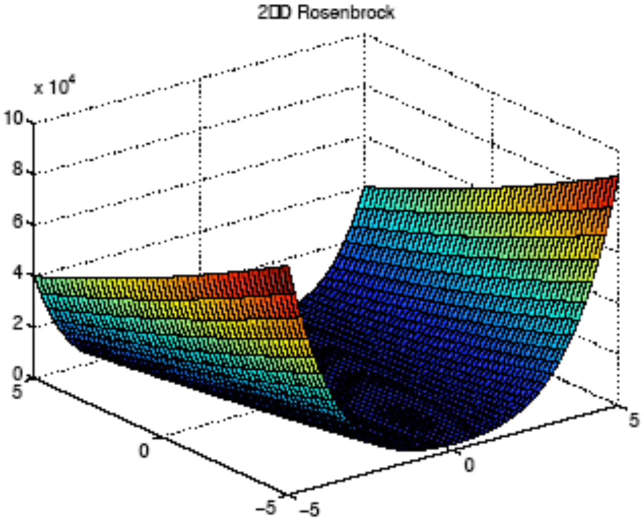


Figure 1: 2-D Rosenbrock function, where $-5 \leq x_i \leq 5,\ i = 1, 2$.

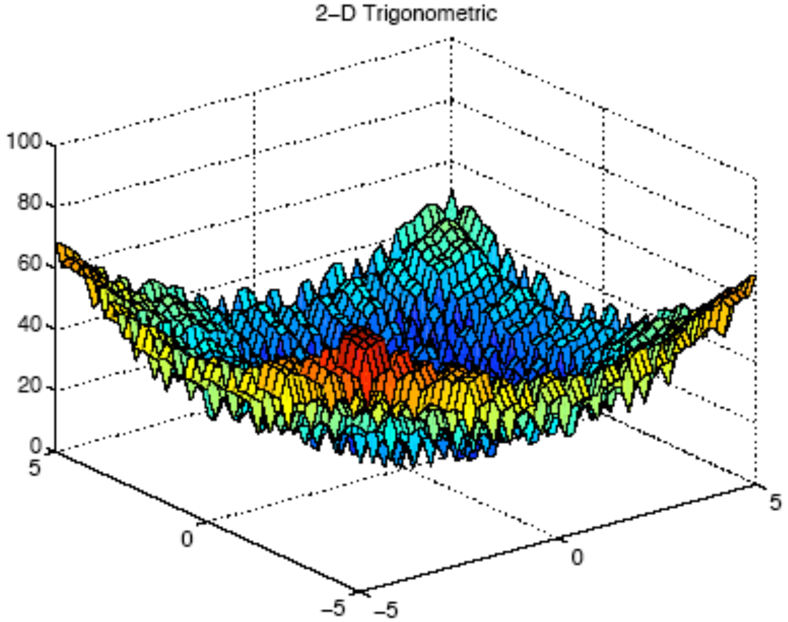


Figure 2: 2-D Trigonometric function, where $-5 \leq x_i \leq 5,\ i = 1, 2$.

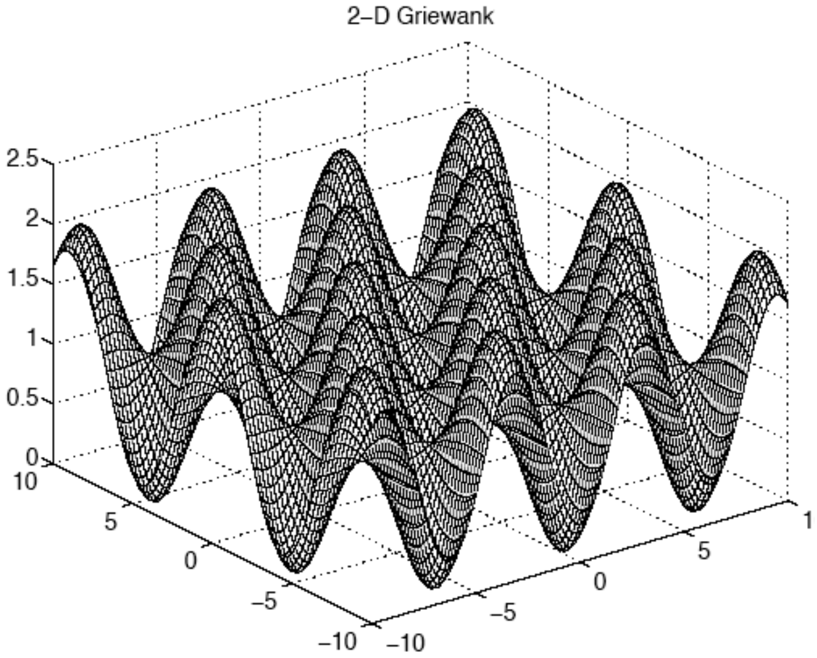


Figure 3: 2-D Griewank function, where $-10 \le x_i \le 10,\ i = 1,$

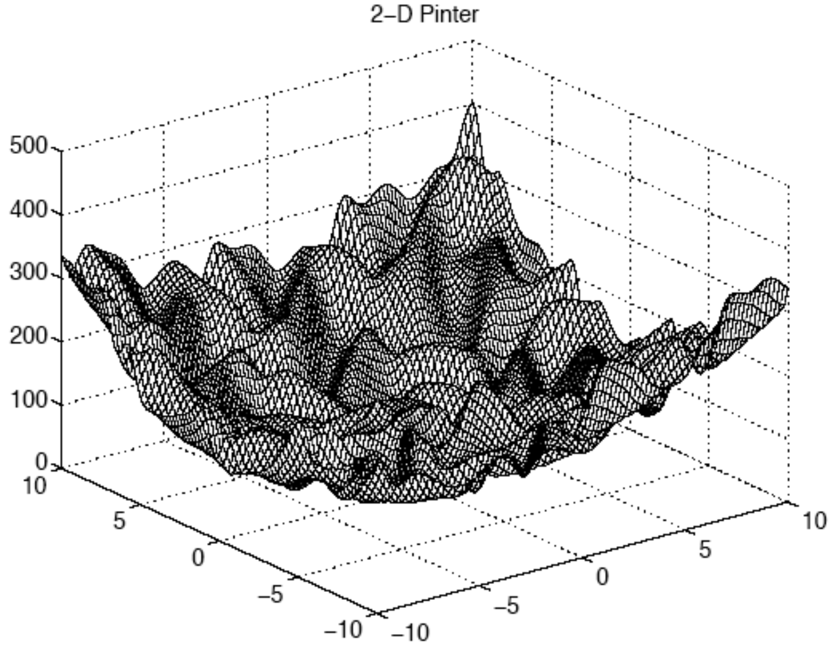
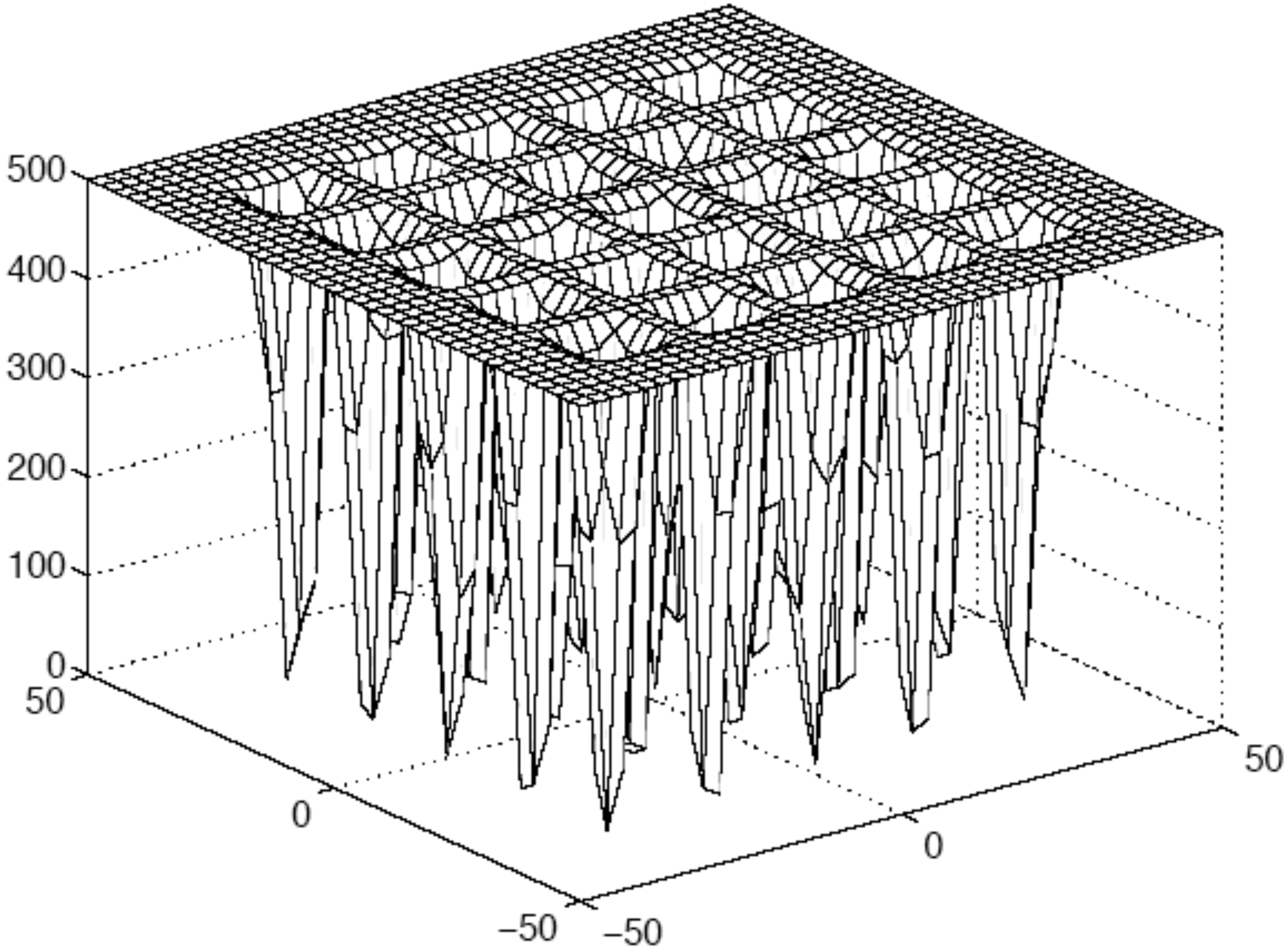
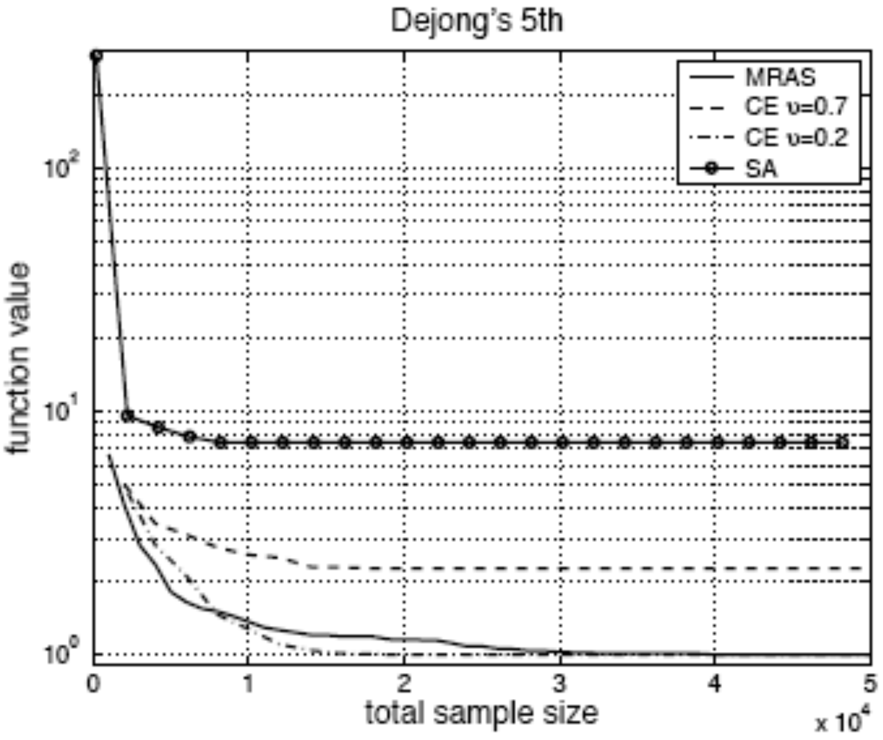
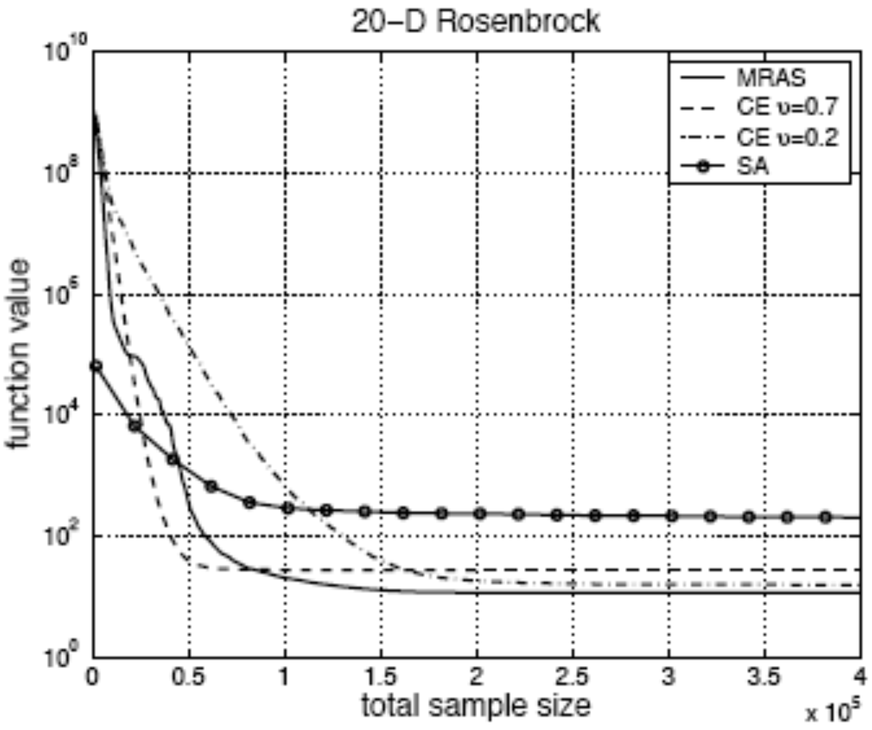


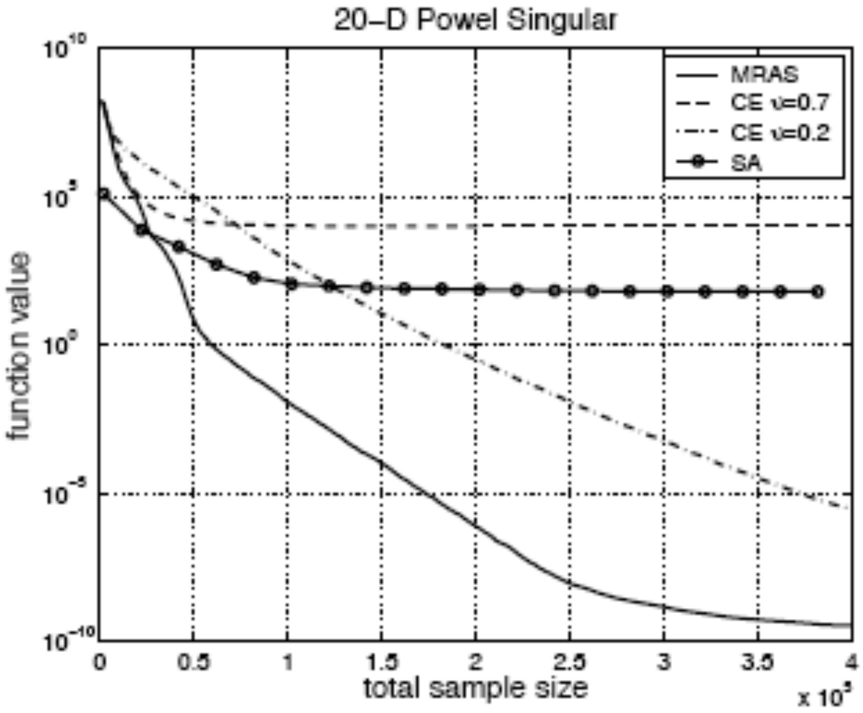
Figure 4: 2-D Pinter function, where $-10 \le x_i \le 10,\ i = 1, 2.$

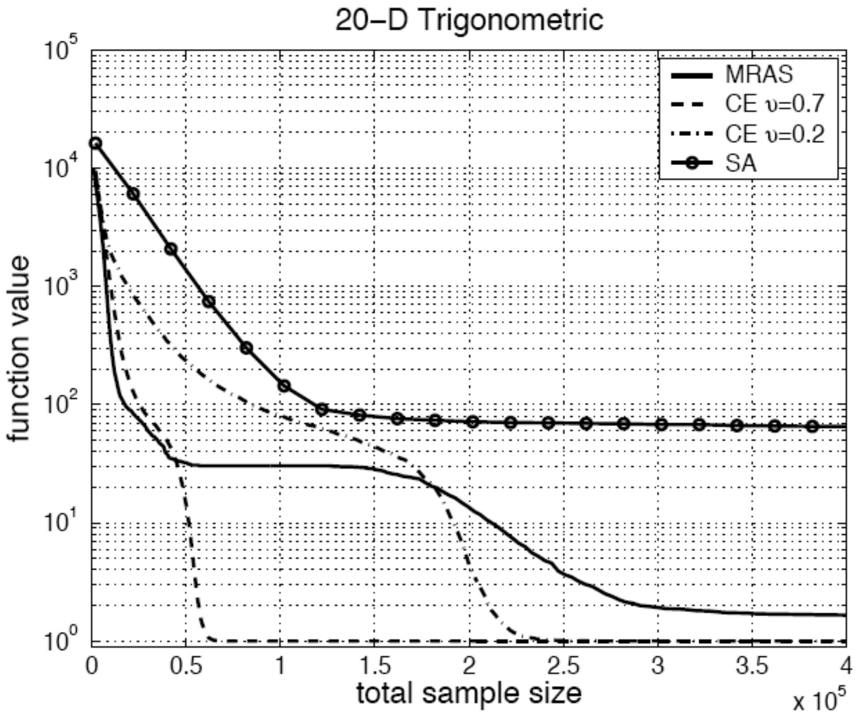


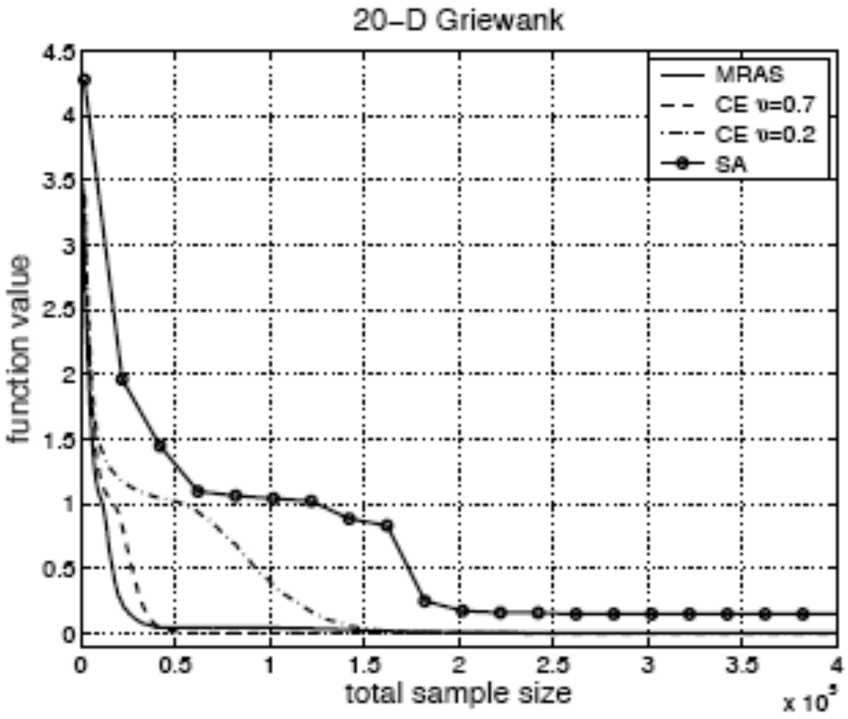
Dejong's 5th function, where $-50 \le x_i \le 50$, i=1,2

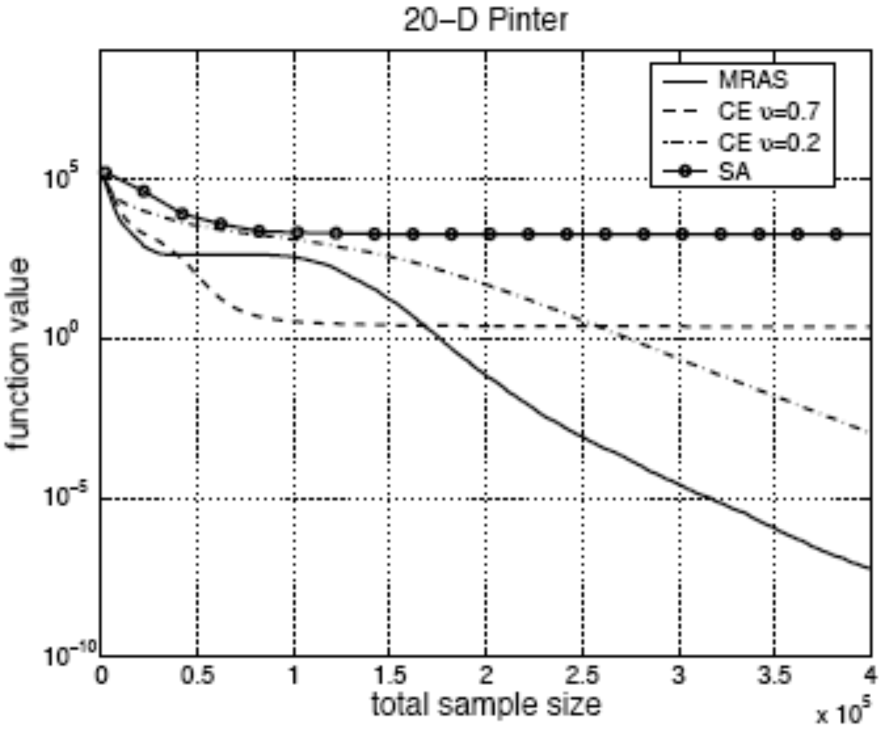

Dejong's 5th
function value
10^2
10^1
10^0
0
1
2
3
4
5
total sample size
x 10^4
MRAS
CE υ=0.7
CE υ=0.2
SA


20−D Rosenbrock
function value
10^{10}
10^8
10^6
10^4
10^2
10^0
0
0.5
1
1.5
2
2.5
3
3.5
4
total sample size
x 10^5
MRAS
CE υ=0.7
CE υ=0.2
SA


20−D Powel Singular
MRAS
CE υ=0.7
CE υ=0.2
SA
function value
10^{10}
10^{5}
10^{0}
10^{-5}
10^{-10}
0
0.5
1
1.5
2
2.5
3
3.5
4
total sample size
x 10^{5}


20−D Trigonometric
MRAS
CE υ=0.7
CE υ=0.2
SA
function value
10^{5}
10^{4}
10^{3}
10^{2}
10^{1}
10^{0}
0
0.5
1
1.5
2
2.5
3
3.5
4
total sample size
x 10^{5}


20−D Griewank


20−D Pinter

Preliminary Comparisons

- MRAS$_1$ performs well on wide variety of continuous optimization problems
- MRAS$_1$ better adapted to optimization of badly scaled problems
- CE works best on problems that are well scaled and with many local optima

Combinatorial Optimization

- **Numerical Results for Asymmetric Travelling Salesman Problems (ATSPs)**

  (http://www.iwr.uniheidelberg.de/groups/comopt/software/TSPLIB95)

  - Performance similar to CE
  - Very good performance with modest number of tours generated

file	N_c	N_{total} *(std err)*	H_*	H^*	H_{best}	$\delta*_{avg}$ *(std err)*
ftv33	**34**	**7.95e+4(3.25e+3)**	**1364**	**1286**	**1286**	**0.023(0.008)**
ftv35	**36**	**1.02e+5(3.08e+3)**	**1500**	**1475**	**1473**	**0.008(0.002)**
ftv38	**39**	**1.31e+5(4.90e+3)**	**1563**	**1530**	**1530**	**0.008(0.003)**
p43	**43**	**1.02e+5(4.67e+3)**	**5637**	**5620**	**5620**	**0.001(2.5e-4)**
ry48p	**48**	**2.62e+5(1.59e+4)**	**14810**	**14446**	**14422**	**0.012(0.003)**
ft53	**53**	**2.94e+5(1.58e+4)**	**7236**	**6973**	**6905**	**0.029(0.005)**
ft70	**70**	**4.73e+5(2.91e+4)**	**39751**	**38744**	**38673**	**0.017(0.003)**

Extension to Stochastic Optimization

- Objective: find optimal $x^* \in \chi$ such that

$$x^* \in \arg\max_{x\in\chi} E_\psi[H(x,\psi)]$$

  - Assumptions: existence, uniqueness (but possibly many local minima)

- Idea: sample average approximation
  - At each iteration *k*, approximate $E_\psi[H(x,\psi)]$ by

$$\overline{H}_k(x) := \frac{1}{M_k}\sum_{i=1}^{M_k} H_{i,k}(x),$$

where $H_{i,k}(x)$ are i.i.d. random observations at *x*.

Extension to Stochastic Optimization

- Parameter updating
  - (1- ρ) quantiles w.r.t. $f(\cdot,\theta_k)$

$$\gamma_{k+1} = \sup_{l}\{l : P_{\theta_k}(\overline{H}_k(X) \geq l) \geq \rho\}$$

  - update θ_{k+1} as

$$\theta_{k+1} = \arg\max_{\theta\in\Theta} \int_{x\in\chi} [S(\overline{H}_k(x))]^k I\{\overline{H}_k(x) > \gamma_{k+1}\} \ln f(x,\theta) dx$$

Extension to Stochastic Optimization

- Convergence issue
  - $M_k \to \infty$ as $k \to \infty$

- Can prove global convergence w.p. 1 under reasonable conditions

- Practical efficiency
  - increase M_k adaptively, i.e., small M_k value at initial search phase, use large M_k when precise estimates are required
  - performs well on initial simple examples

Conclusions and Future Work

- **Summary**
  - generic approach; algorithm performs well
  - guaranteed theoretical convergence (for NEFs)
  - alternative framework to design optimization algorithms
- **Work in Progress**
  - stochastic optimization problems
- **Future Work**
  - high dimensional problems
  - variety of applications
  - more new algorithms in this framework
    - combination with other algorithms

Formal Definition of NEFs

- **Definition:** A family $\{f(\cdot,\theta), \theta \in \Theta \subseteq \Re^{\mathrm{m}}\}$ is said to belong to the natural exponential family (NEF) if there exist $h(\cdot):\Re^n \rightarrow \Re^+$, $\Gamma(\cdot):\Re^n \rightarrow \Re^m$, and $K(\cdot):\Re^m \rightarrow \Re$ such that

$$f(x,\theta) = \exp\{\theta^T \Gamma(x) - K(\theta)\}h(x), \quad \forall \theta \in \Theta.$$

Global convergence: if $\{f(\cdot,\theta), \theta \in \Theta \subseteq \Re^{\mathrm{m}}\}$ belongs to

NEFs, then under some mild regularity conditions, we have

$$\lim_{k\to\infty} E_{\theta_k}[\Gamma(X)] = \Gamma(x^*).$$

- multivariate normal case

$$\lim_{k\to\infty} \mu_k = x^*, \text{ and } \lim_{k\to\infty} \Sigma_k = 0_{n\times n}.$$

- independent univariate case

$$\lim_{k\to\infty} E_{\theta_k}[X] = x^*.$$

For all test problems, the same set of parameters is used to test MRAS$_1$: $\varepsilon = 10^{-5}$, initial sample size $N_0 = 1000$, $\rho_0 = 0.1$, $\lambda = 0.01$, $\alpha = 1.1$, $r = 10^{-4}$, smoothing parameter $\upsilon = 0.2$, and $N_{min} = 5d$, where d is the dimension of the problem. The initial mean vector μ_0 is a d-by-1 vector with each component randomly selected from the interval $[-50, 50]$ according to the uniform distribution, and Σ_0 is a d-by-d diagonal matrix with all diagonal elements equal to 500.

For comparison purposes, we also applied the CE method and the SA algorithm to the above test functions. For CE, we have used the univariate normal p.d.f. with parameter values suggested in Kroese et al. (2004): sample size $N = 2000$, $\rho = 0.01$, smoothing parameter $\upsilon = 0.7$. Again, the initial mean vector μ_0 is randomly selected from $[-50, 50]^d$ according to the uniform distribution, and Σ_0 is a d-by-d diagonal matrix with all elements equal to 500. We found empirically that the above parameters work well for some functions, but in some other cases, the variance matrices in CE may converge too quickly to the zero matrix, which freezes the algorithm at some low quality solutions. To address this issue, for each problem, we also tried CE with different values of the smoothing parameter. In the numerical results reported below, we have used a smaller smoothing parameter value $\upsilon = 0.2$, which gives reasonable performance for all test cases. For SA, we have used the parameters suggested in Corana et al. (1987): initial temperature $T = 50000$, temperature reduction factor $r_T = 0.85$, the search neighborhood of a point x is taken to be $\mathcal{N}(x) = \{y : \|x - y\|_\infty \leq 1\}$, where $\|x\|_\infty := \max_{1\leq i\leq d} |x_i|$, and the initial solution is uniformly selected from $[-50, 50]^d$.