Model-based Annealing Random Search with Stochastic Averaging

Jiaqiao Hu

Department of Applied Mathematics and Statistics
State University of New York, Stony Brook

July 24, 2012
2012 NSF Workshop on Simulation Methodology

STONY BROOK UNIVERSITY

Problem Setting

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

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

- Compact solution space $X \subseteq R^n$
  - $x := (x_1, x_2, \ldots, x_n)^T$ is the vector of *n* decision variables
  - Continuous or discrete (combinatorial)

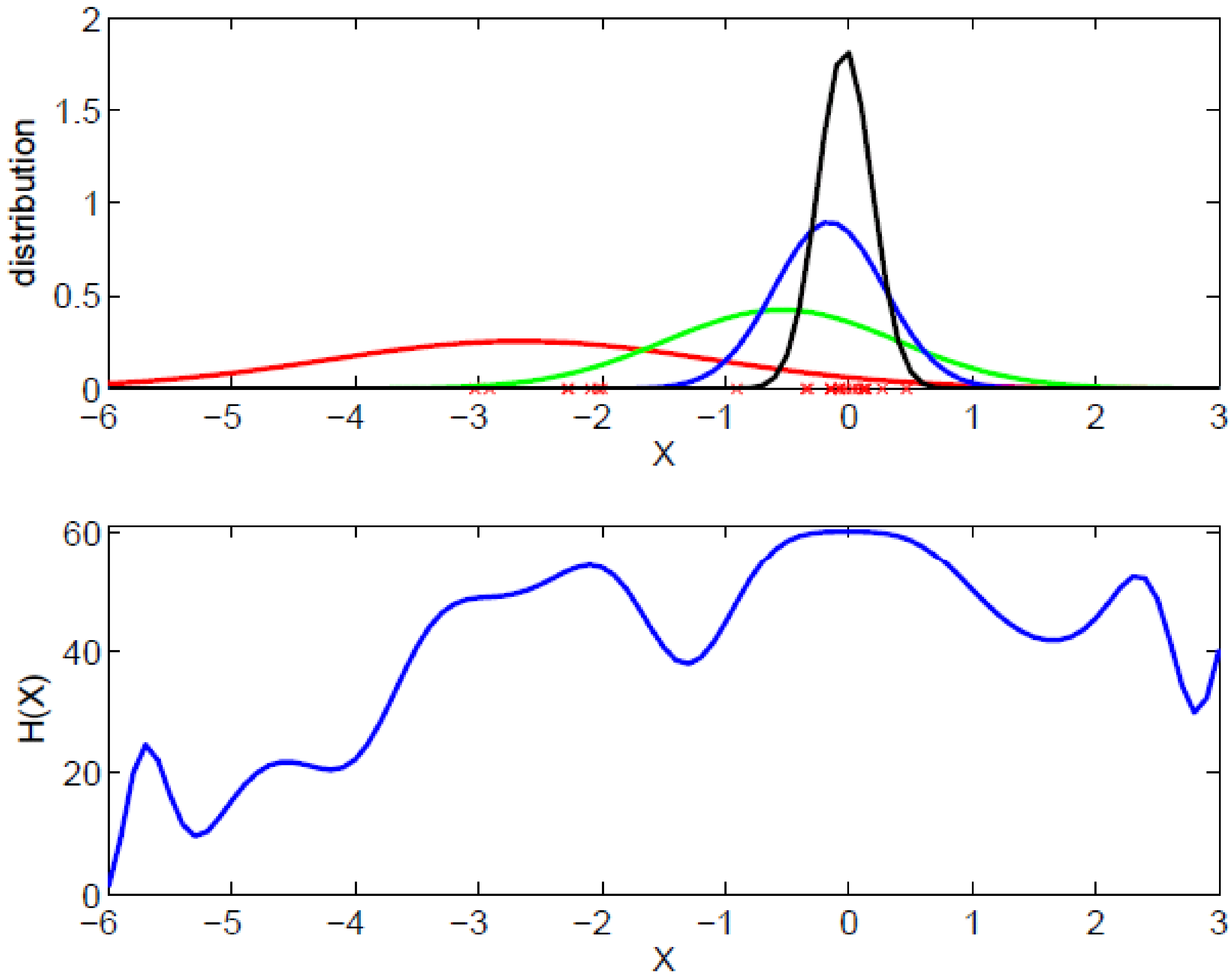
- Assume the uniqueness of x^*, but no structural assumptions on the objective function.

Overview of Related Approaches

- **Instance-based approaches:** generate and use a sequence of candidate solutions to approximate the optimal solution.
  - simulated annealing (Kirkpatrick et al. 1983)
  - genetic algorithms (Srinivas and Patnaik 1994)
  - nested partitions (Shi and Olafsson 2000)
  - tabu search (glover 1990)

- **Model-based approaches:** maintain and update a sequence of probability distributions.
  - annealing adaptive search (AAS) (Romeijn and Smith 1994)
  - estimation of distribution algorithms (EDAs) (Larranaga and Lozano 2002)
  - cross-entropy method (CE) (Rubinstein and Kroese 2004)
  - probability collectives (PCs) (Wolpert 2004)
  - model reference adaptive search (MRAS) (Hu et al. 2007)

Model-Based Approaches



Model-Based Approaches

- Main Steps of Model-based approaches:
  I. Randomly generate solutions by sampling from g_k;
  II. Update g_k to obtain g_{k+1} based on the sampled solutions.

- Idea: Seeking a sequence of distributions $\{g_k\}$ with the hope that

$$g_k \rightarrow g^*,$$

where g^* is a limiting distribution that concentrates on the set of optimal solutions.

Model-based Annealing Random Search

- **Annealing Adaptive Search (AAS) :**
  - Boltzmann distribution: $$g_k(x) = \frac{e^{H(x)/T_k}}{\int_X e^{H(x)/T_k} dx}$$
  - Difficulties:
    - $\{g_k\}$ depends on H, which is unknown.
    - g_k may not have any structure.
  - Solutions:
    - MCMC-based sampling techniques (Zabinsky 2003).
    - Model-based Annealing Random Search (MARS) (Hu and Hu 2010): approximates AAS by sampling from a surrogate distribution that approximates g_k
      - use parameterized distributions $\{f_\theta, \theta \in \Theta\}$;
      - project g_k onto $\{f_\theta, \theta \in \Theta\}$

$$\theta_k = \arg\min_\theta D(g_k \mid f_\theta) := \arg\min_\theta E_{g_k}\left[\ln \frac{g_k(X)}{f_\theta(X)}\right].$$

Model-based Annealing Random Search

- **Model-based Annealing Random Search (a simplified version):**
  - randomly generate solutions by sampling from $f_{\theta k}$;
  - find a new parameter θ_{k+1} of f such that g_{k+1} can be closely approximated by $f_{\theta k+1}$, *i.e.*,

$$\theta_{k+1} = \arg\min_{\theta} D(g_{k+1} \mid f_\theta)$$

$$:= \arg\min_{\theta} E_{g_{k+1}}\left[\ln \frac{g_{k+1}(X)}{f_\theta(X)}\right]$$

$$= \arg\max_{\theta} \left\{ Q_{k+1}(\theta) := \int_X e^{H(x)/T_{k+1}} \ln f_\theta(x) dx \right\}.$$

Model-based Annealing Random Search

- Practical Implementation
  - Estimate the Q-function $Q_{k+1}(\theta)$ by its sample average approximation

    $$\overline{Q}_{k+1}(\theta) := \frac{1}{N}\sum_{l=1}^{N} e^{H(X_k^l)/T_{k+1}} f_{\theta_k}{}^{-1}(X_k^l) \ln f_\theta(X_k^l),$$

    where $X_k^1, \ldots, X_k^N$ are i.i.d. $\sim f_{\theta_k}(x)$.
  - Find the parameter $\overline{\theta}_{k+1}$(an estimate of θ_{k+1}) by solving

    $$\overline{\theta}_{k+1} = \arg\max_{\theta}\left\{\overline{Q}_{k+1}(\theta)\right\}.$$
- Drawbacks:
  - All solutions in previous iterations are discarded.
  - $\overline{\theta}_{k+1}$ is a biased estimator of θ_k because of the ratio bias; require the use of N that grows at least polynomially with k to reduce the bias effect.

MARS with Stochastic Averaging

- Proposed approach: replace the empirical average with a stochastic averaging procedure

$$\hat{Q}_{k+1}(\theta) = (1-\beta_k)\hat{Q}_k(\theta) + \beta_k \frac{1}{N}\sum_{l=1}^{N} e^{H(X_k^l)/T_{k+1}} f_{\theta_k}^{-1}(X_k^l)\ln f_\theta(X_k^l),$$

where $\hat{Q}_1(\theta) = \frac{1}{N}\sum_{l=1}^{N} e^{H(X_0^l)/T_1} f_{\theta_0}^{-1}(X_0^l)\ln f_\theta(X_0^l).$

  - All previous candidate solutions contribute to the estimation of the *Q*-function.
  - Sample size *N* can be held at a small constant value.

- Requires updating an entire function of θ! Focus on the natural exponential family (NEF), i.e.,

$$f_\theta(x) = e^{\theta^T \Gamma(x) - K(\theta)},$$ where

$\Gamma(x)$ is the sufficient statistics, and $K(\theta) = \ln \int_X e^{\theta^T \Gamma(x)} dx.$

MARS with Stochastic Averaging

- An inductive argument shows that $\hat{Q}_{k+1}(\theta)$ can be expressed in the recursive form:

$$\hat{Q}_{k+1}(\theta) = \theta^T S_{k+1} - K(\theta) R_{k+1},$$

where

$$S_{k+1} = (1-\beta_k) S_k + \frac{\beta_k}{N} \sum_{l=1}^{N} e^{H(X_k^l)/T_{k+1}} f_{\theta_k}^{-1}(X_k^l) \Gamma(X_k^l),$$

$$R_{k+1} = (1-\beta_k) R_k + \frac{\beta_k}{N} \sum_{l=1}^{N} e^{H(X_k^l)/T_{k+1}} f_{\theta_k}^{-1}(X_k^l).$$

MARS with Stochastic Averaging

- Replacing $Q_{k+1}(\theta)$ by $\hat{Q}_{k+1}(\theta)$ leads to the optimization problem

$$\hat{\theta}_{k+1} = \arg\max_{\theta} \left\{ \theta^T S_{k+1} - K(\theta) R_{k+1} \right\}.$$

- The closed-form solution is given by

$$m(\hat{\theta}_{k+1}) := E_{\hat{\theta}_{k+1}}[\Gamma(X)] = \frac{S_{k+1}}{R_{k+1}}$$

or equivalently

$$\hat{\theta}_{k+1} = m^{-1}\left(\frac{S_{k+1}}{R_{k+1}}\right).$$

MARS with Stochastic Averaging

- A generalized Boltzmann distribution sequence:

$$\hat{g}_{k+1}(x) = \alpha_k g_{k+1}(x) + (1-\alpha_k) f_{\theta_k}(x), \ \ \alpha_k \in (0,1] \ \forall k.$$

  - use $\hat{g}_{k+1}$ to replace g_{k+1} in calculating the new parameter θ_{k+1} to ensure that the new distribution $f_{\theta_{k+1}}$ does not deviate too much from f_{θ_k} .

- Model-based Annealing Random Search:
  - randomly generate solutions by sampling from f_{θ_k};
  - obtain $f_{\theta_{k+1}}$ by calculating $\theta_{k+1} = \arg\min D(\hat{g}_{k+1} \mid f_\theta)$.

MARS with Stochastic Averaging

- MARS with Stochastic Averaging (MARS-SA):

Step 0: Select $f_{\hat{\theta}_0}$, $\{\alpha_k\}$, $\{\beta_k\}$. Specify $\{T_k\}$, a sample size *N*.

Step 1: Sample *N* solutions $X_k^1, \ldots, X_k^N$ from $f_{\hat{\theta}_k}$.

Step 2: Update S_k and R_k as

$$S_{k+1} = (1-\beta_k)S_k + \frac{\beta_k}{N}\sum_{l=1}^{N} e^{H(X_k^l)/T_{k+1}} f_{\theta_k}{}^{-1}(X_k^l)\Gamma(X_k^l),$$

$$R_{k+1} = (1-\beta_k)R_k + \frac{\beta_k}{N}\sum_{l=1}^{N} e^{H(X_k^l)/T_{k+1}} f_{\theta_k}{}^{-1}(X_k^l).$$

Step 3: Compute the new parameter $\hat{\theta}_{k+1}$

$$m(\hat{\theta}_{k+1}) = \alpha_k \frac{S_{k+1}}{R_{k+1}} + (1\text{-}\alpha_k)m(\hat{\theta}_k).$$

Step 4: Set *k=k+1* and reiterate from Step 1.

Global Convergence of MARS-SA

- *Theorem: Assume*

  1) $\alpha_k > 0, \sum_k \alpha_k = \infty; \beta_k > 0, \sum_k \beta_k = \infty, \sum_k \beta_k^2 < \infty;$

  2) $\alpha_k \beta_k = O(1/k^{3/2});$

  3) $T_k \to 0;$

  4) $\frac{1}{\beta_k}\left(\frac{1}{T_{k+1}} - \frac{1}{T_k}\right) \to 0,$

  then $m(\hat{\theta}_k) \to \Gamma(x^*)$ w.p.1. as $k \to \infty.$

- Intuition: rewrite the recursions for S_k and R_k as:

$$S_{k+1} = S_k + \beta_k \left(\int_X e^{H(x)/T_{k+1}} \Gamma(x) dx - S_k \right) + \beta_k \xi_k,$$

$$R_{k+1} = R_k + \beta_k \left(\int_X e^{H(x)/T_{k+1}} dx - R_k \right) + \beta_k \eta_k,$$

where ξ_k and η_k are martingale difference noises.

Global Convergence of MARS-SA

- Intuition:

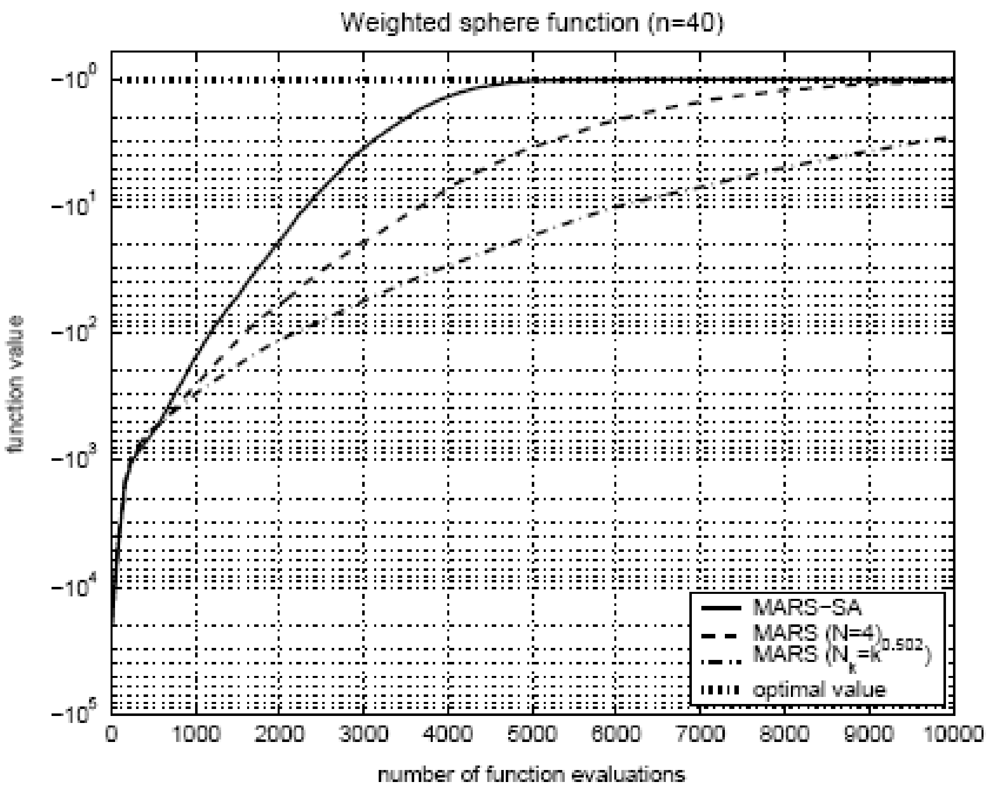
$$
\begin{aligned}
m(\hat{\theta}_{k+1}) &= m(\hat{\theta}_k) + \alpha_k \left(\frac{S_{k+1}}{R_{k+1}} - m(\hat{\theta}_k) \right) \\
&= m(\hat{\theta}_k) + \alpha_k \left(\frac{\int_X e^{H(x)/T_{k+1}} \Gamma(x) dx}{\int_X e^{H(x)/T_{k+1}} dx} - m(\hat{\theta}_k) \right) \\
&\qquad + \alpha_k \left(\frac{S_{k+1}}{R_{k+1}} - \frac{\int_X e^{H(x)/T_{k+1}} \Gamma(x) dx}{\int_X e^{H(x)/T_{k+1}} dx} \right) \\
&= m(\hat{\theta}_k) + \alpha_k \left(E_{g_{k+1}}[\Gamma(X)] - m(\hat{\theta}_k) \right) + \alpha_k \varphi_k .
\end{aligned}
$$

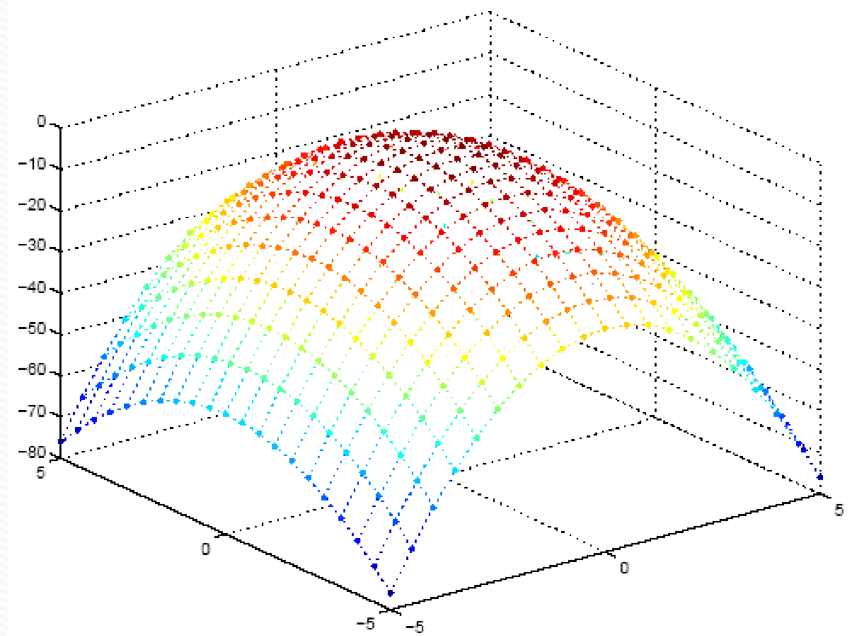
Numerical Examples

Weighted Sphere function ($n = 40,\ -10 \le x_i \le 10,\ i = 1, \ldots, n$)

$$H_1(x) = -1 - \sum_{i=1}^{n} i x_i^2,$$

where $x^* = (0, \ldots, 0)^T$ and $H_1(x^*) = -1$.



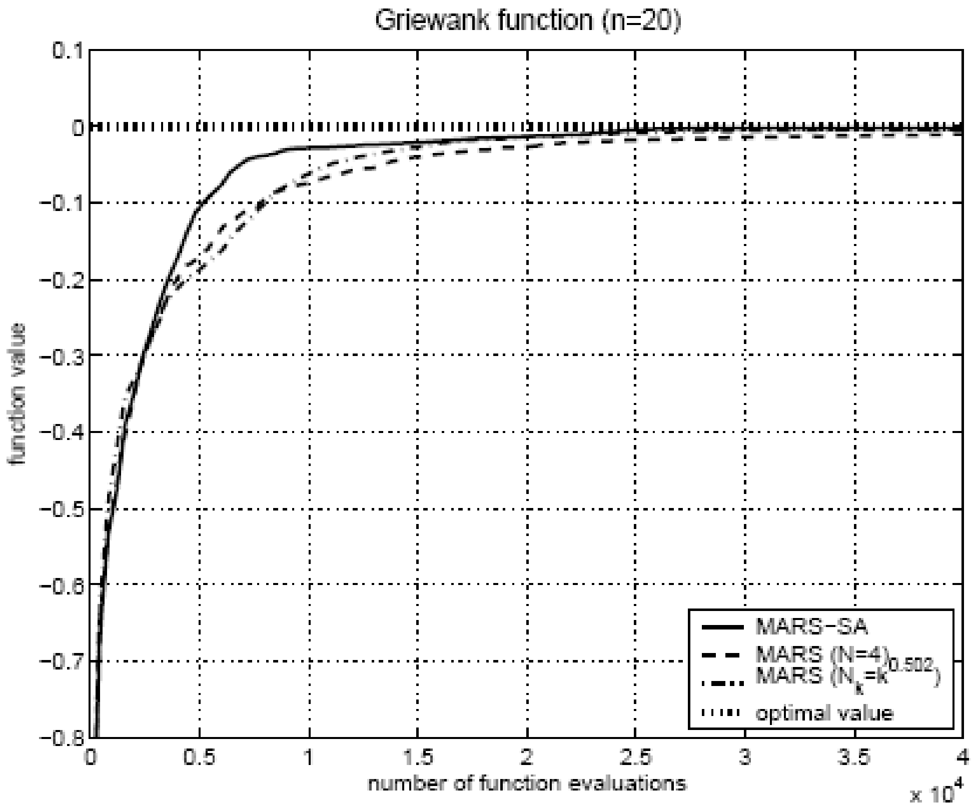


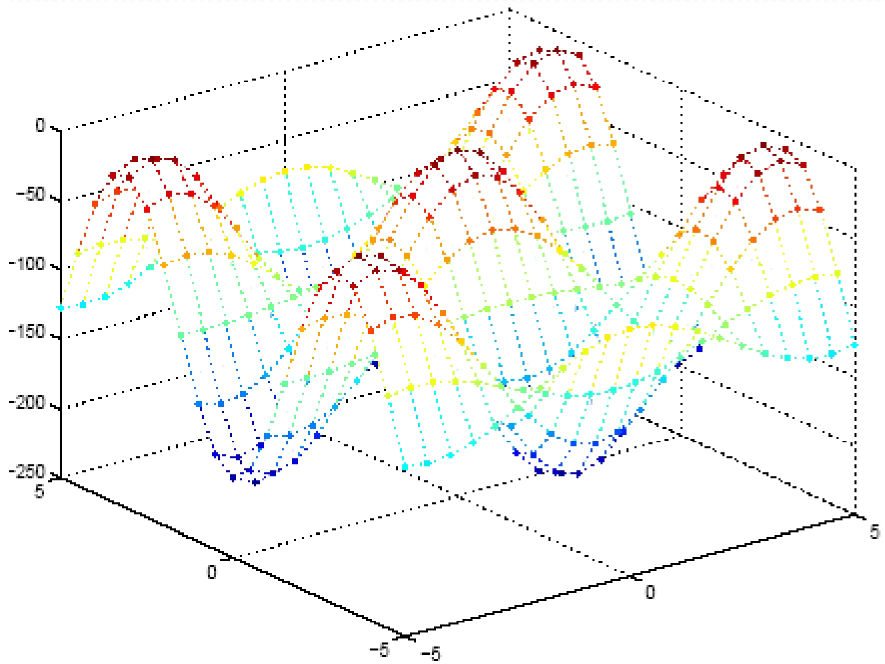
Numerical Examples

Griewank function ($n = 20$, $-10 \le x_i \le 10$, $i = 1, \ldots, n$)

$$H_2(x) = -\frac{1}{4000}\sum_{i=1}^{n} x_i^2 + \prod_{i=1}^{n}\cos\left(\frac{x_i}{\sqrt{i}}\right) - 1,$$

where $x^* = (0, \ldots, 0)^T$, $H_2(x^*) = 0$.



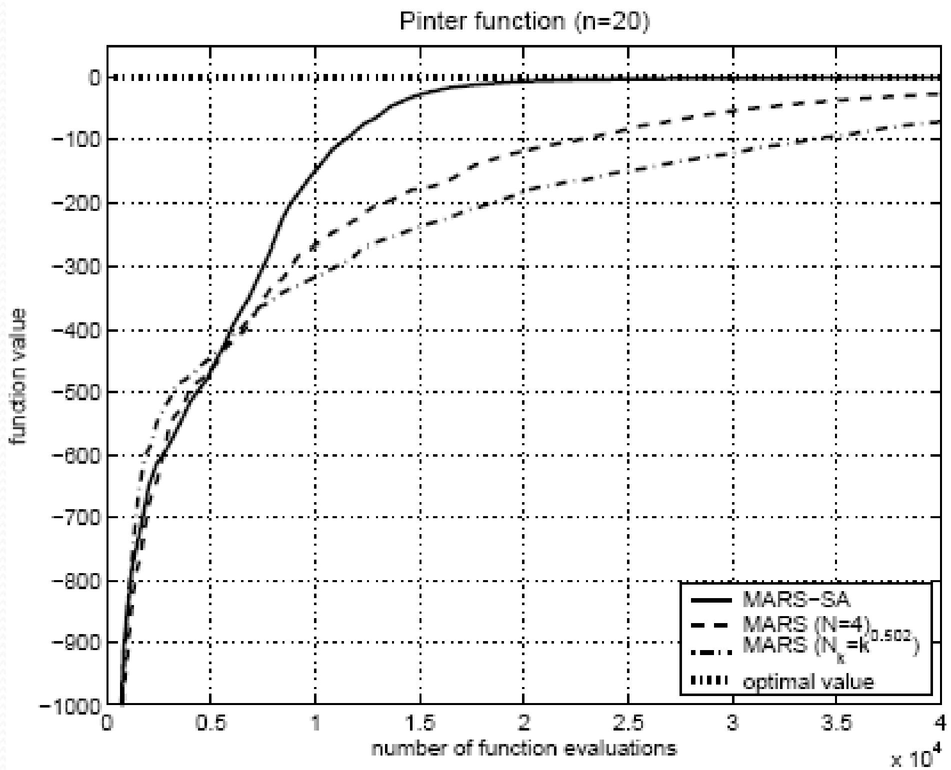


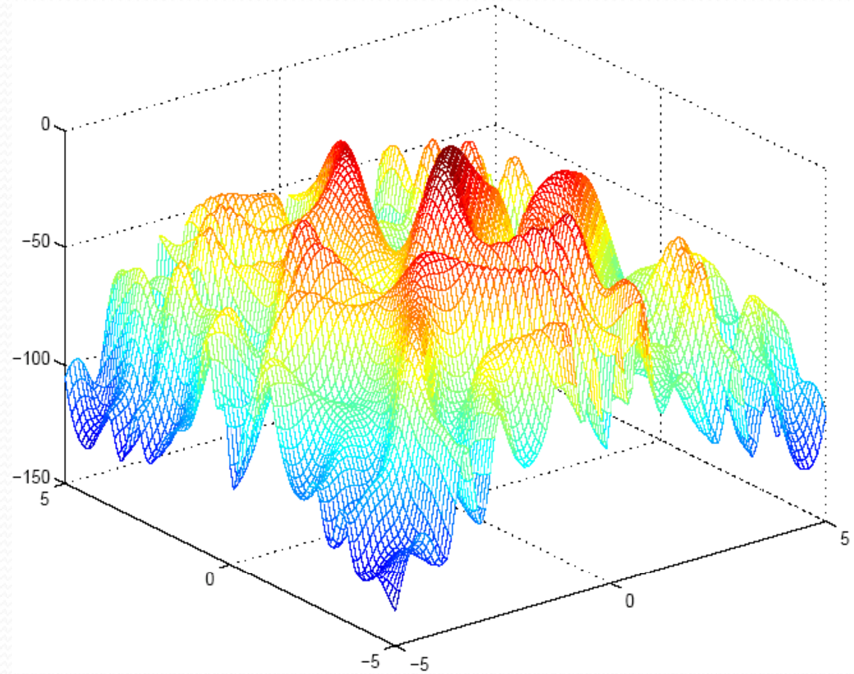
Numerical Examples

Pinter's function ($n = 20,\ -10 \le x_i \le 10,\ i = 1, \ldots, n$)

$$H_3(x) = -\sum_{i=1}^{n} i x_i^2 - \sum_{i=1}^{n} 20 i \sin^2 \left(x_{i-1} \sin x_i - x_i + \sin x_{i+1} \right) - \sum_{i=1}^{n} i \log_{10} \left(1 + i (x_{i-1}^2 - 2x_i + 3x_{i+1} - \cos x_i + 1)^2 \right) - 1,$$

where $x_0 = x_n$, $x_{n+1} = x_1$, $x^* = (0, \ldots, 0)^T$, $H_3(x^*) = -1$.





Conclusions

- Summary
  - a variant of the recently proposed MARS for global optimization, but aims to improve MARS with an additional stochastic averaging procedure.
  - makes more efficient use of past sampling information and thus eliminates the typical polynomially increasing computational requirement.
  - provably convergent to the global optimal solution when the sample size is fixed at a small constant value.