A Model Reference Adaptive Search(MRAS) Method for Global Optimization

Jiaqiao Hu, Michael C.Fu, Steven I.Marcus

Presented by Wenhao Ying
Fudan University

Contents

Overview

- Instance-based: Simulated annealing (SA) (Kirkpatrick et al. 1983), genetic algorithms (GAs) (Srinivas and Patnaik 1994), nested partitions (NP) (Shi and Olafsson 2000)
- Model-based: Cross-entropy (CE) (Rubinstein and Kroese 2004, De Boer et al. 2005), estimation of distribution algorithms (EDAs) (Muhlenbein and Paab 1996)
- Solving global optimization problems works with a parameterized probabilistic model
    - Generate candidate solutions(e.g., random samples)
    - Update the probabilistic model to bias the future search toward "better" solutions

Overview

- Cross-entropy Method
  - A family of parameterized probability distribution
  - Find the parameter that assigns maximum probability to the set of optimal solutions

$$H(p,q) = -\sum_{i=1}^{n} p(x_i) log(q(x_i))$$

- KL divergence
  - Measure how one probability distribution Q is different from a second, reference probability distribution P

$$D_{KL}(p||q) = \sum_{i=1}^{n} p(x_i) log(\frac{p(x_i)}{q(x_i)})$$

$$D_{KL}(p||q) = \sum_{i=1}^{n} p(x_i) log(\frac{p(x_i)}{q(x_i)}) = \sum_{i=1}^{n} p(x_i) log(p(x_i)) - \sum_{i=1}^{n} p(x_i) log(q(x_i)) = -H(p(x)) + [-\sum_{i=1}^{n} p(x_i) log(q(x_i))]$$

MRAS

- Resembles CE in that it works with a family of parameterized distributions on the solution space
- Use a sequence of intermediate reference distributions to facilitate and guide the updating of the parameters associated with the family of parameterized distributions during the search process
- At each iteration of MRAS, candidate solutions are generated from the distribution that possesses the minimum KL divergence with respect to the reference model corresponding to the previous iteration
- Construct the next distribution by minimizing the KL divergence with respect to the current reference model, from which future candidate solutions will be generated

MRAS

- A new framework for global optimization, which allows considerable flexibility in the choices of the reference models
- An instantiation of the MRAS method, which incorporates the key ideas of CE and EDAs
- Explore the relationship between CE and MRAS

MRAS

- We consider the following global optimization problem:

$$x^* \in \arg\max_{x \in \mathscr{X}} H(x), \quad \mathscr{X} \subseteq \Re^n, \qquad (1)$$

where the solution space $\mathscr{X}$ is a nonempty set in $\Re^n$, and $H(\cdot)$: $\mathscr{X} \to \Re$ is a deterministic function that is bounded from below, i.e., $\exists \mathscr{M} > -\infty$ such that $H(x) \geqslant \mathscr{M}$ $\forall x \in \mathscr{X}$.

Throughout this paper, we assume that problem (1) has a unique global optimal solution, i.e., $\exists x^* \in \mathscr{X}$ such that $H(x) < H(x^*)$ $\forall x \neq x^*$, $x \in \mathscr{X}$.

ASSUMPTION A1. *For any given constant* $\xi < H(x^*)$, *the set* $\{x\colon H(x) \geqslant \xi\} \cap \mathscr{X}$ *has a strictly positive Lebesgue or discrete measure.*

Ensures that any neighborhood of the optimal solution x^* with a positive probability of being sampled.

ASSUMPTION A2. *For any given constant* $\delta > 0$, $\sup_{x \in A_\delta} H(x) < H(x^*)$, *where* $A_\delta := \{x\colon \|x - x^*\| \geqslant \delta\} \cap \mathscr{X}$, *and we define the supremum over the empty set to be* $-\infty$.

Since *H(x)* has a unique global optimizer, Assumption 2 satisfied by many functions in practice.

MRAS

- Family of parameterized distributions: $\{f(\cdot,\theta)\theta \in \Theta\}$, where Θ is the parameter space.
- k th iteration $\longrightarrow$ $f(\cdot,\theta_k)$ sampling distribution
- Evaluate candidate solutions and calculate a new parameter vector $\theta_{k+1} \in \Theta$ according to a specific updating rule
- Repeat until a termination criterion is satisfied
- A sequence of distributions $\{g_k(\cdot)\}$ $\longrightarrow$ reference distribution
- Look at the projection of $\{g_k(\cdot)\}$ on the family of distributions
- Compute the new parameter vector θ_{k+1} $\longrightarrow$ minimize KL divergence

MRAS

$$\mathcal{D}(g_k, f(\cdot,\theta)) := E_{g_k}\left[ln\frac{g_k(X)}{f(X,\theta)}\right]$$

$$= \int_{\mathcal{X}} ln\frac{g_k(x)}{f(x,\theta)} g_k(x)\nu d(x)$$

Intuitively, $f(\cdot, \theta_{k+1})$ can be viewed as a compact representation (approximation) of the reference distribution $g_{k(\cdot)}$

Consequently, the performance of this method will largely depend on the choices of reference distributions

$$g_k(x) = \frac{H(x)g_{k-1}(x)}{\int_{\mathscr{X}} H(x)g_{k-1}(x)\,\nu(dx)} \quad \forall x \in \mathscr{X}. \qquad (2)$$

Let $g_0(x) > 0 \forall x \in X$ be an initial p.d.f/p.m.f by tilting the old p.d.f/p.m.f with the performance function *H(x)*.

$$E_{g_k}[H(x)] = \frac{E_{g_{k-1}}\left[H(x)^2\right]}{E_{g_{k-1}}[H(x)]} \geq E_{g_{k-1}}[H(x)]$$

$$E_{g_k}[H(x)] = \int H(x) \cdot \frac{H(x) \cdot g_{k-1}(x)}{\int H(x) \cdot g_{k-1} dx} \cdot dx$$

$$\frac{E_{g_{k-1}}\left[H(x)^2\right]}{E_{g_{k-1}}[H(x)]} = \frac{\int H(x)^2 \cdot g_{k-1}(x) dx}{\int H(x) \cdot g_{k-1}(x) dx}$$

$$E_{g_{k-1}}\left[H(x)^2\right] = \left[E_{g_{k-1}} H(x)\right]^2 + Var(H(x))$$

MRAS Algorithm

Throughout the analysis, we use $P_{\theta_k}(\cdot)$ and $E_{\theta_k}[\cdot]$ to denote the probability and expectation taken with respect to the p.d.f./p.m.f. $f(\cdot, \theta_k)$, and $I_{\{\cdot\}}$ to denote the indicator function, i.e.,

$$I_{\{A\}} := \begin{cases} 1 & \text{if event } A \text{ holds,} \\ 0 & \text{otherwise.} \end{cases}$$

$$P_{\theta_k}(H(X) \geqslant \gamma) = \int_{\mathscr{X}} I_{\{H(x) \geqslant \gamma\}} f(x, \theta_k)\, \nu(dx)$$

$$E_{\theta_k}[H(X)] = \int_{\mathscr{X}} H(x) f(x, \theta_k)\, \nu(dx).$$

Algorithm MRAS_0**—Exact Version**

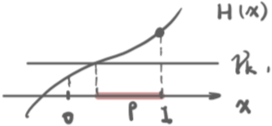
- Initialization:

$\rho \in (0,1]$, $\varepsilon \geq 0$, $S(\cdot) : \mathfrak{R} \rightarrow \mathfrak{R}^{+}, f(x, \theta_0) > 0 \in \forall x \in \mathcal{X}$

Set the iteration counter $k = 0$

- *Step 1.* Calculate the (1-ρ) quantile

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



- *Step 2.* if $k == 0$:

$$\bar{\gamma}_{k+1} = \gamma_{k+1}$$

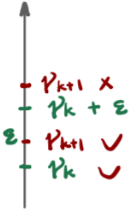
elseif $k \geq 1$:

if $\bar{\gamma}_{k+1} \geq \gamma_k + \varepsilon$:

$$\bar{\gamma}_{k+1} = \gamma_{k+1}$$

else:

$$\bar{\gamma}_{k+1} = \gamma_k$$



MRAS Algorithm

- *Step 3.* Updating the parameter vector

$$\theta_{k+1} := \arg\max_{\theta \in \Theta} E_{\theta_k}\left[\frac{[S(H(X))]^k}{f(X,\theta_k)} I_{\{H(X) \geq \bar{\gamma}_{k+1}\}} \ln f(X,\theta)\right]. \quad (3)$$

$$\theta_{k+1} = \arg\max_{\theta \in \Theta} \int_{\mathscr{X}} [S(H(x))]^k I_{\{H(x) \geq \bar{\gamma}_{k+1}\}} \ln f(x,\theta)\, \nu(dx).$$

- *Step 4.* $k = k + 1$

MRAS Algorithm

LEMMA 1. *The parameter θ_{k+1} computed at the kth iteration of the* MRAS$_0$ *algorithm minimizes the KL-divergence* $\mathcal{D}(g_{k+1}, f(\cdot, \theta))$, *where*

$$g_{k+1}(x) := \frac{S(H(x))I_{\{H(x)\geq\bar{\gamma}_{k+1}\}}g_k(x)}{E_{g_k}[S(H(X))I_{\{H(X)\geq\bar{\gamma}_{k+1}\}}]} \quad \forall x \in \mathcal{X},\ k = 1, 2, \ldots, \quad \textit{and}$$

$$g_1(x) := \frac{I_{\{H(x)\geq\bar{\gamma}_1\}}}{E_{\theta_0}\left[I_{\{H(X)\geq\bar{\gamma}_1\}}/f(X, \theta_0)\right]}.$$

$$g_{k+1}(x) = \frac{H(x) \cdot g_k(x)}{E_{g_k}[H(x)]} = \frac{H(x) \cdot g_k(x)}{\int H(x) \cdot g_k(x)dx}$$

- Global convergence depends on the choice of the parameterized distribution family.
- MRAS focus on a particular family of distributions called the natural exponential family(NEF)(cf., e.g., Morris 1982)

DEFINITION 1. A parameterized family of p.d.f.'s/p.m.f.'s $\{f(\cdot,\theta),\ \theta \in \Theta \subseteq \Re^m\}$ on $\mathscr{X}$ is said to belong to the NEF if there exist functions $h(\cdot)\colon \Re^n \to \Re$, $\Gamma(\cdot)\colon \Re^n \to \Re^m$, and $K(\cdot)\colon \Re^m \to \Re$ such that

$$f(x,\theta) = \exp\{\theta^T \Gamma(x) - K(\theta)\} h(x) \quad \forall \theta \in \Theta, \tag{4}$$

where $K(\theta) = \ln \int_{x \in \mathscr{X}} \exp\{\theta^T \Gamma(x)\} h(x)\, \nu(dx)$, and the superscript T denotes the vector transposition. For the case where $f(\cdot,\theta)$ is a p.d.f., we assume that $\Gamma(\cdot)$ is a continuous mapping.

Convergence Analysis

$$f(x;\theta) = h(x) \cdot c(\theta) \cdot \exp\left\{ \sum_{i=1}^{k} w_i(\theta) t_i(x) \right\}$$

- Possion λ

$$f(x) = e^{-\lambda \cdot \frac{\lambda^x}{x!}} = \exp\{\log \lambda \cdot x - \lambda\} \frac{1}{x!}$$

- Binomial

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x} = \exp\{log(p) \cdot x + nlog(1-p)\} \binom{n}{x}$$

- Normal

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\}$$

$$= \frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{x^2}{2\sigma^2} + \frac{x \cdot \mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \frac{log(\sigma^2)}{2} \right\}$$

ASSUMPTION A3. *There exists a compact set* Π *such that the level set* $\{x: H(x) \geqslant \bar{\gamma}_1\} \cap \mathscr{X} \subseteq \Pi$, *where* $\bar{\gamma}_1 = \sup_l\{l: P_{\theta_0}(H(X) \geqslant l) \geqslant \rho\}$ *is defined as in the* MRAS_0 *algorithm.*

ASSUMPTION A4. *The maximizer of Equation* (3) *is an interior point of* Θ *for all* k.

ASSUMPTION A5. $\sup_{\theta\in\Theta} \|\exp\{\theta^T\Gamma(x)\}\Gamma(x)h(x)\|$ *is integrable/summable with respect to* x, *where* θ, $\Gamma(\cdot)$, *and* $h(\cdot)$ *are defined as in Definition* 1.

- Assumption A3 restricts the search of MRAS to some compact set.
- Assumption A4 is satisfied when MRAS is posed as an unconstrained optimization problem and Assumption A5 is satisfied by most NEFs.

LEMMA 2. *If Assumptions* A3–A5 *hold, then we have*

$$E_{\theta_{k+1}}[\Gamma(X)] = E_{g_{k+1}}[\Gamma(X)] \quad \forall k = 0, 1, \ldots,$$

where $E_{\theta_{k+1}}[\cdot]$ *and* $E_{g_{k+1}}[\cdot]$ *denote the expectations taken with respect to* $f(\cdot, \theta_{k+1})$ *and* $g_{k+1}(\cdot)$, *respectively.*

THEOREM 1. *Let* $\{\theta_k, k = 1, 2, \ldots\}$ *be the sequence of parameters generated by* MRAS_0. *If* $\varepsilon > 0$ *and Assumptions* A1–A5 *are satisfied, then*

$$\lim_{k\to\infty} E_{\theta_k}[\Gamma(X)] = \Gamma(x^*), \tag{5}$$

where the limit is component-wise.

- Remark

$\Gamma(\cdot)$ is a one-to-one mapping ⟶ $\Gamma^{-1}(\lim_{k\to\infty} E_{\theta_k}[\Gamma(x)]) = x^*$

The solution vector *x* will be a component of $\Gamma(x)$ (e.g., multivariate normal distribution ⟶ $\lim_{k\to\infty} E_{\theta_k}[\Gamma(x)] = x^*$

When the components of the random vector X =(X_1, X_2, … X_n) are independent, i.e., each has a univariate p.d.f/p.m.f ⟶ $\Gamma(x) = x$

If take the parameter $\varepsilon = 0$, then Step 2 of MRAS is equivalent to $\bar{\gamma}_{k+1} = max_{1\le i\le k+1}$, thus strict increment is bounded by $\min|H(x) - H(y)|$
Thus, the assumption in Theorem 1 can be relaxed to $\varepsilon \geq 0$

COROLLARY 1 (MULTIVARIATE NORMAL). *For continuous optimization problems in* $\Re^n$, *if multivariate normal p.d.f.'s are used in* MRAS_0, *i.e.,*

$$f(x, \theta_k) = \frac{1}{\sqrt{(2\pi)^n |\Sigma_k|}} \exp\left(-\tfrac{1}{2}(x-\mu_k)^T \Sigma_k^{-1} (x-\mu_k)\right), \qquad (6)$$

where $\theta_k := (\mu_k; \Sigma_k)$, $\varepsilon > 0$, *and Assumptions* A1–A4 *are satisfied, then*

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

where $0_{n\times n}$ *represents an n-by-n zero matrix.*

By Lemma 2, $E_{x\sim MNV(\mu,\Sigma)}(x) = \mu_{k+1} = E_{g_{k+1}}(x)$

$E_{x\sim MNV(\mu,\Sigma)}\left[(x-\mu)(x-\mu)^T\right] = E_{g_{k+1}}(x)\left[(x-\mu_{k+1})(x-\mu_{k+1})^T\right]$

By Theorem 1, $\lim_{k\to\infty} E_{g_k}(X) = x^*$, $\lim_{k\to\infty} E_{g_k}\left[(X-\mu_k)(X-\mu_k)^T\right] = 0$

$$\mu_{k+1} = \frac{E_{\theta_k}[\{[S(H(X))]^k / f(X, \theta_k)\} I_{\{H(X) \geqslant \bar{\gamma}_{k+1}\}} X]}{E_{\theta_k}[\{[S(H(X))]^k / f(X, \theta_k)\} I_{\{H(X) \geqslant \bar{\gamma}_{k+1}\}}]} \tag{7}$$

and

$$\Sigma_{k+1} = \frac{E_{\theta_k}[\{[S(H(X))]^k / f(X, \theta_k)\} I_{\{H(X) \geqslant \bar{\gamma}_{k+1}\}} (X - \mu_{k+1})(X - \mu_{k+1})^T]}{E_{\theta_k}[\{[S(H(X))]^k / f(X, \theta_k)\} I_{\{H(X) \geqslant \bar{\gamma}_{k+1}\}}]}, \tag{8}$$

COROLLARY 2 (INDEPENDENT UNIVARIATE). *If the components of the random vector* $X = (X_1, \ldots, X_n)$ *are independent, each has a univariate p.d.f./p.m.f. of the form*

$$f(x_i, \vartheta_i) = \exp(x_i \vartheta_i - K(\vartheta_i)) h(x_i), \quad \vartheta_i \in \Re \;\; \forall i = 1, \ldots, n,$$

$\varepsilon > 0$, *and Assumptions* A1–A5 *are satisfied, then*

$$\lim_{k \to \infty} E_{\theta_k}[X] = x^*, \quad \textit{where } \theta_k := (\vartheta_1^k, \ldots, \vartheta_n^k).$$

Algorithm CE_0**: Deterministic Version of the CE Method**

Step 1. Choose the initial p.d.f./p.m.f. $f(\cdot, \theta_0)$, $\theta_0 \in \Theta$. Specify the parameter $\rho \in (0, 1]$ and a nondecreasing function $\varphi(\cdot)$: $\Re \rightarrow \Re^+ \cup \{0\}$. Set $k = 0$.

Step 2. Calculate the $(1-\rho)$-quantile γ_{k+1} as

$$\gamma_{k+1} := \sup\{l\colon P_{\theta_k}(H(X) \geqslant l) \geqslant \rho\}.$$

Step 3. Compute the new parameter

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

Step 4. If a stopping rule is satisfied, then terminate; otherwise set $k = k + 1$ and go to Step 2.

Cross-Entropy Method

LEMMA 3. *The parameter* θ_{k+1} *computed at the kth iteration of the* CE_0 *algorithm minimizes the KL-divergence* $\mathcal{D}(g_{k+1}^{ce}, f(\cdot, \theta))$, *where*

$$g_{k+1}^{ce}(x) := \frac{\varphi(H(x)) I_{\{H(x) \geqslant \gamma_{k+1}\}} f(x, \theta_k)}{E_{\theta_k}[\varphi(H(X)) I_{\{H(X) \geqslant \gamma_{k+1}\}}]} \quad \forall x \in \mathscr{X}, k = 0, 1, \ldots . \tag{9}$$

$$E_{g_{k+1}^{ce}}[\varphi(H(X)) I_{\{H(X) \geqslant \gamma_{k+1}\}}] = \frac{E_{\theta_k}[(\varphi(H(X)) I_{\{H(X) \geqslant \gamma_{k+1}\}})^2]}{E_{\theta_k}[\varphi(H(X)) I_{\{H(X) \geqslant \gamma_{k+1}\}}]} \geqslant E_{\theta_k}[\varphi(H(X)) I_{\{H(X) \geqslant \gamma_{k+1}\}}].$$

THEOREM 2. *For the* CE_0 *algorithm, we have*

$$E_{\theta_{k+1}}[\varphi(H(X))I_{\{H(X)\geqslant\gamma_{k+1}\}}]\geqslant E_{\theta_k}[\varphi(H(X))I_{\{H(X)\geqslant\gamma_{k+1}\}}] \quad \forall k=0,1,\ldots.$$

In the standard CE method, Theorem 2 implies the monotonicity of the sequence $\{\gamma_k: k=1,2,\ldots\}$.

LEMMA 4. *For the standard* CE *method (i.e.,* CE_0 *with* $\varphi(H(x))=1$*), we have*

$$\gamma_{k+2}\geqslant\gamma_{k+1} \quad \forall k=0,1,\ldots.$$

PROOF. By Theorem 2, we have

$$E_{\theta_{k+1}}[I_{\{H(X)\geqslant\gamma_{k+1}\}}]\geqslant E_{\theta_k}[I_{\{H(X)\geqslant\gamma_{k+1}\}}],$$

i.e.,

$$P_{\theta_{k+1}}(H(X)\geqslant\gamma_{k+1})\geqslant P_{\theta_k}(H(X)\geqslant\gamma_{k+1})\geqslant\rho.$$

Cross-Entropy Method

LEMMA 5. *Assume that:*

(1) *There exists a compact set* $\bar{\Pi}$ *such that the level set* $\{x\colon H(x) \geqslant \gamma_k\} \cap \mathscr{X} \subseteq \bar{\Pi}$ *for all* $k = 1, 2, \ldots,$ *where* $\gamma_k = \sup_l\{l\colon P_{\theta_{k-1}}(H(X) \geqslant l) \geqslant \rho\}$ *is defined as in the* CE_0 *algorithm.*

(2) *The parameter* θ_{k+1} *computed at Step* 3 *of the* CE_0 *algorithm is an interior point of* Θ *for all* k.

(3) *Assumption* A5 *is satisfied.*

Then,

$$E_{\theta_{k+1}}[\Gamma(X)] = E_{g^{ce}_{k+1}}[\Gamma(X)] \quad \forall k = 0, 1, \ldots .$$

- The difference between CE method and MRAS is that, where the convergence of the reference model is guaranteed, the convergence of the reference model in CE method relies on the choices of the families of distributions and the value of the parameter ρ used.

Monte Carlo version of MRAS

When quantile values and expectations can be valued exactly

$$\theta_{k+1} := \underset{\theta\in\Theta}{argmax} E_{\theta_k}\left[\frac{[S(H(X))]^k}{f(X,\theta_k)} I_{\{H(X)\geq\bar{\gamma}_{k+1}\}} \ln f(x,\theta)\right]$$

In practice we usually resort to its stochastic counterpart, where only a finite number of samples are used and expected values are replaced with their corresponding sample averages.

$$\tilde{\theta}_{k+1} = \underset{\theta\in\Theta}{argmax} \frac{1}{N}\sum_{i=1}^{N}\left[\frac{[S(H(X_i))]^k}{f(X_i,\tilde{\theta}_k)} I_{\{H(X_i)\geq\bar{\gamma}_{k+1}\}} \ln f(x_i,\theta)\right]$$

where $X_1, \ldots, X_N$ are i.i.d. random samples generated from $f(x, \tilde{\theta}_k)$, $\tilde{\theta}_k$ is the estimated parameter vector computed at the previous iteration, and $\bar{\gamma}_{k+1}$ is a threshold determined by the sample $(1-\rho)$-quantile of $H(X_1), \ldots, H(X_N)$.

Monte Carlo version of MRAS

- It is difficult to determined in advance the appropriate number of samples. A sample size too small may cause fail to converge, whereas too large may lead to high computational cost.
- The parameter ρ will affect the performance of the algorithm. Large values of ρ means almost all samples generated will be used to update the probabilistic model, which could slow down the convergence process.
- Small values of ρ will require a large number of samples to be generated at each iteration and result in significant simulation efforts.
- Modified Monte Carlo version of MRAS – sample size N is adaptively increasing and the parameter ρ is adaptively decreasing (Homem-de-Mello 2007).

Monte Carlo version of MRAS

- The parameter ρ and the sample size N may change from one iteration to another.
- The rate of increase in the sample size is controlled by an extra parameter $\alpha > 1$.
- The initial sample size is N_0, then after k increment, the sample size will be approximately $\alpha^k N_0$.

Assumption A3′. *There exists a compact set* Π_ε *such that* $\{x\colon H(x) \geqslant H(x^*) - \varepsilon\} \cap \mathscr{X} \subseteq \Pi_\varepsilon$. *Moreover, the initial density/mass function* $f(x, \theta_0)$ *is bounded away from zero on* Π_ε, *i.e.,* $f_* := \inf_{x\in\Pi_\varepsilon} f(x, \theta_0) > 0$.

Algorithm MRAS_1**—Monte Carlo version**

- **Initialization:** Specify $\rho_0 \in (0, 1]$, an initial sample size $N_0 > 1$, $\varepsilon \geqslant 0$, $\alpha > 1$, a mixing coefficient $\lambda \in (0, 1]$, a strictly increasing function $S(\cdot)\colon \Re \to \Re^+$, and an initial p.d.f. $f(x, \theta_0) > 0\ \forall x \in \mathscr{X}$. Set $\tilde{\theta}_0 \leftarrow \theta_0$, $k \leftarrow 0$.

Monte Carlo version of MRAS

- **Repeat until a specified stopping rule is satisfied:**

Step 1. Generate N_k i.i.d. samples $X_1^k, \ldots, X_{N_k}^k$ according to $\tilde{f}(\cdot, \tilde{\theta}_k) := (1-\lambda) f(\cdot, \tilde{\theta}_k) + \lambda f(\cdot, \theta_0)$.

Step 2. Compute the sample $(1 - \rho_k)$-quantile $\tilde{\gamma}_{k+1}(\rho_k, N_k) := H_{(\lceil (1-\rho_k) N_k \rceil)}$, where $\lceil a \rceil$ is the smallest integer greater than a, and $H_{(i)}$ is the ith order statistic of the sequence $\{H(X_i^k), i = 1, \ldots, N_k\}$.

Step 3. **If** $k = 0$ **or** $\tilde{\gamma}_{k+1}(\rho_k, N_k) \geqslant \bar{\gamma}_k + \varepsilon/2$, **then**

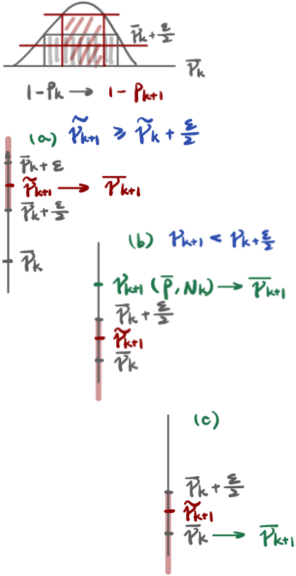
3(a). Set $\bar{\gamma}_{k+1} \leftarrow \tilde{\gamma}_{k+1}(\rho_k, N_k)$, $\rho_{k+1} \leftarrow \rho_k$, $N_{k+1} \leftarrow N_k$.

else, find the largest $\bar{\rho} \in (0, \rho_k)$ such that $\tilde{\gamma}_{k+1}(\bar{\rho}, N_k) \geqslant \bar{\gamma}_k + \varepsilon/2$.

3(b). **If** such a $\bar{\rho}$ exists, **then** set $\bar{\gamma}_{k+1} \leftarrow \tilde{\gamma}_{k+1}(\bar{\rho}, N_k)$, $\rho_{k+1} \leftarrow \bar{\rho}$, $N_{k+1} \leftarrow N_k$.

3(c). **else** (if no such $\bar{\rho}$ exists), set $\bar{\gamma}_{k+1} \leftarrow \bar{\gamma}_k$, $\rho_{k+1} \leftarrow \rho_k$, $N_{k+1} \leftarrow \lceil \alpha N_k \rceil$.

endif



Monte Carlo version of MRAS

Step 4. Compute $\tilde{\theta}_{k+1}$ as

$$\tilde{\theta}_{k+1} = \arg\max_{\theta\in\Theta} \frac{1}{N_k} \sum_{i=1}^{N_k} \frac{[S(H(X_i^k))]^k}{\tilde{f}(X_i^k, \tilde{\theta}_k)} \cdot I_{\{H(X_i^k)\geqslant \bar{\gamma}_{k+1}\}} \ln f(X_i^k, \theta). \quad (12)$$

Step 5. Set $k \leftarrow k+1$.

- In practice, the initial density can be chosen according to some prior knowledge of the problem structure.
- One simple choice is the uniform distribution.
- Forces the algorithm to explore the entire solution space and to maintain a global perspective during the search process.

- Step 3 of MRAS is used to extract a sequence of nondecreasing thresholds and to determine the appropriate values of ρ and N to be used in subsequent iterations.

ASSUMPTION A4′. *The parameter vector* $\tilde{\theta}_{k+1}$ *computed at Step* 4 *of* MRAS$_1$ *is an interior point of* Θ *for all* k.

It is important to note that the set $\{x: H(x) \geqslant \bar{\gamma}_{k+1}, x \in \{X_1^k, \ldots, X_{N_k}^k\}\}$ could be empty if Step 3(c) is visited. If this happens, the right-hand side of (12) will be equal to zero, so any $\theta \in \Theta$ is a maximizer, and we define $\tilde{\theta}_{k+1} := \tilde{\theta}_k$ in this case.

Convergence Analysis

Let $\tilde{g}_{k+1}(\cdot)$, $k = 0, 1, \ldots,$ be defined by

$$\tilde{g}_{k+1}(x) := \begin{cases} \dfrac{[[S(H(x))]^k / \tilde{f}(x, \tilde{\theta}_k)] I_{\{H(x) \geq \bar{\gamma}_{k+1}\}}}{\sum_{i=1}^{N_k} [[S(H(X_i^k))]^k / \tilde{f}(X_i^k, \tilde{\theta}_k)] I_{\{H(X_i^k) \geq \bar{\gamma}_{k+1}\}}} \\ \quad \text{if } \{x \colon H(x) \geq \bar{\gamma}_{k+1}, x \in \{X_1^k, \ldots, X_{N_k}^k\}\} \neq \varnothing, \\ \tilde{g}_k(x), \quad \text{otherwise,} \end{cases} \tag{13}$$

where $\bar{\gamma}_{k+1}$ is given by

$$\bar{\gamma}_{k+1} := \begin{cases} \tilde{\gamma}_{k+1}(\rho_k, N_k) & \text{if Step 3(a) is visited,} \\ \tilde{\gamma}_{k+1}(\bar{\rho}, N_k) & \text{if Step 3(b) is visited,} \\ \bar{\gamma}_k & \text{if Step 3(c) is visited.} \end{cases}$$

Convergence Analysis

LEMMA 6. *If Assumptions* A4′ *and* A5 *hold, then the parameter* $\tilde{\theta}_{k+1}$ *computed at Step* 3 *of* MRAS_1 *satisfies*

$$E_{\tilde{\theta}_{k+1}}[\Gamma(X)] = E_{\tilde{g}_{k+1}}[\Gamma(X)] \quad \forall k = 0, 1, \ldots .$$

The proof Lemma 6 is similar to the proof of Lemma 2.

LEMMA 7. *For any given* $\rho^\dagger \in (0, 1)$, *let* $\gamma_k^\dagger$ *be the set of* $(1-\rho^\dagger)$*-quantiles of* $H(X)$ *with respect to the p.d.f./p.m.f.* $\tilde{f}(\cdot, \tilde{\theta}_k)$, *and let* $\tilde{\gamma}_k^\dagger(\rho^\dagger, N_k)$ *be the corresponding sample quantile of* $H(X_1^k), \ldots, H(X_{N_k}^k)$, *where* $\tilde{f}(\cdot, \tilde{\theta}_k)$ *and* N_k *are defined as in* MRAS_1, *and* $X_1^k, \ldots, X_{N_k}^k$ *are i.i.d. with common density* $\tilde{f}(\cdot, \tilde{\theta}_k)$. *Then, the distance from* $\tilde{\gamma}_k^\dagger(\rho^\dagger, N_k)$ *to* $\gamma_k^\dagger$ *tends to zero as* $k \to \infty$ *w.p.*1.

Convergence Analysis

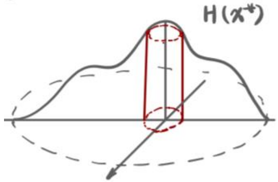
THEOREM 3. *Let* $\varepsilon > 0$, *and define the* ε*-optimal set* $\mathcal{O}_\varepsilon := \{x: H(x) \geqslant H(x^*) - \varepsilon\} \cap \mathcal{X}$. *If Assumptions* A1, A3′, A4′, *and* A5 *are satisfied, then there exists a random variable* $\mathcal{K}$ *such that w.p.*1., $\mathcal{K} < \infty$, *and*

(1) $\bar{\gamma}_k > H(x^*) - \varepsilon \; \forall k \geqslant \mathcal{K}$.

(2) $E_{\tilde{\theta}_{k+1}}[\Gamma(X)] \in \mathrm{CONV}\{\Gamma(\mathcal{O}_\varepsilon)\} \;\; \forall k \geqslant \mathcal{K}$, *where* $\mathrm{CONV}\{\Gamma(\mathcal{O}_\varepsilon)\}$ *indicates the convex hull of the set* $\Gamma(\mathcal{O}_\varepsilon)$.

Furthermore, let β *be a positive constant satisfying the condition that the set* $\{x: S(H(x)) \geqslant 1/\beta\}$ *has a strictly positive Lebesgue/counting measure. If Assumptions* A1, A2, A3′, A4′, *and* A5 *are all satisfied and* $\alpha > (\beta S^*)^2$, *where* $S^* := S(H(x^*))$, *then*

(3) $\lim_{k\to\infty} E_{\tilde{\theta}_k}[\Gamma(X)] = \Gamma(x^*)$ *w.p.*1.



REMARK 4. Roughly speaking, the second result can be understood as finite time ε-optimality. To see this, consider the special case where $H(x)$ is locally concave on the set $\mathscr{O}_\varepsilon$. Let $x, y \in \mathscr{O}_\varepsilon$ and $\eta \in [0, 1]$ be arbitrary. By the definition of concavity, we will have $H(\eta x + (1-\eta)y) \geqslant \eta H(x) + (1-\eta)H(y) \geqslant H(x^*) - \varepsilon$, which implies that the set $\mathscr{O}_\varepsilon$ is convex. If in addition $\Gamma(x)$ is also convex and one-to-one on $\mathscr{O}_\varepsilon$ (e.g., multivariate normal p.d.f.), then $\text{CONV}\{\Gamma(\mathscr{O}_\varepsilon)\} = \Gamma(\mathscr{O}_\varepsilon)$. Thus, it follows that $\Gamma^{-1}(E_{\tilde{\theta}_{k+1}}[\Gamma(X)]) \in \mathscr{O}_\varepsilon \ \forall k \geqslant \mathscr{K}$ w.p.1.

COROLLARY 3 (MULTIVARIATE NORMAL). *For continuous optimization problems in* $\Re^n$*, if multivariate normal p.d.f.'s are used in* MRAS_1*, i.e.,*

$$f(x, \tilde{\theta}_k) = \frac{1}{\sqrt{(2\pi)^n |\widetilde{\Sigma}_k|}} \exp\left(-\tfrac{1}{2}(x - \tilde{\mu}_k)^T \widetilde{\Sigma}_k^{-1} (x - \tilde{\mu}_k)\right),$$

$\varepsilon > 0$, $\alpha > (\beta S^*)^2$, *and Assumptions* A1, A2, A3′, *and* A4′ *are satisfied, then*

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

COROLLARY 4 (INDEPENDENT UNIVARIATE). *If the components of the random vector* $X = (X_1, X_2, \ldots, X_n)$ *are independent, each with a univariate p.d.f./p.m.f. of the form*

$$f(x_i, \vartheta_i) = \exp(x_i\vartheta_i - K(\vartheta_i))h(x_i), \quad \vartheta_i \in \Re \ \ \forall i = 1, \ldots, n,$$

$\varepsilon > 0$, $\alpha > (\beta S^*)^2$, *and Assumptions* A1, A2, A3′, A4′, *and* A5 *are satisfied, then*

$$\lim_{k\to\infty} E_{\tilde{\theta}_k}[X] = x^* \quad \textit{w.p.}1, \quad \textit{where } \tilde{\theta}_k := (\vartheta_1^k, \ldots, \vartheta_n^k).$$

Proof of Theorem 1

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

$$g_{k+1}(x) := \frac{S(H(x))I_{\{H(x)\geqslant\bar{\gamma}_{k+1}\}}g_k(x)}{E_{g_k}[S(H(X))I_{\{H(X)\geqslant\bar{\gamma}_{k+1}\}}]} \quad \forall x\in\mathscr{X}, k=1,2,\ldots.$$

$$E_{g_{k+1}}[S(H(X))I_{\{H(X)\geqslant\bar{\gamma}_{k+1}\}}]$$

$$= \int \left[S(H(x))I_{\{\ldots\}}\right] g_{k+1} dx = \int \left[S(H(x))I_{\{\ldots\}}\right] \cdot \frac{S(H(x))I_{\{\ldots\}} \cdot g_k(x)}{E_{g_k}\left[S(H(x))I_{\{\ldots\}}\right]} dx$$

$$= \int \frac{\left[S(H(x))\right]^2 \cdot I_{\{\ldots\}}}{E_{g_k}[S(H(x))] \cdot I_{\{\ldots\}}} \cdot g_k(x) dx$$

$$= \frac{E_{g_k}[[S(H(X))]^2 I_{\{H(X)\geqslant\bar{\gamma}_{k+1}\}}]}{E_{g_k}[S(H(X))I_{\{H(X)\geqslant\bar{\gamma}_{k+1}\}}]}$$

$$\geqslant E_{g_k}[S(H(X))I_{\{H(X)\geqslant\bar{\gamma}_{k+1}\}}].$$

Convergence Proof

two cases: $\bar{\gamma}_{\mathcal{N}} = H(x^*)$ and $\bar{\gamma}_{\mathcal{N}} < H(x^*)$.

Case 1. If $\bar{\gamma}_{\mathcal{N}} = H(x^*)$

$$g_{k+1}(x) = 0 \quad \forall x \neq x^*$$

and

$$g_{k+1}(x^*) = \frac{[S(H(x^*))]^k I_{\{H(x)=H(x^*)\}}}{\int_{\mathscr{X}} [S(H(x))]^k I_{\{H(x)=H(x^*)\}} \, \nu(dx)} = 1 \quad \forall k \geqslant \mathcal{N}.$$

Hence, it follows immediately that

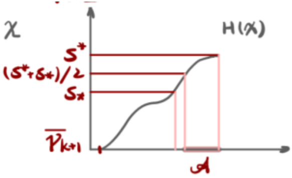
$$E_{g_{k+1}}[\Gamma(X)] = \Gamma(x^*) \quad \forall k \geqslant \mathcal{N}.$$

Case 2. If $\bar{\gamma}_{\mathcal{N}} < H(x^*)$

$$E_{g_{k+1}}[S(H(X))I_{\{H(X)\geqslant\bar{\gamma}_{k+2}\}}] \geqslant E_{g_k}[S(H(X))I_{\{H(X)\geqslant\bar{\gamma}_{k+1}\}}] \quad \forall k \geqslant \mathcal{N}-1, \qquad (19)$$

i.e., the sequence $\{E_{g_k}[S(H(X))I_{\{H(X)\geqslant\bar{\gamma}_{k+1}\}}],\ k=1,2,\ldots\}$ converges.

$$S_* := \lim_{k\to\infty} E_{g_k}[S(H(X))I_{\{H(X)\geqslant\bar{\gamma}_{k+1}\}}] < S^* := S(H(x^*)). \qquad (20)$$



$$\mathscr{A} := \{x\colon H(x) \geqslant \bar{\gamma}_{\mathcal{N}}\} \cap \{x\colon S(H(x)) \geqslant (S^* + S_*)/2\} \cap \mathscr{X}.$$

$$\mathscr{A} = \{x\colon H(x) \geqslant \max\{\bar{\gamma}_{\mathcal{N}}, S^{-1}((S^* + S_*)/2)\}\} \cap \mathscr{X}.$$

$$g_k(x) = \prod_{i=1}^{k-1} \frac{S(H(x))I_{\{H(x)\geqslant\bar{\gamma}_{i+1}\}}}{E_{g_i}[S(H(X))I_{\{H(X)\geqslant\bar{\gamma}_{i+1}\}}]} \cdot g_1(x).$$

Convergence Proof

Because

$$\lim_{k\to\infty} \frac{S(H(x))I_{\{H(x)\geqslant\bar{\gamma}_{k+1}\}}}{E_{g_k}[S(H(X))I_{\{H(X)\geqslant\bar{\gamma}_{k+1}\}}]} = \frac{S(H(x))I_{\{H(x)\geqslant\bar{\gamma}_{\mathcal{N}}\}}}{S_*} > 1$$

$$\forall x \in \mathscr{A},$$

we conclude that

$$\lim_{k\to\infty} g_k(x) = \infty \quad \forall x \in \mathscr{A}.$$

Thus, by Fatou's lemma, we have

$$1 = \liminf_{k\to\infty} \int_{\mathscr{X}} g_k(x)\,\nu(dx) \geqslant \liminf_{k\to\infty} \int_{\mathscr{A}} g_k(x)\,\nu(dx)$$

$$\geqslant \int_{\mathscr{A}} \liminf_{k\to\infty} g_k(x)\,\nu(dx) = \infty,$$

which is a contradiction. Hence, it follows that

$$\lim_{k\to\infty} E_{g_k}[S(H(X))I_{\{H(X)\geqslant\bar{\gamma}_{k+1}\}}] = S^*. \tag{21}$$

Convergence Proof

$$\|E_{g_k}[\Gamma(X)] - \Gamma(x^*)\|$$
$$\leqslant \int_{\mathscr{X}} \|\Gamma(x) - \Gamma(x^*)\| g_k(x)\, \nu(dx)$$
$$= \int_{\mathscr{C}} \|\Gamma(x) - \Gamma(x^*)\| g_k(x)\, \nu(dx), \tag{22}$$

where $\mathscr{C} := \{x\colon H(x) \geqslant \bar{\gamma}_{\mathscr{N}}\} \cap \mathscr{X}$ is the support of $g_k(\cdot)$ $\forall k \geqslant \mathscr{N}$.

By the assumption on $\Gamma(\cdot)$ in Definition 1, for any given $\zeta > 0$, there exists a $\delta > 0$ such that $\|x - x^*\| < \delta$ implies $\|\Gamma(x) - \Gamma(x^*)\| < \zeta$. With A_δ defined from Assumption A2, we have from (22),

$$\|E_{g_k}[\Gamma(X)] - \Gamma(x^*)\|$$
$$\leqslant \int_{A_\delta^c \cap \mathscr{C}} \|\Gamma(x) - \Gamma(x^*)\| g_k(x)\, \nu(dx)$$
$$+ \int_{A_\delta \cap \mathscr{C}} \|\Gamma(x) - \Gamma(x^*)\| g_k(x)\, \nu(dx)$$
$$\leqslant \zeta + \int_{A_\delta \cap \mathscr{C}} \|\Gamma(x) - \Gamma(x^*)\| g_k(x)\, \nu(dx) \quad \forall k \geqslant \mathscr{N}. \tag{23}$$

The rest of the proof amounts to showing that the second term in (23) is also bounded. Clearly, the term $\|\Gamma(x) - \Gamma(x^*)\|$ is bounded on the set $A_\delta \cap \mathscr{C}$. We only need to find a bound for $g_k(x)$.

Define $S_\delta := S^* - S(\sup_{x\in A_\delta} H(x))$. Because $S(\cdot)$ is strictly increasing, we have $S_\delta > 0$. Thus, it follows that

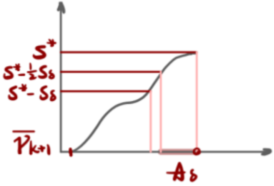
$$S(H(x)) \leqslant S^* - S_\delta \quad \forall x \in A_\delta \cap \mathscr{C}. \tag{24}$$

On the other hand, from (19) and (21), there exists $\bar{\mathcal{N}} \geqslant \mathcal{N}$ such that $\forall k \geqslant \bar{\mathcal{N}}$,

$$E_{g_k}[S(H(X))I_{\{H(X)\geqslant\bar{\gamma}_{k+1}\}}] \geqslant S^* - \tfrac{1}{2}S_\delta. \tag{25}$$

$$g_k(x) = \prod_{i=\bar{\mathcal{N}}}^{k-1} \frac{S(H(x))I_{\{H(x)\geqslant\bar{\gamma}_{i+1}\}}}{E_{g_i}[S(H(X))I_{\{H(X)\geqslant\bar{\gamma}_{i+1}\}}]} \cdot g_{\bar{\mathcal{N}}}(x) \quad \forall k \geqslant \bar{\mathcal{N}}.$$

$$g_k(x) \leqslant \left(\frac{S^* - S_\delta}{S^* - S_\delta/2}\right)^{k-\bar{\mathcal{N}}} \cdot g_{\bar{\mathcal{N}}}(x) \quad \forall x \in A_\delta \cap \mathscr{C}, \forall k \geqslant \bar{\mathcal{N}}.$$



Convergence Proof

$$
\begin{aligned}
&\|E_{g_k}[\Gamma(X)] - \Gamma(x^*)\| \\
&\leqslant \zeta + \sup_{x \in A_\delta \cap \mathscr{C}} \|\Gamma(x) - \Gamma(x^*)\| \int_{A_\delta \cap \mathscr{C}} g_k(x)\, \nu(dx) \\
&\leqslant \zeta + \sup_{x \in A_\delta \cap \mathscr{C}} \|\Gamma(x) - \Gamma(x^*)\| \left(\frac{S^* - S_\delta}{S^* - S_\delta/2} \right)^{k - \bar{\mathcal{N}}} \quad \forall k \geqslant \bar{\mathcal{N}} \\
&= \Big(1 + \sup_{x \in A_\delta \cap \mathscr{C}} \|\Gamma(x) - \Gamma(x^*)\| \Big) \zeta \quad \forall k \geqslant \widehat{\mathcal{N}},
\end{aligned}
$$

where $\widehat{\mathcal{N}}$ is given by

$$
\widehat{\mathcal{N}} := \max\left\{ \bar{\mathcal{N}}, \left\lceil \bar{\mathcal{N}} + \ln \zeta \Big/ \ln\left(\frac{S^* - S_\delta}{S^* - S_\delta/2} \right) \right\rceil \right\}.
$$

Because ζ is arbitrary, we have

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

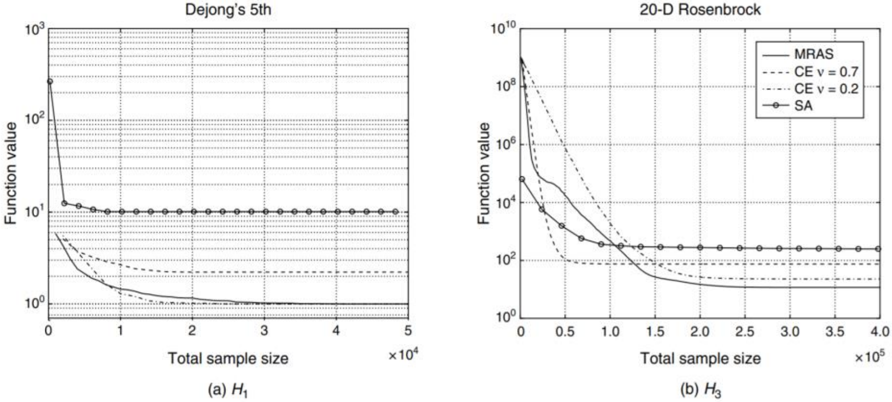
Numerical Examples

Table 1. Performance of different algorithms on benchmark problems $H_1 - H_7$ based on 100 independent replications (standard errors are in parentheses).

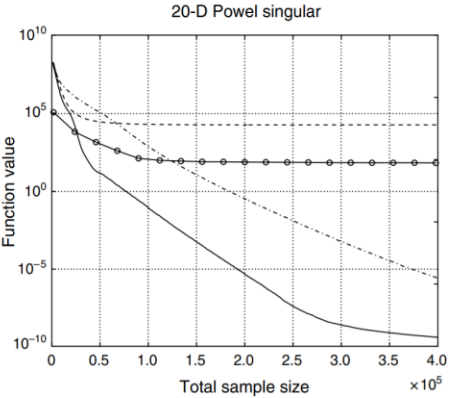
Test problems	MRAS$_1$		CE ($v=0.7$)		CE ($v=0.2$)		SA	
	$\bar{H}_i^*$	M_ε	$\bar{H}_i^*$	M_ε	$\bar{H}_i^*$	M_ε	$\bar{H}_i^*$	M_ε
H_1	0.998 (3.8e−07)	100	2.22 (0.23)	61	0.998 (4.3e−09)	100	10.12 (0.92)	12
H_2	−10.15 (6.6e−07)	100	−8.38 (0.30)	72	−9.12 (0.11)	1	−6.62 (0.35)	1
H_3	11.64 (5.4e−02)	0	74.68 (19.30)	0	22.63 (4.86)	0	248.5 (23.59)	0
H_4	3.2e − 10 (1.8e − 11)	100	1.9e + 04 (2.8e + 03)	0	2.5e − 06 (7.5e − 08)	100	68.19 (2.94)	0
H_5	1.45 (6.4e−02)	47	1.00 (00e−00)	100	1.00 (4.6e−09)	100	75.69 (4.94)	0
H_6	4.7e − 03 (5.8e − 04)	55	1.5e − 04 (1.0e − 04)	98	2.2e − 04 (1.3e − 04)	97	0.12 (9.7e−03)	0
H_7	4.9e − 08 (7.1e − 09)	100	4.75 (1.07)	0	2.1e − 03 (7.5e − 05)	0	1.1e + 03 (93.4)	0

Numerical Examples

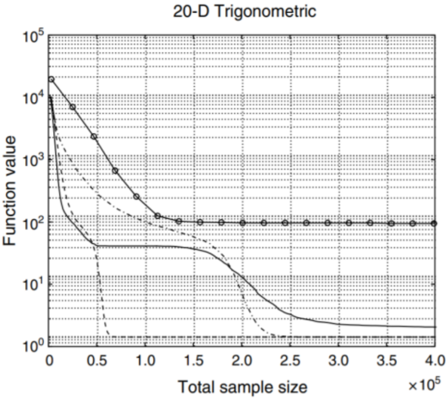
Figure 2. Average performance of $MRAS_1$, CE, and SA on selected benchmark problems.



Numerical Examples

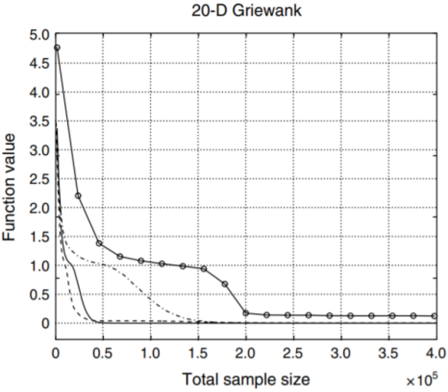


(c) H_4

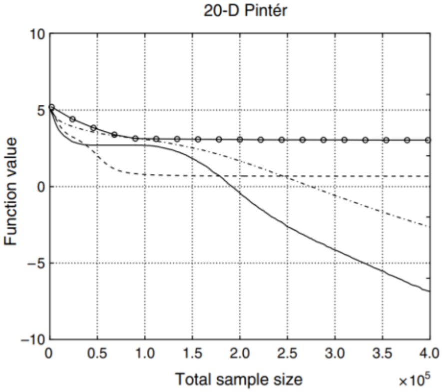


(d) H_5

Numerical Examples



(e) H_6



(f) H_7

Thank you!