Computers and Structures 126 (2013) 107–119

Contents lists available at SciVerse ScienceDirect

Computers and Structures

journal homepage: www.elsevier.com/locate/compstruc

Global optimization using the asymptotically independent Markov sampling method

Konstantin M. Zuev [a,*], James L. Beck [b]

[a] Department of Mathematics, University of Southern California, Los Angeles, CA, United States
[b] Division of Engineering and Applied Science, California Institute of Technology, United States

ARTICLE INFO

Article history:
Received 23 August 2012
Accepted 1 April 2013
Available online 29 April 2013

Keywords:
Global optimization
Advanced simulation techniques
Markov chain Monte Carlo
Importance sampling
Simulated annealing

ABSTRACT

In this paper, we introduce a new efficient stochastic simulation method, AIMS-OPT, for approximating the set of globally optimal solutions when solving optimization problems such as optimal performance-based design problems. This method is based on Asymptotically Independent Markov Sampling (AIMS), a recently developed advanced simulation scheme originally proposed for Bayesian inference. This scheme combines importance sampling, Markov chain Monte Carlo simulation and annealing for efficient sampling from an arbitrary target distribution over a multi-dimensional space. Instead of a single approximation of the optimal solution, AIMS-OPT produces a set of nearly optimal solutions where the accuracy of the near-optimality is controlled by the user. Having a set of nearly optimal system designs, for example, can be advantageous in many practical cases such as when there exists a whole set of optimal designs or in multi-objective optimization where there is a Pareto optimal set. AIMS-OPT is also useful for efficient exploration of the global sensitivity of the objective function to the design parameters. The efficiency of AIMS-OPT is demonstrated with several examples which have different topologies of the optimal solution sets. Comparison is made with the results of applying Simulated Annealing, a well-known stochastic optimization algorithm, to the three two-dimensional problems.

1. Introduction

The ultimate goal of engineering is to design technological systems that satisfy specified performance objectives and constraints over a certain period of time. Usually there are many feasible designs that satisfy the specifications and it is desirable to choose an optimal one according to some criterion such as minimizing life-cycle cost.

Consider the general problem of optimal system design under uncertainty. Suppose that various possibilities for the design of a system are defined by some controllable *design parameters* $\varphi \in \Phi \subset \mathbb{R}^{N_\varphi}$, where Φ denotes the bounded space of possible designs. Assume that a single model class [4,5] is chosen to represent the uncertain system behavior and its uncertain future excitation, where each probability model in the class is specified by the *model parameters* $\theta \in \Theta \subset \mathbb{R}^{N_\theta}$. Since there is uncertainty in which model best describes the real system behavior, a probability density function $\pi(\theta|\varphi)$, which incorporates available knowledge about the system, is assigned to the model parameters. The selection of $\pi(\theta|\varphi)$ should be based on the prior information about the system behavior that may come from the underlying physical theory, past experimental data, observation of similar systems, and expert opinion. The *principle of maximum entropy*, introduced by Jaynes [15], provides a rational approach for the specification of prior distributions. This principle states that the prior distribution $\pi(\theta|\varphi)$ which should be taken to represent the prior state of knowledge is the one that gives the largest measure of uncertainty (largest Shannon entropy) about the values of the model parameters θ, subject to desired specified constraints. Furthermore, if new data $\mathcal{D}$ become available about the behavior of the real system, then this additional information can be used for updating the prior distribution $\pi(\theta|\varphi)$ using Bayes' theorem to get the posterior distribution $\pi(\theta|\varphi,\mathcal{D})$ [4].

Let $h : \mathbb{R}^{N_\varphi} \times \mathbb{R}^{N_\theta} \to \mathbb{R}$ denote the performance measure of the system, e.g., a utility or loss function. In *performance-based design optimization* (PBDO), the goal is then to find the optimal design $\varphi^\star$ that minimizes (maximizes) the expected loss (utility) function

$$\mathcal{H}(\varphi) = \mathbb{E}_\pi[h(\varphi,\theta)] = \int_\Theta h(\varphi,\theta)\pi(\theta|\varphi)d\theta, \tag{1}$$

where $\mathbb{E}_\pi[\cdot]$ denotes the expectation with respect to the distribution $\pi(\theta|\varphi)$ for θ. For example, in reliability-based design optimization (RBDO) [9,10,34,16,17], the performance measure is $h(\varphi,\theta) = I_F(\varphi,\theta)$, where $I_F(\varphi,\theta)$ is the indicator function of the failure domain $F \subset \mathbb{R}^{N_\varphi} \times \mathbb{R}^{N_\theta}$: $I_F(\varphi,\theta) = 1$ if the system model corresponding to θ and φ fails (i.e., the model output is not acceptable according to the performance criteria) and $I_F(\varphi,\theta) = 0$ otherwise. In this case,

* Corresponding author.
E-mail addresses: kzuev@usc.edu (K.M. Zuev), jimbeck@caltech.edu (J.L. Beck).

0045-7949/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.compstruc.2013.04.005

the objective function for minimization is the probability of failure, i.e., $\mathcal{H}(\varphi) = \mathbb{P}(F|\varphi)$.

The performance will depend on the input (excitation) to the system which may be specified either deterministically or probabilistically by, for example, a stochastic process; in the latter case, this will introduce additional uncertain parameters and for brevity in notation, we include them in the model parameters θ. Also, from now on, we assume for definiteness that the performance measure is interpreted as a loss function, i.e., lower values of $h(\varphi,\theta)$ correspond to better performance. The *performance-based design optimization problem* takes then the following form for the *set* of optimal designs:

$$\Phi^\star = \{\varphi^\star\} = \arg\min_{\varphi\in\Phi} \mathbb{E}_\pi[h(\varphi,\theta)] = \arg\min_{\varphi\in\Phi} \mathcal{H}(\varphi). \tag{2}$$

The optimization problem (2) arises in diverse applications, not only optimal system design under uncertainty. For example, it arises in the context of decision making under uncertainty where an optimal action is to be selected from a set of possibilities by minimizing the expected loss, then it is often referred to as a stochastic optimization problem or stochastic programming problem. Problem (2) also arises in deterministic optimal design where uncertainties are ignored.

For optimal design under uncertainty, the integral in (1) must be evaluated. For complex systems, this integral can rarely be calculated analytically. Moreover, the dimension N_θ of the model parameter space Θ is typically very large so the usual numerical quadrature methods are not computationally feasible for evaluating (1). Thus, the objective function $\mathcal{H}(\varphi)$ can be estimated only by stochastic simulation techniques. Unavailability of the exact form of the objective function means, in particular, that no analytical properties of $\mathcal{H}(\varphi)$—such as convexity, boundedness, smoothness—are typically known. In this case, stochastic rather than deterministic optimization methods are preferable.

Solving the optimization problem (2) involves finding the global minimum of $\mathcal{H}(\varphi)$ which is well-known to be very challenging, especially when there may be multiple optimal solutions. Various global optimization algorithms have been devised, including Genetic Algorithm [14], Simulated Annealing [21], Particle Swarm Optimization [8,20], Ant Colony Optimization [27], Generalized Trajectory Methods [38], Annealing Evolutionary Stochastic Approximation Monte Carlo [24], etc. A survey of computational methods in optimization under uncertainties is given in [32].

In this paper, we introduce a new efficient stochastic simulation method for solving the optimization problem (2) that can handle multiple optimal solutions and that was motivated by optimal performance-based design under uncertainty, although it can handle general objective functions $\mathcal{H}(\varphi)$, including those that are specified deterministically. This method is based on Asymptotically Independent Markov Sampling (AIMS), a recently developed advanced simulation scheme originally proposed for Bayesian inference [6]. Instead of a single approximation $\hat{\varphi} \approx \varphi^\star$, the new method, denoted AIMS-OPT, produces a set $\{\hat{\varphi}_1,\ldots,\hat{\varphi}_n\}$ of nearly optimal solutions in the same spirit as Stochastic Subset Optimization [34,35]. Having a set of nearly optimal solutions can be advantageous in many practical situations. First, the solution of problem (2) may not be unique, i.e., there may exist a finite or infinite set $\Phi^\star \subset \Phi$ of solutions, then optimization methods that produce only a single point estimate $\hat{\varphi} \approx \varphi^\star \in \Phi^\star$ would not provide a complete picture of the optimal solution set $\Phi^\star$. Furthermore, there may be a set of nearly optimal solutions in Φ whose objective function values differ by an inconsequential amount from the minimum value $\mathcal{H}(\varphi^\star)$, $\varphi^\star \in \Phi^\star$, that may be of interest for reasons not quantified by the objective function $\mathcal{H}(\varphi)$.

AIMS-OPT generates a set of nearly optimal solutions $\{\hat{\varphi}_1,\ldots,\hat{\varphi}_n\} \subset \Phi_T^\star$ approximately uniformly distributed over a neighborhood $\Phi_T^\star$ of the optimal solution set, $\Phi^\star \subset \Phi_T^\star$, where T is a user-specified parameter controlling the size of $\Phi_T^\star$ ($\lim_{T\to 0}\Phi_T^\star = \Phi^\star$), and therefore defines the meaning of "nearly optimal". AIMS-OPT is also useful for efficient exploration of the global sensitivity of the objective function to the parameters φ: at each stage k, the algorithm generates a set $\left\{\varphi_1^{(k)},\ldots,\varphi_n^{(k)}\right\} \subset \Phi_{T_k}^\star$ that converges to the set of optimal solutions as $k\to\infty$: $\Phi \equiv \Phi_{T_0}^\star \supset \Phi_{T_1}^\star \supset \cdots \supset \Phi_{T_k}^\star \supset \cdots \supset \Phi^\star$. The efficiency of AIMS-OPT depends on the desired accuracy of estimation of the objective function $\mathcal{H}(\varphi)$ when it is not directly available. This accuracy can be reduced to improve computational efficiency but at the expense of not being able to get a set of nearly optimal solutions clustered close to $\Phi^\star$ (i.e., T cannot be made too small).

We note that having a set of nearly optimal designs with respect to a performance-based design objective function $\mathcal{H}(\varphi)$ is very useful in multi-objective optimization (or "pareto optimization"), especially when it is difficult to mathematically define what it means for a design to be optimal with respect to other objectives. The latter often happens when difficult-to-formalize qualities are under consideration such as "aesthetics", "beauty", and "style".

Note also that AIMS-OPT can be used for solving constrained optimization problems, where, in addition to problem (2), inequality constraints $g_i(\varphi) \leqslant 0$, $i = 1,\ldots,m$ are imposed. There are several standard techniques for handling constraints in stochastic optimization algorithms. These techniques include the rejection method, where only the feasible solutions are kept in the search process and infeasible solutions are discarded; the penalty function method, where a constrained problem is transformed into an unconstrained one through the incorporation of a penalty function into the objective function; the multi-objective function method, where a single objective optimization problem is converted into a multi-objective one and then multi-objective techniques are applied to solve it. For a more detailed discussion of the constraint handling techniques, we refer the reader to [23]. These techniques can be used within AIMS-OPT for solving constrained optimization problems. In this paper, we focus on solving the unconstrained optimization problem (2).

The rest of the paper is organized as follows. In Section 2, evaluation of the objective function of the form in (1) is discussed. The two special cases of the Metropolis–Hastings algorithm, Independent Metropolis–Hastings and Random Walk Metropolis–Hastings, that lie at the heart of AIMS-OPT, are briefly reviewed in Section 3. In Section 4, the AIMS-OPT method is introduced. The efficiency of AIMS-OPT is illustrated in Section 5 with several examples involving different topologies for the optimal solution set. Concluding remarks are made in Section 6.

2. Evaluation of the objective function

There are two different types of stochastic simulation methods, called in [12] a "many-samples" method and a "single-sample" method, for evaluation of the objective function of the form in (1). In the first method, *new* Monte Carlo samples are obtained for each evaluation of the integral in (1); that is, whenever we want to estimate $\mathcal{H}(\varphi)$ for some $\varphi \in \Phi$, we generate new samples $\theta_1,\ldots,\theta_N$ from $\pi(\theta|\varphi)$ and use the following Monte Carlo estimator:

$$\widehat{\mathcal{H}}_N(\varphi) = \frac{1}{N}\sum_{i=1}^{N} h(\varphi,\theta_i) \approx \mathcal{H}(\varphi). \tag{3}$$

Note that, although it is preferable in term of efficiency, the samples $\theta_1,\ldots,\theta_N$ need not be independent; they may be MCMC samples generated by an appropriate algorithm (see Section 3). In either case, $\widehat{\mathcal{H}}_N(\varphi) \to \mathcal{H}(\varphi)$ as $N\to\infty$.

The single-sample method involves *importance sampling*. Let $\nu(\theta)$ be a distribution on Θ which does not depend on φ. Then

the objective function can be written as an expectation with respect to $v(\theta)$ as follows:

$$\mathcal{H}(\varphi) = \int_{\Theta} \frac{h(\varphi,\theta)\pi(\theta|\varphi)}{v(\theta)} v(\theta) d\theta = \mathbb{E}_v\left[\frac{h(\varphi,\theta)\pi(\theta|\varphi)}{v(\theta)}\right]. \quad (4)$$

Therefore, if $\theta_1 \ldots, \theta_N$ are samples from $v(\theta)$, then $\mathcal{H}(\varphi)$ can be estimated by

$$\mathcal{H}'_N(\varphi) = \frac{\sum_{i=1}^N \omega_i h(\varphi,\theta_i)}{\sum_{i=1}^N \omega_i} \approx \mathcal{H}(\varphi), \quad (5)$$

where $\omega_i = \pi(\theta_i|\varphi)/v(\theta_i)$ is the importance weight of the ith sample. The estimator $\mathcal{H}'_N(\varphi)$ converges almost surely as $N \to \infty$ to $\mathcal{H}(\varphi)$ by the Strong Law of Large Numbers for any choice of distribution $v(\theta)$, provided supp $(\pi) \subseteq$ supp (v). Note that in the single-sample method, the *same* samples $\theta_1, \ldots, \theta_N \sim v(\theta)$ are used for estimation of $\mathcal{H}(\varphi)$ for different values of φ.

Remark 1. Note that "many (single) samples method" is related to the concept of "interior (exterior)" sampling that is also used for stochastic optimization problems [31,33], with the only difference stemming from the fact that the distribution $\pi(\theta|\varphi)$ is assumed to be dependent on φ.

In terms of computational cost, the single-sample method and the many-samples method are equivalent: both need $N \times n$ performance measure evaluations for estimation of n different values of $\mathcal{H}(\varphi_j)$, $j = 1, \ldots, n$. The main advantage of the single-sample method is that it produces smooth estimates of $\mathcal{H}(\varphi)$ if $h(\varphi,\theta)$ is sufficiently smooth in φ. Note, however, that this property is not important for AIMS-OPT. On the other hand, the single-sample method can be very inefficient and produce extremely poor estimates when the *importance sampling density* (ISD) $v(\theta)$ is chosen without considerable care. Note that the convergence of $\mathcal{H}'_N(\varphi)$ to $\mathcal{H}(\varphi)$ is only theoretical: it may require an enormous number of simulations N to produce an accurate estimate of $\mathcal{H}(\varphi)$. It has been recognized by many researchers that, in general, construction of a good importance sampling density in high dimensions is a very challenging problem [26,3,19]. For importance sampling to work efficiently, the estimator in (5) must have low variance so v must be a good approximation of the integrand in (1). However, this integrand may depend sensitively on the parameters φ so that it is impossible to find $v(\theta)$ that is a good approximation of the integrand for all $\varphi \in \Phi$. This limits the applicability of the single-sample method in the context of the optimal performance-based design problem. For applying AIMS-OPT we adopt the many-samples method for evaluation of the objective function $\mathcal{H}(\varphi)$ of the form in (1).

Remark 2. Using a different approach to the single-sample and many-samples methods, Taflanidis and Beck [34,35] have introduced a very efficient stochastic optimization method called Stochastic Subset Optimization that uses stochastic simulation to produce a nested sequence of decreasing subsets of near-optimal solutions without explicitly evaluating the objective function in (1).

3. MCMC sampling

Markov chain Monte Carlo (MCMC), a family of stochastic simulation algorithms for sampling from arbitrary probability distributions, lies at the heart of AIMS-OPT. These algorithms are based on constructing a Markov chain whose state probability distribution converges to any desired target distribution as its stationary distribution. MCMC sampling originated in statistical physics, and now is widely used for solving problems in statistics and optimization [29,26,30].

The Metropolis–Hastings algorithm [28,13], the most popular MCMC technique, works as follows. Suppose we want to generate samples from a probability distribution $p(\varphi)$ on Φ. Let $q(\xi|\varphi)$ be a distribution for $\xi \in \Phi$, which may or may not depend on $\varphi \in \Phi$. Assume that $q(\xi|\varphi)$ is easy to sample from and it is either computable (up to a multiplicative constant) or symmetric, i.e., $q(\xi|\varphi) = q(\varphi|\xi)$. The sampling distribution $q(\xi|\varphi)$ is called the *proposal distribution*. Starting from essentially any $\varphi_1 \in \Phi$, the Metropolis–Hastings algorithm proceeds by iterating the following two steps.

1. Generate a *candidate* state ξ from the proposal density $q(\xi|\varphi_j)$.
2. Either accept ξ as the next state of the Markov chain, $\varphi_{j+1} = \xi$, with probability

$$\alpha(\xi|\varphi_j) = \min\left\{1, \frac{p(\xi)q(\varphi_j|\xi)}{p(\varphi_j)q(\xi|\varphi_j)}\right\}, \quad (6)$$

or reject ξ and set $\varphi_{j+1} = \varphi_j$ with the remaining probability $1 - \alpha(\xi|\varphi_j)$.

It can be shown (see, for example, [30]), that under fairly weak conditions, $p(\varphi)$ is the stationary distribution of the Markov chain $\varphi_1, \varphi_2, \ldots$, i.e., the distribution of φ_j converges to $p(\varphi)$ as $j \to \infty$.

The two main special cases of the Metropolis–Hastings algorithm are Independent Metropolis–Hastings (IMH), where the proposal distribution $q(\xi|\varphi) = q_g(\xi)$ is independent of φ (so q_g is a *global proposal*), and Random Walk Metropolis–Hastings (RWMH), where the proposal distribution is of the form $q(\xi|\varphi) = q_l(\xi - \varphi)$, i.e., a candidate state is proposed as $\xi = \varphi_j + \varepsilon_j$, where $\varepsilon_j \sim q_l$ is a random perturbation (so q_l is a *local proposal*). In both cases, the choice of the proposal distribution strongly affects the efficiency of the algorithms. For IMH to work well, as with importance sampling, the proposal distribution must be a good approximation of the *target distribution* $p(\varphi)$, otherwise a large fraction of the candidate samples will be rejected and the Markov chain will be too slow in covering the important regions of $p(\varphi)$. When, however, it is possible to find a proposal q_g such that $q_g \approx p$, IMH should always be preferred to RWMH because of better efficiency. Unfortunately, such a proposal is usually difficult to construct when the target distribution $p(\varphi)$ is complex. This limits the applicability of IMH.

Since the random walk proposal q_l is local, it is less sensitive to the target distribution. That is why, in practice, RWMH is more robust and used more frequently than IMH. Nonetheless, there are settings where RWMH also does not work well because of the complexity of the target distribution. Although convergence of the Markov chain $\varphi_1, \varphi_2, \ldots$ to its stationary distribution $p(\varphi)$ is true in theory, in practice, it is often very difficult to check whether the chain has reached its steady-state or not. A potential problem with RWMH (and, in fact, with almost all MCMC algorithms) is that the period of simulation to stationarity, commonly referred to as *burn-in period*, can be very long. This is often the case when the target distribution $p(\varphi)$ contains several widely-separated modes: a chain will move between modes only rarely and it will take a long time before it reaches stationarity.

4. Asymptotically independent Markov sampling for optimization

Recently a new advanced stochastic simulation scheme, called Asymptotically Independent Markov Sampling (AIMS), was developed for computational Bayesian inference [6]. This scheme efficiently combines importance sampling, Markov chain Monte Carlo, and annealing for sampling from any target distribution. In this paper, we extend the applications of AIMS to global optimization problems by introducing AIMS-OPT for solving the optimization problem (2).

The starting point of AIMS-OPT is the concept of *annealing* (or *tempering*), which is based on the following simple but important observation: finding the global minimum of the objective function $\mathcal{H}(\varphi)$ is equivalent to finding the global maximum of $\exp(-\mathcal{H}(\varphi)/T)$ for any given "temperature" $T>0$ [21].

Remark 3. The parameter T is called "temperature" by analogy with the Boltzmann–Gibbs distribution in statistical mechanics.

Let us define a "tempered" distribution on the bounded admissible parameter space Φ as follows:

$$p_T(\varphi) \propto \exp(-\mathcal{H}(\varphi)/T) I_\Phi(\varphi), \tag{7}$$

where $I_\Phi(\varphi)$ is the indicator function of Φ. Note that the tempered PDF (7) becomes flatter as the temperature T increases, i.e., as $p_T(\varphi)$ gets "hotter"; and it becomes spikier as T decreases toward zero, i.e., as $p_T(\varphi)$ gets "cooler". More precisely,

$$\lim_{T\to\infty} p_T(\varphi) = U_\Phi(\varphi) \quad \text{and} \quad \lim_{T\to 0} p_T(\varphi) = U_{\Phi^\star}(\varphi), \tag{8}$$

where $U_\Phi(\varphi)$ and $U_{\Phi^\star}(\varphi)$ are the uniform distributions on the parameter space Φ and the optimal solution set $\Phi^\star$, respectively. $U_{\Phi^\star}(\varphi)$ may be a discrete or continuous distribution. In particular, if the optimization problem (2) has a unique solution, i.e., the optimal solution set $\Phi^\star$ consists only of a single point, $\Phi^\star=\{\varphi^\star\}$, then $\lim_{T\to 0} p_T(\varphi) = \delta_{\varphi^\star}(\varphi)$, where $\delta_{\varphi^\star}(\varphi)$ is the Dirac mass at $\varphi^\star$.

The key idea behind annealing is the following: as the temperature T decreases, the tempered distribution $p_T(\varphi)$ puts more and more of its probability mass (converging to one) into the set of optimal solutions $\Phi^\star$. Therefore, when T is close to zero, a sample drawn from $p_T(\varphi)$ will be in a neighborhood $\Phi_T^\star$ of $\Phi^\star$ with a very high probability. Here, $\Phi_T^\star$ denotes the so-called "practical support" of $p_T(\varphi)$, i.e., the region that contains almost all probability mass of $p_T(\varphi)$ with the property that $\lim_{T\to 0}\Phi_T^\star = \Phi^\star$.

Remark 4. To illustrate the notion of practical support, introduced in [18], consider a sample $z_1,\ldots,z_n$ drawn from the standard Gaussian distribution $\mathcal{N}(0,1)$. It is easy to check that for $n=10^4$ the probability that at least one sample lies outside the interval $[-6,6]$ is smaller than 10^{-4}. Although, strictly speaking, the support of the standard Gaussian distribution is the entire real line (theoretical support), in practice the support may be regarded as just a segment (practical support).

Let $\infty = T_0 > T_1 > \ldots > T_k > \ldots$ be a sequence of monotonically decreasing temperatures with $\lim_{k\to\infty} T_k = 0$, and

$$\begin{aligned} &p_0(\varphi) = U_\Phi(\varphi),\\ &p_k(\varphi) \propto \exp(-\mathcal{H}(\varphi)/T_k) I_\Phi(\varphi), \quad k=1,2,\ldots \end{aligned} \tag{9}$$

be the corresponding sequence of tempered PDFs on Φ. In AIMS-OPT, we sequentially generate samples from the tempered distributions in (9) in the following way. Importance sampling with $p_{k-1}(\varphi)$ as the ISD is used for a construction of an approximation $\hat{p}_{k,n}(\varphi)$ of $p_k(\varphi)$, which is based on samples $\varphi_1^{(k-1)},\ldots,\varphi_n^{(k-1)} \sim p_{k-1}(\varphi)$. This approximation is then employed as the global proposal distribution for sampling from $p_k(\varphi)$ by the IMH algorithm. The tempered distributions in (9) are constructed adaptively, using the essential sample size (ESS) to measure how much $p_{k-1}(\varphi)$ differs from $p_k(\varphi)$ [26,6]. When the number of samples $n\to\infty$, the approximation $\hat{p}_{k,n}(\varphi)$ converges to $p_k(\varphi)$, providing the optimal proposal distribution. In other words, when $n\to\infty$, the corresponding MCMC sampler produces independent samples, which is the reason for "Asymptotically Independent Markov Sampling" in naming the algorithm.

We will refer to k and T_k as the *annealing level* and the *annealing temperature* at level k, respectively. In the next subsection, we assume that T_k is given and therefore the tempered distribution $p_k(\varphi)$ is also known (up to a normalizing constant). In Section 4.2, we describe how to choose the annealing temperatures adaptively.

4.1. AIMS-OPT at annealing level k

First, we describe how AIMS-OPT generates samples $\varphi_1^{(k)},\ldots,\varphi_n^{(k)}$ from $p_k(\varphi)$ based on the samples $\varphi_1^{(k-1)},\ldots,\varphi_n^{(k-1)} \sim p_{k-1}(\varphi)$ obtained at the previous annealing level.

Let $\mathcal{P}_k$ be any Markov transition kernel such that $p_k(\varphi)$ is a stationary distribution with respect to $\mathcal{P}_k$. By definition, this means that

$$p_k(\varphi)d\varphi = \int_\Phi \mathcal{P}_k(d\varphi|\xi) p_k(\xi) d\xi \tag{10}$$

Applying importance sampling with the ISD $p_{k-1}(\varphi)$ to integral (10), we obtain:

$$\begin{aligned} p_k(\varphi)d\varphi &= \int_\Phi \mathcal{P}_k(d\varphi|\xi)\frac{p_k(\xi)}{p_{k-1}(\xi)} p_{k-1}(\xi)d\xi\\ &\approx \sum_{j=1}^n \mathcal{P}_k(d\varphi|\varphi_j^{(k-1)})\bar{\omega}_j^{(k-1)} \stackrel{\text{def}}{=} \hat{p}_{k,n}(d\varphi), \end{aligned} \tag{11}$$

where $\hat{p}_{k,n}(d\varphi)$ will be used as the *global* proposal distribution in the IMH algorithm for sampling from $p_k(\varphi)$, and

$$\begin{aligned} \omega_j^{(k-1)} &= \frac{p_k\left(\varphi_j^{(k-1)}\right)}{p_{k-1}\left(\varphi_j^{(k-1)}\right)} \propto \exp\left[-\mathcal{H}\left(\varphi_j^{(k-1)}\right)\left(\frac{1}{T_k}-\frac{1}{T_{k-1}}\right)\right],\\ \bar{\omega}_j^{(k-1)} &= \frac{\omega_j^{(k-1)}}{\sum_{j=1}^n \omega_j^{(k-1)}},\ j=1,\ldots,n \end{aligned} \tag{12}$$

are the importance weights and normalized importance weights, respectively. Note that to compute $\bar{\omega}_j^{(k-1)}$, we do not need to know the normalizing constants of $p_{k-1}(\varphi)$ and $p_k(\varphi)$. If adjacent tempered distributions $p_{k-1}(\varphi)$ and $p_k(\varphi)$ are sufficiently close (in other words, if the temperature change $\Delta T_k = T_k - T_{k-1}$ is small enough), then the variability of the importance weights (12) will be mild, and, therefore, we can expect that, for reasonably large n, approximation (11) is accurate. A simple illustrative example of approximation (11) is given in [6].

Remark 5. In [7], the stationary condition (10) was used for an analytical approximation of the posterior PDF to evaluate the evidence (marginal likelihood) for a model.

It is important to understand that for any finite n, distribution $\hat{p}_{k,n}(d\varphi)$ will usually have both continuous and discrete parts. This follows from the fact that the transition kernel in Markov chain simulation usually has the following form: $\mathcal{P}(d\varphi|\xi) = f(\varphi|\xi)d\varphi + r(\xi)\delta_\xi(d\varphi)$, where $f(\varphi|\xi)$, the continuous part of $\mathcal{P}(d\varphi|\xi)$, describes transitions from ξ to $\varphi\neq\xi$ and $r(\xi) = 1 - \int_\Phi f(\varphi|\xi)d\varphi$ is the probability that the Markov chain remains at ξ. This is the form, for example, for the Metropolis–Hastings algorithm. Therefore, (11) must be understood as the approximate equality of distributions, not densities. In other words, (11) means that $\mathbb{E}_{\hat{p}_{k,n}}[g] \approx \mathbb{E}_{p_k}[g]$ and $\mathbb{E}_{\hat{p}_{k,n}}[g] \to \mathbb{E}_{p_k}[g]$, when $n\to\infty$, for all integrable functions g.

From now on, we consider a special case where $\mathcal{P}_k$ is the RWMH transition kernel. In this case, it can be written as follows:

$$\mathcal{P}_k(d\varphi|\xi) = q_k(\varphi|\xi)\min\left\{1, \frac{p_k(\varphi)}{p_k(\xi)}\right\}d\varphi + (1-a_k(\xi))\delta_\xi(d\varphi), \tag{13}$$

where $q_k(\varphi|\xi)$ is a symmetric *local* proposal density, and $a_k(\xi)$ is the probability of having a proper transition from ξ to $\Phi\backslash\{\xi\}$:

$$a_k(\xi) = \int_\Phi q_k(\varphi|\xi)\min\left\{1, \frac{p_k(\varphi)}{p_k(\xi)}\right\}d\varphi \tag{14}$$

The approximation $\hat{p}_{k,n}(d\varphi)$ takes then the following form:

$$\hat{p}_{k,n}(d\varphi) = \sum_{j=1}^{n} \bar{\omega}_j^{(k-1)} q_k\left(\varphi|\varphi_j^{(k-1)}\right) \min\left\{1, \frac{p_k(\varphi)}{p_k\left(\varphi_j^{(k-1)}\right)}\right\} d\varphi + \sum_{j=1}^{n} \bar{\omega}_j^{(k-1)} \left(1 - a_k\left(\varphi_j^{(k-1)}\right)\right) \delta_{\varphi_j^{(k-1)}}(d\varphi) \quad (15)$$

For sampling from $p_k(\varphi)$, we will use the Metropolis–Hastings algorithm with the *global* proposal distribution $\hat{p}_{k,n}(d\varphi)$. To accomplish this, we have to be able to compute the ratio $\hat{p}_{k,n}(\varphi)/\hat{p}_{k,n}(\xi)$ for any φ, $\xi \in \Phi$ as a part of the expression for the acceptance probability (6), which, in our case, reduces to

$$\alpha_k(\xi|\varphi) = \min\left\{1, \frac{p_k(\xi)\hat{p}_{k,n}(\varphi)}{p_k(\varphi)\hat{p}_{k,n}(\xi)}\right\} \quad (16)$$

However, as it has been already mentioned, the distribution $\hat{p}_{k,n}(d\varphi)$ does not have a density since it has both continuous and discrete components (the first and the second terms in (15), respectively), and, therefore, the ratio $\hat{p}_{k,n}(\varphi)/\hat{p}_{k,n}(\xi)$ might not make sense. Nevertheless, this technical "lack-of-continuity problem" can be overcome by replacing the sample space Φ with $\Phi \setminus \left\{\varphi_1^{(k-1)}, \ldots, \varphi_n^{(k-1)}\right\}$ (see details in [6]), and this leads to the following algorithm for sampling from the tempered distribution $p_k(\varphi)$.

Algorithm 1. AIMS-OPT at annealing level k

Input:
- ▷ $\varphi_1^{(k-1)}, \ldots, \varphi_n^{(k-1)} \sim p_{k-1}(\varphi)$, samples generated at annealing level $k-1$;
- ▷ $\varphi_1^{(k)} \in \Phi \setminus \left\{\varphi_1^{(k-1)}, \ldots, \varphi_n^{(k-1)}\right\}$, initial state of a Markov chain;
- ▷ $q_k(\varphi|\xi)$, symmetric proposal density associated with the RWMH kernel.

Algorithm:

for $i = 1, \ldots, n-1$ **do**

(1) Generate a candidate state

$$\xi \sim Q_k\left(\xi|\varphi_1^{(k-1)}, \ldots, \varphi_n^{(k-1)}\right) \stackrel{\text{def}}{=} \sum_{j=1}^{n} \bar{\omega}_j^{(k-1)} q_k\left(\xi|\varphi_j^{(k-1)}\right) \quad (17)$$

(a)Select j from $\{1, \ldots, n\}$ with respective probabilities $\bar{\omega}_1^{(k-1)}, \ldots, \bar{\omega}_n^{(k-1)}$ given by (12).

(b)Generate $\xi \sim q_k\left(\xi|\varphi_j^{(k-1)}\right)$.

(2) Update $\varphi_i^{(k)} \to \varphi_{i+1}^{(k)}$ by accepting or rejecting ξ as follows: Set

$$\varphi_{i+1}^{(k)} = \begin{cases} \xi, & \text{with probability } \mathcal{A}\left(\varphi_i^{(k)}, \varphi_j^{(k-1)}, \xi\right), \\ \varphi_i^{(k)}, & \text{with the remaining probability,} \end{cases} \quad (18)$$

where

$$\mathcal{A}\left(\varphi_i^{(k)}, \varphi_j^{(k-1)}, \xi\right) = \min\left\{1, \frac{p_k(\xi)}{p_k\left(\varphi_j^{(k-1)}\right)}\right\} \times \min\left\{1, \frac{p_k(\xi)\sum_{j=1}^{n} \bar{\omega}_j^{(k-1)} q_k\left(\varphi_i^{(k)}|\varphi_j^{(k-1)}\right) \min\left\{1, \frac{p_k\left(\varphi_i^{(k)}\right)}{p_k\left(\varphi_j^{(k-1)}\right)}\right\}}{p_k\left(\varphi_i^{(k)}\right)\sum_{j=1}^{n} \bar{\omega}_j^{(k-1)} q_k\left(\xi|\varphi_j^{(k-1)}\right) \min\left\{1, \frac{p_k(\xi)}{p_k\left(\varphi_j^{(k-1)}\right)}\right\}}\right\} \quad (19)$$

end for

Output:

► $\varphi_1^{(k)}, \ldots, \varphi_n^{(k)}$, n states of a Markov chain with a stationary distribution

$$p_k(\varphi) \propto \exp(-\mathcal{H}(\varphi)/T_k) I_\Phi(\varphi)$$

Schematically, the AIMS-OPT algorithm at annealing level k is shown in Fig. 1. The proof that $p_k(\varphi)$ is indeed a stationary distribution for the Markov chain generated by AIMS-OPT is given in the Appendix. Important ergodic properties of this Markov chain were derived in [6]. In particular, it was shown that, under certain conditions often fulfilled in practice, the generated Markov chain is uniformly ergodic, i.e., it converges to its stationary distribution relatively quickly.

As usual for MCMC algorithms, the fact of convergence of a Markov chain to its stationary distribution does not depend on the initial state; however, the speed of convergence does. One reasonable way to chose the initial state $\varphi_1^{(k)} \in \Phi \setminus \left\{\varphi_1^{(k-1)}, \ldots, \varphi_n^{(k-1)}\right\}$ in practical applications is the following: generate $\varphi_1^{(k)} \sim q_k(\varphi|\varphi_{j^*}^{(k-1)})$, where $j^* = \arg\max_j \bar{\omega}_j^{(k-1)}$, i.e., $\varphi_{j^*}^{(k-1)}$ has the largest normalized importance weight.

The choice of the local proposal density $q_k(\varphi|\xi)$ associated with the RWMH kernel determines how efficiently the Markov chain generated by AIMS-OPT at level k explores local neighborhoods of samples $\varphi_1^{(k-1)}, \ldots, \varphi_n^{(k-1)}$ generated at the previous level. This makes the choice of $q_k(\varphi|\xi)$ very important. It has been observed by many researchers that the efficiency of Metropolis–Hastings based MCMC methods is not sensitive to the type of the proposal density; however, it strongly depends on its spread (e.g., [11,3]). For this reason, we use a truncated Gaussian density as the local proposal:

$$q_k(\varphi|\xi) \propto \mathcal{N}(\varphi|\xi, c_k\mathbb{I}) I_\Phi(\varphi) I_\Phi(\xi), \quad (20)$$

where ξ and $c_k\mathbb{I}$ are the mean and diagonal covariance matrix, respectively. The scaling parameter c_k determines the spread of the local proposal distribution. The optimal values for c_k are, of course, problem dependent. As a general recommendation, c_k should decay with k, since the tempered distributions $p_k(\varphi)$ become more and more concentrated when k increases. Finally, note that for sampling from the truncated Gaussian distribution (20) where $\xi \in \Phi$ is given, the following simple strategy can be used: continue to generate random samples from the "full" Gaussian distribution $\mathcal{N}(\varphi|\xi, c_k\mathbb{I})$ until a sample satisfying $\varphi \in \Phi$ is obtained.

4.2. The full AIMS-OPT method

At the zeroth annealing level, $k = 0$, we generate samples $\varphi_1^{(0)}, \ldots, \varphi_n^{(0)}$ uniformly distributed over the admissible parameter space Φ. Then, using the algorithm described in the previous subsection, we generate samples $\varphi_1^{(1)}, \ldots, \varphi_n^{(1)}$, which are approximately distributed according to tempered distribution $p_1(\varphi) \propto \exp(-\mathcal{H}(\varphi)/T_1)$ and, therefore, are better solutions if compared with $\varphi_1^{(0)}, \ldots, \varphi_n^{(0)}$. We proceed like this until the annealing temperature T_k is small enough so that the corresponding samples $\varphi_1^{(k)}, \ldots, \varphi_n^{(k)} \sim p_k(\varphi) \approx U_{\Phi^\star}(\varphi)$ are approximately uniformly distributed over the optimal solution set $\Phi^\star$. To make the description of AIMS-OPT complete, we have to explain how to choose the annealing temperatures T_k, for $k = 1, 2, \ldots$, and provide the stopping criterion.

4.2.1. Adaptive annealing schedule

It is clear that the choice of the annealing temperatures is very important, since, for instance, it affects the accuracy of the importance sampling approximation (11) and, therefore, the efficiency of the whole AIMS-OPT method. At the same time, it is difficult to make a rational choice of the T_k-values in advance, since this requires some prior knowledge about the optimal solution set $\Phi^\star$ and the global sensitivity of the objective function $\mathcal{H}(\varphi)$ to the parameters φ, which is usually not available. For this reason, we propose an adaptive way of choosing the annealing schedule.

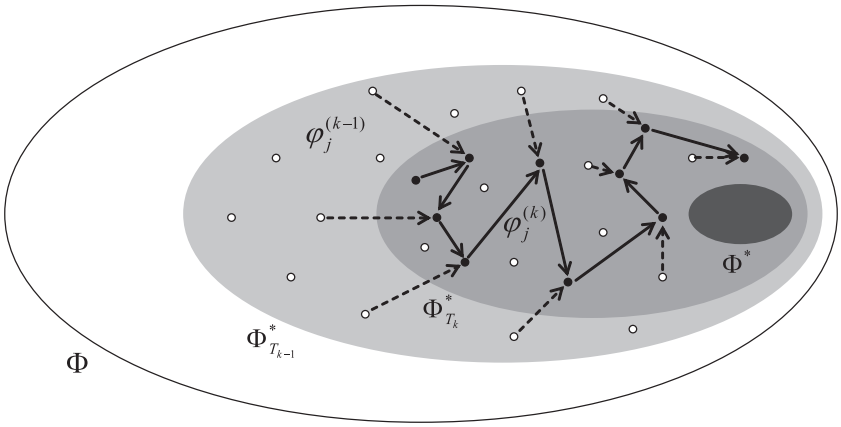

Fig. 1. AIMS-OPT at annealing level k: white and black dots (○ and •) represent $\varphi_1^{(k-1)},\dots,\varphi_n^{(k-1)}$ and $\varphi_1^{(k)},\dots,\varphi_n^{(k)}$, respectively; $\Phi^\star_{T_{k-1}}$ and $\Phi^\star$ (light and dark grey regions) are the corresponding practical supports of the tempered distributions $p_{k-1}(\varphi)$ and $p_k(\varphi)$; dashed arrows show the correspondence between $\varphi_j^{(k-1)}$ that has been chosen in step 1a and the corresponding candidate $\xi \sim q_k\left(\xi|\varphi_j^{(k-1)}\right)$ that has been generated in step 1b. In this schematic picture, all shown candidate states are accepted (in step 2) as new states of the Markov chain.

In importance sampling, a useful measure of degeneracy of the method is the *effective sample size* (ESS) n_{eff} introduced in [22] and [25]. The ESS measures how similar the importance sampling distribution $p_{k-1}(\varphi)$ is to the target distribution $p_k(\varphi)$. Suppose n independent samples $\varphi_1^{(k-1)},\dots,\varphi_n^{(k-1)}$ are generated from $p_{k-1}(\varphi)$, then the ESS of these samples is defined as

$$n_{\text{eff}} = \frac{n}{1+\text{var}_{p_{k-1}}[\omega(\varphi)]} = \frac{n}{\mathbb{E}_{p_{k-1}}[\omega(\varphi)^2]}, \tag{21}$$

where $\omega(\varphi) = p_k(\varphi)/p_{k-1}(\varphi)$. The ESS can be interpreted as implying that n weighted samples $\left(\varphi_1^{(k-1)},\omega_1^{(k-1)}\right),\dots,\left(\varphi_n^{(k-1)},\omega_n^{(k-1)}\right)$ are worth $n_{\text{eff}}(\leqslant n)$ identically and independently distributed (i.i.d.) samples drawn from the target distribution $p_k(\varphi)$. One cannot evaluate the ESS exactly but an estimate $\hat{n}_{\text{eff}}$ of n_{eff} is given by

$$\hat{n}_{\text{eff}} = \frac{1}{\sum_{j=1}^n\left(\bar{\omega}_j^{(k-1)}\right)^2}, \tag{22}$$

where $\bar{\omega}_j^{(k-1)}$ is the normalized importance weight of $\varphi_j^{(k-1)}$.

At annealing level k, when the temperature T_{k-1} is already known, the problem is to define T_k. Let $\gamma = \hat{n}_{\text{eff}}/n \in (0,1)$ be a prescribed threshold that characterizes the "quality" of the weighted sample (the larger γ is, the "better" the weighted sample is). Then we obtain the following equation:

$$\sum_{j=1}^n\left(\bar{\omega}_j^{(k-1)}\right)^2 = \frac{1}{\gamma n} \tag{23}$$

Observe that this equation can be expressed as an equation on T_k by using (12):

$$\frac{\sum_{j=1}^n \exp\left[-2\mathcal{H}(\varphi_j^{(k-1)})\left(\frac{1}{T_k}-\frac{1}{T_{k-1}}\right)\right]}{\left(\sum_{j=1}^n \exp\left[-\mathcal{H}\left(\varphi_j^{(k-1)}\right)\left(\frac{1}{T_k}-\frac{1}{T_{k-1}}\right)\right]\right)^2} = \frac{1}{\gamma n} \tag{24}$$

Solving this equation for T_k gives us the value of the annealing temperature at level k.

Remark 6. It follows from general considerations that the left-hand side of (24) is a strictly decreasing function of T_k on the interval $(0,T_{k-1}]$. Indeed, the smaller T_k, the more different the tempered distributions $p_{k-1}(\varphi)$ and $p_k(\varphi)$ are, and, therefore, the large the variability of the importance weights $\bar{\omega}_j^{(k-1)}$. In particular, the left-hand side of (24) goes to $+\infty$ when $T_k \to 0$, and it converges to $1/n$ when $T_k \to T_{k-1}$. Thus, Eq. (24) can be easily solved for T_k by standard numerical methods, such as the bisection method.

The threshold γ affects the speed of annealing. If γ is very small, i.e., close to zero, then AIMS-OPT will have very few tempered distributions, and this will lead to inaccurate results for a moderate number of samples n. On the other hand, if γ is very large, i.e., close to one, then AIMS-OPT will have too many tempered distributions, which will make the algorithm computationally very expensive. As discussed in [6], in the context of sampling the posterior distribution, $\gamma = 1/2$ is usually a reasonable choice of the threshold. Our simulation results (see Section 5) also show that annealing schedules with γ around 1/2 yield good efficiency.

4.2.2. Stopping criterion

As $k \to \infty$, the adaptively chosen annealing temperature T_k decreases toward zero, the tempered distribution $p_k(\varphi)$ converges to the uniform distribution $U_{\Phi^\star}(\varphi)$ over the optimal solution set $\Phi^\star$, and, therefore, the generated samples $\varphi_1^{(k)},\dots,\varphi_n^{(k)}$ become more and more uniformly distributed over $\Phi^\star$. In practice, however, "absolute zero" $T_k = 0$ cannot be reached, and the algorithm should stop using some stopping rule. The proposed stopping criterion is based on the sample coefficient of variation (COV) of the objective function $\mathcal{H}(\varphi)$.

Remark 7. We can always guarantee positivity of the loss function by replacing $h(\varphi,\theta)$ with $\tilde{h}(\varphi,\theta) = h(\varphi,\theta) - c$, where $c < \min_{\varphi\in\Phi,\theta\in\Theta} h(\varphi,\theta)$, and noting that optimization of the expected value of $h(\varphi,\theta)$ is equivalent to optimization of the expected value of $\tilde{h}(\varphi,\theta)$ since $\mathbb{E}_\pi[\tilde{h}(\varphi,\theta)] = \mathbb{E}_\pi[h(\varphi,\theta)] - c$. Therefore, we assume $h(\varphi,\theta)$ is positive and hence $\mathcal{H}(\varphi)$ is positive.

Let δ_k denote the sample COV of $\mathcal{H}\left(\varphi_1^{(k)}\right),\dots,\mathcal{H}\left(\varphi_n^{(k)}\right)$, i.e.,

$$\delta_k = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^n\left(\mathcal{H}\left(\varphi_i^{(k)}\right)-\frac{1}{n}\sum_{j=1}^n\mathcal{H}\left(\varphi_j^{(k)}\right)\right)^2}}{\frac{1}{n}\sum_{j=1}^n\mathcal{H}\left(\varphi_j^{(k)}\right)} \tag{25}$$

We use δ_k as a measure of sensitivity of the objective function to the parameters φ in the domain $\Phi^\star_{T_k}$, the practical support of $p_k(\varphi)$. If samples $\varphi_1,\dots,\varphi_n \sim U_{\Phi^\star}(\varphi)$, then their COV is zero, since $\mathcal{H}(\varphi_j) = \min_{\varphi\in\Phi}\mathcal{H}(\varphi)$ for all $j = 1,\dots,n$. Therefore, δ_k converges to zero, when $k \to \infty$. This suggests the following stopping criterion: run the algorithm until δ_k becomes less than some fraction α of the initial sample COV δ_0; in other words, stop when the following condition is fulfilled:

$$\delta_k < \alpha\delta_0 \stackrel{\text{def}}{=} \delta_{\text{target}}, \tag{26}$$

where $\alpha \in (0,1)$ and δ_{target} is the target sample COV. The threshold α affects the accuracy of approximation of the optimal solution set $\Phi^\star$ by a set of samples $\{\varphi_1^{(k)},\dots,\varphi_n^{(k)}\}$. Small values of α correspond to better approximation.

Combining the AIMS-OPT algorithm at a given annealing level with the described adaptive annealing schedule and stopping rule gives rise to the following procedure.

Algorithm 2. The AIMS-OPT method

Input:
- ▷ $\mathcal{H}(\varphi)$, objective function in (2) (or a sample-based estimate like $\widehat{\mathcal{H}}_N(\varphi)$ in (3));
- ▷ γ, threshold for the effective sample size (ESS);
- ▷ α, threshold for the stopping rule;
- ▷ n, the number of Markov chain states to be generated at each annealing level;
- ▷ $q_1(\varphi|\xi), q_2(\cdot|\xi),\dots$, where $q_k(\varphi|\xi)$ is the symmetric proposal density associated with the RWMH kernel at annealing level k.

Algorithm:

Set $k = 0$, current annealing level.
Set $T_0 = \infty$, current annealing temperature.
Sample $\varphi_1^{(0)},\dots,\varphi_n^{(0)} \sim U_\Phi(\varphi)$.
Calculate δ_0 using (25).
Set $\delta_{\text{target}} = \alpha\delta_0$, the target sample COV.
while $\delta_k > \delta_{\text{target}}$ **do**
Find T_{k+1} from Eq. (24).
Calculate normalized importance weights $\bar{\omega}_j^{(k)}$, $j = 1,\dots,n$ using (12).
Generate a Markov chain $\varphi_1^{(k+1)},\dots,\varphi_n^{(k+1)}$ with the stationary distribution $p_{k+1}(\varphi)$ using the AIMS-OPT algorithm at annealing level $k+1$.
Calculate δ_{k+1} using (25).
Increment k to $k+1$.
end while
Set $K = k$, the total number of tempered distributions in the annealing schedule.
Set $\tau = T_k$, the smallest annealing temperature.

Output:
- ► $\varphi_1^{(K)},\dots,\varphi_n^{(K)} \sim U_{\Phi^\star}(\varphi)$, samples that are approximately uniformly distributed over the optimal solution set $\Phi^\star$.

It is important to highlight that when the objective function $\mathcal{H}(\varphi)$ is of the form in (1) and so is not directly available, care must be taken when using its sample-based estimate $\widehat{\mathcal{H}}_N(\varphi)$ since AIMS-OPT actually minimizes this approximation. In this case, the sequence of sample COVs $\{\delta_k\}$ may no longer converge to zero as $k \to \infty$, since $\widehat{\mathcal{H}}_N(\varphi) = \mathcal{H}(\varphi) + \epsilon_N$, where ϵ_N is the estimation error in (3) and $\epsilon_N \to 0$ as $N \to \infty$. Therefore, if the target COV δ_{target} in (26) is too small, or, equivalently, the threshold α is too small, the above procedure may never converge because of the approximation error ϵ_N. This pitfall can be avoided by monitoring the relative rate of decrease of $\{\delta_k\}$: if $\frac{\delta_k - \delta_{k+1}}{\delta_k}$ is smaller than a certain threshold, then one should increase the sample size N, to reduce ϵ_N and achieve a better accuracy in (3). Alternatively, if increasing N is not an option, one can simply terminate the algorithm and take the current set of samples $\left\{\varphi_1^{(k)},\dots,\varphi_n^{(k)}\right\}$ as a less accurate (than planned originally) approximation of the optimal solution set $\Phi^\star$.

Finally, note that if a single optimal solution $\hat{\varphi}$ is needed, then a rational choice based on the samples $\varphi_1^{(K)},\dots,\varphi_n^{(K)}$ is:

$$\hat{\varphi} = \varphi_{j^*}^{(K)}, \text{where } j^* = \arg\min_{j=1,\dots,n} \mathcal{H}\left(\varphi_j^{(K)}\right) \tag{27}$$

5. Illustrative examples

To illustrate the effectiveness of AIMS-OPT for solving the optimization problem (2), we consider the following three test functions defining objective functions by (1):

$$\begin{aligned} h_1(\varphi,\theta) &= 1 + \left(\varphi_1 - \frac{a}{2}\right)\theta_1 + \left(\varphi_2 - \frac{a}{2}\right)\theta_2, \\ h_2(\varphi,\theta) &= 1 + \left(\varphi_1 - \frac{a}{2}\right)\left(\varphi_2 - \frac{a}{2}\right)\theta_1\theta_2, \\ h_3(\varphi,\theta) &= 4a - \theta_1 \operatorname{sign}\left(\varphi_1 - \frac{a}{2}\right) - \theta_2 \operatorname{sign}\left(\varphi_2 - \frac{a}{2}\right), \end{aligned} \tag{28}$$

where the admissible parameter space is the square $(\varphi_1, \varphi_2) \in \Phi = [0,a] \times [0,a], a = 10$, and the model parameters $\theta_1 \sim \mathcal{N}(\varphi_1 - \frac{a}{2}, 1)$ and $\theta_2 \sim \mathcal{N}(\varphi_2 - \frac{a}{2}, 1)$.

5.1. Exact optimal solution sets

It is easy to evaluate the corresponding objective functions analytically, $\mathcal{H}_i(\varphi) = \mathbb{E}_\pi[h_i(\varphi,\theta)]$, $i = 1,2,3$,

$$\begin{aligned} \mathcal{H}_1(\varphi) &= 1 + \left(\varphi_1 - \frac{a}{2}\right)^2 + \left(\varphi_2 - \frac{a}{2}\right)^2, \\ \mathcal{H}_2(\varphi) &= 1 + \left(\varphi_1 - \frac{a}{2}\right)^2\left(\varphi_2 - \frac{a}{2}\right)^2, \\ \mathcal{H}_3(\varphi) &= 4a - \left|\varphi_1 - \frac{a}{2}\right| - \left|\varphi_2 - \frac{a}{2}\right|. \end{aligned} \tag{29}$$

The optimal solution sets for these three case are, therefore,

$$\begin{aligned} \Phi_1^\star &= \left(\frac{a}{2}, \frac{a}{2}\right), \\ \Phi_2^\star &= \left\{\varphi \in \Phi | \varphi_1 = \frac{a}{2} \quad \text{or} \quad \varphi_2 = \frac{a}{2}\right\}, \\ \Phi_3^\star &= (0,0) \cup (0,a) \cup (0,a) \cup (a,a). \end{aligned} \tag{30}$$

We will refer to $\Phi_1^\star, \Phi_2^\star$, and $\Phi_3^\star$ as "center", "cross", and "corners", respectively. These optimal subsets of the admissible parameter space are shown in Fig. 2. Note that while "center" is a relatively simple case—$\Phi_1^\star$ consists of a single point located at the center of Φ—"cross" and "corners" are quite challenging cases: $\Phi_2^\star$ has complicated geometry and $\Phi_3^\star$ consists of four different points situated far from each other.

5.2. Approximation of optimal solution sets using AIMS-OPT

To estimate the objective functions $\mathcal{H}_i(\varphi)$, for each value of $\varphi = (\varphi_1,\varphi_2)$, $N = 10^3$ samples of the model parameters $\theta = (\theta_1,\theta_2)$ were used in (3). The left panels of Figs. 3–5 display the scatterplots of $n = 10^3$ samples of the design parameters obtained from AIMS-OPT for "center", "cross", and "corners", respectively. In these figures, k denotes the annealing level. The parameters of the algorithm were chosen as follows: the threshold for the ESS $\gamma = 1/2$; the threshold for the stopping rule $\alpha = 0.05$ for "center" and "corners" and $\alpha = 0.01$ for "cross"; the local proposal density $q_k(\varphi|\xi) \propto \mathcal{N}(\varphi|\xi, c_k\mathbb{I})I_\Phi(\varphi)I_\Phi(\xi)$, where $c_0 = 0.1$ and $c_{k+1} = c_k/4$. This implementation of AIMS-OPT leads to a total number of $K = 6$, 4, and 5 tempered distributions for "center", "cross", and "corners", respectively. The corresponding values of the annealing temperatures are given in Table 1. Note that the total number of tempered distributions K can be considered as a measure of the global sensi-

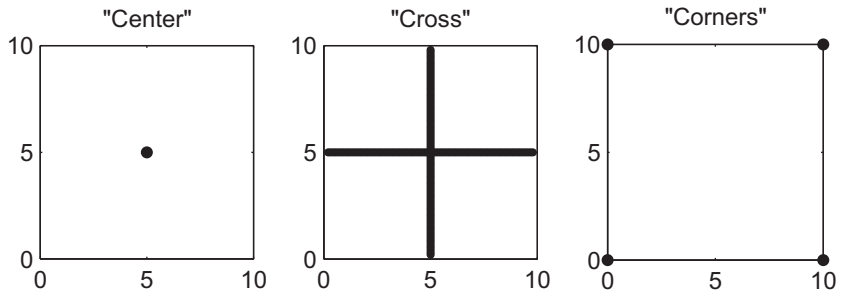

Fig. 2. "Center", "cross", and "corners": optimal solution sets for objective functions (29).

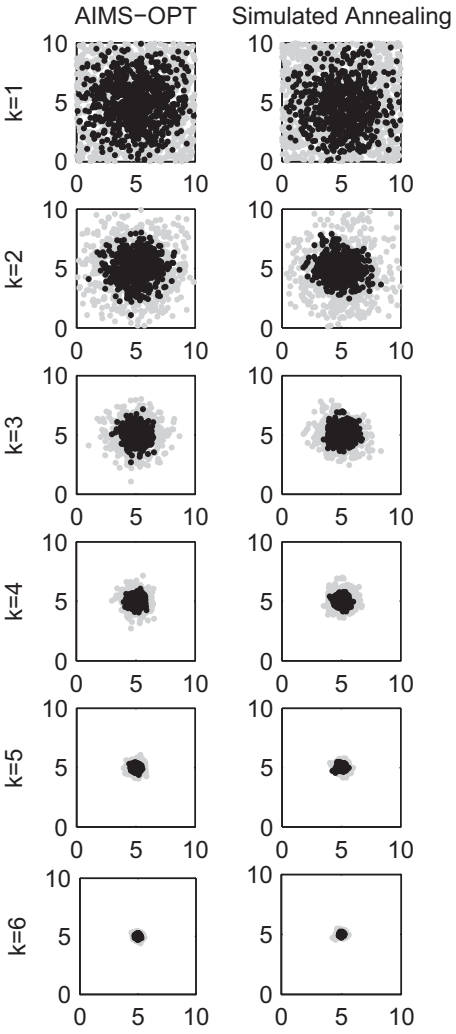

Fig. 3. Case "Center": the left and right panels display the scatterplots of $n = 10^3$ samples obtained from AIMS-OPT and SA, respectively; k denotes the annealing level. Grey and black dots represent $\varphi_1^{(k-1)}, \ldots, \varphi_n^{(k-1)}$ and $\varphi_1^{(k)}, \ldots, \varphi_n^{(k)}$, respectively.

tivity of the objective function to the parameters φ: the larger K, the more sensitive the objective function is.

The theoretical minimum values of the objective functions are

$$\min_{\varphi\in\Phi} \mathcal{H}_1(\varphi) = 1, \min_{\varphi\in\Phi} \mathcal{H}_2(\varphi) = 1, \min_{\varphi\in\Phi} \mathcal{H}_2(\varphi) = 3a = 30. \tag{31}$$

The minimum and the maximum values of the objective functions computed for the samples generated at the last annealing levels ($K = 6$, 4, and 5 for for "center", "cross", and "corners", respectively) are

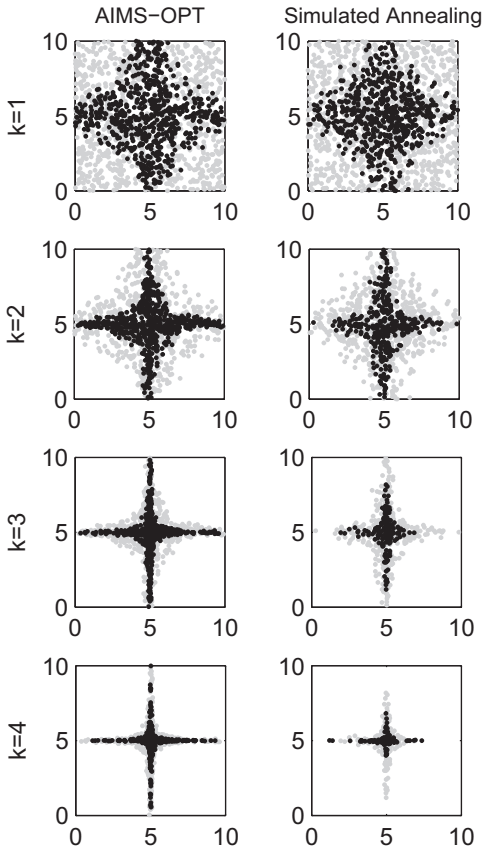

Fig. 4. Case "Cross": the left and right panels display the scatterplots of $n = 10^3$ samples obtained from AIMS-OPT and SA, respectively; k denotes the annealing level. Grey and black dots represent $\varphi_1^{(k-1)}, \ldots, \varphi_n^{(k-1)}$ and $\varphi_1^{(k)}, \ldots, \varphi_n^{(k)}$, respectively.

$$\begin{aligned}
\left[\min_{j=1,\ldots,n} \mathcal{H}_1\left(\varphi_j^{(6)}\right), \max_{j=1,\ldots,n} \mathcal{H}_1\left(\varphi_j^{(6)}\right)\right] &= [1.02, 1.12],\\
\left[\min_{j=1,\ldots,n} \mathcal{H}_2\left(\varphi_j^{(4)}\right), \max_{j=1,\ldots,n} \mathcal{H}_2\left(\varphi_j^{(4)}\right)\right] &= [1.01, 1.03],\\
\left[\min_{j=1,\ldots,n} \mathcal{H}_3\left(\varphi_j^{(5)}\right), \max_{j=1,\ldots,n} \mathcal{H}_3\left(\varphi_j^{(5)}\right)\right] &= [30.012, 31.10].
\end{aligned} \tag{32}$$

Let us now compare the performance of AIMS-OPT with Simulated Annealing (SA) [26,30], a well-known stochastic simulation method originally introduced in [21]. Since we are interested in optimization problems where there may be multiple optimal solutions, we use a population-based SA with adaptive annealing schedule rather than the original algorithm of [21] that uses a single thread where the temperature is changed after each Markov chain update by using a pre-set annealing schedule. Thus, the only difference between AIMS-OPT and a population-based SA is in

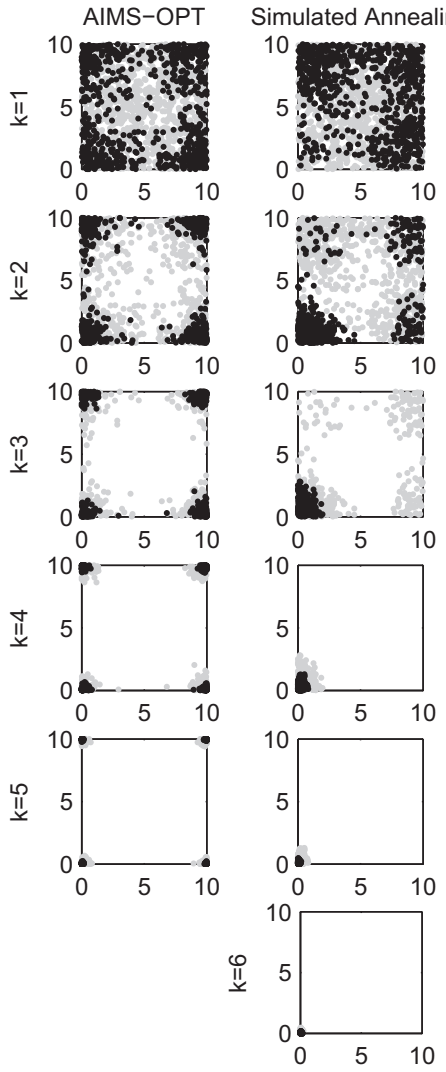

Fig. 5. Case "Corners": the left and right panels display the scatterplots of $n = 10^3$ samples obtained from AIMS-OPT and SA, respectively; k denotes the annealing level. Grey and black dots represent $\varphi_1^{(k-1)},\dots,\varphi_n^{(k-1)}$ and $\varphi_1^{(k)},\dots,\varphi_n^{(k)}$, respectively.

Table 1
Annealing temperatures.

Case	Temperature	T_1	T_2	T_3	T_4	T_5	T_6
"Center"	AIMS-OPT	8.720	2.429	0.771	0.227	0.064	0.021
	Simulated annealing	8.617	2.280	0.598	0.180	0.055	0.016
"Cross"	AIMS-OPT	16.016	1.126	0.085	0.008	–	–
	Simulated annealing	15.971	1.013	0.064	0.003	–	–
"Corners"	AIMS-OPT	2.072	0.828	0.358	0.152	0.045	–
	Simulated annealing	2.169	0.939	0.448	0.221	0.115	0.039

how samples at a given annealing level are generated. In AIMS-OPT, samples $\varphi_1^{(k)},\dots,\varphi_n^{(k)}$ are obtained using IMH with the global proposal distribution $\hat{p}_{k,n}(d\varphi)$ which is constructed based on the samples $\varphi_1^{(k-1)},\dots,\varphi_n^{(k-1)}$ from the previous annealing level. In SA, RWMH with the local proposal distribution $q_{loc,k}(\varphi|\xi)$ is used instead. To capture the global structure of the target distribution, the local proposal distribution is often chosen to be Gaussian of the following form:

$$q_{loc,k}(\varphi|\xi) \propto \mathcal{N}(\varphi|\xi, c\Sigma_k) I_\Phi(\varphi) I_\Phi(\xi), \tag{33}$$

where Σ_k is the sample covariance matrix,

$$\Sigma_k = \sum_{j=1}^{n} \bar{\omega}_j^{(k-1)} \left(\varphi_j^{(k-1)} - \bar{\varphi}^{(k)}\right)\left(\varphi_j^{(k-1)} - \bar{\varphi}^{(k)}\right)^T,$$
$$\bar{\varphi}^{(k)} = \sum_{j=1}^{n} \bar{\omega}_j^{(k-1)} \varphi_j^{(k-1)}, \tag{34}$$

$\bar{\omega}_j^{(k-1)}$ are the normalized importance weights from (12), and c is the scaling factor. In all examples, we used an approximately optimal value of the scaling factor, $c = 0.1$.

Remark 8. Note that in AIMS-OPT, as opposed to SA, we do not need to use the sample covariance matrix for the local proposal density $q_k(\varphi|\xi)$ associated with the RWMH transition kernel (13). A simple diagonal covariance matrix as in (20) will suffice, since the main goal of sampling from $q_k\left(\varphi|\varphi_j^{(k-1)}\right)$ is to explore the neighborhood of $\varphi_j^{(k-1)}$. The global structure of the target distribution $p_k(\varphi)$ will be captured automatically by picking different samples among $\varphi_1^{(k-1)},\dots,\varphi_n^{(k-1)}$(see Step 1a in the description of AIMS-OPT at annealing level k in Section 4.1).

The right panels of Figs. 3–5 display the scatterplots of $n = 10^3$ samples obtained from SA for "center", "cross", and "corners", respectively. In the first (simple) case "center", both AIMS-OPT and SA successfully generate samples in the vicinity of the optimal solution set $\Phi_1^\star$. Note that the corresponding values of the annealing temperatures are also comparable (see Table 1). In a more complicated case "cross", AIMS-OPT approximates $\Phi_2^\star$ more accurately than SA. Finally, in the most challenging case "corners", AIMS-OPT clearly outperforms SA: while the AIMS-OPT algorithm successfully finds all four optimal solutions $(0,0)$, $(0,a)$, $(0,a)$, and (a,a), SA finds only one, $(0,0)$. Note also that in this case, the tempered distributions $p_k(\varphi)$ "cool down" slower under SA, and, as a result, SA requires six annealing levels while AIMS-OPT requires only five.

5.3. Higher dimensional examples

To demonstrate the effectiveness of AIMS-OPT for higher dimensionality, we first consider the case "center" in ten dimensions. Namely, consider the following test function

$$h(\varphi,\theta) = 1 + \sum_{i=1}^{d} \left(\varphi_i - \frac{a}{2}\right)\theta_i, \tag{35}$$

where $d = 10$, the admissible parameter space is the hypercube $\varphi \in \Phi = [0,a]^d$, $a = 10$, and the model parameters $\theta_i \sim \mathcal{N}(\varphi_i - \frac{a}{2}, 1)$, for $i = 1,\dots,d$. As in two dimensions, it is easy to evaluate the corresponding objective function analytically

$$\mathcal{H}(\varphi) = 1 + \sum_{i=1}^{d} \left(\varphi_i - \frac{a}{2}\right)^2. \tag{36}$$

The optimal solution set for this case is the center of the hypercube,

$$\Phi^\star = \left(\frac{a}{2},\dots,\frac{a}{2}\right). \tag{37}$$

Fig. 6 displays the scatterplots of $n = 2000$ samples obtained from AIMS-OPT and projected onto $\varphi_1\varphi_5\varphi_{10}$- subspace. For each value of φ, $N = 10^3$ samples of θ were used in (3) for estimating the objective function $\mathcal{H}(\varphi)$. The AIMS-OPT parameters were chosen as follows: the threshold for the ESS $\gamma = 1/2$; the threshold for the stopping rule $\alpha = 0.05$; the local proposal density $q_k(\varphi|\xi) \propto \mathcal{N}(\varphi|\xi, c_k\mathbb{I}) I_\Phi(\varphi) I_\Phi(\xi)$, where $c_0 = 10$ and $c_{k+1} = c_k/4$. This implementation of AIMS-OPT leads to a total number of $K = 15$

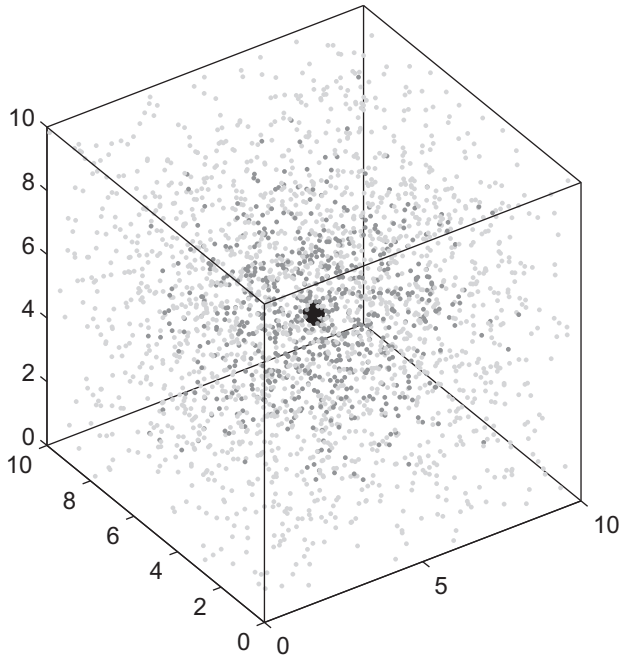

Fig. 6. Case "Center" in 10-D: scatterplots of 2000 samples obtained from AIMS-OPT and projected onto $\varphi_1\varphi_5\varphi_{10}$-subspace. Light grey, dark grey, and black dots represent samples generated at the kth level for $k = 1$, $k = 10$, and $k = 15$, respectively.

tempered distributions. Fig. 7 shows the values of the objective function computed for the samples generated at the last annealing level.

As a more challenging high-dimensional example with multiple optimal solutions, we also considered an analog of the case "corners" in ten dimensions, where the optimal solution set $\Phi^\star$ consists of four corners of the hyper-cube $\Phi = [0,a]^d$, $a = 10$, $d = 10$:

$$\begin{aligned}
\varphi^\star_{(1)} &= (a,0,0,0,0,0,0,0,0,0),\\
\varphi^\star_{(2)} &= (0,0,0,a,0,0,0,0,0,0),\\
\varphi^\star_{(3)} &= (0,0,0,0,0,0,a,0,0,0),\\
\varphi^\star_{(4)} &= (0,0,0,0,0,0,0,0,0,a).
\end{aligned} \tag{38}$$

The same implementation of AIMS-OPT as in the case "center" was able to find all four optimal solutions.

5.4. Reliability-based design optimization of a ductile steel structure subjected to ground motion

Although some of the examples considered in previous sections were quite challenging ("cross" and "corners"), all of them are purely hypothetical in the sense that the test functions (28) did not come from particular applications, but rather were specifically chosen to test the efficiency of the algorithm. To demonstrate the utility of AIMS-OPT for engineering applications, we consider the problem of reliability-based design optimization of a ductile steel structure subjected to ground motion.

Consider a structure that is modeled as a 2D six-story moment-resisting steel frame with two-node beam elements connecting the joints of the frame. The floors are modeled as rigid in-plane and the joints are modeled as rigid-plastic. The yield strength is taken to be 317 MPa for all members. Under service load conditions, the floors and the roof are subjected to a uniformly-distributed static span load of 24.7 kN/m and 13.2 kN/m, respectively. For modeling the horizonal motion of the structure, masses are lumped at the floor levels, which include contributions from live loads and the dead loads from the floors and the frame members. This structural example is taken from [1], and we refer the reader to the original paper for a full description of the structure. This structure is also studied in [39].

In this example, we consider two design parameters that control the structural properties of the frame members: the area moment of inertia, $\varphi_1 = I$, and the cross-sectional area, $\varphi_2 = A$. Suppose that 1000 in^4 $\leqslant I \leqslant$ 2000 in^4 and 40 in^2 $\leqslant A \leqslant$ 50 in^2, i.e., the set of admissible designs Φ is a rectangle $[1000,2000] \times [40,50]$. This design constraint may come, for instance, from economic considerations.

The structure is subject to uncertain earthquake excitations modeled as a nonstationary stochastic process. To simulate a time history of the ground motion acceleration for given moment magnitude M and epicentral distance r, a discrete-time white noise sequence $W_j = \sqrt{2\pi/\Delta t}\,\theta_j$, $j = 1,\ldots,N_\theta$ is first generated, where $\Delta t = 0.05$ s is the sampling time, $N_\theta = 101$ is the number of time instants (which corresponds to a duration of 5 s), and $\theta_1,\ldots,\theta_{N_\theta}$ are i.i.d. standard Gaussian variables. The white noise sequence is then modulated (multiplied) by an envelope function $e(t;M,r)$ at the discrete time instants. The discrete Fourier transform is then applied to the modulated white-noise sequence. The resulting spectrum is multiplied with a radiation spectrum $A(f;M,r)$ [1], after which the discrete inverse Fourier transform is applied to transform the sequence back to the time domain to yield a sample for the ground acceleration time history. The synthetic ground motion $a(t;Z,M,r)$ generated from the model is thus a function of the Gaussian vector

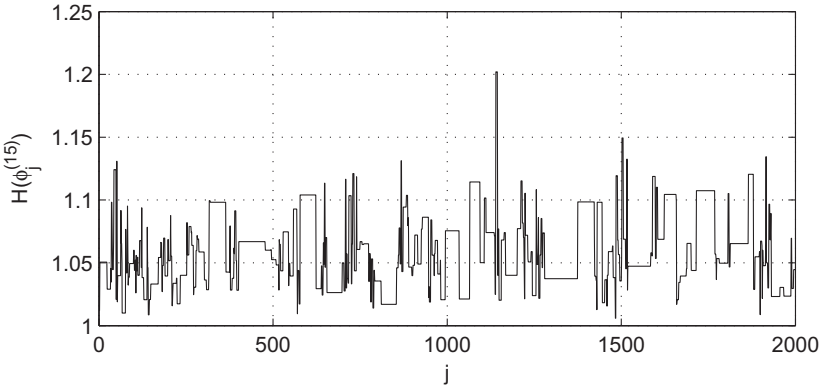

Fig. 7. Case "Center" in 10-D: the values of the objective function computed for the samples $\varphi_1^{(K)},\ldots,\varphi_n^{(K)}$, $n = 2000$ generated at the last annealing level $K = 15$.

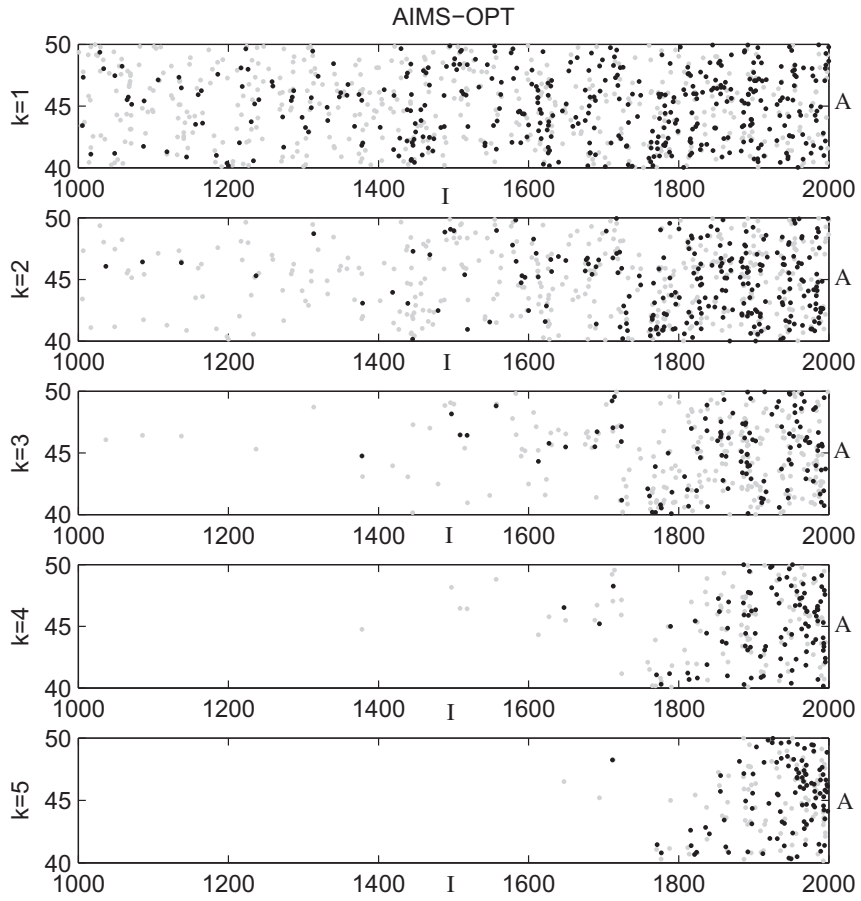

Fig. 8. RBDO: scatterplots of 500 samples obtained from AIMS-OPT; k denotes the annealing level. Grey and black dots represent $\varphi_1^{(k-1)},\ldots,\varphi_n^{(k-1)}$ and $\varphi_1^{(k)},\ldots,\varphi_n^{(k)}$, respectively, where each sample $\varphi_i^{(j)}=\left(I_i^{(j)},A_i^{(j)}\right)$.

$\theta=(\theta_1,\ldots,\theta_{N_\theta})^T$ and stochastic excitation model parameters M and r. We consider a scenario event with $M=7$ and $r=50$ km.

In this example, the uncertainty arises from the model parameters $\theta_1,\ldots,\theta_{N_\theta}$, the i.i.d. Gaussian sequence that generates the synthetic ground motion. The system response of interest, $g(\varphi,\theta)$, is defined to be the peak (absolute) interstory drift ratio $\delta_{\max}=\max_{i=1,\ldots,6}\delta_i$, where δ_i is the maximum absolute interstory drift ratio of the ith story within the duration of study, 5 s. The failure domain $F\subset\mathbb{R}^{N_\varphi}\times\mathbb{R}^{N_\theta}$ is defined as the exceedance of peak interstory drift ratio above a threshold in any one of the stories within the duration of study; that is,

$$F=\{(\varphi,\theta)\in\mathbb{R}^{N_\varphi}\times\mathbb{R}^{N_\theta}:\delta_{\max}>b\},\tag{39}$$

where b is some prescribed critical threshold. In this example, $b=0.2\%$ is considered; according to [37], $\delta_{\max}<0.2\%$ corresponds to no significant damage ("fully operational").

As mentioned in Section 1, in reliability-based design optimization (RBDO), the performance measure is $h(\varphi,\theta)=I_F(\varphi,\theta)$, where $I_F(\varphi,\theta)$ is the indicator function of the failure domain F. In RBDO, the objective function for minimization is the probability of failure, i.e., $\mathcal{H}(\varphi)=\mathbb{E}_\pi[I_F(\varphi,\theta)]=\mathbb{P}(F|\varphi)$. In this example, the distribution π is simply an N_θ-variate standard Gaussian distribution.

Fig. 8 displays the scatterplots of $n=500$ samples obtained from AIMS-OPT. For each value of $\varphi=(I,A)$, $N=10^3$ samples of θ were used in (3) for estimating the failure probability $\mathbb{P}(F|\varphi)$. The AIMS-OPT parameters were chosen as follows: the threshold for ESS: $\gamma=0.9$; the local proposal density: $q_k(\varphi|\xi)\propto\mathcal{N}(\varphi|\xi,c_k\mathbb{I})I_\Phi(\varphi)$ $I_\Phi(\xi)$, where $c_0=10$ and $c_{k+1}=c_k/4$. This implementation of AIMS-OPT leads to a total number of $K=5$ tempered distributions. The minimum and the maximum values of the objective function computed for the samples generated at the first and last annealing level are

$$\begin{aligned}\left[\min_{j=1,\ldots,n}\mathbb{P}\left(F|\varphi_j^{(0)}\right),\max_{j=1,\ldots,n}\mathbb{P}\left(F|\varphi_j^{(0)}\right)\right]&=[0.216,0.313],\\ \left[\min_{j=1,\ldots,n}\mathbb{P}\left(F|\varphi_j^{(5)}\right),\max_{j=1,\ldots,n}\mathbb{P}\left(F|\varphi_j^{(5)}\right)\right]&=[0.201,0.252],\end{aligned}\tag{40}$$

with the minimum value of 0.201 given by $(I,A)=(1996.1,46.5)$. Note that if the failure probability is very small, then the Subset Simulation method [2] can be used instead of plain vanilla Monte Carlo simulation in (3). Simulation results presented in Fig. 8, as expected, show that while the optimal (with respect to the objective function $\mathcal{H}(\varphi)=\mathbb{P}(F|\varphi)$) designs are not sensitive to the cross-sectional area, $40\text{ in.}^2\leqslant A\leqslant 50\text{ in.}^2$, the area moment of inertia I takes larger values that are close to the upper limit of 2000 in.^{4}. The scatter in I values away from 2000 is due to the error in approximating the failure probability for each (I,A) sample (the coefficient of variation of the Monte Carlo estimator is 7%).

6. Conclusions

In this paper, a new stochastic simulation method, denoted AIMS-OPT, is introduced for solving global optimization problems such as those that arise in optimal performance-based design. AIMS-OPT is based on Asymptotically Independent Markov Sampling (AIMS), a recently developed advanced simulation scheme originally proposed for Bayesian inference [6]. The main feature of AIMS-OPT is that instead of a single approximation of the

optimal solution, the algorithm produces a set of nearly optimal solutions. This can be advantageous in many practical cases, e.g., in multi-objective optimization or when there exists whole set of optimal solutions. Also, AIMS-OPT can be used for exploration of the global sensitivity of the objective function. When the objective function is not directly available and must be estimated (e.g., by the Monte Carlo method), the performance of AIMS-OPT depends on the desired accuracy of this estimation.

The efficiency of AIMS-OPT is demonstrated with several examples, including a reliability-based design optimization problem, which have different topologies of the optimal solution sets. A comparison is made with Simulated Annealing (SA). If the optimal solution set has a relatively simple geometry, then the performances of AIMS-OPT and SA are comparable; however, in more complicated cases, AIMS-OPT outperforms SA. The presented simulation results demonstrate that AIMS-OPT works very well in low- and medium-dimensional design parameter spaces. In our future research we plan to investigate the question of efficiency and robustness of AIMS-OPT for solving optimization problems that involve higher-dimensional parameter spaces.

Acknowledgements

This paper is devoted to the memory of Professor Gerhart I. Schuëller, a great scientist and colleague who is sadly missed. This work was supported by the National Science Foundation under award number EAR-0941374 to the California Institute of Technology. This support is gratefully acknowledged. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect those of the National Science Foundation. The authors thank Professor S.K. Au for supplying the structural model for example in Section 5.4. The authors also thank anonymous reviewers of the manuscript for their valuable suggestions and comments that improved the quality of the paper.

Appendix A

In this appendix, we prove that the AIMS-OPT algorithm at annealing level k described in Section 4.1 indeed generates a Markov chain with a stationary distribution $p_k(\varphi)$.

Let $\mathcal{P}(d\varphi_2|\varphi_1) = f(\varphi_2|\varphi_1)d\varphi_2 + r(\varphi_1)\delta_{\varphi_1}(d\varphi_2)$ be a transition kernel of a Markov chain, where $f(\varphi_2|\varphi_1)$ describes transitions from φ_1 to $\varphi_2 \neq \varphi_1$ and $r(\varphi_1) = 1 - \int_\Phi f(\varphi_2|\varphi_1)d\varphi_2$ is the probability that the Markov chain remains at φ_1. It is well know that a sufficient condition for $p(\varphi)$ to be a stationary distribution for the Markov chain is for $f(\varphi_2|\varphi_1)$ to satisfy the detailed balance condition:

$$p(\varphi_1)f(\varphi_2|\varphi_1) = p(\varphi_2)f(\varphi_1|\varphi_2) \tag{41}$$

Indeed, in this case

$$\begin{aligned}
&\int_\Phi \mathcal{P}(d\varphi_2|\varphi_1)p(\varphi_1)d\varphi_1 \\
&= \left(\int_\Phi p(\varphi_1)f(\varphi_2|\varphi_1)d\varphi_1\right)d\varphi_2 + \int_\Phi p(\varphi_1)r(\varphi_1)\delta_{\varphi_1}(d\varphi_2)d\varphi_1 \\
&= p(\varphi_2)\left(\int_\Phi f(\varphi_1|\varphi_2)d\varphi_1\right)d\varphi_2 + \left(\int_\Phi p(\varphi_1)r(\varphi_1)\delta(\varphi_1-\varphi_2)d\varphi_1\right)d\varphi_2 \\
&= p(\varphi_2)(1-r(\varphi_2))d\varphi_2 + p(\varphi_2)r(\varphi_2)d\varphi_2 = p(\varphi_2)d\varphi_2.
\end{aligned} \tag{42}$$

Eq. (42) means that if the current state of the Markov chain φ_1 is distributed as $p(\varphi)$, then the next state $\varphi_2 \sim \mathcal{P}(d\varphi_2|\varphi_1)$ is also distributed according to $p(\varphi)$, and, therefore, $p(\varphi)$ is a stationary distribution of the Markov chain.

Now, let $\varphi_1^{(k)}, \varphi_2^{(k)}, \ldots$ be the Markov chain on $\Phi \setminus \left\{\varphi_1^{(k-1)}, \ldots, \varphi_n^{(k-1)}\right\}$ generated by the AIMS-OPT algorithm at annealing level k described in Section 4.1. To demonstrate that $p_k(\varphi)$ is a stationary distribution of the Markov chain, it is enough to show that the detailed balance condition holds for any $\varphi_1, \varphi_2 \in \Phi \setminus \left\{\varphi_1^{(k-1)}, \ldots, \varphi_n^{(k-1)}\right\}$:

$$p_k(\varphi_1)f_k(\varphi_2|\varphi_1) = p_k(\varphi_2)f_k(\varphi_1|\varphi_2), \tag{43}$$

where $f_k(\varphi_2|\varphi_1)$ describes AIMS-OPT transitions from φ_1 to $\varphi_2 \neq \varphi_1$. From the description of the algorithm, it follows that $f_k(\varphi_2|\varphi_1)$ has the following form:

$$\begin{aligned}
f_k(\varphi_2|\varphi_1) = &\sum_{j=1}^n \bar{\omega}_j^{(k-1)} q_k(\varphi_2|\varphi_j^{(k-1)}) \min\left\{1, \frac{p_k(\varphi_2)}{p_k(\varphi_j^{(k-1)})}\right\} \\
&\times \min\left\{1, \frac{p_k(\varphi_2)\sum_{j=1}^n \bar{\omega}_j^{(k-1)} q_k(\varphi_1|\varphi_j^{(k-1)}) \min\left\{1, \frac{p_k(\varphi_1)}{p_k(\varphi_j^{(k-1)})}\right\}}{p_k(\varphi_1)\sum_{j=1}^n \bar{\omega}_j^{(k-1)} q_k\left(\varphi_2|\varphi_j^{(k-1)}\right) \min\left\{1, \frac{p_k(\varphi_2)}{p_k\left(\varphi_j^{(k-1)}\right)}\right\}}\right\}.
\end{aligned} \tag{44}$$

Using (44) and a simple fact that $a \min\{1, b/a\} = b \min\{1, a/b\}$ for all $a, b > 0$, we have:

$$\begin{aligned}
p_k(\varphi_1)f_k(\varphi_2|\varphi_1) &= p_k(\varphi_1)\sum_{j=1}^n \bar{\omega}_j^{(k-1)} q_k\left(\varphi_2|\varphi_j^{(k-1)}\right) \min\left\{1, \frac{p_k(\varphi_2)}{p_k\left(\varphi_j^{(k-1)}\right)}\right\} \\
&\times \min\left\{1, \frac{p_k(\varphi_2)\sum_{j=1}^n \bar{\omega}_j^{(k-1)} q_k\left(\varphi_1|\varphi_j^{(k-1)}\right) \min\left\{1, \frac{p_k(\varphi_1)}{p_k\left(\varphi_j^{(k-1)}\right)}\right\}}{p_k(\varphi_1)\sum_{j=1}^n \bar{\omega}_j^{(k-1)} q_k\left(\varphi_2|\varphi_j^{(k-1)}\right) \min\left\{1, \frac{p_k(\varphi_2)}{p_k\left(\varphi_j^{(k-1)}\right)}\right\}}\right\} \\
&= p_k(\varphi_2)\sum_{j=1}^n \bar{\omega}_j^{(k-1)} q_k\left(\varphi_1|\varphi_j^{(k-1)}\right) \min\left\{1, \frac{p_k(\varphi_1)}{p_k\left(\varphi_j^{(k-1)}\right)}\right\} \\
&\times \min\left\{1, \frac{p_k(\varphi_1)\sum_{j=1}^n \bar{\omega}_j^{(k-1)} q_k\left(\varphi_2|\varphi_j^{(k-1)}\right) \min\left\{1, \frac{p_k(\varphi_2)}{p_k\left(\varphi_j^{(k-1)}\right)}\right\}}{p_k(\varphi_2)\sum_{j=1}^n \bar{\omega}_j^{(k-1)} q_k\left(\varphi_1|\varphi_j^{(k-1)}\right) \min\left\{1, \frac{p_k(\varphi_1)}{p_k\left(\varphi_j^{(k-1)}\right)}\right\}}\right\} \\
&= p_k(\varphi_2)f_k(\varphi_1|\varphi_2).
\end{aligned} \tag{45}$$

This proves that $p_k(\varphi)$ is a stationary distribution of the AIMS-OPT Markov chain.

A stationary distribution is unique and is the limiting distribution for a Markov chain, if the chain is aperiodic and irreducible (see, for example, [36]). In the case of AIMS-OPT, aperiodicity is guaranteed by the fact that the probability of having a repeated sample $\varphi_{i+1}^{(k)} = \varphi_i^{(k)}$ is not zero. A Markov chain with stationary distribution $p(\varphi)$ is irreducible if, for any initial state, it has positive probability of entering any set to which $p(\varphi)$ assigns positive probability. It is clear that if the proposal distribution $q_k(\varphi|\xi)$ is "standard" (e.g., Gaussian, uniform, log-normal, etc.), then AIMS-OPT generates an irreducible Markov chain. In this case, $\pi_k(\varphi)$ is therefore the unique stationary distribution of the AIMS-OPT Markov chain.

References

[1] Au SK. Reliability based design sensitivity by efficient simulation. Comput Struct 2005;83:1048–61.
[2] Au SK, Beck JL. Estimation of small failure probabilities in high dimensions by subset simulation. Probab Eng Mech 2001;16(4):263–77.
[3] Au SK, Beck JL. Importance sampling in high dimensions. Struct Saf 2003;25(2):139–63.
[4] Beck JL. Bayesian system identification based on probability logic. Struct Contr Health Monit 2010;17:825–47.

[5] Beck JL, Taflanidis AA. Prior and posterior robust stochastic predictions for dynamical systems using probability logic. Int J Uncert Quantif 2013;3(4):271–288.
[6] Beck JL, Zuev KM. Asymptotically independent Markov sampling: a new Markov chain Monte Carlo scheme for Bayesian inference. Int J Uncert Quantif 2013;3(5):445–474.
[7] Cheung SH, Beck JL. Calculation of posterior probabilities for Bayesian model class assessment and averaging from posterior samples based on dynamic system data. J Comput Aided Civil Infrastruct Eng 2010;25(5):304–21.
[8] Eberhart RC, Kennedy J. A new optimizer using particle swarm theory. In: Proceedings of the sixth international symposium on micro machine and human science, 1995.
[9] Enevoldsen I, Sørensen JD. Reliability-based optimization of series systems of parallel systems. J Struct Eng 1993;119(4):1069–84.
[10] Gasser M, Schuëller GI. Reliability-based optimization of structural systems. Math Meth Operat Res 1997;46:287–307.
[11] Gelman A, Roberts GO, Gilks WR. Efficient Metropolis Jumping Rules. Bayesian Stat 1996;5:599–607.
[12] Geyer CJ. Estimation and optimization of function. In: Gilks W, Richardson S, Spiegelhalter D, editors. Markov Chain Monte Carlo in Practice. New York: Chapman and Hall; 1996. p. 241–58.
[13] Hastings WK. Monte Carlo sampling methods using Markov chains and their applications. Biometrika 1970;57(1):97–109.
[14] Holland JH. Adaptation in natural and artificial systems. Ann Arbor, Michigan: University of Michigan Press; 1975.
[15] Jaynes ET. Information theory and statistical mechanics". Phys Rev 1957;106:620–30.
[16] Jensen HA, Valdebenito M, Schuëller G. An efficient reliability-based optimization scheme for uncertain linear systems subject to general Gaussian excitation. Comput Methods Appl Mech Eng 2008;198:72–87.
[17] Jensen HA, Kusanovic D, Valdebenito M, Schuëller G. Reliability-based design optimization of uncertain stochastic systems: a gradient-based scheme. J Eng Mech 2011;138(1):60–70.
[18] Katafygiotis LS, Zuev KM. Estimation of small failure probabilities in high dimensions by Adaptive Linked Importance Sampling, COMPDYN 2007, Rethymno, Crete, Greece, 2007.
[19] Katafygiotis LS, Zuev KM. Geometric insight into the challenges of solving high-dimensional reliability problems. Probab Eng Mech 2008;23(2-3):208–18.
[20] Kennedy J, Eberhart RC. Particle swarm optimization. In: Proceedings of IEEE International Conference on Neural Networks, vol. 4, 1995, pp. 1942–1948.
[21] Kirkpatrick S, Gelatt CD, Vecchi MP. Optimization by simulated annealing. Science 1983;220(4598):671–80.
[22] Kong A, Liu JS, Wong WH. Sequential imputations and Bayesian missing data problems. J Am Stat Assoc 1994;89(425):278–88.
[23] Li HS, Au SK. Design optimization using Subset Simulation algorithm. Struct Saf 2010;32(6):384–92.
[24] Liang F. "Annealing evolutionary stochastic approximation Monte Carlo for global optimization. Stat Comput 2011;21:375–93.
[25] Liu JS. Metropolized independent sampling with comparison to rejection sampling and importance sampling. Stat Comput 1996;6(2):113–9.
[26] Liu JS. Monte Carlo Strategies is Scientific Computing. New York: Springer-Verlag; 2001.
[27] Maniezzo V, Dorigo M, Colorni N. The ant system: Optimization by a colony of cooperating agents. IEEE Trans Syst Man Cybernet B 1996;26(1):29–41.
[28] Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. Equation of state calculations by fast computing machines. J Chem Phys 1953;21(6):1087–92.
[29] Neal RM. Probabilistic Inference Using Markov Chain Monte Carlo Methods, Technical Report CRG-TR-93-1, Dept. of Computer Science, University of Toronto, 1993.
[30] Robert CP, Casella G. Monte Carlo Statistical Methods, second ed. Springer Texts in Statistics, 2004.
[31] Ruszczynski A, Shapiro A. Stochastic Programming. New York: Elsevier; 2003.
[32] Schuëller G, Jensen HA. Computational methods in optimization considering uncertainties – An overview. Comput Meth Appl Mech Eng 2008;198:2–13.
[33] Spall JC. Introduction to Stochastic Search and Optimization. New York: Wiley -Interscience; 2003.
[34] Taflanidis AA, Beck JL. Stochastic Subset Optimization for optimal reliability problems. Probab Eng Mech 2008;23:324–38.
[35] Taflanidis AA, Beck JL. An efficient framework for optimal robust stochastic system design using stochastic simulation. Comput Meth Appl Mech Eng 2008;198(1):88–101.
[36] Tierney L. Markov chains for exploring posterior distributions. Ann Stat 1994;22(4):1701–62.
[37] Vision 2000: Performance based Seismic Engineering of Buildings, Tech. rep., Structural Engineers Association of California, Sacramento, California, 2000.
[38] Yang CM, Beck JL. "Generalized trajectory methods for finding multiple extrema and roots of functions. J Optim Theory Appl 1998;97(1):211–27.
[39] Zuev KM, Beck JL, Au SK, Katafygiotis LS. Bayesian post-processor and other enhancements of Subset Simulation for estimating failure probabilities in high dimensions. Comput Struct 2012;92-93:283–96.