Proceedings of the 2013 Winter Simulation Conference
R. Pasupathy, S.-H. Kim, A. Tolk, R. Hill, and M. E. Kuhl, eds.

POPULATION MODEL-BASED OPTIMIZATION WITH SEQUENTIAL MONTE CARLO

Xi Chen
Enlu Zhou

Department of Industrial & Enterprise Systems Engineering
University of Illinois at Urbana-Champaign
Urbana, IL 61801, USA

ABSTRACT

Model-based optimization algorithms are effective for solving optimization problems with little structure. The algorithms iteratively find candidate solutions by generating samples from a parameterized probabilistic model on the solution space. In order to better capture the multi-modality of the objective function than the traditional model-based methods which use only a *single* model, we propose a framework of using a *population* of models with an adaptive mechanism to propagate the population over iterations. The adaptive mechanism is derived from estimating the optimal parameter of the probabilistic model in a Bayesian manner, and thus provides a proper way to determine the diversity in the population of the models. We develop two practical algorithms under this framework by applying sequential Monte Carlo methods, provide some theoretical justification on the convergence of the proposed methods, and carry out numerical experiments to illustrate their performance.

1 INTRODUCTION

We consider deterministic global optimization problems, where the objective functions have little structure, such as convexity and differentiability, and sometimes can only be assessed by "black box" evaluations. These problems have a wide range of applications and are usually difficult to solve. Stochastic search methods are often effective and promising in solving these problems. One class of stochastic search methods generate new candidate solutions from the neighborhood of the previous solutions, such as simulated annealing (Kirkpatrick et al. 1983, Aarts and Laarhoven 1989), genetic algorithms (Goldberg 1989), tabu search (Glover 1990), nested partitions (Shi and Ólafsson 2000), and sequential Monte Carlo simulated annealing (Zhou and Chen 2013). Another class of stochastic search methods, under the name of model-based methods (Zlochin et al. 2004), generate candidate solutions from a probabilistic model and update the parameter of the model based on the function evaluations of the previous candidate solutions. Examples of model-based methods include annealing adaptive search (Romeijn and Smith 1994, Zabinsky 2003), ant colony optimization (Dorigo and Gambardella 1997), estimation of distribution algorithms (Larranaga and Lozano 2002), cross-entropy method (Rubinstein and Kroese 2004), model reference adaptive search (Hu et al. 2007), and gradient-based adaptive stochastic search (Zhou and Hu 2013, Chen et al. 2013).

In the model-based methods listed above, only one *single* model is used to generate candidate solutions at each iteration. To better capture the multi-modality of the objective function, we may generate candidate solutions from a *population* of probabilistic models at each iteration. There has been very little work on population model-based methods, probably due to the difficulty of propagating multiple models and determining the number of samples to draw from each model. To our knowledge, Hu et al. (2011) is the only work using multiple models in model-based methods. They propose an approach with dynamic sample allocation, which aims at efficiently allocating the budget of samples among several models to achieve better performance. In this paper, we propose a new framework of population model-based optimization

(PMO) by converting the optimization problem to a parameter estimation problem, where we estimate the parameter of the optimal model that is a degenerate distribution concentrating on the optimal solution. The parameter is estimated in a Bayesian manner by tracking the posterior distribution of the parameter given some observations related to the objective function evaluations. In this way, a population of models are generated according to the posterior distribution of the parameter, and the diversity of the population is determined by the spread of the posterior distribution, which is in turn updated based on the function evaluations. In implementation, the parameter is estimated by sequential Monte Carlo (SMC) methods (Doucet et al. 2001) — a class of Monte Carlo methods that empirically approximate and track the posterior distribution of the unobserved state when noisy observations arrive sequentially in time. SMC methods were first introduced into model-based optimization by Zhou et al. (2008) and Zhou et al. (2013). In summary, the contributions of this paper include (1) a new framework of population model-based methods to better capture the shape of the objective function; (2) two practical algorithms: population model-based optimization with sequential Monte Carlo (PMO-SMC) and population model-based optimization with projection sequential Monte Carlo (PMO-PSMC); (3) theoretical justification on the convergence of the proposed methods.

The rest of the paper is organized as follows. In section 2, we introduce the basic idea of our proposed methods. In section 3, we formally propose the framework of PMO, provide the convergence analysis, and develop two numerical algorithms. We present the numerical results in section 4, and finally conclude the paper in section 5.

2 PROBLEM FORMULATION

We consider the global optimization problem:

$$x^* = \arg\max_{x \in \mathscr{X}} H(x), \tag{1}$$

where the solution space $\mathscr{X}$ is a nonempty compact set in $\mathbb{R}^n$, and n is the dimension of the problem. We assume there exists a unique $x^* \in \mathscr{X}$ such that $H(x) < H(x^*) = H^*$, $\forall x \neq x^*$, $x \in \mathscr{X}$. The objective function $H(\cdot)$: $\mathscr{X} \to \mathbb{R}$ is a deterministic real-valued bounded function on $\mathscr{X}$, i.e., there exists a lower bound $H^l > -\infty$ and an upper bound $H^u < \infty$ such that $H^l \leq H(x) \leq H^u$, for any $x \in \mathscr{X}$.

To solve problem (1), a model-based optimization method relies on a parameterized probabilistic model, i.e. a family of parameterized sampling distributions $\{f(\cdot,\theta)|\theta \in \Theta\}$, over the solution space $\mathscr{X}$, where Θ is a compact subset of $\mathbb{R}^m$ and m is the dimension of the parameter. These parameterized distributions characterize the belief about the promising regions of the solution space. A model-based optimization method at iteration k mainly consists of two steps: (1) generate candidate solutions from $f(\cdot,\theta_k)$; (2) compute the updated parameter $\theta_{k+1} \in \Theta$ for the sampling distribution of the next iteration based on the performance of the current candidate solutions. Under a proper parameter updating procedure, the sequence of sampling distributions $\{f(\cdot,\theta_k)\}$ will become more and more concentrated on the promising regions of the solution space. Ideally, the sequence of sampling distributions will eventually converge to $f(\cdot,\theta^*)$, where θ^* is the optimal parameter such that $f(\cdot,\theta^*)$ concentrates on the optimal solution x^*.

Motivated by the formulation of model-based methods in finding the optimal parameter θ^*, we may view the optimization problem as a parameter estimation problem that estimates the optimal parameter θ^* based on the function evaluations. We estimate the parameter in a Bayesian way by first introducing the following dynamic state-space model:

$$\begin{aligned} X_k &\sim f(\cdot;\theta_k), \\ Y_k &= H(X_k) - V_k. \end{aligned} \tag{2}$$

$X_k \in \mathscr{X}$ is the unobserved state that follows the distribution $f(\cdot;\theta_k)$ parameterized by the unknown parameter $\theta_k \in \Theta$. The true value of the unknown parameter θ_k is the optimal parameter θ^*, and thus the underlying value of the unobserved state X_k is the optimal solution x^*. Y_k is the noisy observation of the optimal function

value $H(x_k)$, which is equal to H^*, and V_k is the observation noise. In an optimization algorithm, Y_k comes from the function evaluations of candidate solutions and the distribution of V_k brings in randomization into the algorithm. To estimate the unknown parameter, we track the posterior distribution $b_k(\theta_k) \triangleq p(\theta_k|y_{1:k})$, where $y_{1:k}$ denotes the sequence of the received observations up to iteration k, i.e. $y_{1:k} = \{y_1, \cdots, y_k\}$ and y_k is a realization of the observation Y_k. As the iteration number increases, we gather more information about the true parameter and the true state value. With an appropriate choice of the observation and the distribution of the observation noise, the posterior distribution $b_k(\theta_k)$, which is our belief about the true parameter, will become more and more concentrated on the optimal parameter θ^*. The details about how to choose the observations and the distribution of the noise will be discussed in section 3.

Based on the above idea of parameter estimation, we propose a framework of population model-based optimization methods. At each iteration, the following three steps are carried out:

(1) Generate a population of probabilistic models according to $b_{k-1}(\theta_{k-1})$.
(2) Generate candidate solutions from the population of models yielded in step (1).
(3) Update the posterior distribution on the parameter to $b_k(\theta_k)$ based on the observation y_k, i.e., function evaluation at some candidate solution.

In this framework, the parameter is estimated in terms of the posterior distribution $b_k(\theta_k)$. This means we may get multiple samples of the parameter after sampling from its posterior distribution, and the diversity of the samples is determined by the spread of the posterior distribution. This provides a proper way to propagate the population of models. The use of a population of models helps to capture the shape of the objective function and distribute search in multiple promising regions of the solution space.

3 POPULATION MODEL-BASED OPTIMIZATION

3.1 Framework

As mentioned above in section 2, the optimization problem (1) can be viewed as a parameter estimation problem with the dynamic state-space model (2). Let the probability density function (p.d.f.) of the observation noise V_k be $\varphi(\cdot)$, and we have

$$p(y_k|x_k) = \varphi(H(x_k) - y_k). \tag{3}$$

Thus, the state-space model (2) can be represented in terms of distributions

$$\begin{aligned} X_k &\sim f(\cdot;\theta_k), \\ Y_k &\sim \varphi(H(x_k) - y_k). \end{aligned} \tag{4}$$

Based on the state-space model (4), we can solve the optimization problem by iteratively estimating the optimal parameter θ^* in a Bayesian manner by tracking the posterior distribution $b_k(\theta_k)$. One of the widely-used methods is to treat the unknown parameter as a component of the state vector, with the state equation

$$\theta_k = \theta_{k-1}, \tag{5}$$

where we abuse the notation θ to denote both the state and its realization. Now the state vector becomes (X_k, θ_k). Denote the joint posterior distribution of the state X and the parameter θ by

$$b_k(x_k, \theta_k) \triangleq p(x_k, \theta_k|y_{1:k}).$$

Thus, the posterior distribution of the parameter θ is

$$b_k(\theta_k) \triangleq p(\theta_k|y_{1:k}) = \int_{\mathscr{X}} b_k(x_k, \theta_k) dx_k. \tag{6}$$

We can estimate the optimal parameter in a Bayesian manner by tracking the posterior distribution $b_k(\theta_k)$.

Since the computation of $b_k(\theta_k)$ is usually analytically intractable, we use sequential Mont Carlo (SMC) methods to approximate the posterior distribution $b_k(\theta_k)$ in implementation. The issue of applying SMC to (5) is that there is no evolution on θ, so the candidate samples of θ will only be limited to the initial samples and may cause sample degeneracy. A pragmatic method to overcome this problem is to add an artificial diminishing noise Γ_k (Liu and West 2001):

$$\theta_k = \theta_{k-1} + \Gamma_k. \tag{7}$$

The noise should be small such that (7) does not differ too much from (5). By (7), the samples of the parameter evolve very slowly, and thus the algorithm has slow convergence rate. To accelerate the evolution on θ, we introduce a new method projection SMC for parameter estimation based on the work of Zhou et al. (2010) and Azimi-Sadjadi and Krishnaprasad (2005). This idea is to project $b_{k-1}(\theta_{k-1})$ onto a parameterized distribution $g(\cdot;\lambda_k)$ and generate new samples from $g(\cdot;\lambda_k)$. The details of SMC and projection SMC for parameter estimation will be provided in section 3.3.

In the following, we derive how to propagate $b_{k-1}(x_{k-1},\theta_{k-1})$ to $b_k(x_k,\theta_k)$. Denote

$$\tilde{b}_{k-1}(\theta_k) \triangleq p(\theta_k|y_{1:k-1}).$$

By adding artificial noise,

$$\tilde{b}_{k-1}(\theta_k) = \int_\Theta b_{k-1}(\theta_{k-1}) p(\theta_k|\theta_{k-1}) d\theta_{k-1}, \tag{8}$$

where the transition density $p(\theta_k|\theta_{k-1})$ is induced by the distribution of Γ_k and (7). By projection,

$$\tilde{b}_{k-1}(\theta_k) = g(\theta_k;\lambda_k). \tag{9}$$

Define

$$\tilde{b}_{k-1}(x_k,\theta_k) \triangleq p(x_k,\theta_k|y_{1:k-1}) = p(x_k|\theta_k)p(\theta_k|y_{1:k-1}) = f(x_k;\theta_k)\tilde{b}_{k-1}(\theta_k). \tag{10}$$

According to the Bayes rule and (3), the posterior distribution $b_k(x_k,\theta_k)$ can be expressed by

$$\begin{aligned} b_k(x_k,\theta_k) &= p(x_k,\theta_k|y_{1:k}) \\ &= \frac{p(y_k|x_k,\theta_k,y_{1:k-1})p(x_k,\theta_k,y_{1:k-1})}{p(y_{1:k})} \\ &= \frac{p(y_k|x_k)p(x_k,\theta_k|y_{1:k-1})}{p(y_k|y_{1:k-1})} \\ &= \frac{\varphi(H(x_k)-y_k)\tilde{b}_{k-1}(x_k,\theta_k)}{\int_{\mathscr{X}}\int_\Theta \varphi(H(x_k)-y_k)\tilde{b}_{k-1}(x_k,\theta_k)d\theta_k dx_k}. \end{aligned} \tag{11}$$

Thus, we have

$$b_k(x_k,\theta_k) \propto \varphi(H(x_k)-y_k)\tilde{b}_{k-1}(x_k,\theta_k). \tag{12}$$

Therefore, the posterior distributions are propagated by

$$b_{k-1}(x_{k-1},\theta_{k-1}) \implies b_{k-1}(\theta_{k-1}) \implies \tilde{b}_{k-1}(\theta_k) \implies \tilde{b}_{k-1}(x_k,\theta_k) \implies b_k(x_k,\theta_k).$$

In our proposed optimization algorithms, the noisy observation value y_k is the function evaluation at some candidate solution, and thus at the true value of the unobserved state X_k ($X_k = x^*$) the function value $H(x_k)$ is no less than the observation value y_k, i.e. $H(x_k) \geq y_k$. By (12), $b_k(x_k) \propto \varphi(H(x_k)-y_k)\tilde{b}_{k-1}(x_k)$, where $b_k(x_k)$ and $\tilde{b}_{k-1}(x_k)$ are marginal distributions of x_k associated with the joint distributions $b_k(x_k,\theta_k)$ and $\tilde{b}_{k-1}(x_k,\theta_k)$ respectively. Thus, the support of $b_k(x_k)$ is a subset of $\{x_k \in \mathscr{X} : H(x_k) \geq y_k\}$. To make

sure the support of $b_k(x_k)$ concentrating on more promising regions of the solution space as the iteration number increases, the observation sequence $\{y_k\}$ should be monotonically increasing. One way to obtain the observation is to choose from the $(1-\rho)$-quantile of $H(x)$ with respect to the posterior distribution $\tilde{b}_k(x)$. Denote the quantile by

$$\gamma_k \triangleq \sup_l \{l : P_{\tilde{b}_k}(x \in \mathscr{X} : H(x) \geq l) \geq \rho\},$$

where $P_{\tilde{b}_k}(\cdot)$ denotes the probability with respect to $\tilde{b}_k(x)$. To create an increasing observation sequence, we update $y_k = \gamma_k$, if $\gamma_k \geq y_{k-1} + \varepsilon$, where ε is a small positive constant, and keep the observation the same, i.e., $y_k = y_{k-1}$, otherwise. We use the $(1-\rho)$-quantile, since we want to keep searching the most promising regions of the solution space as well as exploring more to produce better estimation of θ. The parameter ρ serves as a trade-off parameter between exploitation and exploration. With small value of ρ, we exploit more on the current best estimation; with large value of ρ, we explore in a relatively larger area.

In summary, we propose the following framework for population model-based optimization.

Population Model-based Optimization (PMO)

1. Initialization: Set the initial density $b_0(\theta_0)$, and set $k = 1$.
2. Evolution: Obtain $\tilde{b}_{k-1}(\theta_k)$ based on perturbation by (8) or projection by (9) and $\tilde{b}_{k-1}(x_k, \theta_k)$ by (10).
3. Observation: For $k = 1$, $y_1 := \gamma_1$. For $k > 1$, if $\gamma_k \geq y_{k-1} + \varepsilon$, then set $y_k := \gamma_k$; else set $y_k := y_{k-1}$.
4. Updating: Compute $b_k(x_k, \theta_k)$ by (11).
5. Stopping: If a stopping criterion is satisfied, then stop; else, set $k = k+1$ and go to step 2.

3.2 Convergence Analysis

In this section, we show that, under some assumptions,

$$\lim_{k\to\infty} \mathbb{E}_{b_k}[H(X)] = \lim_{k\to\infty} \int_{\mathscr{X}} \int_{\Theta} H(x) b_k(x,\theta) d\theta dx = H^*. \tag{13}$$

In our formulation, we consider the case that $H(x)$ has a unique global optimal solution, thus (13) is equivalent to the fact that the marginal posterior distributions $b_k(x)$ and $b_k(\theta)$ are Dirac delta functions concentrated on the optimal solution x^* and optimal parameter θ^* as $k \to \infty$. We introduce the following assumptions, and the details for the proofs of the theorems in this section can be found in Chen and Zhou (2013).

Assumption 1 The p.d.f. $\varphi(\cdot)$ has support on $[0, H^u - H^l]$, and is continuous, positive, strictly increasing on its support.

Assumption 2 For any constant $H^c < H(x^*)$, the set $\{x \in \mathscr{X} : H(x) \geq H^c\}$ has a strictly positive Lebesgue measure.

Assumption 3 For any $x \in \mathscr{X}$, the parameterized density $f(x;\theta_k) > 0$ for all finite k.

Assumption 4 $\lim_{k\to\infty} \sum_{i=1}^{k} c_i < \infty$, where $c_k \triangleq \mathbb{E}_{b_k}[H(X)] - \mathbb{E}_{\tilde{b}_k}[H(X)]$.

Assumption 5 $\lim_{k\to\infty}(\tilde{b}_k(x) - b_k(x)) = 0$ almost everywhere in $\mathscr{X}$.

Assumption 1 is a general condition on the p.d.f. of the observation noise V_k. Since y_k is the objective function evaluation at some candidate solution, at the true value of the unobserved state X_k, i.e., $x_k = x^*$, we have $H(x_k) \geq y_k$. In addition, since $H(x_k) \leq H^u$ and $y_k \geq H^l$, the noise $V_k = H(X_k) - Y_k \in [0, H^u - H^l]$ by our formulation. With the strictly increasing property of $\varphi(\cdot)$, we have $\varphi(H(x_1) - y_k) > \varphi(H(x_2) - y_k)$, for any $x_1, x_2 \in \mathscr{X}$ satisfying $H(x_1) > H(x_2) \geq y_k$. This ensures that distribution b_k evolves to be more concentrated on the regions with larger function values. Assumption 2 ensures that the neighborhood of the optimal solution x^* has a positive probability to be sampled, and it is satisfied by many functions, such

as a continuous function. Assumption 3 can be satisfied by most of the parameterized distributions, such as normal distributions. Assumptions 4 and 5 can be considered as conditions on the magnitude of the perturbation on the parameter state or on the approximation accuracy of the density projection. Assumptions 4 and 5 require that the artificial noise Γ_k goes to zero or the error of the density projection goes to zero sufficiently fast as k goes to infinity.

Theorem 1 Under Assumptions 1-5, $\lim_{k\to\infty}\mathbb{E}_{b_k}[H(X)] = H^*$.

In the following, we show the convergence of PMO with perturbation on the parameter evolution under some more specific Assumptions 6-8, which can be considered as a special case that satisfies Assumptions 4 and 5. Since y_k is monotonically increasing and upper bounded by H^* and is updated only when $\gamma_k \geq y_{k-1} + \varepsilon$, there exists $K < \infty$ such that $y_k = y_K$, $\forall k \geq K$.

Assumption 6 The perturbation Γ_k is uniformly distributed on the support $[-\delta_k, \delta_k]^m$, where $\delta_k = \delta\alpha^k$, $\delta \geq 0$ and $0 \leq \alpha < \frac{\varphi(0)}{\varphi(H^u - y_K)}$.

Assumption 7 $b_k(\theta)$ is continuous on Θ, and differentiable on $int(\Theta)$, i.e., the interior of Θ.

Assumption 8 There exists a finite constant A, such that $\|\nabla_\theta b_0(\theta)\| \leq A$, $\forall\theta \in int(\Theta)$.

Assumption 6 restricts the magnitude of the perturbation Γ_k. Since $\varphi(\cdot)$ is a strictly increasing function on the support $[0, H^u - H^l]$, we have $\alpha < 1$. Thus, $\delta_k \searrow 0$ as $k \to \infty$. Diminishing perturbation allows more exploration at the beginning and more exploitation on searching the promising solution regions as iteration number increases. We use the uniform distribution for Γ_k because of its simple p.d.f. for further analysis. We can also use other distribution for Γ_k in practice as long as the magnitude of the perturbation diminishes and goes to 0, i.e. normal distributions with mean 0 and variance goes to 0 as $k \to 0$. For Assumption 7, the differentiability of $b_k(\theta)$ mainly depends on the differentiability of $f(x;\theta)$ and $p(\theta|\theta_k)$ with respect to θ. The uniform distribution used in Assumption 6 already ensures the differentiability of $p(\theta|\theta_k)$, and the differentiability of $f(x;\theta)$ is easily satisfied by many parameterized distributions, such as normal distributions. Assumption 8 restricts the initial setting of the distribution of the parameter, and it can be satisfied easily by many distributions.

Theorem 2 Under Assumptions 1-3 and 6-8, $\lim_{k\to\infty}\mathbb{E}_{b_k}[H(X)] = H^*$.

3.3 Implementations

In implementation, we apply sequential Monte Carlo (SMC) to estimate the unobserved state and the unknown parameter jointly based on the observations by tracking the posterior distribution. Given the initial samples $\left\{\left(x_0^i, \theta_0^i\right)\right\}_{i=1}^N$ (N is the sample size) that are i.i.d. from $b_0(x_0, \theta_0)$, SMC methods recursively propagate the pervious samples $\left\{\left(x_{k-1}^i, \theta_{k-1}^i\right)\right\}_{i=1}^N$ that is an empirical approximation of $b_{k-1}(x_{k-1}, \theta_{k-1})$ to samples $\left\{\left(x_k^i, \theta_k^i\right)\right\}_{i=1}^N$ that approximate the posterior distribution $b_k(x_k, \theta_k)$.

We first present how to propagate the samples by adding artificial noise on θ with state equation (7). From samples $\left\{\left(x_{k-1}^i, \theta_{k-1}^i\right)\right\}_{i=1}^N$, we can generate samples $\left\{\left(\tilde{x}_k^i, \tilde{\theta}_k^i\right)\right\}_{i=1}^N$ that empirically approximate $\tilde{b}_{k-1}(x_k, \theta_k)$ based on (8) and (10) by

$$\begin{aligned}\tilde{\theta}_k^i &\sim p(\theta_k|\theta_{k-1}^i), \quad i = 1, \cdots, N,\\ \tilde{x}_k^i &\sim f(x_k; \tilde{\theta}_k^i), \quad i = 1, \cdots, N.\end{aligned}$$

By importance sampling, (11) can be empirically approximated by

$$b_k(x_k, \theta_k) \approx \sum_{i=1}^N \frac{\varphi(H(\tilde{x}_k^i) - y_k)}{\sum_{i=1}^N \varphi(H(\tilde{x}_k^i) - y_k)} \delta((x_k, \theta_k) - (\tilde{x}_k^i, \tilde{\theta}_k^i)),$$

where $\delta(\cdot)$ is the Kronecker delta function. Let the normalized weight for sample $\left(\tilde{x}_k^i, \tilde{\theta}_k^i\right)$ be

$$W_k^i = \frac{\varphi(H(\tilde{x}_k^i) - y_k)}{\sum_{i=1}^N \varphi(H(\tilde{x}_k^i) - y_k)}, \quad i = 1, \cdots, N.$$

Then, the empirical approximation of $b_k(x_k, \theta_k)$ is

$$b_k^N(x_k, \theta_k) = \sum_{i=1}^N W_k^i \delta((x_k, \theta_k) - (\tilde{x}_k^i, \tilde{\theta}_k^i)).$$

The weights W_k^i are proportional to $\varphi(H(x_k^i) - y_k)$. The strictly increasing property of $\varphi(\cdot)$ assigns higher weights on the promising regions of the solutions, and specifically the optimal solution x^* has the strictly largest weight. Therefore, the posterior distribution $b_k(x_k, \theta_k)$ evolves to be more concentrated on the promising regions. We denote the samples and their associated weights by $\left\{\left(\tilde{x}_k^i, \tilde{\theta}_k^i\right), W_k^i\right\}_{i=1}^N$. Then, we can produce the samples with equal weights $\left\{\left(x_k^i, \theta_k^i\right), \frac{1}{N}\right\}_{i=1}^N$ approximately according to $b_k(x_k, \theta_k)$ from the weighted samples $\left\{\left(\tilde{x}_k^i, \tilde{\theta}_k^i\right), W_k^i\right\}_{i=1}^N$ by a resampling step. The resampling procedure is introduced to generate more samples with high weights and less samples with low weights, which helps concentrate more samples on the promising regions as well as avoiding the degeneracy problem of the weights. In summary, we propagate the samples as follows:

$$\left\{\left(x_{k-1}^i, \theta_{k-1}^i\right), \frac{1}{N}\right\}_{i=1}^N \implies \left\{\left(\tilde{x}_k^i, \tilde{\theta}_k^i\right), W_k^i\right\}_{i=1}^N \implies \left\{\left(x_k^i, \theta_k^i\right), \frac{1}{N}\right\}_{i=1}^N.$$

Therefore, given the samples according to $b_0(x_0, \theta_0)$, we may recursively generate random samples to empirically approximate the posterior distribution $b_k(x_k, \theta_k)$.

In projection SMC parameter estimation, we project the empirical posterior distribution $b_{k-1}^N(\theta_{k-1})$ onto a parameterized distribution $g(\cdot; \lambda_k)$, and generate samples $\left\{\theta_k^i\right\}_{i=1}^N$ from $g(\cdot; \lambda_k)$. The projection is conducted by minimizing the Kullback-Leibler (KL) divergence between these two distributions.

$$\lambda_k \triangleq arg\min_{\lambda} D_{KL}\left(b_{k-1}^N(\theta) \| g(\theta; \lambda)\right), \tag{14}$$

where

$$D_{KL}\left(b_{k-1}^N(\theta) \| g(\theta; \lambda)\right) \triangleq \int_\Theta \frac{b_{k-1}^N(\theta)}{g(\theta; \lambda)} b_{k-1}^N(\theta) d\theta.$$

KL divergence is used to measure the distance between two distributions. Lower KL divergence indicates that these two distributions are more similar. When $g(\cdot; \lambda_k)$ is an exponential family distribution, (14) admits an analytical solution. Then, we generate samples of the parameter and candidate solutions by

$$\theta_k^i \sim g(\theta_k; \lambda_k), \quad i = 1, \cdots, N,$$
$$x_k^i \sim f(x_k; \theta_k^i), \quad i = 1, \cdots, N.$$

By importance sampling and (11), the empirical distribution $b_k^N(\theta_k)$ that approximates $b_k(\theta_k)$ is

$$b_k^N(\theta_k) = \sum_{i=1}^N W_k^i \delta(\theta_k - \theta_k^i),$$

where the normalized weights are

$$W_k^i = \frac{\varphi(H(x_k^i) - y_k)}{\sum_{i=1}^N \varphi(H(x_k^i) - y_k)}, \quad i = 1, \cdots, N.$$

Therefore, we propagate the samples as follows:

$$\left\{\left(x_{k-1}^i, \theta_{k-1}^i\right), W_{k-1}^i\right\}_{i=1}^N \implies \left\{\left(x_k^i, \theta_k^i\right), W_k^i\right\}_{i=1}^N.$$

Using projection SMC for parameter estimation not only avoids adding artificial noise to the state of the parameter but also avoids the resampling step. Moreover, projection SMC may also save the computational time, since generating samples from the projected distribution helps to evolve the samples much faster than adding artificial noise.

Applying SMC and projection SMC to empirically approximate the posterior distribution $b_k(x_k, \theta_k)$ in PMO, we propose two numerical algorithms: population model-based optimization with sequential Monte Carlo (PMO-SMC) and population model-based optimization with projection sequential Monte Carlo (PMO-PSMC).

Algorithm 1 Population Model-based Optimization with Sequential Monte Carlo (PMO-SMC)

1. Initialization: Set initial density $b_0(\theta_0)$, and generate samples $\{\theta_0^i\}_{i=1}^N \sim b_0(\theta_0)$. Set $k = 1$.
2. Sampling: For $i = 1, \cdots, N$, generate sample $\tilde{\theta}_k^i \sim p_k(\theta_k | \theta_{k-1}^i)$ and sample $\tilde{x}_k^i \sim f(\cdot; \tilde{\theta}_k^i)$.
3. Observation: Set $\hat{\gamma}_k$ as $(1-\rho)$-quantile of $\{H(\tilde{x}_k^i)\}_{i=1}^N$. For $k = 1$, $y_1 := \hat{\gamma}_1$. For $k > 1$, if $\hat{\gamma}_k \geq y_{k-1} + \varepsilon$, then set $y_k := \hat{\gamma}_k$; else, set $y_k := y_{k-1}$.
4. Updating: Compute weights according to $W_k^i \propto \varphi(H(\tilde{x}_k^i) - y_k)$ and $\sum_{i=1}^N W_k^i = 1$, $i = 1, \cdots, N$.
5. Resampling: Draw samples $\left\{\left(x_k^i, \theta_k^i\right)\right\}_{i=1}^N$ from the empirical distribution $\left\{\left(\tilde{x}_k^i, \tilde{\theta}_k^i\right), W_k^i\right\}_{i=1}^N$.
6. Stopping: If a stopping criterion is satisfied, then stop; else, set $k = k + 1$ and go to step 2.

Algorithm 2 Population Model-based Optimization with Projection Sequential Monte Carlo (PMO-PSMC)

1. Initialization: Set initial density $b_0(\theta_0)$ and initial weights $\left\{W_0^i\right\}_{i=1}^N = \frac{1}{N}$. Generate samples $\{\theta_0^i\}_{i=1}^N \sim b_0(\theta_0)$. Set $k = 1$.
2. Projection: Project the empirical distribution $b_{k-1}^N(\theta_{k-1}) = \sum_{i=1}^N W_{k-1}^i \delta(\theta_{k-1} - \theta_{k-1}^i)$ to a parameterized distribution $g(\cdot; \lambda_k)$ by (14).
3. Sampling: For $i = 1, \cdots, N$, generate sample $\theta_k^i \sim g(\cdot; \lambda_k)$ and sample $x_k^i \sim f(\cdot; \theta_k^i)$.
4. Observation: Set $\hat{\gamma}_k$ as $(1-\rho)$-quantile of $\{H(x_k^i)\}_{i=1}^N$. For $k = 1$, $y_1 := \hat{\gamma}_1$. For $k > 1$, if $\hat{\gamma}_k \geq y_{k-1} + \varepsilon$, then set $y_k := \hat{\gamma}_k$; else, set $y_k := y_{k-1}$.
5. Updating: Compute weights according to $W_k^i \propto \varphi(H(x_k^i) - y_k)$ and $\sum_{i=1}^N W_k^i = 1$, $i = 1, \cdots, N$.
6. Stopping: If a stopping criterion is satisfied, then stop; else, set $k = k + 1$ and go to step 2.

4 NUMERICAL EXPERIMENTS

In this section, we test the performance of PMO-SMC and PMO-PSMC on some well-known continuous and unconstrained benchmark global optimization problems from Hu et al. (2007), and compare their performance with model reference adaptive search (MRAS) (Hu et al. 2007) and multi-start simulated annealing (SA). The problems we consider are listed below with their dimensions in the parentheses.

(1) Powell function (n=20)

$$H_1(x) = -1 - \sum_{i=2}^{n-2} \left[(x_{i-1} + 10x_i)^2 + 5(x_{i+1} - x_{i+2})^2 + (x_i - 2x_{i+1})^4 + 10(x_{i-1} - x_{i+2})^4\right],$$

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

(2) Rosenbrock function (n=10)

$$H_2(x) = -1 - \sum_{i=1}^{n-1} \left[100(x_{i+1} - x_i^2)^2 + (x_i - 1)^2\right],$$

where $x^* = (1, \cdots, 1)^T$, $H_2(x^*) = -1$.

(3) Griewank function (n=20)

$$H_3(x) = -\frac{1}{4000}\sum_{i=1}^{n} x_i^2 + \prod_{i=1}^{n} \cos\left(\frac{x_i}{\sqrt{i}}\right) - 1,$$

where $x^* = (0, \cdots, 0)^T$, $H_3(x^*) = 0$.

(4) Trigonometric function (n=20)

$$H_4(x) = -1 - \sum_{i=1}^{n} \left[8\sin^2(7(x_i - 0.9)^2) + 6\sin^2(14(x_i - 0.9)^2) + (x_i - 0.9)^2\right],$$

where $x^* = (0.9, \cdots, 0.9)^T$, $H_4(x^*) = -1$.

Specifically, Powell (H_1) and Rosenbrock (H_2) are badly-scaled functions; Griewank (H_3) and Trigonometric (H_4) are high-dimensional multi-modal problems with a large number of local optima, and the number of local optima increases exponentially with problem dimension.

In PMO-SMC and PMO-PSMC, we use independent multivariate normal distribution as the parameterized distributions $f(\cdot;\theta_k)$, where $\theta_k = (\mu_k, \sigma_k^2)$ and k is the iteration number. In the experiment, the initial mean μ_0 and the initial standard deviation σ_0 are chosen randomly according to the uniform distribution on $[-50, 50]^n$ and $[0, 50]^n$ respectively. The sample size is $N = 1000$, the quantile parameter ρ is set to be 0.1, and $\varepsilon = 10^{-10}$. Let the p.d.f. of the observation noise $\varphi(\cdot)$ be

$$p(y_k|x_k) = \varphi(H(x_k) - y_k) \propto (H(x_k) - y_k)\mathbb{I}_{\{H(x_k) \geq y_k\}},$$

where $\mathbb{I}_{\{\cdot\}}$ is the indicator function. This choice of $\varphi(\cdot)$ ensures that it is a strictly increasing function on its support. In PMO-SMC, the artificial noise Γ_k is uniformly distributed on $[-\delta_k, \delta_k]^{2n}$, where $\delta_k = \delta\alpha^k$, $\delta = 20$, and $\alpha = 0.995$. With the diminishing noise, the algorithm allows more exploration at the early iterations and more exploitation later. The noise parameter α acts as the trade-off parameter between explorative and exploitative search. In PMO-PSMC, we use independent normal distribution $\mathcal{N}(\mu_\lambda, \Sigma_\lambda)$, where $\Sigma_\lambda = diag(\sigma_\lambda^2)$, as the projection distribution $g(\cdot;\lambda)$ for the parameter θ.

For comparison, we also solve the above benchmark problems with MRAS and multi-start SA. For MRAS, the parameterized exponential family distribution $f(\cdot;\theta_k)$ is also chosen to be independent multivariate normal distributions $\mathcal{N}(\mu_k, \Sigma_k)$. The initial mean μ_0 is generated randomly according to the uniform distribution on $[-50, 50]^n$, and the initial covariance matrix is set to be $\Sigma_0 = 50^2 I_{n\times n}$. The sample size is chosen to be $N = 1000$ and the quantile parameter is $\rho = 0.1$, which are set to be the same as in PMO-SMC and PMO-PSMC. In the implementation, we apply the smoothing parameter updating procedure (Rubinstein and Kroese 2004) to prevent premature convergence. The smoothing parameter is chosen to be $\nu = 0.2$, which is found to work well by trial and error in experiments. Multi-start SA, a naive population-based simulated annealing method, runs simulated annealing independently from multiple initial candidate solutions and picks the best result among these independent runs as the final solution. In the experiment, the initial candidate solutions are chosen according to the uniform distribution on $[-50, 50]^n$, and the sample size is the same as in other methods, $N = 1000$. The initial temperature is $T_0 = 5 \times 10^6$, and the temperature is updated by geometric form $T_k = T_0 r^k$, with reduction parameter $r = 0.995$. The new candidate solution around the point x_k^i is generated by $\mathcal{N}(x_k^i, I_{n\times n})$.

We run each of these four methods 50 times independently, and compare the average optimal values. The average performance is shown in Table 1, where H^* is the true optimal value of $H(\cdot)$, $\bar{H}^*$ is the average optimal value of 50 runs, *std_err* is the standard error of the optimal function values, and P_ε is the percentage of ε-optimal solutions out of 50 runs (ε-optimal solution means the optimal value is within ε difference from the true optimal value H^*). We consider $\varepsilon = 0.01$ in our experiments. In Figure 1, we plot

the average best function values with respect to the total number of function evaluations. The comparison is based on similar computational effort, since the function evaluation dominates the computational time for all these four algorithms.

Table 1: Performance Comparison of PMO-SMC, PMO-PSMC, MRAS and multi-start SA

		PMO-SMC		PMO-PSMC		MRAS		multi-start SA	
	H^*	$\bar{H}^*(std\_err)$	P_ε	$\bar{H}^*(std\_err)$	P_ε	$\bar{H}^*(std\_err)$	P_ε	$\bar{H}^*(std\_err)$	P_ε
H_1	-1	-1(1.07E-5)	100%	-1(1.10E-4)	100%	-1(1.16E-11)	100%	-413.5(10.69)	0%
H_2	-1	-1.041(0.0022)	0%	-8.483(0.0164)	0%	-7.367(1.172)	0%	-38.3(0.876)	0%
H_3	0	-0.0022(5.69E-4))	100%	0(0)	100%	-0.0160(0.003)	56%	-0.277(0.0049)	0%
H_4	-1	-1(5.57E-6)	100%	-1(2.15E-17)	100%	-1(5.09E-16)	100%	-79.65(0.694)	0%

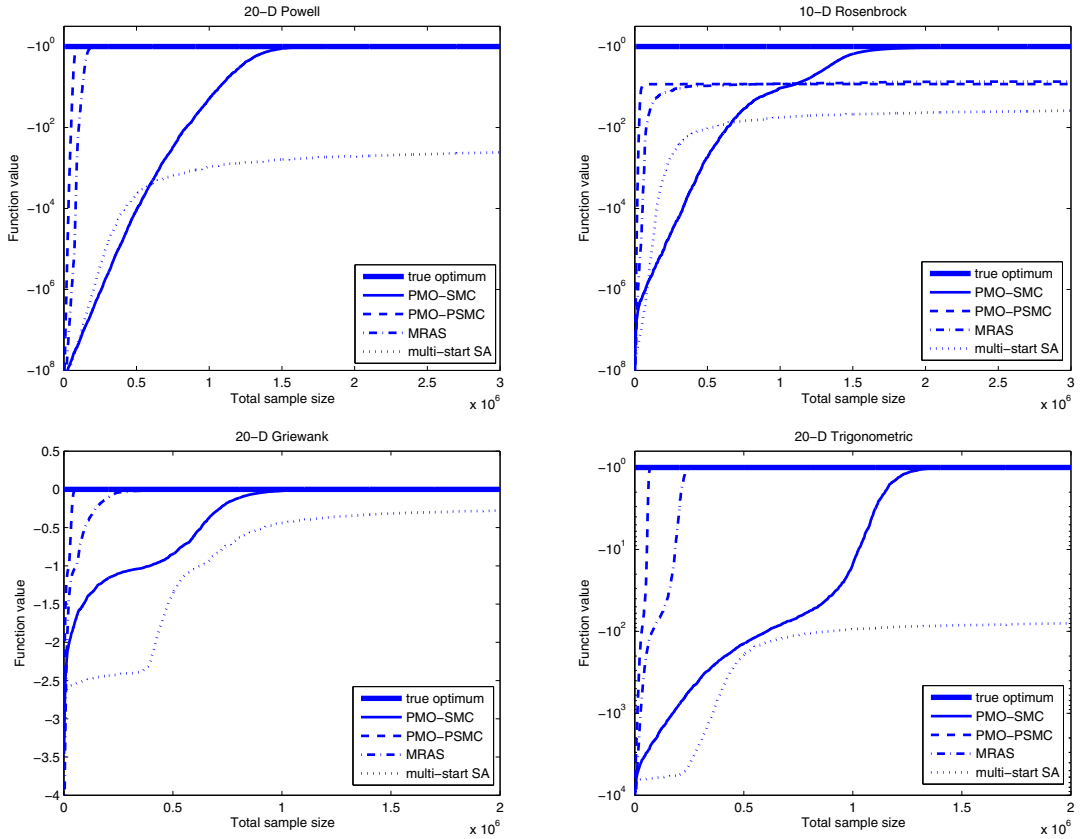

Figure 1: Average Performance of PMO-SMC, PMO-PSMC, MRAS and multi-stat SA

From the results, PMO-SMC and PMO-PSMC find the ε-optimal solutions in all the runs for all the test problems except problem H_2. For MRAS, the accuracy rate is 100% only for problems H_1 and H_4. Multi-start SA dose not provide ε-optimal solutions in any of the test problem. In terms of convergence speed, PMO-PSMC converges faster than MRAS, and MRAS converges faster than PMO-SMC in all the test problems.

5 CONCLUSION

In this paper, we have proposed a framework of population model-based optimization methods, where candidate solutions are generated from a population of models at each iteration. We view the original optimization problem as a parameter estimation problem that estimates the optimal parameter of the probabilistic model.

The parameter estimation is conducted in a Bayesian manner by iteratively approximating the posterior distribution of the parameter given the observations regarding the objective function evaluations, and thus the diversity of the models is determined by the spread of the posterior distribution. Under this framework, we have proposed two practical algorithms, PMO-SMC and PMO-PSMC. Numerical experiments on several benchmark problems have shown their promising performance compared to some other existing stochastic search methods.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation under Grant CMMI-1130273, and Air Force Office of Scientific Research under YIP Grant FA-9550-12-1-0250.

REFERENCES

Aarts, E., and P. V. Laarhoven. 1989. "Simulated Annealing: An Introduction". *Statistica Neerlandica* 43 (1): 43–52.

Azimi-Sadjadi, B., and P. S. Krishnaprasad. 2005. "Approximate Nonlinear Filtering and its Application in Navigation". *Automatica* 41 (6): 945–956.

Chen, X., and E. Zhou. 2013. "Population Model-based Optimization". Working paper, University of Illinois at Urbana-Champaign.

Chen, X., E. Zhou, and J. Hu. 2013. "Discrete Optimization via Gradient-based Adaptive Stochastic Search Methods". Under review.

Dorigo, M., and L. M. Gambardella. 1997. "Ant Colony System: A Cooperative Learning Approach to the Traveling Salesman Problem". *IEEE Transactions on Evolutionary Computation* 1:53–66.

Doucet, A., J. F. G. deFreitas, and N. J. Gordon. (Eds.) 2001. *Sequential Monte Carlo Methods In Practice*. New York: Springer.

Glover, F. W. 1990. "Tabu Search: A Tutorial". *Interfaces* 20:74–94.

Goldberg, D. E. 1989. *Genetic Algorithms in Search, Optimizaion and Machine Learning*. Boston, MA: Addison-Wesley Longman Publishing Co., Inc.

Hu, J., H. S. Chang, M. C. Fu, and S. I. Marcus. 2011. "Dynamic Sample Budget Allocation in Model-based Optimization". *Journal of Global Optimization* 50:575–596.

Hu, J., M. C. Fu, and S. I. Marcus. 2007. "A Model Reference Adaptive Search Method for Global Optimization". *Operations Research* 55:549–568.

Kirkpatrick, S., C. D. Gelatt, and J. M. P. Vecchi. 1983. "Optimization by Simulated Annealing". *Science* 220:671–680.

Larranaga, P., and J. A. Lozano. 2002. *Estimation of Distribution Algorithms: a New Tool for Evolutionary Computation*. Boston, MA: Kluwer Academic Publishers.

Liu, J., and M. West. 2001. "Combined Parameter and State Estimation in Simulation-Based Filtering". In *Sequential Monte Carlo Methods in Practice*, edited by A. Doucet, J. F. G. deFreitas, and N. J. Gordon. New York: Springer-Verlag.

Romeijn, H. E., and R. L. Smith. 1994. "Simulated Annealing and Adaptive Search in Global Optimization". *Probability in the Engineering and Informational Sciences* 8:571–590.

Rubinstein, R. Y., and D. P. Kroese. 2004. *The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation, and Machine Learning*. New York: Springer-Verlag.

Shi, L., and S. Ólafsson. 2000. "Nested Partitions Method for Global Optimization". *Operations Research* 48 (3): 390–407.

Zabinsky, Z. B. 2003. *Stochastic Adaptive Search for Global Optimization*. Nonconvex Optimization and Its Applications. Springer.

Zhou, E., and X. Chen. 2013. "Sequential Monte Carlo Simulated Annealing". *Journal of Global Optimization* 55 (1): 101–124.

Zhou, E., M. C. Fu, and S. I. Marcus. 2008. "A Paricle Filtering Framework for Randomized Optmization Algorithms". In *Proceedings of the 2008 Winter Simulation Conference*, 647–654.

Zhou, E., M. C. Fu, and S. I. Marcus. 2010. "Solving Continuous-state POMDPs via Density Projection". *IEEE Transactions on Automatic Control* 55 (5): 1101–1116.

Zhou, E., M. C. Fu, and S. I. Marcus. 2013. "A Particle Filtering Framework for Randomized Optimization Algorithms". in revision.

Zhou, E., and J. Hu. 2013. "Gradient-Based Adaptive Stochastic Search for Non-Differentiable Optimization". Under review.

Zlochin, M., M. Birattari, N. Meuleau, and M. Dorigo. 2004. "Model-based Search for Combinatorial Optimization: A Critical Survey". *Annals of Operations Research* 131:373–395.

AUTHOR BIOGRAPHIES

XI CHEN is a Ph.D. candidate in the Department of Industrial & Enterprise Systems Engineering at the University of Illinois at Urbana-Champaign. Her research interests lie in the broad areas of simulation optimization. Her email address is xchen37@illinois.edu.

ENLU ZHOU is an Assistant Professor in the Department of Industrial & Enterprise Systems Engineering at the University of Illinois at Urbana-Champaign. She received the B.S. degree with highest honors in electrical engineering from Zhejiang University, China, in 2004, and the Ph.D. degree in electrical engineering from the University of Maryland, College Park, in 2009. Her research interests include stochastic control and simulation optimization, with applications towards Financial engineering. Her email address is enluzhou@illinois.edu.