SMC SAMPLERS FOR BAYESIAN OPTIMAL NONLINEAR DESIGN

Hendrik Kück, Nando de Freitas and Arnaud Doucet

University of British Columbia

ABSTRACT

Experimental design is a fundamental problem in science. It arises in the planning of medical trials, sensor network deployment and control as well as in costly data gathering in physics, chemistry and biology. Bayesian decision theory provides a principled way of treating this problem, but leads to an intractable joint optimization and integration problem. Here, we propose a viable solution to this hard computational problem using sequential Monte Carlo samplers.

1. PROBLEM FORMULATION

We assume that we have a measurement model $p(y|\theta, d)$ of experimental outcomes $y \in \mathcal{Y}$ given a design d as well as a prior $p(\theta)$ on the model parameters $\theta \in \Theta$. The prior could be based on expert knowledge or previous experiments.

The goal is then to choose the optimal design $d^* \in \mathbb{R}^p$, which maximizes the expected utility

$$U(d) = \iint p(\theta)p(y|\theta, d)\, u(y, d, \theta)\, dy\, d\theta \qquad (1)$$

with respect to some measure of utility $u(y, d, \theta)$. When the model parameters are the objects of interest, the negative posterior entropy is commonly chosen as the utility function. That is, one aims to maximize

$$U(d) = \iiint p(\theta)p(y|\theta, d)\, [p(\theta'|y, d) \log p(\theta'|y, d)]\, d\theta' dy\, d\theta.$$

As shown in [1], under the assumptions of stationarity and standard bounds on distributions, this criterion is equivalent to maximizing the marginal entropy of the outcome y

$$U(d) = C - \int p(y|d) \log p(y|d)\, dy, \qquad (2)$$

where C is an arbitrary constant. This transformation reduces the complexity by eliminating one parameter space integral.

2. PREVIOUS WORK

The joint optimization and nested integration problem in equation (2) is computationally challenging. For this reason, most research has focused on the simple linear-normal model, for which closed form solutions exist [2]. However, many design problems encountered in practice are inherently nonlinear. One could linearize around a point estimate $\hat{\theta}$, but this crude approximation often leads to sub-optimal designs.

Another strategy involves discretizing the decision space $\mathbb{R}^p$ and approximating the integrals with direct Monte Carlo methods [3]. However, this approach is expensive and inadequate for high dimensional design spaces.

To eliminate the need for discretizing the decision space, Müller et al. [4] proposed a Markov chain Monte Carlo annealing technique for simultaneous maximization and integration. They define the following artificial target distribution

$$\pi_J(d, \theta_{1:J}, y_{1:J}) \propto \prod_{j=1}^{J} p(y_j, \theta_j|d)\, u(d, \theta_j, y_j). \qquad (3)$$

It is easy to show that this distribution admits $U(d)^J$ (with $U(d)$ as defined in Equation (1)) as its marginal distribution. That is, as J increases the samples concentrate on the modes of $U(d)$.

In [5], this idea has been extended with particle filtering. The intuition here is that interacting multiple chains can provide better exploration of distributions with distant modes. Both approaches however resort to sampling outcomes y^t independently of outcomes y^{t-1} at the previous iteration. This independent sampling is well known to be inefficient [6]. Furthermore, the annealing can only proceed in integer steps.

3. SMC SAMPLERS APPROACH

We adopt the sequential Monte Carlo (SMC) samplers framework of [7]. Our approach is based in particular on the application of SMC samplers to marginal parameter estimation presented in [8], where Müller's algorithm is generalized to non integer annealing steps. In contrast to the algorithms mentioned in the previous section, SMC samplers also enable us to replace the independent proposal distributions with more sophisticated and efficient proposal mechanisms.

SMC samplers [7] are a generalization of SMC methods such as particle filtering. They facilitate efficient sampling from sequences of distributions $\{\pi_t\}_{t\in\mathbb{N}^+}$. The distributions can be defined on the same or different spaces, but subsequent distributions should be close in the sense that efficient proposals for sampling from π_t can be constructed based on samples

from π_{t-1}. The key idea is to define an artificial joint distribution $\tilde{\pi}_t$ on a space of increasing dimension, which admits the distribution of interest π_t as its marginal. More specifically,

$$\tilde{\pi}_t(\mathbf{x}_{1:t}) = \pi_t(x_t) \prod_{i=t-1}^{1} L_i(x_{i+1}, x_i), \tag{4}$$

where L_i is an appropriate backward Markov kernel. Standard SMC methods can be used to sample from this extended growing distribution. Typically, this is done in the framework of sequential importance sampling with resampling. In doing so, new particle locations are proposed according to a forward Markov kernel K_t. These particles are weighted recursively as follows:

$$\frac{w_t}{w_{t-1}} \propto \frac{\pi_t(x^t)}{\pi_{t-1}(x^{t-1})} \frac{L_{t-1}\left(x^t, x^{t-1}\right)}{K_t\left(x^{t-1}, x^t\right)} \tag{5}$$

It is important to carefully choose the kernels K and L in order to achieve good mixing properties and keep the variance of the importance weights small.

4. SMC SAMPLERS FOR MAXIMUM ENTROPY SAMPLING

Motivated by [4] and [8], we define the following artificial target distribution:

$$\pi_{n_t,\nu_t}(\mathbf{y}_{1:n_t}, d) \propto p(d) \left(\prod_{j=1}^{n_t-1} \phi(y_j, d) \right) \phi(y_{n_t}, d)^{\nu_t}$$
$$\text{with } \phi(y_j, d) = p(y_j|d) \left(C - \log p(y_j|d)\right).$$

Unlike the extended target distribution of Equation (3), this distribution allows for real as well as integer annealing steps. The annealing is controlled by $n_t \in \mathbb{N}$ and $\nu_t \in [0, 1]$. In particular, n_t determines the discrete number of simulations of the outcomes y and ν_t is a real valued annealing factor for the last outcome. Both n_t and ν_t are driven by schedules depending on the index t. We restrict ourselves to schedules in which $n_t + \nu_t$ is monotonically increasing and assume that $n_t \leq n_{t-1} + 1$. That is, at most one new outcome is introduced per iteration. Every time the schedule passes through $\nu_t = 1$, the target $\pi_{n_t,1}(d, \mathbf{y}_{1:n_t})$ admits $p(d)\, U(d)^{n_t}$ as its marginal distribution (with $U(d)$ as defined in Equation (2)). The extended artificial distributions with rational values of ν_t are convenient, because they provide a smooth bridge between the distributions of interest, but their marginal distributions are meaningless. Ideally we would want to use a completely uninformative prior $p(d)$ on the design, but in order to ensure that the target is proper, it may be necessary to choose a uniform prior $p(d)$ on a finite interval of the design space instead. We further choose the constant C to be an upper bound on the log marginal likelihood to ensure that $\phi(y_j, d) \geq 0$.

Having defined the target distribution, the incremental importance weights follow from Equation (5):

$$\frac{\pi_{n_t,\nu_t}\left(d^t, \mathbf{y}^t_{1:n_t}\right) L_{t-1}\Big(\left(\mathbf{y}^t_{1:n_t}, d^t\right), \left(\mathbf{y}^{t-1}_{1:n_{t-1}}, d^{t-1}\right)\Big)}{\pi_{n_{t-1},\nu_{t-1}}\left(d^{t-1}, \mathbf{y}^{t-1}_{1:n_{t-1}}\right) K_t\Big(\left(\mathbf{y}^{t-1}_{1:n_{t-1}}, d^{t-1}\right), \left(\mathbf{y}^t_{1:n_t}, d^t\right)\Big)}$$

In [8, Algorithm 3], the forward kernel is the product of a Metropolis-Hastings kernel $\mathcal{K}_{n_t-1,1}$ with invariant distribution $\pi_{n_t-1,1}$, which updates the previously added outcomes $y_{1:n_t-1}$, and an independent proposal distribution q_{ν_t} for the outcome y_{n_t}. In mathematical terms,

$$K_t\Big(\left(\mathbf{y}^{t-1}_{1:n_{t-1}}, d^{t-1}\right), \left(\mathbf{y}^t_{1:n_t}, d^t\right)\Big) =$$
$$\mathcal{K}_{n_t-1,1}\Big(\left(\mathbf{y}^{t-1}_{1:n_t-1}, d^{t-1}\right), \left(\mathbf{y}^t_{1:n_t-1}, d^t\right)\Big) q_{\nu_t}(y^t_{n_t}|d^t)$$

Choosing the following backward kernel

$$L_{t-1}\Big(\left(\mathbf{y}^t_{1:n_t}, d^t\right), \left(\mathbf{y}^{t-1}_{1:n_{t-1}}, d^{t-1}\right)\Big) =$$
$$\mathcal{K}_{n_t-1,1}\Big(\left(\mathbf{y}^{t-1}_{1:n_t-1}, d^{t-1}\right), \left(\mathbf{y}^t_{1:n_t-1}, d^t\right)\Big) \frac{\pi_{n_{t-1},\nu_{t-1}}\left(d^{t-1}, \mathbf{y}^{t-1}_{1:n_t-1}\right)}{\pi_{n_t-1,1}\left(d^t, \mathbf{y}^t_{1:n_t-1}\right)}$$

leads to the incremental weights

$$\frac{w_t}{w_{t-1}} \propto \frac{\phi\left(y^t_{n_t}, d^t\right)^{\nu_t}}{q_{\nu_t}\left(y^t_{n_t}\right)}.$$

Note that if $\mathcal{K}_{n_t-1,1}$ was not a kernel with the correct invariant distribution, the choice of backward kernel above would be invalid. This is because the artifical joint distribution $\tilde{\pi}_t$ in Equation (4) then would no longer admit π_t as its marginal.

The full SMC sampler corresponding to the above choices for K and L is given in Algorithm 1.

To implement the MH kernel $\mathcal{K}_{n_t-1,1}$ we need to evaluate $\pi_{n_t-1,1}$, which in turn requires evaluation of $p(y|d)$. To this end, we draw a large set of samples $\{\theta_j\}_{j=1}^K$ from $p(\theta)$ and use the approximation

$$p(y|d) \approx \frac{1}{K} \sum_{j=1}^{K} p(y|\theta_j). \tag{6}$$

Algorithm 1 has two important shortcomings. First, its performance depends critically on the choice of the proposal distribution q_{ν_t}. We found empirically that the marginal distribution $p(y|d)$ behaves well as proposal when $\nu_t = 1$. Sampling from $p(y|d)$ as approximated in Equation (6) is straightforward for many problems. However, for $\nu_t < 1$, we noticed that $p(y|d)$ is no longer a good proposal and even heavy-tailed and adaptive proposal distributions were not sufficient to keep the variance of the importance weights small.

The second shortcoming is that the outcome y_{n_t} is sampled independently at each annealing step. It is more reasonable to only sample a new outcome y_{n_t} independently when it first gets introduced, that is when $n_t = n_{t-1} + 1$.

Initialization, $t = 1$

- Sample $\left\{\left(d^{(i),1}, y_1^{(i),1}\right)\right\}_{i=1}^{N}$ from proposal $q(\cdot)$
- Compute weights $w_1^{(i)} \propto \frac{\phi(d^{(i),1}, y_1^{(i),1})}{q(d^{(i),1}, y_1^{(i),1})}$
- Resample if ESS $<$ Threshold.

At time $t = 2, 3, \ldots$

- For each particle $i = 1, \ldots, N$:
  - Sample $\left(d^{(i),t}, \mathbf{y}_{1:n_t-1}^{(i),t}\right) \sim \mathcal{K}_{n_t-1,1}\left(\left(d^{(i),t-1}, \mathbf{y}_{1:n_t-1}^{(i),t-1}\right), \left(d^{(i),t}, \mathbf{y}_{1:n_t-1}^{(i),t}\right)\right)$
  - Sample $y_{n_t}^{(i),t}$ from proposal $q_{\nu_t}(\cdot)$
  - Compute incremental weights $\frac{w_t^{(i)}}{w_{t-1}^{(i)}} \propto \frac{\phi\left(y_{n_t}^{(i),t}, d^{(i),t}\right)^{\nu_t}}{q_{\nu_t}\left(y_{n_t}^{(i),t}\right)}$
- Resample if ESS $<$ Threshold.

Algorithm 1: SMC sampler for optimal design. This algorithm corresponds to Algorithm 3 in [8] but the notation is simplified thanks to reasonable assumptions on the schedule. The *effective sample size (ESS)* is a standard measure of the efficiency of the set of particles[9].

To overcome these shortcomings, we propose a different target distribution

$$\tilde{\pi}_{n_t,\nu_t}(d, \mathbf{y}_{1:n_t}) \propto p(d) \left(\prod_{j=1}^{n_t-1} p(y_j|d)\left[C - \log p(y_j|d)\right]\right) \times p(y_{n_t}|d)\left(C - \log p(y_{n_t}|d)\right)^{\nu_t}$$

This distribution still has the same desired property that for $\nu_t = 1$ it admits $p(d)\,U(d)^{n_t}$ as its marginal distribution.

We define different forward and backward kernels for steps in which the number of simulated outcomes increases ($n_t = n_{t-1}+1$) and for those in which it stays constant ($n_t = n_{t-1}$). Let us first look at the case $n_t = n_{t-1}$. As the forward kernel K_t we use a MCMC kernel $\tilde{\mathcal{K}}_{n_t,\nu_t}$ with invariant distribution $\tilde{\pi}_{n_t,\nu_t}$. We further choose the backward kernel as

$$L_{t-1}\left(\left(\mathbf{y}_{1:n_t}^t, d^t\right), \left(\mathbf{y}_{1:n_t}^{t-1}, d^{t-1}\right)\right) = \frac{\tilde{\pi}_{n_t,\nu_t}\left(\mathbf{y}_{1:n_t}^{t-1}, d^{t-1}\right)\tilde{\mathcal{K}}_{n_t,\nu_t}\left(\left(\mathbf{y}_{1:n_t}^{t-1}, d^{t-1}\right), \left(\mathbf{y}_{1:n_t}^{t}, d^{t}\right)\right)}{\tilde{\pi}_{n_t,\nu_t}\left(\mathbf{y}_{1:n_t}^{t}, d^{t}\right)}$$

These kernel choices result in the incremental weights

$$\frac{w_t}{w_{t-1}} \propto \frac{\tilde{\pi}_{n_t,\nu_t}\left(y_{n_t}^{t-1}, d^{t-1}\right)}{\tilde{\pi}_{n_t,\nu_{t-1}}\left(y_{n_t}^{t-1}, d^{t-1}\right)} = \left(C - \log p(y_{n_t}|d)\right)^{\nu_t-\nu_{t-1}}.$$

For steps with $n_t = n_{t-1} + 1$ (and $\nu_t = 0$ and $\nu_{t-1} = 1$), we adopt the forward kernel

$$K_t\left(\left(\mathbf{y}_{1:n_{t-1}}^{t-1}, d^{t-1}\right), \left(\mathbf{y}_{1:n_t}^{t}, d^{t}\right)\right) = \tilde{\mathcal{K}}_{n_{t-1},1}\left(\left(\mathbf{y}_{1:n_t-1}^{t-1}, d^{t-1}\right), \left(\mathbf{y}_{1:n_t-1}^{t}, d^{t}\right)\right) p(y_{n_t}^t|d^t)$$

together with the backward kernel

$$L_{t-1}\left(\left(\mathbf{y}_{1:n_t}^t, d^t\right), \left(\mathbf{y}_{1:n_{t-1}}^{t-1}, d^{t-1}\right)\right) = \tilde{\mathcal{K}}_{n_{t-1},1}\left(\left(\mathbf{y}_{1:n_t-1}^{t-1}, d^{t-1}\right), \left(\mathbf{y}_{1:n_t-1}^{t}, d^{t}\right)\right) \frac{\tilde{\pi}_{n_{t-1},1}\left(\mathbf{y}_{1:n_t-1}^{t-1}, d^{t-1}\right)}{\tilde{\pi}_{n_{t-1},1}\left(\mathbf{y}_{1:n_t-1}^{t}, d^{t}\right)}.$$

These kernels lead to the incremental weights

$$\frac{w_t}{w_{t-1}} \propto \frac{\tilde{\pi}_{n_t,0}\left(\mathbf{y}_{1:n_t}^t, d^t\right)}{\tilde{\pi}_{n_{t-1},1}\left(\mathbf{y}_{1:n_t-1}^t, d^t\right) p(y_{n_t}^t|d^t)} = \frac{p(y_{n_t}^t|d^t)}{p(y_{n_t}^t|d^t)} = 1.$$

That is, the weights do not need to be updated during these steps. The resulting sampler is described in Algorithm 2.

Initialization, $t = 1$

- For each particle $i = 1, \ldots, N$:
  - Sample $d^{(i),1} \sim p(d)$
  - Sample $y_1^{(i),1} \sim p(y|d^{(i),1})$
  - Initialize weight $w_1^{(i)} = \frac{1}{N}$

At time $t = 2, 3, \ldots$

- For each particle $i = 1, \ldots, N$:

If $n_t = n_{t-1}$:

- Sample $\left(d^{(i),t}, \mathbf{y}_{1:n_t}^{(i),t}\right) \sim \tilde{\mathcal{K}}_{n_t,\nu_t}\left(\left(d^{(i),t-1}, \mathbf{y}_{1:n_t}^{(i),t-1}\right), \left(d^{(i),t}, \mathbf{y}_{1:n_t}^{(i),t}\right)\right)$
- Compute incremental weights $\frac{w_t^{(i)}}{w_{t-1}^{(i)}} \propto \left(C - \log p(y_{n_t}|d)\right)^{\nu_t-\nu_{t-1}}$

Else $(n_t = n_{t-1} + 1)$:

- Sample $\left(d^{(i),t}, \mathbf{y}_{1:n_{t-1}}^{(i),t}\right) \sim \tilde{\mathcal{K}}_{n_{t-1},1}\left(\left(d^{(i),t-1}, \mathbf{y}_{1:n_{t-1}}^{(i),t-1}\right), \left(d^{(i),t}, \mathbf{y}_{1:n_{t-1}}^{(i),t}\right)\right)$
- Sample $y_{n_t}^{(i),t} \sim p(y|d^{(i),t})$

- Resample if ESS $<$ Threshold.

Algorithm 2: New SMC sampler for optimal design.

5. EXAMPLE PROBLEM

We study a synthetic problem that, despite its apparent simplicity, exhibits complex multi-modality. In particular, we address the problem of inferring the parameters of a sine wave. This nonlinear experimental design example is motivated by the problem of scheduling expensive astronomical observations [10]. The sine wave is parameterized by its amplitude A, frequency ω and phase ρ as follows

$$y = f(x; A, \omega, \rho) = A \sin\left(2\pi[(d\omega) + \rho]\right).$$

We place Gamma priors on A and ω and a $\mathcal{U}(0,1)$ prior on ρ. The objective is to find the optimal location d^* within a finite interval along the x-axis at which to make the next noisy y measurement. In our example two prior observations have been made. That is, $p(\theta)$ in Equation (1) here is the posterior parameter distribution after these measurements. Sine waves corresponding to samples from $p(\theta)$ are depicted in

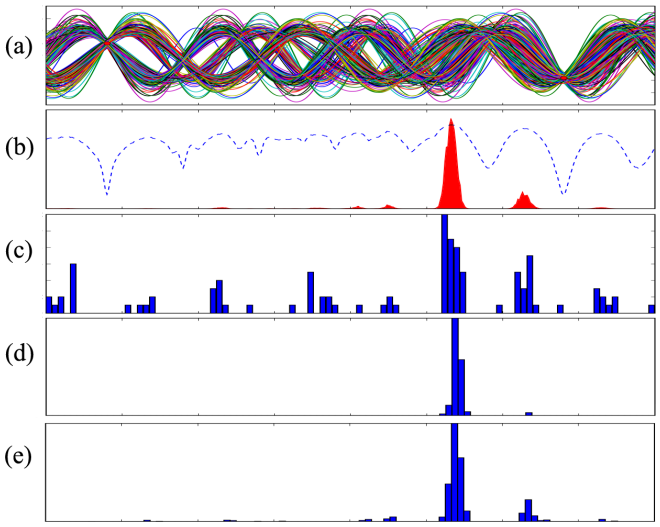

Fig. 1. Plot (a) shows some samples of the stochastic process $p(y|d)$ after 2 observations. The corresponding expected utility $U(d)$ is shown as a dashed blue line in (b) while $U(d)^{50}$ is displayed in solid red. (c) presents a histogram of the final samples of 100 independent chains using the approach of Müller et al. [4] when annealing to $U(d)^{50}$, while (d) and (e) shows the final locations of 100 particles using, respectively, the sampler of Johansen et al. [8] (Algorithm 1) and the proposed new SMC sampler (Algorithm 2).

Figure 1. The same figure shows that the utility $U(d)$ is extremely multi-modal. It also illustrates the performance of the algorithms of Müller et al. [4], Johansen et al. [8] and our new algorithm in approximating the annealed utility $U(d)^{50}$. The same random walk proposal in d and the same amount of computation were used in all cases. While many of the independent MCMC chains get stuck, the interaction in the SMC samplers assists in escaping local minima, thus yielding a better approximation of the target. However, we note that our algorithm maintains a richer particle set than the algorithm of Johansen et al. [8] and, hence, leads to a better approximation of the target. This is due to dramatically smaller variance of the incremental weights. Figure 2 provides clear evidence of this. The new algorithm maintains a higher effective sample size (lower variance). As a result, it requires far fewer resampling steps. Finally, Figure 3 shows that the proposed algorithm does a better job at exploring all modes of the objective as the simulation progresses. This explains the higher quality of the final approximation as seen in Figure 1.

6. CONCLUSION

We have introduced a new SMC algorithm for Bayesian optimal nonlinear design. It behaves well when exploring densely multi-modal target distributions and exhibits lower variance than existing approaches. We believe these two properties will play a crucial role when scaling to real high-dimensional problems.

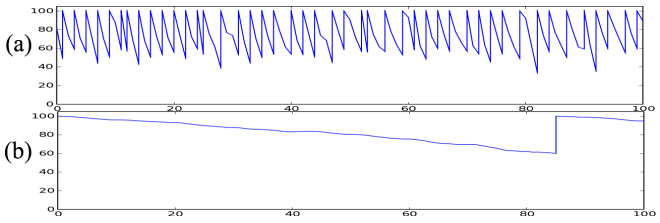

Fig. 2. The effective sample size for the first 100 steps of running (a) Algorithm 1 and (b) Algorithm 2 on the optimal design problem shown in Figure 1. The new algorithm, because of its lower variance, requires far fewer resampling steps (in total 3 compared to 189 out of 500 steps) in order to maintain an acceptable sample size.

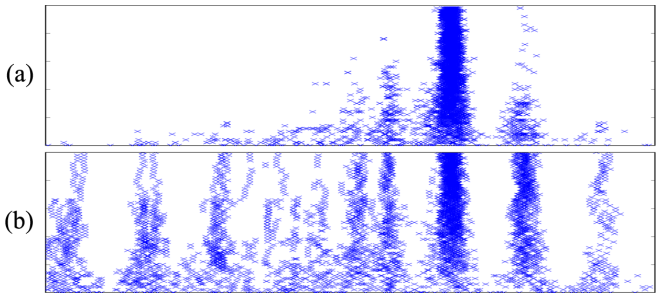

Fig. 3. Location of the particles along the design space (x-axis) over 500 steps (y-axis, increasing upwards). for (a) Algorithm 1 and (b) Algorithm 2. The plots demonstrate the annealing effect. They also show that Algorithm 1 looses track of the minor modes due to excessive resampling.

7. REFERENCES

[1] P. Sebastiani and H. P. Wynn, "Bayesian experimental design and Shannon information," in *Proceedings of the Section on Bayesian Statistical Science*, 1997, pp. 176–181.

[2] K. Chaloner and I. Verdinelli, "Bayesian experimental design: A review," *Statistical Science*, vol. 10, no. 3, pp. 273–304, 1995.

[3] P. Müller, "Simulation based optimal design," in *Bayesian Statistics 6*, 1998.

[4] P. Müller, B. Sansó, and M. de Iorio, "Optimal Bayesian design by inhomogeneous Markov chain simulation," *Journal of the American Statistical Association*, vol. 99, pp. 788–798, 2004.

[5] B. Amzal, F.Y. Bois, E. Parent, and C.P. Robert, "Bayesian optimal design via interacting MCMC," Tech. Rep. 2003-48, Ceremade, Université de Paris, 2003.

[6] K. L. Mengersen and R. L. Tweedie, "Rates of convergence of the Hastings and Metropolis algorithms," *The Annals of Statistics*, vol. 24, pp. 101–121, 1996.

[7] P. del Moral, A. Doucet, and A. Jasra, "Sequential Monte Carlo samplers," *J. Royal Statist. Soc. B*, vol. 68, no. 3, pp. 1–26, 2006.

[8] A. M. Johansen, A. Doucet, and M. Davy, "Maximum likelihood parameter estimation for maximum likelihood models using sequential Monte Carlo," in *Proceedings of ICASSP*, 2006.

[9] Jun S Liu, *Monte Carlo strategies in scientific computing*, Springer, New York, 2001.

[10] T. J. Loredo, "Bayesian adaptive exploration," in *Bayesian Inference And Maximum Entropy Methods In Science And Engineering*, 2003, pp. 330–346.