A FAST HYBRID GENETIC ALGORITHM - GIBBS SAMPLER APPROACH TO ESTIMATE GEOACOUSTIC PARAMETER UNCERTAINTIES

Caglar Yardim, Peter Gerstoft, and William S. Hodgkiss

Marine Physical Laboratory/SIO 0238, UCSD, La Jolla, CA 92093-0238, USA,
Email: gerstoft@ucsd.edu, cyardim@ucsd.edu, whodgkiss@ucsd.edu
Fax No: (1-858) 534 0574

Abstract: *Genetic algorithm (GA) is a fast algorithm that gives very good maximum a posteriori (MAP) estimates. However, it provides poor estimates for the posterior probability distributions (PPD). On the other hand, an exhaustive search can provide the PPD but will require prohibitively large number of samples. Instead, Markov Chain Monte Carlo samplers (MCMC), such as Metropolis-Hastings and Gibbs samplers, are often used to sample the posterior probabilities of the geoacoustic parameters, but still at considerable computational expense.*

Both the speed of GA and the accuracy of MCMC can be achieved using a hybrid method that first uses GA to find the optimal solution and sample the likelihood parameter landscape. Then these samples are used to approximate the likelihood landscape using Voronoi cells. A subsequent Gibbs sampling (GS) not requiring any forward model runs will sharply reduce the number of forward model runs needed to achieve high accuracy.

The hybrid algorithm is used on synthetic vertical array data. The environmental parameters of the ocean floor are estimated using GA, exhaustive search, and the GA-GS hybrid. The results show that the hybrid method requires much fewer forward model runs to accurately obtain the true underlying posterior distribution than a pure GS.

Keywords: *Geoacoustic Inversion, Parameter Uncertainty, Genetic Algorithm, Gibbs Sampler, Nearest Neighborhood Algorithm, Voronoi Cells*

1. INTRODUCTION

A weakness in sonar performance prediction has been the lack of a means for quantifying the impact of uncertainty in estimates of the ocean environment. In order to fully analyze these effects, an accurate knowledge of environmental parameter probability distributions is needed.

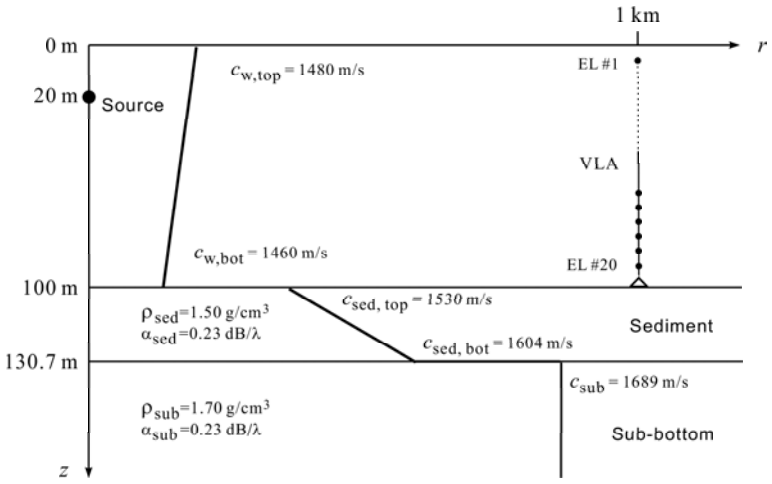

Figure 1: The geoacoustic environmental model used in the simulations.

The inverse problem is formulated using Bayes rule to compute the posterior distribution $\mathrm{p}(\mathbf{m}|\mathbf{d})$, of the environmental model parameters $\mathbf{m}$, for a given ocean acoustic data $\mathbf{d}$ gathered from a vertical or horizontal hydrophone array as shown in Fig. 1 [1,2]. Selection of a Bayesian framework enables us to define the unknown model parameters as random variables so that the inversion results will be in terms of the means, variances and marginal, as well as the n-dimensional joint posterior probability distributions (PPD). This gives the user not only the ability of obtaining the maximum a posteriori (MAP) solution, but also the prospect of performing an uncertainty analysis on the inversion results. These probabilistic properties can be calculated by taking multi-dimensional integrals of the joint PPD. However, these calculations require an accurate estimate of the overall probability distribution.

2. INVERSE PROBLEM FRAMEWORK

For a given forward model $\mathbf{f}(\cdot)$ and source strength S, measured data $\mathbf{d}$ can be written as

$$\mathbf{d} = S\mathbf{f}(\mathbf{m}) + \mathbf{e}. \tag{1}$$

Assuming a zero-mean Gaussian-distributed error term $\mathbf{e}$ with a covariance matrix $\mathbf{C}_D$, the likelihood can be written as

$$\mathcal{L}(\mathbf{m}) \;=\; (\pi|\mathbf{C_D}|)^{-N} \exp\left[-(\mathbf{d} - f(\mathbf{m}))^{\mathrm{H}}\mathbf{C}_{\mathrm{D}}^{-1}(\mathbf{d} - f(\mathbf{m}))\right], \tag{2}$$

where N is number of elements in the array. For convenience, assume the error is IID with $\mathbf{C_D} = \nu\mathbf{I}$. Maximum likelihood estimate (ML) for the source can be found by $\partial \log \mathcal{L}/\partial S = 0$, which gives

$$\widehat{S}_{ML} \;=\; \frac{\mathbf{d}^H\mathbf{f}(\mathbf{m})}{\|\mathbf{f}(\mathbf{m})\|^2} \tag{3}$$

$$\mathcal{L}(\mathbf{m}) \;=\; (\pi\nu)^{-N} \exp\left[-\frac{\phi(\mathbf{m})}{\nu}\right] \tag{4}$$

$$\phi(\mathbf{m}) \;=\; \|\mathbf{d}\|^2 - \left(\frac{\mathbf{d}^{\mathrm{H}}\mathbf{f}(\mathbf{m})}{\|\mathbf{f}(\mathbf{m})\|}\right)^2 \tag{5}$$

The ML estimate for the error variance can be found using $\partial \log \mathcal{L}/\partial\nu = 0$, which gives $\widehat{\nu}_{ML} = \phi(\mathbf{m})/N$. Inserting this back into the likelihood function, it finally becomes

$$\mathcal{L}(\mathbf{m}) = \left(\frac{N}{\pi \mathrm{e}\phi(\mathbf{m})}\right)^N \propto \phi(\mathbf{m})^{-N} \tag{6}$$

The posterior density can then be written as

$$\mathrm{p}(\mathbf{m}|\mathbf{d}) = \frac{\mathcal{L}(\mathbf{m})\mathrm{p}(\mathbf{m})}{\mathrm{p}(\mathbf{d})} \tag{7}$$

In Bayesian statistics $\mathrm{p}(\mathbf{d})$ is called the evidence and it is just a normalizing constant. If, in addition, a non-informative prior density $\mathrm{p}(\mathbf{m})$ is used, the posterior density will only be proportional to the likelihood function.

3. GA, MCMC, AND THE HYBRID GA-GS METHODS

One possible solution is using exhaustive search. For n unknown parameters, the PPD is an n-dimensional joint density. Exhaustive search will grid this n-D volume and calculate the likelihood at each grid point. Although this method gives true distributions, it is very inefficient and not applicable for n more than 4 to 5 parameters due to the huge number of forward model runs needed.

An alternative is genetic algorithm. GA will start with an initial population of points in this n-D space and new points (generations) are created according to the fitness of the individuals in the previous population (parents). It is a very effective and fast global optimizer and the population quickly converges to the highest fitness value (MAP solution). After GA converges, we have a set of samples as a collection of all previous individuals in all previous generations and their fitness values. Without any further processing these samples are usually not very useful to obtain the full distribution. The main reason is that it is designed as a global optimizer focused on finding the maximum likelihood (ML) or MAP point, not obtaining the underlying n-dimensional posterior probability density (PPD). Hence, GA gives good MAP estimates for $\mathbf{m}$, but a poor estimate for the PPD itself.

On the other end, an MCMC sampler like Gibbs sampler will sample the search space in such a way that the number of samples obtained in any region in this n-D space is proportional to the likelihood in that region. Hence, in the end, the set of MCMC samples fully represent the n-D PPD. The posterior means, variances and marginal probability distributions now can easily be found by taking n- or $(n-1)$-dimensional integrals of this PPD using these samples and Monte Carlo integration.

$$\mu_i \;=\; \int \ldots \int_{\mathbf{m}'} m_i' \mathrm{p}(\mathbf{m}'|\mathbf{d}) \mathrm{d}\mathbf{m}' \tag{8}$$

$$\sigma_i^2 \;=\; \int \ldots \int_{\mathbf{m}'} (m_i' - \mu_i)^2 \mathrm{p}(\mathbf{m}'|\mathbf{d}) \mathrm{d}\mathbf{m}' \tag{9}$$

$$\mathrm{p}(m_i|\mathbf{d}) \;=\; \int \ldots \int_{\mathbf{m}'} \delta(m_i' - m_i) \mathrm{p}(\mathbf{m}'|\mathbf{d}) \mathrm{d}\mathbf{m}' \tag{10}$$

MCMC samplers are much faster than exhaustive search, but still need much more forward model runs than GA.

Therefore, it will be very desirable to have a GA-MCMC hybrid method that combines the speed of GA with the accuracy of MCMC. Due to its inherent properties, Gibbs sampler (GS) is the best MCMC algorithm for such a hybrid method. Hence, the hybrid method will be called the GA-GS hybrid. The method consists of three distinct sections:

1. *GA* : Run a classical GA, minimizing the misfit $\phi(\mathbf{m})$, and save all the populations and the misfit values of all generations.

2. *Voronoi Cells & Approximate PPD* : Using the GA samples and their likelihood values construct Voronoi cells (see next section) around each GA point, assigning the likelihood of each GA point to all points inside its own Voronoi cell. This will result in an approximate PPD.

3. *GS* : Run a fast GS [3] on the approximate PPD instead of the real one. Since the conditional densities and the likelihood values required to run GS is known for the approximate PPD, no forward model is needed.

4. VORONOI CELLS AND GS

This process simply is a discretization of the original PPD. It will convert the true analog PPD into a digital one through an A/D converter. The only difference is that, this A/D converter is n-dimensional, and hence, discrete levels are tiny n-dimensional hypercubes. These hypercubes are called Voronoi cells. The shapes and sizes of these cells are determined by the GA sample set. There is only one GA sample in each cell and there does not exist any other GA sample, which is closer to any point inside this hypercube. Hence, for any boundary point between any two adjacent cells, the distances of that point to the two closest GA samples are same. Due to this feature, it is also called the nearest neighborhood (NN) method [4]. In each cell, the likelihood value is assumed to be constant with a value of its GA sample. Therefore, the likelihood of any point anywhere in the entire search space is known and there is no need for any further forward model runs. A very fast GS, without any forward modelling, is used to sample this approximate PPD. The accuracy of the results depends mostly on the quality of the approximate PPD, which means that, GA should gather enough samples from all over the n-dimensional search space to allow the NN algorithm to construct a good enough n-dimensional mesh, hence a good enough approximate PPD.

A simple example is illustrated in Fig. 2 with $n = 2$ only. Voronoi cells are constructed around each dot representing the GA samples and a Gibbs sampling is used on this approximate PPD. GS gets samples by changing one parameter at a time in a circulatory fashion and after all of the parameters are changed once, the result will form the next Gibbs sample, shown as squares in the figure. It selects new parameter values from the 1-D conditional probability densities and as can be seen from the figure, it is very easy to obtain conditional densities using the approximate PPD.

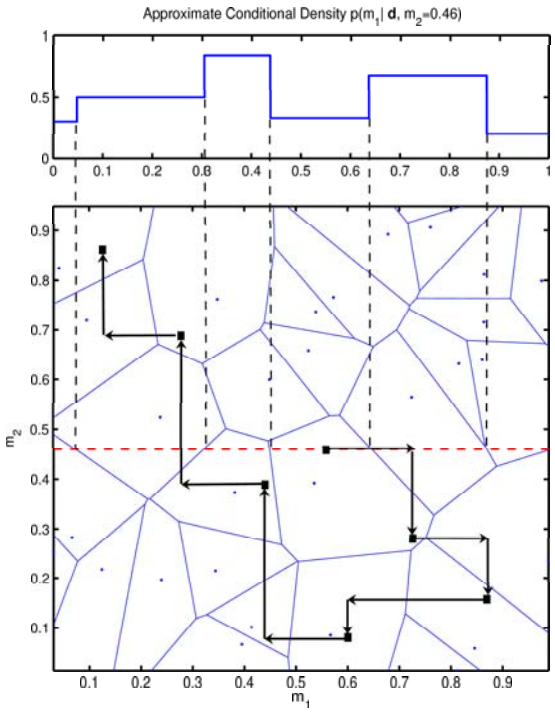

Figure 2: Voronoi cells for a simple 2 parameter search space and an approximate conditional 1-D density for a given cut. Dots represent GA points and squares represent the GS points sampling the approximate 2-D PPD.

5. SIMULATION RESULTS

The hybrid algorithm is illustrated on a synthetic vertical array data given in Fig. 1. SAGA is used to obtain the GA samples and ORCA is used as the normal mode propagation model [5]. All the environmental parameters are assumed to be known, except for the ocean floor (depth and sound speed values for sediment and sub-bottom). Selection of only 4 parameters enables us to compute the exhaustive search results to serve as the true distributions for comparison purposes. Each parameter is divided into a grid of 30 possible values with a total search size of $(30^4 = 810000)$ 810k points. Then, the problem is solved using 20k GA points. Finally, some or all of these GA points are used to create Voronoi cells and subsequent GS are applied with various point sizes (GA-GS hybrid).

1-D and 2-D marginal PPD's are given for exhaustive search and hybrid GA-GS in Fig. 3-4. Comparison of these marginals shows that the hybrid method accurately estimated the n-dimensional PPD using mush fewer forward model runs.

The results given in Fig. 5 show that effects of changing the Gibbs samples, which are obtained without using a forward model. True 1-D marginal densities, results obtained using only 20k GA and no GS, hybrid GA-GS results for 20k-200k, 20k-100k, 20k-50k, and 20k-10k samples are given in Fig. 5(a-f) respectively. The poor results for GA-only case can clearly be seen. The hybrid GA-GS results that are actually obtained using the same initial GA-only points are much better when compared with the true distributions. Even for only 10k GS samples the error in the marginal PPD's is quite limited.

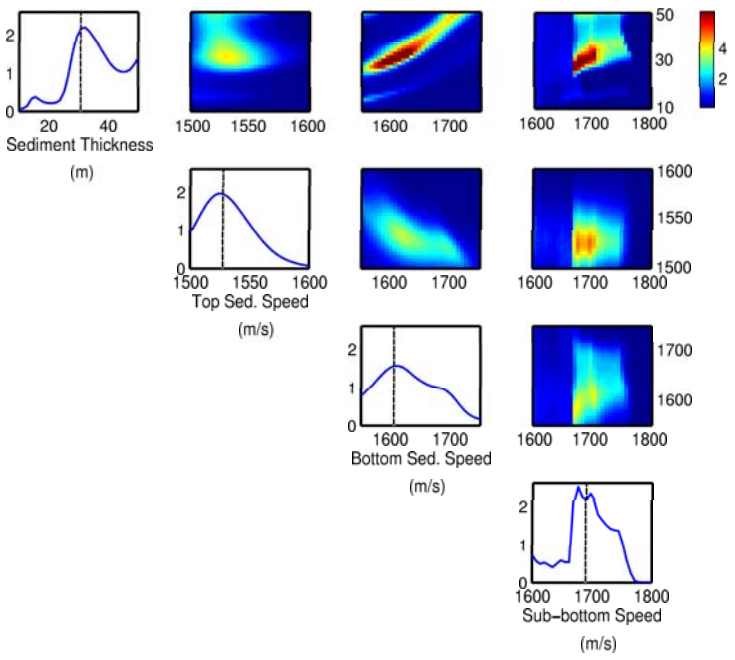

Figure 3: 1-D and 2-D marginal posterior densities obtained by exhaustive search obtained using 810k forward model runs. Vertical lines represent the MAP solution.

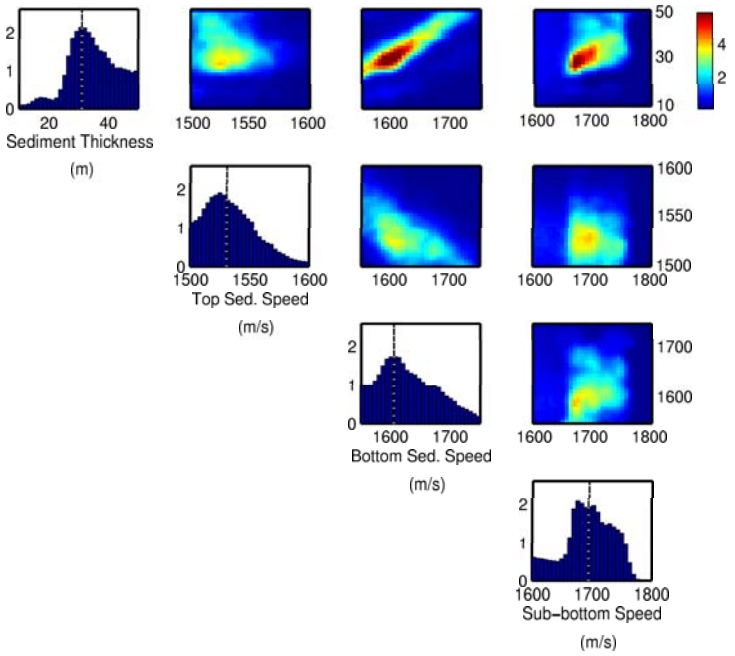

Figure 4: 1-D and 2-D marginal posterior densities obtained by Hybrid GA-GS obtained using 20k GA samples (requiring forward model runs) followed by 100k GS samples (no forward model run). Vertical lines represent the MAP solution.

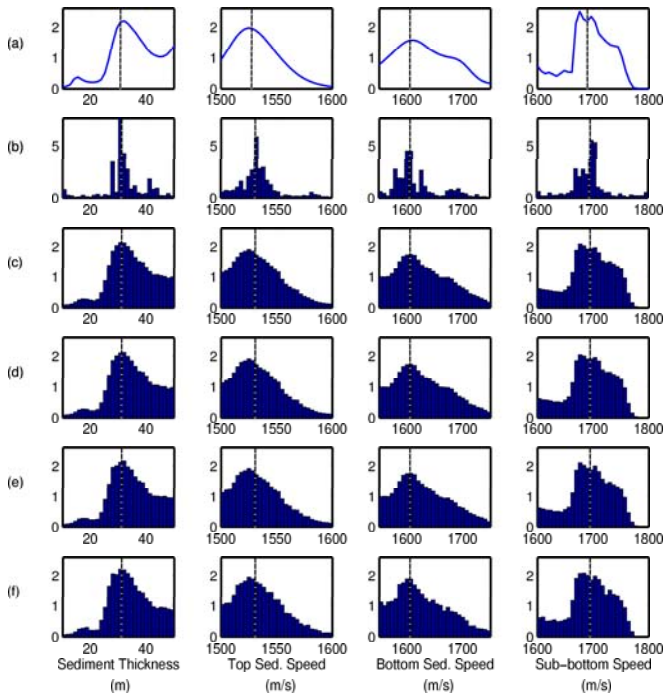

Figure 5: Effects of GS sample size on the 1-D marginal posterior densities for a given GA sample set (20k). Vertical lines represent the MAP solution. Distributions calculated using (a) exhaustive search (true distribution), (b) histogram of 20k GA samples without the hybrid run, (c) 200k GS samples, (d) 100k GS samples, (e) 50k GS samples, and (f) 10k GS samples. All densities are scaled based on a dimensionless [0,1] interval for each parameter.

The effects of changing GA samples on the hybrid method performance is analyzed in Fig. 6. In all cases, the numbers of Gibbs samples are kept the same. True 1-D marginal densities, results obtained using only 20k GA and no GS, hybrid GA-GS results for 20k-200k, 10k-200k, 4k-200k, and 2k-200k samples are given in Fig. 6(a-f) respectively. Again all of the hybrid GA-GS results are much more accurate than the GA-only case. Even starting with only 2k GA points resulted in very similar marginal PPD's. This means the method obtained the PPD using only 2k forward model runs compared to the 810k used by exhausted search. However, one should be careful with the initial GA set, since the following Gibbs sampler is used to sample and obtain the approximate PPD created by the initial GA samples. Any error due to an insufficient sampling in GA section cannot be compensated by increasing the GS sample size.

6. CONCLUSIONS

A hybrid GA-GS method is used to perform geoacoustic inversion. This approach enables us to obtain full n-dimensional posterior probability distributions of the unknown parameters faster than a conventional MCMC sampler or exhaustive search and more accurately than GA.

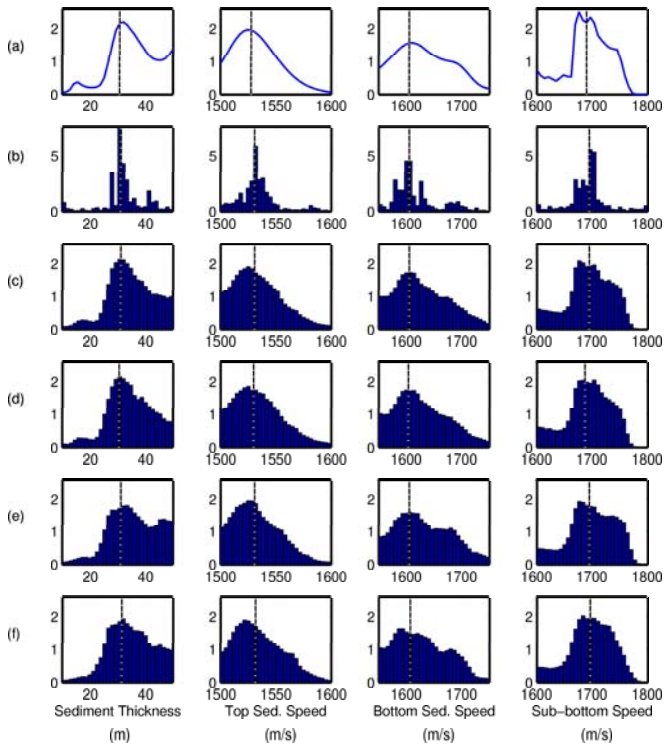

Figure 6: Effects of GA sample size on the 1-D marginal posterior densities for a constant GS sample size (200k). Vertical lines represent the MAP solution. Distributions calculated using (a) exhaustive search (true distribution), (b) histogram of 20k GA samples without the hybrid run, (c) 20k GA samples, (d) 10k GA samples, (e) 4k GA samples, and (f) 2k GA samples. All densities are scaled based on a dimensionless [0,1] interval for each parameter.

ACKNOWLEDGEMENT

This work was supported by the Office of Naval Research.

References

[1] **P. Gerstoft** and **C.F. Mecklenbräuker**, "Ocean acoustic inversion with estimation of *a posteriori* probability distributions," *J. Acoust. Soc. Am.*, 104(2), pp. 808–819, 1998.

[2] **S.E Dosso**, "Quantifying uncertainties in geoacoustic inversion. I: a fast Gibbs sampler approach," *J. Acoust. Soc. Am.*, 111, pp. 129–142, 2002.

[3] **J.K. Ó Ruanaidh** and **W.J. Fitzgerald**, *Numerical Bayesian methods applied to signal processing*, Springer, NY, 1996.

[4] **M. Sambridge**, "Geophysical inversion with a neighborhood algorithm - I. Searching a parameter space," *Geophys. J. Int., 138*, pp 479-494, 1999.

[5] **P. Gerstoft**, "SAGA Users guide 2.0, an inversion software package," *SACLANT Undersea Research Centre*, SM–333, 1997.