Sequential Design of Computer Experiments for the Estimation of a Quantile with Application to Numerical Dosimetry

M. Jala
Orange Labs and Télécom ParisTech
jala@telecom-paristech.fr

C. Lévy-Leduc, E. Moulines and A. Garivier
CNRS/LTCI/Télécom ParisTech/

E. Conil and J. Wiart
Orange Labs and Whist Lab

Abstract

In this paper, we propose a sequential sampling approach for estimating the quantile of $Y = f(X)$, where X has a known distribution in $\mathbb{R}^d$ and f is an unknown, expensive-to-evaluate real-valued function. Our approach aims at estimating the quantile by using as few evaluations of f as possible and consists in modelling f as the sample of a well-chosen Gaussian process. Our method called GPquantile is compared with a more naive methodology on synthetic data, and is also applied to real data coming from numerical dosimetry.

1 Introduction

Over the past 30 years, wireless communication systems have been increasingly used. The number of mobile phones, WIFI boxes, antennas, etc., is growing together with a strong public concern over possible health problems related to the exposure to electromagnetic fields (EMF). Among the questions linked to exposure effects, the assessment of the fetuses exposure to EMF has been recommended as a top priority research topic by the World Health Organization [1]. To deal with this issue, several approaches can be used: long-term epidemiological studies, in vitro and in vivo studies, and numerical dosimetry based methods. Here, we shall consider the latter point of view, which consists in numerically evaluating the absorbed dose of EMF energy with the so-called Specific Absorption Rate (SAR), expressed in watts per kilogram, using 3D pregnant women and fetuses models built from MRI data, see [1]. As the SAR depends on several parameters such as the morphology and the posture of the mother and fetus, the position and the type of the wireless device, we shall model the SAR as $Y = f(\mathbf{X})$, where f is an unknown real-valued function and $\mathbf{X}$ is a random vector of $\mathbb{R}^d$ having a known distribution. However, the evaluation of Y for a given input involve complex and time-consuming computer simulations. This issue of dealing with expensive-to-evaluate functions is not new. For instance, many papers such as those of [11], [12], [2], and references therein have proposed Bayesian optimization methods for finding the maximum of functions which are costly to evaluate. Here, we propose a sequential sampling strategy for estimating the quantile of Y, called GPquantile in the following, using as few evaluations of f as possible. GPquantile is based on modelling f as a sample of a Gaussian process.

This paper is organized as follows: in Section 2, we describe our approach. In Section 3, GPquantile is applied to a one-dimensional function f introduced in [10] and compared with a more naive

[1] see p. 21 of http://whqlibdoc.who.int/publications/2010/9789241599948_eng.pdf

approach. Finally, in Section 4, our approach is applied to assess the exposure to EMF of a 26-week-old fetus in a simplified context.

2 Description of the method

In this section, we shall explain our sequential sampling strategy for estimating the α-quantile q_α of $Y = f(\mathbf{X})$, for a given α in $(0,1)$, where f is an unknown infinitely differentiable real-valued function and $\mathbf{X}$ is a random vector of $\mathbb{R}^d$ having a known distribution. The α-quantile q_α is defined by

$$\mathbb{P}(f(\mathbf{X}) \leq q_\alpha) = \alpha \ .$$

Writing $\mathbf{X}$ as $F^{-1}(\mathbf{U})$, where F is the c.d.f. of X, and $\mathbf{U}$ is a uniform random variable on $[0,1]^d$, we shall assume in the sequel that $\mathbf{X}$ has a uniform distribution on $[0,1]^d$.

Following [9] and [11], we put a GP$(0, k(\mathbf{x}, \mathbf{x}'))$ prior on f, where GP$(0, k(\mathbf{x}, \mathbf{x}'))$ denotes a zero-mean Gaussian process having k as a covariance function. For a sample $\mathbf{y}_T = (Y_1, \ldots, Y_T)^T$, the posterior over f is a GP distribution again with mean $\mu_T(\mathbf{x})$ and covariance $k_T(\mathbf{x}, \mathbf{x}')$ given by

$$\mu_T(\mathbf{x}) = \mathbf{k}_T(\mathbf{x})^T \mathbf{K}_T^{-1} \mathbf{y}_T \ , \tag{1}$$

$$k_T(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}, \mathbf{x}') - \mathbf{k}_T(\mathbf{x})^T \mathbf{K}_T^{-1} \mathbf{k}_T(\mathbf{x}') \ , \tag{2}$$

where $\mathbf{k}_T(\mathbf{x}) = [k(\mathbf{x}_1, \mathbf{x}) \ldots k(\mathbf{x}_T, \mathbf{x})]^T$ and $\mathbf{K}_T = [k(\mathbf{x}_i, \mathbf{x}_j)]_{1 \leq i,j \leq T}$. Since f is assumed to be very smooth, we shall use a squared exponential covariance function defined by

$$k(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{(\mathbf{x} - \mathbf{x}')^2}{2\ell^2} \right) \ , \tag{3}$$

where the characteristic length-scale ℓ is estimated by maximizing with respect to ℓ the posterior log-likelihood given by

$$-\frac{1}{2} \mathbf{y}_T^T \mathbf{K}_T^{-1} \mathbf{y}_T - \frac{1}{2} \log |\mathbf{K}_T| - \frac{T}{2} \log 2\pi \ .$$

Let us now describe the different steps of our method. Let $\mathcal{X}$ be a fine grid of $[0,1]^d$.

- We start with an evaluation of f at a small number T_0 of observations: $\mathcal{X}_0 = \{\mathbf{x}_1, \ldots, \mathbf{x}_{T_0}\}$. Thus, we have the observations $\mathbf{y}_{T_0} = (f(\mathbf{x}_1), \ldots, f(\mathbf{x}_{T_0}))^T$.
- The point of $\mathcal{X}$ to add to the set of observations is $\mathbf{x}_{T_0+1}$ defined by

  $$\mathbf{x}_{T_0+1} = \arg\max_{\mathbf{x} \in \mathcal{X}} \hat{\sigma}_1(\mathbf{x}) \ ,$$

  where $\hat{\sigma}_1(\mathbf{x})$ is an estimator of the standard deviation of the quantile of the observations at the first iteration of the procedure which is computed as follows: we simulate N sample paths of a Gaussian process having a mean and a covariance equal to μ_{T_0} and k_{T_0}, respectively and compute, for each path of this sample, the empirical quantile estimator of q_α at each point $\mathbf{x}$ of $\mathcal{X}$. We recall that the empirical quantile for a sample of n random variables $Z_1, \ldots, Z_n$ is given by $\hat{F}_n^{-1}(\alpha)$, where $\hat{F}_n(t) = n^{-1} \sum_{i=1}^n \mathbf{1}_{\{Z_i \leq t\}}$. The set of observations becomes $\mathcal{X}_1 = \mathcal{X}_0 \cup \{\mathbf{x}_{T_0+1}\}$.
- The point of $\mathcal{X}$ to add to the set of observations is $\mathbf{x}_{T_0+2}$ defined by

  $$\mathbf{x}_{T_0+2} = \arg\max_{\mathbf{x} \in \mathcal{X}} \hat{\sigma}_2(\mathbf{x}) \ ,$$

  where $\hat{\sigma}_2(\mathbf{x})$ is an estimator of the standard deviation of the quantile of the observations at the second iteration of the procedure which is computed as previously except that the mean and the covariance of the Gaussian process are now given by μ_{T_0+1} and k_{T_0+1}. The set of observations becomes $\mathcal{X}_2 = \mathcal{X}_1 \cup \{\mathbf{x}_{T_0+2}\}$.
- We iterate this procedure and stop it at the iteration t when K_{T_0+t} is not invertible.

3 Numerical experiments

In this section, we apply the methodology presented in Section 2 to the following one-dimensional function already used in [10]:

$$f(x) = \sin(2\pi x) + \frac{\cos(8\pi x)}{5} \ , \ x \in [0,1] \ .$$

Here we aim at estimating the 95% quantile of $f(X)$, where $X \sim \mathcal{U}[0,1]$; the real value of this quantile is $q_{0.95} \approx 1.149$.

We start with a set $\mathcal{X}_0$ of $T_0 = 10$ points selected with a maximin LHS obtained with the `lhsdesign` function of Matlab. This set is included in a larger set $\mathcal{X}$ of cardinality 1000 also obtained with a LHS. In Figure 1 (a), μ_{T_0} and the values of f at the points of the set $\mathcal{X}_0$ are displayed. The $N = 1000$ sample paths of the corresponding GP are simulated (20 of them are displayed in the left plot of Figure 2): for each point x in $\mathcal{X} \setminus \mathcal{X}_0$, we have $N = 1000$ possible values of $f(x)$, hence 1000 possible values of the quantile estimator, which enables us to evaluate the empirical standard deviation of the quantile at each point x in $\mathcal{X} \setminus \mathcal{X}_0$, the value of the standard deviation of the quantile at the points x of $\mathcal{X}_0$ being equal to zero. This standard deviation is represented with a dotted line in Figure 1 (a). As we previously explained, the point x_{T_0+1} for which the standard deviation is the highest is added to the set of observations $\mathcal{X}_0$; in Figure 1 (a), this point is around 0.172. Hence in Figure 1 (b), this point has been added to the set, and the GP has been updated, and a new empirical standard deviation function has been computed. The process stops at the iteration $t = 7$ for which the set of observations contains points so close that $\mathbf{K}_{T_0+t}$ is not invertible. In Figures 1 (e) and (f), the empirical standard deviation is almost constant and its values are very small. Note also that the results displayed in Table 1 show that the last values of the quantile estimator are the same.

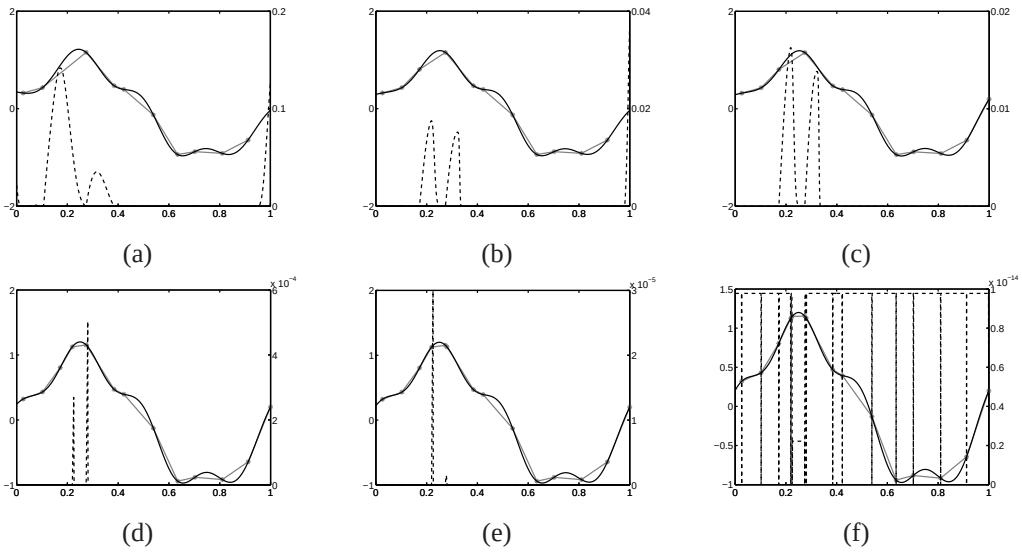

Figure 1: Each plot ((a) to (f)) displays the mean of the Gaussian process (plain black line), the values of the observation points (gray stars) and the standard deviation of the estimated quantile (dotted line) at each iteration. The scale of the standard deviation is displayed in the right y-axis.

Iteration (t)	1	2	3	4	5	6	7
$\hat{q}(t)$	0.469	0.805	0.805	1.125	1.127	1.149	1.149

Table 1: The quantile estimators at the different iteration steps.

We propose now to compare our approach with a more naive method (Random) which consists in randomly adding a new point at each iteration. This comparison is performed through 1000 Monte-Carlo replications (on the choice of the sets $\mathcal{X}_0$ and $\mathcal{X}$) and the corresponding means of the relative errors $|q_{0.95} - \hat{q}(t)|/q_{0.95}$ at each iteration t ($t = 1, \ldots, 20$) are displayed in Figure 3. We can see from this figure that our approach outperforms the Random method.

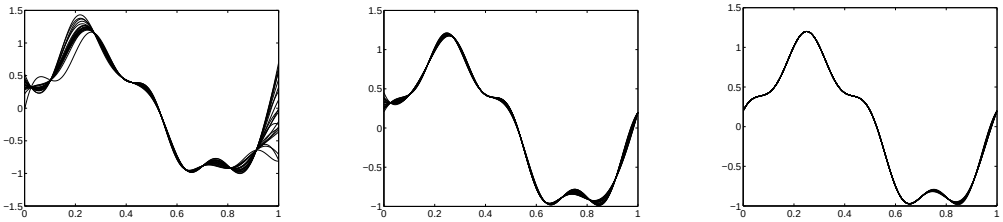

Figure 2: Simulated Gaussian process paths at different steps of the process. From left to right: at the beginning, after 3 iterations and at the last iteration.

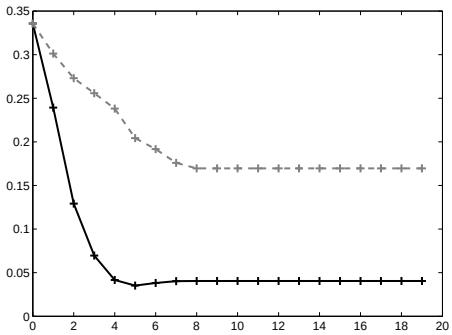

Figure 3: Means of the relative errors at each iteration. GPquantile (black plain line) and Random (gray dotted line).

4 Application to real data

In this section, we shall apply the methodology developed in Section 2 to the estimation of the 95% quantile of the SAR of a 26-week-old fetus. Our application is all the more interesting since most of dosimetric studies are carried out with deterministic approach, which means with one human model in a given posture and one configuration of exposure (such as a frontal incident plane wave). As it has been shown in [6], [3] and [4], that morphology, posture and position of the EMF source have an high influence on the exposure, the JST-ANR Fetus project in which this work is included aims at statistically analyzing the exposure of fetuses. We shall use an anatomically realistic woman model designed by [8] corresponding to the average dimensions of Japanese women, in which a 26-week-old fetus model has been inserted (see [7]); indeed, whole body pregnant woman models do not exist, as medical data needed to build them is not always available. In our application, the pregnant woman model is exposed to 900 MHz vertically polarized electromagnetic plane waves with a 1 Volt per meter amplitude.

The SAR (expressed in W/kg) of the fetus will be considered as a function of only one parameter: the azimuth of the incident wave. The value of the SAR for a given value of the azimuth is computed through the Finite Difference in Time Domain (FDTD) method, which is commonly used in the field of dosimetry, see for instance [5], [3], [13]. Using the same parameters as those used in Section 3, we obtain the values given in Table 2 as 95% quantile estimators at each iteration t.

Iteration (t)	1	2	3	4	5	6	7	8
$\hat{q}(t) \times 10^7$	4.843	5.744	6.300	6.300	6.329	6.336	6.338	6.341

Table 2: The quantile estimators at the different iteration steps.

As a comparison, the SAR of the fetus in the pregnant Japanese model exposed to a frontal incident plane wave is $6.136 * 10^{-7}$ W/kg.

References

[1] L. Bibin, J. Anquez, J. de la Plata Alcalde, T. Boubekeur, E. Angelini, and I. Bloch. Whole-body pregnant woman modeling by digital geometry processing with detailed uterofetal unit based on medical images. *Biomedical Engineering, IEEE Transactions on*, 57(10):2346 –2358, oct. 2010.

[2] E. Brochu, V. M. Cora, and N. de Freitas. A tutorial on bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *CoRR*, abs/1012.2599, 2010.

[3] E. Conil, A. Hadjem, F. Lacroux, M. F. Wong, and J. Wiart. Variability analysis of sar from 20 MHz to 2.4 GHz for different adult and child models using finite-difference time-domain. *Physics in Medicine and Biology*, 53(6):1511, 2008.

[4] A. E. Habachi, E. Conil, A. Hadjem, E. Vazquez, M. F. Wong, A. Gati, G. Fleury, and J. Wiart. Statistical analysis of whole-body absorption depending on anatomical human characteristics at a frequency of 2.1 GHz. *Physics in Medicine and Biology*, 55(7):1875, 2010.

[5] A. Hirata, S. Kodera, J. Wang, and O. Fujiwara. Dominant factors influencing whole-body average SAR due to far-field exposure in whole-body resonance frequency and GHz regions. *Bioelectromagnetics*, 28(6):484–487, 2007.

[6] T. Kientega, E. Conil, A. Hadjem, E. Richalot, A. Gati, M. Wong, O. Picon, and J. Wiart. A surrogate model to assess the whole body SAR induced by multiple plane waves at 2.4 GHz. *Annals of Telecommunications*, 66:419–428, 2011. 10.1007/s12243-011-0261-z.

[7] T. Nagaoka, T. Togashi, K. Saito, M. Takahashi, K. Ito, and S. Watanabe. An anatomically realistic whole-body pregnant-woman model and specific absorption rates for pregnant-woman exposure to electromagnetic plane waves from 10 MHz to 2 GHz. *Physics in Medicine and Biology*, 52(22):6731, 2007.

[8] T. Nagaoka, S. Watanabe, K. Sakurai, E. Kunieda, S. Watanabe, M. Taki, and Y. Yamanaka. Development of realistic high-resolution whole-body voxel models of japanese adult males and females of average height and weight, and application of models to radio-frequency electromagnetic-field dosimetry. *Physics in Medicine and Biology*, 49(1):1, 2004.

[9] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, December 2005.

[10] Robert B. Gramacy and Herbert K. H. Lee. Adaptive Design and Analysis of Supercomputer Experiments. *Technometrics*, 2009.

[11] N. Srinivas, A. Krause, S. M. Kakade, and M. Seeger. Gaussian process bandits without regret: An experimental design approach. *CoRR*, abs/0912.3995, 2009.

[12] E. Vazquez and J. Bect. Convergence properties of the expected improvement algorithm with fixed mean and covariance functions. *Journal of Statistical Planning and Inference*, 140(11):3088 – 3095, 2010.

[13] J. Wiart, A. Hadjem, M. F. Wong, and I. Bloch. Analysis of rf exposure in the head tissues of children and adults. *Physics in Medicine and Biology*, 53(13):3681, 2008.