Differences between estimates of expected computer code output with GEM-SA

John Paul Gosling[†]
University of Sheffield

11th August, 2006

Abstract

This report discusses the use of the Gaussian emulation machine for sensitivity analysis (*GEM-SA*) to produce estimates for expected computer code outputs given that the inputs are unknown and the computer code is only known for a limited number of input combinations. Users have noticed a difference between the expected computer code output given by *GEM-SA* and the result of running the code with the most likely value of the inputs entered. An investigation of main and joint effect plots reveal the reasons behind this difference.

1 Expectation of a function

To understand the results of the Gaussian emulation machine for sensitivity analysis (*GEM-SA*), it is necessary for the user to be familiar with some standard statistical concepts. In this section, the expectation of a random variable and the expectation of a function of that random variable will be briefly reviewed. The sections following this review will concentrate on the use of *GEM-SA* and difference mentioned in the title of this report.

Suppose X is a continuous random variable. The *expectation* of X is given as

$$E(X) \quad = \quad \int xp(x)dx, \tag{1}$$

where $p(x)$ is the probability density function for X. The expectation of a func-

[†]Address for correspondence: John Paul Gosling, Department of Probability and Statistics, University of Sheffield, Hounsfield Road, Sheffield, S3 7RH, UK.
Email: J.P.Gosling@sheffield.ac.uk

tion of x can be calculated using

$$E(g(X)) = \int g(x)p(x)dx. \tag{2}$$

It should be noted that $E(g(X)) \neq g(E(X))$ in general. In the context of *GEM-SA*, g is a computer code, x are inputs and beliefs about x are given by the density function $p(x)$. The *plug-in* estimate is given by $g(E(X))$ and *GEM-SA* calculates $E(g(X))$. The potential inequality between $E(g(X))$ and $g(E(X))$ can be seen in the following example.

Suppose X is a uniform random variable on the range -1 to 1 and $g(x) = x^2$. As $E(x) = 0$, $g(E(x)) = 0$, but $E(g(X)) = 1/3$. Figure 1 gives an illustration of why this is. There is only one point where $g(x) = 0$, and, as X will take values other than just zero, we expect $g(X)$ to be greater than zero. The positive adjustment of $g(E(X))$ to $E(g(X))$ occurs because $g(x)$ is *convex* over the range of uncertainty.

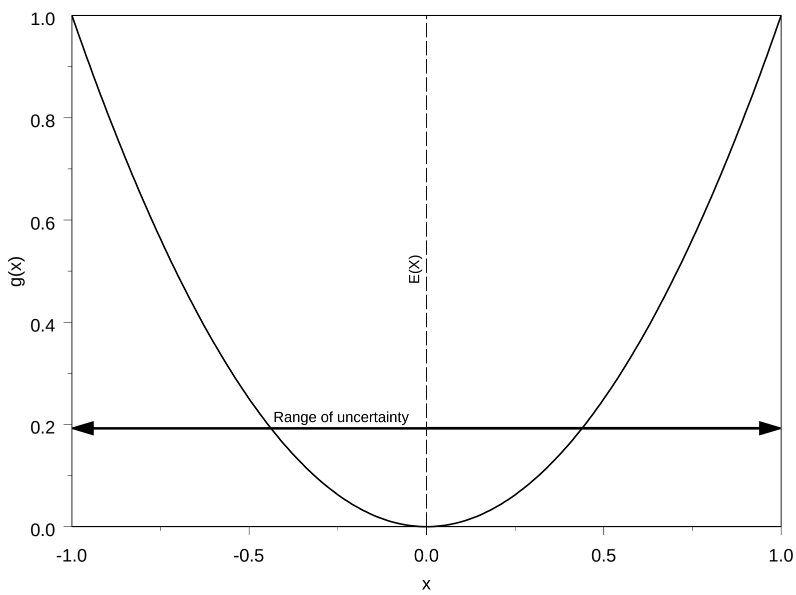

Figure 1: Plot of $g(x) = x^2$.

Now let $g(x) = -e^x$. In this case, $g(E(x)) = -1$, but $E(g(X)) \approx -2.35$. The effect is reversed from what we had before. In this case, $g(x)$ is much less than $g(E(x))$ for x greater than 0 in the range of uncertainty and $g(x)$ is slightly greater than $g(E(x))$ for x less than 0. Over the range, we would expect $g(x)$ to be less than $g(E(X))$. Over the range of uncertainty, $g(x)$ is said to be *concave*.

If $g(x)$ is *linear*, $g(x) = x + 3$ for example, over the range of uncertainty, $E(g(X)) = g(E(X))$. For $g(x) = x + 3$, $g(E(X)) = 3$. For values of x greater than 0, $g(x)$ is greater than 3. For values of x less than 0, $g(x)$ is less than 3. These two effects cancel each other out over the range of uncertainty to give $E(g(X)) = g(E(X))$.

Figure 2 shows the three types of function discussed in this section. The inequality between $E(g(X))$ and $g(E(X))$ is dependent on the curvature over the range of uncertainty given by the distribution for X. If $E(X)$ is in the middle of a concave or convex interval of $g(x)$ and the range of uncertainty spans this interval, then $E(g(X))$ can be far from $g(E(X))$.

2 Uncertainty analysis and GEM-SA

Consider a computer model, represented by the function $\eta(.)$, that takes inputs given by the vector $\mathbf{x}$ and produces a scalar output y. This relationship can be written as

$$y = \eta(\mathbf{x}). \tag{3}$$

The true input to the model $\mathbf{X}$ is unknown. Beliefs about $\mathbf{X}$ are represented by a probability distribution $G(\mathbf{x})$. As $\mathbf{X}$ is unknown, the value of the computer model is unknown when run with the true inputs; the output for the true inputs will be denoted by Y. Uncertainty analysis deals with the distribution of Y conditional

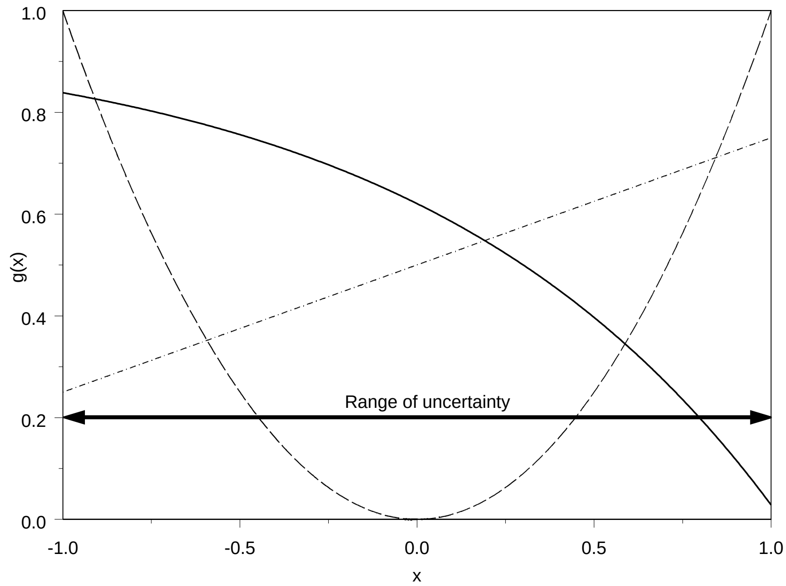

Figure 2: Three types of function curvature over the range of uncertainty: linear (dot-dashed), convex (dashed) and concave (solid).

on $G(\mathbf{x})$. Two aspects of the distribution of interest are its mean and variance:

$$\begin{aligned} M &= E_{\mathbf{X}}(Y|\eta), \\ V &= \mathrm{Var}_{\mathbf{X}}(Y|\eta). \end{aligned} \tag{4}$$

The Gaussian emulation machine for sensitivity analysis (*GEM-SA*) is a program that can calculate the mean and variance of Y conditional on $G(\mathbf{x})$ by building emulators. The program takes inputs $\mathbf{x}_1, \ldots, \mathbf{x}_n$, the respective outputs $y_1 = \eta(\mathbf{x}_1), \ldots, y_n = \eta(\mathbf{x}_n)$ and asks the user to specify beliefs about $\mathbf{X}$. With these specified, *GEM-SA* produces values for the quantities given in (8). It is also capable of calculating the main and joint effects of the computer code inputs. The main effect of input x_i is defined as

$$z_i(x_i) = E_{\mathbf{X}_{\text{-}i}}(Y|x_i) - M, \tag{5}$$

where $\mathbf{X}_{\text{-}i}$ is the vector of inputs with x_i removed. The joint effect of inputs x_i and x_j is defined as

$$z_{i,j}(x_i, x_j) = E_{\mathbf{X}_{\text{-}i,j}}(Y|x_i, x_j) - z_i(x_i) - z_j(x_j) - M. \tag{6}$$

The main and joint effects of the inputs can be thought of the difference from the expected output when the inputs are fixed. Plots of the main and joint effects of the inputs can reveal the behaviour of $\eta(.)$. Note that *GEM-SA* actually plots the expected output when the inputs are fixed rather than the main and joint effects of the inputs.

GEM-SA allows the user to specify uniform or normal distributions to represent beliefs about the inputs of the computer model. In the normal case, the user is required to specify two parameters: the mean and variance. The two parameters can be roughly interpreted as follows: the mean is the most likely value for that input and the input is likely to fall in the interval $(\text{mean}-2\sqrt{\text{variance}}, \text{mean}+2\sqrt{\text{variance}})$. Let $\mathbf{x}_M$ be the vector of user-specified means, then

$$y_M = \eta(\mathbf{x}_M) \tag{7}$$

is the plug-in estimate of Y. The plug-in estimate is the result of running the computer code using the most likely values of the inputs.

Some computer models are computationally expensive: obtaining a single output of the model takes a significant amount of time. If so, a faster surrogate model (an *emulator*) can be used to approximate features of the distribution of Y. The uncertainty about the correct form of $\eta(.)$ needs to be accounted for. This means that we must consider the uncertainty around M and D due to the fact we are not using the true $\eta(.)$:

$$\begin{aligned} & E_\eta(M|D), \\ & \text{Var}_\eta(M|D) \\ \text{and} \quad & E_\eta(V|D), \end{aligned} \tag{8}$$

where D is the data about $\eta(.)$ that we need to build an emulator. The items in (8) represent our beliefs about Y given $G(\mathbf{x})$ and we are emulating $\eta(.)$. An introduction to these concepts and emulator building is given in O'Hagan (2006), and a more detailed and technical description of *GEM-SA* and the underlying statistical theory can be found in Kennedy (2004).

3 A difference between the plug-in and GEM-SA estimates

The plug-in estimate given in equation (7) is often different to M, which is represented by *GEM-SA*'s $E_\eta(M|D)$ and $\text{Var}_\eta(M|D)$. This behaviour can be seen in the following synthetic examples. In the examples, the computer code $\eta(.)$ takes two inputs and produces a scalar output. The data and project files for both the examples are available on the web (http://j-p-gosling.staff.shef.ac.uk/Pub).

3.1 Example 1

In this example, we have

$$\eta(x_1, x_2) = \frac{cos(4\pi x_i)}{4} + x_2 + 2, \tag{9}$$

and our uncertainty about the inputs is represented by uniform distributions. We suppose that $\eta()$ is unknown and can only be evaluated for a limited number of input pairs. Throughout this example, we specify $X_2 \sim U(0,1)$ and alter the distribution for X_1.

3.1.1 Distribution a

We specify $X_1 \sim U(0.25, 0.75)$. This gives us the expected value of 0.5 for both X_1 and X_2. We can calculate $\eta(0.5, 0.5) = 2.75$. From just twenty model outputs, *GEM-SA*'s $E_\eta(M|D)$ is found to be 2.50404, which is markedly different from the plug-in estimate. As the form of $\eta(.)$, is unknown there is uncertainty about this estimate of M. An interval for M, given by $E_\eta(M|D) \pm 2\sqrt{\text{Var}_\eta(M|D)}$, is $(2.4899, 2.5182)$. The reason for the difference between the two estimates can be seen in figure 3. The main effect of x_1 is strongly concave over the range of uncertainty in this example; as we noted in section 1, this leads to the plug-in estimate being greater than the expectation of the function.

3.1.2 Distribution b

We specify $X_1 \sim U(0, 0.5)$; therefore, the expectation of X_1 is 0.25. The plug-in estimate is calculated by $\eta(0.25, 0.5) = 2.25$. Again, from just twenty model outputs, *GEM-SA*'s $E_\eta(M|D) = 2.49716$; this is quite different from the plug-in estimate. An interval for M, given by $E_\eta(M|D) \pm 2\sqrt{\text{Var}_\eta(M|D)}$, is $(2.4815, 2.5128)$. Figure 3 shows that the main effect of x_1 is strongly convex over the range of uncertainty in this example; this leads to the plug-in estimate being less than the expectation of the function.

3.1.3 Distribution c

We specify $X_1 \sim U(0.25, 0.5)$. This gives us the expected value of 0.375 for X_1. We can calculate $\eta(0.375, 0.5) = 2.5$. From just twenty model outputs, *GEM-SA*'s $E_\eta(M|D) = 2.49404$. The interval for M with $X_1 \sim U(0.25, 0.5)$ is $(2.4756, 2.5124)$. The plug-in estimate is well within the interval in this example.

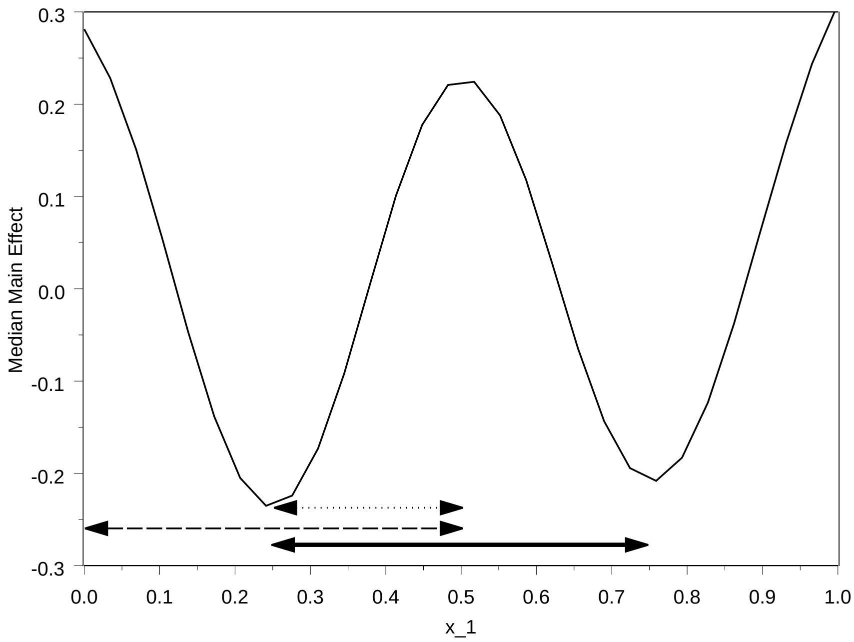

Figure 3: Median main effect plot for x_1 in example 1 showing the three different ranges of uncertainty: a (solid), b (dashed) and c (dotted).

The main effect of x_1 is practically linear over the range of uncertainty; this leads to the plug-in estimate being approximately equal to the expectation of the function.

3.2 Example 2

We have a function that has a greater interaction between the two inputs than in the last example:

$$\eta(x_1, x_2) = \frac{3{x_2}^{\sqrt{x_1}}{x_1}^{x_2}}{\sqrt{x_1 + 0.2}}, \tag{10}$$

and our uncertainty about the inputs is represented by normal distributions. We set a distribution of $N(0.5, 0.05)$ for both inputs. This leads to a plug-in estimate of 1.55309. Again twenty input pairs are calculated and *GEM-SA* is used to give

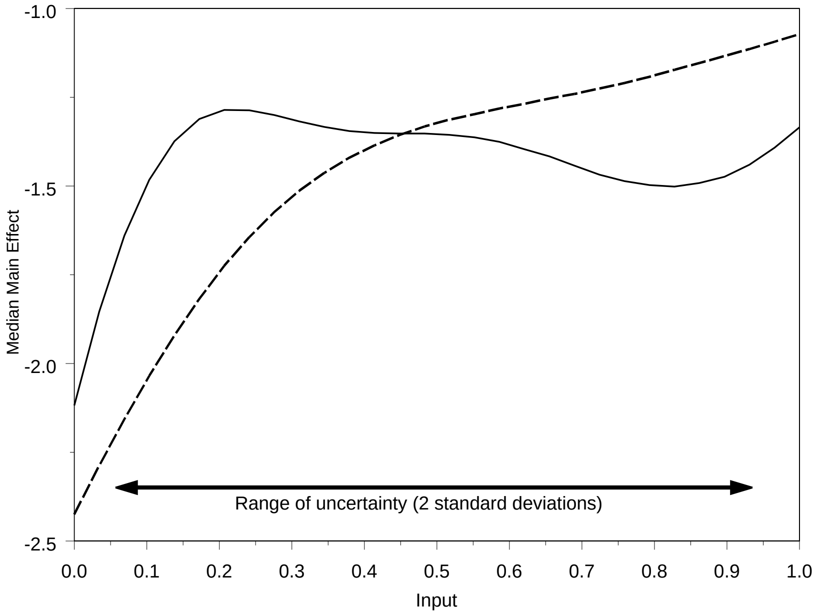

Figure 4: Median main effect plot for the inputs in example 2.

$E_\eta(M|D) = 1.40269$ (with an interval (1.3472,1.4582)). Figure 4 shows that the input's main effect is concave over the range of uncertainty; hence, we should expect the plug-in estimate to be higher than the actual expectation.

Figure 5 is indicative of the input's joint effect. Starting from $\mathbf{x} = (0.5, 0.5)$, the joint effect increases as $\mathbf{x}$ moves towards $(0, 0)$ and $(1, 1)$, and joint effect decreases as $\mathbf{x}$ moves towards $(0, 1)$ and $(1, 0)$. In this case, the joint effect is convex along one diagonal and concave along the other. Thus, the difference in the estimates noted earlier in this example is due to the concave behaviour of the inputs' main effects as the curvature in the joint effect balances out across the range of uncertainty. It should be noted that the opposite can be true: the inputs' main effects can be practically linear and there can still be a difference due to the behaviour of the joint effect.

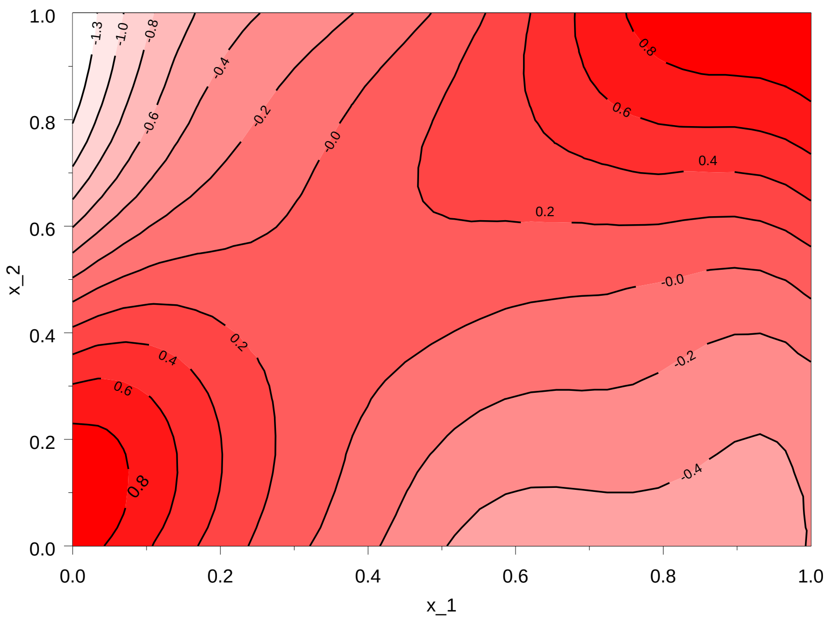

Figure 5: Median joint effect contour plot for example 2.

4 Conclusions

The plug-in estimate for Y will be equal to $E_\eta(M|D)$ if the main effects of each input in a computer code cancel themselves out over the likely intervals specified. This will happen if the main effects of all the inputs are linear as seen in example 2. If an input's main effect is convex over the likely interval, then the plug-in estimate will be lower than $E_\eta(M|D)$ provided the other inputs' main effects do not cancel it out. Conversely, if an input's main effect is concave over the likely interval, then the plug-in estimate will be lower than $E_\eta(M|D)$ provided the other inputs' main effects do not cancel it out. It is also worth noting that main effects may all demonstrate linear behaviour whilst joint effects could be convex or concave.

The main point of this report is to warn users about using the plug-in estimate as an estimate for Y. This is only valid if all the inputs are known for certain or the model is linear. If the inputs are unknown, then using the plug-in estimate could lead to important features of the computer code for likely input combinations being missed.

References

Kennedy, M.C. (2004). Description of the Gaussian process model used in GEM-SA. *GEM-SA help documentation*.

O'Hagan, A. (2006). Bayesian analysis of computer code outputs: a tutorial. *Reliability Engineering and System Safety*, **91**, 1290–1300.