The Human Kernel: Supplementary Materials

Andrew Gordon Wilson CMU
Christoph Dann CMU
Christopher G. Lucas University of Edinburgh
Eric P. Xing CMU

In this appendix we provide some additional experiments regarding the under-fitting property of GP maximum marginal likelihood estimation of kernel length-scales. We also provide instructions and some questions asked in the human experiments. To participate in the exact experiments, and view demonstrations, see http://www.functionlearning.com.

We begin with a brief description of Gaussian processes. For more detail, see Rasmussen and Williams [1].

1 Gaussian Processes

Throughout we assume we have a dataset $\mathcal{D}$ of n input (predictor) vectors $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, each of dimension D, corresponding to a $n \times 1$ vector of targets $\mathbf{y} = (y(\mathbf{x}_1), \ldots, y(\mathbf{x}_n))^\top$.

A Gaussian process (GP) is a collection of random variables, any finite number of which have a joint Gaussian distribution. Using a GP, we can define a distribution over functions $f(\mathbf{x}) \sim \mathcal{GP}(\mu, k)$, meaning that any collection of function values $\mathbf{f}$ has a joint Gaussian distribution:

$$\mathbf{f} = f(X) = [f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n)]^\top \sim \mathcal{N}(\boldsymbol{\mu}, K)\,. \tag{1}$$

The $n \times 1$ mean vector $\boldsymbol{\mu}_i = \mu(\mathbf{x}_i)$, and $n \times n$ covariance matrix $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$, are defined by the user specified mean function $\mu(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})]$ and covariance kernel $k(\mathbf{x}, \mathbf{x}') = \text{cov}(f(\mathbf{x}), f(\mathbf{x}'))$ of the Gaussian process. The smoothness and generalisation properties of the GP are encoded by the covariance kernel and its hyperparameters $\boldsymbol{\theta}$. For example, the popular RBF covariance function, with length-scale hyperparameter ℓ, has the form

$$k_{\text{RBF}}(\mathbf{x}, \mathbf{x}') = \exp(-0.5||\mathbf{x} - \mathbf{x}'||^2/\ell^2)\,. \tag{2}$$

If the targets $y(\mathbf{x})$ are modelled by a GP with additive Gaussian noise, e.g., $y(\mathbf{x})|f(\mathbf{x}) \sim \mathcal{N}(y(\mathbf{x}); f(\mathbf{x}), \sigma^2)$, the predictive distribution at n_* test points X_* is given by

$$\begin{aligned} \mathbf{f}_*|X_*, X, \mathbf{y}, \boldsymbol{\theta}, \sigma^2 &\sim \mathcal{N}(\bar{\mathbf{f}}_*, \text{cov}(\mathbf{f}_*))\,, \\ \bar{\mathbf{f}}_* &= \boldsymbol{\mu}_{X_*} + K_{X_*,X}[K_{X,X} + \sigma^2 I]^{-1}\mathbf{y}\,, \\ \text{cov}(\mathbf{f}_*) &= K_{X_*,X_*} - K_{X_*,X}[K_{X,X} + \sigma^2 I]^{-1}K_{X,X_*}\,. \end{aligned} \tag{3}$$

$K_{X_*,X}$, for example, denotes the $n_* \times n$ matrix of covariances between the GP evaluated at X_* and X. $\boldsymbol{\mu}_{X_*}$ is the $n_* \times 1$ mean vector, and $K_{X,X}$ is the $n \times n$ covariance matrix evaluated at training inputs X. All covariance matrices implicitly depend on the kernel hyperparameters $\boldsymbol{\theta}$.

We can analytically marginalise the Gaussian process $f(\mathbf{x})$ to obtain the marginal likelihood of the data, conditioned only on the kernel hyperparameters $\boldsymbol{\theta}$:

$$\log p(\mathbf{y}|\boldsymbol{\theta}) \propto -[\overbrace{\mathbf{y}^\top (K_{\boldsymbol{\theta}} + \sigma^2 I)^{-1}\mathbf{y}}^{\text{model fit}} + \overbrace{\log |K_{\boldsymbol{\theta}} + \sigma^2 I|}^{\text{complexity penalty}}]\,. \tag{4}$$

Eq. (4) separates into automatically calibrated model fit and complexity terms [2], and can be optimized to learn the kernel hyperparameters $\boldsymbol{\theta}$.

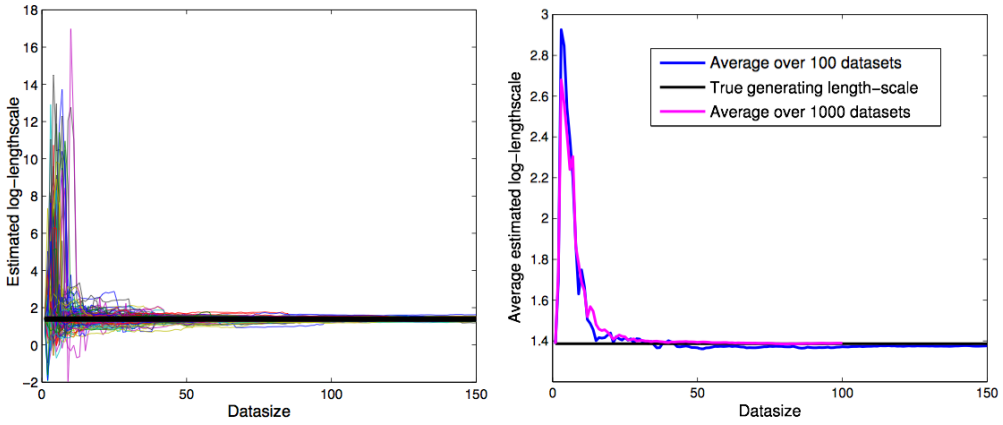

Figure 1: GP Under-fitting via Marginal Likelihood Hyperparameter Estimation. Left: Each curve represents the estimated length-scale of a Gaussian process for a particular dataset, at a given datasize. There are 100 datasets (and thus 100 different coloured curves). Right: The log-lengthscale results in the left plot have been averaged to produce this figure. The results using 1000 datasets are shown in magenta, and they are similar to the results with 100 datasets. These figures consistently show length-scale overestimation, equivalently GP under-fitting, particularly for small $N < 20$ datasets. The standard deviation over the 1000 datasets follows the same trend as the magenta curve with a low of 0 and a high of 7.

2 GP Under-Fitting

To exemplify this surprising under-fitting property of maximum marginal likelihood estimation of kernel hyperparameters, consider the following experiment. We sampled 100 datasets of size N = 150 from a GP with a squared exponential covariance function with a true length-scale of 4, signal standard deviation of 1, and noise standard deviation 0.2, at 1 unit intervals. Using marginal likelihood optimization, we then estimated the kernel hyperparameters (length-scale, signal, and noise stdev) on each of these datasets, as a function of increasing datasize, initializing hyperparameters at their true values. Each curve in Figure 2 (left) shows the estimated log length-scale, for a particular dataset, as a function of datasize. Figure 2 (right) shows the learned length-scale, averaged in log-space, e.g., an estimate of $\mathbb{E}[\log \ell]$ not $\log \mathbb{E}[\ell]$, over the 100 datasets. The truth is in black. The trend is clearest in Figure 2 (right): there is a systematic length-scale overestimation (under-fitting), which is mostly negligable after about $N = 20$ datapoints. In high dimensional input spaces, this under-fitting property may be even more pronounced. As shown in Figure 2 (right) averaging 1000 datasets gives almost exactly the same results (the deviation in these plots is insignificant).

In Figure 2, we show a representative plot of the GP log marginal likelihood as a function length-scale. In this case, there are $N = 5$ datapoints, the true length-scale shown with the dashed green curve, and the mode of the marginal likelihood shown in dashed black. The mode will typically be to the right of the true length-scale, but much of the probability mass will be to the left of the mode. This suggests that sampling, even with a vague (uniform) prior over length-scale, will remove the bias.

We can understand this under-fitting behaviour as follows: If we are unconstrained in estimating the GP covariance matrix, we will converge to the maximum likelihood estimator, $\hat{K} = (\mathbf{y} - \bar{y})(\mathbf{y} - \bar{y})^\top$, which is degenerate and therefore biased. Parametrizing a covariance matrix by a length-scale (for example, by using an RBF kernel), restricts this matrix to a low-dimensional manifold on the full space of covariance matrices. A biased estimator will remain biased when constrained to a lower dimensional manifold, as long as the manifold permits movement along the direction of the bias. Increasing a length-scale moves a covariance matrix towards the degeneracy of the unconstrained maximum likelihood estimator. The low-dimensional manifold becomes more constrained with more data, and less influenced by this under-fitting bias.

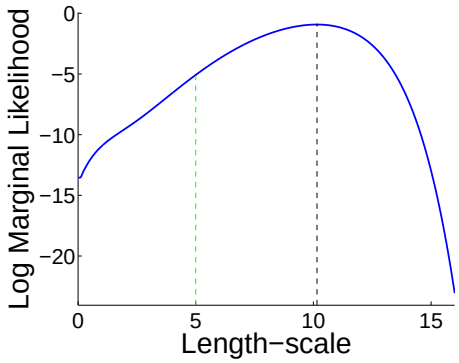

Figure 2: Log marginal likelihood as a function of length-scale. The true length-scale is shown by the dashed green curve, and the mode of the marginal likelihood is shown in dashed black.

3 Materials for Experiment 5.2, Progressive Function Learning

3.1 Introductory text

The goal of this study is to understand what patterns people see in data, and how people think those patterns will continue when they are given incomplete information.

In the following screens, you will try to understand some relationships using plots containing data points.

All of the relationships come from the same underlying system, so there may be similarities between them. Try to use this information to make better predictions.

Based on your understanding of each relationship, you will predict the values of new points.

It should only take a few seconds to make each judgment, and there are approximately 6 relationships requiring approximately 40 judgments each. The experiment is expected to take about 20 minutes in all. Also, please take note of the following:

- There is no single correct pattern in any of the cases you will observe, but your judgments will be reviewed by a human, and your submission may be rejected if it is clear that you did not attempt to find any pattern.
- This study does not currently work with touchscreen devices. If you are using such a device, such as a tablet or smartphone, please return this HIT.
- We apologize for any inconvenience.
- Once you have submitted a judgment, you will not be able to change it – please do not attempt to use the back button during the experiment.

4 Materials for Experiment 2

4.1 Introductory text

The goal of this study is to understand what patterns people see in data, and how people think those patterns will continue when they are given incomplete information.

In the following screens, you will try to understand two distinct relationships using plots containing data points.

The relationships may or may not resemble ones you have seen before.

Based on your understanding of the relationships, you will predict the values of new points.

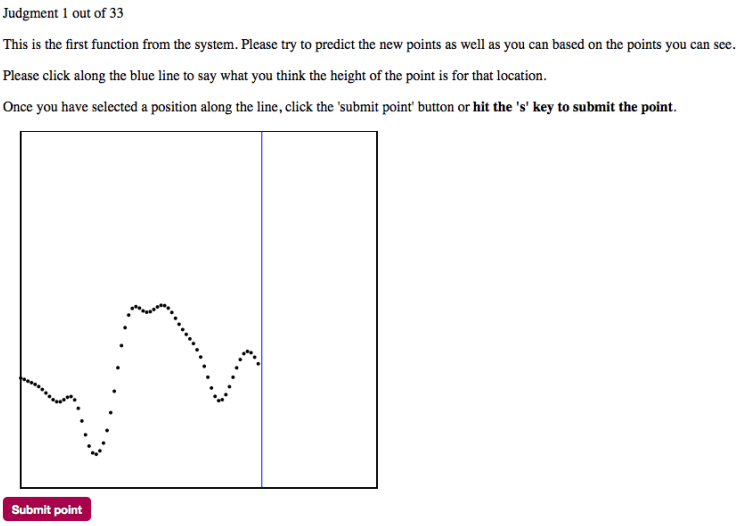

Figure 3: Screenshot of draw 1 from kernel 1

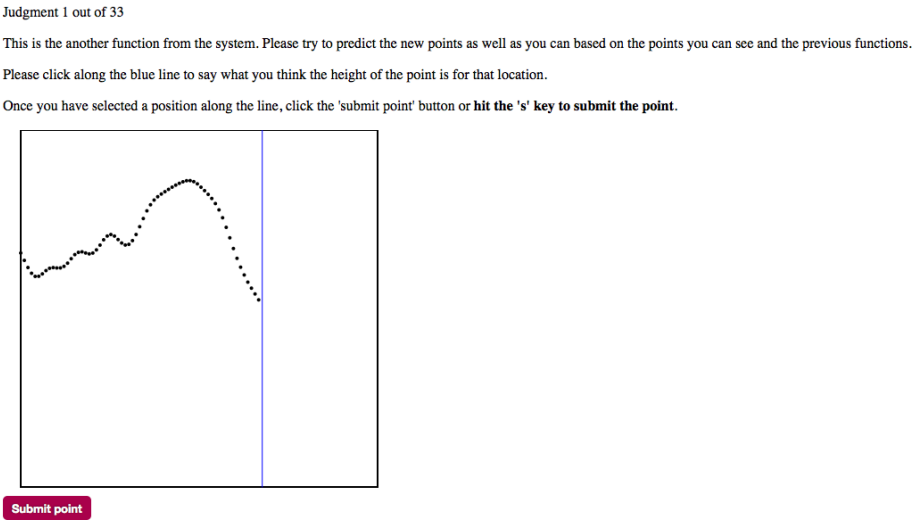

Figure 4: Screenshot of draw 2 from kernel 1

It should only take a few seconds to make each judgment, and the experiment is expected to take fewer than 12 minutes in all. Also, please take note of the following:

- There is no single correct pattern in any of the cases you will observe, but your judgments will be reviewed by a human, and your submission may be rejected if it is clear that you did not attempt to find any pattern.
- This study does not currently work with touchscreen devices. If you are using such a device, such as a tablet or smartphone, please return this HIT.
- We apologize for any inconvenience.
- Once you have submitted a judgment, you will not be able to change it – please do not attempt to use the back button during the experiment.

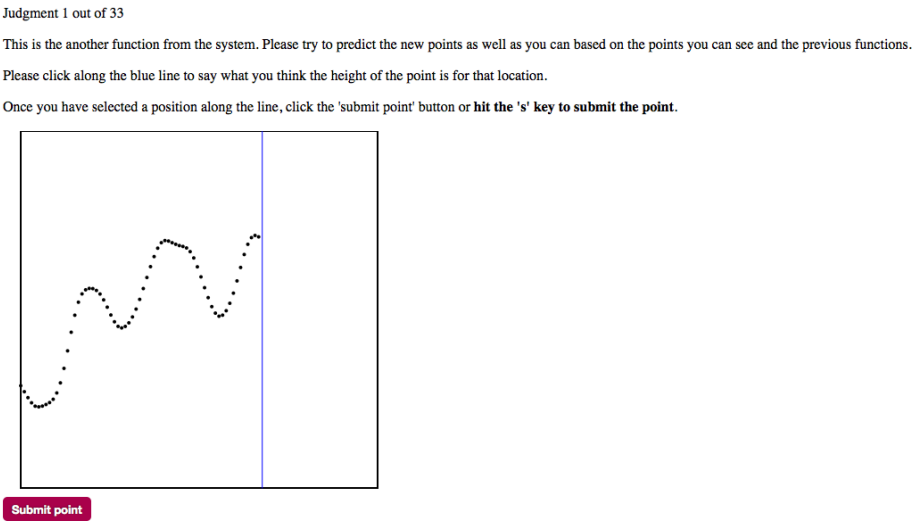

Figure 5: Screenshot of draw 3 from kernel 1

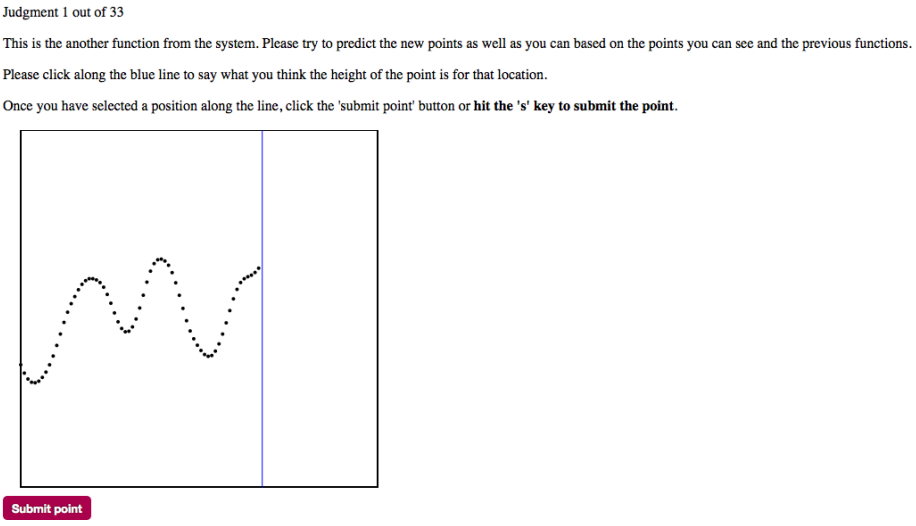

Figure 6: Screenshot of draw 4 from kernel 1

5 Materials for Experiment 5.3, Learning Unconventional Kernels

6 Materials for Experiment 5.4, Human Occam's Razor

References

[1] C. E. Rasmussen and C. K. I. Williams. *Gaussian processes for Machine Learning*. MIT Press, 2006.

[2] Carl Edward Rasmussen and Zoubin Ghahramani. Occam's razor. In *Neural Information Processing Systems (NIPS)*, 2001.

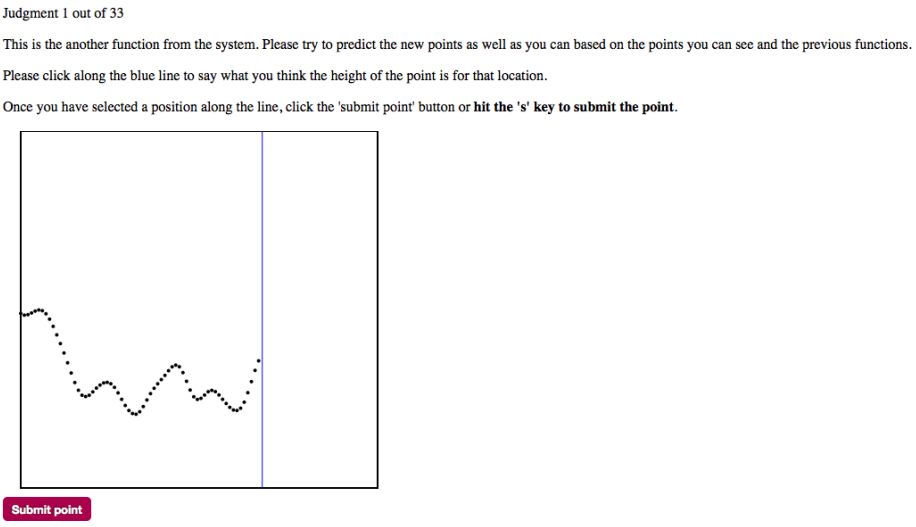

Figure 7: Screenshot of draw 5 from kernel 1

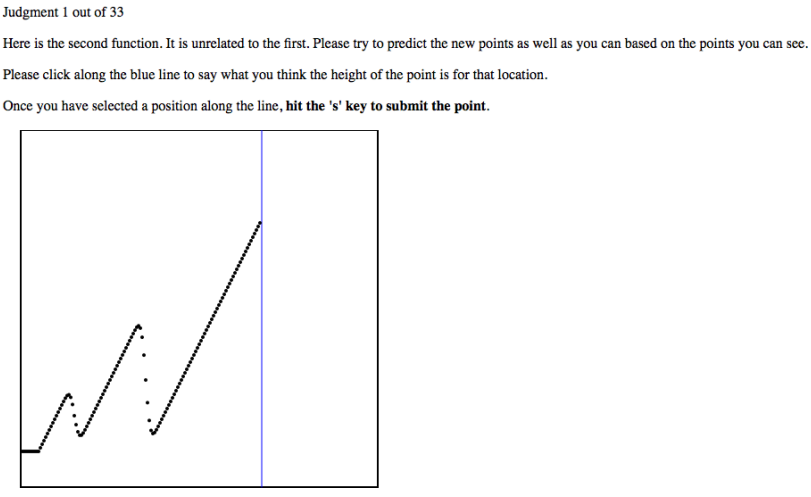

Figure 8: Screenshot of the saw function in Experiment 2.

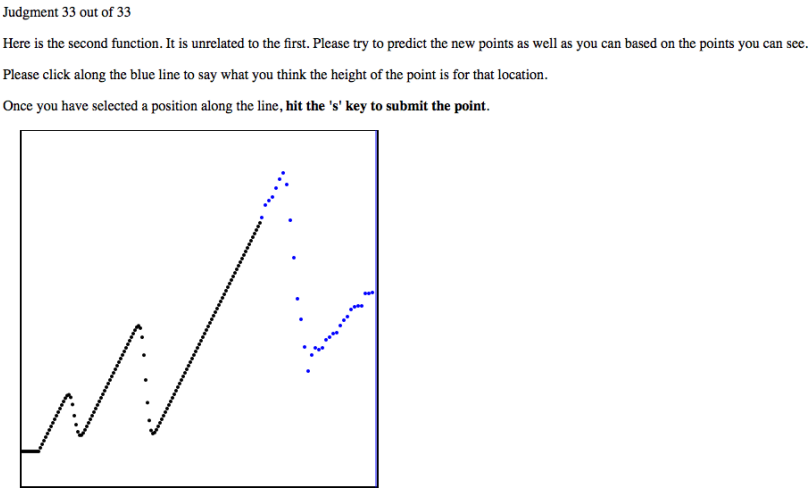

Figure 9: Screenshot of the saw function in Experiment 2 with a set of example judgments.

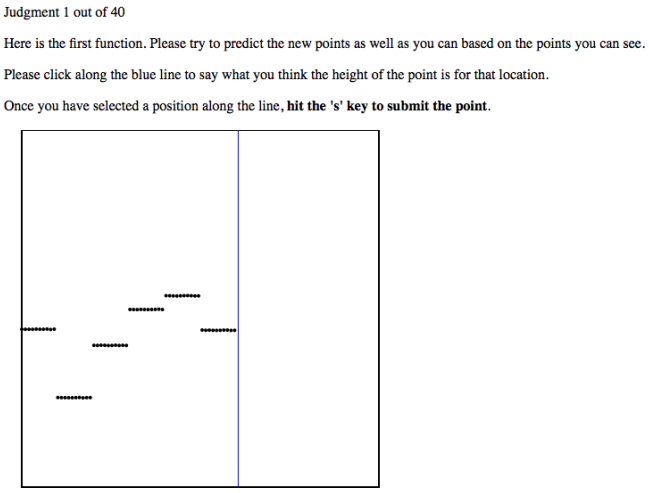

Figure 10: Screenshot of the step function in Experiment 2.

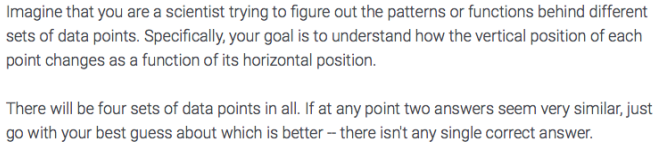

Figure 11: Screenshot of the introduction to Experiment 3.

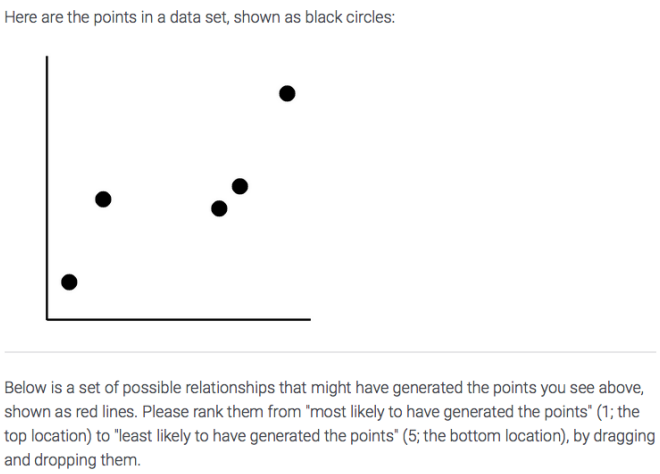

Figure 12: Screenshot of the introduction to one question in Experiment 3.

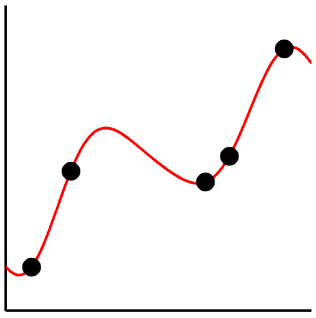

Figure 13: Function option in Experiment 3: fit using maximum marginal likelihood (MM) length scale.

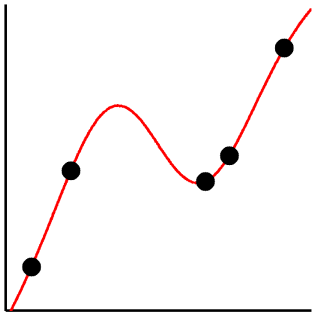

Figure 14: Function option in Experiment 3: Actual point-generating function.

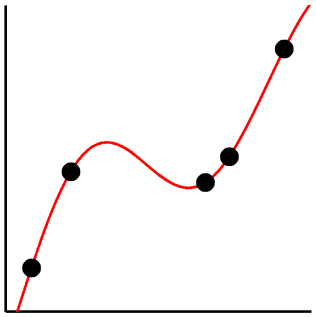

Figure 15: Fit using MM scale times exp(1.0) .

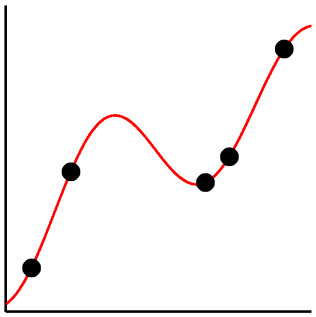

Figure 16: Fit using MM scale times exp$(.5)$.

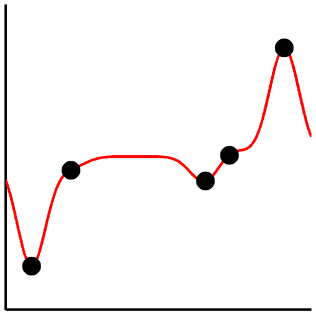

Figure 17: Fit using MM scale times exp(-1.0) .

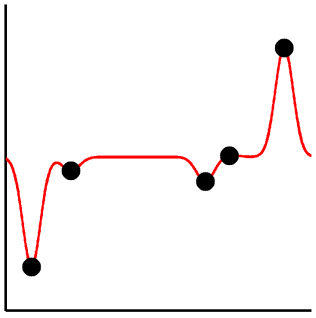

Figure 18: Fit using MM scale times exp(-1.5) .

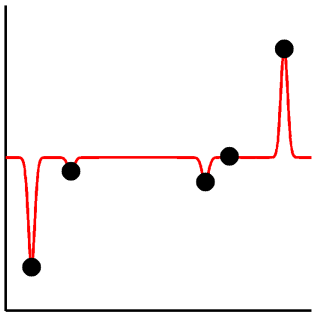

Figure 19: Fit using MM scale times $\exp(-2.5)$.

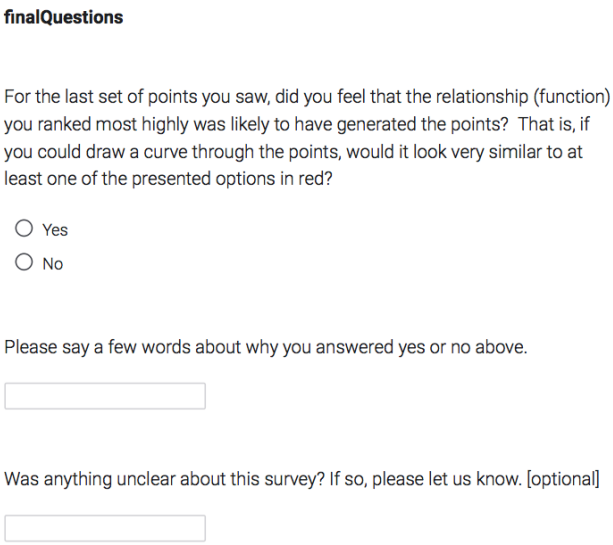

Figure 20: Screenshot of the final page of Experiment 3.