Closed-loop automatic experimentation for Bayesian optimisation

Dave Woods

Southampton Statistical Sciences Research Institute
University of Southampton, UK
D.Woods@southampton.ac.uk
www.southampton.ac.uk/~davew

Outline

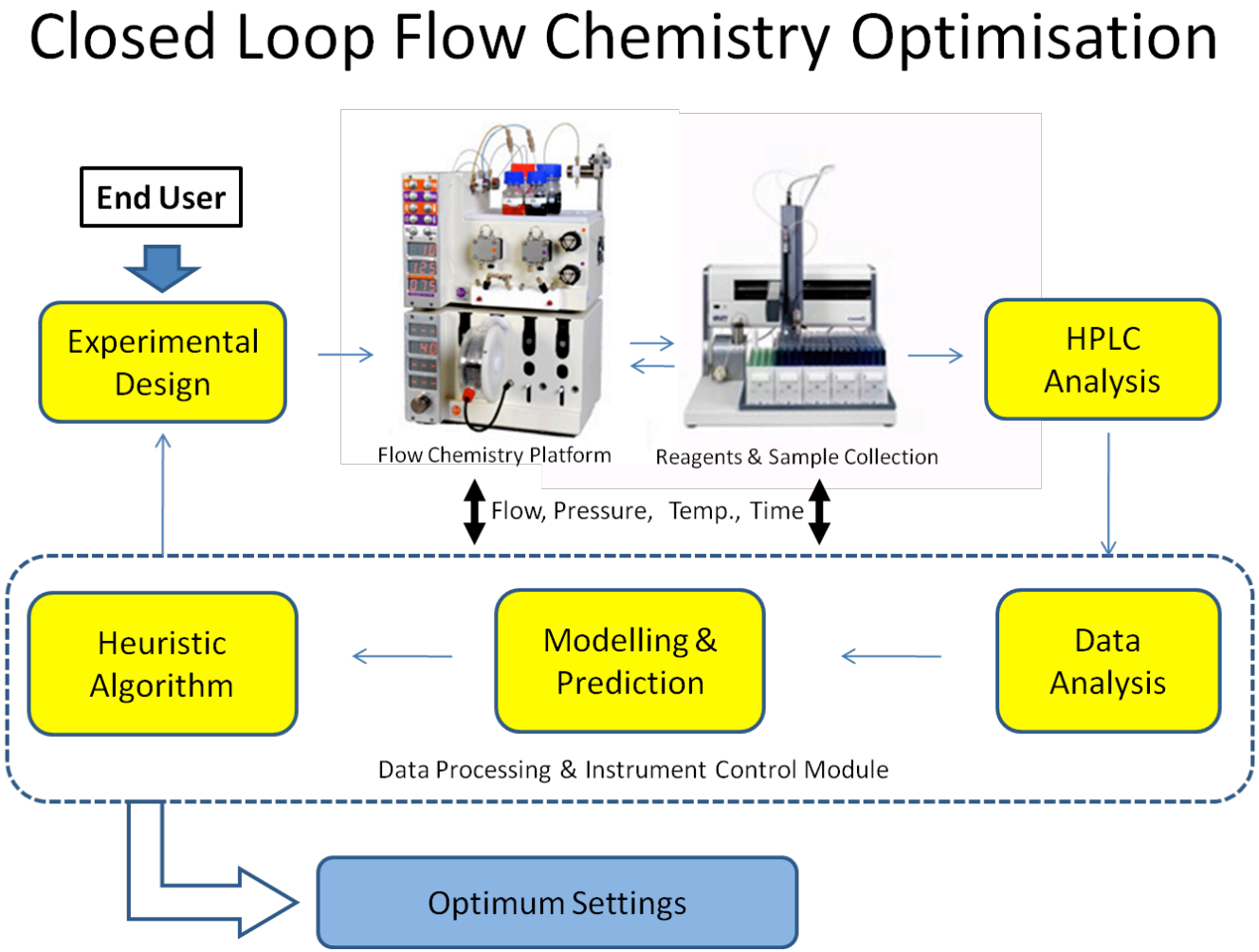
- Motivation: closed-loop optimisation of chemical processes
- Decision-theoretic sequential design of experiments for Bayesian optimisation
- Learning and down-weighting outliers
- Examples

Joint work with Tim Waite (University of Manchester, UK)

Chemistry examples provided by Richard Whitby, Ian Broadwell (University of Southampton) and Xiaoping Tang (Pharmaron)

Supported by the UK Engineering and Physical Sciences Research Council

Motivation


Closed Loop Flow Chemistry Optimisation
End User
Experimental Design
Flow Chemistry Platform
Reagents & Sample Collection
HPLC Analysis
Flow, Pressure, Temp., Time
Heuristic Algorithm
Modelling & Prediction
Data Analysis
Data Processing & Instrument Control Module
Optimum Settings

Automated experimentation

Automation **allows** more (economically, time) efficient experimentation, 24/7 lab operation, reduction in user errors, and wider use of better designed experiments

Automation **requires** principled and computationally efficient algorithms

- Generation of (one-point-at-a-time) sequential designs
- **Robust** modelling of sequentially collected data making minimal assumptions
- Identification of possible outliers

We address these issues through

- Gaussian process regression and Bayesian updating
- Decision-theoretic sequential design for optimisation
- (Re-) weighting of observations to reflect our belief in their reliability

Bayesian optimisation

Assume the **aim** is to maximise an unknown, and expensive, function $g(\boldsymbol{x})$ with respect to q controllable variables $\boldsymbol{x} = (x_1, \ldots, x_q)^{\mathrm{T}} \in \mathcal{X} \subset \mathbb{R}^q$

[Here, expensive will mean an experiment is required to obtain a (noisy) evaluation of $g(\boldsymbol{x})$ at any $\boldsymbol{x}$]

Bayesian optimisation

- places a prior distribution on $g(\boldsymbol{x})$ (i.e. a statistical model)
- collects (noisy) function evaluations (data) at points chosen sequentially via an acquisition function
- updates the prior to a posterior distribution and infers the maximum of $g(\boldsymbol{x})$

Uncertainty in the posterior leads to an exploration/exploitation trade-off

In the statistics literature, **Expected Improvement (EI)** using **Gaussian processes** is the most popular acquisition function

Mockus (1989), Jones et al, (1998), Brochu et al., 2010

Gaussian process regression

Bayesian Gaussian process models: nonparametric modelling of a black box function $g(\boldsymbol{x})$

Assume we can observe $y = g(\boldsymbol{x}) + \varepsilon$, for $\varepsilon \sim N(0, \sigma_\varepsilon^2)$

Gaussian process prior: $g(\boldsymbol{x}) \sim \text{GP}\left\{\mu(\boldsymbol{x}), \kappa(\boldsymbol{x}, \boldsymbol{x}')\right\}$

$$g(\boldsymbol{x}_1), \ldots, g(\boldsymbol{x}_n) \sim \text{MVN}(F\boldsymbol{\beta}, \sigma_g^2 K)$$

F = model matrix for mean **trend**

K = **correlation** matrix with ijth entry $k_{\boldsymbol{\lambda}}(\boldsymbol{x}_i, \boldsymbol{x}_j)$

Correlation function $k_{\boldsymbol{\lambda}}(\boldsymbol{x}, \boldsymbol{x}')$ often assumed stationary and separable in $\boldsymbol{x}$, with parameter $\boldsymbol{\lambda}$

Conditional on the hyperparameters, the posterior for $g(\boldsymbol{x})$ is also a Gaussian process; unconditional posterior inference requires numerical methods

Rasmussen & Williams (2006)

Decision-theoretic approach 1

Classic EI is for the case of non-noisy observations, $\sigma_\varepsilon^2 = 0$.

Most extensions to EI for noisy data are heuristic generalisations that do not take account of parameter updating or inference for the maximum

Huang et al. (2006), Gramacy & Polson (2011)

Problem 1. Estimation: Given data $\boldsymbol{y}_{1:n} = (y_1, \ldots, y_n)^{\mathrm{T}}$, find a point estimate $\boldsymbol{x}_n^\star \in \mathcal{M}_n \subseteq \mathcal{X}$ for the optimum

We assume the utility of a point $\boldsymbol{x}$ is

$$u(\boldsymbol{x}, g) = g(\boldsymbol{x}) \,,$$

with corresponding Bayes decision

$$\boldsymbol{x}_n^\star = \mathrm{argmax}_{\boldsymbol{x} \in \mathcal{M}_n} \, \mathrm{E}_g \left\{ g(\boldsymbol{x}) \mid \boldsymbol{y}_{1:n} \right\}$$

Decision-theoretic approach 2

Now suppose one additional observation, y_{n+1}, will be taken and the posterior mean $\mathrm{E}\{g(\boldsymbol{x})|\boldsymbol{y}_{1:(n+1)}\}$ used to locate $\boldsymbol{x}^{\star}_{n+1} \in \mathcal{M}_{n+1}$

Problem 2. Design: the optimal choice of $\boldsymbol{x}_{n+1}$, the next design point, maximises

$$\begin{aligned} U(\boldsymbol{x}_{n+1}) &= \mathrm{E}_{g,y_{n+1}} \left\{ u(\boldsymbol{x}^{\star}_{n+1}, g) \mid \boldsymbol{y}_{1:n} \right\} \\ &= \mathrm{E}_{y_{n+1}} \left[\max_{\boldsymbol{x} \in \mathcal{M}_{n+1}} \mathrm{E}_g \left\{ g(\boldsymbol{x}) \mid \boldsymbol{y}_{1:(n+1)} \right\} \mid \boldsymbol{y}_{1:n} \right] \end{aligned}$$

For interpretation and presentation, we usually use the **Expected Gain in Utility (EGU)**

$$U(\boldsymbol{x}_{n+1}) - \mathrm{E}_g \left\{ g(\boldsymbol{x}^{\star}_n) \mid \boldsymbol{y}_{1:n} \right\}$$

see also Osbourne et al. (2010)

Two simple results

1. Expected improvement is a special case of EGU ...

... if the responses are deterministic, $y = g(\boldsymbol{x})$, $\mathcal{M}_n = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, and the GP hyperparameters are known

Here, $\boldsymbol{x}_n^\star = \text{argmax}_{i=1\ldots,n} g(\boldsymbol{x}_i)$ and hence

$$\begin{aligned} U(\boldsymbol{x}_{n+1}) &= \mathrm{E}_g \left\{ \max\left(g(\boldsymbol{x}_{n+1}), g(\boldsymbol{x}_n^\star) \right) \mid \boldsymbol{y}_{1:n} \right\} \\ &= g(\boldsymbol{x}_n^\star) + \mathrm{E}_g \left\{ \max\left(g(\boldsymbol{x}_{n+1}) - g(\boldsymbol{x}_n^\star), 0 \right) \mid \boldsymbol{y}_{1:n} \right\} \\ &= g(\boldsymbol{x}_n^\star) + \mathsf{EI}(\boldsymbol{x}_{n+1}) \end{aligned}$$

2. The EGU is non-negative ...

... if $\mathcal{M}_n = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$. Then,

$$U(\boldsymbol{x}_{n+1}) \geq \mathrm{E}_g \left\{ g(\boldsymbol{x}_n^\star) \mid \boldsymbol{y}_{1:n} \right\}$$

Updating the GP and approximating EGU

Updating is via Sequential Monte Carlo with N_θ particle/weight pairs $(\boldsymbol{\theta}_j, w_j)$, where $\boldsymbol{\theta}$ holds values of the hyperparameters and $0 < w_j \leq 1$, $\sum_j w_j = 1$

- Reweight, resample and move steps after observation of y_{n+1}
- Iterated batch importance sampling (IBIS)

Chopin (2002), Gramacy & Polson (2011), Drovandi et al. (2013)

Approximation of EGU presents two main challenges:

1. for given y_{n+1}, we must approximate the posterior for $\boldsymbol{\theta}$;
2. the expectation must be taken with respect to the predictive distribution of y_{n+1}

Approximating EGU 1

Let $u(\boldsymbol{x}_{n+1};\, y_{n+1}) = \max_{x \in \mathcal{M}_{n+1}} \mathrm{E}_g \left\{ g(\boldsymbol{x}) \mid \boldsymbol{y}_{1:(n+1)} \right\}$

Then we need to approximate

$$U(\boldsymbol{x}_{n+1}) = \int_{\mathbb{R}} u(\boldsymbol{x}_{n+1};\, y_{n+1}) \pi(y_{n+1} \mid \boldsymbol{y}_{1:n})\, \mathrm{d}y_{n+1}$$

1. Particle approximation for the predictive distribution:

$$\hat{\pi}(y_{n+1} \mid \boldsymbol{y}_{1:n}) = \sum_j w_j \pi(y_{n+1} \mid \boldsymbol{y}_{1:n}, \boldsymbol{\theta}_j)$$

2. Quadrature approximation for the integral w.r.t. y_{n+1}:

$$\hat{U}(\boldsymbol{x}_{n+1}) = \sum_k v_k \hat{\pi}(y_{n+1}^{(k)} \mid \boldsymbol{y}_{1:n}) u(\boldsymbol{x}_{n+1};\, y_{n+1}^{(k)}),$$

with, v_k, $y_{n+1}^{(k)}$ are integration weights and abscissae

Approximating EGU 2

To estimate future maximised expected utility $u(\boldsymbol{x}_{n+1};\, y_{n+1})$:

1. Given y_{n+1}, calculate an approximate posterior density for $\boldsymbol{\theta}$ via simple reweighting, i.e. using $(\boldsymbol{\theta}_j, \tilde{w}_j)$ with

$$\tilde{w}_j \propto w_j \pi(y_{n+1} \mid \boldsymbol{y}_{1:n}, \boldsymbol{\theta}_j)$$

2. Use these weights to approximate the future posterior expectation $\mathrm{E}_g\left\{g(\boldsymbol{x}) \mid \boldsymbol{y}_{1:(n+1)}\right\}$, giving

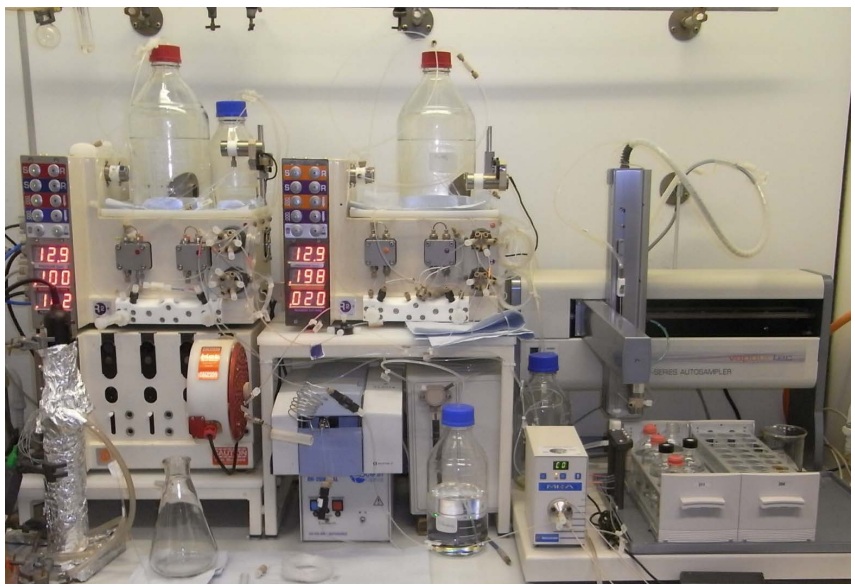
$$\hat{u}(\boldsymbol{x}_{n+1};\, y_{n+1}) = \max_{\boldsymbol{x} \in \mathcal{M}_{n+1}} \left\{ \sum_j \tilde{w}_j \, \mathrm{E}_g\left[g(\boldsymbol{x}) \mid \boldsymbol{y}_{1:(n+1)}, \boldsymbol{\theta}_j\right] \right\}$$

Example 1

Real piece of chemistry run in Southampton Chemistry with online statistical modelling performed at Manchester

Hardware: 4-channel (A–D) Vapourtec flow reactor

Measurement via HPLC/UV



Average time for 1 run is 2 hours; practical upper bound of 30-40 runs

Variables under study

- Conc A - 0.3–1 mol dm^{-3}
- Ratio B - 1–5 eq
- Residence time - 5–30 min
- Temperature - 30–180° C

Ratio D = 0.1 held fixed, Ratio C determined by Conc A

Response: product yield

Some combinations of Conc-Ratio are not physically feasible (as they would imply infeasibly low or negative flow rates)

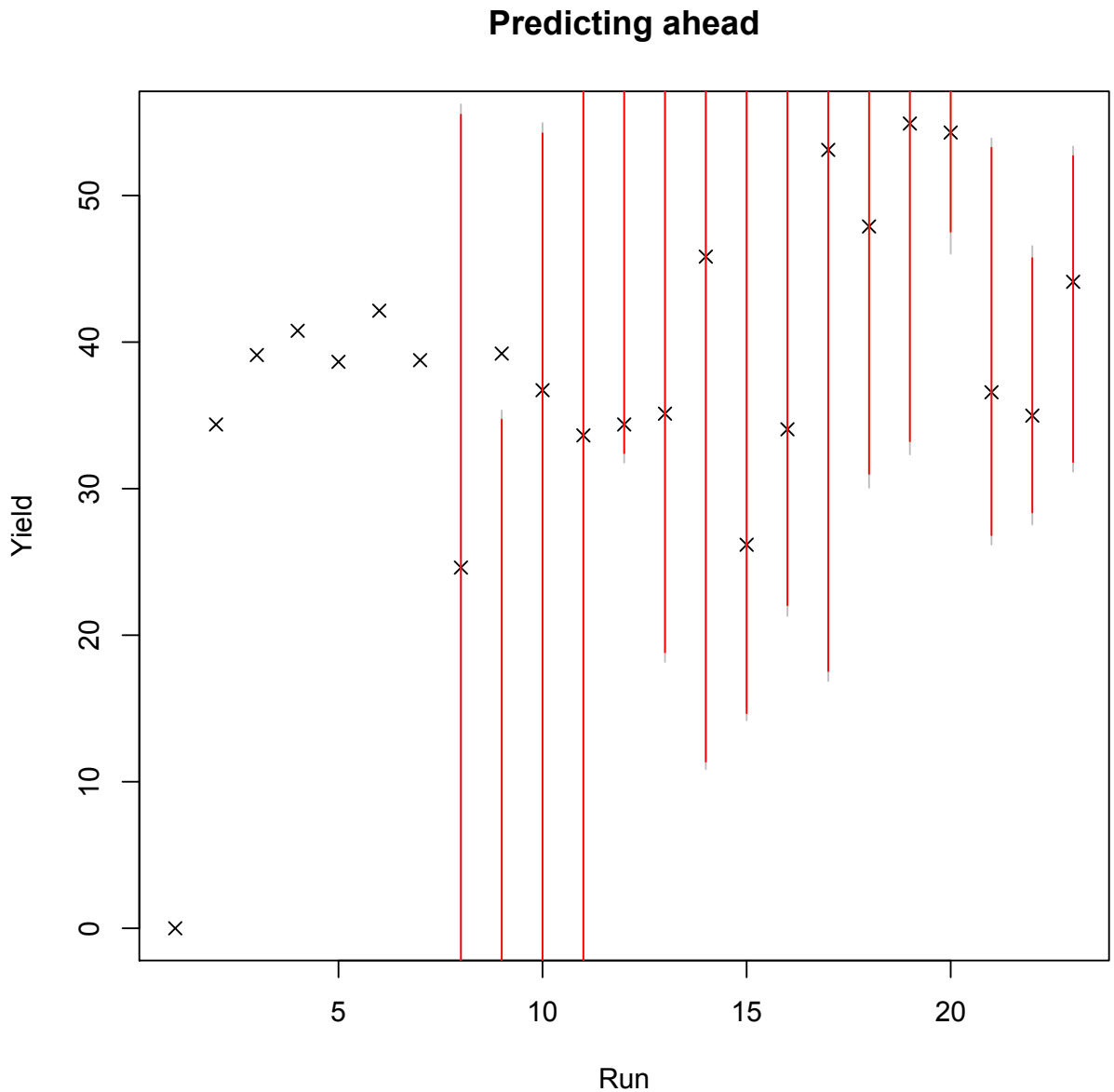
Initial design: 12 run design - 9 space-filling points (maximum-projection) + three centre points

Joseph et al. (2015)

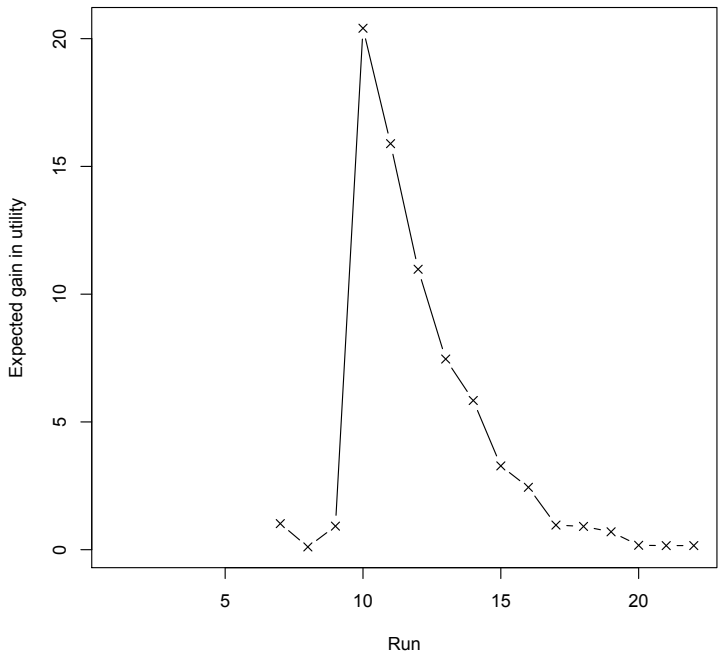
Stopping rule: $U(\boldsymbol{x}_{n+1}) - \mathrm{E}_g\{g(\boldsymbol{x}_n^\star) \mid \boldsymbol{y}_{1:n}\} \leq 0.1$

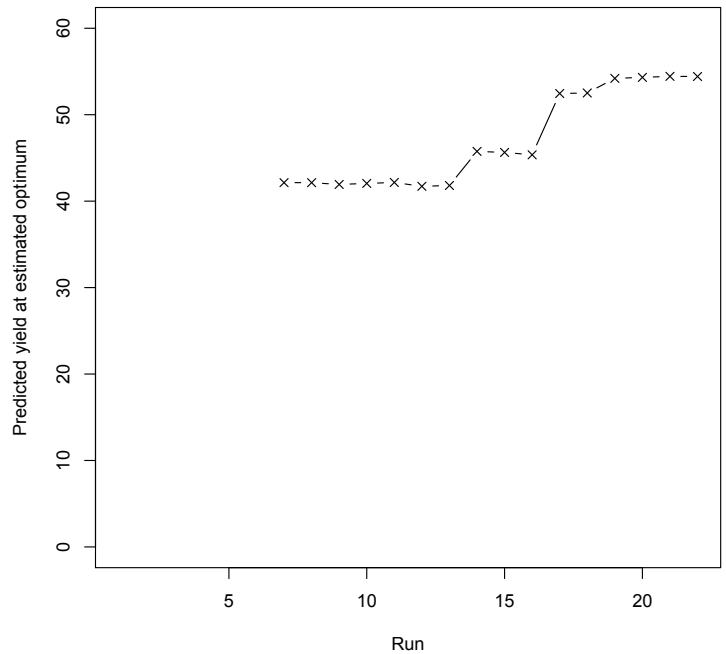
Predicting forward using the sequential model

23 automated runs, with 11 chosen via EGU

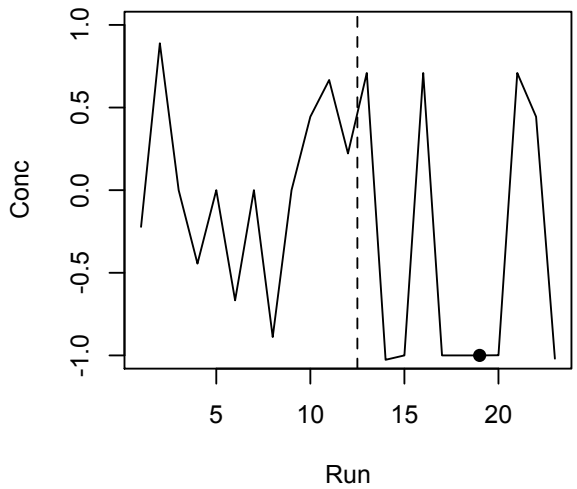


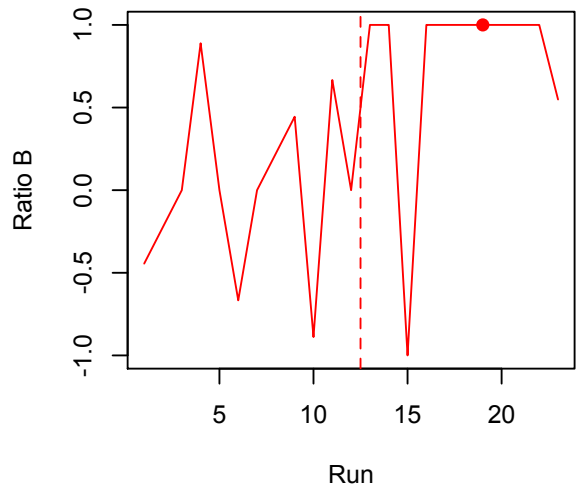
Progress of the algorithm

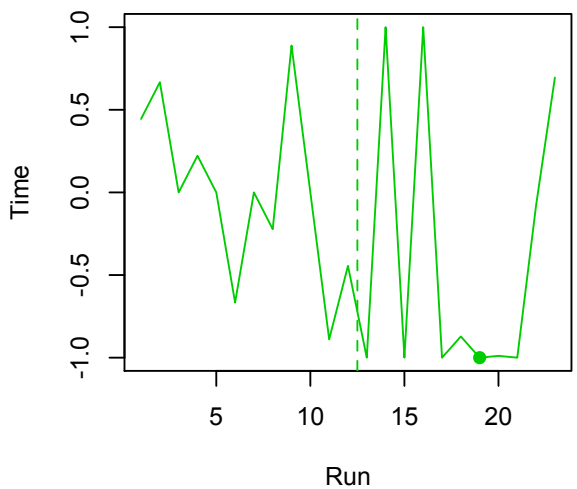


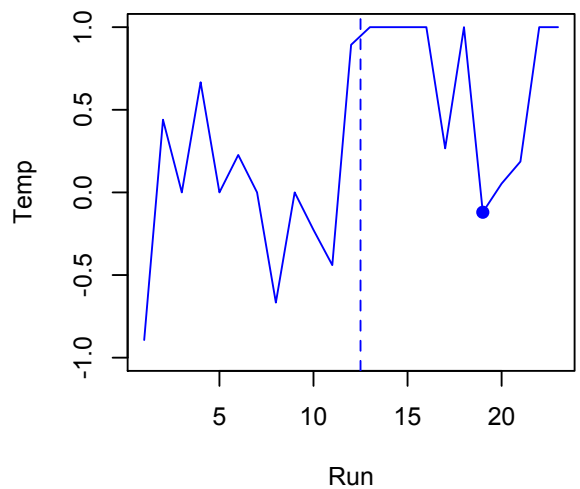


Design points

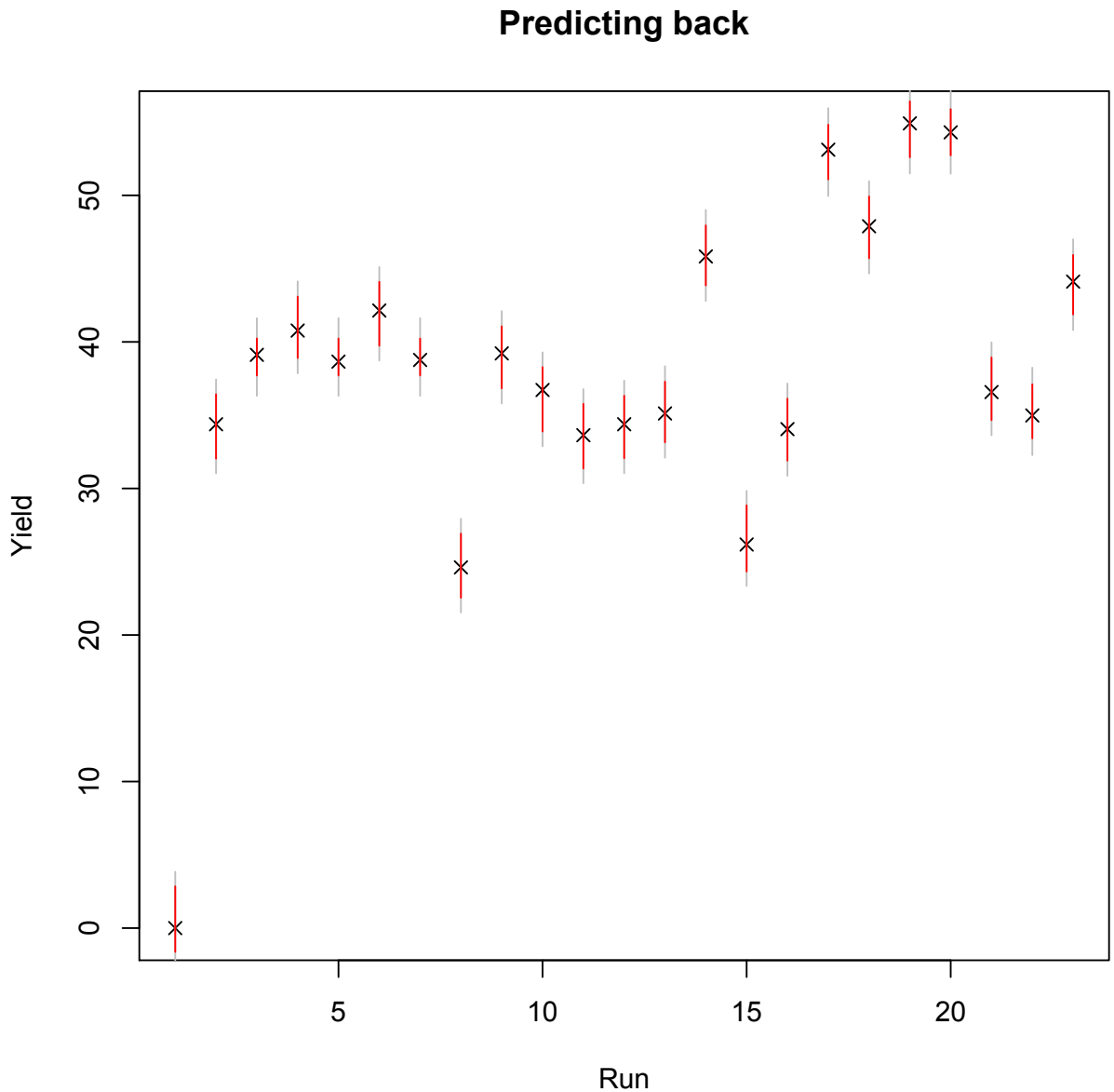








Predicting backwards from the final model



Robustifying the model

Outliers and unusual observations are a fact of life, and are not uncommon in chemistry experiments

Although we may wish to stop the system and investigate outlying observations manually, it is also useful to have automatic methods

To facilitate automatic outlier detection and removal, we extend the model such that

$$y_i \mid g, \sigma_\varepsilon^2, r_i \sim N\left(g(x_i), \sigma_\varepsilon^2 / r_i\right), \quad i = 1, \ldots, n$$

With $r_i \sim \text{Gamma}(a_r/2, b_r/2)$, the response y_i has a non-standardised ***t*-distribution** conditional on g and σ_ε^2

We can use SMC and EGU with this model

cf Neal (1997)

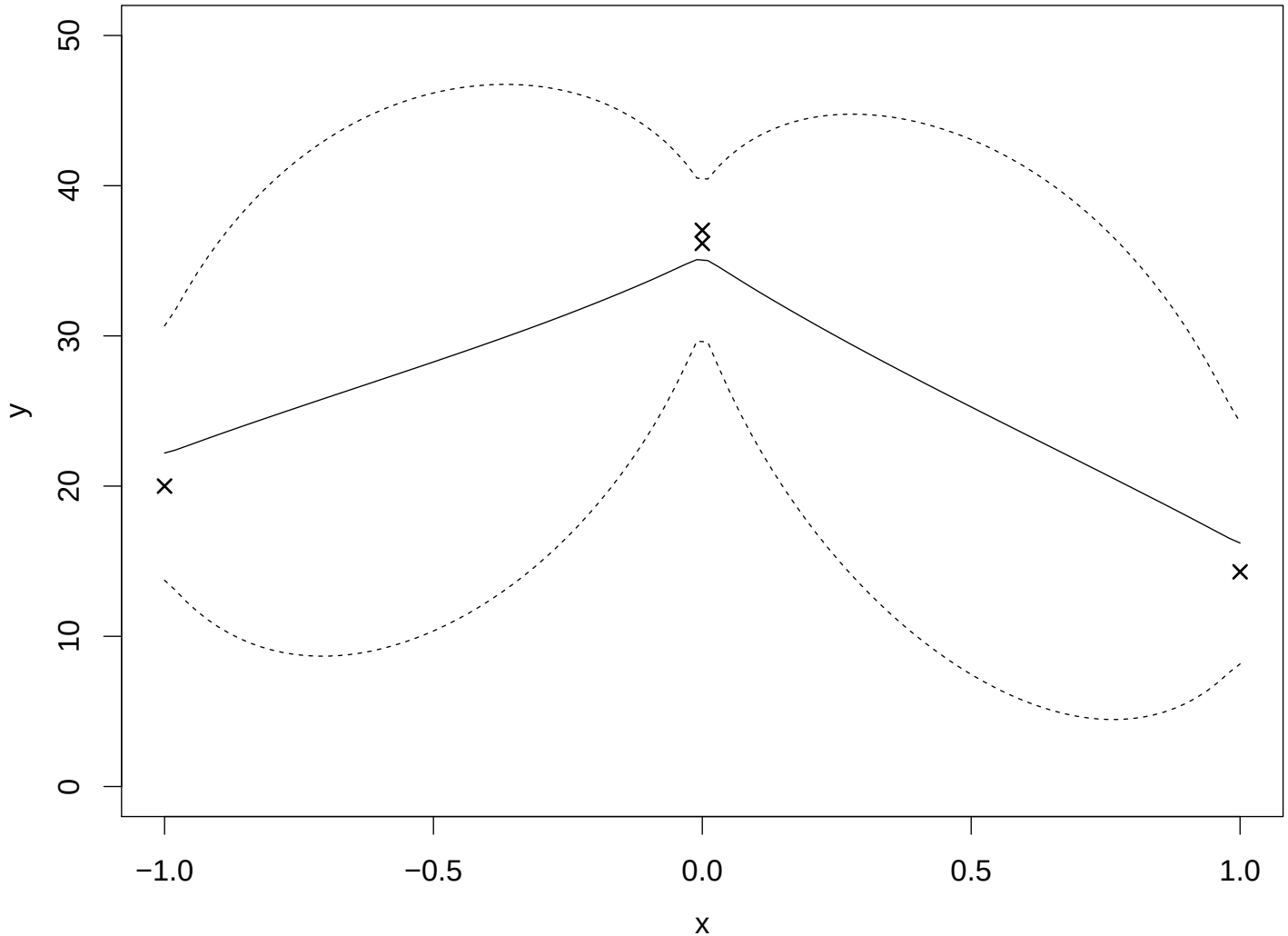
SMC for the t-response model

The parameters now consist of $(\boldsymbol{\theta}, \boldsymbol{r}_{1:n})$, and the dimension increases whenever we observe another data point. Thus IBIS is not applicable

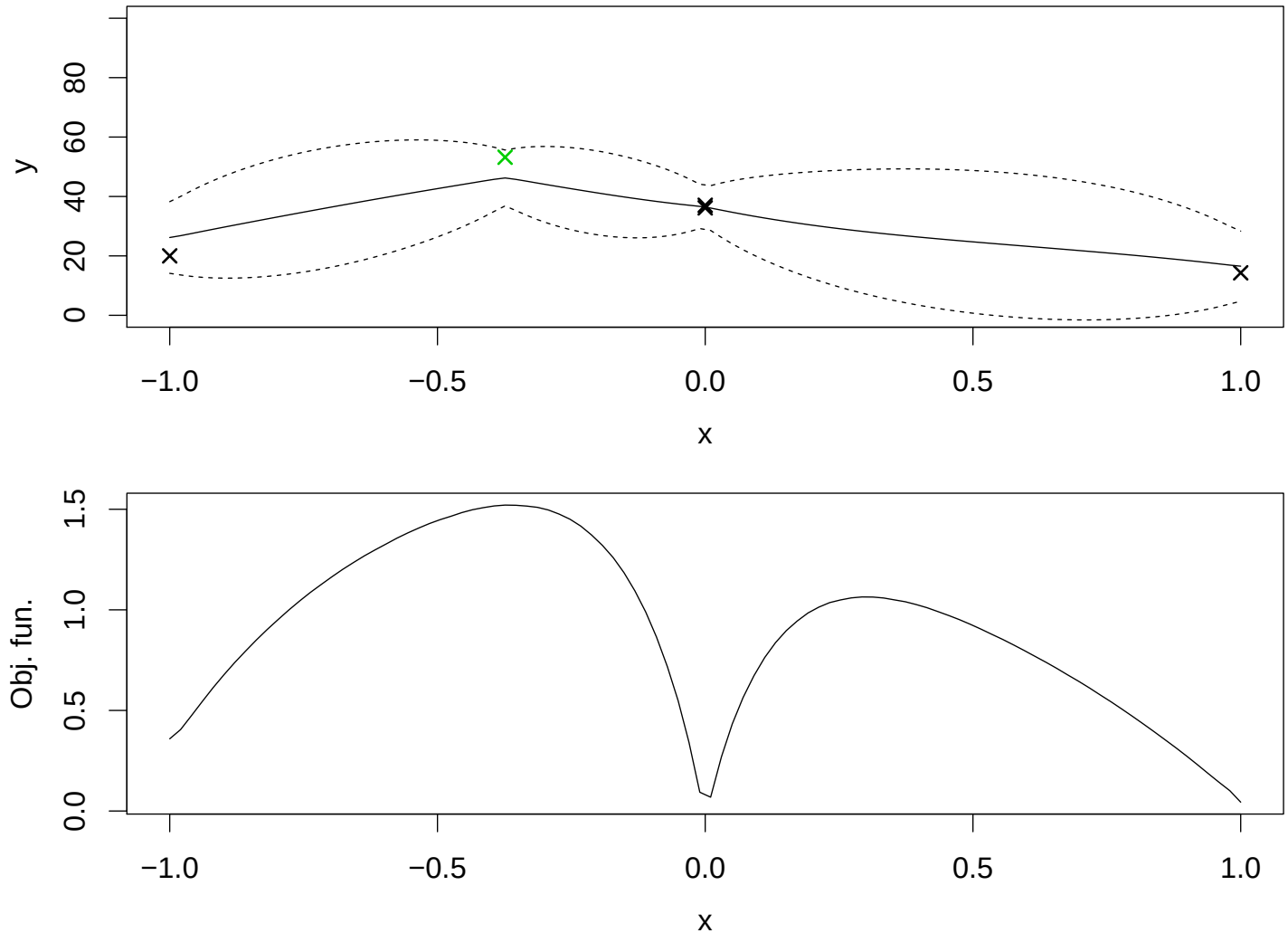
Algorithm on observing y_{n+1},

1. **Resample** particles $(\boldsymbol{\theta}_j, \boldsymbol{r}_{1:n,j})$ with resampling weights proportional to $\pi(y_{n+1}|\boldsymbol{\theta}_j, \boldsymbol{r}_{1:n,j}, \boldsymbol{y}_{1:n})$, approximated using Gauss-Laguerre quadrature
2. **Propagate** each of the resampled particles, by generating a value r_{n+1} from $\pi(r_{n+1} \mid \boldsymbol{\theta}, \boldsymbol{r}_{1:n}, \boldsymbol{y}_{1:(n+1)})$, via a Gibbs sampler
3. **Move** the particles using an MCMC kernel with invariant distribution $\pi(\boldsymbol{\theta}, \boldsymbol{r}_{1:(n+1)} \mid \boldsymbol{y}_{1:(n+1)})$

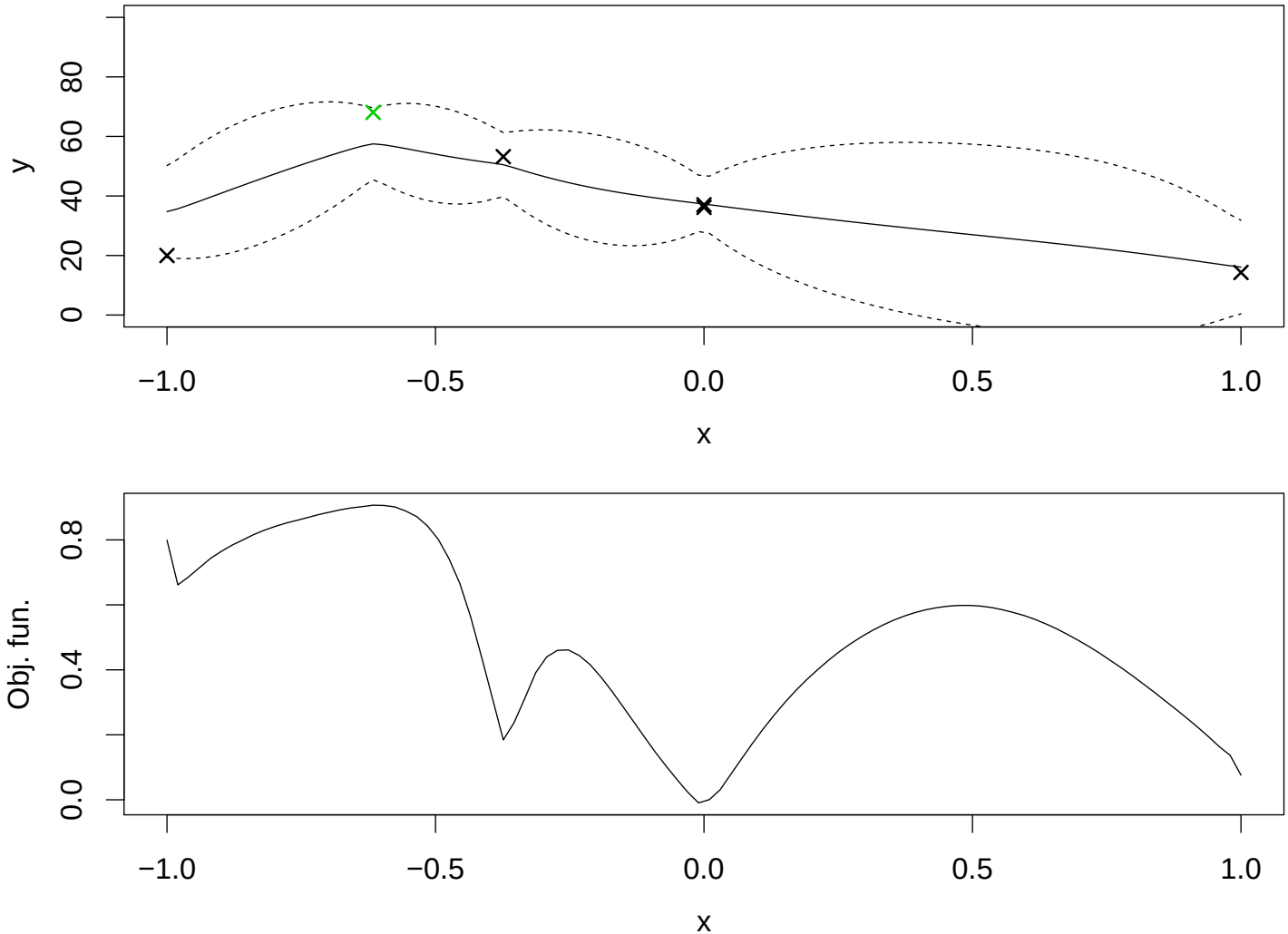
Example 2 - EGU, t-response model



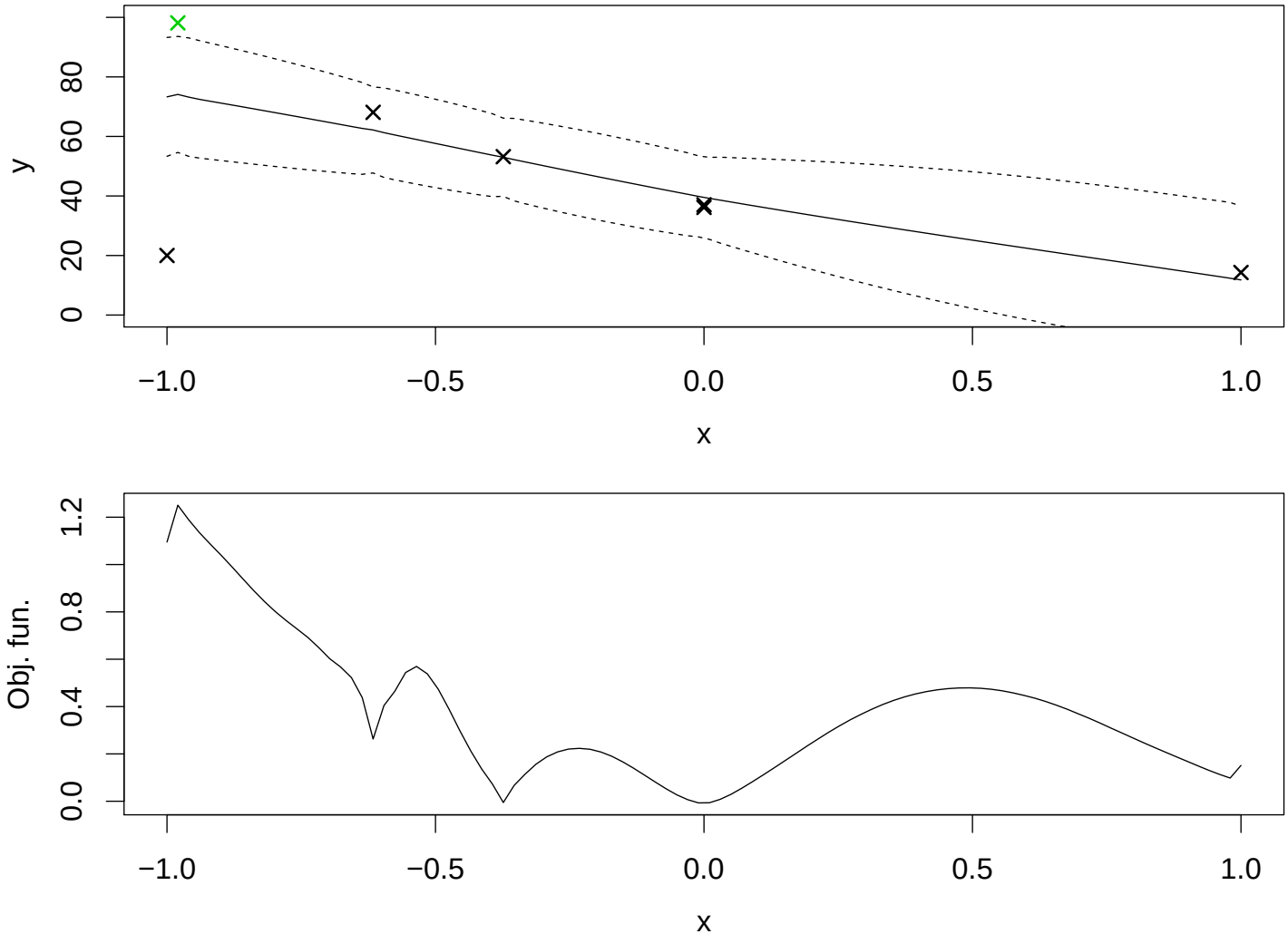
Example 2 - EGU, t-response model



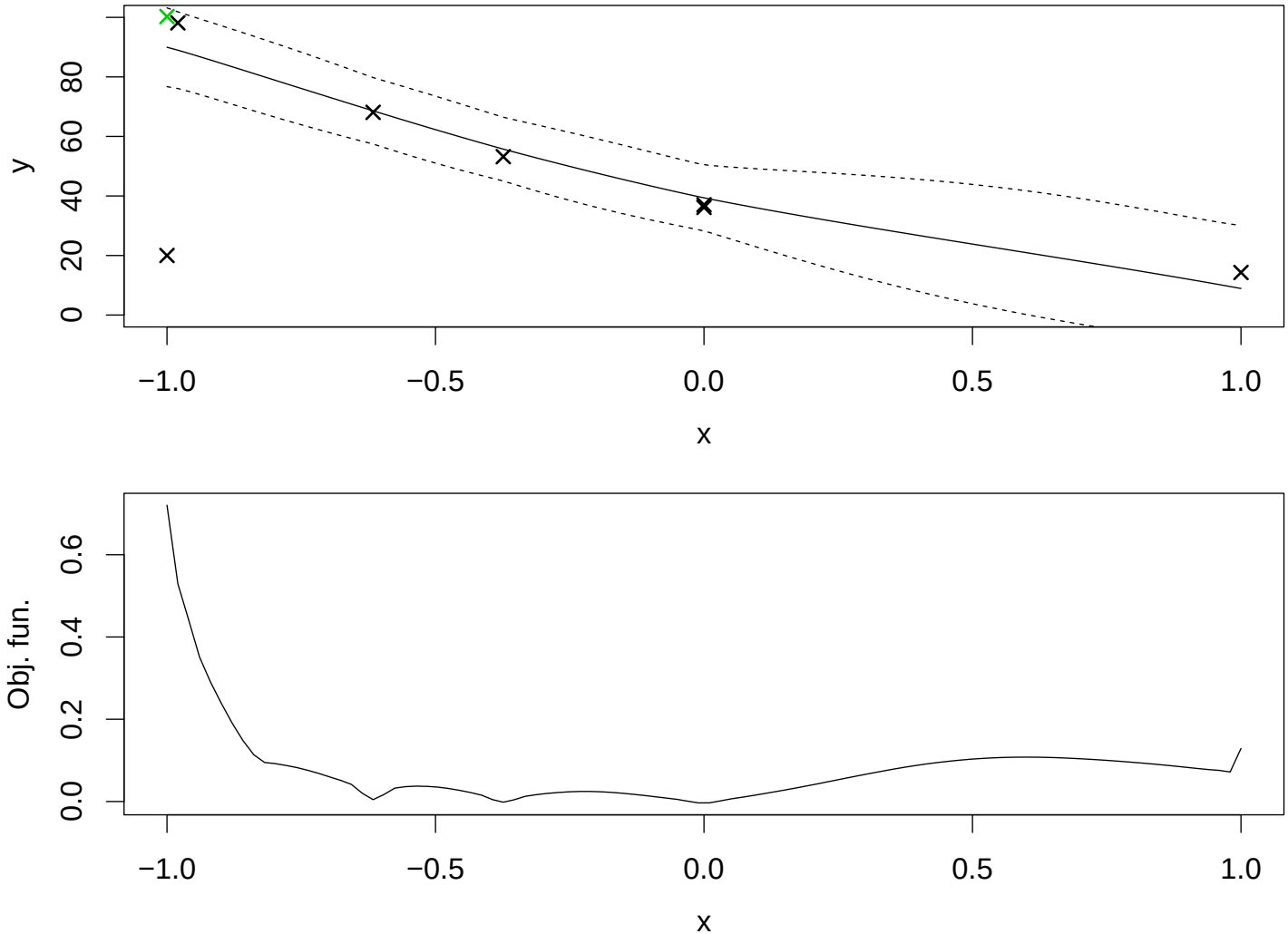
Example 2 - EGU, t-response model



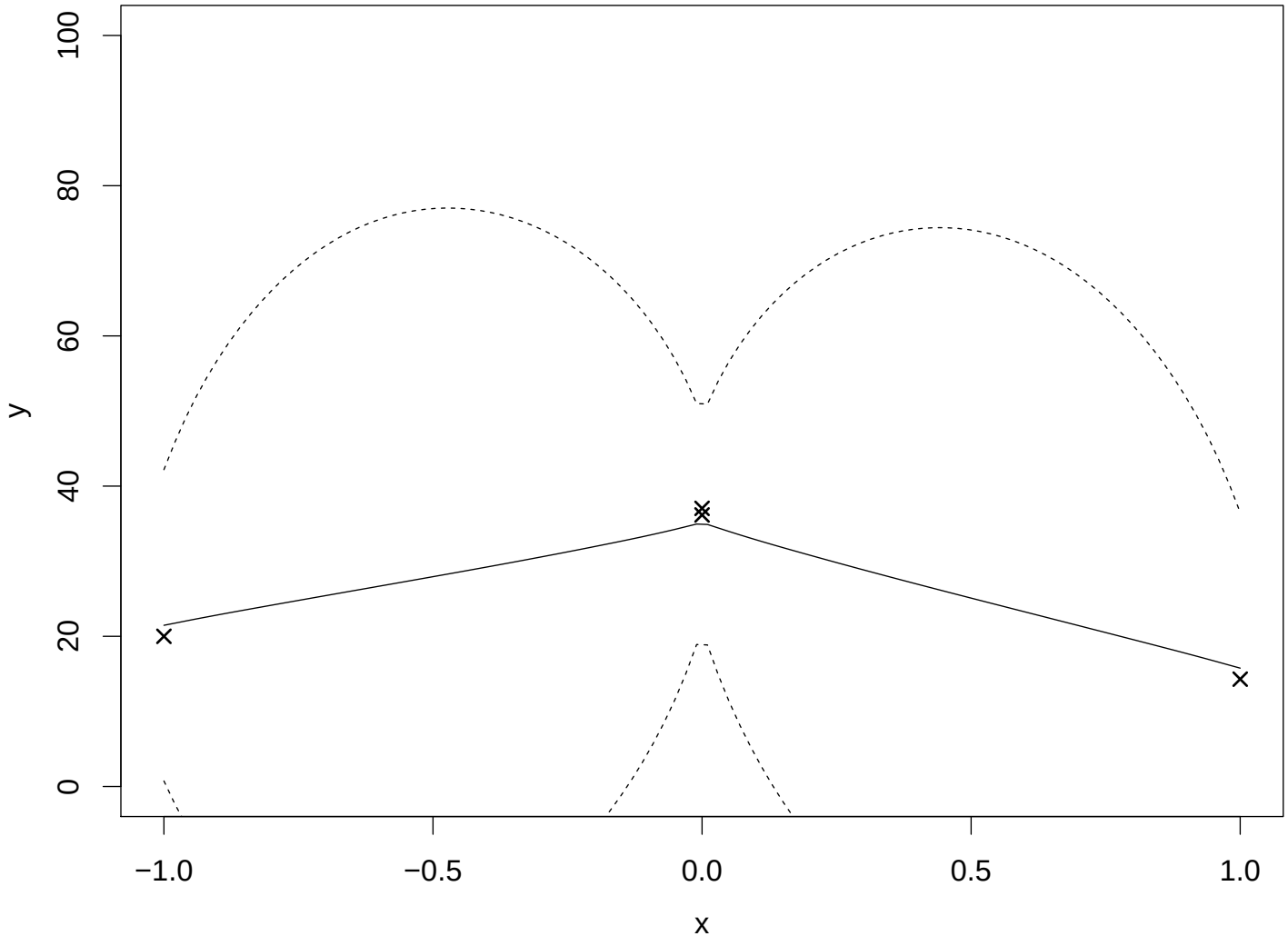
Example 2 - EGU, t-response model



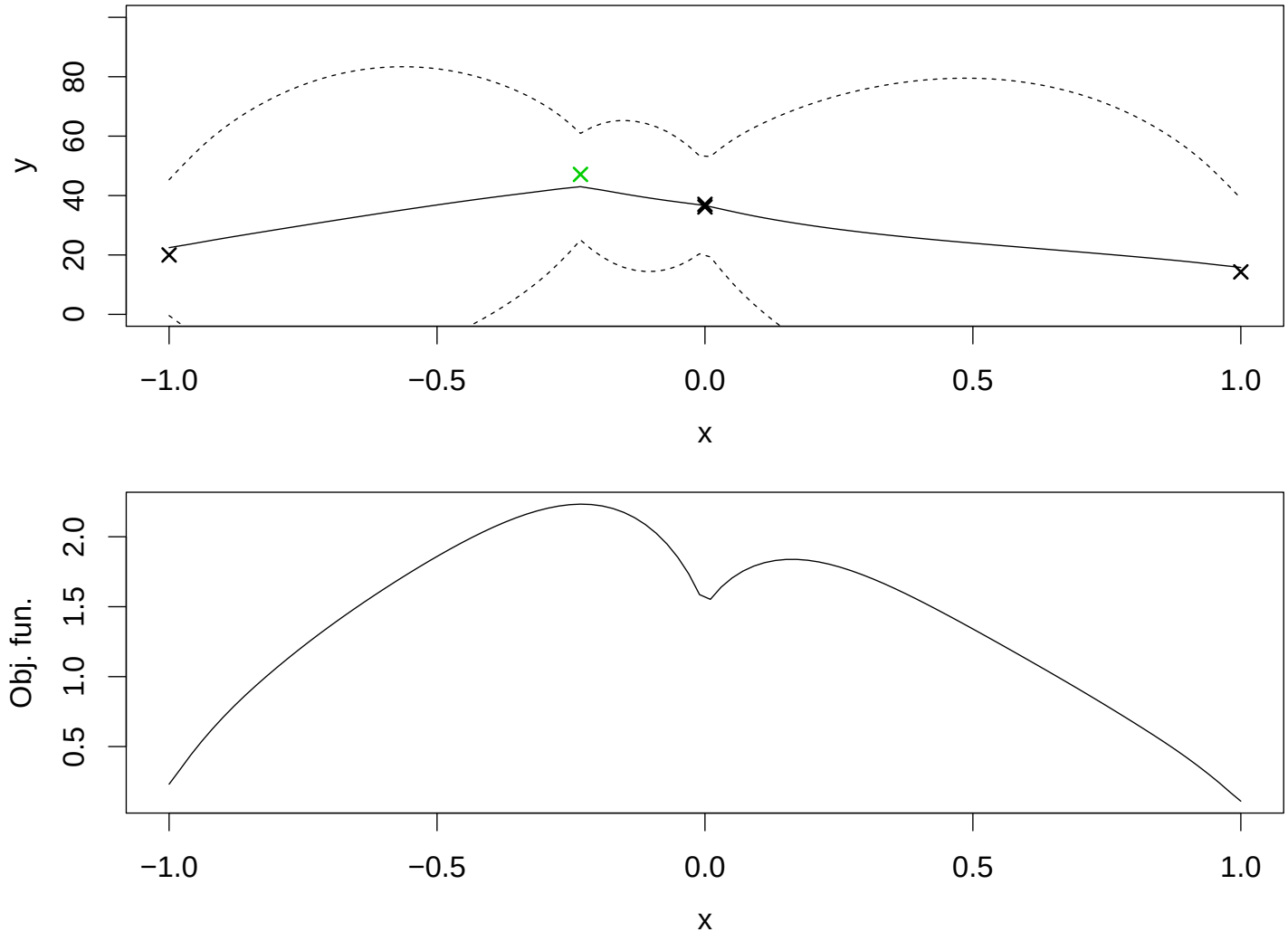
Example 2 - EGU, t-response model



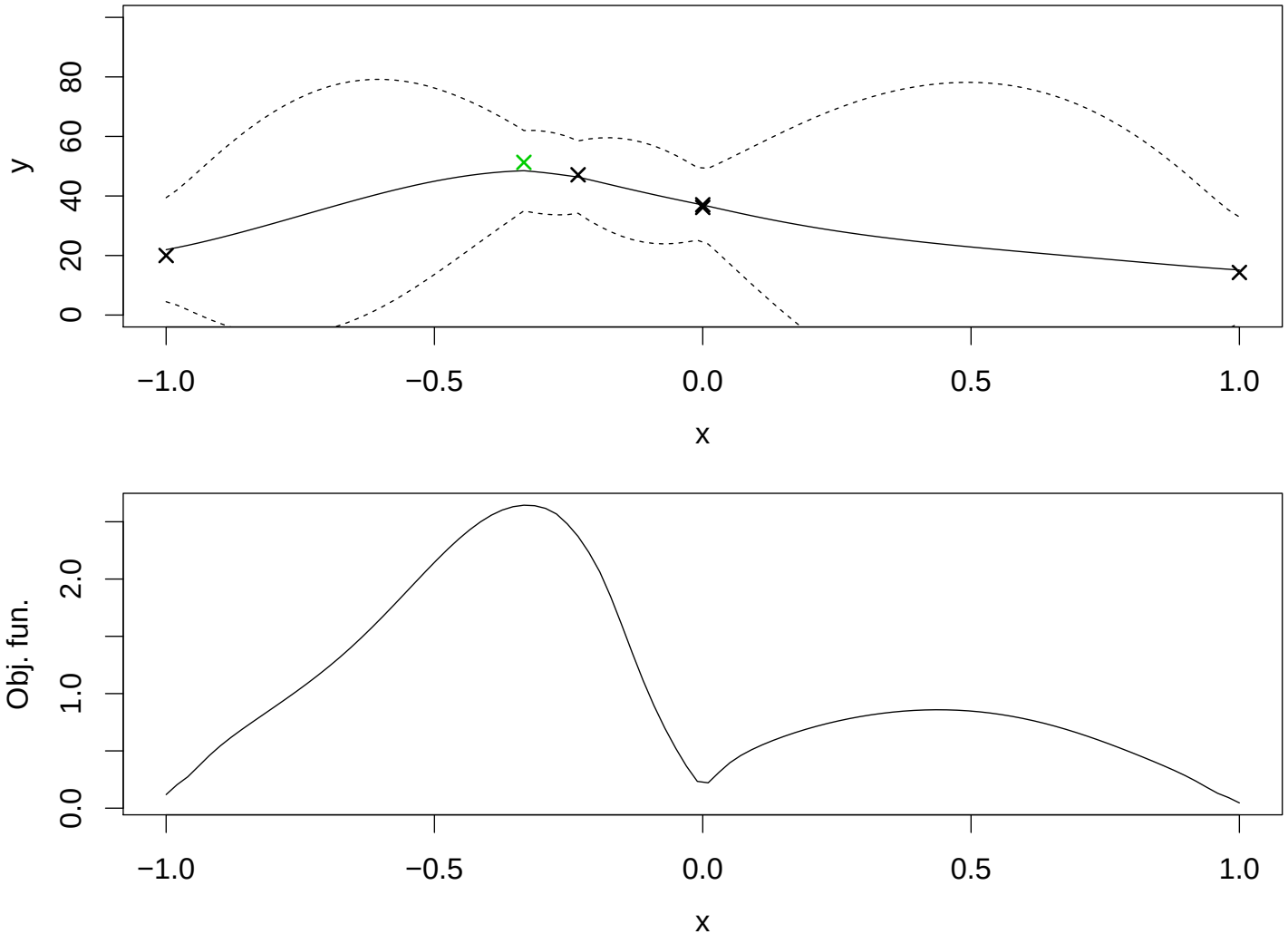
Example 2 - EI, normal-response model



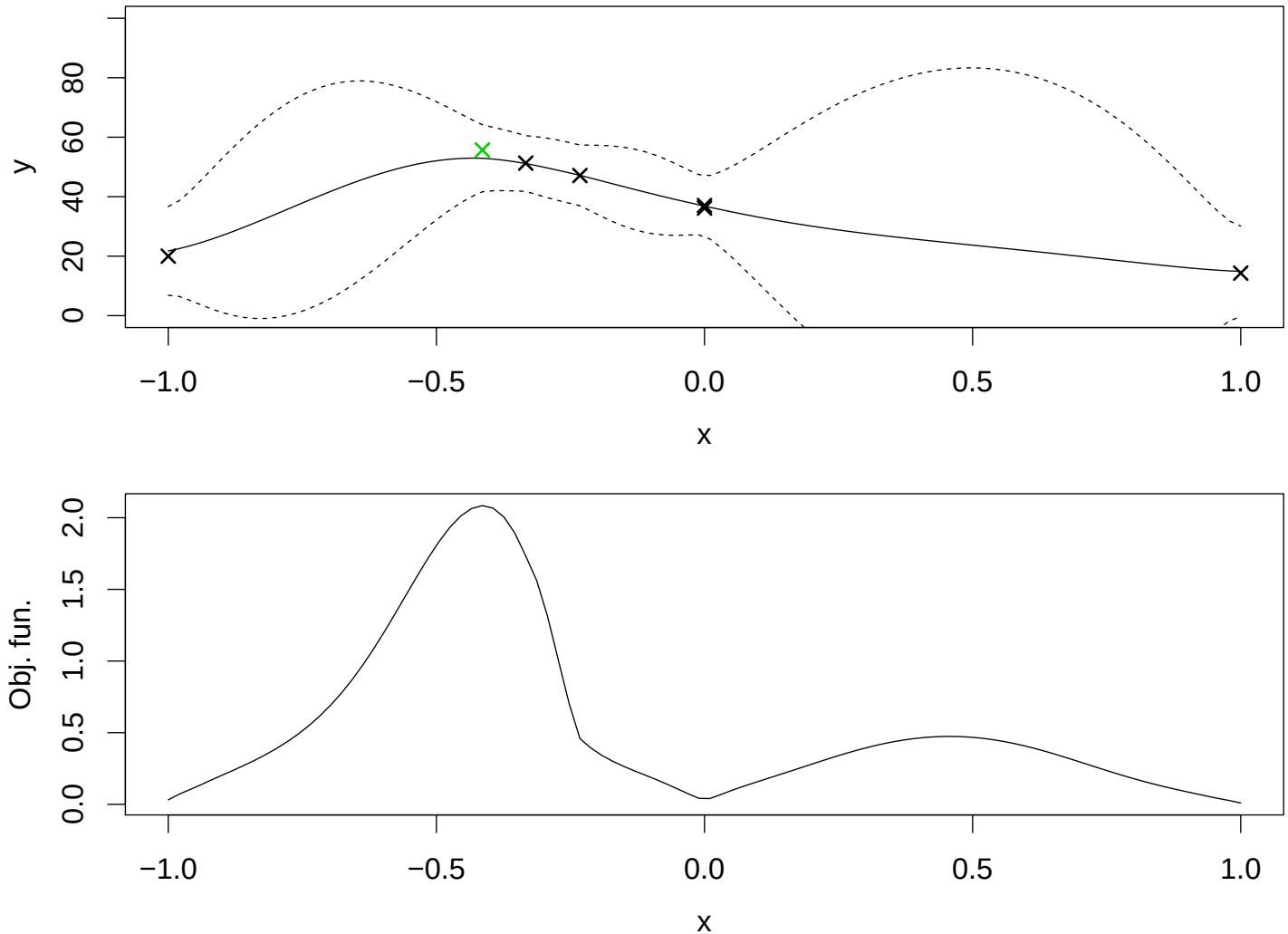
Example 2 - EI, normal-response model



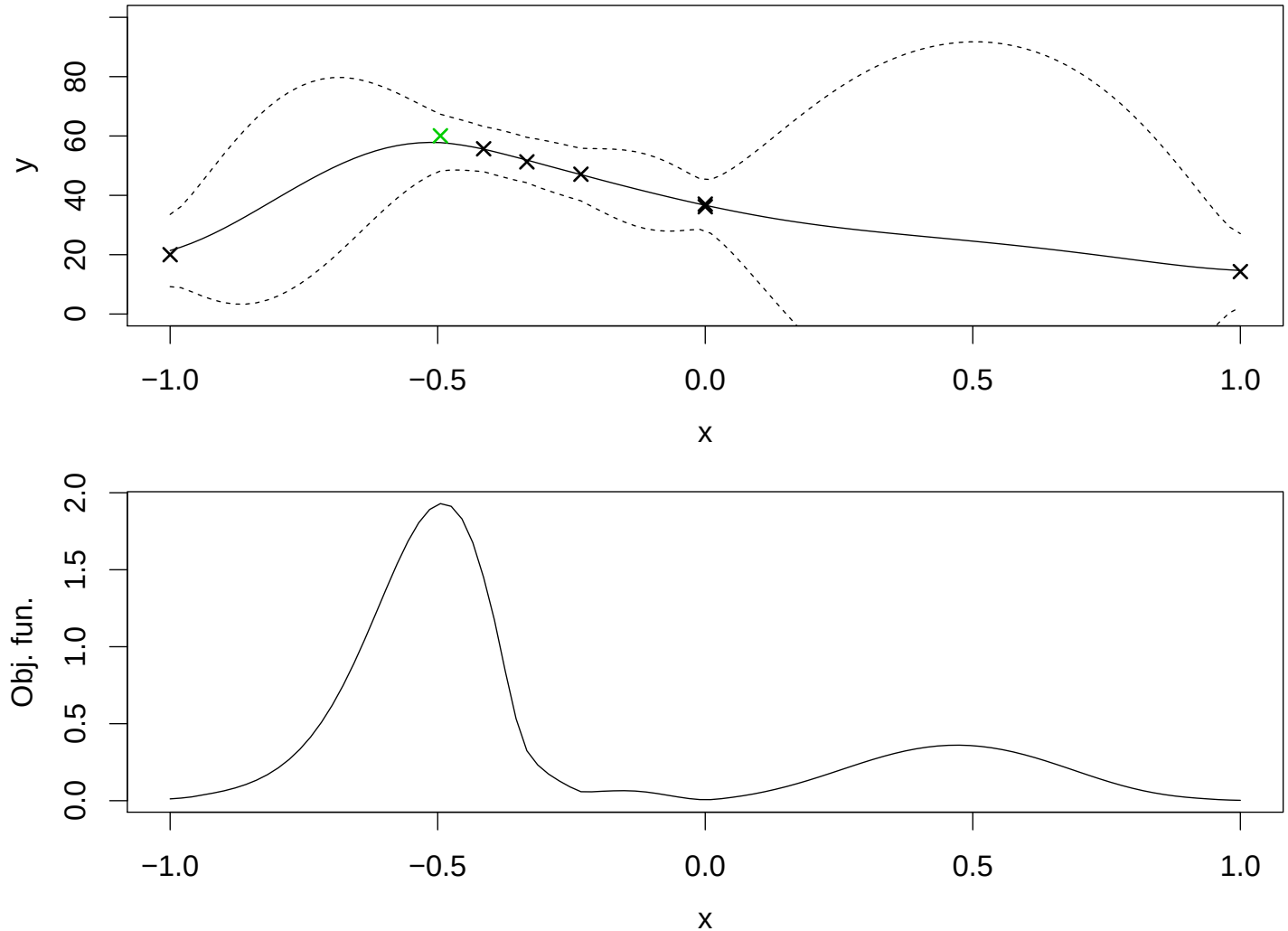
Example 2 - EI, normal-response model



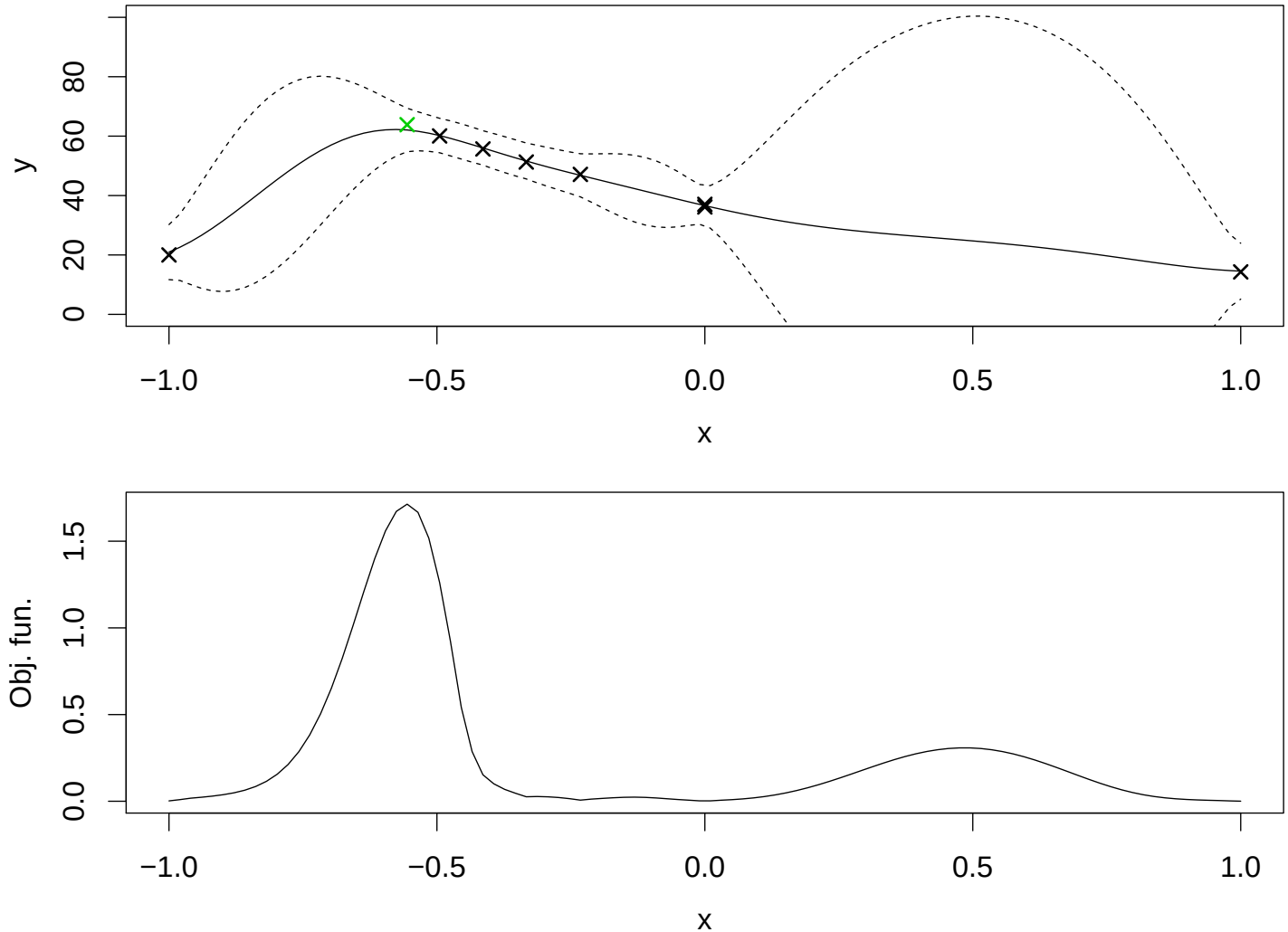
Example 2 - EI, normal-response model



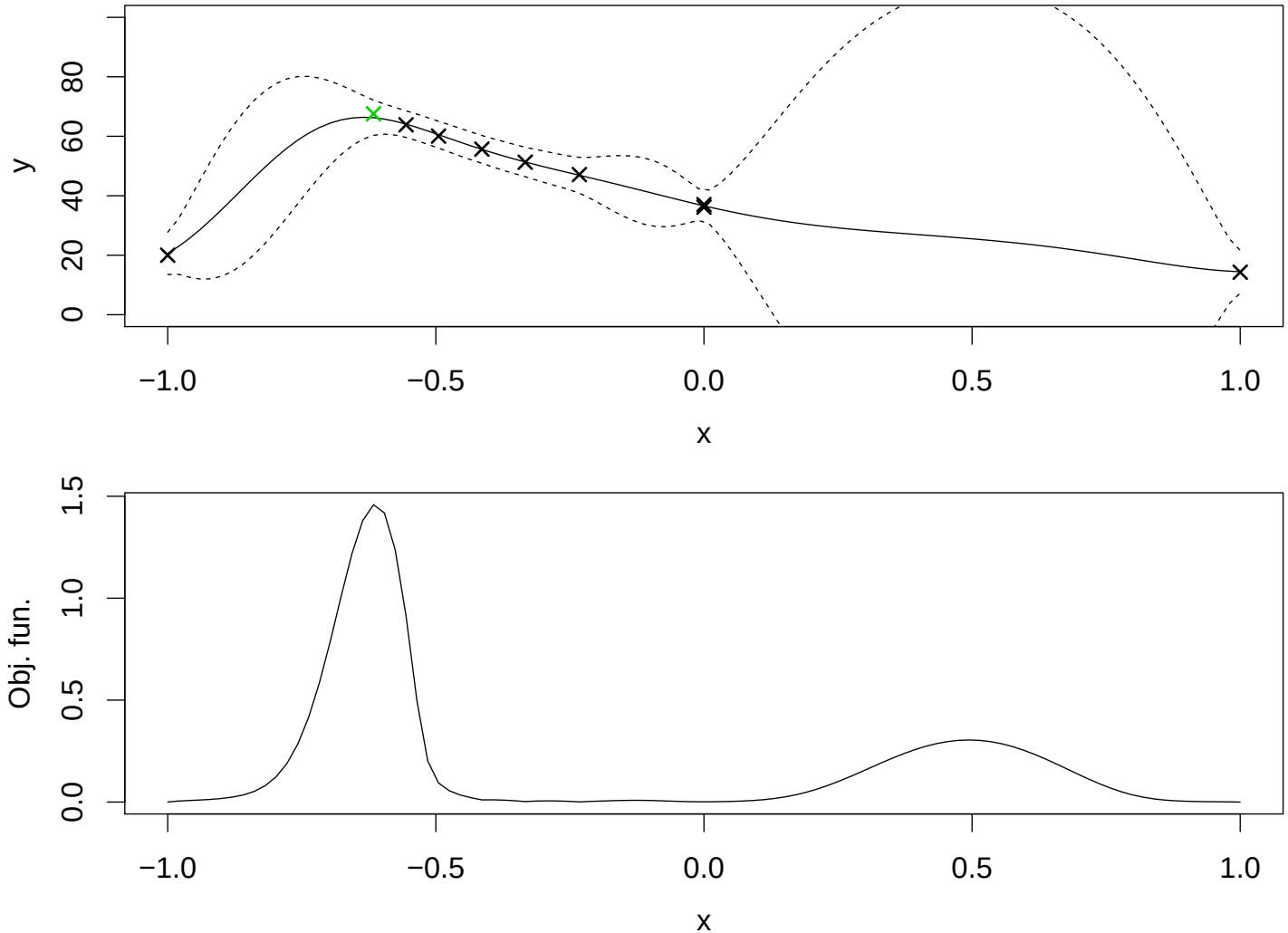
Example 2 - EI, normal-response model



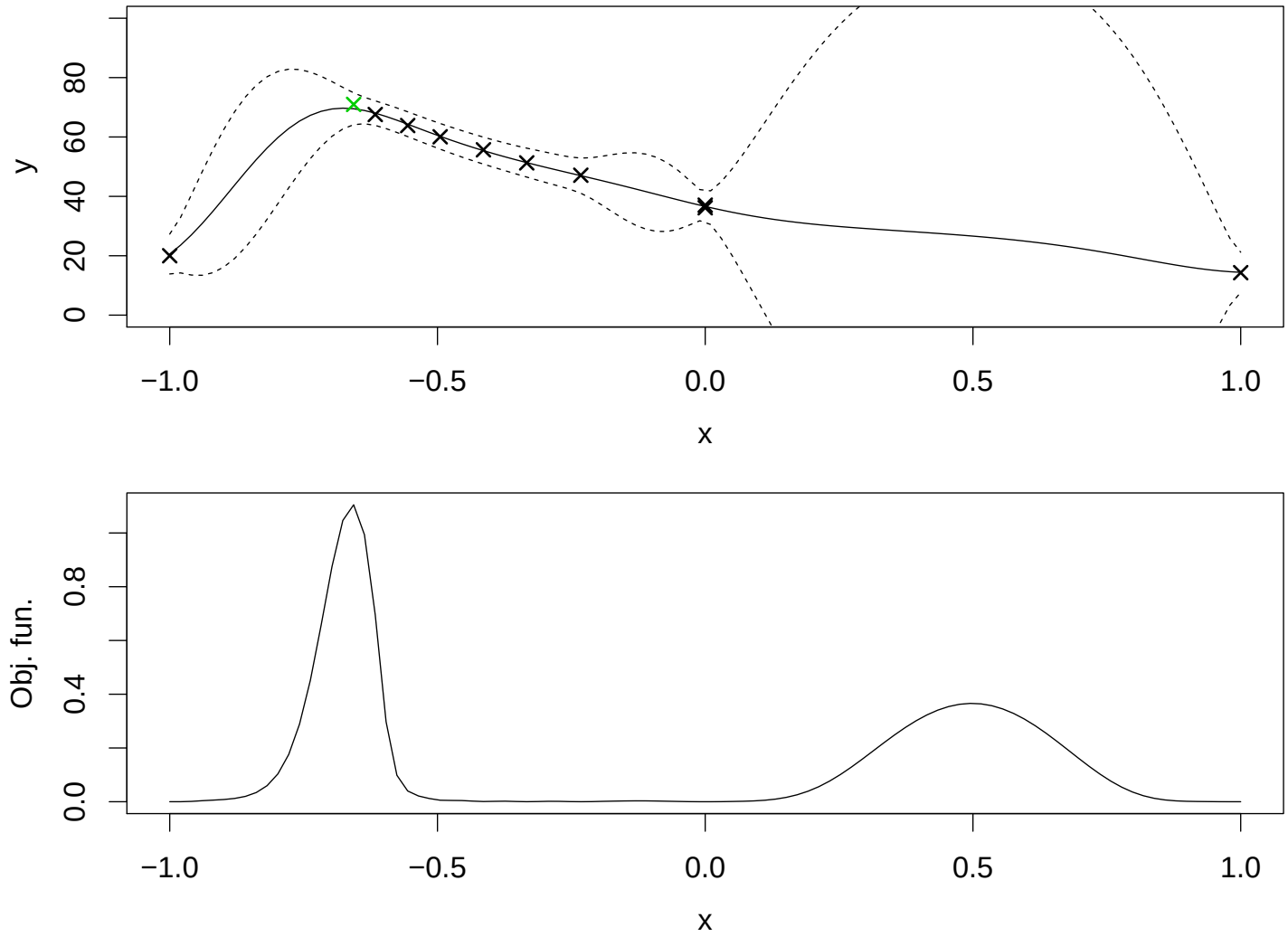
Example 2 - EI, normal-response model



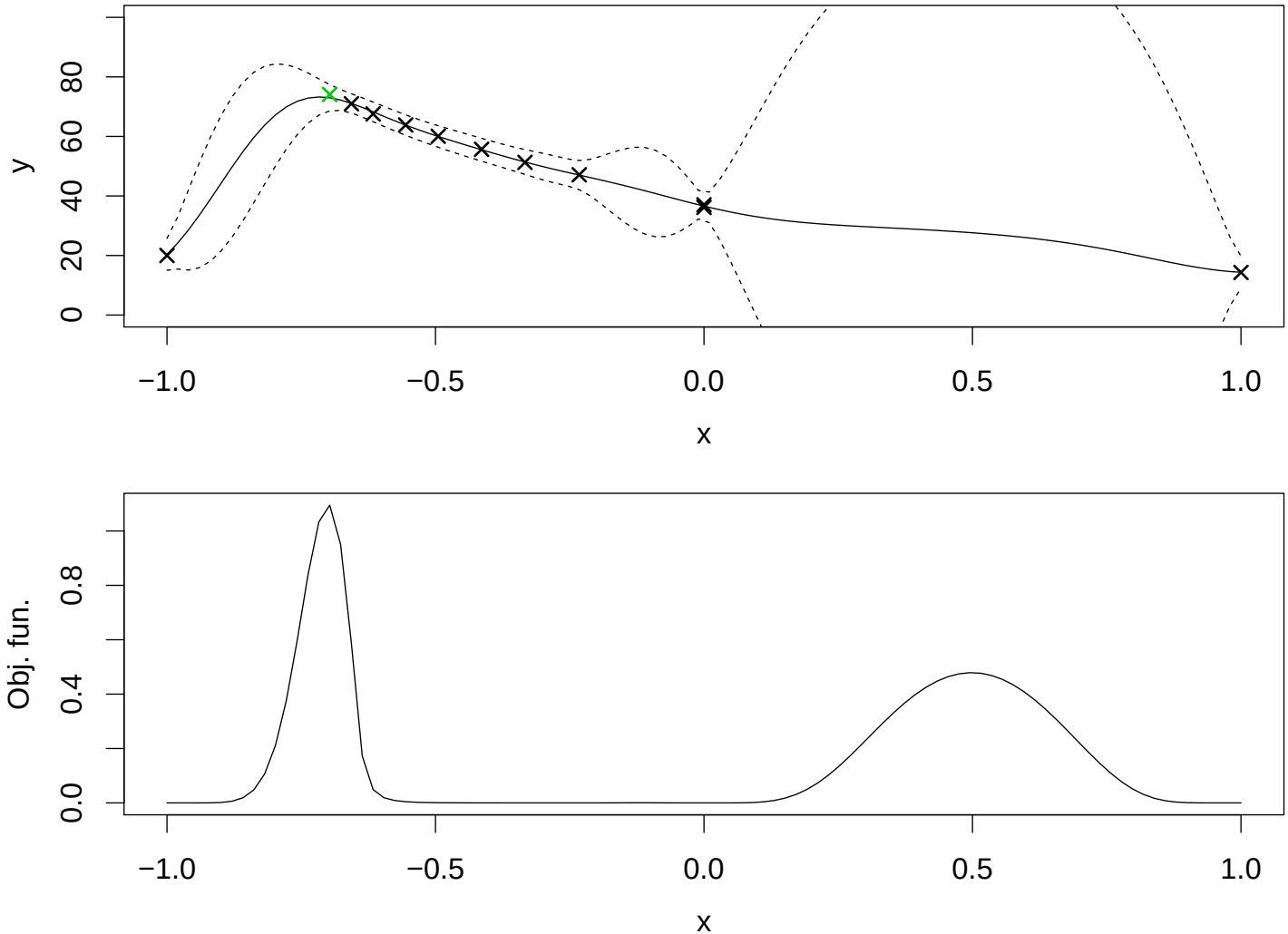
Example 2 - EI, normal-response model



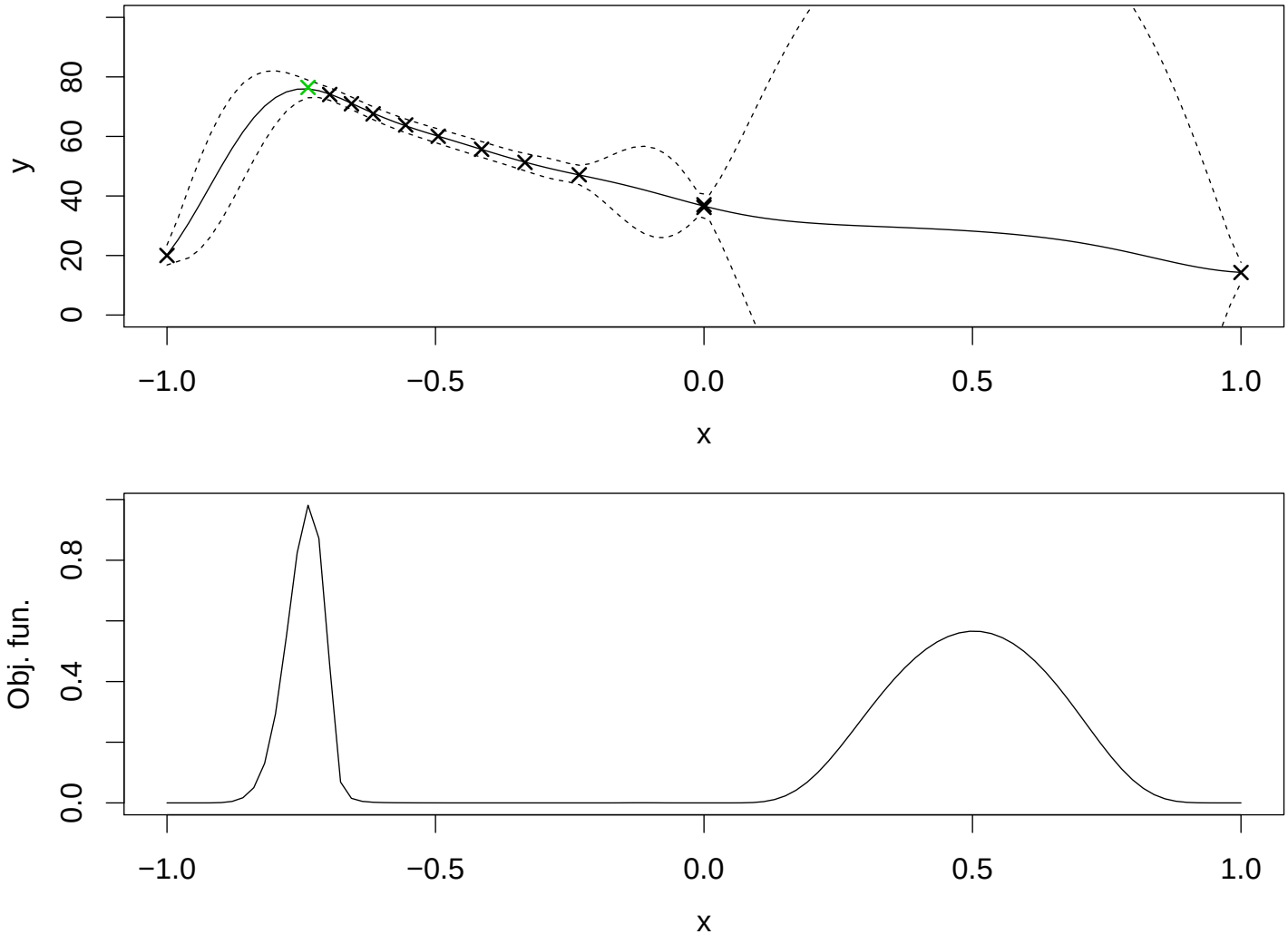
Example 2 - EI, normal-response model



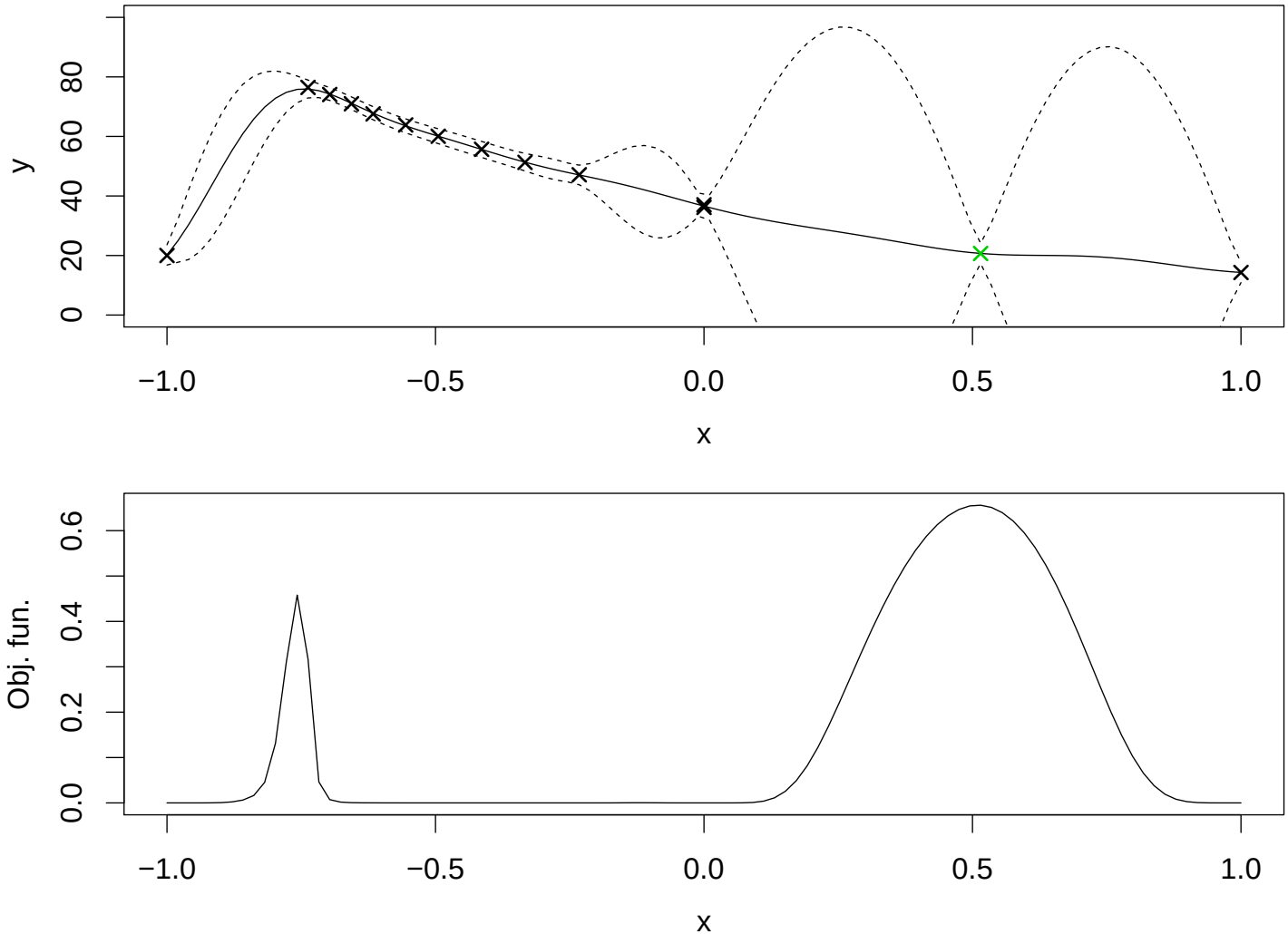
Example 2 - EI, normal-response model



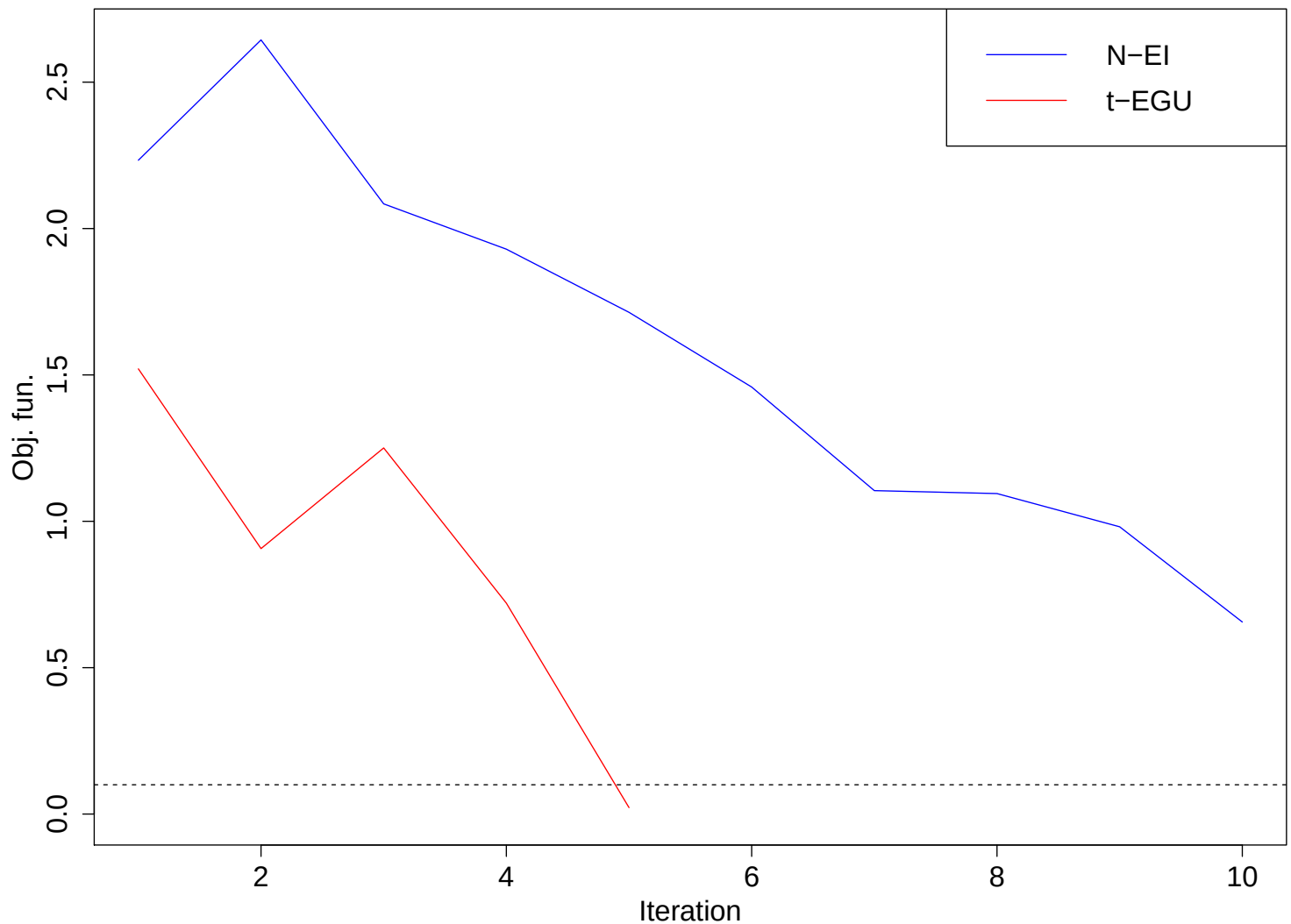
Example 2 - EI, normal-response model



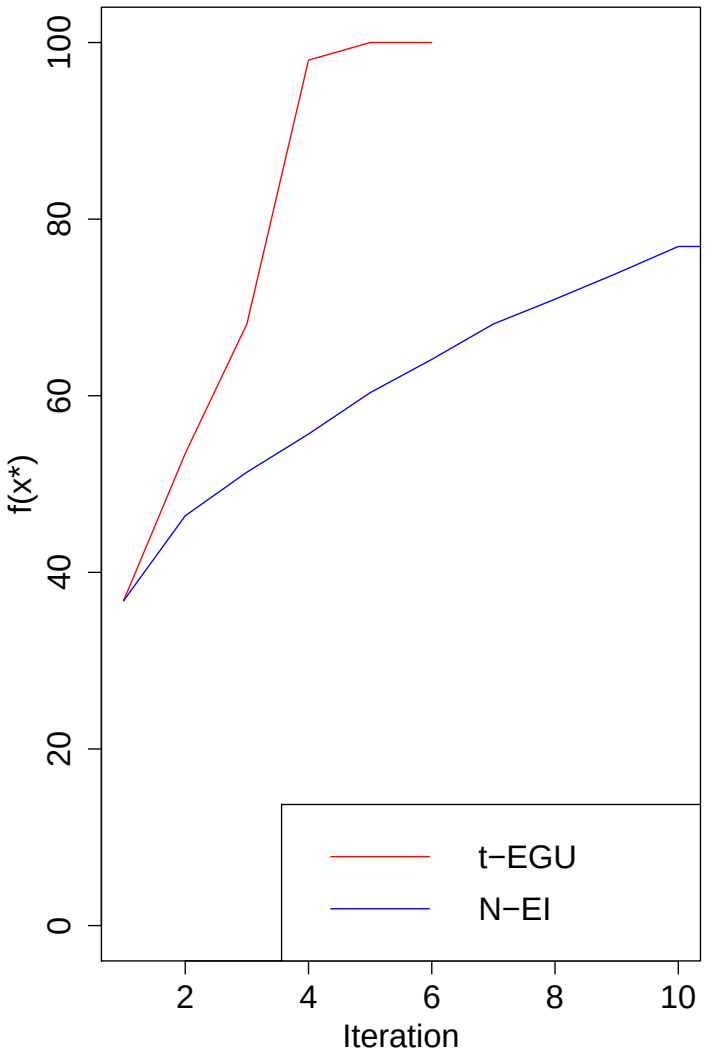
Example 2 - EI, normal-response model

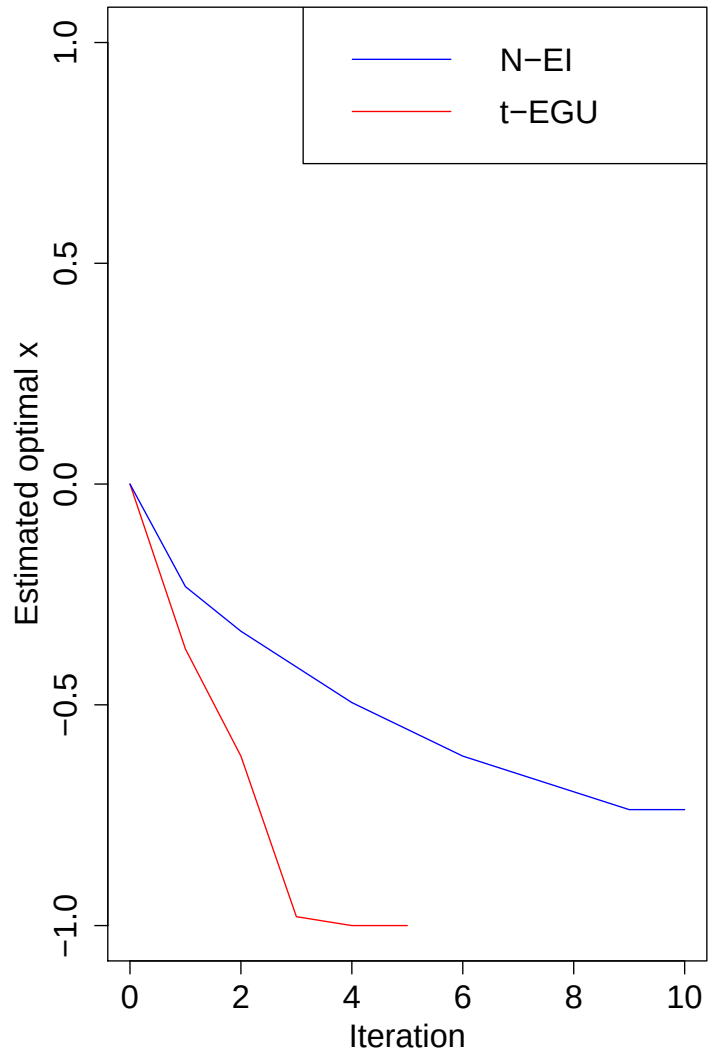


EGU and EI

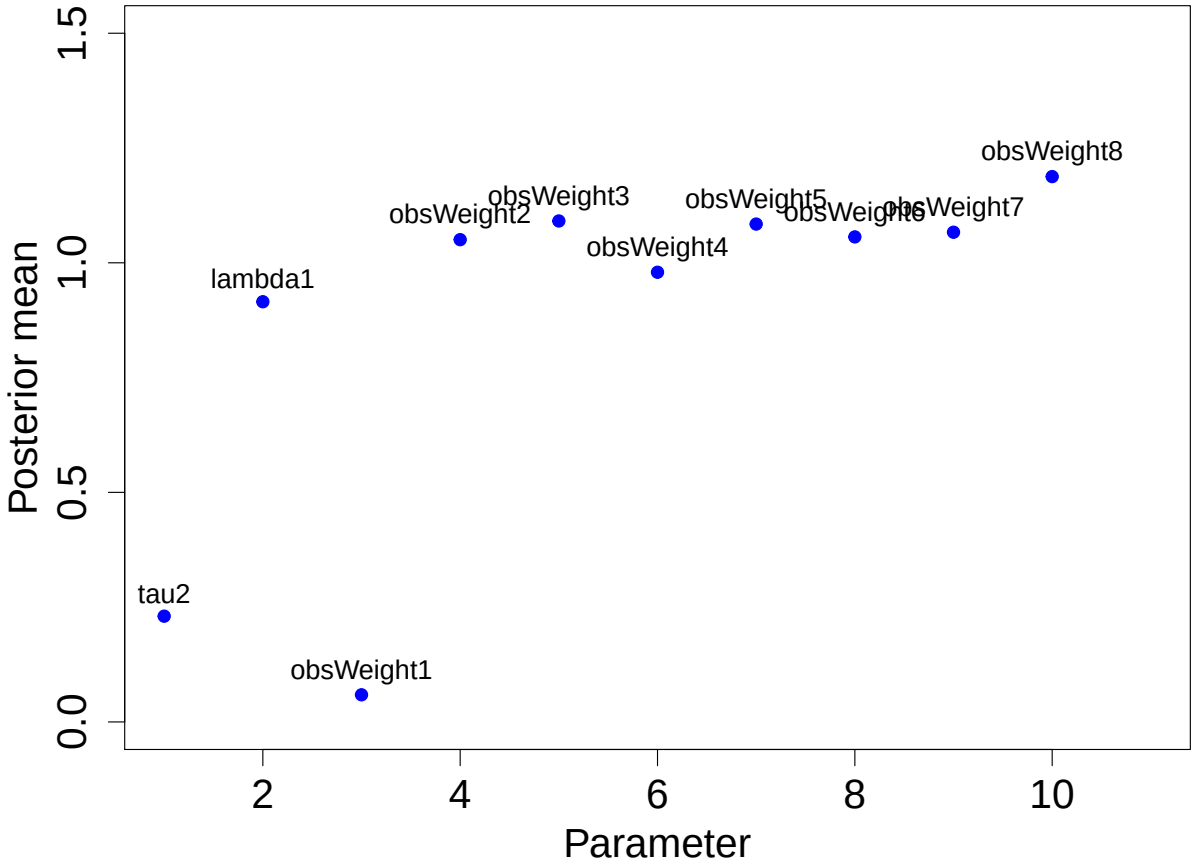


Estimated maximum and its location





Mean posterior estimates from EGU algorithm



λ – correlation function hyperparameter; $\tau^2 = \sigma_\varepsilon^2/\sigma_g^2$

Conclusions and future work

Conclusions

- Decision-theoretic design for (noisy) optimisation, with EI as a special case
- Robust modelling via combination of Gaussian process priors and weighted error variance
- Real examples via a completely automated system

Future work

- Batch sequential and non-myopic design
- More complex chemistry/faster data collection
- `R` package and chemistry software

References

- Brochu, E., Cora, V.M. & de Freitas, N. (2010). arXiv:1012.2599v1.
- Chopin, N. (2002). Biometrika 89, 539–552.
- Drovandi, C.C., McGree, J.M. & Pettitt, A.N. (2013). CSDA, 57, 320–335.
- Gramacy, R.B. & Polson, N.G. (2011). Journal of Computational and Graphical Statistics, 20, 102–118.
- Huang, D., Allen, T.T., Notz, W.I. & Zeng, N. (2006). Journal of Global Optimization 34, 441–466.
- Joseph, V.R., Gul, E. & Ba, S. (2015). Biometrika, in press.
- Jones, D.R., Schonlau, M. & Welch, W.J. (1998). Journal of Global Optimization 13, 455–492.
- Mockus, J. (1989). *Bayesian Approach to Global Optimization: Theory and Applications*. Kluwer.
- Neal, R.M. (1997). arXiv:9701026.
- Osbourne, M.A., Garnett, R. & Roberts, S.J. (2010). In *IEEE International Conference on Advanced Information Networking and Applications*.
- Rasmussen, C.E. & Williams, C.K.I. (2006). *Gaussian Processes for Machine Learning*, MIT Press.