Bayesian Methods for Simulation Optimization

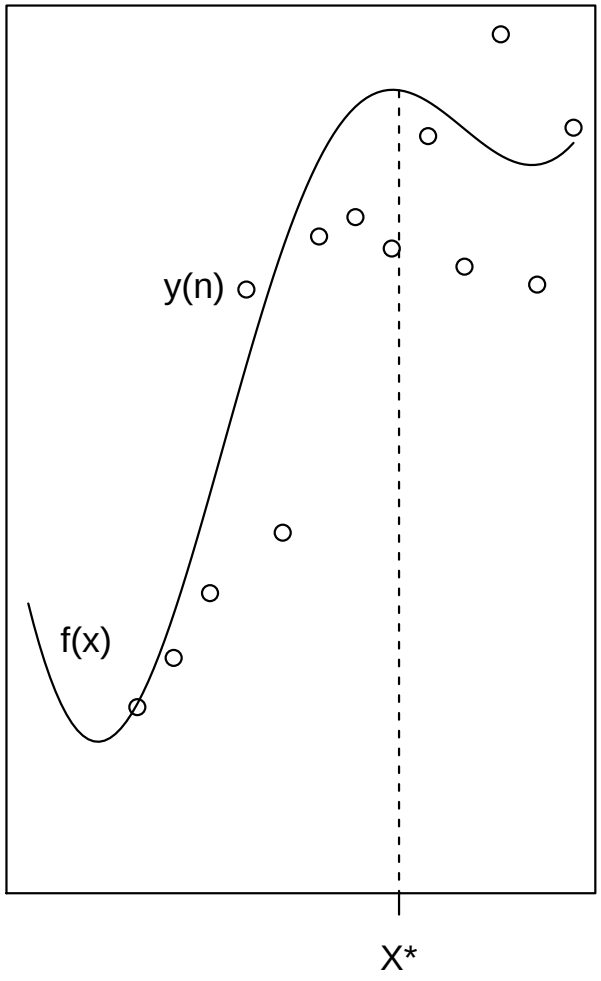
Peter I. Frazier

Operations Research & Information Engineering, Cornell University

Thursday June 12, 2014
Tsinghua University

Research supported by AFOSR and NSF

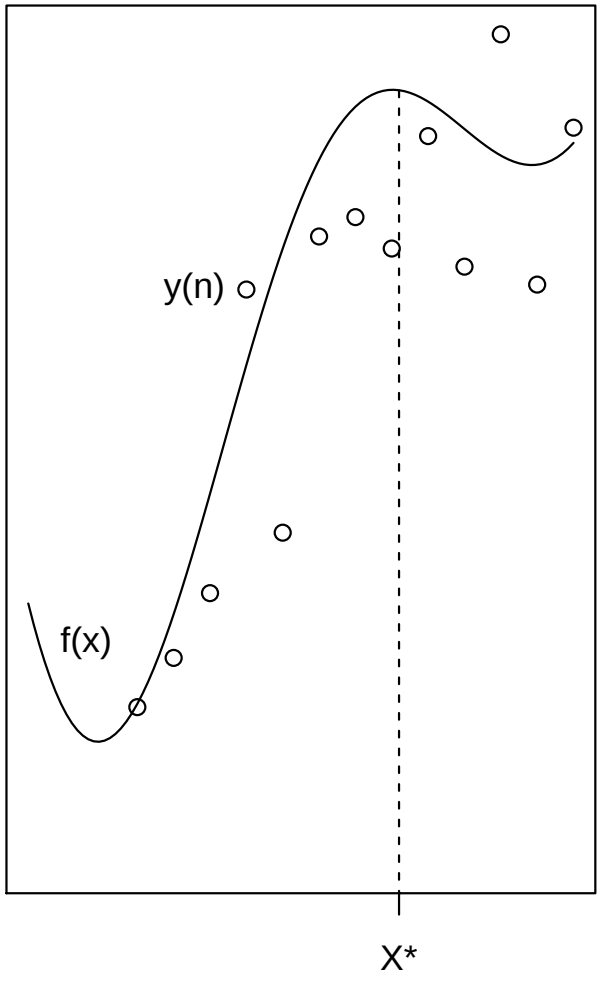
We consider simulation optimization



- Objective function $f : A \mapsto \mathbb{R}$.
- Feasible set A, usually a subset of $\mathbb{R}^d$.
- We cannot evaluate $f(x)$ directly.
- Instead, we have a stochastic simulator that can evaluate $f(x)$ with noise.
- It gives us $y = f(x) + \varepsilon(x)$, where $E[\varepsilon(x)] = 0$.
- Our goal is to find a **global** maximum,

$$\max_{x \in A} f(x)$$

We consider problems with the following characteristics.



- Derivative information is unavailable, and the objective is not necessarily concave.
- We assume a distance measure on A, and that nearby points in A tend to have similar values of $f(x)$.
  - Through most of the talk this comes from the assumption that $A \subseteq \mathbb{R}^d$ and f is continuous on $\mathbb{R}^d$.
- Our simulator takes a long time to run (e.g., minutes or hours), and so we want to minimize the total number of simulation replications required.

There are many problems with these characteristics.

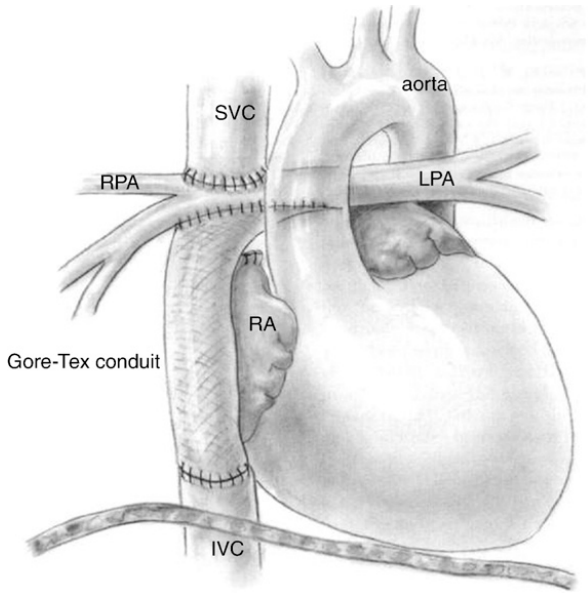


Fig. 1. Extracardiac total cavopulmonary connection. The IVC is disconnected from the right atrium (RA) and connected to the PAs via a Gore-Tex conduit. Figure taken from Reddy et al. [13].

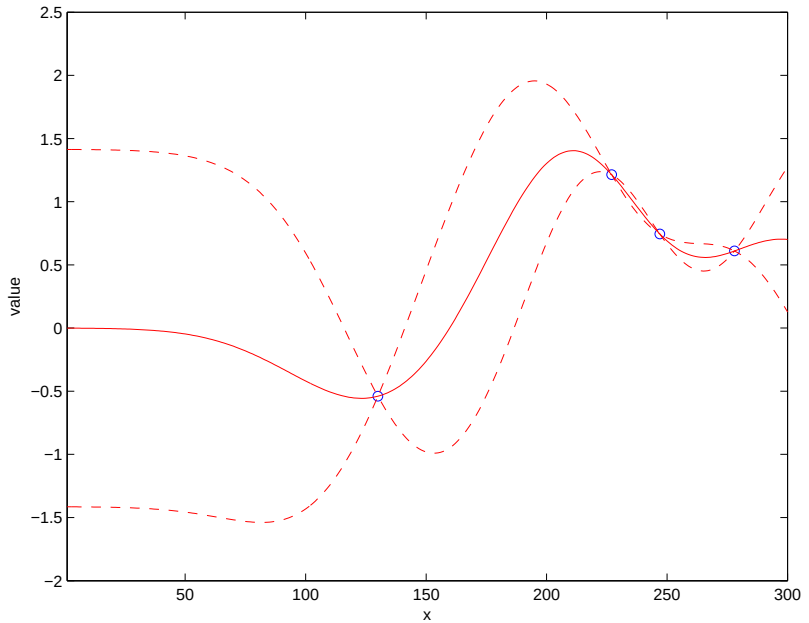
- Tuning algorithm parameters at Yelp. (case study)
- Choose the shape of a graft used in cardiovascular bypass surgery, using a physics-based simulation of blood flow. (case study)
- Calibrate a simulator of a firm's operations to historical data. (case study)
- Choose the chemical composition of a material used in solar cells, to maximize efficiency in converting sunlight to electricity.
- Calibrate a logistics model to historical data (case study).
- Drug development (case study).

What is Bayesian Global Optimization?

- Bayesian Global Optimization (BGO) is a class of algorithms for solving Noise-Free and Noisy Global Optimization problems.
- These algorithms use methods from Bayesian statistics to decide where to sample.

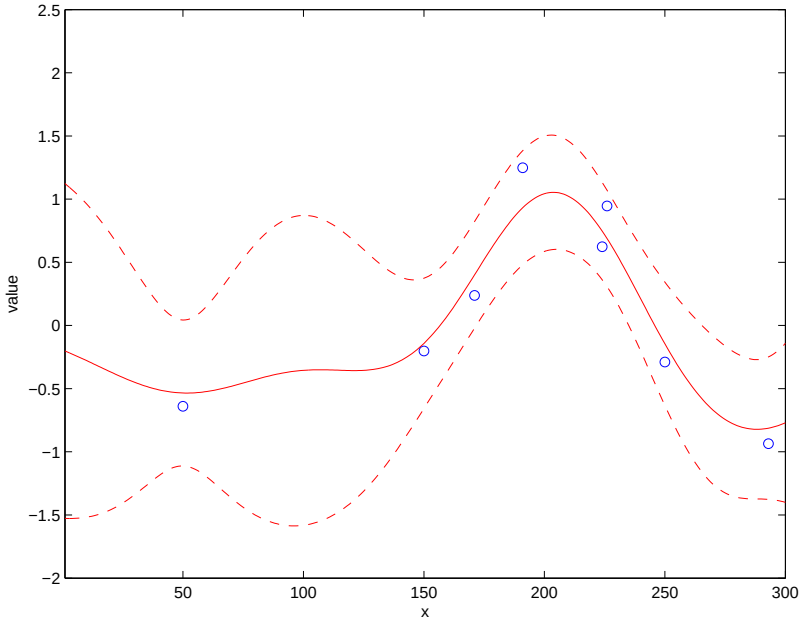
BGO uses Bayesian statistics to decide where to sample.

- Given the function evaluations obtained so for, a BGO algorithm uses Bayesian methods to get:
  - estimates of $f(x)$ over the feasible set.
  - uncertainties in these estimates.
  - together, these are described by the **posterior distribution**.



- BGO uses the posterior distribution to decide where to evaluate next.

BGO uses the posterior in a one-step optimality analysis to decide where to sample.



- We have the posterior distribution pictured at left.
- We decide where to sample next by answering this question:
  - "If I could take just one more sample before choosing a final solution, where would it be **optimal** to take it?"

(Some BGO methods do a multi-step optimality analysis. This will be discussed briefly at the end of the talk.)

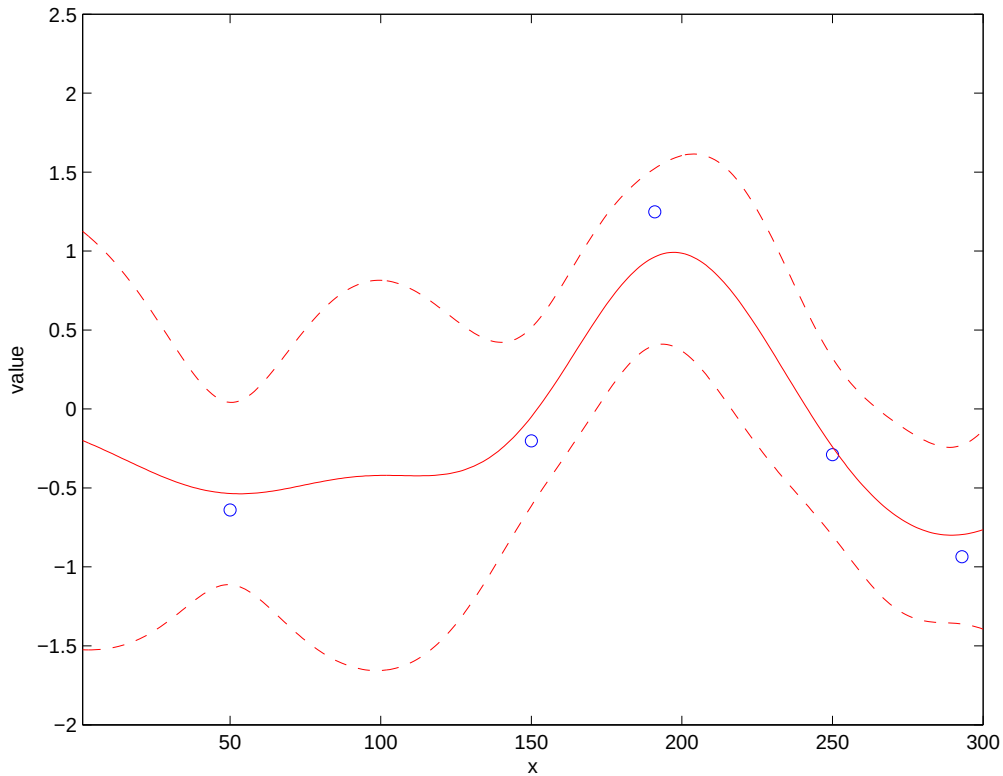
Many other non-Bayesian methods exist for derivative-free global optimization

- Many other derivative-free noise-tolerant global optimization methods exist, e.g.,
  - pattern search, e.g., Nelder-Mead
  - stochastic approximation, e.g., SPSA [Spall 1992].
  - evolutionary algorithms, simulated annealing, tabu search
  - response surface methods. [Myers & Montgomery 2002]
  - Lipschitzian optimization, e.g., DIRECT [Gablonsky et al. 2001]
- Bayesian one-step-optimal methods allow inclusion of application-specific domain knowledge through the prior.
- Bayesian one-step-optimal methods can be adapted in a principled way to novel problem structure (e.g., parallel function evaluations, multiple types of function evaluations, non-stationary function evaluations)

Outline

Gaussian Process Regression is a standard method from Bayesian statistics and machine learning

- The BGO methods in this talk use a standard tool from Bayesian statistics, Gaussian process regression, to calculate the posterior.
- Gaussian process regression estimates, from data, the values taken by an unknown function.
- I will give a **brief** overview, hiding many details.



GP Regression: Formal Definition

A prior distribution on a function $f : \mathbb{R}^d \mapsto \mathbb{R}$ is a **Gaussian Process (GP) prior** with mean function $\mu_0 : \mathbb{R}^d \mapsto \mathbb{R}$ and covariance function $\Sigma_0 : \mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{R}$ if, under the prior for any given set of points $x_1, \ldots, x_k$,

$$\begin{bmatrix} f(x_1) \\ \vdots \\ f(x_k) \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_0(x_1) \\ \vdots \\ \mu_0(x_k) \end{bmatrix}, \begin{bmatrix} \Sigma_0(x_1, x_1) & \ldots & \Sigma_0(x_1, x_k) \\ \vdots & \ddots & \vdots \\ \Sigma_0(x_k, x_1) & \ldots & \Sigma_0(x_k, x_k) \end{bmatrix} \right). \quad (1)$$

- The mean function $\mu_0(\cdot)$ is often a constant, or a linear combination of basis functions estimated from data.
- The covariance function $\Sigma_0(\cdot, \cdot)$ encodes the belief that nearby points have similar function values. A common choice is $\Sigma_0(x, x') = \alpha_0 \exp\left(-\alpha_1 ||x - x'||^2\right)$, with α_0 and α_1 estimated from data.

The posterior can be computed analytically

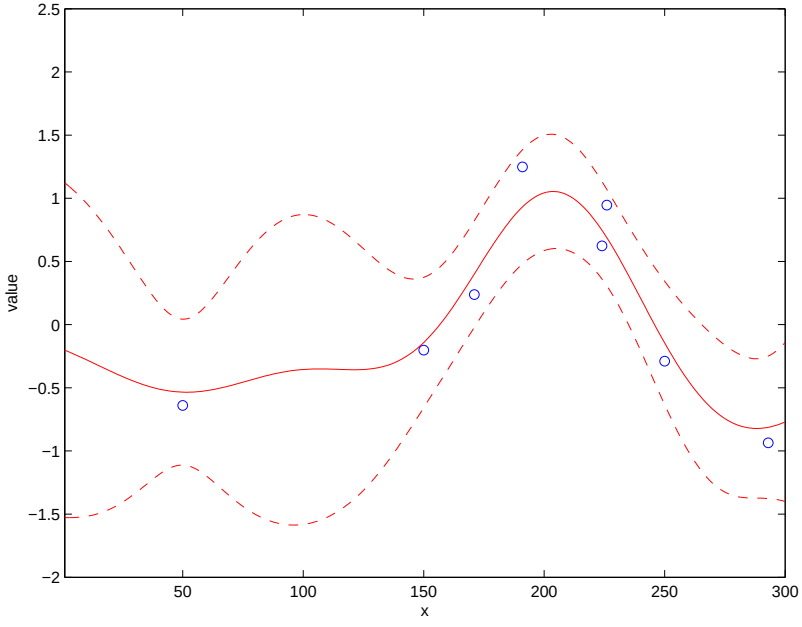
If we have observed conditionally independent $y_1, \ldots, y_n$, where

$$y_i \sim \mathcal{N}(f(x_i), \lambda_i),$$

Then the posterior distribution of $f(x)$ is normal,

$$f(x)|x_{1:n}, y_{1:n}, \lambda_{1:n} \sim \mathcal{N}(\mu_n(x), \sigma_n^2(x))$$

where $\mu_n(x)$ and $\sigma_n^2(x)$ can be computed analytically from $(y_i, \mu_0(x_i), \Sigma_0(x_i, x_j), \lambda_i : i, j = 1, \ldots, n)$.



- Blue circles are (x_n, y_n).
- Solid red line is $\mu_n(x)$.
- Distance between the dashed lines is $2\sigma_n(x)$

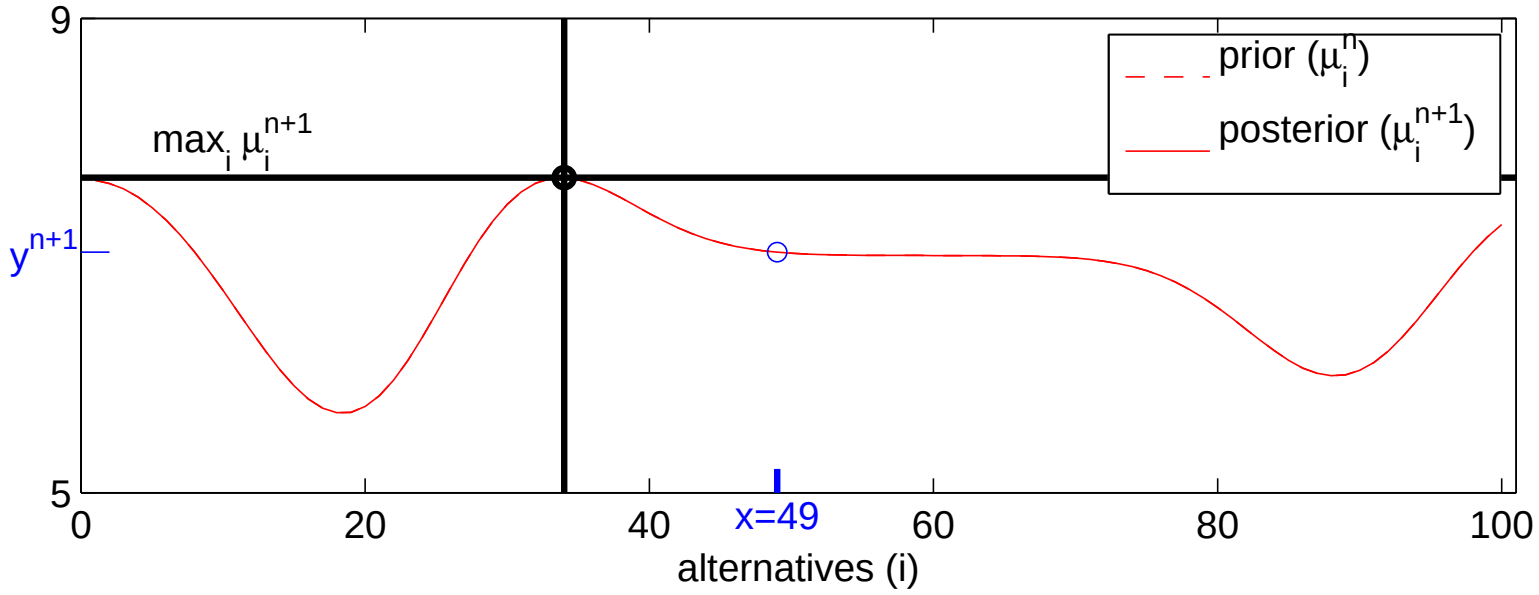
Outline

The Knowledge-Gradient (KG) Method for Correlated Beliefs is designed for the following discrete simulation optimization problem.

- The feasible space A is a finite subset of $\mathbb{R}^d$. Typically, in discrete simulation optimization problems, A is an integer lattice.
- We sample sequentially, choosing each point x_{n+1} to sample next based on all previous samples. We are given a finite sampling budget N.
- Sampling noise is independent.
- Sampling noise is normally distributed, $y_n \sim \mathcal{N}(f(x_n), \lambda(x_n))$. In practice, sampling noise is not normal, but can be made approximately normal through batching.
- The sampling variance $\lambda(x)$ is assumed known. In practice, $\lambda(\cdot)$ is unknown, and is estimated from data.

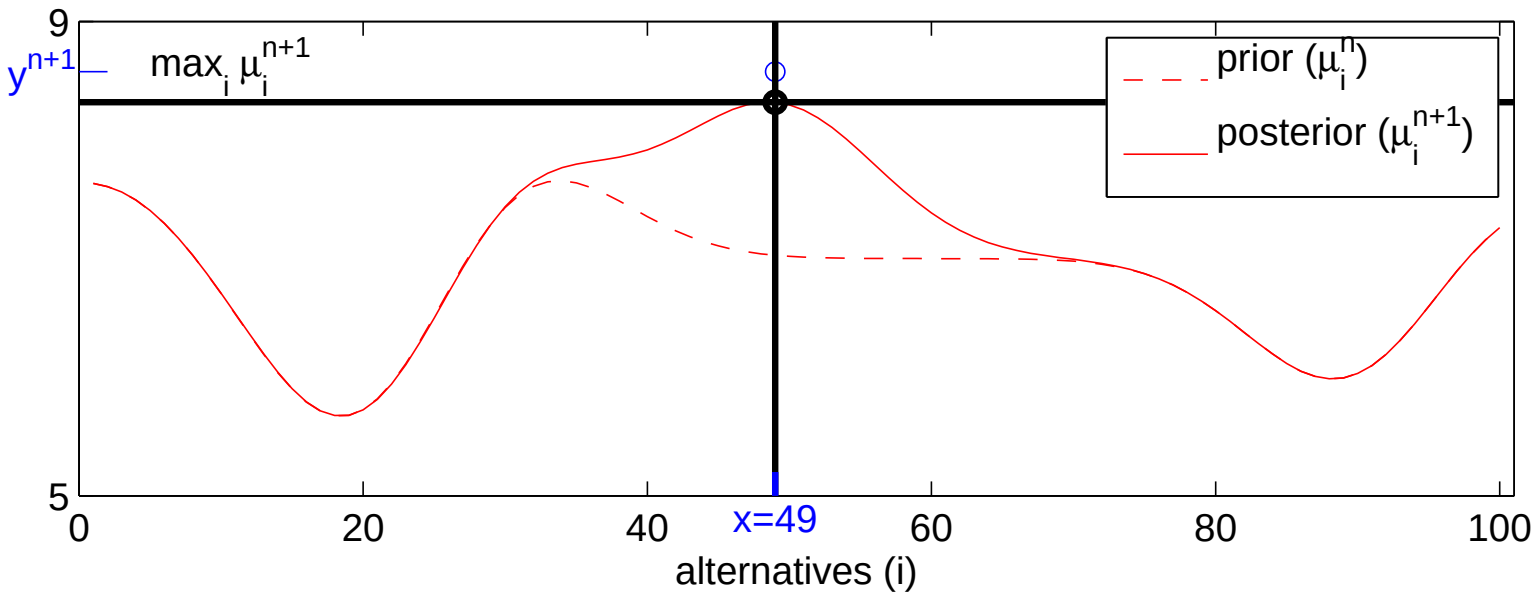
KG Method: μ_n^* is the expected value of the solution we would choose if we stopped after n samples.

- At time n, we've measured $x_1, \ldots, x_n$, and observed $y_1, \ldots, y_n$.
- $\mu_n(x)$ is the expected value of $f(x)$ given these n samples.
- $\mu_n^* = \max_x \mu_n(x)$ is the expected value of the solution we would choose if we stopped after n samples.



KG Method: μ^*_{n+1} is the expected value of the solution we would choose if we stopped after $n+1$ samples.

- Recall from the previous slides:
  - $\mu_n(x)$ is the expected value of $f(x)$ at time n (given $x_{1:n}, y_{1:n}$).
  - $\mu^*_n = \max_x \mu_n(x)$ is the expected value of the solution we would choose if we stopped at time n.
- Similarly, $\mu_{n+1}(x)$ is the expected value of $f(x)$ at time $n+1$ (given $x_{1:n+1}, y_{1:n+1}$).
- Also, $\mu^*_{n+1} = \max_x \mu_{n+1}(x)$ is the expected value of the solution we would choose if we stopped at time $n+1$.



The knowledge-gradient factor quantifies a sample's value.

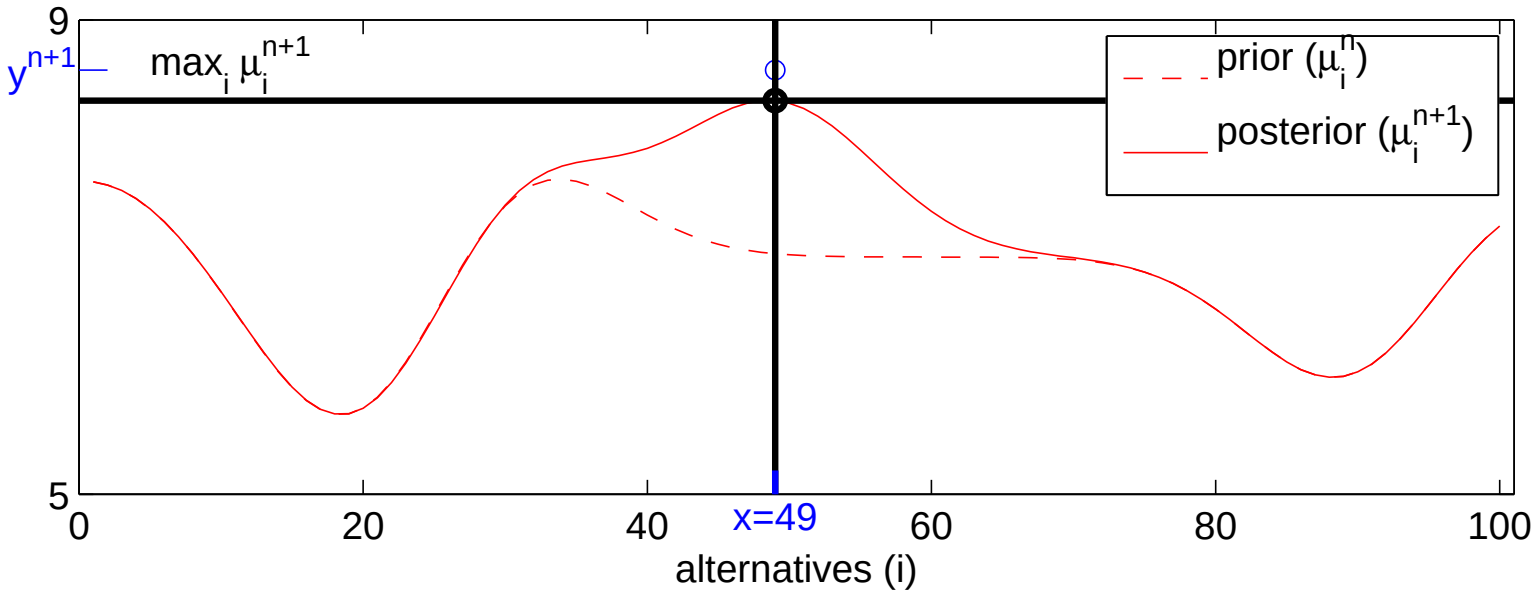
- The additional sample x_{n+1}, y_{n+1} has increased our solution's value by

$$\mu^*_{n+1} - \mu^*_n.$$

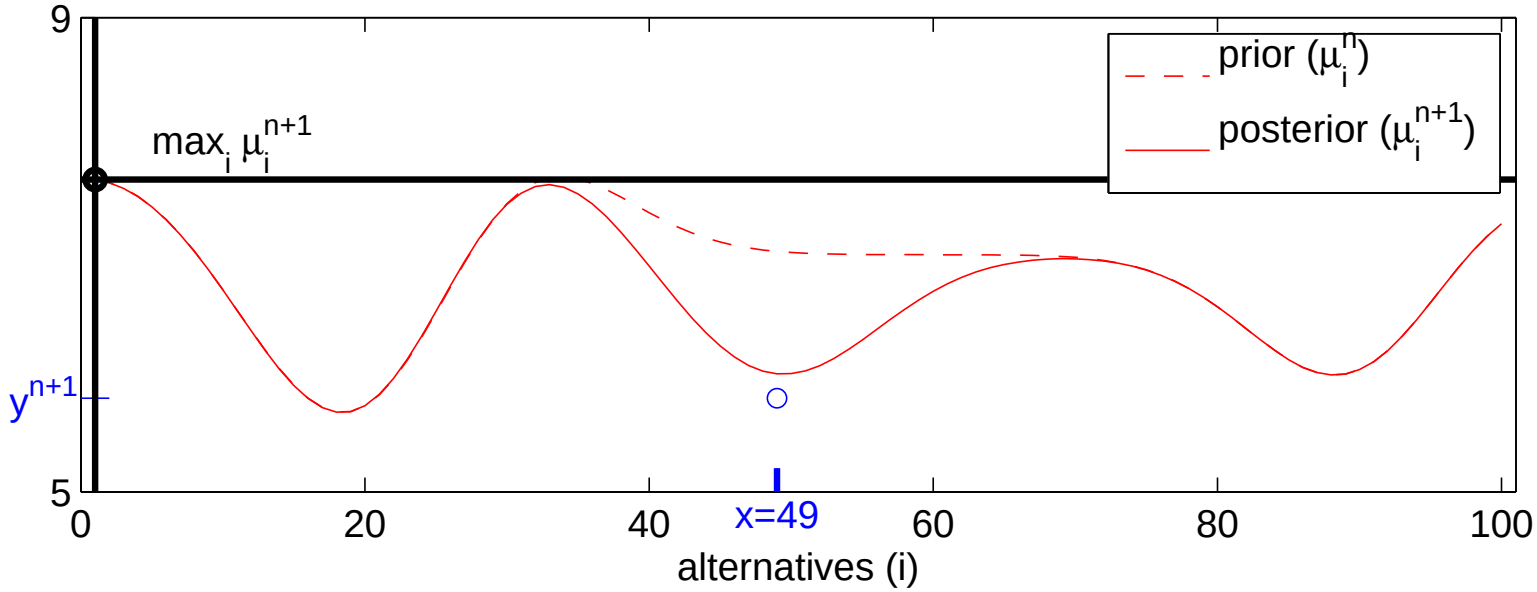
- At time n, we don't know y_{n+1}, so we can't compute this quantity.
- We can, however, compute its expected value,

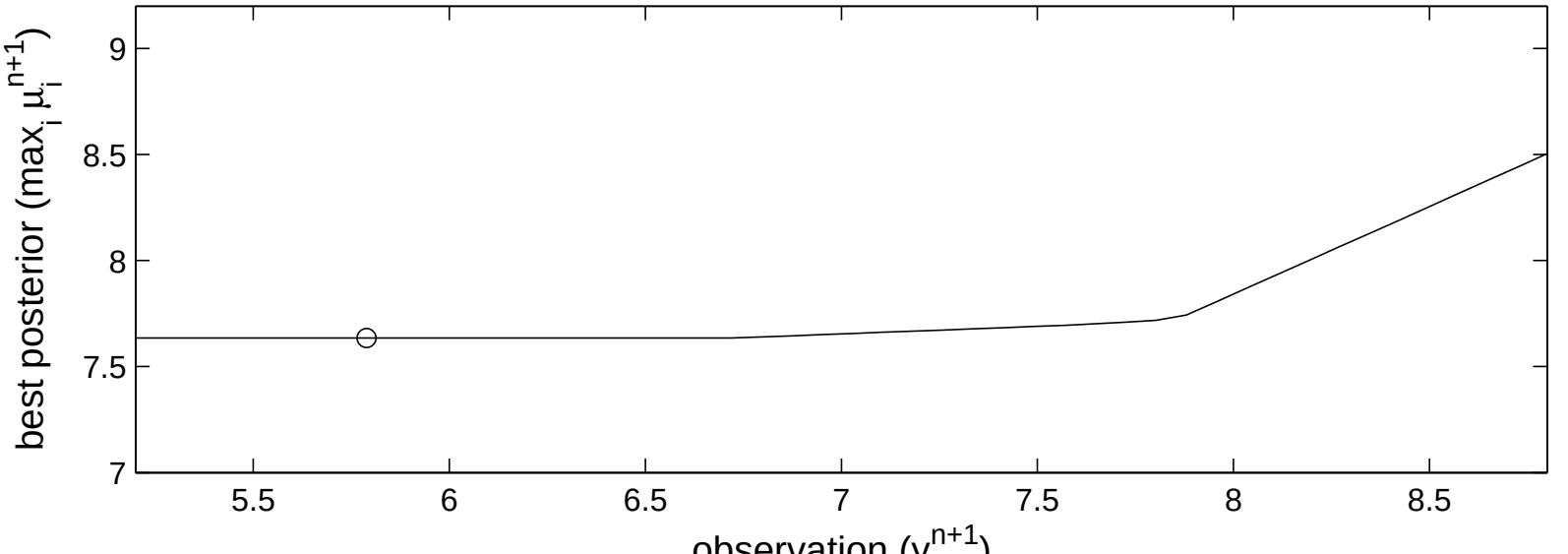
$$\mathrm{KG}_n(x) = \mathbb{E}_n \left[\mu^*_{n+1} - \mu^*_n \mid x_{n+1} = x \right].$$

- We call this quantity the *knowledge-gradient (KG) factor*, because it measures the change in the value of our knowledge.

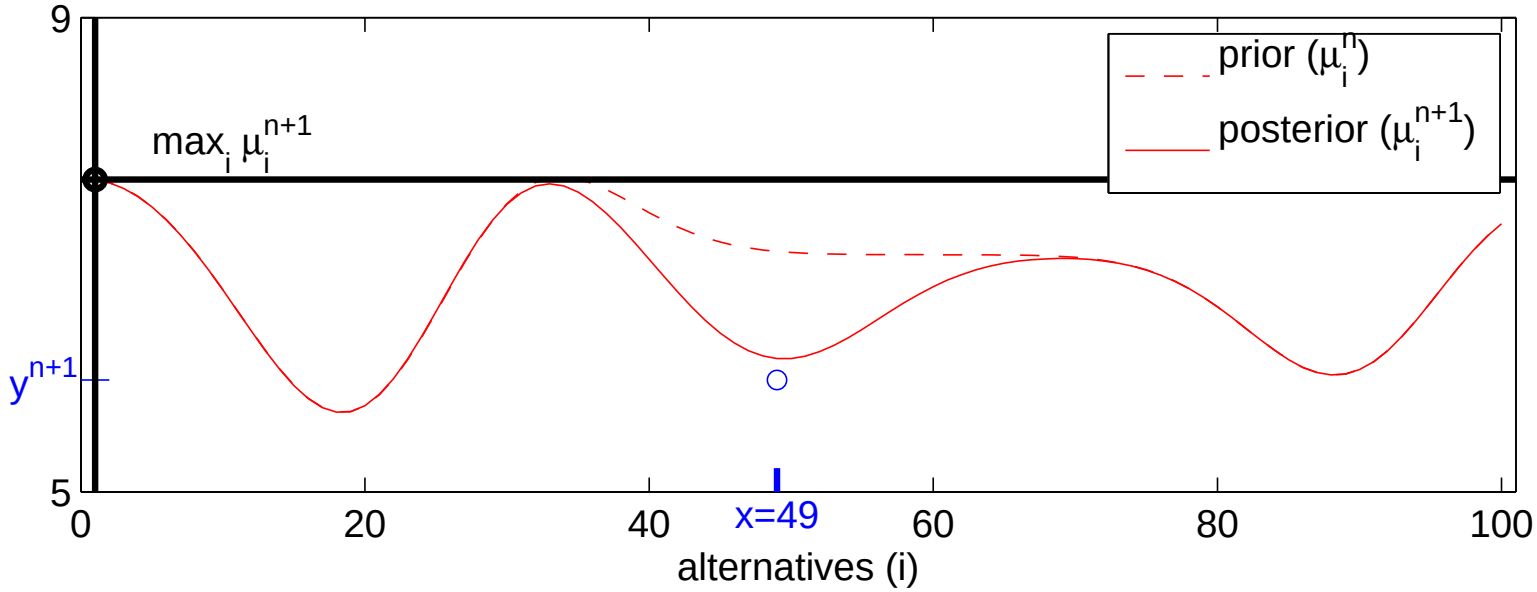


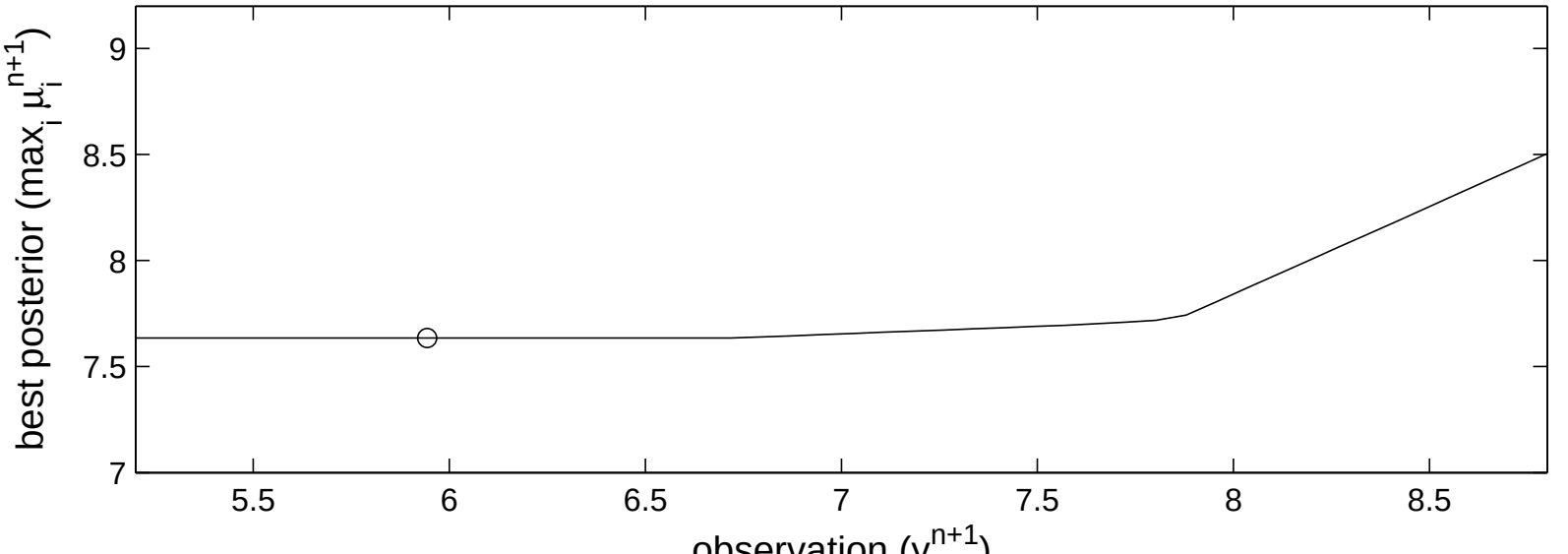
Computing the KG factor requires us to think about how the next measurement will change our posterior.



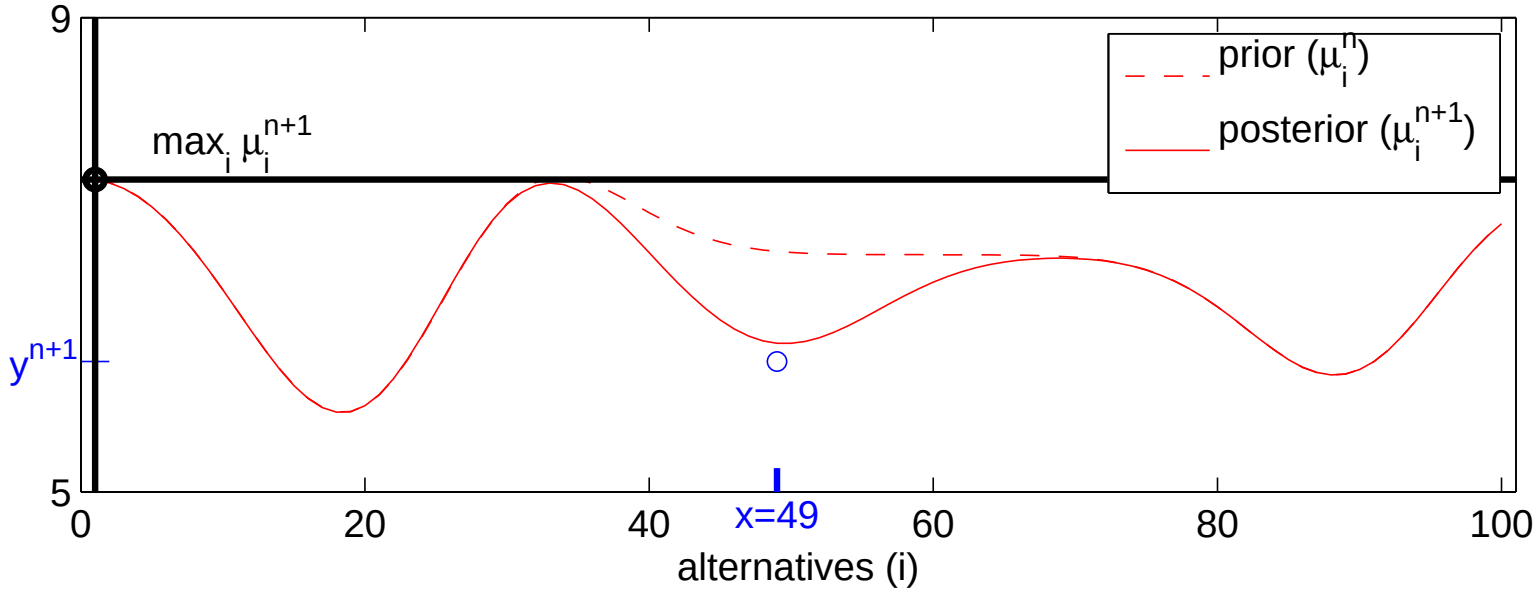


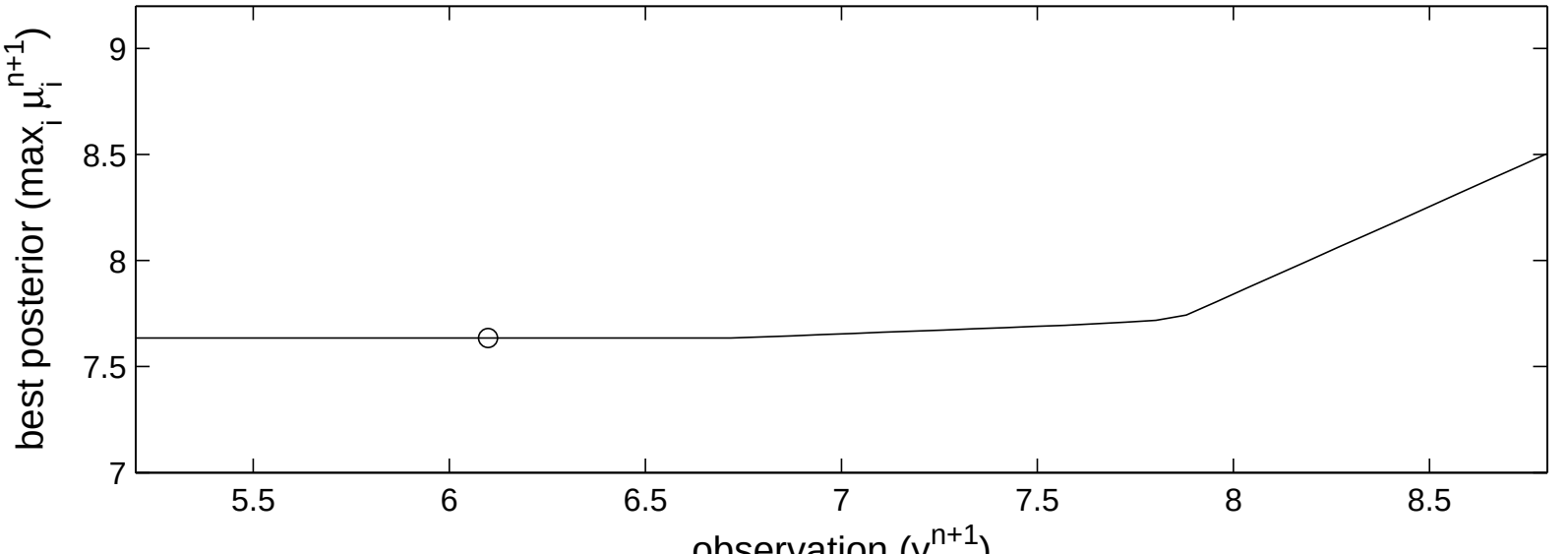
Computing the KG factor requires us to think about how the next measurement will change our posterior.



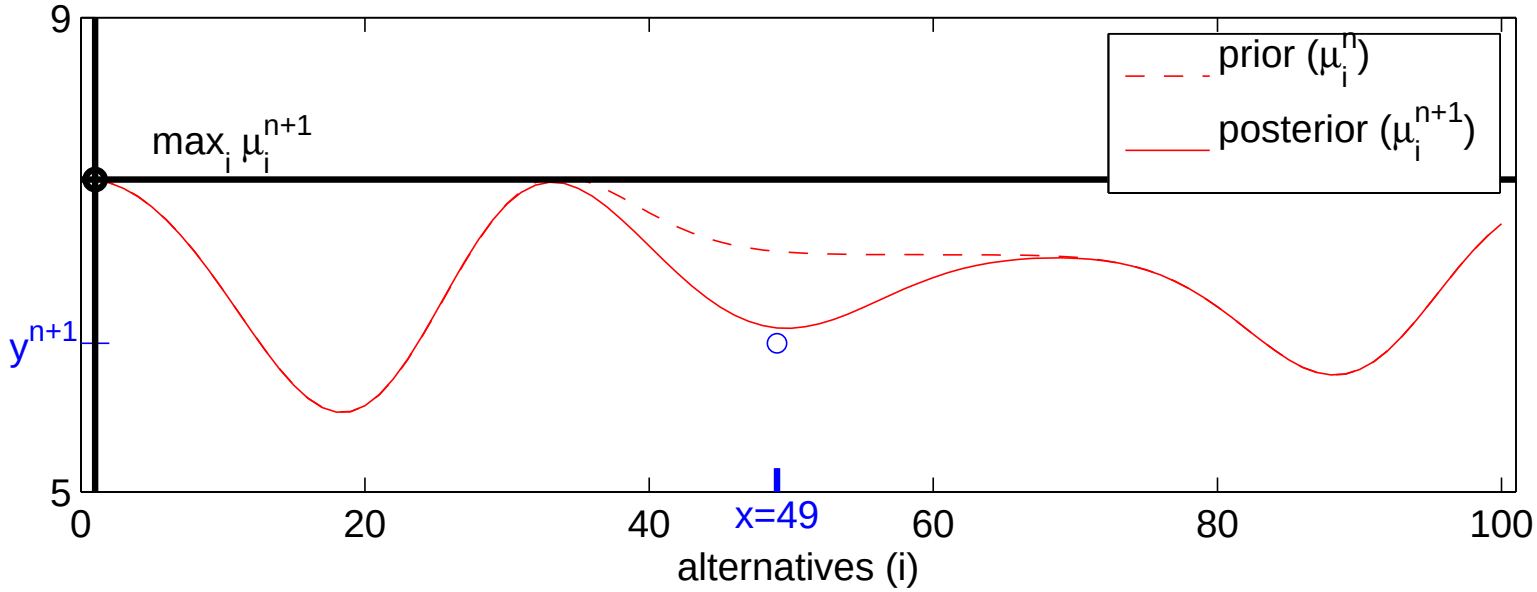


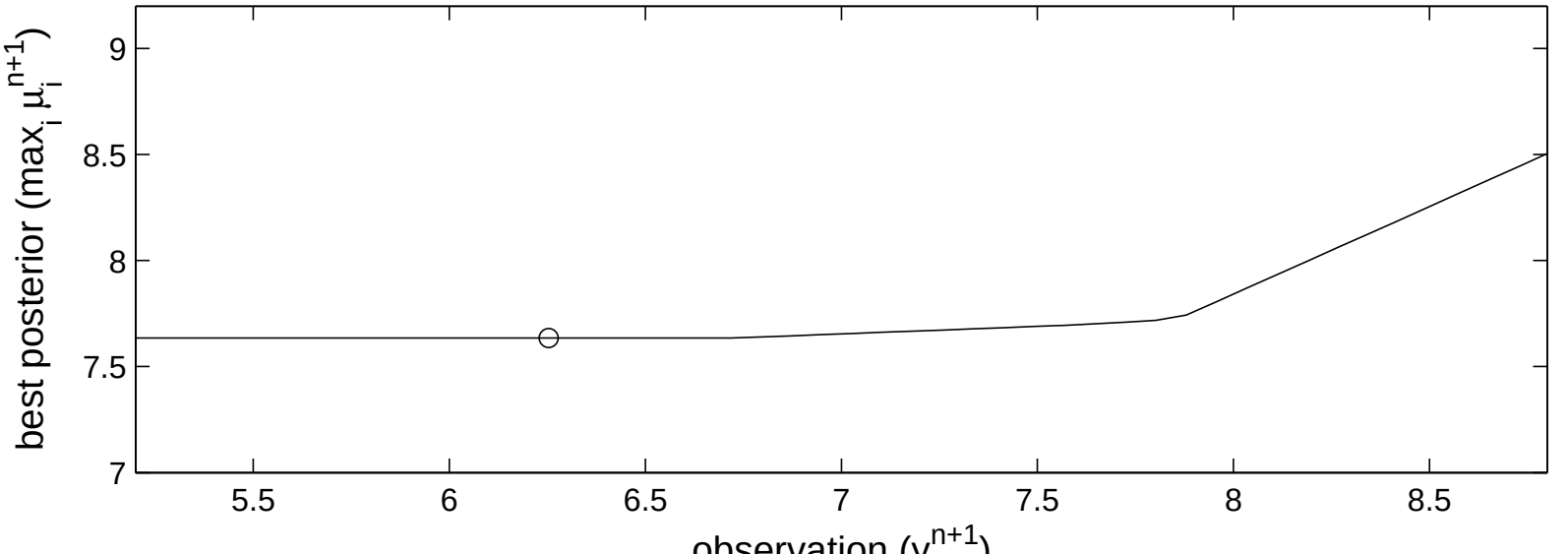
Computing the KG factor requires us to think about how the next measurement will change our posterior.



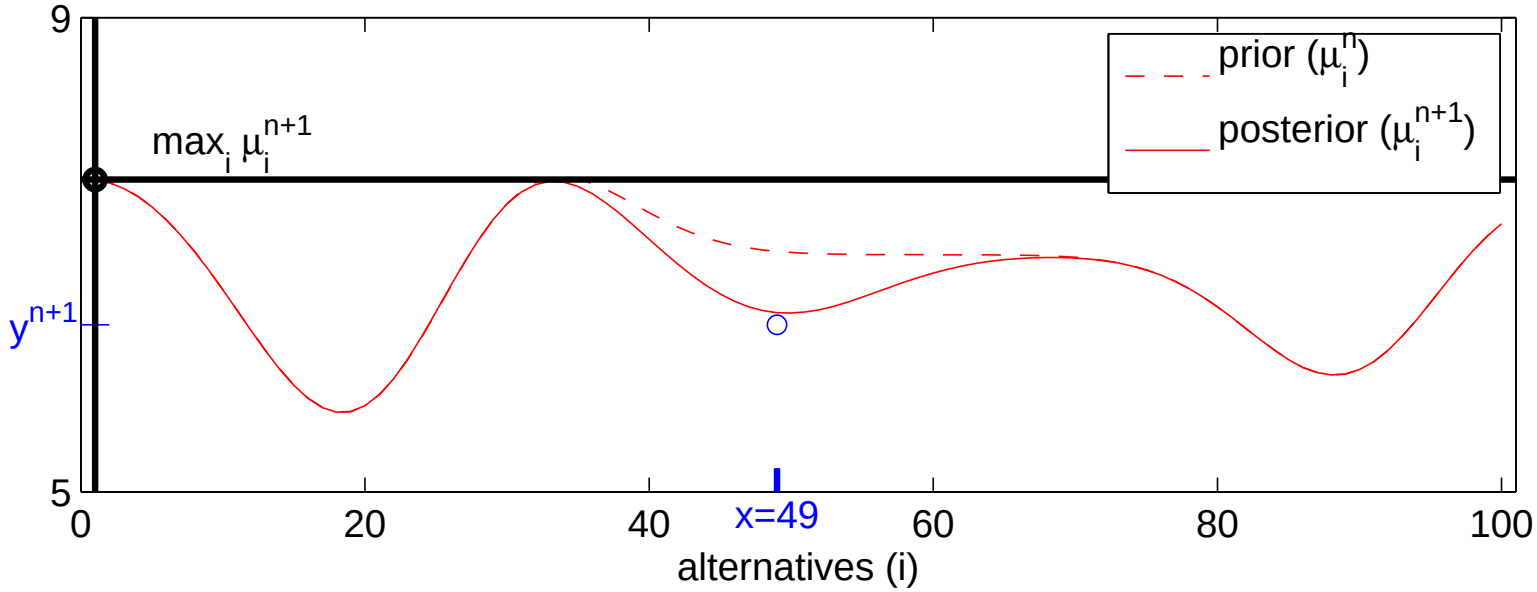


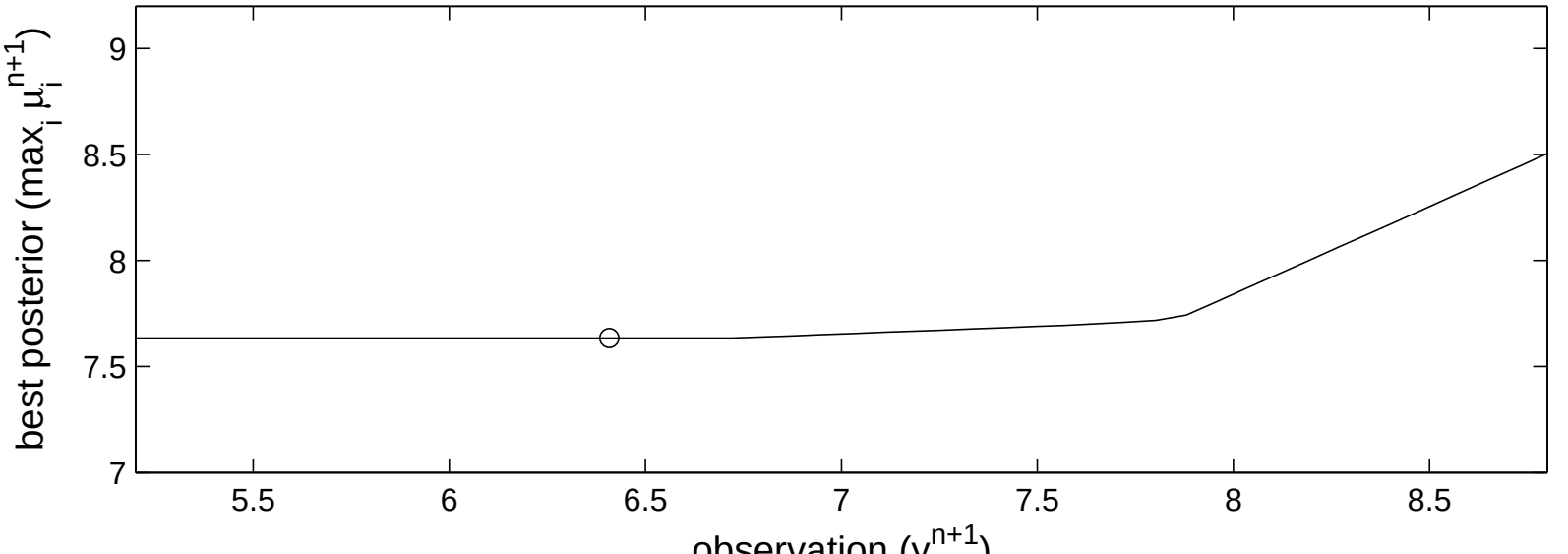
Computing the KG factor requires us to think about how the next measurement will change our posterior.



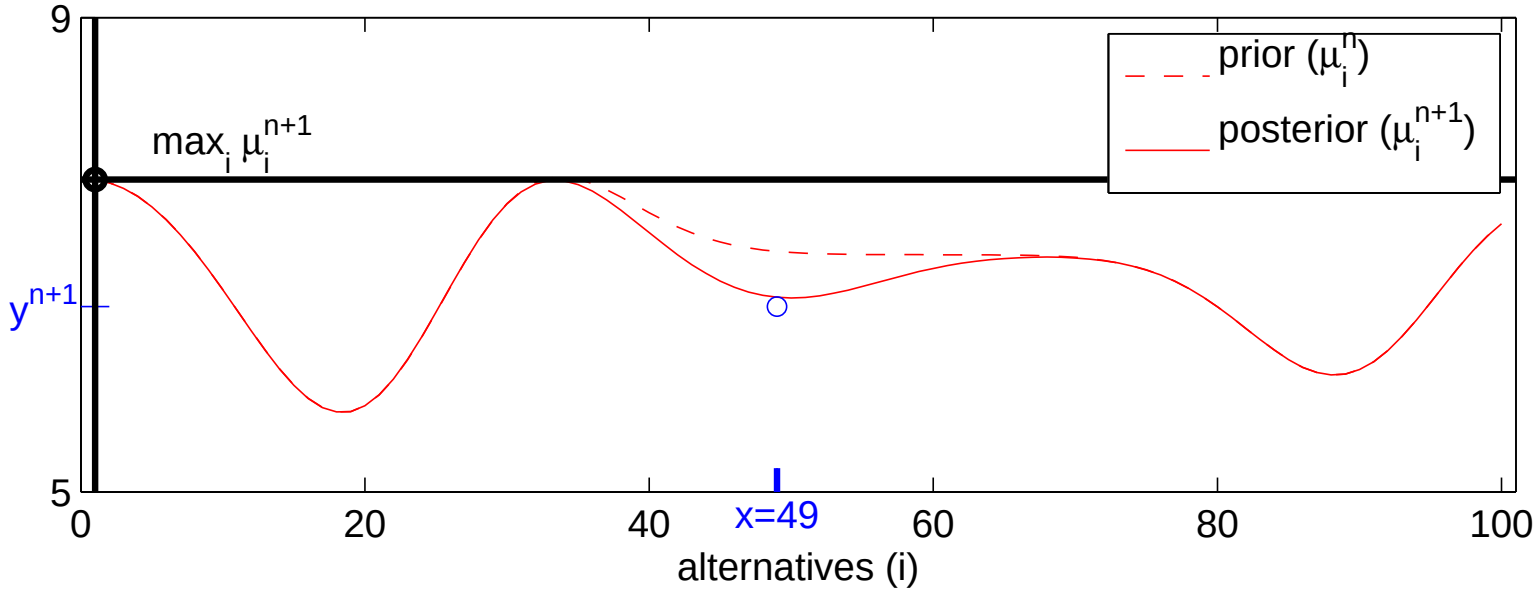


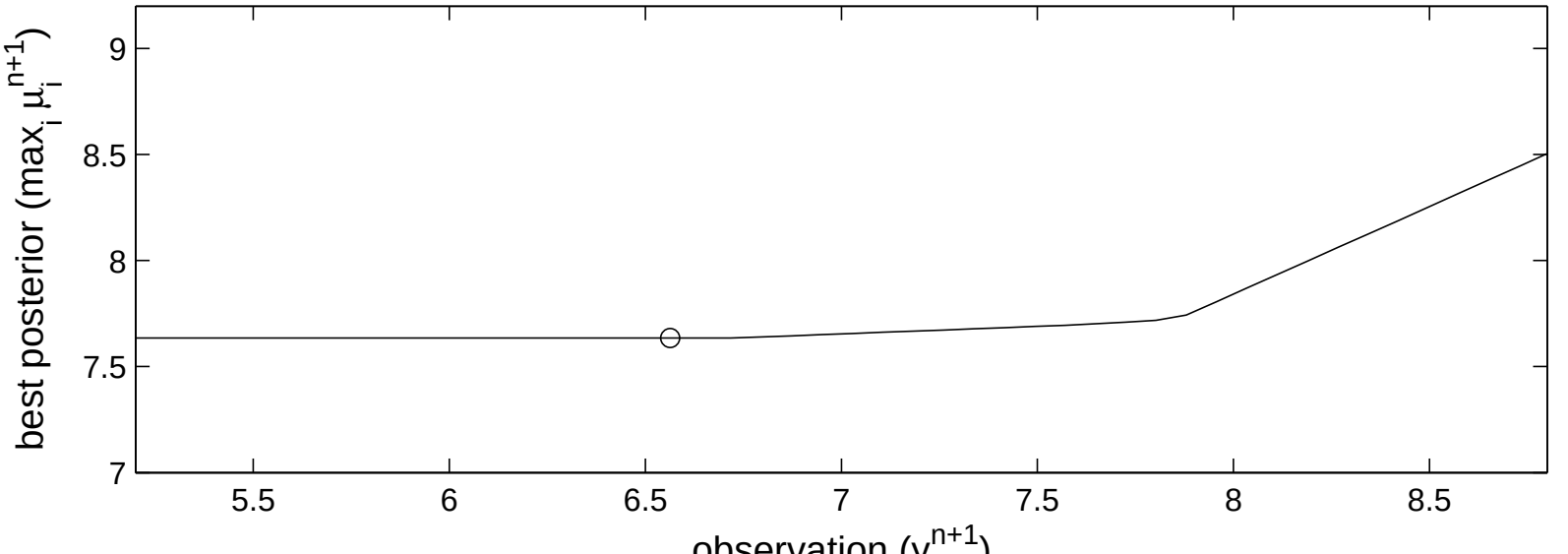
Computing the KG factor requires us to think about how the next measurement will change our posterior.



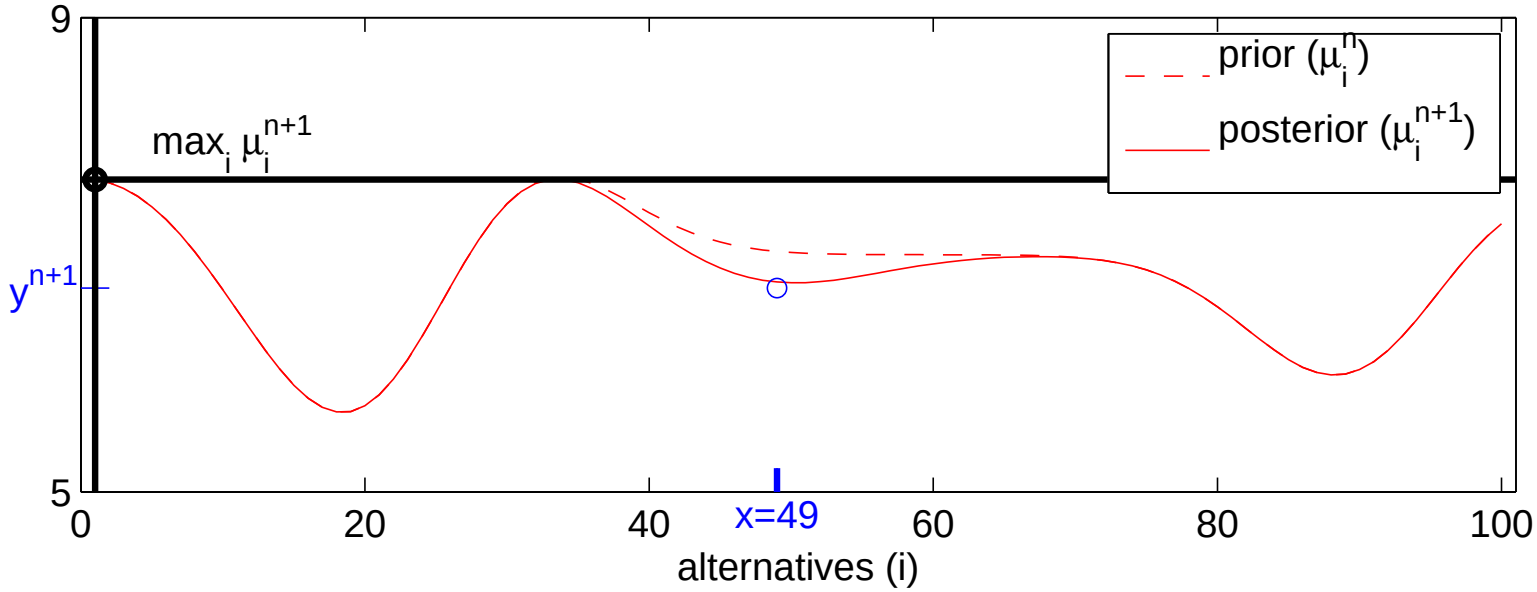


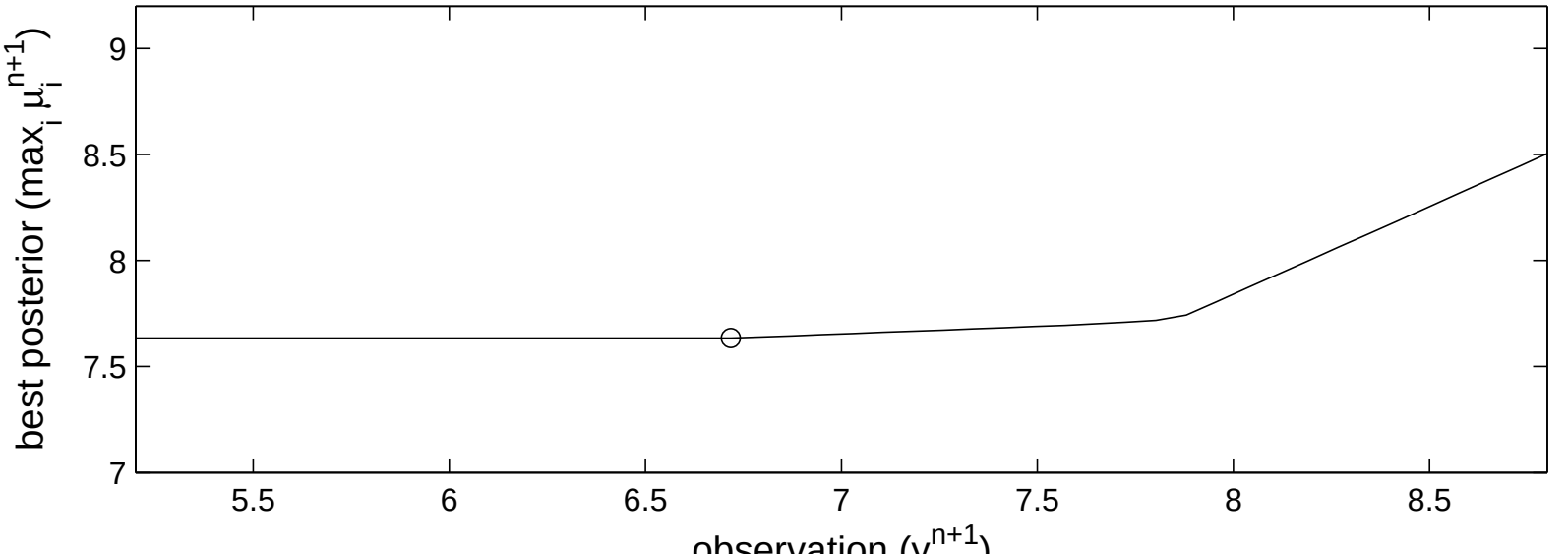
Computing the KG factor requires us to think about how the next measurement will change our posterior.



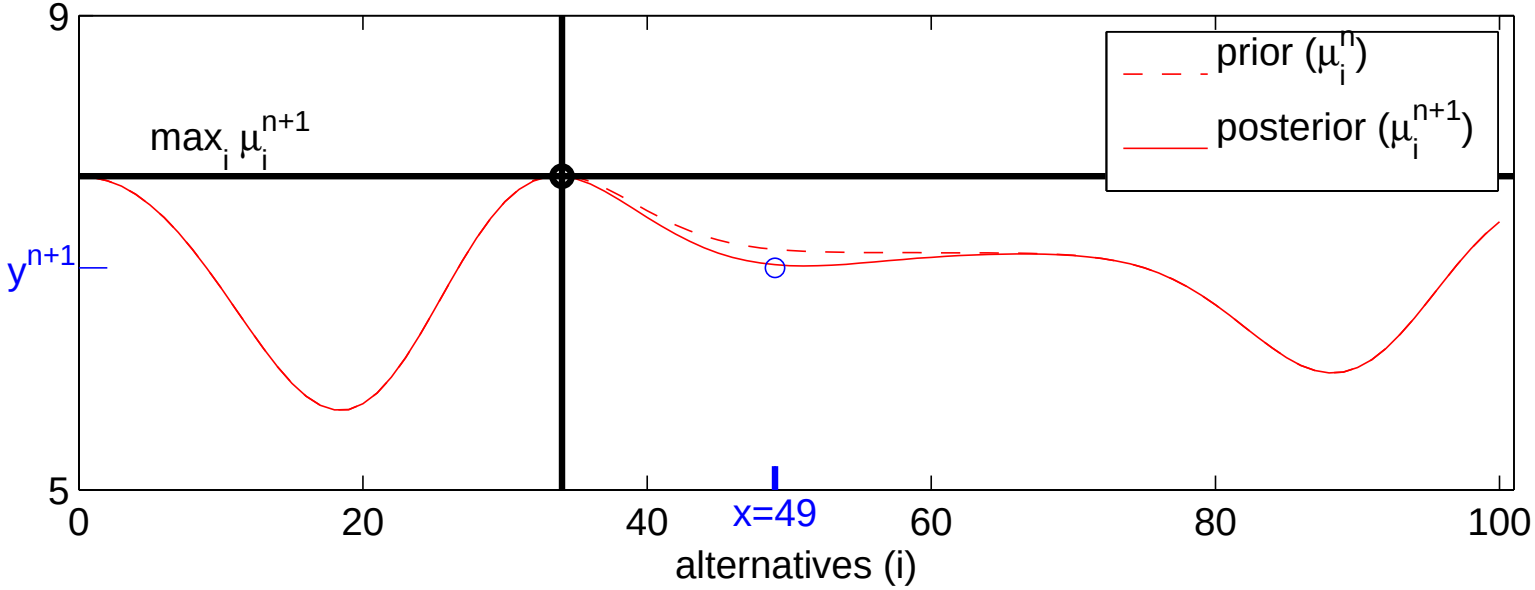


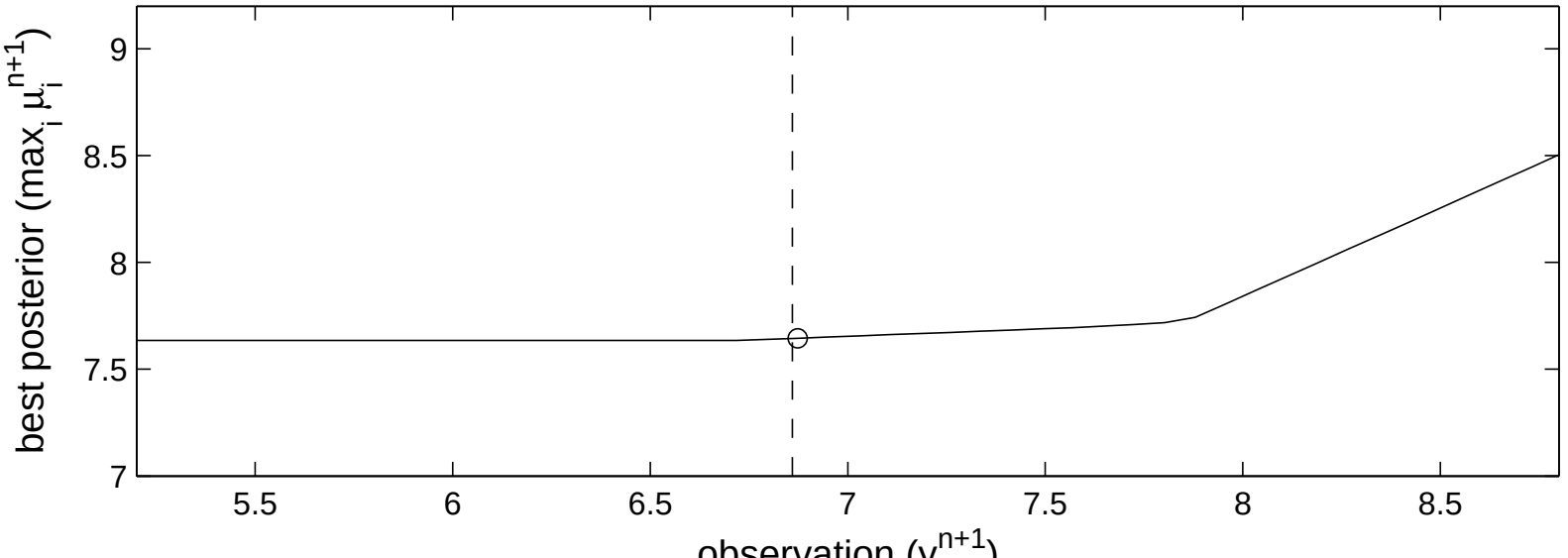
Computing the KG factor requires us to think about how the next measurement will change our posterior.



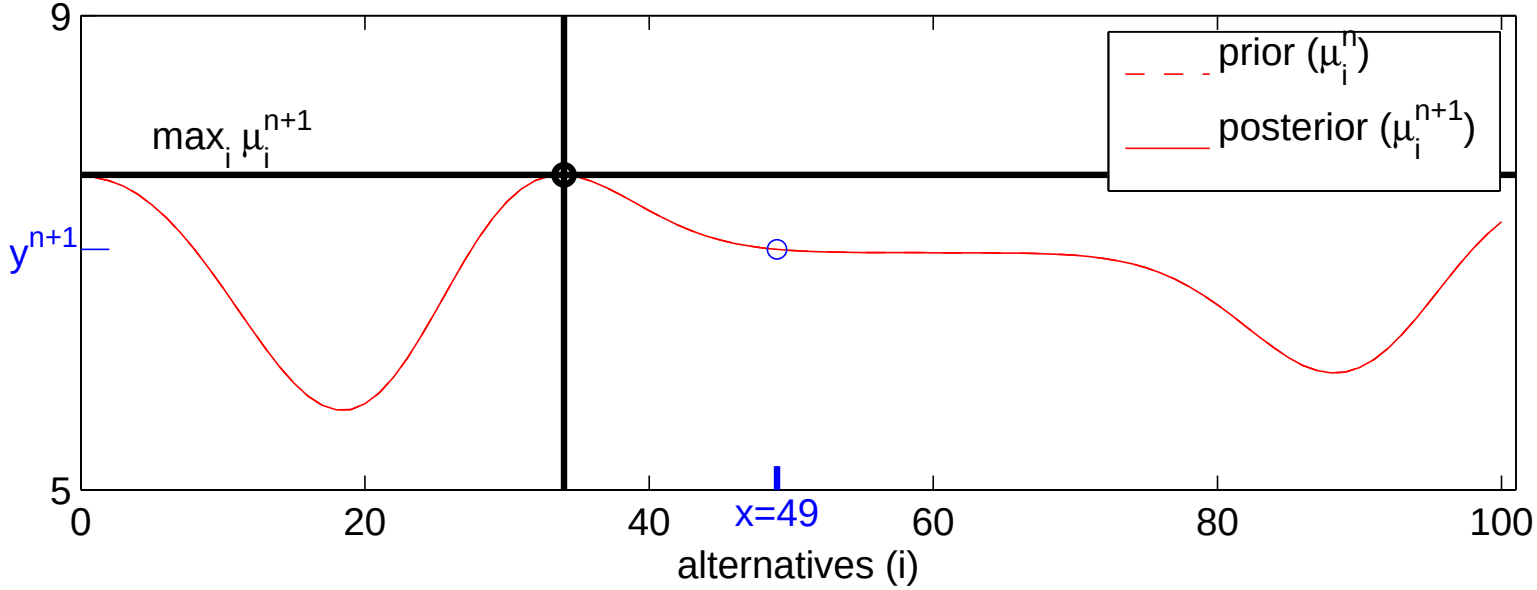


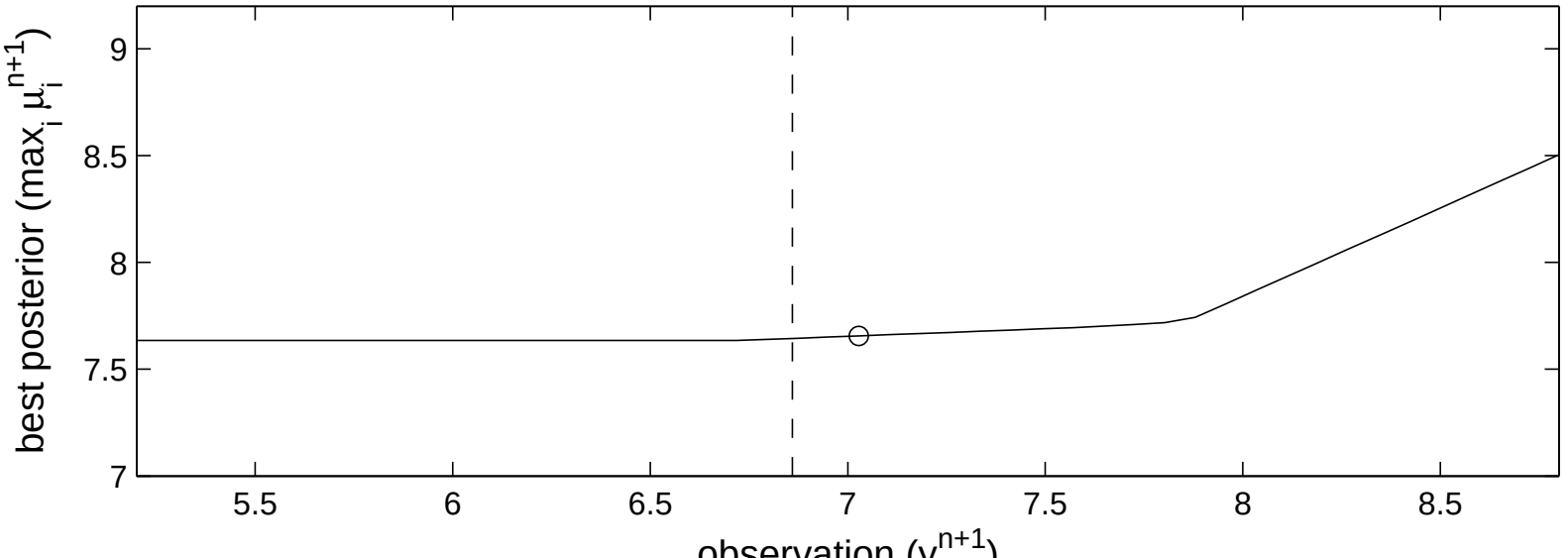
Computing the KG factor requires us to think about how the next measurement will change our posterior.



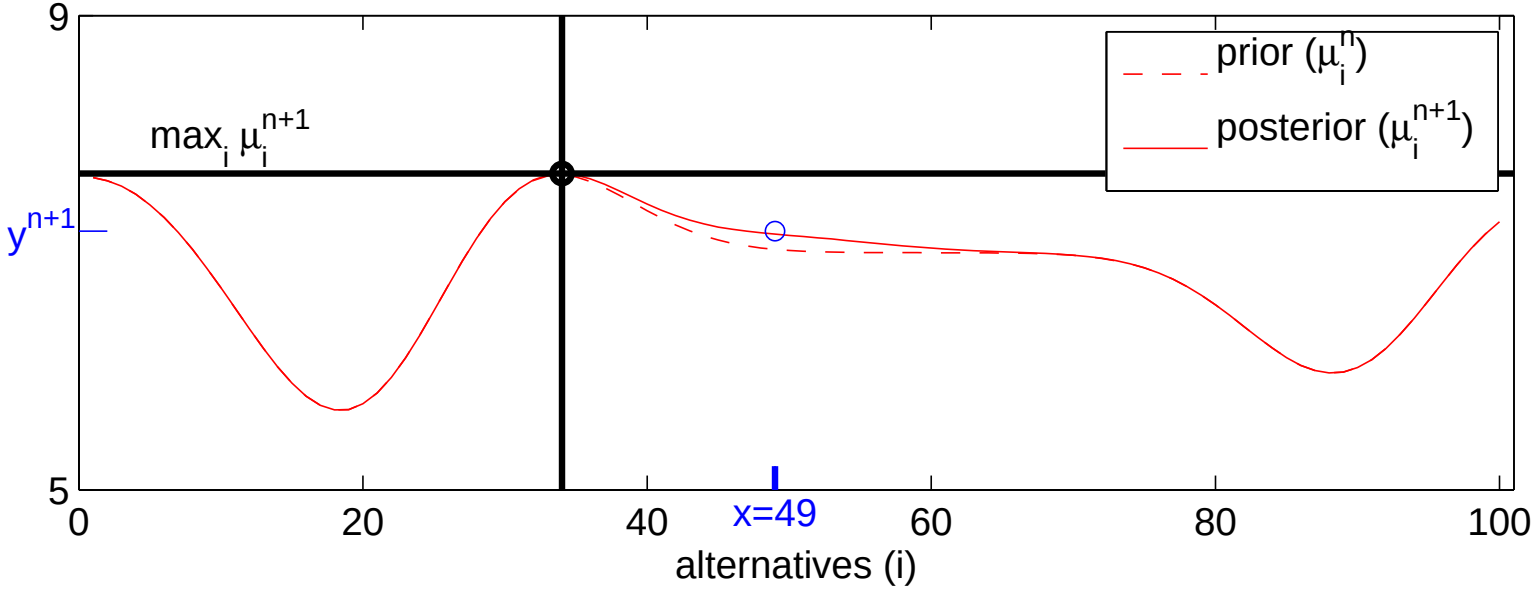


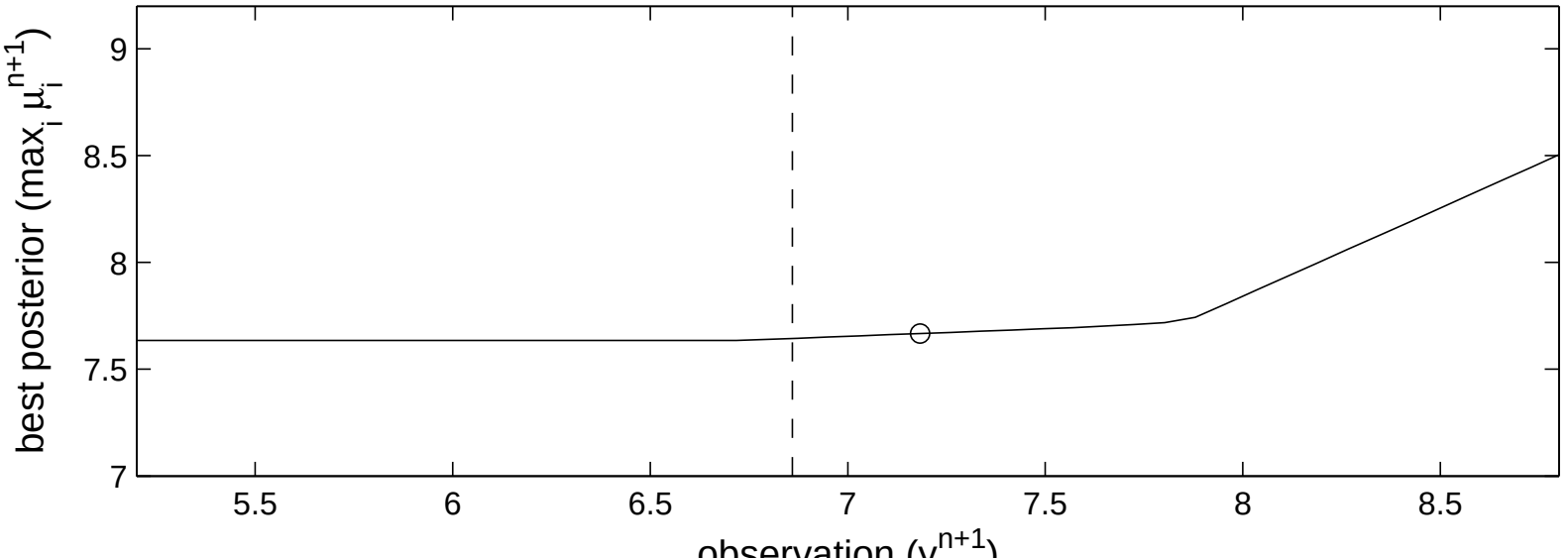
Computing the KG factor requires us to think about how the next measurement will change our posterior.



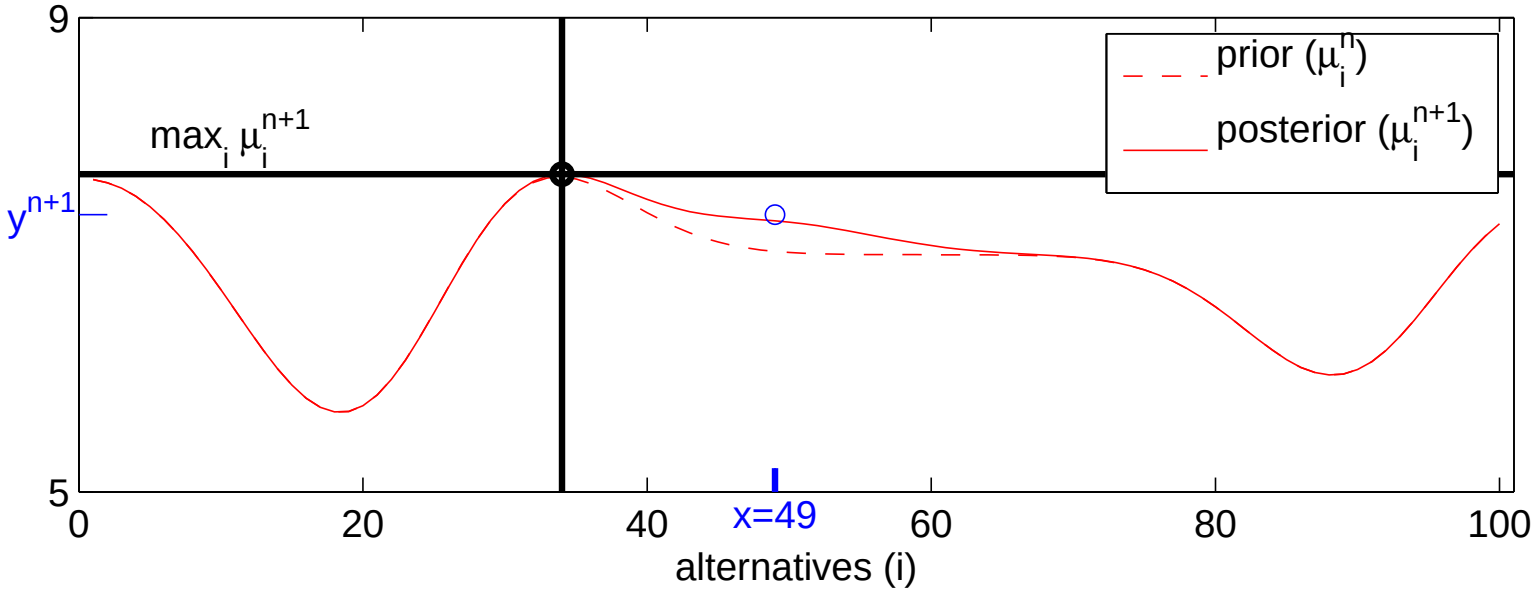


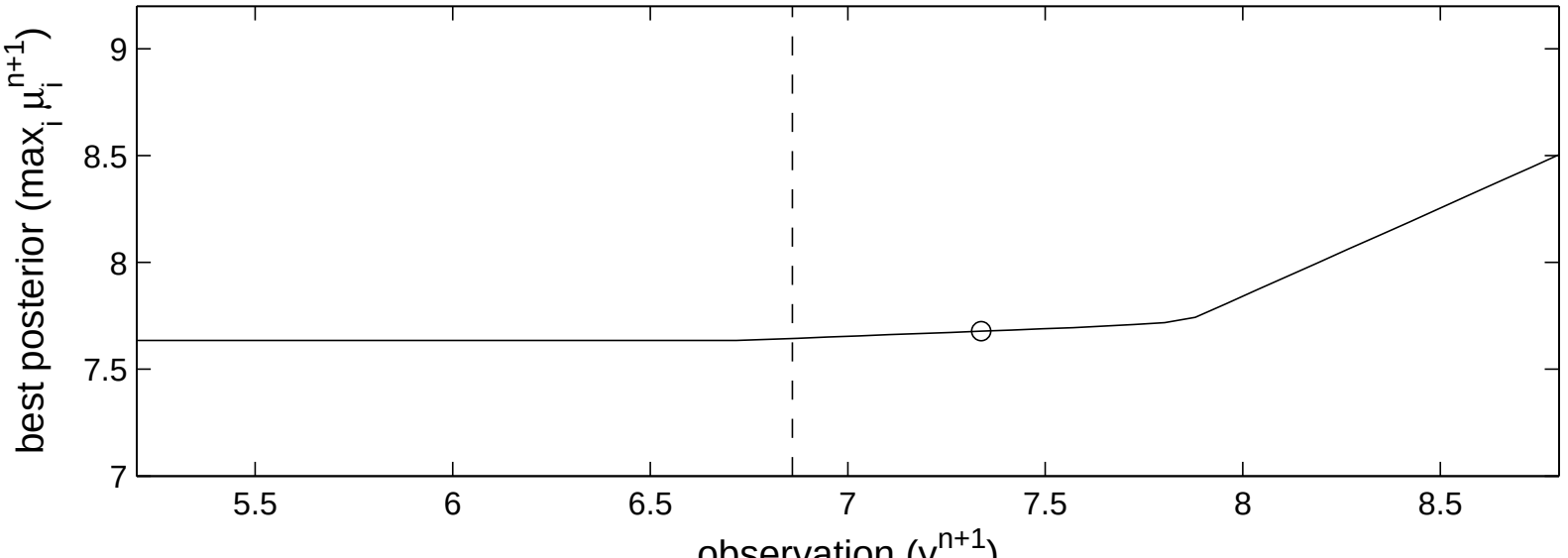
Computing the KG factor requires us to think about how the next measurement will change our posterior.



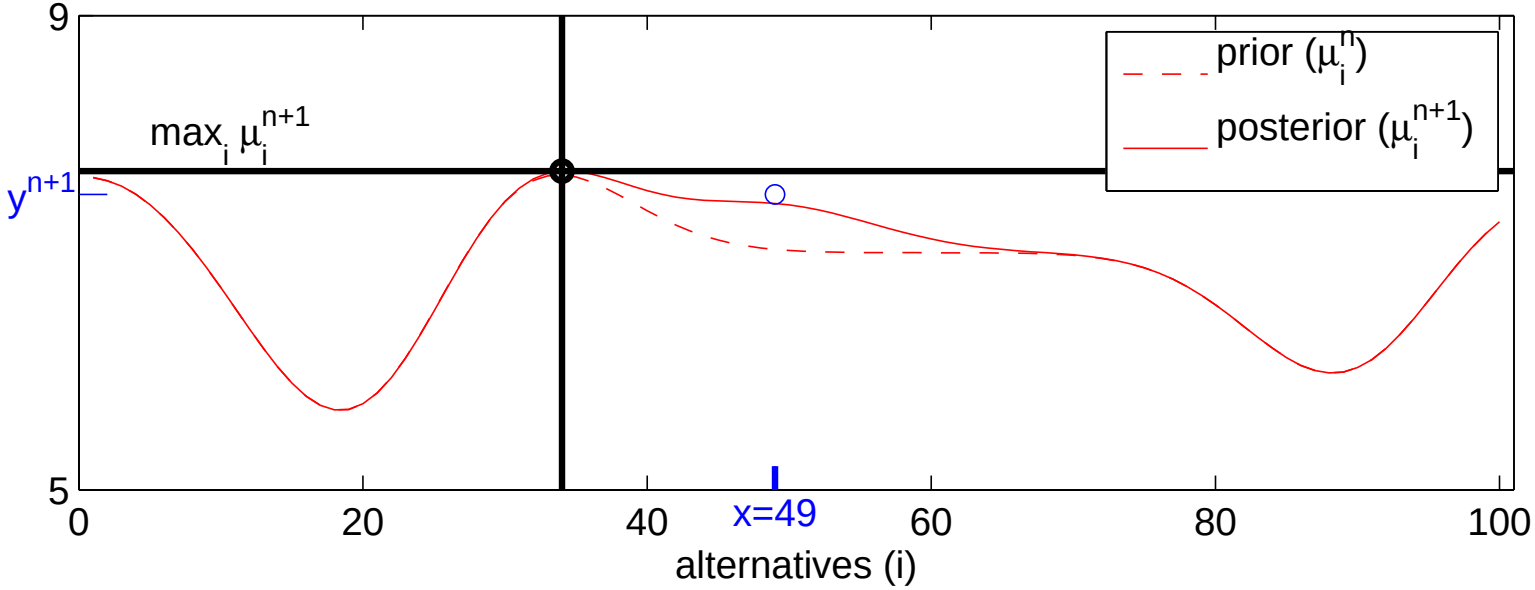


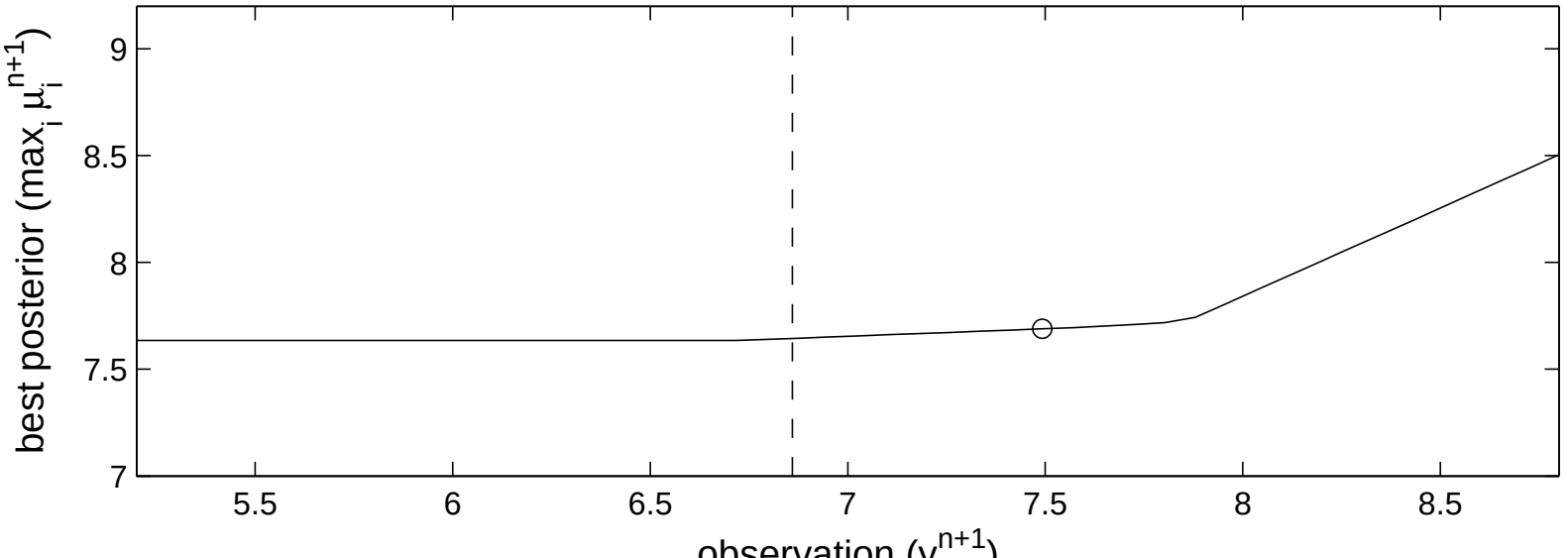
Computing the KG factor requires us to think about how the next measurement will change our posterior.



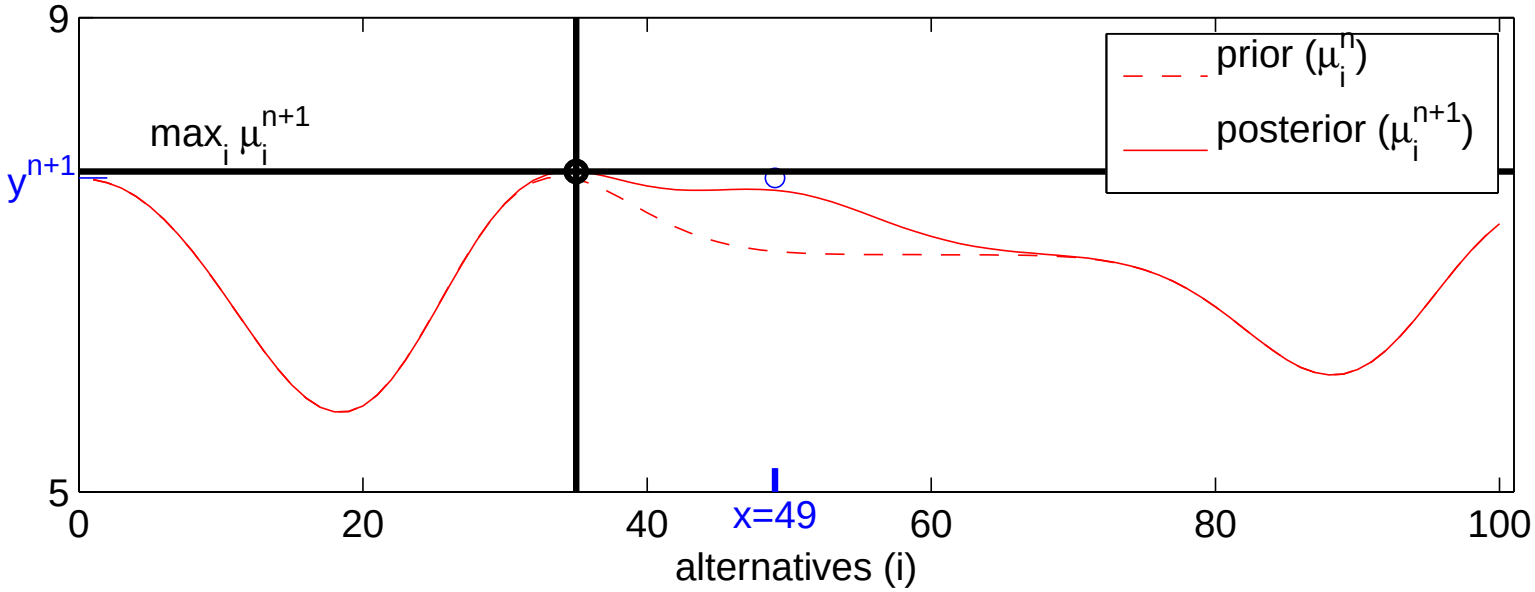


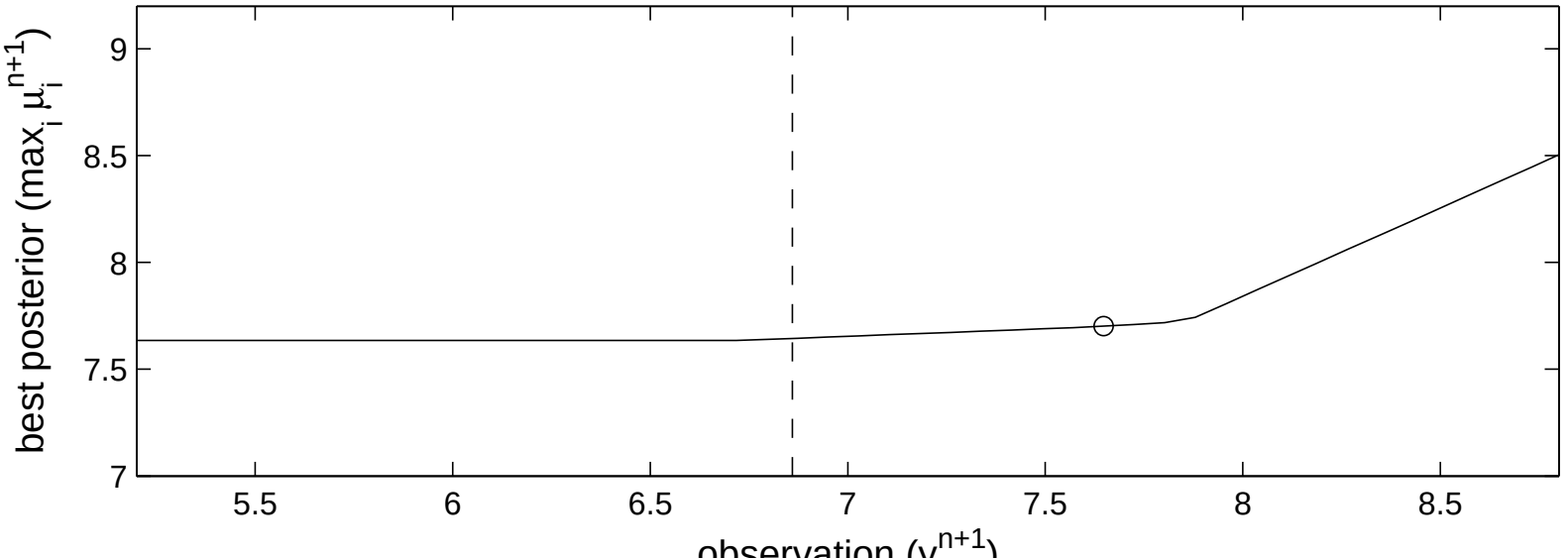
Computing the KG factor requires us to think about how the next measurement will change our posterior.



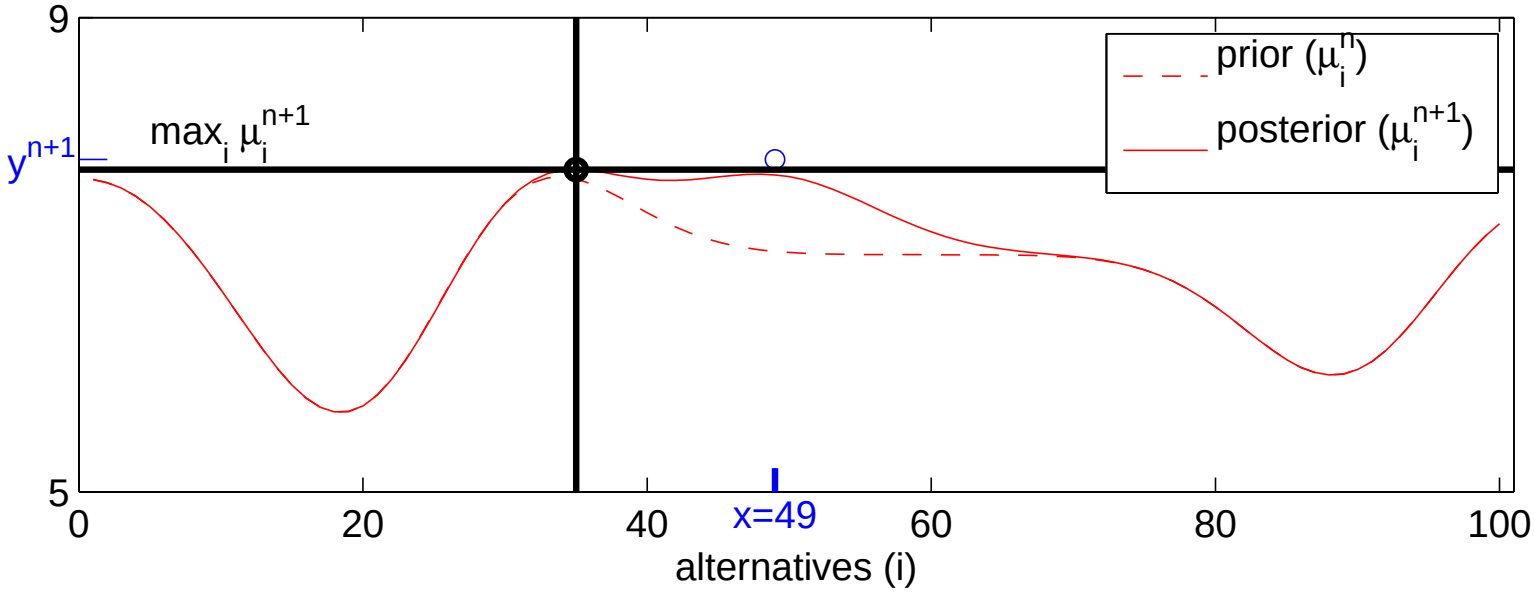


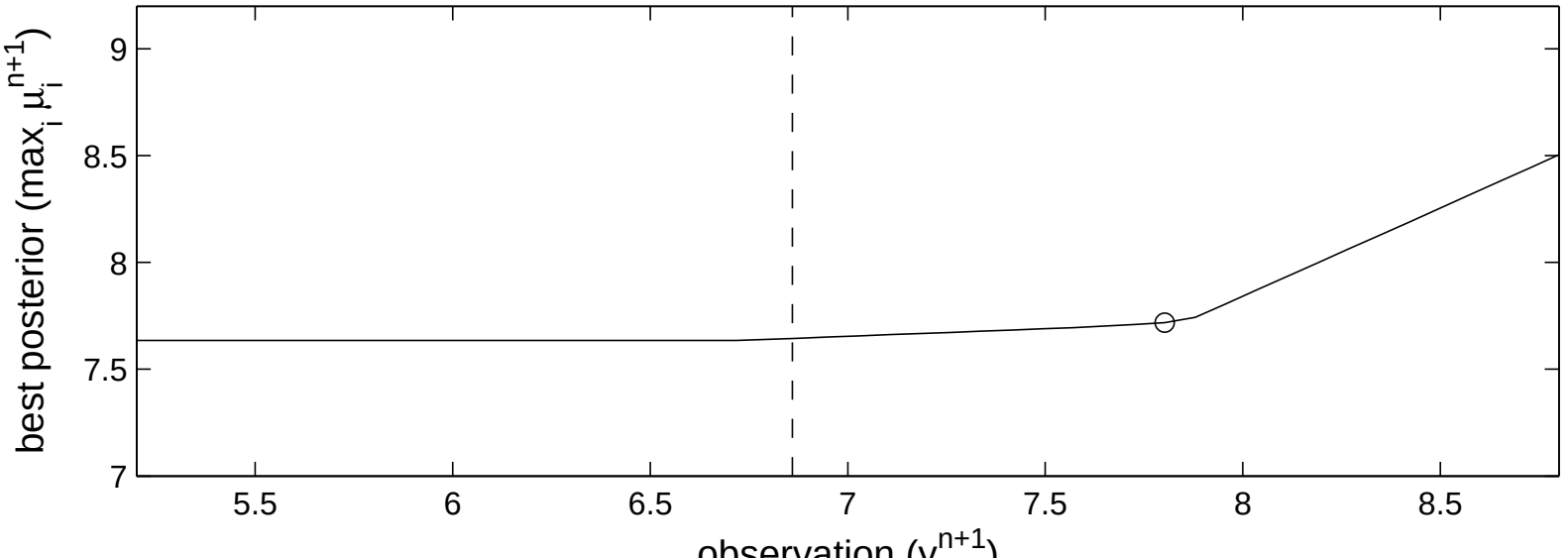
Computing the KG factor requires us to think about how the next measurement will change our posterior.



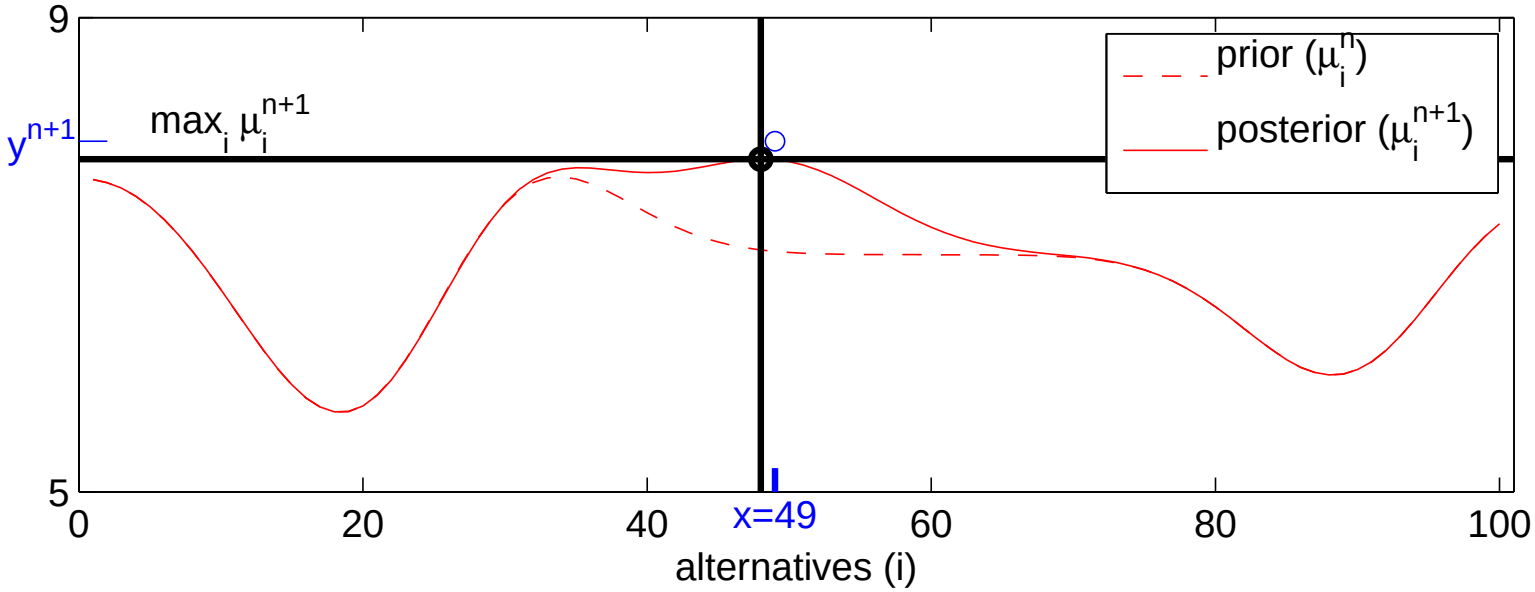


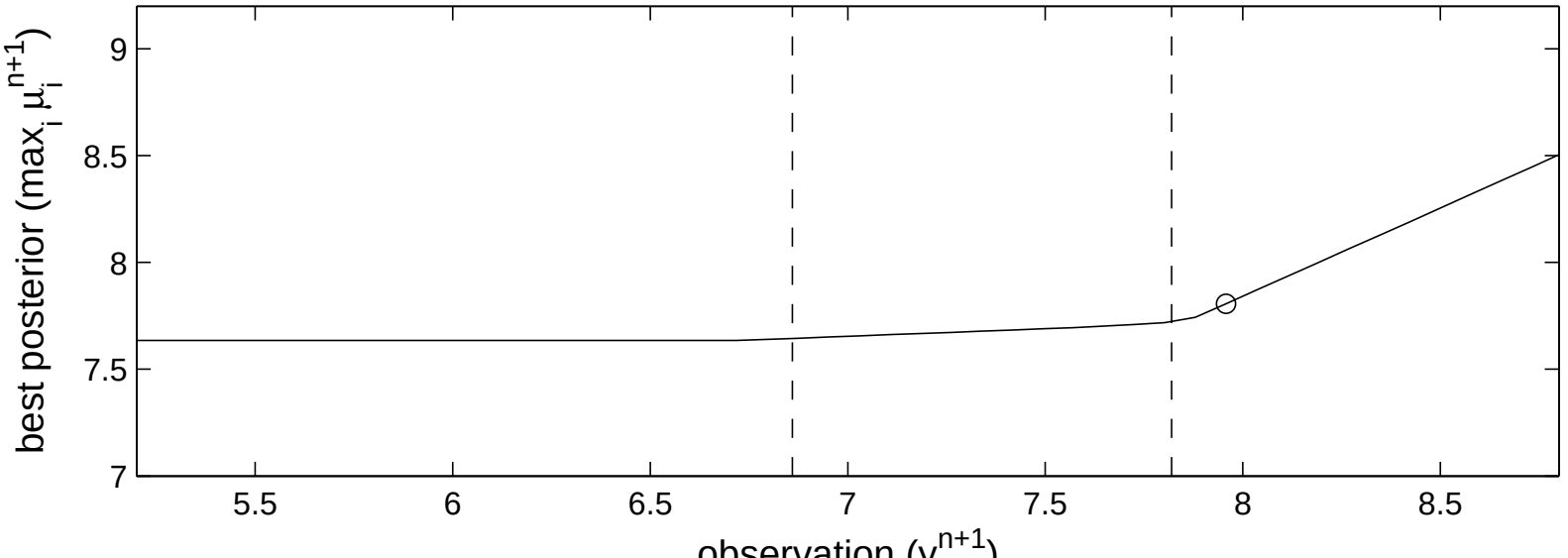
Computing the KG factor requires us to think about how the next measurement will change our posterior.



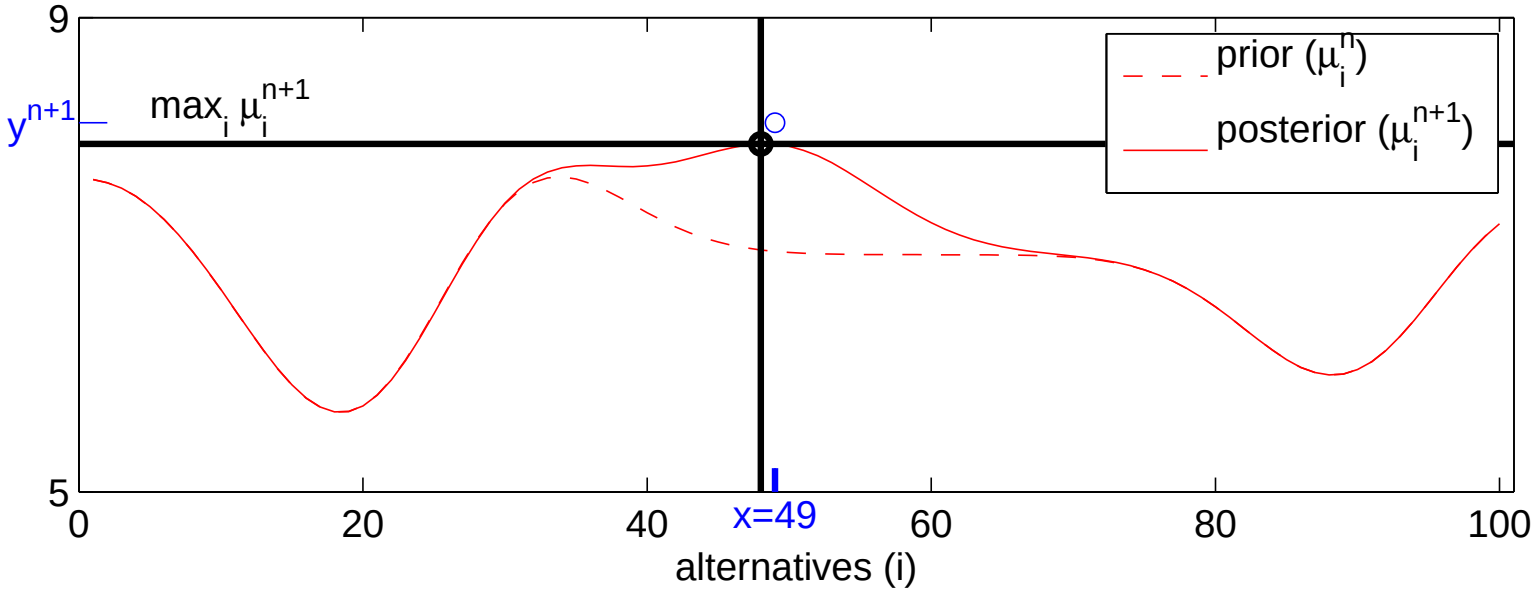


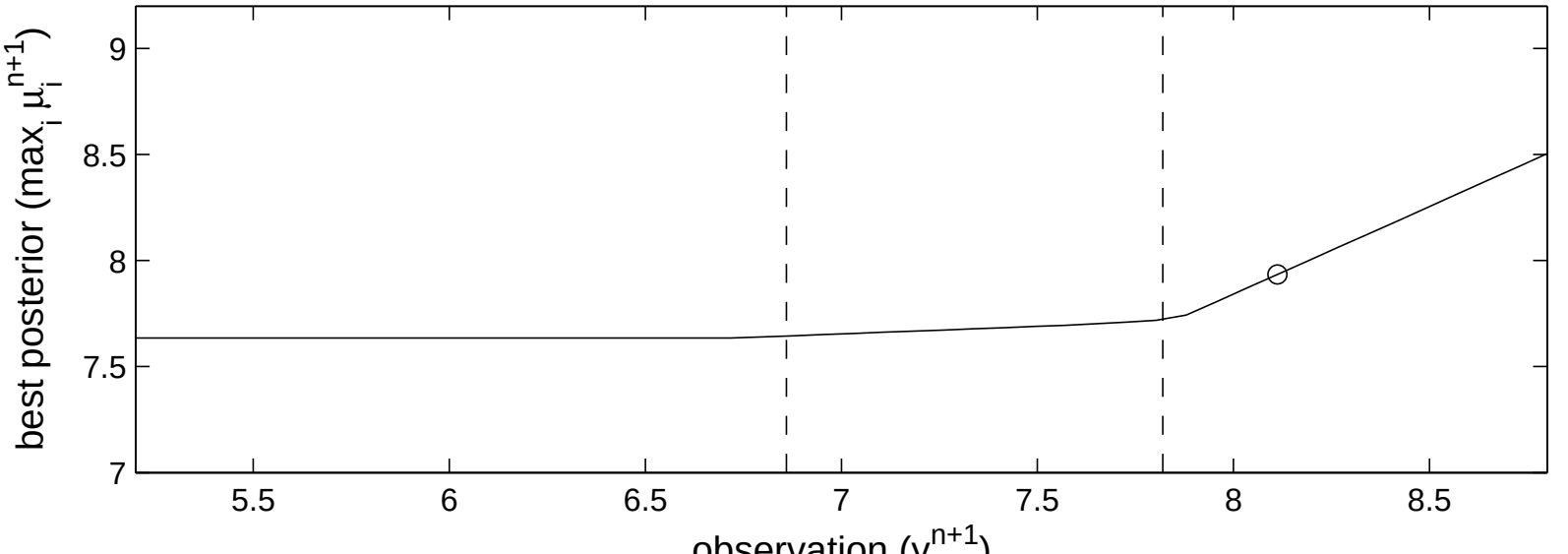
Computing the KG factor requires us to think about how the next measurement will change our posterior.



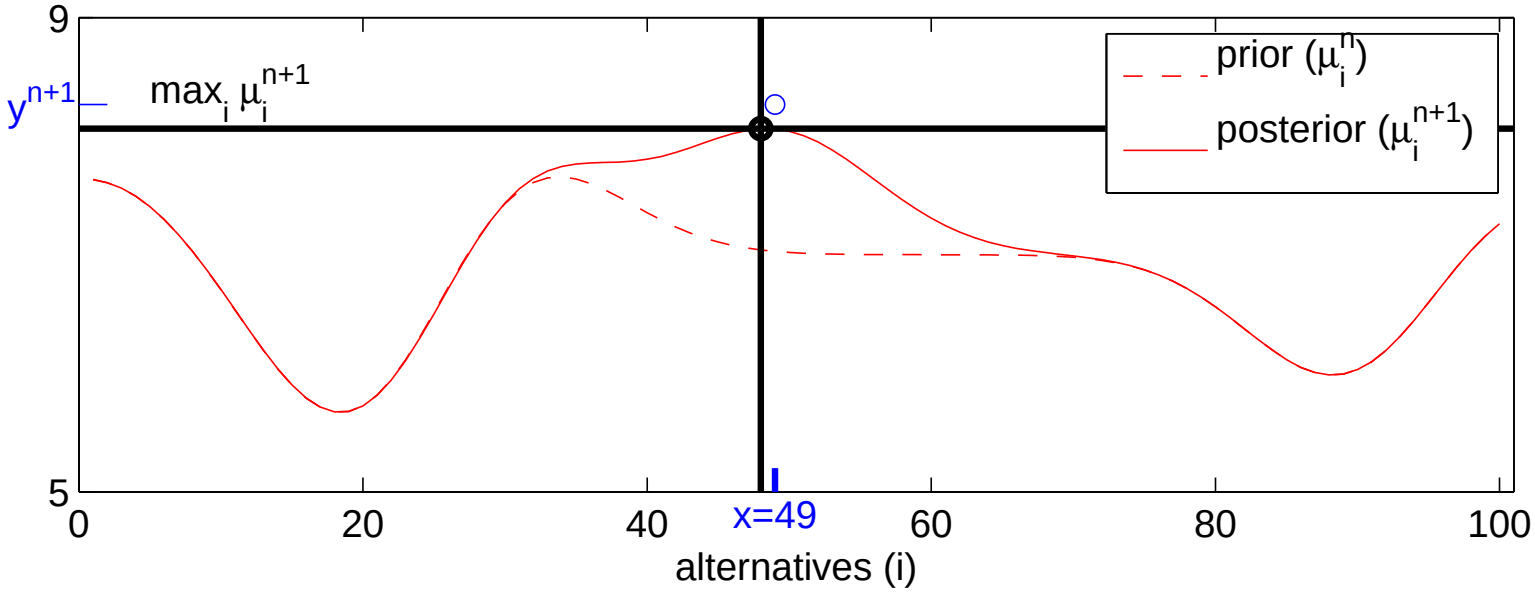


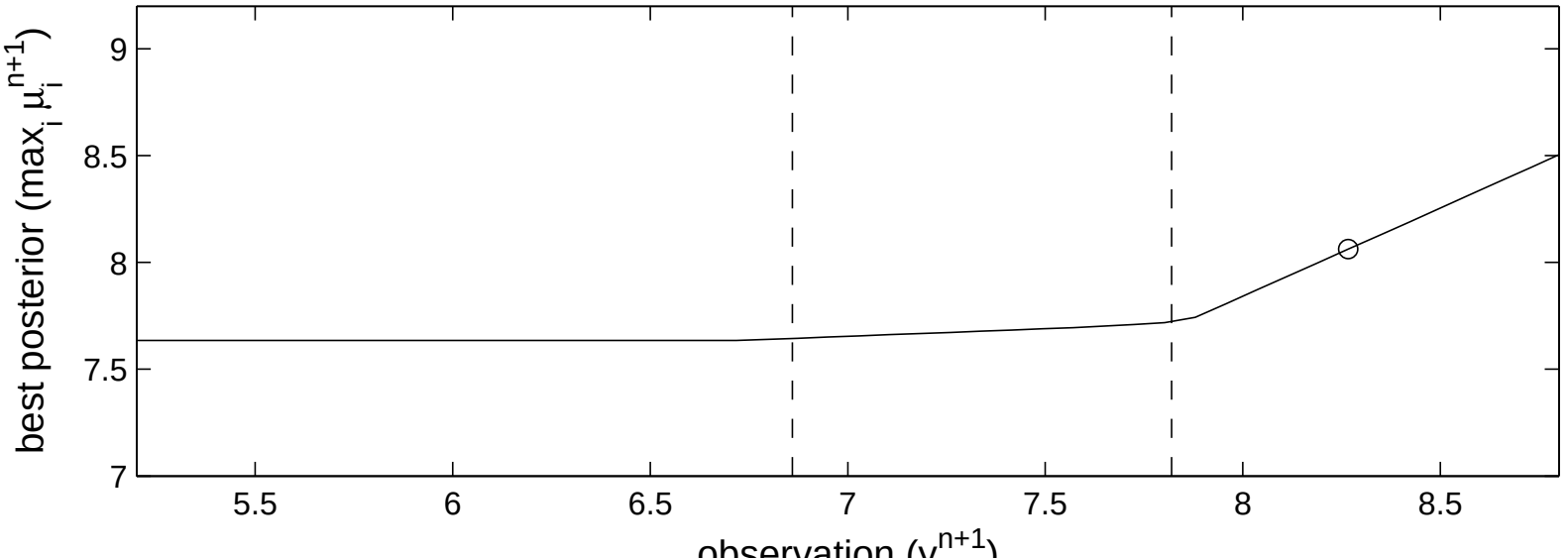
Computing the KG factor requires us to think about how the next measurement will change our posterior.



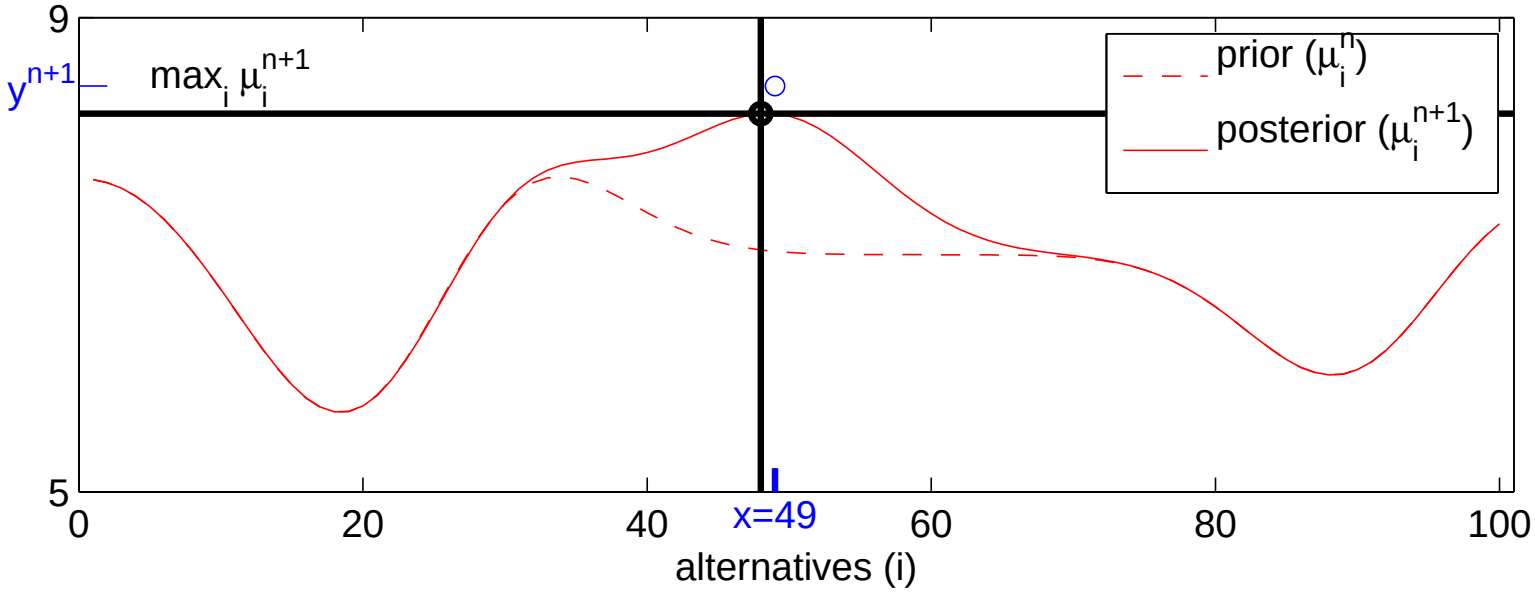


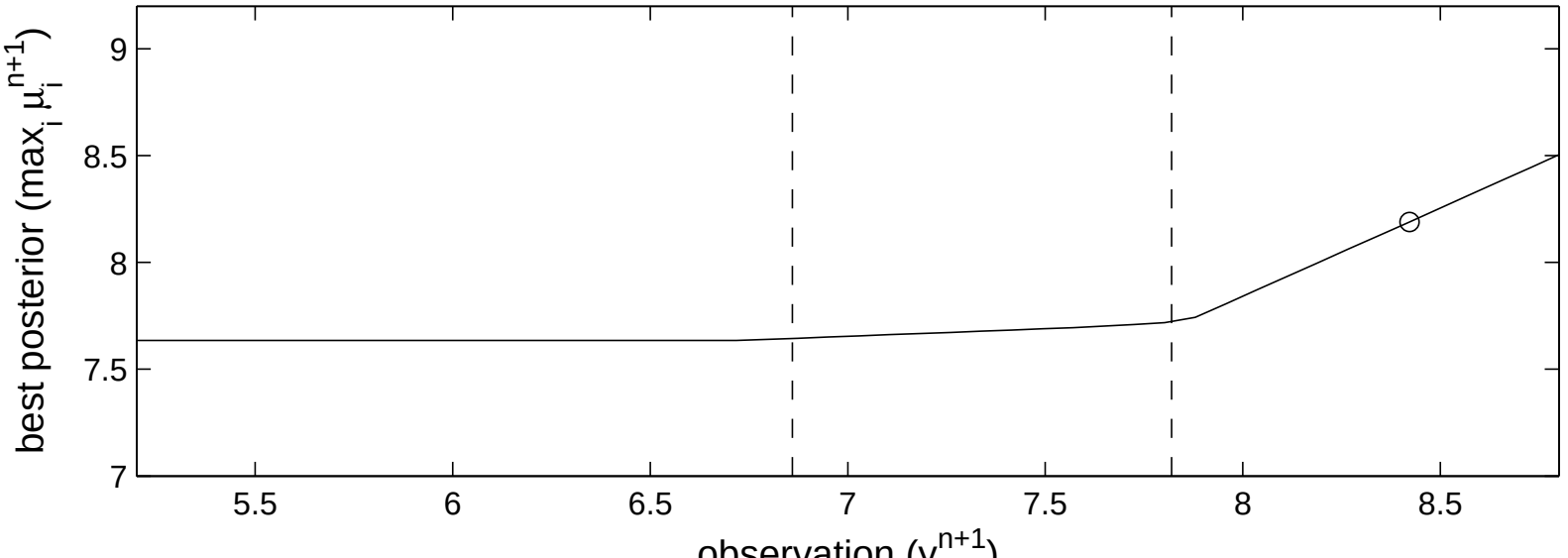
Computing the KG factor requires us to think about how the next measurement will change our posterior.



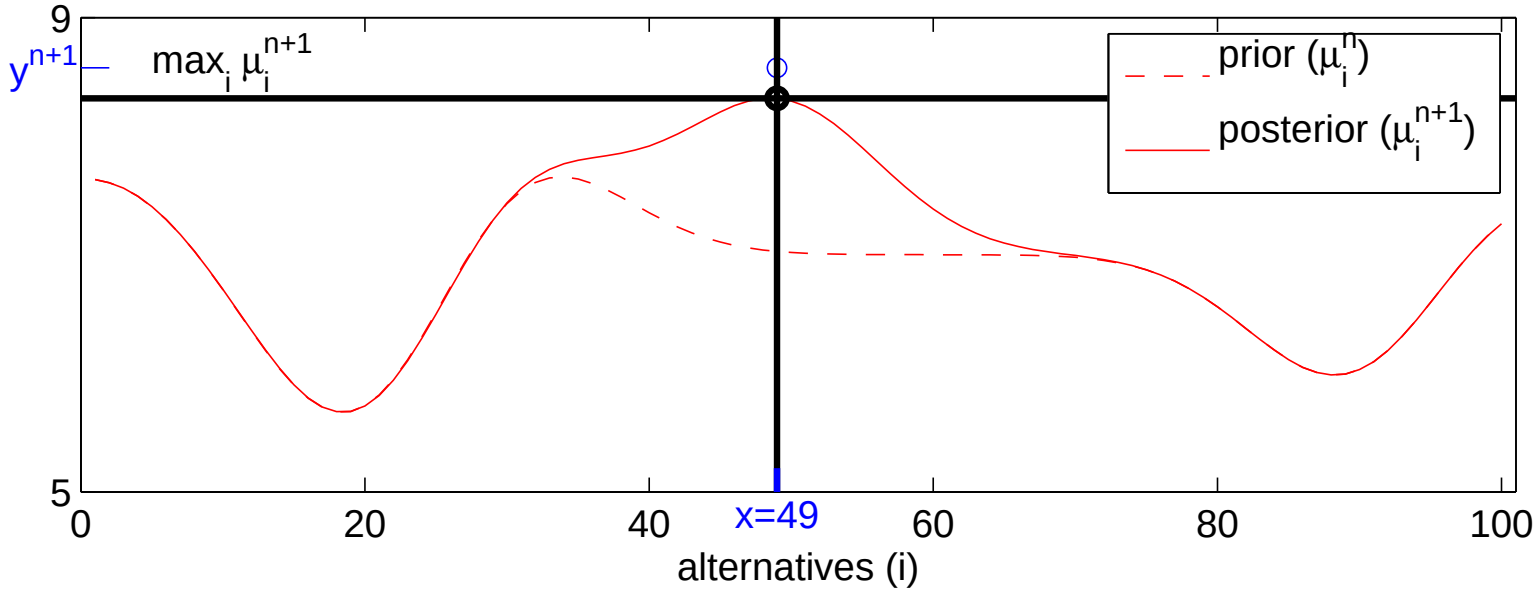


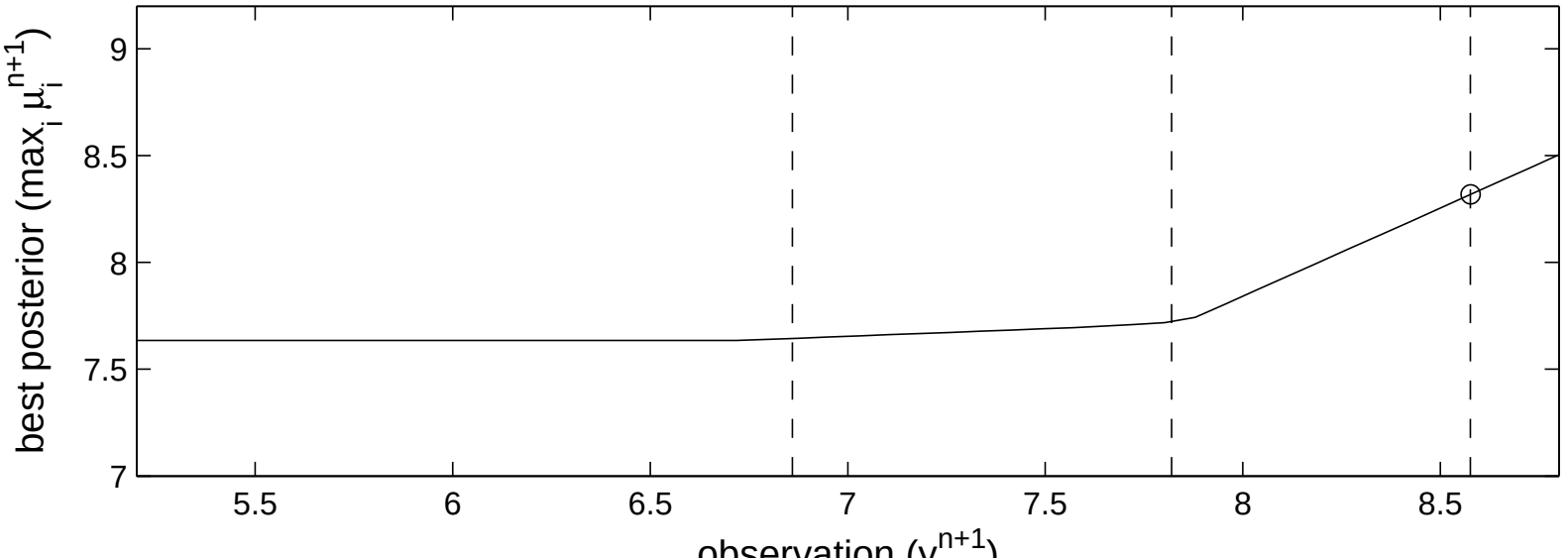
Computing the KG factor requires us to think about how the next measurement will change our posterior.





Computing the KG factor requires us to think about how the next measurement will change our posterior.





Here's how we compute the knowledge-gradient factor

In general, to compute the KG factor for a candidate measurement x:

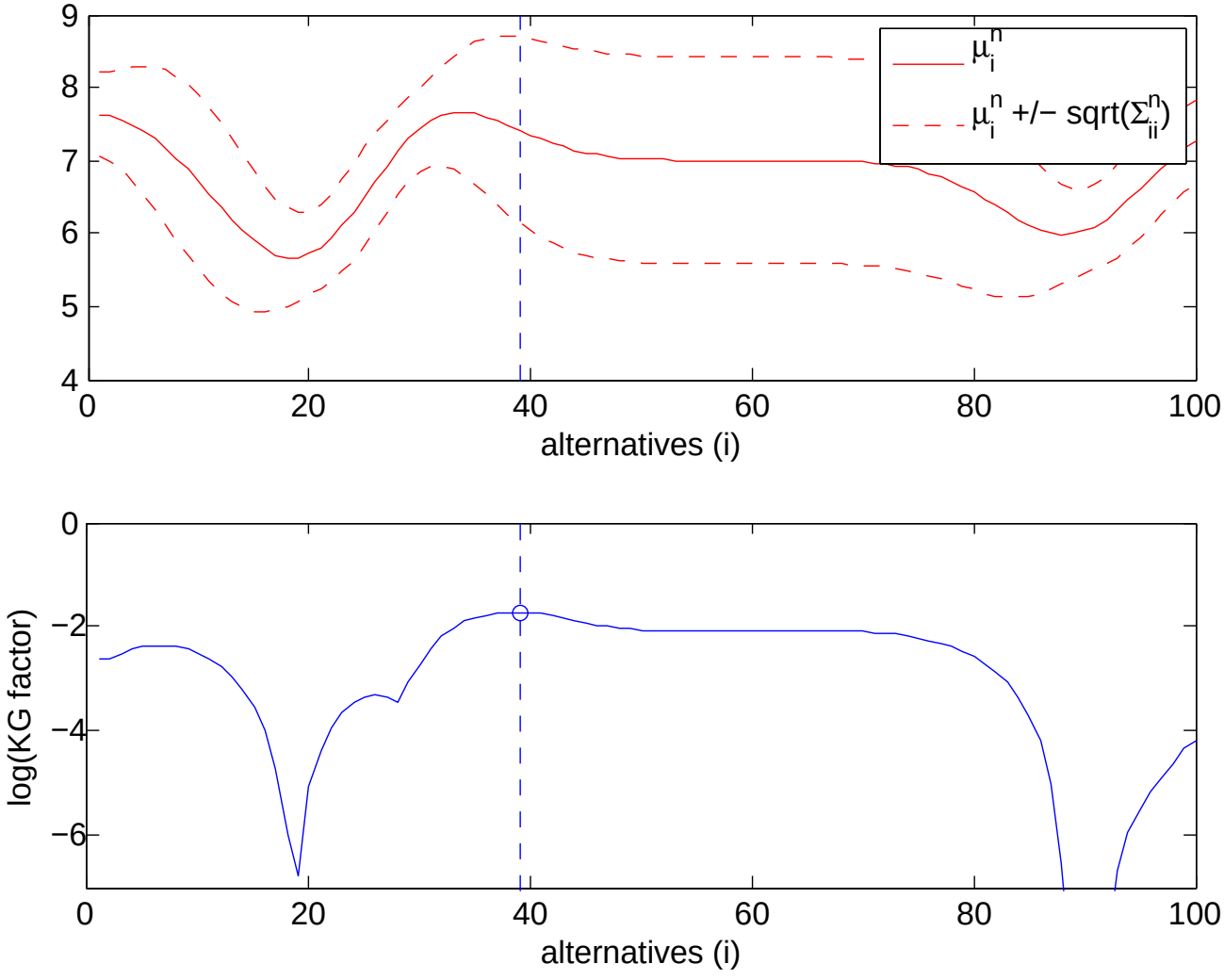
- Let A contain those alternatives that are best under the posterior with nonzero probability.
- Let $x(j)$ denote the j^{th} entry in A.
- Let $a_j = \mu_n(x(j))$ and $b_j = \sqrt{\text{Var}_n\left[\mu_{n+1}(x(j)) \mid x_n = x\right]}$.
- Sort A in order of increasing b_j.
- The KG factor is

$$\text{KG}_n(x) = \sum_{j=1}^{|A|-1} (b_{i+1} - b_i) f\left(\frac{-|a_{i+1} - a_i|}{b_{i+1} - b_i}\right),$$

  where $f(z) = \varphi(z) + z\Phi(z)$, and φ and Φ are the normal pdf and cdf.
- When A is large, or infinite, we replace A with a subset. For details, see [Scott et al., 2011].

We can compute the KG factor at each x, and plot vs. x.

The upper plot shows the posterior after n samples.
The lower plot shows the KG factor $\mathbf{KG}_n(x)$ versus x.



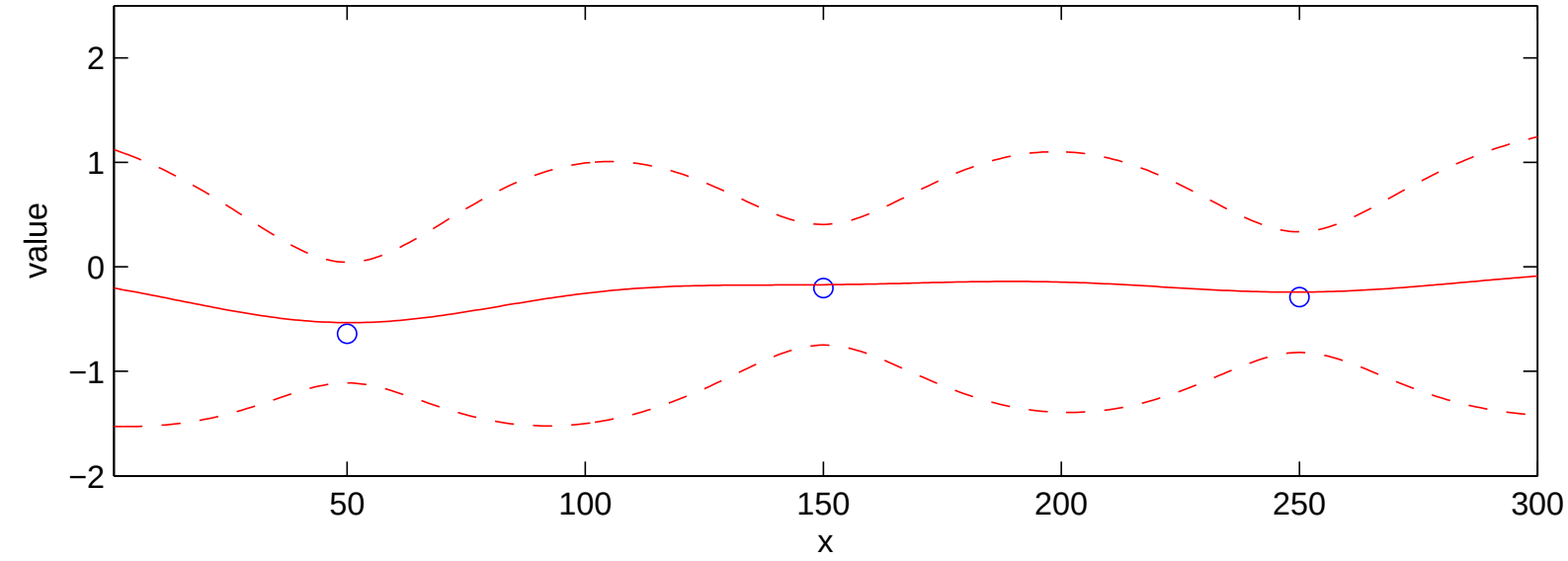
The knowledge-gradient method measures at the point with the largest knowledge-gradient factor.

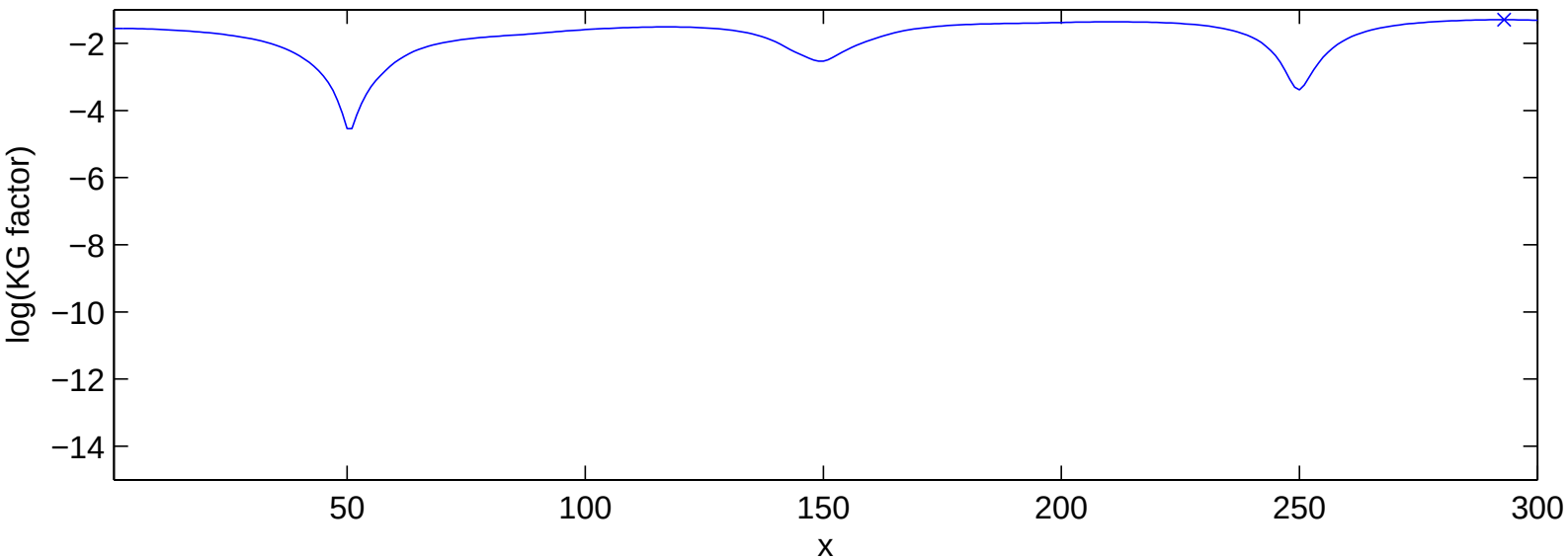
The knowledge-gradient method chooses the next point to sample via,

$$x_{n+1} \in \arg\max_x \mathrm{KG}_n(x).$$

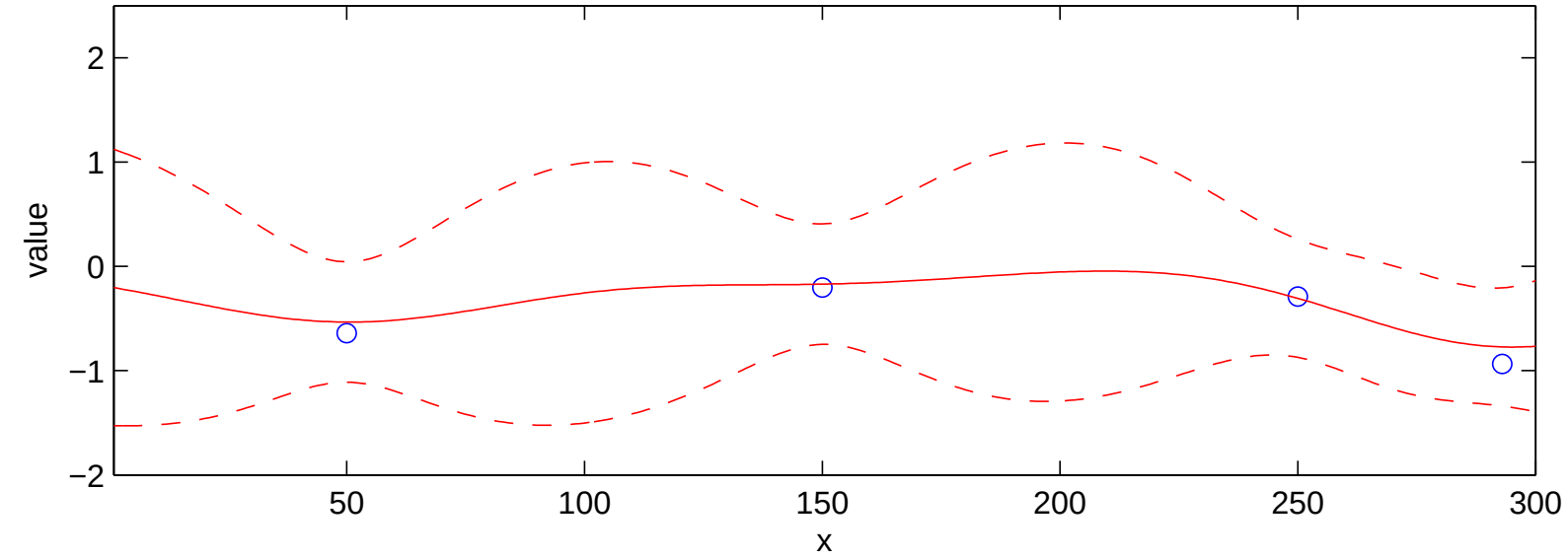
- This can be computed via enumeration if the feasible set A is not too big (1000s of points), or via nonlinear optimization (for details, see [Xie et al., 2013]).
- We have replaced an optimization problem with very expensive function evaluations (evaluation f takes minutes or hours) by one with cheap function evaluations (evaluating $\mathrm{KG}_n(x)$ takes milliseconds).

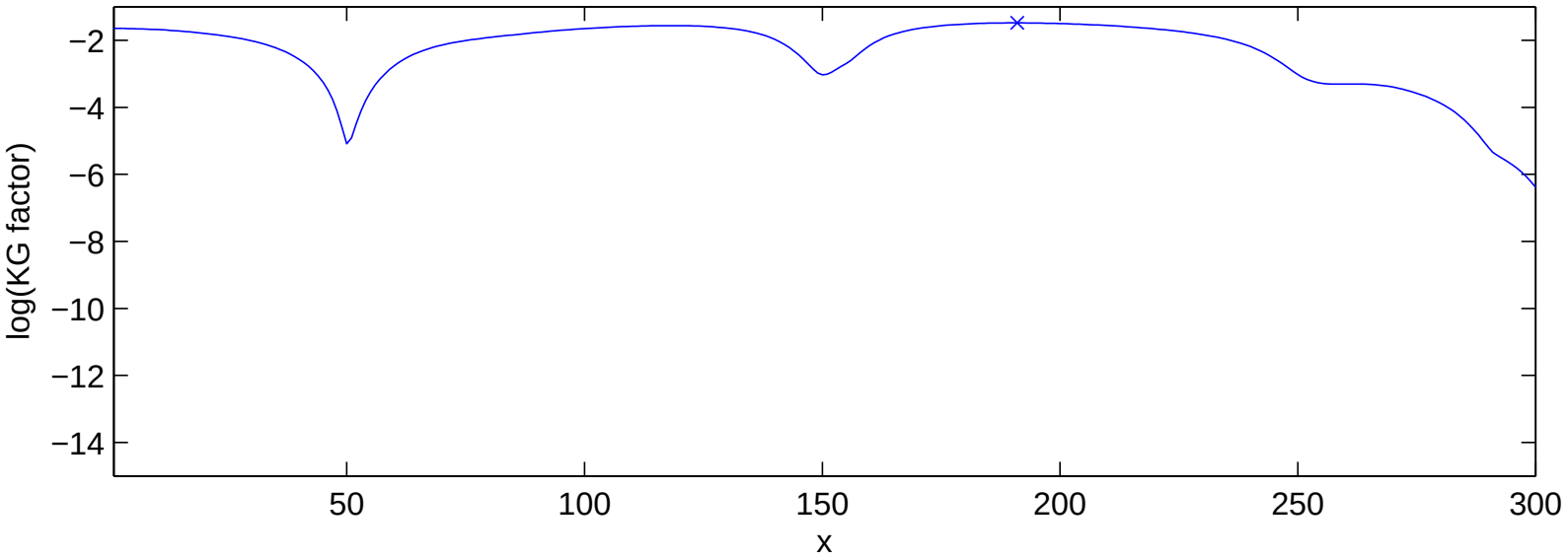
Illustrative 1D Example with Noise (KG)



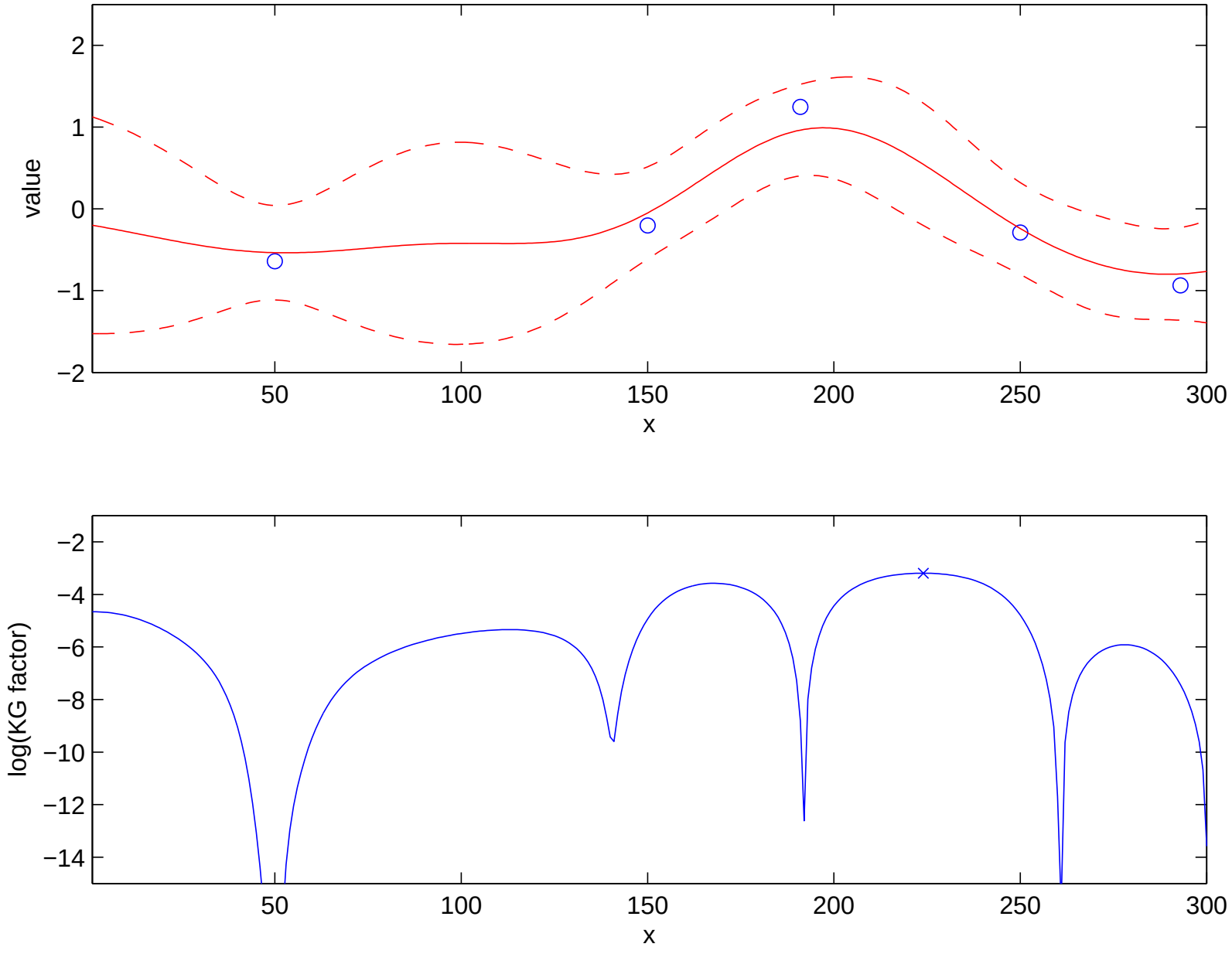


Illustrative 1D Example with Noise (KG)

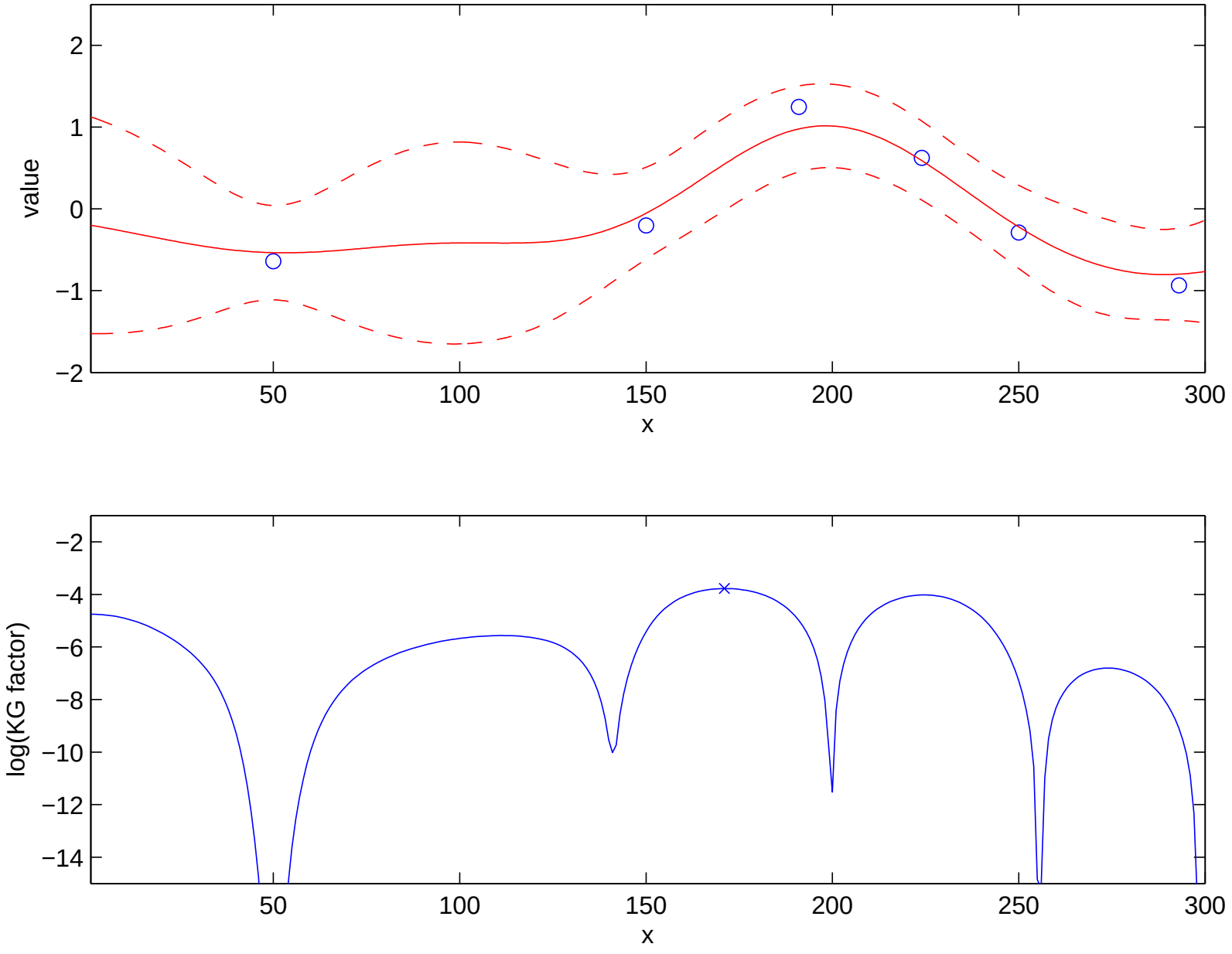




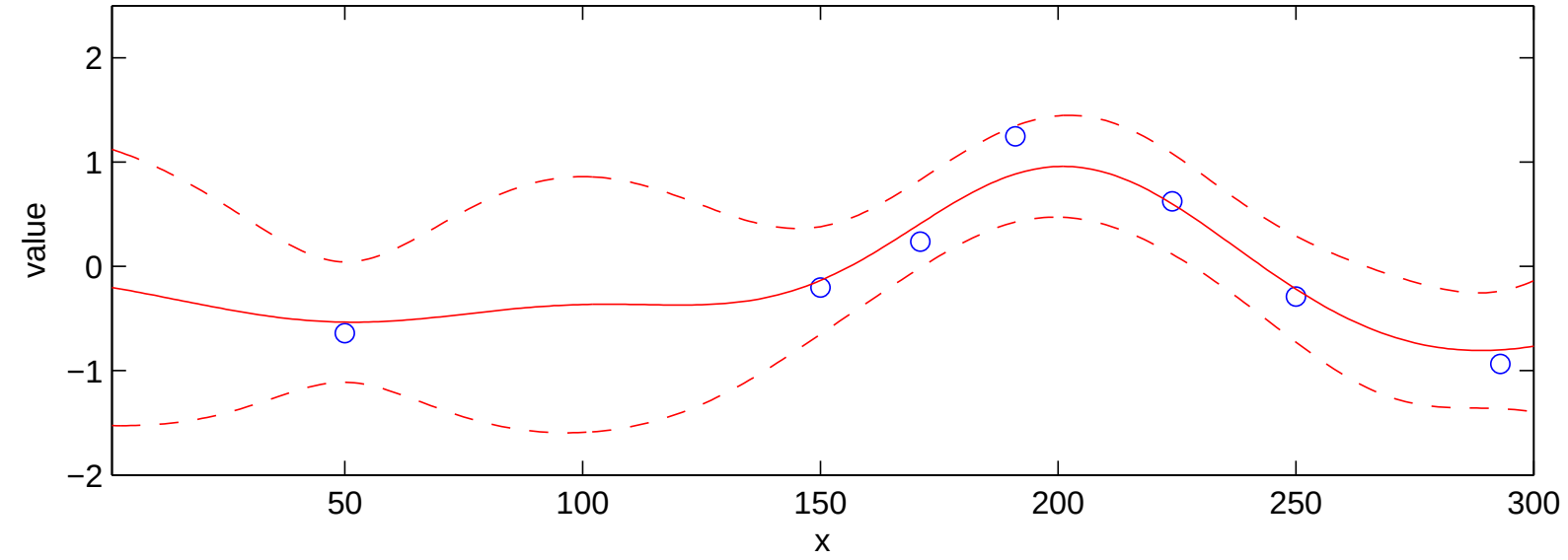
Illustrative 1D Example with Noise (KG)

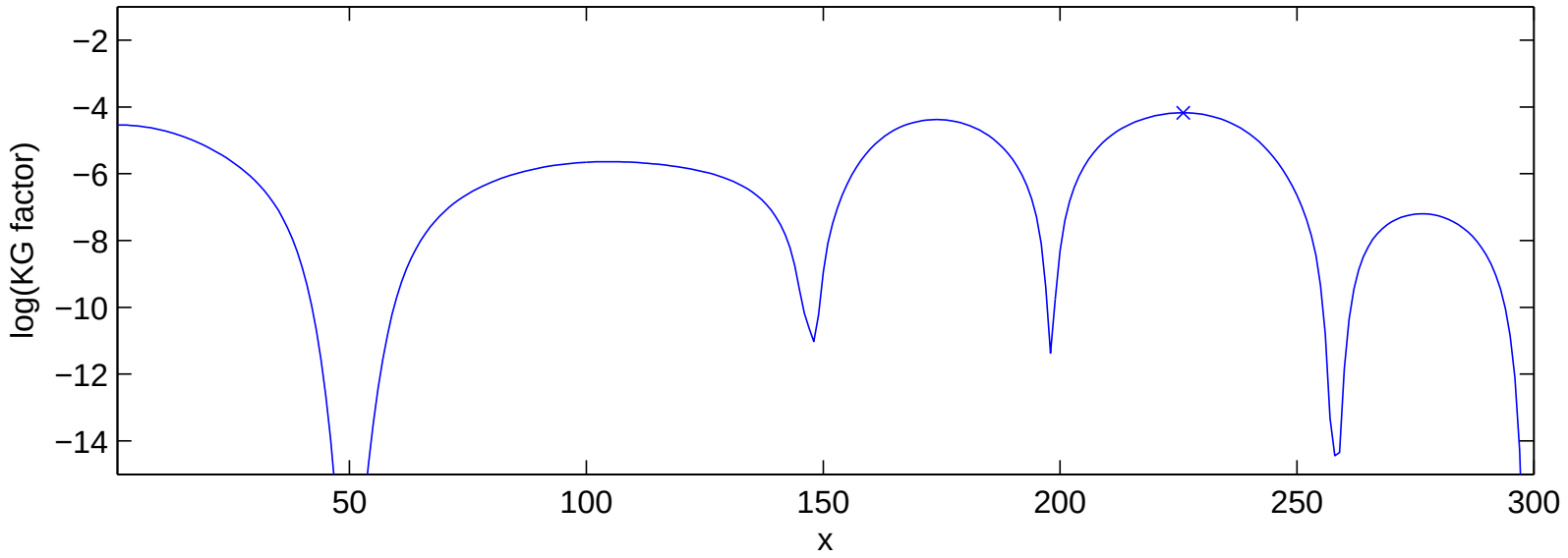


Illustrative 1D Example with Noise (KG)

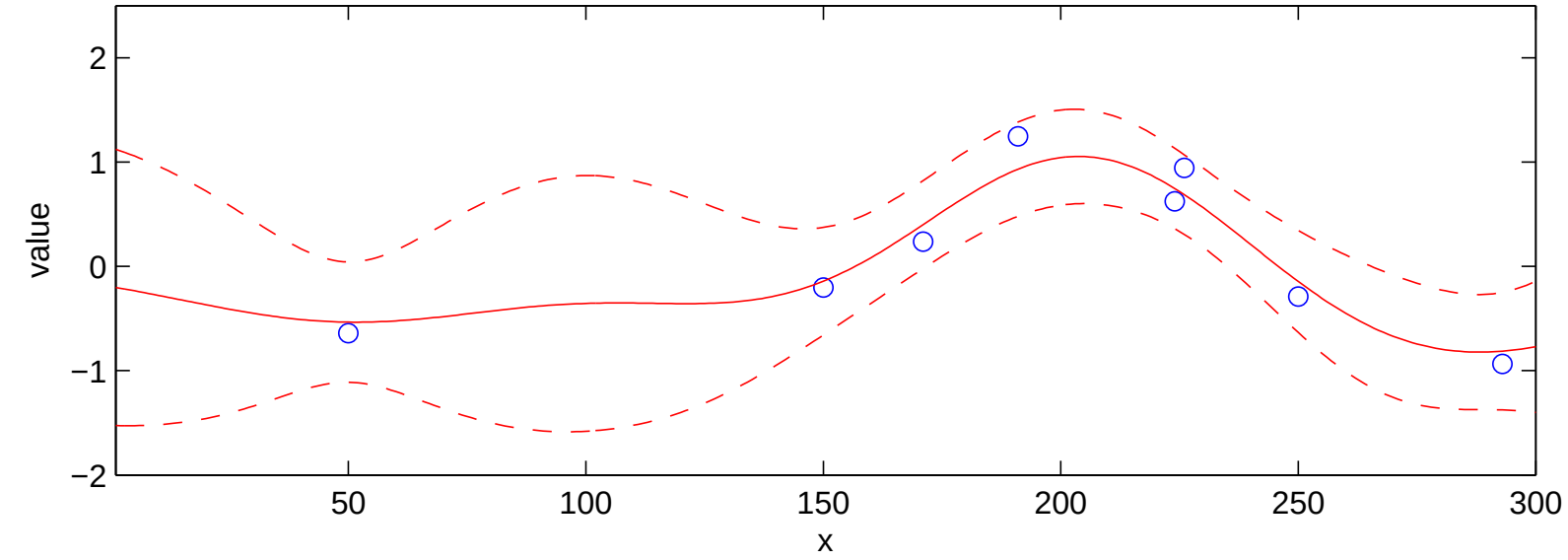


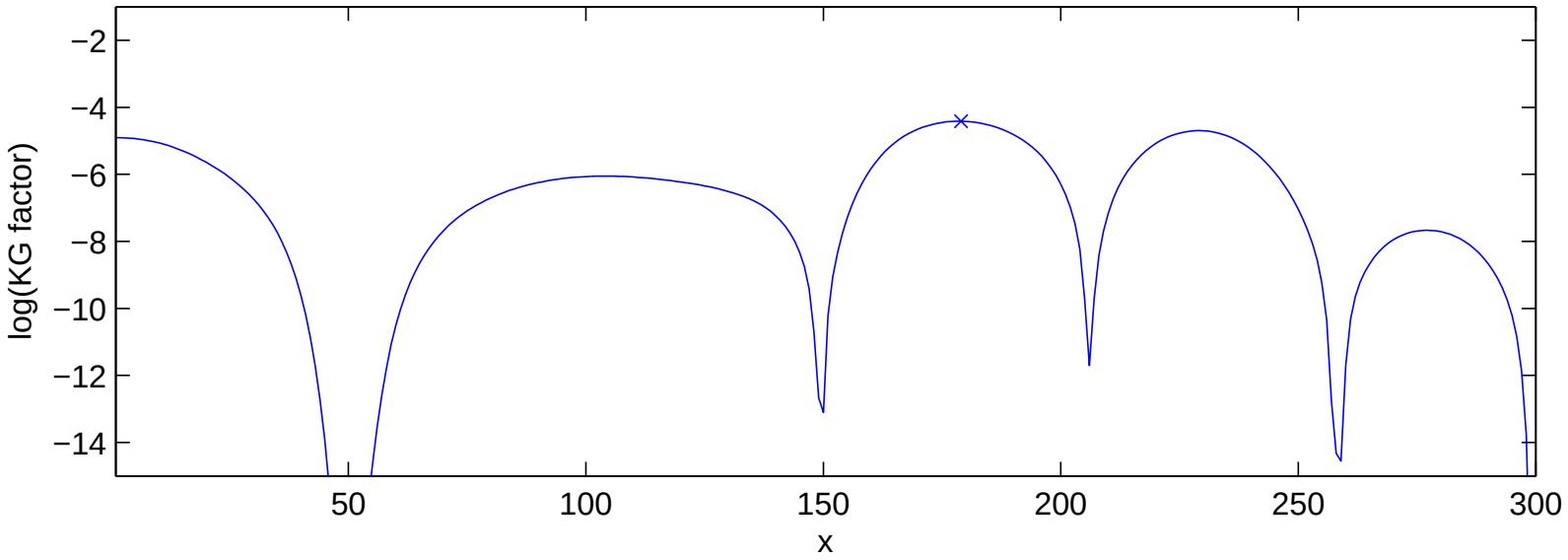
Illustrative 1D Example with Noise (KG)



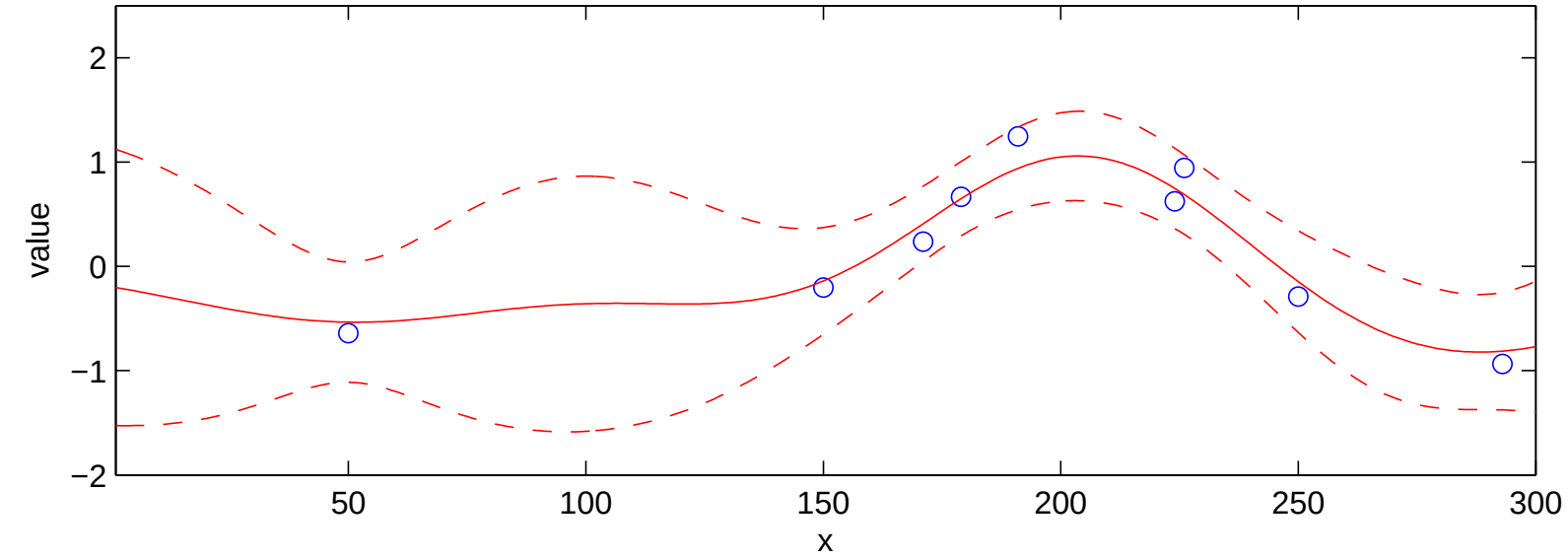


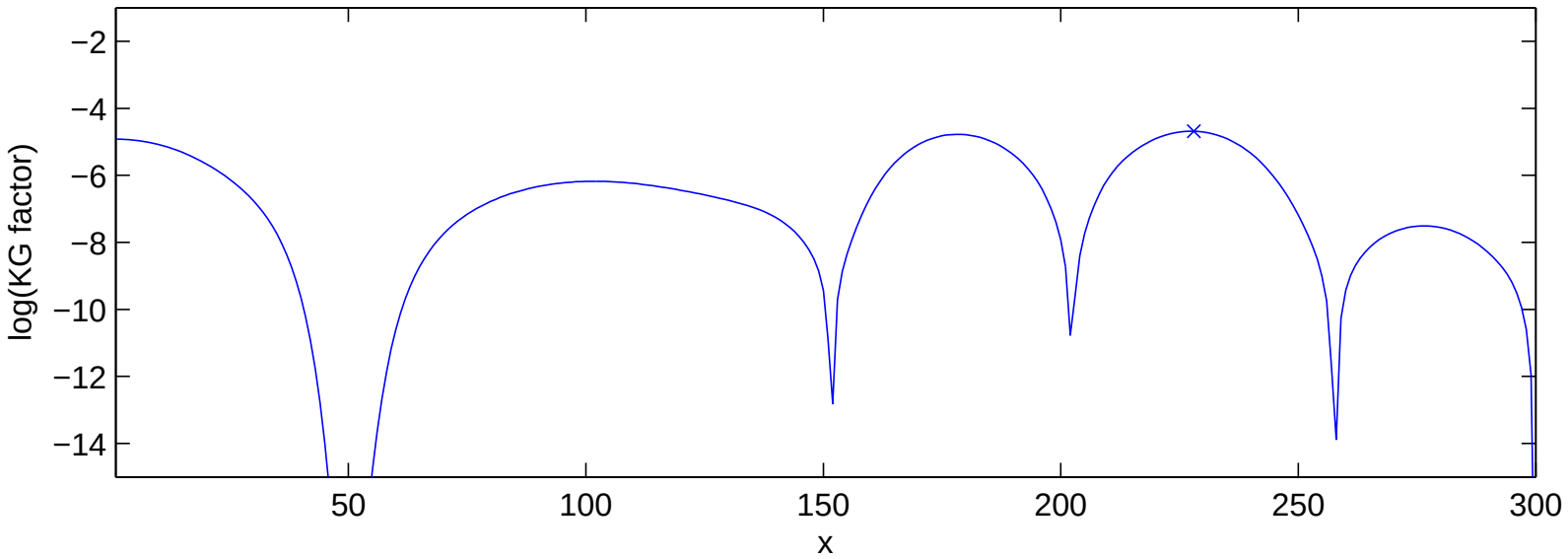
Illustrative 1D Example with Noise (KG)





Illustrative 1D Example with Noise (KG)





Optimality Results

The knowledge-gradient method is optimal when either:

- $N = 1$ (this is by construction);
- $N \to \infty$ (i.e., it is a consistent method).

The method's suboptimality is bounded in the remaining cases by

$$V^n(\mu^n, \Sigma^n) - V^{n,KG}(\mu^n, \Sigma^n) \leq \frac{1}{\sqrt{2\pi}} \max_{x^n,\ldots,x^{N-2}} \sum_{k=n+1}^{N-1} ||\tilde{\sigma}(\Sigma^k, \cdot)||$$

where V^n is the optimal value function at time n, $V^{n,KG}$ is the value of the KG method function at time n, $||u|| := \max_i u_i - \min_j u_j$ and $||\tilde{\sigma}(\Sigma, \cdot)|| := \max_x ||\tilde{\sigma}(\Sigma, x)||$,

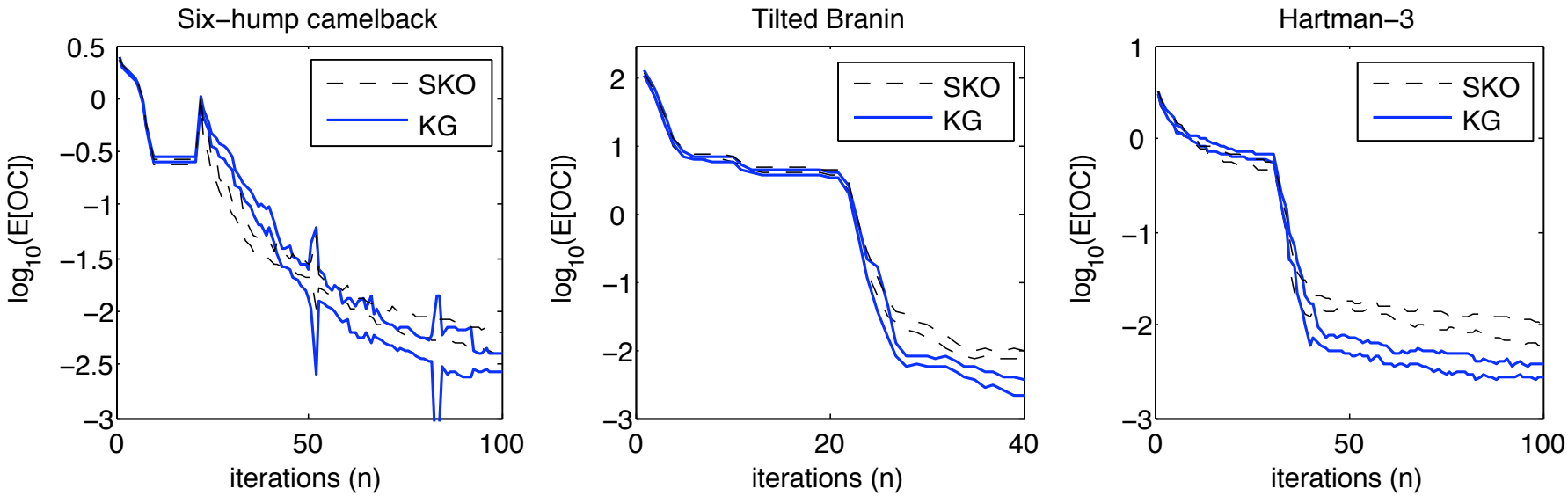
The KG method is closely related to Expected Improvement [Mockus 1989; Jones, Schonlau, Welch 1998]

If we follow the one-step analysis used by the KG method, making the following changes,

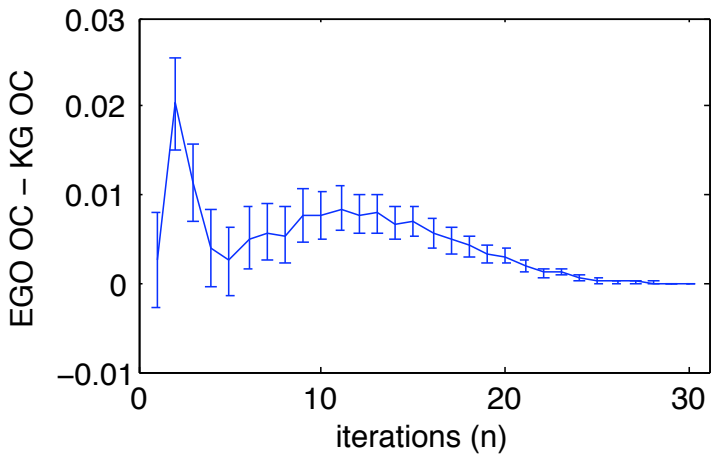
- require samples to be noise free;
- redefine $\mu_n^* = \max_{m=1,\ldots,n} \mu_n(x_m)$, so that our solution must be at a previously evaluated point;
- and use a continuous domain;

then we recover the expected improvement method of [Mockus, 1989, Jones et al., 1998].

KG improves over other BGO methods in both noisy and noise-free problems



- Upper three graphs compares KG and SKO [Huang et al., 2006] on problems with independent noise.
- Lower right graph compares KG and EI on noise-free problems.



There are many other methods use the posterior to decide where to sample

Here are a few methods that use the posterior to decide where to sample.

- Ranking and Selection:
  [Gupta and Miescke, 1996, Chick and Inoue, 2001, Fu et al., 2004].
- Noise-free continuous global optimization:
  [Kushner, 1964, Mockus et al., 1978, Stuckman, 1988,
  Jones et al., 1998, Calvin and Zilinskas, 2002,
  Calvin and Zilinskas, 2005].
- Noisy continuous global optimization:
  [Huang et al., 2006, Forrester et al., 2006, Taddy et al., 2009,
  Villemonteix et al., 2009, Kleijnen et al., 2011, Scott et al., 2011].

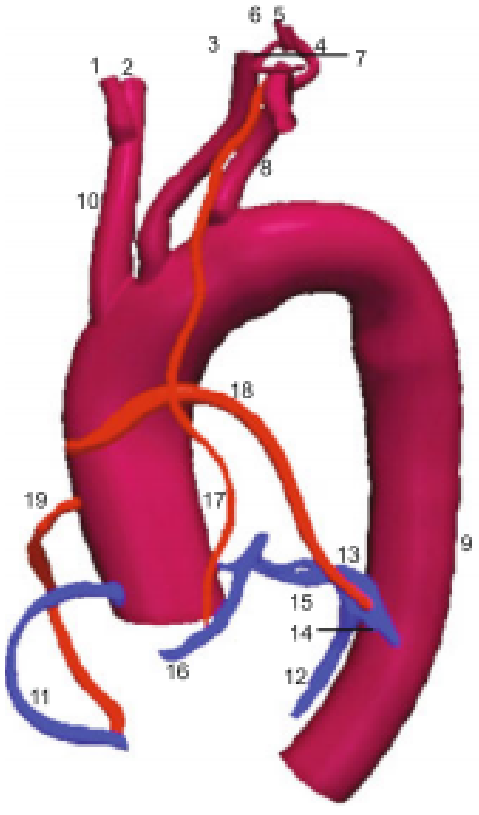
Outline

1. Background: Gaussian Process Regression
2. The Knowledge Gradient (KG) Method [F., Powell, Dayanik 2009]
3. Applications and Variations
   - Design of Cardiovascular Bypass Grafts [Xie, F., Sankaran, Marsden, Elmohamed 2012]
   - Simulation Calibration at Schneider National [F., Powell, Simao 2009]
   - Improving Customer Experience and Revenue at Yelp
   - Exploiting Common Random Numbers [Xie, F., Chick 2013]
   - Developing Inexpensive Organic Photovoltaic Cells
   - Drug Development for Ewing's Sarcoma [Negoescu, F., Powell 2011]
4. Conclusion

Outline

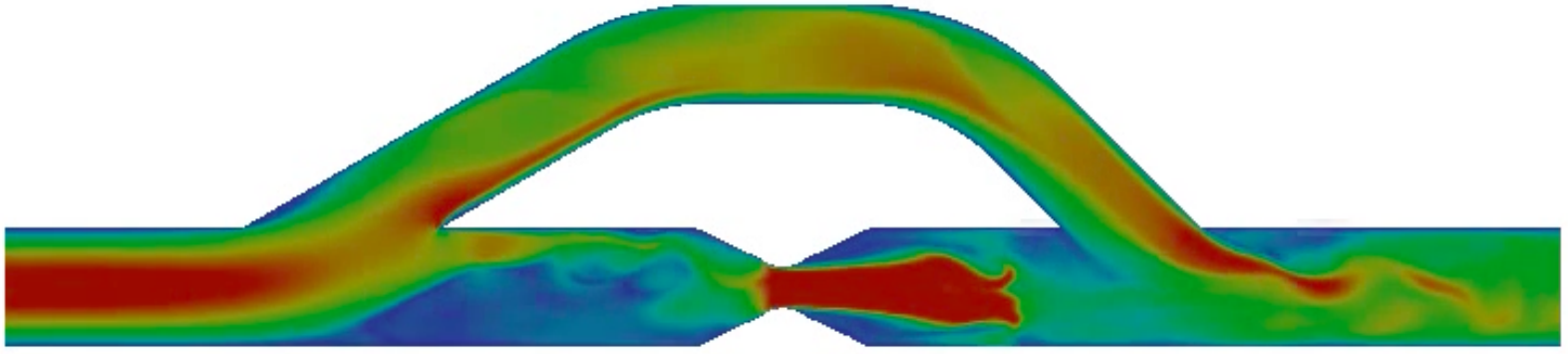
1. Background: Gaussian Process Regression
2. The Knowledge Gradient (KG) Method [F., Powell, Dayanik 2009]
3. Applications and Variations
   - Design of Cardiovascular Bypass Grafts [Xie, F., Sankaran, Marsden, Elmohamed 2012]
   - Simulation Calibration at Schneider National [F., Powell, Simao 2009]
   - Improving Customer Experience and Revenue at Yelp
   - Exploiting Common Random Numbers [Xie, F., Chick 2013]
   - Developing Inexpensive Organic Photovoltaic Cells
   - Drug Development for Ewing's Sarcoma [Negoescu, F., Powell 2011]
4. Conclusion

Design of Cardiovascular Bypass Grafts



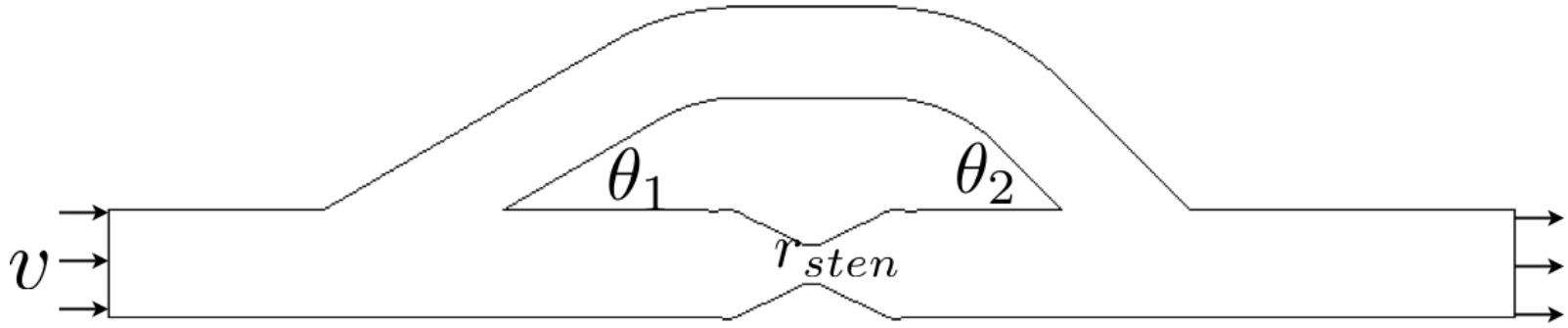
Joint work with Alison Marsden, Sethuraman Sankaran (UCSD, Mechanical and Aerospace Engineering), Jing Xie, Saleh Elmohamed (Cornell). [Xie et al., 2012]

Design of Cardiovascular Bypass Grafts



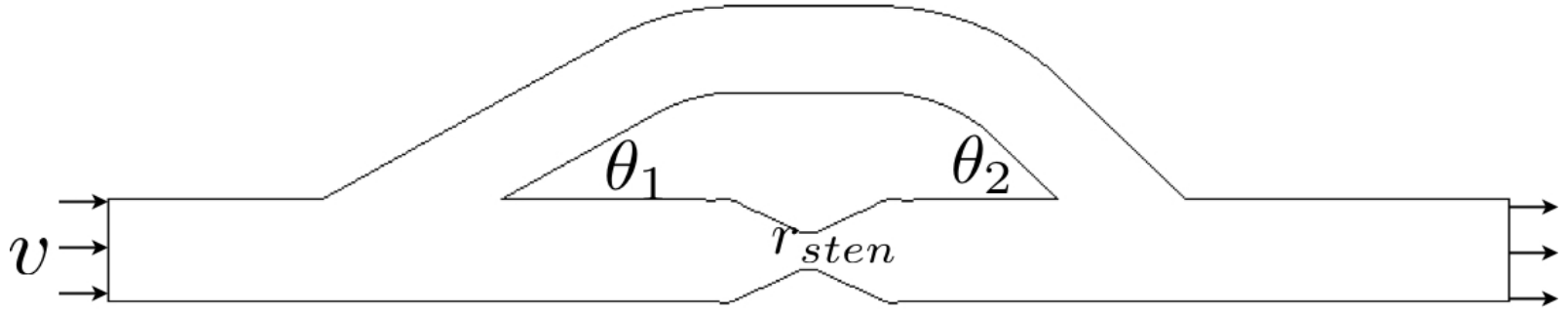
- We work with an idealized model of a cardiovascular bypass graft.
- Our goal is to choose the attachment angles to minimize the area of low wall-shear stress, subject to uncertainty about graft implementation, and environmental conditions within the body.
- To evaluate the area of low wall-shear stress for a particular set of implemented angles, and a particular set of environmental conditions, we have a fluid-flow simulation.

Design of Cardiovascular Bypass Grafts



- Target attachment angles are given to the surgeon $x = (x_1, x_2)$.
- Actual attachment angles constructed in surgery are $\theta = (\theta_1, \theta_2) = x + \delta$, where $\delta = (\delta_1, \delta_2)$ are the implementation errors introduced during surgery.
- Stenosis radius r and blood inflow velocity v are environmental variables.

Design of Cardiovascular Bypass Grafts



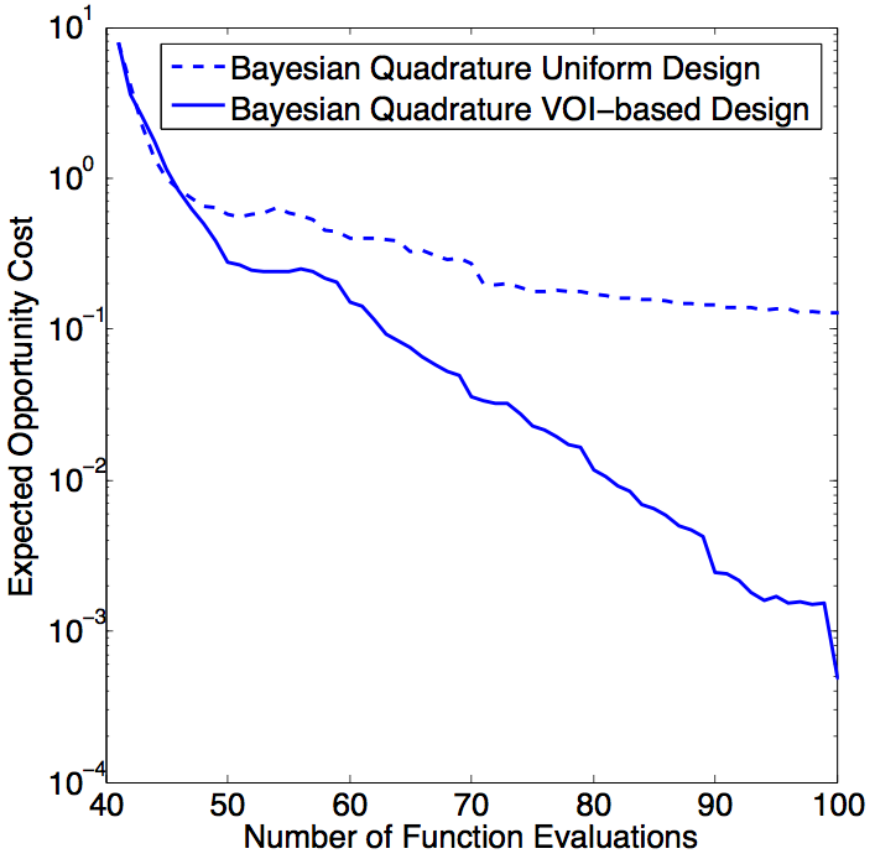
- The area of low wall-shear stress (WSS) given θ and $\omega = (r, v)$ is $f(\theta, \omega)$. f can be evaluated exactly through expensive simulation.
- The joint probability density of (δ, ω) is assumed known, and is denoted $p(\delta, \omega)$.
- Our goal is to find x_1 and x_2 to minimize the average area of low WSS,

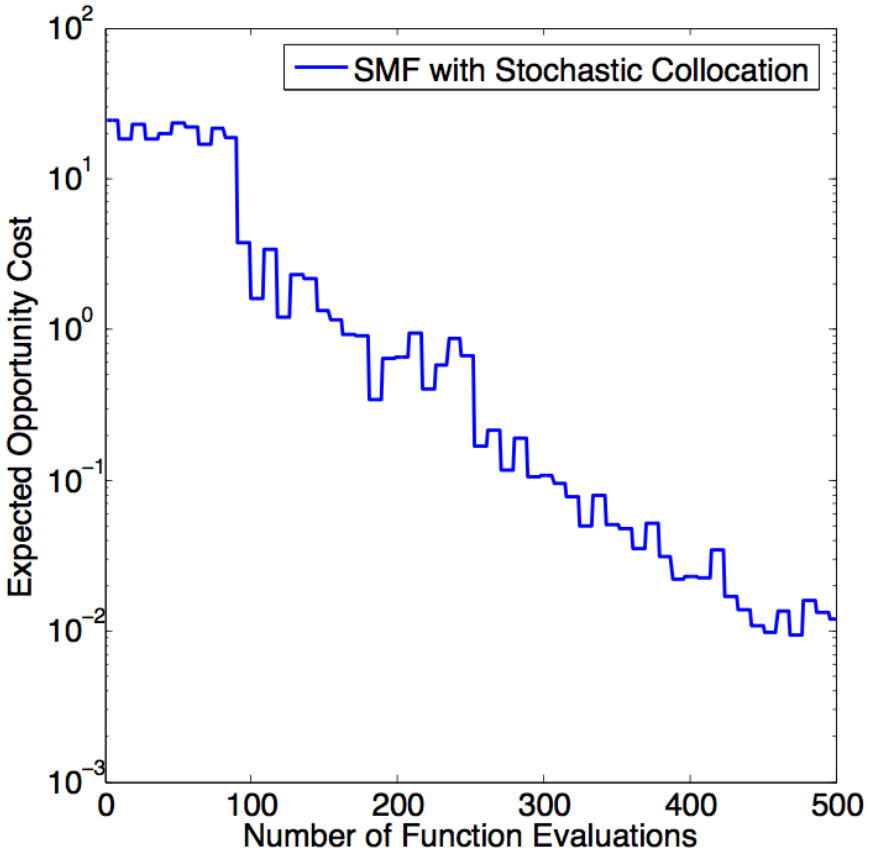
$$\min_{x_1, x_2} \int p(\delta, \omega) f(x + \delta, \omega)\, d\delta\, d\omega.$$

Design of Cardiovascular Bypass Grafts

- The algorithm chooses not just which x to evaluate, but also which δ and ω.
- The problem setting is a bit different from typical simulation optimization problems, but we still use value of information calculations to decide where to sample next.
- We compare against the method from [Sankaran and Marsden, 2011], which uses stochastic collocation within the surrogate management framework, and which was designed for problems of this type.
- We compare on a test problem which is faster than the true fluid-flow simulation to run. Comparison on the fluid flow simulation is in process.

Design of Cardiovascular Bypass Grafts



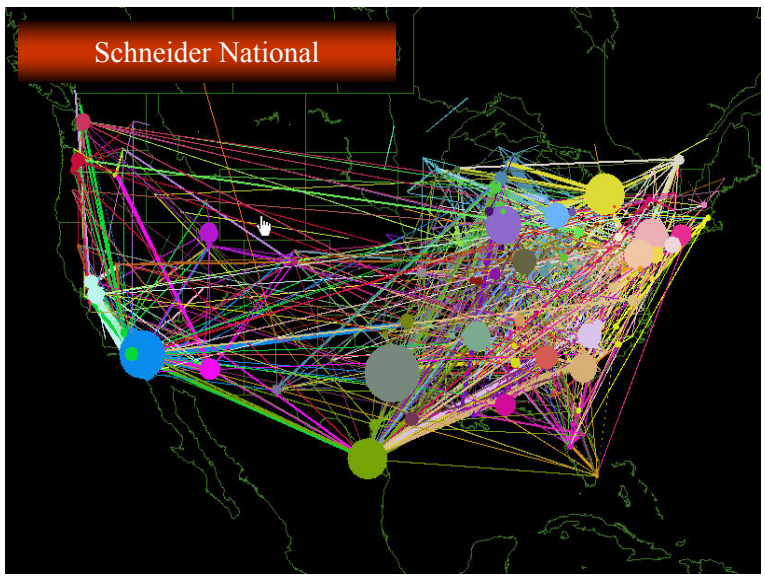


Outline

1. Background: Gaussian Process Regression
2. The Knowledge Gradient (KG) Method [F., Powell, Dayanik 2009]
3. Applications and Variations
   - Design of Cardiovascular Bypass Grafts [Xie, F., Sankaran, Marsden, Elmohamed 2012]
   - Simulation Calibration at Schneider National [F., Powell, Simao 2009]
   - Improving Customer Experience and Revenue at Yelp
   - Exploiting Common Random Numbers [Xie, F., Chick 2013]
   - Developing Inexpensive Organic Photovoltaic Cells
   - Drug Development for Ewing's Sarcoma [Negoescu, F., Powell 2011]
4. Conclusion

Simulation Model Calibration at Schneider National

- The logistics company Schneider National uses a large simulation-based optimization model to try "what if" scenarios.
- The model has several input parameters that must be tuned to make its behavior match reality before it can be used.
- The model is tuned by hand once per year on the most recent data. Each tuning effort requires between 1 and 2 weeks.



(Joint work with Warren B. Powell and Hugo Simão, Princeton University, [Frazier et al., 2009a])

Model Parameters

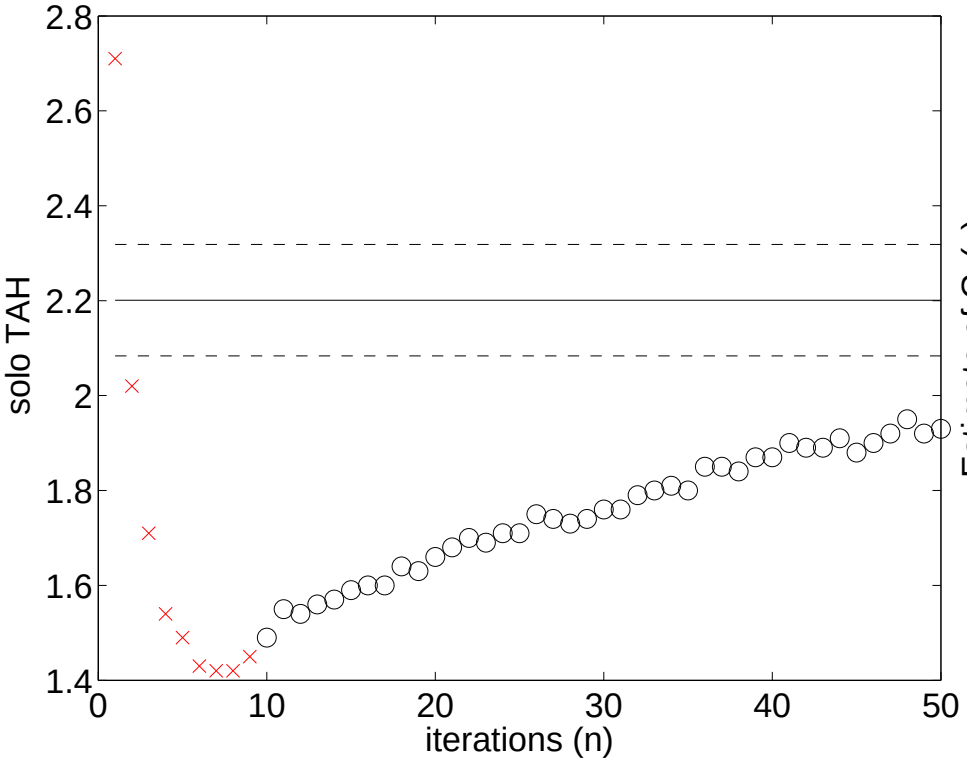
- Input parameters to the model include:
  - time-at-home bonuses.
  - "pacing" parameters describing how fast and far drivers drive per day.
  - gas prices
  - ...
- Output parameters from the model include:
  - billed miles
  - driver utilization
  - average number of trips home per driver per 4 weeks.
  - proportion of drivers without time at home over 4 weeks.
  - ...
- Some of these inputs are known (e.g., gas prices), but some are unknown (e.g. time-at-home bonuses).
- Goal: adjust the inputs to make the optimal solution found by the model match current practice.

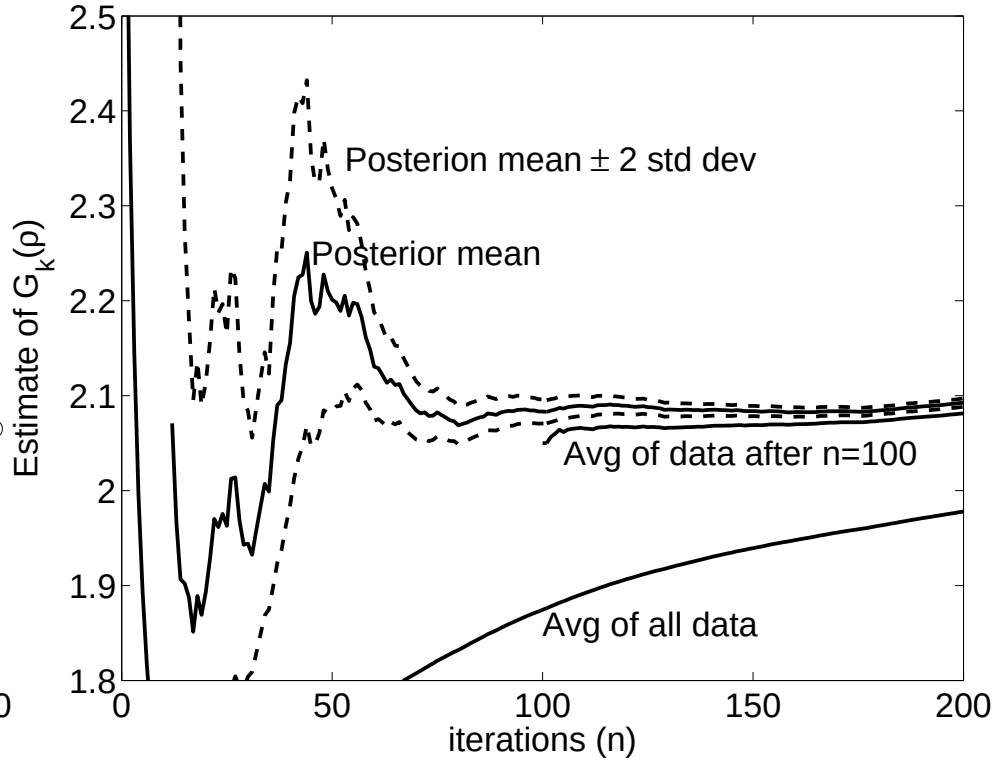
Simulation Model Calibration

- Goal: adjust the inputs to make the optimal solution found by the ADP model match current practice.
  - x is a set of inputs to the simulator.
  - $f(x)$ is how closely the simulator output matches history.
- Running the simulator for one set of bonuses takes 3 days, making calibration difficult.
- The model may be run for shorter periods of time, e.g. 12 hours, to obtain noisy output estimates.

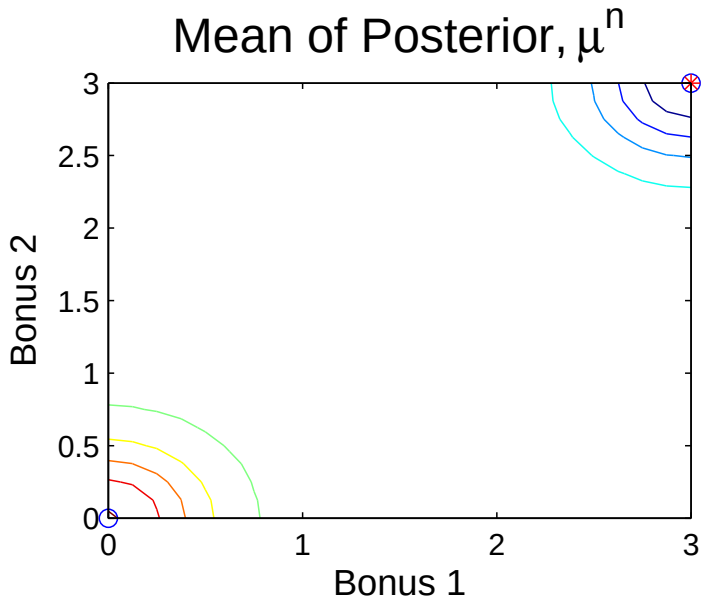
BGO is Flexible Enough to Handle Non-stationary Output

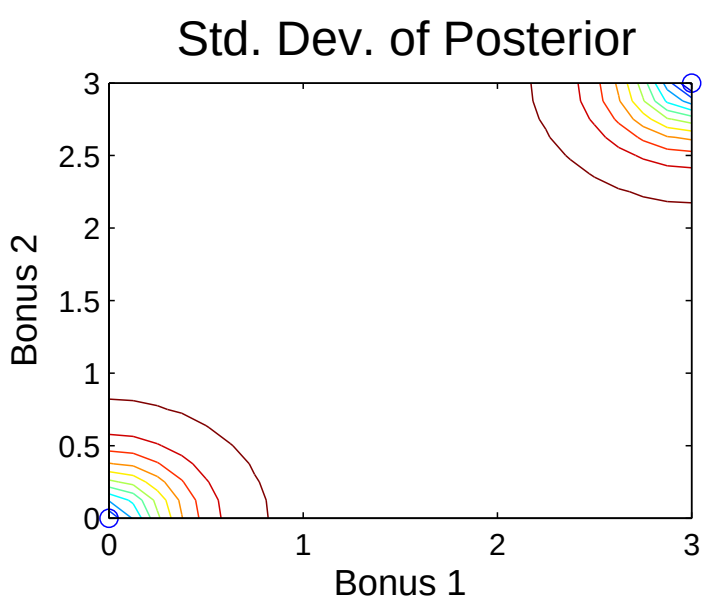
- The output of the simulator is non-stationary.
- Running the simulator to convergence takes too long (3 days).
- With just 12 hours of samples, we can use Bayesian statistics to get a noisy estimate of where the path is going.

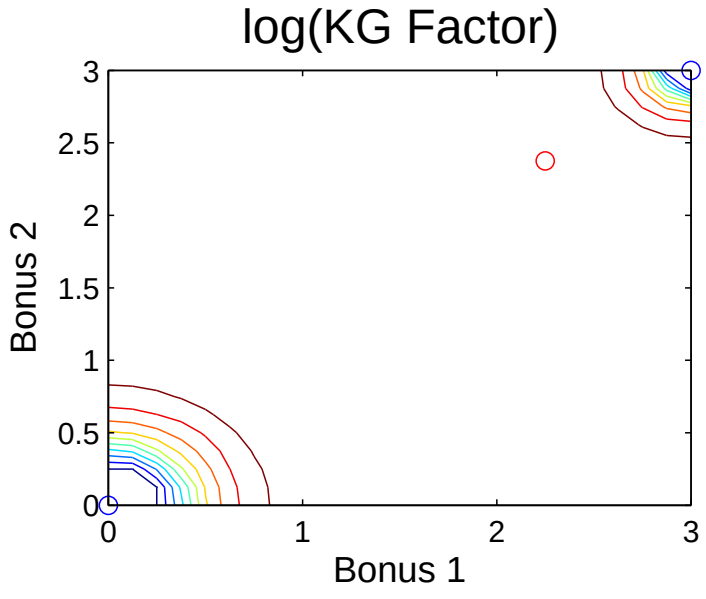


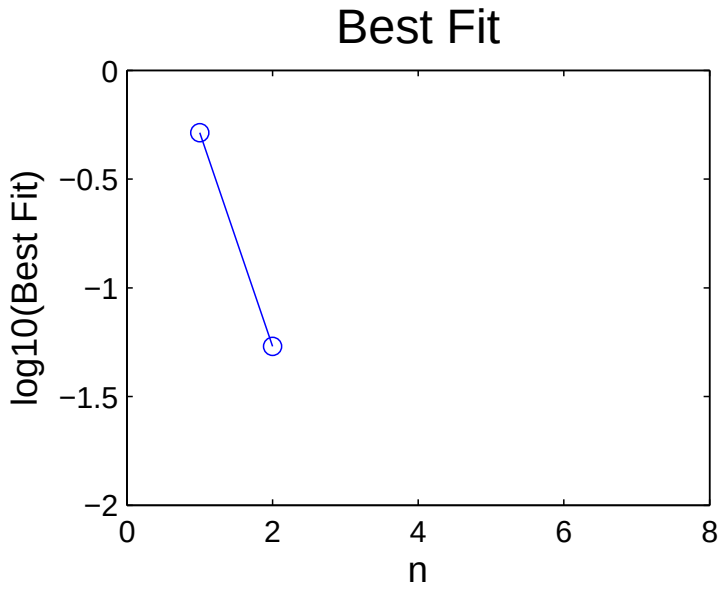


Simulation Model Calibration Results

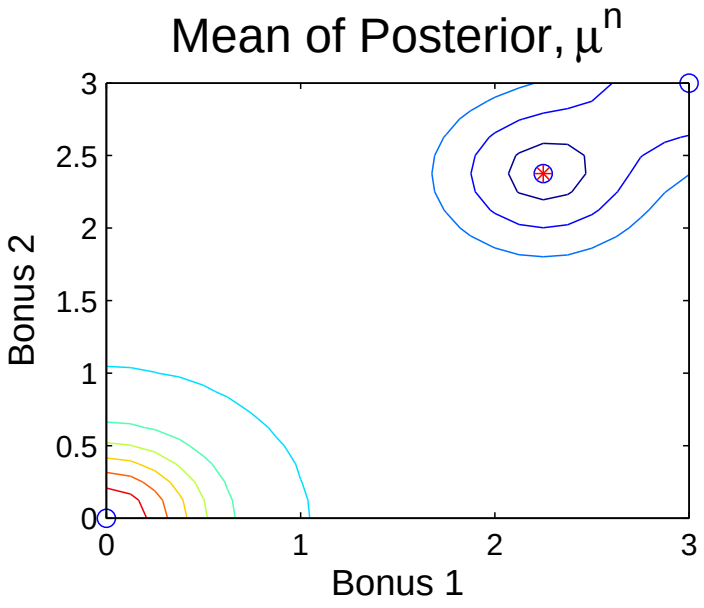


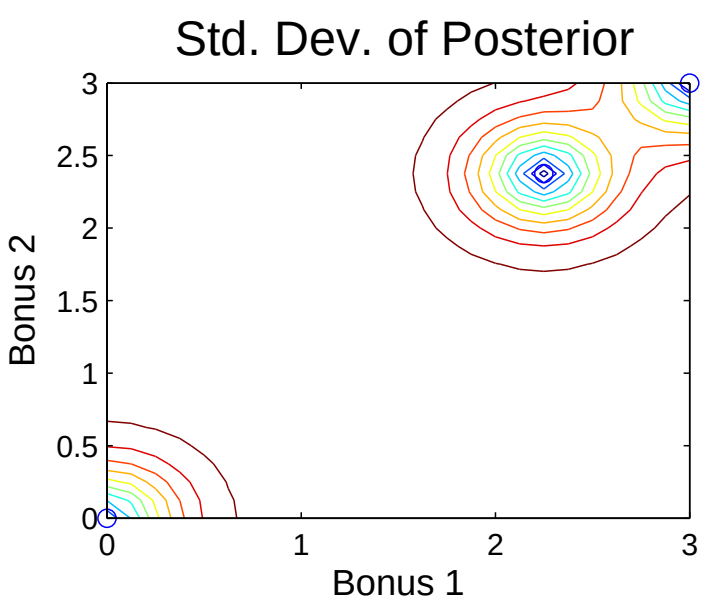


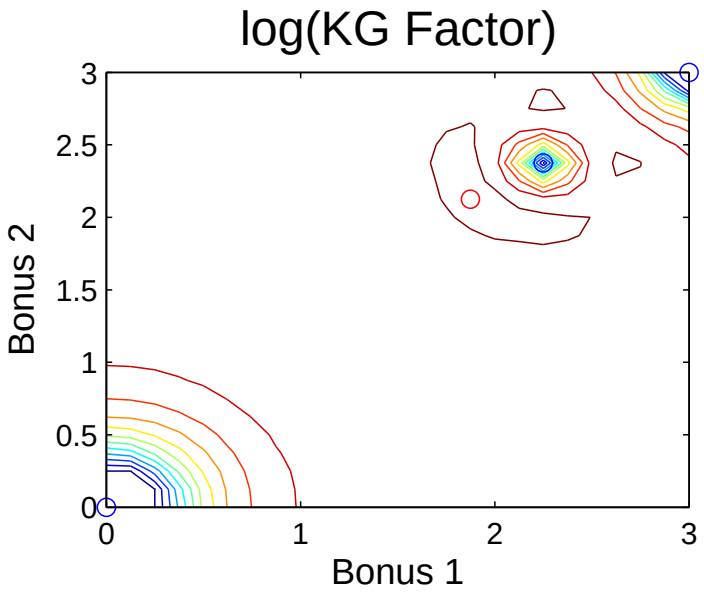


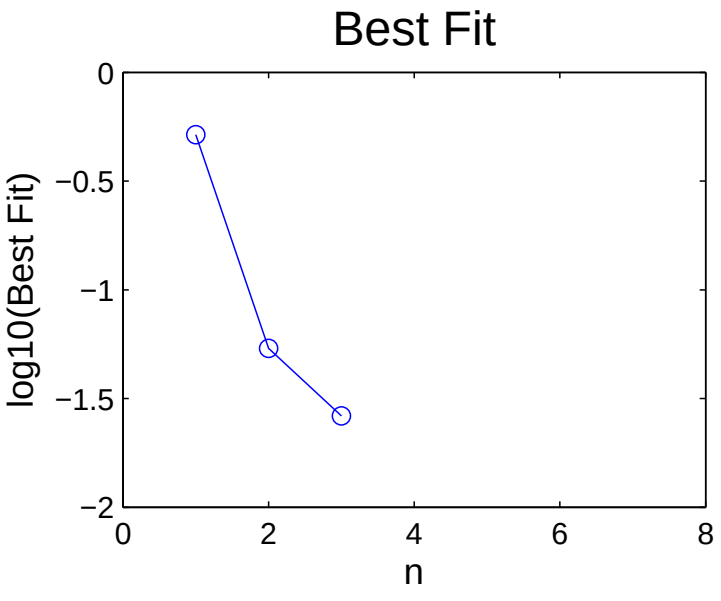


Simulation Model Calibration Results

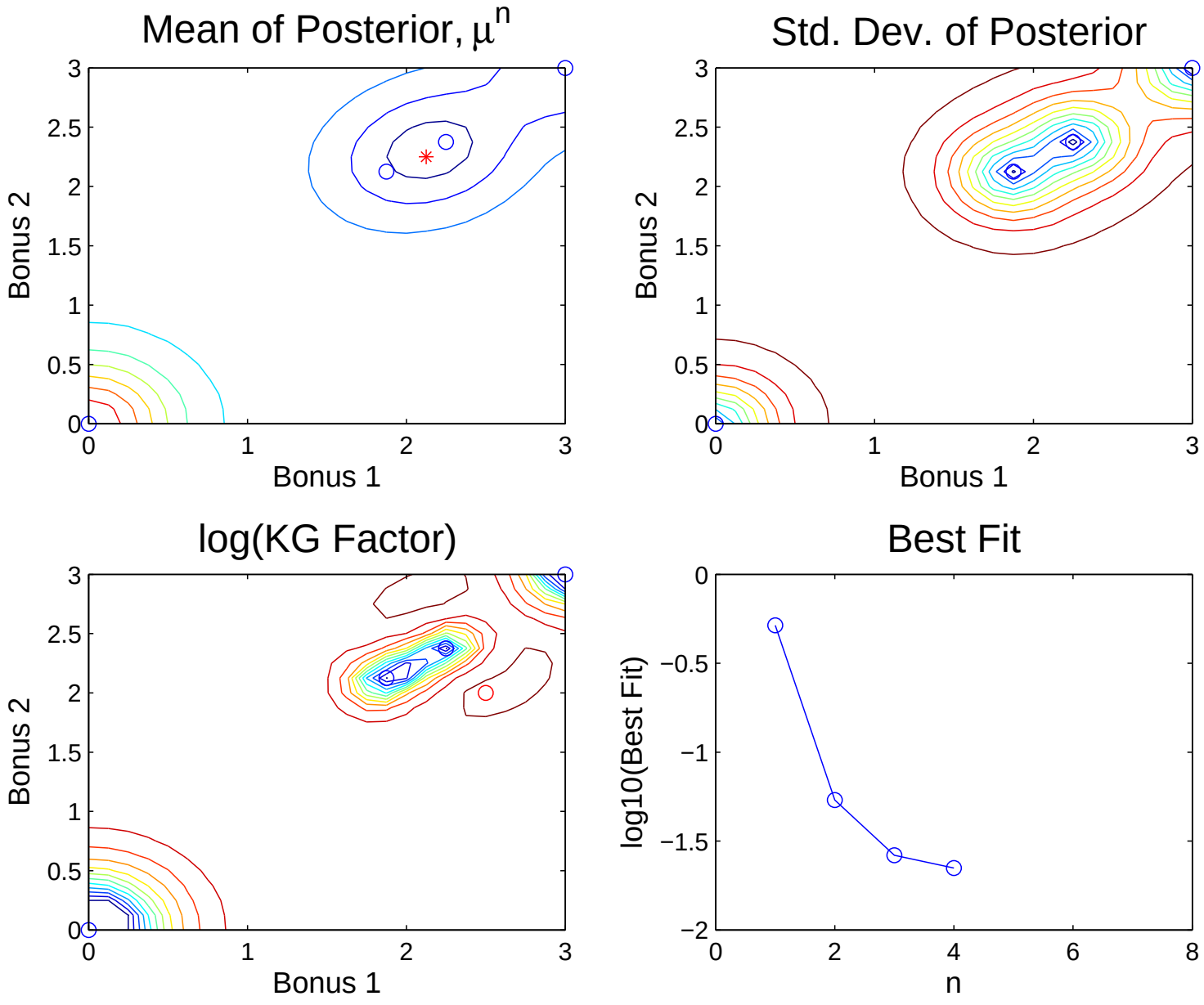




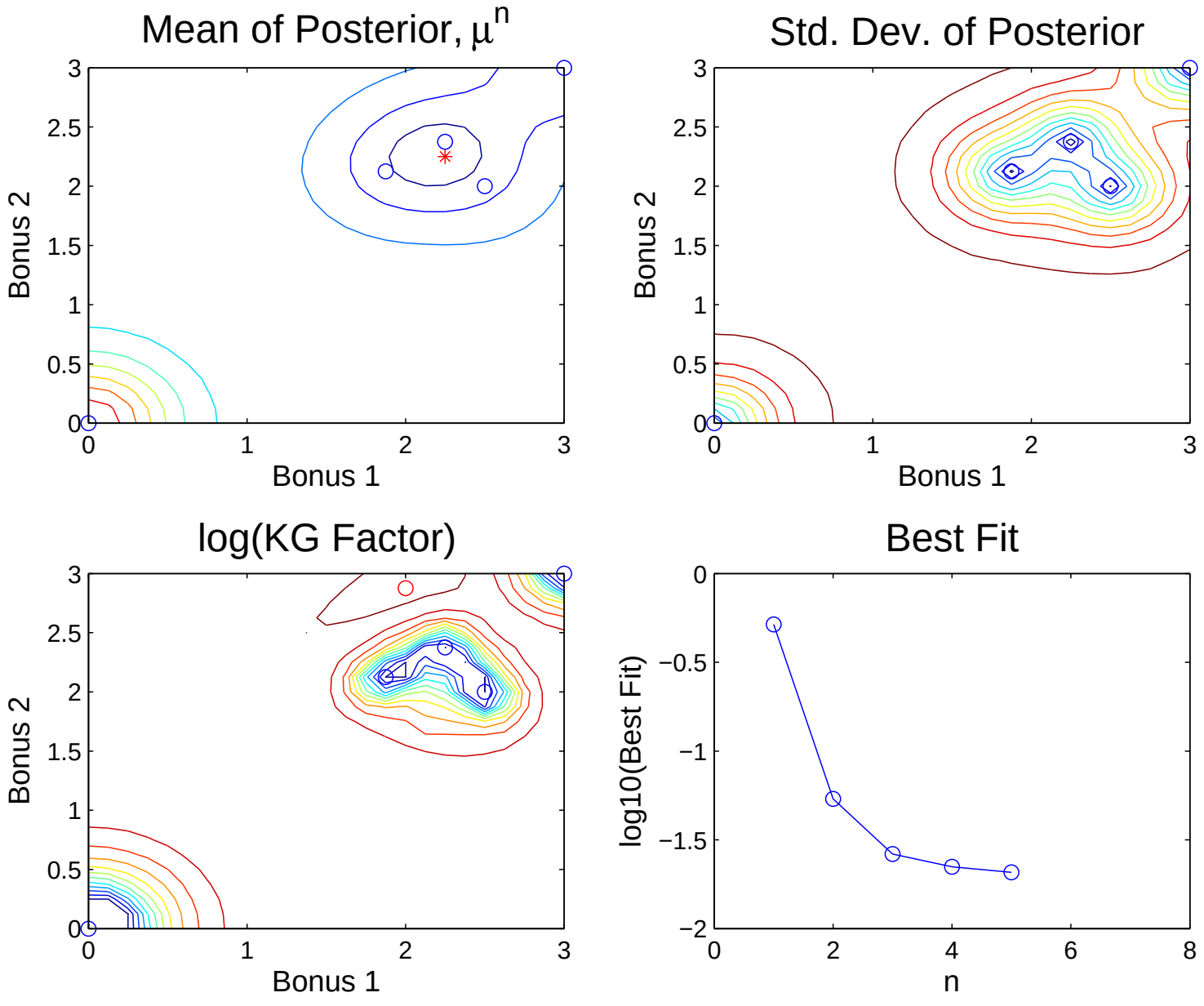




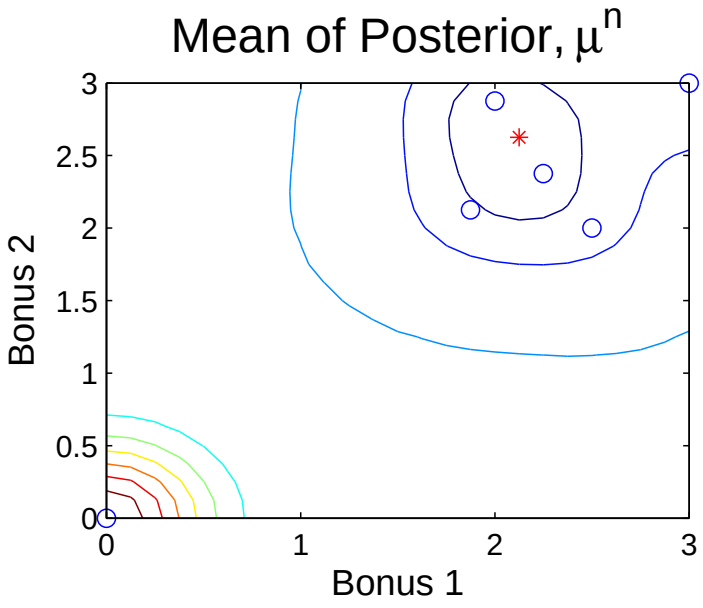
Simulation Model Calibration Results

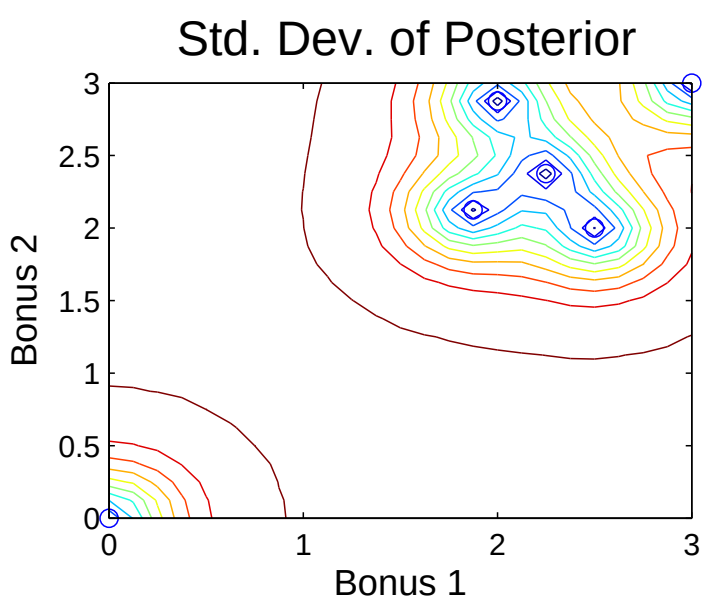


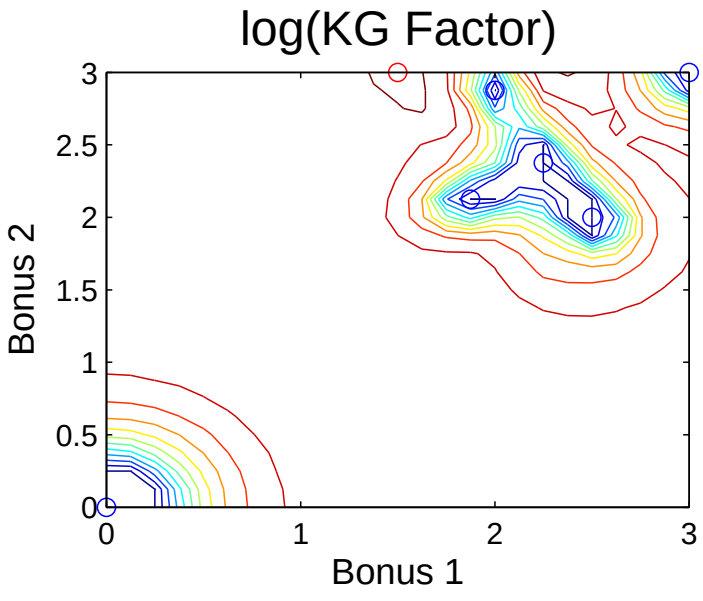
Simulation Model Calibration Results

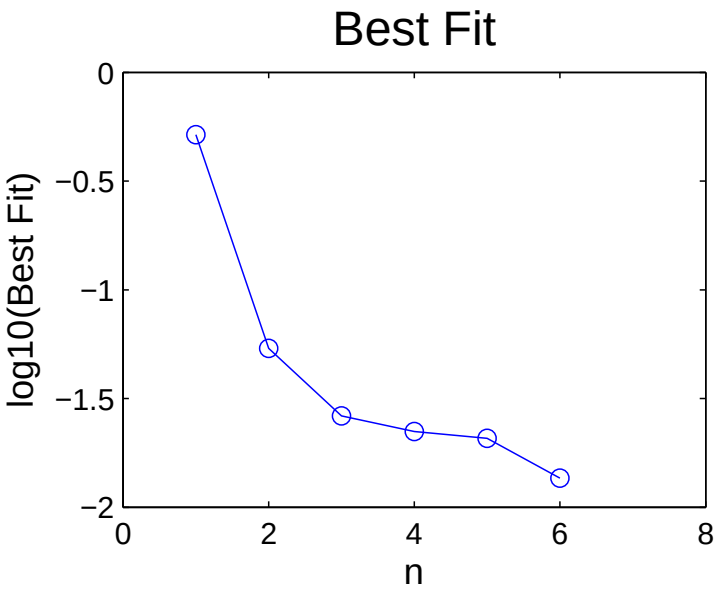


Simulation Model Calibration Results

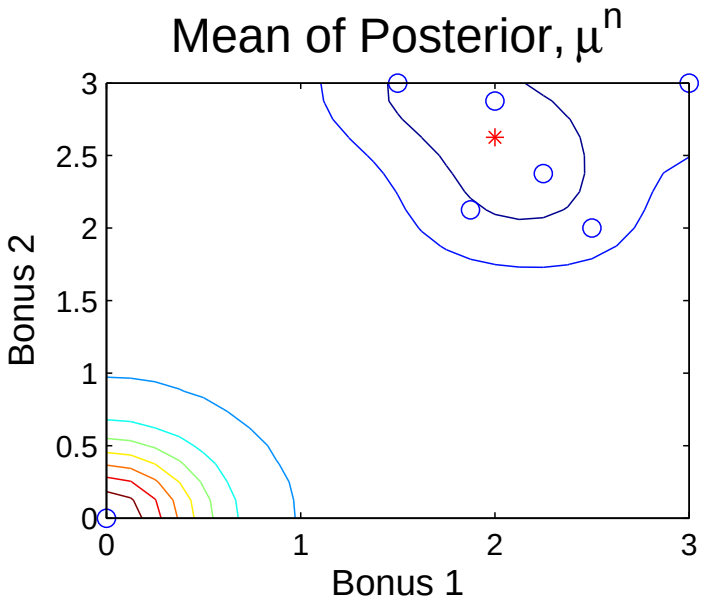


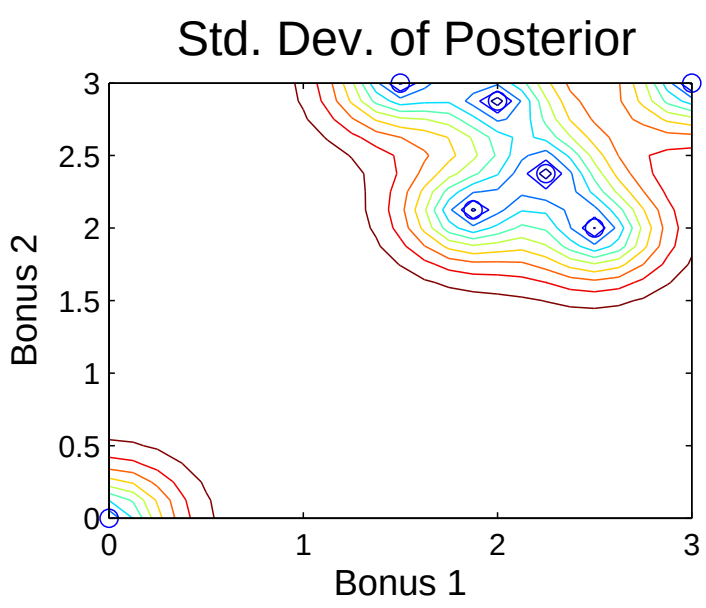


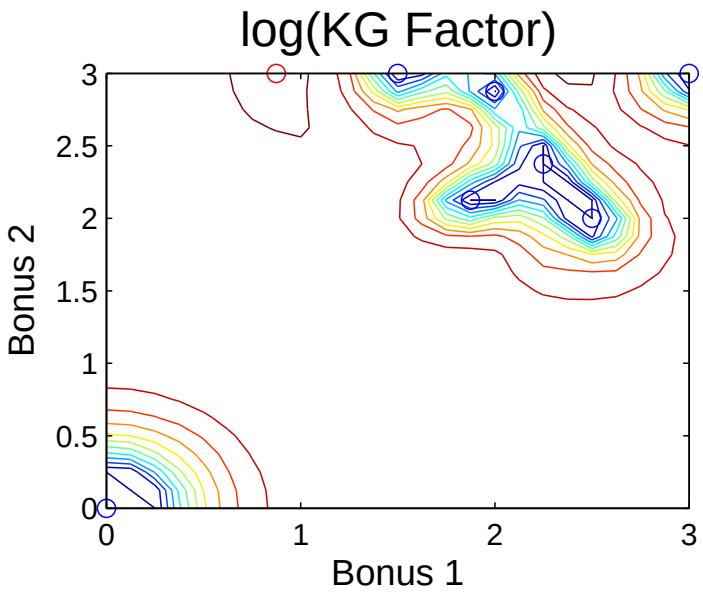


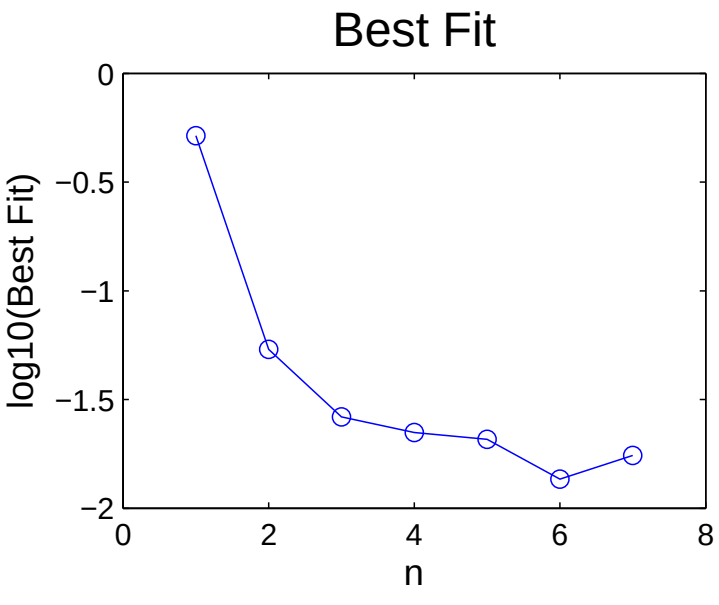


Simulation Model Calibration Results









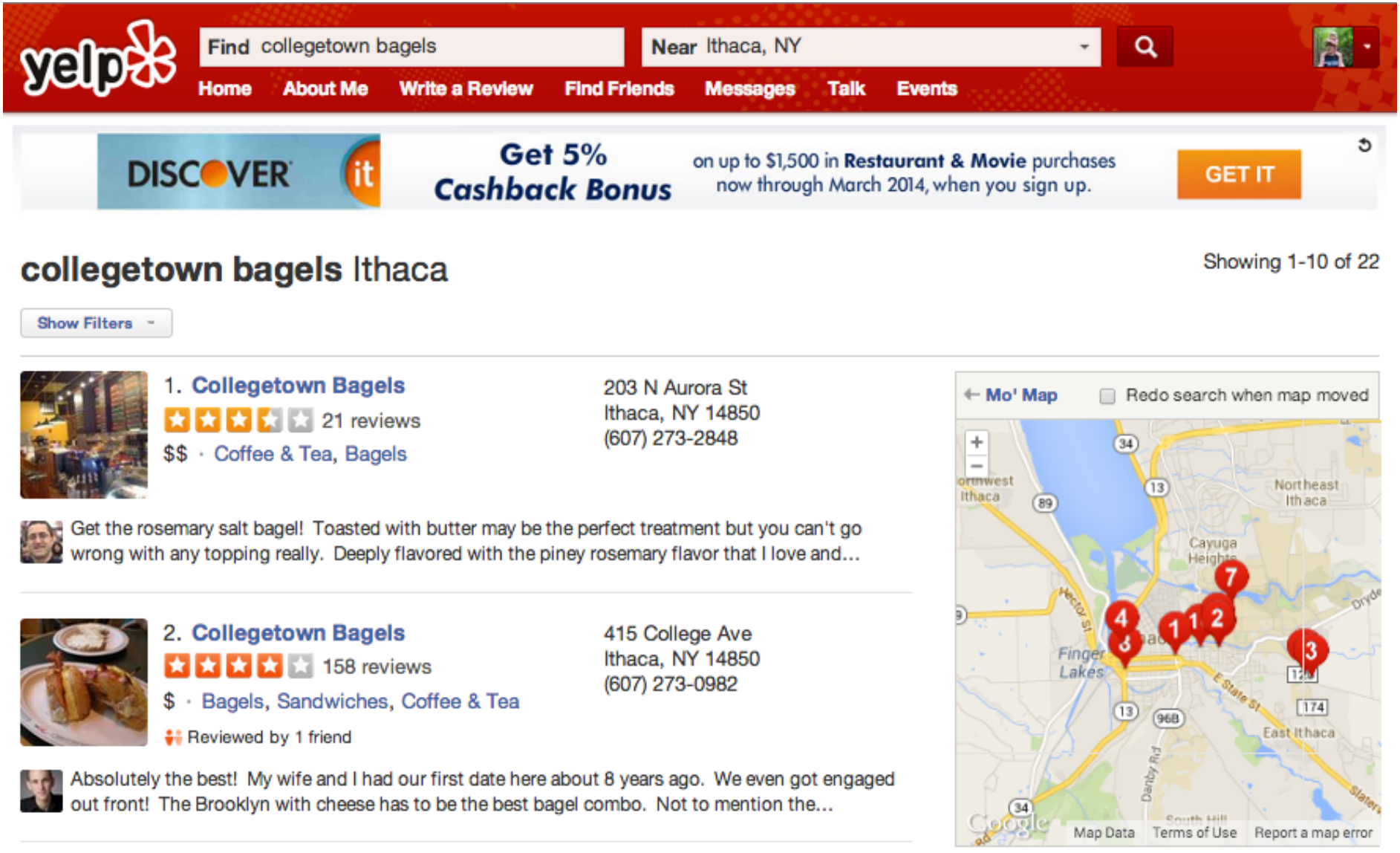
Simulation Model Calibration Results

- The KG method calibrates the model in approximately 3 days, compared to $7-14$ days when tuned by hand.
- The calibration is automatic, freeing the human calibrator to do other work.
- The KG method calibrates as accurately or better than does by-hand calibration.
- Current practice uses the year's calibrated bonuses for each new "what if" scenario, but to enforce the constraint on driver at-home time it would be better to recalibrate the model for each scenario. Automatic calibration with the KG method makes this feasible.

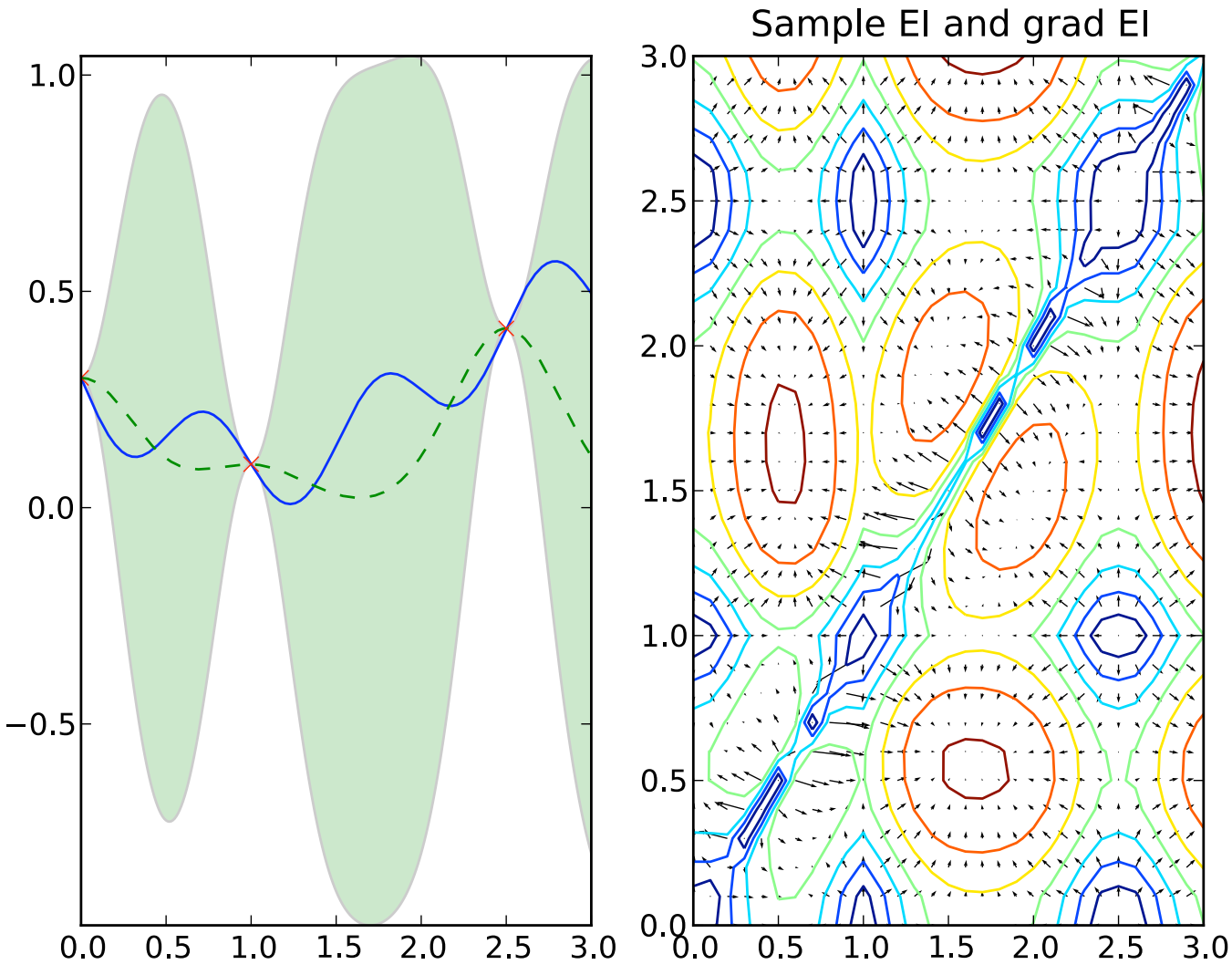
Outline

1. Background: Gaussian Process Regression
2. The Knowledge Gradient (KG) Method [F., Powell, Dayanik 2009]
3. Applications and Variations
   - Design of Cardiovascular Bypass Grafts [Xie, F., Sankaran, Marsden, Elmohamed 2012]
   - Simulation Calibration at Schneider National [F., Powell, Simao 2009]
   - Improving Customer Experience and Revenue at Yelp
   - Exploiting Common Random Numbers [Xie, F., Chick 2013]
   - Developing Inexpensive Organic Photovoltaic Cells
   - Drug Development for Ewing's Sarcoma [Negoescu, F., Powell 2011]
4. Conclusion

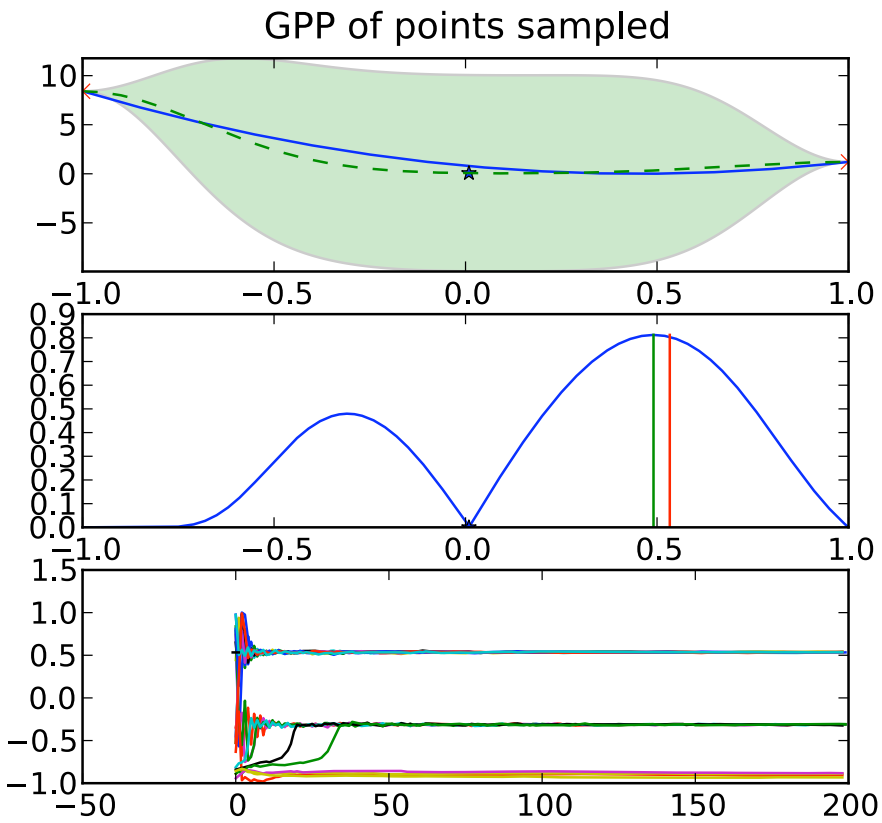
We are using BGO at Yelp

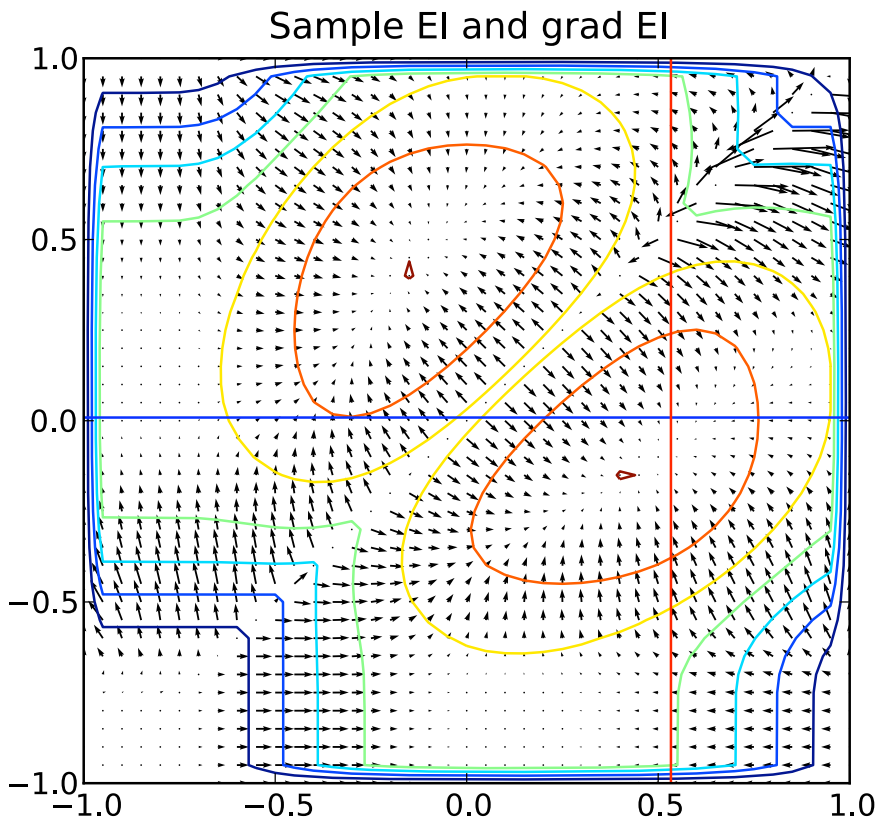


Yelp can do multiple simultaneous function evaluations

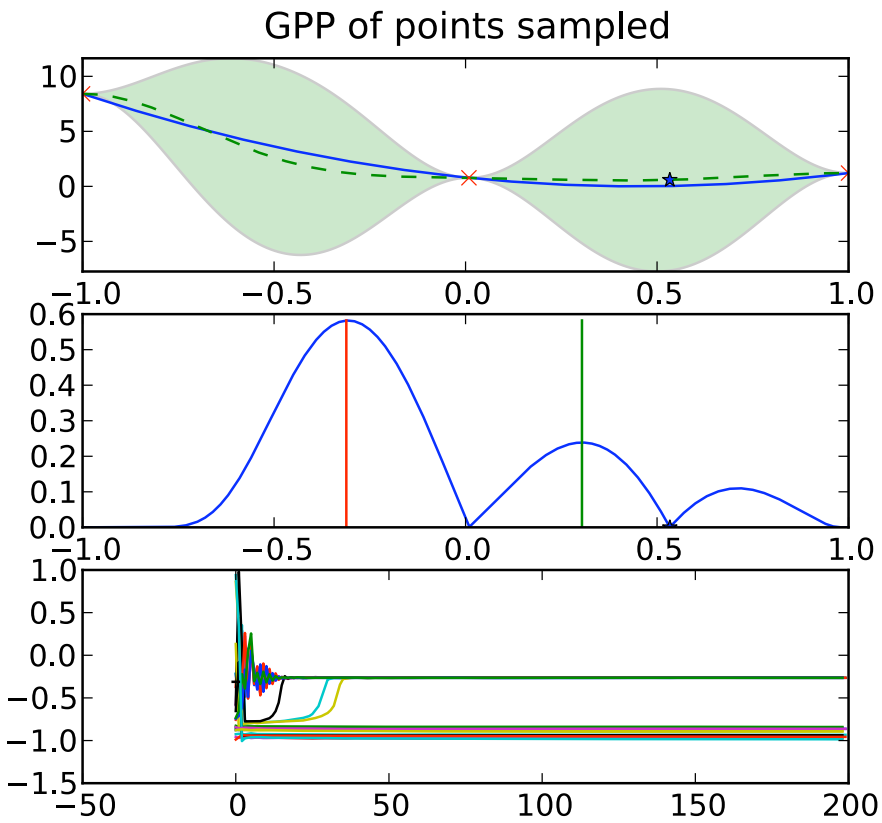


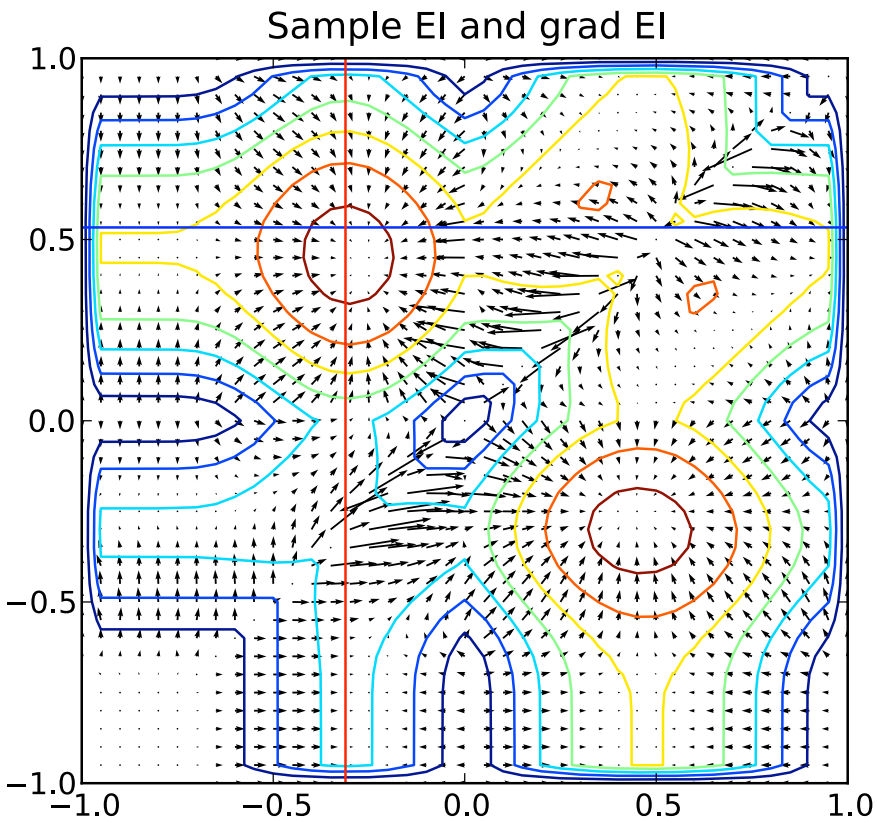
We find *sets* measurements that optimize the value of information.



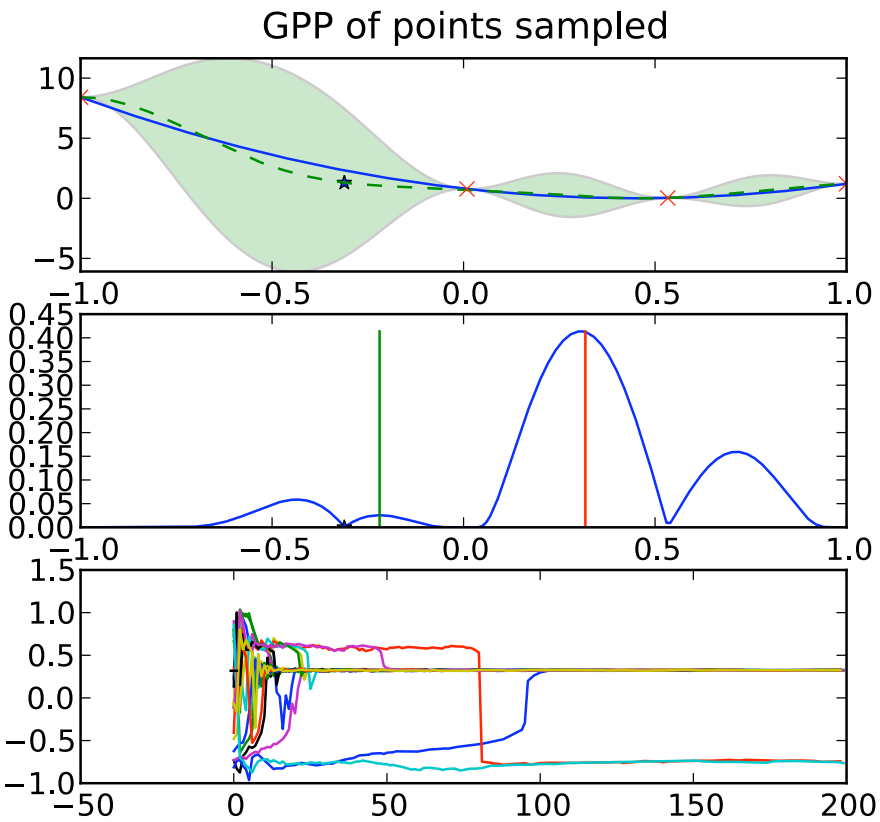


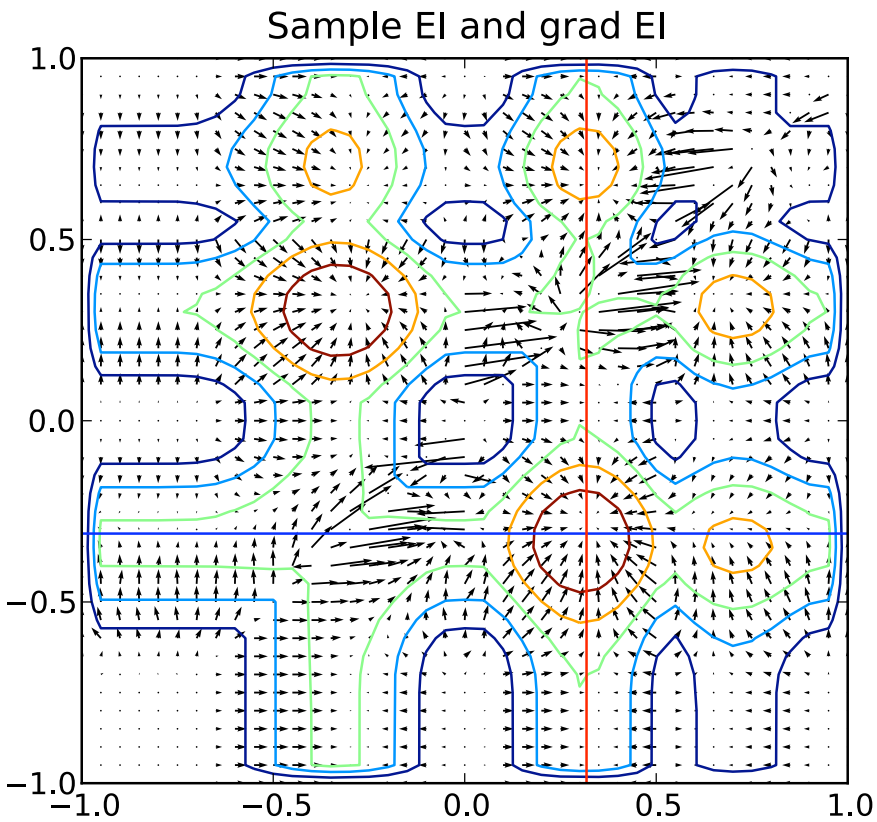
We find *sets* measurements that optimize the value of information.





We find *sets* measurements that optimize the value of information.

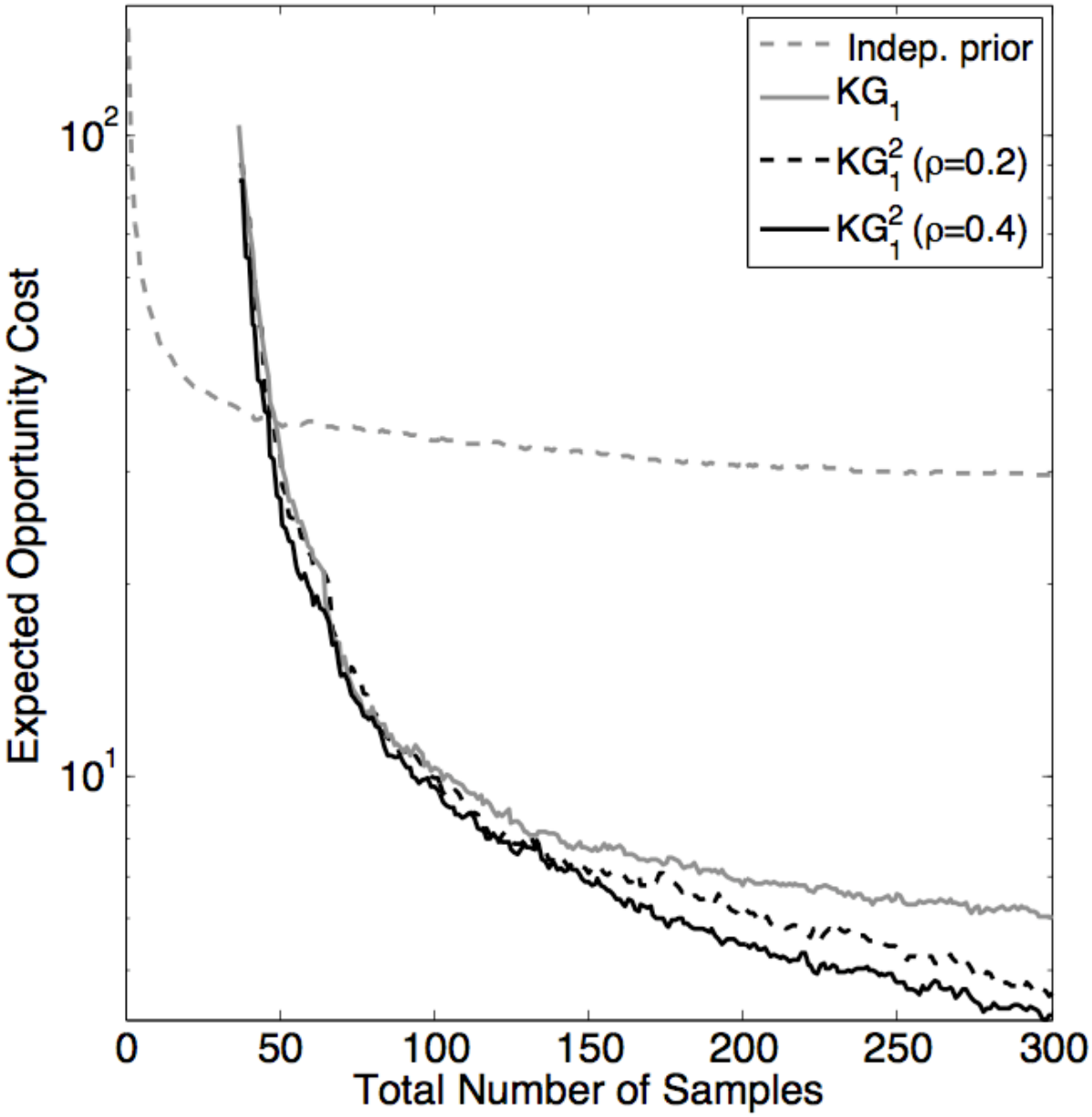




Outline

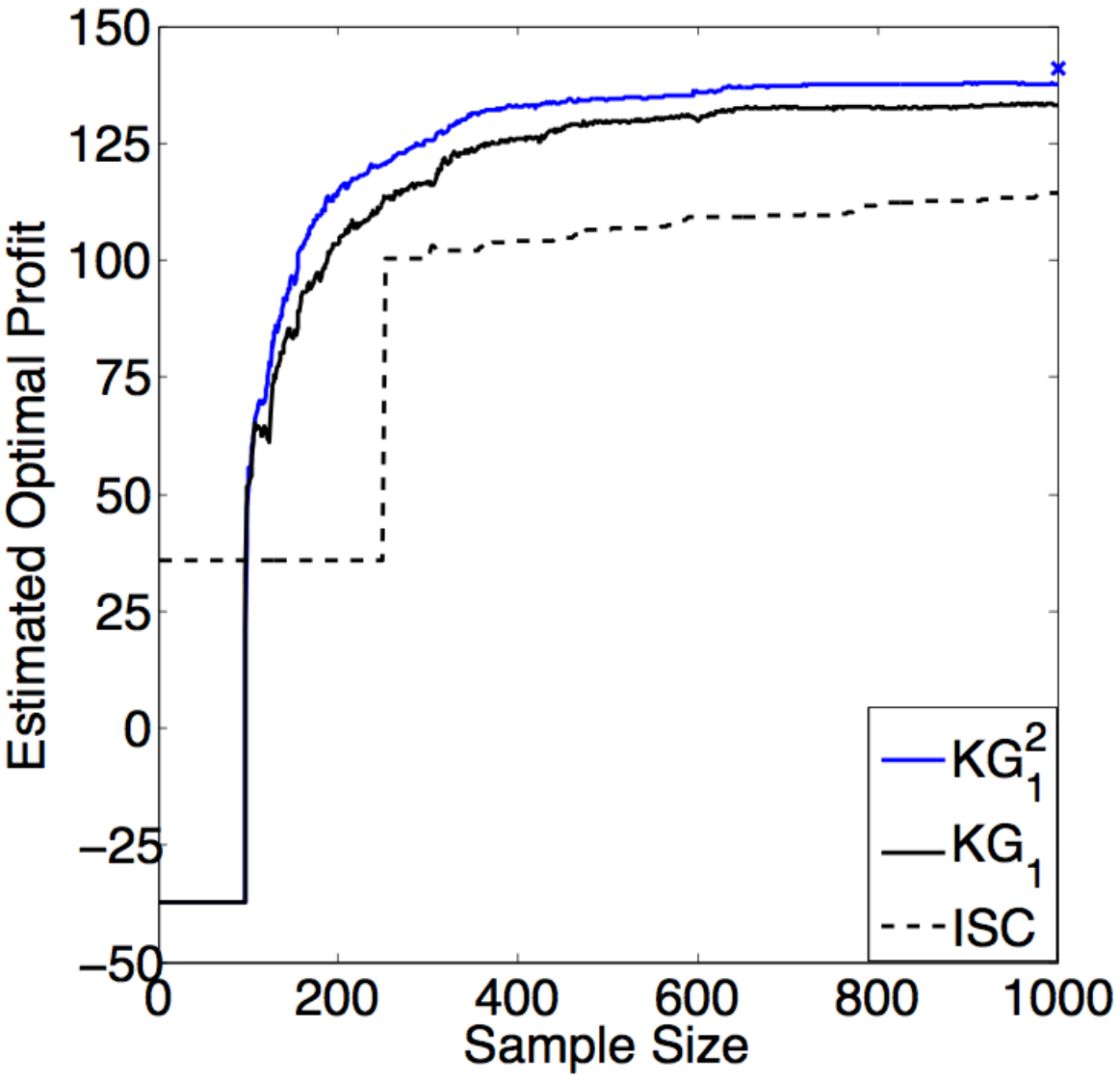
1. Background: Gaussian Process Regression
2. The Knowledge Gradient (KG) Method [F., Powell, Dayanik 2009]
3. Applications and Variations
   - Design of Cardiovascular Bypass Grafts [Xie, F., Sankaran, Marsden, Elmohamed 2012]
   - Simulation Calibration at Schneider National [F., Powell, Simao 2009]
   - Improving Customer Experience and Revenue at Yelp
   - Exploiting Common Random Numbers [Xie, F., Chick 2013]
   - Developing Inexpensive Organic Photovoltaic Cells
   - Drug Development for Ewing's Sarcoma [Negoescu, F., Powell 2011]
4. Conclusion

Using Common Random Numbers Improves Performance



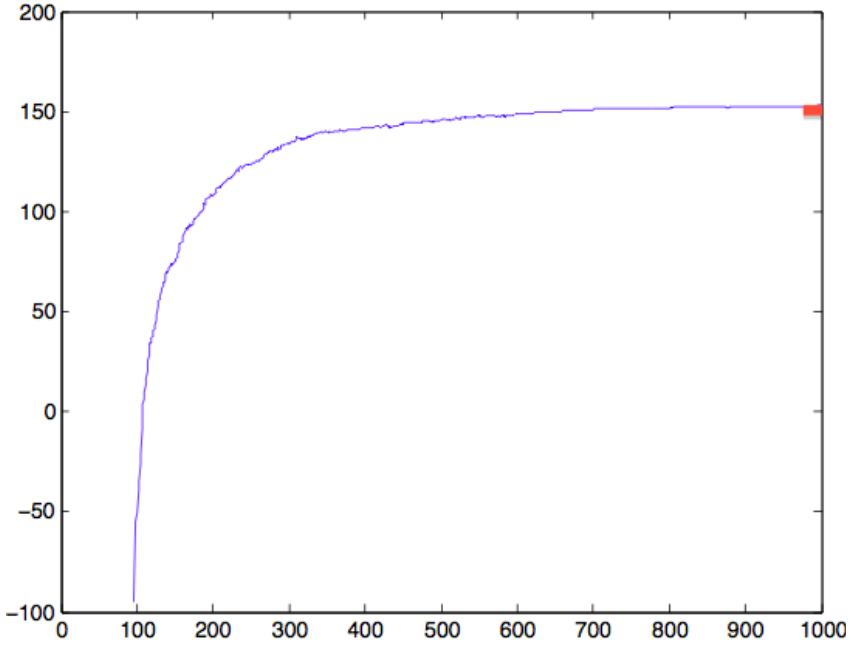
- f is the 3-dimensional Rosenbrock function.
- Our feasible set has $|A| = 4096$ points.
- KG_1 uses independent sampling noise.
- KG_1^2 can use either independent samples or common random numbers with sampling correlation ρ.
- Increasing ρ decreases opportunity cost, and improves performance.

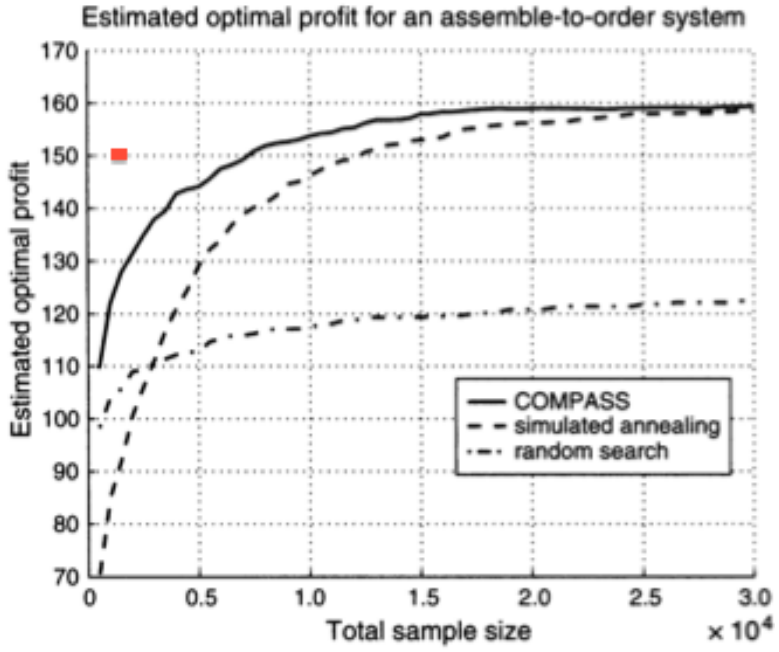
KG does well against Industrial Strength COMPASS (developed in [Xie, Nelson & Hong '10])



- Figure shows performance (value vs. # samples) on an 8-dim assemble-to-order problem from simopt.org
- KG_1^2 uses common random numbers; KG_1 uses independent sampling; ISC is Industrial Strength COMPASS.
- ISC took 27 min to do 1000 samples. KG took 9 min (independent sampling) and 5.5 min (correlated sampling) to achieve the same avg. solution quality.

KG does well against COMPASS [Hong & Nelson '06]





- Left shows KG's performance (value vs. # samples), from 1 to 1,000 samples on an 8-dimensional assemble-to-order problem from [Hong and Nelson, 2006].
- Right shows COMPASS [Hong and Nelson, 2006], simulated annealing [Alrefaei and Andradóttir, 1999], and random search from 1 to 30,000 samples on the same problem.

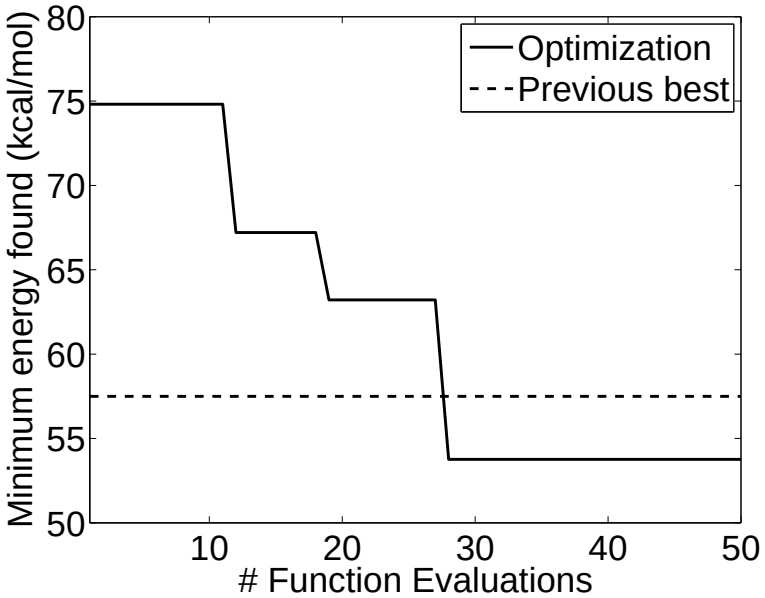
What do we know from a theoretical point of view?

- In a very broad set of problems, these one-step methods are **consistent**, i.e., as our sampling budget grows to infinity, our error in finding the optimum shrinks to 0.
  - For convergence proofs in specific problem settings, see [Frazier et al., 2008, Frazier et al., 2009b].
  - For general sufficient conditions, see [Frazier and Powell, 2011].
- In some special problems, one-step methods are actually **optimal**.
  - Some special variants of the independent normal ranking and selection problem, [Frazier et al., 2008].
  - A stylized version of stochastic root-finding, [Jedynak et al., 2012]
- In other special problems, we can compute the optimal method exactly using **dynamic programming**, and compare the quality of one-step methods against optimal.
  - Multiple comparisons with a known standard [Xie and Frazier, 2013].

Outline

1. Background: Gaussian Process Regression
2. The Knowledge Gradient (KG) Method [F., Powell, Dayanik 2009]
3. Applications and Variations
   - Design of Cardiovascular Bypass Grafts [Xie, F., Sankaran, Marsden, Elmohamed 2012]
   - Simulation Calibration at Schneider National [F., Powell, Simao 2009]
   - Improving Customer Experience and Revenue at Yelp
   - Exploiting Common Random Numbers [Xie, F., Chick 2013]
   - Developing Inexpensive Organic Photovoltaic Cells
   - Drug Development for Ewing's Sarcoma [Negoescu, F., Powell 2011]
4. Conclusion

We are using BGO to solve crystal structures, to design better photovoltaic cells

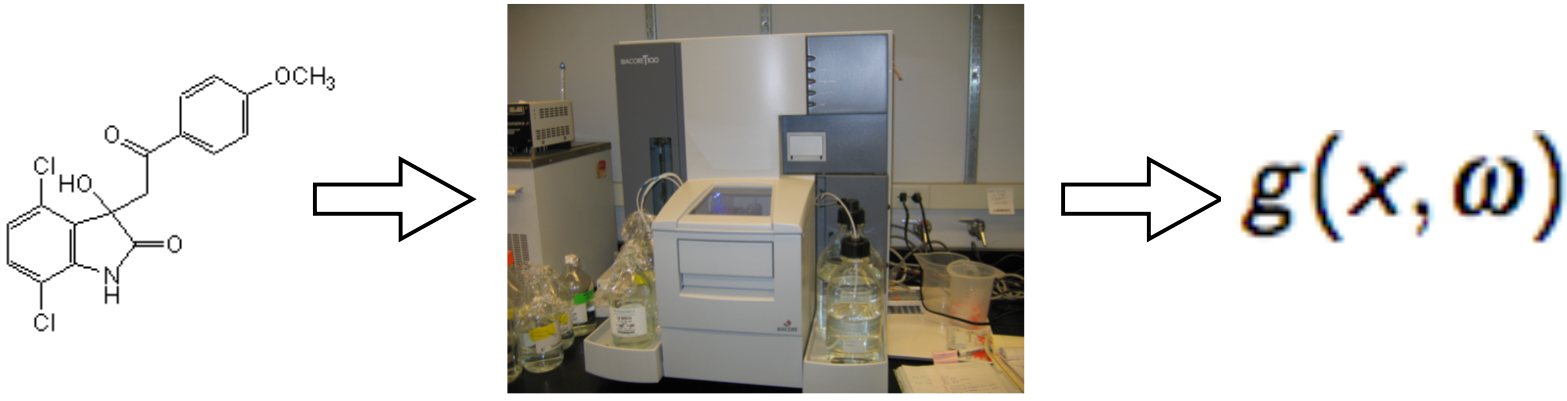


- With Paulette Clancy and Kristina Lenn (Cornell Chemical Eng.).
- Each function evaluation evaluates the energy of a particular crystal configuration, and requires 1 to 2 minutes of computation.
- The dashed line shows the lowest energy configuration found by a human, using 10 hours CPU time and weeks of manual inspection.
- The optimization algorithm was able to find a lower energy configuration, automatically, and using only 1 hour of CPU time.

Outline

1. Background: Gaussian Process Regression
2. The Knowledge Gradient (KG) Method [F., Powell, Dayanik 2009]
3. Applications and Variations
   - Design of Cardiovascular Bypass Grafts [Xie, F., Sankaran, Marsden, Elmohamed 2012]
   - Simulation Calibration at Schneider National [F., Powell, Simao 2009]
   - Improving Customer Experience and Revenue at Yelp
   - Exploiting Common Random Numbers [Xie, F., Chick 2013]
   - Developing Inexpensive Organic Photovoltaic Cells
   - Drug Development for Ewing's Sarcoma [Negoescu, F., Powell 2011]
4. Conclusion

Drug Development is Global Optimization

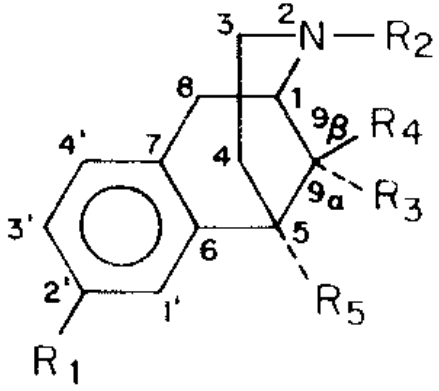


- We have a large number of chemically related small molecules, some of which might make a good drug.
- We can synthesize and test the quality of these molecules, but each molecule tested takes days of effort.
- $f(x)$ is the quality of molecule x, and $g(x, \omega)$ is the test result.
- We would like to find a good drug with a limited number of tests.

Joint work with Jeffrey Toretsky, M.D. (Georgetown), Diana Negoescu (Stanford), Warren Powell (Princeton), [Negoescu et al., 2011])

We Use a Gaussian Process Prior

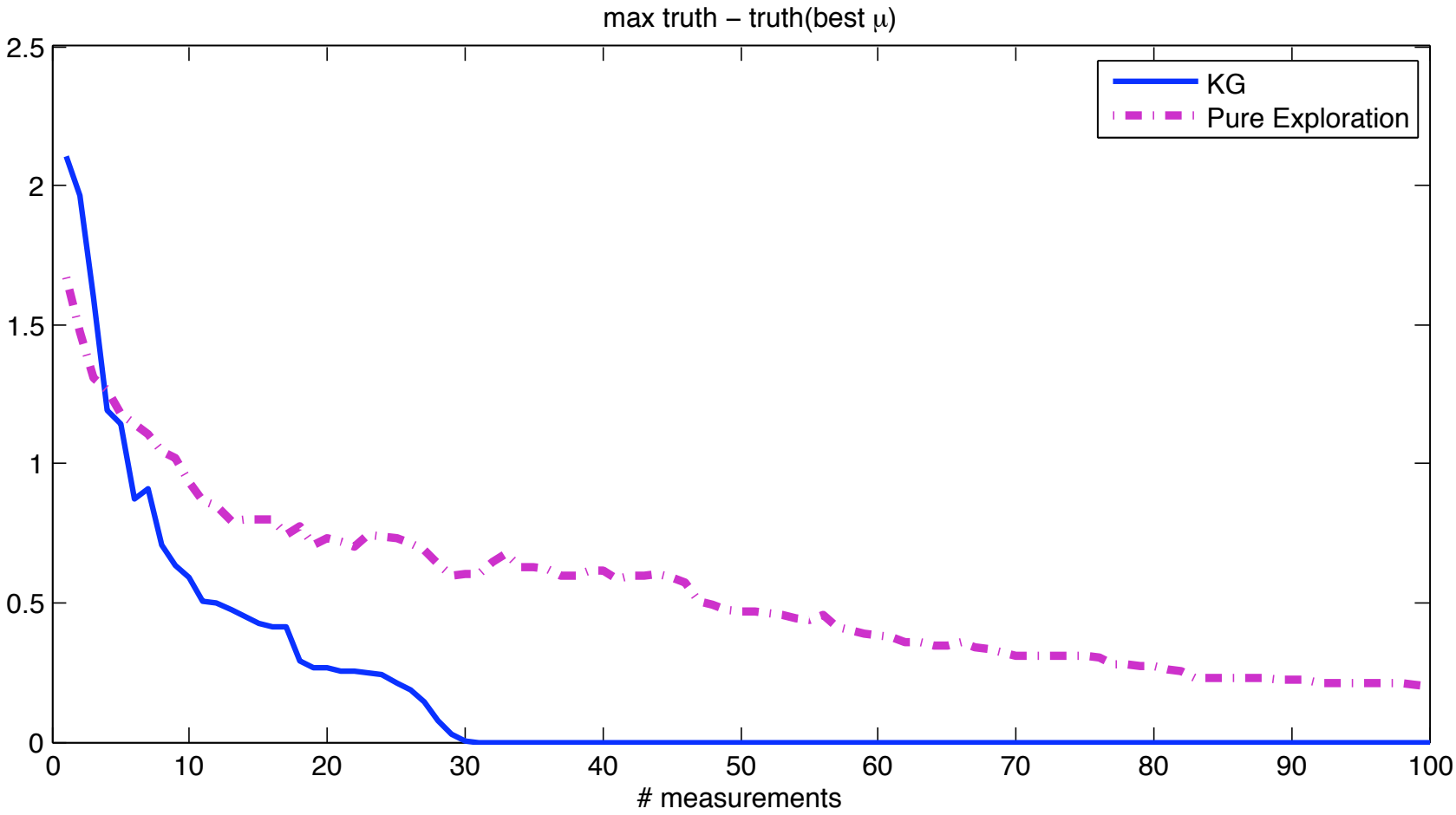
- The molecules we consider share a common skeleton, and are described by which substituents are present at each location.



Substituent
$R_1 = F$
$R_1 = OH$
$R_1 = OCOCH_3$
$R_1 = OCOCH_2CH_3$
$R_1 = OCOCH_2CH_2CH_3$
$R_1 = OCOCH_2(CH_2)_2CH_3$
$R_1 = OCH_3$
$R_1 = NO_2$
$R_1 = OCOAr(3\text{-}C_5H_4N)$
$R_1 = Cl$

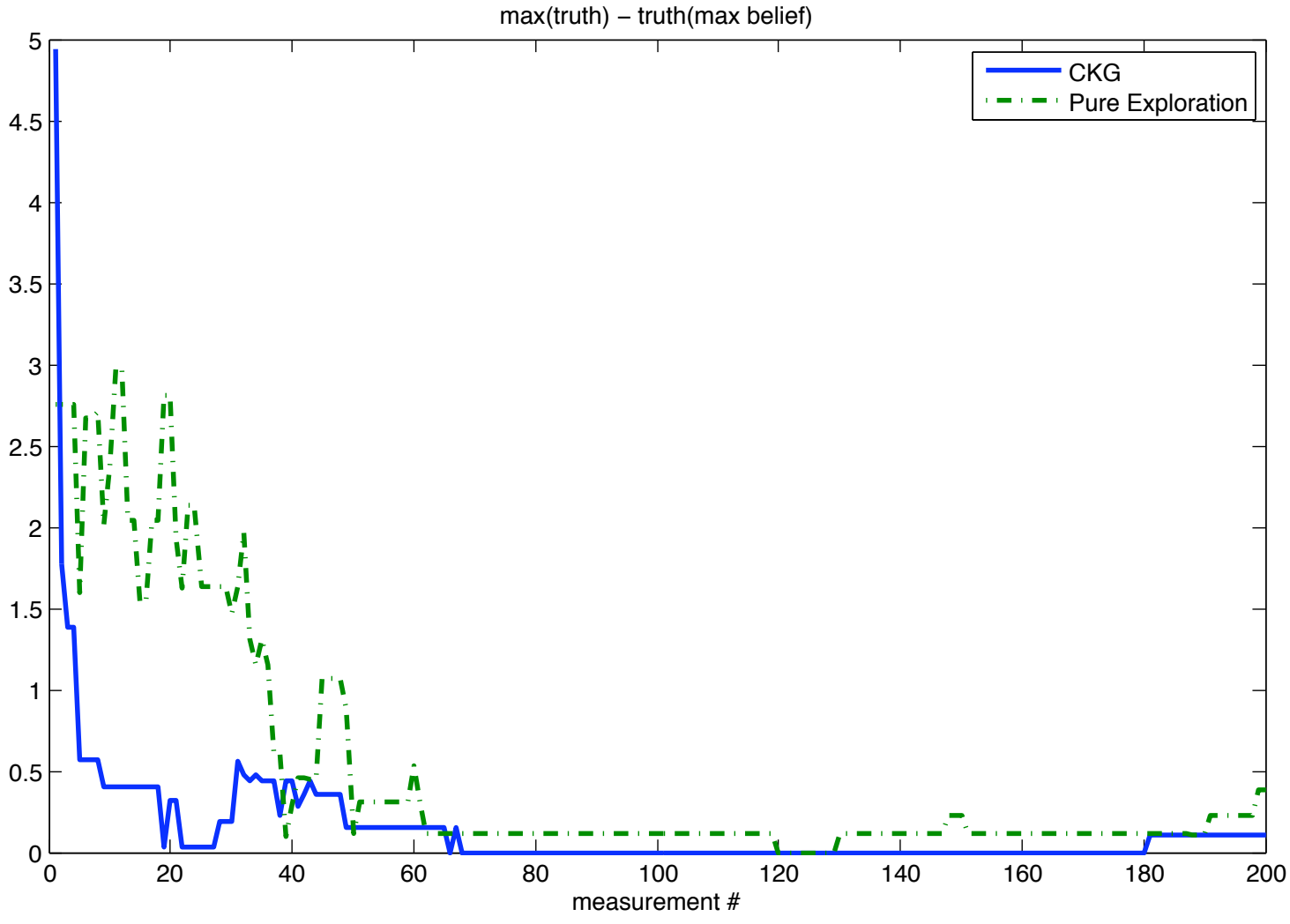
- We use Gaussian Process regression over the discrete, combinatorial, space of molecules.
  (this could also be called Bayesian linear regression).
- The covariance $\Sigma_0(x, x')$ of two molecules x and x' is larger when the two molecules have more substituents in common
  [Free and Wilson, 1964].
- It is also possible to include molecular features like molecular weight, hydrophilicity, etc. [Hansch and Fujita, 1964].

KG works well on a curated dataset from the literature



Average over 100 sample paths on randomly selected subsets of benzomorphan compounds of size 99. Data from [Katz et al., 1977].

KG Works Well in Tests



One sample path on the full set of 87,120 benzomorphan compounds.

Discussion: KG Works Well So Far...

- KG works well in test problems using a chemical dataset from the literature. [Negoescu, Frazier, Powell 2011]
- We are applying this method to a collection of small molecules that inhibit a protein interaction required for metastisis of Ewing's Sarcoma.

Outline

Conclusion

- We have discussed methods that use the Bayesian posterior and a one-step optimality analysis to decide where to sample next.
- These methods aim to reduce the number of simulation replications required, in the **average case**.
- This approach is **flexible**, and can be adapted to new applications (non-stationary output, combinatorial feasible set, . . .)

Thank You

Any questions?

References I

Alrefaei, M. and Andradóttir, S. (1999).
A simulated annealing algorithm with constant temperature for discrete stochastic optimization.
Management science, 45(5):748–764.

Calvin, J. and Zilinskas, A. (2002).
One-dimensional Global Optimization Based on Statistical Models.
Nonconvex Optimization and its Applications, 59:49–64.

Calvin, J. and Zilinskas, A. (2005).
One-Dimensional global optimization for observations with noise.
Computers & Mathematics with Applications, 50(1-2):157–169.

Chick, S. and Inoue, K. (2001).
New two-stage and sequential procedures for selecting the best simulated system.
Operations Research, 49(5):732–743.

Forrester, A., Keane, A., and Bressloff, N. (2006).
Design and Analysis of" Noisy" Computer Experiments.
AIAA Journal, 44(10):2331–2339.

Frazier, P. and Powell, W. (2011).
Consistency of sequential Bayesian sampling policies.
SIAM Journal on Control and Optimization, 49:712–731.

Frazier, P., Powell, W., and Simão, H. (2009a).
Simulation model calibration with correlated knowledge-gradients.
In *Winter Simulation Conference Proceedings, 2009*, Piscataway, New Jersey. Institute of Electrical and Electronics Engineers, Inc.

References II

Frazier, P., Powell, W. B., and Dayanik, S. (2008).

A knowledge gradient policy for sequential information collection.
SIAM Journal on Control and Optimization, 47(5):2410–2439.

Frazier, P., Powell, W. B., and Dayanik, S. (2009b).

The knowledge gradient policy for correlated normal beliefs.
INFORMS Journal on Computing, 21(4):599–613.

Free, S. and Wilson, J. (1964).

Mathematical contribution to structure-activity studies.
Journal of Medicinal Chemistry, 7(4):395.

Fu, M., Hu, J., Chen, C., and Xiong, X. (2004).

Optimal computing budget allocation under correlated sampling.
Proceedings of the 36th conference on Winter simulation, pages 595–603.

Gupta, S. and Miescke, K. (1996).

Bayesian look ahead one-stage sampling allocations for selection of the best population.
Journal of statistical planning and inference, 54(2):229–244.

Hansch, C. and Fujita, T. (1964).

ρ-σ-π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure.
J. Am. Chem. Soc, 86(8):1616–1626.

Hong, L. and Nelson, B. (2006).

Discrete optimization via simulation using compass.
OPERATIONS RESEARCH, 54(1):115.

References III

Huang, D., Allen, T., Notz, W., and Miller, R. (2006).

Sequential kriging optimization using multiple-fidelity evaluations.
Structural and Multidisciplinary Optimization, 32(5):369–382.

Jedynak, B., Frazier, P. I., and Sznitman, R. (2012).

Twenty questions with noise: Bayes optimal policies for entropy loss.
Journal of Applied Probability, 49(1):114–136.

Jones, D., Schonlau, M., and Welch, W. (1998).

Efficient Global Optimization of Expensive Black-Box Functions.
Journal of Global Optimization, 13(4):455–492.

Katz, R., Osborne, S., and Ionescu, F. (1977).

Application of the Free-Wilson technique to structurally related series of homologues. Quantitative structure-activity relationship studies of narcotic analgetics.
J Med Chem, 20(11):1413–9.

Kleijnen, J., van Beers, W., and van Nieuwenhuyse I. (2011).

Expected improvement in efficient global optimization through bootstrapped kriging.
Journal of Global Optimization, pages 1–15.

Kushner, H. J. (1964).

A new method of locating the maximum of an arbitrary multi- peak curve in the presence of noise.
Journal of Basic Engineering, 86:97–106.

Mockus, J. (1989).

Bayesian approach to global optimization: theory and applications.
Kluwer Academic, Dordrecht.

References IV

Mockus, J., Tiesis, V., and Zilinskas, A. (1978).

The application of Bayesian methods for seeking the extremum.
In Dixon, L. and Szego, G., editors, *Towards Global Optimisation*, volume 2, pages 117–129. Elsevier Science Ltd., North Holland, Amsterdam.

Negoescu, D., Frazier, P., and Powell, W. (2011).

The knowledge gradient algorithm for sequencing experiments in drug discovery.
INFORMS Journal on Computing, 23(1).

Sankaran, S. and Marsden, A. (2011).

A stochastic collocation method for uncertainty quantification in cardiovascular simulations.
Journal of Biomechanical Engineering, 133:031001.

Scott, W., Frazier, P. I., and Powell, W. B. (2011).

The correlated knowledge gradient for simulation optimization of continuous parameters using gaussian process regression.
SIAM Journal on Optimization, 21:996–1026.

Stuckman, B. (1988).

A global search method for optimizing nonlinear systems.
Systems, Man and Cybernetics, IEEE Transactions on, 18(6):965–977.

Taddy, M., Lee, H., Gray, G., and Griffin, J. (2009).

Bayesian guided pattern search for robust local optimization.
Technometrics, 51(4):389–401.

Villemonteix, J., Vazquez, E., and Walter, E. (2009).

An informational approach to the global optimization of expensive-to-evaluate functions.
Journal of Global Optimization, 44(4):509–534.

References V

Xie, J. and Frazier, P. (2013).

Sequential bayes-optimal policies for multiple comparisons with a known standard.
Operations Research.
to appear.

Xie, J., Frazier, P., and Chick, S. (2013).

Bayesian optimization via simulation with pairwise sampling and correlated prior beliefs.
in review.

Xie, J., Frazier, P., Sankaran, S., Marsden, A., and Elmohamed, S. (2012).

Gaussian-process-based black-box optimization for cardiovascular bypass graft design under parameter uncertainty.
In *The 50th Annual Allerton Conference on Communication, Control and Computing*.

Computation of the KG Policy for $|A| < \infty$

- Let B contain those points in A that are best under the time $n+1$ posterior with strictly positive probability.
- Let $x_{[j]}$ denote the j^{th} entry in B.
- Let $a_j = \mu_n(x_{[j]})$ and $b_j = \sqrt{\text{Var}_n\left[\mu_{n+1}(x_{[j]})\right]}$. These values, and the set B, depend on $x_{n+1} = x$.
- Sort B in order of increasing b_j.
- The KG factor is

$$\text{KG}(x) = \sum_{j=1}^{|B|-1} (b_{j+1} - b_j) g\left(\frac{-|a_{j+1} - a_j|}{b_{j+1} - b_j}\right),$$

where $g(z) = \varphi(z) + z\Phi(z)$, φ is the normal pdf and Φ is the normal cdf.