Gaussian Process Response Surface Optimization

Dan Lizotte

Department of Statistics
University of Michigan

Russ Greiner, Dale Schuurmans

Department of Computing Science
University of Alberta

Response Surface Methods for Noisy Functions

- Review of response surface methods for optimizing deterministic functions
- New methodology for algorithm evaluation
- Applying our methodology to response surface methods for noisy functions

Response Surface Methods

- Methods for optimizing a function f(x) that is
  - At least somewhat continuous/differentiable/regular
    - i.e., not thinking about combinatorial problems
  - Non-convex, multiple local optima
  - Expensive to evaluate

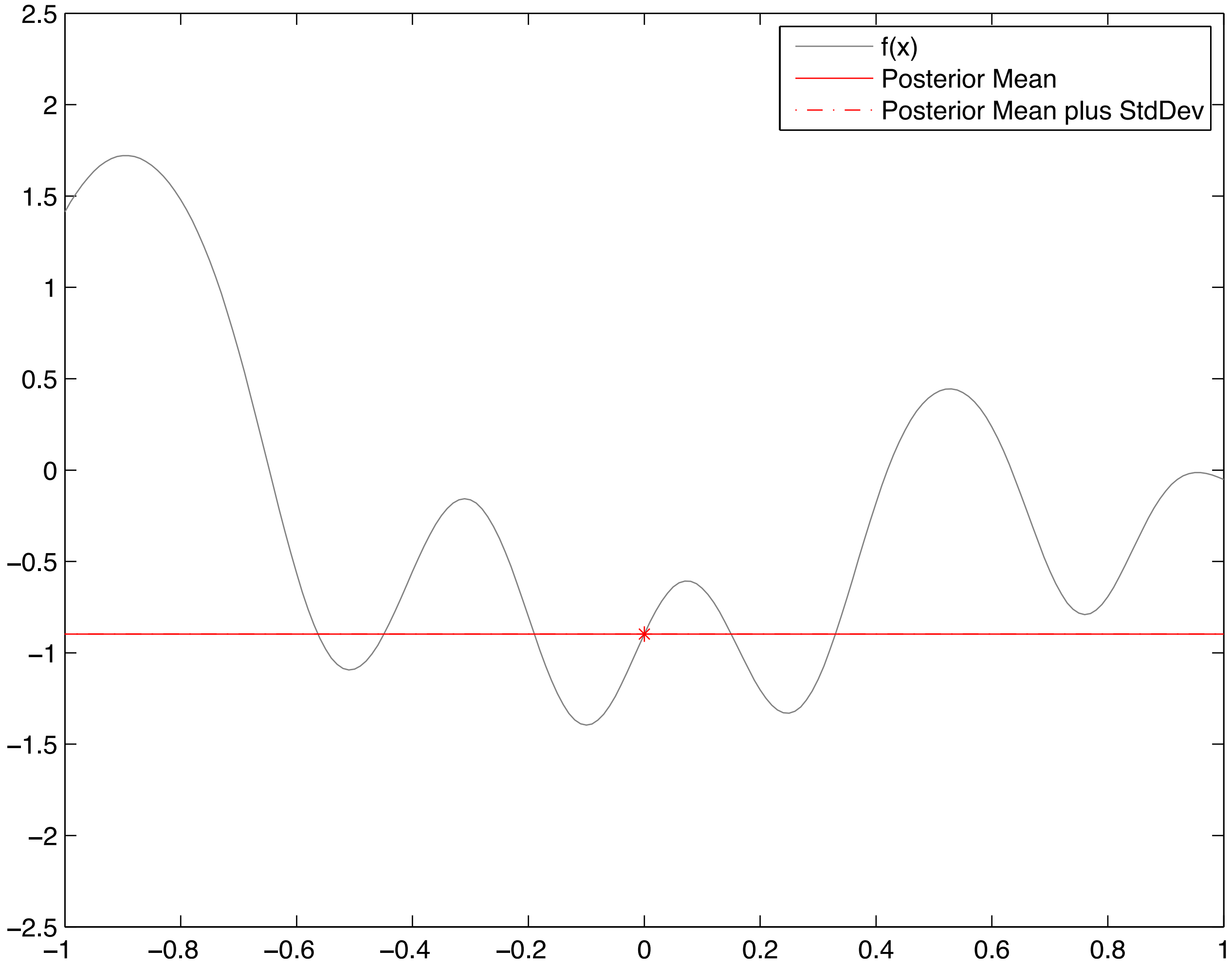
Response Surface Methods

- Two main components:
  - Response Surface Model
    - Makes a prediction $\mu(x)$ about $f(x)$ at any point x
    - Provides uncertainty information $\sigma(x)$ about predictions
  - Acquisition Criterion
    - A function of $\mu(x)$ and $\sigma(x)$
    - Expresses our desire to observe $f(x)$ versus $f(z)$ next
    - Prefers points x that, with high confidence, are predicted to have larger $f(x)$ than we have already observed

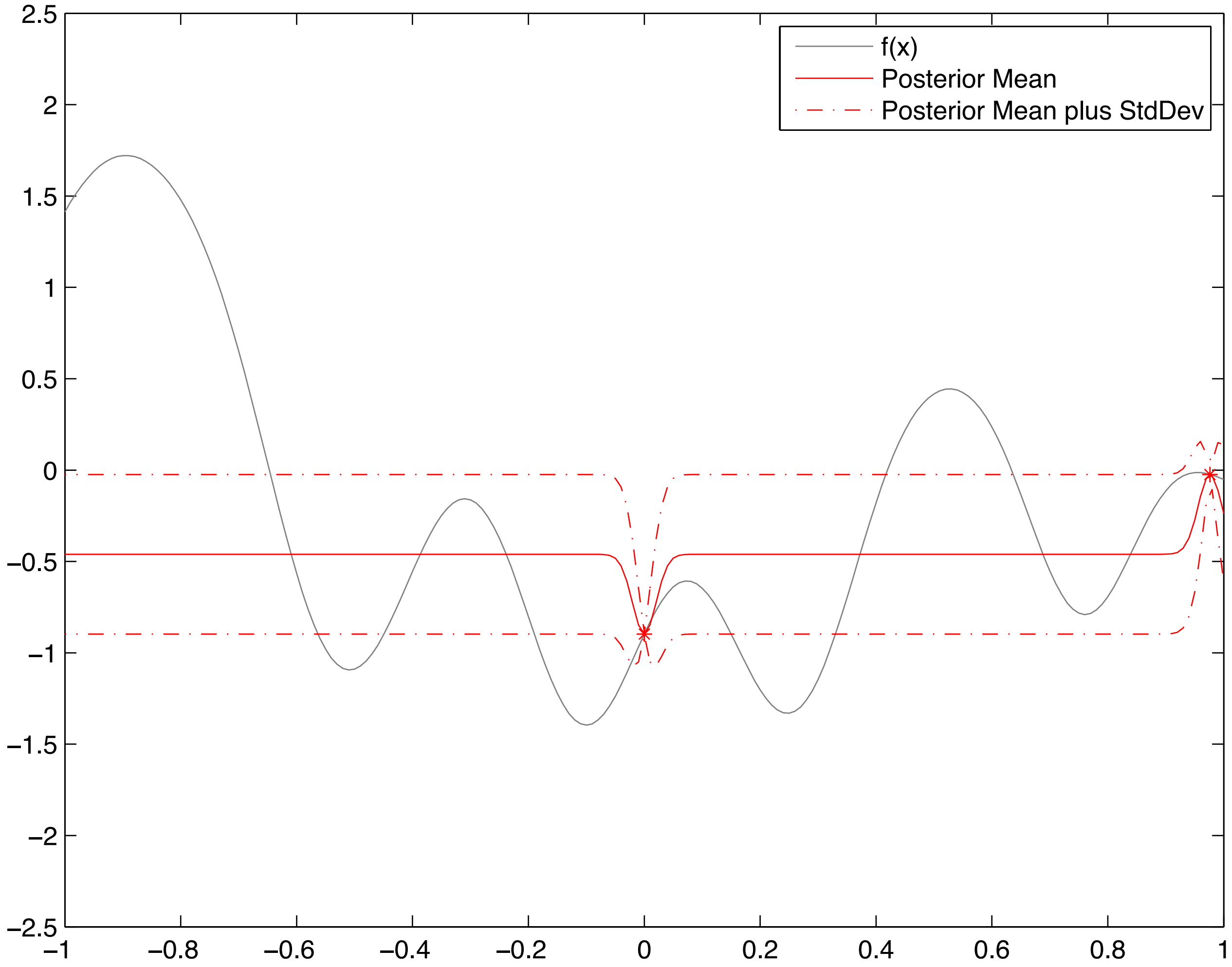
Response Surface Methods

- DO
  - Construct a model of f(x) using Data, giving μ(x) and σ(x)
    - Model is probabilistic; can accommodate noisy f
  - Find the optimum of the acquisition criterion, giving x^+
  - Evaluate $f(x^+)$, add observation to our pool of Data
- UNTIL “bored” (e.g. number of samples >= N),
  or “hopeless” (e.g. probability of improvement less than ε)

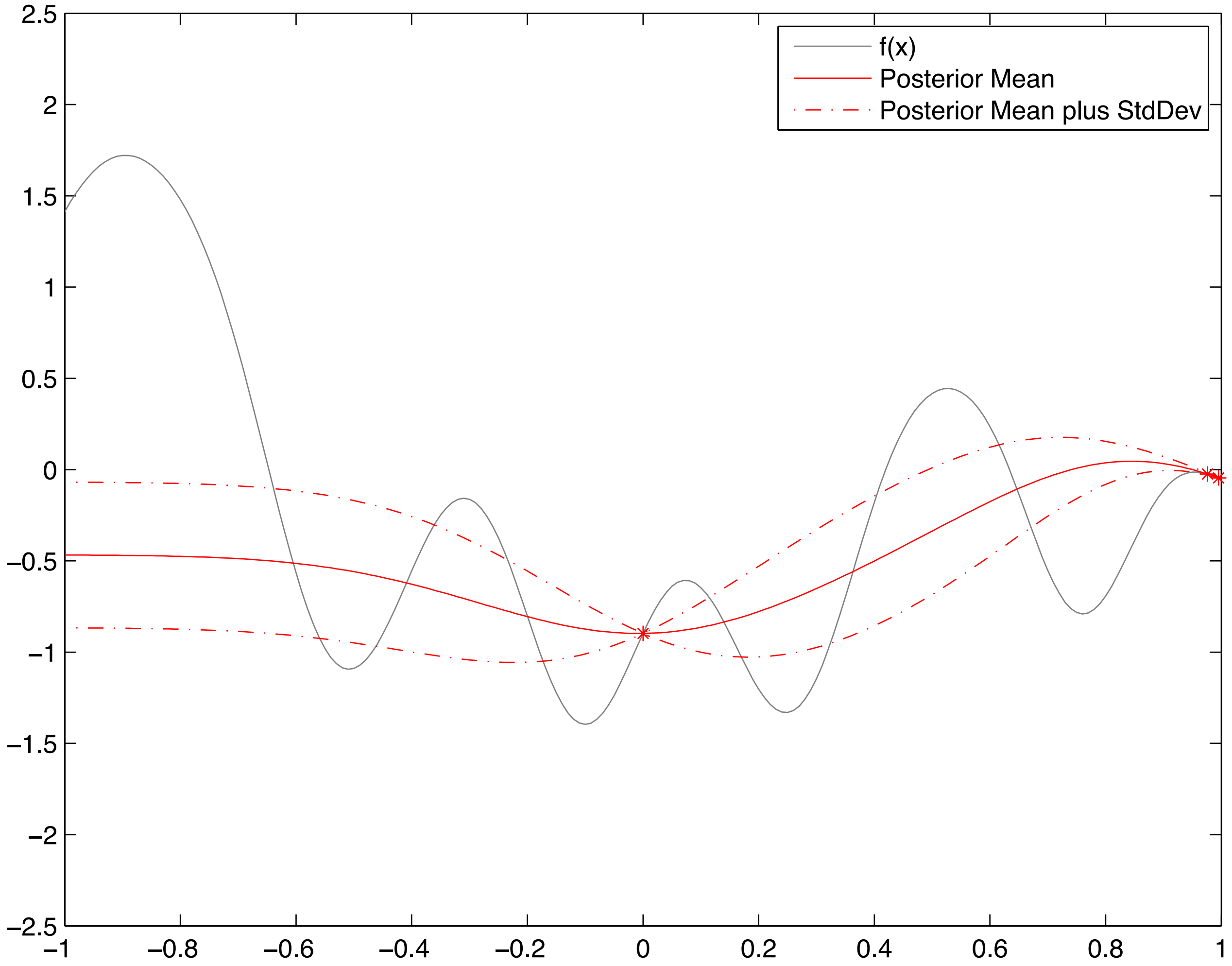
Response Surface Methods



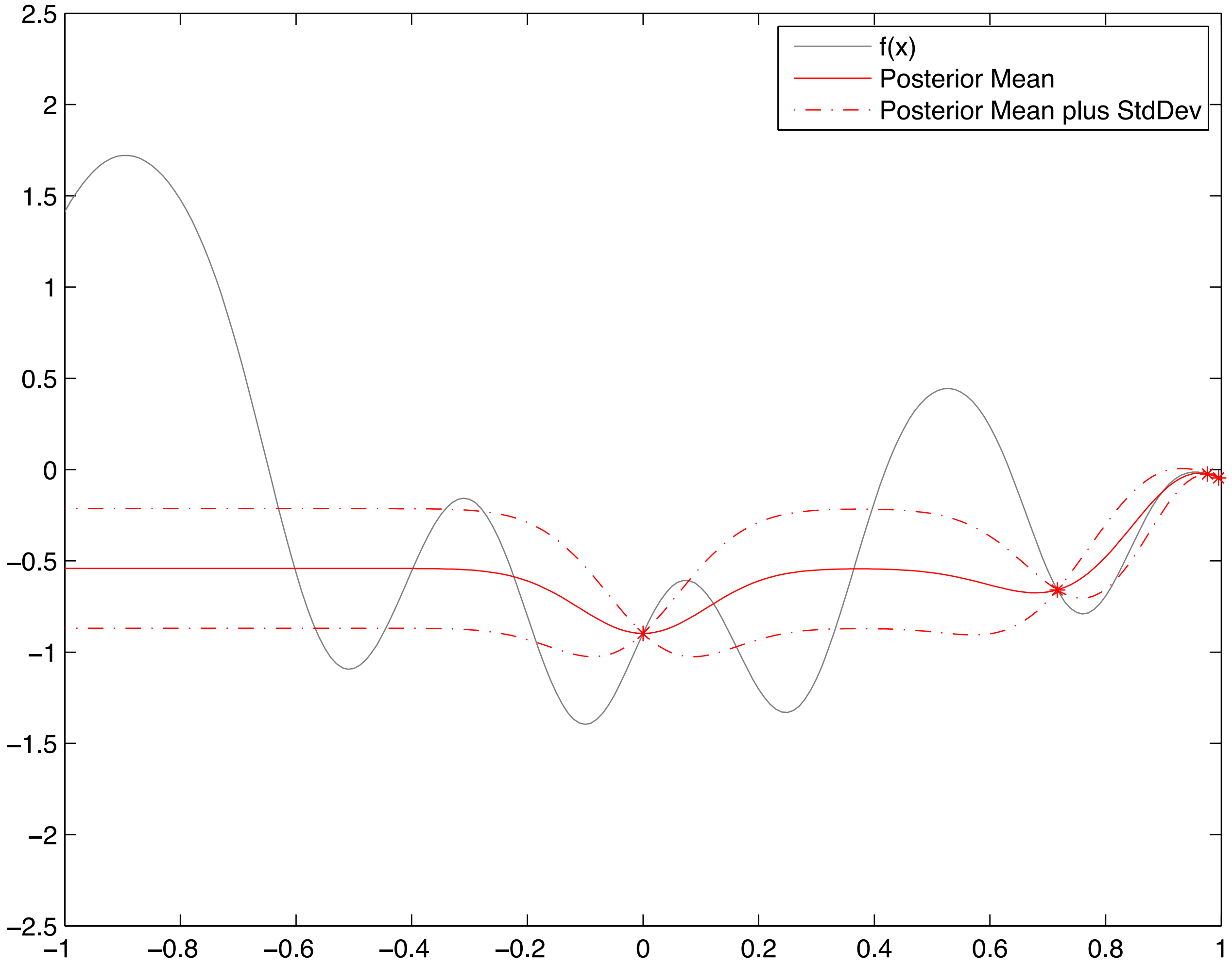
Response Surface Methods



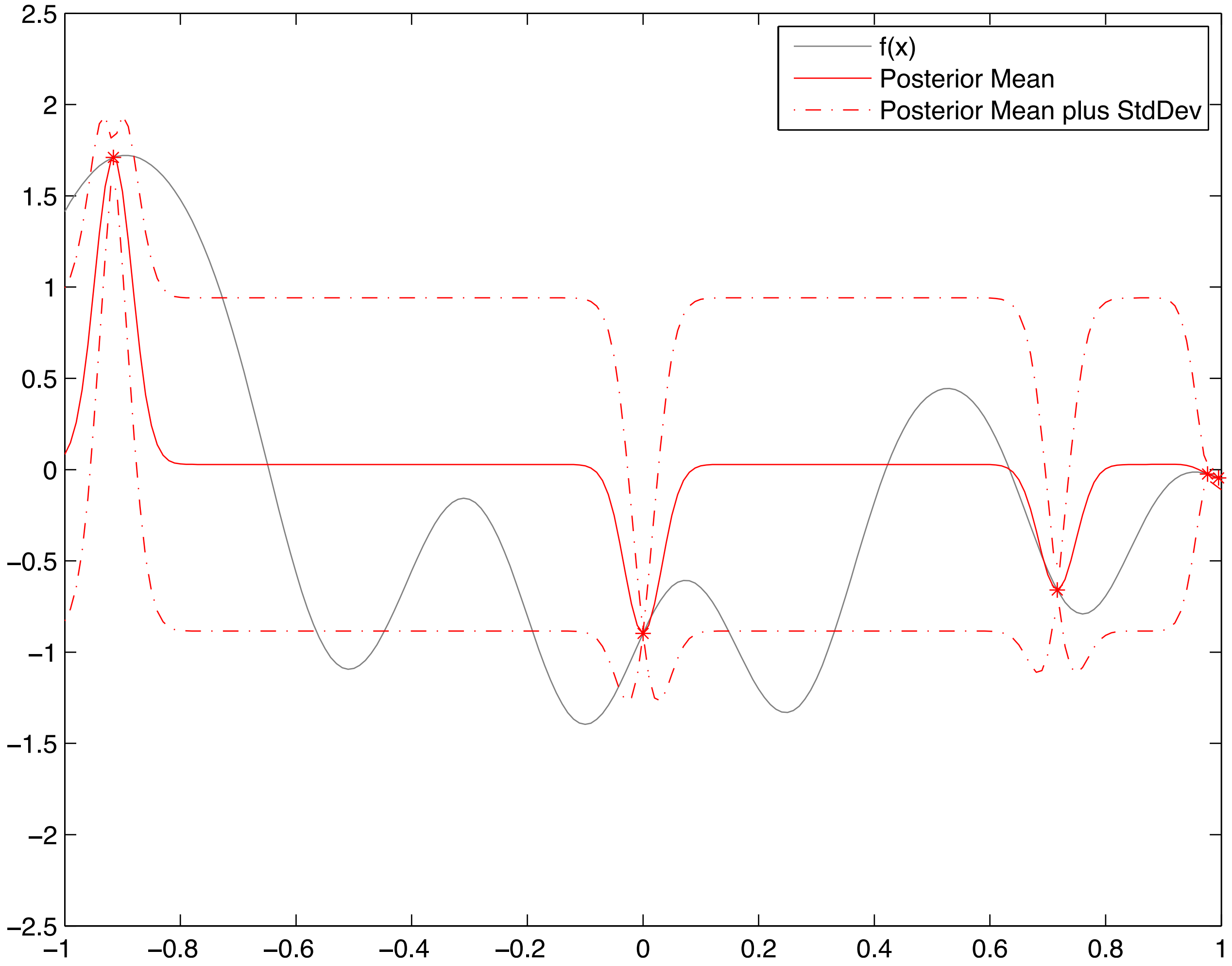
Response Surface Methods



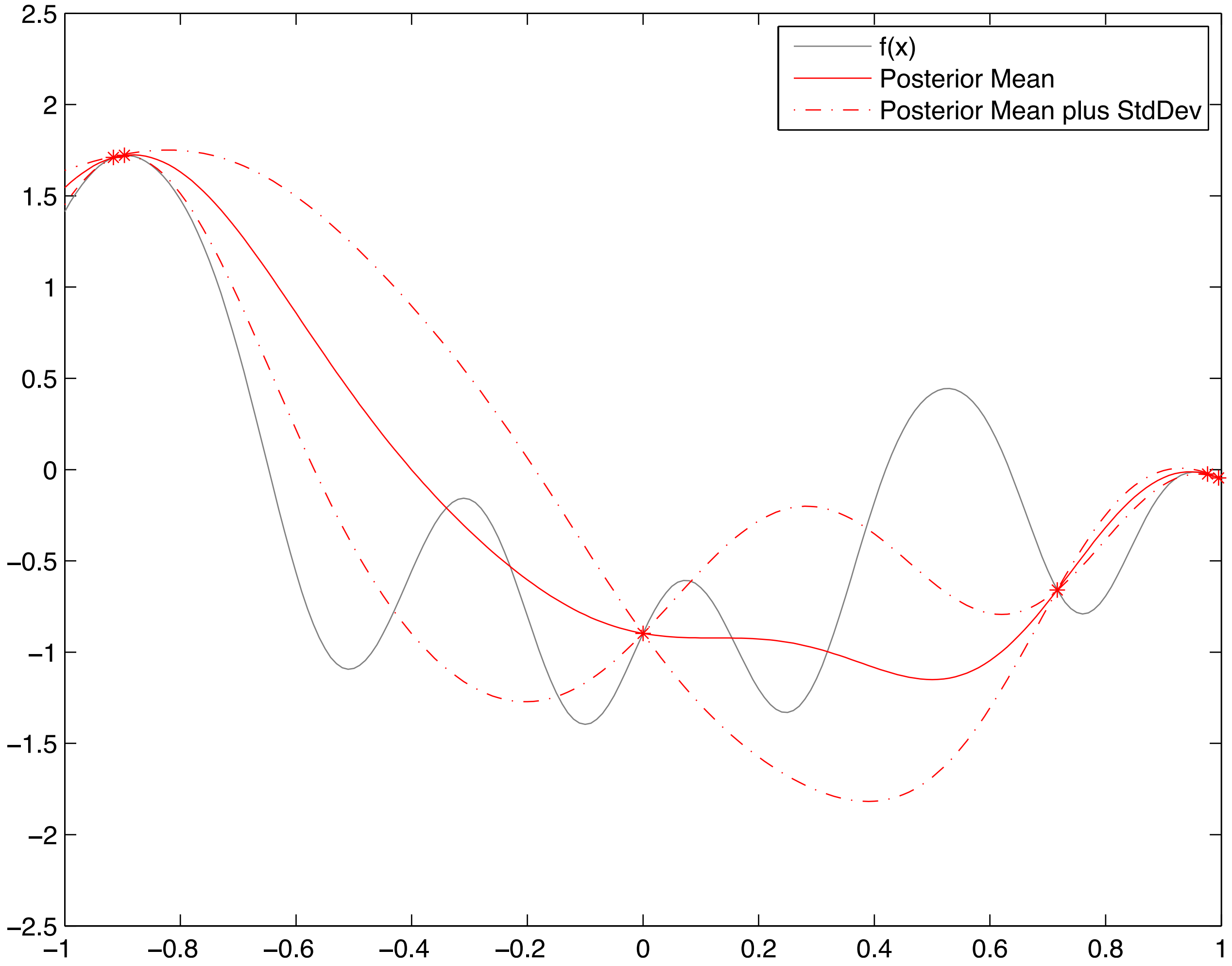
Response Surface Methods



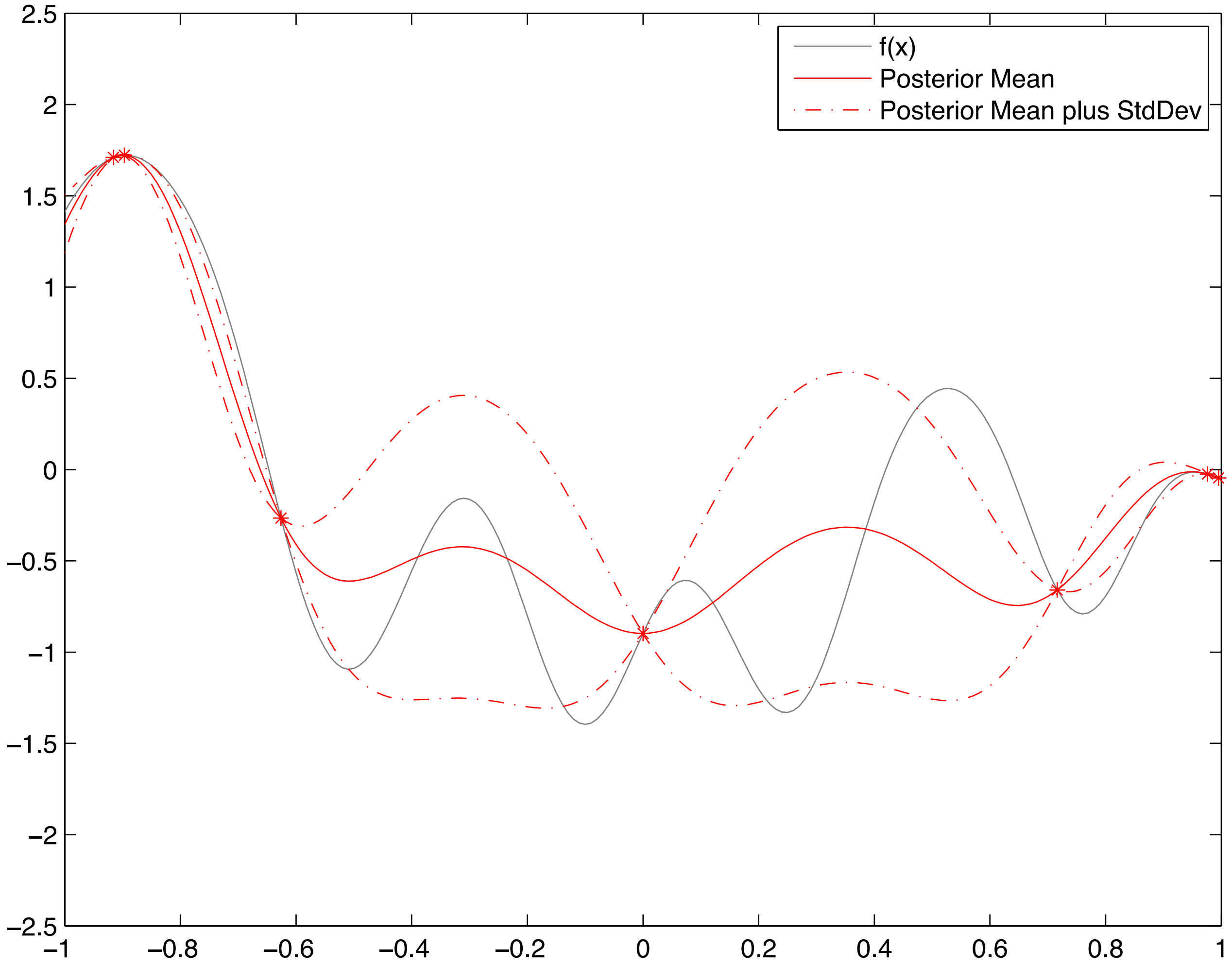
Response Surface Methods



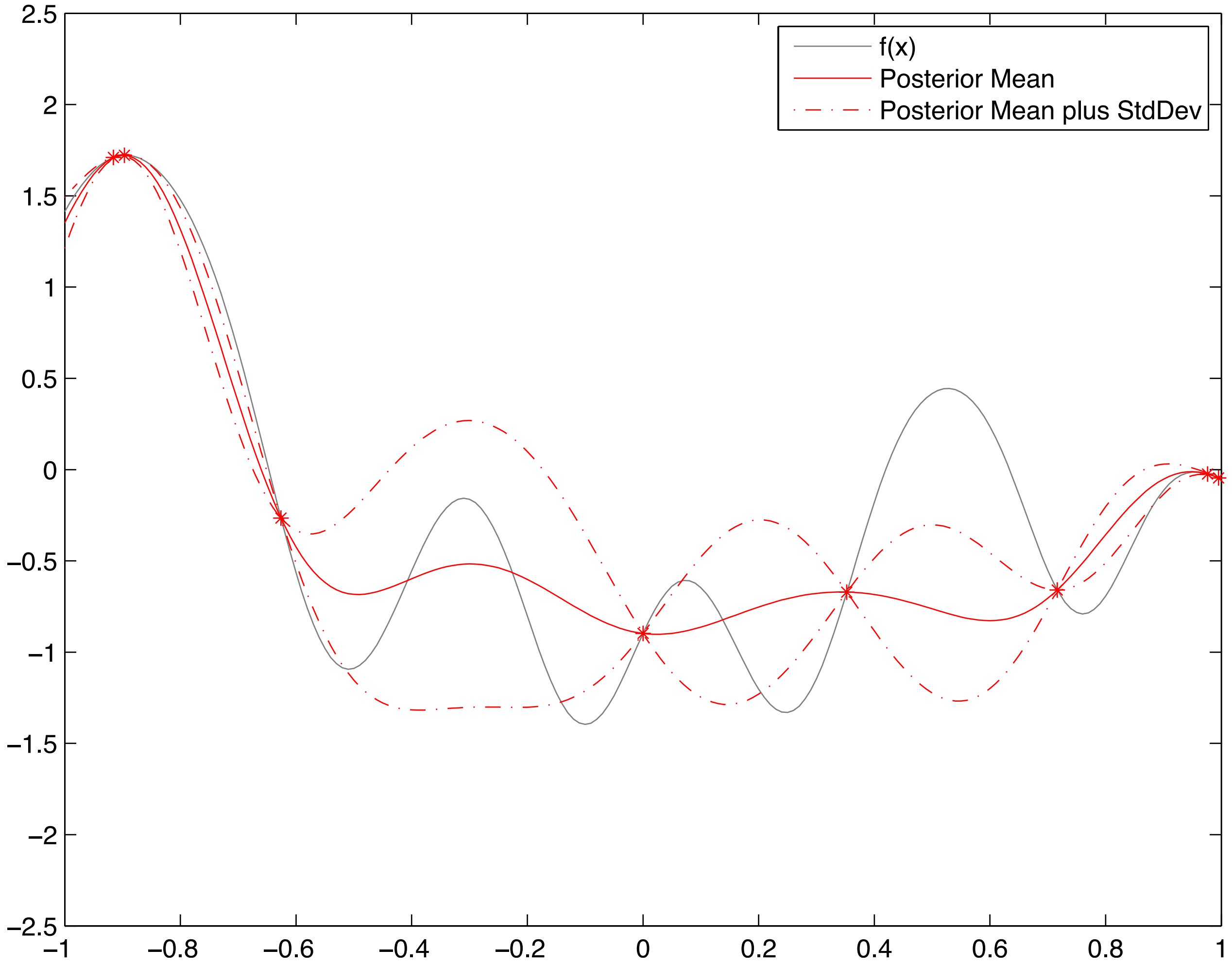
Response Surface Methods



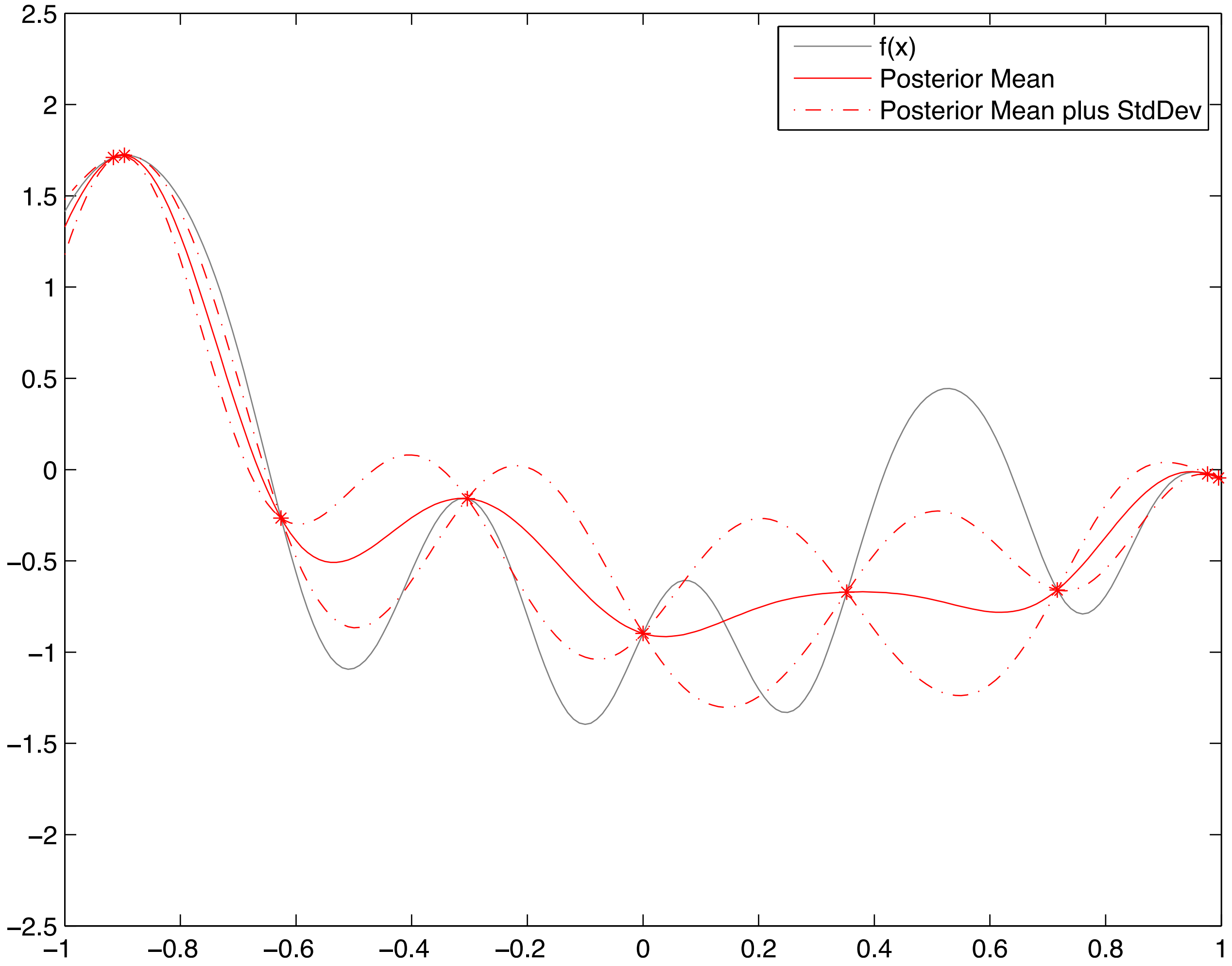
Response Surface Methods



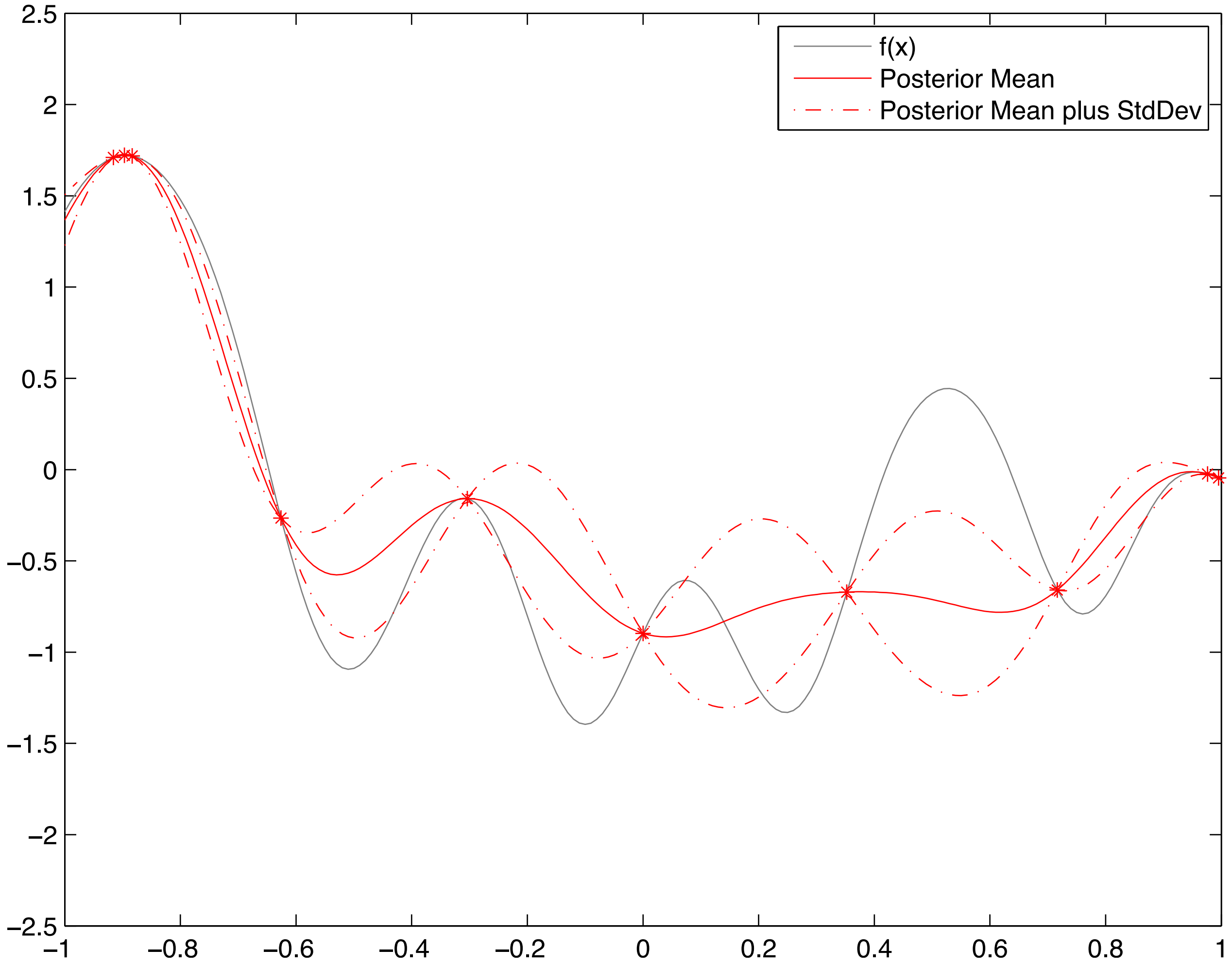
Response Surface Methods



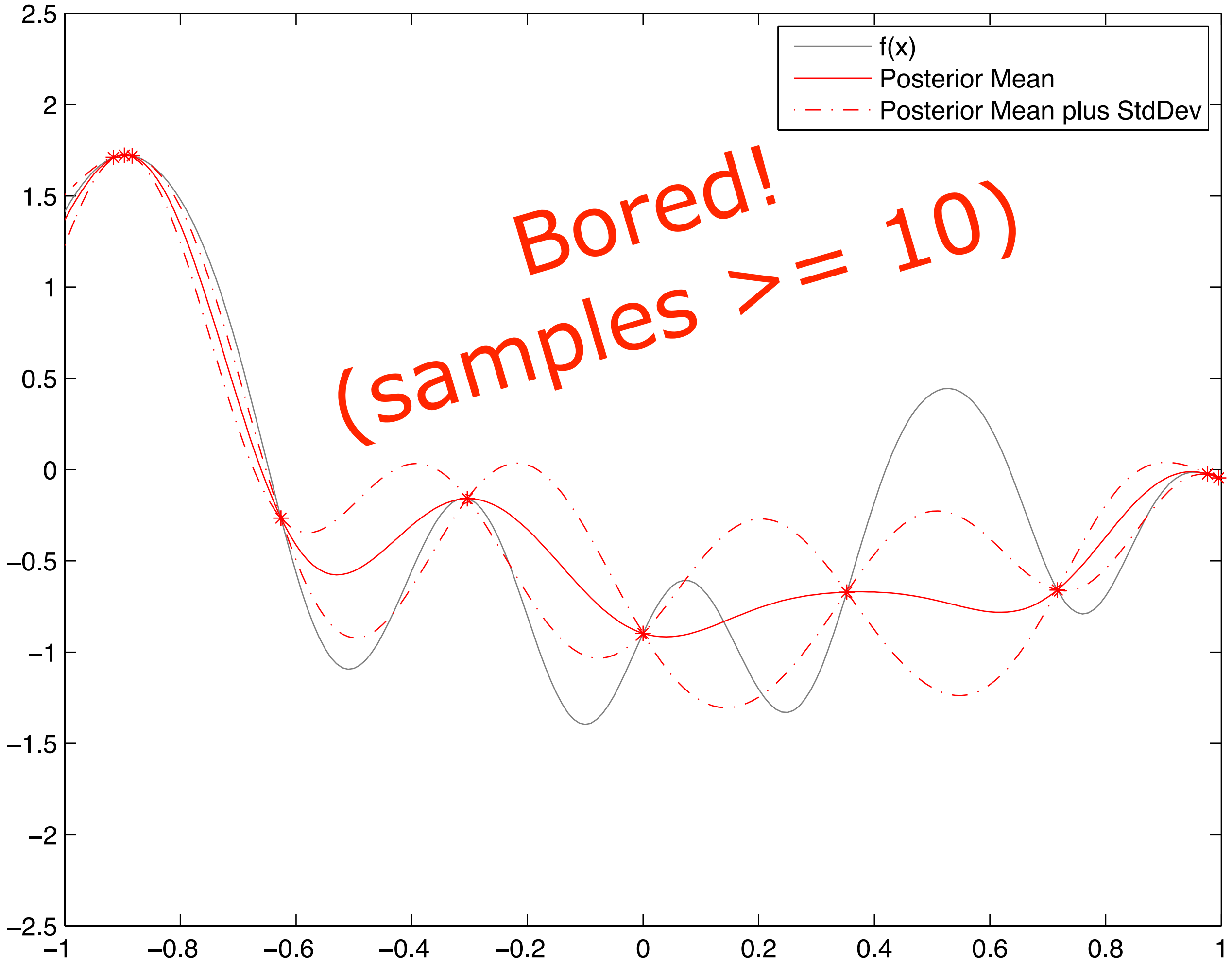
Response Surface Methods



Response Surface Methods



Response Surface Methods



Bored!
(samples >= 10)

Example Application: Robot Gait Optimization

- Gait is controlled by ~12 parameters
- "f(x)" is walk speed at parameters x
  - Expensive - 30s per

Response Surface Model Choice

- We will consider Gaussian process regression
  - Subsumes linear and polynomial regression, Kriging, splines, wavelets, other semi-parametric models…
  - But there are certainly other possible choices
- Still many modeling choices to be made within Gaussian process regression

Gaussian Process Regression

- Bayesian; have prior/posterior over function values
- Posterior of f(z) is a normal random variable $F_z|$Data

$$\begin{aligned}
\mu(F_z|\mathbf{F}_\mathbf{x}) &= \mu_0(\overset{\text{query point}}{z}) + k(z, \overset{\text{domain points}}{\mathbf{x}})k(\mathbf{x},\mathbf{x})^{-1}(\overset{\text{observations}}{\mathbf{f}} - \mu_0(\mathbf{x})) \\
\sigma^2(F_z|\mathbf{F}_\mathbf{x}) &= \underset{\text{scalar}}{k(z,z)} - \underset{\text{1-by-N}}{k(z,\mathbf{x})}\underset{\text{N-by-N}}{k(\mathbf{x},\mathbf{x})^{-1}}\underset{\text{N-by-1}}{k(\mathbf{x},z)}
\end{aligned}$$

Gaussian Process Regression

- The kernel *k*(x,z) gives covariance between F_x and F_z
  - *k*(x,x) can be augmented to accommodate observation noise
- Prior mean μ_0(x) is ‘baseline’

$$
\begin{aligned}
\mu(F_z|\mathbf{F}_\mathbf{x}) &= \mu_0(\overset{\substack{\text{query point}\\ \downarrow}}{z}) + k(z,\overset{\substack{\text{domain points}\\ \downarrow}}{\mathbf{x}})k(\mathbf{x},\mathbf{x})^{-1}(\overset{\substack{\text{observations}\\ \downarrow}}{\mathbf{f}} - \mu_0(\mathbf{x})) \\
\sigma^2(F_z|\mathbf{F}_\mathbf{x}) &= \underset{\substack{\uparrow\\ \text{scalar}}}{k(z,z)} - \underset{\substack{\uparrow\\ \text{1-by-N}}}{k(z,\mathbf{x})}\,\underset{\substack{\uparrow\\ \text{N-by-N}}}{k(\mathbf{x},\mathbf{x})^{-1}}\,\underset{\substack{\uparrow\\ \text{N-by-1}}}{k(\mathbf{x},z)}
\end{aligned}
$$

Example Kernel

$$k(x,z) = \sigma_f \cdot \mathrm{e}^{-\frac{1}{2}\sum_{i=1}^{d}\left(\frac{x_i - z_i}{\ell_i}\right)^2}$$

- Signal variance, length scales are free parameters
- Can use maximum likelihood, MAP, CV, to learn parameters
- Parametric form of *k* is one choice among many

Acquisition Criteria

- Two main criterion choices:
  - MPI - Maximum Probability of Improvement
    - Acquire observation at point x^+ where $f(x^+)$ is most likely to be better than (best_obs + ξ)
  - MEI - Maximum Expected Improvement
    - Acquire observation at point x^+ where the expectation of $[\text{best\_obs} - (F(x^+) + \xi)]_+$ is maximized.
- In both cases, greater ξ means more 'exploration'

Parameters So Far

- Parametric form of kernel function
  - Plus parameter estimation method
- Choice of acquisition criterion
  - Plus choice of ξ

Potential Drawbacks to the Response Surface Approach

- Model choice not obvious
  - Free parameters in the definition of the RS model
- Acquisition criterion not obvious
  - Different proposals, each with free parameters also

How do I choose these for my problem?

- Traditionally, such questions are answered with a small set of test functions
- Choices are adjusted to get reasonable behavior
- Alternative methodology: Use 1000s or 10000s of test functions, not 10s of test functions

Gaussian Process as Generative Model

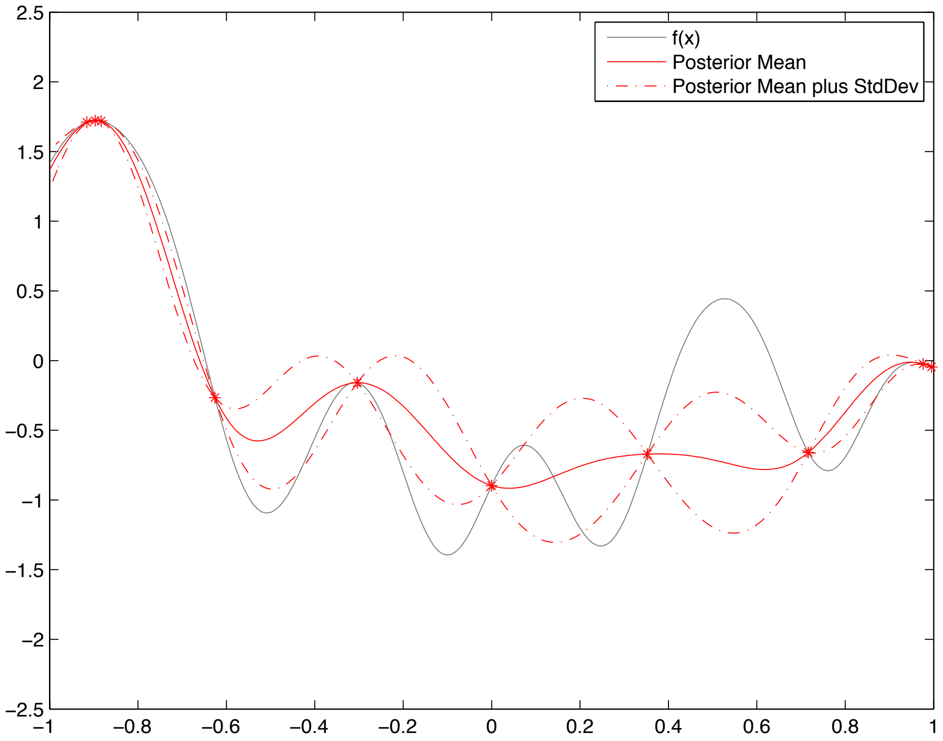
- Can also draw sample functions from this model

$$\mathbf{F}_{\mathbf{x}} \sim \mathcal{N}(\mu_0(\mathbf{x}),\ k(\mathbf{x},\mathbf{x}))$$

- In practice, we take a grid of x, and sample F_x
- In this way, we can sample as many test functions as we wish.
- We hope algorithms designed by testing on many different objective functions will be more robust.

Example

Grey: $\mu_0(x) = 0.00, \quad k(x,z) = 1.0 \cdot \mathrm{e}^{-\frac{1}{2}\left(\frac{x-z}{0.13}\right)^2}$



Red: $\mu_0(x) = 0.14, \quad k(x,z) = 0.77 \cdot \mathrm{e}^{-\left(\frac{x-z}{0.22}\right)^2}$

Simulation Study Goals

We wanted good choices for:

- Kernel parameter learning
  - ML, MAP
- Acquisition criterion
  - MPI, MEI, ξ

Regardless of, or tailored to:

- Signal variance
- Vertical shifting
- Dimension
- Length scales
- Observation budget

- Tests on over 100 000 functions
  Results forthcoming

Acquisition Criterion for Noisy Functions

- MEI - Maximum Expected Improvement
  - Acquire observation at point x^+ where the expectation of [best_obs - $F(x^+)]_+$ is maximized.
    - No concern for producing an accurate estimate of the optimum
- Augmented MEI
  - Huang et al. (2006)
  - Find points that has a large predicted value, but penalize the uncertainty in that value
  - Introduces yet another parameter *c*

How do I pick *c*?

How do I pick *c*?

- Authors chose $c = 1.0$, ran test on 5 functions
- Results look encouraging

How do I pick *c*?

- Authors chose *c* = 1.0, ran test on 5 functions
- Results look encouraging

- We can apply our test problem generation strategy to explore the relationship between
  - Test model parameters
  - New parameter *c*
  - Measures of algorithm performance

Summary

- Response Surface optimization seems well-suited to optimizing noisy functions
- Most work to date has focussed on deterministic functions
- Good ideas for the noisy case, but perhaps under-explored
- Our evaluation methodology can help to more rigorously identify where RS algorithms will work and not work

Thank you

- Dan Lizotte, danjl@umich.edu

- Supported in part by NSERC of Canada and US NIH grants R01 MH080015 and P50 DA10075

- C. E. Rasmussen and C. K. I. Williams. Gaussian Processes for Machine Learning. MIT Press, 2006a.
- Daniel J. Lizotte, Russell Greiner, Dale Schuurmans. An Experimental Methodology for Response Surface Optimization Methods. (e-mail Dan)
- D. Huang, T. T. Allen, W. I. Notz, and N. Zeng. Global optimization of stochastic black-box systems via sequential kriging meta-models. Journal of Global Optimization, 34:441–466, 2006.