New One-Step Bayes-Optimal Algorithms for Global Optimization: Parallel Computing and Common Random Numbers

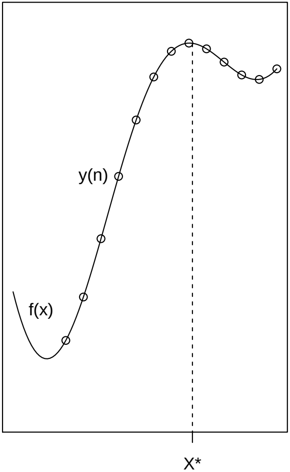
Peter I. Frazier

School of Operations Research & Information Engineering, Cornell University

April 18-20, 2012
AFOSR Optimization and Discrete Mathematics Program Review
Arlington, VA

Supported by AFOSR YIP FA9550-11-1-0083

Derivative-Free Black-box Global Optimization



- Objective function $f : \mathbb{R}^d \mapsto \mathbb{R}$, continuous but not concave.
- Feasible set $A \subseteq \mathbb{R}^d$.
- Our goal is to solve

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

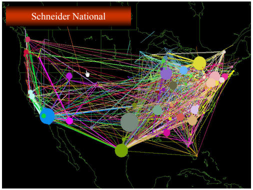
- Assumptions: f is time-consuming to evaluate (hours or days), and derivative information is unavailable.

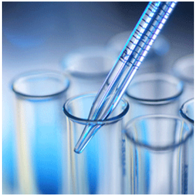
This is joint work with Scott C. Clark (Cornell PhD student).

Bayesian Global Optimization is a class of methods for Derivative-Free Black-Box Global Optimization

- One class of methods for derivative-free black-box global optimization is the class of **Bayesian Global Optimization (BGO)** methods.
- In these methods, we place a **Bayesian prior distribution** on the objective function f. (This is typically a Gaussian process prior).
- Ideally, we would find an algorithm with optimal average-case performance under this prior.
- We will settle for an algorithm with good average-case performance.
- (There are many other types of DFO methods. We do not discuss these in this talk.)

BGO is useful for optimizing computational models and physical experiments





- BGO is often used for optimizing large-scale computational models.
  - Example: Design of grafts to be used in heart surgery. [Yang et al., 2010]
  - Example: Calibration of a simulation-based logistics model. [Frazier et al., 2009b].
- BGO can also be used for optimization problems where "evaluating the objective function" means running a physical experiment
  - Example: Optimizing the concentrations of chemicals used to manufacture a material.
  - (Typically, physical experiments are noisy. We do not consider noise in this talk.)

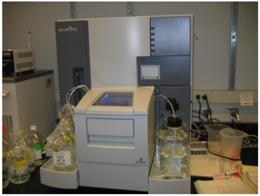
Almost all existing BGO methods are sequential

- Early work: [Kushner, 1964, Mockus et al., 1978, Mockus, 1989]
- Convergence analysis: [Calvin, 1997, Calvin and Zilinskas, 2002, Vazquez and Bect, 2010].
- Perhaps the most well-known method is Efficient Global Optimization (EGO) [Schonlau, 1997, Jones et al., 1998], which uses the notion of expected improvement.
- Recently many metheds have been developed that allow noise: [Calvin and Zilinskas, 2005, Villemonteix et al., 2009, Frazier et al., 2009a, Huang et al., 2006]

These methods are all fully sequential (one function evaluation at a time).

How can we extent BGO to multiple simultaneous function evaluations?

Cornell Tardis Cluster



BIAcore machine

- What if we can perform multiple function evaluations simultaneously?
- This is the case with parallel computing, and in many experimental settings (particularly in biology).
- We explore an idea that follows naturally from a decision-theoretic analysis.
- This idea was previously suggested by [Ginsbourger et al., 2007].

We generalize to multiple function evaluations using a decision-theoretic approach

- We've evaluated $\vec{x}^{(1)}, \dots, \vec{x}^{(n)}$, and observed $f(\vec{x}^{(1)}), \dots, f(\vec{x}^{(n)})$.
- Once sampling stops, we will select the best point evaluated so far.
- **What would be the Bayes-optimal way to choose the set of points $\vec{x}_1, \dots, \vec{x}_q$ to evaluate next?**
- In general, the optimal points are given by the solution to a dynamic program. (Difficult to solve)
- When this is the last stage of measurements, the dynamic program becomes a more straightforward optimization problem.

Generalizing EI to multiple function evaluations

- We've evaluated $\vec{x}^{(1)}, \ldots, \vec{x}^{(n)}$, and observed $f(\vec{x}^{(1)}), \ldots, f(\vec{x}^{(n)})$.
- Let $f_n^* = \max_{m=1,\ldots,n} f(\vec{x}_m)$ be the best value observed so far.
- If we measure at new points $\vec{x}_1, \ldots, \vec{x}_q$, and if that is our last stage of measurements, then the expected value of our solution is

$$\mathbb{E}_n \left[\max\left(f_n^*, \max_{i=1,\ldots,q} f(\vec{x}_i) \right) \right]$$

- This can be rewritten as $\mathrm{EI}_n(\vec{x}_1, \ldots, \vec{x}_q) + f_n^*$ where

$$\mathrm{EI}_n(\vec{x}_1, \ldots, \vec{x}_q) = \mathbb{E}_n \left[\left(\max_{i=1,\ldots,q} f(\vec{x}_i) - f_n^* \right)^+ \right]$$

  is given the name **q-EI** (and also "multipoints expected improvement") by [Ginsbourger et al., 2007].

q-EI gives the single-stage Bayes-optimal set of evaluations

- If we have one stage of function evaulations left to take, and must take our final solution from the set of points that have been evaluated, then evaluating

$$\arg\max_{\vec{x}_1,\ldots,\vec{x}_q} \mathrm{EI}_n(\vec{x}_1,\ldots,\vec{x}_q)$$

  is **Bayes optimal**, i.e., optimal with respect to average case performance under posterior.
- If we have more than one stage left to go, it is a heuristic.

q-EI has no general closed form expression

$\mathrm{EI}_n(\vec{x}_1, \ldots, \vec{x}_q) = \mathbb{E}_n\left[\left(\max_{i=1,\ldots,q} f(\vec{x}_i) - f_n^*\right)^+\right]$

- When $q = 1$ (no parallelism), this reduces to the expected improvement of [Jones et al., 1998], which has a closed form.
- When $q = 2$, [Ginsbourger et al., 2007] provides an expression in terms of bivariate normal cdfs.
- When $q > 2$, there is no analytic expression.
  [Ginsbourger et al., 2007] proposes estimation through Monte Carlo.

q-EI is hard to optimize

- From [Ginsbourger, 2009], "*directly optimizing the q-EI becomes extremely expensive as q and d (the dimension of inputs) grow.*"
- Rather than actually solving $\arg\max_{\vec{x}_1,\ldots,\vec{x}_q} \mathrm{EI}(\vec{x}_1,\ldots,\vec{x}_q)$ when $q > 2$, [Ginsbourger et al., 2007] proposes other heuristic schemes.

Our Contribution

- Our contribution is an efficient method for solving

$$\arg\max_{\vec{x}_1,\ldots,\vec{x}_q} \mathrm{EI}(\vec{x}_1,\ldots,\vec{x}_q)$$

- This transforms the Bayes optimal function evaluation plan, previously considered to be a purely conceptual algorithm, into something implementable.

Our approach to solving $\arg\max_{\vec{x}_1,\ldots,\vec{x}_q} \mathrm{EI}(\vec{x}_1,\ldots,\vec{x}_q)$

1. Construct an unbiased estimater of

$$\nabla \mathrm{EI}(\vec{x}_1,\ldots,\vec{x}_q)$$

   using infinitessimal perturbation analysis (IPA).

2. Use multistart stochastic gradient ascent to find an approximate solution to $\max_{\vec{x}_1,\ldots,\vec{x}_q} \mathrm{EI}(\vec{x}_1,\ldots,\vec{x}_q)$.

We construct an estimator of the gradient

- Using sufficient conditions described on the next slide, we switch ∇ and expectation to obtain our unbiased estimator of the gradient,

$$\begin{aligned}\nabla \mathrm{EI}(\vec{x}_1,\ldots,\vec{x}_q) &= \nabla \mathbb{E}_n\left[\left(\max \vec{\mu}_n(\vec{x}_1,\ldots,\vec{x}_q) + C_n(\vec{x}_1,\ldots,\vec{x}_q)\vec{Z} - f_n^*\right)^+\right] \\ &= \mathbb{E}_n\left[g(\vec{x}_1,\ldots,\vec{x}_q,\vec{Z})\right],\end{aligned}$$

  where

$$g(\vec{x}_1,\ldots,\vec{x}_q,\vec{Z}) = \nabla\left(\max \vec{\mu}_n(\vec{x}_1,\ldots,\vec{x}_q) + C_n(\vec{x}_1,\ldots,\vec{x}_q)\vec{Z} - f_n^*\right)^+$$

  when this gradient exists, and 0 otherwise.
- $g(\vec{x}_1,\ldots,\vec{x}_q,\vec{Z})$ can be computed using results from [Smith, 1995] on differentiation of the Cholesky decomposition.

Our gradient estimator is unbiased, given sufficient conditions

Theorem

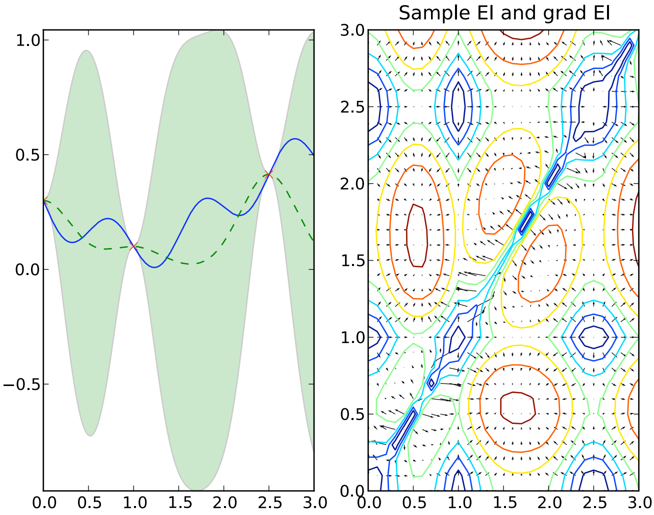
If the following conditions hold

- $\vec{\mu}_n(\vec{x}_1,\ldots,\vec{x}_q)$ *and* $\Sigma_n(\vec{x}_1,\ldots,\vec{x}_q)$ *are continuously differentiable in a neighborhood of* $\vec{x}_1,\ldots,\vec{x}_q$
- $\vec{x}_i \neq \vec{x}_j$ *for all* $i \neq j$*. (Don't propose measuring the same point twice)*
- $\vec{x}_i \neq x^{(j)}$ *for all* i,j*. (Don't measure a previously measured points)*
- $P_n(f(x') \neq f(x)) = 1$ *for all* $x' \neq x$ *and all* $x \notin \{\vec{x}^{(1)},\ldots,\vec{x}^{(n)}\}$ *(Our model is not degenerate)*

then

$$\nabla \mathrm{EI}(\vec{x}_1,\ldots,\vec{x}_q) = \mathbb{E}_n\left[g(\vec{x}_1,\ldots,\vec{x}_q,\vec{Z})\right].$$

Example of Estimated Gradient

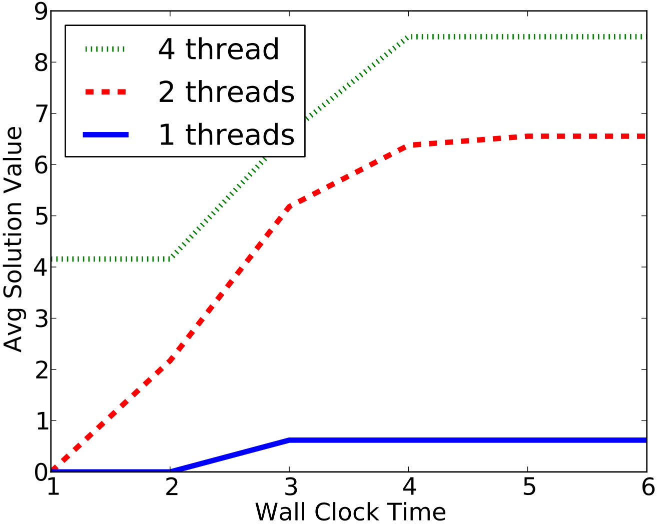


We can handle asynchronous function evaluations

- As previously described, if there are no function evaluations currently in progress, we solve
$$\max_{\vec{x}_1,\ldots,\vec{x}_q} \mathrm{EI}(\vec{x}_1,\ldots,\vec{x}_q)$$
to get the set to run next.
- If there are function evaluations already in progress, say $\vec{x}_1,\ldots,\vec{x}_k$, we take these as given and optimize the rest $\vec{x}_{k+1},\ldots,\vec{x}_q$.

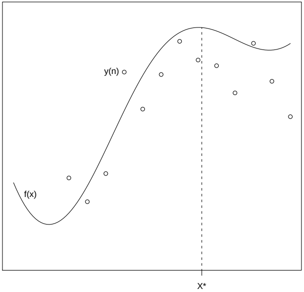
$$\max_{\vec{x}_{k+1},\ldots,\vec{x}_q} \mathrm{EI}(\vec{x}_1,\ldots,\vec{x}_q)$$

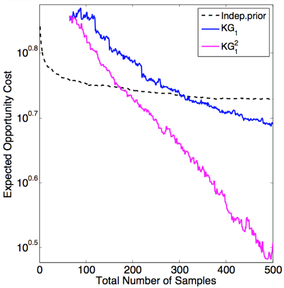
Initial Results



Other Work: Stochastic Global Optimization Using Common Random Numbers

Maximize a noisy non-concave function in $\mathbb{R}^d$. Use common random numbers to reduce the variance of comparisons.

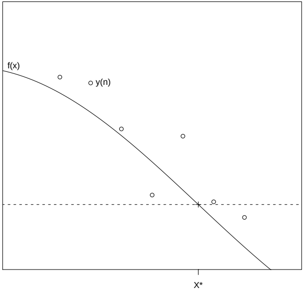


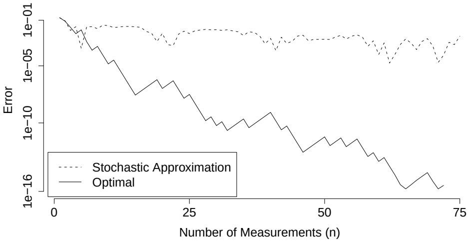


- P.I. Frazier, J. Xie, and S.E. Chick, "Bayesian Optimization via Simulation with Correlated Sampling and Correlated Prior Beliefs," Winter Simulation Conference, 2011.
- J. Xie, P.I. Frazier and S.E. Chick, "Large-Scale Bayesian Optimization via Simulation with Common Random Numbers," in preparation.

Other Work: Stochastic Root Finding

Find the root of a monotone function whose values can only be measured with noise. Use as few function evaluations as possible.

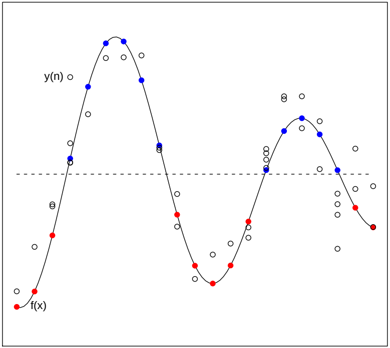


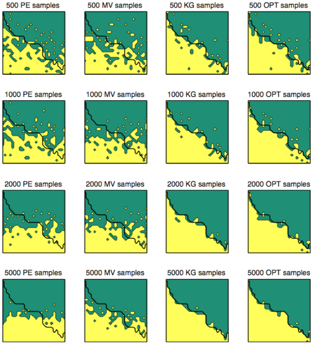


- R. Waeber, P.I. Frazier and S.G. Henderson, "Bisection Search with Noisy Responses." in review at SIAM Journal on Control and Optimization.
- R. Waeber, P.I. Frazier, and S.G. Henderson, "A Bayesian Approach to Stochastic Root-Finding," Winter Simulation Conference, 2011. Winner of the Best Student Paper (OR/MS focused).

Other Work: Finding Level Sets

Identify those points whose value $f(x)$ is above a threshold. Use as few function evaluations as possible.

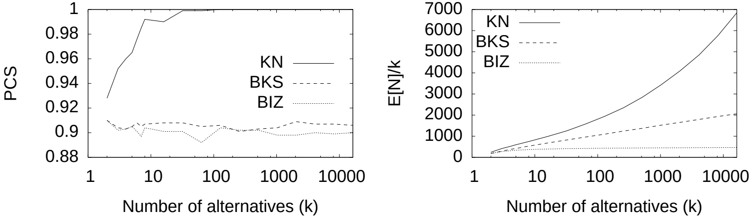




- J. Xie and P.I. Frazier, "Sequential Bayes-Optimal Policies for Multiple Comparisons with a Control." in revision for Operations Research. Finalist, INFORMS Junior Faculty Interest Group (JFIG) Paper Competition, 2011.

Other Work: Ranking & Selection

Find the alternative in a finite discrete set with the largest sampling mean. Do so with a guarantee on probability of correct selection, while taking as few samples as possible.



We provide the first fully sequential elimination procedure with tight bounds on probability of correct selection.

- P.I. Frazier, "A Fully Sequential Elimination Procedure for Indifference-Zone Ranking and Selection with Tight Bounds on Probability of Correct Selection." in revision for Operations Research.

References I

Calvin, J. M. (1997).
Average performance of a class of adaptive algorithms for global optimization.
The Annals of Applied Probability, 7(3):711–730.

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

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

Frazier, P. I., Powell, W. B., and Dayanik, S. (2009a).
The Knowledge Gradient Policy for Correlated Normal Beliefs.
INFORMS Journal on Computing, 21(4):599–613.

Frazier, P. I., Powell, W. B., and Simão, H. P. (2009b).
Simulation Model Calibration with Correlated Knowledge-Gradients.
In *Winter Simulation Conference Proceedings, 2009*. Winter Simulation Conference.

Ginsbourger, D. (2009).
Two advances in Gaussian Process-based prediction and optimization for computer experiments.
In *MASCOT09 Meeting*, pages 1–2.

Ginsbourger, D., Le Riche, R., and Carraro, L. (2007).
A Multi-points Criterion for Deterministic Parallel Global Optimization based on Kriging.
In *Intl. Conf. on Nonconvex Programming, NCP07*, page ..., Rouen, France.

References II

Huang, D., Allen, T. T., Notz, W. I., and Zeng, N. (2006).
Global Optimization of Stochastic Black-Box Systems via Sequential Kriging Meta-Models.
Journal of Global Optimization, 34(3):441–466.

Jones, D. R., Schonlau, M., and Welch, W. J. (1998).
Efficient Global Optimization of Expensive Black-Box Functions.
Journal of Global Optimization, 13(4):455–492.

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.

Mockus, J., Tiesis, V., and Zilinskas, A. (1978).
The application of Bayesian methods for seeking the extremum.
In Dixon, L. C. W. and Szego, G. P., editors, *Towards Global Optimisation*, volume 2, pages 117–129. Elsevier Science Ltd., North Holland, Amsterdam.

Schonlau, M. (1997).
Computer experiments and global optimization.
PhD thesis, University of Waterloo.

Smith, S. (1995).
Differentiation of the Cholesky algorithm.
Journal of Computational and Graphical Statistics, pages 134–147.

References III

Vazquez, E. and Bect, J. (2010).
Convergence properties of the expected improvement algorithm with fixed mean and covariance functions.
Journal of Statistical Planning and Inference, 140(11):3088–3095.

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.

Yang, W., Feinstein, J. A., and Marsden, A. L. (2010).
Constrained optimization of an idealized Y-shaped baffle for the Fontan surgery at rest and exercise.
Computer methods in applied mechanics and engineering, 199(33-36):2135–2149.

Backup

The Gradient Estimator

- We can rewrite our gradient estimator $g(\vec{x}_1, \ldots, \vec{x}_q, \vec{Z})$ more clearly.
- Let e_* be the unit vector corresponding to the maximal strictly positive component of

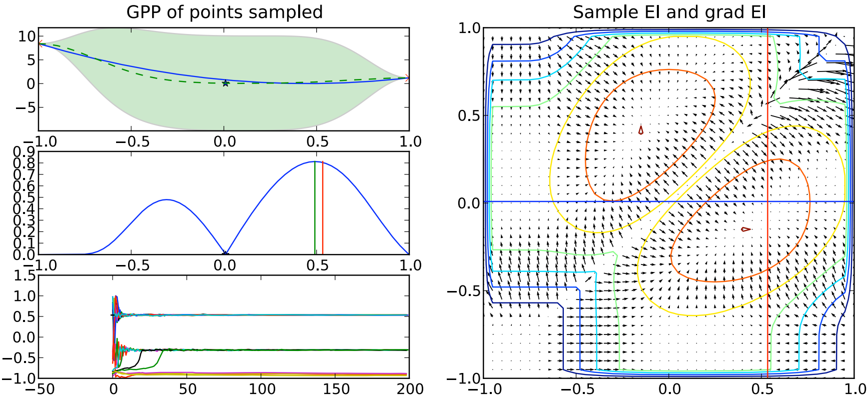
$$\vec{\mu}_n(\vec{x}_1, \ldots, \vec{x}_q) + C_n(\vec{x}_1, \ldots, \vec{x}_q)\vec{Z} - f_n^*,$$

  or 0 if all components are non-negative.
- Then,

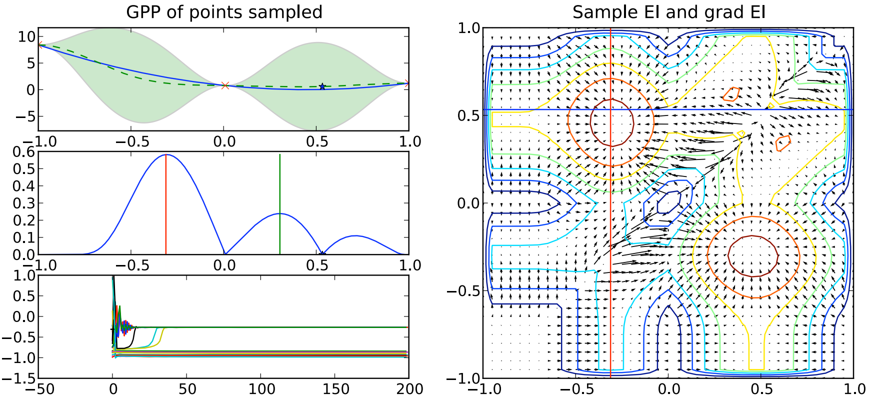
$$g(\vec{x}_1, \ldots, \vec{x}_q, \vec{Z}) = \nabla \left[e_* \vec{\mu}_n(\vec{x}_1, \ldots, \vec{x}_q) + e_* C_n(\vec{x}_1, \ldots, \vec{x}_q)\vec{Z} \right]$$

- $g(\vec{x}_1, \ldots, \vec{x}_q, Z)$ can be computed using results from [Smith, 1995] on differentiation of the Cholesky decomposition.

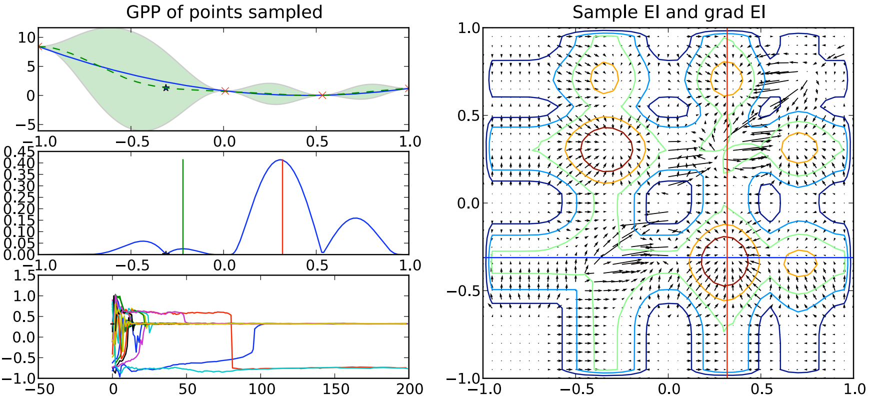
Animation



Animation



Animation


GPP of points sampled
10
5
0
−5
−1.0
−0.5
0.0
0.5
1.0
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
1.5
0.5
−0.5
−50
50
100
150
200
Sample EI and grad EI

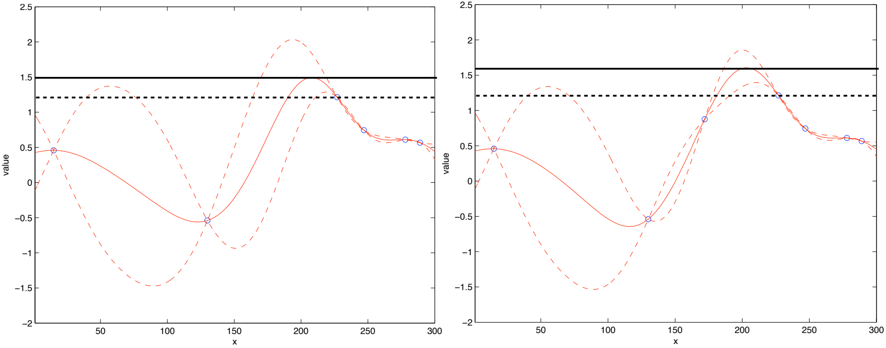
Best estimated overall value might be at an unmeasured point

- The improvement considered by EI is:

$$[f(x) - f_n^*]^+ = \max(f(x), f_n^*) - f_n^* = f_{n+1}^* - f_n^*$$

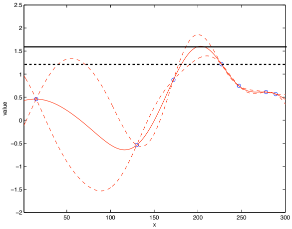
where $f_n^* = \max_{m \leq n} f(x_m)$ is the best point we've measured by time n.

- But the point with the best estimated value might not be a point we've measured.



We can measure improvement w.r.t. the best overall value

$$\text{Replace } f_n^* = \max_{m \leq n} f(x_m) = \max_{m \leq n} \mu_n(x_m) \text{ with } \mu_n^* = \max_{x \in A} \mu_n(x).$$



- The corresponding improvement is $\mu_{n+1}^* - \mu_n^*$.
- The corresponding value for taking a sample is

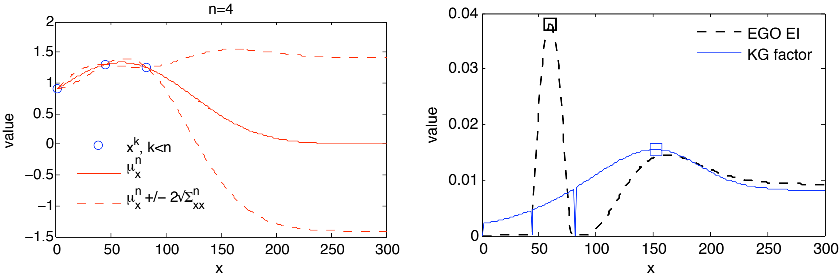
$$\mathbb{E}_n \left[\mu_{n+1}^* - \mu_n^* \mid x_{n+1} = x \right].$$

- The policy that measures at the x with the largest such value is called the **knowledge-gradient with correlated beliefs** (KGCB) policy.

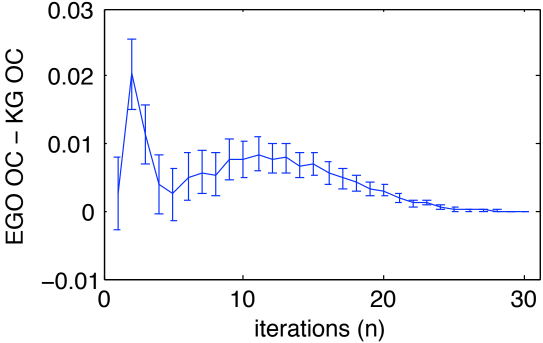
Knowledge-Gradient with Correlated Beliefs (KGCB)

- Call this modified expected improvement the knowledge-gradient (KG) factor
$$\mathrm{KG}_n(x) = \mathbb{E}_n\left[\mu^*_{n+1} - \mu^*_n \mid x_{n+1} = x\right].$$
- The KGCB policy measures at the point with the largest KG factor.

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



KGCB Requires Fewer Function Evaluations than EGO, but More Computation



- Graph shows the difference in expected solution quality between KGCB and EGO, on noise-free problems.
- KGCB needs fewer function evaluations to find a good solution, but more computation to decide where to evaluate.