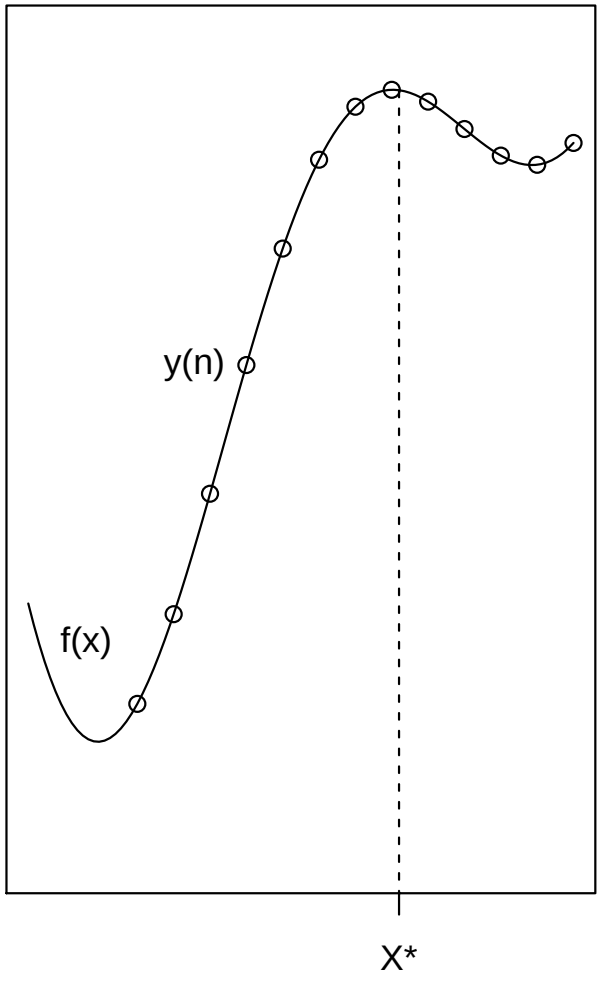
Parallel Global Optimization Using An Improved Multi-points Expected Improvement Criterion

Scott C. Clark, Peter I. Frazier

Center for Applied Math, Cornell University
School of Operations Research & Information Engineering, Cornell University

Saturday February 25, 2012
INFORMS Optimization Society Conference
University of Miami

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

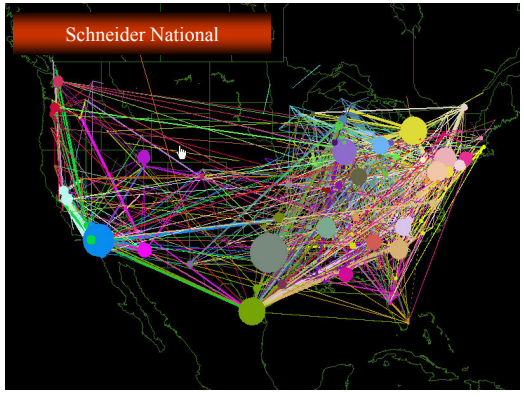
$$\max_{x \in A} f(x)$$

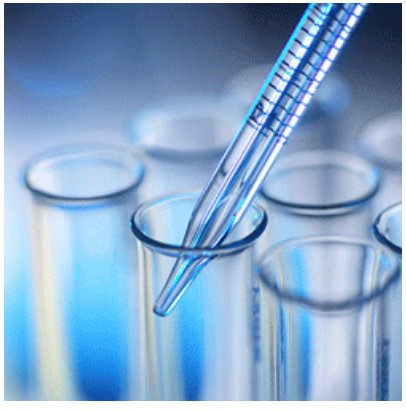
- Assumptions: f is time-consuming to evaluate (hours or days), derivative information or other structural information is unavailable.

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).
- Our goal is to design an algorithm with good average-case performance under this prior.
- (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.)

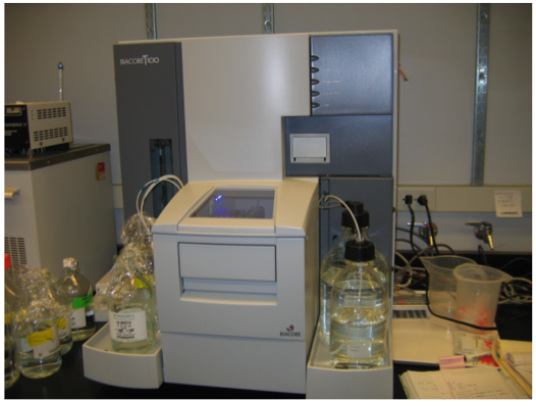
Almost all existing BGO methods are sequential

- 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).
- An idea from [Ginsbourger et al., 2007, Ginsbourger et al., 2010] is to
  - (1) calculate an expected improvement for multiple points evaluated in parallel
  - (2) choose sets of points to evaluate by optimizing this multipoints expected improvement.

Multi-points expected improvement [Ginsbourger et al., 2007]

- 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 $\vec{x}_1, \ldots, \vec{x}_q$, the **improvement** in our objective function is

$$\left(\max_{i=1,\ldots,q} f(\vec{x}_i) - f_n^* \right)^+$$

- The expected improvement is

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

  where $\mathbb{E}_n$ is taken with respect to the time-n posterior distribution (which is multivariate normal).
- A natural algorithm would be to evaluate, at time n,

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

Multipoints Expected Improvement 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 reasonable heuristic.

Multipoints Expected Improvement has no general closed form expression

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

- 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.

Multipoints Expected Improvement 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 $\max_{\vec{x}_1,\ldots,\vec{x}_q} \mathrm{EI}(\vec{x}_1,\ldots,\vec{x}_q)$ when $q > 2$, [Ginsbourger et al., 2007] proposes other heuristic schemes.

Our Contribution

- Our contribution is an efficient method for solving

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

- This transforms the Bayes optimal function evaluation plan, suggested as a purely conceptual algorithm by [Ginsbourger et al., 2007], into something implementable.

Our Approach

1. Construct an unbiased estimater of

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

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

Constructing the Gradient Estimator

The posterior distribution of f evaluated at $\vec{x}_1, \ldots, \vec{x}_q$ is

$$\begin{bmatrix} f(\vec{x}_1) \\ \vdots \\ f(\vec{x}_q) \end{bmatrix} \sim N\left(\vec{\mu}_n(\vec{x}_1, \ldots, \vec{x}_q), \Sigma_n(\vec{x}_1, \ldots, \vec{x}_q)\right),$$

where $\vec{\mu}_n(\cdot)$ and $\Sigma_n(\cdot)$ can be calculated in closed form via standard results in Gaussian process regression.

Constructing the Gradient Estimator

- Recall $[f(\vec{x}_1), \ldots, f(\vec{x}_q)]^T \sim N(\vec{\mu}_n(\vec{x}_1, \ldots, \vec{x}_q), \Sigma_n(\vec{x}_1, \ldots, \vec{x}_q))$.
- Let $C_n(\vec{x}_1, \ldots, \vec{x}_q)$ be the Cholesky decomposition of $\Sigma_n(\vec{x}_1, \ldots, \vec{x}_q)$.
- Let $\vec{Z}$ be a q-dimensional standard normal random vector.
- Then $[f(\vec{x}_1), \ldots, f(\vec{x}_q)]^T$ is equal in distribution to

$$\vec{\mu}_n(\vec{x}_1, \ldots, \vec{x}_q) + C_n(\vec{x}_1, \ldots, \vec{x}_q)\vec{Z}.$$

Constructing the Gradient Estimator

Using this equality in distribution, we can write,

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

Constructing the Gradient Estimator

- We switch ∇ and expectation to obtain our unbiased estimator of the gradient,

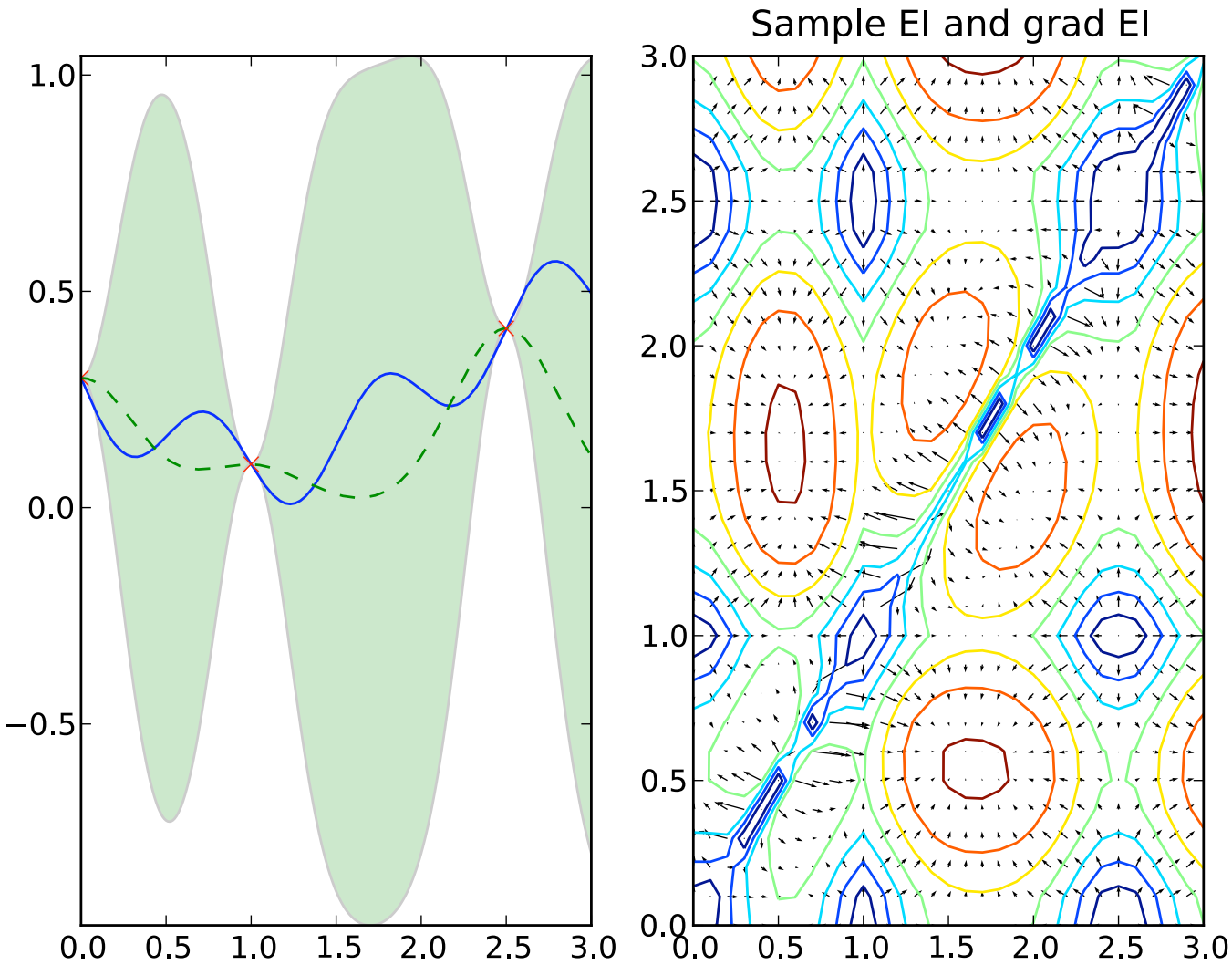
$$\nabla \mathrm{EI}(\vec{x}_1, \ldots, \vec{x}_q) = \nabla \mathbb{E}_n \left[\left(\max \vec{\mu}_n(\vec{x}_1, \ldots, \vec{x}_q) + C_n(\vec{x}_1, \ldots, \vec{x}_q)\vec{Z} - f_n^* \right)^+ \right]$$
$$= \mathbb{E}_n \left[g(\vec{x}_1, \ldots, \vec{x}_q, \vec{Z}) \right],$$

  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. The gradient exists with probability 1.
- We have equality when switching ∇ and $\mathbb{E}_n$ by sufficient conditions [L'Ecuyer, 1990] from infinitessimal perturbation analysis (so far, only checked for a GP prior with mean 0 and a squared exponential kernel).
- $g(\vec{x}_1, \ldots, \vec{x}_q, Z)$ can be computed using results from [Smith, 1995] on differentiation of the Cholesky decomposition.

Example of Estimated Gradient



Our Approach

1. Construct an unbiased estimater of

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

2. Use multistart stochastic gradient ascent to find an approximate solution to $\max_{\vec{x}_1,\ldots,\vec{x}_q} \mathrm{EI}(\vec{x}_1, \ldots, \vec{x}_q)$.
   - Multistart stochastic gradient ascent is a standard simulation optimization technique for finding good extrema, using unbiased estimates of the gradient
     [Robbins and Monro, 1951, Blum, 1954, Kushner and Yin, 1997].

This approach can handle asynchronous function evaluations

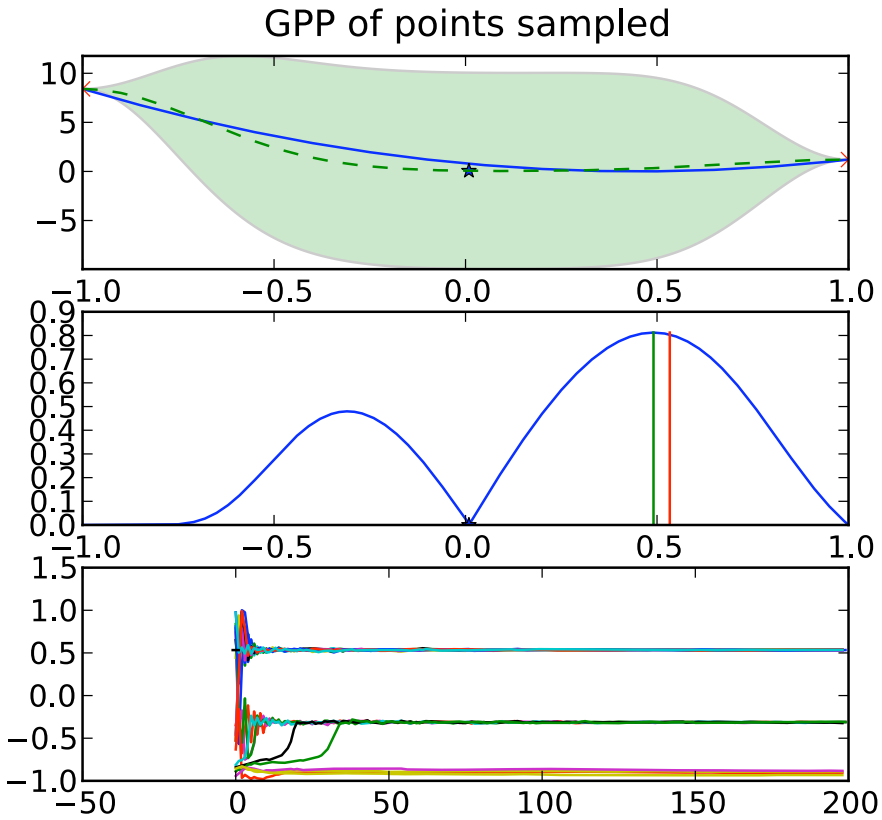
- As previously described, if there are no function evaluations currently in progress, we solve

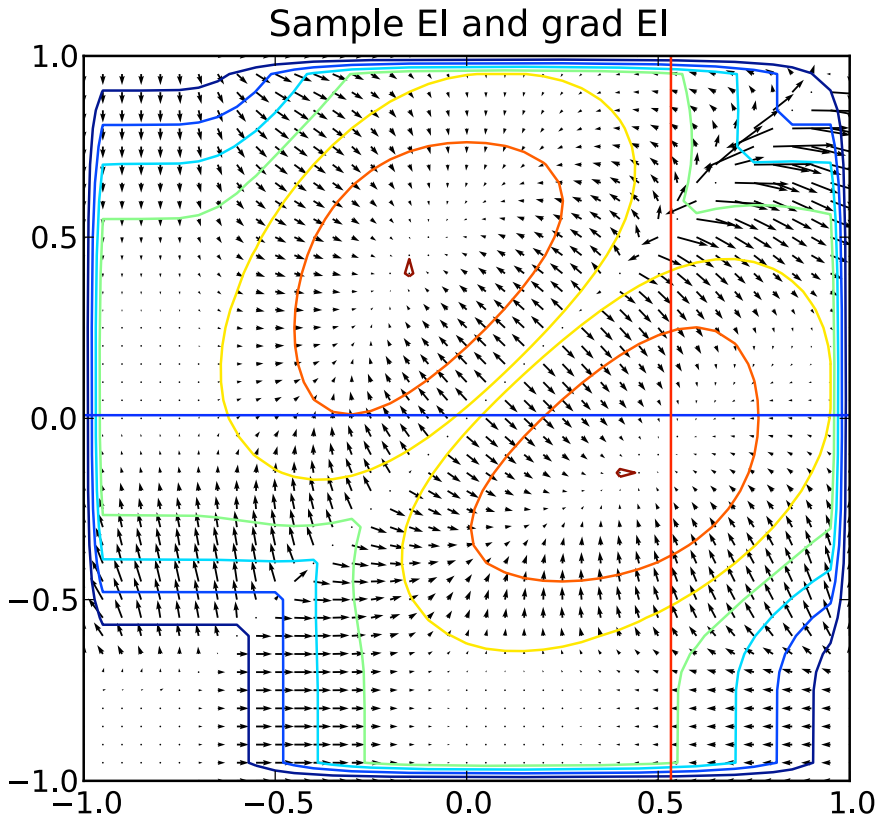
$$\max_{\vec{x}_1,\ldots,\vec{x}_q} \mathrm{EI}(\vec{x}_1,\ldots,\vec{x}_q)$$

  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 $\vec{x}_{k+1},\ldots,\vec{x}_q$.

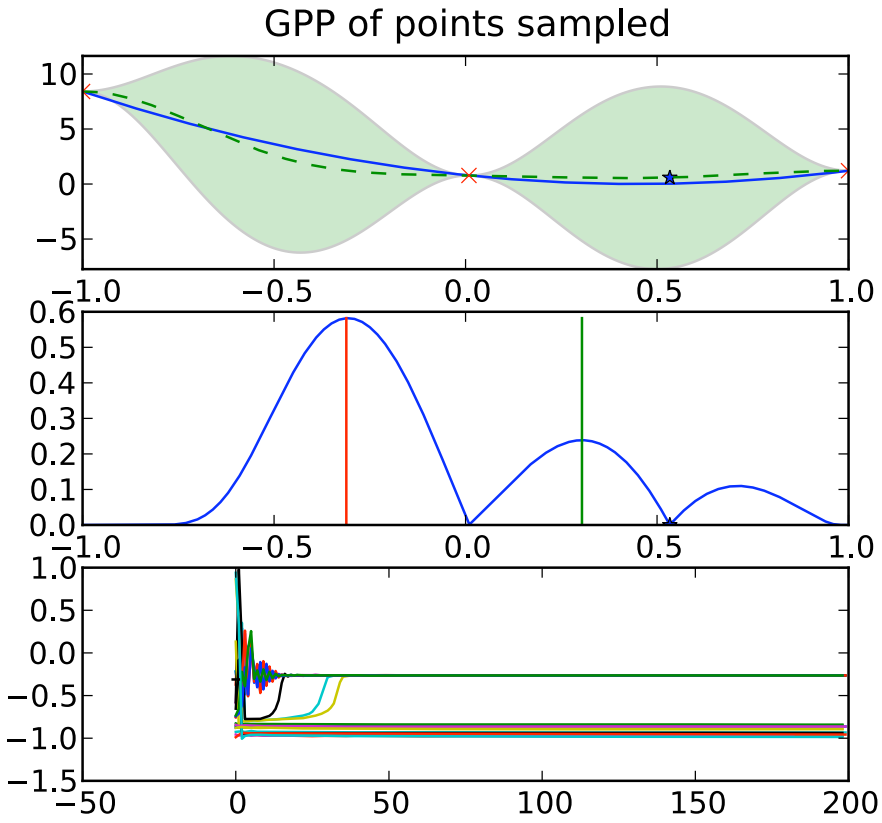
$$\max_{\vec{x}_{k+1},\ldots,\vec{x}_q} \mathrm{EI}(\vec{x}_1,\ldots,\vec{x}_q)$$

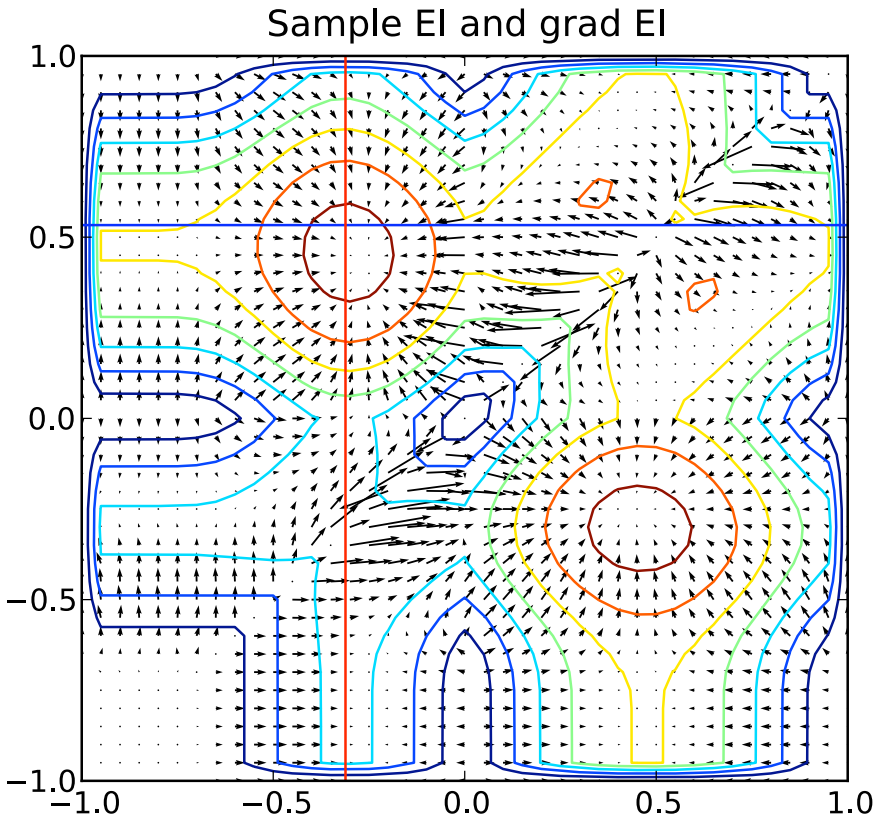
Animation


GPP of points sampled
10
5
0
−5
−1.0
−0.5
0.0
0.5
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−50
0
50
100
150
200

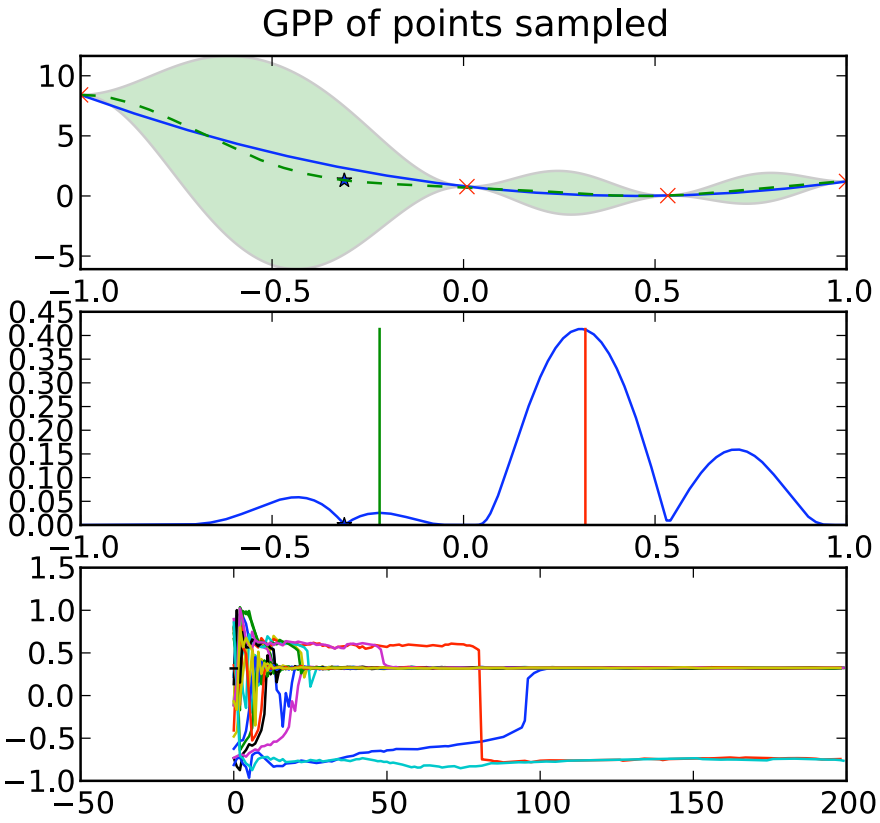

Sample EI and grad EI
1.0
0.5
0.0
−0.5
−1.0
−1.0
−0.5
0.0
0.5
1.0

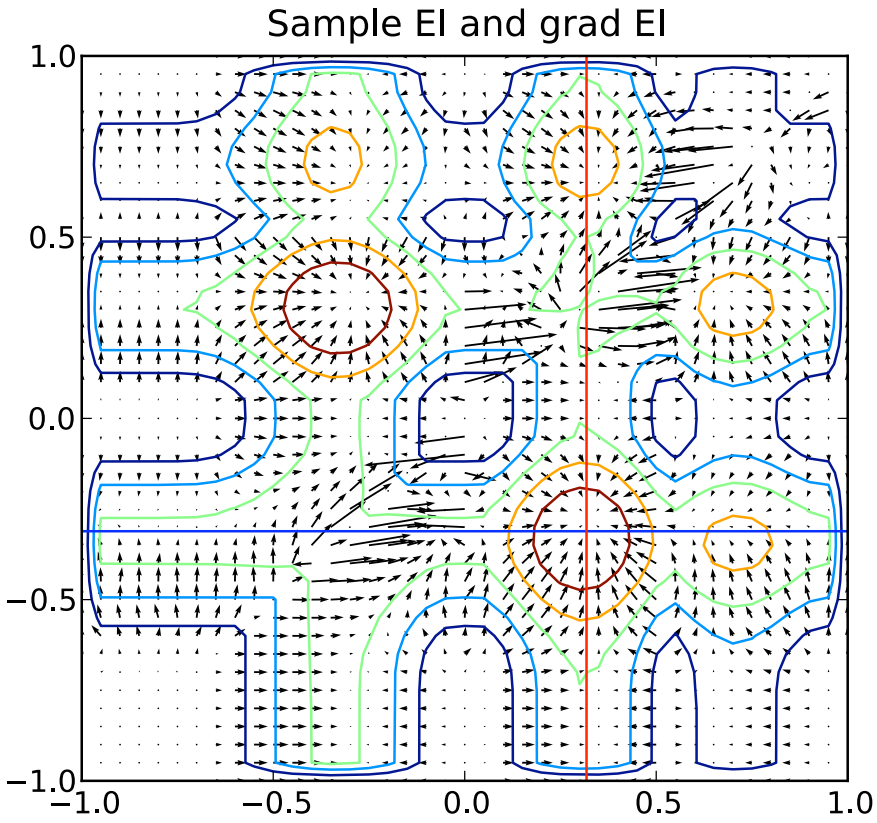
Animation


GPP of points sampled
10
5
0
−5
−1.0
−0.5
0.0
0.5
1.0
0.6
0.5
0.4
0.3
0.2
0.1
0.0
1.0
0.5
0.0
−0.5
−1.0
−1.5
−50
0
50
100
150
200

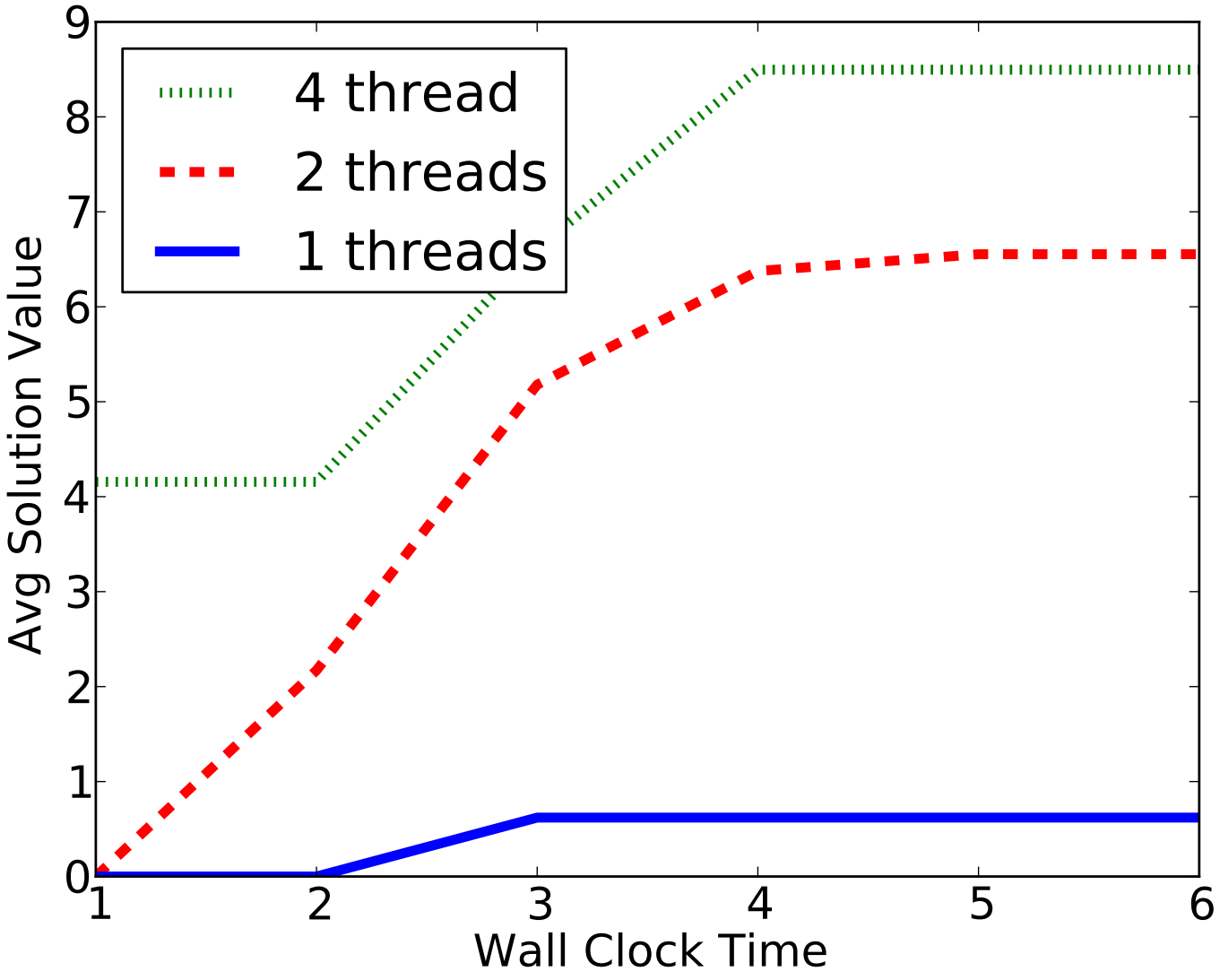

Sample EI and grad EI
1.0
0.5
0.0
−0.5
−1.0
−1.0
−0.5
0.0
0.5
1.0

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
1.0
0.5
0.0
−0.5
−1.0
−50
0
50
100
150
200


Sample EI and grad EI
1.0
0.5
0.0
−0.5
−1.0
−1.0
−0.5
0.0
0.5
1.0

Initial Results



Conclusion

- We considered a previously proposed conceptual method for parallel Bayesian global optimization (BGO).
- This conceptual algorithm was difficult to implement in systems with a large degree of parallelism.
- We used methods from simulation optimization to provide a better implementation.
- Future work: comparisons to other parallel methods; improved gradient estimators; improved selection of seed locations; tests in systems with more parallelism; extensions to evaluations with noise.

References I

Blum, J. R. (1954).
Multidimensional stochastic approximation methods.
The Annals of Mathematical Statistics, 25(4):737–744.

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.

References II

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.

Ginsbourger, D., Le Riche, R., and Carraro, L. (2010).

Kriging is well-suited to parallelize optimization.
In *Computational Intelligence in Expensive Optimization Problems*, volume 2, pages 131–162. Springer.

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.

Kushner, H. J. and Yin, G. G. (1997).

Stochastic Approximation Algorithms and Applications.
Springer-Verlag, New York.

L'Ecuyer, P. (1990).

A unified view of the IPA, SF, and LR gradient estimation techniques.
Management Science, pages 1364–1383.

References III

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.

Robbins, H. and Monro, S. (1951).

A Stochastic Approximation Method.
Annals of Math. Stat., 22:400–407.

Schonlau, M. (1997).

Computer experiments and global optimization.
PhD thesis, University of Waterloo.

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.

Smith, S. (1995).

Differentiation of the Cholesky algorithm.
Journal of Computational and Graphical Statistics, pages 134–147.

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.

References IV

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.

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.

Multistart Stochastic Gradient Ascent

1. Select several starting points, uniformly at random.
2. From each starting point, iterate using the stochastic gradient method until convergence.

$$(\vec{x}_1, \ldots, \vec{x}_q) \leftarrow (\vec{x}_1, \ldots, \vec{x}_q) + \alpha_n g(\vec{x}_1, \ldots, \vec{x}_q, \omega),$$

   where (α_n) is a stepsize sequence.
3. For each starting point, average the iterates to get an estimated stationary point. (Polyak-Ruppert averaging)
4. Select the estimated stationary point with the best estimated value as the solution.