A Taxonomy of Global Optimization Methods Based on Response Surfaces

Don Jones
General Motors
don.jones@gm.com

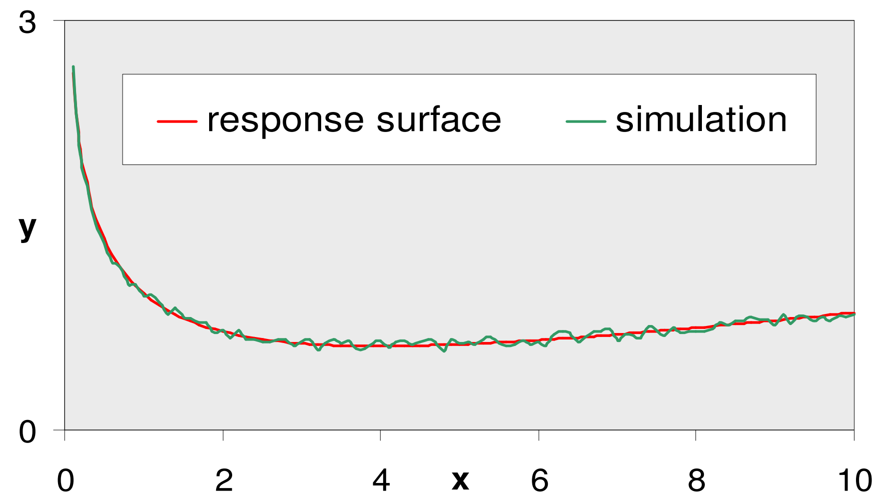
Presented at the Fields Institute, February 6, 2007

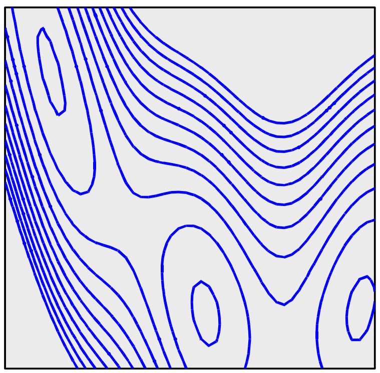
Why use response surfaces for optimization?

- Smooth “numerical jiggle”
- Identify important variables and visualize input-output relationships via functional ANOVA
- Solve variations of problem (e.g., different bounds, objective function) quickly, without additional simulations.
- Exploit parallel processing
- Fast computation transmitted variance for robust design

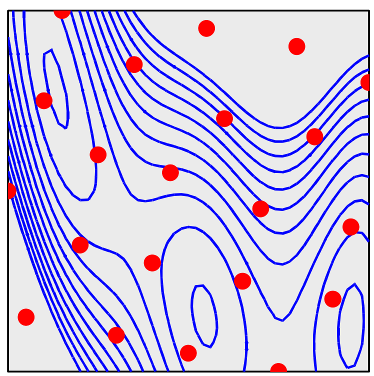
☞ Reduce evaluations for optimization

☞ Increase the likelihood of finding a <u>global</u> optimum





Sample Nonlinear Function



Response Surface Based on 21 Points

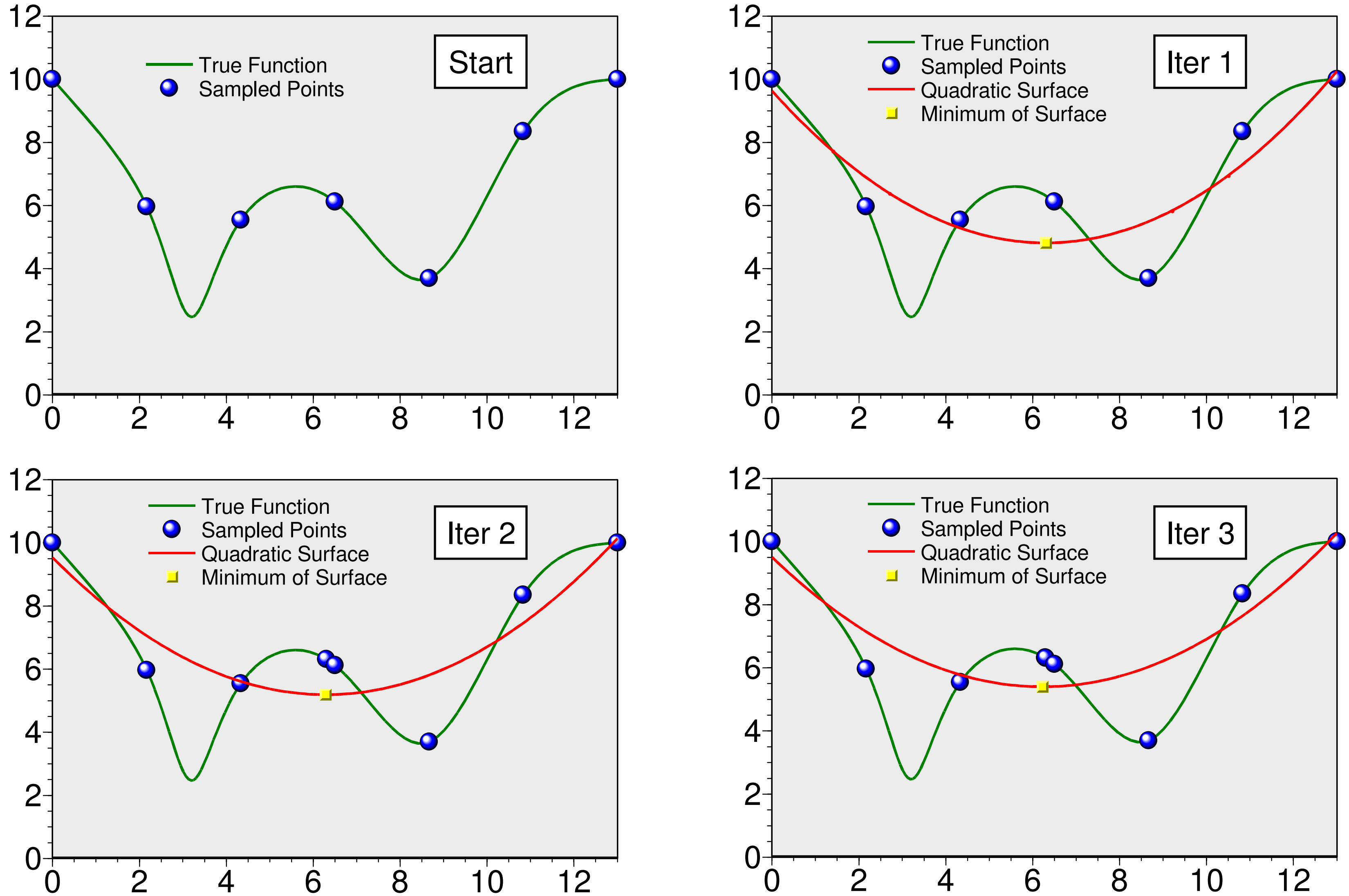
Today's talk

- We will present a taxonomy of existing approaches to using response surfaces for global optimization
- Key messages:
  - Methods that seem "reasonable" often have non-obvious failure modes
  - Developing a method that delivers on the intuitive promise of response surfaces is non-trivial
  - The area abounds with interesting and exciting research issues

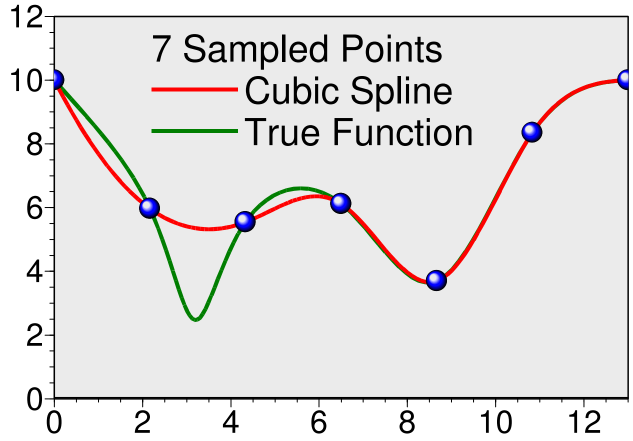
The taxonomy

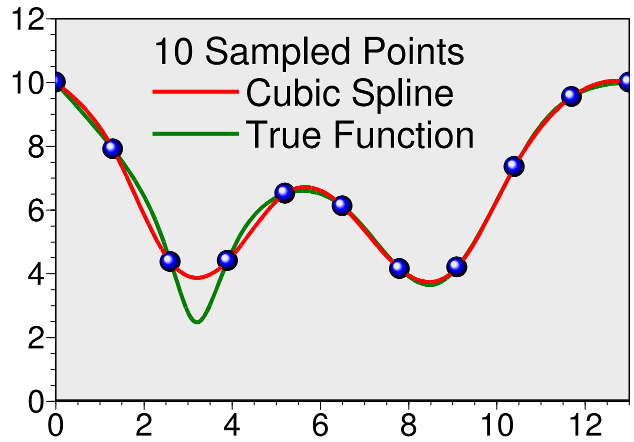
<table>
<tr><td colspan="3" rowspan="3">Kind of Response Surface</td><td colspan="7">Method for selecting search points</td></tr>
<tr><td colspan="5">Two-stage approach: first fit a surface, then find the next iterate by optimizing an auxiliary function based on the surface</td><td colspan="2">One stage approach: evaluate hypotheses about optimum based on implications for the response surface</td></tr>
<tr><td>Minimize the Response Surface</td><td>Mimimize a Lower Bounding Function</td><td>Maximize the Probability of Improvement</td><td>Maximize Expected Improvement</td><td>Goal seeking: find point that achieves a given target</td><td>Optimization: find point that minimizes an objective</td></tr>
<tr><td>Not interpolating (smoothing)</td><td colspan="2">Quadratic polynomials and other regression models</td><td>1</td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td rowspan="2">Interpolating</td><td>Fixed basis functions. NO statistics.</td><td>Thin-plate splines, Hardy multiquadrics</td><td rowspan="2">2</td><td></td><td></td><td></td><td rowspan="2">6</td><td rowspan="2">7</td></tr>
<tr><td>Tuned basis functions. Statistical interpretation</td><td>riging</td><td>3</td><td>4</td><td>5</td></tr>
</table>

❶ Fit surface with regression, optimize surface, sample minimum, update surface, iterate



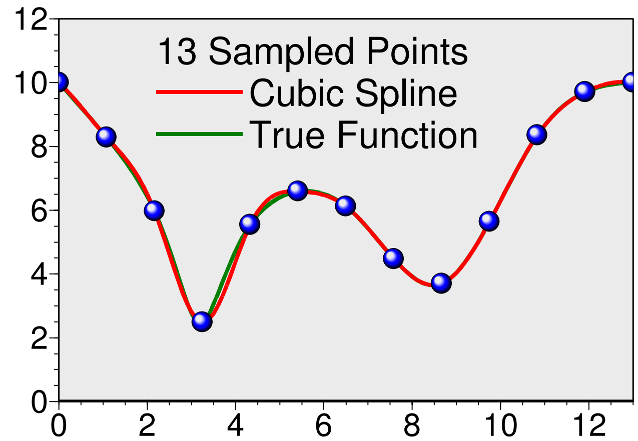
Method 1 fails when the assumed functional form is wrong.   We therefore need a data-adaptive surface.



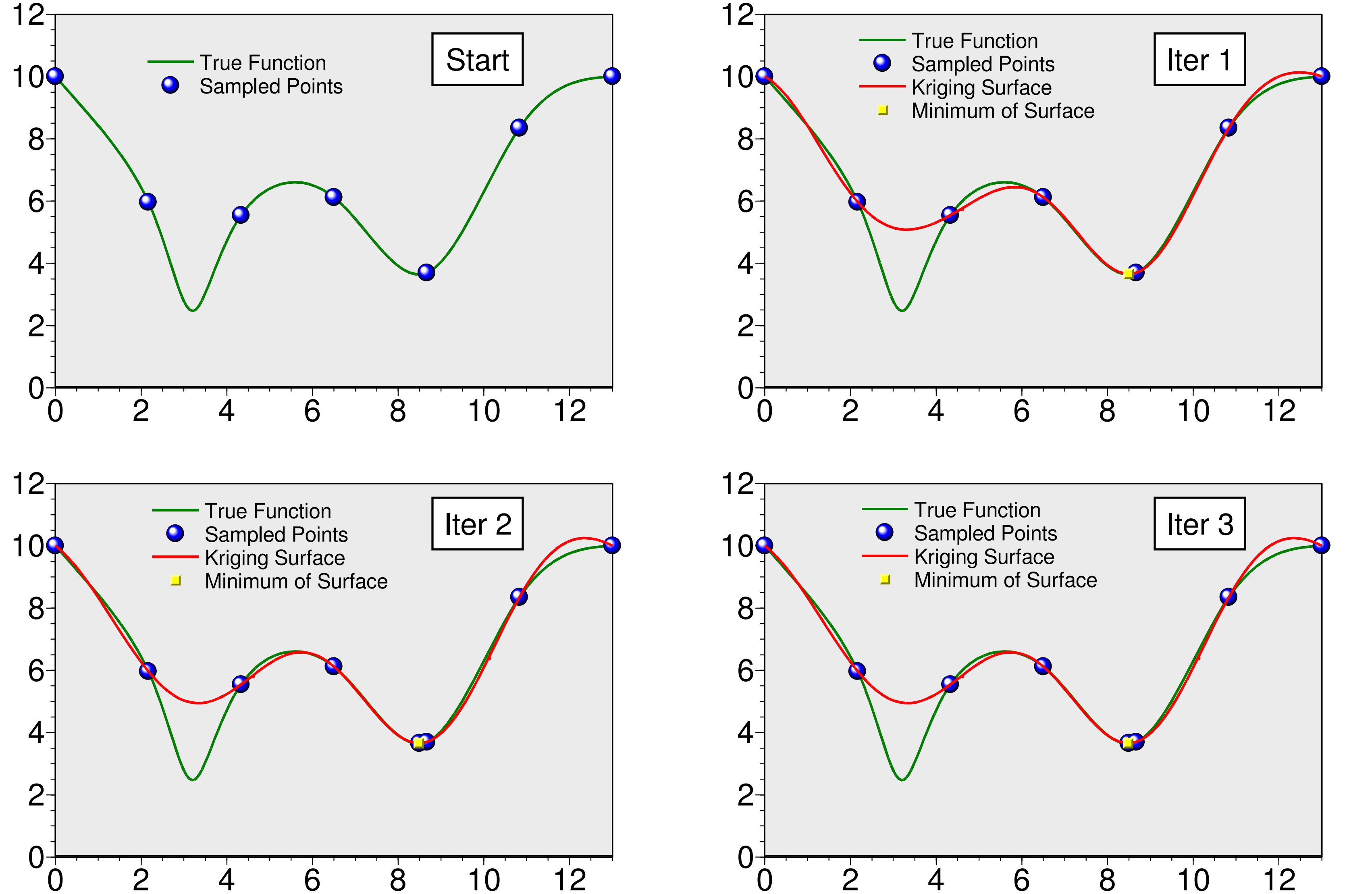


Splines, radial-basis functions, and kriging are data-adaptive techniques that find a linear combination of n "basis functions" that interpolate the n data points.

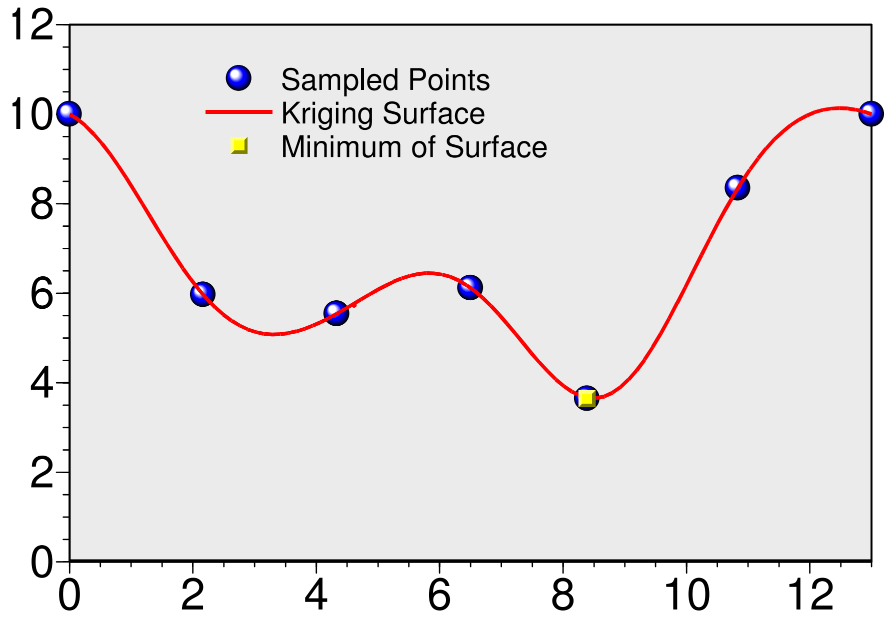
Kriging stands out because: (1) the basis functions are tuned to the data and (2) its statistical derivation allows the estimation of confidence intervals.



❷ Fit kriging surface, optimize surface, sample minimum, update surface, iterate

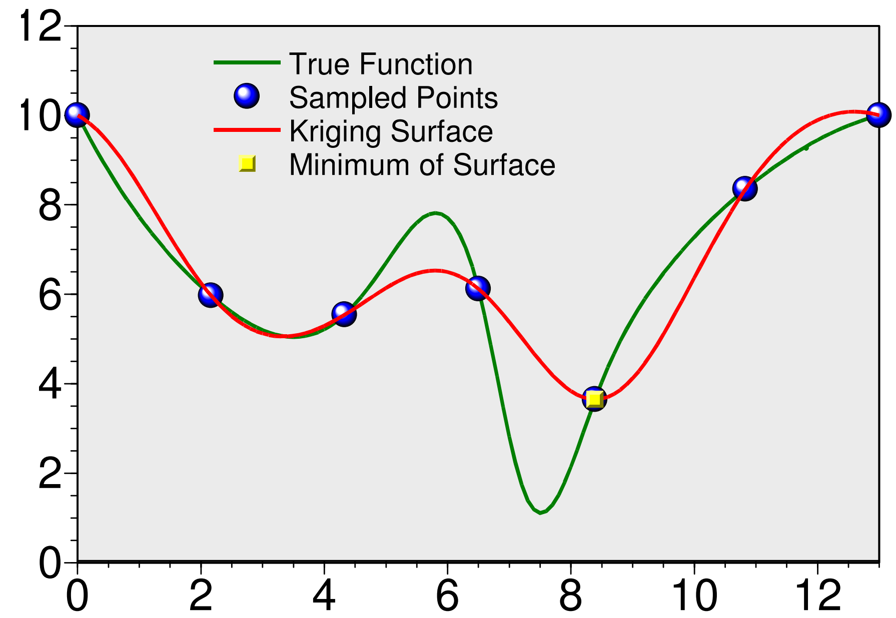


Method 2 may fail because critical point of surface may not be critical point of the true function



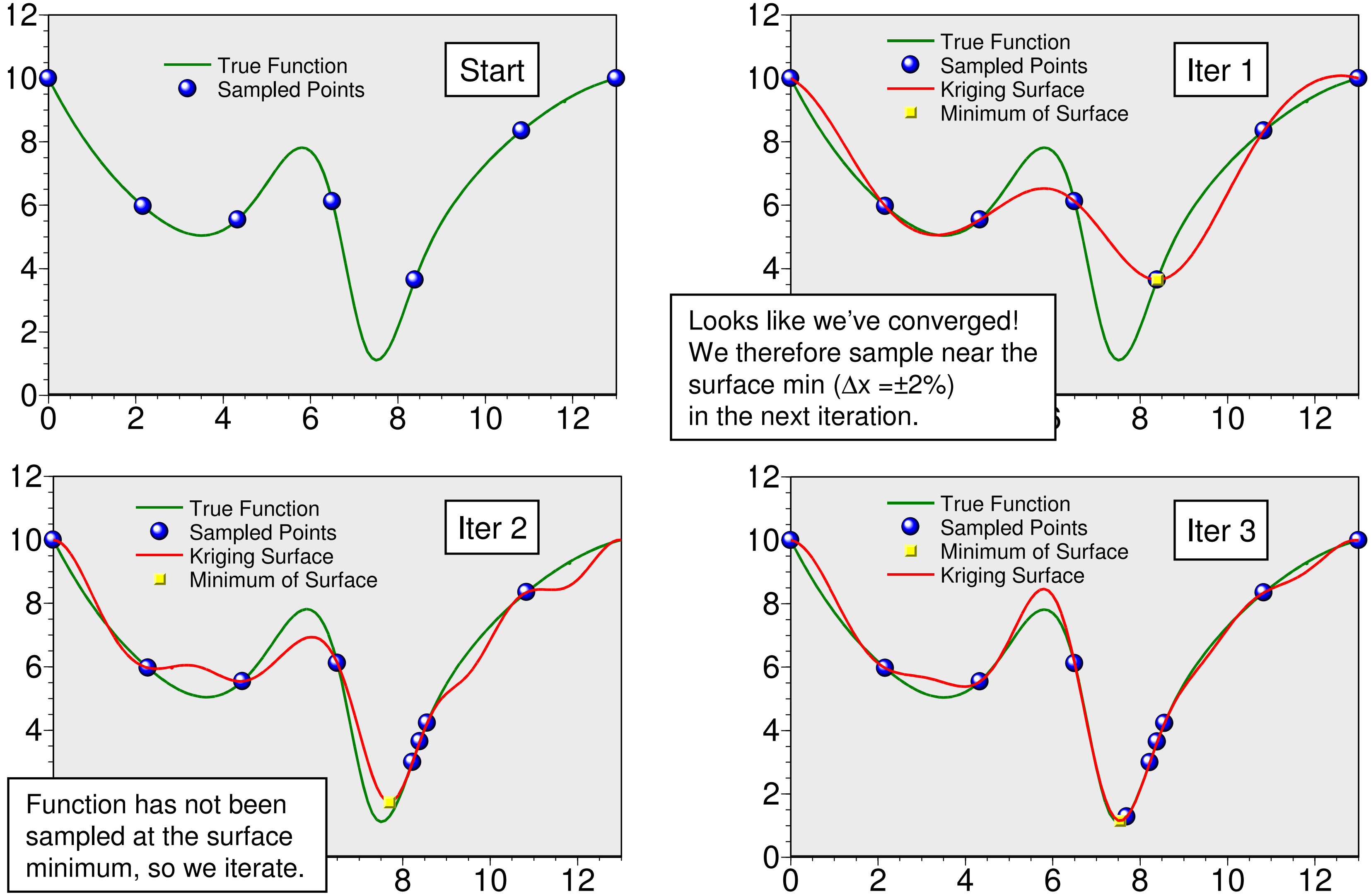
Minimum of surface agrees with a data point, so further iterations will not change anything.

We have converged, but do we have a local?

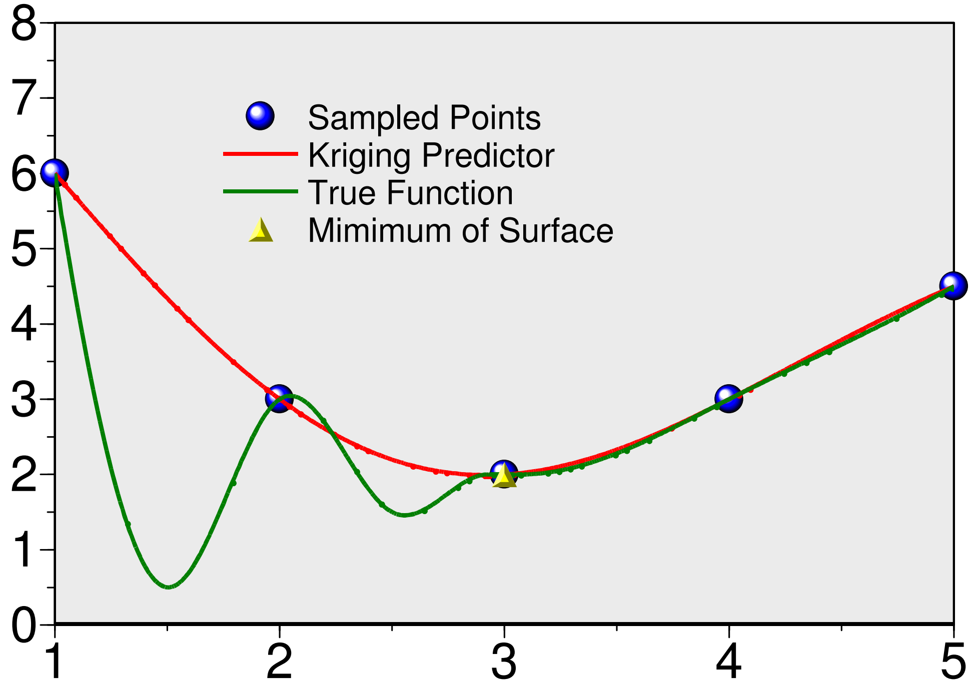


No, we may not have a local.

Enhanced ❷ — Same as ② but force surface gradient to match that of true function whenever we tentatively converge



Enhanced Method ❷ may not only miss a global optimum—it may also fail to find a local optimum



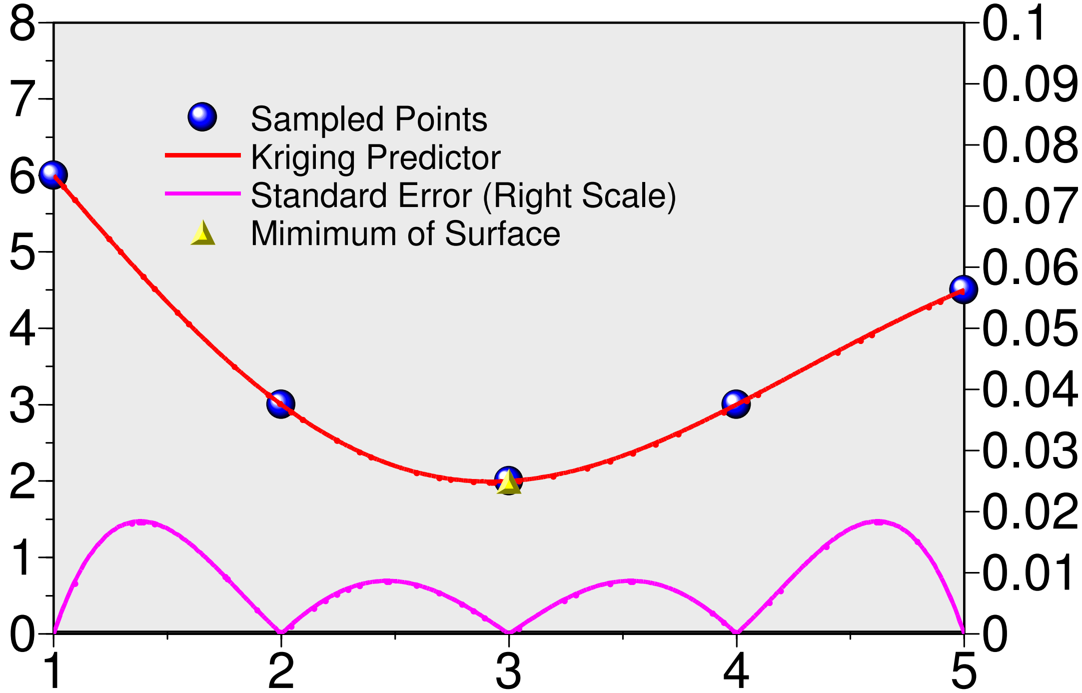
Possible causes of failure:

- Saddle points (as shown)
- Poor local approximation
  - Without analytic gradient, proof of convergence must somehow insure that added points give good local approximation

Relevant papers that show how can guarantee convergence to a local minimum:

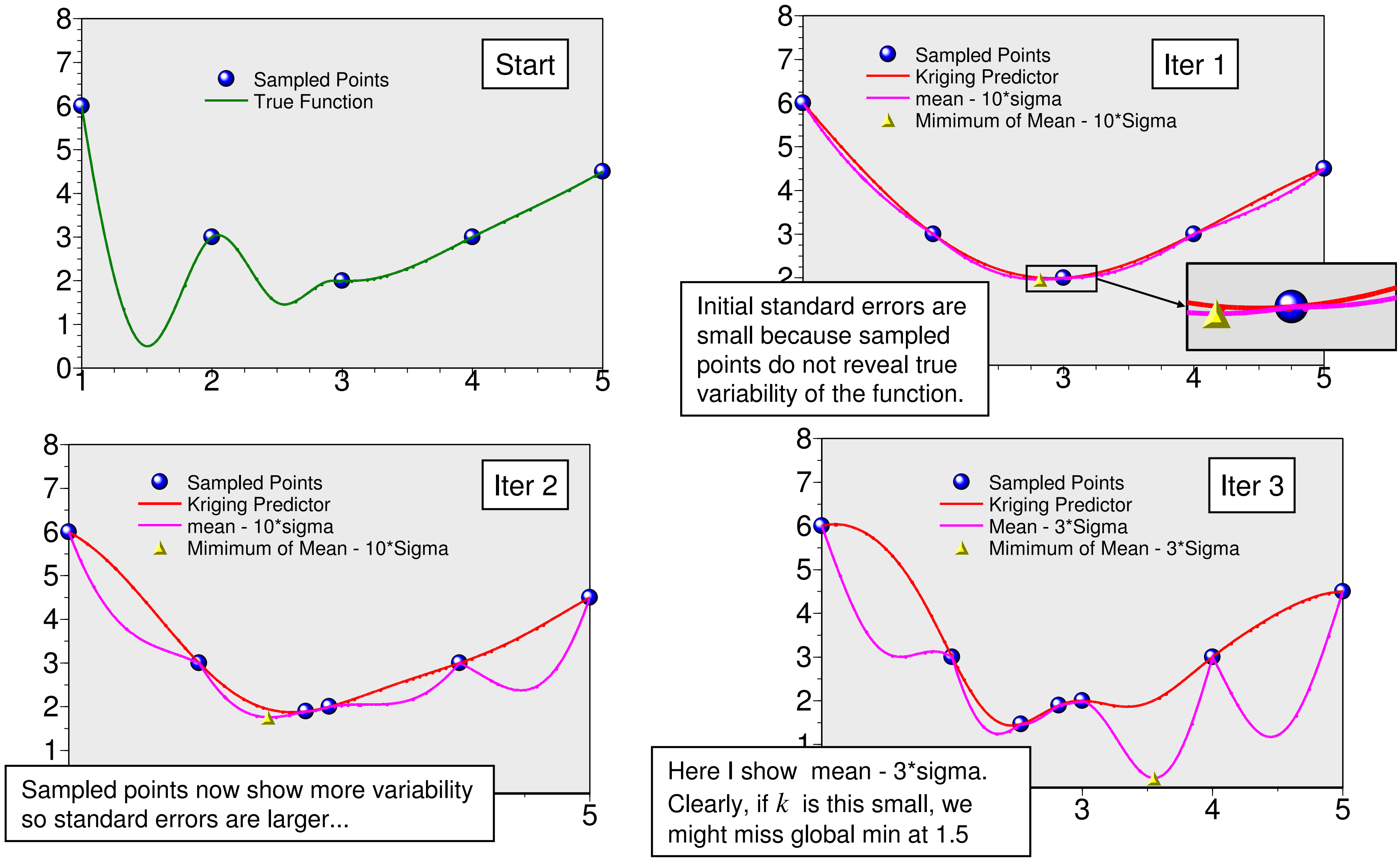
- N. M. Alexandrov et. al. "A trust region framework for managing the use of approximation models in optimization." *Structural Optimization*, 15(1):16-23, 1998.
- M.J.D. Powell. "A direct search optimization method that models the objective and constraint functions by linear approximation." Dept. of Applied Math. and Theoretical Physics, Cambridge U., England, paper DAMTP 1992/NA5, April 1992.

Key to making search global is to put some emphasis on sampling where the surface may be inaccurate

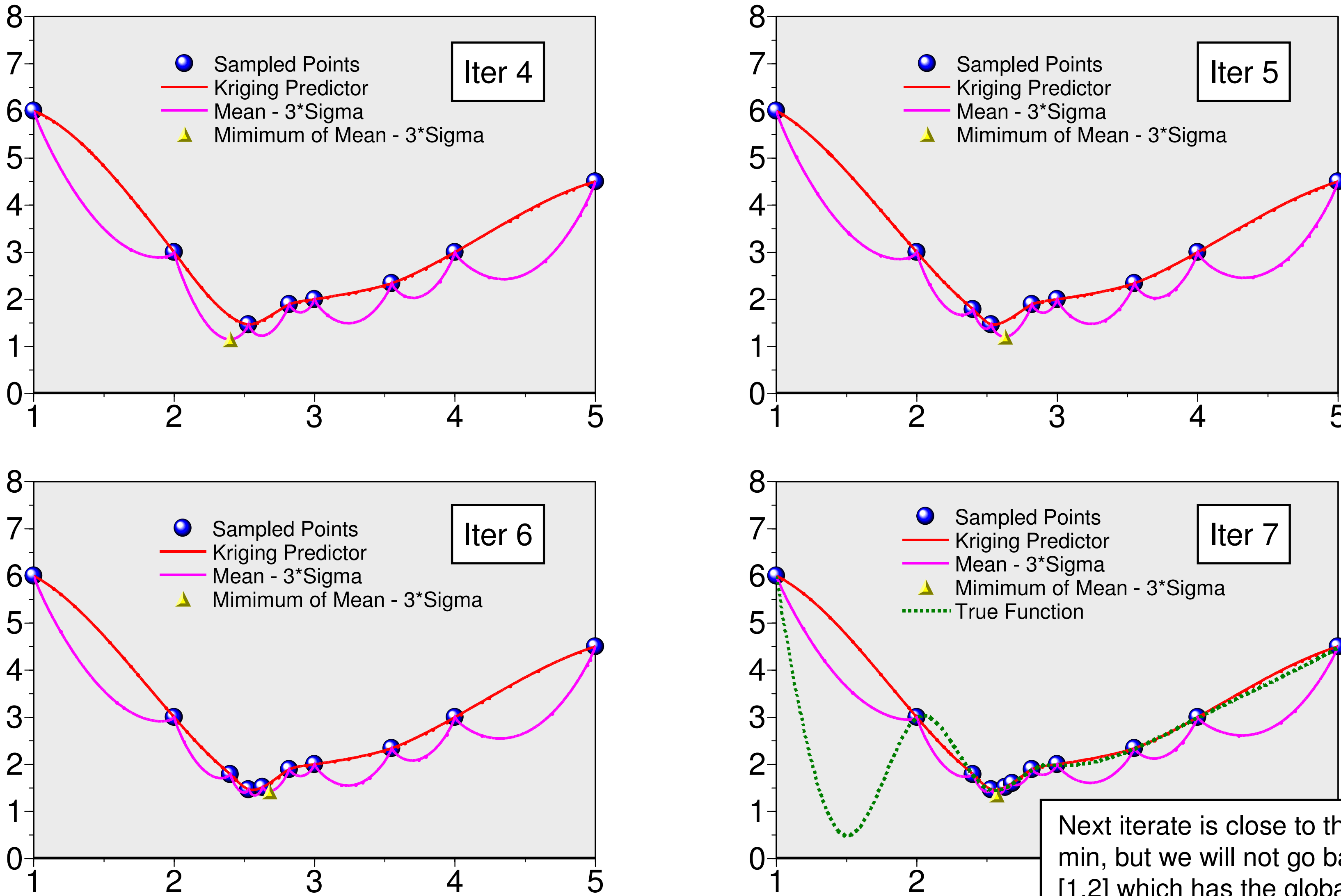


- The kriging predictor at x* is the best guess of y*
- The kriging standard error reflects much error there may be in the predictor, based on how far we are from the sampled points.
- To be global, we must balance sampling where:
  - surface is minimized (focus on predictor)
  - surface may be inaccurate (focus on standard error)

❸ Fit kriging surface, find minimum of $y(x)\text{-}k\sigma(x)$, sample this point, update surface, iterate

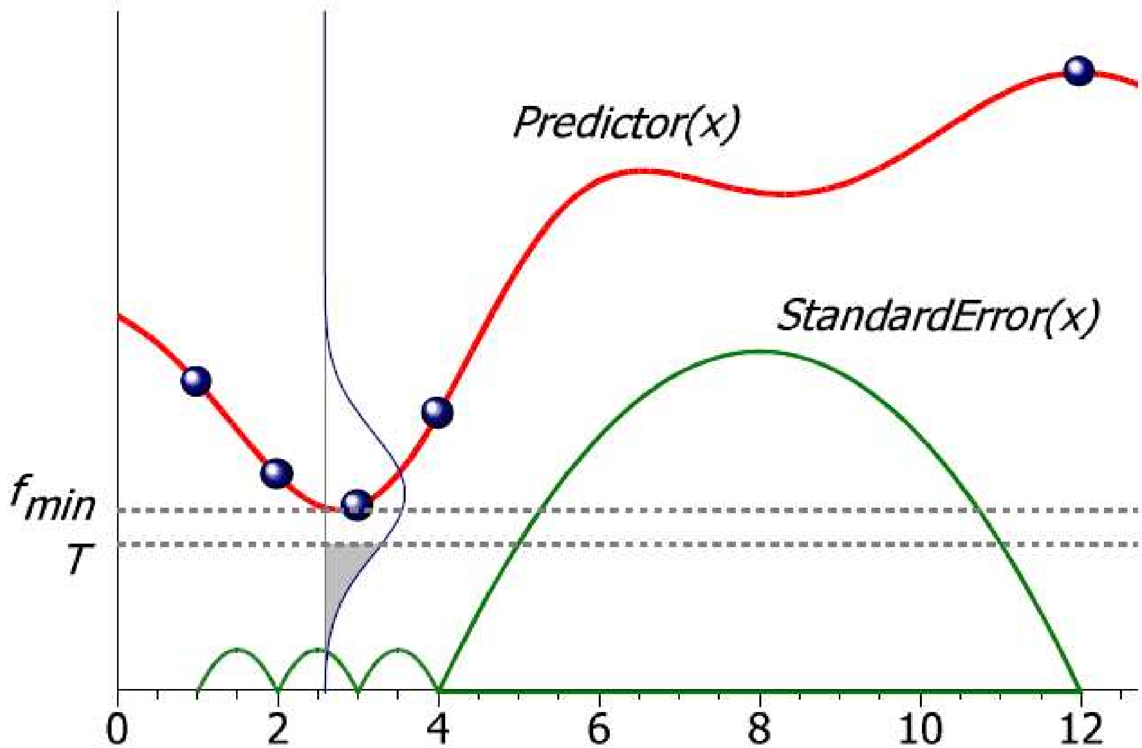


❸ Fit kriging surface, find minimum of $y(x)\text{-}k\sigma(x)$, sample this point, update surface, iterate *(continued)*



Next iterate is close to the local min, but we will not go back to [1,2] which has the global min.

Method 3 fails because it prematurely rules out entire regions. We need the search points to fill space

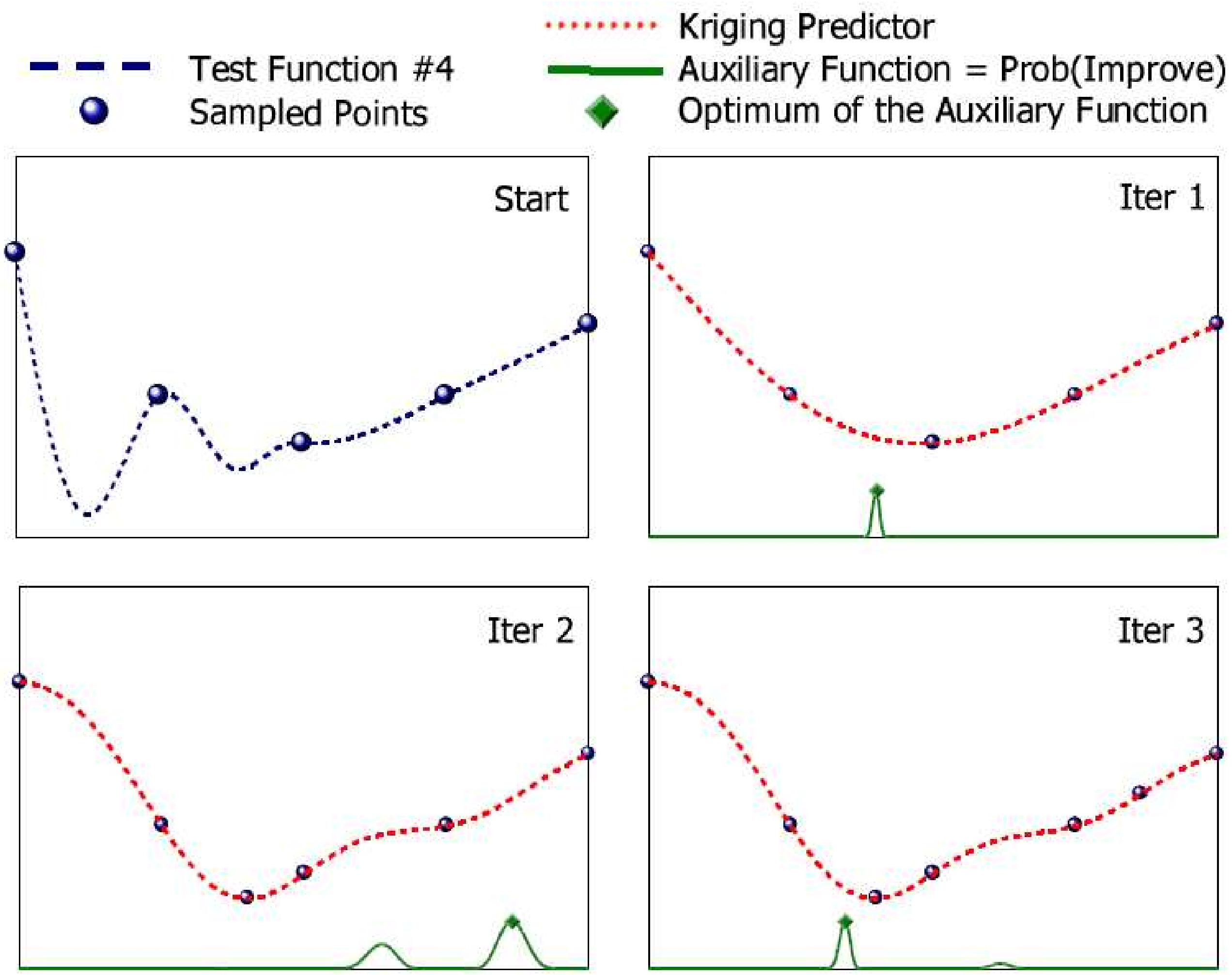


The **Probability of Improvement** criterion specifies a target value of the objective T (better than the current best point) and samples where the probability of exceeding T is maximized..

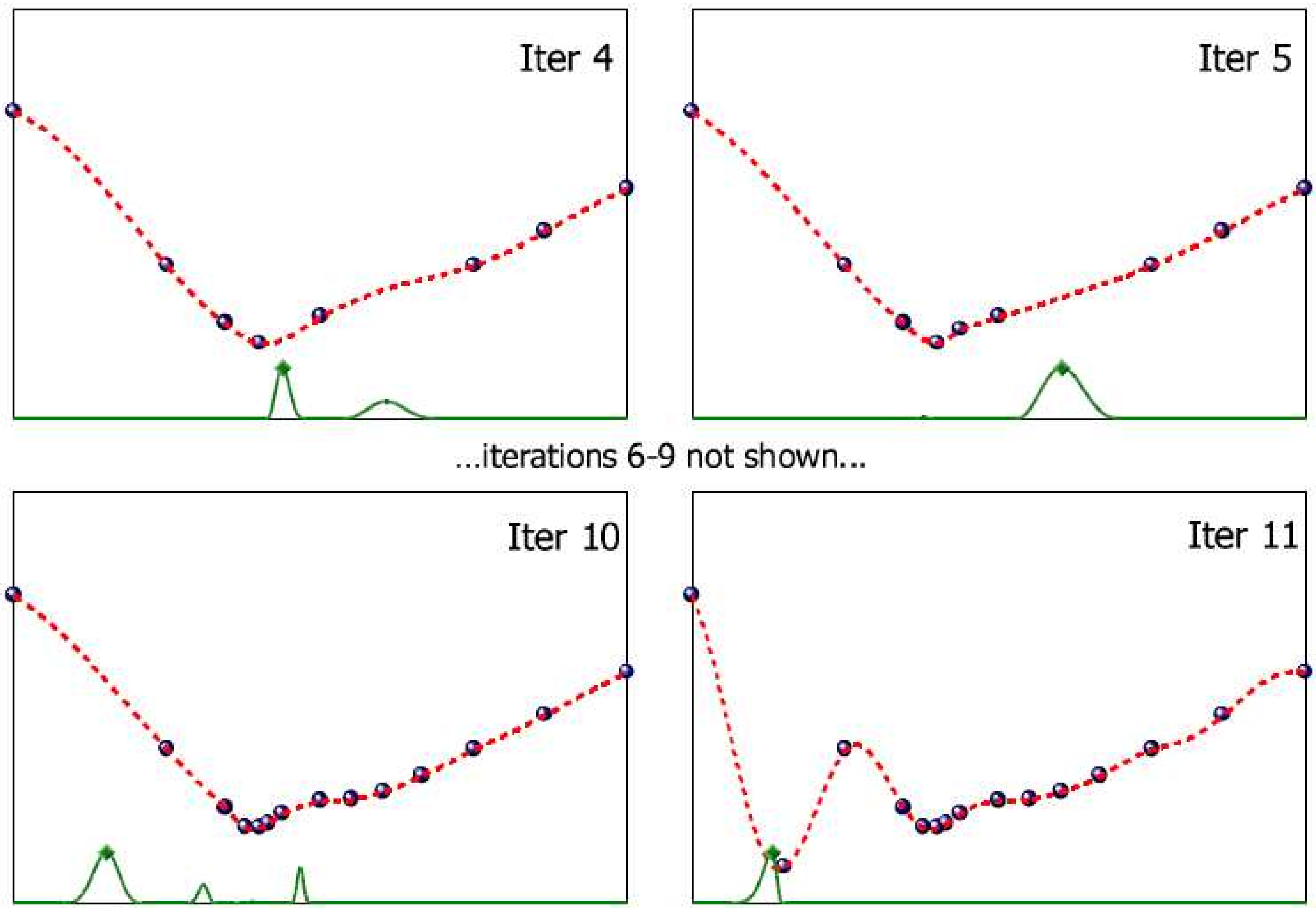
Under mild conditions, it will generate search points that fill the space (i.e., are dense). Hence it guarantees ultimate convergence.

❹ Fit kriging surface, find maximum of Probability of Improvement, sample this point, update surface, iterate

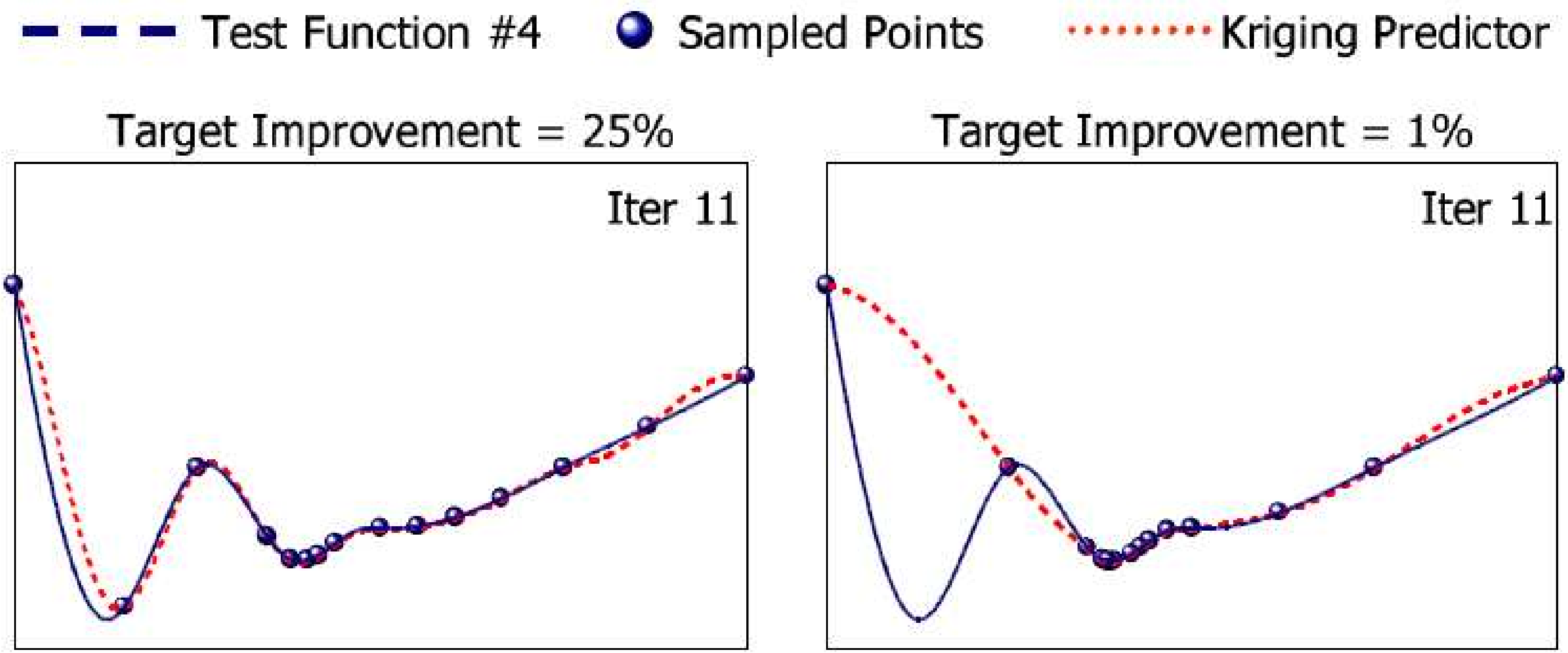
Target objective function is computed using: $T = s_{\min} - \alpha(f_{\max} - f_{\min})$ ***with*** α ***= 0.25***



❹ Fit kriging surface, find maximum of Probability of Improvement, sample this point, update surface, iterate



❹ Problem with Method 4 is the sensitivity to the target function value T (determined by choice of α)

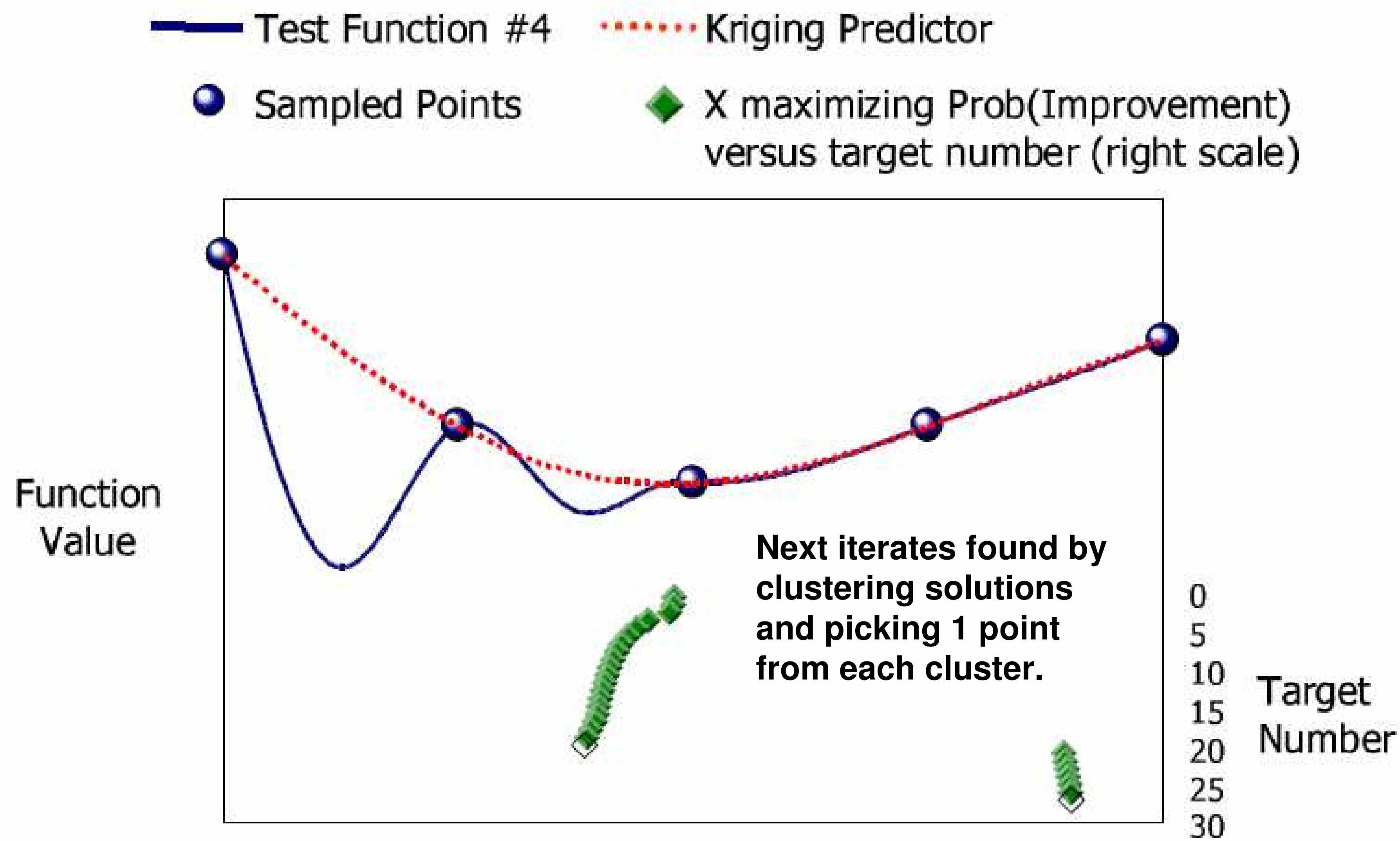


Results of first 11 iterations of Method 4 using two values for α

Enhanced ❹ — Find the point that maximizes the probability of improvement for <u>many</u> targets

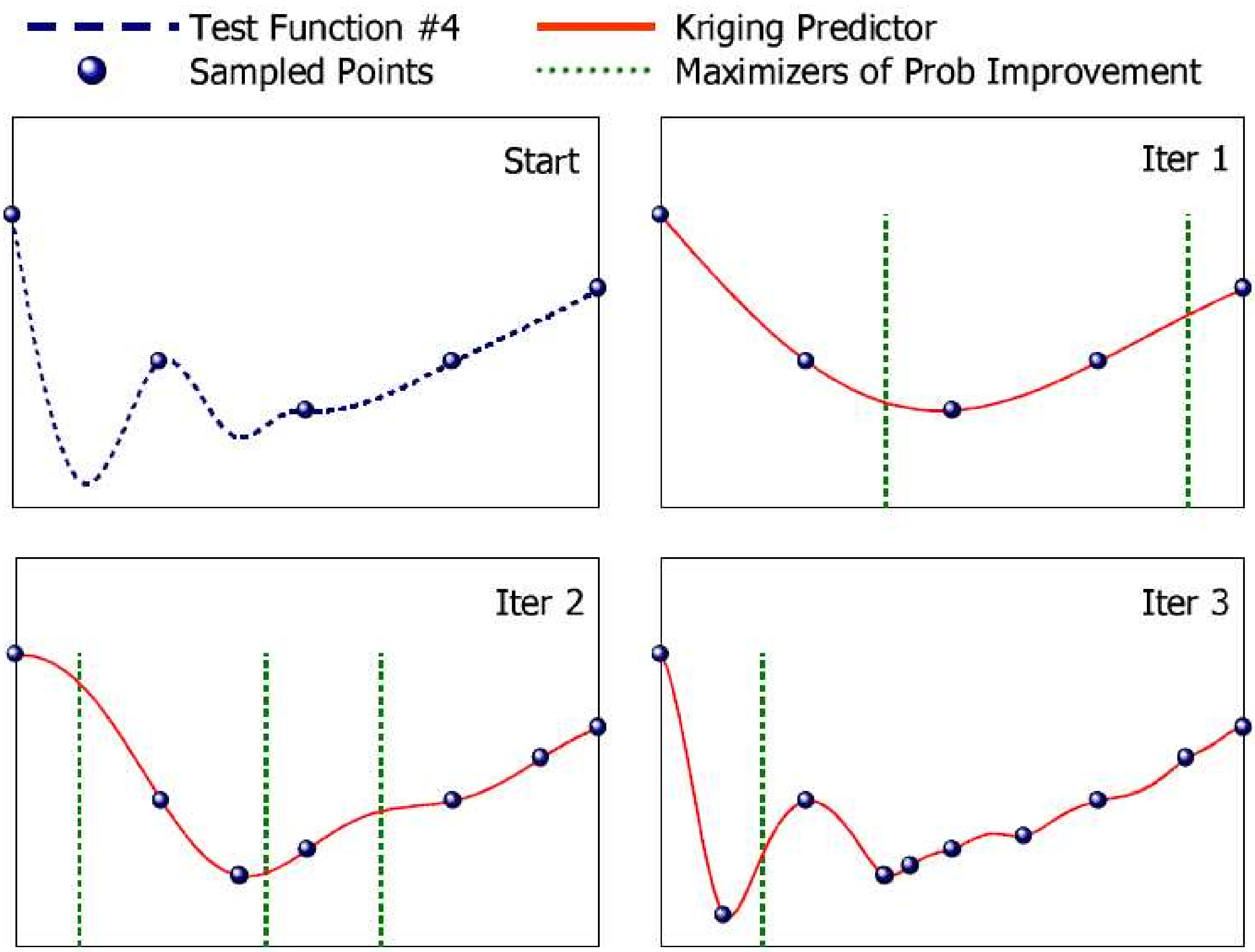
Target Number	α	Target Number	α	Target Number	α
1	0.0	10	0.07	19	0.25
2	0.0001	11	0.08	20	0.30
3	0.001	12	0.09	21	0.40
4	0.01	13	0.10	22	0.50
5	0.02	14	0.11	23	0.75
6	0.03	15	0.12	24	1.00
7	0.04	16	0.13	25	1.50
8	0.05	17	0.15	26	2.00
9	0.06	18	0.20	27	3.00

Enhanced ❹ — Find the point that maximizes the probability of improvement for many targets

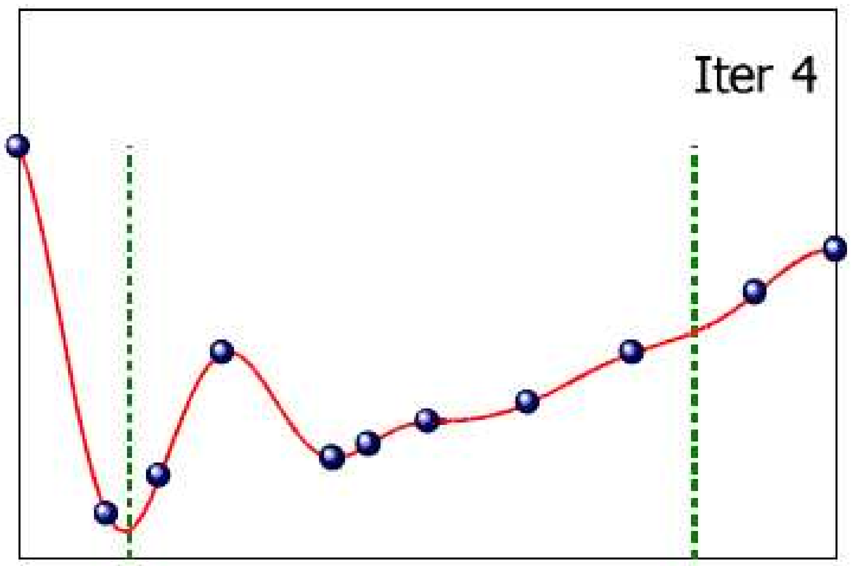


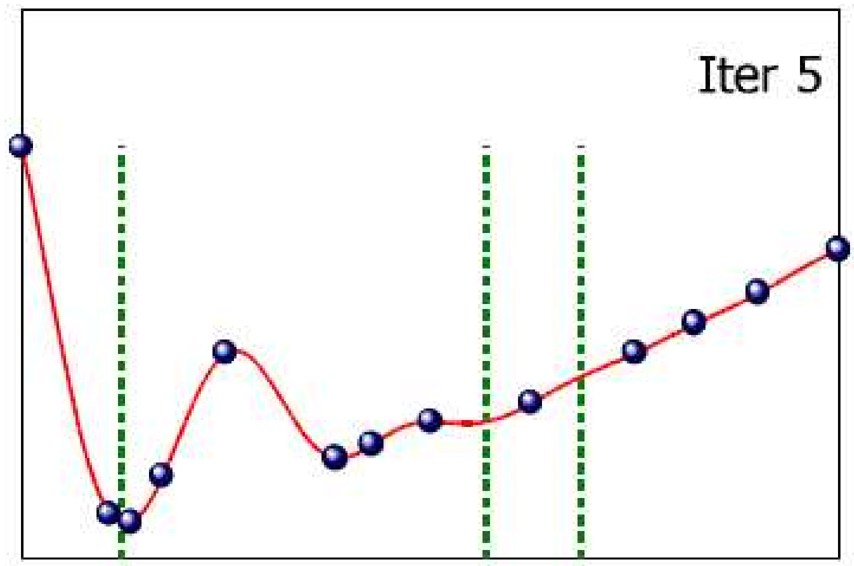
Next iterates found by clustering solutions and picking 1 point from each cluster.

Enhanced ❹ — Find the point that maximizes the probability of improvement for <u>many</u> targets

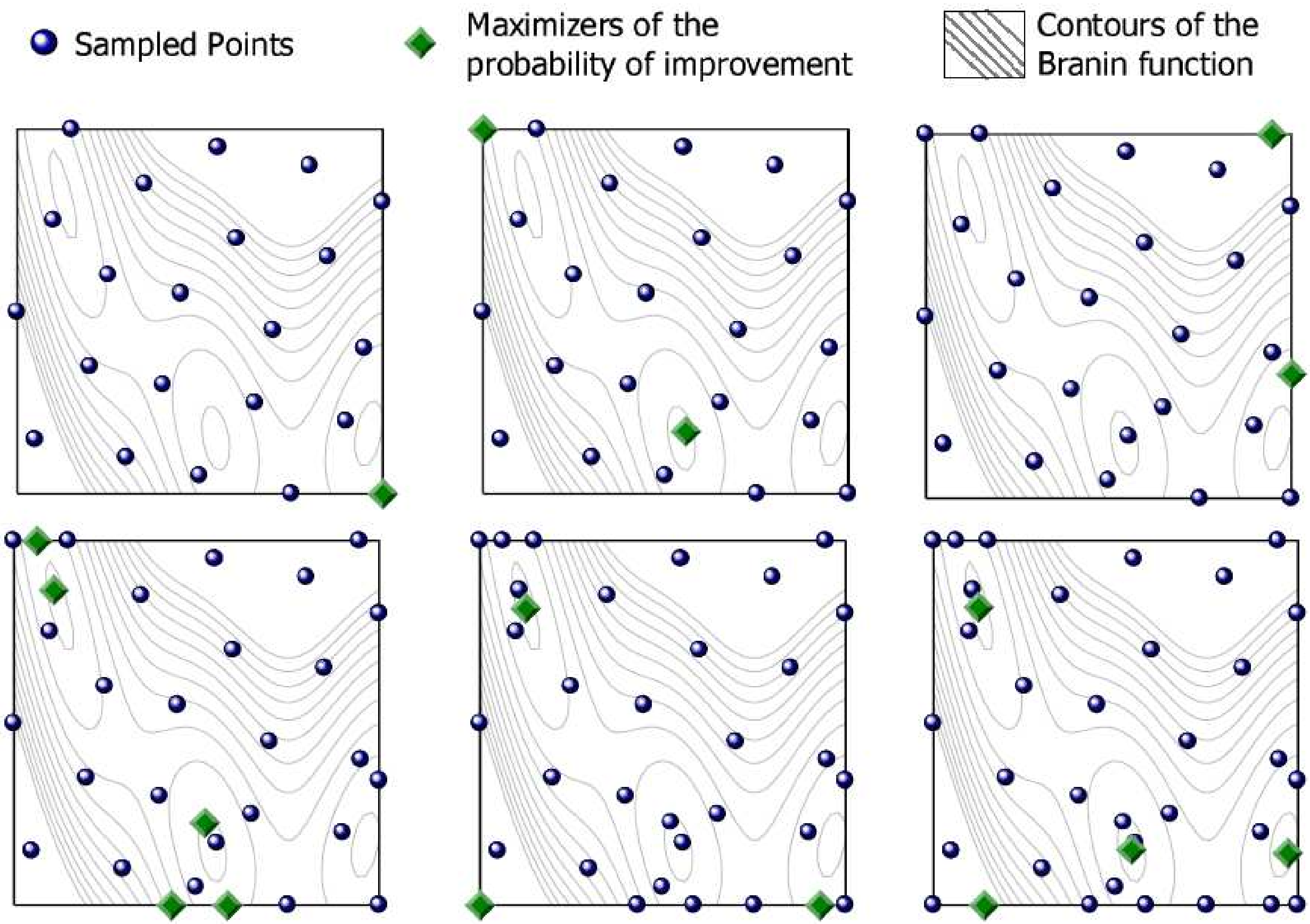


Enhanced ❹ — Find the point that maximizes the probability of improvement for <u>many</u> targets

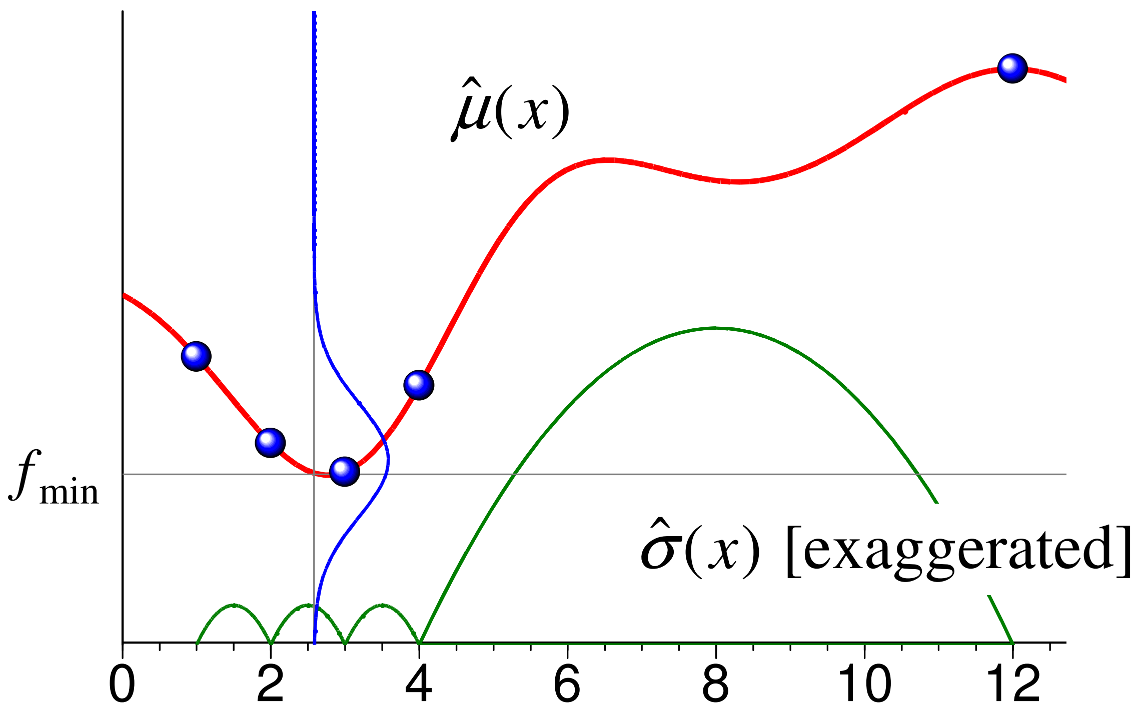




Enhanced ❹ — Example on two-variable problem



Expected Improvement

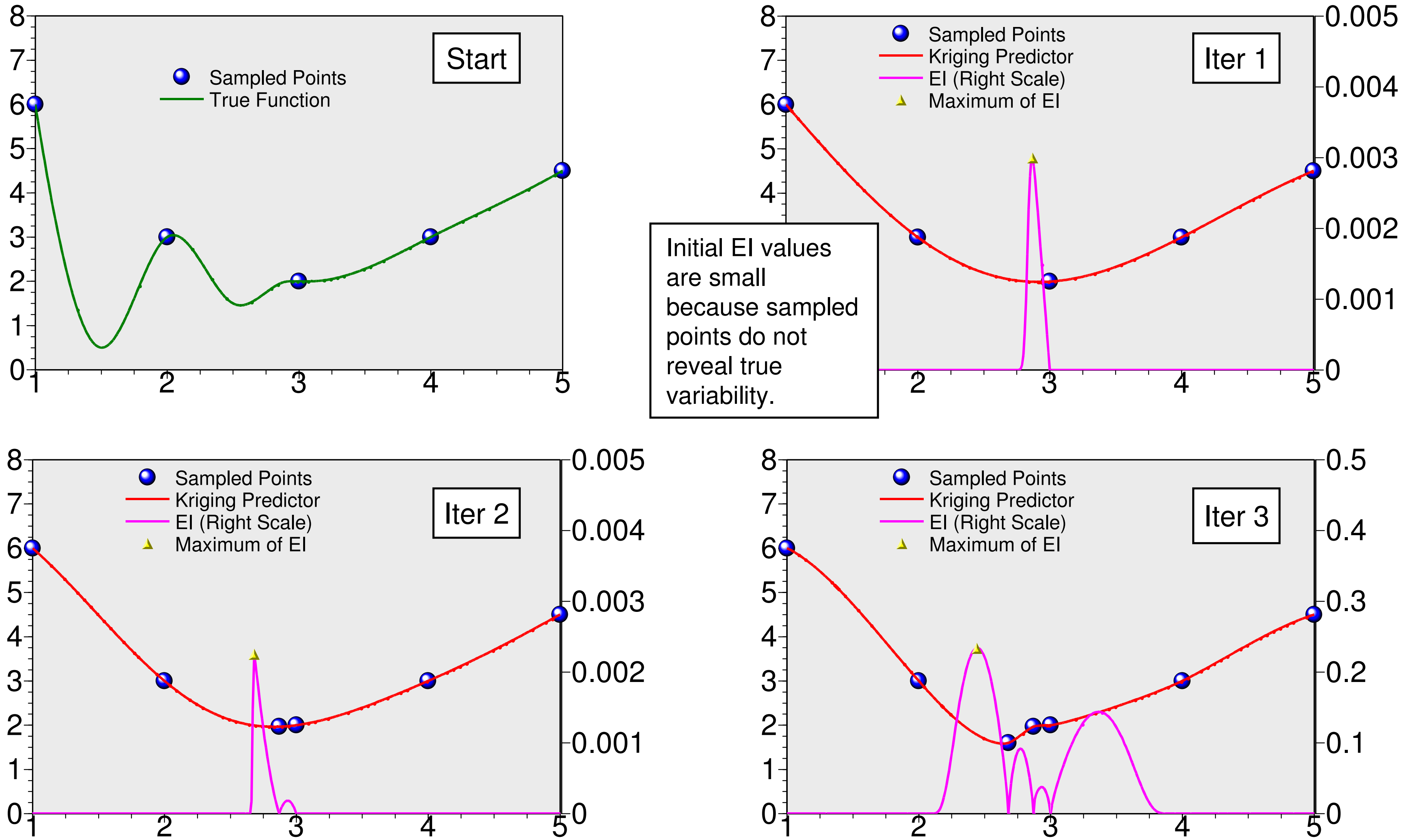


The **Expected Improvement** (EI) criterion does not force us to choose a single value for the target T in Method 4. It is an alternative to using many values of T as just discussed.

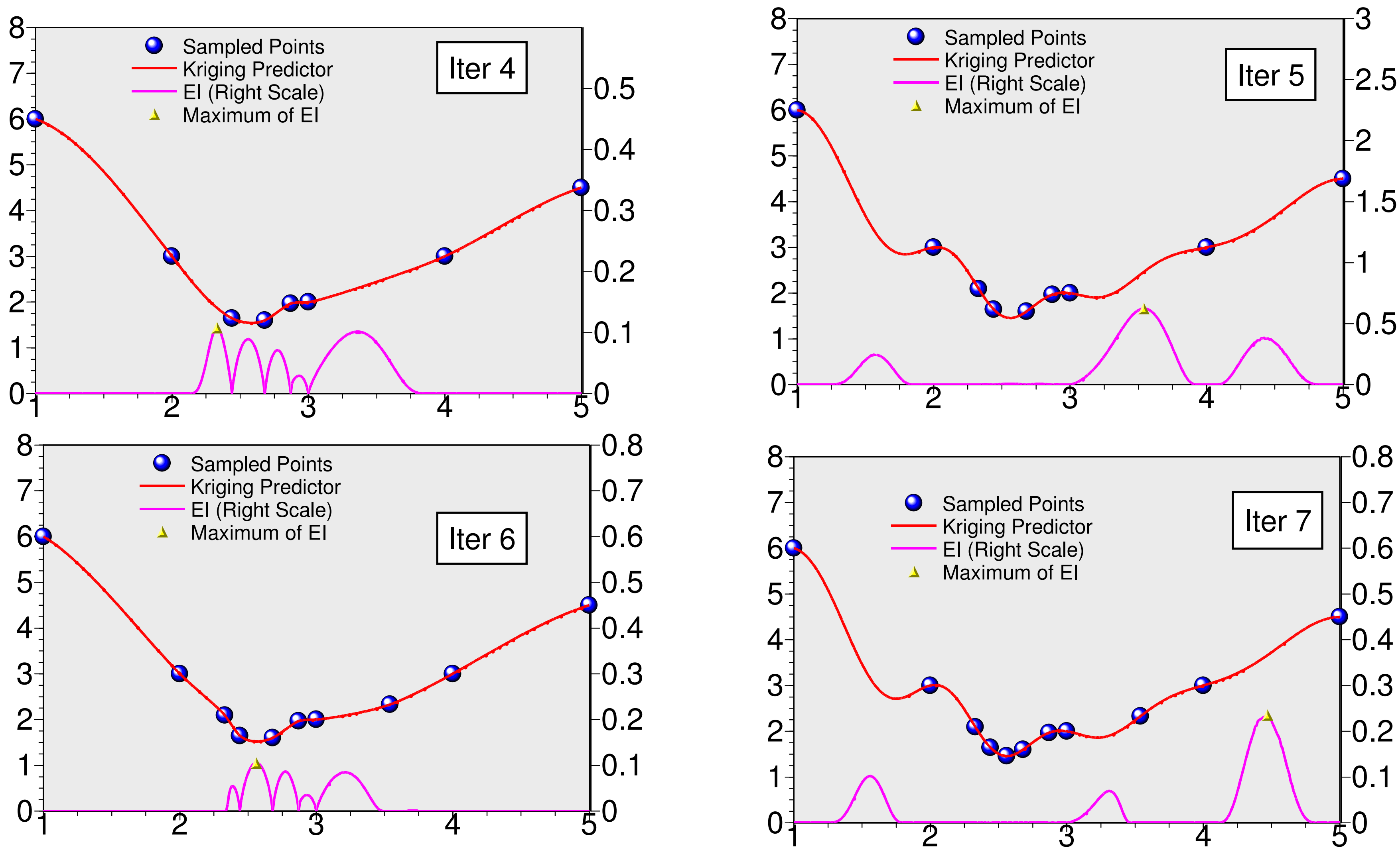
Under mild conditions, it will generate search points that fill the space (i.e., are dense). Hence it guarantees ultimate convergence.

$$\mathrm{EI}(x) = \int_{t=0}^{+\infty} t \times \left(\frac{1}{\sqrt{2\pi}\hat{\sigma}}\right) \exp\left(-\frac{(f_{\min} - t - \hat{\mu})}{2\hat{\sigma}}\right) dt$$

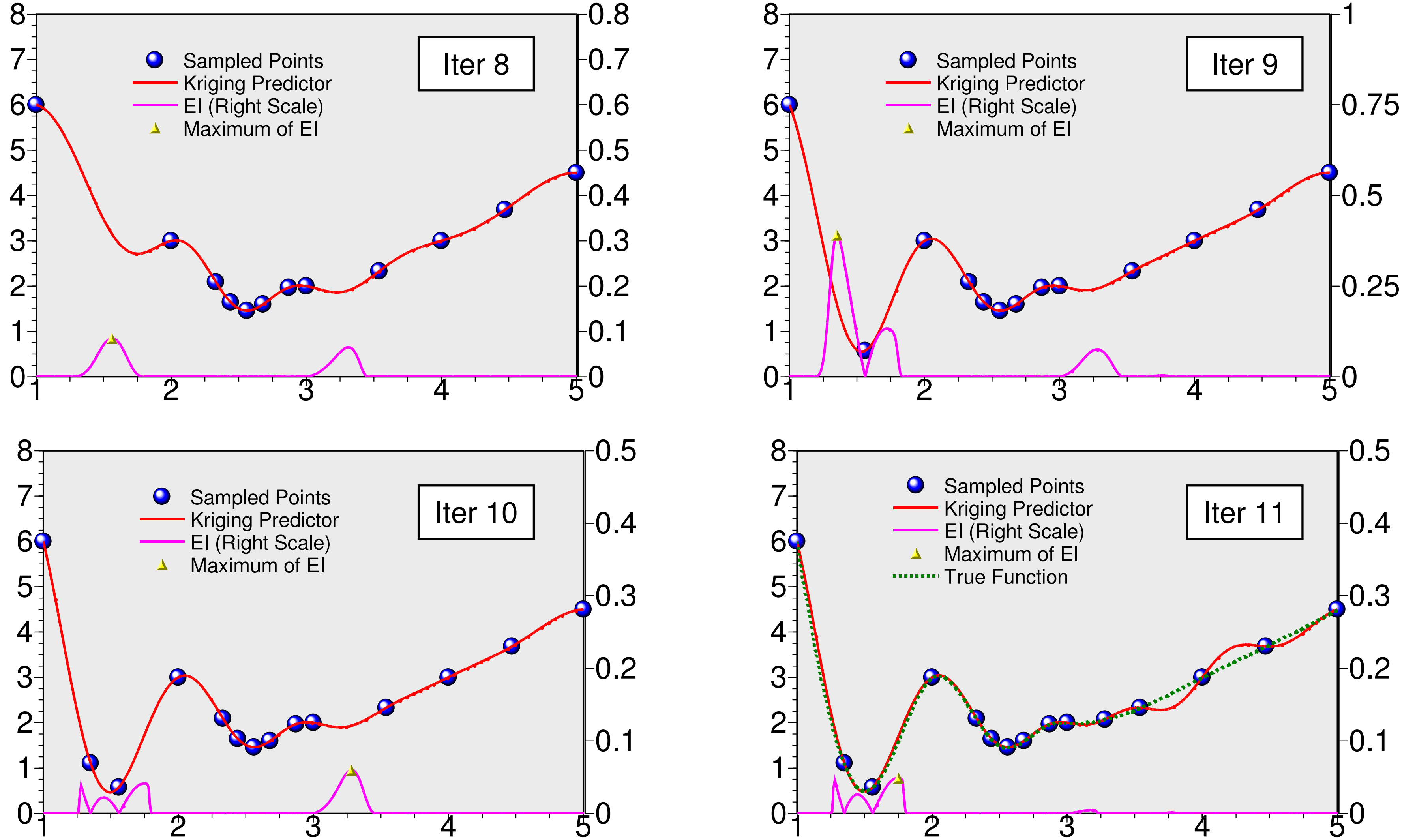
❺ Fit kriging surface, find maximum of EI, sample this point, update surface, iterate



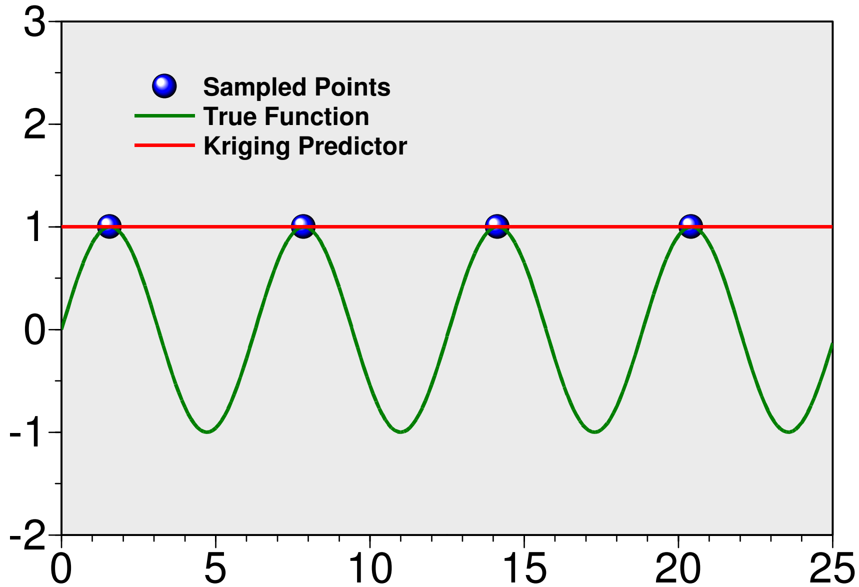
❺ Fit kriging surface, find maximum of EI, sample this point, update surface, iterate *(continued)*



❺ Fit kriging surface, find maximum of EI, sample this point, update surface, iterate *(continued)*



Method 5 can be slow or prematurely converge if initial sample causes gross underestimation of standard error

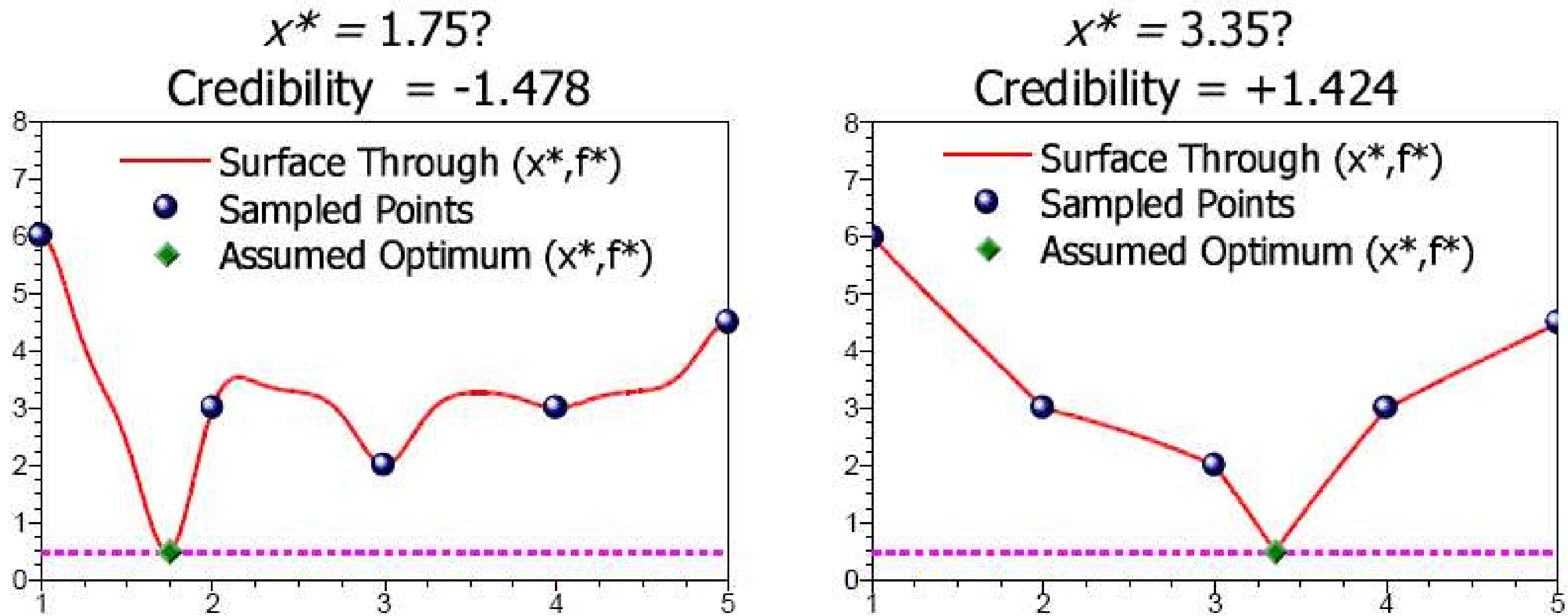


Relevant papers on the Expected Improvement algorithm

- D. Jones, M. Schonlau, W. Welch. “Efficient global optimization of expensive black-box functions.” *Journal of Global Optimization*, 13:455-492, 1998.
- M. Schonlau, W. Welch, D. Jones. “Global versus local search in constrained optimization of computer models.” In *New Developments and Applications in Experimental Design*, Lecture Notes—Monograph Series Volume 34, Institute for Mathematical Statistics, 1998.

- In the example, the standard error— and hence EI—would be 0 everywhere.
- To guarantee convergence, must force strictly positive standard errors estimates
- Can be slow if initial sample is deceptive and errors therefore underestimated
- Maximization of EI is difficult but can be done exactly via a branch-and-bound algorithm in low dimensions (< 6)
- Can vary local/global balance by maximizing $\mathrm{E}[\, I^{g}\,]$ and varying g
- Fundamental problem
  - it is a **two-stage** method: fit surface, then use surface to compute iterate
  - Possible errors in first stage are not acknowledged in the second stage

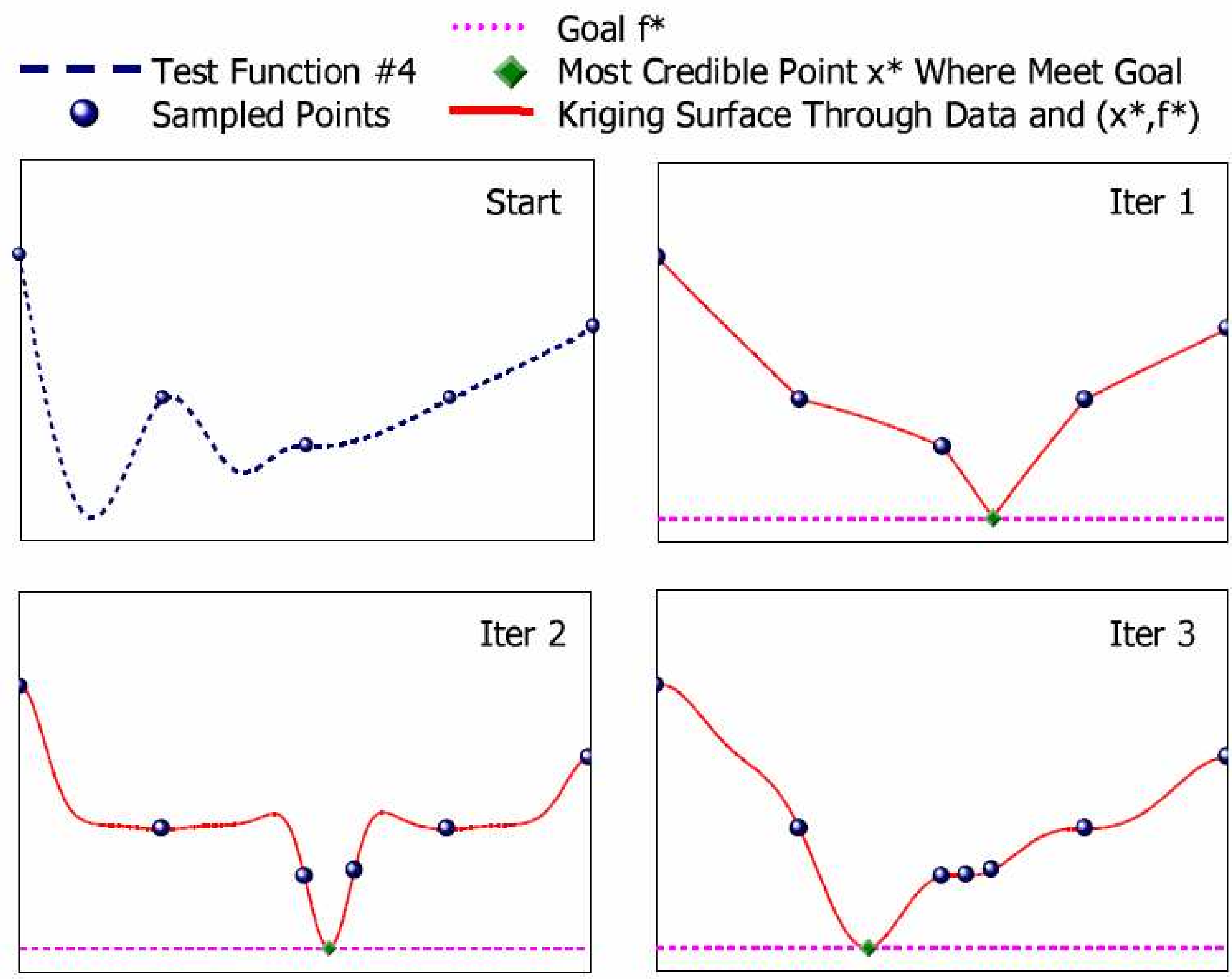
❻ Assuming optimal value $f*$ known, find point $(x, f*)$ that gives best surface fit if added to the sample



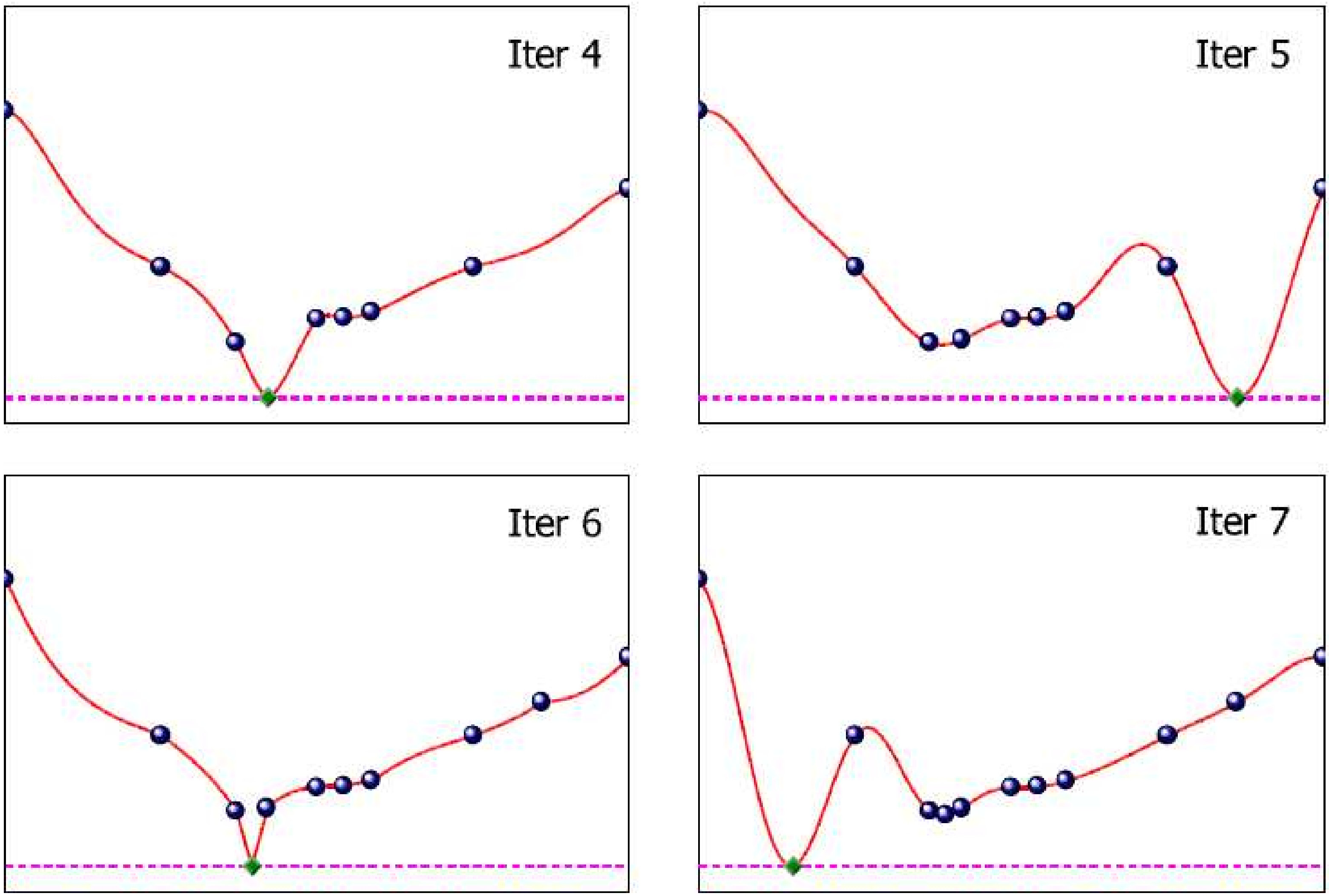
Possible measures of "credibility":

— Likelihood of sampled points conditional on surface passing through (*x, f **)
— Leave-one-out cross-validated error when point (*x, f **) is added to sample

❻ Assuming optimal value f^* known, find point (x, f^*) that gives best surface if added to sample *(continued)*



❻ Assuming optimal value f^* known, find point (x, f^*) that gives best surface if added to sample *(continued)*



The one-stage approach in a sound bite

Ask not what the surface implies about the minimum.

Ask what the minimum implies about the surface.

Obvious problem with Method 6

What if we don't know the value f?*

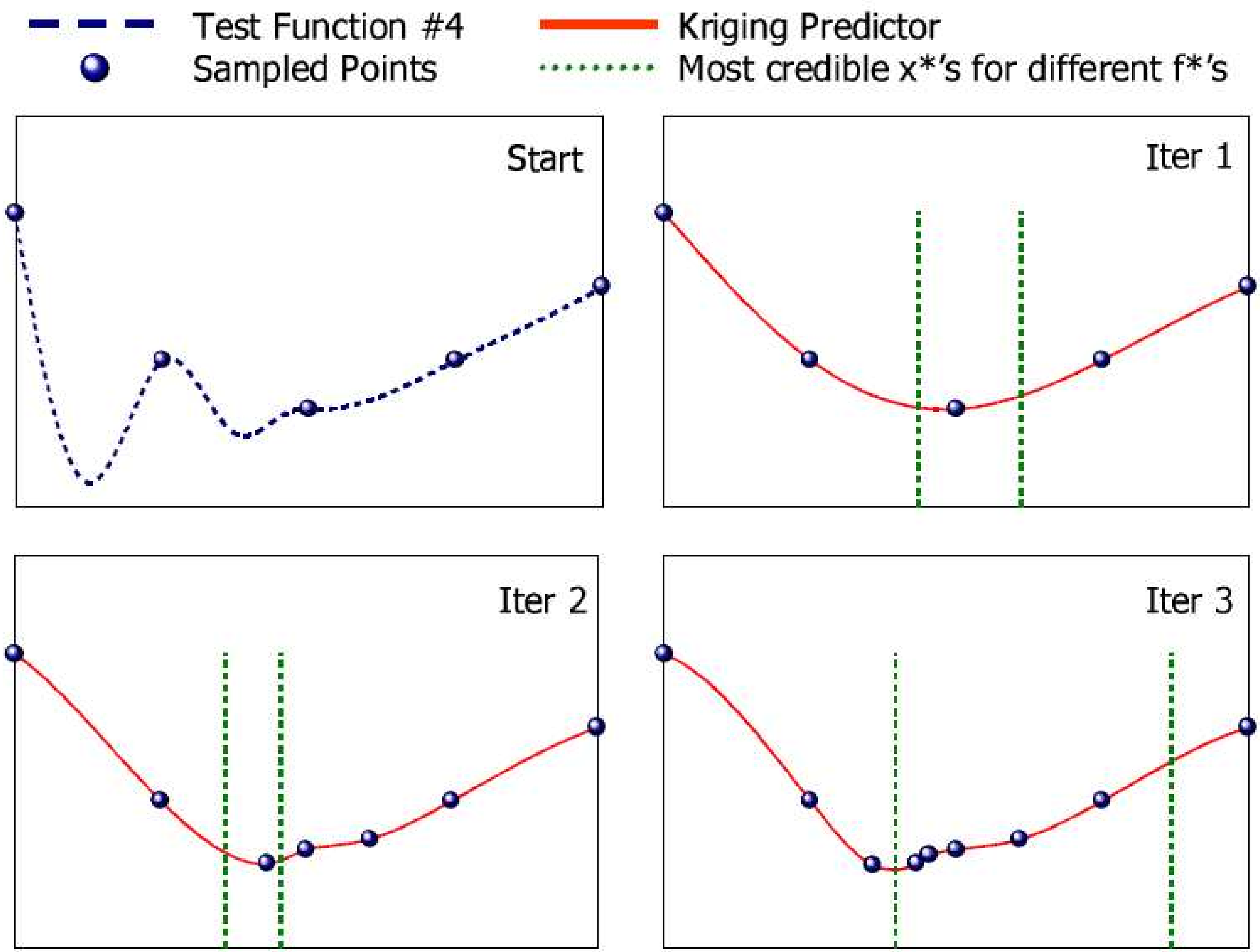
The answer should now be obvious:

Use several values for f^*!

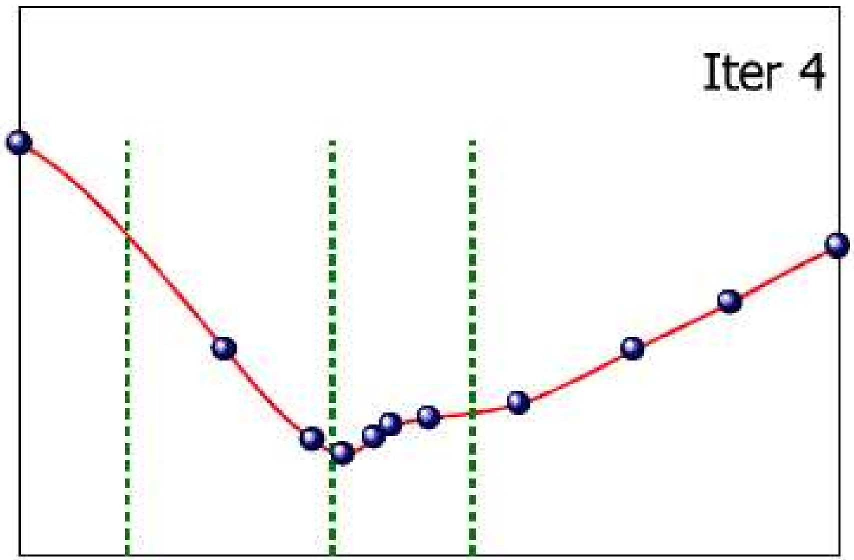
Maximize "credibility" for each one, getting a candidate next point. Then cluster all the candidates and sample one point from each cluster.

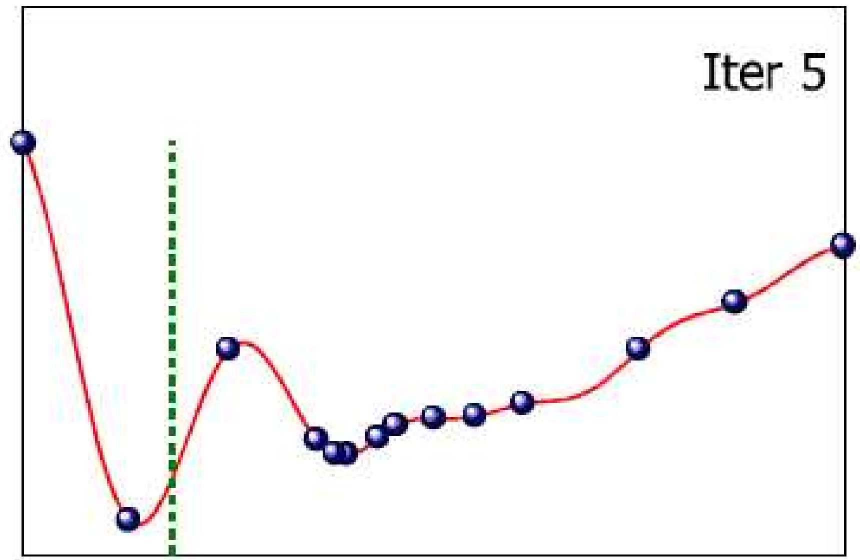
This gives Method 7

❼ Maximize "credibility" for many values of f^*, cluster solutions, and sample one point from each cluster



❼ Maximize "credibility" for many values of f^*, cluster solutions, and sample one point from each cluster





New developments

- Local optimization using radial basis functions and trust region, with convergence proof
  – R. Oeuvray and M. Bierlaire, “A New Derivative-free Algorithm for the Medical Image Registration Problem,” *From Proceeding* (429) Modeling, Simulation, and Optimization - 2004
- Study using kriging for local optimization with trust region approach. Promising numerical results!
  – Rommel Regis, Stefan Wild, Christine Shoemaker, “A Derivative-Free Trust-Region Method for Engineering Optimization,” presented at INFORMS Pittsburg 2006.

New developments

- Fresh and extremely promising way to do *constrained* global optimization with radial basis functions, without using the kriging standard error.
  - Rommel Regis and Christine Shoemaker, (2005), "Constrained Global Optimization sing Radial Basis Functions," Journal of Global Optimization, vol. 31, 153-171.
- Extension of Expected Improvement approach to handle noisy functions that require non-interpolating (smoothing) suraces
  - A. I. J. Forrester, N. W. Bressloff, A. J. Keane, "Design and analysis of 'noisy' computer experiments," *AIAA journal*, 44(10), 2331-2339

New developments

- Extremely interesting approach where next point is selected not to be the most likely to “be” the optimum as in this presentation. Instead, the next point is selected to be the one that is expected to provide the “most information” on where the optimum is.
  - J. Villemontieix, E. Vazquez, and E. Walter. “An informational approach to the global optimization of expensive to evaluate functions.” Submitted to the Journal of Global Optimization. Available on the web at http://arxiv.org/PS_cache/cs/pdf/0611/0611143.pdf

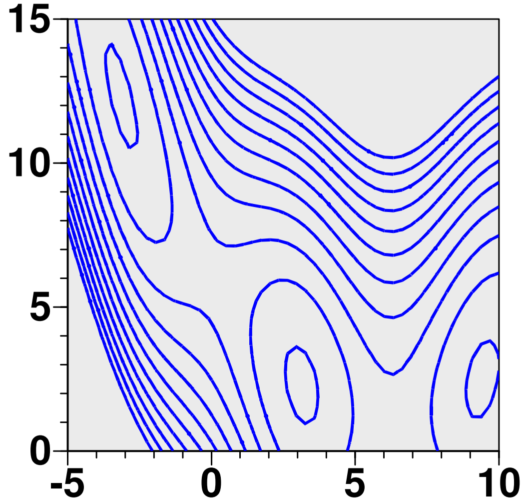
Summary

- The use of radial basis functions or kriging for efficient local search is very promising
- For global search, two-stage methods can perform poorly if the initial surface underestimates the error. This is especially true for the expected improvement approach.
- Drawback of two-stage approaches seems to be greatly reduced by forcing the surface to suggest iterates that look promising under some constraints — even if it “thinks” these values are unlikely. All the promising approaches do this in some way other.
- Handling constraints is difficult, because it is not clear how to combine the objective and constraints into a single criterion that can be optimized to find the next iterate. Regis/Shoemaker nicely finesse this issue, but more options are possible and need to be explored

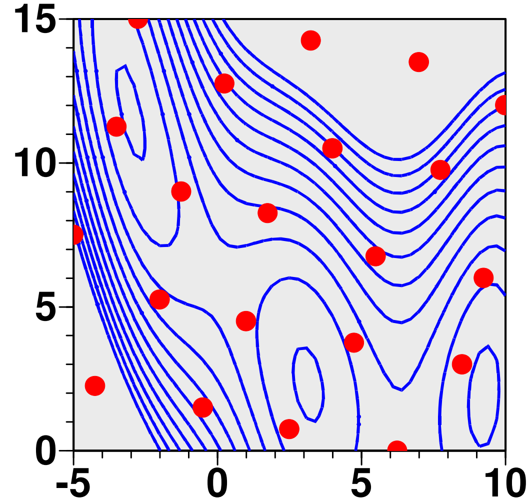
BACKUP SLIDES

Kriging compared to other methods

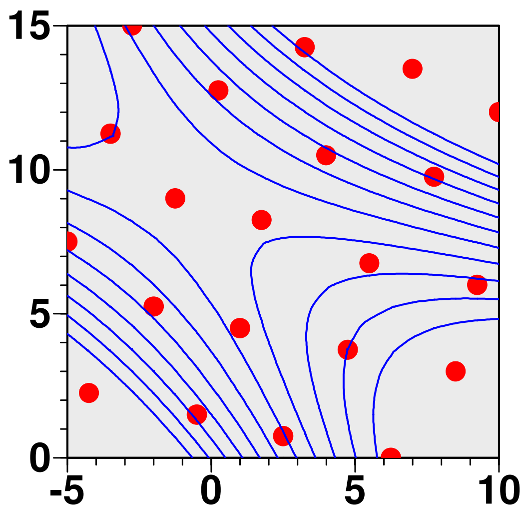
Sample Nonlinear Function



Kriging



Quadratic Surface



Thin Plate Spline