Part 2: Global Optimization

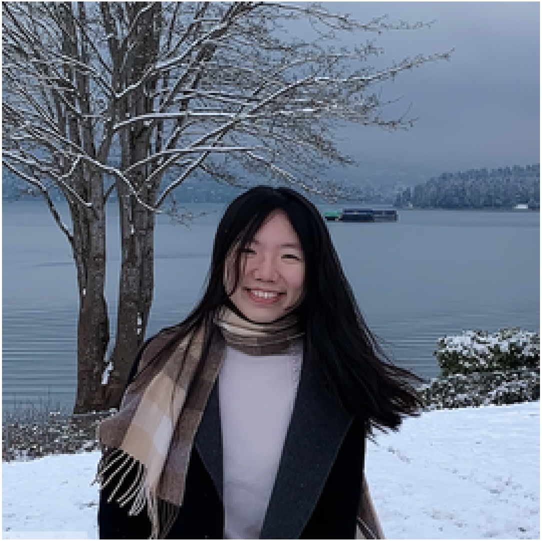
William J. Welch

University of British Columbia

Joint Work With Xiaomeng (Jasmine) Ju, New York University

School on Artificial Intelligence for Materials Science
in the Exascale Era
May 22, 2023

Xiaomeng (Jasmine) Ju



Outline of Topics

1. Strategy
2. GPs Automatically Work for Optimization? (Diagnostics)
3. Expected Improvement
4. Efficient Global Optimization (EGO)

Bayesian Optimization: The Big Picture

Brochu, Cora, and de Freitas (2010) and Frazier (2018) give reviews.

1. Make function evaluations of $y(\mathbf{x})$ at an initial experimental design of n points $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}$ in the input space $\mathbf{x}$.
2. Iterate
   1. Train a GP model using the sample of size n to give:
      - $\hat{y}(\mathbf{x})$, the predictive mean at $\mathbf{x}$
      - $s(\mathbf{x})$, the predictive standard deviation at $\mathbf{x}$.
   2. Combine $\hat{y}(\mathbf{x})$ and $s(\mathbf{x})$ in an acquisition function $a(\mathbf{x})$.
      - $a(\mathbf{x})$ is large implies an evaluation at $\mathbf{x}$ is "good" for minimization.
   3. Maximize the acquisition function:

$$\mathbf{x}^{(n+1)} = \operatorname*{argmax}_{\mathbf{x}} a(\mathbf{x}).$$

   4. Make the new evaluation $y(\mathbf{x}^{(n+1)})$ and add it to the training data.
   5. n is increased by 1.

Big Picture 2: Local Versus Global Search

- Choosing $\mathbf{x}^{(n+1)}$ for the next evaluation by minimizing $\hat{y}$:
  - Will tend to acquire evaluations in a neighbourhood of the minimum found so far;
  - Can get trapped at a local optimum.
- Choosing $\mathbf{x}^{(n+1)}$ for the next evaluation by maximizing $s(\mathbf{x})$:
  - Would eventually fill the entire space and be inefficient;
  - Global search.
- Hence an acquisition function to balance local versus global search.

The Strategy Depends on a Statistical Model

- For efficient search (few function evaluations) in high-dimensional space:
    - Much of the space has to be ruled out as not giving the minimum.
    - This is a probabilistic argument based on the GP model.
- Needs a reliable GP model!
- Check the GP model
    - Using the data from the initial design;
    - Fortunately we have diagnostic tools for checking a GP model;
    - Via cross validation.

Cross Validation (CV)

Let $\mathbf{x}^{(i)}$ denote $\mathbf{x}$ for run i in the data $(i = 1, \ldots, n)$. For run i:

- The cross validated prediction of $y(\mathbf{x}^{(i)})$ is
$$\hat{y}_{-i}(\mathbf{x}^{(i)}),$$
i.e., $\hat{y}(\mathbf{x}) = \hat{m}(\mathbf{x})$ computed from the $n-1$ runs excluding run i.
- The cross validated standard deviation of $\hat{y}_{-i}(\mathbf{x}^{(i)})$ is
$$s_{-i}(\mathbf{x}^{(i)}),$$
i.e., $s(\mathbf{x}) = \sqrt{\hat{v}(\mathbf{x})}$ computed from the $n-1$ runs excluding run i.
- The cross-validated residual for run i is
$$y(\mathbf{x}^{(i)}) - \hat{y}_{-i}(\mathbf{x}^{(i)}).$$
- The standardized cross-validated residual for run i is
$$\frac{y(\mathbf{x}^{(i)}) - \hat{y}_{-i}(\mathbf{x}^{(i)})}{s_{-i}(\mathbf{x}^{(i)})}.$$
    - Should be drawn from an approximately standard normal distribution.

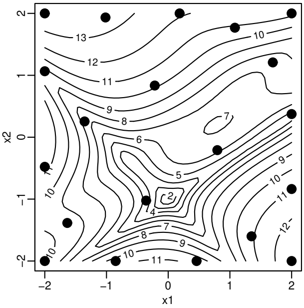
Diagnostic Plots (Jones, Schonlau, and Welch, 1998)

- Plot the observations versus the cross-validated predictions to assess the overall magnitude of error.
- Plot the standardized cross-validated residuals to assess the validity of the standard deviation for individual predictions.

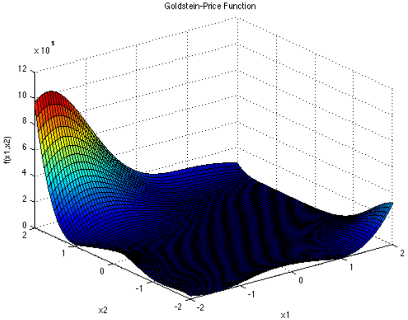
(There are others.)

Goldstein-Price: Is the GP Model Reasonable (for Bayesian Optimization)?

Recall the Goldstein-Price initial design (evaluate the function at these points)



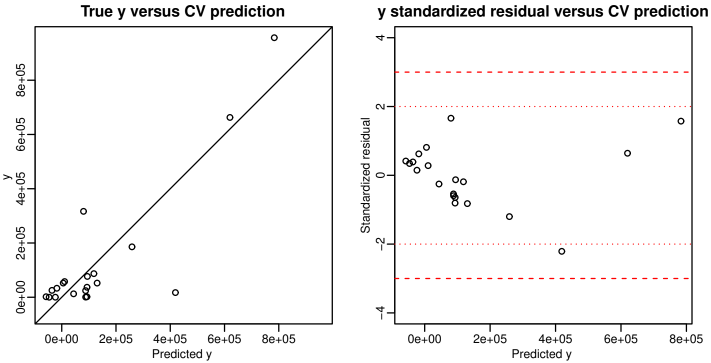
Contours in units of 10^5



https://www.sfu.ca/~ssurjano/goldpr.html

Goldstein-Price: CV Diagnostics for GP Model of y

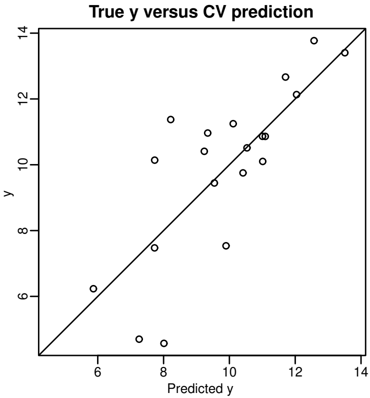
Model $y(\mathbf{x})$ as a Gaussian process

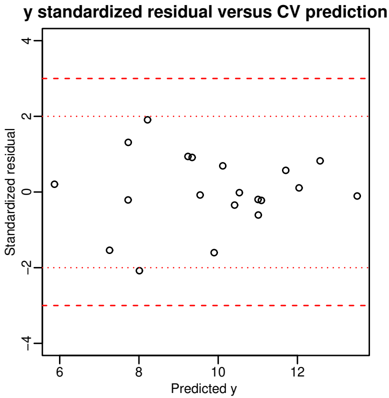


Right-side plot looks like the "funnel" diagnostic in regression.

Goldstein-Price: Diagnostics for GP Model of $\ln y$

Model $\ln y(\mathbf{x})$ as a Gaussian process





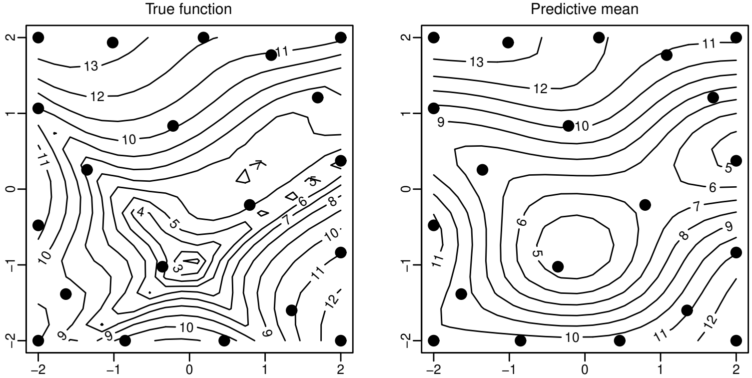
Statistical properties look much more valid!

Aquire a New Function Evaluation Using Expected Improvement

Let's tackle the Goldstein-Price minimization.

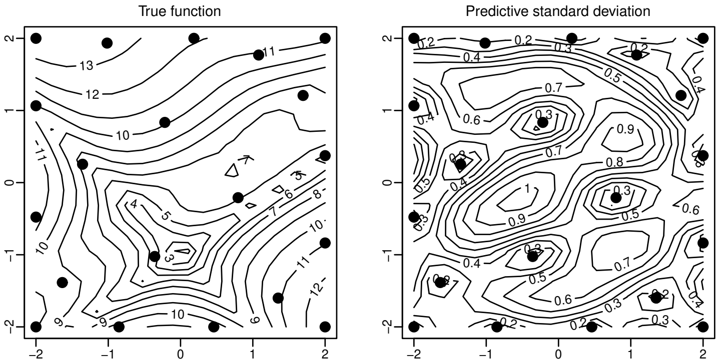
- We have a statistical model with reasonable properties.
- Iteration 1: acquire a new evaluation $y(\mathbf{x})$ at an $\mathbf{x}$ chosen by expected improvement.
- First, let's visualize what expected improvement is.

Goldstein-Price: Predictive Mean (Initial Design)



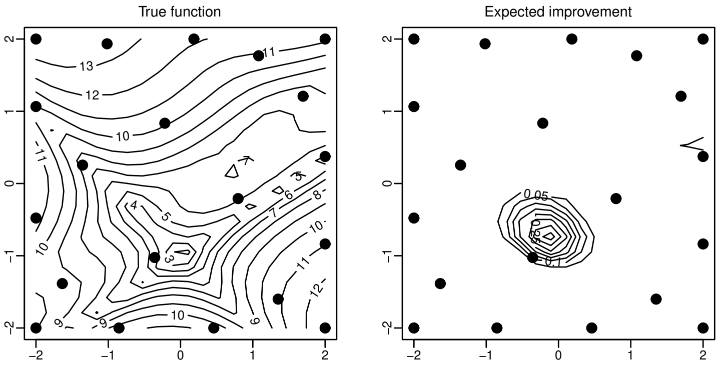
Promising locations have smaller predictive mean (we are minimizing).

Goldstein-Price: Predictive Standard Deviation



Promising locations have larger predictive standard deviation (larger uncertainty gives a chance of beating the best y value so far).

Goldstein-Price: Expected Improvement (Initial Design)



Expected improvement combines smaller predictive mean and larger predictive standard deviation.

Improvement

(See Schonlau, Welch, and Jones (1998) for more details.)

- $\hat{y}(\mathbf{x})$ is the predictive mean at $\mathbf{x}$
- $s(\mathbf{x})$ is the predictive standard deviation at $\mathbf{x}$
- The unknown $y(\mathbf{x})$ is approximately $\mathrm{N}\left(\hat{y}(\mathbf{x}), s(\mathbf{x})\right)$.
- Improvement:
  - Let $f_{\min}$ be the minimum y observed so far;
  - If $y(\mathbf{x})$ is a new evaluation, the improvement in $f_{\min}$ is

$$I(\mathbf{x}) = \max\left(f_{\min} - y(\mathbf{x}), 0\right).$$

Expected Improvement Acquisition Function

- The expected improvement has a closed form:

$$\mathrm{EI}(\mathbf{x}) = \mathrm{E}(I(\mathbf{x})) = (f_{\min} - \hat{y}(\mathbf{x}))\, \Phi\left(\frac{f_{\min} - \hat{y}(\mathbf{x})}{s(\mathbf{x})}\right) + s(\mathbf{x})\phi\left(\frac{f_{\min} - \hat{y}(\mathbf{x})}{s(\mathbf{x})}\right),$$

  where the expectation is with respect to the above normal predictive distribution, and $\Phi(\cdot)$ and $\phi(\cdot)$ are respectively the cumulative distribution function and the probability density function of the standard normal.

- Acquire the next function evaluation at $\mathbf{x}^*$, where

$$\mathbf{x}^* = \underset{\mathbf{x}}{\operatorname{argmax}}\, \mathrm{EI}(\mathbf{x}),$$

  i.e., maximize the expected improvement.

Efficient Global Optimization (EGO)

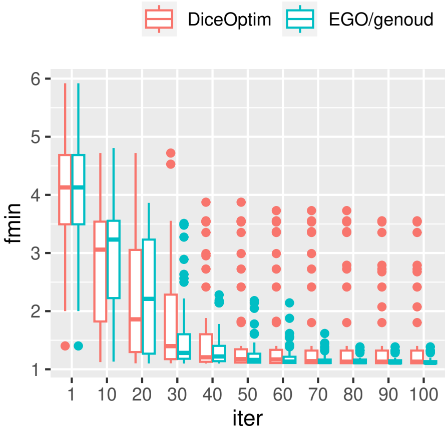
- The idea of a statistical (Bayesian) model to guide optimization has a long history, in several disciplines.
- Expected improvement goes back at least to Mockus, Tiesis, and Zilinskas (1978).
- Jones et al. (1998) implemented a more complex model (the one described today) to make Bayesian optimization more effective.
- Schonlau et al. (1998) extended the method to include constraint functions that are also expensive to compute and modelled by further GPs.

How to Maximize EI at Each Iteration?

- Bayesian optimization replaces the orginal optimization problem with another: maximize EI!
  - But EI is cheap to compute relative to a computationally expensive (exascale?) objective function.
- The original Jones et al. (1998) paper included a branch and bound method to optimize EI.
  - The `EGO` package by Ju and Welch will eventually include it.
- The following results maximize EI with the `rgenoud` package, an idea borrowed from `DiceOptim` (Roustant, Ginsbourger, and Deville, 2012).

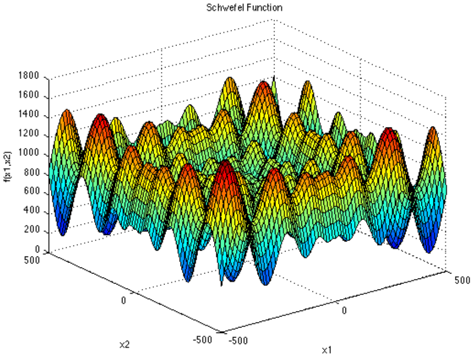
Goldstein-Price Function: Acquire 100 Evaluations

`DiceOptim` and `EGO` each repeated 50 times:



Mathematical minimum is $\ln(3) = 1.10$.
See the simple illustrative script for `EGO`.

A Bigger Challenge: Schwefel Function



https://www.sfu.ca/~ssurjano/schwef.html

$f(\mathbf{x}) = 418.9829 \times d - \sum_{i=1}^{d} \sin\left(\sqrt{|x_i|}\right)$ has minimum 0.

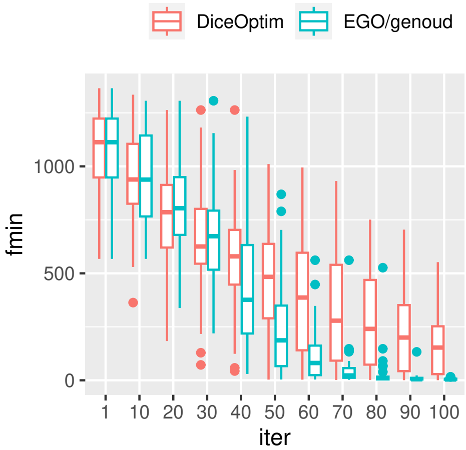
We will minimize this for $d = 5$, i.e., over $x_1, \ldots, x_5$.

Initial design has $n = 10d = 50$ evaluations

(Loeppky, Sacks, and Welch, 2009).

5-Dimensional Schwefel Function: Acquire 100 Evaluations

`DiceOptim` and `EGO` each repeated 50 times:



Mathematical minimum is 0.

Summary

- Optimization via a GP model can be efficient in terms of function evaluations.
- But it is not magic!
  - It needs a reasonably sized starting design ($n = 10d$).
  - The method relies heavily on a valid measure of uncertainty.
  - Check the GP model!
- GP calculations are subtle and need a careful implementation to train the model.

Summary

- Optimization via a GP model can be efficient in terms of function evaluations.
- But it is not magic!
  - It needs a reasonably sized starting design ($n = 10d$).
  - The method relies heavily on a valid measure of uncertainty.
  - Check the GP model!
- GP calculations are subtle and need a careful implementation to train the model.

Thank you for your attention!

References I

Brochu, E., Cora, V. M., and de Freitas, N. (2010), "A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning," *arXiv:1012.2599*.

Frazier, P. I. (2018), "A Tutorial on Bayesian Optimization," *arXiv:1807.02811*.

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

Loeppky, J. L., Sacks, J., and Welch, W. J. (2009), "Choosing the Sample Size of a Computer Experiment: A Practical Guide," *Technometrics*, 51, 366–376.

References II

Mockus, J., Tiesis, V., and Zilinskas, A. (1978), *The Application of Bayesian Methods for Seeking the Extremum*, Amsterdam: North Holland, vol. 2, p. 117129.

Roustant, O., Ginsbourger, D., and Deville, Y. (2012), "DiceKriging, DiceOptim: Two R Packages for the Analysis of Computer Experiments by Kriging-Based Metamodeling and Optimization," *Journal of Statistical Software*, 51, 1–55.

Schonlau, M., Welch, W. J., and Jones, D. R. (1998), "Global Versus Local Search in Constrained Optimization of Computer Models," in *New Developments and Applications in Experimental Design*, eds. Flournoy, N., Rosenberger, W. F., and Wong, W. K., Hayward, California: Institute of Mathematical Statistics, pp. 11–25.