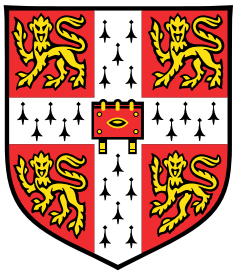
Minimising a Noisy Expensive Function Using Active Learning



Ross Brown

Supervisor: Dr. O. Orhobor

Department of Chemical Engineering and Biotechnology
University of Cambridge

This dissertation is submitted for the degree of
Master of Engineering

Robinson College September 2021

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 65,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Ross Brown
September 2021

Abstract

This is where you write your abstract ...

Table of contents

List of figures

List of tables

Chapter 1

Introduction

Finding the global minimum of a function within a set of boundaries is a problem of major import. From optimising a synthetic pathway in drug development, to minimising the error in a neural network, minimisation is vitally important to mathematics. Within the numerical field, the goal is usually two fold: reduce the error, ε, to the true value *and* reduce the processing time. With these goals in mind, the majority of algorithms exploit the commonality of the cheapness of the target function. However, this is not always the case. Take as an example an experimentation of sand grain size, d, on the strength of concrete, τ. An underlying function of the form $\tau = f(d)$ exists, but each call to this function takes at least a day, and is labour and material expensive. The target of this paper is to explore how to minimise such a fucntion with the fewest function calls.

1.1 Problem Definition

$$y = f(\boldsymbol{x}) \tag{1.1}$$

Given an equation $y = f(\boldsymbol{x})$ where $\boldsymbol{x}$ is a vector with $x_i[\alpha_i, \beta_i]$, and y is scalar, find the solution to argmin$[f(\boldsymbol{x})]$. The algorithm will be able to invoke $g(\boldsymbol{x})$ as shown in [] with ε representing an unknown random error.

$$g(\boldsymbol{x}) = f(\boldsymbol{x}) + \varepsilon \tag{1.2}$$

Chapter 2

Simple 1-Dimentional Problem

2.1 Outlining of Basic Principles

By constraining $\boldsymbol{x}$ to one dimention allows for the problem to be simplified. Suppose $f(x) = \sin(x) + 0.05x^2$ with $x \in [-10, 10]$ as shown in Figure 2.1. In this range, there are multiple minima with only one global minima. The task here is to successfully locate the minima situated at -1.428 (found through ananlytical differentiation and solving $10\cos(x) + x = 0$). This will be compared against two benchmarks: linear spacing and the fminbound method available in the scipy package given the same number of function calls. ε will be chosen to be independent of x and y and fit a normal distribution such that $\varepsilon \sim N(0, 0.2)$.

2.1.1 Algorithms

Linear Selection

The least inteligent method while not being deliberately obtuse is linearly spacing sampled values and choosing x equal to the lowest sampled value. This does have the advantage that all experimentations may be construed asynchronously. Thus, where material and labour cost is low, this may be beneficial.

Active Learning

This methodolgy has two underlying core principles: sparse areas reveal the most information and minimal areas reveal information to the location of the minima. Combining these allows for better decision making with regards to the next sample to choose.

There are several methods that may be used to enhance this strategy. Firstly, a smoothing spline between points allows for a non-parametric fit of the data to be used. Alternatively, a

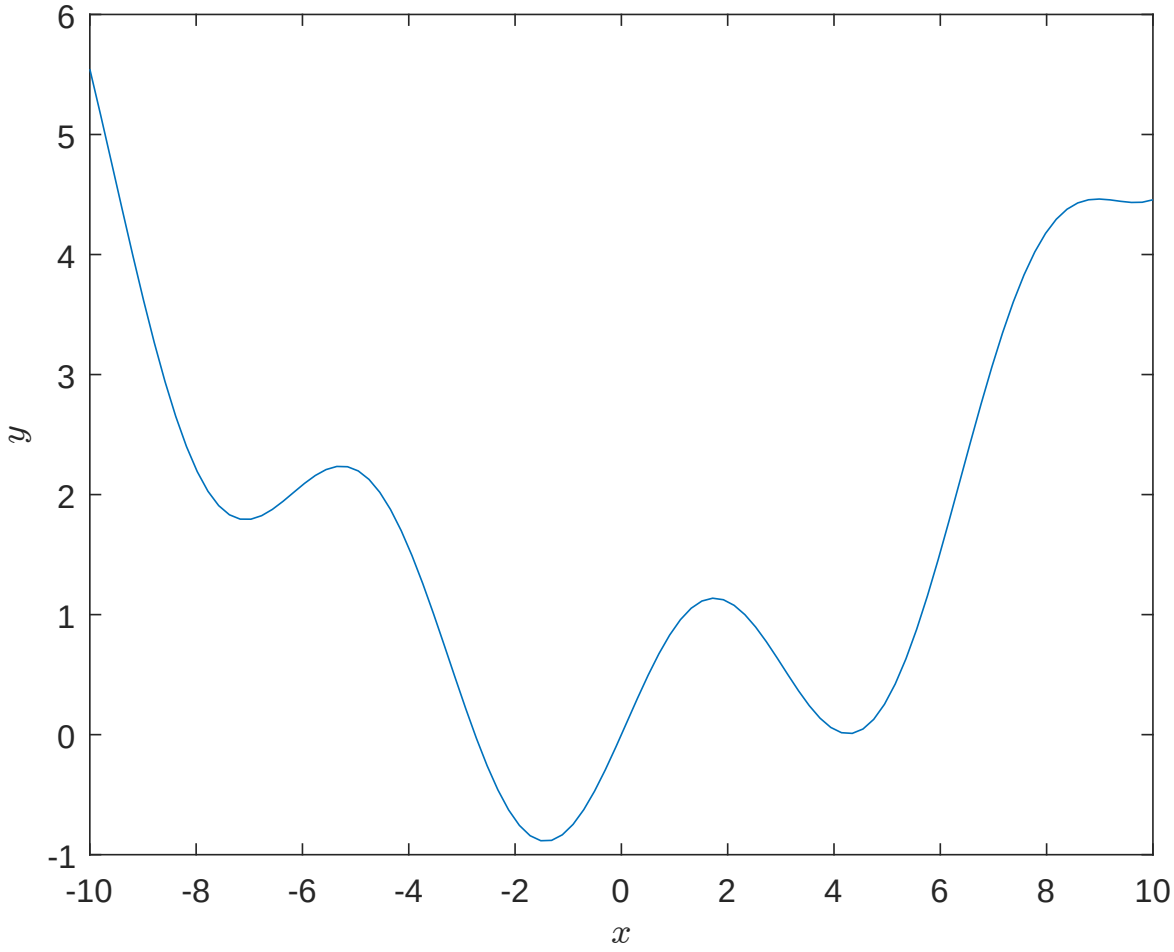


Fig. 2.1 $y = \sin(x) + 0.005x^2$ with $x \in [-10, 10]$

local regression fit may be used allowing for changes in sample density. Advanced methods using Baysian Statistics and advanced information theory may be used, although have been ommitted due to time restraints.

In this paper, three functions are found: $e(x)$, $p(x)$ and $h(x)$, denoting a guessed fit, the sparcity, and the height respectively. $p(x)$ is defined between adjacent samples, s_i.

$$h(x) = -e(x) + \max[e(x)] \tag{2.1}$$

$$p(s_i \leq x \leq s_{i+1}) = \min[x - s_i, s_{i+1} - x] \tag{2.2}$$

The next sample is then taken as $\text{argmax}[h(x)p(x)]$.

fminbound

fminbound is a function included in the scipy optimisation library. It uses Brent's method allowing it to be quick in situations where labeling is quick and error is low.

2.1.2 Comparison on $\sin(x) + 0.005x^2$

Each method discussed in [] was executed 50 times for each sample size between 2 and 15.

Itemize

- The first topic is dull
- The second topic is duller
  - The first subtopic is silly
  - The second subtopic is stupid
- The third topic is the dullest

Description

The first topic is dull

The second topic is duller

The first subtopic is silly

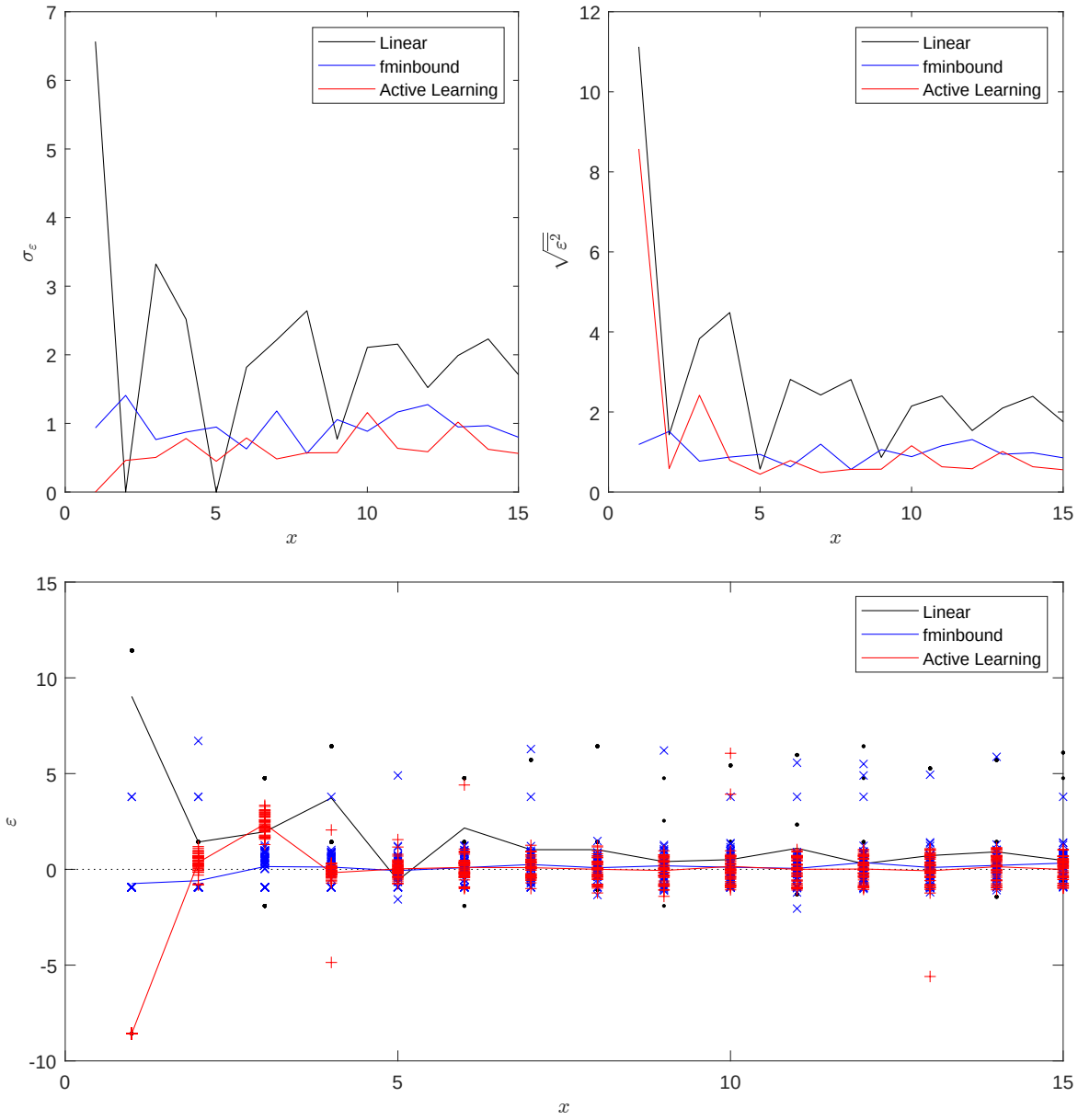
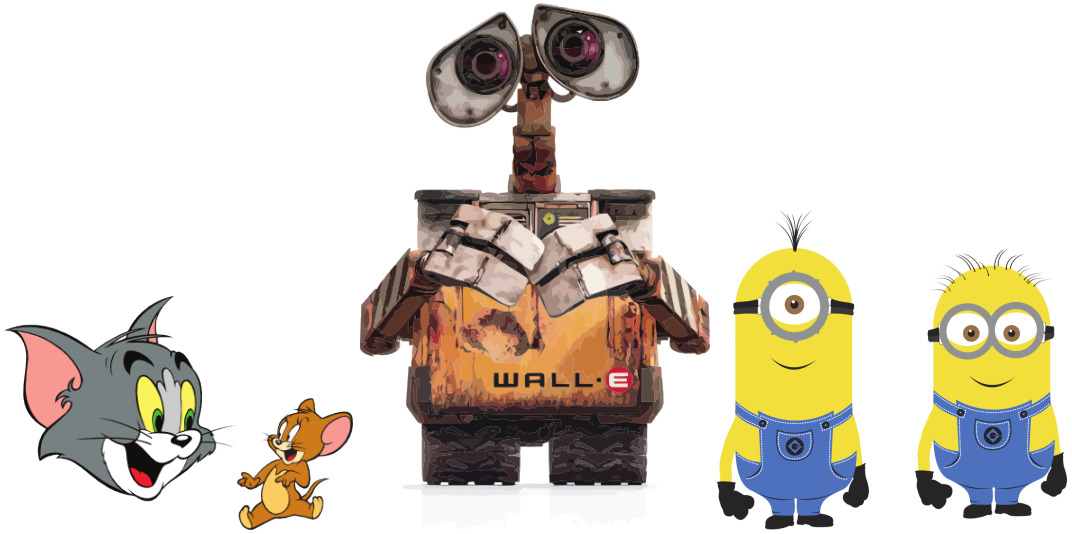


Fig. 2.2 Comparison of the three methods discussed. The top left shows the standard deviation of each method, the top right shows the square of the residues, and the bottom shows the result of each sample with an average of the residue drawn on to guide the eye.

The second subtopic is stupid

The third topic is the dullest

2.2 Hidden section



(a) Tom and Jerry (b) Wall-E (c) Minions

Fig. 2.3 Best Animations

Subplots

I can cite Wall-E (see Fig. 2.3b) and Minions in despicable me (Fig. 2.3c) or I can cite the whole figure as Fig. 2.3

Chapter 3

My third chapter

3.1 First section of the third chapter

And now I begin my third chapter here . . .

And now to cite some more people Read [2], Ancey et al. [1]

3.1.1 First subsection in the first section

. . . and some more

3.1.2 Second subsection in the first section

. . . and some more . . .

First subsub section in the second subsection

. . . and some more in the first subsub section otherwise it all looks the same doesn't it? well we can add some text to it . . .

3.1.3 Third subsection in the first section

. . . and some more . . .

First subsub section in the third subsection

. . . and some more in the first subsub section otherwise it all looks the same doesn't it? well we can add some text to it and some more and some more and some more and some more and some more and some more and some more . . .

Second subsub section in the third subsection

. . . and some more in the first subsub section otherwise it all looks the same doesn't it? well we can add some text to it . . .

3.2 Second section of the third chapter

and here I write more . . .

3.3 The layout of formal tables

This section has been modified from "Publication quality tables in LaTeX*" by Simon Fear.

The layout of a table has been established over centuries of experience and should only be altered in extraordinary circumstances.

When formatting a table, remember two simple guidelines at all times:

1. Never, ever use vertical rules (lines).
2. Never use double rules.

These guidelines may seem extreme but I have never found a good argument in favour of breaking them. For example, if you feel that the information in the left half of a table is so different from that on the right that it needs to be separated by a vertical line, then you should use two tables instead. Not everyone follows the second guideline:

There are three further guidelines worth mentioning here as they are generally not known outside the circle of professional typesetters and subeditors:

3. Put the units in the column heading (not in the body of the table).
4. Always precede a decimal point by a digit; thus 0.1 *not* just .1.
5. Do not use 'ditto' signs or any other such convention to repeat a previous value. In many circumstances a blank will serve just as well. If it won't, then repeat the value.

A frequently seen mistake is to use '\begin{center}' . . . '\end{center}' inside a figure or table environment. This center environment can cause additional vertical space. If you want to avoid that just use '\centering'

Table 3.1 A badly formatted table

	Species I		Species II	
Dental measurement	mean	SD	mean	SD
I1MD	6.23	0.91	5.2	0.7
I1LL	7.48	0.56	8.7	0.71
I2MD	3.99	0.63	4.22	0.54
I2LL	6.81	0.02	6.66	0.01
CMD	13.47	0.09	10.55	0.05
CBL	11.88	0.05	13.11	0.04

Table 3.2 A nice looking table

Dental measurement	Species I		Species II	
	mean	SD	mean	SD
I1MD	6.23	0.91	5.2	0.7
I1LL	7.48	0.56	8.7	0.71
I2MD	3.99	0.63	4.22	0.54
I2LL	6.81	0.02	6.66	0.01
CMD	13.47	0.09	10.55	0.05
CBL	11.88	0.05	13.11	0.04

Table 3.3 Even better looking table using booktabs

Dental measurement	Species I		Species II	
	mean	SD	mean	SD
I1MD	6.23	0.91	5.2	0.7
I1LL	7.48	0.56	8.7	0.71
I2MD	3.99	0.63	4.22	0.54
I2LL	6.81	0.02	6.66	0.01
CMD	13.47	0.09	10.55	0.05
CBL	11.88	0.05	13.11	0.04

References

[1] Ancey, C., Coussot, P., and Evesque, P. (1996). Examination of the possibility of a fluid-mechanics treatment of dense granular flows. *Mechanics of Cohesive-frictional Materials*, 1(4):385–403.

[2] Read, C. J. (1985). A solution to the invariant subspace problem on the space l_1. *Bull. London Math. Soc.*, 17:305–317.

Appendix A

How to install LaTeX

Windows OS

TeXLive package - full version

1. Download the TeXLive ISO (2.2GB) from
   https://www.tug.org/texlive/
2. Download WinCDEmu (if you don't have a virtual drive) from
   http://wincdemu.sysprogs.org/download/
3. To install Windows CD Emulator follow the instructions at
   http://wincdemu.sysprogs.org/tutorials/install/
4. Right click the iso and mount it using the WinCDEmu as shown in
   http://wincdemu.sysprogs.org/tutorials/mount/
5. Open your virtual drive and run setup.pl

or

Basic MikTeX - TeX distribution

1. Download Basic-MiKTeX(32bit or 64bit) from
   http://miktex.org/download
2. Run the installer
3. To add a new package go to Start » All Programs » MikTex » Maintenance (Admin)
   and choose Package Manager

4. Select or search for packages to install

TexStudio - TeX editor

1. Download TexStudio from
   http://texstudio.sourceforge.net/#downloads

2. Run the installer

Mac OS X

MacTeX - TeX distribution

1. Download the file from
   https://www.tug.org/mactex/

2. Extract and double click to run the installer. It does the entire configuration, sit back and relax.

TexStudio - TeX editor

1. Download TexStudio from
   http://texstudio.sourceforge.net/#downloads

2. Extract and Start

Unix/Linux

TeXLive - TeX distribution

Getting the distribution:

1. TexLive can be downloaded from
   http://www.tug.org/texlive/acquire-netinstall.html.

2. TexLive is provided by most operating system you can use (rpm,apt-get or yum) to get TexLive distributions

Installation

1. Mount the ISO file in the mnt directory

   ```
   mount -t iso9660 -o ro,loop,noauto /your/texlive####.iso /mnt
   ```

2. Install wget on your OS (use rpm, apt-get or yum install)

3. Run the installer script install-tl.

   ```
   cd /your/download/directory
   ./install-tl
   ```

4. Enter command 'i' for installation

5. Post-Installation configuration:
   http://www.tug.org/texlive/doc/texlive-en/texlive-en.html#x1-320003.4.1

6. Set the path for the directory of TexLive binaries in your .bashrc file

For 32bit OS

For Bourne-compatible shells such as bash, and using Intel x86 GNU/Linux and a default directory setup as an example, the file to edit might be

```
edit $~/.bashrc file and add following lines
PATH=/usr/local/texlive/2011/bin/i386-linux:$PATH;
export PATH
MANPATH=/usr/local/texlive/2011/texmf/doc/man:$MANPATH;
export MANPATH
INFOPATH=/usr/local/texlive/2011/texmf/doc/info:$INFOPATH;
export INFOPATH
```

For 64bit OS

```
edit $~/.bashrc file and add following lines
PATH=/usr/local/texlive/2011/bin/x86_64-linux:$PATH;
export PATH
MANPATH=/usr/local/texlive/2011/texmf/doc/man:$MANPATH;
export MANPATH
```

```
INFOPATH=/usr/local/texlive/2011/texmf/doc/info:$INFOPATH;
export INFOPATH
```

Fedora/RedHat/CentOS:

```
sudo yum install texlive
sudo yum install psutils
```

SUSE:

```
sudo zypper install texlive
```

Debian/Ubuntu:

```
sudo apt-get install texlive texlive-latex-extra
sudo apt-get install psutils
```

Appendix B

Installing the CUED class file

LaTeX.cls files can be accessed system-wide when they are placed in the <texmf>/tex/latex directory, where <texmf> is the root directory of the user's TeXinstallation. On systems that have a local texmf tree (<texmflocal>), which may be named "texmf-local" or "localtexmf", it may be advisable to install packages in <texmflocal>, rather than <texmf> as the contents of the former, unlike that of the latter, are preserved after the LaTeXsystem is reinstalled and/or upgraded.

It is recommended that the user create a subdirectory <texmf>/tex/latex/CUED for all CUED related LaTeXclass and package files. On some LaTeXsystems, the directory look-up tables will need to be refreshed after making additions or deletions to the system files. For TeXLive systems this is accomplished via executing "texhash" as root. MIKTeXusers can run "initexmf -u" to accomplish the same thing.

Users not willing or able to install the files system-wide can install them in their personal directories, but will then have to provide the path (full or relative) in addition to the filename when referring to them in LaTeX.