Optimization Application Diversity and Complexity

Thursday, April 4, 2019 1:25 PM

Optimization is a procedure that seeks the best and is relevant to a wide range of applications, but uses a common statement:

$$\min_{\{DV\}} J = OF \tag{2.1}$$

However, there is a wide diversity of applications presenting a variety of characteristics that should be considered when choosing an optimizer.

Diverse applications of optimizations may involve:

Nonlinear
Min–max (or so)
Integer or other discretization

Stochastic
Underspecified
Rank or Categorization of OF
Constraints
Discontinuous Models
Saddle Points
Multiple Optima
Economic
Regression
Reliability
Path Analysis
Dynamic Model
Conditionals

Nonlinearity

Consider the simple example of determining the dimensions of rectangular box to meet a desired volume while minimizing surface area. If the sides have dimensions of height, width, and length, but no thickness, and if the sides can be constructed from ideal planes, then the volume is $V = hwl$ and the surface area is $S = 2hw + 2lw + 2hl$. Since choosing two dimensions constrains the third to meet the volume constraint, the optimization statement is

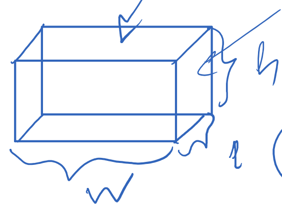
$$\min_{\{h,w\}} J = S = 2hw + \frac{2V}{h} + \frac{2V}{w} \tag{2.2}$$

Solving analytically, by setting the derivatives of the OF w.r.t. each DV to zero, results in the following set of equations:

$$wl = \frac{V}{h}$$

$$A = l \cdot h = \frac{V}{w}$$

$$0 = 2w - \frac{2V}{h^2}$$
$$0 = 2h - \frac{2V}{w^2} \tag{2.3}$$



$$V = w \cdot l \cdot h$$

$$\cancel{2}wh^2 - \cancel{2V} = \cancel{2}hw^2 - \cancel{2V}$$

$$h = w$$

$h = w$

$2w - \frac{2V}{w^2} = 0$

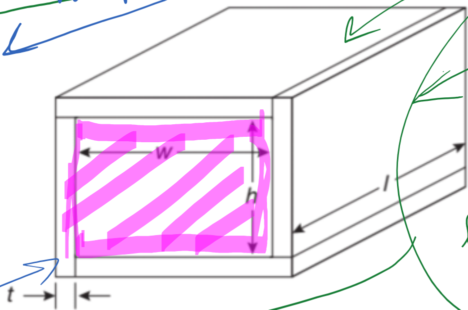
$w^3 = V$

$w = \sqrt[3]{V} = h \Rightarrow$ cube!

$l = \frac{V}{wh} = \frac{V}{\sqrt[3]{V}\cdot\sqrt[3]{V}}$

$l = \sqrt[3]{V}$

More accurate approach:



$t(w+2t)$

$l\cdot(h+2t)$

Now: $l = \frac{V}{wh} + 2t$

The optimization statement is

$$\min_{\{h,w\}} J = S = 2hw + 2(h+w+4t)\left(\frac{V}{hw}+2t\right) + 4t(h+w+2t) \tag{2.4}$$

Now, taking the calculus derivatives without error will be a challenge for many, and if done, then attempting to solve the system of nonlinear relations is another challenge.

Whether mathematical complexity is the result of nonlinear relations or not, for realistic versions of even simple applications, the mathematics often become intractable—or effectively so for an engineer who is not also a hobbyist mathematician.

Min-Max, Max-Min, etc.

Examples:

1. Getting from point A to point B along a path.

   Minimizes the steepest slope in addition to the overall path length. The procedure would be to look at all of the slopes along the path, find the maximum, and seek a path that minimizes the maximum.

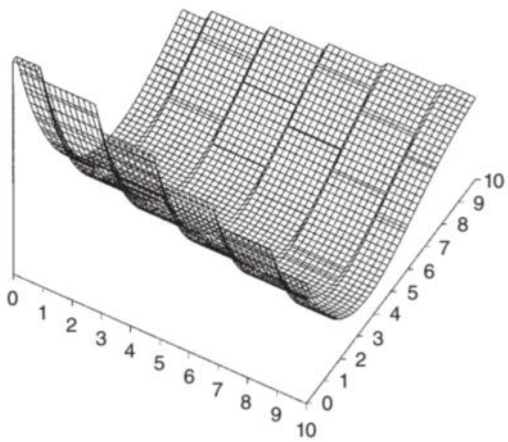
2. Business:

   Rather than choose DV values to maximize the profitability index at the nominal situation forecast, a better objective would be to search over the range of the external influence values, then to see what the worst possible (the minimum possible) outcome is, and then to maximize the minimum over all situations.

3. Allocation of investments for retirement.

Integers and other discretization

Variables can be integers. They might come from a list of possible values.



Conditionals and Discontinuities: Cliff Ridges/Valleys

Conditionals lead to either cliff or sharp valley/ridge discontinuities in the OF response to the DV. These invalidate local derivative evaluation and invalidate the basis of gradient-based, Newton-type, and surrogate-model optimizers that presume continuum-valued surfaces.

Whether gradient-based or direct search algorithms, sharp valleys tend to make the trial solution jump back and forth from side to side across the valley, rather than to follow it downhill.

Procedures not equations

Contrasting most textbook examples, applications of utility are usually not convenient, one-line, mathematical statements of either the OF or the constraints. Although simple examples are important to understanding, in my experience what we seek to optimize for either career or personal life does not lead to simple one-line equations. What I encounter are procedures—computer simulations that are solved by numerical techniques.

One issue is constraints. Procedures often have a myriad of internal variables, which can lead to execution errors such as divide by zero, log of a negative number, or subscript out of range. There may be path integrals and unknowable constraints until the procedure finds it on the path. It may not be possible to express such constraints as a simple function of the DVs.

Another issue is obtaining derivatives and second derivatives. Without a closed-form, analytically tractable relation between OF and DV, optimizers need to estimate derivatives numerically. Here, choices of increment size and whether forward, backward, or central difference can affect the solution.

Static and Dynamic Models

Static models are most often used in process design and process analysis. But dynamic models are essential to process control and scheduling, in which understanding and shaping the path from one state to another is the objective. Since a dynamic model will settle to the steady-state value, the dynamic model must contain the functionality of the static model. Additionally, dynamic models include the time dependence. The issues related to how something evolves over time adds several aspects to optimization.

Path Integrals

When the objective is to maximize or minimize the accumulation or consumption along a path, one has an integral along a curve. Perhaps it is to choose a car accelerator position schedule that minimizes the consumption of fuel from point A to point B along a road. Perhaps it is to choose a time of day for travel that minimizes the number of bug splats accumulated on a windshield along a particular road. The path does not need to be a road or a physical path through Cartesian space. The objective might be to determine control action for a process to transition from one product to another with minimum total waste. Here the path is how the process state (temperature, composition, etc.) changes in time. Chapter 5 shows how to analyze along a path.

Economic Optimization and non-additive cost functions

$A = P(1+r)^i$

$FV = PV(1+r)^i$

Generally: want to balance benefit with the initial cost and future expenses.
Typically, minimize the capital investment and maximize the cash flow.
Compare the present value of everything, using compound interest formulas.
See p. 39 for more details.

$\dfrac{F}{(1+r)^i}$ ← # of conv. periods

rate per conv. period

Reliability

A light bulb might fail. If there is only one bulb in the room, when the bulb fails, the room goes dark. So, place two bulbs in the light fixture, and when one bulb fails, one bulb is left to provide light to keep the room functional and to provide light so that the bad bulb can be changed. But the second bulb might fail before the first can be fixed. There are different questions that can be asked. A simple one is "How many bulbs are needed to have the room functional 99.999% of the instances of use?" The situation may be that the chance of failure of each bulb is independent and that when one of more bulbs is operating, the room is functional. Then the probability of at least one out of N bulbs being operable is the same as the probability of either 1 or 2 or 3 or 4 or … N out of N being operational, which is the same as the complement the probability that none is operable. Using $p_{\text{individual}}$ as the probability that a bulb works on demand,

$$P_{\text{desired system reliability}} = P(n > 1|N) = 1 - \overbrace{(1 - p_{\text{individual}})}^{\text{lamps fails}}{}^{N} \tag{2.5}$$

known fields

$$P_{\text{desired system reliability}} = P(n > 1|N) = 1 - \left(1 - p_{\text{individual}}\right)^N \tag{2.5}$$

Solving for the number of bulbs,

0.9999

$$N = \frac{\ln\left(1 - P_{\text{desired for system}}\right)}{\ln\left(1 - p_{\text{individual}}\right)} \tag{2.6}$$

given

$(1-p_{ind})^N = 1 - P_{des}$

$N \ln(1-p_{ind}) = \ln(1-P_{d})$

This is a simple introductory probability example, which does not need optimization to solve for the DV value, the number of bulbs.

Optimization might aim to balance system reliability with cost. If c_1 is the price of one bulb, then $c_1 N$ is the cost of N bulbs. If c_2 is the opportunity cost of a nonfunctional room, then $c_2\left(1 - P_{\text{system functionality}}\right)$ is the risk. The optimization might be to determine N to minimize the initial and opportunity cost:

$$\min_{\{N\}} J = c_1 N + c_2\left(1 - P_{\text{system functionality}}\right) = c_1 N + c_2\left(1 - p_{\text{individual}}\right)^N \tag{2.7}$$

Although this is a simple application, it reveals that the decision variable is an integer, not a continuum-valued number, which is a difficulty for many optimization algorithms.

Regression - as illustrated last week

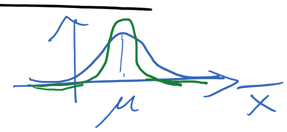
Deterministic vs stochastic models

$\bar{x} \to$ normal

Sometimes, in very idealized situations we can derive the average or statistical expectation. However, in complicated applications (such as in a game or business projection in which the outcome is subject to many rules, influences, and conditional inputs) an analytical average is not possible to derive. Another approach to identify the expectation is to increase the number of realizations and average the results. But, although variation reduces, it is not eliminated. The central limit theorem indicates that the range on the variable drops with the square root of the number of realizations.

$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

$\bar{x}$:



In structuring a business, for instance, we might forecast possible outcomes considering the vagaries of possible future tax changes, raw material and energy costs, customer demand, etc., to see the result of our business choices. We might have a thousand possible realizations and look at either the average or 95% possible worst outcome. In any case, the value for 1000 simulations is not the certain truth, it has statistical vagaries. In designing a reservoir, for instance, we might consider the vagaries of rainfall and water use over a 100-year simulation to see the impact of choices on location and dam height. In designing an airplane, for instance, one might model the response to 1000 possible arrangements of baggage distribution and weight, passenger seating, fuel weight, air properties (humidity, temperature, density), wind, flight path changes, and pilot response.

focus on it

If considering minimizing the 95% worst case (not the worst possible, but of all possible outcomes the one for which a worse outcome only happens 5% of the instances) and optimizing a design to maximize its value, the optimization statement might appear as

ignore it

the one for which a worse outcome only happens 5% of the instances) and optimizing a design to maximize its value, the optimization statement might appear as

Figure it

$$\max_{\{\mathbf{DV}\}} J = \mathrm{OF}_{0.05} \tag{2.12}$$

Experimental vs modeled OF

Models have various degrees of fidelity to nature. Some match (predict, forecast) very well, but in my experience, none is perfect. Improved sensors, better experimental techniques, and more data always seem to come along and reveal some inadequacy in the model. If there is strong belief that the model is an adequate representation of nature, then models are used as a surrogate for experimentation to guide optimization.

However, often the models are not available, or when a model exists, it does not match the data with desired fidelity to make results from its use credible. In such cases, we can use experimental results to provide the OF response to DVs.

Single and multiple optima

Figure 2.5a and b reveals objective functions of one variable with several optima.

Each figure illustrates three local minima. The left-hand figure might illustrate delivery costs for a shipping company as cost changes with business size. With small size, buy a small truck. The left portion of Figure 2.5a illustrates this. The more you ship, the less is the truck cost allocated per delivery. However, when shipping volume exceeds the small truck capacity, you need to either trade it for a larger truck or buy a second small truck. This extra capital increases cost per delivery but permits you toward larger volumes.

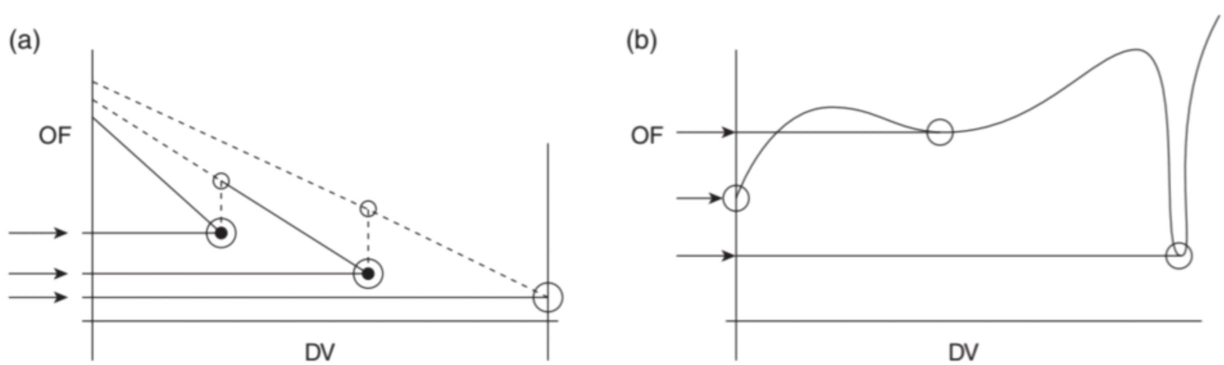
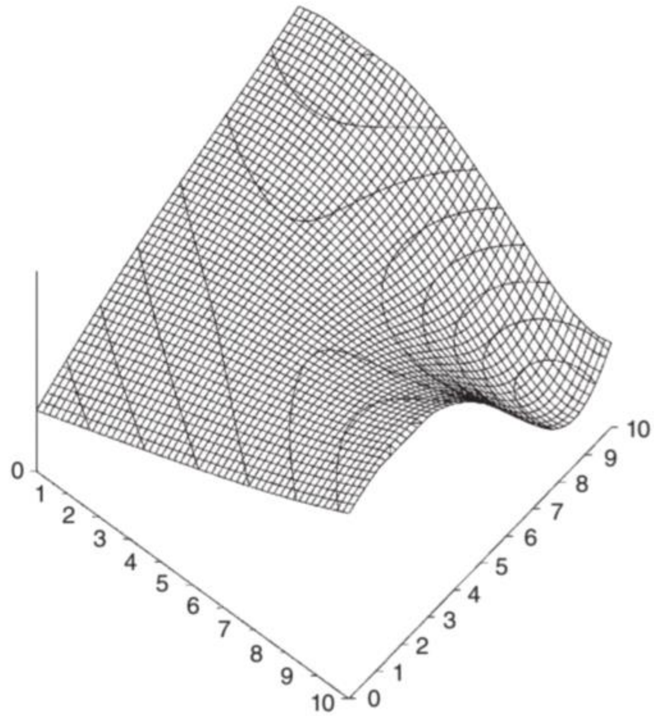


Figure 2.5 Illustrations of multiple optima: (a) discontinuous, (b) continuum. *Source:* Rhinehart 2016. Reproduced with permission of Wiley.

Saddle Points

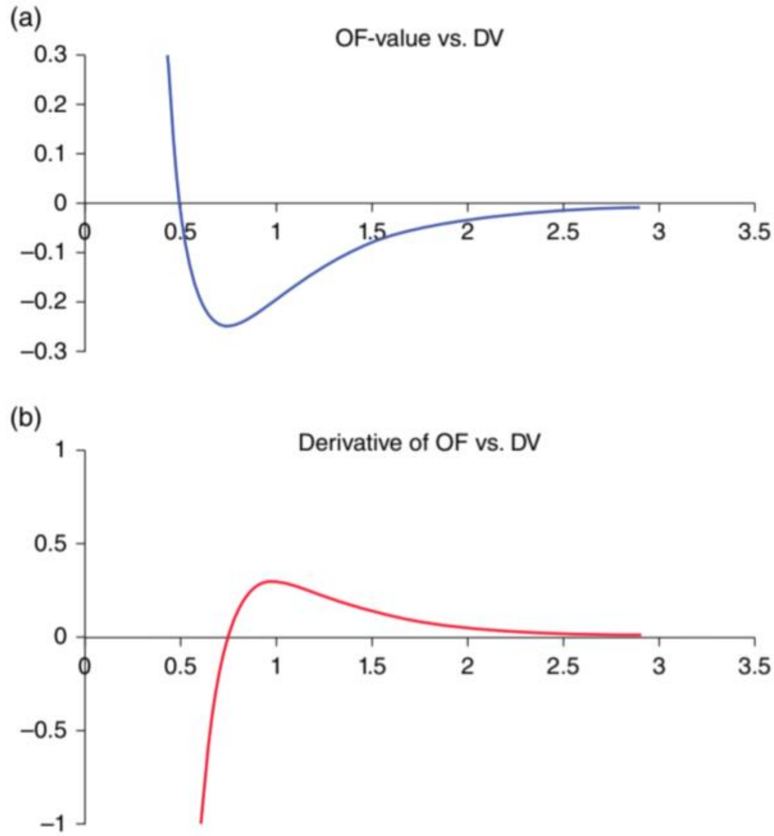
$$\nabla f = \mathbf{0}$$

Figure 2.6 Illustration of a saddle point.



Inflections

Any of the several Newton's method optimizers seek to root-find on the derivative. Start at a DV value of 1.5, and root finding on the derivative will send a search toward higher DV values, not back to the optimum. In general for Newton-type algorithms, if the trial solution is on the other side of an inflection point from the optimum, then the search will move further away from the optimum



Constraints and Penalty functions

Hard constraints are values or events that cannot be accepted. They may be on a DV value, which for physical implementation must be a positive, non-complex value. Or the constraint might be on a secondary consequence of a DV choice that happens nearly immediately (it violates an explosive limit) or in the future (such as the tank cannot overflow or the cash flow must remain positive). These are illustrated as the subject to (ST) portion of this generic optimization statement:

$$\begin{aligned} &\min_{\{\boldsymbol{x}\}} J = f(\boldsymbol{x}) \\ &\text{S.T.}: \quad a < x_3 < b \\ &\qquad\quad g(\boldsymbol{x}) > c \end{aligned} \tag{2.15}$$

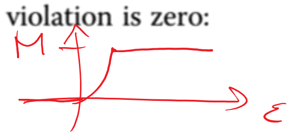
There are other possible solutions to handle hard constraints, but a common approach is to convert them into a penalty function and add the penalty for a constraint violation to the OF. First, consider the STs as desirables, but not absolute requirements. Examples of rules that are desirable are "Wash hands prior to eating." "Obey the speed limit." "Floss daily." "Keep the tank level between 25 and 85%." "No more than two fertilizer bags on any single pallet will leak."

A soft constraint acknowledges that there is a penalty for the rule violation, but that violating the rule is not a catastrophic disaster. Normally, the penalty value is proportional to the square of the constraint violation and is scaled by a coefficient to make it a reasonable balance with the other elements in the objective function.

λ

In the example earlier, a constraint is $g(\boldsymbol{x}) > c$. If the constraint is violated then the amount of violation is $\varepsilon = c - g(\boldsymbol{x})$; otherwise the magnitude of the violation is zero:

$$\varepsilon(x) = \begin{cases} \text{if } g(\boldsymbol{x}) \le c & \text{then } \varepsilon = c - g(\boldsymbol{x}) \\ \text{if } g(\boldsymbol{x}) > c & \text{then } \varepsilon = 0 \end{cases} \tag{2.16}$$

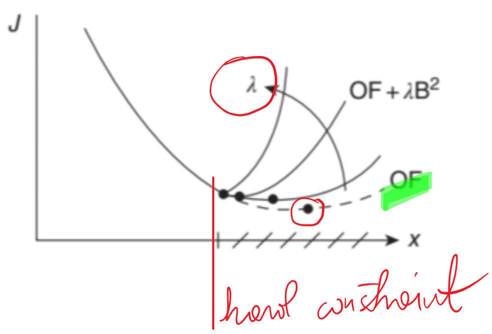


The penalty is added to the OF, where λ is the factor for equilibrating the importance of the constraint violation to the original OF:

$$\begin{aligned} &\min_{\{\boldsymbol{x}\}} J = f(\boldsymbol{x}) + \lambda\varepsilon^2 \\ &\text{S.T.}: \quad a < x_3 < b \end{aligned} \tag{2.17}$$

(play with λ)

Note that the $g(\boldsymbol{x}) > c$ constraint is no longer indicated in the S.T. list. It is now part of the OF calculation.



Underspecified OF

If the optimization application is underspecified, there will be multiple DV values with the same OF. This means that you have a choice in the DV value. Should you use the large one or the small one? If you can make the choice, then something makes one choice better than the other. Identify that rationale and include the new relation in the OF so that the optimizer finds it.

A common instance in regression modeling is to structure a model with redundant coefficients. For instance, it might seem that these three $\{a, b, c\}$ coefficients are independent in a model of the form $y = f(x) = (a + bx)/c$. However, dividing each term by coefficient c reveals that there are only two coefficients: $y = f(x) = (a/c) + (b/c)x = \alpha + \beta x$. If there is a unique $\{\alpha, \beta\}$ set, then there are an infinite number of $\{a, b, c\}$ sets that are equivalent, and values for these will be correlated. In the event of parameter correlation, remove a parameter and reformulate the model.

Some applications are effectively underspecified. Instead of the minimum being identical for a range of DV values, the OF appears as a gently sloped valley between steep walls.

Some ways of checking that the answer is sensible:

1) Have expectations about an equation, rule, theory, or procedure. If it is right, what do you expect to see, and not to see? Know how to test, assess, and evaluate the knowledge.
2) Don't just accept the procedure, recipe, formula, or rule. Understand the basis, assumptions, and context. Know the why about it. Be able to explicitly and quantitatively express the cause-and-effect mechanism.
3) Don't just accept computer output. Test it over a range of inputs and givens, and be confident that the output is consistent with your expectations.
4) Test your understanding by creating your own exercises. Explore alternate examples, values, assumptions, and the inverse relation. Be sure that trends are as expected.
5) What else does it apply to? What if you extrapolate it? Does the application make sense?
6) Compare a next-step better model to ideal calculations. Does it approach the ideal in the limit? Does it fit expected trends and homologous trends? When you adjust a parameter value, does the result behave as expected?
7) Compare to alternate methods such as an old-style handbook graphical method, software product A, software product B, or prior work.
8) Seek knowledge from product bulletins, handbooks, trade magazines, vendor's white papers, and the Internet. Textbooks are often an acceptable source of fundamentals and procedures. However, in spite of its importance to the academic community, avoid the scientific journal literature when you are seeking practical knowledge.
9) Think of analysis, synthesis, and evaluation in terms of Bloom's taxonomy of cognitive skill (see Section 3.3). Critically question the basis and assumptions in your knowledge. See how to apply it and how to integrate it in context. Consider how stakeholders will see the outcomes. Reveal how you know that it is correct by providing assessments of multiple, comprehensive, and competing criteria.
10) Be your own devil's advocate. Take the perspective of those who could claim to have an alternate opinion about the thing (maintenance, purchasing, labor, community, politicians, scientists, operator, opposition, etc.), and consider what aspects they could find and claim are undesired.
11) Learn by doing, not by studying. Don't just read or follow. Extrapolate on your own. Prepare by doing, not by intellectualizing. I tell my students the secret to success is this: "Do not study." To be sure they get the message, I write it on the board. Then, I pretend to be surprised at their not understanding, look at the statement, feign puzzlement, and agree that it doesn't make sense. Then pretend to have overlooked the comma when writing the secret to learning and add it, converting the sentence to say, "Do, not study."

12) Test and evaluate your own learning. Make your own quiz problems. If you can't, you don't understand it yet. Solve them. Implement your procedures in a spread sheet or structured code, and explore the validity of the outcomes.
13) Guide learning by what is needed to be able to do some task, not by what is interesting, or by the basic body of knowledge or by what everyone else knows.
14) Test it on simple ideal cases. Show that it gives the right answer. Then test it on more complicated cases, and show that it gives the answers that several experts agree on. Be sure to challenge it. Don't choose one or two cases that are simple to implement because that can provide a false affirmation. For example, my favorite pretend claim is that addition is the same as multiplication. I use $2 \times 2 = 2 + 2$ and $0 + 0 = 0 \times 0$ as examples to defend that claim. There are an infinite number of supporting examples. Further, the claim has a sophisticated name "Theory of Positional Invariance," which states that regardless of the observer's viewpoint the object retains its properties. There are many examples: Whether observed from the north or south poles, the moon has the same mass, although the moon appears upside down to one observer. Whether you look at a person from top or back or front, it is still that person with the same personality and color of eyes. Applying the principle, observe that except for their 45° rotation, the $\times$ and the $+$ symbols are the same, so the theory claims that $2 + 2 = 2 \times 2$. There you are! The theory is intuitively logical, has a sophisticated name, and is confirmed by data, which has an infinite number of cases. (Given any x-value, the constraint $x + y = xy$ can be solved for $y = x/(x-1)$. This even works for complex numbers.) So, the claim must be true. Challenge your knowledge and understanding with situations that might reveal the error.

15) Test it on real data, not just by calculations. It should provide a good enough match to the real data.
16) Accept your new knowledge on a tentative basis until you come to know better. You will have a tendency to want to accept your self-learned knowledge. You created it. It is your progeny. It may be difficult for you to see its inadequacy. Realize that even mankind's best theories have been proven false. We once thought that the magical substance called the ether transmitted electromagnetic waves. That led to Maxwell's equations, which were affirmed by data of that era. The caloric theory of heat led to the diffusion equations, again, affirmed by data. At one time, data and logic seemed to support the flat Earth concept as the center of the universe. Perfection in knowledge is elusive.
17) If it is a numerical procedure, see if smaller step sizes or convergence criteria change the answer. If right, calculated values should not change.
18) Extrapolate the application to large chronological time or large sizes and to initial values or very small dimensions. Test parameter and coefficient value extremes of 0, 1, or ∞. Take variable values to extreme conditions (dilute or concentrated, hot or cold, high or low flow rate, short or long tube, early or long time) and look at asymptotic limits of the model terms. Do they reduce to ideal conditions? Does it still make sense?
19) When the data is functionally transformed, is the trend as expected? For instance, when a power law model is log transformed, the data trend should be linear.
20) Consider and report the uncertainty in any value. There is uncertainty in the givens, in the coefficients, and in the models. How do these sources affect the output? How does uncertainty on the application impact a decision?
21) Seek challenges to your skill or knowledge that are sufficient, meaning that all relevant cases considered, testing is complete w.r.t. your context.
22) Also, seek challenges that are credible, meaning that it is tested with meaningful, known cases and returns "right" answers.

sidered, testing is complete w.r.t. your context.

22) Also, seek challenges that are credible, meaning that it is tested with meaningful, known cases and returns "right" answers.

Univariate Search Techniques

Tuesday, April 9, 2019 11:17 AM

The term "line search" is often used to mean that the optimization application has a single decision variable. I will use the term univariate, one variable, to indicate a single DV:

$$\min_{\{x\}} J = f(x) \tag{4.1}$$

one var.

The variable x could be a primary, fundamental variable, such as what might be graphed in an $f(x)$ w.r.t. x presentation.

However, the decision variable might be distance along a line that goes through space, and the objective might be to find the point along the line to maximize proximity (minimize distance) to a planet. Using a parametric notation for a line to calculate the (x, y, z) position as a function of parameter "s" from the origin:

$$\begin{aligned} x &= x_0 + as \\ y &= y_0 + bs \\ z &= z_0 + cs \end{aligned} \tag{4.2}$$

Note: At some position (x, y, z) distance along the line from the origin (x_0, y_0, z_0) is S, not s:

$$S = \sqrt{(x-x_0)^2 + (y-y_0)^2 + (z-z_0)^2} = \sqrt{(as)^2 + (bs)^2 + (cs)^2} = s\sqrt{a^2 + b^2 + c^2} \tag{4.3}$$

However, here the objective is not to minimize distance along the line, but the distance from the line to the point. Then, with the OF to minimize distance to a planet, a point object, at (x_p, y_p, z_p), the optimization statement becomes

$$\min_{\{s\}} J = \sqrt{\left(x(s)-x_p\right)^2 + \left(y(s)-y_p\right)^2 + \left(z(s)-z_p\right)^2} \tag{4.4}$$

Here the DV is not distance along an axis in Cartesian space, or even distance along the line, but a parameter representing distance along a line.

Although this is a 3-D application, there is a single DV, which is related to distance along a line, supporting the "line search" terminology.

Alternately, the path might not be a straight line; it might be a curve in space. Consider a path through mountains on a Cartesian map in (x, y) space, where elevation is a function of the x–y position, $z(x, y)$. The decision variable might be the distance along a path, and the objective might be to find it from the point of origin to the point where the maximum slope occurs:

$$\min_{\{S\}} J = \frac{dz(x,y)}{dS} \tag{4.5}$$

Here the rate of change of the z-elevation w.r.t. path distance, S, is the objective, and the elevation depends on the (x, y) position, which depends on distance along the path. Although this path through (x, y, z) space might be curved (not a straight line), and although the path may change in three dimensions, it is still a single DV search.

The term line search means a single DV search, whether the search path is along an axis, along a line, along a curved path in physical space, along a straight or curved path in N-dimensional DV space, or described by a related parameter. However, the single path must have unique model influence values for the single DV value.

Analytical Method of Optimization

$$\min_{\{x\}} J = f(x)$$

Determine the x^* that makes $df/dx|_{x^*} = 0$.

Example 1 Quadratic Function

$ax^2 + bx + c$

What is the value of x^*, if $y = a + bx + cx^2$? $(c = 5)$
Set the derivative to zero, $0 = b + 2cx^*$, and solve for $x^* = -b/2c$.

$y' = b + 2cx$

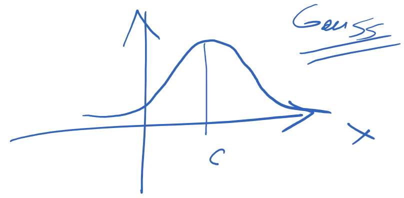
Example 2 Radial Basis Function

What is the value of x^*, if $y = e^{-((x-c)/s)^2}$?

Set the derivative to zero, $0 = \underbrace{-2((x^*-c)/s)e^{-((x^*-c)/s)^2}}_{y'}$, and solve for $x^* = c$.

y is maximized when: $z = [(x-c)/s]^2$ is minimized at $x = c$

$z(c) = 0$



Example 3 Saddle Point Function

What is the value of x^*, if $y = a + bx + cx^2 + dx^3$? $(b = 12, c = -6, d = 1)$

Set the derivative to zero, $0 = b + 2cx^* + 3dx^{*2}$, and solve for $x^* = 2$.

$y' = b + 2cx + 3dx^2 = 0$

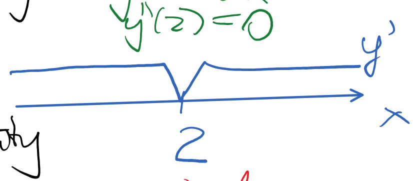
$y' = 3x^2 - 12x + 12 = 3(x^2 - 4x + 4) = 3(x-2)^2$

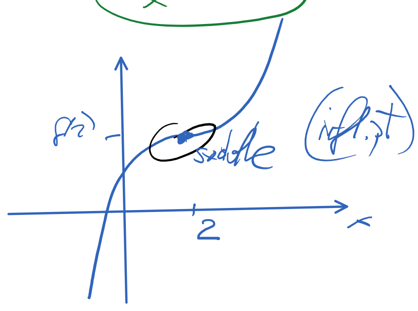
$x^* = 2$

y' controls monotonicity

$y'' = 6x - 12$

$y''(2) = 0$





y'' controls concavity

4 basic shapes:

sign changes

y'	$-$	$+$	$+$	$-$
y''	$+$	$+$	$-$	$-$
y				

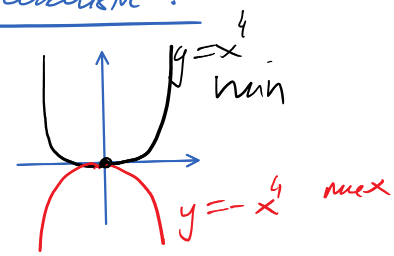
min: $y'' > 0$ max: $y'' < 0$

$y'' = 0$ at $x = x^*$, $\Rightarrow$ inconclusive!

Ex: $y = -x^4$

$y' = -4x^3 = 0 \Rightarrow x = 0$

$y'' = -12x^2$ $\quad y''(0) = 0$



Example 4 Nonlinear Function

What is the value of x^* if $y = a + b\sqrt{x} + c^x$?
Set the derivative to zero, and after rearrangement to isolate DV*, $\sqrt{x^*}c^{x^*} = -b/(2\ln(c))$. Solve for x^*?

A method to solve this nonlinear relation is to use some form of root finding such as Newton's method. Define

$$g(x^*) = \sqrt{x^*}c^{x^*} + \frac{b}{2\ln(c)} = 0 \tag{4.8}$$

$y' = \frac{1}{2}bx^{-\frac{1}{2}} + c^x \cdot \ln c =$

$= \frac{b}{2\sqrt{x}} + c^x \ln c = 0$

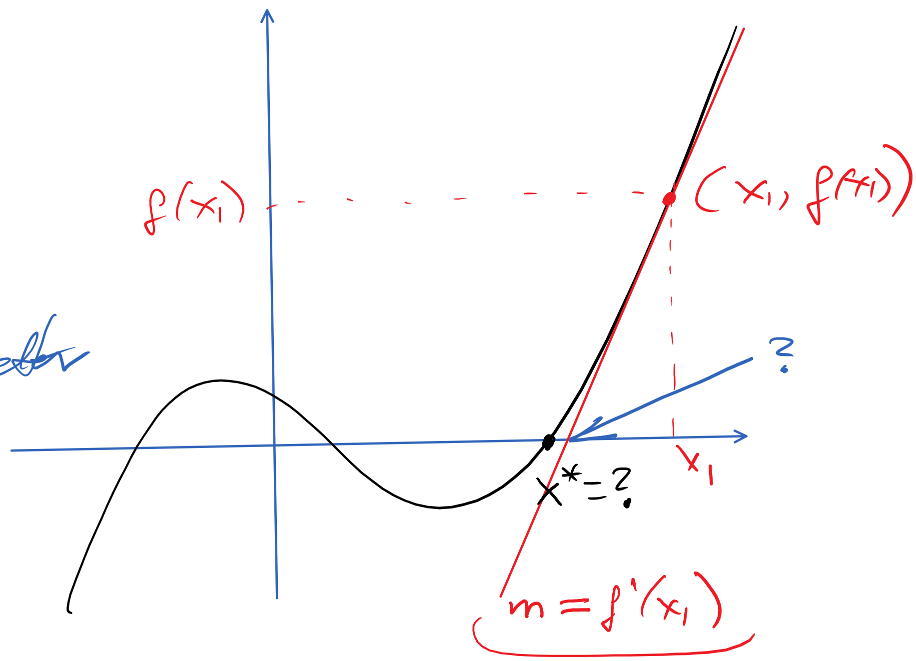
good luck!

Then iteratively estimate the value of x^* from

$$x^*_{k+1} = x^*_k - \frac{g(x^*_k)}{dg/dx|_{x^*_k}} \tag{4.9}$$

where k represents the iteration number and x^*_{k+1} indicates the sequential estimate of x^*.

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$



$y = mx + b \leftarrow$ o.k

$y = m(x - x_1) + y_1 \leftarrow$ better

$m = \frac{y - y_1}{x - x_1}$

In our case:

$y = f'(x_1)(x - x_1) + f(x_1)$

$0 = f'(x_1)(x - x_1) + f(x_1) \Rightarrow \boxed{x_2 = x_1 - \frac{f(x_1)}{f'(x_1)}}$

$y' = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$ $\to 0$, $\to 0$

$y = 2^x$ $\quad 2^{x+h} - 2^x$ $\quad 2^x \cdot 2^h - 2^x$

$$y = 2^x$$

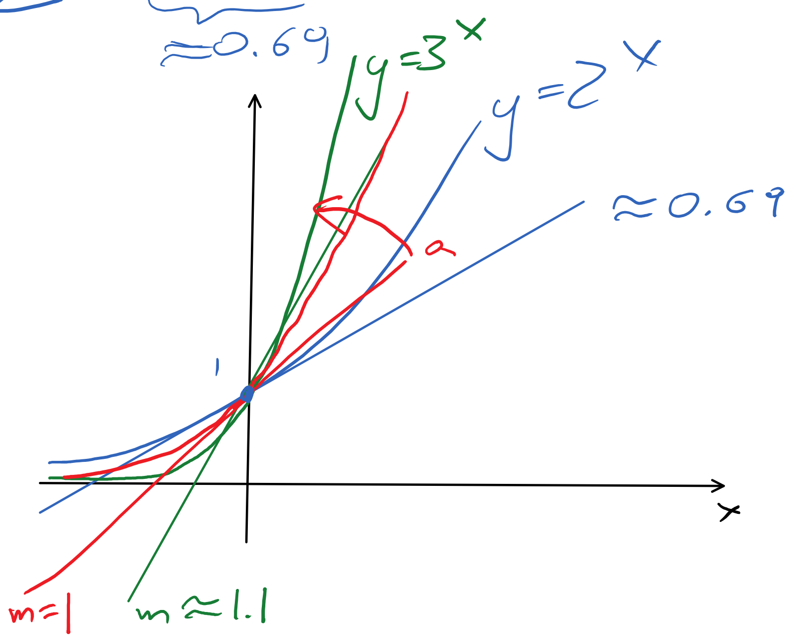
$$y' = \lim_{h\to 0} \frac{2^{x+h} - 2^x}{h} = \lim_{h\to 0} \frac{2^x \cdot 2^h - 2^x}{h} =$$

$$= 2^x \cdot \lim_{h\to 0} \frac{2^h - 1}{h} \approx 0.69$$

$2^h - 1 \to 0$, $h \to 0$

$\frac{0}{0}$ – indeterminate form

$$(2^x)' = 2^x \cdot \underbrace{\text{const}}_{\approx 0.69} = 2^x \cdot y'(0)$$



$y = a^x$

$$(a^x)' = a^x \cdot \cancel{\text{const}} \cdot \quad \ln a$$

$a = e \approx 2.71828\ldots$

$$(e^x)' = e^x$$

Newton's Methods - continued

Tuesday, April 9, 2019 11:16 AM

Another way of looking at the Newton's method:

There are several ways to develop a Newton's iterative method for optimization with a single DV. The methods provide a common functionality but reveal the possibility for flexible (human choice in) coefficient values. As a first derivation, choose a point x_0, hopefully near x^*, and use a Taylor series quadratic expansion for the f value at x^*. Eliminate high-order terms as having negligible impact when x-values are near x^*. (This step is grounded in the value of x_0 being sufficiently near x^*, so that the quadratic approximation is locally valid.)

$f(x) = f(x_0) + f'(x_0)(x-x_0) + \frac{1}{2}f''(x_0)\cdot(x-x_0)^2 + \frac{1}{6}f'''(x_0)\cdot(x-x_0)^3 + \dots$

$\dots + \frac{1}{n!}f^{(n)}(x_0)(x-x_0)^n + \dots$

$$f(x^*) \cong f(x_0) + \left.\frac{df}{dx}\right|_{x_0} (x^* - x_0) + \frac{1}{2}\left.\frac{d^2f}{dx^2}\right|_{x_0} (x^* - x_0)^2 \qquad (4.10)$$

$x^* - x_0$

At x^*, at the minimum, the slope $df/dx|_{x^*}$ is zero. (Note: This is an analytical concept and assumes that the function has a continuous value and first and second derivatives in the region of x^* and x_0.) Rearrange the truncated Taylor series model of the function to approximate the derivative at x_0 with a backward finite difference. (Subtract $f(x_0)$ from both sides of the equation, and then divide by $(x^* - x_0)$.)

$$0 = \left.\frac{df}{dx}\right|_{x^*} \cong \frac{f(x^*) - f(x_0)}{x^* - x_0} \cong \left.\frac{df}{dx}\right|_{x_0} + \frac{1}{2}\left.\frac{d^2f}{dx^2}\right|_{x_0} (x^* - x_0) \qquad (4.11)$$

Note the coefficient on the second derivative is ½, which generates 2 when Equation (4.11) is rearranged to solve for x^*:

$$x^* \cong x_0 - 2\frac{df/dx|_{x_0}}{d^2f/dx^2|_{x_0}} \qquad (4.12)$$

Recognize that x^* is just the next approximation. So, use the formula recursively:

$$\boxed{x_{k+1} = x_k - 2\frac{df/dx|_{x_k}}{d^2f/dx^2|_{x_k}}} \qquad (4.13)$$

similar but not same as before

2

This is one version of an analytical Newton's method. An alternate legitimate relation can be generated from a similar analysis. Repeat the Taylor series quadratic expansion about x_0, but have it predict $f(x)$, not $f(x^*)$:

$$f(x) \cong f(x_0) + \left.\frac{df}{dx}\right|_{x_0}(x - x_0) + \frac{1}{2}\left.\frac{d^2f}{dx^2}\right|_{x_0}(x - x_0)^2 \tag{4.14}$$

Take the derivative of $f(x)$

$$\frac{df}{dx} \cong \frac{d}{dx}\left(f(x_0) + \left.\frac{df}{dx}\right|_{x_0}(x - x_0) + \frac{1}{2}\left.\frac{d^2f}{dx^2}\right|_{x_0}(x - x_0)^2\right) = 0 + \left.\frac{df}{dx}\right|_{x_0} + \frac{1}{2}\cdot 2\left.\frac{d^2f}{dx^2}\right|_{x_0}(x - x_0) \tag{4.15}$$

Note that the coefficient of ½ is normalized by the coefficient of 2.
At the minimum, at $x = x^*$, the derivative is zero:

$$\left.\frac{df}{dx}\right|_{x^*} \cong 0 = \left.\frac{df}{dx}\right|_{x_0} + \left.\frac{d^2f}{dx^2}\right|_{x_0}(x^* - x_0) \tag{4.16}$$

which can be rearranged to solve for x^*

$$x^* \cong x_0 - \frac{df/dx|_{x_0}}{d^2f/dx^2|_{x_0}} \tag{4.17}$$

matches what we had before!

The recursion formula is

$$x_{k+1} = x_k - \frac{df/dx|_{x_k}}{d^2f/dx^2|_{x_k}} \tag{4.18}$$

This is the standard Newton's method for optimization. Again, it is a recursive, successive approximation method.

The analytical version of Newton's root-finding technique on the function $g(x)$ is

$$x^* \cong x_0 - \left.\frac{g}{dg/dx}\right|_{x_0} \tag{4.19}$$

With df/dx substituted for $g(x)$, it is identical to Equation (4.17):

$$x^* \cong x_0 - \left.\frac{g}{dg/dx}\right|_{x_0} = x_0 - \left.\frac{df/dx}{d^2f/dx^2}\right|_{x_0} \tag{4.20}$$

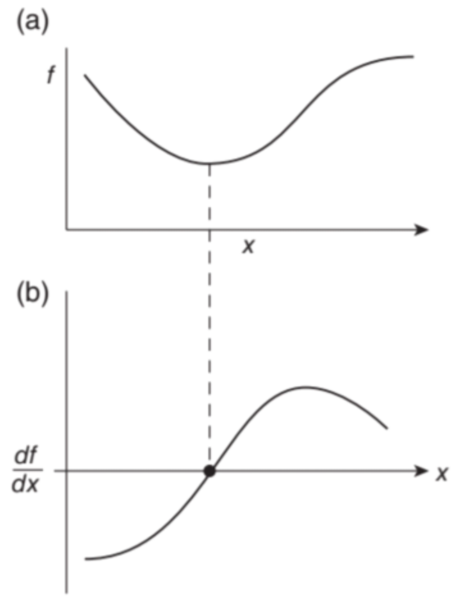


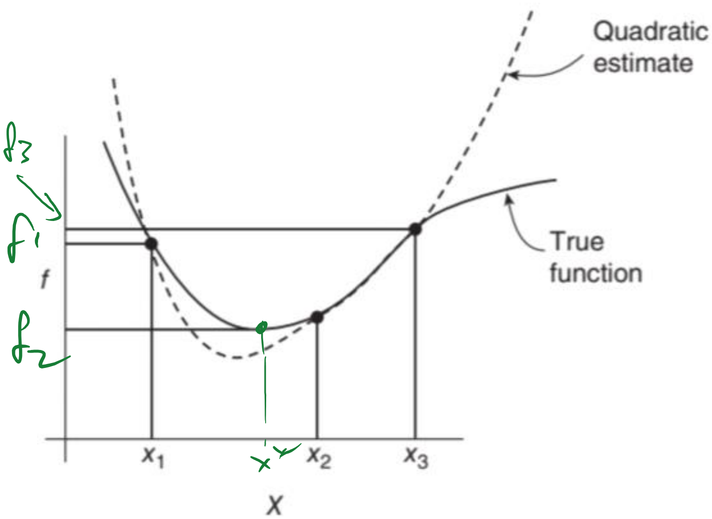
Figure 4.1 (a) A continuous function of a single variable and (b) its derivative.

Note: In the illustration in Figure 4.1, there is an inflection point on the function, which corresponds to the maximum value of the derivative. At an inflection point the second derivative changes sign, and the function transitions from convex to concave. This is the point of maximum rate of change and slope of the function. With an initial trial solution, within the x^* proximity of the inflection point, Newton's method of root finding on the derivative, the desire to get the derivative closer to zero, will move toward x^*. However, on the other side of the inflection point, the downward trend of the derivative indicates that moving further away leads to g-values that might approach zero.

Chapter 9 - more details

Successive Quadratic Method

Tuesday, April 9, 2019 2:45 PM



The method starts with three $(x, f(x))$ pairs and fits them with a quadratic estimate of the function. The model, illustrated as a dashed line, should perfectly fit the three data points, but the model is just an approximation to the function, $f(x) \cong \tilde{f}(x) = a + bx + cx^2$.

How does one obtain the model a, b, c values? The surrogate model is valid for each of the three data pairs, which yields three equations linear in the three model coefficients of unknown values (a, b, c):

$$\begin{cases} f_1 = a + bx_1 + cx_1^2 \\ f_2 = a + bx_2 + cx_2^2 \\ f_3 = a + bx_3 + cx_3^2 \end{cases} \qquad (4.28)$$

$f_1 = f(x_1)$

This linear algebra problem of determining coefficient values can be placed in matrix–vector form:

$$\mathbf{Mc} = \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ 1 & x_3 & x_3^2 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \\ f_3 \end{bmatrix} = \boldsymbol{f} \tag{4.29}$$

solve for a,b,c

After solving for the values of coefficients (a, b, c) (Gaussian elimination is one method), use the analytical approach to get x^* for the approximating function (the surrogate model):

$$\left.\frac{df}{dx}\right|_{x^*_{\text{function}}} \cong \left.\frac{d\tilde{f}}{dx}\right|_{x^*_{\text{estimate}}} = 0 = \frac{d}{dx}\left(a + bx + cx^2\right)\Big|_{x^*} = b + 2cx^* \tag{4.30}$$

which defines

$$\boxed{x^* \cong -\frac{b}{2c}} \tag{4.31}$$

Since the value of coefficient a is not needed in estimating x^*, the final step in the backward solution stage of the Gaussian elimination algorithm is not needed.

Since this gives the value of x^* for the surrogate model, which is not necessarily x^* for the function, the procedure needs to be repeated. The recursion relation is

$$x_{k+1} = -\frac{b_k}{2c_k} \tag{4.32}$$

One initializes the procedure with three trial solutions. The successive quadratic (SQ) procedure leads to a 4th x-value, then a 5th, and so on. Which three of the past N $(x, f(x))$ pairs should be used in defining the next surrogate model? One good rule is to use the new point and most recent past two points. It is claimed to be a good rule because it works and is simple to implement. Alternately, rules that are a bit more burdensome, but which may converge faster, are to use the new point and (i) the past two points with closest x-values to new or (ii) the past two points with lowest OF values.

Regardless of the selection rule, each iteration requires (i) a new function evaluation and (ii) solution of the linear equations.

The method is predicated on the local shape of the function being adequately approximated by a quadratic model. There are many cases in which this is not valid. Such as a discontinuous function, which might arise with discretized DVs, constraints, conditionals, etc.

The successive quadratic method provides sequential estimates. Hopefully, each new estimate is better than the prior estimate, and provides a better base point for the next estimate. They are successive, or iterative, or recursive. A recursion formula gives the next value from past values, and while the values may change, the formula is invariant.

How many iterations to apply? One method to claim convergence is to decide on a Δx value (ε, a precision) for which any x-value is close enough to the true (but unknowable) value of x^*. Then, when sequential estimates are closer than a 10th of that precision, claim that the last calculated value is close enough to x^*. If $|x_k - x_{k-1}| < 0.1\varepsilon$, then $x_k \cong x^*$.

Is there a bound on the number of iterations to converge? Usually there are not too many, but like Newton's methods some functions make it exceedingly difficult for SQ to converge, and it might diverge if initial trial solutions are beyond the inflection from x^*.

Did it find a minimum, maximum, or saddle point? Also, like Newton's methods, SQ seeks the point of zero slope in the derivative, which may be a minimum, maximum, or saddle point. To be sure, investigate the second derivative of the function w.r.t. DV at x^*.

4.4.1 Bisection Method

The bisection method does not rely on models or presumed surface behavior. As a result it is more generally applicable than are the analytical or second-order methods, but it is not necessarily as fast or efficient in jumping to the optimum.

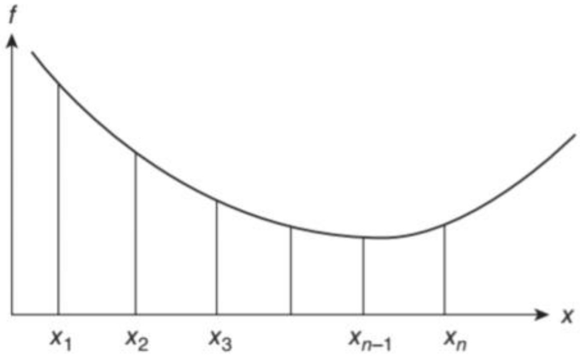


Figure 4.4 Bisection search stage 1—marching.

The bisection method is a two-stage procedure. First, bound the location of the optimum by a marching method, and then hone in on the optimum with interval halving. The marching method stage is illustrated in Figure 4.4.

First, start at one trial solution value that might be an extreme, perhaps the lower extreme, with a value of $x = x_1$. Evaluate $df/dx|_{x_1}$. If $df/dx|_{x_1}$ is negative, then increasing the value of x is a move toward the minimum. Increment to x_2, where $x_2 = x_1 + \Delta x$. Evaluate $df/dx|_{x_2}$. Continue until df/dx changes sign; if $df/dx|_{x_{n-1}} df/dx|_{x_n} < 0$, then the values of x_n and x_{n-1} bound the minimum.

Note: If an analytical equation for the derivative is not available, the derivative can be estimated numerically, $df/dx|_{x_k} = g_k \cong [f(x_k + \delta) - f(x_k)]/\delta$. However, precise derivative values are unnecessary. Just the sign is needed to direct the search. So, a coarse estimate of the derivative is all that is justified. Using $g_k \cong [f(x_k) - f(x_{k-1})]/\Delta x$ is fully adequate and avoids an additional function evaluation.

Note: Choose Δx, the marching increment, to have the opposite sign of the slope, or else this algorithm will find the maximum.

Note: If the value for Δx is too small, it may take an excessive number of marching iterations (and function evaluations hence computational time) to bound the optimum. But, if too large a value for Δx, the marching method might skip over an important feature. The right value for Δx typically has a large range but does require user understanding of the application.

Note: The marching method could be stopped by observing the function evaluations, not the derivative. This cuts the number of function evaluations in half. Observe $f(x_j)$. As long as $f(x_j) < f(x_{j-1})$, the search is approaching the minimum. But in this case, when $f(x_n) > f(x_{n-1})$, the minimum is not necessarily between x_n and x_{n-1}. It might be between x_{n-1} and x_{n-2}. So start the next stage with bounds of x_n and x_{n-2}.

Select a midpoint value in the range. $x_M = (x_L + x_R)/2$. If the slope at x_M has the same sign as slope at x_L, then reject the x_L to x_M region and set x_M as x_L. Else, do the complementary. With new x-values of the right-most and left-most bounds, recalculate the new midpoint $x_M = (x_L + x_R)/2$. Repeat until the x-interval is small enough to claim that the midpoint is close enough to the optimum, and claim that x^* is near the midpoint of the interval. If $|x_L - x_R| < \varepsilon$, then $x^* \cong (x_L + x_R)/2$.

Interval halving discards 50% of the DV range at each stage in the iterations. If there is an analytical equation for the derivative, then each iteration only requires one function evaluation, but I like to also observe the function value as well as the derivative, so I like two function evaluations per iteration. But, if there is not an analytical derivative, it takes two function evaluations at each x_M to determine the slope. Since two function calculations reduce the x-range by 50%, this implies a 25% range reduction per function evaluation.

If the marching method Δx defines the initial range for interval halving, and each iteration cuts the range in half, then after N interval halving iterations $|x_L - x_R| = \Delta x(1/2)^N$. If convergence is based on $|x_L - x_R| < \varepsilon$, then $\Delta x(1/2)^N < \varepsilon$, and solving for N, $N \cong \ln(\varepsilon/\Delta x)/\ln(1/2)$. The approximately equal sign indicates that N must be an integer, the rounded-up value of $\ln(\varepsilon/\Delta x)/\ln(1/2)$. This indicates that the bisection stage of the optimization is guaranteed to find the optimum in at least N iterations, but the number of trials in the marching stage cannot be predicted.

For the interval halving method, the user must specify (i) the initial x-value, (ii) the Δx value for the marching method (the sign on Δx can be determined by the computer by the derivative evaluation at the initial x-value), and (iii) the convergence ε (or the equivalent N for the interval halving stage).

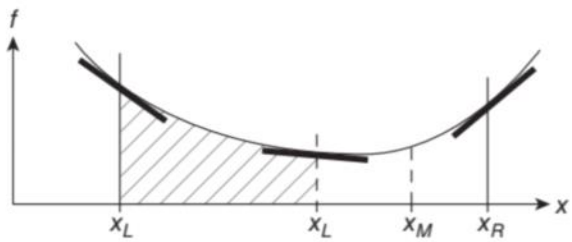


Figure 4.5 Illustration of the bisection method.

Golden Section Method ← check Wikipedia

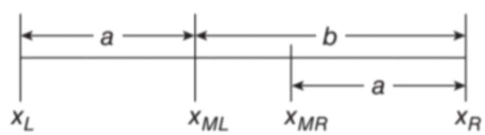


Figure 4.7 Golden Section apportionment.

The Golden Section method starts with four x-values, a left extreme, a right extreme, and two interior locations mid-left and mid-right. (The interior points do not equally divide the DV range; they are not at one-third and two-thirds values. The rule for interior point placement will be revealed in a bit.) Evaluate f at each point. Keep the exterior region with the lowest of all four f-values. In this case x_{MR} has the lowest OF value, so the right-hand extreme is kept. Discard the other exterior region, the one that is shaded. In this case point x_L is discarded, and three points x_{ML}, x_{MR}, and x_R are retained. Place the new fourth point, x_{new}, to preserve the original spacing pattern (the trick is to be revealed), relabel the two points that changed relative positions, and label the new point appropriately. In this case x_{ML} becomes x_L, x_{MR} becomes x_{ML}, and x_{New} becomes x_{MR}. Repeat, until the range remaining satisfies convergence. If $|x_L - x_R| < \varepsilon$, then stop and report $x^* = (x_L - x_R)/2$.

Note: The rule is to keep the extreme region that contains the lowest of the four OF values, not to discard the region that contains the worst OF value.

Note: The rule is to keep the extreme region that contains the lowest of the four OF values, not to discard the region that contains the worst OF value.

The trick with GS is to make each new point preserve the pattern within the spacing of the four points. Figure 4.7 illustrates the points and uses $a + b$ to indicate the total range and the relative positions of the points.

The desire is that when the left portion of the DV range is excluded and a new point is placed in the remaining DV range that the new a–b proportionality is preserved:

$$\frac{x_R - x_{ML}}{x_R - x_L} = \frac{x_R - x_{MR}}{x_R - x_{ML}} \tag{4.33}$$

which means $b/range = b/(a+b) = a/b$, which can be rearranged to the quadratic form, $a^2 + ab - b^2 = 0$. Then, applying the quadratic formula, the ratio of b to *range* or the ratio of a to b is

$$\frac{a}{b} = \frac{b}{a+b} = \frac{-1 \pm \sqrt{5}}{2} \tag{4.34}$$

Since only the "+" root is meaningful

$$\frac{a}{b} = 0.6180339888\ldots = \gamma \tag{4.35}$$

Gamma, γ, is called the golden ratio.

Once x_R and x_L values are chosen, determine initial x_{MR} and x_{MR} values from these equivalent relations:

$$x_{MR} = x_L + \gamma(x_R - x_L) = x_L(1-\gamma) + \gamma x_R \tag{4.36a}$$

$$x_{ML} = x_R - \gamma(x_R - x_L) = x_R(1-\gamma) + \gamma x_L \tag{4.36b}$$

Then, each time a region is discarded and new x_R and x_L values are updated, determine the new point:

$$x_{MR} = x_L + \gamma(x_R - x_L) \quad \text{if left is discarded, or} \tag{4.37a}$$

$$x_{ML} = x_R - \gamma(x_R - x_L) \quad \text{if right is discarded} \tag{4.37b}$$

This procedure preserves the pattern in the points with only one point added at each iteration.

Each new point requires one new function evaluation (the function *f*, not its derivative *df*/*dx*) and discards $(1-\gamma) = 0.381966\ldots$ fraction of area. This is a 38.2% range reduction per function evaluation. GS (at 38.2% DV range reduction per function evaluation) is more efficient in discarding space than interval halving (at 25% per function evaluation).

Further, there is only one iterative procedure in the GS approach, as contrasting the two (marching to bound the root, then interval halving) in the interval halving approach. GS is simpler to program.

Additionally, GS does not depend on derivative information, making it more robust to surface aberrations.

After *N* iterations GS reduces the DV region by a factor of γ^N. If the convergence threshold, ε, is based on the remaining DV range, then the relation between initial range, *R*, *N*, and convergence threshold ε is

$$N = 1 + \text{RoundUp}\left[\frac{\ln(\varepsilon/R)}{\ln(\gamma)}\right] \tag{4.38}$$

The 1 represents the first iteration to initialize the four DVs. The RoundUp function takes a non-integer value to the next higher integer. If such a function is not available, add 1 and take the INT value. Since there are four function evaluations on the first initialization, and one for each remaining iteration, the number of function evaluations for GS is

$$\text{NOFE} = 4 + \text{Int}\left[1 + \frac{\ln(\varepsilon/R)}{\ln(\gamma)}\right] \tag{4.39}$$

Golden Section is efficient, and the number of iterations or function evaluations to converge in this one-rule one-stage procedure is guaranteed.

But, GS is not perfect. If the initial x_R to x_L range does not include the optimum, then it will converge on either the initial x_R or the initial x_L value. Further, if the range includes a local optimum, not the global, it will find the local. Also, if the optimum is a pinhole in the extreme high function side, then it could be in the discarded region. Finally, GS does not scale to higher dimensions; it is for single DV searches only.

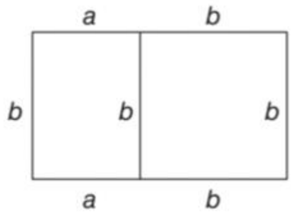


Figure 4.8 Squaring a rectangle.

4.4.2.1 Side Note on γ

The golden ratio is a value that frequently arises in geometry, math, art, and folklore. One classic origin is the question, "What is the aspect ratio of a rectangle such that when the large side is squared, the new rectangle preserves the original aspect ratio?" This is illustrated in Figure 4.8. The original rectangle has short and long side dimensions of a and b. When a square is placed on the long side b, the new rectangle dimensions are long side $(a+b)$ and short side b.

Specify that $a/b = b/(a+b)$ leads to $a/b = 0.6180339888\ldots$.

Continuing the side note, and mystique, about the golden ratio, consider a Fibonacci series: 0, 1, 1, 2, 3, 5, 8, 13, 21, … in which the recursion formula is

$$\text{Term}_N = \text{Term}_{N-1} + \text{Term}_{N-2} \tag{4.40}$$

In the limit of a large N $\text{Term}_{N-1}/\text{Term}_N = \gamma$.

I've often been entertained by much more of the mystique related to how the golden ratio appears throughout nature, math, art, folklore, and architecture.

A direct search only uses function values. It does not use derivatives. “FAIL” can be a function value. This aspect makes handling hard constraints easy.

The heuristic search only goes downhill (or uphill if seeking to maximize). It is not seeking a zero derivative, so it is not confounded by an inflection and is not drawn to a saddle or maximum.

Figure 4.9 shows a univariate search (sometimes termed a line search) example. The plot is J (OF or y) w.r.t. DV (or x). The initial trial solution is x_{base}, and the shaded area represents a constrained or forbidden range of x-values.

For the heuristic direct search, initialize by choosing x_{initial} (the starting value) and Δx (the initial search direction and size). Set $x_{\text{base}} = x_{\text{initial}}$, and evaluate $J(x_{\text{base}})$ and assign that value to J_{base}. Then

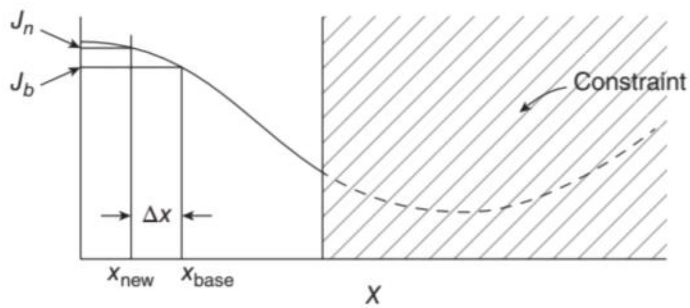


Figure 4.9 Illustration of the beginning of a heuristic direct search.

define a new trial solution as $x_{\text{new}} = x_{\text{base}} + \Delta x$. Evaluate $J(x_{\text{new}})$ and assign the value to J_{new}. If J_{new} or x_{new} or some other aspect of the calculation violates a constraint, set the value of J_{new} as “Fail.”

Now there are two possible outcomes: either J_{new} indicates that x_{new} is a better trial solution, or it is not better. Not better could mean that the OF is either worse or equivalent, or the TS is constraint violating. If the trial solution is not better, it could mean that the search was in the wrong direction (uphill) and that it led to a constraint violation, or it was too large a step in the right direction (which crossed over the minimum and started up the other side) or that it is on a flat spot and made no OF improvement. If so, contract and reverse the search direction. IF J_{new} = “Fail” or $J_{\text{new}} \geq J_{\text{base}}$, THEN $\Delta x = -\text{contract} \cdot \Delta x$. Otherwise, it was not worse; x_{new} was a step in the right direction and found a better spot. If so, assign the new conditions to the base case, $J_{\text{base}} = J_{\text{new}}$, $x_{\text{base}} = x_{\text{new}}$, and with confidence that you are moving in the right direction expand the search step size $\Delta x = \Delta x \cdot \text{expand}$.

Repeat until convergence, perhaps based on the step size factor IF $|\Delta x| < \varepsilon$ THEN stop and claim that $x^* = x_{\text{base}}$.

Figure 4.10a and b indicates the sequence for the first few iterations of unconstrained and constrained searches.

The expansion factor makes Δx larger as long as J keeps improving. Perhaps make "expand" = 1.2, 20% larger. (1.05 works and creates a more cautious expansion, 1.25 works and makes a more aggressive search, ….) Within this reasonable range, the value of the expansion factor is not critical to the search. Too small a value, and the search does not accelerate when going in the right direction. With too large an expansion value, Δx can become large, and once in the vicinity of the optimum, it takes

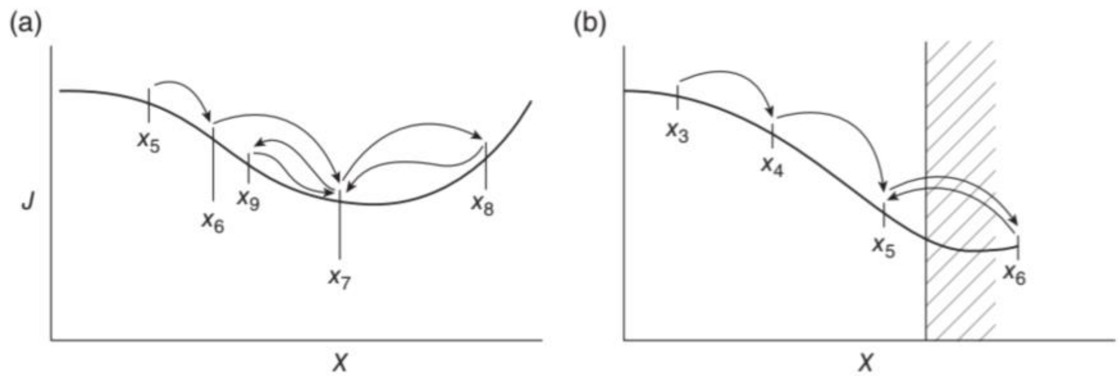


Figure 4.10 Illustration of a heuristic direct search: (a) unconstrained and (b) constrained.

Leapfrogging

Leapfrogging is a simple and effective multiplayer direct search. Start by choosing several (about 10) independent trial solutions, randomly scattered within the DV feasible range. Consider them as players on a surface. There is the worst player (the highest OF value if seeking a minimum) and the best (the lowest OF value). If there is a tie for the best (perhaps there are flat spots on the surface), just take the first in the player index list to represent best. In Leapfrogging, the worst leaps over the best (like the children's game of leapfrog), and the leaping player lands in a random spot on the other side of the best. For simplicity, the leap-into window is equal to the leap-from distance between worst and best. The placement for the leaping player, the new DV value, is calculated by

$$x_{w,\text{new}} = x_b + r(x_b - x_{w,\text{old}}) \tag{4.41}$$

where r is a random number (uniformly and independently distributed on the interval or 0–1, UID [0,1]) and subscripts b and w represent best and worst player x-values.

If the new player position is better than the prior best, IF $J(x_{w,\text{new}}) < J(x_b)$, and no constraints were encountered with the $x_{w,\text{new}}$-value, that player becomes the best, and the remaining worst of the other players leaps over it. If the leap-to position is not the best, the worst is found among the remaining team of players, and it leaps over the best. If the leap-to position violates a constraint, then it is the worst, and it leaps from that infeasible spot back over the best.

Figure 4.12 illustrates several situations. For simplicity only five players are shown. The initialization is indicated in Figure 4.12a, where the player DV values are indicated on the x-axis as $x_1, x_2, \ldots, x_5$, and

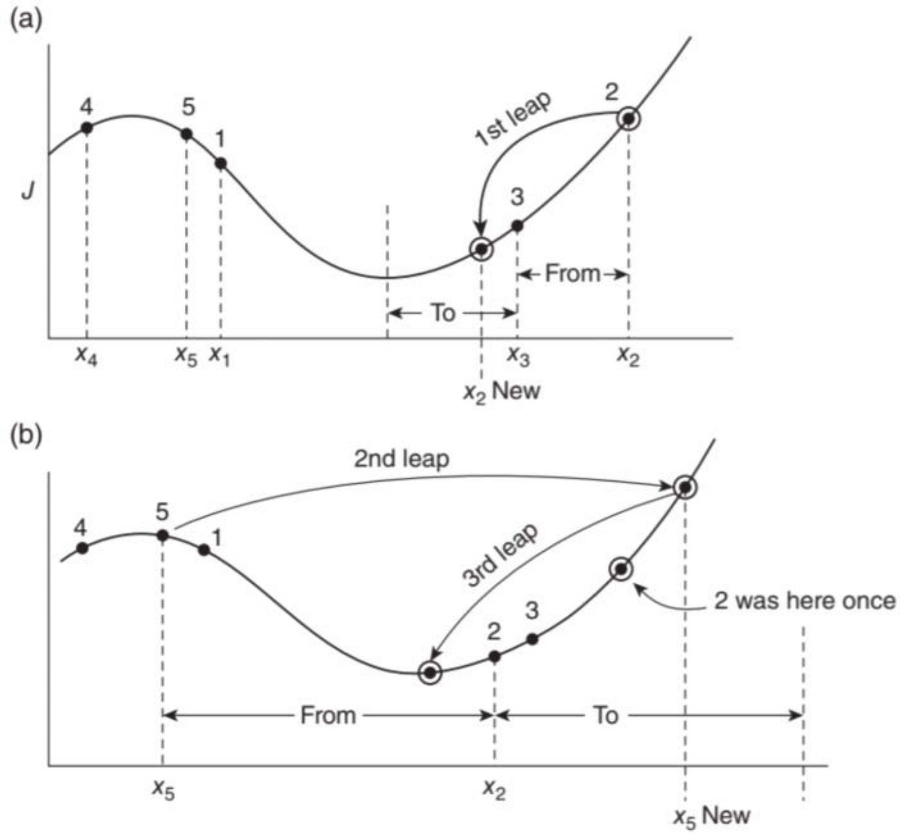


Figure 4.12 Illustration of leapfrogging: (a) first leap and (b) second and third leaps.

the player location on the surface is indicated by the player index 1,2, …, 5. Note that the index is not the order in either DV or OF value, it is the chronological order of initial randomized player placement in the game. Also note that Player 4 is on a portion of the function that slopes in the direction away from the global optimum toward a local optimum at the left-side of the feasible range.

After initialization, Player 3 is the best, because it has the lowest OF value, while Player 2 is the worst. So, 2 leaps over 3. Figure 4.12a also indicates the leap-from distance and the leap-into window. Player 2 moves to a random location in the leap-to DV location and has an OF value indicated by the open circle in the figure. This location will make Player 2 the best.

Figure 4.12b illustrates the second and third leaps. For the second leap, Player 2 is the best, while 5 is the worst, so 5 leaps over 2, and 5 happens to land on a DV value that has a worse yet OF value. So, Player 5 remains the worst, and on Leap 3 it again leaps over Player 2. As illustrated, 5 becomes the new best.

Note: Not every leap goes to a better spot, but in moving toward the local best, there is an overall tendency to find the optimum.

Note: Eventually Player 4 will be identified as the worst, and it will leap out of the vicinity of the local optimum into the region of the global optimum.

Note: On average each leap cuts the distance between the worst and best in half. But the illustration reveals that some leaps are into the far side of the leap-to window and some to the near side. The stochastic nature of the UID r overrides a deterministic path.

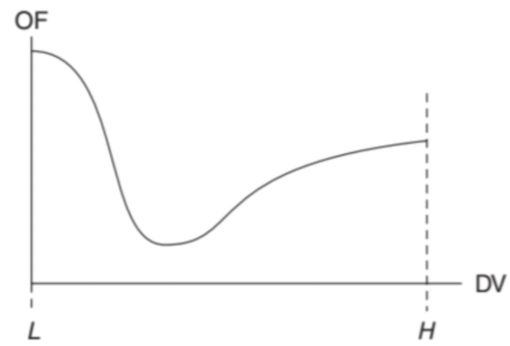
Note: Randomization of the initial player placement may seem less efficient than uniform placement to characterize the surface. In random placement some areas will be relatively unexplored and others more densely characterized. Further, simple placement on a grid would be computationally more efficient than calling the random number generator function. However, nature often devises patterns that would surprise a human planner or human convenience and evade discovery. I like using the random initialization.

Problems

Tuesday, April 9, 2019 4:46 PM

8 Figure 4.19 is a sketch of a function with low and high DV expectations marked. Sketch and explain the first two iterations of the Golden Section method to find the minimum.

Figure 4.19 Illustration for Exercises 8 and 9.



9 Use Figure 4.19 and create your own sketch of the derivative of the OF over the entire DV range, and starting near the high DV value, sketch and explain the first move of a Newton-type search for the optimum.

10 Write the computer code to perform univariate searches to find the minimum of $J = y = 4 + 8x - 2x^2$. Use x as the DV. Alternately, use $J = a + bx + cx^2$, and you pick values for the coefficients a, b, c. Use each of these methods: successive quadratic, Golden Section, secant (or Newton's), and heuristic direct.

a) Graph OF versus DV.
b) For each algorithm, reveal the path to the optimum, so that you can "see" the algorithm logic and know that your code is working correctly. You choose the starting point.
c) Show your computer code for each search method.
d) Choose the method you liked best. Explain and quantify your evaluation criteria.

11 Repeat Exercise 10 with $J = y = (x_1 - 8)^2 + (x_2 - 6)^2 + 15\,\text{Abs}((x_1 - 2)(x_2 - 4)) - 300\,\text{Exp}(-((x_1 - 9)^2 + (x_2 - 9)^2))$ search along the line defined by $x_2 = a + bx_1$, and use distance along the line as the DV. You pick values for "a" and "b" and for the line origin.

12 Repeat 10 with $J = \cos(2\pi x/10)$.

13 Repeat 10 with $J = [5 + (x - 5)^2][1 - \exp(-10(x - 9)^2)]$.

14 Repeat 10 with $J = 5 - x + 0.45x^2 - 0.08x^3 + 0.005x^4$.

15 Repeat 10 with $J = (1/x) - 5e^{-x}$

16 Repeat 10 with $J = \left[1 + 0.2*(x - 0.5)^2\right] * \left[1 - e^{-200(x-4)^2}\right]$.

17 Repeat 10 with $J = \left(2 - 2x^2 e^{-(x/2)}\right)^2$.

18 Repeat 10 with $J = (50 + \ln(x) - 5\sqrt{x} - x)^2$.

19 Minimize this function of D. You choose numerical values for coefficients ($a > 0$ and $b > 0$). $J = aD^{-5} + bD^{1.2}$.

20 Here are three functions. The DV is x, the OF is y.

a) $y = 5 - x + 0.45x^2 - 0.08x^3 + 0.005x^4$

b) $y = e^{-\frac{1}{2}\left(\frac{x-5}{2}\right)^2}$

c) $y = -2\left(\frac{1}{x-5}\right)^2$

$\boxed{\text{Hole 2}}$ only by a) analytical method. b) Newton's method.

For each, the DV range is 0–6. Function a) has a minimum and a saddle point. Function b) has two minima. The minimum for Function c) has a y-value of $-\infty$. Minimize each using four techniques: analytical (set the derivative equal to zero, and then root-find for the value of x^*), either Newton's using numerical first and second derivatives or ~~secant~~, successive quadratic, and golden section. This is a $3 * 4 = 12$-part exercise. However, depending on the starting location, the initial trial solution, you will get different results for the first two functions. So, for each of the 12 combinations, start at places which reveal the different results. Explain why the optimizer outcomes are what they are. A printout of the sequential trial solutions is not an explanation. Describe why each optimizer method created that DV sequence. I think that you can best explain this by hand illustrations on a printout of the graph of y w.r.t. x for each function.

section