Techniques for Simple Problems

CE 377K

Stephen D. Boyles

Spring 2015

These notes describe ways to solve two types of simple optimization problems: problems with no constraints, and problems with only a single variable. Most realistic problems of any real interest involve constraints and more than one variable. However, solution techniques for those types of problems are necessarily more complicated. These simple settings let us develop building blocks of solution methods which can be modified for more complicated problems.

1 Unconstrained Optimization

1.1 One-dimensional unconstrained optimization

In this section, we are considering an optimization problem with no constraints whatsoever, so that each decision variable can take any real number. By the convention described in the previous notes, there is no harm in assuming that this is a minimization problem. In the simplest case, there is only one decision variable, so the optimization problem is simply

$$\min_x f(x) \tag{1}$$

where x can take any real number. (You have probably seen this type of problem before in calculus. We are going through it again both to review and to preview what will be done for trickier problems.)

Assume that f is a differentiable function, that is, the derivative f' exists everywhere. (This implies that f is continuous.) We say that the point x is *stationary point* of f if $f'(x) = 0$. As in the previous notes, x^* is a global optimum if $f(x^*) \leq f(x)$ for all x. If $f(x^*) \leq f(x)$ only when x is close to x^*, then x^* is a *local optimum.* More specifically, x^* is a local optimum of f if there is some open interval (ℓ, h) containing x^* such that $f(x^*) \leq f(x)$ for all $x \in (\ell, h)$.

Theorem 1. *Assume that f is differentiable. Then every global optimum of f is a local optimum, and every local optimum is a stationary point.*

Proof. Assume x^* to be a global optimum, so $f(x^*) \leq f(x)$ for all real x. Then certainly $f(x^*) \leq f(x)$ for all x in any interval, so x^* is a local optimum as well.

Now assume that x^* is any local optimum, and by contradiction assume that x^* is not a stationary point, so $f'(x^*) \neq 0$. If $f'(x^*) > 0$, then f is strictly increasing at x^*, so for points $x' < x^*$ sufficiently close to x^* we have $f(x') < f(x^*)$, contradicting the assumption that x^* is a local optimum. If $f'(x^*) < 0$, then f is strictly decreasing at x^*, so for points $x' > x^*$ sufficiently close to x^* we have $f(x') < f(x^*)$, again a contradiction. □

The theorem does not apply in the reverse direction: it is not hard to come up with examples where there are local optima which are not global optima, and stationary points which are not local optima. As a result, the intuitive-sounding procedure of finding all the stationary points and choosing the one with the least f value is *not* guaranteed to work. If $f(x) = -x^2$, then $f'(x) = -2x$ and the only stationary point is $x = 0$. However, this point actually corresponds to a global *maximum*, not a global minimum. Or, if $f(x) = x^3$, then $f'(x) = x^2$ and the only stationary point is $x = 0$, however this point is neither a global maximum nor a global minimum. For this result to be helpful we need another condition.

We say that the function f is *coercive* if $\lim_{|x|\to\infty} f(x) = +\infty$, that is, as x grows large in either the positive or negative direction, $f(x)$ grows large as well. In this case we have the following result:

Theorem 2. *If f is differentiable and coercive. Then the global minimum of f is the stationary point of f with the least objective function value.*

Proof. If f is differentiable and coercive, then there is at least one global minimum. After all, if f does not have a global minimum, then it must take arbitrary large negative values. Since f is coercive this cannot happen in the limit as x becomes very large or very small; and since f is differentiable it is continuous and exists everywhere, so it cannot have any vertical asymptotes. By Theorem 1, each such global minimum must be a stationary point, and by definition must be a stationary point with the least objective function value. □

Example 1. *Find the value of x which minimizes $f(x) = x^2 - 3x + 5$.*

Solution. $f'(x) = 2x - 3$, which vanishes if $x = 3/2$, so this is a stationary point. Furthermore f is coercive since it $f(x)$ tends to $+\infty$ as x grows very positive or negative. Therefore $x = 3/2$ is the global minimum of $f(x)$. ■

1.2 Multi-dimensional unconstrained optimization

The following results transfer in a similar way when there is more than one decision variable. If there are n decision variables with no constraints, then let the vector $\mathbf{x}$ be an n-dimensional vector, and $f(\mathbf{x})$ a function of n variables. For a function of more than one variable, the analogue of the derivative is the *gradient* ∇f, which is the vector of all first partial derivatives:

$$\nabla f(\mathbf{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(\mathbf{x}) & \frac{\partial f}{\partial x_2}(\mathbf{x}) & \cdots & \frac{\partial f}{\partial x_n}(\mathbf{x}) \end{bmatrix} \tag{2}$$

A *stationary point* of a function of multiple variables is a point where the gradient is the zero vector: $f(\mathbf{x}) = \mathbf{0}$, that is, a point where all of the partial derivatives simultaneously vanish. Furthermore, the same definition of a coercive function works in higher dimensions if the absolute value in the limit $|\mathbf{x}| \to \infty$ is taken to mean the length of the vector, $|\mathbf{x}| = \sqrt{x_1^2 + x_2^2 + \cdot + x_n^2}$.

All of the results for single-dimension functions given in Section 1.1 generalize to multi-dimensional functions with the new definition of a stationary point: global optima are all local optima, and all local optima are stationary points, and if the objective function is coercive then its global optima are the stationary points with the least objective function values.

2 Line Search Techniques

The previous section explored the case when there were no constraints, even though there could be multiple decision variables. This section explores the opposite case, when there is a single decision variable, but potentially constraints. In one-dimensional optimization, it is very common that the only constraint is that the decision variable x must lie in some closed interval $[a, b]$. (Can you see why open intervals are problematic?)

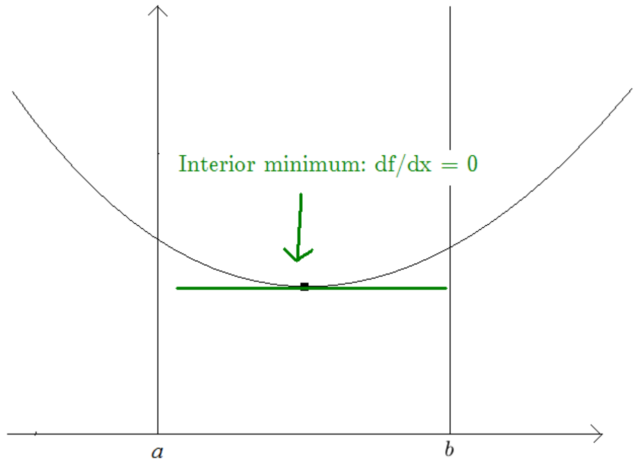
In this case, we can use a *line search technique* to find the optimal solution. Such techniques are *iterative*, which means that a particular step is repeated over and over. In general, line search techniques are not guaranteed to find the optimal solution, but will converge to it in the limit as more and more iterations are performed. For line search to work, a function needs to be fairly well-behaved. A one-dimensional function defined on an interval $[a, b]$ is *unimodal* if it has only one local minimum in $[a, b]$. Since this section adds the constraint that $x \in [a, b]$, the definition of a local minimum shold be changed slightly as well: x is a local minimum of f on $[a, b]$ if there is some open interval (ℓ, h) containing x^* such that $f(x^*) \leq f(x)$ for all $x \in (\ell, h) \cap [a, b]$. (In other words, points outside of $[a, b]$ don't count when checking for a local optimum.) Throughout this section, assume that the function f is unimodal. If this is not the case, the techniques can still be used, but there is no guarantee that they will find the global optimum. Since f is unimodal, it has only one local optimum on $[a, b]$, so this must be the global optimum as well. In this section, we also assume that f is differentiable. There are other line search techniques which can be used when this is not the case.

As shown in Figure 2, this one optimum (call it x^*) can occur in one of three ways. In the first case, if x^* lies in the interior of the feasible region, that is, $x^* \in (a, b)$ so there are feasible points in both directions around x^*, then the derivative of f must be zero at x^*, because the constraints are essentially irrelevant at this point and the unconstrained optimality condition applies. However, if x^* lies at one of the boundary points a or b, this analysis must be modified. If $x^* = a$, then for x^* to be an optimum solution, f must be increasing at this point (or at least nondecreasing), that is, $f'(x^*) \geq 0$. Otherwise, f could be decreased by choosing an x value slightly larger than a. Likewise, if $x^* = b$, then for x^* to be an optimum solution, f must be nonincreasing at this point, so $f'(x^*) \leq 0$.

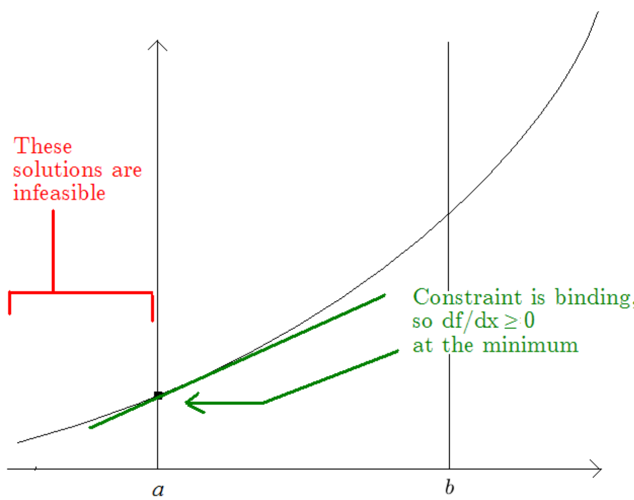
This section discusses two line search techniques that can be used. The first is an adaptation of Newton's method, which tends to be very fast but requires that the function f be twice differentiable and convex (defined below). The second is the bisection method, which is somewhat slower but only requires f to have one derivative, not two. In each of these cases, we need to show that the method converges to one of the three cases in the previous paragraph: either a point where the derivative is zero; or the lower bound a if the derivative there is nonnegative; or the upper bound b if the derivative there is nonpositive.

2.1 Newton's method

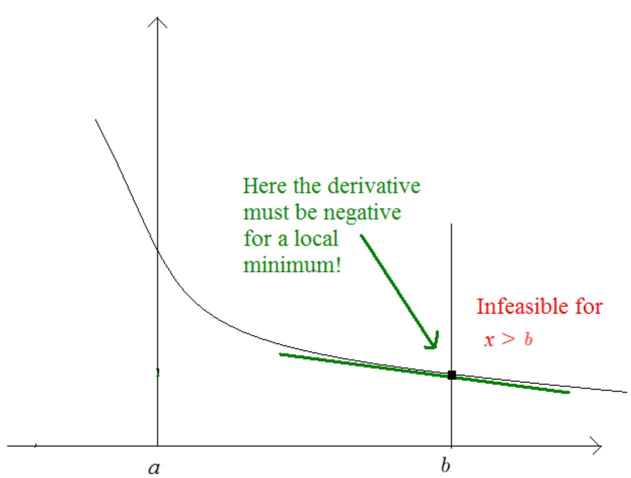
Assume that f is twice differentiable. If the optimal solution x^* lies in the interior of the feasible set $[a, b]$, then x^* must be a stationary point, so $f'(x^*) = 0$. Newton's method is a simple procedure for finding the zeroes of a function, which often converges quite quickly. Recall that in general, to



(a) Convex function where neither constraint is binding.



(b) Convex function where the constraint is binding at the lower bound.



(c) Convex function where the constraint is binding at the upper bound.

find a zero of a function $g(x)$, Newton's method involves repeating the following steps:

1. Make an initial guess x
2. Calculate $g(x)$ and its derivative $g'(x)$
3. If $g(x) \approx 0$, stop.
4. Update $x \leftarrow x - g(x)/g'(x)$ and return to step 2.

The last step is the key step. The idea is to approximate the (possibly nonlinear) function $g(x)$ with its linear approximation, since the zero of a linear function can be found in closed form. Newton's method does not in general work; however it can be shown to converge for convex functions. If f is twice differentiable, it is *convex* if its second derivative $f''(x)$ is nonnegative on the interval $[a, b]$. (There are other, more general definitions of convexity which we will see later in this course.)

Furthermore, we need to make sure that the search stays within the feasible region $[a, b]$. The easiest way to do this is to "truncate" the Newton search by projecting the updated x value onto the set $[a, b]$. That is, if the x value is less than the lower bound a, change it to a; if it is greater than the upper bound b set it to b; and otherwise leave it as is. So, the modified Newton's method for minimizing a function of a single variable on the interval $[a, b]$ is as follows:

1. Make an initial guess x
2. Calculate $f'(x)$ and $f''(x)$
3. If $f'(x) \approx 0$, stop.
4. Update $x \leftarrow x - f'(x)/f''(x)$
5. If $x < a$, set $x \leftarrow a$; if $x > b$ set $x \leftarrow b$.
6. Return to step 2.

In step 4, if $f''(x) = 0$, then you can treat the division by zero as resulting in either $+\infty$ or $-\infty$, depending on the sign of $f'(x)$. After performing the projection in the next step, this essentially means that if $f'(x) > 0$ but $f''(x) = 0$ (f is increasing), move as far as possible to the left so $x \leftarrow a$; if $f'(x) < 0$ but $f''(x) = 0$ (f is decreasing), move as far as possible to the right so $x \rightarrow b$.

Example 2. *Use Newton's method to find the minimum of* $f(x) = (x - 1)^4 + e^x$ *on the interval* $x \in [0, 3]$*, performing seven iterations.*

Solution. You may find it useful to follow along in Table 1. The first and second derivatives of $f(x)$ are $f'(x) = 4(x-1)^3 + e^x$ and $f''(x) = 12(x-1)^2 + e^x$. As an initial guess, choose the midpoint: $x = 1.5$. At this point, $f'(x) = 4.982$ and $f''(x) = 7.482$, so the update is $x \leftarrow 1.5 - 4.982/7.482$, so the new value is $x = 0.834$. At this point, $f'(x) = 2.285$ and $f''(x) = 2.633$, so the formula gives $x \leftarrow 0.834 - 2.285/2.633 = -0.034$. Since this value is less than the lower limit 0, "project" x back to the feasible region and set $x = 0$. Continuing on for seven iterations, at $x = 0.30307$ the value of

Table 1: Demonstration of Newton's method with $f(x) = (x-1)^4 + e^x$, $x \in [0, 3]$.

Iteration	x	$f'(x)$	$f''(x)$	$x - f'(x)/f''(x)$
1	1.5	4.982	7.482	0.834
2	0.834	2.285	2.633	-0.034
3	0	-3.000	13.000	0.231
4	0.231	-0.561	8.360	0.298
5	0.298	-0.0375	7.263	0.30304
6	0.30304	-2.058×10^{-4}	7.1829	0.3030725347
7	0.30307	-6.307×10^{-9}	7.1825	0.3030725355

the first derivative is $f'(x) = -6.307 \times 10^{-9}$ which is very, very close to zero. Therefore this value of x is very close to the minimum value on $[0, 3]$.

How can we show that Newton's method always works? A full proof is difficult at this point in the class, but we can give some intuition. The "direction" that Newton's method moves in is determined by the update step $x \leftarrow x - f'(x)/f''(x)$. Assume first that the optimum solution is an interior point, so $f'(x^*) = 0$. Then Newton's method works as it always does, moving towards a zero of $f'(x)$. If $f'(x) > 0$ and $f''(x) \geq 0$, then f is increasing and convex at that point, so Newton's method will decrease x. If $f'(x) < 0$ and $f''(x) \geq 0$, then f is decreasing and convex at that point, so Newton's method will increase x. In either case, we are moving toward the point where $f'(x) = 0$. In the second case, the optmium solution is at the lower bound and $f'(a) \geq 0$. If f is convex, then this means that $f'(x) \geq 0$ for *all* points $x \in [a, b]$ as well as $f'(x) \geq 0$, so Newton's method will always decrease x at each step until the lower bound is reached. Similarly, if the optimum solution is at the upper bound and $f'(b) \leq 0$, then since f is convex $f'(x) \leq 0$ for all $x \in [a, b]$ and Newton's method will increase x until the upper bound is reached.

2.2 Bisection Method

The bisection is another line search method. It is not as fast as Newton's method, but it has several other advantages: (a) it does not require f to have two derivatives, just one; (b) it will always work when f is unimodal, not just when f is convex; and (c) it is easier to prove that it works.

The bisection method works by constantly narrowing down the region where the optimal solution lies. It is an *iterative* algorithm, in that it repeats the same steps over and over until a *termination criterion* is reached. After the k-th iteration, the bisection method will tell you that the optimum solution lies in the interval $[a_k, b_k]$, with this interval shrinking over time (that is, $b_k - a_k < b_{k-1} - a_{k-1}$). A natural termination criterion is to stop when the interval is sufficiently small, that is, when $b_k - a_k < \epsilon$, where ϵ is the precision you want for the final solution. (For example, phase lengths for many signal controllers have to be in multiples of half a second, so there is no point in determining the optimal phase length to any greater precision than this.)

Here's how the algorithm works.

Step 0: Initialize. Set the iteration counter $k = 1$, $a_1 = a$, $b_1 = b$.

Step 1: Evaluate midoint. Calculate the derivative of f at the middle of this interval, $d_k =$

Table 2: Demonstration of the bisection algorithm with $f(x) = (x-1)^4 + e^x$, $x \in [0,3]$.

k	a_k	b_k	$(a_k+b_k)/2$	d_k
1	0	3	1.5	$4.98 > 0$
2	0	1.5	0.75	$2.05 > 0$
3	0	0.75	0.375	$0.478 > 0$
4	0	0.375	0.1875	$-0.939 < 0$
5	0.1875	0.375	0.28125	$-0.160 < 0$
6	0.28125	0.375	0.328125	$0.175 > 0$
7	0.28125	0.328125	0.3046875	$0.0116 > 0$

$f'((a_k + b_k)/2)$

Step 2: Bisect. If $d_k > 0$, set $a_{k+1} = a_k$, $b_{k+1} = (a_k + b_k)/2$. Otherwise, set $a_{k+1} = (a_k + b_k)/2$, $b_{k+1} = b$.

Step 3: Iterate. Increase the counter k by 1 and check the termination criterion. If $b_k - a_k < \epsilon$, then terminate; otherwise, return to step 1.

Example 3. *Use the bisection algorithm to find the minimum of $f(x) = (x-1)^4 + e^x$ on the interval $x \in [0,3]$ for seven iterations.*

Solution. You may find it useful to follow along in Table 2. We start off with $k = 1$, $a_1 = 0$, and $b_1 = 3$. We calculate the derivative at the midpoint: $f'(x) = 4(x-1)^3 + e^x$, so $f'(1.5) = 4.98$, which is positive. Since f' is positive at $x = 1.5$, the minimum must occur to the *left* of this point, that is, somewhere in the interval $[0, 1.5]$. We set a_2 and b_2 equal to these new values, and repeat. The new midpoint is 0.75, and $f'(0.75) = 2.05$ is positive, so the minimum must occur to the *left* of this point. Thus, the new interval is $[0, 0.75]$, and we repeat as shown in Table 2. At the end of the seventh iteration, our best guess of the optimum as the midpoint of the interval at this point: $x^* \approx 0.3047$. The true minimum point occurs at $x^* = 0.3031$; if we had chosen a smaller tolerance ϵ, the algorithm would have narrowed the interval further, with both ends converging towards this value. Note that the value of the derivative at this point is 0.01158, which is zero at the true minimum. By comparison, after the same number of iterations Newton's method returned a point with derivative -6.3×10^{-9}, which is much closer to zero. This is typical performance: Newton's method converges faster in terms of number of iterations, but each iteration requires more calculation.

How do we know the bisection algorithm always works? We can prove it using the optimality conditions derived in the previous section. Remember that x^* is an optimal solution to $\min f(x)$, $x \in [a, b]$ if one of these three conditions is true:

- $f'(x^*) = 0$ and $x \in [a, b]$
- $f'(x^*) \geq 0$ and $x = a$
- $f'(x^*) \leq 0$ and $x = b$

Now, it is clear that the two endpoints a_k and b_k converge towards a common point as $k \to \infty$. Mathematically, this happens because $b_k - a_k = (b_{k-1} - a_{k-1})/2$, which implies $b_k - a_k \to 0$, which implies $a_k \to \hat{x}$ and $b_k \to \hat{x}$ for some point $\hat{x} \in [a, b]$. We need to show that $\hat{x}$ satisfies one of the three conditions above. This is formalized as the following theorem.

Theorem 3. *Let a_k and b_k be determined by the bisection method, and let $\hat{x}$ be the point a_k and b_k converge to. Then $\hat{x}$ is a minimum for the optimization problem* $\min f(x)$ *subject to $x \in [a, b]$.*

Proof. Recall our assumption that f is convex and continuously differentiable. This implies that $f'(a) \leq f'(b)$, so one of three cases must be true.

Case I. $f'(a) \geq 0$ and $f'(b) \geq 0$, that is, f is increasing at both endpoints of $[a, b]$. This is the situation in Figure 1(b). Because f is convex, $f'((a_k + b_k)/2) \geq 0$ no matter what a and b are. Then, according to Step 2 of the bisection method, $a_k = a_{k-1}$ at every iteration k. Thus $a_k = a$ for all k, so certainly $a_k \to a$, implying $\hat{x} = a$. Since $f(a) \geq 0$, optimality condition 2 is satisfied and $\hat{x}$ is optimal.

Case II. $f'(a) < 0$ and $f'(b) < 0$, that is, f is decreasing at both endpoints of $[a, b]$. This is the situation in Figure 1(c). Because f is convex, $f'((a_k + b_k)/2) < 0$ no matter what a and b are. Then, according to Step 2 of the bisection method, $b_k = b_{k-1}$ at every iteration k. Thus $b_k = a$ for all k, so certainly $b_k \to b$, implying $\hat{x} = b$. Since $f(b) \leq 0$, optimality condition 3 is satisfied and $\hat{x}$ is optimal.

Case III. $f'(a) < 0$ and $f'(b) \geq 0$, that is, f is decreasing at a but increasing at b. This is the situation in Figure 1(a). After initialization of the algorithm, we have $f'(a_0) < 0 \leq f'(b_0)$. Consider each iteration k, starting with $k = 0$, and assume that $f'(a_k) < 0 \leq f'(b_k)$. If $f'((a_k + b_k)/2) < 0$, then $a_{k+1} = (a_k + b_k)/2$, $b_{k+1} = b_k$, so we have $f'(a_{k+1}) < 0 \leq f'(b_{k+1})$. On the other hand, if $f'((a_k + b_k)/2) > 0$, then $a_{k+1} = a_k$, $b_{k+1} = (a_k + b_k)/2$, so again we have $f'(a_{k+1}) < 0 \leq f'(b_{k+1})$. Thus $f'(a_{k+1}) < 0 \leq f'(b_{k+1})$ for all k by induction. Since $f'(a_k) < 0$ for all k, by continuity we have $f'(\hat{x}) = \lim_{k\to\infty} f'(a_k) \leq 0$. Since $f'(b_k) \geq 0$ for all k, we also have $f'(\hat{x}) = \lim_{k\to\infty} f'(b_k) \geq 0$. Since $f'(\hat{x}) \leq 0$ and $f'(\hat{x}) \geq 0$, we must have $f'(\hat{x}) = 0$, so $\hat{x}$ is optimal by condition 1. □