Local Descent

Up to this point, we have focused on optimization involving a single design variable. This chapter introduces a general approach to optimization involving *multivariate* functions, or functions with more than one variable. The focus of this chapter s on how to use *local methods* to incrementally improve a design point until some convergence criterion is met. We being by discussing methods that, at each iteration, choose a descent direction based on a local model and then choose a step size. We then discuss methods that restrict the step to be within a region where the local model is believed to be valid. This chapter concludes with a discussion of convergence conditions.

-1. Descent Direction Iteration

A common approach to optimization is to incrementally improve a design point x by taking a step that minimizers the objective value based on a local model. The local model may be obtained, for example, from a first-or second order Taylor approximation. Optimization algorithms that follow this general approach are referred to as *descent direction methods.* They start with a design point $x^{(1)}$ and then generates a sequence of points, sometimes called *iterates,* to converge to a local minimum.

The iterative descent direction procedure involves the following steps:

1. Check whether $x^{(k)}$ satisfies the termination conditions. If it does, terminate; otherwise proceed to the next step.

2. Determine the descent direction $d^{(k)}$ using local information such as the gradient or Hessian. Some algorithms assume $||d^{(k)}|| = 1,$ but others do note.

3. Determine the step size or learning rate $\alpha^{(k)}$. Some algorithms attempts to optimize the step size so that the step maximally decreases f.

4. Compute the next design point according to:

$$x^{(k+1)} \leftarrow x^{(k)} + \alpha^{(k)} d^{(k)} \tag{1}$$

(We use step size to refer to the magnitude of the overall step. Obtaining a new iterate using equation (1) with a step size $\alpha^{(k)}$ implies that the descent direction $d^{(k)}$ has unit length. We use learning rate to refer to a scalar multiple used on a descent direction vector which does not necessarily have unit length.)

There are many different optimization methods, each of their own ways of determining α and d.

-2. Line Search

For the moment, assume we have chosen a descent direction d. We need to choose the step factor α to obtain our next design point. One approach is to use $line\ search,$ which selects the step factor that minimizer the one-dimensional function:

$$minimize_{\alpha}\ f(x+\alpha d) \tag{2}$$

Line search is a univariate optimization problem. We can apply the univariate optimization method of our choice. To inform the search, we can use the derivative of the line search objective, which is simply the directional derivative along $d\ at\ x+\alpha d$.

Algorithm 4.1

```
def function line_search(f, x, d):
  objective = alpha -> f(x+alpha*d)
  a,b = bracket_minimum(objective)
  alpha = minimize(objective,a,b)
  return x + alpha*d
end.
```

We will be using Bisection method to find the roots of the function.

Link to Colab workbook : Exact Line search

-3. Bisection Method

The bisection method can be used to find *roots* of a function, or points where the function is zero. Such root-finding methods can be used for optimization by applying them to the derivative of the objective, locating where $f'(x) = 0$. In general, we must ensure that the resulting points are indeed local minima.

The *bisection method* maintains a bracket $[a, b]$ in which at least one root is known to exist. If f is continuous on $[a, b]$, and there is some $y \in [f(a), f(b)]$, then the *intermediate value theorem* stipulates that there exists at least one $x \in [a, b]$, such that $f(x) = y$.

The bisection method cuts the bracketed region in half with every iteration. The midpoint $(a + b)/2$ is evaluated, and the new bracket is formed from the midpoint and whichever side that continues to bracket a zero. We can terminate immediately if the midpoint evaluates to zero. Otherwise we can terminate if the midpoint evaluates to zero.
Root-finding algorithms like the bisection method require starting intervals $[a, b]$ on opposite sides

of a zero. That is, $sign(f'(a)) \neq sign(f'(b))$, or equivalently, $f'(a)f'(b) \leq 0$.

The bisection algorithm,where f' is the derivative of the univariate function we seem to optimize. We have $a < b$ that bracket a zero of f'. The interval width tolerance is ϵ. Calling bisection returns the new bracketed interval $[a, b]$ as a tuple.

```
function bisection(f',a,b,eps)
  if a > b; a,b = b,a ; end # ensure a < b

  ya, yb = f'(a), f'(b)
  if ya == 0; b = a; end
  if yb == 0; a = b; end

  while b - a > eps
    x = (a + b)/2
    y = f'(x)
    if y == 0
      a,b = x,x
    else if sign(y) == sign(ya)
      a = x
    else
      b = x
   end
end
  return (a,b)
end
```

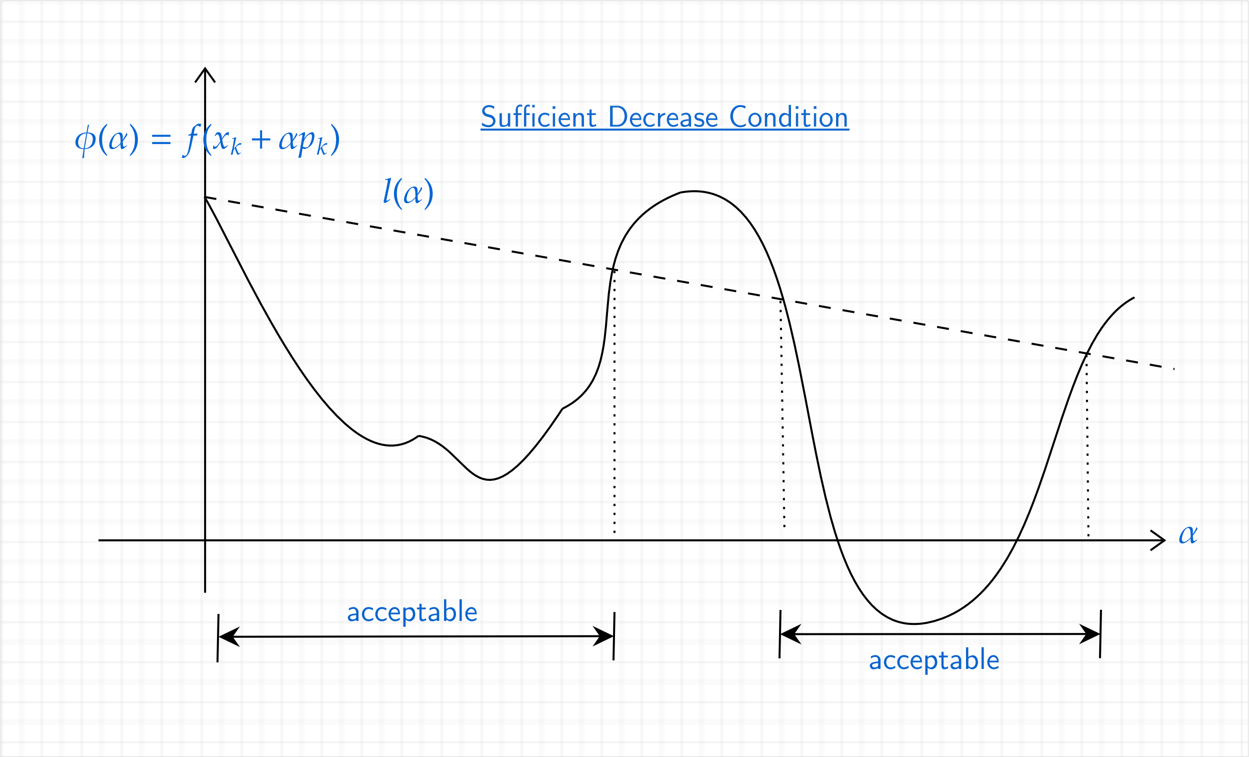
-4. The Wolfe Conditions

A popular inexact line search condition stipulates that α_k should first of all give *sufficient decrease* in the objective function f, as measured by the following inequality:

$$f(x_k + \alpha p_k) \leq f(x_k) + c_1 \alpha \nabla f_k^T p_k, \tag{3}$$

For some constant $c_1 \in (0, 1)$. In other words, the reduction in f should be proportional to both the step length α_k and the directional derivative $\nabla f_k^T p_k$. Inequality (3) is sometimes called the *Armijo condition*.

The sufficient decrease condition is illustrated below.

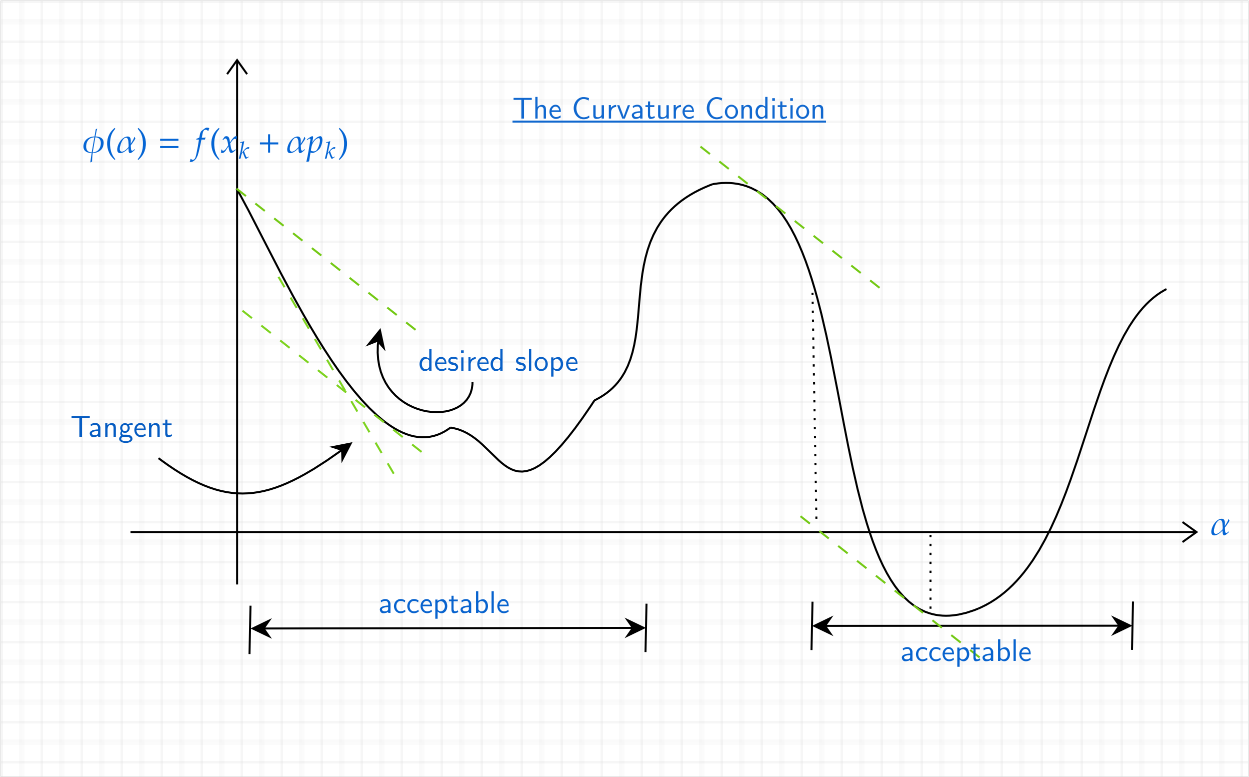

Sufficient Decrease Condition
$\phi(\alpha) = f(x_k + \alpha p_k)$
$l(\alpha)$
α
acceptable
acceptable

The right-hand side of (3), which is a linear function, can be denoted by $l(\alpha)$. The function $l(.)$ has negative slope $c_1 \nabla f_k^T p_k$, but because $c_1 \in (0, 1)$, it lies above the graph of ϕ for small positive values of α. The sufficient decrease condition states that α is acceptable only if $\phi(\alpha) \leq l(\alpha)$. The intervals on which this condition is satisfied are shown in the diagram. In practice, c_1 is chosen to be quite small say $c_1 = 10^{-4}$.

The sufficient decrease condition is not enough by itself to ensure that the algorithm makes reasonable progress because, as we can see, it is satisfied for all sufficiently small values of α. To rule out unacceptably short steps we introduce a second requirement, called the *curvature condition*, which requires α_k to satisfy

$$\nabla f(x_k + \alpha_k p_k)^T p_k \geq c_2 \nabla f_k^T p_k, \tag{4}$$

for some constant $c_2 \in (c_1, 1)$, where c_1 is the constant from equation (3). Note that the left-hand side is simply the derivative $\phi'(\alpha_k)$, so the curvature condition ensures that the slope of ϕ at α_k is greater than c_2 times the initial slope $\phi'(0)$. This makes sense because if the slope $\phi'(\alpha)$ is strongly negative, we have an indication that we can reduce f significantly by moving further along the chosen direction.



On the other hand, if $\phi'(\alpha_k)$ is only slightly negative or even positive, it is a sign that we cannot expect much more decrease in f in this direction, so it makes sense to terminate the line search. Typical values of c_2 are 0.9 when the search direction p_k is chosen by a Newton or quasi-Newton, and 0.1 when p_k is obtained from a non-linear conjugate gradient method.

Backtracking Line Search