1. Basic knowledge beforehand
2. Gradient (of a scalar over a vector)

Note: is the direction that **increases** *f* the fastest!

1. Hessian matrix

Note:

1. we can usually express a quadratic function in the term of:

In the above expression, *A* (a N by N symmetric matrix) would be the hessian matrix (that also explains why quadratic programming is faster than fmincon: no need to calculate Hessian again!).

Note: If *f* is a vector instead of scalar, the Hessian matrix becomes a 3D matrix.

1. Sometimes we also use to denote the Hessian matrix.
2. Jacobian matrix

A Jacobian matrix is the gradient of a vector (M by 1) over another vector (N by 1).

In fact the gradient in (1) is just a special case for Jacobian.

1. Taylor Series
2. Newton method in function solving

Based on Taylor series if a function is differentiable then

So we can iteratively solve the problem:

Usually we also add a coefficient *α* to control the rate of iteration.

**For following discussions we define the optimization target as .**

1. Gradient descent

Gradient descent (steepest descent) method tries to move the variable vector in the **opposite** direction of the Jacobian matrix, such that the value of *f* keeps decreasing in each iteration:

Here *α­k* is also called “learning rate”. Some algorithms will adjust this rate to make the process fast in the beginning and avoid wiggling in the end.

Note: pay attention to the relationship between gradient descent and Newton’s method (the math is exactly the same)! If we approximate to the first order derivative and then take the derivative of the function over , we can also get the above equation.

1. Newton’s method in optimization

Again, based on Taylor series, this time we can expand the target function to the second order:

If we fix and take the derivative of the function over and set it to zero:

As with gradient descent method, we also usually use “learning rate” to control the step size.

Note: intuitively Newton’s method is better than simple gradient descent because it uses second-order derivative information. However, **not every Hessian function is invertible**. Also the inverting can take extra time. Some also argue that if the initial value is too far from the optimal point, the Taylor series may not be close enough to the target function hence the method can fail.

1. Levenberg-Marquadt
2. Conjugate gradients