Second Order Optimization

Second order optimization is more accurate than first order optimization like gradient descent. The paper represents second order optimization methods for non-linear equations using least squares.

- **What is a non-linear equation and a set of non-linear equation?**
  Equations **not** having form $f(x) = ax + b$ or $f(x) = constant$ are considered as non-linear equations.
  i.e. $f(x) = e^x + x^2$ is a non-liner equation.
  Consider a set of m equations in n variables $x_1, \dots, x_n$.
  $f_i(x) = 0; i = 1, \dots, m$
  Here i^{th} equation $f_i(x)$ is also a i^{th} residual. And $x = (x_1, \dots, x_n)$ is a vector of unknowns.
  Collectively $f(x) = (f_1(x), \dots, f_m(x))$, where $f(x)$ is a vector of residuals.

  Here our goal is to find $\widehat{x}$ which minimizes $||f(x)||^2 \ = f_1(x)^2 + \dots + f_m(x)^2$.
  Optimality condition for any $\widehat{x}$ being a solution is to satisfy $\nabla ||f(x)||^2 \ = 0$.
  Important: The optimality condition is necessary but not sufficient condition. Values satisfying the condition may not be a solution.
- **Convexity**
  Now, if the function is **convex** then it reduces computation to find the minimum value of the function. Convexity prevents two local minima and hence if function is convex then it would have usually one minimum value or set of minimum values lying on a line.
  e.g. $f(x) = x^2$
  Mathematically, function $F(x_1, \dots, x_n)$ is convex if its Hessian matrix (second derivative matrix) is positive semidefinite at all x.

- **Newton Algorithm**
  It is a powerful heuristic algorithm for the nonlinear least squares problem.For multi-variables, Newton's method for minimizing f(x) is defined as:
  $$x^{k+1} = x^k - (\bigtriangledown f(x)^T \bigtriangledown f(x))^{-1} \bigtriangledown f(x)^T * f(x);\ H = \bigtriangledown f(x)^T \bigtriangledown f(x)$$
  This iteration gives the Newton algorithm where H is the **Hessian Matrix.**
  SHORTCOMINGS
  The basic Newton algorithm can diverge and the iterations terminate if the derivative matrix is not invertible. So we use another algorithm called **Levenberr-Marquardt Algorithm** which can remove the above drawbacks.

- **levenberg-Marquardt Algorithm**
  This is another powerful heuristic algorithm and also the advanced Version of Newton Algorithm. Here we have two objectives The first objective is an approximation of what we really want to minimize; the second objective expresses the idea that we should not move so far that we cannot trust the affine approximation.
  $$x^{k+1} = x^k - (\bigtriangledown f(x)^T \bigtriangledown f(x) + \lambda I)^{-1} \bigtriangledown f(x)^T * f(x)$$

  Here we are using a new parameter λ **trust parameter.**

- **Data Fitting**
  As in linear model fitting, we choose the parameter θ (approximately) by minimizing the sum of the squares of the prediction residuals
  $$\sum_{i=1}^{N} \left(\widehat{f} \ \left(x^{(i)}; \theta\right) - y^{(i)} \right)^2 where f\,(x; \theta) = \theta_1 e^{\theta_2 x} \cos(\theta_3 x + \theta_4)$$