Contents

This fantastic blog post gives an intuitive introduction to the Hessian.

0.1 Basic Concepts

- The quadratic form:

$$f(\mathbf{x}) = \mathbf{x}^T A\mathbf{x} - \mathbf{b}^T\mathbf{x} + c$$

- A matrix A is positive-definite: $\mathbf{x}^T A\mathbf{x} > 0$ holds for every nonzero vector $\mathbf{x}$.
  - if A is positive-definite, then the surface defined by $f(\mathbf{x})$ is shaped like a paraboloid bowl.
- Gradients: the gradient is the rate of change of some function (in deep learning, this is generally the loss function) in various directions.
  - the gradient is simply the slope of the function at that point.

$$f'(\mathbf{x}) = \frac{1}{2}A^T\mathbf{x} + \frac{1}{2}A\mathbf{x} - \mathbf{b} \tag{1}$$

- With equation(1) in hand, if A is positive-definite, and:
  - if A is symmetric, the above derivative is reduced to $f'(\mathbf{x}) = A\mathbf{x} + b$. $f(\mathbf{x})$ is minimized by the solution to $A\mathbf{x} = b$.
  - If A is not symmetric, gradient descent will find a solution to the system $\frac{1}{2}(A^T + A)\mathbf{x} = \mathbf{b}$
- The second-order derivative (Hessian) is simply the derivative of the derivative. It is the rate of change of the slope.
  - in high-dimensional space, there is a different rate of change = slope for each direction.
  - The second-order derivative is a matrix.
- The Hessian represents the curvature.
  - The Hession affect the paraboloid's shape.
  - The values of $\mathbf{b}$ and c determine where the minimum point of the paraboloid, but do not affact its shape.
- Possiblities of A:
  1. positive-definite
  2. negative-definite
  3. singular
  4. indefinite: gradient descent methods will fail.
- The intuition of A is positive-definite: the quadratic form $f(\mathbf{x})$ is a paraboloid.
- The Hessian matrix of local minimizer is positive definite, while the Hessian matrix of saddle points are indefinite.

Why symmetric positive-definite matrices have this nice property?

Consider the relationship between f at some arbitrary point $\mathbf{p}$ and at the solution point $\mathbf{x} = A^{-1}\mathbf{b}$, we can obtain:

$$f(\mathbf{p}) = f(\mathbf{x}) + \frac{1}{2}(\mathbf{p} - \mathbf{x})^T A(\mathbf{p} - \mathbf{x})$$

If A is positive-definite, the latter term is positive for all $\mathbf{p} \neq \mathbf{x}$. It follows that $\mathbf{x}$ is a global minimum of f.

0.2 Hessian and Optimization

Eigenvalues and Eigenvectors: $M\mathbf{v}_i = \lambda_i \mathbf{v}_i$, where $\mathbf{v}_i$ is eigenvector and λ_i is eigenvalue.

- Eigenvectors and eigenvalues tell us a lot about the nature of the matrix.

In the case of the Hessian, the eigenvectors and eigenvalues have the following important properties:

- Each eigenvector represents a direction where the curvature is independent of the other directions.
- The curvature in the direction of the eigenvector is determined by the eigenvalue.
  - If the eigenvalue is larger, there is a larger curvature, and if it positive, the curvature will be positive, and vice-versa.
- the larger the eigenvalue, the faster the convergence from the direction of its corresponding eigenvector.