E9 211: Adaptive Signal Processing

Newton's Method

Outline

1. Newton's Method
2. Stability condition
3. Convergence rate

Newton's method

- Let us consider the quadratic cost function

$$J(\mathbf{w}) = J_0 + (\mathbf{w} - \mathbf{w}_0)^{\mathrm{H}} \mathbf{R}_x (\mathbf{w} - \mathbf{w}_0)$$

  and an iterative algorithm where the update equation is given by

$$\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} + \mu \mathbf{p}$$

- Let us recall that a necessary condition for $J(\mathbf{w}^{(k+1)}) < J(\mathbf{w}^{(k)})$ is to choose the descent direction $\mathbf{p}$ such that

$$\mathrm{Re}[\nabla J(\mathbf{w}^{(k)})^{\mathrm{H}} \mathbf{p}] < 0$$

- This can be obtained by choosing

$$\mathbf{p} = -\mathbf{B} \nabla J(\mathbf{w}^{(k)}) \quad \text{for any } \mathbf{B} > \mathbf{0}$$

Newton's method

- For the steepest descent, we simply chose $\mathbf{B} = \mathbf{I}$.
- Instead, we can choose $\mathbf{B} = [\nabla^2 J(\mathbf{w}^{(k)})]^{-1}$ where $\nabla^2 J(\mathbf{w}^{(k)})$ is the *Hessian* matrix of the cost function $J(\mathbf{w})$ evaluated at $\mathbf{w} = \mathbf{w}^(k)$.
- This leads to the Newton's method.
- Update equation for the Newton's method is given by

$$\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} - \mu[\nabla^2 J(\mathbf{w}^{(k)})]^{-1} \nabla J(\mathbf{w}^{(k)})$$

Newton's method

- For the quadratic cost function

$$J(\mathbf{w}) = J_0 + (\mathbf{w} - \mathbf{w}_0)^{\mathrm{H}} \mathbf{R}_x (\mathbf{w} - \mathbf{w}_0),$$

  we have

$$\nabla J(\mathbf{w}^{(k)}) = \mathbf{R}_x \mathbf{w}^{(k)} - \mathbf{r}_{xy} \quad \text{and} \quad \nabla^2 J(\mathbf{w}^{(k)}) = \mathbf{R}_x$$

- Newton's method update equation is given by

$$\begin{aligned}
\mathbf{w}^{(k+1)} &= \mathbf{w}^{(k)} - \mu \mathbf{R}_x^{-1} (\mathbf{R}_x \mathbf{w}^{(k)} - \mathbf{r}_{xy}) \\
&= \mathbf{w}^{(k)} - \mu \mathbf{R}_x^{-1} \mathbf{R}_x \mathbf{w}^{(k)} + \mu \mathbf{R}_x^{-1} \mathbf{r}_{xy} \\
&= \mathbf{w}^{(k)} - \mu \mathbf{w}^{(k)} + \mu \mathbf{w}_0
\end{aligned}$$

Newton's method - stability

- Let us define the weight error $\mathbf{e}^{(k)} = \mathbf{w}^{(k)} - \mathbf{w}_0$. Then,

$$\begin{array}{r l} \mathbf{w}^{(k+1)} &= \mathbf{w}^{(k)} - \mu(\mathbf{w}^{(k)} - \mu\mathbf{w}_0) \\ \mathbf{w}_0 &= \mathbf{w}_0 \\ \hline \mathbf{e}^{(k+1)} &= \mathbf{e}^{(k)} - \mu\mathbf{e}^{(k)} \end{array}$$

- We obtain the first-order matrix difference equation as

$$\begin{aligned} \mathbf{e}^{(k+1)} &= (\mathbf{I} - \mu\mathbf{I})\mathbf{e}^{(k)} \\ &= (\mathbf{I} - \mu\mathbf{I})^{(k+1)}\mathbf{e}^{(0)} \\ &= (1 - \mu)^{(k+1)}\mathbf{e}^{(0)} \end{aligned}$$

which is stable if $(1 - \mu)^{(k)} \to 0$.

Newton's method - stability

- The iterations will converge to optimum (i.e., $\mathbf{w}^{(k)} \rightarrow \mathbf{w}_0$) if

$$||\mathbf{e}^{(k)}|| \rightarrow 0 \implies |1-\mu| < 1.$$

- This means that the Newton's method will converge if

$$-1 < 1-\mu < 1 \implies 0 < \mu < 2$$

- Unlike the SGD, the choice of step size μ for the convergence of Newton's method is not depending on the eigen values of $\mathbf{R}_x$.

- In other words, the choice of step size μ to ensure convergence, does not depend on data.

- Special case : Newton's method converges in 1 iteration for quadratic cost functions if we choose $\mu = 1$.

Newton's method - convergence rate

- All entries of $\mathbf{e}^{(k)}$ converge at the same rate.
- The time constant can be computed as

$$|1-\mu|^{\tau} = \frac{1}{e} \implies \tau = \frac{-1}{\ln|1-\mu|} \approx \frac{1}{\mu} \quad \text{(For small } \mu\text{)}$$

- Time constant is indepnedent of data.