E9 211: Adaptive Signal Processing

Steepest Gradient Descent
Outline

1. Steepest gradient descent
2. Stability condition
3. Convergence rate
Suppose we would like to estimate a scalar $s_k : p \times 1$ based on vector valued observations $x_k : M \times 1$

$$x_k = a s_k + n_k, \quad k = 1, 2, \ldots$$

with $a : M \times 1$ and $n_k : M \times 1$ is the noise vector.

The linear estimator (equalizer or beamformer) is given by $\hat{s}_k = w^H x$
Assume source has unit power, i.e., $E(|s_k|^2) = 1$. Also, Let $R_x = E(x_k x_k^H)$ and $r_{xs} = E(x s_k^*)$.

To find the beamformer $w : M \times 1$ by minimizing the output error using the cost function

$$J(w) = E(|w^H x - s_k|^2) = w^H R_x w - w^H r_{xs} - r_{xs}^H w + 1$$

The gradient vector will be

$$\nabla J(w) = R_x w - r_{xs}$$
Linear least-mean-squares estimator

Let the optimum that minimizes $J(w)$ be $w_0$. At the optimum, $J(w_0) = 0$:

$$R_x w_0 - r_{xs} = 0 \Rightarrow w_0 = R_x^{-1} r_{xs}$$

Also, $J(w) = 0$ implies

$$E(x_k x_k^H w - x_k s_k^*) = 0 \Rightarrow E(x_k e_k^*) = 0$$

where the error signal $e_k = w^H x - s_k$

The cost at the optimum is

$$J(w_0) = J_0 = 1 - r_{xs}^H R_x^{-1} r_{xs}$$

The optimum estimator involved $R_x^{-1}$. To avoid this inversion, we compute the optimum iteratively.
The cost function is quadratic in $\mathbf{w}$ and can be expressed as

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

with $\mathbf{w}_0$ being the minimizer.
Steepest gradient descent method

To minimize $f(x)$

- Take initial point $x^{(1)}$ with gradient $\nabla f^{(1)}$

- For a point $x^{(2)}$ close to $x^{(1)}$, we can write the slope of the tangent

$$\nabla f^{(1)} \approx \frac{f(x^{(2)}) - f(x^{(1)})}{x^{(2)} - x^{(1)}} \quad \Rightarrow \quad f(x^{(2)}) \approx f(x^{(1)}) + (x^{(2)} - x^{(1)})\nabla f^{(1)}$$

- Suppose we choose

$$x^{(2)} = x^{(1)} - \mu \nabla f^{(1)}$$

with a small number $\mu$, referred to as the step size.

- Then, $f(x^{(2)}) \approx f(x^{(1)}) - \mu (\nabla f^{(1)})^2 < f(x^{(1)})$.

- At the minimum, $\nabla f^{(1)} = 0$ and $x^{(2)} = x^{(1)}$

- Taking small steps in the direction of the negative gradient, the value of the function becomes smaller.
Steepest gradient descent method

- Let us focus on our objective function $J(w)$ and use the update direction $p$ to get the update equation

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

- Then, we have

$$J(w^{(k+1)}) = (w^{(k)} + \mu p)^H R_x (w^{(k)} + \mu p) - r_{xs}^H (w^{(k)} + \mu p)$$

$$- (w^{(k)} + \mu p)^H r_{xs} + 1$$

$$= J(w^{(k)}) + 2\mu \text{Re}[\nabla J(w^{(k)})^H p] + \mu^2 p^H R_x p$$

From the above equation, the necessary condition for $J(w^{(k+1)}) < J(w^{(k)})$ is

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

- This can be obtained by choosing

$$p = -B \nabla J(w^{(k)})$$

for any $B > 0$

- For steepest gradient descent method, we simply choose $B = I$
Since we have $\nabla J(w) = R_x w - r_{xs}$, the steepest gradient descent iterations are

$$w^{(k+1)} = w^{(k)} + \mu [R_x w^{(k)} - r_{xs}].$$

The iteration is initialized (usually) with $w^0 = 0$. 

The choice of $\mu$ is important for stability and convergence of this technique.
Steepest gradient descent method - stability

Let us define the weight error $e^{(k)} = w^{(k)} - w_0$. Then,

$$
w^{(k+1)} = w^{(k)} - \mu (R_x w^{(k)} - r_{xs})
$$

$$
w_0 = w_0 - \mu (R_x w_0 - r_{xs})
$$

$$
e^{(k+1)} = e^{(k)} - \mu R_x e^{(k)}
$$

We obtain the first-order matrix difference equation

$$
e^{(k+1)} = (I - \mu R_x) e^{(k)} = \cdots = (I - \mu R_x)^{(k+1)} e^{(0)}
$$

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

Let the eigenvalue decomposition $R_x =: U \Lambda U^H$ and

$$
I - \mu R_x =: U \Lambda \mu U^H \Rightarrow (I - \mu R_x)^k = U \Lambda^k \mu U^H = U[I - \Lambda]^k U^H.
$$

Also, let $v^{(k)} = U^H e^{(k)}$, so that $v^{(k)} = [I - \Lambda]^k v^{(0)}$. 
Then the condition for stability of the recursion is
\[ \|e^{(k)}\| = \|v^{(k)}\| \rightarrow 0 \iff |1 - \mu \lambda_i| < 1 \quad i = 1, 2, \ldots, M \]

Since \( \lambda_{\min} = \lambda_1 \leq \lambda_2 \cdots \leq \lambda_M = \lambda_{\max} \), the steepest gradient descent is stable if
\[ 0 \leq \mu \lambda_{\max} \leq 2 \iff 0 \leq \mu \leq \frac{2}{\lambda_{\max}} \]
Steepest gradient descent method - convergence rate

Transient behaviour:

- Since $v_i^{(k)} = (1 - \mu \lambda_i)^k v_i^{(0)}$, different entries of $v^{(k)}$ converge at different rates.

- Modes with $0 < 1 - \mu \lambda_i < 1$ monotonically decay to 0.

- Modes with $-1 < 1 - \mu \lambda_i < 0$ oscillate.

- Mode with the largest magnitude (close to 1) decays at the slowest rate. Suppose $1 - \mu \lambda_{\text{max}} > 0$, the slowest mode is determined by $\lambda_{\text{min}}$. 
Convergence rate:

- Mode with the largest magnitude (close to 1) decays at the slowest rate. Suppose $1 - \mu \lambda_{\text{max}} > 0$, the slowest mode is determined by $\lambda_{\text{min}}$.

- For a function $f(t) = e^{-t/\tau}$, $\tau$ is the time constant, which is the time required for the value of the function to decay by a factor $e$ as $f(t + \tau) = f(t)/e$.

- For $f(\tau) = \|\mathbf{v}^{(\tau)}\| = \|\mathbf{v}^{(0)}\|/e$, the time constant is

$$
\tau = \frac{-1}{\ln(1 - \mu \lambda_{\text{min}})}
$$

For small $\mu$, $\tau \approx \frac{1}{\mu \lambda_{\text{min}}}$

- If $\mu = 1/\lambda_{\text{max}}$, then

$$
\tau \approx \frac{\lambda_{\text{max}}}{\lambda_{\text{min}}} =: \text{cond}(\mathbf{R}_x)
$$

If $\mathbf{R}_x$ is ill-conditioned, then the convergence will be slow.