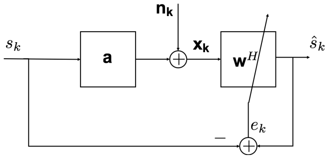
E9 211: Adaptive Signal Processing

Steepest Gradient Descent

Outline

1. Steepest gradient descent
2. Stability condition
3. Convergence rate

Linear least-mean-squares estimator



- Suppose we would like to estimate a scalar $s_k : p \times 1$ based on vector valued observations $\mathbf{x}_k : M \times 1$

$$\mathbf{x}_k = \mathbf{a}s_k + \mathbf{n}_k, \quad k = 1, 2, \dots$$

with $\mathbf{a} : M \times 1$ and $\mathbf{n}_k : M \times 1$ is the noise vector.

- The linear estimator (equalizer or beamformer) is given by $\hat{s}_k = \mathbf{w}^{\mathrm{H}}\mathbf{x}$

Linear least-mean-squares estimator

- Assume source has unit power, i.e., $E(|s_k|^2) = 1$. Also, Let $\mathbf{R}_x = E(\mathbf{x}_k\mathbf{x}_k^{\mathrm{H}})$ and $\mathbf{r}_{xs} = E(\mathbf{x}s_k^*)$.
- To find the beamformer $\mathbf{w} : M \times 1$ by minimizing the output error using the cost function

$$J(\mathbf{w}) = E(|\mathbf{w}^{\mathrm{H}}\mathbf{x} - s_k|^2) = \mathbf{w}^{\mathrm{H}}\mathbf{R}_x\mathbf{w} - \mathbf{w}^{\mathrm{H}}\mathbf{r}_{xs} - \mathbf{r}_{xs}^{\mathrm{H}}\mathbf{w} + 1$$

- The gradient vector will be

$$\nabla J(\mathbf{w}) = \mathbf{R}_x\mathbf{w} - \mathbf{r}_{xs}$$

Linear least-mean-squares estimator

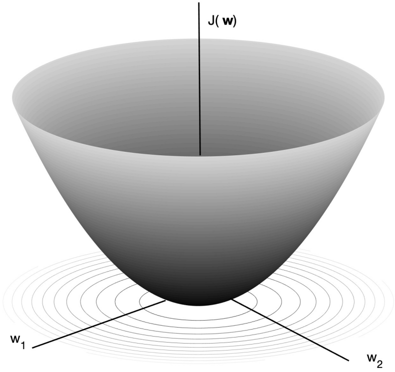
- Let the optimum that minimizes $J(\mathbf{w})$ be $\mathbf{w}_0$. At the optimum, $J(\mathbf{w}_0) = 0$:
$$\mathbf{R}_x \mathbf{w}_0 - \mathbf{r}_{xs} = \mathbf{0} \Rightarrow \mathbf{w}_0 = \mathbf{R}_x^{-1} \mathbf{r}_{xs}$$

- Also, $J(\mathbf{w}) = 0$ implies
$$E(\mathbf{x}_k \mathbf{x}_k^{\mathrm{H}} \mathbf{w} - \mathbf{x}_k s_k^*) = 0 \Rightarrow E(\mathbf{x}_k e_k^*) = 0$$
where the error signal $e_k = \mathbf{w}^{\mathrm{H}} \mathbf{x} - s_k$

- The cost at the optimum is
$$J(\mathbf{w}_0) = J_0 = 1 - \mathbf{r}_{xs}^{\mathrm{H}} \mathbf{R}_x^{-1} \mathbf{r}_{xs}$$

- The optimum estimator involved $\mathbf{R}_x^{-1}$. To avoid this inversion, we compute the optimum *iteratively*.

Linear least-mean-squares objective function



- The cost function is quadratic in $\mathbf{w}$ and can be expressed as

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

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 μ, 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 $J(\mathbf{w})$ and use the update direction $\mathbf{p}$ to get the update equation

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

- Then, we have

$$\begin{aligned} J(\mathbf{w}^{(k+1)}) &= (\mathbf{w}^{(k)} + \mu\mathbf{p})^{\mathrm{H}}\mathbf{R}_x(\mathbf{w}^{(k)} + \mu\mathbf{p}) - \mathbf{r}_{xs}^{\mathrm{H}}(\mathbf{w}^{(k)} + \mu\mathbf{p}) \\ &\qquad - (\mathbf{w}^{(k)} + \mu\mathbf{p})^{\mathrm{H}}\mathbf{r}_{xs} + 1 \\ &= J(\mathbf{w}^{(k)}) + 2\mu\mathrm{Re}[\nabla J(\mathbf{w}^{(k)})^{\mathrm{H}}\mathbf{p}] + \mu^2\mathbf{p}^{\mathrm{H}}\mathbf{R}_x\mathbf{p} \end{aligned}$$

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

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

- This can be obtained by choosing

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

- For steepest gradient descent method, we simply choose $\mathbf{B} = \mathbf{I}$

Steepest gradient descent method

- Since we have $\nabla J(\mathbf{w}) = \mathbf{R}_x \mathbf{w} - \mathbf{r}_{xs}$, the steepest gradient descent iterations are

$$\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} + \mu[\mathbf{R}_x \mathbf{w}^{(k)} - \mathbf{r}_{xs}].$$

  The iteration is initialized (usually) with $\mathbf{w}^0 = \mathbf{0}$.

- The choice of μ is important for stability and convergence of this technique

Steepest gradient descent method - stability

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

$$\begin{array}{cccc} \mathbf{w}^{(k+1)} & = & \mathbf{w}^{(k)} & - & \mu(\mathbf{R}_x\mathbf{w}^{(k)} - \mathbf{r}_{xs}) \\ \mathbf{w}_0 & = & \mathbf{w}_0 & - & \mu(\mathbf{R}_x\mathbf{w}_0 - \mathbf{r}_{xs}) \\ \hline \mathbf{e}^{(k+1)} & = & \mathbf{e}^{(k)} & - & \mu\mathbf{R}_x\mathbf{e}^{(k)} \end{array}$$

- We obtain the first-order matrix difference equation

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

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

- Let the eigenvalue decomposition $\mathbf{R}_x =: \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{\mathrm{H}}$ and

$$\mathbf{I} - \mu\mathbf{R}_x =: \mathbf{U}\mathbf{\Lambda}_\mu\mathbf{U}^{\mathrm{H}} \quad \Rightarrow \quad (\mathbf{I} - \mu\mathbf{R}_x)^k = \mathbf{U}\mathbf{\Lambda}_\mu^k\mathbf{U}^{\mathrm{H}} = \mathbf{U}[\mathbf{I} - \mathbf{\Lambda}]^k\mathbf{U}^{\mathrm{H}}.$$

  Also, let $\mathbf{v}^{(k)} = \mathbf{U}^{\mathrm{H}}\mathbf{e}^{(k)}$, so that $\mathbf{v}^{(k)} = [\mathbf{I} - \mathbf{\Lambda}]^k\mathbf{v}^{(0)}$.

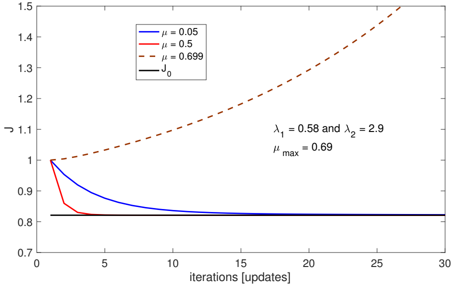
Steepest gradient descent method - stability

- Then the condition for stability of the recursion is

$$\|\mathbf{e}^{(k)}\| = \|\mathbf{v}^{(k)}\| \to 0 \quad \Leftrightarrow \quad |1 - \mu\lambda_i| < 1 \quad i = 1, 2, \ldots, M$$

- Since $\lambda_{\min} = \lambda_1 \leq \lambda_2 \cdots \leq \lambda_M = \lambda_{\max}$, the steepest gradient descent is stable if

$$0 \leq \mu\lambda_{\max} \leq 2 \quad \Leftrightarrow \quad 0 \leq \mu \leq \frac{2}{\lambda_{\max}}$$



Steepest gradient descent method - convergence rate

Transient behaviour:

- Since $v_i^{(k)} = (1 - \mu\lambda_i)^k v_i^{(0)}$, different entries of $\mathbf{v}^{(k)}$ 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 $1 - \mu\lambda_{\max} > 0$, the slowest mode is determined by $\lambda_{\min}$.

Steepest gradient descent method - convergence rate

Convergence rate:

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

- For a function $f(t) = e^{-t/\tau}$, τ 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

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

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

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

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

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