신경망 네트워크와 수학적 기반

Least squares problem

- suppose $m \times n$ matrix A is tall, so $Ax = b$ is over-determined
- for most choices of b, there is no x that satisfies $Ax = b$
- *residual* is $r = Ax - b$
- *least squares problem* : choose x to minimize $\|Ax - b\|^2$
- $\|Ax - b\|^2$ is the *objective function*
- $\hat{x}$ is a *solution* of least squares problem for any n-vector x if

$$\|A\hat{x} - b\|^2 \leq \|Ax - b\|^2$$

- idea: $\hat{x}$ makes residual as small as possible

Least squares problem

- $\hat{x}$ called *least squares approximate solution* of $Ax = b$
- $\hat{x}$ need not (and usually does not) satisfy $A\hat{x} = b$
- but if $\hat{x}$ does satisfy $A\hat{x} = b$, then it solves least squares problem

Least squares

Column interpretation

- suppose a_1, ..., a_n are columns of A, then

$$\|Ax - b\|^2 = \|(x_1 a_1 + \cdots + x_n a_n) - b\|^2$$

- so least squares problem is to find a linear combination of columns of A

  that is closest to b

- if $\hat{x}$ is a solution of least squares problem, the m-vector is closest to b among all linear combinations of columns of A

$$A\hat{x} = \hat{x}_1 a_1 + \cdots + \hat{x}_n a_n$$

Least squares
Row interpretation

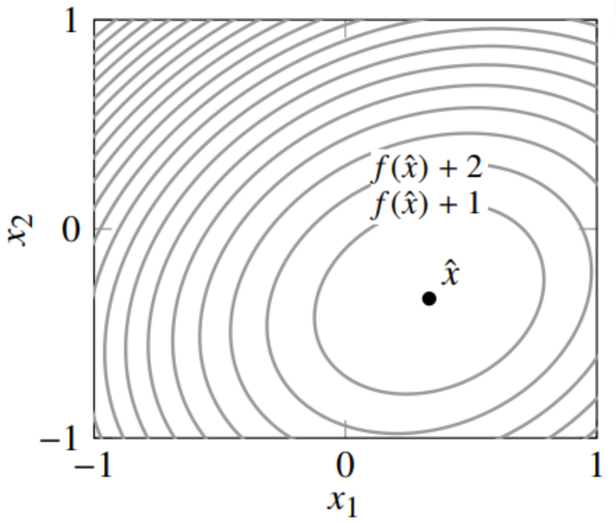
- suppose $\tilde{a}_1^T, ..., \tilde{a}_m^T$ are rows of A
- residual components are $r_i = \tilde{a}_1^T x - b_i$
- least squares objective is the sum of squares of the residuals

$$\|Ax - b\|^2 = (\tilde{a}_1^T x - b_1)^2 + \cdots + (\tilde{a}_m^T x - b_m)^2$$

- so least squares minimizes sum of squares of residuals
  - solving $Ax = b$ is making all residuals zero
  - least squares attempts to make them all small

Least squares
Example

$$A = \begin{bmatrix} 2 & 0 \\ -1 & 1 \\ 0 & 2 \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}$$



- $Ax = b$ has no solution
- least squares problem is to choose x to minimize

$$\|Ax - b\|^2 = (2x_1 - 1)^2 + (-x_1 + x_2)^2 + (2x_2 + 1)^2$$

- least squares approximate solution is $\hat{x} = (1/3, -1/3)$ (say, via calculus)
- $\|A\hat{x} - b\|^2 = 2/3$ is smallest possible value of $\|Ax - b\|^2$
- $A\hat{x} = (2/3, -2/3, -2/3)$ is linear combination of columns of A closest to b

Least squares

Solution of least squares problem

- we make one assumption: A has linearly independent columns
- this implies that Gram matrix A^TA is invertible
- unique solution of least squares problem is

$$\hat{x} = (A^TA)^{-1}A^Tb = A^{\dagger}b$$

- $x = A^{-1}b$ is a solution of square invertible system $Ax = b$

Least squares

Derivation via calculus

- define

$$f(x) = \|Ax - b\|^2 = \sum_{i=1}^{m}\left(\sum_{j=1}^{n} A_{ij}x_j - b_i\right)^2$$

- Solution $\hat{x}$ satisfies

$$\frac{\partial f}{\partial x_k}(\hat{x}) = \nabla f(\hat{x})_k = 0, \quad k = 1, \ldots, n$$

- taking partial derivatives we get $\nabla f(x)_k = (2A^T(Ax - b))_k$
- in matrix-vector notation: $\nabla f(\hat{x}) = 2A^T(A\hat{x} - b) = 0$
- so $\hat{x}$ satisfies *normal equations* $(A^TA)\hat{x} = A^T b$
- and therefore $\hat{x} = (A^TA)^{-1}A^Tb$

Direct verification

- let $\hat{x} = (A^TA)^{-1}A^Tb$, so $A^T(A\hat{x} - b) = 0$
- for any n-vector x we have

$$\begin{aligned}
\|Ax - b\|^2 &= \|(Ax - A\hat{x}) + (A\hat{x} - b)\|^2 \\
&= \|A(x - \hat{x})\|^2 + \|A\hat{x} - b\|^2 + 2(A(x - \hat{x}))^T(A\hat{x} - b) \\
&= \|A(x - \hat{x})\|^2 + \|A\hat{x} - b\|^2 + 2(x - \hat{x})^T A^T (A\hat{x} - b) \\
&= \|A(x - \hat{x})\|^2 + \|A\hat{x} - b\|^2
\end{aligned}$$

- so for any x, $\|Ax - b\|^2 \geq \|A\hat{x} - b\|^2$
- if equality holds, $A(x - \hat{x}) = 0$, which implies $x = \hat{x}$ since columns of A are linearly independent

Computing least squares approximate solutions

- compute QR factorization of A: $A = QR$
- QR factorization exists since columns of A are linearly independent
- to compute $\hat{x} = A^{\dagger}b = R^{-1}Q^{T}b$
- identical to algorithm for solving $Ax = b$ for square invertible A
- but when A is tall, gives least squares approximate solution