Machine Learning: Optimization 1

Hiroshi Shimodaira and Hao Tang

2025

Ver. 1.0.2

Optimisation problems we saw so far

- Squared error in Linear regression

$$\min_{\boldsymbol{w}} L$$

$$L = \sum_{i=1}^{N} (\boldsymbol{w}^\top \boldsymbol{x}_i - y_i)^2$$

Optimisation problems we saw so far

- Squared error in Linear regression

$$\min_{\boldsymbol{w}} L$$

$$L = \sum_{i=1}^{N} (\boldsymbol{w}^\top \boldsymbol{x}_i - y_i)^2$$

- Log-loss in logistic regression

$$\min_{\boldsymbol{w}} \text{NLL}$$

$$\text{NLL} = \sum_{i=1}^{N} \log\left(1 + \exp(-y_i \boldsymbol{w}^\top \boldsymbol{x}_i)\right)$$

Optimisation problems we saw so far

- Squared error in Linear regression

$$\min_{\boldsymbol{w}} L$$

$$L = \sum_{i=1}^{N} (\boldsymbol{w}^\top \boldsymbol{x}_i - y_i)^2$$

- Log-loss in logistic regression

$$\min_{\boldsymbol{w}} \text{NLL}$$

$$\text{NLL} = \sum_{i=1}^{N} \log\left(1 + \exp(-y_i \boldsymbol{w}^\top \boldsymbol{x}_i)\right)$$

  - How do we find the optimal solution?

Optimisation problems we saw so far

- Squared error in Linear regression

$$\min_{\boldsymbol{w}} L$$

$$L = \sum_{i=1}^{N} (\boldsymbol{w}^\top \boldsymbol{x}_i - y_i)^2$$

- Log-loss in logistic regression

$$\min_{\boldsymbol{w}} \text{NLL}$$

$$\text{NLL} = \sum_{i=1}^{N} \log\left(1 + \exp(-y_i \boldsymbol{w}^\top \boldsymbol{x}_i)\right)$$

  - How do we find the optimal solution?
  - How do we know the solution is optimal?

Optimisation problems we saw so far

- Squared error in Linear regression

$$\min_{\boldsymbol{w}} L$$

$$L = \sum_{i=1}^{N} (\boldsymbol{w}^\top \boldsymbol{x}_i - y_i)^2$$

- Log-loss in logistic regression

$$\min_{\boldsymbol{w}} \text{NLL}$$

$$\text{NLL} = \sum_{i=1}^{N} \log\left(1 + \exp(-y_i \boldsymbol{w}^\top \boldsymbol{x}_i)\right)$$

  - How do we find the optimal solution?
  - How do we know the solution is optimal?
  - How do we know the solution is unique?

A toy example of constrained optimisation problem

Question: Given a rope of length L meters, find the rectangle with the largest area that can be formed using the rope.

A toy example of constrained optimisation problem

Question: Given a rope of length L meters, find the rectangle with the largest area that can be formed using the rope.

Optimisation problem: Letting the area $S(x, y) = x\,y$,

$$\begin{aligned} \max_{x,y} \quad & S(x, y) \\ \text{subject to:} \quad & 2(x + y) = L \end{aligned}$$

A toy example of constrained optimisation problem

Question: Given a rope of length L meters, find the rectangle with the largest area that can be formed using the rope.

Optimisation problem: Letting the area $S(x, y) = x\,y$,

$$\begin{aligned} \max_{x,y} \quad & S(x, y) \\ \text{subject to:} \quad & 2(x + y) = L \end{aligned}$$

$$\begin{aligned} S = xy &= x(\frac{L}{2} - x) \\ &= -x^2 + \frac{L}{2}x \\ &= -(x - \frac{L}{4})^2 + \frac{L^2}{16} \end{aligned}$$

A toy example of constrained optimisation problem

Question: Given a rope of length L meters, find the rectangle with the largest area that can be formed using the rope.

Optimisation problem: Letting the area $S(x, y) = x\,y$,

$$\begin{aligned} \max_{x,y} \quad & S(x, y) \\ \text{subject to:} \quad & 2(x + y) = L \end{aligned}$$

$$\begin{aligned} S = xy &= x(\frac{L}{2} - x) \\ &= -x^2 + \frac{L}{2}x \\ &= -(x - \frac{L}{4})^2 + \frac{L^2}{16} \end{aligned}$$

$$\frac{dS}{dx} = -2x + \frac{L}{2}, \quad \frac{d^2S}{dx^2} = -2$$

Notation

- Vector $\boldsymbol{x} \in \mathbb{R}^d$

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix} = \begin{bmatrix} x_1 & x_2 & \cdots & x_d \end{bmatrix}^\top$$

Notation

- Vector $\boldsymbol{x} \in \mathbb{R}^d$

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix} = \begin{bmatrix} x_1 & x_2 & \cdots & x_d \end{bmatrix}^\top$$

- Matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

$$= \begin{bmatrix} \boldsymbol{a}_1 & \boldsymbol{a}_2 & \cdots & \boldsymbol{a}_n \end{bmatrix}$$

$$= \begin{bmatrix} \tilde{\boldsymbol{a}}_1^\top \\ \tilde{\boldsymbol{a}}_2^\top \\ \vdots \\ \tilde{\boldsymbol{a}}_m^\top \end{bmatrix}$$

Notation (*cont.*)

- Data matrix $\boldsymbol{X} \in \mathbb{R}^{N \times d}$

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1^\top \\ \boldsymbol{x}_2^\top \\ \vdots \\ \boldsymbol{x}_N^\top \end{bmatrix}$$

Notation (*cont.*)

- Data matrix $\boldsymbol{X} \in \mathbb{R}^{N \times d}$

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1^\top \\ \boldsymbol{x}_2^\top \\ \vdots \\ \boldsymbol{x}_N^\top \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{Nd} \end{bmatrix}$$

Vector matrix multiplication

$$\boldsymbol{A} \in \mathbb{R}^{m \times d}, \ \boldsymbol{x} \in \mathbb{R}^{d}$$

$$\boldsymbol{A}\boldsymbol{x} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1d} \\ a_{21} & a_{22} & \cdots & a_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{md} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix}$$

$$= \begin{bmatrix} a_{11}x_1 + \cdots + a_{1d}x_d \\ \vdots \\ a_{m1}x_1 + \cdots + a_{md}x_d \end{bmatrix} = \begin{bmatrix} \tilde{\boldsymbol{a}}_1^\top \boldsymbol{x} \\ \vdots \\ \tilde{\boldsymbol{a}}_m^\top \boldsymbol{x} \end{bmatrix}$$

Gradients and Hessians

$$f(\boldsymbol{x}) : \mathbb{R}^d \to \mathbb{R}$$

- Gradient of $f(\boldsymbol{x})$

$$\frac{\partial f}{\partial \boldsymbol{x}} = \nabla f = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_d} \end{bmatrix} \tag{1}$$

NB: See Wikipedia: Matrix Calculus for layout conventions

Gradients and Hessians

$$f(\boldsymbol{x}) : \mathbb{R}^d \to \mathbb{R}$$

- Gradient of $f(\boldsymbol{x})$

$$\frac{\partial f}{\partial \boldsymbol{x}} = \nabla f = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_d} \end{bmatrix} \tag{1}$$

- Hessian of $f(\boldsymbol{x})$

$$\boldsymbol{H}_f = \frac{\partial^2 f}{\partial \boldsymbol{x}^2} = \nabla^2 f = \begin{bmatrix} \frac{\partial f}{\partial x_1^2} & \cdots & \frac{\partial f}{\partial x_1 x_d} \\ & \vdots & \\ \frac{\partial f}{\partial x_d x_1} & \cdots & \frac{\partial f}{\partial x_d^2} \end{bmatrix} \tag{2}$$

NB: See Wikipedia: Matrix Calculus for layout conventions

Useful formula 1

$$\frac{\partial \boldsymbol{a}^\top \boldsymbol{x}}{\partial \boldsymbol{x}} = \boldsymbol{a} \tag{3}$$

Useful formula 1

$$\frac{\partial \boldsymbol{a}^\top \boldsymbol{x}}{\partial \boldsymbol{x}} = \boldsymbol{a} \tag{3}$$

$$\boldsymbol{a}^T \boldsymbol{x} = a_1 x_1 + a_2 x_2 + \cdots + a_d x_d$$

Useful formula 1

$$\frac{\partial \boldsymbol{a}^\top \boldsymbol{x}}{\partial \boldsymbol{x}} = \boldsymbol{a} \tag{3}$$

$$\boldsymbol{a}^T \boldsymbol{x} = a_1 x_1 + a_2 x_2 + \cdots + a_d x_d$$

$$\frac{\partial \boldsymbol{a}^T \boldsymbol{x}}{\partial \boldsymbol{x}} = \begin{bmatrix} \frac{\partial \boldsymbol{a}^T \boldsymbol{x}}{\partial x_1} \\ \vdots \\ \frac{\partial \boldsymbol{a}^T \boldsymbol{x}}{\partial x_d} \end{bmatrix}$$

Useful formula 1

$$\frac{\partial \boldsymbol{a}^\top \boldsymbol{x}}{\partial \boldsymbol{x}} = \boldsymbol{a} \tag{3}$$

$$\boldsymbol{a}^T \boldsymbol{x} = a_1 x_1 + a_2 x_2 + \cdots + a_d x_d$$

$$\frac{\partial \boldsymbol{a}^T \boldsymbol{x}}{\partial \boldsymbol{x}} = \begin{bmatrix} \frac{\partial \boldsymbol{a}^T \boldsymbol{x}}{\partial x_1} \\ \vdots \\ \frac{\partial \boldsymbol{a}^T \boldsymbol{x}}{\partial x_d} \end{bmatrix} = \begin{bmatrix} a_1 \\ \vdots \\ a_d \end{bmatrix} = \boldsymbol{a}$$

Useful formula 2

$$\frac{\partial \boldsymbol{b}^\top \boldsymbol{A}\boldsymbol{x}}{\partial \boldsymbol{x}} = \boldsymbol{A}^\top \boldsymbol{b}$$

Useful formula 2

$$\frac{\boldsymbol{b}^\top \boldsymbol{A}\boldsymbol{x}}{\partial \boldsymbol{x}} = \boldsymbol{A}^\top \boldsymbol{b}$$

Let $\boldsymbol{a}^\top = \boldsymbol{b}^\top \boldsymbol{A}$

$$\frac{\partial \boldsymbol{b}^\top \boldsymbol{A}\boldsymbol{x}}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{a}^\top \boldsymbol{x}}{\partial \boldsymbol{x}}$$

Useful formula 2

$$\frac{\partial \boldsymbol{b}^\top \boldsymbol{A}\boldsymbol{x}}{\partial \boldsymbol{x}} = \boldsymbol{A}^\top \boldsymbol{b}$$

Let $\boldsymbol{a}^\top = \boldsymbol{b}^\top \boldsymbol{A}$

$$\frac{\partial \boldsymbol{b}^\top \boldsymbol{A}\boldsymbol{x}}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{a}^\top \boldsymbol{x}}{\partial \boldsymbol{x}} = \boldsymbol{a} = \left(\boldsymbol{b}^\top \boldsymbol{A}\right)^\top = \boldsymbol{A}^\top \boldsymbol{b}$$

Useful formula 3

$$\frac{\partial \boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x}}{\partial \boldsymbol{x}} = (\boldsymbol{A} + \boldsymbol{A}^\top)\boldsymbol{x}$$

Useful formula 3

$$\frac{\partial \boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x}}{\partial \boldsymbol{x}} = (\boldsymbol{A} + \boldsymbol{A}^\top)\boldsymbol{x}$$

$$\boldsymbol{x}^\top \boldsymbol{A} \boldsymbol{x} = \boldsymbol{x}^\top \begin{bmatrix} \tilde{\boldsymbol{a}}_1^\top \boldsymbol{x} \\ \vdots \\ \tilde{\boldsymbol{a}}_d^\top \boldsymbol{x} \end{bmatrix}$$

Useful formula 3

$$\frac{\partial \boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x}}{\partial \boldsymbol{x}} = (\boldsymbol{A} + \boldsymbol{A}^\top)\boldsymbol{x}$$

$$\boldsymbol{x}^\top \boldsymbol{A} \boldsymbol{x} = \boldsymbol{x}^\top \begin{bmatrix} \tilde{\boldsymbol{a}}_1^\top \boldsymbol{x} \\ \vdots \\ \tilde{\boldsymbol{a}}_d^\top \boldsymbol{x} \end{bmatrix}$$

$$= x_1 \tilde{\boldsymbol{a}}_1^\top \boldsymbol{x} + x_2 \tilde{\boldsymbol{a}}_2^\top \boldsymbol{x} + \cdots + x_d \tilde{\boldsymbol{a}}_d^\top \boldsymbol{x}$$

Useful formula 3

$$\frac{\partial \boldsymbol{x}^T \boldsymbol{A}\boldsymbol{x}}{\partial \boldsymbol{x}} = (\boldsymbol{A} + \boldsymbol{A}^\top)\boldsymbol{x}$$

$$\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} = \boldsymbol{x}^\top \begin{bmatrix} \tilde{\boldsymbol{a}}_1^\top \boldsymbol{x} \\ \vdots \\ \tilde{\boldsymbol{a}}_d^\top \boldsymbol{x} \end{bmatrix}$$

$$= x_1 \tilde{\boldsymbol{a}}_1^\top \boldsymbol{x} + x_2 \tilde{\boldsymbol{a}}_2^\top \boldsymbol{x} + \cdots + x_d \tilde{\boldsymbol{a}}_d^\top \boldsymbol{x}$$

$$\frac{\partial \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}}{\partial \boldsymbol{x}} = \begin{bmatrix} \frac{\partial \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}}{\partial x_1} \\ \vdots \\ \frac{\partial \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}}{\partial x_d} \end{bmatrix}$$

Useful formula 3

$$\frac{\partial \boldsymbol{x}^T \boldsymbol{A}\boldsymbol{x}}{\partial \boldsymbol{x}} = (\boldsymbol{A} + \boldsymbol{A}^\top)\boldsymbol{x}$$

$$\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} = \boldsymbol{x}^\top \begin{bmatrix} \tilde{\boldsymbol{a}}_1^\top \boldsymbol{x} \\ \vdots \\ \tilde{\boldsymbol{a}}_d^\top \boldsymbol{x} \end{bmatrix}$$

$$= x_1 \tilde{\boldsymbol{a}}_1^\top \boldsymbol{x} + x_2 \tilde{\boldsymbol{a}}_2^\top \boldsymbol{x} + \cdots + x_d \tilde{\boldsymbol{a}}_d^\top \boldsymbol{x}$$

$$\frac{\partial \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}}{\partial \boldsymbol{x}} = \begin{bmatrix} \frac{\partial \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}}{\partial x_1} \\ \vdots \\ \frac{\partial \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}}{\partial x_d} \end{bmatrix} = \begin{bmatrix} \tilde{\boldsymbol{a}}_1^\top \boldsymbol{x} + \boldsymbol{a}_1^\top \boldsymbol{x} \\ \vdots \\ \tilde{\boldsymbol{a}}_d^\top \boldsymbol{x} + \boldsymbol{a}_d^\top \boldsymbol{x} \end{bmatrix} = \left(\boldsymbol{A} + \boldsymbol{A}^\top\right) \boldsymbol{x}$$

MSE in linear regression

$$L = \frac{1}{N} \sum_{i=1}^{N} (\boldsymbol{w}^\top \boldsymbol{x}_i - y_i)^2$$

MSE in linear regression

$$
\begin{aligned}
L &= \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{w}^\top \boldsymbol{x}_i - y_i)^2 \\
&= \frac{1}{N}\begin{bmatrix} (\boldsymbol{w}^\top \boldsymbol{x}_1 - y_1) & \cdots & (\boldsymbol{w}^\top \boldsymbol{x}_N - y_N) \end{bmatrix} \begin{bmatrix} (\boldsymbol{w}^\top \boldsymbol{x}_1 - y_1) \\ \vdots \\ (\boldsymbol{w}^\top \boldsymbol{x}_N - y_N) \end{bmatrix}
\end{aligned}
$$

MSE in linear regression

$$
\begin{aligned}
L &= \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{w}^\top \boldsymbol{x}_i - y_i)^2 \\
&= \frac{1}{N}\begin{bmatrix} (\boldsymbol{w}^\top \boldsymbol{x}_1 - y_1) & \cdots & (\boldsymbol{w}^\top \boldsymbol{x}_N - y_N) \end{bmatrix} \begin{bmatrix} (\boldsymbol{w}^\top \boldsymbol{x}_1 - y_1) \\ \vdots \\ (\boldsymbol{w}^\top \boldsymbol{x}_N - y_N) \end{bmatrix} \\
&= \frac{1}{N}(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y})^\top (\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y})
\end{aligned}
$$

MSE in linear regression

$$L = \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{w}^\top \boldsymbol{x}_i - y_i)^2$$

$$= \frac{1}{N}\begin{bmatrix} (\boldsymbol{w}^\top \boldsymbol{x}_1 - y_1) & \cdots & (\boldsymbol{w}^\top \boldsymbol{x}_N - y_N) \end{bmatrix} \begin{bmatrix} (\boldsymbol{w}^\top \boldsymbol{x}_1 - y_1) \\ \vdots \\ (\boldsymbol{w}^\top \boldsymbol{x}_N - y_N) \end{bmatrix}$$

$$= \frac{1}{N}(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y})^\top(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}) = \frac{1}{N}\|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|^2$$

MSE in linear regression

$$
\begin{aligned}
L &= \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{w}^\top \boldsymbol{x}_i - y_i)^2 \\
&= \frac{1}{N}\begin{bmatrix} (\boldsymbol{w}^\top \boldsymbol{x}_1 - y_1) & \cdots & (\boldsymbol{w}^\top \boldsymbol{x}_N - y_N) \end{bmatrix} \begin{bmatrix} (\boldsymbol{w}^\top \boldsymbol{x}_1 - y_1) \\ \vdots \\ (\boldsymbol{w}^\top \boldsymbol{x}_N - y_N) \end{bmatrix} \\
&= \frac{1}{N}(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y})^\top(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}) = \frac{1}{N}\|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|^2 \\
&= \frac{1}{N}\left(\boldsymbol{w}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{w} - 2\boldsymbol{y}^\top \boldsymbol{X}\boldsymbol{w} + \boldsymbol{y}^\top \boldsymbol{y}\right)
\end{aligned}
$$

Gradient of MSE in linear regression

$$L = \frac{1}{N}\left(\boldsymbol{w}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{w} - 2\boldsymbol{y}^\top \boldsymbol{X}\boldsymbol{w} + \boldsymbol{y}^\top \boldsymbol{y}\right)$$

Gradient of MSE in linear regression

$$L = \frac{1}{N}\left(\boldsymbol{w}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{w} - 2\boldsymbol{y}^\top \boldsymbol{X}\boldsymbol{w} + \boldsymbol{y}^\top \boldsymbol{y}\right)$$

$$\frac{\partial L}{\partial \boldsymbol{w}} = \frac{1}{N}\left(\frac{\partial \boldsymbol{w}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{w}}{\partial \boldsymbol{w}} - 2\frac{\partial \boldsymbol{y}^\top \boldsymbol{X}\boldsymbol{w}}{\partial \boldsymbol{w}}\right)$$

Gradient of MSE in linear regression

$$L = \frac{1}{N}\left(\boldsymbol{w}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{w} - 2\boldsymbol{y}^\top \boldsymbol{X}\boldsymbol{w} + \boldsymbol{y}^\top \boldsymbol{y}\right)$$

$$\begin{aligned}\frac{\partial L}{\partial \boldsymbol{w}} &= \frac{1}{N}\left(\frac{\partial \boldsymbol{w}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{w}}{\partial \boldsymbol{w}} - 2\frac{\partial \boldsymbol{y}^\top \boldsymbol{X}\boldsymbol{w}}{\partial \boldsymbol{w}}\right) \\ &= \frac{1}{N}\left(\left(\boldsymbol{X}^\top \boldsymbol{X} + (\boldsymbol{X}^\top \boldsymbol{X})^\top\right)\boldsymbol{w} - 2\boldsymbol{X}^\top \boldsymbol{y}\right)\end{aligned}$$

Gradient of MSE in linear regression

$$L = \frac{1}{N}\left(\boldsymbol{w}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{w} - 2\boldsymbol{y}^\top \boldsymbol{X}\boldsymbol{w} + \boldsymbol{y}^\top \boldsymbol{y}\right)$$

$$\begin{aligned}\frac{\partial L}{\partial \boldsymbol{w}} &= \frac{1}{N}\left(\frac{\partial \boldsymbol{w}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{w}}{\partial \boldsymbol{w}} - 2\frac{\partial \boldsymbol{y}^\top \boldsymbol{X}\boldsymbol{w}}{\partial \boldsymbol{w}}\right) \\ &= \frac{1}{N}\left(\left(\boldsymbol{X}^\top \boldsymbol{X} + (\boldsymbol{X}^\top \boldsymbol{X})^\top\right)\boldsymbol{w} - 2\boldsymbol{X}^\top \boldsymbol{y}\right) \\ &= \frac{2}{N}\left(\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{w} - \boldsymbol{X}^\top \boldsymbol{y}\right)\end{aligned}$$

Gradient of MSE in linear regression

$$L = \frac{1}{N}\left(\boldsymbol{w}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{w} - 2\boldsymbol{y}^\top \boldsymbol{X}\boldsymbol{w} + \boldsymbol{y}^\top \boldsymbol{y}\right)$$

$$\begin{aligned}
\frac{\partial L}{\partial \boldsymbol{w}} &= \frac{1}{N}\left(\frac{\partial \boldsymbol{w}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{w}}{\partial \boldsymbol{w}} - 2\frac{\partial \boldsymbol{y}^\top \boldsymbol{X}\boldsymbol{w}}{\partial \boldsymbol{w}}\right) \\
&= \frac{1}{N}\left(\left(\boldsymbol{X}^\top \boldsymbol{X} + (\boldsymbol{X}^\top \boldsymbol{X})^\top\right)\boldsymbol{w} - 2\boldsymbol{X}^\top \boldsymbol{y}\right) \\
&= \frac{2}{N}\left(\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{w} - \boldsymbol{X}^\top \boldsymbol{y}\right) \\
&= \frac{2}{N}\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y})
\end{aligned}$$

Gradient of MSE in linear regression

$$L = \frac{1}{N}\left(\boldsymbol{w}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{w} - 2\boldsymbol{y}^\top \boldsymbol{X}\boldsymbol{w} + \boldsymbol{y}^\top \boldsymbol{y}\right)$$

$$\begin{aligned}
\frac{\partial L}{\partial \boldsymbol{w}} &= \frac{1}{N}\left(\frac{\partial \boldsymbol{w}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{w}}{\partial \boldsymbol{w}} - 2\frac{\partial \boldsymbol{y}^\top \boldsymbol{X}\boldsymbol{w}}{\partial \boldsymbol{w}}\right) \\
&= \frac{1}{N}\left(\left(\boldsymbol{X}^\top \boldsymbol{X} + (\boldsymbol{X}^\top \boldsymbol{X})^\top\right)\boldsymbol{w} - 2\boldsymbol{X}^\top \boldsymbol{y}\right) \\
&= \frac{2}{N}\left(\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{w} - \boldsymbol{X}^\top \boldsymbol{y}\right) \\
&= \frac{2}{N}\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y})
\end{aligned}$$

Letting $\frac{\partial L}{\partial \boldsymbol{w}} = \boldsymbol{0}$ yields

$$\begin{aligned}
\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}) &= \boldsymbol{0} \\
\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{w} &= \boldsymbol{X}^\top \boldsymbol{y} \\
\boldsymbol{w} &= (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y}
\end{aligned}$$

Log loss in logistic regression training

Letting $(1, \boldsymbol{x})$ and $(b, \boldsymbol{w})$ denoted as $\boldsymbol{x}$ and $\boldsymbol{w}$, respectively,

$$\begin{aligned} \mathrm{LL} = \log \prod_{i=1}^{N} p(y_i|\boldsymbol{x}_i, \boldsymbol{\theta}) &= \sum_{i=1}^{N} \log \frac{1}{1 + \exp(-y_i \boldsymbol{w}^\top \boldsymbol{x}_i)} \\ &= \sum_{i=1}^{N} -\log \left(1 + \exp(-y_i \boldsymbol{w}^\top \boldsymbol{x}_i) \right) \end{aligned}$$

$$\mathrm{NLL} = \sum_{i=1}^{N} \log \left(1 + \exp(-y_i \boldsymbol{w}^\top \boldsymbol{x}_i) \right)$$

Gradient of log loss

$$\text{NLL} = \sum_{i=1}^{N} \log\left(1 + \exp(-y_i \boldsymbol{w}^\top \boldsymbol{x}_i)\right)$$

$$\begin{aligned}\frac{\partial\, \text{NLL}}{\partial \boldsymbol{w}} &= \sum_{i=1}^{N} \frac{\exp(-y_i \boldsymbol{w}^\top \boldsymbol{x}_i)}{1 + \exp(-y_i \boldsymbol{w}^\top \boldsymbol{x}_i)}(-y_i \boldsymbol{x}_i) \\ &= \sum_{i=1}^{N} \left(1 - \frac{1}{1 + \exp(-y_i \boldsymbol{w}^\top \boldsymbol{x}_i)}\right)(-y_i \boldsymbol{x}_i)\end{aligned}$$

What if we use 0/1 labels instead of -1/+1?

$$\text{NLL} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log s_i + (1-y_i)\log(1-s_i)$$

$$\text{where } s_i = \sigma(\boldsymbol{w}^\top \boldsymbol{x}_i) = \frac{1}{1+\exp(-(\boldsymbol{w}^\top \boldsymbol{x}_i))}.$$

$$\frac{\partial \text{NLL}}{\partial \boldsymbol{w}} = -\frac{1}{N}\sum_{i=1}^{N} y_i \frac{1}{s_i}\frac{\partial s_i}{\partial \boldsymbol{w}} + (1-y_i)\frac{-1}{1-s_i}\frac{\partial s_i}{\partial \boldsymbol{w}}$$

$$\text{Since } \frac{\partial s_i}{\partial \boldsymbol{w}} = \sigma(\boldsymbol{w}^\top \boldsymbol{x}_i)(1-\sigma(\boldsymbol{w}^\top \boldsymbol{x}_i))\boldsymbol{x}_i \;=\; s_i(1-s_i)\boldsymbol{x}_i$$

$$\frac{\partial \text{NLL}}{\partial \boldsymbol{w}} = -\frac{1}{N}\sum_{i=1}^{N} \left(y_i(1-s_i) - (1-y_i)s_i\right)\boldsymbol{x}_i$$

$$= \frac{1}{N}\sum_{i=1}^{N} (s_i - y_i)\boldsymbol{x}_i$$

Vectorisation in Python

- Consider a linear regression model $\boldsymbol{w}^\top \boldsymbol{x} + b$ that has been trained already. How are the outputs calculated for a test samples $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$?
- Consider a logistic regression model $p(y=1 \mid \boldsymbol{x}) = \frac{1}{1+\exp(\boldsymbol{w}^\top \boldsymbol{x}+b)}$. How are the outputs calculated for the test samples?

Quizzes

- Show $\dfrac{\partial \boldsymbol{x}^T \boldsymbol{A}\boldsymbol{x}}{\partial \boldsymbol{x}} = (\boldsymbol{A} + \boldsymbol{A}^\top)\boldsymbol{x}$
- Find $\dfrac{\partial \boldsymbol{x}^T \boldsymbol{B}^\top \boldsymbol{B}\boldsymbol{x}}{\partial \boldsymbol{x}}$
- Letting $\boldsymbol{X} \in \mathbb{R}^{N \times d}$ and $\boldsymbol{S} = \boldsymbol{X}^\top \boldsymbol{X}$,
  - show that $\boldsymbol{S}$ is a symmetric matrix,
  - show that $\boldsymbol{x}^\top \boldsymbol{S}\boldsymbol{x} \geq 0$ for any $\boldsymbol{x} \in \mathbb{R}^d$.