Machine Learning: Optimization 2

Hiroshi Shimodaira and Hao Tang

2025

Ver. 1.0

Topics

- Optimal solution / minimiser
- Convex functions and strictly convex functions
- Optimality condition
- Positive semi-definite and positive definite matrix

- For mean-squared error

$$L = \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{w}^\top \boldsymbol{x}_i - y_i)^2 = \frac{1}{N}\|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2, \tag{1}$$

  we know that

$$\boldsymbol{w}^* = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y} \tag{2}$$

  is the solution of $\nabla_{\boldsymbol{w}} L = \mathbf{0}$.

- How do we know $\boldsymbol{w}^*$ is the optimal point?

- For log loss

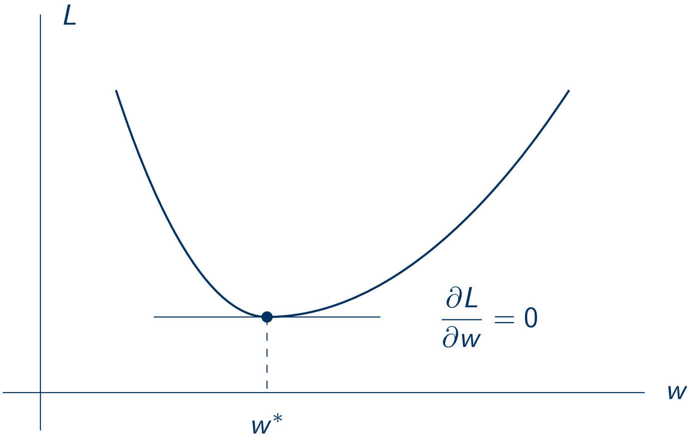
$$\text{NLL} = \sum_{i=1}^{N} \log \left(1 + \exp(-y_i \boldsymbol{w}^\top \phi(\boldsymbol{x}_i)) \right) \tag{3}$$

  we cannot even solve $\nabla_{\boldsymbol{w}} L = \mathbf{0}$.

- How do we find the optimal solution?

- Could we find an approximate solution?

Convex optimisation



Optimisation

- Suppose $f : \mathbb{R}^d \to \mathbb{R}$.
- The goal is solve

$$\min_{\boldsymbol{x}} f(\boldsymbol{x}). \tag{4}$$

Optimisation

- Suppose $f : \mathbb{R}^d \to \mathbb{R}$.
- The goal is solve

$$\min_{\boldsymbol{x}} f(\boldsymbol{x}). \tag{4}$$

- It means $\min_{\boldsymbol{x}} f(\boldsymbol{x}) \leq f(\boldsymbol{y})$ for any $\boldsymbol{y}$.

Optimisation

- Suppose $f : \mathbb{R}^d \to \mathbb{R}$.
- The goal is solve

$$\min_{\boldsymbol{x}} f(\boldsymbol{x}). \tag{4}$$

- It means $\min_{\boldsymbol{x}} f(\boldsymbol{x}) \leq f(\boldsymbol{y})$ for any $\boldsymbol{y}$.
- We want to find $\boldsymbol{x}^*$ such that $f(\boldsymbol{x}^*) = \min_{\boldsymbol{x}} f(\boldsymbol{x})$.
- The point x^* is called the **optimal solution** or the **minimiser** of f.

Optimisation

- Suppose $f : \mathbb{R}^d \to \mathbb{R}$.
- The goal is solve

$$\min_{\boldsymbol{x}} f(\boldsymbol{x}). \tag{4}$$

- It means $\min_{\boldsymbol{x}} f(\boldsymbol{x}) \leq f(\boldsymbol{y})$ for any $\boldsymbol{y}$.
- We want to find $\boldsymbol{x}^*$ such that $f(\boldsymbol{x}^*) = \min_{\boldsymbol{x}} f(\boldsymbol{x})$.
- The point x^* is called the **optimal solution** or the **minimiser** of f.
- There might not be a minimiser or there might have many, not just one. (In most case, we are content with finding one.)

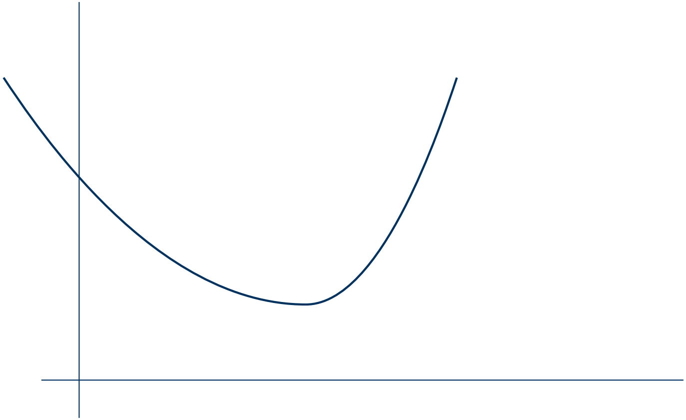
Global vs local minimum / optimal

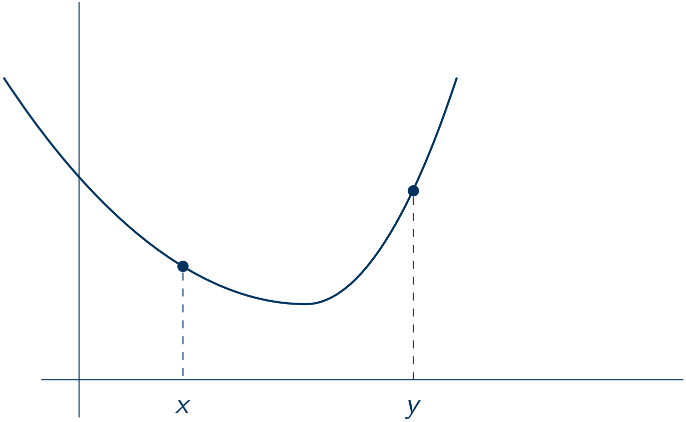
Convex functions

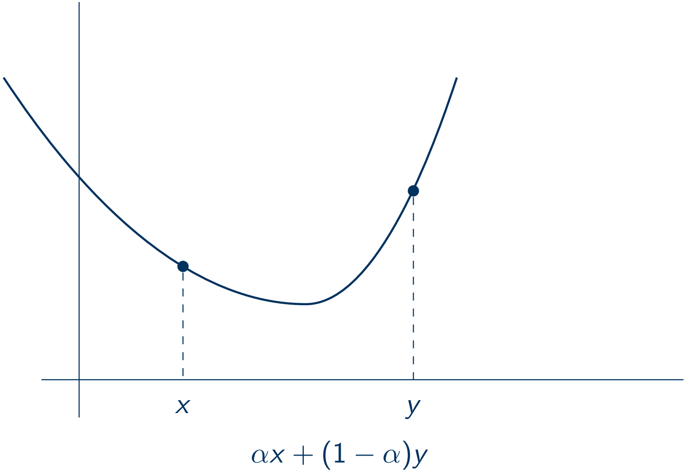
A function f is **convex** if

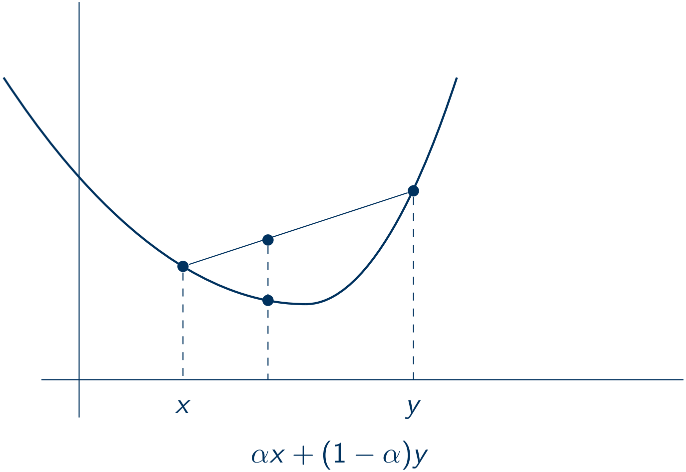
$$f(\alpha \boldsymbol{x} + (1-\alpha)\boldsymbol{y}) \leq \alpha f(\boldsymbol{x}) + (1-\alpha) f(\boldsymbol{y}), \tag{5}$$

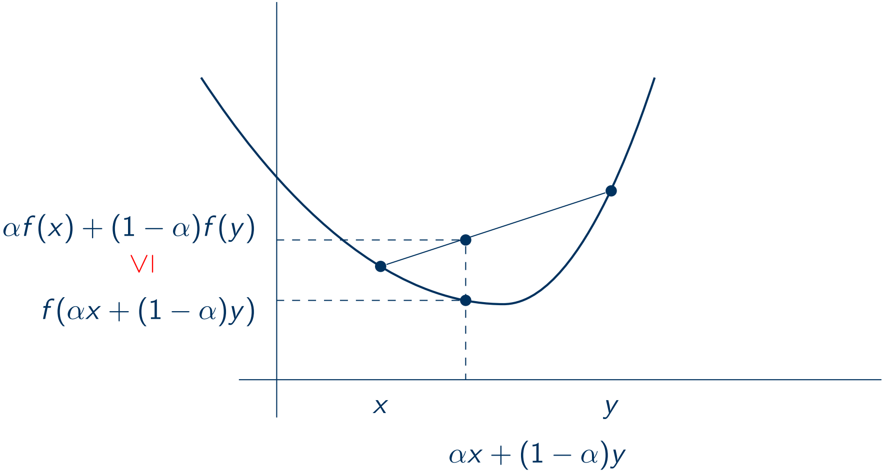
for every $\boldsymbol{x}$, $\boldsymbol{y}$, and $0 \leq \alpha \leq 1$.




x
y


x
y
$\alpha x + (1 - \alpha) y$


x
y
$\alpha x + (1 - \alpha) y$



$\alpha f(x) + (1-\alpha) f(y)$
$\geq$
$f(\alpha x + (1-\alpha)y)$
x
y
$\alpha x + (1-\alpha)y$

Properties of convex functions

If f is convex, then

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}), \tag{6}$$

for any $\boldsymbol{x}$ and $\boldsymbol{y}$.

Properties of convex functions

If f is convex, then

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}), \tag{6}$$

for any $\boldsymbol{x}$ and $\boldsymbol{y}$.

Proof: considering a dimension

$$f((1-\alpha)x + \alpha y) \leq (1-\alpha)f(x) + \alpha f(y) \tag{7}$$

Properties of convex functions

If f is convex, then

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}), \tag{6}$$

for any $\boldsymbol{x}$ and $\boldsymbol{y}$.

Proof: considering a dimension

$$f((1-\alpha)x + \alpha y) \leq (1-\alpha) f(x) + \alpha f(y) \tag{7}$$

$$\alpha f(y) \geq f(x + \alpha(y - x)) - (1-\alpha) f(x) \tag{8}$$

Properties of convex functions

If f is convex, then

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}), \tag{6}$$

for any $\boldsymbol{x}$ and $\boldsymbol{y}$.

Proof: considering a dimension

$$f((1-\alpha)x + \alpha y) \leq (1-\alpha)f(x) + \alpha f(y) \tag{7}$$

$$\alpha f(y) \geq f(x + \alpha(y - x)) - (1-\alpha)f(x) \tag{8}$$

$$f(y) \geq f(x) + \frac{f(x + \alpha(y-x)) - f(x)}{\alpha} \tag{9}$$

Properties of convex functions

If f is convex, then

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}), \tag{6}$$

for any $\boldsymbol{x}$ and $\boldsymbol{y}$.

Proof: considering a dimension

$$f((1-\alpha)x + \alpha y) \leq (1-\alpha) f(x) + \alpha f(y) \tag{7}$$

$$\alpha f(y) \geq f(x + \alpha(y - x)) - (1-\alpha) f(x) \tag{8}$$

$$f(y) \geq f(x) + \frac{f(x + \alpha(y-x)) - f(x)}{\alpha} \tag{9}$$

$$f(y) \geq f(x) + \frac{f(x+h) - f(x)}{h}(y - x) \quad \text{where } h = \alpha(y-x) \tag{10}$$

Properties of convex functions

If f is convex, then

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}), \tag{6}$$

for any $\boldsymbol{x}$ and $\boldsymbol{y}$.

Proof: considering a dimension

$$f((1-\alpha)x + \alpha y) \leq (1-\alpha) f(x) + \alpha f(y) \tag{7}$$

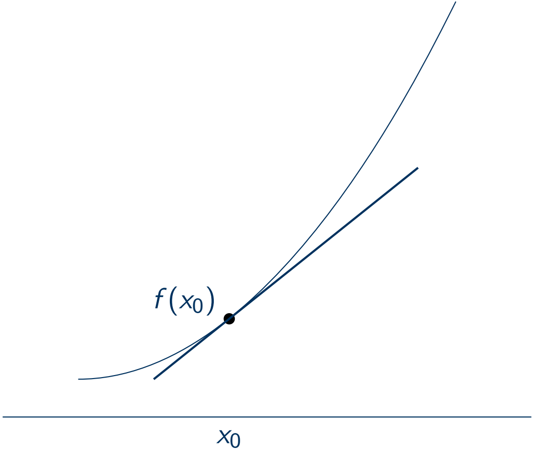
$$\alpha f(y) \geq f(x + \alpha(y - x)) - (1-\alpha) f(x) \tag{8}$$

$$f(y) \geq f(x) + \frac{f(x + \alpha(y - x)) - f(x)}{\alpha} \tag{9}$$

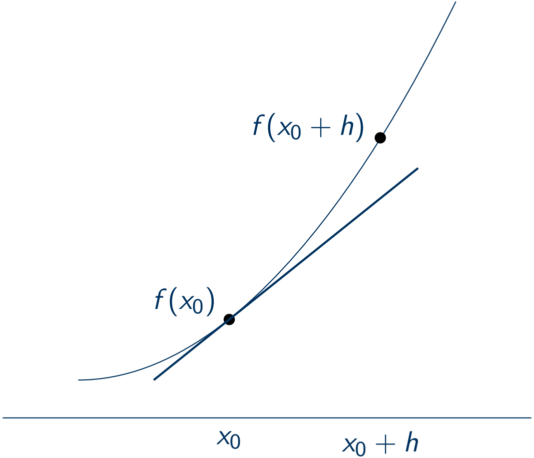
$$f(y) \geq f(x) + \frac{f(x+h) - f(x)}{h}(y - x) \quad \text{where } h = \alpha(y - x) \tag{10}$$

$$f(y) \geq f(x) + f'(x)(y - x) \qquad h \to 0 \tag{11}$$

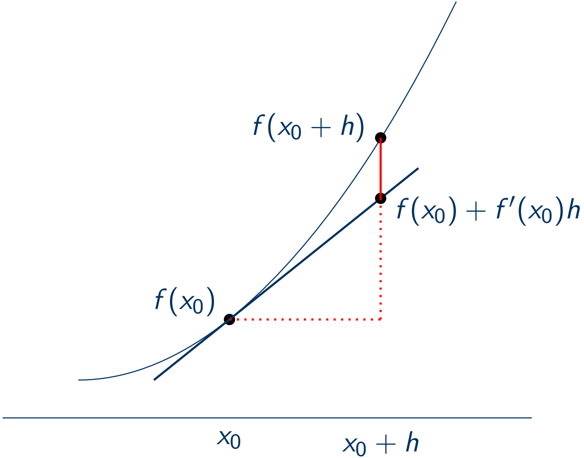
Properties of convex functions *(cont.)*



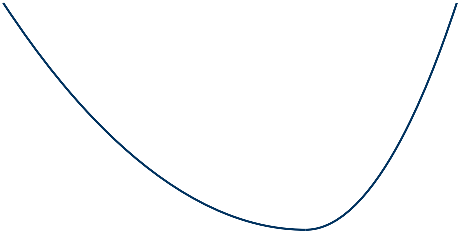
Properties of convex functions *(cont.)*



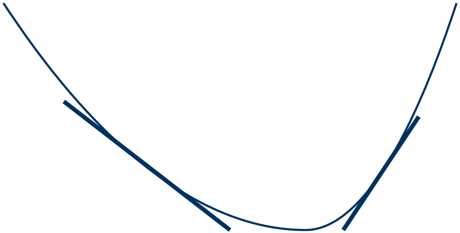
Properties of convex functions *(cont.)*



Supporting hyperplanes



Supporting hyperplanes



- Is the mean-squared error

$$L = \frac{1}{N}\|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2 \tag{12}$$

  convex in $\boldsymbol{w}$?

- The definition itself is not always easy to use for checking convexity.

A sufficient condition: Second derivative

- Suppose $f(\boldsymbol{x})$ is twice differentiable for any $\boldsymbol{x}$.
- $f(\boldsymbol{x})$ is convex iff the Hessian $\mathbf{H} = \nabla^2 f(\boldsymbol{x})$ is positive semi definite for any $\boldsymbol{x}$.

$$\mathbf{H} = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_d} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_d} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_d \partial x_1} & \frac{\partial^2 f}{\partial x_d \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_d^2} \end{bmatrix} \tag{13}$$

- A matrix $\boldsymbol{H}$ is positive semi definite if $\boldsymbol{x}^\top \boldsymbol{H}\boldsymbol{x} \geq \mathbf{0}$ for any $\boldsymbol{x}$.

Convexity of squared distance

- The squared distance $\ell(s) = (s - s')^2$ is convex in s.

Convexity of squared distance

- The squared distance $\ell(s) = (s - s')^2$ is convex in s.

$$\frac{\partial^2 \ell}{\partial s^2} = 2 \geq 0 \tag{14}$$

Convexity of the ℓ_2 norm

- Show that $f(\boldsymbol{x}) = \|\boldsymbol{x}\|_2^2 = \boldsymbol{x}^\top \boldsymbol{x}$ is convex in $\boldsymbol{x}$.

Convexity of the ℓ_2 norm

- Show that $f(\boldsymbol{x}) = \|\boldsymbol{x}\|_2^2 = \boldsymbol{x}^\top \boldsymbol{x}$ is convex in $\boldsymbol{x}$.

$$\frac{\partial^2 \ell}{\partial x_i \partial x_j} = 0 \qquad \frac{\partial^2 \ell}{\partial x_i^2} = 2 \tag{15}$$

Affine transform preserves convexity

- If f is convex, then $g(\boldsymbol{x}) = f(\mathbf{A}\boldsymbol{x} + b)$ is also convex.

Affine transform preserves convexity

- If f is convex, then $g(\boldsymbol{x}) = f(\mathbf{A}\boldsymbol{x} + b)$ is also convex.

$$g(\alpha \boldsymbol{x} + (1-\alpha)\boldsymbol{y}) = f(\alpha(\mathbf{A}\boldsymbol{x} + b) + (1-\alpha)(\mathbf{A}\boldsymbol{y} + b)) \tag{16}$$

Affine transform preserves convexity

- If f is convex, then $g(\boldsymbol{x}) = f(\mathbf{A}\boldsymbol{x} + b)$ is also convex.

$$g(\alpha\boldsymbol{x} + (1-\alpha)\boldsymbol{y}) = f(\alpha(\mathbf{A}\boldsymbol{x} + b) + (1-\alpha)(\mathbf{A}\boldsymbol{y} + b)) \quad (16)$$

$$\leq \alpha f(\mathbf{A}\boldsymbol{x} + b) + (1-\alpha) f(\mathbf{A}\boldsymbol{y} + b) = \alpha g(\boldsymbol{x}) + (1-\alpha) g(\boldsymbol{y}) \quad (17)$$

Non-negative weighted sum of convex functions

- If $f_1, \ldots, f_k$ are convex, then $f = \beta_1 f_1 + \cdots + \beta_k f_k$ is also convex when $\beta_1, \ldots, \beta_k \geq 0$

Non-negative weighted sum of convex functions

- If $f_1, \ldots, f_k$ are convex, then $f = \beta_1 f_1 + \cdots + \beta_k f_k$ is also convex when $\beta_1, \ldots, \beta_k \geq 0$

$$
\begin{aligned}
f(\alpha \boldsymbol{x} + (1-\alpha)\boldsymbol{y}) &= \beta_1 f_1(\alpha \boldsymbol{x} + (1-\alpha)\boldsymbol{y}) + \cdots + \beta_k f_k(\alpha \boldsymbol{x} + (1-\alpha)\boldsymbol{y}) && (18)\\
&\leq \beta_1 \alpha f_1(\boldsymbol{x}) + \beta_1 (1-\alpha) f(\boldsymbol{y}) + \cdots + \beta_k \alpha f_k(\boldsymbol{x}) + \beta_k (1-\alpha) f_k(\boldsymbol{y}) && (19)\\
&= \alpha(\beta_1 f_1(\boldsymbol{x}) + \cdots + \beta_k f_k(\boldsymbol{x})) + (1-\alpha)(\beta_1 f_1(\boldsymbol{y}) + \cdots + \beta_k f_k(\boldsymbol{y})) && (20)\\
&= \alpha f(\boldsymbol{x}) + (1-\alpha) f(\boldsymbol{y}) && (21)
\end{aligned}
$$

Convexity of MSE

- The mean-squared error is

$$L = \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{w}^\top \boldsymbol{x}_i - y_i)^2 = \frac{1}{N}\|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2. \tag{22}$$

- We know that the squared distance is convex.
- Use the affine transform and non-negative weighted sum to obtain the mean-squared error.

Optimality condition

If f is convex and

$$\nabla f(\boldsymbol{x}^*) = \mathbf{0} \tag{23}$$

at $\boldsymbol{x}^*$, then $\boldsymbol{x}^*$ is the minimiser of f.

Optimality condition

If f is convex and

$$\nabla f(\boldsymbol{x}^*) = \mathbf{0} \tag{23}$$

at $\boldsymbol{x}^*$, then $\boldsymbol{x}^*$ is the minimiser of f.

Proof: Suppose $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$. For any $\boldsymbol{x}$,

$$f(\boldsymbol{x}) \geq f(\boldsymbol{x}^*) + \nabla f(\boldsymbol{x}^*) \top (\boldsymbol{x} - \boldsymbol{x}^*) = f(\boldsymbol{x}^*). \tag{24}$$

Optimal solution of MSE

- The mean-squared error is

$$L = \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{w}^\top \phi(\boldsymbol{x}_i) - y_i)^2 = \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2. \tag{25}$$

- The solution to $\nabla_{\boldsymbol{w}} L = \mathbf{0}$ is $\boldsymbol{w}^* = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y}$.

- Because L is convex in $\boldsymbol{w}$, $\boldsymbol{w}^*$ is a minimiser of L.

Convexity of log loss in logistic regression

- The log loss in the binary case is

$$L = \sum_{i=1}^{N} \log\left(1 + \exp(-y_i \boldsymbol{w}^\top \boldsymbol{x}_i)\right). \tag{26}$$

- We just need to show $\ell(s) = \log(1 + \exp(-s))$ is convex in s.

- Use affine transform and non-negative weighted sum to obtain the log loss.

$$\frac{\partial \ell}{\partial s} = \frac{-\exp(-s)}{1+\exp(-s)} = \frac{1}{1+\exp(-s)} - 1 \tag{27}$$

$$\frac{\partial^2 \ell}{\partial s^2} = \frac{1}{1+\exp(-s)} \frac{\exp(-s)}{1+\exp(-s)} = \frac{1}{1+\exp(-s)} \left(1 - \frac{1}{1+\exp(-s)}\right) \geq 0 \tag{28}$$

Strictly convex functions

A function f is **strictly convex** if

$$f(\alpha \boldsymbol{x} + (1-\alpha)\boldsymbol{y}) < \alpha f(\boldsymbol{x}) + (1-\alpha) f(\boldsymbol{y}), \tag{29}$$

for every $\boldsymbol{x} \neq \boldsymbol{y}$, and $0 \leq \alpha \leq 1$.

Properties of strictly convex functions

- If f is strictly convex, then

$$f(\boldsymbol{x}) > f(\boldsymbol{y}) + \nabla f(\boldsymbol{y})^\top(\boldsymbol{x} - \boldsymbol{y}), \tag{30}$$

  for any $\boldsymbol{x} \neq \boldsymbol{y}$.

Properties of strictly convex functions

- If f is strictly convex, then

$$f(\boldsymbol{x}) > f(\boldsymbol{y}) + \nabla f(\boldsymbol{y})^\top (\boldsymbol{x} - \boldsymbol{y}), \tag{30}$$

  for any $\boldsymbol{x} \neq \boldsymbol{y}$.

- A matrix $\boldsymbol{H}$ is positive definite if $\boldsymbol{x}^\top \boldsymbol{H}\boldsymbol{x} > 0$ for any $\boldsymbol{x} \neq \boldsymbol{0}$.

- If the Hessian of f is positive definite, then f is strictly convex.

Uniqueness of minimisers for strictly convex functions

A strictly convex function f has a unique minimiser.

Uniqueness of minimisers for strictly convex functions

A strictly convex function f has a unique minimiser.

Proof: Suppose x^* is a minimiser of f, i.e., $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$. Since f is strictly convex,

$$f(\boldsymbol{x}) > f(\boldsymbol{y}) + \nabla f(\boldsymbol{y})^\top (\boldsymbol{x} - \boldsymbol{y}) \tag{31}$$

for any $\boldsymbol{x} \neq \boldsymbol{y}$. In particular, if we let $\boldsymbol{y} = \boldsymbol{x}^*$

Quizzes

- Show the convexity for the following functions.
  - $f(x) = x^2$
  - $f(x) = |x|^p$ for $p \geq 1$
  - $f(x) = \exp(ax)$
  - $f(x) = x \log x$
  - $f(x, y) = \log(e^x + e^y)$
- Find the condition(s) under which the following function $f(\boldsymbol{x})$ is convex in $\boldsymbol{x}$.

$$f(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}^\top \boldsymbol{x} + c$$

- Consider a function $f(x) = \frac{1}{x^2}$.
  - Find the first and second derivatives.
  - Discuss the convexity of the function.