Machine Learning

Lecture 7: Optimization 1

Hao Tang

October 6, 2022

- For mean-squared error

$$L = \frac{1}{N} \sum_{i=1}^{N} (w^\top \phi(x_i) - y_i)^2, \tag{1}$$

  we know that

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

  is the solution of $\frac{\partial L}{\partial w} = 0$.

- How do we know w^* is the optimal point?

- For log loss

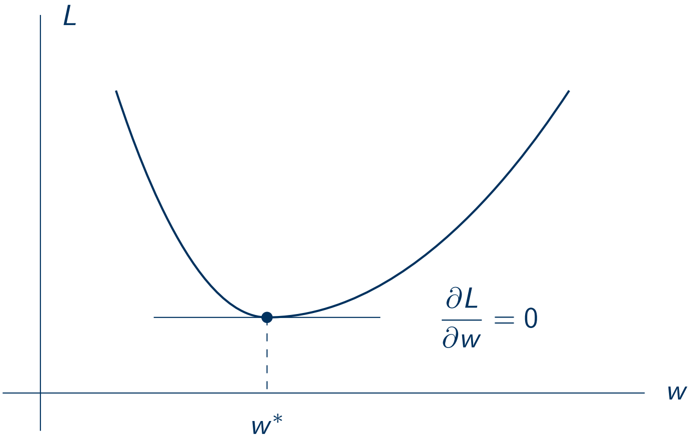
$$L = \sum_{i=1}^{N} \log \left(1 + \exp(-y_i w^\top \phi(x_i)) \right) \tag{3}$$

we cannot even solve $\frac{\partial L}{\partial w} = 0$.

- How do we find the optimal solution?

- Could we find an approximate solution?

Convex optimization



Optimization

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

$$\min_x f(x). \tag{4}$$

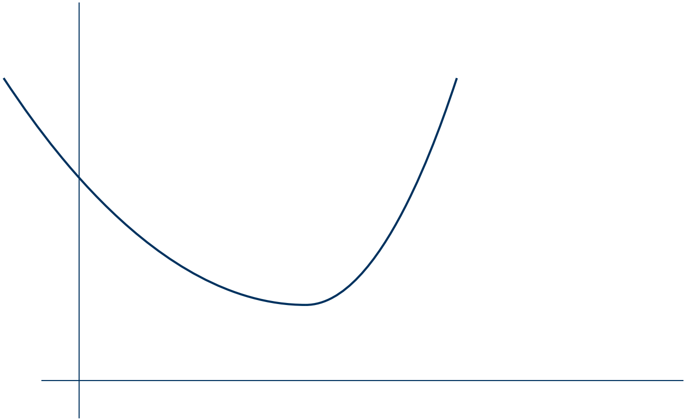
- Note $\min_x f(x) \leq f(y)$ for any y.
- We want to find x^* such that $f(x^*) = \min_x f(x)$.
- The point x^* is called the **optimal solution** or the **minimizer** of f.
- There might not be a minimizer or there might have many, not just one. (In most case, we are content with finding one.)

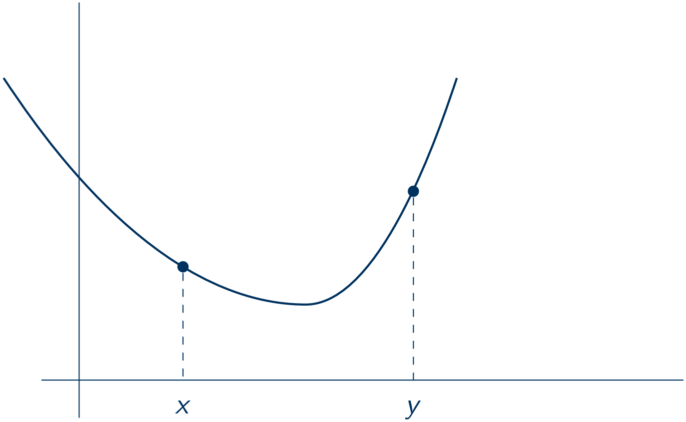
Convex functions

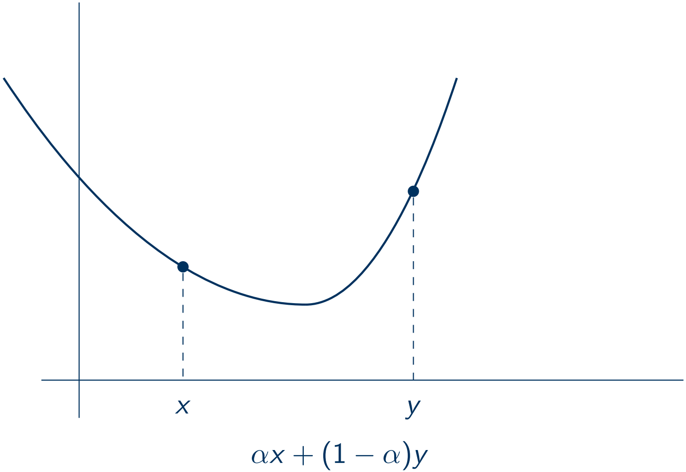
A function f is **convex** if

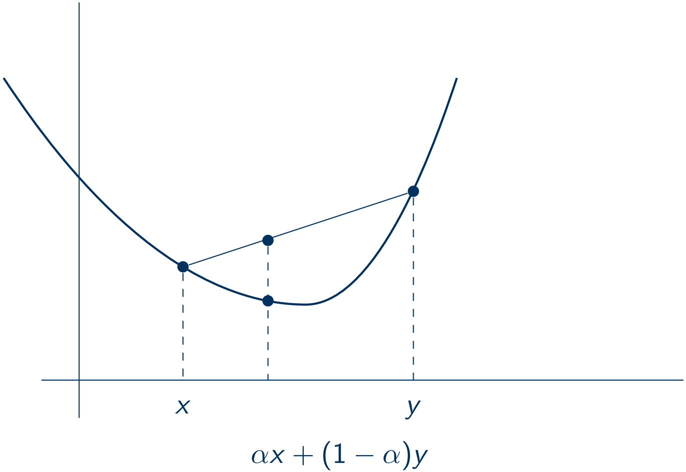
$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y), \tag{5}$$

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

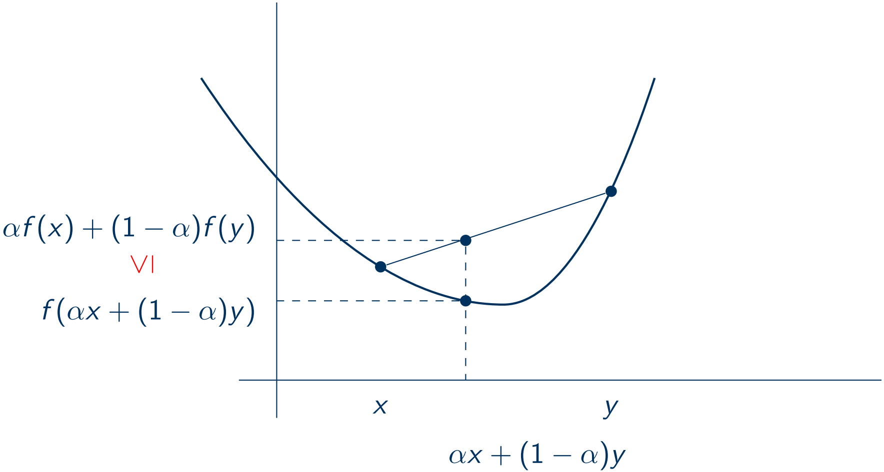



x
y


x
y
αx + (1 − α)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

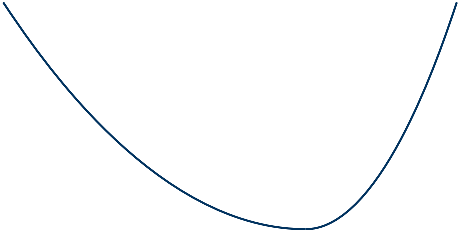
$$f(x) \geq f(y) + (x-y)^\top \nabla f(y), \tag{6}$$

for any x and y.

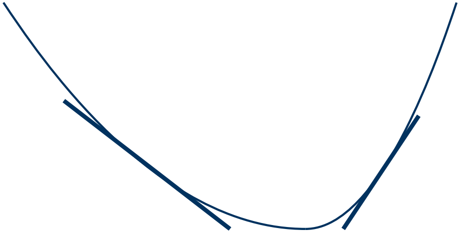
Proof:

$$\begin{aligned} f(\alpha x + (1-\alpha)y) &\leq \alpha f(x) + (1-\alpha) f(y) \\ \alpha f(y) + f(y + \alpha(x-y)) - f(y) &\leq \alpha f(x) \\ f(y) + \frac{f(y + \alpha(x-y)) - f(y)}{\alpha} &\leq f(x) \\ f(y) + (x-y)^\top \nabla f(y) &\leq f(x) \end{aligned}$$

Supporting hyperplanes



Supporting hyperplanes



- Is the mean-squared error

$$L = \frac{1}{N} \sum_{i=1}^{N} (w^\top \phi(x_i) - y_i)^2 \tag{7}$$

  convex in w?

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

A sufficient condition: Second derivative

- If $\nabla^2 f(x)$ exists and $\nabla^2 f(x) \succeq 0$ for all x, then f is convex.
- When we write $\nabla^2 f(x) \succeq 0$, we say that $\nabla^2 f(x)$ is positive semi-definite.
- A matrix H is positive semi-definite, if $v^\top H v \geq 0$ for every v.

Convexity of squared distance

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

$$\frac{\partial^2 \ell}{\partial s^2} = 2 \geq 0 \tag{8}$$

Affine transform preserves convexity

- If f is convex, then $g(x) = f(Ax + b)$ is also convex.

$$
\begin{aligned}
g(\alpha x + (1-\alpha)y) &= f(\alpha(Ax + b) + (1-\alpha)(Ay + b)) && (9)\\
&\leq \alpha f(Ax + b) + (1-\alpha)f(Ay + b) = \alpha g(x) + (1-\alpha)g(y) && (10)
\end{aligned}
$$

Nonnegative 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 $\beta_1, \ldots, \beta_k \geq 0$

$$
\begin{aligned}
f(\alpha x + (1-\alpha)y) &= \beta_1 f_1(\alpha x + (1-\alpha)y) + \cdots + \beta_k f_k(\alpha x + (1-\alpha)y) && (11)\\
&\leq \beta_1 \alpha f_1(x) + \beta_1(1-\alpha)f(y) + \cdots + \beta_k \alpha f_k(x) + \beta_k(1-\alpha)f_k(y) && (12)\\
&= \alpha(\beta_1 f_1(x) + \cdots + \beta_k f_k(x)) + (1-\alpha)(\beta_1 f_1(y) + \cdots + \beta_k f_k(y)) && (13)\\
&= \alpha f(x) + (1-\alpha)f(y) && (14)
\end{aligned}
$$

Convexity of MSE

- The mean-squared error is

$$L = \frac{1}{N}\sum_{i=1}^{N}(w^\top \phi(x_i) - y_i)^2. \tag{15}$$

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

Optimality condition

If f is convex and

$$\nabla f(x^*) = 0 \tag{16}$$

at x^*, then x^* is the minimizer of f.

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

$$f(x) \geq f(x^*) + (x - x^*)^\top \nabla f(x^*) = f(x^*). \tag{17}$$

Optimal solution of MSE

- The mean-squared error is

$$L = \frac{1}{N}\sum_{i=1}^{N}(w^\top \phi(x_i) - y_i)^2. \tag{18}$$

- The solution to $\frac{\partial L}{\partial w} = 0$ is $w^* = (\Phi\Phi^\top)^{-1}\Phi y$.

- Because L is convex in w, w^* is the global minimum.

Convexity of log loss

- The log loss in the binary case is

$$L = \sum_{i=1}^{N} \log \left(1 + \exp(-y_i w^\top \phi(x_i)) \right). \tag{19}$$

- We just need to show $\ell(s) = \log(1 + \exp(-s))$ is convex in s.
- Use affine transform and nonnegative weighted sum to obtain the log loss.

$$\frac{\partial \ell}{\partial s} = \frac{-\exp(-s)}{1+\exp(-s)} = \frac{1}{1+\exp(-s)} - 1 \tag{20}$$

$$\frac{\partial^2 \ell}{\partial s^2} = \frac{1}{1+\exp(-s)} \frac{\exp(-s)}{1+\exp(-s)} = \frac{1}{1+\exp(-s)} \left(1 - \frac{1}{1+\exp(-s)}\right) \geq 0 \tag{21}$$

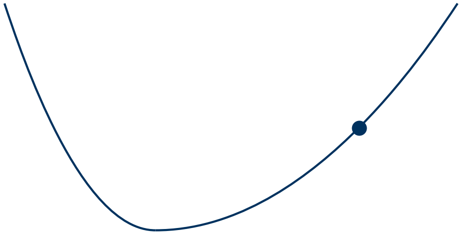
Strong convexity

- A function f is μ-strongly convex if

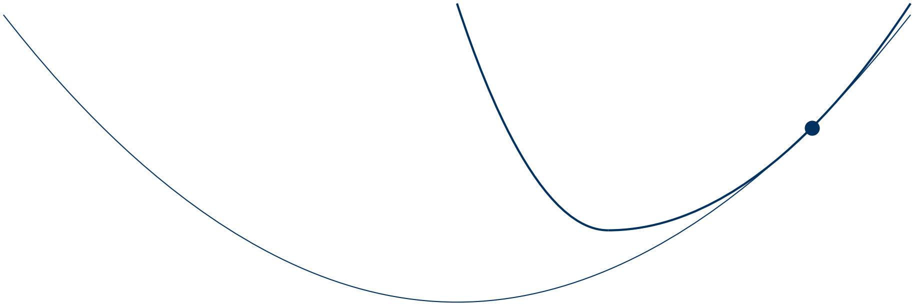
$$f(y) \geq f(x) + (y - x)^\top \nabla f(x) + \frac{\mu}{2}\|y - x\|^2 \qquad (22)$$

for any x and y.

Quadratic lower bound



Quadratic lower bound



Lipschitz continuous

- A function is L-Lipschitz if

$$|f(x) - f(y)| \leq L\|x - y\| \tag{23}$$

  for any x and y.

- In words, the function values can only change so much for points that are close.

Smoothness

- When the gradient of f is L-Lipschitz, then we say that f is L-smooth.
- In other words, f is L-smooth if

$$\|\nabla f(y) - \nabla f(x)\| \leq L\|y - x\| \tag{24}$$

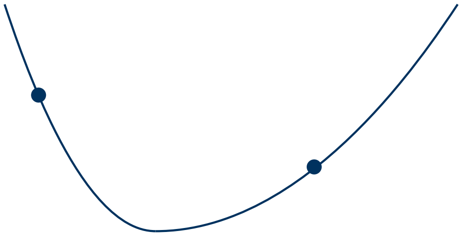
  for any x and y.

- L-smoothness also implies

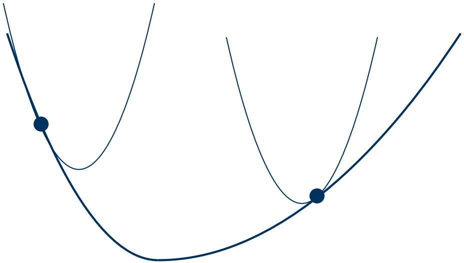
$$f(y) \leq f(x) + (y - x)^\top \nabla f(x) + L\|x - y\|_2^2. \tag{25}$$

$$
\begin{aligned}
f(y) - f(x) - (y-x)^\top \nabla f(x) & \qquad (26)\\
\leq \nabla f(y)^\top (y-x) - \nabla f(x)^\top (y-x) & \qquad (27)\\
\leq (\nabla f(y) - \nabla f(x))^\top (y-x) & \qquad (28)\\
\leq \|\nabla f(y) - \nabla f(x)\| \|y-x\| & \qquad (29)\\
\leq L\|y-x\|^2 & \qquad (30)
\end{aligned}
$$

Quadratic upper bound



Quadratic upper bound



Check your understanding

- What is the definition of convex functions?
- Can you show that a convex function is supported by hyperplanes everywhere?
- Can you show that mean-squared error is convex in w?
- Can you show that log loss is convex in w?
- How does a function being convex help us do optimization?
- What are strongly convex functions
- What are Lipschitz continuous functions?
- What are Lipschitz smooth functions?