Machine Learning: Optimization 1

Hao Tang

January 30, 2026

Recap: Maximum likelihood estimation of the Gaussian mean

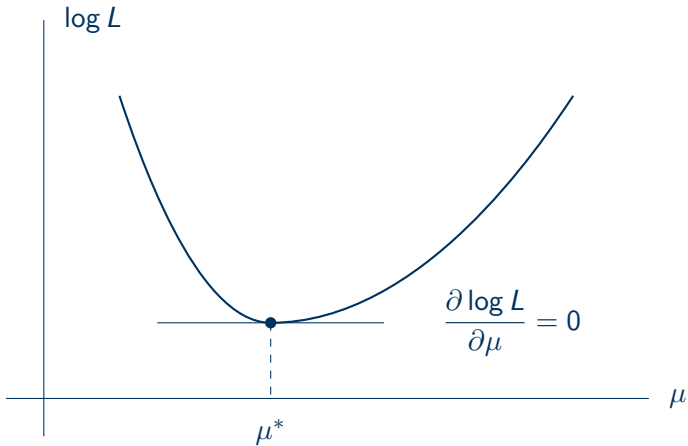
- We find the maximum solution by solving

$$\frac{\partial}{\partial \mu} \log L(\mu) = 0, \tag{1}$$

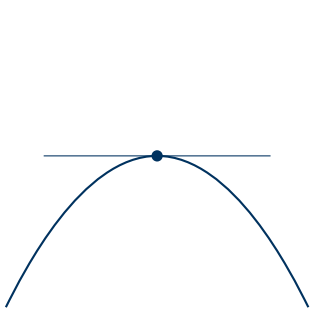
  where $L(\mu)$ is the likelihood function.

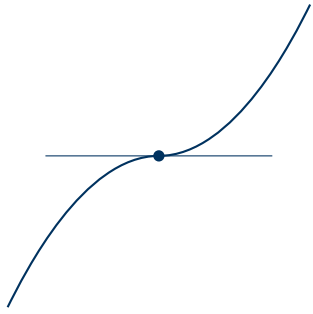
- Why does this work? When does this work?

An ideal scenario



Not so ideal scenarios





Questions

- Points with derivative 0 can be a maximal, a minimal, or a saddle point.
- The "nice" functions are the ones where the points with derivative 0 are all minimal solutions.
- How do we know our functions are "nice"?
- How do we exactly characterize these "nice" functions?
- Convex functions are a family of "nice" functions that we are looking for.

Optimization

- We are mostly interested in functions of the type $\mathbb{R}^d \to \mathbb{R}$.
- The goal is solve

$$\min_x f(x). \tag{2}$$

Optimization

- We are mostly interested in functions of the type $\mathbb{R}^d \to \mathbb{R}$.
- The goal is solve

$$\min_x f(x). \tag{2}$$

- The term $\min_x f(x)$ is a value, and it means

$$\min_x f(x) \leq f(y) \qquad \text{for any } y. \tag{3}$$

Optimization

- We want to find x^* such that $f(x^*) = \min_x f(x)$.
- The point x^* is called an **optimal solution** or a **minimizer** of f.
- For some functions, there might be many minimizers. (In most cases, we are content with finding one.)
- For some functions, a minimizer might not even exist.

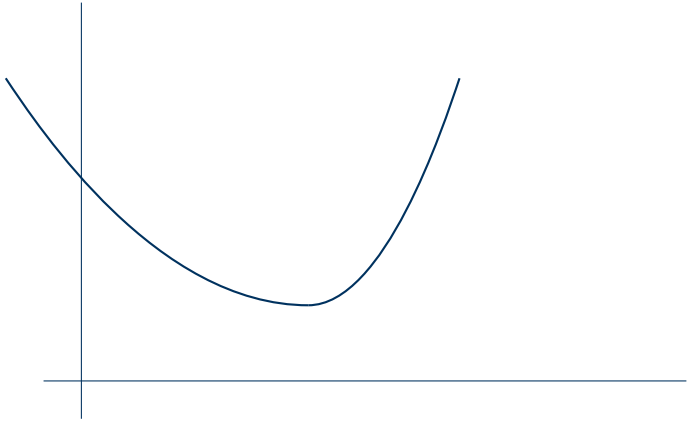
Convex functions

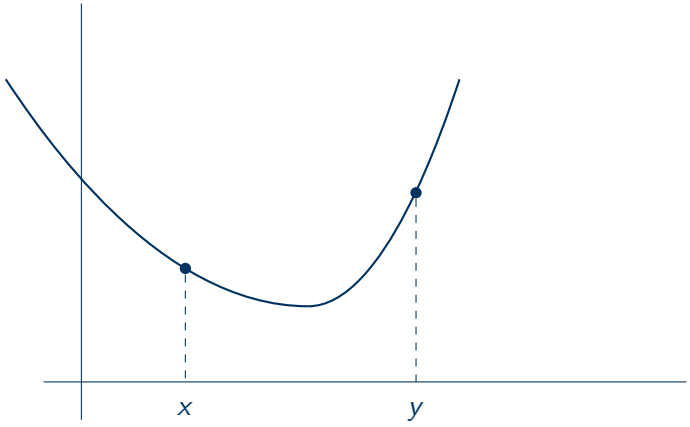
- A function f is **convex** if

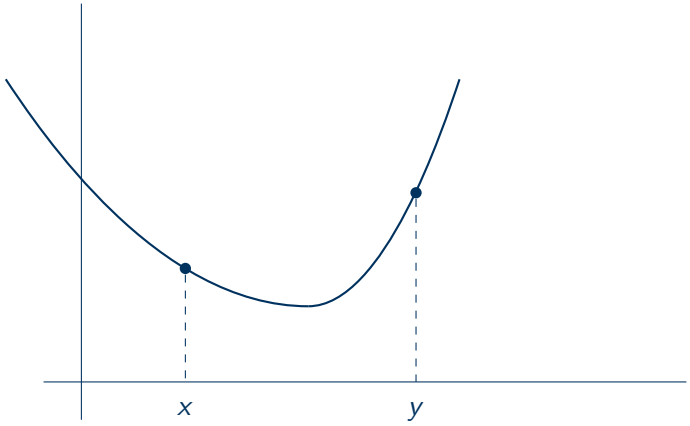
$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y), \tag{4}$$

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

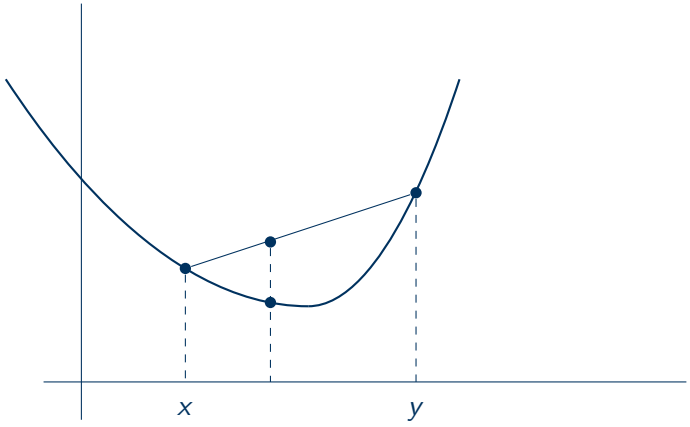
- A function f is **concave** if $-f$ is convex.



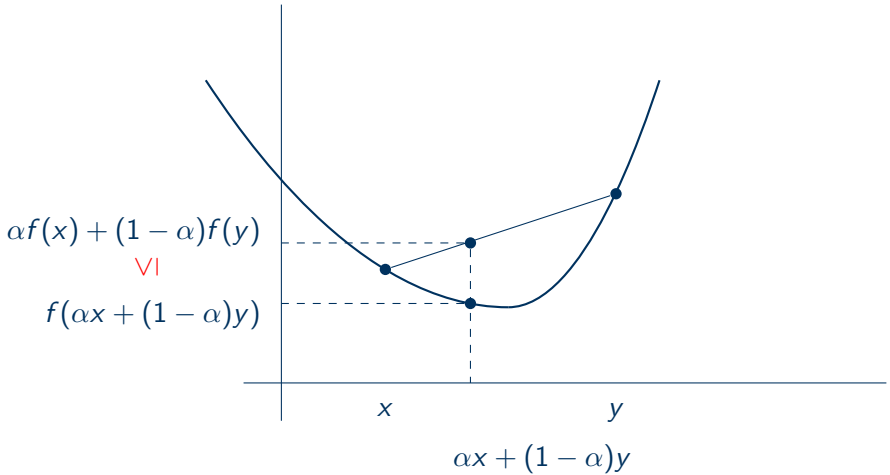

x
y



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



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


$\alpha f(x) + (1-\alpha) f(y)$
$\vee$|
$f(\alpha x + (1-\alpha) y)$
x
y
$\alpha x + (1-\alpha) y$

Affine functions are both convex and concave

- The function $f(x) = w^\top x + b$ is both convex and concave.
- Proof:

$$
\begin{aligned}
f(\alpha x + (1-\alpha y)) &= w^\top(\alpha x + (1-\alpha y)) + b && (5)\\
&= \alpha(w^\top x + b) + (1-\alpha)(w^\top y + b) && (6)\\
&= \alpha f(x) + (1-\alpha) f(y) && (7)
\end{aligned}
$$

A convex function is supported by hyperplanes

If f is convex, then

$$f(x) \geq f(y) + \nabla f(y)^\top (x - y), \tag{8}$$

for any x and y.

Hyperplanes

Hyperplanes

- A hyperplane is a set of points perpendicular to the normal vector.

Hyperplanes

- A hyperplane is a set of points perpendicular to the normal vector.
- In math, a hyperplane is written as

$$\{x : \mathbb{R}^d \mid w^\top (x - u) = 0\}, \tag{9}$$

where u is vector that shifts the plane and w is the normal vector.

Hyperplanes

- A hyperplane is a set of points perpendicular to the normal vector.
- In math, a hyperplane is written as

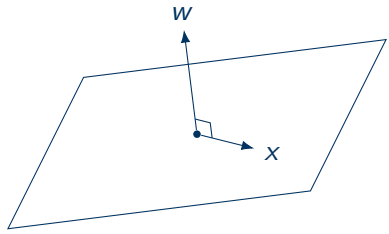
$$\{x : \mathbb{R}^d \mid w^\top (x - u) = 0\}, \tag{9}$$

  where u is vector that shifts the plane and w is the normal vector.
- Alternatively, we can also write a hyperplane as

$$\{x : \mathbb{R}^d \mid w^\top x + b = 0\}, \tag{10}$$

  for some constant b.

Hyperplanes



A convex function is supported by hyperplanes

- For some y, the set of points x that satisfy

$$f(y) + \nabla f(y)^\top (x - y) = 0 \tag{11}$$

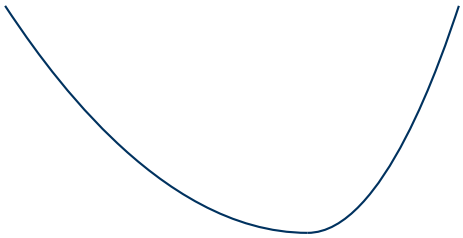
  forms a hyperplane with a normal vector $\nabla f(y)$.

- The function value of a convex function $f(x)$ has to be greater than the hyperplane

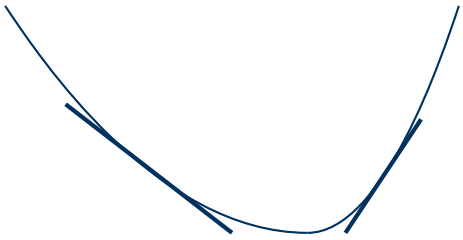
$$f(x) \geq f(y) + \nabla f(y)^\top (x - y). \tag{12}$$

- In other words, a convex function is supported by hyperplanes everywhere.

Supporting hyperplanes



Supporting hyperplanes



A convex function is supported by hyperplanes

- If f is convex, then

$$f(x) \geq f(y) + \nabla f(y)^\top (x - y), \tag{13}$$

  for any x and y.

A convex function is supported by hyperplanes

- If f is convex, then

$$f(x) \geq f(y) + \nabla f(y)^\top (x - y), \tag{13}$$

  for any x and y.

- Proof:

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

Optimality condition

- If f is convex and

$$\nabla f(x^*) = 0 \tag{14}$$

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

Optimality condition

- If f is convex and

$$\nabla f(x^*) = 0 \tag{14}$$

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

- Proof: For any x,

$$f(x) \geq f(x^*) + \nabla f(x^*)^\top (x - x^*) = f(x^*). \tag{15}$$

A definition is not enough

- Is the log likelihood of Gaussian mean convex?

$$\log L(\mu) = \sum_{i=1}^{n} \left[-\frac{d}{2} \log 2\pi - \frac{1}{2} \log |\Sigma| - \frac{1}{2} (x_i - \mu)^\top \Sigma^{-1} (x_i - \mu) \right] \tag{16}$$

- We could check for convexity using the definition but it's not straightforward.

Second-order derivative

- If the Hessian of f exists and is positive semidefinite everywhere, then f is convex.
- If the Hessian of f is not positive semidefinite somewhere, then f is not convex.
- Just as a reminder, a matrix H is positive semidefinite if $v^\top Hv \geq 0$ for any v.
- The proof amounts to reducing to the 1D case by taking slices of the function and showing that the second-order derivative is positive.

Convexity of squared distance

- The squared distance $f(x) = (x - x')^2$ is convex in x.

Convexity of squared distance

- The squared distance $f(x) = (x - x')^2$ is convex in x.
- Proof:

$$\frac{\partial^2 f}{\partial x^2} = 2 \geq 0 \tag{17}$$

Convexity of the ℓ_2 norm

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

Convexity of the ℓ_2 norm

- Show that $f(x) = \|x\|_2^2$ is convex in x.
- Proof:

$$\frac{\partial^2 f}{\partial x_i \partial x_j} = 0 \qquad \frac{\partial^2 f}{\partial x_i^2} = 2 \tag{18}$$

Saddle point

- Show that $f(x, y) = x^2 - y^2$ is not convex.

Saddle point

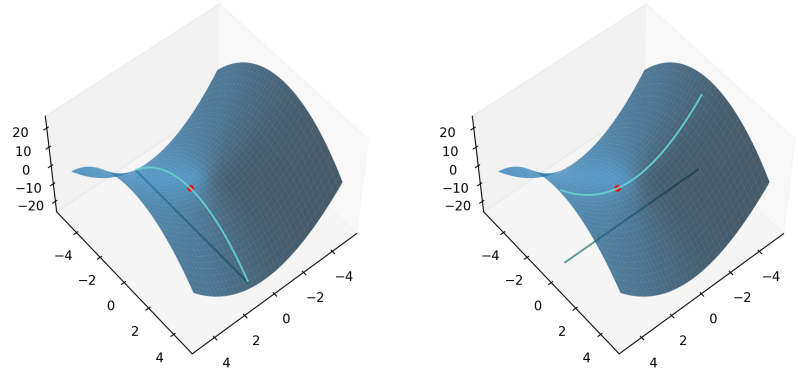
- Show that $f(x, y) = x^2 - y^2$ is not convex.
- The Hessian is $\begin{bmatrix} 2 & 0 \\ 0 & -2 \end{bmatrix}$.

Saddle point

- Show that $f(x, y) = x^2 - y^2$ is not convex.
- The Hessian is $\begin{bmatrix} 2 & 0 \\ 0 & -2 \end{bmatrix}$.
- To show that the Hessian is not positive semidefinite, we just need to find one vector that breaks.

$$\begin{bmatrix} 0 & 1 \end{bmatrix} \begin{bmatrix} 2 & 0 \\ 0 & -2 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} \leq 0 \tag{19}$$

Saddle point



The likelihood of the Gaussian mean

- Is the likelihood of the Gaussian mean convex?

$$\log L(\mu) = \sum_{i=1}^{n} \left[-\frac{d}{2} \log 2\pi - \frac{1}{2} \log |\Sigma| - \frac{1}{2} (x_i - \mu)^\top \Sigma^{-1} (x_i - \mu) \right] \tag{20}$$

- The Hessian of the log likelihood is $-n\Sigma^{-1}$.

- Σ^{-1} is positive semidefinite, so the log likelihood of the Gaussian mean is concave.

The Hessian is still not enough

- Is the log loss convex?

$$L = \sum_{i=1}^{N} \log \left(1 + \exp(-y_i w^\top x_i) \right) \tag{21}$$

- We could derive the Hessian but it's going to be hairy.

Preserving convexity

- How can we compose functions to maintain convexity?
- Once we know that, we can break up a function into smaller functions and only check the convexity of small functions.

Preserving convexity

- Affine transform preserves convexity.

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

- Nonnegative weighted sum preserves convexity.

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

Convexity of log loss

- The log loss in the binary case is

$$L = \sum_{i=1}^{N} \log\left(1 + \exp(-y_i w^\top x_i)\right). \tag{22}$$

- 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.

$$\frac{\partial \ell}{\partial s} = \frac{-\exp(-s)}{1+\exp(-s)} = \frac{1}{1+\exp(-s)} - 1 \tag{23}$$

$$\frac{\partial^2 \ell}{\partial s^2} = \frac{1}{1+\exp(-s)} \frac{\exp(-s)}{1+\exp(-s)} = \frac{1}{1+\exp(-s)} \left(1 - \frac{1}{1+\exp(-s)}\right) \geq 0 \tag{24}$$