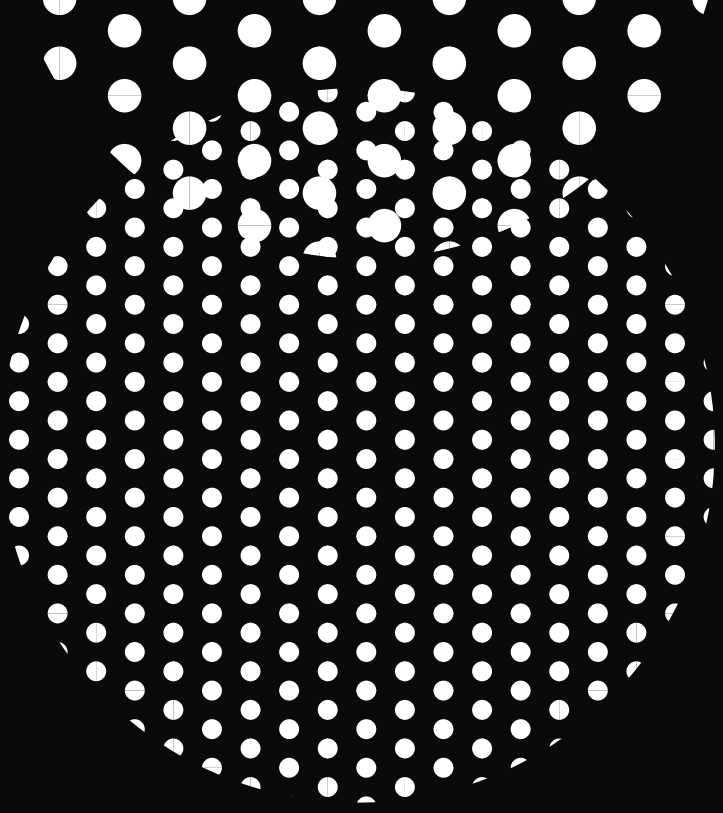


Proving convexity

1) $f(\lambda x + (1-\lambda) y) \leq \lambda f(x) + (1-\lambda) f(y)$ $\quad \forall x,y \in D \ \forall \lambda \in (0,1)$

2) f is 1-diff

$$f(x) \geq f(y) + (y-x)^T \nabla f(y) \quad \forall x,y \in D$$

3) f is 2-diff

$$\nabla^2 f(x) \geq 0 \iff z^T \nabla^2 f(x) z \geq 0 \quad \forall x \in D$$

$\nabla^2 f(x) \geq 0$ def'd

$\Leftrightarrow$

$f(x) \geq 0$

4) Rules

Disproving convexity

1) 2) 3) $\rightarrow$ violated $\Rightarrow$ no convexity

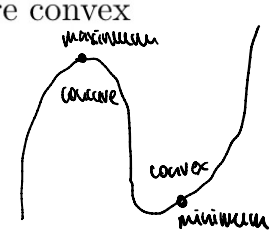
counterexamples

Problem 1: Prove or disprove whether the following functions $f : D \to \mathbb{R}$ are convex

a) $D = (1, \infty)$ and $f(x) = \log(x) - x^3$,

b) $D = \mathbb{R}^+$ and $f(x) = -\min(\log(3x + 1), -x^4 - 3x^2 + 8x - 42)$,

c) $D = (-10, 10) \times (-10, 10)$ and $f(x, y) = y \cdot x^3 - y \cdot x^2 + y^2 + y + 4$.



$f'' < 0 \rightarrow$ concave

$f'' > 0 \rightarrow$ convex

a) The second derivative of f is:

$$\frac{d^2 f(x)}{dx^2} = \frac{d}{dx}\left(\frac{1}{x} - 3x^2\right) = -\frac{1}{x^2} - 6x$$, which is negative on the given set D and therefore f is not convex.

$$-\frac{1}{x^2} - 6x = 0 \; ; \; -\frac{1}{x^2} = 6x \; ; \; 1 = -6x^3 \; ; \; x = \sqrt[3]{-1/6}$$

b) Transform min to max

$$-\min\{\log(3x+1), -x^4 - 3x^2 + 8x - 42\} =$$
$$\max\{-\log(3x+1), x^4 + 3x^2 - 8x + 42\}$$

$\max(g_1(x), g_2(x))$ is convex if both g_1 and g_2 are convex on $D = \mathbb{R}^+$

$g_1(x) = -\log(3x + 1)$ is convex since the second derivative is positive on $\mathbb{R}^+$:

$$\frac{d^2}{dx^2}(-\log(3x+1)) = \frac{d}{dx}\left(-\frac{3}{3x+1}\right) = \frac{9}{(3x+1)^2} > 0$$

$g_2(x) = x^4 + 3x^2 - 8x + 42$ is also convex:

$$\frac{d^2}{dx^2}(x^4 + 3x^2 - 8x + 42) = \frac{d}{dx}(4x^3 + 6x - 8) = 12x^2 + 6 > 0$$

So, f is convex

c) $D = (-10, 10) \times [-10, 10)$ and $f(x,y) = y \cdot x^3 - y \cdot x^2 + y^2 + y + 4$

For the function $f(x,y)$ to be convex (on D) it has to hold for all $x_1, x_2 \in D$ and $\lambda \in (0,1)$ that

$$\lambda f(x_1, y) + (1-\lambda) f(x_2, y) \geq f(\lambda x_1 + (1-\lambda) x_2, y).$$

If does not hold in our case, consider $y = 1$, $x_1 = -4$, $x_2 = 0$, $\lambda = 0.5$:

$$0.5\, f(-4, 1) + 0.5 f(0, 1) = 0.5\,(-74) + 0.5\,(6) = -34$$

$$f[0.5\,(-4) + 0.5 \cdot 0,\ 0.5 \cdot 1 + 0.5 \cdot 1)) = f(-2, 1) = -6 > -34$$

Thus $f(x,y)$ is not convex

Problem 2: Prove that the following function (the loss function of logistic regression) $f : \mathbb{R}^d \to \mathbb{R}$ is convex:

$$f(w) = -\ln p(\boldsymbol{y} \mid \boldsymbol{w}, \boldsymbol{X}) = -\sum_{i=1}^{N} \left(y_i \ln \sigma(\boldsymbol{w}^T \boldsymbol{x}_i) + (1 - y_i) \ln(1 - \sigma(\boldsymbol{w}^T \boldsymbol{x}_i)) \right) .$$

First, let's simplify the above expression

$$\sigma(z) = \frac{1}{1+e^{-z}} = \frac{e^z}{1+e^z} \qquad 1-\sigma(z) = \sigma(-z) = \frac{1}{1+e^z}$$

$$\frac{1}{1+e^{-z}} \cdot \left(\frac{e^z}{e^z}\right) = \frac{e^z}{e^z + e^{-z} \cdot e^z} = \frac{e^z}{1+e^z}$$

which implies that

$$\ln \sigma(z) = \ln\left(\frac{e^z}{1+e^z}\right) = z - \ln(1+e^z)$$

$\ln\left(\frac{a}{b}\right) = \ln(a) - \ln(b)$

$$\ln(1-\sigma(z)) = -\ln(1+e^z)$$

Plugging this into the definition of the loss function we obtain

$$f(w) = -\sum_{i=1}^{N} \left(y_i \ln \sigma(w^T x_i) + (1-y_i) \ln(1-\sigma(w^T x_i)) \right)$$

$$= -\sum_{i=1}^{N} \left(y_i \left(w^T x_i - \ln\left(1+e^{w^T x_i}\right)\right) - (1-y_i) \ln\left(1+e^{w^T x_i}\right) \right)$$

$$= \sum_{i=1}^{N} \left(-y_i (w^T x_i) + \ln\left(1+e^{w^T x_i}\right) \right)$$

f - convex
g - convex & nondecreasing $\Rightarrow g(f(x))$ - convex

We know that $w^T x_i$ is a convex (and concave) function of w. Therefore,

the first term $-y_i (w^T x_i)$ is also convex.

Now if we show that $\ln(1+e^z)$ is a nondecreasing and convex function of z on $\mathbb{R}$, we will be able to use the convexity preserving operations to prove that $f(w)$ is convex

The first derivative: $\frac{d}{dz} \ln(1+e^z) = \frac{e^z}{1+e^z} = \sigma(z)$, which is positive for all $z \in \mathbb{R}$, which means that $\ln(1+e^z)$ is a nondecreasing function.

The second derivative: $\frac{d^2}{dz^2} \ln(1+e^z) = \frac{d}{dz} \sigma(z) = \sigma(z)\sigma(-z)$,

which is also positive for all $z \in \mathbb{R}$, which means that $\ln(1+e^z)$ is

a convex function

Using the following 2 facts:

1) Sum of convex functions is convex

2) Composition of a convex function with a convex non decreasing function is convex

we can verify that $f(u)$ is indeed convex in w on $\mathbb{R}^d$

Problem 3: Prove that for differentiable convex functions each local minimum is a global minimum. More specifically, given a differentiable convex function $f : \mathbb{R}^d \to \mathbb{R}$, prove that

a) if $\boldsymbol{x}^*$ is a local minimum, then $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$.

b) if $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$, then $\boldsymbol{x}^*$ is a global minimum.

We will show that if the gradient at some point x^* is not equal to zero, then this point cannot be a local optimum - we could simply follow the direction of the negative gradient and end up in a point with a lower value of the function.

More formally, suppose $\nabla f(x^*) \neq 0$ for some x^*. Then by Taylor's theorem for a sufficiently small $\varepsilon > 0$ we get

$$f(x^* - \varepsilon \nabla f(x^*)) = f(x^*) - (\varepsilon \nabla f(x^*))^T \nabla f(x^*) + O(\varepsilon^2 \|\nabla^2 f(x^*)\|_2^2)$$

$$= f(x^*) - \varepsilon \|\nabla f(x^*)\|_2^2 + O(\varepsilon^2 \|\nabla f(x^*)\|_2^2)$$

$$< f(x^*)$$

Which means that x^* is not a local optimum. Therefore, the gradient must be equal to 0 for any local optimum x^*.

We will prove (b) using the first-order criterion for convexity:

$$f(y) \geq f(x) + (y - x)^T \nabla f(x)$$

If we plug in x^* and use the fact that $\nabla f(x^*) = 0$ we get: $f(y) \geq f(x^*)$ for all y, meaning x^* is a global minimum

Problem 4: Assume that $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$ are convex functions. Prove or disprove the following statements:

a) The function $h(x) = g(f(x))$ is convex. False

b) The function $h(x) = g(f(x))$ is convex if g is non-decreasing. True

a) Counterexample: $f(x) = x^2$

$g(z) = -x$

Since the function $h(x) = g(f(x)) = -x^2$ is twice differentiable, we can inspect its 2nd derivative: $\frac{d^2}{dx^2} h(x) = -2$

Since the second derivative is negative for all x, we conclude that the function h is not convex

b) Suppose $x_0, x_1 \in \mathbb{R}$ and $\lambda \in (0,1)$. We will use a shorthand notation

$x_\lambda = \lambda x_1 + (1-\lambda) x_0$

We will prove the convexity of h using the definition of convexity and the properties of f and g:

f convex $\Rightarrow f(x_\lambda) \leq \lambda f(x_1) + (1-\lambda) f(x_0)$

g non-decreasing $\Rightarrow g(f(x_\lambda)) \leq g(\lambda f(x_1) + (1-\lambda) f(x_0))$ (1)

g convex $\Rightarrow g(\lambda f(x_1) + (1-\lambda) f(x_0)) \leq$

$\lambda g(f(x_1)) + (1-\lambda) g(f(x_0))$ (2)

(1) and (2) $\Rightarrow g(f(x_\lambda)) \leq \lambda g(f(x_1)) + (1-\lambda) g(f(x_0))$

$\Leftrightarrow h(x_\lambda) \leq \lambda h(x_1) + (1-\lambda) h(x_0)$

Therefore h is convex

2 Optimization / Gradient descent

Problem 5: You are given the following objective function $f : \mathbb{R}^2 \to \mathbb{R}$

$$f(x_1, x_2) = 0.5x_1^2 + x_2^2 + 2x_1 + x_2 + \cos(\sin(\sqrt{\pi})).$$

a) Compute the minimizer $\boldsymbol{x}^*$ of f analytically.

As f is a sum of convex functions, it is convex. To find the global minimizer, we compute the gradient and set it to 0

$$\nabla f(x_1, x_2) = \begin{pmatrix} x_1 + 2 \\ 2x_2 + 1 \end{pmatrix} \overset{!}{=} \begin{pmatrix} 0 \\ 0 \end{pmatrix} \Rightarrow \begin{pmatrix} x_1^* \\ x_2^* \end{pmatrix} = \begin{pmatrix} -2 \\ -\frac{1}{2} \end{pmatrix}$$

b) Perform 2 steps of gradient descent on f starting from the point $\boldsymbol{x}^{(0)} = (0, 0)$ with a constant learning rate $\tau = 1$.

first step $\begin{pmatrix} x_1^{(1)} \\ x_2^{(1)} \end{pmatrix} = \begin{pmatrix} x_1^{(0)} \\ x_2^{(0)} \end{pmatrix} - \tau \begin{pmatrix} x_1^{(0)} + 2 \\ 2x_2^{(0)} + 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} - 1 \begin{pmatrix} 0+2 \\ 0+1 \end{pmatrix} = \begin{pmatrix} -2 \\ -1 \end{pmatrix}$

second step $\begin{pmatrix} x_1^{(2)} \\ x_2^{(2)} \end{pmatrix} = \begin{pmatrix} x_1^{(1)} \\ x_2^{(1)} \end{pmatrix} - \tau \begin{pmatrix} x_1^{(1)} + 2 \\ 2x_2^{(1)} + 1 \end{pmatrix} = \begin{pmatrix} -2 \\ -1 \end{pmatrix} - 1 \begin{pmatrix} 0 \\ -1 \end{pmatrix} = \begin{pmatrix} -2 \\ 0 \end{pmatrix}$

c) Will the gradient descent procedure from Problem b) ever converge to the true minimizer $\boldsymbol{x}^*$? Why or why not? If the answer is no, how can we fix it?

By performing one more iteration of gradient descent we observe that

$$\begin{pmatrix} x_1^{(3)} \\ x_2^{(3)} \end{pmatrix} = \begin{pmatrix} x_1^{(2)} \\ x_2^{(2)} \end{pmatrix} - \tau \begin{pmatrix} x_1^{(2)} + 2 \\ 2x_2^{(2)} + 1 \end{pmatrix} = \begin{pmatrix} -2 \\ 0 \end{pmatrix} - 1 \begin{pmatrix} 0 \\ +1 \end{pmatrix} = \begin{pmatrix} -2 \\ -1 \end{pmatrix} = \begin{pmatrix} x_1^{(1)} \\ x_2^{(1)} \end{pmatrix}$$

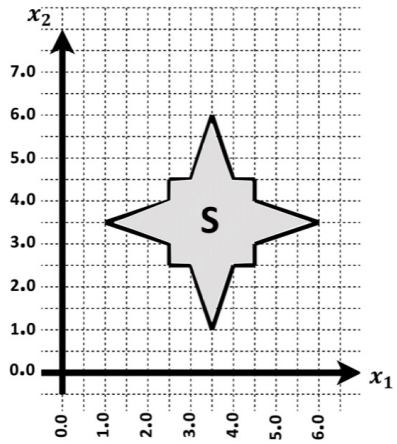
That is, we are stuck iterating between $x^{(1)}$ and $x^{(2)}$ forever. We can fix this by decreasing the learning rate.

Problem 7: Let $f : \mathbb{R}^2 \to \mathbb{R}$ be the following convex function:

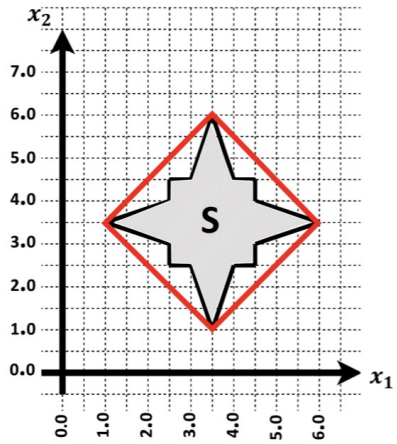
$$f(x_1, x_2) = e^{x_1 + x_2} - 5 \cdot \log(x_2)$$

a) Consider the following shaded region $S \subset \mathbb{R}^2$. Is this region convex? Why?

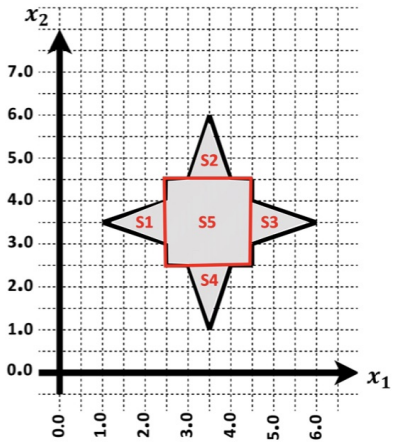
b) Assume that we are given an algorithm `ConvOpt`(f, D) that takes as input a convex function f and convex region D, and returns the minimum of f over D. Using the `ConvOpt` algorithm, how would you find the global minimum of f over the shaded region S?



(a) initial non-convex set



(b) convex hull



(c) union of convex sets

a) It is not because we can choose 2 points in S such that the line connecting the points does not completely resides in S. for example $(1.0, 3.5)^T$ and $(3.5, 6.0)^T$

b)

We can partition the shaded region S to the following five convex regions $S_1, \ldots, S_5$ (Figure 1c)

Afterwards, run the ConvOpt algorithm separately for the 5 regions and obtain

$$m_i = \min_{x \in S_i} f(x) = \text{ConvOpt}(f, S_i)$$

Finally, the minimum over the whole S can be computed as the smallest of these values, that is

$$\min_{x \in S} f(x) = \min(m_1, \ldots, m_5).$$