Lecture 07: Optimization

Introduction to Machine Learning [25737]

Sajjad Amini

Sharif University of Technology

Contents

References

The material in the slides except cited are inspired from the following reference:

- Murphy, K. P. (2022). *Probabilistic machine learning: an introduction.* MIT press.

Section 1

Basic Definitions

Optimization Problem

Training as Optimization Problem

Assume function $\mathcal{L} : \Theta \rightarrow \mathbb{R}$. An optimization problem is the process of finding the value for vector $\boldsymbol{\theta} \in \Theta$, denoted $\boldsymbol{\theta}^\star$ that minimizes L. We write this process as:

$$\theta^\star \in \overbrace{\underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argmin}} \; \mathcal{L}(\boldsymbol{\theta})}^{Set}$$

where:

$\mathcal{L}(\boldsymbol{\theta})$	Loss function or cost function
$R(\boldsymbol{\theta}) = -\mathcal{L}(\boldsymbol{\theta})$	Score function or reward function
$R(\boldsymbol{\theta}), \mathcal{L}(\boldsymbol{\theta})$	Objective function
$\Theta \subseteq \mathbb{R}^D$	Parameter space
D	Number of Variables

Global vs Local Optimization

Global Minimum

The set for global minimum is:

$$\{\boldsymbol{\theta}^\star : \forall \boldsymbol{\theta} \in \Theta, \mathcal{L}(\boldsymbol{\theta}^\star) \leq \mathcal{L}(\boldsymbol{\theta})\}$$

The set for strict global minimum is:

$$\{\boldsymbol{\theta}^\star : \forall \boldsymbol{\theta} \in \Theta, \mathcal{L}(\boldsymbol{\theta}^\star) < \mathcal{L}(\boldsymbol{\theta})\}$$

Local Minimum

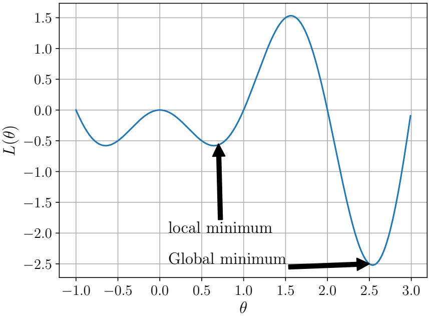
The set for local minimum is:

$$\{\boldsymbol{\theta}^\star : \exists \delta > 0, \forall \boldsymbol{\theta} \in \Theta, \boldsymbol{\theta} \neq \boldsymbol{\theta}^\star, \text{ if } \|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\| < \delta \text{ then } \mathcal{L}(\boldsymbol{\theta}^\star) \leq \mathcal{L}(\boldsymbol{\theta})\}$$

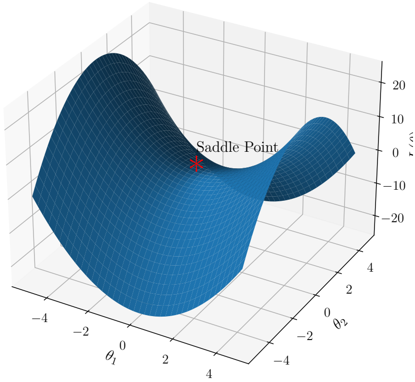
The set for strict local minimum is:

$$\{\boldsymbol{\theta}^\star : \exists \delta > 0, \forall \boldsymbol{\theta} \in \Theta, \boldsymbol{\theta} \neq \boldsymbol{\theta}^\star, \text{ if } \|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\| < \delta \text{ then } \mathcal{L}(\boldsymbol{\theta}^\star) < \mathcal{L}(\boldsymbol{\theta})\}$$

Illustration of Local and Global Minimum



(a) Local minimum vs global minimum



(b) Saddle point

Optimality Conditions

Local Minimum

Assume $\mathcal{L}$ to be twice differentiable and $\Theta = \mathbb{R}^D$. Consider a point $\boldsymbol{\theta}^\star \in \mathbb{R}^D$ and let $\boldsymbol{g}^\star = \boldsymbol{g}(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}^\star}$ and $\boldsymbol{H}^\star = \boldsymbol{H}(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}^\star}$ to be gradient vector and Hessian matrix at $\boldsymbol{\theta}^\star$. Then:

- *Necessary condition*: If $\boldsymbol{\theta}^\star$ is a local minimum, then we must have $\boldsymbol{g}^\star = \mathbf{0}$ and $\boldsymbol{H}^\star \succeq 0$.
- *Sufficient condition*: If $\boldsymbol{g}^\star = \mathbf{0}$ and $\boldsymbol{H}^\star \succ 0$, then $\boldsymbol{\theta}^\star$ is a local minimum.

Constrained vs Unconstrained Optimization

Feasible Set

Feasible set is the subset of the parameter space that satisfies the constraints over the parameter vector as:

$$\mathcal{C} = \{\boldsymbol{\theta} : g_j(\boldsymbol{\theta}) \leq 0, j \in \mathcal{I} \text{ and } h_k(\boldsymbol{\theta}) = 0, k \in \epsilon\} \subseteq \mathbb{R}^D$$

where:

$g_j(\boldsymbol{\theta})$	Inequality constraints
$h_k(\boldsymbol{\theta}) = 0$	Equality constraints
$\mathcal{I}$	Index set for Inequality constraints
ϵ	Index set for Equality constraints

Constrained vs Unconstrained Optimization

$$\theta^\star \in \underset{\boldsymbol{\theta} \in \mathcal{C}}{\operatorname{argmin}} \ \mathcal{L}(\boldsymbol{\theta})$$

The above optimization problem is unconstrained if $\mathcal{C} \in \mathbb{R}^D$, otherwise it is constrained.

Section 2

First Order Methods

First Order Methods

General Properties

First order methods are methods that:

- Leverage first order derivatives of the objective function
- Ignore curvature (higher order derivatives)

Procedure

- Specify starting point $\boldsymbol{\theta}_0$
- Perform update by:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \eta_t \boldsymbol{d}_t$$

where:

η_t	Step size or learning rate
$\boldsymbol{d}_t$	Descent direction

First Order Methods

Gradient Direction

Using Taylor expansion we have:

$$\mathcal{L}(\boldsymbol{\theta} + \epsilon\boldsymbol{\lambda}) \simeq \mathcal{L}(\boldsymbol{\theta}) + \epsilon\boldsymbol{g}^T(\boldsymbol{\theta})\boldsymbol{\lambda} + \mathcal{O}(\epsilon^2)$$

Thus if we assume $\boldsymbol{\lambda} = -\boldsymbol{g}(\boldsymbol{\theta})$, then for a small enough ϵ, we have:

$$\mathcal{L}(\boldsymbol{\theta} + \epsilon\boldsymbol{\lambda}) - \mathcal{L}(\boldsymbol{\theta}) \simeq -\epsilon\|\boldsymbol{g}(\boldsymbol{\theta})\|^2 \leq 0$$

So $-\boldsymbol{g}(\boldsymbol{\theta})$ is a descent direction.

First Order Methods

Gradient Descent

Gradient Descent (GD) method uses the following direction:

$$\boldsymbol{d}_t = -\boldsymbol{g}(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}_t}$$

Momentum

Momentum method uses the following direction:

$$\boldsymbol{m}_t = \beta \boldsymbol{m}_{t-1} + \boldsymbol{g}_{t-1}$$
$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} - \eta_t \boldsymbol{m}_t$$

where $\boldsymbol{m}_t$ is the momentum vector and $\beta < 1$

Momentum as Generalization of GD

For $\beta = 1$, momentum method degenerated to GD method.

Learning Rate Schedule

Learning Rate Schedule

The sequence of step sizes $\{\eta_t\}$ is called the learning rate schedule.

Sample Schedules

- Constant: $\eta_t = \eta$
  - Too large values may fail convergence
  - Too small values lead to low convergence rate
- Armijo-Goldstein: Assume $m = \langle \boldsymbol{g}(\boldsymbol{\theta}_t), \boldsymbol{d}_t \rangle$ and select $\tau \in (0, 1)$, $c \in (0, 1)$ and η_0, then:
  - $\gamma = -cm$ and $j = 0$
  - Until $\mathcal{L}(\boldsymbol{\theta}_t) - \mathcal{L}(\boldsymbol{\theta}_t + \eta_{tj}\boldsymbol{d}_t) \geq \eta_{t,j}\gamma$, increament j and set $\eta_{t,j} = \tau\eta_{t,(j-1)}$
  - $\eta_{t,j}$ is the learning rate at iteration t.

Section 3

Second Order Methods

Newton's Method

Descent Direction

Assume $\boldsymbol{H}_t \triangleq \nabla^2 \mathcal{L}(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}_t} \succ 0$. Then Second order approximation of $\mathcal{L}(\cdot)$ in $\boldsymbol{\theta} = \boldsymbol{\theta}_t$ is:

$$\mathcal{L}(\boldsymbol{\theta}) \simeq \mathcal{L}(\boldsymbol{\theta}_t) + \boldsymbol{g}_t^T(\boldsymbol{\theta} - \boldsymbol{\theta}_t) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_t)^T \boldsymbol{H}_t(\boldsymbol{\theta} - \boldsymbol{\theta}_t)$$

The minimizer for the above approximation is: $\boldsymbol{\theta}_t - \boldsymbol{H}_t^{-1}\boldsymbol{g}_t$. Thus:

$$\Rightarrow \boldsymbol{d}_t = -\boldsymbol{H}_t^{-1}\boldsymbol{g}_t$$

Algorithm 0: Optimization based on descent direction

Input : $t_{\max}$ (Maximum iterations),
$f_d(\cdot)$ (direction function),
$f_l(\cdot)$ (learning rate function)

Initialization: $t = 0$, $\boldsymbol{\theta}_0$, $flag_c = \text{True}$

begin
- **while** $flag_c$ **do**
  - $\boldsymbol{d}_t = f_d(\boldsymbol{\theta}_t)$
  - $\eta_t = f_l(\boldsymbol{\theta}_t)$
  - $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta_t \boldsymbol{d}_t$
  - $t \leftarrow t + 1$
  - **if** $\|\boldsymbol{g}_{t+1}\| \leq \delta$ *or* $t > t_{\max}$ **then**
    - $flag_c \leftarrow \text{False}$
  - **end**
- **end**

end

Output : $\boldsymbol{\theta}_t$

Section 4

Stochastic Gradient Descent

Stochastic Gradient Descent

Loss Measurement Limitation

Previously we have seen that Gradient Descent (GD) method uses $\boldsymbol{d}_t = -\frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}_t}$.

Now assume you only have access to a noisy version of loss function, denoted $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{z}_t)$, where $\boldsymbol{z}_t \sim q$ and we have:

$$\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{q(\boldsymbol{z})}[\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{z})]$$

Stochastic gradient descent is a solution to the aforementioned problem.

Stochastic Gradient Descent

The update rule for stochastic gradient descent is:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta_t \nabla \mathcal{L}(\boldsymbol{\theta}_t, \boldsymbol{z}_t) = \boldsymbol{\theta}_t - \eta_t \boldsymbol{g}_t$$

The sequence $\{\boldsymbol{\theta}_t\}$ is guaranteed to converge to a stationary point provided:

- The step size η_t is decayed at a certain rate
- $\boldsymbol{z}$ is independent of $\boldsymbol{\theta}$

Section 5

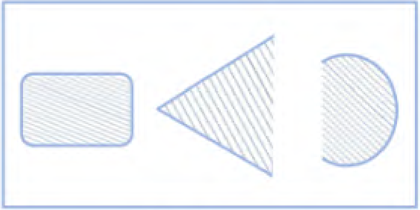
Constrained Optimization

Convex Set

Convex Set

Set $\mathcal{S}$ is convex if, for any $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{S}$, we have:

$$\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{x}' \in \mathcal{S}, \forall \lambda \in [0,1]$$



(a) Convex sets

(b) Nonconvex sets

Convex function

Convex Function

Function $f(\boldsymbol{x})$ is convex if it is defined on a convex set $\mathcal{S}$ and if, for any $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{S}$, and for any $0 \leq \lambda \leq 1$, we have:

$$f(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y}) \leq \lambda f(\boldsymbol{x}) + (1-\lambda) f(\boldsymbol{y})$$

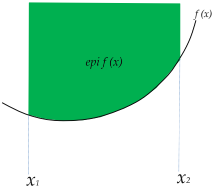


Figure: Convexity check based on epigraph

Properties of Convex Functions

Hessian Matrix of Convex Function

A twice differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if and only if the Hessian $\nabla^2 f(\boldsymbol{x})$ is positive semi-definite for all $\boldsymbol{x} \in \mathbb{R}^n$ [1].

Hessian Matrix of Convex Function

Assume that $f : \mathbb{R}^n \to R$ is convex and differentiable. Then $\boldsymbol{x}^\star$ is a global minimizer of $f(\cdot)$, if and only if $\nabla f(\boldsymbol{x}^\star) = \mathbf{0}$ [1].

Constrained vs Unconstrained Optimization

Constrained Optimization

A constrained optimization is defined as:

$$\theta^\star \in \underset{\boldsymbol{\theta} \in \mathcal{C}}{\operatorname{argmin}} \; \mathcal{L}(\boldsymbol{\theta})$$

where $\mathcal{C} = \{\boldsymbol{\theta} : g_j(\boldsymbol{\theta}) \leq 0, j \in \mathcal{I} \text{ and } h_k(\boldsymbol{\theta}) = 0, k \in \epsilon\} \subseteq \mathbb{R}^D$. The above optimization problem is unconstrained if $\mathcal{C} \in \mathbb{R}^D$, otherwise it is constrained.

Lagrange Multiplier

Simple Case with One Equality Constraint

Assume we have $\theta^\star \in \text{argmin}_{h\left(\boldsymbol{\theta}\right)=0}\ \mathcal{L}(\boldsymbol{\theta})$. Then:

- $\nabla h(\boldsymbol{\theta})$ is orthogonal to constraint surface because:

$$\begin{cases} h(\boldsymbol{\theta}+\boldsymbol{\epsilon}) \simeq h(\boldsymbol{\theta}) + \boldsymbol{\epsilon}^T \nabla h(\boldsymbol{\theta}) \\ h(\boldsymbol{\theta}) = h(\boldsymbol{\theta}+\boldsymbol{\epsilon}) \\ \boldsymbol{\epsilon} \parallel \text{ constraint surface} \end{cases} \Rightarrow \nabla h(\boldsymbol{\theta}) \perp \text{ constraint surface}$$

- If $\boldsymbol{\theta}^\star$ is optimizer then $\nabla \mathcal{L}(\boldsymbol{\theta}^\star) \perp$ constraint surface

Altogether: $\nabla \mathcal{L}(\boldsymbol{\theta}^\star) = \boldsymbol{\lambda}^\star \nabla h(\boldsymbol{\theta}^\star), \boldsymbol{\lambda}^\star \in \mathbb{R}$

Lagrange Multiplier

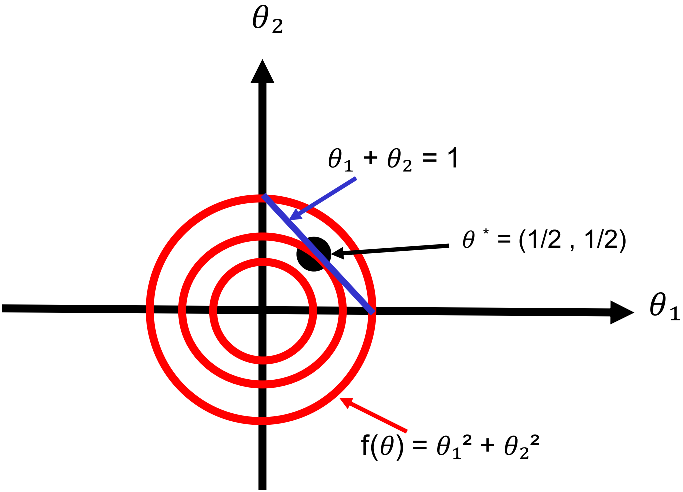


Figure: Solving problem $\theta^\star \in \text{argmin}_{\theta_1+\theta_2=0}\ \theta_1^2 + \theta_2^2$

Lagrangian

Lagrangian

Assume Lagrangian as:

$$L(\boldsymbol{\theta}, \lambda) \triangleq \mathcal{L}(\boldsymbol{\theta}) + \lambda h(\boldsymbol{\theta})$$

Then we have:

$$\nabla_{\boldsymbol{\theta},\lambda} L(\boldsymbol{\theta}, \lambda) = 0 \Leftrightarrow \begin{cases} \lambda \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}) = -\nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) \\ h(\boldsymbol{\theta}) = 0 \end{cases}$$

Thus the stationary points of Lagrangian satisfy constraints and lead to parallel gradient vectors.

M Equality constraints

For this case we simply find the stationary points of Lagrangian defined as:

$$L(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \mathcal{L}(\boldsymbol{\theta}) + \sum_{j=1}^{m} \lambda_j h_j(\boldsymbol{\theta})$$

Generalizing the Results (M Equality constraints)

Constrained Optimization with M Equality Constraints

Assume the following optimization problem:

$$\theta^\star \in \underset{\boldsymbol{\theta}\in\mathcal{C}}{\operatorname{argmin}} \; \mathcal{L}(\boldsymbol{\theta})$$
$$\mathcal{C} = \{\boldsymbol{\theta} : g_j(\boldsymbol{\theta}) \leq 0, j \in \mathcal{I} \text{ and } h_k(\boldsymbol{\theta}) = 0, k \in \epsilon\} \subseteq \mathbb{R}^D$$

- The necessary condition for $\boldsymbol{\theta}^\star$ is $L(\boldsymbol{\theta}^\star, \boldsymbol{\lambda}^\star) = \mathbf{0}$
- If $\mathcal{L}(\boldsymbol{\theta})$ is convex and equality constraints are Affine $(h_k(\boldsymbol{\theta}) = \boldsymbol{a}_k\boldsymbol{\theta} = 0)$, then the optimization problem is convex and condition $L(\boldsymbol{\theta}^\star, \boldsymbol{\lambda}^\star) = \mathbf{0}$ is sufficient.

KKT Conditions

Constrained Optimization

Assume the following optimization problem:

$$\theta^\star \in \underset{\boldsymbol{\theta}\in\mathcal{C}}{\text{argmin}} \ \mathcal{L}(\boldsymbol{\theta})$$

$$\mathcal{C} = \{\boldsymbol{\theta} : g_j(\boldsymbol{\theta}) \leq 0, j \in \mathcal{I} \text{ and } h_k(\boldsymbol{\theta}) = 0, k \in \epsilon\} \subseteq \mathbb{R}^D$$

We define the generalized Lagrangian as:

$$L(\boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\lambda}) = \mathcal{L}(\boldsymbol{\theta}) + \sum_i \mu_i g_i(\boldsymbol{\theta}) + \sum_j \lambda_j h_j(\boldsymbol{\theta})$$

Then the KKT (Karush–Kuhn–Tucker) conditions are:

- $\nabla\mathcal{L}(\boldsymbol{\theta}) + \sum_i \mu_i \nabla g_i(\boldsymbol{\theta}) + \sum_j \lambda_j \nabla h_j(\boldsymbol{\theta}) = 0$ (Stationary point of Lagrangian)
- $\boldsymbol{g}(\boldsymbol{\theta}) \leq 0, \boldsymbol{h}(\boldsymbol{\theta}) = 0$ (Feasibilty)
- $\boldsymbol{\mu} \geq 0$ (Dual feasibility)
- $\boldsymbol{\mu} \odot \boldsymbol{g} = \mathbf{0}$ (Complementary Slackness)

KKT Conditions

Constrained Optimization

Again assume the following optimization problem:

$$\theta^\star \in \underset{\boldsymbol{\theta} \in \mathcal{C}}{\operatorname{argmin}} \ \mathcal{L}(\boldsymbol{\theta})$$

$$\mathcal{C} = \{\boldsymbol{\theta} : g_j(\boldsymbol{\theta}) \leq 0, j \in \mathcal{I} \text{ and } h_k(\boldsymbol{\theta}) = 0, k \in \epsilon\} \subseteq \mathbb{R}^D$$

Then KKT conditions are:

- Necessary for $\boldsymbol{\theta}$
- Sufficient for $\boldsymbol{\theta}$ if above problem is convex ($\mathcal{L}(\boldsymbol{\theta})$ and $\{g_j(\boldsymbol{\theta})\}_j \in \mathcal{I}$ are convex functions and $\{h_k(\boldsymbol{\theta})\}_{k \in \epsilon}$ are Affine transforms).

KKT Conditions

KKT Conditions [2]

Consider the following convex optimization problem:

$$\min_{(x,y)\in\mathcal{S}} \frac{1}{x+y}$$

$$\text{subject to } \begin{cases} 2x+y^2-6 \leq 0 \\ 1-x \leq 0 \\ 1-y \leq 0 \end{cases}$$

where $\mathcal{S} = \{(x,y) : x, y > 0\}$. Find the optimal point.
Solution: From zero Lagrangian gradient we have:

$$\begin{bmatrix} -\frac{1}{(x+y)^2} \\ -\frac{1}{(x+y)^2} \end{bmatrix} + \mu_1 \begin{bmatrix} 2 \\ 2y \end{bmatrix} + \mu_2 \begin{bmatrix} -1 \\ 0 \end{bmatrix} + \mu_3 \begin{bmatrix} 0 \\ -1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

From complementary slackness equations we have:

$$\mu_1(2x+y^2-6) = \mu_2(1-x) = \mu_3(1-y) = 0$$

KKT Conditions [2] (Continue)

Assume $\mu_1 = 0$, then:

$$\begin{bmatrix} -\frac{1}{(x+y)^2} \\ -\frac{1}{(x+y)^2} \end{bmatrix} + \mu_2 \begin{bmatrix} -1 \\ 0 \end{bmatrix} + \mu_3 \begin{bmatrix} 0 \\ -1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Rightarrow \mu_2 = \mu_3 = -\frac{1}{(x+y)^2} \leq 0$$

$$\Rightarrow \begin{cases} \text{Contradiction} \\ 2x + y^2 - 6 = 0 \end{cases}$$

No we assume $x = 1$ then:

$$x = 1 \Rightarrow y = \left\{ +2\ (valid) - 2\ (invalid) \right. \quad \Rightarrow \begin{cases} \mu_1 = \frac{1}{36} \\ \mu_2 = -\frac{1}{18} \\ \mu_3 = 0 \end{cases} \quad \Rightarrow \begin{cases} \text{Contradiction} \\ x \neq 0 \end{cases}$$

No we assume $y = 1$ then:

$$y = 1 \Rightarrow x = 2.5 \Rightarrow \begin{cases} \mu_1 = \frac{2}{49} \\ \mu_2 = 0 \\ \mu_3 = 0 \end{cases} \quad \Rightarrow \begin{cases} \boldsymbol{\theta}^\star = (2.5, 1) \\ \boldsymbol{\mu} = (\frac{2}{49}, 0, 0) \end{cases}$$

References I

Markus Grasmair,
"Basic properties of convex functions,"
Department of Mathematics, Norwegian University of Science and Technology, 2016.

"Chapter 5, lecture 6: Kkt theorem, gradient form,"
https://faculty.math.illinois.edu/~mlavrov/docs/484-spring-2019/ch5lec6.pdf,
Accessed: 2022-10-26.