Week 9

Lecture 1

Constrained Optimization with One Inequality Constraint

Problem Setup

Given functions
$f : \mathbb{R}^d \to \mathbb{R}$
$g : \mathbb{R}^d \to \mathbb{R}$
Consider the optimization problem

$$\min_x f(x)$$

subject to

$$g(x) \leq 0$$

Necessary Conditions for Optimality

Suppose a point x^* is claimed to be optimal.
To verify optimality, we check necessary conditions.

1. Feasibility Condition

The point must satisfy the constraint:

$$g(x^*) \leq 0$$

If

$$g(x^*) > 0$$

then x^* is infeasible and cannot be optimal.

---

2. No Descent Direction Should Be Feasible

If x^* is optimal, then:
There must not exist a direction that

- decreases f

- and maintains feasibility

Descent Directions

Taylor Expansion

For small step size η and direction d,

$$f(x + \eta d) \approx f(x) + \eta d^T \nabla f(x)$$

A direction d is a descent direction for f at x if

$$d^T \nabla f(x) < 0$$

This condition depends only on f.
Thus, at x^*:

$$d^T \nabla f(x^*) < 0$$

characterizes descent directions.
Geometrically, descent directions are all directions making an obtuse angle with $\nabla f(x^*)$.

Feasible Directions

Assume

$$g(x^*) \leq 0$$

A direction d is feasible if, for sufficiently small $\eta > 0$,

$$g(x^* + \eta d) \leq 0$$

Using Taylor expansion:

$$g(x^* + \eta d) \approx g(x^*) + \eta d^T \nabla g(x^*)$$

If

$$d^T \nabla g(x^*) < 0$$

then g decreases further and feasibility is preserved.
Thus, any descent direction of g at x^* is a feasible direction.
This condition depends only on g.

Combined Condition

A direction d is
Descent for f if

$$d^T \nabla f(x^*) < 0$$

Feasible if

$$d^T \nabla g(x^*) < 0$$

If there exists d such that

$$d^T \nabla f(x^*) < 0$$

and

$$d^T \nabla g(x^*) < 0$$

then

- we can move in direction d
- decrease f
- remain feasible

Therefore,
x^* cannot be optimal.

Geometric Interpretation

At x^*:

- $\nabla f(x^*)$ determines descent half space
- $\nabla g(x^*)$ determines feasible half space

If the descent region of f and the feasible region of g intersect, then
there exists a direction that is both descent and feasible.
In that case, x^* is not optimal.

Key Necessary Condition

If x^* is optimal and satisfies $g(x^*) \leq 0$, then
There must not exist any direction d such that

$$d^T \nabla f(x^*) < 0$$

and

$$d^T \nabla g(x^*) < 0$$

This condition will lead to a structural relationship between

$$\nabla f(x^*)$$

and

$$\nabla g(x^*)$$

which characterizes constrained optimality.

****

Lecture 2

Relationship Between Gradients at Optimality

Consider the constrained problem

$$\min_x f(x)$$

subject to

$$g(x) \leq 0$$

Let x^* be a feasible point:

$$g(x^*) \leq 0$$

We seek a necessary condition relating

$$\nabla f(x^*) \quad \text{and} \quad \nabla g(x^*)$$

such that no descent direction is feasible.

Case 1: Gradients Point in Same Direction

Assume

$$\nabla f(x^*) \parallel \nabla g(x^*)$$

Descent directions satisfy

$$d^T \nabla f(x^*) < 0$$

Feasible directions satisfy

$$d^T \nabla g(x^*) < 0$$

If the gradients are parallel, then any direction satisfying

$$d^T \nabla f(x^*) < 0$$

automatically satisfies

$$d^T \nabla g(x^*) < 0$$

Thus every descent direction is feasible.
Therefore, there exists a whole half space of directions that both decrease f and preserve feasibility.
Conclusion:
x^* cannot be optimal.

Case 2: Gradients Anti Parallel

Assume

$$\nabla f(x^*) \quad \text{and} \quad \nabla g(x^*)$$

point in opposite directions.
Descent directions satisfy

$$d^T \nabla f(x^*) < 0$$

Feasible directions satisfy

$$d^T \nabla g(x^*) < 0$$

Since the gradients are anti parallel, the descent half space of f lies on one side, while feasible directions lie on the opposite side.
Thus no direction d can satisfy both

$$d^T \nabla f(x^*) < 0$$

and

$$d^T \nabla g(x^*) < 0$$

Therefore no descent direction is feasible.

Necessary Condition for Inequality Constraint

The above geometric argument implies:
There exists a scalar $\lambda > 0$ such that

$$\nabla f(x^*) = -\lambda \nabla g(x^*)$$

This ensures anti parallel alignment.
This is a necessary condition for optimality for

$$\min_x f(x) \quad \text{subject to} \quad g(x) \leq 0$$

The scalar λ is called the Lagrange multiplier.

Equality Constraint Case

Consider

$$\min_x f(x)$$

subject to

$$g(x) = 0$$

Feasible Directions

Using Taylor expansion,

$$g(x^* + \eta d) \approx g(x^*) + \eta d^T \nabla g(x^*)$$

Since $g(x^*) = 0$, feasibility requires

$$d^T \nabla g(x^*) = 0$$

Thus feasible directions lie on the hyperplane perpendicular to $\nabla g(x^*)$.

Descent Directions

Descent directions satisfy

$$d^T \nabla f(x^*) < 0$$

Optimal Configuration

For optimality, no direction must satisfy both

$$d^T \nabla f(x^*) < 0$$

and

$$d^T \nabla g(x^*) = 0$$

This occurs when $\nabla f(x^*)$ is orthogonal to the feasible hyperplane, i.e.,

$$\nabla f(x^*) \parallel \nabla g(x^*)$$

or

$$\nabla f(x^*) \text{ is anti parallel to } \nabla g(x^*)$$

Thus,
there exists scalar λ such that

$$\nabla f(x^*) = -\lambda \nabla g(x^*)$$

where λ is any real scalar.

Comparison

Inequality Constraint

$$\nabla f(x^*) = -\lambda \nabla g(x^*)$$

with

$$\lambda > 0$$

Equality Constraint

$$\nabla f(x^*) = -\lambda \nabla g(x^*)$$

with

$$\lambda \in \mathbb{R}$$

Key Observation

For constrained optimality, the gradient of the objective must lie in the span of the gradient of the constraint.
The scalar λ is called the Lagrange multiplier.
This condition provides the foundation for solving constrained optimization problems using the method of Lagrange multipliers.

****

Lecture 3

Method of Lagrange Multipliers

Consider the equality constrained optimization problem

$$\min_x f(x)$$

subject to

$$g(x) = 0$$

Necessary Conditions for Optimality

If x^* is optimal, then:

1. Feasibility

$$g(x^*) = 0$$

2. Gradient Alignment

There exists scalar λ such that

$$\nabla f(x^*) = -\lambda \nabla g(x^*)$$

For equality constraints,

$$\lambda \in \mathbb{R}$$

The scalar λ is called the Lagrange multiplier.

Example

Let

$$f(x_1, x_2) = x_1^2 + 2x_2 + 4x_2^2$$

Constraint:

$$g(x_1, x_2) = x_1^2 + x_2^2 - 1$$

Thus feasible points lie on the unit circle:

$$x_1^2 + x_2^2 = 1$$

Step 1: Compute Gradients

Gradient of f:

$$\nabla f(x_1, x_2) = \begin{pmatrix} 2x_1 \\ 2 + 8x_2 \end{pmatrix}$$

Gradient of g:

$$\nabla g(x_1, x_2) = \begin{pmatrix} 2x_1 \\ 2x_2 \end{pmatrix}$$

Step 2: Apply Lagrange Condition

We impose

$$\nabla f = -\lambda \nabla g$$

This gives the system

$$\begin{pmatrix} 2x_1 \\ 2 + 8x_2 \end{pmatrix} = -\lambda \begin{pmatrix} 2x_1 \\ 2x_2 \end{pmatrix}$$

Which yields componentwise equations:

$$2x_1 = -\lambda 2x_1$$

$$2 + 8x_2 = -\lambda 2x_2$$

Step 3: Solve System

Equation 1

$$2x_1 + 2\lambda x_1 = 0$$

$$2x_1(1 + \lambda) = 0$$

Thus either
Case A:

$$\lambda = -1$$

or
Case B:

$$x_1 = 0$$

Case A: $\lambda = -1$

Substitute into Equation 2:

$$2 + 8x_2 = -(-1)2x_2$$

$$2 + 8x_2 = 2x_2$$

$$2 + 6x_2 = 0$$

$$x_2 = -\frac{1}{3}$$

Feasibility condition:

$$x_1^2 + x_2^2 = 1$$

$$x_1^2 + \frac{1}{9} = 1$$

$$x_1^2 = \frac{8}{9}$$

$$x_1 = \pm\frac{\sqrt{8}}{3}$$

Thus two candidate points:

$$\left(\frac{\sqrt{8}}{3}, -\frac{1}{3}\right)$$

$$\left(-\frac{\sqrt{8}}{3}, -\frac{1}{3}\right)$$

Case B: $x_1 = 0$

Feasibility:

$$x_2^2 = 1$$

$$x_2 = \pm 1$$

Thus two candidate points:

$$(0, 1)$$

$$(0, -1)$$

Step 4: Evaluate Objective Function

Recall

$$f(x_1, x_2) = x_1^2 + 2x_2 + 4x_2^2$$

At $(0, 1)$

$$f = 0 + 2 + 4 = 6$$

At $(0, -1)$

$$f = 0 - 2 + 4 = 2$$

At $\left(\pm\frac{\sqrt{8}}{3}, -\frac{1}{3}\right)$

$$x_1^2 = \frac{8}{9}$$

$$x_2 = -\frac{1}{3}$$

$$x_2^2 = \frac{1}{9}$$

$$f = \frac{8}{9} - \frac{2}{3} + \frac{4}{9}$$

$$= \frac{12}{9} - \frac{2}{3}$$

$$= \frac{4}{3} - \frac{2}{3}$$

$$= \frac{2}{3}$$

Conclusion

Function values:

$$6, \quad 2, \quad \frac{2}{3}, \quad \frac{2}{3}$$

Minimum value:

$$\frac{2}{3}$$

Minimizers:

$$\left(\frac{\sqrt{8}}{3}, -\frac{1}{3} \right)$$

$$\left(-\frac{\sqrt{8}}{3}, -\frac{1}{3} \right)$$

Maximum value:

$$6$$

at

$$(0, 1)$$

Projected Gradient Descent

When analytic solution is not feasible, use iterative methods.

Standard Gradient Descent

Initialize:

$$x_0$$

Iterate:

$$x_{t+1} = x_t - \eta \nabla f(x_t)$$

This may leave the feasible region.

Projection Operator

Let feasible set

$$S = \{y : g(y) \leq 0\}$$

Define projection operator:

$$\Pi_S(x) = \arg\min_{y \in S} \|x - y\|^2$$

Projection finds the closest feasible point.

Projected Gradient Descent Algorithm

Initialization:

$$x_0$$

For $t = 0, 1, \ldots, T$:
Gradient step:

$$z_t = x_t - \eta \nabla f(x_t)$$

Projection step:

$$x_{t+1} = \Pi_S(z_t)$$

Key Observations

1. Every iterate remains feasible.
2. Requires efficient computation of projection.
3. If feasible set is convex, convergence to optimal solution can be guaranteed under suitable conditions.

4. Convex objective functions improve convergence guarantees.
   Projected gradient descent extends gradient descent to constrained optimization.

****

Lecture 4

Introduction to Convexity

Optimization algorithms such as gradient descent typically converge to local minima.
To guarantee global optimality, we restrict attention to special classes of functions and sets.
These are based on the notion of convexity.

Convex Sets

Definition

Let $S \subseteq \mathbb{R}^d$.
The set S is said to be convex if for all $x_1, x_2 \in S$ and for all $\lambda \in [0,1]$,

$$\lambda x_1 + (1-\lambda)x_2 \in S$$

Interpretation

For any two points in the set, the entire line segment joining them must lie inside the set.
The point

$$\lambda x_1 + (1-\lambda)x_2$$

is called a convex combination of x_1 and x_2.
Special cases:
If $\lambda = 1$,

$$\lambda x_1 + (1-\lambda)x_2 = x_1$$

If $\lambda = 0$,

$$\lambda x_1 + (1-\lambda)x_2 = x_2$$

If $\lambda = \frac{1}{2}$,

$$\frac{1}{2}x_1 + \frac{1}{2}x_2$$

is the midpoint of the segment joining x_1 and x_2.
As λ varies from 0 to 1, the point traces the entire line segment.

Examples of Convex Sets

1. Line in $\mathbb{R}^2$

Any line in $\mathbb{R}^2$ is convex.
Given two points on the line, the segment joining them lies entirely on the line.

---

2. Ellipse or Circle

Any filled ellipse or circle is convex.
For any two points inside the ellipse, the entire line segment lies inside the ellipse.

---

3. Non Convex Set

A set with an indentation or gap is not convex.
If there exist $x_1, x_2 \in S$ such that the line segment between them exits the set, then S is not convex.

Convex Sets in One Dimension

If $S \subseteq \mathbb{R}$, then convex sets are exactly intervals of the form

$$[a, b]$$

or

$$(a, b)$$

or infinite intervals such as

$$(-\infty, b]$$

Union of two disjoint intervals is not convex.
If

$$S = [a, b] \cup [c, d]$$

with $b < c$, then S is not convex.

Hyperplanes

Definition

Given $w \in \mathbb{R}^d$ and scalar b, define

$$S = \{x \in \mathbb{R}^d : w^T x = b\}$$

This set is called a hyperplane.

Claim

Hyperplanes are convex sets.

Proof

Take any $x_1, x_2 \in S$.
Then

$$w^T x_1 = b$$

$$w^T x_2 = b$$

Consider any $\lambda \in [0, 1]$.
Compute:

$$w^T \big(\lambda x_1 + (1 - \lambda) x_2\big)$$

Using linearity:

$$= \lambda w^T x_1 + (1 - \lambda) w^T x_2$$

Substitute values:

$$= \lambda b + (1 - \lambda) b$$

$$= b$$

Therefore,

$$\lambda x_1 + (1-\lambda)x_2 \in S$$

Hence S is convex.

Half Spaces

Definition

Given $w \in \mathbb{R}^d$ and scalar b, define

$$S = \{x \in \mathbb{R}^d : w^T x \leq b\}$$

This set is called a half space.

Convexity of Half Spaces

Take $x_1, x_2 \in S$.
Then

$$w^T x_1 \leq b$$

$$w^T x_2 \leq b$$

For $\lambda \in [0,1]$,

$$w^T\big(\lambda x_1 + (1-\lambda)x_2\big) = \lambda w^T x_1 + (1-\lambda) w^T x_2$$

Since both terms are less than or equal to b,

$$\leq \lambda b + (1-\lambda) b$$

$$= b$$

Therefore,

$$\lambda x_1 + (1-\lambda)x_2 \in S$$

Hence half spaces are convex.

Important Property

Intersection of Convex Sets

Let $S_1, S_2 \subseteq \mathbb{R}^d$ be convex.
Define

$$S = S_1 \cap S_2$$

Take any $x_1, x_2 \in S$.
Then

$$x_1, x_2 \in S_1$$

and

$$x_1, x_2 \in S_2$$

Since both sets are convex,

$$\lambda x_1 + (1-\lambda)x_2 \in S_1$$

and

$$\lambda x_1 + (1-\lambda)x_2 \in S_2$$

Thus,

$$\lambda x_1 + (1-\lambda)x_2 \in S$$

Therefore, intersection of convex sets is convex.
Convex sets play a fundamental role in optimization and machine learning, as many feasible regions are intersections of half spaces, and therefore convex.

****

Lecture 5

Properties of Convex Sets

Intersection of Convex Sets

Let $S_1, S_2 \subseteq \mathbb{R}^d$ be convex sets.
Define

$$S_{12} = S_1 \cap S_2$$

Then

$$S_{12} = \{x \in \mathbb{R}^d : x \in S_1 \text{ and } x \in S_2\}$$

Claim

Intersection of convex sets is convex.

Proof

Take arbitrary $x_1, x_2 \in S_{12}$.
Then

$$x_1, x_2 \in S_1$$

and

$$x_1, x_2 \in S_2$$

Since S_1 is convex, for any $\lambda \in [0, 1]$,

$$\lambda x_1 + (1 - \lambda) x_2 \in S_1$$

Since S_2 is convex, for any $\lambda \in [0, 1]$,

$$\lambda x_1 + (1 - \lambda) x_2 \in S_2$$

Therefore,

$$\lambda x_1 + (1 - \lambda) x_2 \in S_1 \cap S_2$$

Hence S_{12} is convex.

Application: Solution Set of Linear Systems

Consider

$$S = \{x \in \mathbb{R}^d : Ax = b\}$$

where

$$A \in \mathbb{R}^{m \times d}, \quad b \in \mathbb{R}^m$$

Write A row wise:

$$A = \begin{pmatrix} a_1^T \\ a_2^T \\ \vdots \\ a_m^T \end{pmatrix}$$

Then

$$Ax = b$$

is equivalent to the system

$$a_1^T x = b_1$$

$$a_2^T x = b_2$$

$$\vdots$$

$$a_m^T x = b_m$$

Thus,

$$S = \bigcap_{i=1}^{m} \{x : a_i^T x = b_i\}$$

Each set

$$\{x : a_i^T x = b_i\}$$

is a hyperplane and hence convex.
Since intersection of convex sets is convex,

$$S$$

is convex.

Convex Combinations

Let

$$S = \{x_1, x_2, \ldots, x_n\} \subseteq \mathbb{R}^d$$

A point $z \in \mathbb{R}^d$ is called a convex combination of points in S if there exist scalars

$$\lambda_1, \lambda_2, \ldots, \lambda_n$$

such that

$$\lambda_i \geq 0$$

and

$$\sum_{i=1}^{n} \lambda_i = 1$$

and

$$z = \sum_{i=1}^{n} \lambda_i x_i$$

Convex Hull

Definition

The convex hull of S is defined as

$$\mathrm{ch}(S) = \left\{ \sum_{i=1}^{n} \lambda_i x_i : \lambda_i \geq 0, \sum_{i=1}^{n} \lambda_i = 1 \right\}$$

Thus, convex hull is the set of all convex combinations of points in S.

Geometric Interpretation

For finite points in $\mathbb{R}^2$, the convex hull is the smallest polygon containing all the points.
It consists of all points that can be written as convex combinations of the original points.

Convexity of Convex Hull

Claim:

$$\mathrm{ch}(S)$$

is convex.

Proof outline:
Take two convex combinations:

$$z_1 = \sum_{i=1}^{n} \lambda_i x_i$$

$$z_2 = \sum_{i=1}^{n} \mu_i x_i$$

Let $\theta \in [0, 1]$.
Then

$$\theta z_1 + (1-\theta) z_2 = \sum_{i=1}^{n} \big(\theta \lambda_i + (1-\theta)\mu_i\big) x_i$$

Coefficients satisfy

$$\theta \lambda_i + (1-\theta)\mu_i \geq 0$$

and

$$\sum_{i=1}^{n} \big(\theta \lambda_i + (1-\theta)\mu_i\big) = 1$$

Hence it is also a convex combination.
Therefore convex hull is convex.

Alternate Definition of Convex Hull

The convex hull of S can also be defined as
the intersection of all convex sets that contain S.
Formally,

$$\mathrm{ch}(S) = \bigcap \{C : C \text{ convex and } S \subseteq C\}$$

Since intersection of convex sets is convex,
this definition also yields a convex set.

Euclidean Balls

Define Euclidean ball in $\mathbb{R}^d$:

$$B = \{x \in \mathbb{R}^d : \|x\|_2 \leq \theta\}$$

where

$$\|x\|_2 = \sqrt{\sum_{i=1}^{d} x_i^2}$$

Convexity of Euclidean Ball

Take $x_1, x_2 \in B$.
Then

$$\|x_1\|_2 \leq \theta$$

$$\|x_2\|_2 \leq \theta$$

For $\lambda \in [0,1]$,
by triangle inequality and homogeneity,

$$\|\lambda x_1 + (1-\lambda)x_2\|_2 \leq \lambda\|x_1\|_2 + (1-\lambda)\|x_2\|_2$$

$$\leq \lambda\theta + (1-\lambda)\theta$$

$$= \theta$$

Thus

$$\lambda x_1 + (1-\lambda)x_2 \in B$$

Therefore Euclidean balls are convex.
Convex sets, convex combinations, convex hulls, and intersections form foundational tools for optimization and machine learning.

****

Lecture 6

Convex Functions

Let $f : \mathbb{R}^d \to \mathbb{R}$.
Assume the domain of f is a convex set.
We present four equivalent definitions of convex functions.

Definition 1: Epigraph Definition

Epigraph

The epigraph of f is defined as

$$\operatorname{epi}(f) = \{(x, z) \in \mathbb{R}^{d+1} : z \geq f(x)\}$$

Thus the epigraph consists of all points lying on or above the graph of f.

Convexity via Epigraph

The function f is convex if and only if

$$\operatorname{epi}(f)$$

is a convex set in $\mathbb{R}^{d+1}$.

Definition 2: Jensen Inequality Form

A function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if and only if
for all $x_1, x_2 \in \mathbb{R}^d$ and all $\lambda \in [0, 1]$,

$$f\big(\lambda x_1 + (1 - \lambda) x_2\big) \leq \lambda f(x_1) + (1 - \lambda) f(x_2)$$

Geometric Interpretation

For any two points on the graph,
the function value at the convex combination of inputs
lies below the straight line joining the function values.
Thus the graph of a convex function lies below all secant lines.

Definition 3: First Order Condition

Assume f is differentiable.
Then f is convex if and only if
for all $x, y \in \mathbb{R}^d$,

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

Interpretation

The linear approximation of f at point x

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

is a global under-estimator of the function.
Thus every tangent hyperplane lies below the function.

Relation to Taylor Expansion

Using first order Taylor expansion,

$$f(x + \epsilon d) = f(x) + \epsilon d^T \nabla f(x) + \text{higher order terms}$$

Convexity implies

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

for all y,
without requiring ϵ to be small.

Definition 4: Second Order Condition

Assume f is twice differentiable.
Define the Hessian matrix

$$H(x) = \nabla^2 f(x) \in \mathbb{R}^{d \times d}$$

with entries

$$H_{ij}(x) = \frac{\partial^2 f}{\partial x_i \partial x_j}$$

Convexity via Hessian

The function f is convex if and only if

$$H(x)$$

is positive semi definite for all x,
that is,
all eigenvalues of $H(x)$ are greater than or equal to zero.

Example

Let

$$f(x) = x^2$$

Then

$$f'(x) = 2x$$

$$f''(x) = 2$$

Since

$$f''(x) > 0$$

for all x,
f is convex.

Summary

A function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if any one of the following equivalent conditions holds:

1. Its epigraph is a convex set.
2. It satisfies the Jensen inequality.
3. Its tangent hyperplanes globally underestimate it.
4. Its Hessian is positive semi definite.

These characterizations provide multiple tools to verify convexity in optimization and machine learning.

****

Lecture 7

Local Minima and Global Minima of Convex Functions

Theorem

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function.
Then every local minimum of f is also a global minimum.

Definitions

Local Minimum

A point x^* is a local minimum if there exists $\delta > 0$ such that
for all z satisfying

$$\|z - x^*\| \leq \delta$$

we have

$$f(x^*) \leq f(z)$$

Thus within a ball of radius δ around x^*, the function attains its smallest value at x^*.

Global Minimum

A point z is a global minimum if
for all $y \in \mathbb{R}^d$,

$$f(z) \leq f(y)$$

A global minimum satisfies the local minimum condition for all $\delta > 0$.

Proof of Theorem

We prove by contradiction.
Assume:
x^* is a local minimum but not a global minimum.
Let z be a global minimum.
Since x^* is not global,

$$f(z) < f(x^*)$$

Step 1: Use Local Minimum Property

Since x^* is a local minimum, there exists $\delta > 0$ such that
for all y satisfying

$$\|y - x^*\| \leq \delta$$

we have

$$f(x^*) \leq f(y)$$

Step 2: Move Toward Global Minimum

Consider points on the line segment joining x^* and z:

$$y_\lambda = \lambda x^* + (1 - \lambda) z$$

For sufficiently small $\lambda > 0$,
y_λ lies within the δ ball around x^*.
Thus by local minimality,

$$f(x^*) \leq f(y_\lambda)$$

Step 3: Use Convexity

Since f is convex,

$$f(y_\lambda) \leq \lambda f(x^*) + (1 - \lambda) f(z)$$

Substitute into previous inequality:

$$f(x^*) \leq \lambda f(x^*) + (1 - \lambda) f(z)$$

Step 4: Use Strict Inequality

Since

$$f(z) < f(x^*)$$

we obtain

$$\lambda f(x^*) + (1 - \lambda) f(z) < \lambda f(x^*) + (1 - \lambda) f(x^*)$$
$$= f(x^*)$$

Thus

$$f(x^*) < f(x^*)$$

Contradiction

This is impossible.
Therefore the assumption that x^* is a local minimum but not global must be false.
Hence every local minimum is a global minimum.

Consequences

If f is convex:

1. Any algorithm that finds a local minimum automatically finds a global minimum.
2. Gradient descent, which guarantees convergence to a local minimum under suitable conditions, finds the global minimum for convex functions.
3. Convexity removes the difficulty of multiple bad local minima.
   Convexity ensures that optimization is globally well behaved.

****
