Week 10

Lecture 1

Convex Functions: Optimality and Additional Properties

Reminder: Local vs Global Minima

If f is convex, then
all local minima are global minima.
This does not imply uniqueness of the global minimum.

Non-Uniqueness of Global Minima

For $f : \mathbb{R} \to \mathbb{R}$ convex,
the set of all global minima can be an interval.
Fact:
The set of all global minima of a convex function is a convex set.
For $f : \mathbb{R} \to \mathbb{R}$, this set must be an interval.

Necessary and Sufficient Condition for Optimality

We now study optimality conditions for differentiable convex functions.

Problem

Minimize

$$f(x)$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is convex and differentiable.

Theorem

Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable and convex.
Then

$$x^* \text{ is a global minimum} \quad \iff \quad \nabla f(x^*) = 0$$

Proof Sketch

Necessity

Assume x^* is a global minimum.
Then x^* is also a local minimum.
Suppose

$$\nabla f(x^*) \neq 0$$

Then consider moving in direction

$$d = -\nabla f(x^*)$$

Using first-order Taylor approximation,
for sufficiently small $\eta > 0$,

$$f(x^* + \eta d) < f(x^*)$$

This contradicts global minimality.
Hence

$$\nabla f(x^*) = 0$$

Note: convexity is not required for this direction.

Sufficiency

Assume

$$\nabla f(x^*) = 0$$

Use first-order convexity condition:
For all $x, y \in \mathbb{R}^d$,

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

Set $x = x^*$:

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

Since

$$\nabla f(x^*) = 0$$

we obtain

$$f(y) \geq f(x^*) \quad \forall y$$

Thus x^* is a global minimum.
Convexity is essential for this direction.

Conclusion

For differentiable convex functions:

$$\nabla f(x^*) = 0$$

is both necessary and sufficient for global optimality.

Implication for Gradient Descent

Gradient descent converges to a point where

$$\nabla f(x) = 0$$

For convex functions, this guarantees global optimality.

Additional Properties of Convex Functions

Property 1: Sum of Convex Functions

Let

$$f, g : \mathbb{R}^d \to \mathbb{R}$$

be convex.
Define

$$h(x) = f(x) + g(x)$$

Then h is convex.

Proof

For any x, y and $\lambda \in [0, 1]$:

$$h(\lambda x + (1-\lambda)y) = f(\lambda x + (1-\lambda)y) + g(\lambda x + (1-\lambda)y)$$

Using convexity of f and g:

$$\leq \lambda f(x) + (1-\lambda) f(y) + \lambda g(x) + (1-\lambda) g(y)$$
$$= \lambda h(x) + (1-\lambda) h(y)$$

Hence h is convex.

Property 2: Composition with Non-Decreasing Convex Function

Let

$$f : \mathbb{R} \to \mathbb{R}$$

be convex and non-decreasing.
Let

$$g : \mathbb{R}^d \to \mathbb{R}$$

be convex.
Define

$$h(x) = f(g(x))$$

Then h is convex.

Proof Idea

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

$$h(\lambda x + (1-\lambda) y) = f(g(\lambda x + (1-\lambda) y))$$

By convexity of g:

$$g(\lambda x + (1-\lambda) y) \leq \lambda g(x) + (1-\lambda) g(y)$$

Since f is non-decreasing:

$$f(g(\lambda x + (1-\lambda) y)) \leq f(\lambda g(x) + (1-\lambda) g(y))$$

By convexity of f:

$$\leq \lambda f(g(x)) + (1-\lambda) f(g(y))$$

Thus h is convex.

Property 3: Composition with Linear Function

Let

$$f : \mathbb{R} \to \mathbb{R}$$

be convex.
Let

$$g : \mathbb{R}^d \to \mathbb{R}$$

be linear.
Define

$$h(x) = f(g(x))$$

Then h is convex.

Proof

Linearity implies

$$g(\lambda x + (1-\lambda)y) = \lambda g(x) + (1-\lambda)g(y)$$

Then convexity of f gives

$$h(\lambda x + (1-\lambda)y) \leq \lambda h(x) + (1-\lambda)h(y)$$

Important Remark

In general,
if both f and g are convex,
the composition $f \circ g$ need not be convex.

Counterexample

Let

$$g(x) = x^2$$

which is convex.
Let

$$f(x) = e^{-x}$$

Since e^x is convex and $-x$ is linear,
$f(x) = e^{-x}$ is convex.
But

$$f(g(x)) = e^{-x^2}$$

This function is not convex.
Thus composition of convex functions is not convex in general.

Concave Functions

A function g is concave if

$$-g$$

is convex.
Example:

$$g(x) = -x^2$$

is concave.
These structural properties allow us to build complex convex functions from simpler ones, which is essential in machine learning optimization.

****

Lecture 2

Applications of Optimization to Machine Learning

Linear Regression

Dataset

Training data consists of

$$\{(x_i, y_i)\}_{i=1}^{n}$$

where

- $x_i \in \mathbb{R}^d$
- $y_i \in \mathbb{R}$
  Goal: Learn a function

  $$h : \mathbb{R}^d \to \mathbb{R}$$

  such that for a new test point x_{test},

$$\hat{y}_{\text{test}} = h(x_{\text{test}})$$

Linear Hypothesis Class

In linear regression, we restrict h to be linear:

$$h(x) = w^\top x$$

where $w \in \mathbb{R}^d$.
Thus learning h reduces to learning w.

Performance Measure

Define the sum of squared errors:

$$f(w) = \sum_{i=1}^{n} (w^\top x_i - y_i)^2$$

Often scaled as

$$f(w) = \frac{1}{2} \sum_{i=1}^{n} (w^\top x_i - y_i)^2$$

Optimization problem:

$$\min_{w \in \mathbb{R}^d} f(w) = \min_{w \in \mathbb{R}^d} \frac{1}{2} \sum_{i=1}^{n} (w^\top x_i - y_i)^2$$

Convexity of the Objective

Write

$$f(w) = \sum_{i=1}^{n} h_i(w)$$

where

$$h_i(w) = (w^\top x_i - y_i)^2$$

Each $h_i(w)$ is a composition:

- $g(w) = w^\top x_i - y_i$ which is linear in w
- $\phi(z) = z^2$ which is convex
  Since composition of convex function with linear function is convex,

$$h_i(w) \text{ is convex}$$

  Since sum of convex functions is convex,

$$f(w) \text{ is convex}$$

Matrix Formulation

Define

- $X \in \mathbb{R}^{n \times d}$ with rows $x_i^\top$
- $w \in \mathbb{R}^d$
- $y \in \mathbb{R}^n$

Then

$$f(w) = \frac{1}{2} \|Xw - y\|_2^2$$

Equivalent quadratic form:

$$f(w) = \frac{1}{2} (Xw - y)^\top (Xw - y)$$

Expand:

$$f(w) = \frac{1}{2} \left(w^\top X^\top X w - 2 w^\top X^\top y + y^\top y \right)$$

Gradient Computation

Compute gradient:

$$\nabla f(w) = X^\top X w - X^\top y$$

Optimality Condition

Since f is convex and differentiable,
global minimum satisfies

$$\nabla f(w^*) = 0$$

Thus

$$X^\top X w^* = X^\top y$$

If $X^\top X$ is invertible:

$$w^* = (X^\top X)^{-1} X^\top y$$

If not invertible:

$$w^* = (X^\top X)^\dagger X^\top y$$

where † denotes pseudo inverse.
This is the analytical solution.

Computational Considerations

Matrix inversion cost:

$$O(d^3)$$

When d is large, this becomes computationally expensive.

Gradient Descent Alternative

Iterative update:

$$w_{t+1} = w_t - \eta_t \nabla f(w_t)$$

Using

$$\nabla f(w_t) = X^\top X w_t - X^\top y$$

No matrix inversion required.
Convexity guarantees convergence to global optimum.

Stochastic Gradient Descent

Full gradient requires entire dataset.
When n is large:

- Sample small subset of data uniformly at random.
- Compute approximate gradient.
- Update:

$$w_{t+1} = w_t$$

- \eta_t \tilde \nabla f(w_t)

$$where \tilde{\nabla} f is gradient using sampled data. Under suitable conditions,$$

\frac{1}{T}
\sum_{t=1}^T w_t
\to
w^*

$$as T \to \infty. - - -$$

Key Takeaways

1. Sum of squared errors is convex.
2. Convexity guarantees global optimality via first order condition.
3. Closed form solution exists.
4. For large scale problems, gradient descent avoids expensive matrix inversion.
5. Stochastic gradient descent scales to massive datasets.
   Convex optimization provides both theoretical guarantees and practical algorithms for machine learning.

****

Lecture 3

Constrained Optimization and Duality

Unconstrained Convex Optimization

If $f : \mathbb{R}^n \to \mathbb{R}$ is convex and differentiable, then

$$\nabla f(x^*) = 0$$

is a first order optimality condition.
For convex f, this condition is sufficient:

$$\nabla f(x^*) = 0 \quad \Longrightarrow \quad x^* \text{ is global optimum}$$

Constrained Optimization Problem

Consider

$$\min_{x \in \mathbb{R}^n} f(x)$$

subject to

$$h(x) \leq 0$$

This is a single inequality constrained optimization problem.

Lagrangian Function

Define the Lagrangian

$$L(x, \lambda) = f(x) + \lambda h(x)$$

where

$$\lambda \geq 0$$

- $x \in \mathbb{R}^n$
- $\lambda \in \mathbb{R}$

Key Observation

Fix x and consider

$$\max_{\lambda \geq 0} L(x, \lambda) = \max_{\lambda \geq 0} \left(f(x) + \lambda h(x) \right)$$

Case 1: Feasible Point

If

$$h(x) \leq 0$$

Since $\lambda \geq 0$,

$$\lambda h(x) \leq 0$$

Thus maximum occurs at

$$\lambda = 0$$

and

$$\max_{\lambda \geq 0} L(x, \lambda) = f(x)$$

Case 2: Infeasible Point

If

$$h(x) > 0$$

Then

$$f(x) + \lambda h(x) \to +\infty$$

as $\lambda \to +\infty$.
Thus

$$\max_{\lambda \geq 0} L(x, \lambda) = +\infty$$

Reformulation of Original Problem

Original problem:

$$\min_{x} f(x) \quad \text{s.t.} \quad h(x) \leq 0$$

Equivalent to

$$\min_{x} \max_{\lambda \geq 0} L(x, \lambda)$$

Reason:

- If x is feasible, value equals $f(x)$.
- If x is infeasible, value equals $+\infty$.

Thus minimizing automatically enforces constraint.

Primal Problem

Define

$$\min_{x} \max_{\lambda \geq 0} L(x, \lambda)$$

Let x^* denote primal optimal solution.
This problem is equivalent to the original constrained problem.

Dual Problem

Swap min and max:

$$\max_{\lambda \geq 0} \min_{x} L(x, \lambda)$$

Define dual function

$$g(\lambda) = \min_{x} L(x, \lambda)$$

Thus dual problem becomes

$$\max_{\lambda \geq 0} g(\lambda)$$

Properties of Dual Function

For each fixed λ,

$$g(\lambda) = \min_{x} \left(f(x) + \lambda h(x) \right)$$

- For fixed x, $L(x, \lambda)$ is affine in λ.
- $g(\lambda)$ is pointwise minimum of affine functions.

Therefore

$$g(\lambda)$$

is a concave function.
Thus dual problem is:

$$\max_{\lambda \geq 0} g(\lambda)$$

which is a concave maximization problem.

Key Insights

1. Original constrained problem can be written as

$$\min_{x} \max_{\lambda \geq 0} L(x, \lambda)$$

2. Swapping order gives dual problem

$$\max_{\lambda \geq 0} \min_{x} L(x, \lambda)$$

3. Dual problem is often easier:
   - Inner minimization is unconstrained.
   - Outer maximization is over $\lambda \geq 0$.
   - Dual function is concave.
     Next step: relate primal and dual solutions and understand when their optimal values coincide.

****

Lecture 4

Relation Between Primal and Dual Problem

Primal Problem

Given the constrained optimization problem

$$\min_{x \in \mathbb{R}^n} f(x) \quad \text{subject to} \quad h(x) \leq 0$$

define the Lagrangian

$$L(x, \lambda) = f(x) + \lambda h(x), \quad \lambda \geq 0$$

The primal formulation is

$$\min_{x} \max_{\lambda \geq 0} L(x, \lambda)$$

Dual Problem

Define

$$g(\lambda) = \min_x L(x, \lambda)$$

The dual problem is

$$\max_{\lambda \geq 0} g(\lambda) = \max_{\lambda \geq 0} \min_x L(x, \lambda)$$

Auxiliary Function

Define

$$J(x) = \begin{cases} f(x) & \text{if } h(x) \leq 0 \\ \infty & \text{otherwise} \end{cases}$$

For any fixed $\lambda \geq 0$,

$$L(x, \lambda) \leq J(x) \quad \forall x$$

Hence,

$$\min_x L(x, \lambda) \leq \min_x J(x)$$

But

$$\min_x J(x) = f(x^*)$$

where x^* is the primal optimal solution.
Taking maximum over λ:

$$\max_{\lambda \geq 0} \min_x L(x, \lambda) \leq f(x^*)$$

Weak Duality

Let λ^* be dual optimal and x^* primal optimal.
Then

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

The value at dual optimum is less than or equal to value at primal optimum.
This is called **weak duality**.

Strong Duality

If

- f is convex
- h is convex
- regularity conditions hold

then

$$\min_x \max_{\lambda \geq 0} L(x, \lambda) = \max_{\lambda \geq 0} \min_x L(x, \lambda)$$

Hence,

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

This is called **strong duality**.

Necessary Conditions Under Strong Duality

Assume f and h are convex and strong duality holds.
Let x^*, λ^* be optimal solutions.

Condition 1: Stationarity

Since

$$f(x^*) = \min_x \left(f(x) + \lambda^* h(x) \right)$$

the gradient must vanish:

$$\nabla f(x^*) + \lambda^* \nabla h(x^*) = 0$$

Condition 2: Complementary Slackness

From equality

$$f(x^*) = f(x^*) + \lambda^* h(x^*)$$

we obtain

$$\lambda^* h(x^*) = 0$$

Condition 3: Primal Feasibility

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

Condition 4: Dual Feasibility

$$\lambda^* \geq 0$$

KKT Conditions (Single Inequality Constraint)

For convex f, h, x^*, λ^* are optimal if and only if:

1. Stationarity

$$\nabla f(x^*) + \lambda^* \nabla h(x^*) = 0$$

2. Complementary Slackness

$$\lambda^* h(x^*) = 0$$

3. Primal Feasibility

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

4. Dual Feasibility

$$\lambda^* \geq 0$$

General KKT Conditions

Consider

$$\min_x f(x)$$

subject to

$$h_i(x) \leq 0, \quad i = 1, \ldots, m$$

$$\ell_j(x) = 0, \quad j = 1, \ldots, n$$

Lagrangian

Introduce multipliers

- $u_i \geq 0$ for inequality constraints
- v_j for equality constraints

$$L(x, u, v) = f(x)$$

- \sum_{i=1}^m u_i h_i(x)
- \sum_{j=1}^n v_j \ell_j(x)

$$- - -$$

KKT Conditions (General Case)

1. Stationarity

$$\nabla f(x^*) + \sum_{i=1}^{m} u_i^* \nabla h_i(x^*) + \sum_{j=1}^{n} v_j^* \nabla \ell_j(x^*) = 0$$

---

2. Complementary Slackness

$$u_i^* h_i(x^*) = 0 \quad \forall i$$

---

3. Primal Feasibility

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

$$\ell_j(x^*) = 0$$

---

4. Dual Feasibility

$$u_i^* \geq 0$$

Interpretation

For convex problems:

- These conditions are **necessary and sufficient**
- Any (x^*, u^*, v^*) satisfying KKT conditions is a global optimum

These are called the **Karush-Kuhn-Tucker conditions**.

****

Lecture 5

Relevance of KKT Conditions in Machine Learning

Support Vector Machine Optimization

Given a dataset

$$\{(x_i, y_i)\}_{i=1}^n$$

where

- $x_i \in \mathbb{R}^d$
- $y_i \in \mathbb{R}$

consider the optimization problem

$$\min_w \frac{1}{2}\|w\|^2$$

subject to

$$y_i\, w^\top x_i \geq 1 \quad \forall i$$

Structure of the Optimization Problem

Objective

$$\frac{1}{2}\|w\|^2 = \frac{1}{2}\sum_{j=1}^d w_j^2$$

This is a quadratic function in w.
Quadratic functions with positive definite Hessian are convex.
Hence:

- Objective is convex.

Constraints

Each constraint

$$y_i \, w^\top x_i \geq 1$$

can be written as

$$1 - y_i w^\top x_i \leq 0$$

This is linear in w.
Linear functions are convex.
Hence:

- Constraints are convex.

Convex Quadratic Program

The SVM formulation is:

- Quadratic objective
- Linear constraints

Therefore:

- Convex optimization problem
- Strong duality holds

Consequence of Strong Duality

Because:

- Objective is convex
- Constraints are convex

Strong duality applies.
Hence:

$$\min_{w} \max_{\lambda \geq 0} L(w, \lambda) = \max_{\lambda \geq 0} \min_{w} L(w, \lambda)$$

Thus:

- Primal and dual optimal values coincide
- Dual problem can be solved instead of primal

Importance of Dual Formulation

Solving the dual:

- Often computationally easier
- Enables kernel methods
- Converts linear models into nonlinear models implicitly

The KKT conditions:

- Characterize optimality
- Provide necessary and sufficient conditions in convex setting
- Used directly in deriving SVM solution

Summary of Optimization Framework

Unconstrained Optimization

If f is convex:

$$\nabla f(x^*) = 0 \Rightarrow x^* \text{ is global minimum}$$

Gradient descent converges to global optimum.

Constrained Optimization

Given convex objective and convex constraints:

- Convert to Lagrangian formulation
- Derive primal and dual problems
- Strong duality holds
- KKT conditions characterize optimality

KKT Conditions Recap

For general problem

$$\min_x f(x)$$

subject to

$$h_i(x) \leq 0, \quad i = 1, \ldots, m$$
$$\ell_j(x) = 0, \quad j = 1, \ldots, n$$

Lagrangian:

$$L(x, u, v) = f(x) + \sum_{i=1}^{m} u_i h_i(x) + \sum_{j=1}^{n} v_j \ell_j(x)$$

Optimal (x^*, u^*, v^*) satisfy:

---

1. Stationarity

$$\nabla f(x^*) + \sum_{i=1}^{m} u_i^* \nabla h_i(x^*) + \sum_{j=1}^{n} v_j^* \nabla \ell_j(x^*) = 0$$

---

2. Complementary Slackness

$$u_i^* h_i(x^*) = 0$$

---

3. Primal Feasibility

$$h_i(x^*) \leq 0$$
$$\ell_j(x^*) = 0$$

---

4. Dual Feasibility

$$u_i^* \geq 0$$

Final Takeaways

1. Convexity simplifies optimization.
2. Duality provides alternative formulations.
3. KKT conditions characterize optimality.
4. Linear regression illustrates unconstrained convex optimization.
5. Support Vector Machines illustrate constrained convex optimization.
6. Dual viewpoint enables powerful extensions such as kernel methods.
   This completes the optimization framework and its connection to machine learning.

****
