Machine Learning and Optimization

Lecture 5: Oracle methods for optimization, Gradient Descent, and Convexity

CSCI 4140/6960 — Spring 2026

Course website: `cs.rpi.edu/~gittea/.../mlandopt.html`

Learning Objectives

By the end of this lecture, you should be able to:

1. **Explain** the oracle framework and how Taylor series enable local approximations
2. **Derive** gradient descent from first-order Taylor approximations using Cauchy-Schwarz
3. **Distinguish** three convergence notions: parameter distance, objective gap, gradient norm
4. **Prove** first-order and second-order characterizations of convex functions
5. **Establish** that norms, affine functions, and log-sum-exp are convex

Lecture Roadmap

Motivation: ML as Optimization

Oracle Methods and Taylor Series

Deriving Gradient Descent

Convergence Notions

Why Convexity Matters

Convex Sets

Convex Functions

Summary and Looking Ahead

Machine Learning as Optimization

Key insight: Most ML algorithms reduce to optimization problems

Three canonical examples:

1. **Linear Regression**:

$$\min_{\mathbf{w}} \frac{1}{n} \sum_{i=1}^{n} (y_i - \mathbf{w}^T \mathbf{x}_i)^2$$

2. **Logistic Regression**:

$$\min_{\mathbf{w}} \frac{1}{n} \sum_{i=1}^{n} \log(1 + e^{-y_i \mathbf{w}^T \mathbf{x}_i})$$

3. **Support Vector Machine**:

$$\min_{\mathbf{w}} \frac{1}{n} \sum_{i=1}^{n} \max\{0, 1 - y_i \mathbf{w}^T \mathbf{x}_i\} + \frac{\lambda}{2} \|\mathbf{w}\|_2^2$$

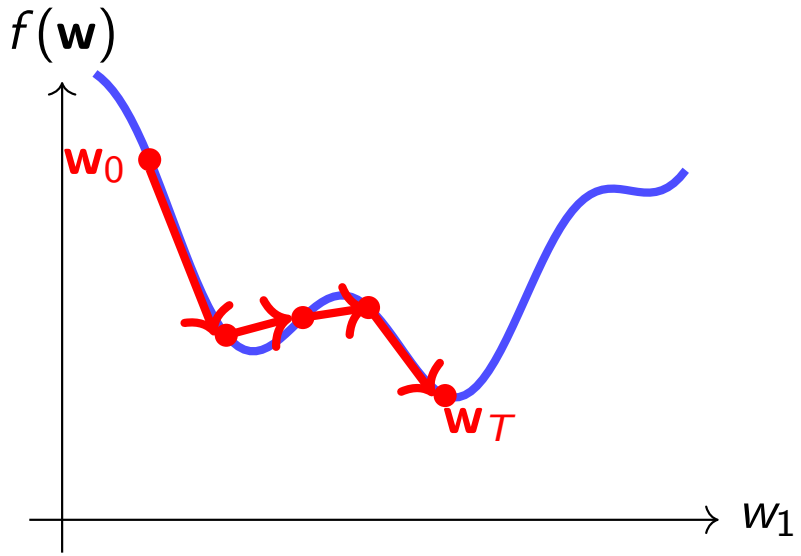
Goal: Find $\mathbf{w}^*$ that minimizes loss function $f(\mathbf{w})$

The Challenge: High-Dimensional Optimization

Problem: Minimize $f(\mathbf{w})$ where $\mathbf{w} \in \mathbb{R}^d$

When d is large (e.g., $d = 10^6$ for image classification):

- **Grid search**: Evaluate f on grid $\rightarrow O(k^d)$ evaluations
  - Even with $k = 2$ points per dimension: $2^{1000000}$ evaluations!
- **Exhaustive search**: Impossible for continuous domains
- **Need**: Iterative methods using *local information* efficiently



Use local info to guide steps

Idea: No global view $\Rightarrow$ model f locally using Taylor series!

Oracle Methods of Optimization

Algorithm structure:

1. Start with initial guess $\mathbf{w}_0 \in \mathbb{R}^d$
2. At iteration t, query **oracle** about f at $\mathbf{w}_t$:
   - **Zeroth-order**: Returns $f(\mathbf{w}_t)$
   - **First-order**: Returns $f(\mathbf{w}_t)$ and $\nabla f(\mathbf{w}_t)$
   - **Second-order**: Returns $f(\mathbf{w}_t)$, $\nabla f(\mathbf{w}_t)$, $\nabla^2 f(\mathbf{w}_t)$
3. Build **local model** of f around $\mathbf{w}_t$
4. Choose next iterate $\mathbf{w}_{t+1}$ based on local model
5. Repeat until convergence

Question: How do we build accurate local models?

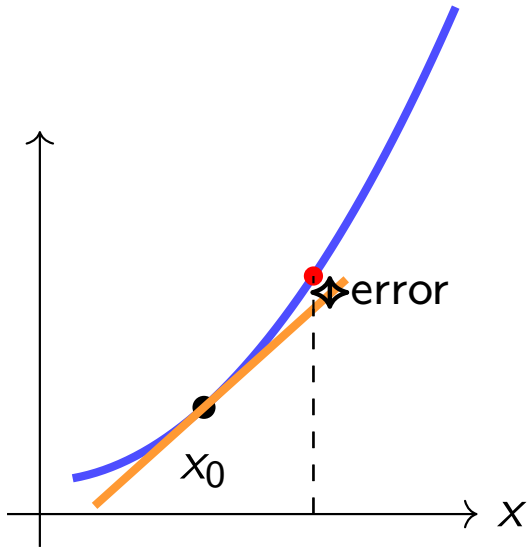
Answer: Multivariate Taylor Series!

Taylor Series: Univariate Case

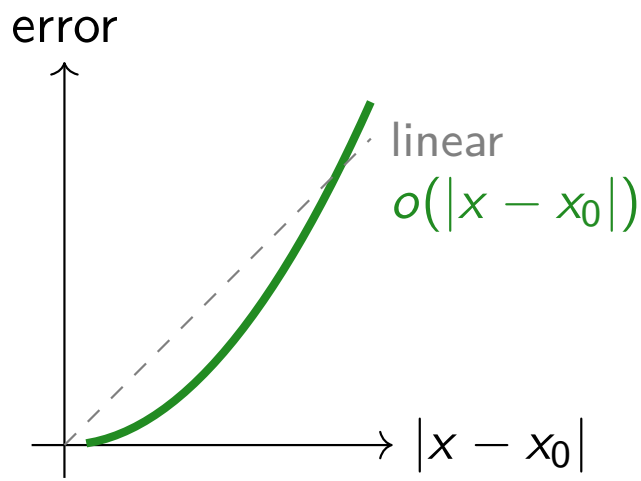
First-order Taylor approximation:

$$f(x) \approx f(x_0) + f'(x_0)(x - x_0)$$

Locally approximate f using the **tangent line** at x_0.



Function vs tangent



Error vanishes superlinearly

Error guarantee:

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + o(|x - x_0|)$$

equivalently, $\lim_{x \to x_0} \frac{|f(x) - f(x_0) - f'(x_0)(x - x_0)|}{|x - x_0|} = 0$

First-Order Multivariate Taylor Series

Goal: Approximate $f(\mathbf{y})$ near $\mathbf{w}$ using f and derivatives at $\mathbf{w}$

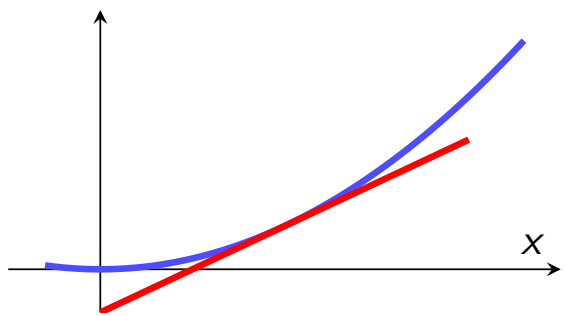
Gradient (first derivatives):

$$\nabla f(\mathbf{w}) = \begin{bmatrix} \frac{\partial f}{\partial w_1} \\ \vdots \\ \frac{\partial f}{\partial w_d} \end{bmatrix} \in \mathbb{R}^d$$

First-order Taylor expansion:

$$\boxed{f(\mathbf{y}) \approx f(\mathbf{w}) + \langle \nabla f(\mathbf{w}), \mathbf{y} - \mathbf{w} \rangle}$$

- Linear approximation — doesn't capture curvature
- Computational cost: $O(d)$ to compute gradient



Error: $f(\mathbf{y}) = f(\mathbf{w}) + \langle \nabla f(\mathbf{w}), \mathbf{y} - \mathbf{w} \rangle + o(\|\mathbf{y} - \mathbf{w}\|_2)$

Second-Order Multivariate Taylor Series

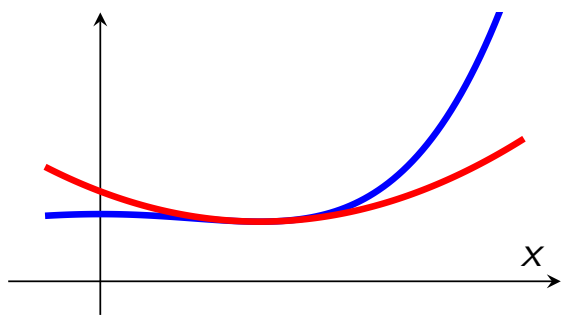
Hessian (second derivatives):

$$\nabla^2 f(\mathbf{w}) = \begin{bmatrix} \frac{\partial^2 f}{\partial w_1 \partial w_1} & \cdots & \frac{\partial^2 f}{\partial w_1 \partial w_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial w_d \partial w_1} & \cdots & \frac{\partial^2 f}{\partial w_d \partial w_d} \end{bmatrix} \in \mathbb{R}^{d \times d}$$

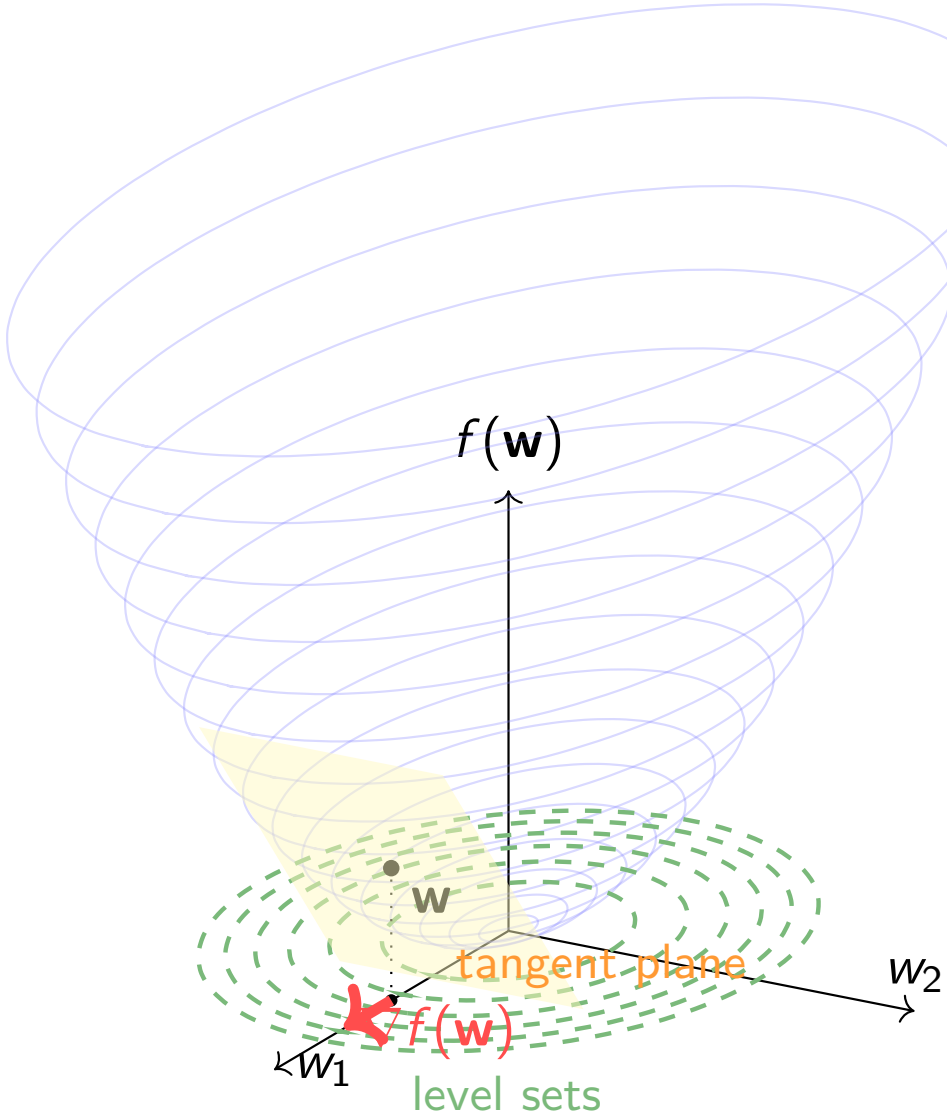
Second-order Taylor expansion:

$$f(\mathbf{y}) \approx f(\mathbf{w}) + \langle \nabla f(\mathbf{w}), \mathbf{y} - \mathbf{w} \rangle + \frac{1}{2} \langle \mathbf{y} - \mathbf{w}, \nabla^2 f(\mathbf{w})(\mathbf{y} - \mathbf{w}) \rangle$$

- Quadratic approximation — captures curvature
- Computational cost: $O(d^2)$ to compute gradient + Hessian, but gives better local model



Visualizing First-Order Taylor Series



First order TS: $f(\mathbf{y}) \approx f(\mathbf{w}) + \langle \nabla f(\mathbf{w}), \mathbf{y} - \mathbf{w} \rangle$

From Taylor Series to Gradient Descent

Setup: Minimize $f(\mathbf{w})$, currently at $\mathbf{w}_t$

First-order TS: $f(\mathbf{w}_{t+1}) \approx f(\mathbf{w}_t) + \langle \nabla f(\mathbf{w}_t), \mathbf{w}_{t+1} - \mathbf{w}_t \rangle$
Let $\boldsymbol{\Delta} = \mathbf{w}_{t+1} - \mathbf{w}_t$. Then:

$$f(\mathbf{w}_{t+1}) \approx f(\mathbf{w}_t) + \langle \nabla f(\mathbf{w}_t), \boldsymbol{\Delta} \rangle$$

Which direction Δ decreases f most?

Minimize $\langle \nabla f(\mathbf{w}_t), \boldsymbol{\Delta} \rangle$ subject to $\|\boldsymbol{\Delta}\|_2 = \alpha$ (Why do we introduce α?)

By Cauchy-Schwarz:

$$\langle \nabla f(\mathbf{w}_t), \boldsymbol{\Delta} \rangle \geq -\|\nabla f(\mathbf{w}_t)\|_2 \cdot \|\boldsymbol{\Delta}\|_2 = -\alpha \|\nabla f(\mathbf{w}_t)\|_2$$

with equality when $\boldsymbol{\Delta} = -\alpha \frac{\nabla f(\mathbf{w}_t)}{\|\nabla f(\mathbf{w}_t)\|_2}$

Absorbing normalization into the stepsize: $\boldsymbol{\Delta} = -\alpha \nabla f(\mathbf{w}_t)$ is the best direction to move in.

Gradient Descent Algorithm

Problem: Minimize $f(\mathbf{w})$ where $\mathbf{w} \in \mathbb{R}^d$

Algorithm 1: Gradient Descent (GD)

Input: Starting point $\mathbf{w}_0 \in \mathbb{R}^d$, step sizes $\{\alpha_t\}_{t=1}^T > 0$, iterations T, gradient oracle ∇f

Output: $\mathbf{w}_T$

for $t = 1, 2, \ldots, T$ **do**

  $\mathbf{w}_t \leftarrow \mathbf{w}_{t-1} - \alpha_t \nabla f(\mathbf{w}_{t-1})$

end

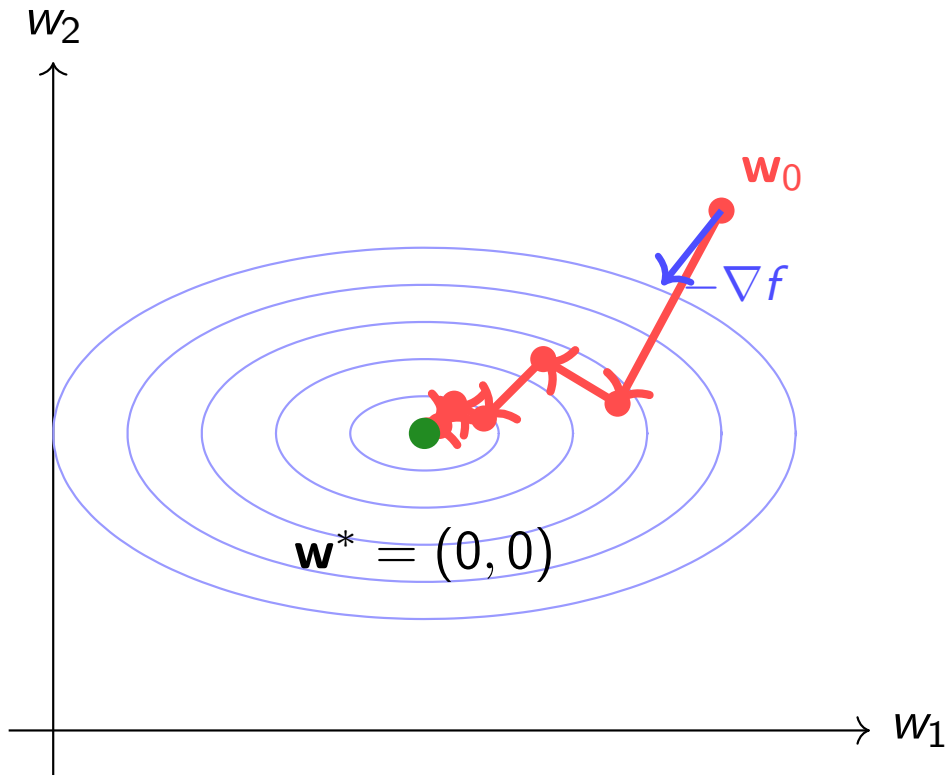
return $\mathbf{w}_T$

Key properties:

- Extremely simple to implement
- Only requires first-order oracle
- Step size α controls movement magnitude

Gradient Descent in Action

Example: $f(w_1, w_2) = w_1^2 + 4w_2^2$ (elliptical bowl)



Zigzag: steps perpendicular to contours

Gradient perpendicular to level sets $\Rightarrow$ GD moves perpendicular to contours

Three Ways to Measure Convergence

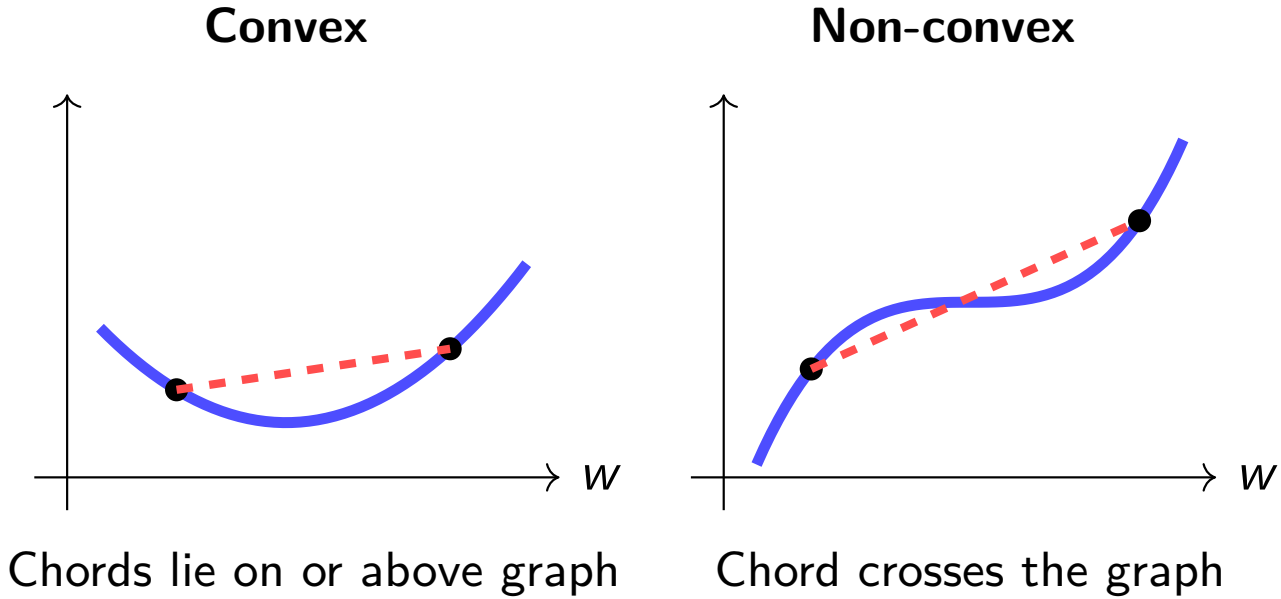
When running GD, what do we want?

Three different measures:

1. **Parameter Distance**: $\|\mathbf{w}_t - \mathbf{w}^*\|_2 < \varepsilon$
   - How close in $\mathbb{R}^d$?
   - Relevant when $\mathbf{w}^*$ has meaning
2. **Objective Gap (Suboptimality)**: $f(\mathbf{w}_t) - f(\mathbf{w}^*) < \varepsilon$
   - How close is objective value to optimal?
   - Most relevant for convex optimization
3. **Gradient Norm**: $\|\nabla f(\mathbf{w}_t)\|_2 < \varepsilon$
   - How close to stationary point? ($\nabla f(\mathbf{w}^*) = 0$ at any local min)
   - Relevant for non-convex optimization, like learning neural networks

Convexity: Local Minimizers = Global Minimizers

Key property: **Convex functions have no spurious local minima**
Every local minimum is a global minimum!

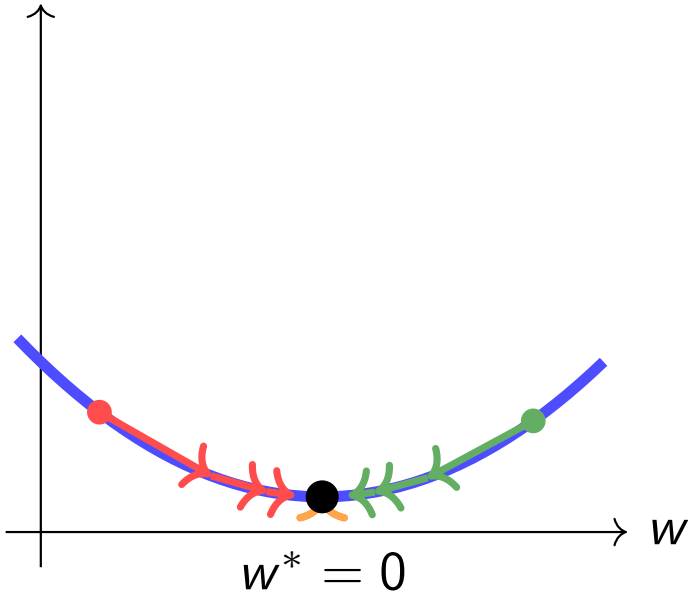


Consequences:

- GD finds global optimum (under mild conditions)
- Can prove convergence rates
- No dependence on initialization (for convex)

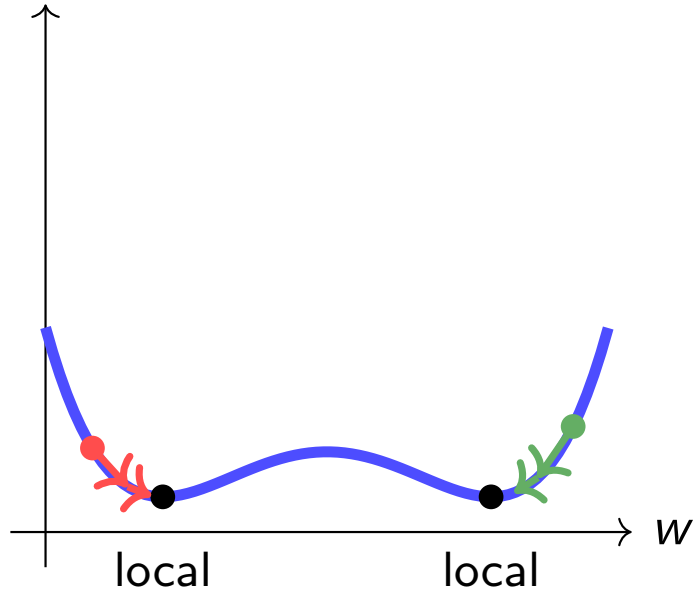
GD on Convex vs Non-Convex Functions

Convex: $f(w) = w^2$



All converge to
global minimum

Non-convex



Outcome depends
on initialization

Convexity enables optimization with strong guarantees

Convex Sets: Definition

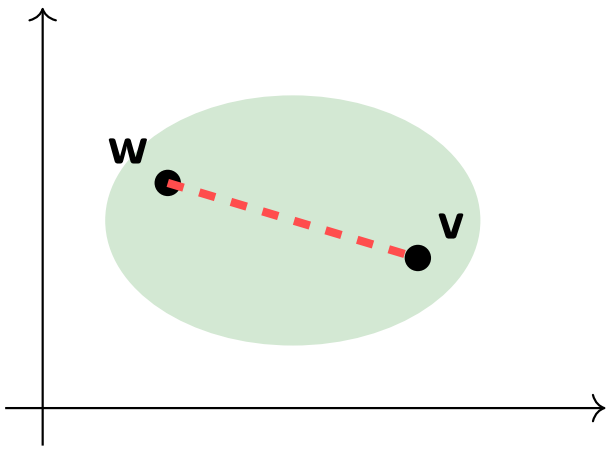
Motivation: Constraint sets in optimization (minimize $f(\mathbf{w})$ s.t. $\mathbf{w} \in C$)

Definition: Set $C \subseteq \mathbb{R}^d$ is **convex** if:

$$\forall \mathbf{w}, \mathbf{v} \in C, \ \forall \alpha \in [0,1]: \quad \alpha\mathbf{w} + (1-\alpha)\mathbf{v} \in C$$

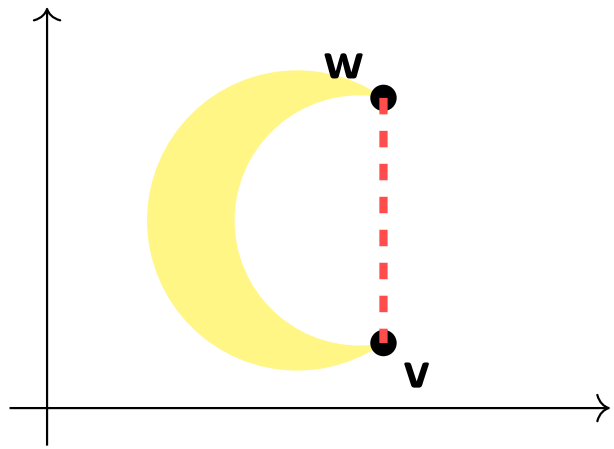
Line segment between any two points stays in C

Convex



Segment inside

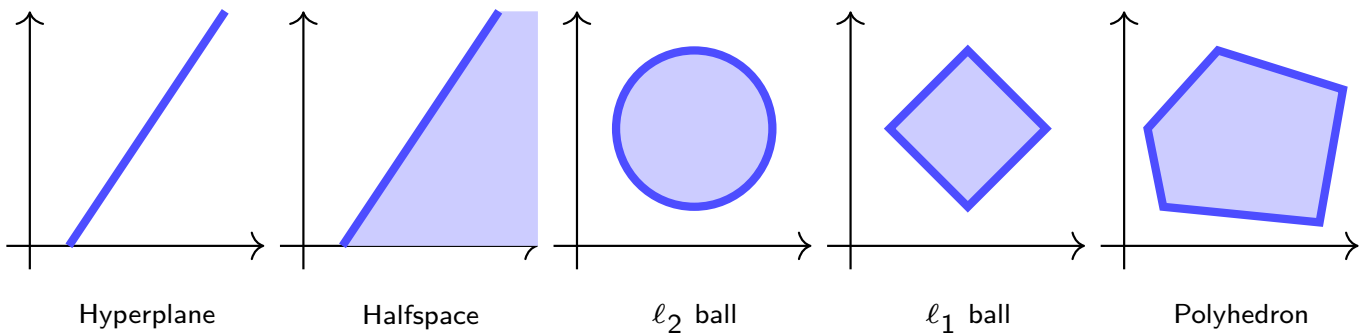
Non-convex



Segment exits

Examples of Convex Sets

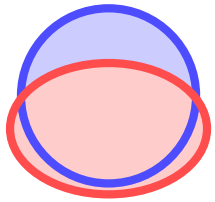
1. **Hyperplanes**: $H = \{\mathbf{w} : \mathbf{a}^T\mathbf{w} = b\}$
2. **Halfspaces**: $H_+ = \{\mathbf{w} : \mathbf{a}^T\mathbf{w} \leq b\}$
3. **Norm balls**: $B_r = \{\mathbf{w} : \|\mathbf{w}\| \leq r\}$ for any norm
   - ℓ_2 ball: $\|\mathbf{w}\|_2 \leq 1$ (Euclidean)
   - ℓ_1 ball: $\|\mathbf{w}\|_1 \leq 1$ (diamond in 2D)
   - ℓ_∞ ball: $\|\mathbf{w}\|_\infty \leq 1$ (hypercube)
4. **Polyhedra**: $P = \{\mathbf{w} : A\mathbf{w} \leq \mathbf{b}\}$ (intersection of halfspaces)



Operations Preserving Convexity

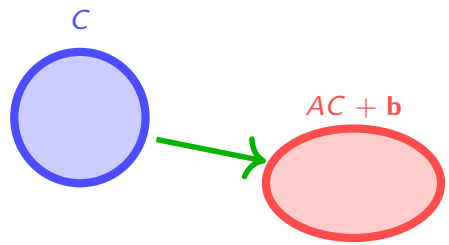
Intersection: $C_1, \ldots, C_n$ convex $\Rightarrow$ $\cap_{i=1}^n C_i$ convex

- Works for infinite intersections
- Polyhedron = intersection of halfspaces



Affine transformation: C convex, $A \in \mathbb{R}^{m \times d}, \mathbf{b} \in \mathbb{R}^m$

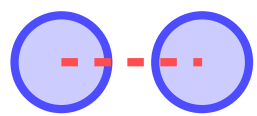
- Image: $\{A\mathbf{w} + \mathbf{b} : \mathbf{w} \in C\}$ convex
- Preimage: $\{\mathbf{w} : A\mathbf{w} + \mathbf{b} \in C\}$ convex



Scaling/translation: C convex, $\alpha \geq 0$ $\Rightarrow \alpha C$, $C + \mathbf{b}$ convex

Special cases of affine transformation, e.g. $A = \alpha I$

Warning: Union of convex sets is usually NOT convex!

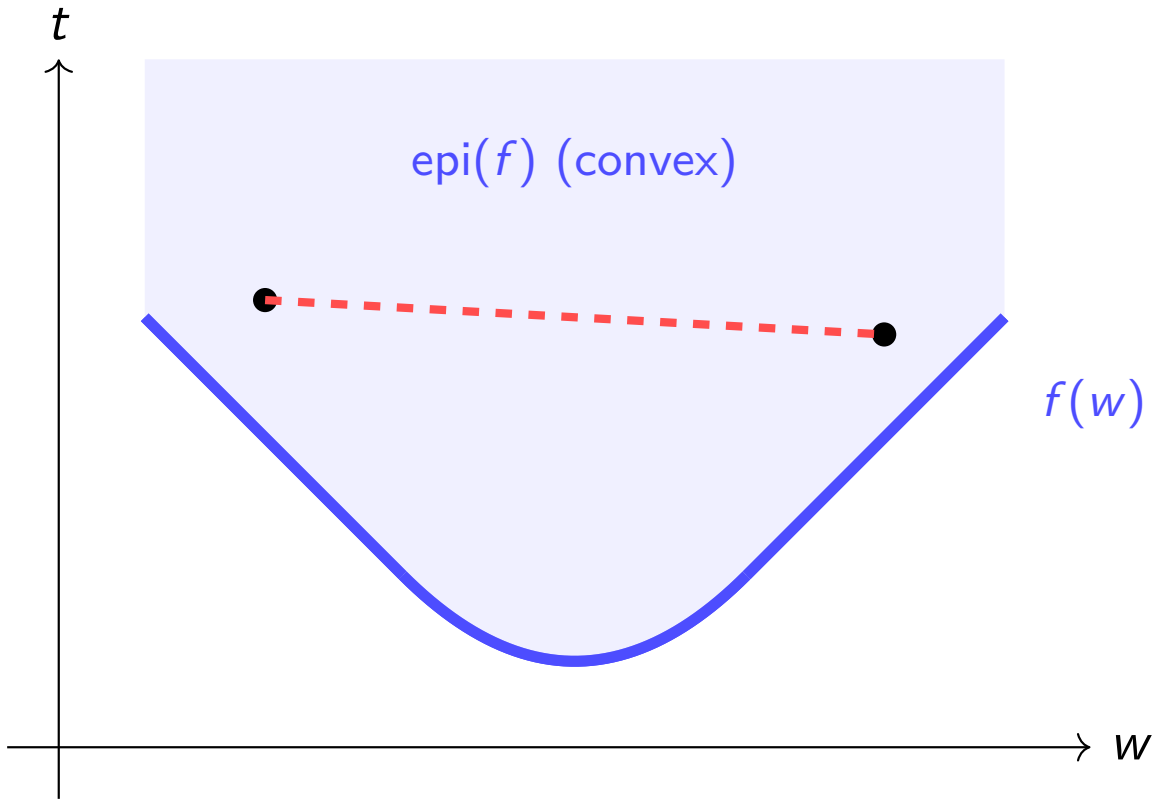


Convex Functions: Epigraph Definition

Definition: The **epigraph** of $f : \mathbb{R}^d \to \mathbb{R}$ is:

$$\text{epi}(f) = \{(\mathbf{w}, t) : \mathbf{w} \in \text{dom}(f),\ t \geq f(\mathbf{w})\}$$

Definition: f is **convex** if $\text{epi}(f)$ is a convex set.



$$f(w) = \begin{cases} \frac{1}{2}w^2 & |w| \leq 1 \\ |w| - \frac{1}{2} & |w| > 1 \end{cases}$$

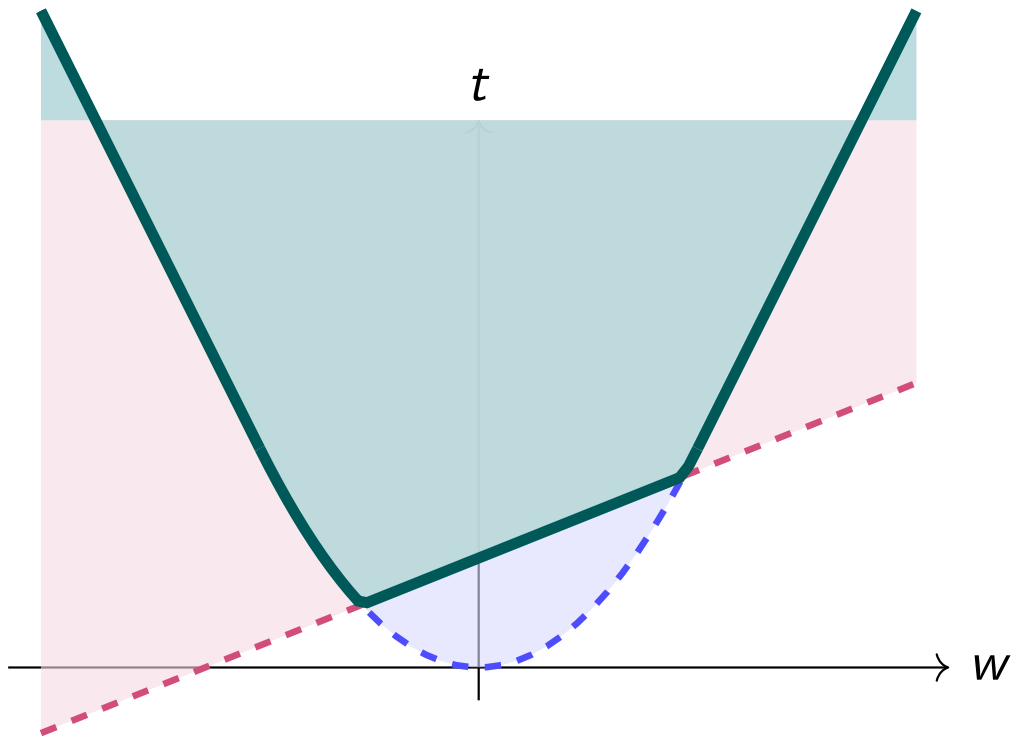
Example: Pointwise Maximum Preserves Convexity

Claim: If $f_1, \ldots, f_n$ are convex, then $f(\mathbf{w}) = \max\{f_1(\mathbf{w}), \ldots, f_n(\mathbf{w})\}$ is convex.

Proof: Use the epigraph characterization.

$$\text{epi}(f) = \{(\mathbf{w}, t) : t \geq \max_i f_i(\mathbf{w})\} = \{(\mathbf{w}, t) : t \geq f_i(\mathbf{w}) \; \forall i\} = \bigcap_{i=1}^{n} \text{epi}(f_i)$$

Since each $\text{epi}(f_i)$ is convex and intersection preserves convexity, $\text{epi}(f)$ is convex.

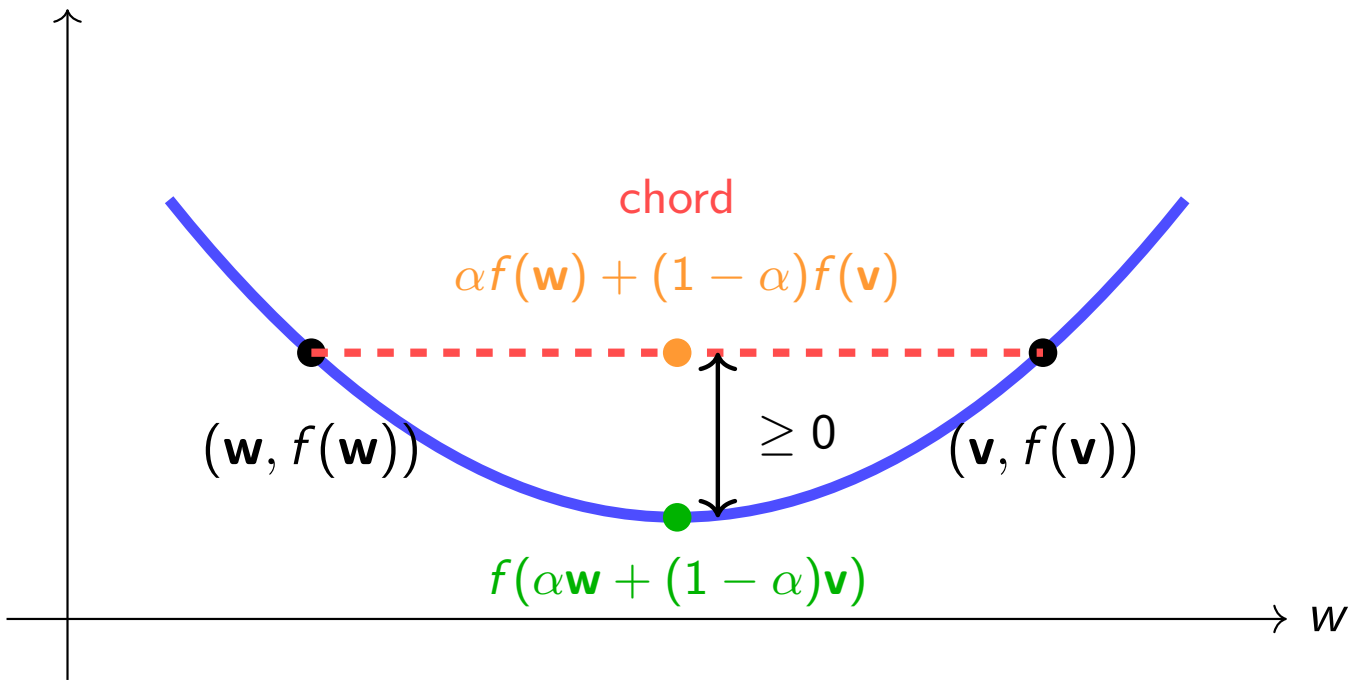


Convex Functions: Equivalent Characterization

Theorem: $f : \mathbb{R}^d \to \mathbb{R}$ is convex iff $\text{dom}(f)$ is convex and:

$$\forall \mathbf{w}, \mathbf{v} \in \text{dom}(f),\ \forall \alpha \in [0,1] : \quad f(\alpha \mathbf{w} + (1-\alpha)\mathbf{v}) \leq \alpha f(\mathbf{w}) + (1-\alpha) f(\mathbf{v})$$

Interpretation: Function lies below its chords.



Example: All Norms are Convex

Claim: Any norm $\|\cdot\|$ is a convex function

Proof via triangle inequality and homogeneity:
For $\alpha \in [0, 1]$:

$$\begin{aligned} \|\alpha \mathbf{w} + (1-\alpha)\mathbf{v}\| &\leq \|\alpha \mathbf{w}\| + \|(1-\alpha)\mathbf{v}\| && \text{(triangle inequality)} \\ &= \alpha\|\mathbf{w}\| + (1-\alpha)\|\mathbf{v}\| && \text{(positive homogeneity)} \end{aligned}$$

Therefore $f(\mathbf{w}) = \|\mathbf{w}\|$ satisfies the convexity inequality. □

Examples:

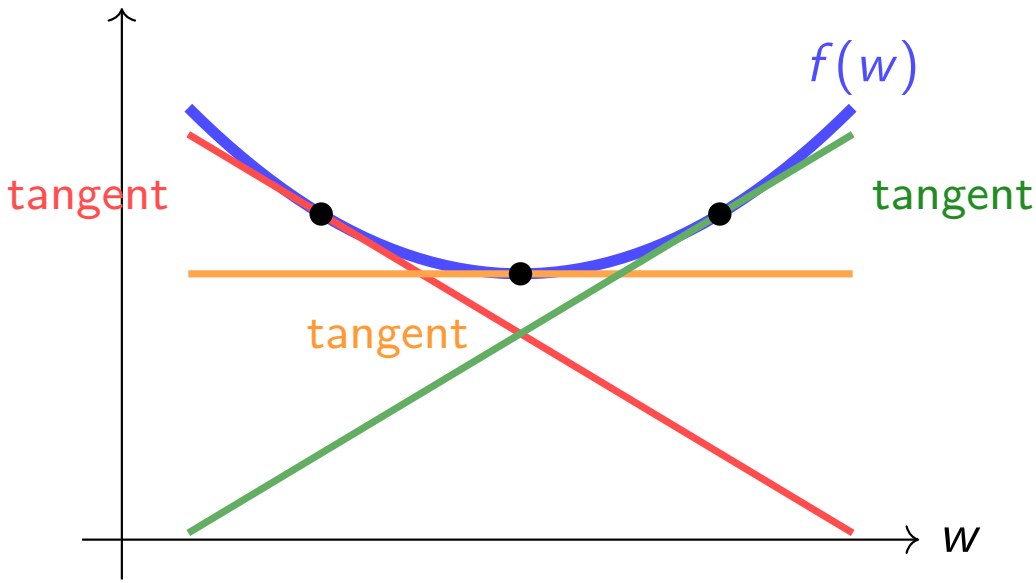
- ℓ_1 norm: $\|\mathbf{w}\|_1 = \sum_{i=1}^{d} |w_i|$ (convex)
- ℓ_2 norm: $\|\mathbf{w}\|_2 = \sqrt{\sum_{i=1}^{d} w_i^2}$ (convex)
- ℓ_∞ norm: $\|\mathbf{w}\|_\infty = \max_i |w_i|$ (convex)

First-Order Characterization of Convexity

Theorem: Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable on convex domain. Then f is convex if and only if:

$$\boxed{\forall \mathbf{w}, \mathbf{v} \in \mathrm{dom}(f): \quad f(\mathbf{v}) \geq f(\mathbf{w}) + \langle \nabla f(\mathbf{w}), \mathbf{v} - \mathbf{w} \rangle}$$

Geometric interpretation: Function lies above all tangent hyperplanes
First-order Taylor is global underestimator!



Function above all tangents

Second-Order Characterization of Convexity

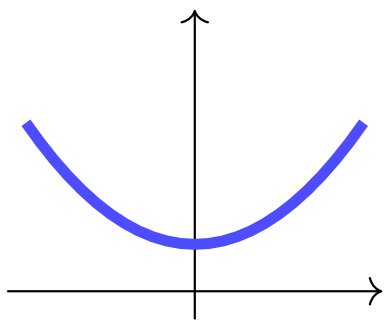
Theorem: Let $f : \mathbb{R}^d \to \mathbb{R}$ be twice differentiable on open convex domain. Then f is convex if and only if:

$$\boxed{\forall \mathbf{w} \in \text{dom}(f) : \quad \nabla^2 f(\mathbf{w}) \succeq 0}$$

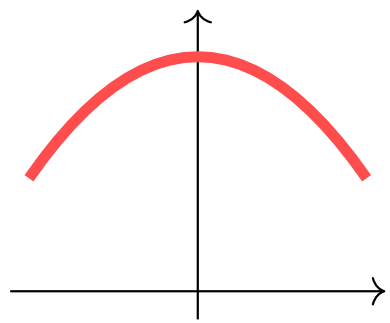
Hessian is positive semidefinite (PSD) everywhere

Recall: $M \succeq 0$ means $\mathbf{v}^T M \mathbf{v} \geq 0$ for all $\mathbf{v} \in \mathbb{R}^d$

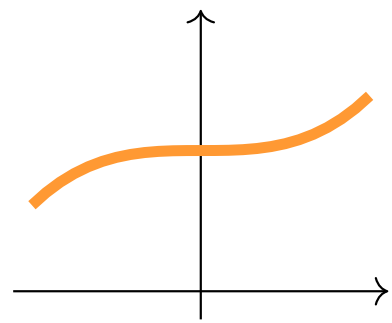
1D case: f convex $\Leftrightarrow f''(w) \geq 0$ for all w



$f''(w) = 2 > 0$



$f''(w) = -2 < 0$



$f''(w) = 6w$ varies

Example: Log-Sum-Exp is Convex

Claim: $f(\mathbf{w}) = \log(\sum_{i=1}^{d} e^{w_i})$ is convex

Proof via second-order condition:

Gradient:

$$\frac{\partial f}{\partial w_j} = \frac{e^{w_j}}{\sum_i e^{w_i}} \quad \text{(softmax)}$$

Hessian:

$$\frac{\partial^2 f}{\partial w_j \partial w_k} = \frac{\delta_{jk} e^{w_j} \sum_i e^{w_i} - e^{w_j} e^{w_k}}{(\sum_i e^{w_i})^2}$$

Let $p_i = e^{w_i} / \sum_j e^{w_j}$ (softmax probabilities). Then:

$$\nabla^2 f(\mathbf{w}) = \text{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^T$$

For any $\mathbf{v} \in \mathbb{R}^d$:

$$\begin{aligned}
\mathbf{v}^T \nabla^2 f(\mathbf{w}) \mathbf{v} &= \mathbf{v}^T \text{diag}(\mathbf{p}) \mathbf{v} - \mathbf{v}^T \mathbf{p}\mathbf{p}^T \mathbf{v} \\
&= \sum_i p_i v_i^2 - (\sum_i p_i v_i)^2 \\
&= \mathbb{E}_{\mathbf{p}}[V^2] - (\mathbb{E}_{\mathbf{p}}[V])^2 = \text{Var}_{\mathbf{p}}(V) \geq 0
\end{aligned}$$

where V is r.v. with $\mathbb{P}(V = v_i) = p_i$. Therefore $\nabla^2 f \succeq 0$. □

Examples of Convex Functions

1. **Affine**: $f(\mathbf{w}) = \mathbf{a}^T\mathbf{w} + b$ (convex AND concave: $\nabla^2 f = 0$)
2. **Quadratic**: $f(\mathbf{w}) = \frac{1}{2}\mathbf{w}^T Q\mathbf{w} + \mathbf{b}^T\mathbf{w} + c$ (convex if $Q \succeq 0$)
3. **Norms**: $f(\mathbf{w}) = \|\mathbf{w}\|$ for any norm
4. **Exponential**: $f(w) = e^w$ on $\mathbb{R}$ ($f''(w) = e^w > 0$)
5. **Negative log**: $f(w) = -\log(w)$ on $\mathbb{R}_{++}$ ($f''(w) = 1/w^2 > 0$)
6. **Log-sum-exp**: $f(\mathbf{w}) = \log(\sum_{i=1}^d e^{w_i})$

All appear in ML loss functions and regularizers! You have everything you need to prove their convexity.

Proof: Affine Functions are Convex

Claim: $f(\mathbf{w}) = \mathbf{a}^T\mathbf{w} + b$ is convex (and concave)

Proof via second-order condition:

$$\nabla f(\mathbf{w}) = \mathbf{a}, \quad \nabla^2 f(\mathbf{w}) = \mathbf{0}$$

Since $\mathbf{v}^T\mathbf{0}\mathbf{v} = 0$ for all $\mathbf{v}$, we have $\nabla^2 f(\mathbf{w}) \succeq 0$ (and $\preceq 0$).
Therefore f is convex (and concave).

Alternative proof via inequality definition:
For $\alpha \in [0, 1]$:

$$\begin{aligned} f(\alpha\mathbf{w} + (1-\alpha)\mathbf{v}) &= \mathbf{a}^T(\alpha\mathbf{w} + (1-\alpha)\mathbf{v}) + b \\ &= \alpha(\mathbf{a}^T\mathbf{w} + b) + (1-\alpha)(\mathbf{a}^T\mathbf{v} + b) \\ &= \alpha f(\mathbf{w}) + (1-\alpha) f(\mathbf{v}) \end{aligned}$$

Equality holds, so both convex and concave. □

Operations Preserving Convexity

1. **Non-negative weighted sum**: $f_1, \ldots, f_n$ convex, $\alpha_i \geq 0 \Rightarrow \sum \alpha_i f_i$ convex
2. **Pointwise maximum**: $f_1, \ldots, f_n$ convex $\Rightarrow \max_i f_i(\mathbf{w})$ convex
   - Example: Hinge loss $(1 - \hat{y}y)_+ = \max\{0, 1 - \hat{y}y\}$
3. **Composition with affine**: f convex $\Rightarrow g(\mathbf{w}) = f(A\mathbf{w} + \mathbf{b})$ convex
4. **Composition rules**: f convex, g convex + non-decreasing on the codomain of $f \Rightarrow g \circ f$ convex
   - Example: $e^{f(\mathbf{w})}$ if f convex

Exercise: These are very useful. Prove these with the various characterizations of convexity!

Many ML Loss Functions are Convex in $\mathbf{w}$

1. **Squared loss**: $\frac{1}{n}\sum_i(\mathbf{w}^T\mathbf{x}_i - y_i)^2$
   Convex (quadratic with PSD Hessian)
2. **Logistic loss**: $\frac{1}{n}\sum_i \log(1 + e^{-y_i\mathbf{w}^T\mathbf{x}_i})$
   Convex ($\nabla^2 f \succeq 0$, composition of convex function with affine)
3. **Hinge loss**: $\frac{1}{n}\sum_i \max\{0, 1 - y_i\mathbf{w}^T\mathbf{x}_i\}$
   Convex (max of affine functions)
4. **With regularization**:
   - $\text{loss}(\mathbf{w}) + \lambda\|\mathbf{w}\|_2^2$ (ridge)
   - $\text{loss}(\mathbf{w}) + \lambda\|\mathbf{w}\|_1$ (lasso)

   Sum of convex $\rightarrow$ convex

Implication: GD guaranteed to find global minimum!

Summary: What We've Established

From Local Models to Gradient Descent

- Oracle methods use local Taylor approximations
- GD derived from steepest descent direction via Cauchy-Schwarz
- Three convergence notions; focus on objective gap for convex

Convexity - When GD Works

- Convex sets
- Rules for recognizing convex sets
- Convex functions: epigraph, convex inequality, first-order (above tangents), second-order ($\nabla^2 f \succeq 0$)
- Proved affine, norms, log-sum-exp are convex
- Rules for recognizing convex functions

Questions?

Next time: convergence of gradient descent on smooth convex functions

Appendix: Proof of First Order Characterization (1/2)

Proof ($\Rightarrow$): Convexity implies first-order condition

Setup: Fix arbitrary points $\mathbf{w}, \mathbf{v} \in \text{dom}(f)$ and $\alpha \in (0, 1)$.

Step 1: Apply convexity definition

$$f(\mathbf{w} + \alpha(\mathbf{v} - \mathbf{w})) \leq (1 - \alpha)f(\mathbf{w}) + \alpha f(\mathbf{v})$$

Step 2: Rearrange to isolate $f(\mathbf{v})$ term

$$f(\mathbf{w} + \alpha(\mathbf{v} - \mathbf{w})) - f(\mathbf{w}) \leq \alpha(f(\mathbf{v}) - f(\mathbf{w}))$$

Step 3: Divide by $\alpha > 0$ (inequality preserved)

$$\frac{f(\mathbf{w} + \alpha(\mathbf{v} - \mathbf{w})) - f(\mathbf{w})}{\alpha} \leq f(\mathbf{v}) - f(\mathbf{w})$$

Step 4: Take limit $\alpha \to 0$. The left side is the directional derivative in direction $\mathbf{v} - \mathbf{w}$:

$$\langle \nabla f(\mathbf{w}), \mathbf{v} - \mathbf{w} \rangle \leq f(\mathbf{v}) - f(\mathbf{w})$$

Since $\mathbf{w}, \mathbf{v}$ were arbitrary, this holds for all points. □

Appendix: Proof of First Order Characterization (2/2)

Proof ($\Leftarrow$): First-order condition implies convexity

Goal: Show $f(\alpha \mathbf{w} + (1-\alpha)\mathbf{v}) \leq \alpha f(\mathbf{w}) + (1-\alpha) f(\mathbf{v})$ for all $\alpha \in [0,1]$.

Step 1: Define the convex combination. Let

$$\mathbf{u} = \alpha \mathbf{w} + (1-\alpha)\mathbf{v}$$

be the point we want to bound.

Step 2: Apply first-order condition at $\mathbf{u}$ to both $\mathbf{w}$ and $\mathbf{v}$

$$f(\mathbf{w}) \geq f(\mathbf{u}) + \langle \nabla f(\mathbf{u}), \mathbf{w} - \mathbf{u} \rangle$$
$$f(\mathbf{v}) \geq f(\mathbf{u}) + \langle \nabla f(\mathbf{u}), \mathbf{v} - \mathbf{u} \rangle$$

Step 3: Take convex combination of these inequalities. Multiply first by α, second by $(1-\alpha)$, then add:

$$\alpha f(\mathbf{w}) + (1-\alpha) f(\mathbf{v}) \geq f(\mathbf{u}) + \langle \nabla f(\mathbf{u}), \alpha(\mathbf{w} - \mathbf{u}) + (1-\alpha)(\mathbf{v} - \mathbf{u}) \rangle.$$

Step 4: Observe that the red term equals $\mathbf{0}$: since $\mathbf{u} = \alpha \mathbf{w} + (1-\alpha)\mathbf{v}$, we have $\alpha(\mathbf{w} - \mathbf{u}) + (1-\alpha)(\mathbf{v} - \mathbf{u}) = \mathbf{0}$.

Therefore: $\alpha f(\mathbf{w}) + (1-\alpha) f(\mathbf{v}) \geq f(\mathbf{u})$. $\square$

Appendix: Proof of Second-Order Characterization (1/3)

Proof ($\Rightarrow$): Assume f convex, twice differentiable. Show $\nabla^2 f(\mathbf{w}) \succeq 0$.

Step 1: Apply first-order condition twice

Since f is convex, for any $\mathbf{w}, \mathbf{v}$:

$$f(\mathbf{v}) \geq f(\mathbf{w}) + \langle \nabla f(\mathbf{w}), \mathbf{v} - \mathbf{w} \rangle$$
$$f(\mathbf{w}) \geq f(\mathbf{v}) + \langle \nabla f(\mathbf{v}), \mathbf{w} - \mathbf{v} \rangle$$

Step 2: Add the inequalities

$$f(\mathbf{v}) + f(\mathbf{w}) \geq f(\mathbf{w}) + \langle \nabla f(\mathbf{w}), \mathbf{v} - \mathbf{w} \rangle + f(\mathbf{v}) + \langle \nabla f(\mathbf{v}), \mathbf{w} - \mathbf{v} \rangle$$

Canceling $f(\mathbf{v}) + f(\mathbf{w})$ from both sides:

$$0 \geq \langle \nabla f(\mathbf{w}), \mathbf{v} - \mathbf{w} \rangle + \langle \nabla f(\mathbf{v}), \mathbf{w} - \mathbf{v} \rangle$$

Equivalently:

$$0 \geq \langle \nabla f(\mathbf{w}) - \nabla f(\mathbf{v}), \mathbf{v} - \mathbf{w} \rangle$$

Appendix: Proof of Second-Order Characterization (2/3)

Step 3: Introduce local coordinates

Let $\boldsymbol{\Delta} = \mathbf{v} - \mathbf{w}$, so $\mathbf{v} = \mathbf{w} + \boldsymbol{\Delta}$.

The inequality becomes:

$$0 \geq \langle \nabla f(\mathbf{w}) - \nabla f(\mathbf{w} + \boldsymbol{\Delta}), \boldsymbol{\Delta} \rangle$$

Step 4: Apply Taylor expansion of gradient

Since f is twice differentiable:

$$\nabla f(\mathbf{w} + \boldsymbol{\Delta}) = \nabla f(\mathbf{w}) + \nabla^2 f(\mathbf{w})\boldsymbol{\Delta} + o(\|\boldsymbol{\Delta}\|_2)$$

Substituting into our inequality:

$$\begin{aligned} 0 &\geq \langle \nabla f(\mathbf{w}) - (\nabla f(\mathbf{w}) + \nabla^2 f(\mathbf{w})\boldsymbol{\Delta} + o(\|\boldsymbol{\Delta}\|_2)), \boldsymbol{\Delta} \rangle \\ &= \langle -\nabla^2 f(\mathbf{w})\boldsymbol{\Delta} - o(\|\boldsymbol{\Delta}\|_2), \boldsymbol{\Delta} \rangle \\ &= -\langle \nabla^2 f(\mathbf{w})\boldsymbol{\Delta}, \boldsymbol{\Delta} \rangle + o(\|\boldsymbol{\Delta}\|_2^2) \end{aligned}$$

Appendix: Proof of Second-Order Characterization (3/3)

Step 5: Normalize and take limit

From the previous slide:

$$0 \geq -\langle \nabla^2 f(\mathbf{w})\boldsymbol{\Delta}, \boldsymbol{\Delta} \rangle + o(\|\boldsymbol{\Delta}\|_2^2)$$

Let $\mathbf{d} = \boldsymbol{\Delta}/\|\boldsymbol{\Delta}\|_2$ (unit direction vector).
Divide both sides by $\|\boldsymbol{\Delta}\|_2^2 > 0$:

$$0 \geq -\langle \nabla^2 f(\mathbf{w})\mathbf{d}, \mathbf{d} \rangle + \frac{o(\|\boldsymbol{\Delta}\|_2^2)}{\|\boldsymbol{\Delta}\|_2^2}$$

As $\|\boldsymbol{\Delta}\|_2 \to 0$, the second term vanishes:

$$0 \geq -\langle \nabla^2 f(\mathbf{w})\mathbf{d}, \mathbf{d} \rangle$$

Therefore: $\langle \nabla^2 f(\mathbf{w})\mathbf{d}, \mathbf{d} \rangle \geq 0$.

Since $\mathbf{v}$ was arbitrary, $\mathbf{d}$ represents any unit direction.
Thus $\nabla^2 f(\mathbf{w}) \succeq 0$ for all $\mathbf{w}$. □

Appendix: Proof of Second-Order Characterization ($\Leftarrow$, 1/2)

Proof ($\Leftarrow$): Assume $\nabla^2 f(\mathbf{w}) \succeq 0$ for all $\mathbf{w}$. Show f is convex.

Strategy: Prove the first-order condition using Taylor's theorem with integral remainder.

Step 1: Taylor expansion with integral remainder
For any $\mathbf{w}, \mathbf{v}$, by the fundamental theorem of calculus:

$$f(\mathbf{v}) = f(\mathbf{w}) + \int_0^1 \langle \nabla f(\mathbf{w} + t(\mathbf{v} - \mathbf{w})), \mathbf{v} - \mathbf{w} \rangle \, dt$$

This follows from: $\frac{d}{dt} f(\mathbf{w} + t(\mathbf{v} - \mathbf{w})) = \langle \nabla f(\mathbf{w} + t(\mathbf{v} - \mathbf{w})), \mathbf{v} - \mathbf{w} \rangle$

Step 2: Expand the gradient using Taylor's theorem
For each $t \in [0, 1]$:

$$\nabla f(\mathbf{w} + t(\mathbf{v} - \mathbf{w})) = \nabla f(\mathbf{w}) + \int_0^t \nabla^2 f(\mathbf{w} + s(\mathbf{v} - \mathbf{w}))(\mathbf{v} - \mathbf{w}) \, ds$$

Appendix: Proof of Second-Order Characterization ($\Leftarrow$, 2/2)

Step 3: Combine the expansions

Substituting the gradient expansion into the integral for $f(\mathbf{v})$:

$$f(\mathbf{v}) = f(\mathbf{w}) + \int_0^1 \langle \nabla f(\mathbf{w}), \mathbf{v} - \mathbf{w} \rangle \, dt + \int_0^1 \left\langle \int_0^t \nabla^2 f(\mathbf{w} + s(\mathbf{v} - \mathbf{w}))(\mathbf{v} - \mathbf{w}) \, ds, \mathbf{v} - \mathbf{w} \right\rangle dt$$

The first integral equals $\langle \nabla f(\mathbf{w}), \mathbf{v} - \mathbf{w} \rangle$.

Step 4: Use PSD assumption

The second term can be written:

$$\int_0^1 \int_0^t \langle \nabla^2 f(\mathbf{w} + s(\mathbf{v} - \mathbf{w}))(\mathbf{v} - \mathbf{w}), \mathbf{v} - \mathbf{w} \rangle \, ds \, dt$$

Since $\nabla^2 f(\mathbf{w} + s(\mathbf{v} - \mathbf{w})) \succeq 0$ for all $s \in [0, 1]$:

$$\langle \nabla^2 f(\mathbf{w} + s(\mathbf{v} - \mathbf{w}))(\mathbf{v} - \mathbf{w}), \mathbf{v} - \mathbf{w} \rangle \geq 0$$

Therefore: $f(\mathbf{v}) \geq f(\mathbf{w}) + \langle \nabla f(\mathbf{w}), \mathbf{v} - \mathbf{w} \rangle$.