

MATH 565: Lecture 5 (01/27/2026)

Today: * More results on convexity
* Details of gradient descent

Recall: $f: \Omega \rightarrow \mathbb{R}$ for convex $\Omega \subseteq \mathbb{R}^d$ is convex if $\forall \bar{w}_1, \bar{w}_2 \in \Omega, \lambda \in [0, 1]$,

$$f(\lambda \bar{w}_1 + (1-\lambda) \bar{w}_2) \leq \lambda f(\bar{w}_1) + (1-\lambda) f(\bar{w}_2) \quad (*)$$

We can use convexity to show that every local minimum of a convex function must be a global minimum!

Proof

Let \bar{w}_0 be a local minimum of $f(\bar{w})$.

$$\Rightarrow f(\bar{w}) \geq f(\bar{w}_0) \quad \forall \|\bar{w} - \bar{w}_0\| \leq \epsilon \quad (1)$$

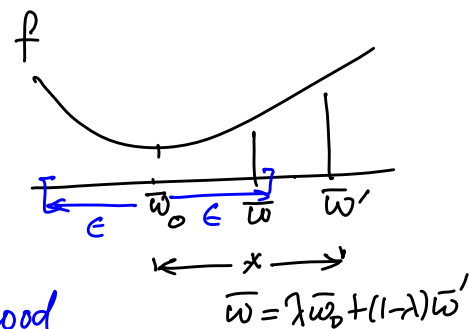
Consider \bar{w}' such that $\|\bar{w}' - \bar{w}_0\| > \epsilon$.

Let $\lambda \in (0, 1)$ be such that

$$\bar{w} = \lambda \bar{w}_0 + (1-\lambda) \bar{w}' \quad \text{has} \quad \|\bar{w} - \bar{w}_0\| \leq \epsilon.$$

small enough neighborhood of \bar{w}_0

\bar{w}' is outside the local neighborhood of \bar{w}_0 , but \bar{w} is within the same.



f is convex \Rightarrow

$$f(\underbrace{\lambda \bar{w}_0 + (1-\lambda) \bar{w}'}_{\bar{w}}) \leq \lambda f(\bar{w}_0) + (1-\lambda) f(\bar{w}') \quad (2)$$

by (*)

Combining (1) and (2) \Rightarrow

$$f(\bar{w}_0) \leq f(\bar{w}) \leq \lambda f(\bar{w}_0) + (1-\lambda) f(\bar{w}')$$

$$\Rightarrow (1-\lambda) f(\bar{w}_0) \leq (1-\lambda) f(\bar{w}')$$

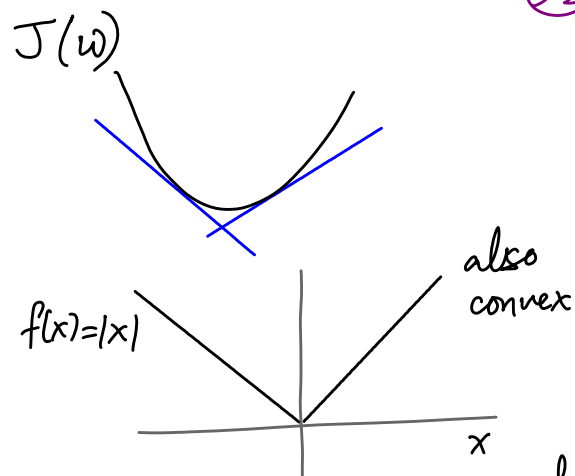
$$\Rightarrow f(\bar{w}') \geq f(\bar{w}_0)$$

$\Rightarrow \bar{w}_0$ is a global minimum.

□

An Observation

A convex function f lies above any tangent to f . We can use this observation to derive a characterization of convex functions in terms of ∇f .



Lemma 4 A differentiable function $f(\bar{w}) : \Omega \rightarrow \mathbb{R}$ for convex $\Omega \subseteq \mathbb{R}^d$ is a convex function if and only if

$$f(\bar{w}) \geq f(\bar{w}_0) + [\nabla f(\bar{w}_0)]^T (\bar{w} - \bar{w}_0) \quad \text{--- (I)}$$

$\forall \bar{w}, \bar{w}_0 \in \Omega$

\hookrightarrow first derivative condition of convexity

Proof (\Rightarrow) Assume f is convex \Rightarrow

$$\forall \bar{w}, \bar{w}_0 \in \Omega, \lambda \in [0, 1]$$

$$f((1-\lambda)\bar{w}_0 + \lambda\bar{w}) \leq (1-\lambda)f(\bar{w}_0) + \lambda f(\bar{w})$$

\nearrow same as (*), but with λ and $1-\lambda$ flipped

$$\Rightarrow f(\bar{w}_0 + \lambda(\bar{w} - \bar{w}_0)) - f(\bar{w}_0) \leq \lambda (f(\bar{w}) - f(\bar{w}_0))$$

Hence, for $\lambda > 0$, we get

$$\frac{f(\bar{w}_0 + \lambda(\bar{w} - \bar{w}_0)) - f(\bar{w}_0)}{\lambda} \leq f(\bar{w}) - f(\bar{w}_0).$$

Taking the limit of LHS as $\lambda \rightarrow 0^+$ gives the directional derivative of f at \bar{w}_0 in the direction of $\bar{w} - \bar{w}_0$.

$$\Rightarrow [\nabla f(\bar{w}_0)]^T (\bar{w} - \bar{w}_0) \leq f(\bar{w}) - f(\bar{w}_0)$$

$$\Rightarrow f(\bar{w}) \geq f(\bar{w}_0) + \nabla f(\bar{w}_0)^T (\bar{w} - \bar{w}_0)$$

(\Leftarrow) proof: you can do it yourselves!

□

Note: This definition assumes that f is differentiable.
 $\rightarrow f(x)=|x|$ is not differentiable, but is convex!


But still, going one step further, we can give a characterization based on the second derivative of f .

Lemma 5 A twice differentiable function $f: \Omega \rightarrow \mathbb{R}$ for convex $\Omega \subseteq \mathbb{R}^d$ is convex iff $Hf(\bar{w}) \succeq 0 \quad \forall \bar{w} \in \Omega$. \rightarrow second derivative condition of convexity


Note: Lemma 4 says that if $\nabla f(\bar{w}_0) = \bar{0}$, then $f(\bar{w}) \geq f(\bar{w}_0) \quad \forall \bar{w}$, directly implying that \bar{w}_0 is a global optimum.

Strict Convexity \rightarrow a stronger notion of convexity

Def $f: \Omega \rightarrow \mathbb{R}$ for convex $\Omega \subseteq \mathbb{R}^d$ is strictly convex if $f(\lambda \bar{w}_1 + (1-\lambda)\bar{w}_2) < \lambda f(\bar{w}_1) + (1-\lambda)f(\bar{w}_2) \quad \forall \bar{w}_1, \bar{w}_2 \in \Omega, \lambda \in [0, 1]$.



convex



strictly convex

Lemma 6 A strictly convex function can contain at most one critical point. If such a critical point exists, it is the global minimum of the function. $y = e^x$: strictly convex but has no critical point.

Lemma 7 Let $f: \Omega \rightarrow \mathbb{R}$ be convex and $g: \Omega \rightarrow \mathbb{R}$ be strictly convex for convex $\Omega \subseteq \mathbb{R}^d$. Then $h: \Omega \rightarrow \mathbb{R}$ defined as $h = f + g$ is strictly convex.

This property is used in many ML settings — we add a strictly convex term to convex loss J to make it strictly convex.

For instance, Tikhonov regularization:

$$J = \underbrace{\frac{1}{2} \|D\bar{w} - \bar{y}\|^2}_{\text{convex}} + \underbrace{\frac{1}{2} \|\bar{w}\|^2}_{\text{strictly convex}}$$

J is strictly convex.

Details of Gradient descent

$$\bar{w} \leftarrow \bar{w} - \underbrace{\alpha}_{\text{learning rate}} \underbrace{\nabla J(\bar{w})}_{\text{gradient}}$$

We consider both these components in detail.

How to compute ∇J ?

- * analytically (when simple/possible) and hard code it into the computation
 - automatic differentiation (in NNs) ...
 - we'll talk about it later...

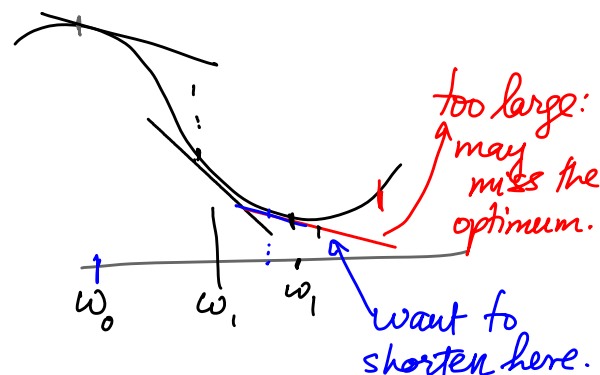
- * finite difference approximation: $\nabla J = \begin{bmatrix} \frac{\partial J}{\partial w_1} \\ \vdots \\ \frac{\partial J}{\partial w_d} \end{bmatrix}$

We compute $\frac{\partial J}{\partial w_i} \approx \frac{J(w_1, \dots, w_i + \Delta, \dots, w_d) - J(\bar{w})}{\Delta_i}$ for "small" Δ

- can do this calculation for only a subset of $i \in \{1, \dots, d\}$, especially when $d \gg 1$.
 - When d is in the billions, we cannot afford to take the partial derivative in each dimension.
- We could also use different Δ for each i , (as long as Δ_i is "small").

Changing α

Intuitively, we want to start with a large α so that J can decrease rapidly at start. And then make α smaller when we are "close to" a critical point. (so, we "decay" α)



$$\bar{w}_{t+1} = \bar{w}_t - \alpha_t \nabla J(\bar{w}_t)$$

t : time (\equiv iteration k as used before)

1. Exponential decay

$$\alpha_t = \alpha_0 e^{(-\mu t)}$$

μ controls the rate of decay

2. inverse decay:

$$\alpha_t = \frac{\alpha_0}{1 + \mu t}$$

3. Step decay: \rightarrow decrease α_t every m steps

$$\alpha_t = \frac{\alpha_{t-m+1}}{\mu} \quad \text{when } t = zm \text{ for } z \in \mathbb{N}.$$

4. Bold driver algorithm:
 α_0 : (initial value)

step t : $\alpha_t = \alpha_{t-1} \times 1.05$

if $J_t < J_{t-1}$
 else $\alpha_t = \alpha_{t-2}/2$
 compute J_t again
 end

\rightarrow increase α_t by 5% to keep the decrease of J going

\rightarrow better to go one step back and then halve α_t from that step.