Goal: use Lyapunov method for convergence analysis of optimization algorithms

- Assume, you like to find the minimizer of the function $J: \mathbb{R}^n \to \mathbb{R}$

$$\min_{x \in \mathbb{R}^n} J(x) \quad \to \text{training NN}$$

- A classical alg. is gradient descent: (we assume J is differentiable)

$$x_{k+1} = x_k - \underbrace{\eta}_{\text{step-size}} \nabla J(x_k)$$

- In the cont. time limit, as $\eta \to 0$, this update converges to gradient flow:

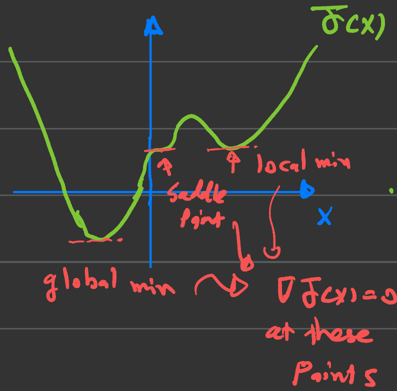
$$\dot{x} = -\nabla J(x)$$

- We will use Lyapunov method to analyze convergence of gradient flow $\to$ gives insight about conv. of grad. descent.

- Eqlb. points:

$$E = \{ x \in \mathbb{R}^n ; \nabla J(x) = 0 \}$$

→ Critical/extreme points



- Any global minimizer x^* is a critical point

$\nabla J(x^*) = 0$ → 1st-order necessary condition for optimality.

- Our goal is to analyze the convergence of gradient flow under different assumptions on the obj. function $J(x)$

Case 1: No assumption
Case 2: radially unbdd
Case 3: Convex
Case 4: gradient dominance
Case 5: strongly convex.

Case 1: (No assumption on J)

- let $V(x) = J(x) - \underbrace{J(x^*)}_{\min_x J(x)}$, then

$$\frac{d}{dt} V(x(t)) = \nabla V(x(t)) \left(-\nabla J(x(t)) \right)$$

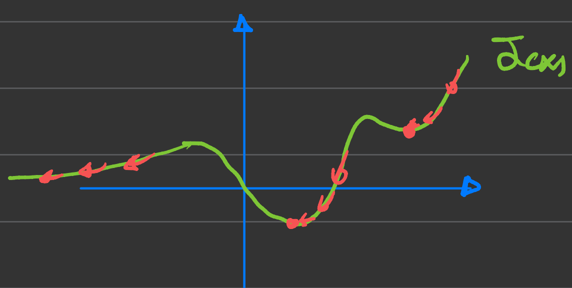
$$= -\|\nabla J(x(t))\|_2^2$$

- Integrating with respect to time:

$$V(x(t)) - V(x(0)) = -\int_0^t \|\nabla J(x(s))\|_2^2 \, ds \leq -t \min_{s \in [0,t]} \|\nabla J(x(s))\|^2$$

$$\Leftrightarrow \min_{s \in [0,t]} \|\nabla J(x(s))\|^2 \leq \frac{V(x(0)) - V(x(t))}{t} \leq \frac{J(x(0)) - J(x^*)}{t}$$

- Easiest conv. result under no assumption $O(\frac{1}{t})$ convergence



Case 2: (J is radially unbdd)

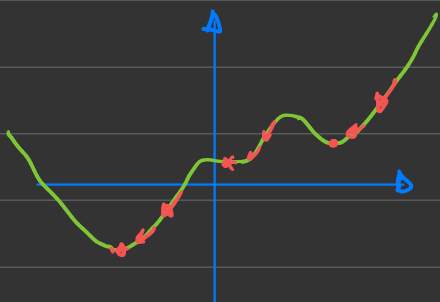
- In this case, $V(x) = J(x) - J(x^*)$ is a radially unbdd function and $\dot{V}(x) = -\|\nabla J(x)\|_2^2 \leq 0 \quad \forall x$.

$\Longrightarrow$ all solutions are bounded

LaSalle

$\Longrightarrow X(t) \to E = \{x \in \mathbb{R}^n ; \nabla J(x) = 0\} \to$ largest invariant set in E is E

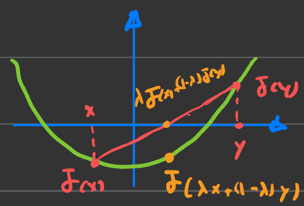
$\Longrightarrow$ X(t) converges to a critical point!



- We can do the linearization procedure to obtain local convergence regims, and local convergence rates.

Convex analysis : [Rockafellar]

- $f: \mathbb{R}^n \to \mathbb{R}$ is convex if



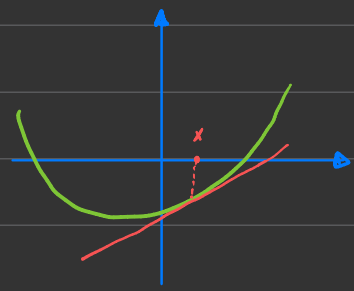
$$f(\lambda x+(1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y), \quad \forall x,y \in \mathbb{R}^n$$
$$\forall \lambda \in [0,1]$$

- A convex func is always locally Lip. on its domain (diff almost everywhere)

- If f is diff. everywhere, then

$$\underbrace{f(y) \geq f(x) + \nabla f(x)(y-x)}_{}, \quad \forall x,y \in \mathbb{R}^n$$

→ graph lies above its tangents



- For convex functions :

all critical points are global minimizers

why?

$\nabla f(x) = 0 \Rightarrow f(y) \geq f(x) \quad \forall y \Rightarrow$ x is global minimizer

Case 3 : (J is convex)

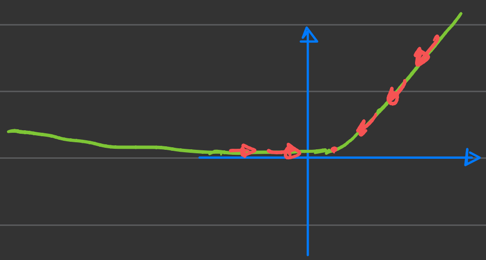
- Take $V(t,x) = \frac{1}{2}\|x - x^*\|^2 + t(J(x) - J(x^*))$

- Then, $\frac{d}{dt} V(t, X(t)) = (X(t) - x^*)^T(-\nabla J(X(t))) + J(X(t)) - J(x^*)$
$- t\|\nabla J(X(t))\|^2 \leq 0$

- Integrating : $V(t, X(t)) \leq V(0, X(0)) = \frac{1}{2}\|X(0) - x^*\|^2$

$\Rightarrow \quad J(X(t)) - J(x^*) \leq \frac{1}{2t}\|X(0) - x^*\|^2$

$O(\frac{1}{t})$ convergence of value of obj. func.



Polyak–Lojasiewicz

Gradient dominance or PL condition:

$$J(x) - J(x^*) \leq \frac{1}{2\mu} \|\nabla J(x)\|_2^2, \qquad \forall x$$

-

Case 4: (J is not convex, but satisfies PL condition)

- Take $V(x) = J(x) - J(x^*)$, then

$$\frac{d}{dt} V(x(t)) = -\|\nabla J(x)\|_2^2 \leq -2\mu V(x(t))$$

$$V(x(t)) \leq e^{-2\mu t} V(x(0))$$

$\xrightarrow[\text{lemma}]{\text{Comparison}}$

$$\Rightarrow J(x(t)) - J(x^*) \leq e^{-2\mu t} \left(J(x(0)) - J(x^*)\right)$$

$\downarrow$

exp. convergence

Strongly convex functions:

- $J: \mathbb{R}^n \to \mathbb{R}$ is μ-strongly convex if

$$J(y) \geq J(x) + \nabla J(x)(y-x) + \frac{\mu}{2}\|y-x\|^2$$

Case 5: (J is strongly convex)

- let $V(x) = \frac{1}{2}\|x-x^*\|^2 \Rightarrow$

$$\frac{d}{dt} V(x(t)) = \nabla J(x(t))(x^* - x(t))$$

- Strong convexity condition for $y = x^*$ and $x = x(t)$ implies

$$\nabla J(x(t))(x^* - x(t)) \leq \underbrace{J(x^*) - J(x(t))}_{\leq 0} - \frac{\mu}{2}\underbrace{\|x(t) - x^*\|^2}_{V(x(t))}$$

$$\leq -\mu V(x(t))$$

$$\Rightarrow \frac{d}{dt} V(x(t)) \leq -\mu V(x(t)) \Rightarrow V(x(t)) \leq e^{-t\mu} V(x(0))$$

$$\Rightarrow \underbrace{\|x(t) - x^*\|^2 \leq e^{-\mu t}\|x(0) - x^*\|^2}_{\text{exp. convergence!}}$$

Remark:

- Strong convexity $\Longrightarrow$ PL

- Some PL functions are not convex, and convex function not PL

Summary: