Loss landscapes and optimization in over-parameterized non-linear systems and neural networks

[Liu, Zhu, Belkin 2022]

Presenter: Shirley
Xiaoqi Liu
27 Nov 2024

- Karimi, Nutini, Schmidt 16
- Liu, Zhu, Belkin 21
- Frei, Muthukumar, Yang 23 (Neurips tutorial)
- Frei 22 (Tutorial at Simons)

Today's plan

1/23

Today's plan

① Introduce Polyak-Lojasiewicz (PL) condition

Today's plan

① Introduce Polyak-Lojasiewicz (PL) condition

② Can we establish PL for large non-linear systems?

- What do their loss landscapes look like?
- Whether PL is satisfied depends on the conditioning of the tangent kernel

Today's plan

1. Introduce Polyak-Lojasiewicz (PL) condition
2. Can we establish PL for large non-linear systems?
   - What do their loss landscapes look like?
   - Whether PL is satisfied depends on the conditioning of the tangent kernel
3. Application to neural networks
4. Summary and extensions

Today's plan

1. Introduce Polyak-Lojasiewicz (PL) condition
2. Can we establish PL for large non-linear systems?
   - What do their loss landscapes look like?
   - Whether PL is satisfied depends on the conditioning of the tangent kernel
3. Application to neural networks
4. Summary and extensions

Try to answer:

Why over-parameterized NNs are non-convex yet easy to optimize?

2/23

2/23

① Introduce Polyak-Lojasiewicz (PL) condition

3/23

3/23

- Strong convexity (SC) $L : \mathbb{R}^m \to \mathbb{R}$

$$L(y) \geq L(x) + \langle \nabla L(x), y-x \rangle + \frac{\mu}{2} \| y-x \|^2 \quad \forall x, y \in \mathbb{R}^n$$

3/23

- Strong convexity (SC) $\qquad L : \mathbb{R}^m \to \mathbb{R}$

$$L(y) \geq L(x) + \langle \nabla L(x), y-x \rangle + \frac{\mu}{2} \| y-x \|^2 \qquad \forall x, y \in \mathbb{R}^n$$

- Polyak- Lojasiewicz (PL) condition

for some $\mu > 0$

$$\frac{1}{2} \| \nabla L(x) \|^2 \geq \mu \big(L(x) - \overbrace{L(x^*)}^{\text{global minimum } (L^*)} \big) \qquad \forall x \in \mathbb{R}^m$$

- Strong convexity (SC) $\qquad L : \mathbb{R}^m \to \mathbb{R}$

$$L(y) \geq L(x) + \langle \nabla L(x), y-x \rangle + \frac{\mu}{2}\|y-x\|^2 \qquad \forall x, y \in \mathbb{R}^n$$

- Polyak-Lojasiewicz (PL) condition

for some $\mu > 0$

$$\frac{1}{2}\|\nabla L(x)\|^2 \geq \mu\big(L(x) - \overbrace{L(x^*)}^{\text{global minimum } (L^*)}\big) \qquad \forall x \in \mathbb{R}^m$$

- Strong convexity (SC) $\quad L : \mathbb{R}^m \to \mathbb{R}$

$$L(y) \geq L(x) + \langle \nabla L(x), y - x \rangle + \frac{\mu}{2} \|y - x\|^2 \quad \forall x, y \in \mathbb{R}^n$$

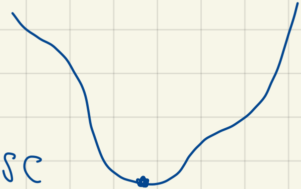
- Polyak-Lojasiewicz (PL) condition

for some $\mu > 0$

$$\frac{1}{2} \|\nabla L(x)\|^2 \geq \mu \big(L(x) - \overbrace{L(x^*)}^{\text{global minimum } (L^*)}\big) \quad \forall x \in \mathbb{R}^m$$

Remarks:

- SC ⇒ PL
- All stationary points are global minima
- PL is somewhat easier to verify than SC



4/23

4/23

Smoothness + PL $\Rightarrow$ Linear convergence of gradient descent

4/23

Smoothness + PL $\Rightarrow$ Linear convergence of gradient descent

Recall: μ-PL gives $\frac{1}{2}\|\nabla L(x)\|^2 \geq \mu\left(L(x)-L(x^*)\right)$

GD $\quad x^{k+1} = x^k - \eta \nabla L(x^k)$

Smoothness + PL ⇒ Linear convergence of gradient descent

Recall: μ-PL gives $\frac{1}{2}\|\nabla L(x)\|^2 \geq \mu\left(L(x)-L(x^*)\right)$

GD $\quad x^{k+1} = x^k - \eta \nabla L(x^k)$

Proof:

Smoothness + PL $\Longrightarrow$ Linear convergence of gradient descent

Recall: μ-PL gives $\frac{1}{2}\|\nabla L(x)\|^2 \geq \mu\left(L(x)-L(x^*)\right)$

GD $\quad x^{k+1} = x^k - \eta \nabla L(x^k)$

Proof:

$$L(x^{k+1}) - L(x^*)$$

Smoothness + PL $\Rightarrow$ Linear convergence of gradient descent

Recall: μ-PL gives $\frac{1}{2}\|\nabla L(x)\|^2 \geq \mu\left(L(x)-L(x^*)\right)$

GD $\quad x^{k+1} = x^k - \eta \nabla L(x^k)$

Proof:

$$L(x^{k+1}) - L(x^*)$$

β-smoothness

$$\leq L(x^k) + \langle \nabla L(x^k), x^{k+1} - x^k \rangle + \frac{1}{2}\beta \|x^{k+1} - x^k\|^2 - L(x^*)$$

$$= L(x^k) + \langle \nabla L(x^k), -\eta \nabla L(x^k) \rangle + \frac{1}{2}\beta \|\eta \nabla L(x^k)\|^2 - L(x^*)$$

Smoothness + PL $\Rightarrow$ Linear convergence of gradient descent

Recall: μ-PL gives $\frac{1}{2}\|\nabla L(x)\|^2 \geq \mu\left(L(x)-L(x^*)\right)$

GD $\quad x^{k+1} = x^k - \eta \nabla L(x^k)$

Proof:

$$L(x^{k+1}) - L(x^*)$$

β-smoothness

$$\leq L(x^k) + \langle \nabla L(x^k), x^{k+1}-x^k \rangle + \frac{1}{2}\beta\|x^{k+1}-x^k\|^2 - L(x^*)$$

$$= L(x^k) + \langle \nabla L(x^k), -\eta\nabla L(x^k) \rangle + \frac{1}{2}\beta\|\eta\nabla L(x^k)\|^2 - L(x^*)$$

$\eta = 1/\beta$

$$= L(x^k) - \frac{1}{2}\eta\|\nabla L(x^k)\|^2 - L(x^*)$$

4/23

Smoothness + PL ⟹ Linear convergence of gradient descent

Recall: μ-PL gives $\frac{1}{2}\|\nabla L(x)\|^2 \geq \mu\left(L(x)-L(x^*)\right)$

GD $\quad x^{k+1} = x^k - \eta \nabla L(x^k)$

Proof:

$$L(x^{k+1}) - L(x^*)$$

β-smoothness

$$\leq L(x^k) + \langle \nabla L(x^k), x^{k+1} - x^k \rangle + \frac{1}{2}\beta \|x^{k+1} - x^k\|^2 - L(x^*)$$

$$= L(x^k) + \langle \nabla L(x^k), -\eta \nabla L(x^k) \rangle + \frac{1}{2}\beta \|\eta \nabla L(x^k)\|^2 - L(x^*)$$

$\eta = 1/\beta$

$$= L(x^k) - \frac{1}{2}\eta \|\nabla L(x^k)\|^2 - L(x^*)$$

μ-PL

$$\leq L(x^k) - \eta\mu\left(L(x^k) - L(x^*)\right) - L(x^*)$$

$$= (1-\eta\mu)\left(L(x^k) - L(x^*)\right) \quad \square$$

5/23

Aside:

Aside:

- Similar proofs for stochastic gradient descent, randomized coordinate descent, greedy coordinate descent...

Aside:

- Similar proofs for stochastic gradient descent, randomized coordinate descent, greedy coordinate descent...
- Proximal gradient generalization (via proximal-PL)

$$\min_{x} \tilde{L}(x) \equiv \underbrace{L(x)}_{\beta\text{-smooth}} + \underbrace{g(x)}_{\substack{\text{non-smooth}\\ \text{convex}}}$$

Aside:

- Similar proofs for stochastic gradient descent, randomized coordinate descent, greedy coordinate descent...
- Proximal gradient generalization (via proximal-PL)

$$\min_x \tilde{L}(x) \equiv \underbrace{L(x)}_{\beta\text{-smooth}} + \underbrace{g(x)}_{\substack{\text{non-smooth}\\ \text{convex}}}$$

- Other conditions for obtaining linear convergence:
  - Weak SC: $L(x^*) \geq L(x) + \langle \nabla L(x), x^* - x \rangle + \frac{\mu}{2} \| x^* - x \|^2$
  - Quadratic Growth: $L(x) - L(x^*) \geq \frac{\mu}{2} \| x^* - x \|^2$

⋮

6/23

(2) Can we establish PL for large non-linear systems?

Setting

7/23

Setting

- Training data $\{x_i, y_i\}_{i=1}^{n}$ $x_i \in \mathbb{R}^d$ $y_i \in \mathbb{R}$

Setting

- Training data $\{x_i, y_i\}_{i=1}^{n}$ $x_i \in \mathbb{R}^d$ $y_i \in \mathbb{R}$
- A parametric model $f(w; x_i)$ e.g. a neural network

Setting

- Training data $\{x_i, y_i\}_{i=1}^{n}$ $x_i \in \mathbb{R}^d$ $y_i \in \mathbb{R}$
- A parametric model $f(w; x_i)$ e.g. a neural network

Define shorthand: $F(w) = \begin{bmatrix} f(w; x_1) \\ f(w; x_2) \\ \vdots \\ f(w; x_n) \end{bmatrix} \in \mathbb{R}^n$ $F: \mathbb{R}^m \to \mathbb{R}^n$

$$\nabla F = \frac{\partial F}{\partial w} \in \mathbb{R}^{n \times m}$$

$$H_{F_i} = \frac{\partial^2 F_i}{\partial w^2} \in \mathbb{R}^{m \times m}$$

Setting

- Training data $\{x_i, y_i\}_{i=1}^{n}$ $x_i \in \mathbb{R}^d$ $y_i \in \mathbb{R}$
- A parametric model $f(w; x_i)$ e.g. a neural network

Define shorthand: $F(w) = \begin{bmatrix} f(w; x_1) \\ f(w; x_2) \\ \vdots \\ f(w; x_n) \end{bmatrix} \in \mathbb{R}^n$ $F: \mathbb{R}^m \to \mathbb{R}^n$

$\nabla F = \frac{\partial F}{\partial w} \in \mathbb{R}^{n \times m}$

$H_{F_i} = \frac{\partial^2 F_i}{\partial w^2} \in \mathbb{R}^{m \times m}$

β_F - smooth, L_F - Lipschitz

Setting

- Training data $\{x_i, y_i\}_{i=1}^{n}$ $\quad x_i \in \mathbb{R}^d \quad y_i \in \mathbb{R}$
- A parametric model $f(w; x_i)$ e.g. a neural network

Define shorthand: $F(w) = \begin{bmatrix} f(w; x_1) \\ f(w; x_2) \\ \vdots \\ f(w; x_n) \end{bmatrix} \in \mathbb{R}^n \qquad F : \mathbb{R}^m \to \mathbb{R}^n$

$$\nabla F = \frac{\partial F}{\partial w} \in \mathbb{R}^{n \times m}$$

$$H_{F_i} = \frac{\partial^2 F_i}{\partial w^2} \in \mathbb{R}^{m \times m}$$

β_F-smooth, L_F-Lipschitz

- Loss function for optimization

$$L(w) \overset{\text{e.g.}}{=} \frac{1}{2} \| F(w) - y \|^2$$

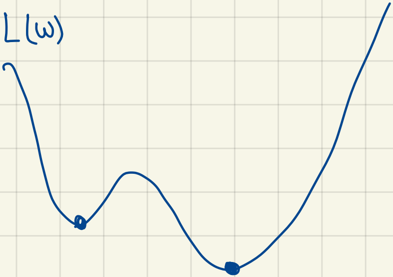
$$H_L = \frac{\partial^2 L}{\partial w^2} \in \mathbb{R}^{m \times m}$$

Loss landscapes

8/23

Loss landscapes

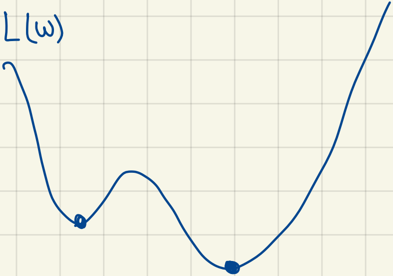
8/23



under-parameterized $m<n$

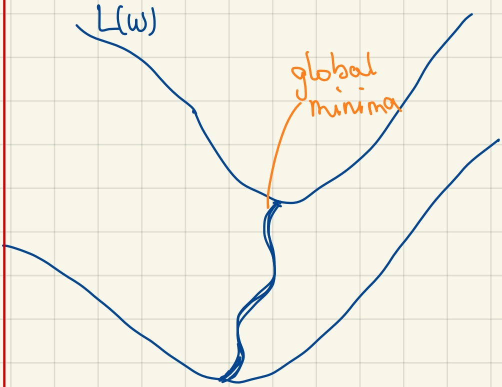
isolated local minima

Loss landscapes



under-parameterized $m < n$

isolated local minima



over-parameterized $m \gg n$

- non-convex, even locally
- manifolds of global minima

Loss landscape in overparameterized case.

9/23

Loss landscape in overparameterized case.

Prop 1. (Local non-convexity). can be generalized

Consider $L(w)=\frac{1}{2}\|F(w)-y\|^2$. Suppose $\nabla F(w^*)\neq 0$ and a mild technical assumption on $\{H_{F_i}\}_{i=1}^{n}$, then $L(w)$ isn't convex in any neighbourhood of w^*

9/23

Loss landscape in overparameterized case.

Prop 1. (Local non-convexity).

can be generalized

Consider $L(w) = \frac{1}{2}\|F(w)-y\|^2$. Suppose $\nabla F(w^*) \neq 0$ and a mild technical assumption on $\{H_{F_i}\}_{i=1}^{n}$, then $L(w)$ isn't convex in any neighbourhood of w^*

Proof idea:

9/23

Loss landscape in overparameterized case.

Prop 1. (Local non-convexity).

can be generalized

Consider $L(w)=\frac{1}{2}\|F(w)-y\|^2$. Suppose $\nabla F(w^*)\neq 0$ and a mild technical assumption on $\{H_{F_i}\}_{i=1}^{n}$, then $L(w)$ isn't convex in any neighbourhood of w^*

Proof idea:

- $H_L(w) = \nabla F(w)^T \frac{\partial^2 L}{\partial F^2} \nabla F(w) + \sum_{i=1}^{n} \left(F(w)-y\right)_i H_{F_i}(w)$

9/23

Loss landscape in overparameterized case.

Prop 1. (Local non-convexity).

can be generalized

Consider $L(w) = \frac{1}{2}\|F(w)-y\|^2$. Suppose $\nabla F(w^*) \neq 0$ and a mild technical assumption on $\{H_{F_i}\}_{i=1}^n$, then $L(w)$ isn't convex in any neighbourhood of w^*

Proof idea:

- $$\underbrace{H_L(w)}_{m\times m} = \underbrace{\nabla F(w)^T}_{m\times n} \underbrace{\frac{\partial^2 L}{\partial F^2}}_{n\times n} \underbrace{\nabla F(w)}_{n\times m} + \sum_{i=1}^{n} \underbrace{\big(F(w)-y\big)_i}_{\text{scalar}} \underbrace{H_{F_i}(w)}_{m\times m}$$

for $m \gg n$, first term is generally low-rank, but second term isnt

9/23

Loss landscape in overparameterized case.

> **Prop 1.** (Local non-convexity).
> Consider $L(w) = \underbrace{\frac{1}{2}\|F(w)-y\|^2}_{\text{can be generalized}}$. Suppose $\nabla F(w^*) \neq 0$ and a mild technical assumption on $\{H_{F_i}\}_{i=1}^{n}$, then $L(w)$ isn't convex in any neighbourhood of w^*

Proof idea:

- $$\underbrace{H_L(w)}_{m\times m} = \underbrace{\nabla F(w)^T}_{m\times n} \underbrace{\frac{\partial^2 L}{\partial F^2}}_{n\times n} \underbrace{\nabla F(w)}_{n\times m} + \sum_{i=1}^{n} \underbrace{(F(w)-y)_i}_{\text{scalar}} \underbrace{H_{F_i}(w)}_{m\times m}$$

  for $m \gg n$, first term is generally low-rank, but second term isn't

- Consider the special case where $n=1$, we have

$$H_L(w) = \underbrace{\nabla F(w)^T \nabla F(w)}_{\text{rank-1}} + \underbrace{(F(w)-y)}_{=0 \text{ at } w=w^*} \underbrace{H_F(w)} \quad \text{Assume} > \text{rank-1}$$

9/23

Loss landscape in overparameterized case.

Prop 1. (Local non-convexity).
Consider $L(w) = \frac{1}{2}\underbrace{\|F(w)-y\|^2}_{}$ can be generalized. Suppose $\nabla F(w^*) \neq 0$ and a mild technical assumption on $\{H_{F_i}\}_{i=1}^{n}$, then $L(w)$ isn't convex in any neighbourhood of w^*

Proof idea:

- $$\underbrace{H_L(w)}_{m\times m} = \underbrace{\nabla F(w)^T}_{m\times n} \frac{\partial^2 L}{\underbrace{\partial F^2}_{n\times n}} \underbrace{\nabla F(w)}_{n\times m} + \sum_{i=1}^{n} \underbrace{\big(F(w)-y\big)_i}_{\text{scalar}} \underbrace{H_{F_i}(w)}_{m\times m}$$

for $m \gg n$, first term is generally low-rank, but second term isnt

- Consider the special case where $n=1$, we have

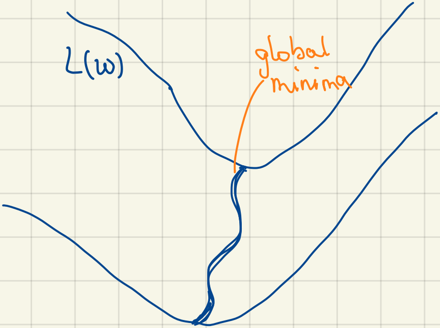
$$H_L(w) = \underbrace{\nabla F(w)^T \nabla F(w)}_{\text{rank-1}} + \underbrace{\big(F(w)-y\big)}_{=0 \text{ at } w=w^*} \underbrace{H_F(w)}_{} \quad \text{Assume} > \text{rank-1}$$

$H_L(w+\delta)$ and/or $H_L(w-\delta)$ are not PSD $\therefore$ no local convexity $\square$

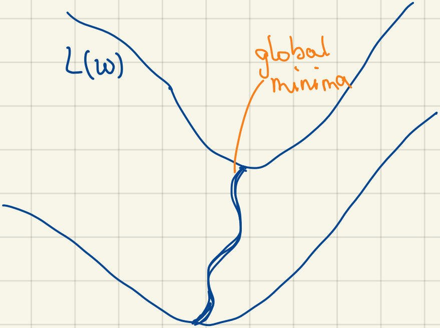
10/23

Convexity → PL

20/23



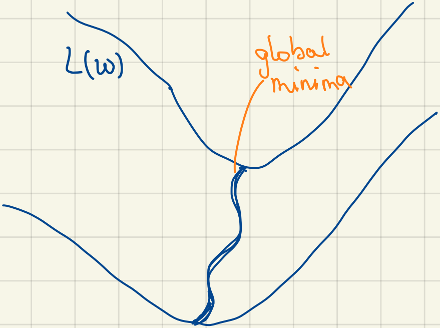
Convexity → PL



- Recall: PL requires

$$\frac{1}{2}\|\nabla L(w)\|^2 \geq \mu\left(L(w) - \underbrace{L(w^*)}_{\text{global optimum}}\right) \quad \forall w \in \mathbb{R}^m$$

Convexity → PL → Local PL



- Recall: PL requires

global optimum

$$\frac{1}{2}\|\nabla L(w)\|^2 \geq \mu\big(L(w) - \underbrace{L(w^*)}\big) \qquad \forall w \in \mathbb{R}^m$$

- Local PL / PL* :

A non-negative function L satisfies μ-PL* on a set $S \subset \mathbb{R}^m$ if

$$\|\nabla L(w)\|^2 \geq \mu L(w) \qquad \forall\, w \in S$$

11/23

Can we get PL* for our overparam. system? 11/23

Can we get PL* for our overparam. system? 11/23

Theorem 1 (Uniform conditioning $\Rightarrow$ PL*)

The square loss $L(w) = \frac{1}{2}\|F(w) - y\|^2$ satisfies μ-PL* on $S \subset \mathbb{R}^m$ if $\lambda_{\min}(K(w)) \geq \mu$ on S

Can we get PL* for our overparam. system? 11/23

Theorem 1 (Uniform conditioning $\Rightarrow$ PL*)

The square loss $L(w) = \frac{1}{2}\|F(w) - y\|^2$ satisfies μ-PL* on $S \subset \mathbb{R}^m$ if $\lambda_{\min}(K(w)) \geq \mu$ on S

Proof:

Can we get PL* for our overparam. system? 11/23

Theorem 1 (Uniform conditioning ⇒ PL*)

The square loss $L(w) = \frac{1}{2}\|F(w) - y\|^2$ satisfies μ-PL* on $S \subset \mathbb{R}^m$ if $\lambda_{\min}(K(w)) \geq \mu$ on S

Proof:

$$\frac{1}{2}\|\nabla L(w)\|^2 = \frac{1}{2}\left((F(w)-y)\,\nabla F(w)\,\nabla F(w)^T(F(w)-y)^T\right)$$

Can we get PL* for our overparam. system? 1/23

Theorem 1 (Uniform conditioning ⇒ PL*)

The square loss $L(w) = \frac{1}{2}\|F(w) - y\|^2$ satisfies μ-PL* on $S \subset \mathbb{R}^m$ if $\lambda_{\min}(K(w)) \geq \mu$ on S

Proof:

$$\frac{1}{2}\|\nabla L(w)\|^2 = \frac{1}{2}\Big(\underbrace{(F(w)-y)}_{\text{len-}n\text{ vector}}\underbrace{\nabla F(w)}_{n\times m}\underbrace{\nabla F(w)^T}_{m\times n}(F(w)-y)^T\Big)$$

Can we get PL* for our overparam. system? 11/23

Theorem 1 (Uniform conditioning $\Rightarrow$ PL*)

The square loss $L(w) = \frac{1}{2}\|F(w) - y\|^2$ satisfies μ-PL* on $S \subset \mathbb{R}^m$ if $\lambda_{\min}(K(w)) \geq \mu$ on S

Proof:

$$\frac{1}{2}\|\nabla L(w)\|^2 = \frac{1}{2}\Big(\underbrace{(F(w)-y)}_{\text{len-}n\text{ vector}}\overbrace{\underbrace{\nabla F(w)}_{n\times m}\,\underbrace{\nabla F(w)^T}_{m\times n}}^{K(w)\,:\ \text{tangent kernel}}(F(w)-y)^T\Big)$$

Can we get PL* for our overparam. system? 1/23

Theorem 1 (Uniform conditioning ⇒ PL*)
The square loss $L(w) = \frac{1}{2}\|F(w)-y\|^2$ satisfies μ-PL* on $S \subset \mathbb{R}^m$ if $\lambda_{min}(K(w)) \geq \mu$ on S

Proof:

$$\frac{1}{2}\|\nabla L(w)\|^2 = \frac{1}{2}\left(\underbrace{(F(w)-y)}_{\text{len-n vector}}\overbrace{\underbrace{\nabla F(w)}_{n\times m}\,\underbrace{\nabla F(w)^T}_{m\times n}}^{K(w)\text{ : tangent kernel}}(F(w)-y)^T\right)$$

$$\geq \frac{1}{2}\lambda_{min}(K(w))\,\|F(w)-y\|^2$$

Can we get PL* for our overparam. system? 4/23

Theorem 1 (Uniform conditioning $\Rightarrow$ PL*)

The square loss $L(w) = \frac{1}{2}\|F(w)-y\|^2$ satisfies μ-PL* on $S \subset \mathbb{R}^m$ if $\lambda_{\min}(K(w)) \geq \mu$ on S

Proof:

$$\frac{1}{2}\|\nabla L(w)\|^2 = \frac{1}{2}\Big(\underbrace{(F(w)-y)}_{\text{len-}n\text{ vector}}\overbrace{\underbrace{\nabla F(w)}_{n\times m}\,\underbrace{\nabla F(w)^T}_{m\times n}}^{K(w):\ \text{tangent kernel}}(F(w)-y)^T\Big)$$

$$\geq \frac{1}{2}\lambda_{\min}(K(w))\,\|F(w)-y\|^2 \geq \mu L(w) \quad \square$$

Can we get PL* for our overparam. system? 11/23

Theorem 1 (Uniform conditioning ⇒ PL*)
The square loss $L(w) = \frac{1}{2}\|F(w)-y\|^2$ satisfies μ-PL* on $S \subset \mathbb{R}^m$ if $\lambda_{\min}(K(w)) \geq \mu$ on S

Proof:

$$\frac{1}{2}\|\nabla L(w)\|^2 = \frac{1}{2}\Big(\underbrace{(F(w)-y)}_{\text{len-}n\text{ vector}}\overbrace{\underbrace{\nabla F(w)}_{n\times m}\,\underbrace{\nabla F(w)^T}_{m\times n}}^{K(w)\text{ : tangent kernel}}(F(w)-y)^T\Big)$$

$$\geq \frac{1}{2}\lambda_{\min}(K(w))\,\|F(w)-y\|^2 \geq \mu L(w) \quad \square$$

Can we get PL* for our overparam. system? 1/23

Theorem 1 (Uniform conditioning $\Rightarrow$ PL*)

The square loss $L(w) = \frac{1}{2}\|F(w)-y\|^2$ satisfies μ-PL* on $S \subset \mathbb{R}^m$ if $\lambda_{\min}(K(w)) \geq \mu$ on S

Proof:

$$\frac{1}{2}\|\nabla L(w)\|^2 = \frac{1}{2}\Big(\underbrace{(F(w)-y)}_{\text{len-n vector}}\overbrace{\underbrace{\nabla F(w)}_{n\times m}\,\underbrace{\nabla F(w)^T}_{m\times n}}^{K(w)\,:\ \text{tangent kernel}}(F(w)-y)^T\Big)$$

$$\geq \frac{1}{2}\lambda_{\min}(K(w))\,\|F(w)-y\|^2 \geq \mu L(w) \quad \square$$

---

① For Gaussian random matrix $M \in \mathbb{R}^{n\times m}$, the condition number κ of MM^T satisfies $\mathbb{E}[\log \kappa] \sim \log \frac{m}{|m-n|+1}$ [Chen, Dongarra 05]

Can we get PL* for our overparam. system? 11/23

Theorem 1 (Uniform conditioning ⇒ PL*)

The square loss $L(w) = \frac{1}{2}\|F(w)-y\|^2$ satisfies μ-PL* on $S \subset \mathbb{R}^m$ if $\lambda_{min}(K(w)) \geq \mu$ on S

Proof:

$$\frac{1}{2}\|\nabla L(w)\|^2 = \frac{1}{2}\Big(\underbrace{(F(w)-y)}_{\text{len-}n\text{ vector}}\overbrace{\underbrace{\nabla F(w)}_{n\times m}\,\underbrace{\nabla F(w)^T}_{m\times n}}^{K(w):\ \text{tangent kernel}}(F(w)-y)^T\Big)$$

$$\geq \frac{1}{2}\lambda_{min}(K(w))\,\|F(w)-y\|^2 \geq \mu L(w) \quad \square$$

① For Gaussian random matrix $M \in \mathbb{R}^{n\times m}$, the condition number κ of MM^T satisfies $\mathbb{E}[\log \kappa] \sim \log \frac{m}{|m-n|+1}$ [Chen, Dongarra 05]

② Standard initialization $w_0 \overset{iid}{\sim} N(0,1)$, for wide NN $F(w)$, $\lambda_{min}(K(w_0)) = O(1)$ w.h.p. assuming data is not degenerate

12/23

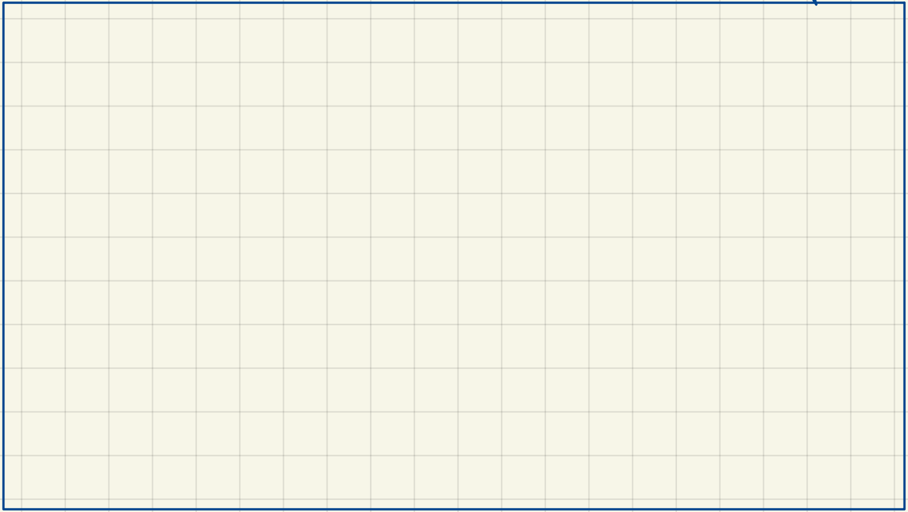
Picture so far

12/23

12/23

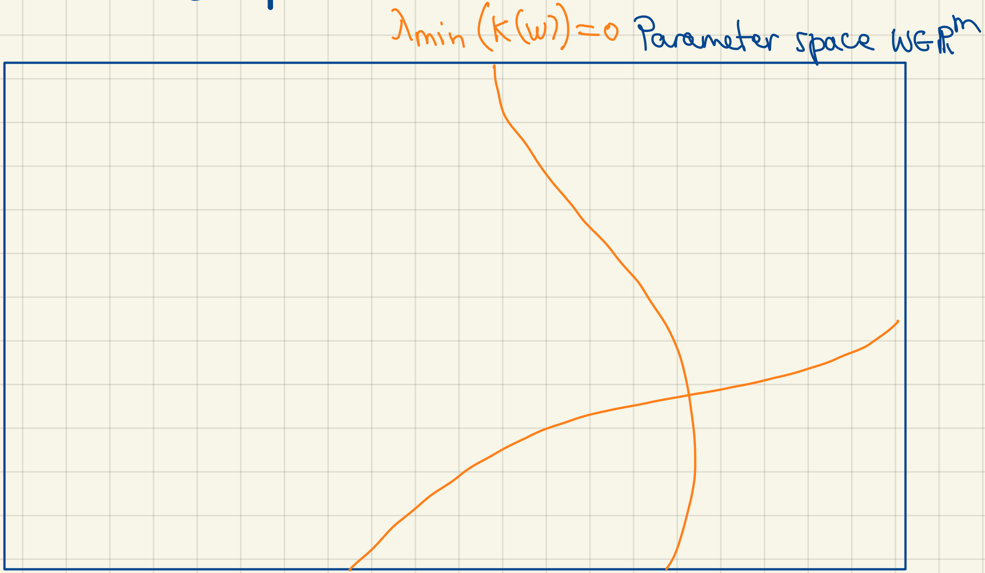
Picture so far

Parameter space $W \in \mathbb{R}^m$



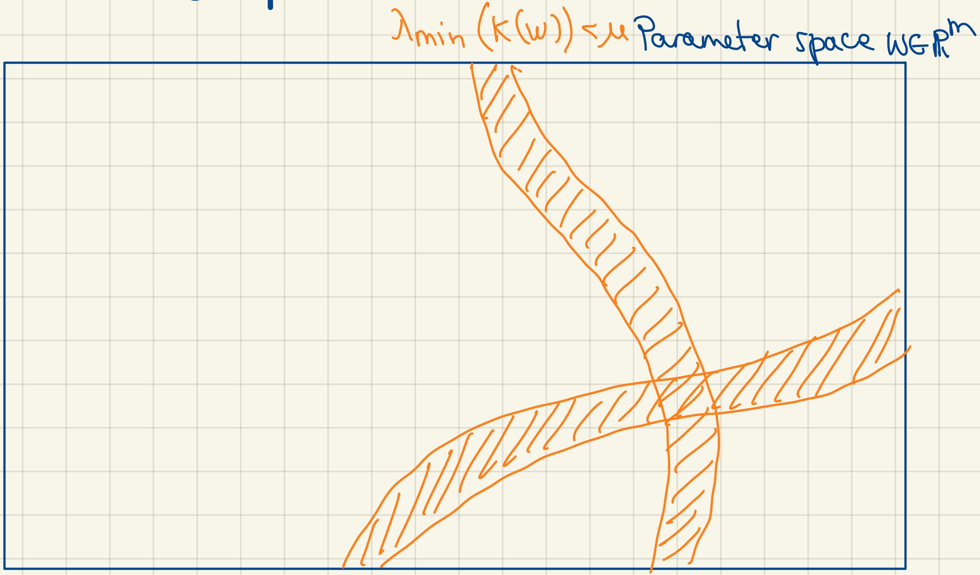
Picture so far

12/23

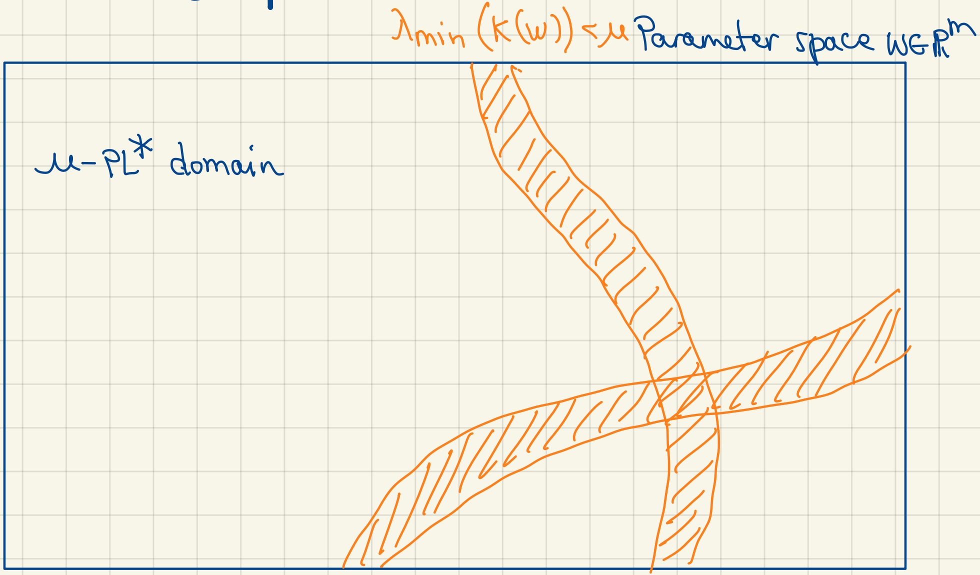


Picture so far

12/23

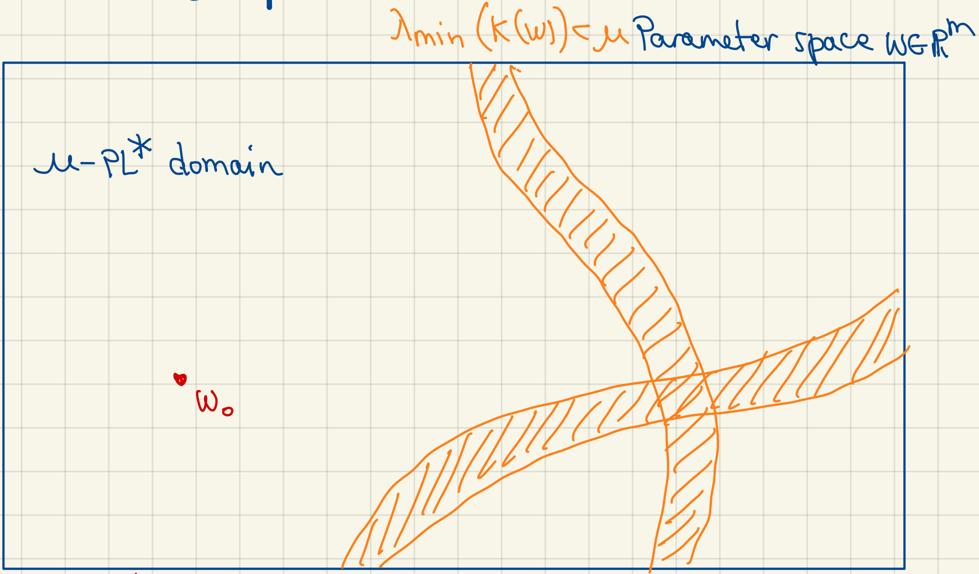

$\lambda_{\min}(K(w)) < \mu$
Parameter space $W \in \mathbb{R}^m$

Picture so far


$\lambda_{min}(K(w)) < \mu$
Parameter space $w \in \mathbb{R}^m$
μ-PL* domain

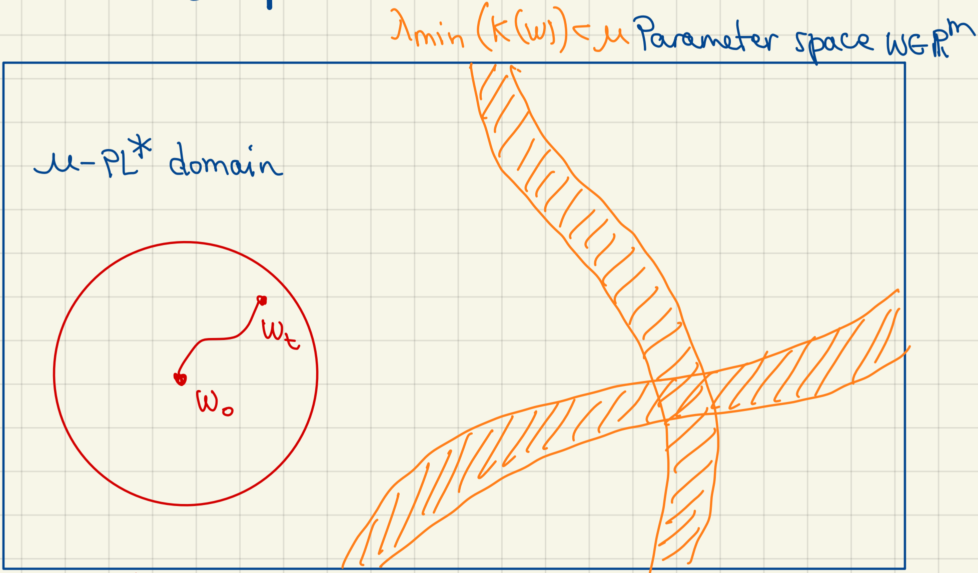
Picture so far

12/23


$\lambda_{min}(K(w)) < \mu$
Parameter space $W \in \mathbb{R}^m$
μ-PL* domain
w_0

w.h.p. the initializer w_0 is good (i.e $\lambda_{min}(K(w_0)) \geq \mu$)

Picture so far


$\lambda_{\min}(K(w)) < \mu$
Parameter space $w \in \mathbb{R}^m$
μ-PL* domain
w_t
w_0

w.h.p. the initializer w_0 is good (i.e. $\lambda_{\min}(K(w_0)) \geq \mu$)

- Next steps: want to show $w \in B(w_0, R)$ is good too, and adjust R to cover training trajectory in the ball.

13/23

Theorem 2 (Small Hessian $\Rightarrow$ PL* in $B(w_0, R)$)

Theorem 2 (Small Hessian $\Rightarrow$ PL* in $B(w_0,R)$)

Suppose $\lambda_{\min}(K(w_0)) = \lambda_0 > 0$. If $\boxed{\|H_{F_i}(w)\|_2 \leq \frac{\lambda_0 - \mu}{2 L_F \sqrt{n}\, R}}$ for $w \in B(w_0, R)$, then $K(w)$ is μ-uniformly conditioned in the ball.

Hence, $L(w) = \frac{1}{2}\|F(w) - y\|^2$ satisfies μ-PL* for $w \in B(w_0, R)$.

Theorem 2 (Small Hessian $\Rightarrow$ PL* in $B(w_0, R)$)

Suppose $\lambda_{\min}(K(w_0)) = \lambda_0 > 0$. If $\boxed{\|H_{F_i}(w)\|_2 \leq \frac{\lambda_0 - \mu}{2 L_F \sqrt{n} R}}$ for $w \in B(w_0, R)$, then $K(w)$ is μ-uniformly conditioned in the ball.

Hence, $L(w) = \frac{1}{2}\|F(w) - y\|^2$ satisfies μ-PL* for $w \in B(w_0, R)$.

Proof idea:

Theorem 2 (Small Hessian $\Rightarrow$ PL* in $B(w_0, R)$)

Suppose $\lambda_{\min}(K(w_0)) = \lambda_0 > 0$. If $\boxed{\|H_{F_i}(w)\|_2 \leq \frac{\lambda_0 - \mu}{2 L_F \sqrt{n} R}}$ for $w \in B(w_0, R)$, then $K(w)$ is μ-uniformly conditioned in the ball.

Hence, $L(w) = \frac{1}{2}\|F(w) - y\|^2$ satisfies μ-PL* for $w \in B(w_0, R)$.

Proof idea: fundamental theorem of calculus

$$\nabla F_i(w) = \nabla F_i(w_0) + \int_0^1 \underbrace{H_{F_i}(w_0 + \tau(w - w_0))}\underbrace{(w - w_0)}\, d\tau$$

Cauchy-Schwartz ...

Theorem 2 (Small Hessian $\Rightarrow$ PL* in $B(w_0, R)$)

Suppose $\lambda_{\min}(K(w_0)) = \lambda_0 > 0$. If $\boxed{\|H_{F_i}(w)\|_2 \leq \frac{\lambda_0 - \mu}{2 L_F \sqrt{n} R}}$ for $w \in B(w_0, R)$, then $K(w)$ is μ-uniformly conditioned in the ball.

Hence, $L(w) = \frac{1}{2}\|F(w) - y\|^2$ satisfies μ-PL* for $w \in B(w_0, R)$.

Proof idea: fundamental theorem of calculus

$$\nabla F_i(w) = \nabla F_i(w_0) + \int_0^1 \underbrace{H_{F_i}(w_0 + \tau(w - w_0))}\underbrace{(w - w_0)}\, d\tau$$

Cauchy-Schwartz ...

- How do results so far apply to neural nets?
- When do we have small $\|H_{F_i}(w)\|_2$?

14/23

③ Application to neural networks

A shallow neural network.

15/23

A shallow neural network.

- We'll show that for large enough width m, $\|H_{F_i}(w)\|_2$ is small

A shallow neural network.

- We'll show that for large enough width m, $\| H_{F_i}(w) \|_2$ is small

- For simplicity, let's consider data $x_i \in \mathbb{R}$, $|x_i| \leq 1$ $\forall i \in [n]$

A shallow neural network.

- We'll show that for large enough width m, $\|H_{F_i}(w)\|_2$ is small

- For simplicity, let's consider data $x_i \in \mathbb{R}$, $|x_i| \leq 1$ $\forall i \in [n]$

$$\boxed{F_i(w) = f(w; x_i) = \frac{1}{\sqrt{m}} \sum_{j=1}^{m} v_j \, \sigma(w_j x_i)}$$

A shallow neural network.

- We'll show that for large enough width m, $\|H_{F_i}(w)\|_2$ is small

- For simplicity, let's consider data $x_i \in \mathbb{R}$, $|x_i| \leq 1$ $\forall i \in [n]$

$$F_i(w) = f(w; x_i) = \frac{1}{\sqrt{m}} \sum_{j=1}^{m} v_j \, \sigma(w_j x_i)$$

- Fix $v_j \in \{1, -1\}$
- Initialize $w_j \sim N(0,1)$ then train
- σ: β_σ-smooth, L_σ-Lipschitz
  e.g. tanh, sigmoid

A shallow neural network.

- We'll show that for large enough width m, $\| H_{F_i}(w) \|_2$ is small

- For simplicity, let's consider data $x_i \in \mathbb{R}$, $|x_i| \le 1$ $\forall i \in [n]$

$$F_i(w) = f(w; x_i) = \frac{1}{\sqrt{m}} \sum_{j=1}^{m} v_j \, \sigma(w_j x_i)$$

- Fix $v_j \in \{1, -1\}$
- Initialize $w_j \sim N(0,1)$ then train
- σ: β_σ-smooth, L_σ-Lipschitz

  e.g. tanh, sigmoid

① $$[H_{F_i}(w)]_{jk} = \frac{\partial^2 f(w; x_i)}{\partial w_j \partial w_k} = \frac{1}{\sqrt{m}} v_j \, \sigma''(w_j x_i) \, x_i^2 \, \mathbb{1}\{j = k\}$$

$$\Rightarrow \| H_{F_i}(w) \|_2 \le \frac{1}{\sqrt{m}} \beta_\sigma = O\left(\frac{1}{\sqrt{m}}\right)$$

② $$\| \nabla F_i(w) \|^2 = \frac{1}{m} \sum_{j=1}^{m} x_i^2 \left(\sigma'(w_j x_i)\right)^2 \le L_\sigma^2 = \Theta(1)$$

Deeper neural nets

Deeper neural nets

L-layer NN:

$$\alpha^{(0)} = x$$

$$\alpha^{(\ell)} = \sigma_\ell\left(\frac{1}{\sqrt{m_{\ell-1}}} W^{(\ell)} \alpha^{(\ell-1)}\right) \quad \forall\ \ell = 1, \cdots L+1$$

$$f(W; x) = \alpha^{(L+1)}$$

Deeper neural nets

L-layer NN:

$$\alpha^{(0)} = x$$

$$\alpha^{(l)} = \sigma_l\left(\frac{1}{\sqrt{m_{l-1}}} W^{(l)} \alpha^{(l-1)}\right) \quad \forall\, l = 1, \dots L+1$$

$$f(W; x) = \alpha^{(L+1)}$$

Theorem 3 (Wide neural nets w. linear output layer satisfy PL*)

Deeper neural nets

L-layer NN:

$$\alpha^{(0)} = x$$

$$\alpha^{(l)} = \sigma_l\left(\frac{1}{\sqrt{m_{l-1}}} W^{(l)} \alpha^{(l-1)}\right) \quad \forall\ l = 1, \cdots L+1$$

$$f(W; x) = \alpha^{(L+1)}$$

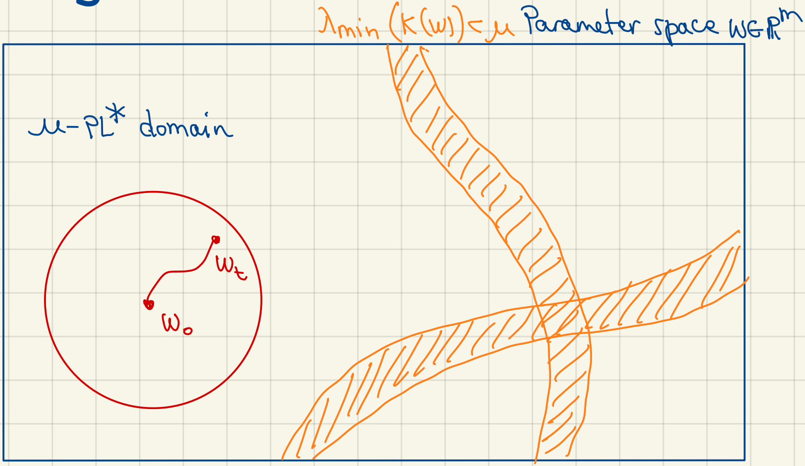
Theorem 3 (Wide neural nets w. linear output layer satisfy PL*)

Consider $W_0^{(l)} \sim N(0, I)$ $\forall\ l \in [L+1]$, and $\sigma_{l+1}(z) = z$. Suppose $\lambda_0 = \lambda_{\min}(K(W_0)) > 0$. Then $L(w) = \frac{1}{2}\|F(w) - y\|^2$ satisfies μ-PL* on $B(W_0, R)$ if

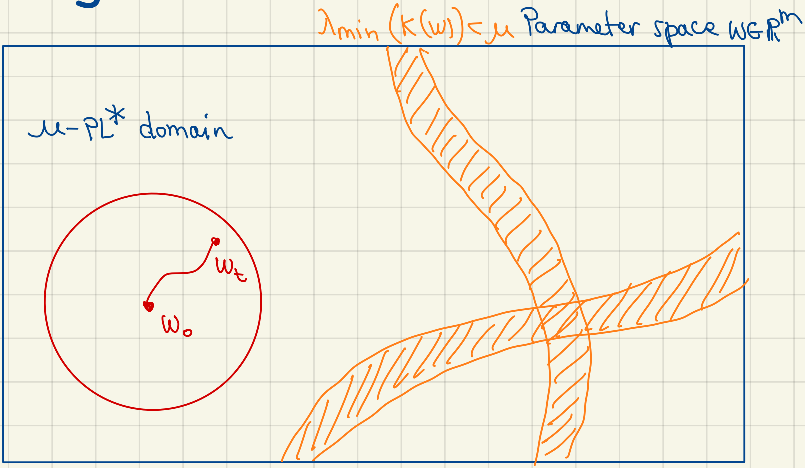
$$\boxed{m = \tilde{\Omega}\left(\frac{nR^{6L+2}}{(\lambda_0 - \mu\rho^{-2})^2}\right)}$$

Summary so far

Summary so far



Summary so far

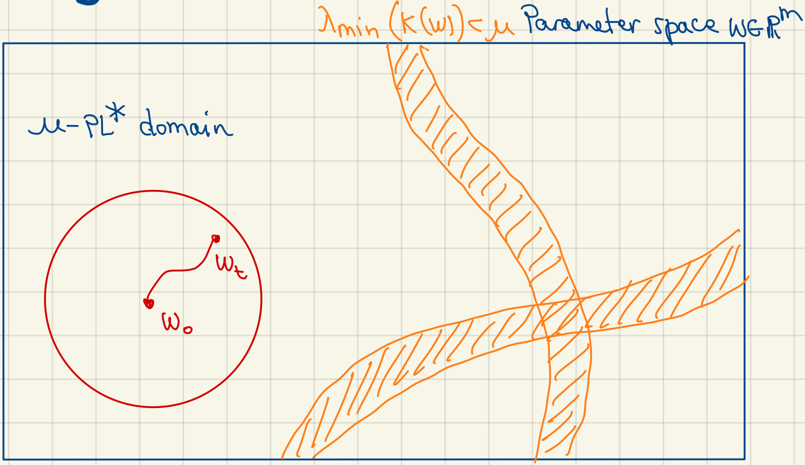

$\lambda_{min}(K(w)) < \mu$
Parameter space $w \in \mathbb{R}^m$
μ-PL* domain
w_t
w_0

Neural net sufficiently wide

$\Rightarrow K(w) = \nabla F(w) \nabla F(w)^T$ well-conditioned in $B(w_0, R)$

$\Rightarrow L(w) = \frac{1}{2} \| F(w) - y \|^2$ satisfies PL* in $B(w_0, R)$

Summary so far


$\lambda_{min}(K(w)) < \mu$
Parameter space $w \in \mathbb{R}^m$
μ-PL* domain
w_t
w_0

Neural net sufficiently wide

$\Rightarrow$ $K(w) = \nabla F(w) \nabla F(w)^T$ well-conditioned in $B(w_0, R)$

$\Rightarrow$ $L(w) = \frac{1}{2} \| F(w) - y \|^2$ satisfies PL* in $B(w_0, R)$

$\Rightarrow$ Linear convergence of (S)GD

18/23

Theorem 4 (PL^* ⇒ existence of solution + linear convergence)

Theorem 4 ($PL^* \Rightarrow$ existence of solution + linear convergence)
Consider the neural net and random initialization W_0 as described in Theorem 3. For large enough m, then with an appropriate step size η,

Theorem 4 ($PL^* \Rightarrow$ existence of solution + linear convergence)

Consider the neural net and random initialization W_0 as described in Theorem 3. For large enough m, then with an appropriate step size η,

① GD conveges to a global minimizer in $B(w_0, R)$ with $R = O(\frac{1}{\mu})$

w^* s.t. $F(w^*) = y$

Theorem 4 ($PL^* \Rightarrow$ existence of solution + linear convergence)

Consider the neural net and random initialization W_0 as described in Theorem 3. For large enough m, then with an appropriate step size η,

① GD conveges to a global minimizer in $B(w_0, R)$ with $R = O(\frac{1}{\mu})$

w^* s.t. $F(w^*) = y$

② with exponential rate: $\boxed{L(w_t) \leq (1-\eta\mu)^t L(w_0)}$

Theorem 4 ($PL^* \Rightarrow$ existence of solution + linear convergence)

Consider the neural net and random initialization w_0 as described in Theorem 3. For large enough m, then with an appropriate step size η,

① GD conveges to a global minimizer in $B(w_0, R)$ with $R = O(\frac{1}{\mu})$

w^* s.t. $F(w^*) = y$

② with exponential rate: $\boxed{L(w_t) \leq (1-\eta\mu)^t L(w_0)}$

Remarks:

- Note $R = O(\frac{1}{\mu})$

Theorem 4 ($PL^* \Rightarrow$ existence of solution + linear convergence)
Consider the neural net and random initialization W_0 as described in Theorem 3. For large enough m, then with an appropriate step size η,

① GD converges to $\underbrace{\text{a global minimizer}}_{w^* \text{ s.t. } F(w^*) = y}$ in $B(w_0, R)$ with $R = O(\frac{1}{\mu})$

② with exponential rate: $\boxed{L(w_t) \leq (1 - \eta\mu)^t L(w_0)}$

Remarks:

- Note $R = O(\frac{1}{\mu})$
- Condition number $\kappa := \frac{1}{\eta\mu} = \sup_{w \in B(w_0, R)} \frac{\lambda_{max}(H_L(w))}{\mu}$

Theorem 4 (PL* ⇒ existence of solution + linear convergence)

Consider the neural net and random initialization w_0 as described in Theorem 3. For large enough m, then with an appropriate step size η,

① GD conveges to a global minimizer in $B(w_0, R)$ with $R = O(\frac{1}{\mu})$

w^* s.t. $F(w^*) = y$

② with exponential rate: $\boxed{L(w_t) \leq (1-\eta\mu)^t L(w_0)}$

Remarks:

- Note $R = O(\frac{1}{\mu})$
- Condition number $\kappa := \frac{1}{\eta\mu} = \sup_{w \in B(w_0,R)} \frac{\lambda_{\max}(H_L(w))}{\mu}$
- In the special case of a linear system $F(w) = Aw$ with $L(w) = \frac{1}{2}\|Aw - y\|^2$, then $\kappa = \frac{\lambda_{\max}(AA^T)}{\lambda_{\min}(AA^T)}$

19/23

Aside:

19|23

Aside:

This theory covers:

- Wide NN with linear output layer (ie. $\sigma_{L+1}(z) = z$)
- CNN, resnet

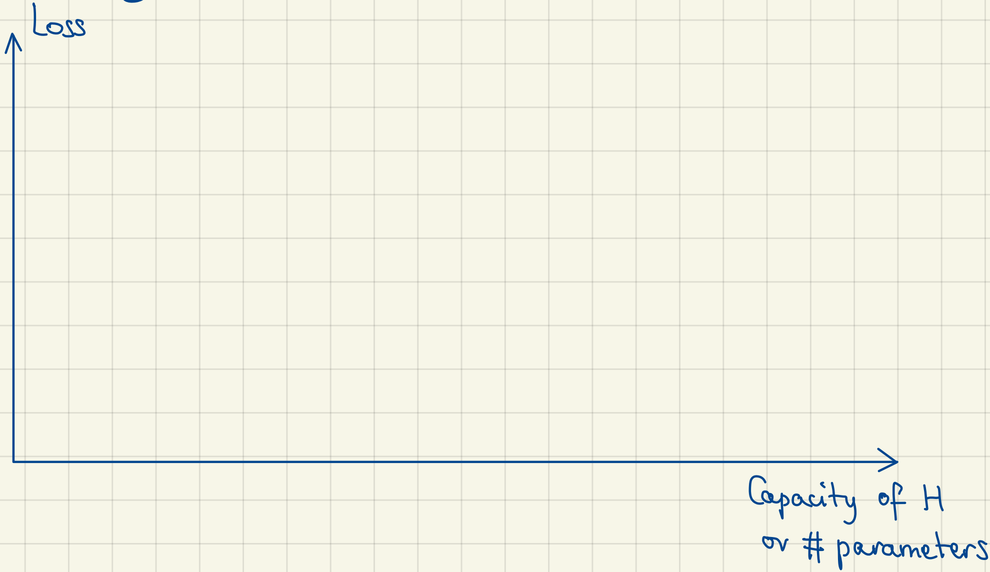
Doesnt cover

- NN with non-linear output layer
- Bottleneck layers.

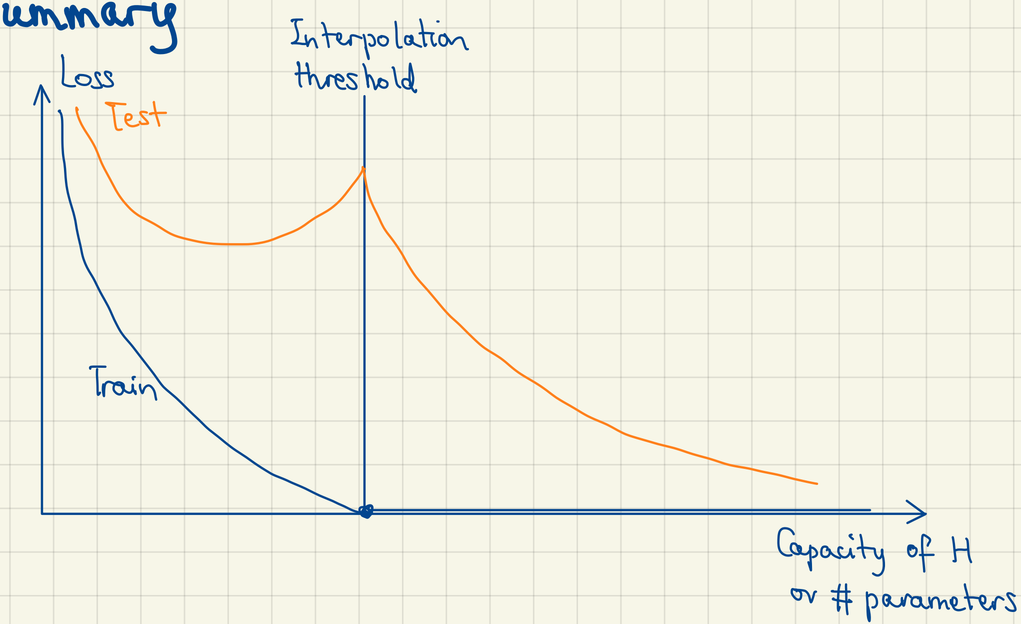
④ Summary and extensions

Summary

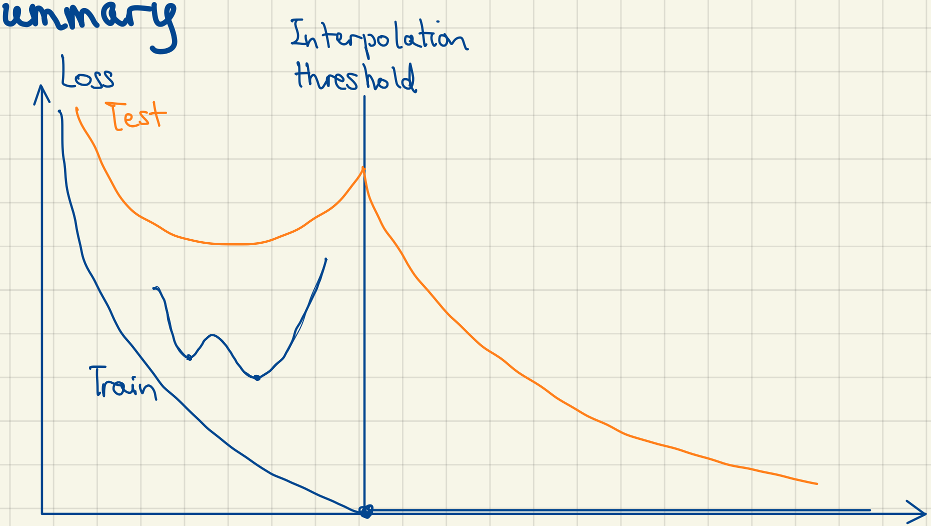
Summary



Summary


Loss
Test
Interpolation threshold
Train
Capacity of H or # parameters

Summary

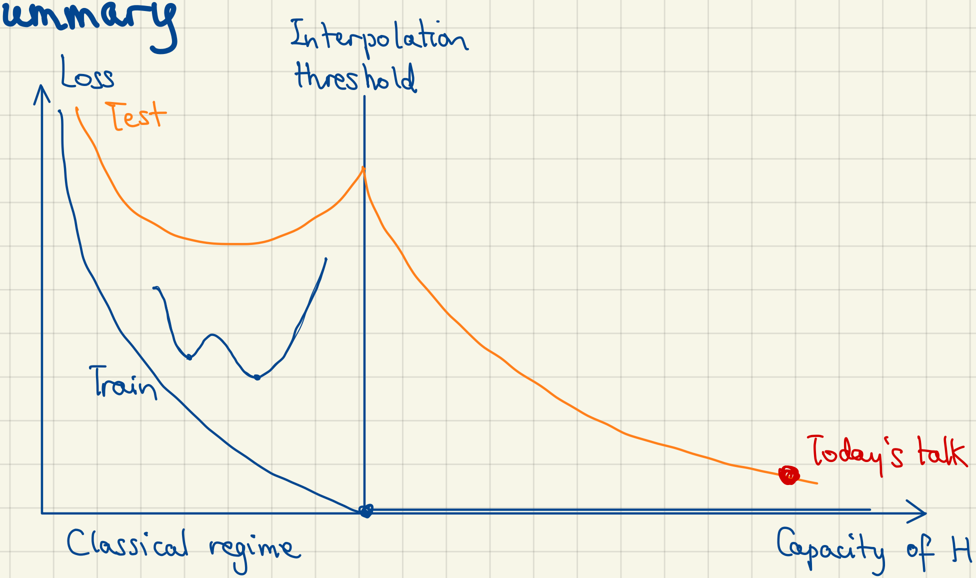


Classical regime

Capacity of H or # parameters

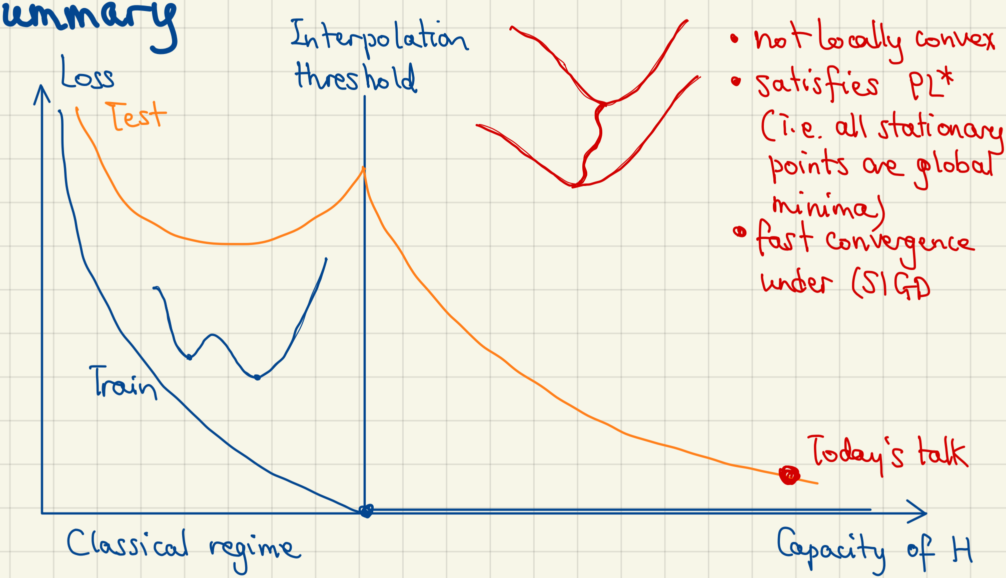
- capacity control required
- many non-global minima
- local convexity, GD oscillates

Summary



- capacity control required
- many non-global minima
- local convexity, GD oscillates

Summary



- not locally convex
- satisfies PL* (i.e. all stationary points are global minima)
- fast convergence under (S)GD

Classical regime

- capacity control required
- many non-global minima
- local convexity, GD oscillates

Limitations

Limitations

- m unrealistically lage. $m = \tilde{\Omega}(nR^{6L+2})$

Limitations

- m unrealistically lage. $m = \tilde{\Omega}(n R^{6L+2})$
- still in the NTK regime $\nabla F(w_0) \approx \nabla F(w_t)$

  this ignores the network's ability to learn features.

Limitations

- m unrealistically lage. $m = \tilde{\Omega}(n R^{6L+2})$
- still in the NTK regime $\nabla F(w_0) \approx \nabla F(w_t)$

  this ignores the network's ability to learn features.

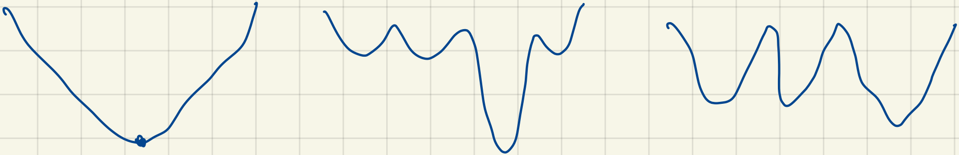
- beyond PL ?

Limitations

- m unrealistically lage. $m = \tilde{\Omega}(nR^{6L+2})$
- still in the NTK regime $\nabla F(w_0) \approx \nabla F(w_t)$

  this ignores the network's ability to learn features.

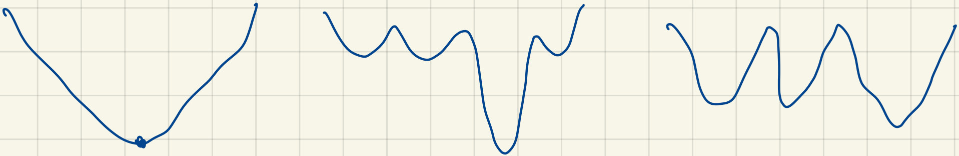
- beyond PL? to incorporate (good) non-global minima?



Limitations

- m unrealistically lage. $m = \tilde{\Omega}(nR^{6L+2})$
- still in the NTK regime $\nabla F(w_0) \approx \nabla F(w_t)$

  this ignores the network's ability to learn features.
- beyond PL? to incorporate (good) non-global minima?

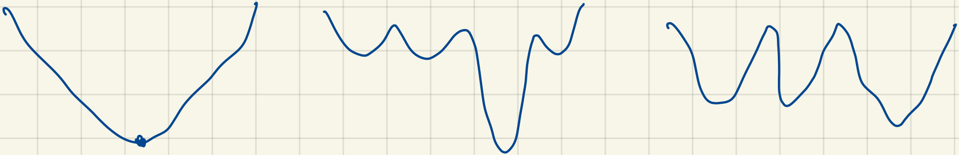


- square Loss $L(w) = \frac{1}{2}\|F(w) - y\|^2$. isn't the most commonly used loss function. What about other loss functions?

Limitations

- m unrealistically lage. $m = \tilde{\Omega}(nR^{6L+2})$
- still in the NTK regime $\nabla F(w_0) \approx \nabla F(w_t)$

  this ignores the network's ability to learn features.
- beyond PL ? to incorporate (good) non-global minima ?



- square Loss $L(w) = \frac{1}{2}\|F(w)-y\|^2$. isn't the most commonly used loss function. What about other loss functions ?
- model $F(w)$ may not be smooth, e.g. ReLU

  [Oymak, Soltanolkotabi 20]

Remedies.

23/23

Remedies. [Frei, Gu 21]

Remedies. [Frei, Gu 21]

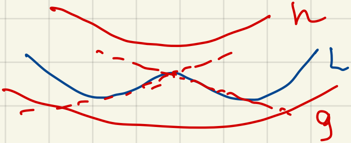
- Proxy convexity

Remedies. [Frei, Gu 21]

- Proxy convexity

$L:\mathbb{R}^m \to \mathbb{R}$ is (g,h)-proxy-PL if $\exists\, g,h:\mathbb{R}^m \to \mathbb{R}$ s.t. $\forall\, w, v,$

$$h(v) \geq g(w) + \langle \nabla L(w), v-w \rangle$$

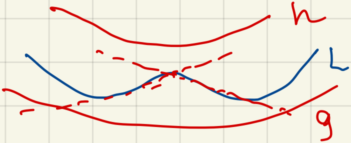


Remedies. [Frei, Gu 21]

- Proxy convexity

  $L : \mathbb{R}^m \to \mathbb{R}$ is (g,h)-proxy-PL if $\exists\, g, h : \mathbb{R}^m \to \mathbb{R}$ s.t. $\forall\, w, v$,

  $$h(v) \geq g(w) + \langle \nabla L(w), v - w \rangle$$



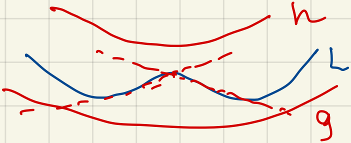
- Proxy PL

Remedies. [Frei, Gu 21]

- Proxy convexity

  $L : \mathbb{R}^m \to \mathbb{R}$ is (g,h)-proxy-PL if $\exists\, g, h : \mathbb{R}^m \to \mathbb{R}$ s.t. $\forall\, w, v,$

  $$h(v) \geq g(w) + \langle \nabla L(w), v - w \rangle$$



- Proxy PL

  $L : \mathbb{R}^m \to \mathbb{R}$ satisfies (h, ξ)-proxy PL inequality if $\exists\, h : \mathbb{R}^m \to \mathbb{R}$ and $\xi \geq 0$, $\alpha, \mu > 0$ s.t. $\forall\, w$

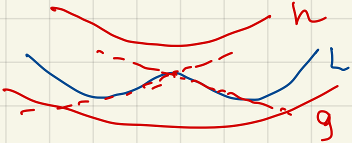
  $$\|\nabla L(w)\|^{\alpha} \geq \mu\,(h(w) - \xi)$$

Remedies. [Frei, Gu 21]

- Proxy convexity

$L:\mathbb{R}^m\to\mathbb{R}$ is (g,h)-proxy-PL if $\exists\, g,h:\mathbb{R}^m\to\mathbb{R}$ s.t. $\forall\, w,v$,

$$h(v)\geq g(w)+\langle\nabla L(w), v-w\rangle$$

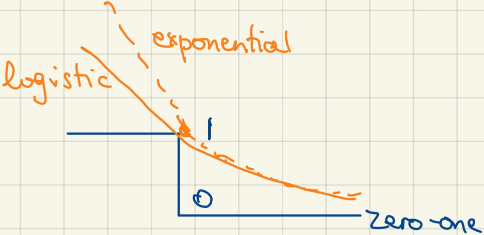


- Proxy PL

$L:\mathbb{R}^m\to\mathbb{R}$ satisfies (h,ξ)-proxy PL inequality if $\exists\, h:\mathbb{R}^m\to\mathbb{R}$ and $\xi\geq 0$, $\alpha,\mu>0$ s.t. $\forall\, w$

$$\|\nabla L(w)\|^{\alpha}\geq\mu\left(h(w)-\xi\right)$$

- think of $h(w)\geq L(w)$

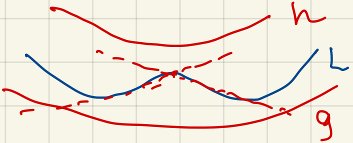


Remedies. [Frei, Gu 21]

- Proxy convexity

  $L:\mathbb{R}^m\to\mathbb{R}$ is (g,h)-proxy-PL if $\exists\, g,h:\mathbb{R}^m\to\mathbb{R}$ s.t. $\forall\, w,v,$

  $$h(v)\geq g(w)+\langle \nabla L(w), v-w\rangle$$

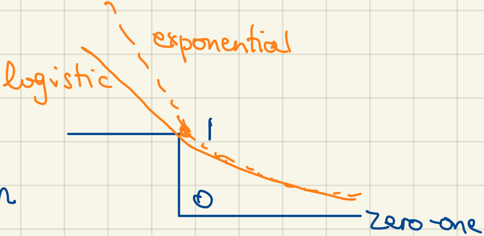


- Proxy PL

  $L:\mathbb{R}^m\to\mathbb{R}$ satisfies (h,ξ)-proxy PL inequality if $\exists\, h:\mathbb{R}^m\to\mathbb{R}$ and $\xi\geq 0$, $\alpha,\mu>0$ s.t. $\forall\, w$

  $$\|\nabla L(w)\|^\alpha \geq \mu\,(h(w)-\xi)$$

  - think of $h(w)\geq L(w)$
  - $\xi\geq 0$ allows good local minimum



Thanks.

Non-linear vs linear output layer

$$g(w;x) = \overset{\text{last layer activation}}{\phi}(f(w;x))$$

$$(\phi(f))'' = \underbrace{\phi''(f)}_{\begin{cases} =0 & \text{For linear output layer} \\ \neq 0 & \text{For non-linear output layer.} \end{cases}} \underbrace{f'}_{O(1)} + \underbrace{\phi'(f)}_{O(1)} \underbrace{f''}_{O\left(\frac{1}{\sqrt{m}}\right)}$$

Classical statistical learning theory

Expected risk $$R(f) = \mathbb{E}_{p(x,y)}\, \ell(f(x), y)$$

Bayes-optimal $$f^* = \operatorname*{argmin}_{f:\mathbb{R}^d\to\mathbb{R}} R(f)$$

Empirical risk minimization
$$R_{emp}(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(x_i), y_i)$$
$$f_{emp} = \operatorname*{argmin}_{f\in H} R_{emp}(f)$$

⇒ w.h.p. over the randomness in data, for any $f\in H$,

$$\underbrace{R(f)}_{\approx \text{Test}} - \underbrace{R_{emp}(f)}_{\text{Train}} < \tilde{O}\left(\sqrt{\frac{cap(H)}{n}}\right)$$

— e.g. VC dim, covering numbe